音声対話に基づく情報検索システムの研究開発と評価

(課題番号：12680406)

平成12年度～平成15年度科学研究費補助金　基盤研究(C)(2)研究成果報告書

平成16年3月

研究代表者　石川　徹也
(筑波大学　図書館情報学系　教授)

音声対話に基づく情報検索システムの研究開発と評価

平成12年度～平成15年度科学研究費補助金 基盤研究(C)(2) 研究成果報告書

はしがき

近年，従来のキーボード入力・画面出力を前提としない新しい形態のコンピュータ情報システムに対する需要が高まっている．これらの背景には，携帯端末の小型化，カーナビゲーションシステムのように他の作業を同時に行わなければならない環境での利用，非晴眼者のような視覚情報を利用できないユーザへの補助などが挙げられる．

Webに代表されるオンライン情報の普及によって，自宅やオフィスにいながらにして，世界中の情報にアクセスが可能である．しかし他方において，キーボードリテラシーのない人や非晴眼者は情報弱者となりつつある．そこで，コンピュータ利用におけるキーボード入力・画面表示というバリアーを克服する方式の提案は社会的意義が大きい．

本研究は，上記問題に対する解決策の1つとして，音声入出力を用いた情報検索システムの研究開発を行い，実験によってその有効性を評価した．

従来の音声認識・対話に関する研究は，主に認識用辞書の規模の問題のために，比較的限られた分野に適用されることが多かった．それに対して，テキスト情報検索の研究では，Webのような異種分野文書群を対象に，一般的文書から専門的文書まで幅広い分野を扱う研究が行われている．本研究は，音声研究と情報検索における語彙の規模に関する齟齬を埋める手法を開発した．

Web上の外国語文書へのアクセスは，コンピュータ利用者にとっては今や日常的になっている．しかし，外国語文書を検索するための有効なキーワードを特定することは一般ユーザにとって必ずしも容易な作業ではないため，情報を十分に活用できていないのが現状である．本研究では，多言語検索・言語横断検索の基礎技術を導入することで，上記問題の解消を行った．

わが国では，音声処理・情報検索・自然言語処理などの分野が「言葉」という共通の対象を扱っているにも関わらず，研究者相互の協力体制が諸外国に比べると希薄である．そこで，個々の分野での成果をどのように統合すればよいのか，そもそも統合する際にどのような問題があるのかさえも研究者ベースでは分かりにくくなっている．本研究を通して，音声処理と情報検索を統合する上での研究課題を特定し，以降の広い意味での言語処理研究に貢献できる．

研究組織

研究代表者: 石川 徹也 （筑波大学 図書館情報学系 教授）
研究分担者: 藤井 敦 （筑波大学 図書館情報学系 助教授）
研究分担者: 伊藤 克亘 （名古屋大学 大学院情報科学研究科 助教授）
研究分担者: 秋葉 友良 （独立行政法人 産業技術総合研究所 主任研究員）

交付金額

平成12年度	800千円
平成13年度	700千円
平成14年度	1,200千円
平成15年度	800千円
総計	3,500千円

研究発表

(1) 学会誌等

1. 秋葉 友良, 伊藤 克亘, 藤井 敦. N-gram の部分的強調による定型表現への言語モデル適応手法. 電子情報通信学会論文誌 D-II, J86-D-II, No.12, pp.1727–1736, 2003.
2. 関 和広, 藤井 敦, 石川 徹也. 確率モデルを用いた日本語ゼロ代名詞の照応解析. 自然言語処理, Vol.9, No.3, pp.63–85, 2002.
3. Atsushi Fujii and Tetsuya Ishikawa. Japanese/English Cross-Language Information Retrieval: Exploration of Query Translation and Transliteration. Computers and the Humanities, Vol.35, No.4, pp.389–420, 2001.
4. 藤井 敦, 石川 徹也. 質問翻訳と文書翻訳を統合した日英言語横断情報検索. 電子情報通信学会論文誌 D-II, Vol.J84-D-II, No.2, pp.362–369, 2001.

(2) 口頭発表

1. Atsushi Fujii, Katunobu Itou, Tomoyosi Akiba, and Tetsuya Ishikawa. A Cross-media Retrieval System for Lecture Videos. Proceedings of the 8th European Conference on Speech Communication and Technology, pp.1149–1152, 2003.
2. Atsushi Fujii and Katunobu Itou. Building a Test Collection for Speech-Driven Web Retrieval. Proceedings of the 8th European Conference on Speech Communication and Technology, pp.1153–1156, 2003.
3. Tomoyosi Akiba, Katunobu Itou, and Atsushi Fujii. Adapting Language Models for Frequent Fixed Phrases by Emphasizing N-gram Subsets. Proceedings of the 8th European Conference on Speech Communication and Technology, pp.1469–1472, 2003.
4. Atsushi Fujii, Katunobu Itou, and Tetsuya Ishikawa. LODEM: A Multilingual Lecture-on-demand System. Proceedings of the 2003 ISCA Workshop on Multilingual Spoken Document Retrieval, pp.13–18, 2003.
5. Atsushi Fujii and Katunobu Itou. Evaluating Speech-Driven Web Retrieval in the Third NTCIR Workshop. AAAI-03 Spring Symposium on Intelligent Multimedia Knowledge Management, 2003.
6. Atsushi Fujii, Katunobu Itou, and Tetsuya Ishikawa. A System for On-demand Video Lectures. AAAI-03 Spring Symposium on Intelligent Multimedia Knowledge Management, 2003.
7. Tomoyosi Akiba, Katunobu Itou, Atsushi Fujii, and Tetsuya Ishikawa. Selective Back-off Smoothing for Incorporating Grammatical Constraints into the N-gram language model. Proceedings of the 7th International Conference on Spoken Language Processing, pp.881–884, 2002.
8. Kazuhiro Seki, Atsushi Fujii, and Tetsuya Ishikawa. A Probabilistic Method for Analyzing Japanese Anaphora Integrating Zero Pronoun Detection and Resolution. Proceedings of the 19th International Conference on Computational Linguistics, pp.911–917, 2002.

9. Atsushi Fujii, Katunobu Itou, and Tetsuya Ishikawa. A Method for Open-Vocabulary Speech-Driven Text Retrieval. Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, pp.188–195, 2002.

10. Katunobu Itou, Atsushi Fujii, and Tetsuya Ishikawa. Language Modeling for Multi-Domain Speech-Driven Text Retrieval. IEEE Automatic Speech Recognition and Understanding Workshop, 2001.

11. Kazuhiro Seki, Atsushi Fujii, and Tetsuya Ishikawa. A Probabilistic Model for Japanese Zero Pronoun Resolution Integrating Syntactic and Semantic Features. Proceedings of the 6th Natural Language Processing Pacific Rim Symposium, pp.403–410, 2001.

12. Atsushi Fujii, Katunobu Itou, and Tetsuya Ishikawa. Speech-Driven Text Retrieval: Using Target IR Collections for Statistical Language Model Adaptation in Speech Recognition. ACM SIGIR'01 Workshop on Information Retrieval Techniques for Speech Applications, 2001.

13. Atsushi Fujii and Tetsuya Ishikawa. Applying Machine Translation to Two-Stage Cross-Language Information Retrieval. Proceedings of the 4th Conference of the Association for Machine Translation in the Americas, pp.13–24, 2000.

14. Atsushi Fujii and Tetsuya Ishikawa. A Novelty-based Evaluation Method for Information Retrieval. Proceedings of the 2nd International Conference on Language Resources and Evaluation, pp.1637–1641, 2000.

15. 坂本 尚実, 藤井 敦, 石川 徹也, 秋葉 友良, 伊藤 克亘. 講演音声の認識における言語モデルのタスク適応. 日本音響学会講演論文集, pp.205–206, 2003.

16. 秋葉 友良, 伊藤 克亘, 藤井 敦. 音声入力型情報検索のための自由発話収録. 日本音響学会講演論文集, pp.59–60, 2003.

17. 藤井 敦. 音声による言語バリアフリーな多言語情報アクセス. 電子情報通信学会技術研究報告, SP2002-153, pp.31–36, 2002.

18. 伊藤 克亘, 藤井 敦. NTCIR-3 ワークショップにおける音声入力型ウェブ検索タスク. 情報処理学会研究報告, 2002-SLP-43, pp.25–32, 2002.

19. 伊藤 克亘, 秋葉 友良, 藤井 敦. WWW は大語彙連続音声認識の学習データとして使えるか? 日本音響学会講演論文集, pp.131–132, 2002.

20. 秋葉 友良, 伊藤 克亘, 藤井 敦. 単語連鎖を重視した N-gram 言語モデル平滑化手法の検討. 日本音響学会講演論文集, pp.135–136, 2002.

21. 秋葉 友良, 伊藤 克亘, 藤井 敦, 石川 徹也. 部分 N-gram 頻度情報を利用した質問応答定型表現への言語モデル適応. 情報処理学会研究報告, 2002-SLP-42, pp.31–38, 2002.

22. 藤井 敦, 伊藤 克亘, 石川 徹也. 音声文書検索の応用によるオンデマンド講演システム. 言語処理学会第 8 回年次大会発表論文集, pp.192–195, 2002.

23. 秋葉 友良, 伊藤 克亘, 藤井 敦, 石川 徹也. 音声入力による質問応答システムのための音声認識用言語モデルの検討. 言語処理学会第 8 回年次大会発表論文集, pp.244–247, 2002.

24. 藤井 敦, 伊藤 克亘, 石川 徹也. ユーザ発話中の未知語を自動補完する音声入力型検索システム. 言語処理学会第 8 回年次大会発表論文集, pp.487–490, 2002.

25. 関 和広, 藤井 敦 石川 徹也. ゼロ代名詞の検出と補完を統合した確率的照応解消モデル. 言語処理学会第 8 回年次大会発表論文集, pp.591–594, 2002.

26. 伊藤 克亘, 藤井 敦, 石川 徹也. 未知語検出を用いた語彙統制のない音声検索. 日本音響学会講演論文集, pp.103–104, 2002.

27. 伊藤 克亘, 藤井 敦, 石川 徹也. 音声文書検索を用いたオンデマンド講義システム. 情報処理学会研究報告, 2001-SLP-39, pp.165–170, 2001.

28. 伊藤 克亘, 秋葉 友良, 藤井 敦, 石川 徹也. 音声入力型テキスト検索システムのための音声認識. 日本音響学会講演論文集, pp.193–194, 2001.

29. 藤井 敦, 伊藤 克亘, 秋葉 友良, 石川 徹也. 音声入力型文書検索システムの開発とテストコレクションの構築. 情報処理学会研究報告 2001-FI-63, pp.65–72, 2001.

30. 関 和広, 藤井 敦, 石川 徹也. 確率モデルに基づく日本語ゼロ代名詞の照応解消. 言語処理学会第 7 回年次大会発表論文集, pp.510–513, 2001.

31. 藤井 敦, 伊藤 克亘, 秋葉 友良, 石川 徹也. 音声言語データの構造化に基づく講演発表の自動要約. 話し言葉の科学と工学ワークショップ講演予稿集, pp.173–177, 2001.

(3) 出版物

1. Atsushi Fujii, Katunobu Itou, and Tetsuya Ishikawa. Speech-Driven Text Retrieval: Using Target IR Collections for Statistical Language Model Adaptation in Speech Recognition. Anni R. Coden and Eric W. Brown and Savitha Srinivasan (Eds.), Information Retrieval Techniques for Speech Applications (LNCS 2273), pp.94–104, Springer, 2002.

特許出願

1. 音声入力によるテキスト検索方法およびその装置. 特開 2003-271629.

2. 音声データ検索システム. 特開 2003-067388.

3. 音声入力検索システム. 特開 2003-036093.

研究成果

本研究課題では，音声発話による情報検索システムの研究開発を評価を行った．開発した検索システムを図1に示す．システムへの入力は，ユーザが音声で発話した検索質問である．音声発話には以下に示すような区分がある．

- 発話形式による区分
  - 読み上げ音声（read speech）
    あらかじめ用意された質問内容を読み上げるため，発話が整っており高い精度で音声認識が可能である．
  - 自由発話（spontaneous speech）
    質問を考えながら発話するため，言い間違いや言い直しが多く音声認識精度が低下する．また，質問内容の言い忘れがあると，検索精度が低下する．
- 質問形式による区分
  - 検索要求
    検索エンジンに入力するような短いキーワードや複合語などのフレーズなどで表現された検索要求である．
  - 質問文
    「カナダの首都はどこですか?」というような WH（5W1H）質問文である．既存の検索エンジンには,「カナダ」や「首都」などのキーワードを入力することしかできないのに対して,「どこ」という場所に関する要素を質問に導入することができる．
  - 対話型質問
    ある質問に対するシステムの回答を見て，副次的に様々なことを連続して質問したくなる場合を想定した質問である．カナダの首都が分かったら，観光名所や歴史についても続けて質問するような場合である．そこで，文脈の解析が必要になる．

本システムでは,「読み上げ音声」については全ての質問形式に対応している．しかし,「自由発話」については，自然言語理解，音声言語理解の側面で未解決の問題が多いため,「検索要求」のみに対処した．

音声認識は，音響モデルと言語モデルを利用して，ユーザの発話をテキストに転記する．言語モデルは，検索対象のテキストデータから抽出した語や語順に関する統計情報である．そこで，当テキストデータ中の情報を検索することを目的とした発話を高い精度で音声認識することができる．また，未知語モデルによって，音声認識辞書にない語の認識を可能する．さらに，質問文に使用される定型表現を記述した文法規則を併用することで，質問文に対する音声認識の精度を向上させる．

外国語情報を検索する場合は，認識された内容を対象の言語に機械翻訳する．

情報検索によって，ユーザの要求に適合する文書群を検索する．質問応答では，自然言語解析によって検索された文書内容を解析し，特定の回答を抽出する．情報検索では，テキスト情報だけでなく音声情報を対象にすることも可能である．この場合は，音声認識によって音声データを事前にテキストに転記しておく．

本研究課題では，研究開発したシステムの性能を評価するためのテストコレクション（ベンチマークデータ）を構築した．テストコレクションは，検索質問となるユーザ発話の音声データ，検索対象の文書集合，質問ごとの正解判定（適合文書）情報で構成される．文書集合と適合判定の構築は，それだけで研究テーマになるため，対象外とした．その代りに，既存のテキスト検索テストコレクションと連携して使用できるユーザ発話の音声データを収録し，整備した．上述した発話形式や質問形式によって個別の音声データを構築し，多面的な観点からの評価を可能にした．

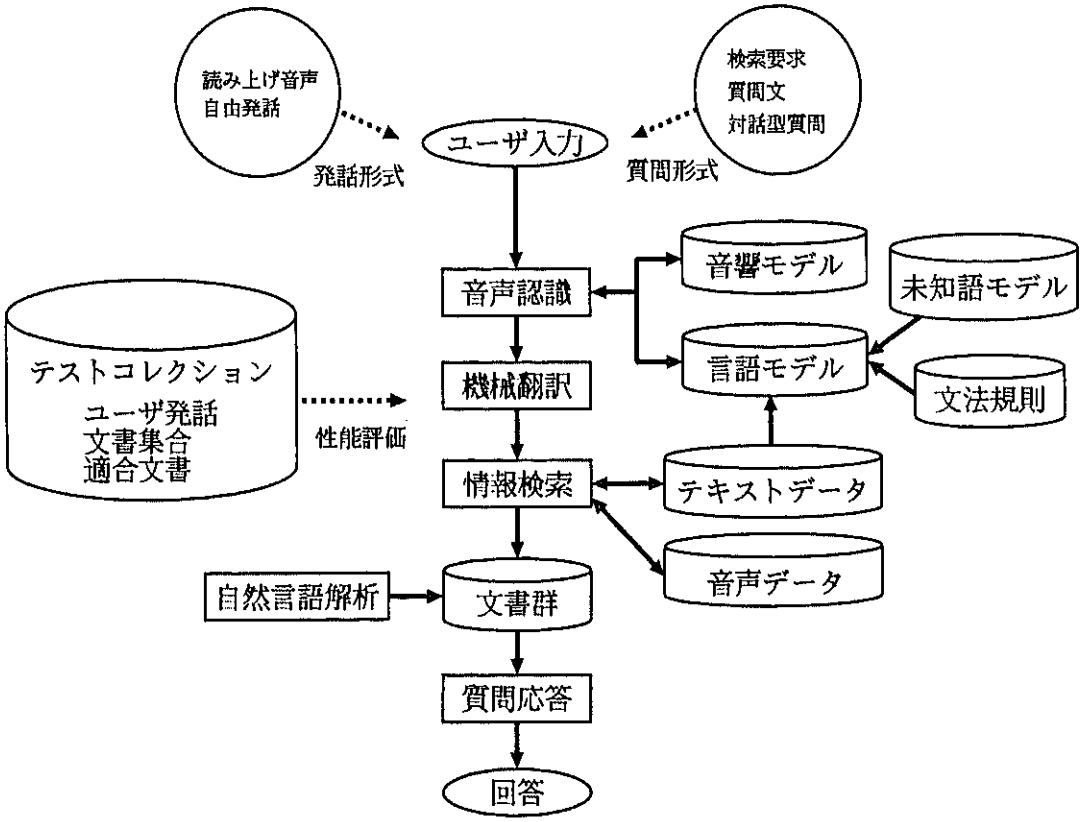

図 1: 研究開発した検索システム

研究成果論文集

目次

論　文

N-gramの部分的強調による定型表現への言語モデル適応手法

秋葉　友良[†]　　伊藤　克亘[††,†††]　　藤井　敦[†††,††††]

A Language Model Adaptation Method for Fixed Phrases by Emphasizing N-gram Subsets

Tomoyosi AKIBA[†], Katunobu ITOU[††,†††], and Atsushi FUJII[†††,††††]

あらまし　音声入力に対応した質問応答システムのためのN-gram言語モデルを作成する手法を提案する．質問応答の入力は，検索トピックに関する前半部分と，質問文に使われる定型的な文末表現（定型表現）から構成されることが多い．新聞記事などの検索対象コーパスから作成した言語モデルでは，前半をうまくモデル化する一方，定型表現を十分にモデル化できない．本論文では，この問題を一般コーパスから作成した言語モデルを定型表現に適応する問題として扱い，収集が困難な適応コーパスを用いず，人手で作成した定型表現のリストだけを利用する言語モデル適応手法を提案する．提案手法は，定型表現リストを用いて，一般コーパスから得たN-gramの定型表現に該当する箇所を特定し，そのN-gramを強調することでモデルの適応を行う手法であり，一般コーパス自身の部分的なN-gramを事後知識とみなした最大事後確率推（MAP推定）を行うことに相当する．質問応答システムの入力となる質問文音声の認識実験を行い，本手法の有効性を確認した．

キーワード　大語彙連続音声認識，N-gram言語モデル，タスク適応，質問応答，定型表現

1. まえがき

近年の大語彙音声認識技術の発展に伴い，様々なアプリケーションに音声入力手段を利用する可能性が広がっている．中でも，音声入力による情報検索[3],[6]は，キーボードリテラシーがないユーザ，非晴眼者，携帯端末ユーザなどを支援する手段として有効である．

一方，情報検索の分野では，より高度な情報アクセス手段として，質問応答(Question Answering: QA)技術への関心が高まっている．キーワードを入力し関連文書を出力する従来の情報検索と異なり，質問応答では自然言語による質問文を入力し，文書より細かな単位の回答を出力する．質問応答は，1999年のTREC-8[16]で再利用可能な大規模テストコレクション（ベンチマークデータ）が構築されたことを主たる背景とし，我が国でも2001年からNTCIR[7]で大規模な評価が始まるなど，近年盛んに研究されている．

質問応答の入力は話し言葉に近い質問文であることから，キーワード入力が中心の情報検索よりも音声入力に適したタスクである．質問応答を指向した音声研究として音声対話[1]がある．音声対話の研究がドメインを限定し高度に組織化された小規模のデータベース（例えば，リレーショナルデータベース）を対象としてきたのに対し，質問応答はドメインを限定しない大規模な（組織化されていない）文書を対象にするため，応用範囲の広い技術である．このような背景の中，我々は音声入力による質問応答システムを開発中である．

TRECやNTCIRで評価に使用される質問文は，いわゆる5W1Hのうち3W (Who, When, Where)に関する質問が中心である．そこで，質問文は検索トピックと質問文に使われる典型的な表現（定型表現）の二つで構成されることが多い．例えば，次のような質問文[19]を考える．

† 産業技術総合研究所，つくば市
National Institute of Advanced Industrial Science and Technology, TsukubaCentral 2, 1-1-1 Umezono, Tsukuba-shi, 305-8568 Japan
†† 名古屋大学大学院情報科学研究科，名古屋市
Graduate School of Information Science, Nagoya University, Furo-cho, Chigusa-ku, Nagoya-shi, 464-8601 Japan
††† 科学技術振興事業団 CREST，川口市
CREST JST (Japan Science and Technology Corporation), 4-1-8 Honmachi, Kawaguchi-shi, 332-0012 Japan
†††† 筑波大学図書館情報学系，つくば市
Institute of Library and Information Science, University of Tsukuba, 1-2 Kasuga, Tsukuba-shi, 305-8550 Japan

「1976 年に火星に軟着陸した探査機は何という名前でしたか」

前半の「1976 年に火星に軟着陸した探査機は」の部分は検索トピックであり，回答を探すコーパス（新聞記事や辞典など）から作成した言語モデルでモデル化できる．後半の「何という名前でしたか」の部分は，質問文に典型的に現れる文末表現であり，新聞記事や辞典では比較的稀な表現であるため，モデル化が困難である．しかし，何を尋ねているのかを特定する重要な情報を運ぶため，質問応答において重要な役割を担っている．そのため，音声入力による質問応答システムでは，検索トピックとともに定型表現を適切にモデル化することが必要となる．

大量のコーパスが入手困難な特定のタスクに対して，入手が容易な大量のタスク非依存のコーパス（以下，一般コーパス）と，少量の特定タスク用コーパス（以下，適応コーパス）から，特定タスクの言語モデルを推定する，言語モデルのタスク適応の手法が提案されている [4]．この方法を用いれば，一般コーパスと，定型表現を含む質問文から構成される適応コーパスを用いて，言語モデルを作成することができる．しかし，この手法の問題は，適応コーパスの収集が必要となる点にある．

一方，質問応答の質問文の特徴を考慮すると，この問題に特化した適切な手法が可能である．質問文定型表現の多様性はそれほど大きくない．そのため，定型表現を人手で列挙したり文法を記述することが可能である．また，質問文の定型表現は一般的な語で構成されるため，高頻度ではないものの一般コーパスにも含まれる．そこで，一般コーパスに含まれる N-gram の確率分布を利用することが可能である．

本論文では，一般コーパスと人手で列挙した定型表現のリストを用いて，質問応答の質問文をモデル化する手法を提案する．提案手法は，定型表現リストを用いて，一般コーパスから獲得した N-gram において定型表現に該当する箇所を特定し，特定した N-gram を強調する．理論的には，一般コーパスを事前知識とし，一般コーパス自身の部分的な N-gram を事後知識とみなして，最大事後確率推定（MAP 推定）によって言語モデルのタスク適応を行うことに相当する．従来の言語モデル適応手法が文を単位とした手法であったのに対し，本手法は文を構成する表現（定型表現）に適応する点が異なる．

以下，**2.** で提案手法とその理論的背景を説明する．**3.** で提案手法を質問文の定型表現に適用した実験によってその有効性を示す．**4.** で言語モデルのタスク適応に関する従来の手法を概説し，本手法の位置付けを明確にする．

2. 定型表現へのタスク適応手法

2.1 概　　要

提案する手法の概要を図 1 に示す．新聞記事から得た N-gram 頻度のうち，定型表現を含む文の確率が相対的に高くなるように該当する部分の N-gram を強調（γ 倍）する．

2.2 定型表現に対応する N-gram の強調

新聞記事から得た N-gram がモデル化（カバー）する文の集合 S のうち，定型表現リストに記述された単語列を含む文だけから構成される部分集合 $S_{FP}(\subset S)$ を考える．また，S 中の文を生成する確率を P とする．このとき定型表現へのタスク適応の目的は，$\hat{s} \in S_{FP}$ なる文 $\hat{s}$ に割り当てる確率 $P'(\hat{s})$ を相対的に高くし，$s \in \overline{S_{FP}}(= S - S_{FP})$ なる文 s に割り当てる確率 $P'(s)$ に対しては $P(s)$ の性質（順序関係）をできるだけ保持した P' を求めることである．以下，$P(\hat{s})$ から $P'(\hat{s})$ への修正に伴う $P(s)$ の性質の損失を，適応の副作用と呼ぶ．

定型表現 $\hat{w}_p \cdots \hat{w}_q$ を含む文

$$\hat{s} = w_1 w_2 \cdots \hat{w}_p \hat{w}_{p+1} \cdots \hat{w}_q w_{q+1} \cdots w_m$$

の生成確率は，次のように近似できる．

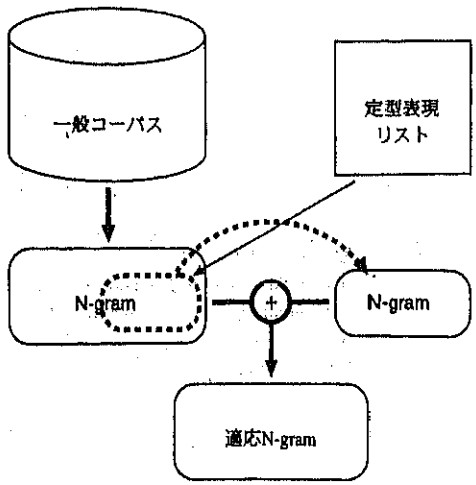

図 1　部分的な N-gram の強調によるタスク適応
Fig. 1　Task adaptation by emphasizing selected N-gram subsets.

$$
\begin{aligned}
P(\hat{s}) &= \prod_{i=1}^{m} P(w_i|w_1 \cdots w_{i-1}) \\
&\approx P(w_1)P(w_2|w_1)\cdots P(w_i|w_{i-N+1}^{i-1})\cdots \\
&\quad \cdots P(\hat{w}_p|w_{p-N+1}^{p-1})P(\hat{w}_{p+1}|w_{p-N+2}^{p-1}\hat{w}_p)\cdots \\
&\quad \cdots P(\hat{w}_{p+N-2}|w_{p-1}\hat{w}_p^{p+N-3})\cdots \\
&\quad \cdots P(\hat{w}_{p+N-1}|\hat{w}_p^{p+N-2})\cdots \\
&\quad \cdots P(\hat{w}_q|\hat{w}_{q-N+1}^{q-1})P(w_{q+1}|\hat{w}_{q-N+2}^{q})\cdots \\
&\quad \cdots P(w_m|w_{m-N+1}^{m-1})
\end{aligned}
$$

$\hat{s}$ の確率を相対的に高くするのと同時に s への副作用を抑えるためには，定型表現内の単語 $\hat{w}_p \cdots \hat{w}_q$ について，その N-gram 確率が相対的に高くなるように N-gram の確率分布を修正すればよい．すなわち，$P(\hat{s})$ を構成する N-gram のうち，次の式 (1) と (2) に該当する N-gram の値を強調する．

$$P(\hat{w}_p|w_{p-N+1}^{p-1})\cdots P(\hat{w}_{p+N-2}|w_{p-1}\hat{w}_p^{p+N-3}) \quad (1)$$

$$P(\hat{w}_{p+N-1}|\hat{w}_p^{p+N-2})\cdots P(\hat{w}_q|\hat{w}_{q-N+1}^{q-1}) \quad (2)$$

式 (1) は，長さ N の N-gram について，定型表現の先頭から N より短い接頭部分の確率であり，条件部の接頭部分には定型表現以外の単語が現れる．式 (2) は，先頭から N 以降に現れる部分の確率であり，条件部は定型表現の単語のみで構成される．

以上の確率を再推定するために，まず最ゆう推定値の修正を行い，次にバックオフスムージングを適用する．

2.3 最ゆう推定値の修正

本論文で扱う質問文の定型表現は，十分に一般的な言い回しであるため，一般コーパスに含まれていると仮定できる(注1)．したがって，定型表現の部分単語列については N-gram が存在するため，それらの単語列に対応する N-gram 確率は，バックオフスムージングによって，より短い N-gram にバックオフされることはない．そこで，定型表現に跨る長さ以上の（接頭部分は接頭単語長以上の，それ以降は最高次 N の）N-gram の修正だけを行えば，定型表現を含む文の確率を高くして，同時に定型表現を含まない文への副作用を抑えることができる．

一般コーパスから得た長さ $n(1 \le n \le N)$ の N-gram の最ゆう推定値 $P_{ML(n)}(w_i|w_{i-n+1}^{i-1})$ に対し，以下に示す順序で条件に合致するものを $\gamma(\geq 1)$ 倍して修正，正規化し，$P'_{ML(n)}(w_i|w_{i-n+1}^{i-1})$ とする．

（1） 定型表現の接頭単語列に対しては，接頭単語列長以上の長さの N-gram を強調する．接頭単語長を $k(1 \le k < N \wedge k \le n)$ とすると，すべての文脈単語列 w_{p-n+k}^{p-1} について，以下の式で最ゆう推定値を強調する．

$$
\begin{aligned}
&P'_{ML(n)}(\hat{w}_{p+k-1}|w_{p-n+k}^{p-1}\hat{w}_p^{p+k-2}) \\
&\quad = \beta_n(w_{p-n+k}^{p-1}\hat{w}_p^{p+k-2}) \cdot \\
&\quad \gamma P_{ML(n)}(\hat{w}_{p+k-1}|w_{p-n+k}^{p-1}\hat{w}_p^{p+k-2})
\end{aligned}
$$

tri-gram モデルの場合，すべての文脈単語列 $w_{p-2}w_{p-1}$ について，次のように各長さの最ゆう推定値を強調する．

$$
\begin{aligned}
&P'_{ML(3)}(\hat{w}_{p+1}|w_{p-1}\hat{w}_p) \\
&\quad = \beta_3(w_{p-1}\hat{w}_p)\cdot\gamma P_{ML(3)}(\hat{w}_{p+1}|w_{p-1}\hat{w}_p) \\
&P'_{ML(2)}(\hat{w}_{p+1}|\hat{w}_p) \\
&\quad = \beta_2(\hat{w}_p)\cdot\gamma P_{ML(2)}(\hat{w}_{p+1}|\hat{w}_p) \\
&P'_{ML(3)}(\hat{w}_p|w_{p-2}w_{p-1}) \\
&\quad = \beta_3(w_{p-2}w_{p-1})\cdot\gamma P_{ML(3)}(\hat{w}_p|w_{p-2}w_{p-1}) \\
&P'_{ML(2)}(\hat{w}_p|w_{p-1}) \\
&\quad = \beta_2(w_{p-1})\cdot\gamma P_{ML(2)}(\hat{w}_p|w_{p-1}) \\
&P'_{ML(1)}(\hat{w}_p) = \beta_1(\epsilon)\cdot\gamma P_{ML(1)}(\hat{w}_p)
\end{aligned}
$$

（2） 定型表現の接頭部分以外にのみ現れる単語列に対しては，最も長い N-gram だけを強調する．

$$
\begin{aligned}
&P'_{ML(N)}(\hat{w}_i|\hat{w}_{i-N+1}^{i-1}) \\
&\quad = \beta_N(\hat{w}_{i-N+1}^{i-1})\cdot\gamma P_{ML(N)}(\hat{w}_i|\hat{w}_{i-N+1}^{i-1})
\end{aligned}
$$

例えば tri-gram モデルを構築する場合，tri-gram だけを γ 倍する．

$$
\begin{aligned}
&P'_{ML(3)}(\hat{w}_i|\hat{w}_{i-2}\hat{w}_{i-1}) \\
&\quad = \beta_3(\hat{w}_{i-2}\hat{w}_{i-1})\cdot\gamma P_{ML(3)}(\hat{w}_i|\hat{w}_{i-2}\hat{w}_{i-1})
\end{aligned}
$$

（3） それ以外の N-gram は，もとの N-gram と同じとする．各 $n(1 \le n \le N)$ について，

$$
\begin{aligned}
&P'_{ML(n)}(w_i|w_{i-n+1}^{i-1}) \\
&\quad = \beta_n(w_{i-n+1}^{i-1})\cdot P_{ML(n)}(w_i|w_{i-n+1}^{i-1})
\end{aligned}
$$

(注1)：コーパスに含まれていることが仮定できない定型表現に適用する場合，定型表現を含む模擬の文を一般コーパスに混合する，定型表現に対応する N-gram を直接フロアリングする，などの方法で仮定を成立させることができる．

ここで，$\beta_n(w_{i-n+1}^{i-1})(0<n)$ は，同じ条件部 w_{i-n+1}^{i-1} をもつ確率 $P(w|w_{i-n+1}^{i-1})$ の総和を 1 とするための正規化係数である．

以上の方法で得られる修正後の最ゆう推定値 $P'_{ML(n)}$ は，定型表現に関する N-gram 頻度（図 2）を γ 倍して得られる長さごとの N-gram 頻度 $\mathbf{C}_n$ から，次の式で N-gram を最ゆう推定する場合と等価である．

$$P'_{ML(n)}(w_i|w_{i-n+1}^{i-1}) = \frac{\mathbf{C}_n(w_{i-n+1}^{i})}{\sum_{w_i}\mathbf{C}_n(w_{i-n+1}^{i})}$$

よって本手法は，一般コーパスから得た長さごとの最ゆう推定モデルを，図 2 に示した該当する自分自身の部分的な N-gram を事後知識として重み $(\gamma-1)$ で N-gram 混合を行うこと（MAP 推定による言語モデル適応 [17]）に相当する．

2.4 バックオフスムージング

長さごとに得られた修正後の最ゆう推定値 $P'_{ML(n)}$ を統合して，バックオフ N-gram を作成する．バックオフスムージングの一般式は，次のように表される．

$$P(w_i|w_{i-n+1}^{i-1}) = \begin{cases} d_{w_{i-n+1}^{i}} P_{ML}(w_i|w_{i-n+1}^{i-1}) & C(w_{i-n+1}^{i})>0 \\ \alpha(w_{i-n+1}^{i-1})P(w_i|w_{i-n+2}^{i-1}) & C(w_{i-n+1}^{i})=0 \end{cases}$$

ここで $d_{w_{i-n+1}^{i}}$，P_{ML}，α は，それぞれ，ディスカウント係数，最ゆう推定による N-gram 確率，確率の総和を 1 とするための正規化係数である．

このうち，P_{ML} はそのまま $P'_{ML(n)}$ で置き換え可能であり，α は他の値から自動的に求まる．一方，ディスカウント係数 d は頻度から求めることが前提となるため，事前知識である頻度 C を用いて求める．

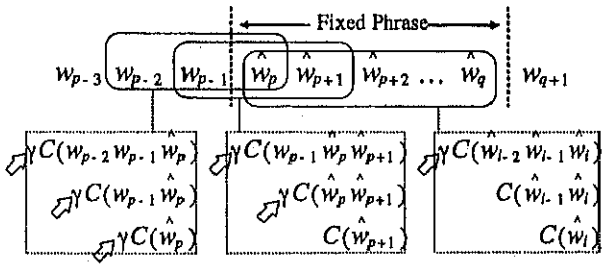

図 2 部分 N-gram 頻度による定型表現の強調 (tri-gram の場合)

Fig. 2 Emphasizing N-gram counts corresponding to fixed phrases (in the case of tri-gram).

Witten-Bell スムージング法 [14] では，単語コンテキスト w_{i-n+1}^{i-1} の直後に現れる単語の異なり語数 $r(w_{i-n+1}^{i-1})$ を用いて，ディスカウント係数 d を次のように求める．

$$d_{w_{i-n+1}^{i}} P_{ML}(w_i|w_{i-n+1}^{i-1}) = \frac{C(w_{i-n+1}^{i})}{C(w_{i-n+1}^{i-1})+r(w_{i-n+1}^{i-1})}$$

提案手法による最ゆう推定値 $P'_{ML(n)}(w_i|w_{i-n+1}^{i-1})$ に対して，$r(w_{i-n+1}^{i-1})$ を用いてディスカウントを行うと，ディスカウント係数 $d'_{w_{i-n+1}^{i}}$ は以下の式で求められる．

$$\begin{aligned} & d'_{w_{i-n+1}^{i}} P'_{ML(n)}(w_i|w_{i-n+1}^{i-1}) \\ &= \frac{\mathbf{C}_n(w_{i-n+1}^{i})}{\{\sum_{w_i}\mathbf{C}_n(w_{i-n+1}^{i})\}+r(w_{i-n+1}^{i-1})} \\ &= \frac{\sum_{w_i}\mathbf{C}_n(w_{i-n+1}^{i})}{\{\sum_{w_i}\mathbf{C}_n(w_{i-n+1}^{i})\}+r(w_{i-n+1}^{i-1})} \\ &\quad \cdot P'_{ML(n)}(w_i|w_{i-n+1}^{i-1}) \end{aligned}$$

3. 実　　験

提案手法の有効性を評価するために，新聞記事と定型表現リストを用いて種々の手法で言語モデルを作成し，質問文の音声認識実験を行った．

一般コーパスとして新聞記事 111 か月分を用い，高頻度 2 万語の N-gram 頻度を抽出し，各適応手法の基本言語モデルとして使用した．適応のための知識として，質問文の定型表現を受理する文法（単語ネットワーク）（図 3）を人手で作成した．更に図 3 の単語ネットワークから全文生成を行い，定型表現のリストを作成した．単語の表記揺れを含めると，172 文が得られた．

比較した言語モデルは以下のとおりである．それぞれ，bi-gram と tri-gram を作成した．

・*BASE*　一般コーパスだけから作成した基本モデル．

・*MIX1*　終助詞「か」を文末にもつ文を一般コーパスから抽出して適応コーパスとみなし，重み w で N-gram の混合を行い作成したモデル．

・*MIX2*　定型表現リストのいずれかの表現を含む文を一般コーパスから抽出して適応コーパスとみなし，重み w で N-gram の混合を行い作成したモデル．

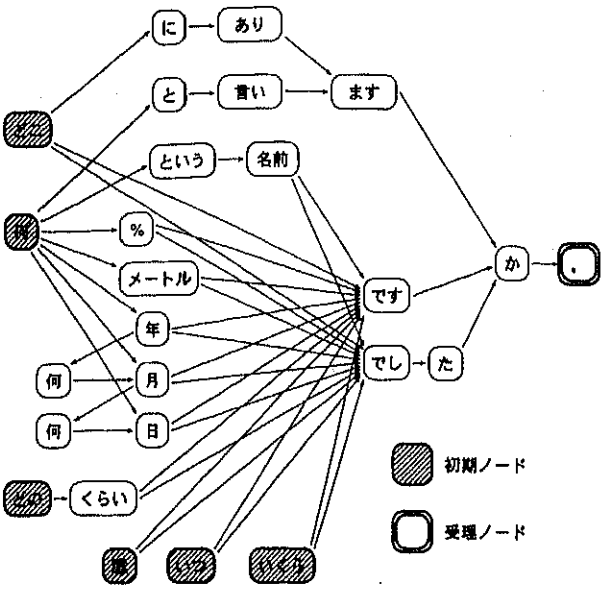

図 3　QA タスク定型表現の単語ネットワーク
Fig. 3　The word network for the fixed phrases used in QA task.

・*MIX3*　定型表現リスト自体を「文」(注2)集合として適応コーパスとみなし，重み w で N-gram の混合を行い作成したモデル．

・*EMP*　本論文の提案手法を用いて，新聞記事から得た N-gram のうち定型表現に関連する部分を γ 倍して強調したモデル．

・*NET*　N-gram 言語モデルに記述文法の制約を埋め込む手法 [2] によって，新聞記事 N-gram と定型表現の単語ネットワークを統合したモデル．N-gram から定型表現への遷移確率を重み γ で強調した．

各手法が *BASE* の認識精度をどれだけ改善できるか比較した．言語モデルのスムージングには，共通して Witten-Bell 法を用いた．

MIX1 と *MIX2* は，それぞれ，文末の終助詞「か」及びリスト中の定型表現を含むことを選択の条件として，新聞記事から質問文の集合を抽出して適応コーパスとみなし，N-gram の重み付混合によるタスク適応手法で作成したモデルである．一般コーパスから適応するタスクに関連する部分集合を抽出し，それを適応コーパスとして利用する従来手法 [9], [12] に相当する．文単位で抽出するため，定型表現以外の表現も適応データとみなされることになる．新聞記事の 10,725,886 文から，*MIX1* は 187,222 文，*MIX2* は 785 文が抽出された．*MIX3* は，利用できる適応コーパスが定型表現リストのみという条件のもとで，従来法の N-gram の重み付き混合によるタスク適応手法で作成したモデルである．人手で記述した例文を適応コーパスとして利用する同様の手法が岡登ら [18] によって提案されている．しかし，本実験では文の一部（定型表現）を単位とする点が異なる．コーパスには定型表現とそれ以外の部分の連接部分を含まないため，連接部分のモデル化に問題がある．*NET* は，文の一部（定型表現）への言語モデルのタスク適応手法として，筆者らが以前提案した手法である．

各手法はそれぞれ，以下に挙げる種々の視点から位置付けされる．第 1 に，適応のために用意する知識源に着目すると，定型表現のリストや文法を使わない手法 (*MIX1*) と使う手法 (*MIX2*, *MIX3*, *EMP*, *NET*) に分類される．第 2 に，事後知識の抽出法に着目すると，一般コーパスから文の抽出を行い事後知識とする手法 (*MIX1*, *MIX2*) と定型表現や文法をそのまま事後知識とする手法 (*MIX3*, *EMP*, *NET*) に分類される．第 3 に，N-gram の抽出法に着目すると，「文」の集合であるコーパスから N-gram を得る手法 (*MIX1*, *MIX2*, *MIX3*) と N-gram を直接抽出する手法 (*EMP*, *NET*) に分類される．第 4 に，適応手法に着目すると，N-gram の重み付き混合による手法 (*MIX1*, *MIX2*, *MIX3*, *EMP*) と N-gram に文法的制約を組み込む手法 (*NET*) に分類される．

入力として，新聞記事 100 文 (*NP*) と質問文 50 件 [19](*QA*) を，男女各 2 人によって読み上げた音声データを用いた．作成したネットワーク文法は 29 単語と比較的小規模であるものの，質問文のうち 72%の 36 文 (*QA'*) がこの文法でモデル化されていた．

大語彙音声認識デコーダ julius [20] のバージョン 3.2 を使用し，音響モデルには 2000 状態 16 混合性別非依存 triphone を用いた，bi-gram の比較には第 1 パスの結果を，tri-gram の比較には第 2 パスの結果を用いた．

実験結果を図 4 (bi-gram) と図 5 (tri-gram) に示す．各手法における重みパラメータ（横軸）と単語誤り率 (WER)（縦軸）との関係を示した．*MIX1*, *MIX2*, *MIX3*, *NET* での重み 0, *EMP* の重み 1 が，*BASE* の結果と一致する点に注意を要する．ベースラインは，対象が *QA* と *NP*，それぞれの場合について示した．

各手法において質問文 (*QA* や *QA'*) を対象に最も単

(注2)：以降，「適応コーパスを構成する文」を括弧付の「文」と記し，認識対象の発話に相当する文と区別する．例えば，定型表現は必ずしも文と一致しないが，手法 *MIX3* では定型表現を「文」として扱う．

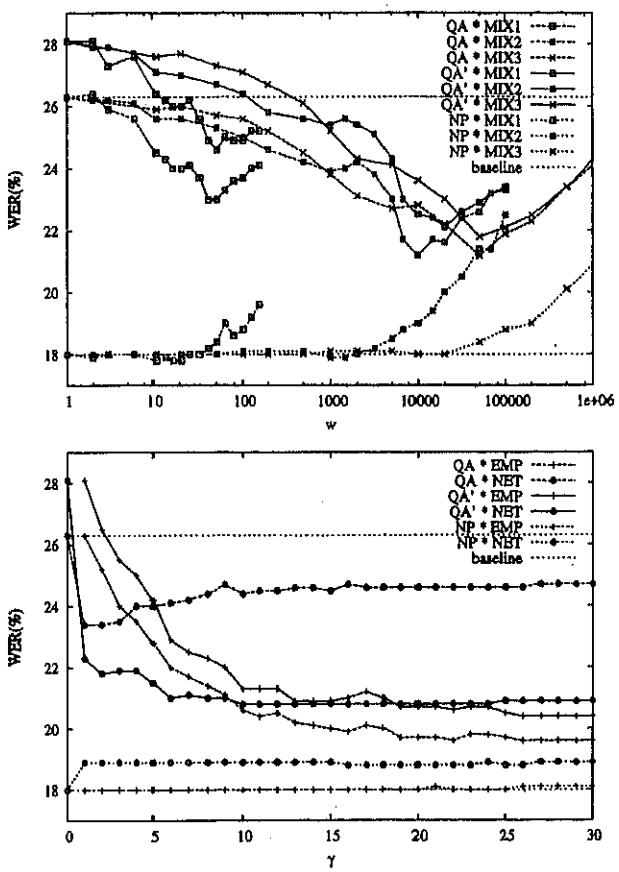

図 4 bi-gram 単語誤り率
Fig. 4 Word error rate for bi-gram.

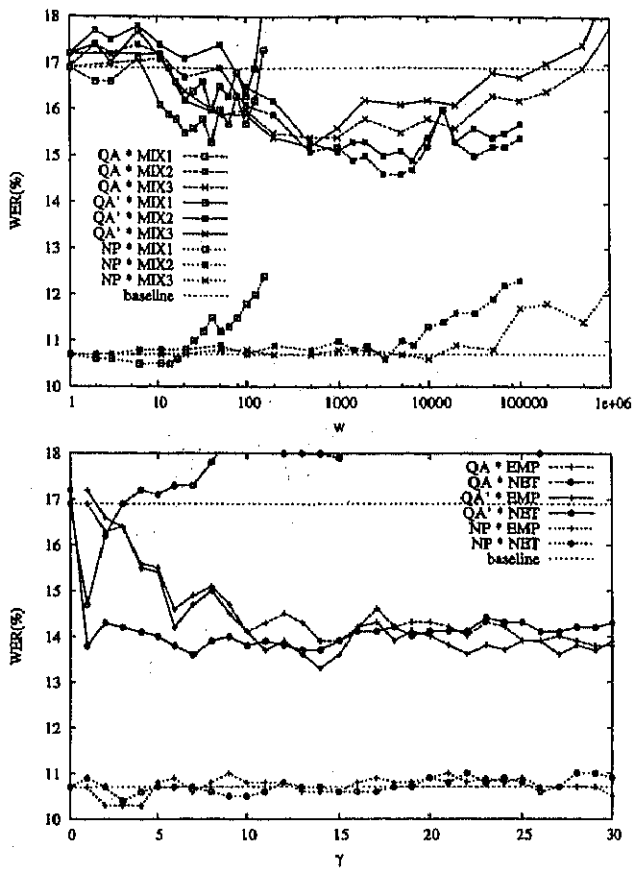

図 5 tri-gram 単語誤り率
Fig. 5 Word error rate for tri-gram.

表 1 重みを最適化した場合の単語誤り率 (%)
Table 1 WERs by optimizing the weights.

target (# of sent.)	language model	文 2-gram	文 3-gram	定型表現 3-gram
QA(50)	*BASE*	26.3	16.9	11.5
	MIX1	23.0	15.3	8.3
	MIX2	21.2	14.6	8.0
	MIX3	21.2	15.4	6.0
	EMP	19.6	13.8	4.5
	NET	23.4	14.7	9.0
QA'(36)	*BASE*	28.1	17.2	11.2
	MIX1	24.6	16.0	8.1
	MIX2	22.1	14.9	5.5
	MIX3	21.8	15.2	3.4
	EMP	20.4	13.3	1.5
	NET	20.8	13.6	1.9

語誤り率を改善している重みに注目する．結果を表 1 の行「文 2-gram」及び「文 3-gram」に示す．どの適応化手法もベースラインの新聞記事モデルに比べて，質問文 (*QA*, *QA'*) に対する単語誤り率を改善していることがわかる．特に提案手法 (*EMP*) は，N-gram の長さ（bi-gram 及び tri-gram）や，定型表現が認識対象をカバーしているか（*QA* あるいは *QA'*）にかかわらず，比較した手法の中で最も単語誤り率を改善した．

各手法の性質を詳しく調査するため，質問文（*QA* 及び *QA'*）に対する tri-gram の結果について，入力の前半（検索トピック）と後半（定型表現）それぞれの認識率を調べた．前後半の境界は，疑問詞（「何」「だれ」等）の前後で判定した．*QA* 及び *QA'* を対象とした結果を，それぞれ図 6，図 7 に示す．*-FP* は疑問詞より前に現れる前半の単語列，*FP* は疑問詞以降の後半の単語列を示す．それぞれ，定型表現以外の表現，定型表現に対応する．また *FP*（定型表現）について，各手法の最も単語誤り率を改善している重みに注目した結果を表 1 の行「定型表現 3-gram」に示す．

まず，定型表現がカバーする質問文 *QA'*（図 7）について，定型表現 (*FP*) に対する単語誤り率を比較する．単語誤り率の小さい順に並べると，*EMP* (1.5%)，*NET* (1.9%)，*MIX3* (3.4%)，*MIX2* (5.5%)，*MIX1* (8.1%) となった．この結果から，適応の知識源として定型表現リストや文法を用いる手法 *MIX2*, *MIX3*, *EMP*, *NET* の方が用いない手法 *MIX1* よりも，事後

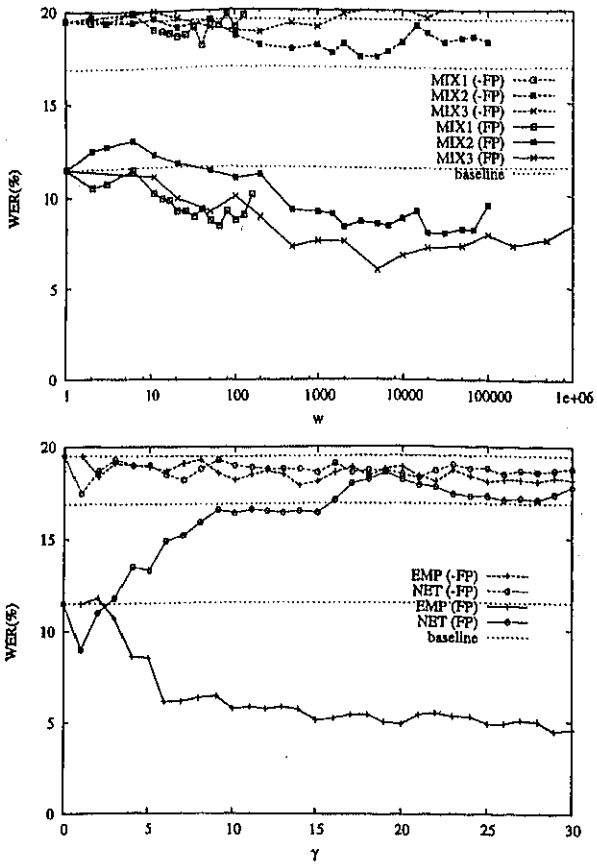

図 6　文の前半と後半でみた単語誤り率 (tri-gram, QA)
Fig. 6　Word error rates of first and latter half of sentences (tri-gram, QA).

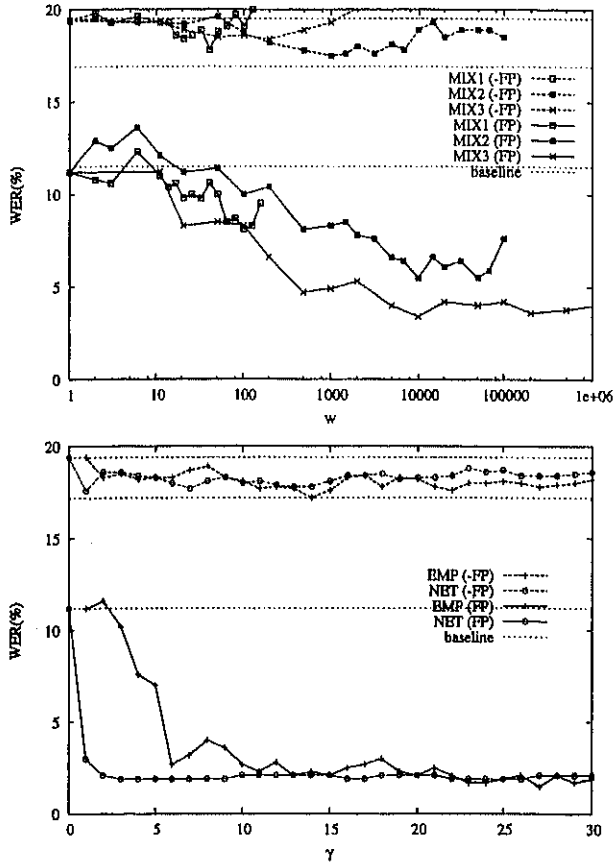

図 7　文の前半と後半でみた単語誤り率 (tri-gram, QA')
Fig. 7　Word error rates of first and latter half of sentences (tri-gram, QA').

知識に定型表現に対する N-gram だけを用いる (*MIX3*, *EMP*, *NET*) の方が定型表現以外の表現の N-gram を含む手法 (*MIX1*, *MIX2*) よりも，また N-gram を直接抽出する手法 (*EMP*, *NET*) の方が「文」単位のコーパスから N-gram を得る手法 (*MIX1*, *MIX2*, *MIX3*) よりもより良い結果を示している．定型表現の認識率改善だけに着目すれば，この順に適応の効果があることがわかる．

次に，定型表現以外の部分 (*-FP*) の認識率に着目する．*MIX3* は定型表現部の単語誤り率を改善する一方，定型表現以外の単語誤り率を悪化させた．これは，手法 *MIX3* の適応コーパスが定型表現とそれ以外の連接部分を含まないことによる影響と考えられる．事後知識に連接部分の N-gram を含む *EMP*, *NET*, *MIX1*, *MIX2* ではこのような傾向は認められない．

定型表現が完全にカバーしない類似した質問文が含まれる質問文 QA (図 6) に着目する．*NET* 以外の手法については，単語誤り率の小さい順に *EMP* (4.5%), *MIX3* (6.0%), *MIX2* (8.0%), *MIX1* (8.3%) と，QA' と同様の傾向であった．一方 *NET* モデルは，文法で扱える表現だけの発話 (QA') に対しては，重み γ によらず単語誤り率を引き下げるが，文法でモデル化されていない類似した発話を含む場合 (QA)，重みを上げることで定型表現部 *FP* の単語誤り率が増加する．ところが，文法の表現とは大きく異なる発話 (*NP* や *-FP*) の場合は，この傾向は認められない．これは，類似した発話を文法で扱える表現で誤認識してしまうことが原因と考えられる．*NET* では，対象とする発話をすべてモデル化するような文法記述の質が重要であることを示唆している．一方，*EMP* モデルではこのような傾向は認められず，より頑健な手法である．

以上の結果から，一般コーパスから文の抽出を行い事後知識とする手法 (*MIX1*, *MIX2*) 及び「文」単位のコーパスから N-gram を得る手法 (*MIX1*, *MIX2*, *MIX3*) などの，文を単位とした従来法と比較して，N-gram を直接強調する提案手法の有効性が明らかに

なった．また，N-gram に文法の制約を埋め込む手法 (*NET*) に比べ，定型表現を逸脱した発話に対する頑健性が明らかになった．

4. 提案手法の位置付け

4.1 言語モデルのタスク適応手法

大量の適応コーパスの入手が困難な特定タスク用に，一般コーパスと少量の適応コーパスから特定タスクの言語モデルを推定する，言語モデルのタスク適応の手法が提案されている [4]．適応モデルの推定法としては，削除補間法 [11]，最大エントロピー法 [15]，最小判別情報量による手法 [10] などが提案されている．中でも比較的単純で効果的な N-gram 言語モデルの適応化手法として，一般的コーパスと適応コーパスの N-gram 頻度を重み付きで足し合わせ，単一の N-gram モデルを作成する手法がある [5], [17]．理論的には，一般テキストから獲得した N-gram 確率を事前分布，対象テキストからの確率を事後分布と見て MAP 推定を行うことに相当する．このとき混合重みは，事前分布をベータ分布とみなした場合の分布のパラメータを決定するために用いられる [17]．

提案手法は，一般コーパスを事前知識とし，一般コーパス自身の部分的な N-gram 頻度を事後知識とみなして，MAP 推定によって言語モデルのタスク適応を行うことに相当する．また，バックオフ言語モデルを作成することを前提とし，定型表現を含む文に割り当てる確率だけが相対的に高くなるように，長さごとに N-gram の推定を行う点に特長がある．

4.2 適応コーパスの収集

言語モデル適応化手法では，対象タスク用の適応コーパスが入手可能であることが前提である．しかし，適応コーパスがあらかじめ入手できない場合，新規にコーパスを収集する必要があり，高価である．この問題に対し，コーパス収集を避ける代替手法や少量の適応コーパスを拡張する手法が提案されている．

適応コーパスの収集を避ける代替手法として，人手で知識を記述する手法と認識結果を利用する手法がある．前者として，対象タスクの文法を記述してテキストデータを自動生成する方法 [8]，対象タスクの典型的な例文を用いる方法 [18]，対象タスクに典型的に現れる表現を記述した文法を N-gram に統合する手法 [2] がある．後者は，一般コーパスから作成した言語モデルを用いて対象音声の認識を行い，認識結果単語列を適応コーパスとする手法 [12], [13] である．これらの手法では，以下で述べる手法を用いて，少量の認識結果を拡張することが多い．

適応コーパスの拡張手法は，大量の一般コーパスから適応するタスクに関連する部分集合を抽出しそれを適応コーパスとして利用する手法である．最初に与える初期適応コーパスは，タスクとの関連性（類似性）の判定に利用される．関連性の尺度としては，初期適応コーパスに対するパープレキシティ減少の基準 [9] や情報検索の分野で広く利用されている tf-idf [12] が用いられる．

提案手法は，人手で作成した例文を利用し，一般コーパスから適応コーパスの拡張を行う手法に相当する．しかし，従来の手法の多くが文や文章を単位にした適応コーパスの拡張を行っているのに対し，提案手法では N-gram を単位に拡張を行っている点が異なる．

N-gram を単位とする利点は次の 2 点にある．第 1 に，適応の対象をより特化できることにある．本論文で扱った質問文では，文の一部である定型表現に対する誤り率だけを大きく引き下げることができた．第 2 に，より多くの学習データが得られる点である．例えば本論文の実験において，文単位で選択できた定型表現を含む文は約 1 千万文中わずか 785 文であり，得られた N-gram は少ない．提案法では，文よりずっと小さな N-gram 単位で選択することで，多くの N-gram を学習データとすることができる．

4.3 言語モデルにおける定型表現の扱い

定型表現を利用して言語モデルの改善を行う手法として，学習コーパスから頻度やエントロピー減少などの基準から定型表現（単語連鎖）を自動的に抽出し，それを新たな一単語とみなし，N-gram を再計算することでモデルの性能を向上させる手法が提案されている [21], [22]．これらの手法は，特定タスクのコーパスが入手可能という前提で，定型表現を適切にモデル化する．これに対し，提案手法は，定型表現が頻出ではない一般コーパスを用いたタスク適応手法であり，目的が異なる．

5. むすび

音声入力による質問応答システムを目指し，音声入力の質問文を高精度で認識する言語モデルを作成する手法を提案した．本手法の特長は，回答抽出の対象となる一般コーパス（新聞記事や辞典）から作成した N-gram を，人手で作成した質問文固有の定型表現を列挙したリストを用いて適応化する点にあった．理論

的には，一般コーパスを事前知識とし，一般コーパス自身の部分的な N-gram 頻度を事後知識とみなして，最大事後確率推定（MAP 推定）によって言語モデルのタスク適応を行うことに相当する．従来の言語モデル適応手法が文を単位とした適応の手法であったのに対し，本手法は文を構成する部分的な表現（定型表現）に適応する手法である．評価実験の結果，文を単位とした従来法に比べ，本手法は単語誤り率を改善することが示された．また本手法は，入力音声中の定型表現の認識率を改善し，それ以外の部分の認識率を悪化することはなかった．本手法は，質問応答以外でも，比較的多様性が少ない（人手で記述できる程度の多様性をもつ）定型表現を伴うタスク一般に応用可能である．

文　献

[1] 小特集 —音声対話システムの実力と課題—，音響誌，vol.54, no.11, 1998.

[2] T. Akiba, K. Itou, A. Fujii, and T. Ishikawa, "Selective back-off smoothing for incorporating grammatical constraints into the n-gram language model," Proc. International Conference on Spoken Language Processing, vol.2, pp.881–884, Denver, Colorado, Sept. 2002.

[3] J. Barnett, S. Anderson, J. Broglio, M. Singh, R. Hudson, and S.W. Kuo, "Experiments in spoken queries for document retrieval," Proc. European Conference on Speech Communication and Technology, pp.1323–1326, Rhodes, Greece, Sept. 1997.

[4] J.R. Bellegarda, "An overview of statistical language model adaptation," Proc. Isca ITR Workshop on Adaptation Methods for Speech Recognition, Sophia-Antipolis, France, Aug. 2001.

[5] M. Federico, "Bayesian estimation methods for n-gram language model adaptation," Proc. International Conference on Spoken Language Processing, pp.240–243, Philadelphia, USA, 1996.

[6] A. Fujii, K. Itou, and T. Ishikawa, "Speech-driven text retrieval: Using target IR collections for statistical language model adaptation in speech recognition," Information Retrieval Techniques for Speech Applications (LNCS 2273), ed. A.R. Coden, E. W. Brown, and S. Srinivasan, pp.94–104, Springer, 2002.

[7] J. Fukumoto, T. Kato, and F. Masui, "Question answering challenge (QAC-1) question answering evaluation at NTCIR workshop 3," Working Notes of the Third NTCIR Workshop Meeting, pp.1–6, Tokyo, Japan, Oct. 2002.

[8] L. Galescu, E. Ringger, and J. Allen, "Rapid language model development for new task domains," Proc. International Conference on Language Resources and Evaluation, pp.807–812, Granada, Spain, May 1998.

[9] D. Klakow, "Selecting articles from the language model training corpus," Proc. International Conference on Acoustics Speech and Signal Processing, vol.3, pp.1695–1698, Istanbul, Turkey, June 2000.

[10] R. Kneser, J. Peters, and D. Klakow, "Language model adaptation using dynamic marginals," Proc. European Conference on Speech Communication and Technology, pp.1971–1974, Rodos, Greece, Sept. 1997.

[11] R. Kneser and V. Steinbiss, "On the dynamic adaptation of stochastic language models," Proc. International Conference on Acoustics Speech and Signal Processing, pp.586–589, Minneapolis, USA, April 1993.

[12] M. Mahajan, D. Beeferman, and X.D. Huang, "Improved topic-dependent language modeling using information retrieval techniques," Proc. International Conference on Acoustics Speech and Signal Processing, Phoenix, Arizona, March 1999.

[13] T. Niesler and D. Willett, "Unsupervised language model adaptation for lecture speech transcription," Proc. International Conference on Spoken Language Processing, vol.2, pp.1413–1416, Denver, Colorado, Sept. 2002.

[14] P. Placeway, R. Schwartz, P. Fung, and L. Nguyen, "The estimation of powerful language models from small and large corpora," Proc. International Conference on Acoustics Speech and Signal Processing, vol.2, pp.33–36, Minneapolis, USA, April 1993.

[15] R. Rosenfeld, "A maximum entropy approach to adaptive statistical language modeling," Comput. Speech Lang., vol.10, no.3, pp.155–186, July 1996.

[16] E. Voorhees and D. Tice, "The TREC-8 question answering track evaluation," Proc. 8th Text Retrieval Conference, pp.83–106, Gaithersburg, Maryland, 1999.

[17] 伊藤彰則，好田正紀，"N-gram 出現回数の混合によるタスク適応の性能解析，" 信学論（D-II），vol.J83-D-II, no.11, pp.2418–2427, Nov. 2000.

[18] 岡登洋平，石井　純，花沢利行，"タスクの例文を用いた自由発話音声認識のための言語モデルの構築，" 音響講論集，pp.73–74, Oct. 2001.

[19] 佐々木裕，磯崎秀樹，平　博順，廣田啓一，賀沢秀人，中島裕之，"質問応答システムの比較と評価，" 信学技報，NLC-24, 2000.

[20] 鹿野清宏，伊藤克亘，河原達也，武田一哉，山本幹雄（編），音声認識システム，オーム社，2001.

[21] 中川聖一，赤松裕隆，西崎博光，"音声認識用言語モデルのためのタスク適応化と定型表現の利用，" 自然言語処理，vol.6, no.2, pp.97–115, 1999.

[22] 和田陽介，小林紀彦，中野裕一郎，小林哲則，"大語彙連続音声認識における連鎖語の追加による語彙拡大の効果，" 情処学論，vol.40, no.4, pp.1413–1420, 1999.

（平成 14 年 12 月 27 日受付，15 年 5 月 6 日再受付）

Speech-Driven Text Retrieval: Using Target IR Collections for Statistical Language Model Adaptation in Speech Recognition

Atsushi Fujii[1], Katunobu Itou[2], and Tetsuya Ishikawa[1]

[1] University of Library and Information Science
1-2 Kasuga, Tsukuba, 305-8550, Japan
{fujii,ishikawa}@ulis.ac.jp
[2] National Institute of Advanced Industrial Science and Technology
1-1-1 Chuuou Daini Umezono, Tsukuba, 305-8568, Japan
itou@ni.aist.go.jp

Abstract. Speech recognition has of late become a practical technology for real world applications. Aiming at speech-driven text retrieval, which facilitates retrieving information with spoken queries, we propose a method to integrate speech recognition and retrieval methods. Since users speak contents related to a target collection, we adapt statistical language models used for speech recognition based on the target collection, so as to improve both the recognition and retrieval accuracy. Experiments using existing test collections combined with dictated queries showed the effectiveness of our method.

1 Introduction

Automatic speech recognition, which decodes human voice to generate transcriptions, has of late become a practical technology. It is feasible that speech recognition is used in real world computer-based applications, specifically, those associated with human language. In fact, a number of speech-based methods have been explored in the information retrieval community, which can be classified into the following two fundamental categories:

- spoken document retrieval, in which written queries are used to search speech (e.g., broadcast news audio) archives for relevant speech information [5, 6, 15–17, 19, 20],
- speech-driven (spoken query) retrieval, in which spoken queries are used to retrieve relevant textual information [2, 3].

Initiated partially by the TREC-6 spoken document retrieval (SDR) track [4], various methods have been proposed for spoken document retrieval. However, a relatively small number of methods have been explored for speech-driven text retrieval, although they are associated with numerous keyboard-less retrieval applications, such as telephone-based retrieval, car navigation systems, and user-friendly interfaces.

Barnett et al. [2] performed comparative experiments related to speech-driven retrieval, where an existing speech recognition system was used as an input interface for the INQUERY text retrieval system. They used as test inputs 35 queries collected from the TREC 101-135 topics, dictated by a single male speaker. Crestani [3] also used the above 35 queries and showed that conventional relevance feedback techniques marginally improved the accuracy for speech-driven text retrieval.

These above cases focused solely on improving text retrieval methods and did not address problems of improving speech recognition accuracy. In fact, an existing speech recognition system was used with no enhancement. In other words, speech recognition and text retrieval modules were fundamentally independent and were simply connected by way of an input/output protocol.

However, since most speech recognition systems are trained based on specific domains, the accuracy of speech recognition across domains is not satisfactory. Thus, as can easily be predicted, in cases of Barnett et al. [2] and Crestani [3], a relatively high speech recognition error rate considerably decreased the retrieval accuracy. Additionally, speech recognition with a high accuracy is crucial for interactive retrieval.

Motivated by these problems, in this paper we integrate (not simply connect) speech recognition and text retrieval to improve both recognition and retrieval accuracy in the context of speech-driven text retrieval.

Unlike general-purpose speech recognition aimed to decode any spontaneous speech, in the case of speech-driven text retrieval, users usually speak contents associated with a target collection, from which documents relevant to their information need are retrieved. In a stochastic speech recognition framework, the accuracy depends primarily on acoustic and language models [1]. While acoustic models are related to phonetic properties, language models, which represent linguistic contents to be spoken, are strongly related to target collections. Thus, it is intuitively feasible that language models have to be produced based on target collections.

To sum up, our belief is that by adapting a language model based on a target IR collection, we can improve the speech recognition and text retrieval accuracy, simultaneously.

Section 2 describes our prototype speech-driven text retrieval system, which is currently implemented for Japanese. Section 3 elaborates on comparative experiments, in which existing test collections for Japanese text retrieval are used to evaluate the effectiveness of our system.

2 System Description

2.1 Overview

Figure 1 depicts the overall design of our speech-driven text retrieval system, which consists of speech recognition, text retrieval and adaptation modules. We explain the retrieval process based on this figure.

In the off-line process, the adaptation module uses the entire target collection (from which relevant documents are retrieved) to produce a language model, so that user speech related to the collection can be recognized with a high accuracy. On the other hand, an acoustic model is produced independent of the target collection.

In the on-line process, given an information need spoken by a user, the speech recognition module uses the acoustic and language models to generate a transcription for the user speech. Then, the text retrieval module searches the collection for documents relevant to the transcription, and outputs a specific number of top-ranked documents according to the degree of relevance, in descending order.

These documents are fundamentally final outputs. However, in the case where the target collection consists of multiple domains, a language model produced in the off-line adaptation process is not necessarily precisely adapted to a specific information need. Thus, we optionally use top-ranked documents obtained in the initial retrieval process for an on-line adaptation, because these documents are associated with the user speech more than the entire collection. We then reperform speech recognition and text retrieval processes to obtain final outputs.

In other words, our system is based on the two-stage retrieval principle [8], where top-ranked documents retrieved in the first stage are intermediate results, and are used to improve the accuracy for the second (final) stage. From a different perspective, while the off-line adaptation process produces the *global* language model for a target collection, the on-line adaptation process produces a *local* language model based on the user speech.

In the following sections, we explain speech recognition, adaptation, and text retrieval modules in Figure 1, respectively.

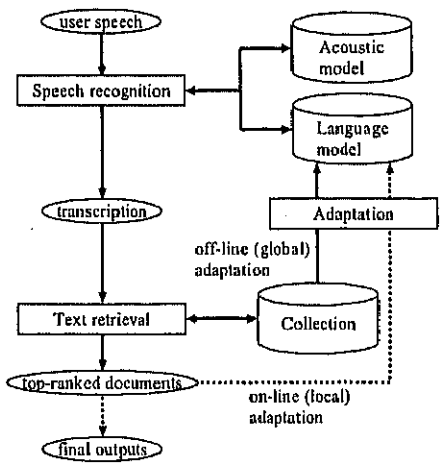

Fig. 1. The overall design of our speech-driven text retrieval system.

2.2 Speech Recognition

The speech recognition module generates word sequence W, given phoneme sequence X. In the stochastic speech recognition framework, the task is to output the W maximizing $P(W|X)$, which is transformed as in equation (1) through use of the Bayesian theorem.

$$\arg\max_W P(W|X) = \arg\max_W P(X|W) \cdot P(W) \quad (1)$$

Here, $P(X|W)$ models a probability that word sequence W is transformed into phoneme sequence X, and $P(W)$ models a probability that W is linguistically acceptable. These factors are usually called acoustic and language models, respectively.

For the speech recognition module, we use the Japanese dictation toolkit [7][1], which includes the "Julius" recognition engine and acoustic/language models trained based on newspaper articles. This toolkit also includes development softwares, so that acoustic and language models can be produced and replaced depending on the application. While we use the acoustic model provided in the toolkit, we use new language models produced by way of the adaptation process (see Section 2.3).

2.3 Language Model Adaptation

The basis of the adaptation module is to produce a word-based N-gram (in our case, a combination of bigram and trigram) model by way of source documents.

In the off-line (global) adaptation process, we use the ChaSen morphological analyzer [10] to extract words from the entire target collection, and produce the global N-gram model.

On the other hand, in the on-line (local) adaptation process, only top-ranked documents retrieved in the first stage are used as source documents, from which word-based N-grams are extracted as performed in the off-line process. However, unlike the case of the off-line process, we do not produce the entire language model. Instead, we re-estimate only statistics associated with top-ranked documents, for which we use the MAP (Maximum A-posteriori Probability) estimation method [9].

Although the on-line adaptation theoretically improves the retrieval accuracy, for real-time usage, the trade-off between the retrieval accuracy and computational time required for the on-line process has to be considered.

Our method is similar to the one proposed by Seymore and Rosenfeld [14] in the sense that both methods adapt language models based on a small number of documents related to a specific domain (or topic). However, unlike their method, our method does not require corpora manually annotated with topic tags.

[1] http://winnie.kuis.kyoto-u.ac.jp/dictation/

2.4 Text Retrieval

The text retrieval module is based on an existing probabilistic retrieval method [13], which computes the relevance score between the transcribed query and each document in the collection. The relevance score for document i is computed based on equation (2).

$$\sum_t \left(\frac{TF_{t,i}}{\frac{DL_i}{avglen} + TF_{t,i}} \cdot \log \frac{N}{DF_t} \right) \tag{2}$$

Here, t's denote terms in transcribed queries. $TF_{t,i}$ denotes the frequency that term t appears in document i. DF_t and N denote the number of documents containing term t and the total number of documents in the collection. DL_i denotes the length of document i (i.e., the number of characters contained in i), and $avglen$ denotes the average length of documents in the collection.

We use content words extracted from documents as terms, and perform a word-based indexing. For this purpose, we use the ChaSen morphological analyzer [10] to extract content words. We extract terms from transcribed queries using the same method.

3 Experimentation

3.1 Test Collections

We investigated the performance of our system based on the NTCIR workshop evaluation methodology, which resembles the one in the TREC ad hoc retrieval track. In other words, each system outputs 1,000 top documents, and the TREC evaluation software was used to plot recall-precision curves and calculate non-interpolated average precision values.

The NTCIR workshop was held twice (in 1999 and 2001), for which two different test collections were produced: the NTCIR-1 and 2 collections [11, 12][2]. However, since these collections do not include spoken queries, we asked four speakers (two males/females) to dictate information needs in the NTCIR collections, and simulated speech-driven text retrieval.

The NTCIR collections include documents collected from technical papers published by 65 Japanese associations for various fields. Each document consists of the document ID, title, name(s) of author(s), name/date of conference, hosting organization, abstract and author keywords, from which we used titles, abstracts and keywords for the indexing. The number of documents in the NTCIR-1 and 2 collections are 332,918 and 736,166, respectively (the NTCIR-1 documents are a subset of the NTCIR-2).

The NTCIR-1 and 2 collections also include 53 and 49 topics, respectively. Each topic consists of the topic ID, title of the topic, description, narrative. Figure 2 shows an English translation for a fragment of the NTCIR topics[3],

[2] http://research.nii.ac.jp/~ntcadm/index-en.html

[3] The NTCIR-2 collection contains Japanese topics and their English translations.

where each field is tagged in an SGML form. In general, titles are not informative for the retrieval. On the other hand, narratives, which usually consist of several sentences, are too long to speak. Thus, only descriptions, which consist of a single phrase and sentence, were dictated by each speaker, so as to produce four different sets of 102 spoken queries.

```
<TOPIC q=0118>
<TITLE>TV conferencing</TITLE>
<DESCRIPTION>Distance education support systems using TV
conferencing</DESCRIPTION>
<NARRATIVE>A relevant document will provide information on
the development of distance education support systems using TV
conferencing. Preferred documents would present examples of using
TV conferencing and discuss the results. Any reported methods
of aiding remote teaching are relevant documents (for example,
ways of utilizing satellite communication, the Internet, and ISDN
circuits).</NARRATIVE>
</TOPIC>
```

Fig. 2. An English translation for an example topic in the NTCIR collections.

In the NTCIR collections, relevance assessment was performed based on the pooling method [18]. First, candidates for relevant documents were obtained with multiple retrieval systems. Then, for each candidate document, human expert(s) assigned one of three ranks of relevance: "relevant," "partially relevant" and "irrelevant." The NTCIR-2 collection also includes "highly relevant" documents. In our evaluation, "highly relevant" and "relevant" documents were regarded as relevant ones.

3.2 Comparative Evaluation

In order to investigate the effectiveness of the off-line language model adaptation, we compared the performance of the following different retrieval methods:

- text-to-text retrieval, which used written descriptions as queries, and can be seen as the perfect speech-driven text retrieval,
- speech-driven text retrieval, in which a language model produced based on the NTCIR-2 collection was used,
- speech-driven text retrieval, in which a language model produced based on ten years worth of *Mainichi Shimbun* Japanese newspaper articles (1991-2000) was used.

The only difference in producing two different language models (i.e., those based on the NTCIR-2 collection and newspaper articles) are the source documents.

In other words, both language models have the same vocabulary size (20,000), and were produced using the same softwares.

Table 1 shows statistics related to word tokens/types in two different source corpora for language modeling, where the line "Coverage" denotes the ratio of word tokens contained in the resultant language model. Most of word tokens were covered in both language models.

Table 1. Statistics associated with source words for language modeling.

	NTCIR	News
# of Types	454K	315K
# of Tokens	175M	262M
Coverage	97.9%	96.5%

In cases of speech-driven text retrieval methods, queries dictated by four speakers were used individually. Thus, in practice we compared nine different retrieval methods. Although the Julius decoder outputs more than one transcription candidate for a single speech input, we used only the one with the greatest probability score. The results did not significantly change depending on whether or not we used lower-ranked transcriptions as queries.

Table 2 shows the non-interpolated average precision values and word error rate in speech recognition, for different retrieval methods. As with existing experiments for speech recognition, word error rate (WER) is the ratio between the number of word errors (i.e., deletion, insertion, and substitution) and the total number of words. In addition, we also investigated error rate with respect to query terms (i.e., keywords used for retrieval), which we shall call "term error rate (TER)."

In Table 2, the first line denotes results of the text-to-text retrieval, which were relatively high compared with existing results reported in the NTCIR workshops [11, 12].

The remaining lines denote results of speech-driven text retrieval combined with the NTCIR-based language model (lines 2-5) and the newspaper-based model (lines 6-9), respectively. Here, "Mx" and "Fx" denote male/female speakers, respectively. Suggestions which can be derived from these results are as follows.

First, for both language models, results did not significantly change depending on the speaker. The best average precision values for speech-driven text retrieval were obtained with a combination of queries dictated by a male speaker (M1) and the NTCIR-based language model, which were approximately 80% of those with the text-to-text retrieval.

Second, by comparing results of different language models for each speaker, one can see that the NTCIR-based model significantly decreased WER and TER obtained with the newspaper-based model, and that the retrieval method using

Table 2. Results for different retrieval methods (AP: average precision, WER: word error rate, TER: term error rate).

Method	NTCIR-1			NTCIR-2		
	AP	WER	TER	AP	WER	TER
Text	0.3320	—	—	0.3118	—	—
M1 (NTCIR)	0.2708	0.1659	0.2190	0.2504	0.1532	0.2313
M2 (NTCIR)	0.2471	0.2034	0.2381	0.2114	0.2180	0.2799
F1 (NTCIR)	0.2276	0.1961	0.2857	0.1873	0.1885	0.2500
F2 (NTCIR)	0.2642	0.1477	0.2222	0.2376	0.1635	0.2388
M1 (News)	0.1076	0.3547	0.5143	0.0790	0.3594	0.5149
M2 (News)	0.1257	0.4044	0.5460	0.0691	0.5022	0.6343
F1 (News)	0.1156	0.3801	0.5238	0.0798	0.4418	0.5709
F2 (News)	0.1225	0.3317	0.5016	0.0917	0.4080	0.5858

the NTCIR-based model significantly outperformed one using the newspaper-based model. In addition, these results were observable, irrespective of the speaker. Thus, we conclude that adapting language models based on target collections was quite effective for speech-driven text retrieval.

Third, TER was generally higher than WER irrespective of the speaker. In other words, speech recognition for content words was more difficult than functional words, which were not contained in query terms.

We analyzed transcriptions for dictated queries, and found that speech recognition error was mainly caused by the out-of-vocabulary problem. In the case where major query terms are mistakenly recognized, the retrieval accuracy substantially decreases. In addition, descriptions in the NTCIR topics often contain expressions which do not appear in the documents, such as "I want papers about..." Although these expressions usually do not affect the retrieval accuracy, misrecognized words affect the recognition accuracy for remaining words including major query terms. Consequently, the retrieval accuracy decreases due to the partial misrecognition.

Finally, we investigated the trade-off between recall and precision. Figures 3 and 4 show recall-precision curves of different retrieval methods, for the NTCIR-1 and 2 collections, respectively. In these figures, the relative superiority for precision values due to different language models in Table 2 was also observable, regardless of the recall.

However, the effectiveness of the on-line adaptation remains an open question and needs to be explored.

4 Conclusion

Aiming at speech-driven text retrieval with a high accuracy, we proposed a method to integrate speech recognition and text retrieval methods, in which target text collections are used to adapt statistical language models for speech

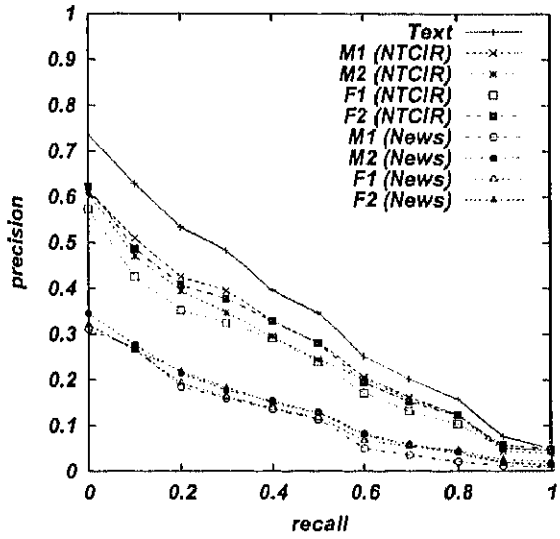

Fig. 3. Recall-precision curves for different retrieval methods using the NTCIR-1 collection.

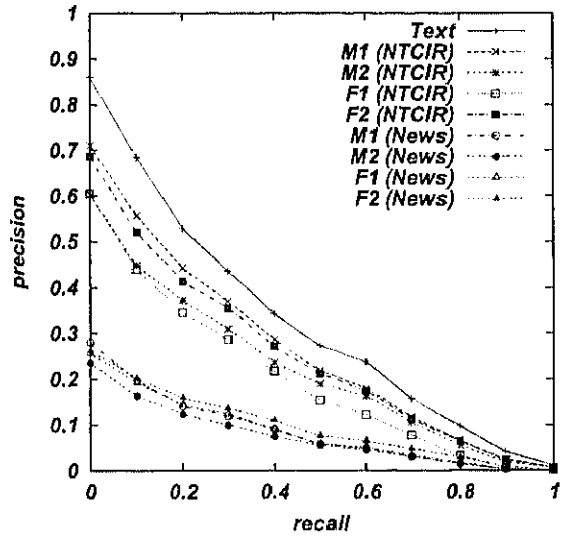

Fig. 4. Recall-precision curves for different retrieval methods using the NTCIR-2 collection.

recognition. We also showed the effectiveness of our method by way of experiments, where dictated information needs in the NTCIR collections were used as queries to retrieve technical abstracts. Future work would include experiments on various collections, such as newspaper articles and Web pages.

5 Acknowledgments

The authors would like to thank the National Institute of Informatics for their support with the NTCIR collections.

References

1. L. R. Bahl, F. Jelinek, and R. L. Mercer. A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(2):179–190, 1983.
2. J. Barnett, S. Anderson, J. Broglio, M. Singh, R. Hudson, and S. W. Kuo. Experiments in spoken queries for document retrieval. In *Proceedings of Eurospeech97*, pages 1323–1326, 1997.
3. F. Crestani. Word recognition errors and relevance feedback in spoken query processing. In *Proceedings of the Fourth International Conference on Flexible Query Answering Systems*, pages 267–281, 2000.
4. J. S. Garofolo, E. M. Voorhees, V. M. Stanford, and K. S. Jones. TREC-6 1997 spoken document retrieval track overview and results. In *Proceedings of the 6th Text REtrieval Conference*, pages 83–91, 1997.
5. S. Johnson, P. Jourlin, G. Moore, K. S. Jones, and P. Woodland. The Cambridge University spoken document retrieval system. In *Proceedings of ICASSP'99*, pages 49–52, 1999.
6. G. Jones, J. Foote, K. S. Jones, and S. Young. Retrieving spoken documents by combining multiple index sources. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 30–38, 1996.
7. T. Kawahara, A. Lee, T. Kobayashi, K. Takeda, N. Minematsu, S. Sagayama, K. Itou, A. Ito, M. Yamamoto, A. Yamada, T. Utsuro, and K. Shikano. Free software toolkit for Japanese large vocabulary continuous speech recognition. In *Proceedings of the 6th International Conference on Spoken Language Processing*, pages 476–479, 2000.
8. K. Kwok and M. Chan. Improving two-stage ad-hoc retrieval for short queries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 250–256, 1998.
9. H. Masataki, Y. Sagisaka, K. Hisaki, and T. Kawahara. Task adaptation using MAP estimation in n-gram language modeling. In *Proceedings of ICASSP'97*, pages 783–786, 1997.
10. Y. Matsumoto, A. Kitauchi, T. Yamashita, Y. Hirano, H. Matsuda, and M. Asahara. Japanese morphological analysis system ChaSen version 2.0 manual 2nd edition. Technical Report NAIST-IS-TR99009, NAIST, 1999.
11. National Center for Science Information Systems. *Proceedings of the 1st NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, 1999.

12. National Institute of Informatics. *Proceedings of the 2nd NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization*, 2001.
13. S. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 232–241, 1994.
14. K. Seymore and R. Rosenfeld. Using story topics for language model adaptation. In *Proceedings of Eurospeech97*, 1997.
15. P. Sheridan, M. Wechsler, and P. Schäuble. Cross-language speech retrieval: Establishing a baseline performance. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 99–108, 1997.
16. A. Singhal and F. Pereira. Document expansion for speech retrieval. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 34–41, 1999.
17. S. Srinivasan and D. Petkovic. Phonetic confusion matrix based spoken document retrieval. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 81–87, 2000.
18. E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 315–323, 1998.
19. M. Wechsler, E. Munteanu, and P. Schäuble. New techniques for open-vocabulary spoken document retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 20–27, 1998.
20. S. Whittaker, J. Hirschberg, J. Choi, D. Hindle, F. Pereira, and A. Singhal. SCAN: Designing and evaluating user interfaces to support retrieval from speech archives. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 26–33, 1999.

A Cross-media Retrieval System for Lecture Videos

Atsushi Fujii[†,†††], *Katunobu Itou*[††,†††], *Tomoyosi Akiba*[††], *Tetsuya Ishikawa*[†]

[†] Institute of Library and Information Science, University of Tsukuba
[††] National Institute of Advanced Industrial Science and Technology
[†††] CREST, Japan Science and Technology Corporation

Abstract

We propose a cross-media lecture-on-demand system, in which users can selectively view specific segments of lecture videos by submitting text queries. Users can easily formulate queries by using the textbook associated with a target lecture, even if they cannot come up with effective keywords. Our system extracts the audio track from a target lecture video, generates a transcription by large vocabulary continuous speech recognition, and produces a text index. Experimental results showed that by adapting speech recognition to the topic of the lecture, the recognition accuracy increased and the retrieval accuracy was comparable with that obtained by human transcription.

1. Introduction

The growing number of multimedia contents available via the World Wide Web, CD-ROMs, and DVDs has made information technologies incorporating speech, image, and text processing crucial. Of the various types of contents, lectures (audio/video) are typical and a valuable multimedia resource, in which speeches (i.e., oral presentations) are usually organized based on text materials, such as resumes, slides, and textbooks. In lecture videos, image information, such as flip charts, is often also used. In other words, a single lecture consists of different types of compatible multimedia contents.

Because a single lecture often refers to several topics and takes a long time, it is useful to obtain specific segments (passages) selectively so that the audience can satisfy their information needs at minimum cost. To resolve this problem, in this paper we propose a lecture-on-demand system that retrieves relevant video/audio passages in response to user queries. For this purpose, we utilize the benefits of different media types to improve retrieval performance.

On the one hand, text has the advantage that users can view/scan the entire contents quickly and can easily identify relevant passages using the layout information (e.g., text structures based on sections and paragraphs). In other words, text contents can be used for random-access purposes. On the other hand, speech is used mainly for sequential-access purposes. Therefore, it is difficult to identify relevant passages unless target video/audio data includes additional annotation, such as indexes. Even if the target data are indexed, users are not necessarily able to provide effective queries. To resolve this problem, textbooks are desirable materials from which users can extract effective keywords and phrases. However, while textbooks are usually concise, speech has a high degree of redundancy and therefore is easier to understand than textbooks, especially where additional image information is provided.

In view of the above, we model our lecture-on-demand (LOD) system as follows. A user selects text segments (keywords, phrases, sentences, and paragraphs) that are relevant to their information needs from a textbook for a target lecture. By using selected segments, a text query is generated automatically. That is, queries can be formulated even if users cannot provide effective keywords. Users can also submit additional keywords as queries, if necessary. Video passages relevant to a given query are retrieved and presented to the user. To retrieve the video passages in response to text queries, we extract the audio track from a lecture video, generate a transcription by means of large vocabulary continuous speech recognition, and produce a text index, prior to system use. Our system is a cross-media system in the sense that users can retrieve video and audio information by means of text queries.

2. System Description

2.1. Overview

Figure 1 depicts the overall design of our lecture-on-demand system, in which the left and right regions correspond to the on-line and off-line processes, respectively. Although our system is currently implemented for Japanese, our methodology is fundamentally language independent. For the purpose of research and development, we tentatively target lecture programs on TV for which textbooks are published. We explain the basis of our system using Figure 1.

In the off-line process, given the video data of a target lecture, audio data are extracted and segmented into a number of passages. Then, a speech recognition system transcribes each passage. Finally, the transcribed passages are indexed as in conventional text retrieval systems, so that each passage can be retrieved efficiently in response to text queries. To adapt speech recognition to a specific lecturer, we perform unsupervised speaker adaptation using an initial speech recognition result (i.e., a transcription). To adapt speech recognition to a specific topic, we perform language model adaptation, for which we search a general corpus for documents relevant to the textbook related to a target lecture. Then, retrieved documents (i.e., a topic-specific corpus) are used to produce a word-based N-gram language model. We also perform image analysis to extract text (e.g., keywords and phrases) from flip charts. These contents are also used to improve our language model.

In the on-line process, a user can view specific video passages by submitting any text queries, i.e., keywords, phrases, sentences, and paragraphs, extracted from the textbook. Any queries not in the textbook can also be used. The current implementation is based on a client-server system on the Web. Both the off-line and on-line processes are performed on servers, but users can access our system using Web browsers on their own PCs.

Figure 2 depicts a prototype interface of our LOD system, in which a lecture associated with "nonlinear multivariate anal-

ysis" is given. In this interface, an electronic version of a textbook is displayed on the left side, and a lecture video is played on the right side. In addition, users can submit any text queries in the input box, which is not depicted in Figure 2. In this scenario, a text paragraph related to "discriminant analysis" was copied and pasted into the query input box, and top-ranked transcribed passages for the query were listed according to the degree of relevance (in the lower part of Figure 2). Users can select (click on) transcriptions to play the corresponding video passage.

It should be noted that unlike conventional keyword-based retrieval systems, in which users usually submit a small number of keywords, in our system users can easily submit longer queries using textbooks. Where submitted keywords are misrecognized in transcriptions, the retrieval accuracy decreases. However, longer queries are relatively robust for speech recognition errors, because the effect of misrecognized words is overshadowed by the large number of words correctly recognized.

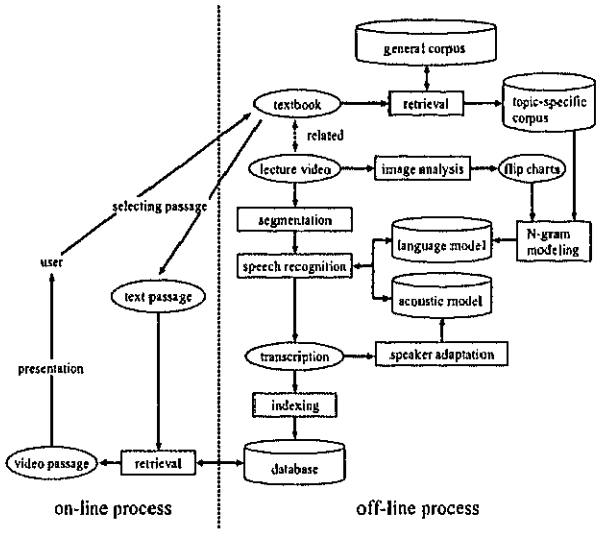

Figure 1: An overview of our lecture-on-demand system.

2.2. Passage Segmentation

The basis of passage segmentation is to divide the entire video data for a single lecture into more than one unit to be retrieved. We call these smaller units "passages". For this purpose, both speech and image data can provide promising clues. However, in lecture TV programs, it is often the case that a lecturer sitting still is the main focus and a small number of flip charts are used occasionally. In such cases, image data is less informative. Therefore, tentatively we use only speech data for the passage segmentation process. However, segmentation can potentially vary depending on the user query. Thus, it is difficult to predetermine a desirable segmentation in the off-line process.

Because of the above problems, we first extract the audio track from a target video and use a simple pause-based segmentation method to obtain minimal speech units, such as sentences and long phrases. In other words, speech units are continuous audio segments that do not include pauses longer than a certain threshold. Finally, we generate variable-length passages from one or more speech units. To put it more precisely, we combine

Figure 2: The interface of our LOD system over the Web.

N speech units into a single passage, with N ranging from 1 to 5 in the current implementation.

2.3. Speech Recognition

The speech recognition module generates word sequence W, given phone sequence X. In a stochastic framework, the task is to select the W maximizing $P(W|X)$, which is transformed as in Equation (1) through the Bayesian theorem.

$$\arg\max_W P(W|X) = \arg\max_W P(X|W) \cdot P(W) \quad (1)$$

$P(X|W)$ models the probability that the word sequence W is transformed into the phone sequence X, and $P(W)$ models the probability that W is linguistically acceptable. These factors are called the acoustic and language models, respectively.

We use the Japanese dictation toolkit[1], which includes the Julius decoder and acoustic/language models. Julius performs a two-pass (forward-backward) search using word-based forward bigrams and backward trigrams. The acoustic model was produced from the ASJ speech database, which contains approximately 20,000 sentences uttered by 132 speakers including both gender groups. A 16-mixture Gaussian distribution triphone Hidden Markov Model, in which states are clustered into 2,000 groups by a state-tying method, is used. We adapt the provided acoustic model by means of an MLLR-based unsupervised speaker adaptation method, for which in practice we use the HTK toolkit[2].

Existing methods to adapt language models can be classified into two fundamental categories. In the first category – the *integration* approach – general and topic-specific corpora are integrated to produce a topic-specific language model [1, 2]. Because the sizes of those corpora differ, N-gram statistics are calculated using the weighted average of the statistics extracted independently from those corpora. However, it is difficult to determine the optimal weight depending on the topic. In the second category – the *selection* approach – a topic-specific subset is selected from a general corpus and is used to produce a language model. This approach is effective if general corpora contain documents associated with target topics, but N-gram statis-

[1] http://winnie.kuis.kyoto-u.ac.jp/dictation/
[2] http://htk.eng.cam.ac.uk/

tics in those documents are overshadowed by other documents in resultant language models.

We followed the selection approach, because the 10M Web page corpus [3] containing mainly Japanese pages associated with various topics was publicly available. The quality of the selection approach depends on the method of selecting topic-specific subsets. An existing method [4] uses hypotheses in the initial speech recognition phase as queries to retrieve topic-specific documents from a general corpus. However, errors in the initial hypotheses have the potential to decrease the retrieval accuracy. Instead, we use textbooks related to target lectures as queries to improve the retrieval accuracy and consequently the quality of the language model adaptation.

2.4. Retrieval

Given transcribed passages and text queries, the basis of the retrieval module is the same as that for text retrieval. We use an existing probabilistic text retrieval method [5] to compute the relevance score between the query and each passage in the database. The relevance score for passage p is computed by Equation (2).

$$\sum_t f_{t,q} \cdot \frac{(K+1) \cdot f_{t,p}}{K \cdot \{(1-b) + \frac{dl_p}{b \cdot avgdl}\} + f_{t,p}} \cdot \log \frac{N - n_t + 0.5}{n_t + 0.5} \tag{2}$$

where $f_{t,q}$ and $f_{t,p}$ denote the frequency with which term t appears in query q and passage p, respectively. N and n_t denote the total number of passages in the database and the number of passages containing term t, respectively. dl_p denotes the length of passage p, and $avgdl$ denotes the average length of passages in the database. We empirically set $K = 2.0$ and $b = 0.8$, respectively. We use content words, such as nouns, extracted from transcribed passages as index terms, and perform word-based indexing. We use the ChaSen morphological analyzer[3] to extract content words. The same method is used to extract terms from queries.

However, retrieved passages are not disjoint, because top-ranked passages often overlap with one another in terms of the temporal axis. It is redundant simply to list the top-ranked retrieved passages as they are. Therefore, we reorganize those overlapped passages into a single passage. The relevance score for a group (a new passage) is computed by averaging the scores of all passages belonging to the group. New passages are sorted according to the degree of relevance and are presented to users as the final result.

3. Experimentation

3.1. Methodology

To evaluate the performance of our LOD system, we produced a test collection (as a benchmark data set) and performed experiments partially resembling a task performed in the TREC spoken document retrieval (SDR) track [6]. Five lecture programs on TV (each lecture was 45 minutes long), for which printed textbooks were also published, were videotaped in DV and were used as target lectures. Each lecture was manually transcribed and sentence boundaries with temporal information (i.e., correct speech units) were also identified manually. Each paragraph in the corresponding textbook was used as a query independently. For each query, a human assessor (a graduate student not an author of this paper) identified one or more relevant sentences in the human transcription.

[3] http://chasen.aist-nara.ac.jp/

Using our test collection, we evaluated the accuracy of speech recognition and passage retrieval. For the five lectures, our system used the sentence boundaries in human transcriptions to identify speech units, and performed speech recognition. We also used human transcriptions as perfect speech recognition results and investigated the extent to which speech recognition errors affect the retrieval accuracy. Our system retrieved top-ranked passages in response to each query. Note that the passages here are those grouped based on the temporal axis, which should not be confused with those obtained from the passage segmentation method.

3.2. Results

To evaluate the accuracy of speech recognition, we used the word error rate (WER), which is the ratio of the number of word errors (deletion, insertion, and substitution) to the total number of words. We also used test-set out-of-vocabulary rate (OOV) and trigram test-set perplexity (PP) to evaluate the extent to which our language model adapted to the target topics. We used human transcriptions as test set data. For example, OOV is the ratio of the number of word tokens not contained in the language model for speech recognition to the total number of word tokens in the transcription. Note that smaller values of OOV, PP, and WER are obtained with better methods.

The final outputs (i.e., retrieved passages) were evaluated based on recall and precision, averaged over all queries. Recall (R) is the ratio of the number of correct speech units retrieved by our system to the total number of correct speech units for the query in question. Precision (P) is the ratio of the number of correct speech units retrieved by our system to the total number of speech units retrieved by our system. To summarize recall and precision into a single measure, we used the F-measure (F).

Table 1 shows the accuracy of speech recognition (WER) and passage retrieval (R, P, and F), for each lecture. In this table, the columns "HUM" and "ASR" correspond to the results obtained with human transcriptions and automatic speech recognition, respectively. The column "+LA" denotes results for ASR combined with language model adaptation. The column "Topic" denotes topics for the five lectures.

To adapt language models, we used the textbook corresponding to a target lecture and searched the 10M Web page corpus for 2,000 relevant pages, which were used as a source corpus. In the case where the language model adaptation was not performed, all 10M Web pages were used as a source corpus. In either case, 20,000 high frequency words were selected from a source corpus to produce a word-based trigram language model. We used the ChaSen morphological analyzer to extract words (morphemes) from the source corpora, because Japanese sentences lack lexical segmentation.

In passage retrieval, we regarded the top N passages as the final outputs. In Table 1, the value of N ranges from 1 to 3. As the value of N increases, the recall improves, but potentially sacrificing precision.

3.3. Discussion

By comparing the results of ASR and +LA in Table 1, for some cases OOV and PP increased by adapting language models. However, WER decreased by adapting language models to target topics, irrespective of the lecture.

The values of OOV, PP, and WER for lecture #1 were generally smaller than those for the other lectures. One possible reason is that the lecturer of #1 spoke more fluently and made fewer erroneous utterances than the other lecturers.

Table 1: Experimental results for speech recognition and passage retrieval.

ID		#1			#2			#3			#4			#5		
Topic		Criminal law			Greek history			Domestic relations			Food and body			Solar system		
		HUM	ASR	+LA	HUM	ASR	+LA	HUM	ASR	+LA	HUM	ASR	+LA	HUM	ASR	+LA
OOV		—	.044	.020	—	.073	.082	—	.039	.049	—	.053	.041	—	.051	.053
PP		—	48.9	43.2	—	122	96.7	—	136	132	—	89.3	108	—	163	130
WER		—	.209	.133	—	.516	.423	—	.604	.543	—	.488	.416	—	.637	.482
	R	.695	.726	.732	.449	.258	.551	.632	.291	.505	.451	.220	.357	.296	.138	.241
N=1	P	.534	.548	.519	.377	.319	.386	.479	.362	.464	.414	.277	.337	.529	.358	.436
	F	.604	.624	.607	.410	.286	.454	.545	.322	.484	.432	.245	.347	.379	.200	.311
	R	.847	.858	.832	.663	.360	.674	.791	.464	.677	.655	.380	.463	.482	.228	.421
N=2	P	.441	.448	.458	.301	.211	.314	.372	.273	.353	.321	.247	.239	.462	.332	.409
	F	.580	.588	.591	.414	.266	.429	.506	.343	.464	.431	.300	.316	.472	.270	.415
	R	.879	.868	.874	.764	.438	.708	.827	.495	.718	.718	.392	.604	.637	.289	.527
N=3	P	.410	.405	.401	.269	.163	.252	.363	.215	.318	.297	.188	.235	.466	.280	.385
	F	.560	.553	.550	.398	.237	.372	.505	.300	.441	.420	.254	.338	.538	.285	.445

Recall, precision, and F-measure increased by adapting language models for lectures #2-5, irrespective of the number of passages retrieved. For lecture #1, the retrieval accuracy did not significantly differ whether or not we adapted the language model to the topic. One possible reason is that the WER of lecture #1 without language model adaptation (20.9%) was sufficiently small to obtain a retrieval accuracy comparable with the text retrieval [7]. The difference between HUM and ASR was marginal in terms of the retrieval accuracy. Therefore, the effect of the language model adaptation method was overshadowed in passage retrieval.

The retrieval accuracy for lecture #1 was higher than those for the other lectures. The story of lecture #1 was organized based primarily on the textbook, when compared with the other lectures. This suggests that the performance of our LOD system is dependent of the organization of target lectures.

Surprisingly, for lectures #1 and #2, recall, precision, and F-measure of +LA were better than those of HUM. This means that the automatic transcription was more effective than human transcription for passage retrieval purposes. One possible reason is the existence of Japanese variants (i.e., more than one spelling form corresponding to the same word), such as "*girisha*/*girishia* (Greece)". Because the language model was adapted by means of the textbook for a target lecture, the spelling in automatic transcriptions systematically resembled that in the queries extracted from the textbooks. In contrast, it is difficult to standardize the spelling in human transcriptions. Therefore, relevant passages in automatic transcriptions were more likely to be retrieved than passages in the human transcriptions.

We conclude that our language model adaptation method was effective for both speech recognition and passage retrieval.

4. Conclusion

Reflecting the rapid growth in the use of multimedia contents, information technologies appropriate to speech, image, and text processing are crucial. Of the various content types in this paper we focused on the video data of lectures with their organization based on textbooks, and proposed a system for cross-media on-demand lectures, in which users can formulate text queries using the textbook for a target lecture to retrieve specific video passages.

To retrieve video passages in response to text queries, we extract the audio track from a lecture video, generate a transcription by large vocabulary continuous speech recognition, and produce a text index, prior to system use.

We evaluated the performance of our system experimentally, for which five TV lecture programs in various topics were used. The experimental results showed that the accuracy of speech recognition varied depending on the topic and presentation style of the lecturers. However, the accuracy of speech recognition and passage retrieval was improved by adapting language models to the topic of the target lecture. Even if the word error rate was approximately 40%, the accuracy of retrieval was comparable with that obtained by human transcription.

5. References

[1] C. Auzanne, J. S. Garofolo, J. G. Fiscus, and W. M. Fisher, "Automatic language model adaptation for spoken document retrieval," in *Proceedings of RIAO 2000 Conference on Content-Based Multimedia Information Access*, 2000.

[2] K. Seymore and R. Rosenfeld, "Using story topics for language model adaptation," in *Proceedings of Eurospeech97*, 1997, pp. 1987–1990.

[3] K. Eguchi, K. Oyama, K. Kuriyama, and N. Kando, "The Web retrieval task and its evaluation in the third NTCIR workshop," in *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2002, pp. 375–376.

[4] L. Chen, J.-L. Gauvain, L. Lamel, G. Adda, and M. Adda, "Language model adaptation for broadcast news transcription," in *Proceedings of ISCA Workshop on Adaptation Methods For Speech Recognition*, 2001.

[5] S. Robertson and S. Walker, "Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval," in *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1994, pp. 232–241.

[6] J. S. Garofolo, E. M. Voorhees, V. M. Stanford, and K. S. Jones, "TREC-6 1997 spoken document retrieval track overview and results," in *Proceedings of the 6th Text REtrieval Conference*, 1997, pp. 83–91.

[7] P. Jourlin, S. E. Johnson, K. S. Jones, and P. C. Woodland, "Spoken document representations for probabilistic retrieval," *Speech Communication*, vol. 32, pp. 21–36, 2000.

Building a Test Collection for Speech-Driven Web Retrieval

Atsushi Fujii[†,†††], *Katunobu Itou*[††,†††]

[†] Institute of Library and Information Science, University of Tsukuba
[††] National Institute of Advanced Industrial Science and Technology
[†††] CREST, Japan Science and Technology Corporation

Abstract

This paper describes a test collection (benchmark data) for retrieval systems driven by spoken queries. This collection was produced in the subtask of the NTCIR-3 Web retrieval task, which was performed in a TREC-style evaluation workshop. The search topics and document collection for the Web retrieval task were used to produce spoken queries and language models for speech recognition, respectively. We used this collection to evaluate the performance of our retrieval system. Experimental results showed that (a) the use of target documents for language modeling and (b) enhancement of the vocabulary size in speech recognition were effective in improving the system performance.

1. Introduction

Automatic speech recognition, which decodes the human voice to generate transcriptions, has recently become a practical technology. A number of speech-based methods have been explored in the information retrieval (IR) community, which can be classified into the following two fundamental categories:

- spoken document retrieval, in which written queries are used to search speech (e.g., broadcast news audio) archives for relevant speech information,
- speech-driven retrieval, in which spoken queries are used to retrieve relevant textual information.

Initiated partially by the TREC-6 spoken document retrieval (SDR) track [1], various methods have been proposed for spoken document retrieval. However, a relatively small number of methods [2, 3, 4] have been explored for speech-driven text retrieval, although they are associated with numerous keyboardless retrieval applications, such as telephone-based retrieval, car navigation systems, and user-friendly interfaces.

In the NTCIR-3 workshop[1], which is a TREC-style evaluation workshop, the Web retrieval main task was organized to promote text-based Web IR [5]. Additionally, *optional* subtasks were also invited, in which a group of researchers voluntarily organized a subtask to promote their common research area. We made use of this opportunity and organized the "speech-driven retrieval" subtask to produce a reusable test collection for experimental of Web retrieval driven by spoken queries.

Section 2 describes the test collection produced for the speech-driven retrieval subtask. Section 3 describes our speech-driven retrieval system, and Section 4 elaborates on comparative experiments, in which we evaluated our system in terms of the speech recognition and retrieval accuracy.

[1] http://research.nii.ac.jp/ntcir/index-en.html

2. Test Collection for Speech-Driven IR

2.1. Overview

The purpose of the speech-driven retrieval subtask was to produce reusable and publicly available test collections and tools, so that researchers in the information retrieval and speech processing communities can develop technologies and share scientific knowledge concerning speech-driven information retrieval. In principle, as with conventional IR test collections, test collections for speech-driven retrieval are required to include test queries, target documents, and relevance assessment for each query. However, unlike conventional text-based IR, queries are speech data uttered by humans. In practice, because producing the entire collection is prohibitive, we produced speech data related to the Web retrieval main (text-based) task. Thus, target documents and relevance assessment in the main task can be used for the purpose of speech-driven retrieval.

However, participants for the NTCIR workshop are mainly researchers in the information retrieval and natural language processing communities, and are not necessarily experts in developing and operating speech recognition systems. Therefore, we also produced language models that can be used with an existing speech recognition engine (decoder), which helps researchers to perform experiments similar to those described in this paper. All above data are included in the NTCIR-3 Web retrieval test collection, which is publicly available.

2.2. Spoken Queries

For the Web retrieval main task, 105 search topics were produced, for each of which relevance assessment was performed with respect to two different document sets: the 10GB and 100GB collections. The 10GB and 100GB collections correspond approximately to 1M and 10M documents, respectively.

Each topic is in SGML-style form and consists of the topic ID (<NUM>), title of the topic (<TITLE>), description (<DESC>), narrative (<NARR>), list of synonyms related to the topic (<CONC>), sample of relevant documents (<RDOC>), and a brief profile of the user who produced the topic (<USER>). Figure 1 depicts a translation of an example topic. Although Japanese topics were used in the main task, English translations are also included in the Web retrieval collection mainly for publication purposes.

Participants in the main task were allowed to submit more than one retrieval result using one or more fields. However, participants were required to submit results obtained with the title and description fields independently. Titles are lists of keywords, and descriptions are phrases and sentences.

From the viewpoint of speech recognition, titles and descriptions can be used to evaluate *word* and *continuous* recognition methods, respectively. Because state-of-the-art speech

```
<TOPIC>
<NUM>0010</NUM>
<TITLE CASE="b">Aurora, conditions, obser-
vation</TITLE>
<DESC>For observation purposes, I want to
know the conditions that give rise to an
aurora</DESC>
<NARR><BACK>I want to observe an aurora
so I want to know the conditions neces-
sary for its occurrence and the mechanism
behind it.</BACK><RELE>Aurora observation
records, etc. list the place and time
so only documents that provide additional
information such as the weather and tem-
perature at the time of occurrence are
relevant. </RELE></NARR>
<CONC>Aurora, occurrence, conditions,
observation, mechanism</CONC>
<RDOC>NW003201843, NW001129327,
NW002699585</RDOC>
<USER>1st year Master's student, female,
2.5 years search experience</USER>
</TOPIC>
```

Figure 1: An example topic in the Web retrieval collection.

recognition is based on a continuous recognition framework, we used only the description field. For the first speech-driven retrieval subtask, we focused on *dictated* (*read*) speech, although our ultimate goal is to recognize *spontaneous* speech. We asked ten speakers (five adult males and five adult females) to dictate descriptions in the 105 topics. The ten speakers also dictated 50 sentences in the ATR phonetic-balanced sentence set as reference data, which can potentially be used for speaker adaptation. However, we did not use this additional data for the purpose of the experiments described in this paper. The above-mentioned spoken queries and sentences were recorded with the same close-talk microphone in a noiseless office. Speech waves were digitized at a 16KHz sampling frequency and a quantization of 16 bits. The resulting data are in the RIFF format.

2.3. Language Models

Unlike general-purpose speech recognition, in speech-driven text retrieval, users usually speak contents associated with a target collection, from which documents relevant to user needs are retrieved. In a stochastic speech recognition framework, the accuracy depends primarily on acoustic and language models. Whereas acoustic models are related to phonetic properties, language models, which represent linguistic contents to be spoken, are related to target collections. Therefore, it is feasible that language models have to be produced based on target collections. In summary, our belief is that by adapting a language model to a target IR collection, we can improve the speech recognition accuracy and, consequently, the retrieval accuracy. Motivated by this background, we used target documents for the main task to produce the language models. For this purpose, we used only the 100GB collection, because the 10GB collection is a subset of the 100GB collection.

We produced two language models of different vocabulary sizes so that the relation between the vocabulary size and system performance can be investigated. In practice, 20K and 60K high frequency words were used independently to produce word-based trigram models. We shall call these models "Web20K" and "Web60K", respectively. We used the ChaSen morphological analyzer[2] to extract words from the 100GB collection. To resolve the data sparseness problem, we used a back-off smoothing method, in which the Witten-Bell discounting method was used to compute back-off coefficients. In addition, through preliminary experiments, cut-off thresholds were empirically set at 20 and 10 for the Web20K and Web60K models, respectively. Trigrams whose frequency was above the threshold were used for language modeling. Language models and dictionaries are in the ARPA and HTK formats, respectively.

Table 1 shows the statistics related to word tokens/types in the 100GB collection and ten years of "Mainichi Shimbun" newspaper articles from 1991 to 2000. We shall use the term "word token" to refer to occurrences of words, and the term "word type" to refer to vocabulary items. The size of the 100G collection ("Web") is approximately 10 times that of 10 years of newspaper articles ("News"), which was one of the largest Japanese corpora available for the purpose of research and development in language modeling. This means that the Web is a vital, as yet untapped, corpus for language modeling.

Table 1: The statistics of corpora for language modeling.

	Web (100GB)	News (10 years)
# of Word types	2.57M	0.32M
# of Word tokens	2.44G	0.26G

3. System Description

3.1. Overview

Figure 2 depicts the overall design of our speech-driven text retrieval system, which consists of speech recognition and text retrieval modules. In the off-line process, a target IR collection is used to produce a language model, so that user speech related to the collection can be recognized with high accuracy. However, an acoustic model was produced independently of the target collection. In the on-line process, given an information request spoken by a user (i.e., a spoken query), the speech recognition module uses acoustic and language models to generate a transcription of the user speech. Then, the text retrieval module searches the target IR collection for documents relevant to the transcription, and outputs a specific number of top-ranked documents according to the degree of relevance in descending order. In the following two sections, we describe the speech recognition and text retrieval modules.

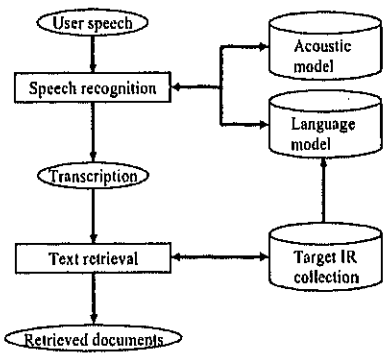

Figure 2: An overview of our speech-driven retrieval system.

[2]http://chasen.aist-nara.ac.jp/

3.2. Speech Recognition

We used the Japanese dictation toolkit[3] including the Julius decoder and acoustic/language models. Julius performs a two-pass (forward-backward) search using word-based forward bigrams and backward trigrams. The acoustic model was produced from the ASJ speech database, which contains 20,000 sentences uttered by 132 speakers including both genders. A 16-mixture Gaussian distribution triphone Hidden Markov Model, in which the states are clustered into 2,000 groups by a state-tying method, is used. The language model is a word-based trigram model produced from 60,000 high frequency words in 10 years of Mainichi Shimbun newspaper articles. This toolkit also includes development software so that acoustic and language models can be produced depending on the application. While we used the acoustic model provided in the toolkit, we used new language models produced from the 100GB collections, that is, the Web20K and Web60K models.

3.3. Text Retrieval

The retrieval module is based on an existing retrieval method [6], which computes the relevance score between the transcribed query and each document in the collection. The relevance score for document d is computed by Equation (1).

$$\sum_t f_{t,q} \cdot \frac{(K+1) \cdot f_{t,d}}{K \cdot \{(1-b) + \frac{dl_d}{b \cdot avgdl}\} + f_{t,d}} \cdot \log \frac{N - n_t + 0.5}{n_t + 0.5} \quad (1)$$

where $f_{t,q}$ and $f_{t,d}$ denote the frequency that term t appears in query q and document d, respectively; N and n_t denote the total number of documents in the collection and the number of documents containing term t, respectively; dl_d denotes the length of document d, and $avgdl$ denotes the average length of documents in the collection. We empirically set $K = 2.0$ and $b = 0.8$, respectively.

Given transcriptions (i.e., speech recognition results for spoken queries), the retrieval module searches a target IR collection for relevant documents and sorts them in descending order according to the score. We used content words, such as nouns, extracted from documents as index terms, and performed word-based indexing. We used the ChaSen morphological analyzer to extract content words. We also extracted terms from transcribed queries using the same method. We used words and bi-words (i.e., word-based bigrams) as index terms.

4. Experimentation

In the Web retrieval main task, different types of text retrieval were performed. The first type was "Topic Retrieval" resembling the TREC ad hoc retrieval. The second type was "Similarity Retrieval", in which documents were used as queries instead of keywords and phrases. The third type was "Target Retrieval", in which systems with a high precision were highly valued. This feature provided a salient contrast to the first two retrieval types, in which both recall and precision were used equally as evaluation measures.

Although the spoken queries produced can be used for the first and third task types, we focused solely on Topic Retrieval for the sake of simplicity. We used the 47 topics for the Topic Retrieval task to retrieve the 1,000 top documents, and we used the TREC evaluation software to calculate the mean average precision (MAP) values (i.e., non-interpolated average precision values, averaged over the 47 topics).

[3]http://winnie.kuis.kyoto-u.ac.jp/dictation/

Relevance assessment was performed based on four ranks of relevance: highly relevant, relevant, partially relevant and irrelevant. In addition, unlike conventional retrieval tasks, documents hyperlinked from retrieved documents were optionally used for relevance assessment. In summary, the following four assessment types were available to calculate the MAP values:

- (highly) relevant documents were regarded as correct answers, and hyperlink information was not used (RC),
- (highly) relevant documents were regarded as correct answers, and hyperlink information was used (RL),
- partially relevant documents were also regarded as correct answers, and hyperlink information was not used (PC),
- partially relevant documents were also regarded as correct answers, and hyperlink information was used (PL).

In the formal run for the main task, we submitted results obtained with different methods for the 10GB and 100GB collections. The best performance was obtained when we used description (<DESC>) fields as queries and we used a combination of words and bi-words as index terms.

The purpose of the experiments for speech-driven retrieval was two-fold. First, we investigated the extent to which a language model based on a target document collection contributes to an improvement in performance. Second, we investigated the impact of the vocabulary size for speech recognition on speech-driven retrieval. Therefore, we compared the performance of the following four retrieval methods:

- text-to-text retrieval, which used written queries, and can be seen as the perfect speech-driven text retrieval method ("Text"),
- speech-driven text retrieval, in which the Web60K model was used ("Web60K"),
- speech-driven text retrieval, in which a language model produced from 60,000 high frequency words in ten years of Mainichi Shimbun newspaper articles was used ("News60K"),
- speech-driven text retrieval, in which the Web20K model was used ("Web20K").

For text-to-text retrieval, we used descriptions (<DESC>) as queries, because the spoken queries used for speech-driven retrieval methods were descriptions dictated by speakers.

For speech-driven text retrieval methods, queries dictated by the ten speakers were used independently, and the final result was obtained by averaging the results for all speakers. Although the Julius decoder used in the speech recognition module generated more than one transcription candidate (hypothesis) for a single speech, we used only that with the greatest probability score. All language models were produced by means of the same softwares, but they were different in terms of the vocabulary size and the source documents. Table 2 shows the MAP values with respect to the four relevance assessment types and the word error rate in speech recognition, for different retrieval methods targeting the 10GB and 100GB collections.

As with existing experiments for speech recognition, the word error rate (WER) is the ratio between the number of word errors (i.e., deletion, insertion, and substitution) and the total number of words. In addition, we investigated the error rate with respect to query terms (i.e., keywords used for retrieval), which we shall call the term error rate (TER). Note that unlike MAP, smaller values of WER and TER are obtained with better methods. Table 2 also shows the test-set out-of-vocabulary

Table 2: Experimental results for different retrieval methods targeting the 10GB and 100GB collections (OOV: test-set out-of-vocabulary rate, WER: word error rate, TER: term error rate, MAP: mean average precision).

				MAP (10GB)				MAP (100GB)			
Method	OOV	WER	TER	RC	RL	PC	PL	RC	RL	PC	PL
Text	—	—	—	.1470	.1286	.1612	.1476	.0855	.0982	.1257	.1274
Web60K	.0073	.1311	.2162	.0966	.0916	.0973	.1013	.0542	.0628	.0766	.0809
News60K	.0157	.1806	.2991	.0701	.0681	.0790	.0779	.0341	.0404	.0503	.0535
Web20K	.0423	.1642	.2757	.0616	.0628	.0571	.0653	.0315	.0378	.0456	.0485

rate (OOV), which is the ratio of the number of words not included in the speech recognition dictionary to the total number of words in the spoken queries. Suggestions that can be derived from the results in Table 2 are as follows.

Looking at the WER and TER columns, News60K and Web20K were comparable in speech recognition performance, but Web60K outperformed in both cases. However, the difference between News60K and Web20K in OOV did not affect WER and TER. In addition, TER was greater than WER, because in computing TER, functional words, which are generally recognized with a high accuracy, were excluded.

Whereas the MAP values of News60K and Web20K were comparable, the MAP values of Web60K, which were approximately 60–70% of those obtained with Text, were greater than those for News60K and Web20K, irrespective of the relevance assessment type. These results were observed for both the 10GB and 100GB collections.

The only difference between News60K and Web60K was the source corpus for language modeling in speech recognition, and therefore we conclude that the use of target collections to produce a language model was effective for speech-driven retrieval. In addition, by comparing the MAP values of Web20K and Web60K, we conclude that the vocabulary size for speech recognition was also influential for the performance of speech-driven retrieval.

We analyzed speech recognition errors, focusing mainly on those attributed to the out-of-vocabulary problem. Table 3 shows the ratio of the number of out-of-vocabulary words to the total number of misrecognized words (or terms) in transcriptions. However, it should be noted that the actual ratio of errors due to the OOV problem can potentially be higher than those figures, because non-OOV words collocating with OOV words are often misrecognized. The remaining reasons for speech recognition errors are associated with insufficient N-gram statistics and the acoustic model. As predicted, the ratio of OOV words (terms) in Web20K was much higher than the ratios in Web60K and News60K. However, by comparing News60K and Web20K, WER and TER of News60K in Table 2 were higher than those of Web20K. This suggests that insufficient N-gram statistics were more problematic in News60K, compared to Web20K.

Table 3: The ratio of the number of OOV words/terms to the total number of misrecognized words/terms.

	Word	Term
Web60K	.0704	.1838
News60K	.0966	.2143
Web20K	.2855	.5049

5. Conclusion

In the NTCIR-3 Web retrieval task, we organized the speech-driven retrieval subtask and produced 105 spoken queries dictated by ten speakers. We also produced word-based trigram language models using approximately 10M documents in the 100GB collection used for the main task. We used those queries and language models to evaluate the performance of our speech-driven retrieval system. Experimental results showed that (a) the use of target documents for language modeling and (b) enhancement of the vocabulary size in speech recognition were effective in improving the system performance. Future work will include experiments using spontaneous spoken queries.

6. Acknowledgments

The authors thank the organizers of the NTCIR-3 Web retrieval task for their support to the speech-driven retrieval subtask.

7. References

[1] J. S. Garofolo, E. M. Voorhees, V. M. Stanford, and K. S. Jones, "TREC-6 1997 spoken document retrieval track overview and results," in *Proceedings of the 6th Text REtrieval Conference*, 1997, pp. 83–91.

[2] J. Barnett, S. Anderson, J. Broglio, M. Singh, R. Hudson, and S. W. Kuo, "Experiments in spoken queries for document retrieval," in *Proceedings of Eurospeech97*, 1997, pp. 1323–1326.

[3] F. Crestani, "Word recognition errors and relevance feedback in spoken query processing," in *Proceedings of the Fourth International Conference on Flexible Query Answering Systems*, 2000, pp. 267–281.

[4] A. Fujii, K. Itou, and T. Ishikawa, "A method for open-vocabulary speech-driven text retrieval," in *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, 2002, pp. 188–195.

[5] K. Eguchi, K. Oyama, K. Kuriyama, and N. Kando, "The Web retrieval task and its evaluation in the third NTCIR workshop," in *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2002, pp. 375–376.

[6] S. Robertson and S. Walker, "Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval," in *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1994, pp. 232–241.

Adapting Language Models for Frequent Fixed Phrases by Emphasizing N-gram Subsets

Tomoyosi AKIBA[†], *Katunobu ITOU*[†], *Atsushi FUJII*[‡]

[†] National Institute of Advanced Industrial Science and Technology (AIST)
1-1-1 Umezono, Tsukuba, 305-8568, JAPAN, E-mail: t-akiba@aist.go.jp
[‡] University of Tsukuba
1-2 Kasuga, Tsukuba, 305-8550, JAPAN

Abstract

In support of speech-driven question answering, we propose a method to construct N-gram language models for recognizing spoken questions with high accuracy. Question-answering systems receive queries that often consist of two parts: one conveys the query topic and the other is a fixed phrase used in query sentences. A language model constructed by using a target collection of QA, for example, newspaper articles, can model the former part, but cannot model the latter part appropriately. We tackle this problem as task adaptation from language models obtained from background corpora (e.g., newspaper articles) to the fixed phrases, and propose a method that does not use the task-specific corpus, which is often difficult to obtain, but instead uses only manually listed fixed phrases. The method emphasizes a subset of N-grams obtained from a background corpus that corresponds to fixed phrases specified by the list. Theoretically, this method can be regarded as maximizing a posteriori probability (MAP) estimation using the subset of the N-grams as a posteriori distribution. Some experiments show the effectiveness of our method.

1. Introduction

Question answering (QA) was first evaluated largely at TREC-8[11]. The goal in the QA task is to retrieve small snippets of text that contain the actual answer to a question rather than the document lists traditionally returned by text retrieval systems. We are trying to extend question-answering systems as traditional text retrieval systems[3] that accept spoken queries. In this paper, we address issues related to language modeling for the speech recognition subsystem of speech-driven question-answering systems.

Question-answering systems receive queries that often consist of a part that conveys various query contents about, for example, newspaper articles, and a part that represents a fixed phrase for query sentences. For example, the following query may be submitted.

> *seN / kyu-/ hyaku/ nana / ju- / roku / neN / ni / kasei / ni / naN / chakuriku / shita / taNsaki / wa / naN / to / yu- / namae /desu / ka*
> (What was the name of the spacecraft that landed safely on Mars in 1976?)

The first half of the query, i.e., "seN kyu- hyaku nana ju- roku neN ni kasei ni naN chakuriku shita taNsaki wa (the spacecraft that landed safely on Mars in 1976)", conveys the topic of the retrieval, and is best dealt with by using an N-gram model trained with the target documents of QA systems. In this paper, newspaper articles are used for the target documents[4]. On the other hand, the latter half of the query, i.e., "naN to yu- namae desu ka (What was the name?)", is a fixed phrase typically used in interrogative questions, but is not very frequent in newspaper articles. Thus, we require language models that can deal with both types of fragments.

Note that recognizing the fixed phrases with high accuracy is crucial to success in question answering, because they convey clues to determine the query type[6]. For example, a fixed phrase might indicate that the answer should be a name of some object as in the last example, while another might indicate that the answer should be a date of some event (e.g., in English, "On what date was...").

There has been work on language model adaptation in which language models for a specific task were constructed from both a large general-purpose corpus and a relatively small task-specific corpus. Using this approach, we can construct a language model for question answering from both a large number of generic newspaper articles and a small number of query sentences for QA.

One issue that should be considered when using this approach is how the task-specific corpus should be acquired. If the corpus does not exist already, it must be collected somehow or other, and collecting a new corpus directly from practical use is always expensive, even if the resulting corpus is small. Alternative methods have been proposed to obtain a considerable amount of task-specific corpus indirectly, including such methods as: automatically generating sentences from a hand-made task-specific grammar[5]; incorporating a task-specific grammar-based model into the background N-gram[1]; and utilizing the results of speech recognition using a general-purpose language model[8, 9].

In our case, the number of the fixed phrases used in QA is small enough for all the patterns to be enumerated by hand. Thus we can inexpensively prepare a list of phrases instead of collecting a corpus of query sentences. In this paper, we propose a method of constructing language models for question answering from a target collection (e.g., newspaper articles) and a list of the fixed phrases typically used in interrogative questions. The method emphasizes N-gram subsets corresponding to the fixed phrases and can be considered as a variant of a maximum a posteriori probability (MAP) estimation using the N-gram subsets of a background corpus as an a posteriori distribution.

2. The Method

Figure 1 illustrates our proposed method of adapting a language model for fixed phrases. The list of fixed phrases is used to select the subset of N-grams related to the phrases. Then, adding the subset to the original N-grams produces the adapted model.

The second and third authors are also members of CREST, Japan Science and Technology Corporation.

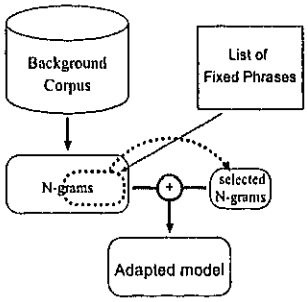

Figure 1: Language model adaptation using a list of fixed phrases

2.1. Language model adaptation for fixed phrases

Let S be a set of sentences. Let S_{FP} be a subset of S that consists only of sentences that have the fixed phrases specified by the list. Let P be a language model for generating sentences $s(\in S)$ obtained from a general-purpose background corpus. The aim of the language model adaptation for the fixed phrases is to obtain the adapted language model P', which gives relatively high probabilities to the sentences $\hat{s} \in S_{FP}$ but preserves the order relations on the sentences $s \in S - S_{FP}$ as much as possible.

The generative probability that the sentence $\hat{s}$ includes a fixed phrase $\hat{w}_p \cdots \hat{w}_q$ is:

$$\begin{aligned} P(\hat{s}) &= \prod_{i=1}^{n} P(w_i|w_1 \cdots w_{i-1}) \\ &\approx P(w_1)P(w_2|w_1)\cdots P(w_i|w_{i-N+1}^{i-1})\cdots \\ &\quad \cdots P(\hat{w}_p|w_{p-N+1}^{p-1})P(\hat{w}_{p+1}|w_{p-N+2}^{p-1}\hat{w}_{p-1})\cdots \\ &\quad \cdots P(\hat{w}_{p+N-2}|w_{p-1}\hat{w}_p^{p+N-3})\cdots \\ &\quad \cdots P(\hat{w}_{p+N-1}|\hat{w}_p^{p+N-2})\cdots \\ &\quad \cdots P(\hat{w}_q|\hat{w}_{q-N+1}^{q-1})P(w_{q+1}|\hat{w}_{q-N+2}^{q})\cdots \\ &\quad \cdots P(w_m|w_{m-N+1}^{m-1}) \end{aligned}$$

The following components of the above equation are important in obtaining P':

$$P(\hat{w}_p|w_{p-N+1}^{p-1})\cdots P(\hat{w}_{p+N-2}|w_{p-1}\hat{w}_p^{p+N-3}), \quad \text{(a)}$$

$$P(\hat{w}_{p+N-1}|\hat{w}_p^{p+N-2})\cdots P(\hat{w}_q|\hat{w}_{q-N+1}^{q-1}). \quad \text{(b)}$$

The component (a) corresponds to the generative probabilities of the prefix words of the fixed phrases, each of which, in its condition part, has one or more words other than those that consist of the fixed phrases. The component (b) corresponds to the generative probabilities of the intermediate words of the fixed phrases, each of which has only the words of the fixed phrases in its condition part.

The adapted model P' is calculated using the following two steps.

i. Revise the maximum likelihood estimates of P:

$$P_{ML(1)}(w_i)P_{ML(2)}(w_i|w_i-1)\cdots P_{ML(N)}(w_i|w_{i-N+1}^{i-1}),$$

which are calculated for each length $n(1 \le n \le N)$.

ii. Apply back-off smoothing to integrate the revised ML estimates $P'_{ML(n)}(w_i|w_{i-n+1}^{i-1})(1 \le n \le N)$.

The proposed method emphasizes only the carefully selected $P_{ML(n)}$s that are meaningful for following back-off smoothing calculation, to make the produced model harmless to the other generative probabilities assigned to the sentences that do not have the fixed phrases.

2.2. Revision of the maximum likelihood estimate

For all lengths $n(1 \le n \le N)$, the maximum likelihood estimates $P_{ML(n)}(w_i|w_{i-n+1}^{i-1})$ of the N-gram probability P obtained from the background corpus are revised to P'_{ML} by the following procedure.

(1). *If the postfix $w_{i-k+1}\cdots w_i(1 \le k < n)$ of the word sequence $w_{i-n+1}\cdots w_i$ is equal to the prefix $\hat{w}_p \cdots \hat{w}_{p+k-1}$ of one of the fixed phrases $\hat{w}_p \cdots \hat{w}_q$, then emphasize the P_{ML} as follows:*

$$\begin{aligned} &P'_{ML(n)}(\hat{w}_{p+k-1}|w_{p-n+k}^{p-1}\hat{w}_p^{p+k-2}) = \\ &\quad \beta_n(w_{p-n+k}^{p-1}\hat{w}_p^{p+k-2}) \cdot \\ &\quad \gamma P_{ML(n)}(\hat{w}_{p+k-1}|w_{p-n+k}^{p-1}\hat{w}_p^{p+k-2}) \end{aligned}$$

Otherwise, go to step (2).

For example, for tri-grams, for all context word sequences $w_{p-2}w_{p-1}$, we have:

$$\begin{aligned} &P'_{ML(3)}(\hat{w}_{p+1}|w_{p-1}\hat{w}_p) = \\ &\quad \beta_3(w_{p-1}\hat{w}_p)\cdot\gamma P_{ML(3)}(\hat{w}_{p+1}|w_{p-1}\hat{w}_p) \\ &P'_{ML(2)}(\hat{w}_{p+1}|\hat{w}_p) = \\ &\quad \beta_2(\hat{w}_p)\cdot\gamma P_{ML(2)}(\hat{w}_{p+1}|\hat{w}_p) \\ &P'_{ML(3)}(\hat{w}_p|w_{p-2}w_{p-1}) = \\ &\quad \beta_3(w_{p-2}w_{p-1})\cdot\gamma P_{ML(3)}(\hat{w}_p|w_{p-2}w_{p-1}) \\ &P'_{ML(2)}(\hat{w}_p|w_{p-1}) = \\ &\quad \beta_2(w_{p-1})\cdot\gamma P_{ML(2)}(\hat{w}_p|w_{p-1}) \\ &P'_{ML(1)}(\hat{w}_p) = \beta_1(\epsilon)\cdot\gamma P_{ML(1)}(\hat{w}_p) \end{aligned}$$

(2). *If the word sequence $w_{i-n+1}\cdots w_i$ is equal to the subsequence $\hat{w}_{i-n+1}\cdots\hat{w}_i$ of one of the fixed phrases $\hat{w}_p\cdots\hat{w}_q$ then emphasize only the longest N-gram probability $P_{ML(N)}$ as follows:*

$$\begin{aligned} &P'_{ML(N)}(\hat{w}_i|\hat{w}_{i-N+1}^{i-1}) = \\ &\quad \beta_N(\hat{w}_{i-N+1}^{i-1})\cdot\gamma P_{ML(N)}(\hat{w}_i|\hat{w}_{i-N+1}^{i-1}) \end{aligned}$$

Otherwise, go to step (3).

For example, for tri-grams, only the tri-gram probability should be emphasized:

$$\begin{aligned} &P'_{ML(3)}(\hat{w}_i|\hat{w}_{i-2}\hat{w}_{i-1}) = \\ &\quad \beta_3(\hat{w}_{i-2}\hat{w}_{i-1})\cdot\gamma P_{ML(3)}(\hat{w}_i|\hat{w}_{i-2}\hat{w}_{i-1}) \end{aligned}$$

(3). *For all n ($1 \le n \le N$), the revised probability is:*

$$\begin{aligned} &P'_{ML(n)}(w_i|w_{i-n+1}^{i-1}) = \\ &\quad \beta_n(w_{i-n+1}^{i-1})\cdot P_{ML(n)}(w_i|w_{i-n+1}^{i-1}) \end{aligned}$$

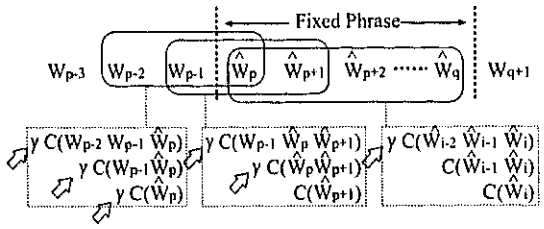

Figure 2: Emphasizing N-gram Counts (for tri-grams)

where $\gamma(\geq 1)$ is a multiplier that emphasizes the selected N-grams, and $\beta_1(\epsilon)\cdots\beta_N(w_{i-N+1}^{i-1})$ are normalized coefficients that make the probabilities sum to one.

This can be seen as the task adaptation process by maximum a posteriori probability (MAP) estimation[2] using the N-gram subsets corresponding to the fixed phrases as task specific data for adaptation, because P'_{ML} is equivalent to the maximum likelihood estimate calculated from the N-gram counts C'_n of each length $n(1 \leq n \leq N)$ obtained by emphasizing the selected subset of the original N-gram counts C, as shown in Fig. 2.

$$P'_{ML(n)}(w_i|w_{i-n+1}^{i-1}) = \frac{C'_n(w_{i-n+1}^{i})}{\sum_{w_i} C'_n(w_{i-n+1}^{i})}$$

2.3. Back-off smoothing

Back-off smoothing integrates the revised ML probabilities $P'_{ML(n)}$ of each length n to produce the final adapted language model. Any back-off smoothing method can be applied, except that the discount coefficient should be calculated using the a priori knowledge of the adaptation, i.e., the N-gram counts obtained from the background corpus.

For example, for Witten-Bell smoothing [10], the following discount coefficient $d'_{w_{i-n+1}^{i}}$ should be used for the proposed method.

$$d'_{w_{i-n+1}^{i}} P'_{ML(n)}(w_i|w_{i-n+1}^{i-1}) = \frac{C_n(w_{i-n+1}^{i})}{\{\sum_{w_i} C_n(w_{i-n+1}^{i})\} + r(w_{i-n+1}^{i-1})}$$

where r is the number of different words appearing after the word context w_{i-n+1}^{i-1} in the background corpus.

3. Experimental Results

We extracted N-gram counts of 20,000 words that were obtained from newspaper articles collected over 111 months. As task-specific training data, we developed a word network for the Japanese fixed phrases used for question answering. From the network, we extracted a list of all the 172 fixed phrases that were acceptable to the network. Then we compared several adaptation methods including that mentioned in this paper. We applied Witten-Bell discounting[10] for all methods.

We first made the N-gram model from only the newspaper articles as the baseline (referred to as the *BASE* model). As a conventional MAP adaptation method[2], we mixed two sets of N-gram counts obtained from newspaper articles and the list of fixed phrases (magnified by w), and obtained the adapted model referred to as *MIX*. As the method proposed in this paper, we magnified N-gram counts corresponding to the fixed phrases in the N-gram of newspaper articles (by γ), and obtained the adapted model referred to as *EMP*. Finally, using the method that we had previously proposed [1], we integrated the N-gram of newspaper articles and the word network for fixed phrases (magnified by γ), and obtained the adapted model referred to as *NET*.

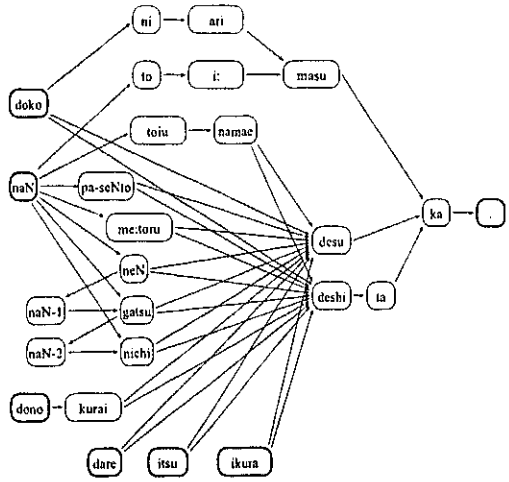

Figure 3: A word network for the Japanese fixed phrases frequently used in queries for QA

We prepared 100 sentences from the newspaper articles (referred to as *NP*) and 50 query sentences for the QA system (referred to as *QA*), and these were recorded for four speakers (two men and two women). Though the word network was relatively small and had only 33 nodes (31 words), 36 of the 50 queries had the fixed phrases characterized in the network.

We used an existing N-gram decoder [7] for the recognition experiments. The language model weight and the insertion penalty were set to the best values for the newspaper (*BASE*) model. The results are shown in Fig. 4 and Fig. 5.

Figure 4 shows the relations between word error rate (WER) and the parameter (w or γ) with respect to both the target *QA* and *NP*. The best results with respect to *QA* of *BASE*, *MIX*, *EMP* and *NET* are obtained by adjusting the parameter to 16.9, 15.4, 13.8 and 14.7, respectively. The proposed model *EMP* outperformed the other models, while it did not worsen the WER for the sentences that did not have the fixed phrases (*NP*).

Figure 5 shows the difference between WERs for the first (referred to as *FH*) and latter (referred to as *LH*) half of the interrogative sentences (*QA*). We divided each sentence of *QA* into first and latter half by using a Japanese WH-word as the boundary (the latter half included the WH-word), and investigated the WERs of both halves separately. Note that the latter halves roughly correspond to the fixed phrases. It indicated that the proposed method (*EMP*) best reduced the WER corresponding to the fixed phrases (*LH*), while it did not worsen the WER for the other part of the input sentences (*FH*).

4. Conclusion

We have proposed methods for language model adaptation that enable recognition of spoken queries submitted to QA systems with high accuracy. The method does not require a task-specific corpus but, instead, uses a list of fixed phrases enumerated by hand. Our experiments showed that the method outperformed a conventional language model adaptation method in terms of the recognition accuracy. The proposed methods can be used for other task-adaptation problems in language modeling where the variation in expressions to be adapted is relatively small allowing for these expressions to be enumerated by hand without collecting a new text corpus.

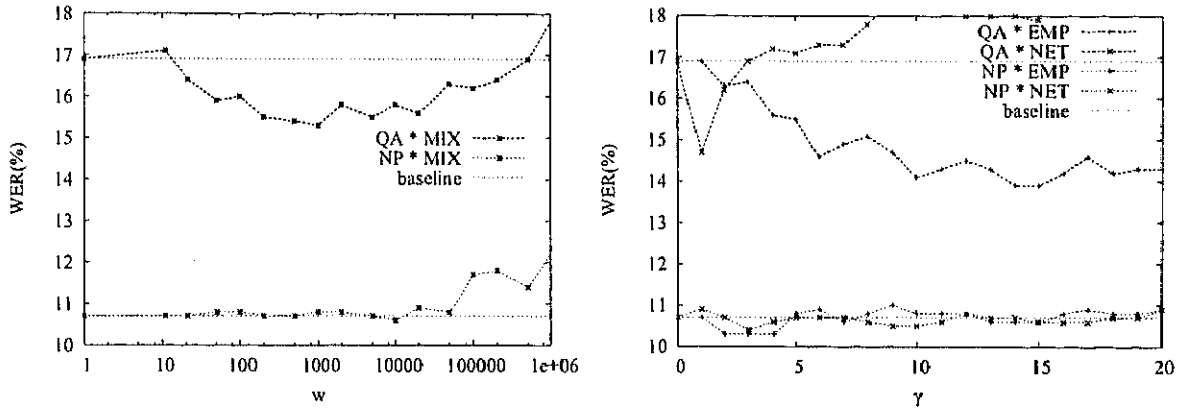

Figure 4: The relation between word error rate and the parameter.

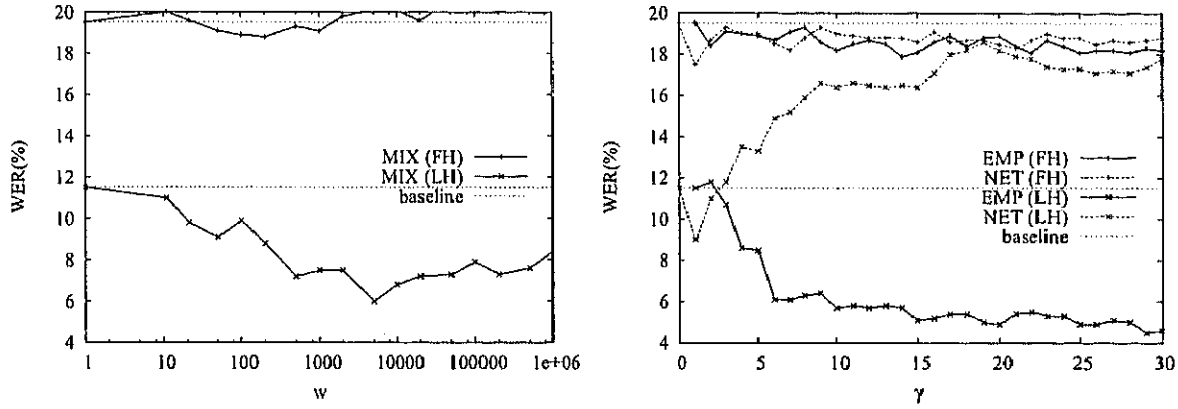

Figure 5: Word error rates for first and latter halves of sentences (*QA*).

5. References

[1] T. Akiba, K. Itou, A. Fujii, and T. Ishikawa. Selective back-off smoothing for incorporating grammatical constraints into the n-gram language model. In *Proceedings of International Conference on Spoken Language Processing*, volume 2, pages 881–884, Denver, Colorado, Sept. 2002.

[2] M. Federico. Bayesian estimation methods for n-gram language model adaptation. In *Proceedings of International Conference on Spoken Language Processing*, pages 240–243, Philadelphia, USA, 1996.

[3] A. Fujii, K. Itou, and T. Ishikawa. Speech-driven text retrieval: Using target IR collections for statistical language model adaptation in speech recognition. In A. R. Coden, E. W. Brown, and S. Srinivasan, editors, *Information Retrieval Techniques for Speech Applications (LNCS 2273)*, pages 94–104. Springer, 2002.

[4] J. Fukumoto, T. Kato, and F. Masui. Question answering challenge (QAC-1) question answering evaluation at NTCIR workshop 3. In *Working Notes of the Third NTCIR Workshop Meeting*, pages 1–6, Tokyo, Japan, Oct. 2002.

[5] L. Galescu, E. Ringger, and J. Allen. Rapid language model development for new task domains. In *Proceedings of International Conference on Language Resources and Evaluation*, pages 807–812, Granada, Spain, May 1998.

[6] A. Ittycheriah, M. Franz, W.-J. Zhu, and A. Ratnaparkhi. IBM's statistical question answering system. In *Proceedings of the 9th Text Retrieval Conference*, pages 229–234, Maryland, 2000.

[7] A. Lee, T. Kawahara, and K. Shikano. Julius — an open source real-time large vocabulary recognition engine. In *Proceedings of European Conference on Speech Communication and Technology*, pages 1691–1694, Aalborg, Denmark, Sept.

[8] M. Mahajan, D. Beeferman, and X. D. Huang. Improved topic-dependent language modeling using information retrieval techniques. In *Proceedings of International Conference on Acoustics Speech and Signal Processing*, Phoenix, Arizona, March 1999.

[9] T. Niesler and D. Willett. Unsupervised language model adaptation for lecture speech transcription. In *Proceedings of International Conference on Spoken Language Processing*, volume 2, pages 1413–1416, Denver, Colorado, Sept. 2002.

[10] P. Placeway, R. Schwartz, P. Fung, and L. Nguyen. The estimation of powerful language models from small and large corpora. In *Proceedings of International Conference on Acoustics Speech and Signal Processing*, volume 2, pages 33–36, Minneapolis, USA, April 1993.

[11] E. Voorhees and D. Tice. The TREC-8 question answering track evaluation. In *Proceedings of the 8th Text Retrieval Conference*, pages 83–106, Gaithersburg, Maryland, 1999.

LODEM: A multilingual lecture-on-demand system

Atsushi Fujii[†,†††] Katunobu Itou[††,†††] Tetsuya Ishikawa[†]
[†] Institute of Library and Information Science
University of Tsukuba
1-2 Kasuga, Tsukuba, 305-8550, Japan
[††] National Institute of Advanced Industrial Science and Technology
1-1-1 Chuuou Daini Umezono, Tsukuba, 305-8568, Japan
[†††] CREST, Japan Science and Technology Corporation
fujii@slis.tsukuba.ac.jp

Abstract

We propose a multilingual lecture-on-demand system, which searches lecture videos for segments relevant to user information needs across languages. We utilize the benefits of textbooks and audio/video data corresponding to a single lecture. We extract the audio track from a target lecture video, generate a transcription by large vocabulary continuous speech recognition, and produce a textual index. Users can view specific video segments by selecting paragraphs in the textbook for the target lecture, machined translated into the user language. Experimental results showed that by adapting speech recognition to the lecture topic, the recognition accuracy increased and the retrieval accuracy was comparable with that obtained by human transcriptions. Our system is implemented as a client-server system over the Web to facilitate e-education.

1 Introduction

Reflecting the rapid growth in the utilization of multimedia contents available via the World Wide Web and CD-ROMs, information technologies across speech, image, and text processing have of late become crucial. Among various types of contents, lectures (audio/video) are typical and valuable multimedia contents, in which oral presentations are usually organized based on textual materials, such as resumes, slides, and textbooks. In lecture videos, image information, such as flip charts, is often additionally used. Thus, a single lecture consists of different types of compatible multimedia contents.

However, since a single lecture often includes multiple stories and takes long time, it is useful to selectively obtain specific segments (passages) so that audience can satisfy their information needs with a minimal cost. In addition, since each lecture is usually provided in a single language, it is feasible that users are interested in retrieving and viewing foreign lecture videos by their native languages.

To resolve this problem, in this paper we propose a lecture-on-demand system, "LODEM" (Lecture-On-DEMand), which retrieves relevant video/audio passages in response to user queries. For this purpose, we utilize the benefits of different media types to improve retrieval performance.

On the one hand, textual contents are advantageous in the sense that users can view/scan the entire contents quickly and easily identify relevant passages using layout information (e.g., text structures based on sections and paragraphs). In other words, textual contents can be used for random-access purposes.

On the other hand, speech contents are primarily used for sequential-access purposes. It is difficult to identify relevant passages unless target video/audio data include additional annotations, such as indexes. Even if target data are indexed, users are not necessarily able to come up with effective queries. To resolve this problem, textbooks are desirable materials, from which users can extract effective keywords and phrases.

However, while textbooks are usually concise, speeches are relatively redundant and thus are easy to understand more than textbooks, specifically in the case where additional image information is provided.

In view of the above discussion, we model LODEM system as follows. A user, who browses the textbook for a target lecture machine translated into the user language, selects text passages (i.e., paragraphs) relevant to their information needs from the translated textbook. Then, source passages (from which translations of selected passages were generated) are used to formulate a textual query. In other words, queries can be formulated even if users cannot come up with effective keywords and phrases. Users can also submit additional keywords as queries, if necessary. Video passages relevant to a given query are retrieved and presented to the user, in which transcriptions of speech data are machine translated into the user language.

To retrieve video passages in response to textual queries, we extract the audio track from a lecture video, generate a transcription by means of large vocabulary continuous speech recognition, and produce a textual in-

dex, prior to the system usage.

Our on-demand system should not be confused with video-on-demand (VOD) systems, which search video archives for specific videos in response to user requests. While in VOD systems, minimal unit for retrieval is the entire program, in our system, retrieval units are passages smaller than the entire program.

2 System Description

2.1 Overview

Figure 1 depicts the overall design of our lecture-on-demand system, in which left/right-hand regions correspond to the on-line and off-line processes, respectively. Although we implemented an English-to-Japanese system, our methodology is fundamentally language-independent. For research and development purposes, we tentatively target lecture programs on TV for which textbooks are published. We explain the basis of our system using this figure.

In the off-line process, given the video data of a target lecture, the audio data are extracted and segmented into more than one passage. Then, speech recognition transcribes each passage. Finally, the transcribed passages are indexed as performed in conventional text retrieval systems, so that each passage can be retrieved efficiently in response to textual queries.

To adapt speech recognition to a specific lecturer, we perform unsupervised speaker adaptation using an initial speech recognition result (i.e., a transcription). To adapt speech recognition to a specific topic, we perform language model adaptation, for which we search a general corpus for documents relevant to the textbook related to the target lecture. Then, the retrieved documents (i.e., a topic-specific corpus) are used to produce a word-based N-gram language model. We also perform image analysis to extract textual contents (e.g., keywords and phrases) in flip charts. These contents are also used later to improve our language model.

In the on-line process, a user can view specific video passages by selecting paragraphs in a translation of the textbook. In video passages, automatic transcriptions are machine translated into the user language. For the purpose of machine translation, we use the PC-transer MT system[1], which uses a Japanese/English bilingual dictionary consisting of approximately 1M entries for 19 technical fields.

The current implementation is based on a client-server system over the Web. While both the off-line and on-line processes are performed on servers, users can utilize our system by means of Web browsers on their own PCs. Figure 2 depicts the interface of LODEM, where a lecture associated with "nonlinear multivariate analysis" is given. In Figure 2, an electronic version of a translated textbook is displayed in the left-hand side, and a lecture video is played in the right-hand side.

[1] http://www.nova.co.jp/english/

Users can copy paragraphs in the textbook and paste them into the query input box, which are not depicted in Figure 2. Alternatively, users can submit keywords to retrieve paragraphs in the textbook. In the bottom of Figure 2, top-ranked transcribed Japanese passages and their English translations are listed according to the degree of relevance. Users can select (click) one of transcriptions to play the corresponding video passage.

Ideally, full translations are more useful to improve the browsing efficiency. However, since machine translation for misrecognized transcriptions is problematic and speech translation still remains a difficult problem, we perform morphological analysis to extract content words from transcriptions, and translate only those words. Thus, our system is practical for users who can understand Japanese to a certain extent, with translations of textbooks and partial transcriptions.

2.2 Passage Segmentation

The basis of the passage segmentation module is to divide the entire video data for a single lecture into more than one minimal unit to be retrieved. We shall call those units passages. For this purpose, both speech and image data can be promising clues. However, in lecture TV programs, it is often the case that a lecturer sitting still is mainly focused and a small number of flip charts are occasionally used. In such cases, image data is less informative. Thus, we tentatively use only speech data for the passage segmentation process.

However, unlike the case where a target speech (e.g., a news program) consists of multiple different topics [10], topic segmentation for lectures is relatively difficult, because a single lecture consists of sub-topics closely related to one another, and thus topic boundaries are not necessarily clear. Additionally, for LODEM, segmentation can potentially vary depending on the user query. Thus, it is difficult to predetermine a desirable segmentation in the off-line process.

In view of the above problems, we first extract the audio track from a target video, and perform a simple pause-based segmentation method to obtain minimal speech units, such as sentences and long phrases. In other words, speech units are continuous audio segments that do not include pauses longer than a certain threshold. Finally, we generate variable-length passages from one or more speech units. To put it more precisely, we combine N speech units into a single passage, with N ranging from 1 to 5 in the current implementation.

Figure 3 shows an example of variable-length passages, in which any sequences of speech units that are 1-5 in length are identified as passages.

2.3 Speech Recognition

The speech recognition module generates word sequence W, given phone sequence X. In a stochastic speech recognition framework, the task is to select the W maximizing $P(W|X)$, which is transformed as in

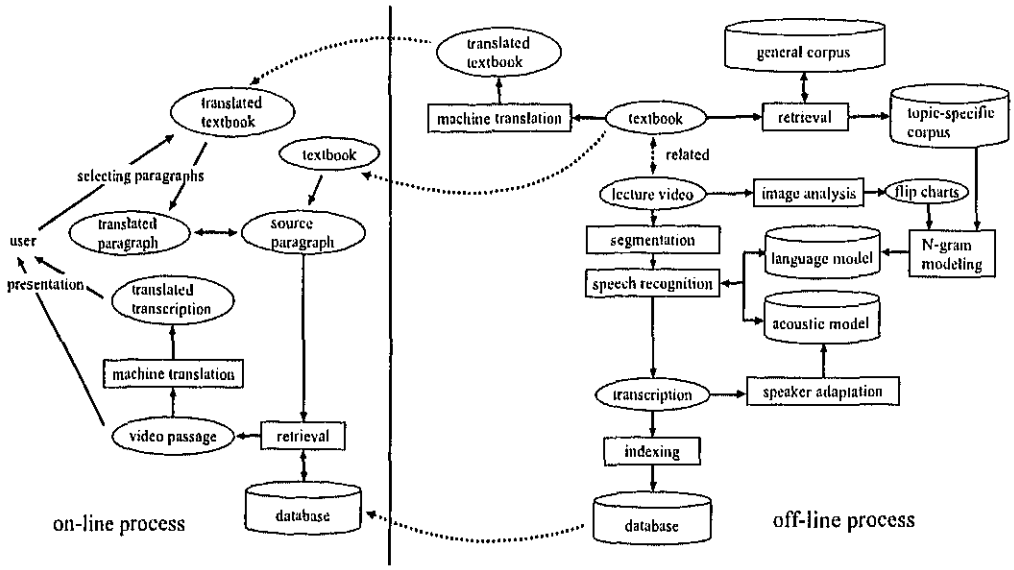

Figure 1. The overall design of our lecture-on-demand system (LODEM).

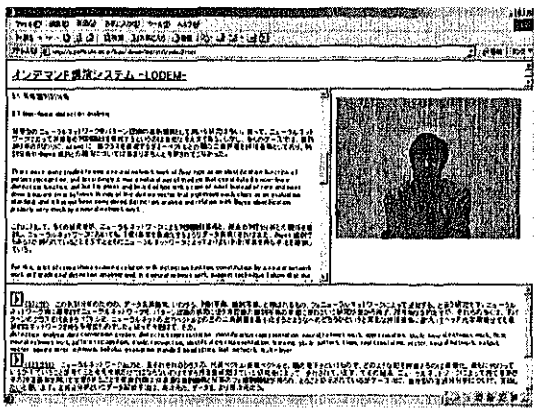

Figure 2. The interface of LODEM.

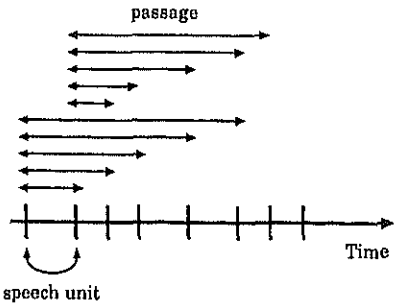

Figure 3. An example of the passage segmentation process.

Equation (1) through the Bayesian theorem.

$$\arg\max_W P(W|X) = \arg\max_W P(X|W) \cdot P(W) \quad (1)$$

$P(X|W)$ models the probability that word sequence W is transformed into phone sequence X, and $P(W)$ models the probability that W is linguistically acceptable. These are acoustic and language models, respectively.

We use the Japanese dictation toolkit[2] including the Julius decoder and acoustic/language models. Julius performs a two-pass (forward-backward) search using word-based forward bigrams and backward trigrams.

2.4 Retrieval

We use an existing probabilistic text retrieval method [8] to compute the relevance score between the textual query and each passage in the database. The relevance score for passage p is computed by Equation (2).

$$\sum_t f_{t,q} \cdot \frac{(K+1) \cdot f_{t,p}}{K \cdot \{(1-b) + \frac{dl_p}{b \cdot avgdl}\} + f_{t,p}} \cdot \log \frac{N - n_t + 0.5}{n_t + 0.5} \quad (2)$$

Here, $f_{t,q}$ and $f_{t,p}$ denote the frequency that term t appears in query q and passage p, respectively. N and n_t denote the total number of passages in the database and the number of passages containing term t, respectively. dl_p denotes the length of passage p, and $avgdl$ denotes the average length of passages in the database. We empirically set $K = 2.0$ and $b = 0.8$, respectively.

It should be noted that in Equation (2), the score is normalized with the length of each passage. Thus, longer passages, which potentially include more index terms, are not necessarily assigned with a higher score. This property is important, because variable-length passages are considerably different in terms of length.

[2] http://winnie.kuis.kyoto-u.ac.jp/dictation/

We use content words, such as nouns, extracted from transcribed passages as index terms, and perform word-based indexing. We use the ChaSen morphological analyzer[3] to extract content words. We also extract terms from queries using the same method.

However, retrieved passages are not disjoint, because top-ranked passages often overlap with one another in terms of the temporal axis. It is redundant to simply list the top-ranked retrieved passages as they are. Thus, we reorganize those overlapped passages into a single passage. In Figure 4, which uses the same basic notation as Figure 3, illustrates an example scenario. In this figure, top-ranked passages are organized into three groups.

The relevance score for a group (a new passage) is computed by averaging scores for all passages belonging to the group. New passages are sorted according to the degree of relevance and are presented to users as the final result.

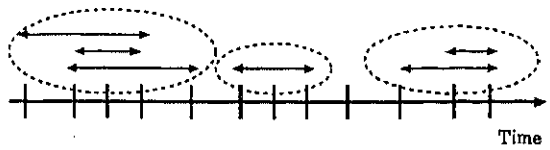

Figure 4. An example of grouping retrieved passages.

3 Experimentation

3.1 Methodology

Since LODEM consists of a number of different components, it is difficult evaluate its performance on the basis of a single evaluation method. In this paper, we focus on the performance of LODEM in terms of speech recognition and passage retrieval. For this, we produced a test collection and performed experiments partially resembling one performed in the TREC spoken document retrieval (SDR) track [2].

Two lecture programs on TV, for which printed textbooks were also published, were videotaped in DV and were used as target lectures. Both lectures were manually transcribed and sentence boundaries with temporal information (i.e., correct speech units) were also manually identified. The textbooks for the two target lectures were read by an OCR software and were manually revised. The accuracy of the OCR software was roughly 97% on a word-by-word basis. For both lectures, each paragraph in the corresponding textbook was used as a query independently. For each query, a human assessor (a graduate student other than authors of this paper) identified one or more relevant sentences in the human transcription.

Table 1 shows details of our test collection, in which lectures #1 and #2 were related to the criminal law and histories of ancient Greece, respectively. Each lecture was 45 minutes long. In this table, we shall use the term "word token" to refer to occurrences of words, and the term "word type" to refer to vocabulary items. The column "# of Fillers" denoting the number of interjections in speech partially shows the fluency of each lecturer.

Table 1. Details of our test collection.

ID	#1	#2
Topic	Law	History
# of Word tokens in lecture	6917	8092
# of Word types in lecture	1029	1219
# of Fillers in lecture	3	953
# of Sentences in lecture	181	191
# of Queries	25	13
Avg. # of relevant sentences per query	7.6	6.8
Avg. length of queries (Avg. # of words)	154	247

By using our test collection, we evaluated the accuracy of speech recognition and passage retrieval. It may be argued that passage segmentation should also be evaluated. However, to evaluate the extent to which the accuracy of passage segmentation affects the entire system performance, relevance assessment for passage retrieval has to be performed for multiple segmentations.

our system used the sentence boundaries in human transcriptions to identify speech units, and performed speech recognition. We also used human transcriptions as perfect speech recognition results and investigated the extent to which speech recognition errors affect the retrieval accuracy. Our system retrieved top-ranked passages in response to each query. It should be noted that passages here are those grouped based on the temporal axis, which should not be confused with those obtained in the passage segmentation method.

For lecture #1, we adapted the acoustic model to the lecturer by means of the MLLR-based method[4]. However, for lecture #2 we did not perform acoustic model adaptation, because the speech data contained constant background noise and the sound quality was not good enough to adapt the acoustic model. For both lectures #1 and #2, we did not use flip chart information obtained by means of image analysis.

3.2 Results

To evaluate the accuracy of speech recognition, we used word error rate (WER), which is the ratio between the number of word errors (i.e., deletion, insertion, and substitution) and the total number of words. We also used test-set out-of-vocabulary rate (OOV) and trigram test-set perplexity (PP) to evaluate the extent to which our language model was adapted to target topics.

We used human transcriptions as test set data. For example, OOV is the ratio between the number of word tokens not contained in the language model for speech

[3] http://chasen.aist-nara.ac.jp/

[4] http://htk.eng.cam.ac.uk/

recognition and the total number of word tokens in the transcription. It should be noted that smaller values of OOV, PP, and WER are obtained with better methods.

The final outputs (i.e., retrieved passages) were evaluated based on recall and precision, averaged over all queries. Recall (R) is the ratio between the number of correct speech units retrieved by our system and the total number of correct speech units for the query in question. Precision (P) is the ratio between the number of correct speech units retrieved by our system and the total number of speech units retrieved by our system. To summarize recall and precision into a single measure, we used F-measure (F).

Table 2 shows the accuracy of speech recognition (WER) and passage retrieval (R, P, and F), for each lecture. In this table, the columns "HUM" and "ASR" correspond to the results obtained with human transcriptions and automatic speech recognition, respectively. The results for ASR are also divided into those obtained with/without language model adaptation (LA).

To adapt language models, we used the textbook corresponding to a target lecture and searched the 10M Web page corpus [1][5] for 2,000 relevant pages, which were used as a source corpus. In the case where the language model adaptation was not performed, all 10M Web pages were used as a source corpus. In either case, 20,000 high frequency words were selected from a source corpus to produce word-based trigram language model. We used the ChaSen morphological analyzer to extract words (morphemes) from source corpora, because Japanese sentences lack lexical segmentation.

In passage retrieval, we regarded the top N passages as the final outputs. In Table 2, the value of N ranges from 1 to 3. As the value of N increases, the recall improves, but potentially sacrificing the precision.

3.3 Discussion

By comparing the results of ASR with/without LA in Table 2, OOV, PP, and WER decreased by adapting language models to target topics, irrespective of the lecture. Thus, our language model adaptation method was effective to improve the quality of speech recognition.

The values of OOV, PP, WER for lecture #2 were generally greater than those for lecture #1. One possible rationale is that the lecturer of #1 spoke more fluently and the number of erroneous utterances were smaller, when compared with the lecturer of #2. This tendency was partially observable in the column "# of Fillers in lecture" of Table 1. Additionally, the acoustic model was not adapted to the lecturer of #2, because the sound quality of the speech data for lecture #2 was not good enough to perform acoustic model adaptation.

By comparing the results of ASR with/without LA in Table 2, recall, precision, and F-measure increased by adapting language models to the topic of lecture #2, irrespective of the number of passages retrieved. This suggests that our language model adaptation method was effective to improve the retrieval accuracy.

For lecture #1, the retrieval accuracy did not differ whether or not we adapted the language model to the topic. One possible rationale is that WER of lecture #1 without language model adaptation (20.9%) was small enough to obtain the retrieval accuracy comparable with text retrieval [7]. In fact, the difference between HUM and ASR was marginal in terms of the retrieval accuracy. The effect of the language model adaptation method was overshadowed in passage retrieval.

Surprisingly, for lecture #2, recall, precision, and F-measure of ASR with LA were better than those of HUM except for the case of $N = 3$. In other words, the automatic transcription was more effective than the human transcription for passage retrieval purposes.

One possible rationale is related to Japanese variants (i.e., more than one spelling form corresponding to the same word), such as "*girisha*/*girishia* (Greece)." Since the language model was adapted by means of the textbook corresponding to a target lecture, the spelling in automatic transcriptions systematically resembled one in queries extracted from textbooks. In contrast, it is difficult to standardize the spelling in human transcriptions. Thus, relevant passages in automatic transcriptions were retrieved more likely than those in human transcriptions.

For all cases, recall was better than precision. This is attributed to our retrieval method. Since passages (one or more sentences) obtained by the initial phase were grouped into a single passage based on the temporal axis, irrelevant sentences were often contained in the retrieval results.

The retrieval accuracy for lecture #1 was generally higher than those for lecture #2. While the story of lecture #1 was organized based primarily on the textbook, the story of lecture #2 was relatively independent of the contents in the textbook. This suggests that the performance of LODEM is dependent of the organization of target lectures.

At the same time, since our test collection includes only two lectures, experiments using larger test collections in various topics should be further explored.

Table 2. Experimental results for speech recognition and passage retrieval.

ID		#1			#2		
			ASR			ASR	
		HUM	w/o LA	w/ LA	HUM	w/o LA	w/ LA
OOV		—	.0444	.0203	—	.0729	.0821
PP		—	48.91	43.27	—	122.1	96.69
WER		—	.2088	.1335	—	.5161	.4232
N=1	R	.6947	.7263	.7316	.4494	.2584	.5506
	P	.5344	.5476	.5187	.3774	.3194	.3858
	F	.6041	.6244	.6070	.4103	.2857	.4537
N=2	R	.8474	.8579	.8316	.6629	.3596	.6742
	P	.4411	.4478	.4580	.3010	.2105	.3141
	F	.5802	.5884	.5907	.4140	.2656	.4286
N=3	R	.8789	.8684	.8737	.7640	.4382	.7079
	P	.4103	.4054	.4010	.2688	.1625	.2520
	F	.5595	.5528	.5497	.3977	.2371	.3717

[5] http://research.nii.ac.jp/ntcir/index-en.html

4 Related Work

Spoken document retrieval (SDR), in which textual queries are used to search speech archives for relevant information, is primarily related to our research. Initiated partially by the TREC-6 SDR track [2], various SDR methods targeting broadcast news have been proposed [5, 6, 9]. State-of-the-art SDR methods, with WER being approximately 20%, are comparable with text retrieval methods in performance [7], and thus are already practical.

However, as shown in Table 2 (lecture #2), the speech recognition accuracy for lectures was not necessarily high when compared with broadcast news. While the TREC conference concluded that SDR in English was a solved problem, SDR for lectures remains unsolved and should be further explored, specifically for languages other than English.

Sheridan et al. [9] evaluated cross-language speech retrieval, in which French queries (manually produced based on German newspaper articles) were used to retrieve German news speech data. Informedia [3] retrieves video passages of TV news programs in response to textual queries. In the above two cases, users have to type the entire queries. However, in our case, users can utilize paragraphs in the textbook associated with a lecture, to formulate queries in foreign languages. We also use a number of adaptation methods to improve the quality of speech recognition.

Our research is also related to speech summarization [4], because a specific number of passages extracted from the entire speech data are organized so that users can understand important contents with a minimal cost. However, unlike existing methods targeting generic summaries, our method can be classified as a query-biased (user-focused) speech summarization, in which different summaries are generated depending on the user information need.

Finally, our research is crucial for e-education purposes, in which educational contents, such as lecture video/audio data are provided in real-time over computer networks. For example, in the WIDE University, School of Internet[6], lecture video data manually associated (synchronized) with presentation slides are available to the public over the Web. Our method is expected to reduce a cost required for manual annotation.

5 Conclusion

We proposed a multilingual lecture-on-demand system, in which users can view foreign video segments by submitting textual queries in their native language. To formulate effective foreign queries, users can utilize the textbook corresponding to a target lecture, machine translated into the user language. For the purpose of indexing video data, we used speech recognition, which were adapted to a target lecture in terms of acoustic and language models, to transcribe lecture speech information and to produce a textual index.

We evaluated the performance of our system by means of experiments, for which two TV lecture programs were used. The accuracy of speech recognition varied depending on the domain and presentation style of lectures. However, the accuracy of speech recognition and passage retrieval was improved by adapting language models to the topic of a target lecture. Even if the word error rate was approximately 40%, the accuracy of retrieval was comparable with that obtained by human transcriptions.

Future work will include improvement of each component in our system and extensive experiments using larger test collections related to various domains.

[6] http://www.soi.wide.ad.jp/

References

[1] K. Eguchi, K. Oyama, K. Kuriyama, and N. Kando. The Web retrieval task and its evaluation in the third NTCIR workshop. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 375–376, 2002.

[2] J. S. Garofolo, E. M. Voorhees, V. M. Stanford, and K. S. Jones. TREC-6 1997 spoken document retrieval track overview and results. In *Proceedings of the 6th Text REtrieval Conference*, pages 83–91, 1997.

[3] A. G. Hauptmann and M. J. Witbrock. Informedia: News-on-demand multimedia information acquisition and retrieval. In M. T. Maybury, editor, *Intelligent Multimedia Information Retrieval*, pages 215–239. AAAI Press/MIT Press, 1997.

[4] C. Hori and S. Furui. Automatic speech summarization based on word significance and linguistic likelihood. In *Proceedings of ICASSP2000*, pages 1579–1582, 2000.

[5] S. Johnson, P. Jourlin, G. Moore, K. S. Jones, and P. Woodland. The Cambridge University spoken document retrieval system. In *Proceedings of ICASSP'99*, pages 49–52, 1999.

[6] G. Jones, J. Foote, K. S. Jones, and S. Young. Retrieving spoken documents by combining multiple index sources. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 30–38, 1996.

[7] P. Jourlin, S. E. Johnson, K. S. Jones, and P. C. Woodland. Spoken document representations for probabilistic retrieval. *Speech Communication*, 32:21–36, 2000.

[8] S. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 232–241, 1994.

[9] P. Sheridan, M. Wechsler, and P. Schäuble. Cross-language speech retrieval: Establishing a baseline performance. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 99–108, 1997.

[10] S. Takao, J. Ogata, and Y. Ariki. Topic segmentation of news speech using word similarity. In *Proceedings of the Eighth ACM International Conference on Multimedia*, pages 442–444, 2000.

Evaluating Speech-Driven Web Retrieval in the Third NTCIR Workshop

Atsushi Fujii[†,†††] **and Katunobu Itou**[††,†††]
[†] Institute of Library and Information Science
University of Tsukuba
1-2 Kasuga, Tsukuba, 305-8550, Japan
[††] National Institute of Advanced Industrial Science and Technology
1-1-1 Chuuou Daini Umezono, Tsukuba, 305-8568, Japan
[†††] CREST, Japan Science and Technology Corporation
fujii@slis.tsukuba.ac.jp

Abstract

Speech recognition has of late become a practical technology for real world applications. For the purpose of research and development in speech-driven retrieval, which facilitates retrieving information with spoken queries, we organized the speech-driven retrieval subtask in the NTCIR-3 Web retrieval task. Search topics for the Web retrieval main task were dictated by ten speakers and were recorded as collections of spoken queries. We used those queries to evaluate the performance of our speech-driven retrieval system, in which speech recognition and text retrieval modules were integrated. The text retrieval module, which is based on a probabilistic model, indexed only textual contents in documents (Web pages), but did not use HTML tags and hyperlink information in documents. Experimental results showed that a) the use of target documents for language modeling and b) enhancement of the vocabulary size in speech recognition were effective to improve the system performance.

Introduction

Automatic speech recognition, which decodes human voice to generate transcriptions, has of late become a practical technology. It is feasible that speech recognition is used in real world computer-based applications, specifically, those associated with human language. In fact, a number of speech-based methods have been explored in the information retrieval (IR) community, which can be classified into the following two fundamental categories:

- spoken document retrieval, in which written queries are used to search speech (e.g., broadcast news audio) archives for relevant speech information (Johnson *et al.* 1999; Jones *et al.* 1996; Sheridan, Wechsler, & Schäuble 1997; Singhal & Pereira 1999; Srinivasan & Petkovic 2000; Wechsler, Munteanu, & Schäuble 1998; Whittaker *et al.* 1999),
- speech-driven retrieval, in which spoken queries are used to retrieve relevant textual information (Barnett *et al.* 1997; Crestani 2000; Fujii, Itou, & Ishikawa 2002a; 2002b; Itou, Fujii, & Ishikawa 2001; Kupiec, Kimber, & Balasubramanian 1994).

Initiated partially by the TREC-6 spoken document retrieval (SDR) track (Garofolo *et al.* 1997), various methods have been proposed for spoken document retrieval. However, a relatively small number of methods have been explored for speech-driven text retrieval, although they are associated with numerous keyboard-less retrieval applications, such as telephone-based retrieval, car navigation systems, and user-friendly interfaces.

Barnett et al. (1997) performed comparative experiments related to speech-driven retrieval, in which the DRAGON speech recognition system was used as an input interface for the INQUERY text retrieval system. They used as test inputs 35 queries collected from the TREC topics and dictated by a single male speaker. Crestani (2000) also used the above 35 queries and showed that conventional relevance feedback techniques marginally improved the accuracy of speech-driven text retrieval.

These above cases focused solely on improving text retrieval methods and did not address problems in improving speech recognition accuracy. In fact, an existing speech recognition system was used with no enhancement. In other words, speech recognition and text retrieval modules were fundamentally independent and were simply connected by means of an input/output protocol.

However, since most speech recognition systems are trained based on specific domains, the accuracy of speech recognition across domains is not satisfactory. As can easily be predicted, in cases of Barnett et al. (1997) and Crestani (2000), the speech recognition error rate was relatively high and decreased the retrieval accuracy. Additionally, speech recognition with a high accuracy is important for interactive retrieval, such as dialog-based retrieval.

Kupiec et al. (1994) proposed a method based on *word* recognition, which accepts only a small number of keywords, derives multiple transcription hypotheses (i.e., possible word combinations), and uses a target collection to determine the most plausible word combination. In other words, word combinations that frequently appear in the target collection can be recognized with a high accuracy. However, for longer queries, such as phrases and sentences, the number of hypotheses increases, and thus the searching cost is prohibitive. Thus, their method cannot easily be used for *continuous* speech recognition methods.

Motivated by these problems, we integrated continuous speech recognition and text retrieval to improve both recognition and retrieval accuracy in speech-driven text retrieval (Fujii, Itou, & Ishikawa 2002a; 2002b; Itou, Fujii, & Ishikawa 2001). In brief, our method used target documents to adapt language models and to recognize out-of-vocabulary words for speech recognition. However, a number of issues still remain open questions before speech-driven retrieval can be used as a practical (real-world) application. For example, extensive experiments using large test collections have not been performed for speech-driven retrieval. This stimulated us to further explore this exciting research area.

In the NTCIR-3 Web retrieval task[1], the *main* task was organized to promote conventional text-based retrieval (Eguchi *et al.* 2002). Additionally, *optional* subtasks were also invited, in which a group of researchers voluntarily organized a subtask to promote their common research area. To make use of this opportunity, we organized the "speech-driven retrieval" subtask, and produced a reusable test collection for experiments of Web retrieval driven by spoken queries. Since we also participated in the main task, we performed comparative experiments to evaluate the performance of text-based and speech-driven retrieval systems.

Test Collection for Speech-Driven Retrieval

Overview

The purpose of the speech-driven retrieval subtask was to produce reusable test collections and tools available to the public, so that researchers in the information retrieval and speech processing communities can develop technologies and share the scientific knowledge inherent in speech-driven information retrieval.

In principle, as with conventional IR test collections, test collections for speech-driven retrieval are required to include test queries, target documents, and relevance assessment for each query. However, unlike conventional text-based IR, queries are speech data uttered by human speakers.

In practice, since producing the entire collection is prohibitive, we produced speech data related to the Web retrieval main task. Therefore, target documents and relevance assessment in the main task can be used for the purpose of speech-driven retrieval. It should be noted that in the main task, retrieval results driven by spoken queries were not used for pooling, which is a method to reduce the number of relevant document candidates by using retrieval results of multiple IR systems (Voorhees 1998).

However, participants for the NTCIR workshop are mainly researchers in the information retrieval and natural language processing communities, and are not necessarily experts in developing and operating speech recognition systems. Thus, we also produced dictionaries and language models that can be used with an existing speech recognition engine (decoder), which helps researchers to perform similar experiments described in this paper.

All above data are included in the NTCIR-3 Web retrieval test collection, which is available to the public.

[1] http://research.nii.ac.jp/ntcir/index-en.html

Spoken Queries

For the NTCIR-3 Web retrieval main task, 105 search topics were manually produced, for each of which relevance assessment was manually performed with respect to two different document sets, i.e., 10GB and 100GB collections. The 10GB and 100GB collections translate approximately to 1M and 10M documents, respectively.

Each topic is in SGML-style form and consists of the topic ID (<NUM>), title of the topic (<TITLE>), description (<DESC>), narrative (<NARR>), list of synonyms related to the topic (<CONC>), sample of relevant documents (<RDOC>), and brief profile of the user who produced the topic (<USER>).

Figure 1 depicts a translation of an example topic. Although Japanese topics were used in the main task, English translations are also included in the Web retrieval collection mainly for publication purposes.

```
<TOPIC>
<NUM>0010</NUM>
<TITLE CASE="b">Aurora, conditions, ob-
servation</TITLE>
<DESC>For observation purposes, I want
to know the conditions that give rise to
an aurora</DESC>
<NARR><BACK>I want to observe an aurora
so I want to know the conditions neces-
sary for its occurrence and the mecha-
nism behind it.</BACK><RELE>Aurora ob-
servation records, etc. list the place
and time so only documents that pro-
vide additional information such as the
weather and temperature at the time of
occurrence are relevant. </RELE></NARR>
<CONC>Aurora, occurrence, conditions,
observation, mechanism</CONC>
<RDOC>NW003201843, NW001129327,
NW002699585</RDOC>
<USER>1st year Master's student, female,
2.5 years search experience</USER>
</TOPIC>
```

Figure 1: An example topic in the Web retrieval collection.

Participants for the main task were allowed to submit more than one retrieval result using one or more fields. However, participants were required to submit results obtained with the title and description fields independently. Titles are a list of keywords, and descriptions are phrases and sentences.

From the viewpoint of speech recognition, titles and descriptions can be used to evaluate *word* and *continuous* recognition methods, respectively. Since the state-of-the-art speech recognition is based on a continuous recognition framework, we used only the description field. For the first speech-driven retrieval subtask, we focused on *dictated* (*read*) speech, although our ultimate goal is to recognize *spontaneous* speech. We asked ten speakers (five adult males/females) to dictate descriptions in the 105 topics.

The ten speakers also dictated 50 sentences in the ATR phonetic-balanced sentence set as reference data, which can potentially be used for speaker adaptation (however, we did not use this additional data for the purpose of experiments described in this paper).

These above spoken queries and sentences were recorded with the same close-talk microphone in a noiseless office. Speech waves were digitized at a 16KHz sampling frequency and were quantized at 16 bits. The resultant data are in the RIFF format.

Language Models

Unlike general-purpose speech recognition, in speech-driven text retrieval, users usually speak contents associated with a target collection, from which documents relevant to user needs are retrieved.

In a stochastic speech recognition framework, the accuracy depends primarily on acoustic and language models (Bahl, Jelinek, & Mercer 1983). While acoustic models are related to phonetic properties, language models, which represent linguistic contents to be spoken, are related to target collections. Thus, it is intuitively feasible that language models have to be produced based on target collections. To sum up, our belief is that by adapting a language model to a target IR collection, we can improve the speech recognition accuracy and consequently the retrieval accuracy.

Motivated by this background, we used target documents for the main task to produce language models. For this purpose, we used only the 100GB collection, because the 10GB collection is a subset of the 100GB collection.

State-of-the-art speech recognition systems still have to limit the vocabulary size (i.e., the number of words in a dictionary), due to problems in estimating statistical language models (Young 1996) and constraints associated with hardware, such as memory. In addition, computation time is crucial for a real-time usage, including speech-driven retrieval. Consequently, for many languages the vocabulary size is limited to a couple of ten thousands (Itou *et al.* 1999; Paul & Baker 1992; Steeneken & van Leeuwen 1995).

We produced two language models of different vocabulary sizes, for which 20,000 and 60,000 high frequency words were independently used to produce word-based trigram models, so that researchers can investigate the relation between the vocabulary size and system performance. We shall call these models "Web20K" and "Web60K", respectively. We used the ChaSen morphological analyzer[2] to extract words from the 100GB collection.

To resolve the data sparseness problem, we used a back-off smoothing method, in which the Witten-Bell discounting method was used to compute back-off coefficients. In addition, through preliminary experiments, cut-off thresholds were empirically set 20 and 10 for the Web20K and Web60K models, respectively. Trigrams whose frequency was above the threshold were used for language modeling. Language models and dictionaries are in the ARPA and HTK formats, respectively.

[2]http://chasen.aist-nara.ac.jp/

Table 1 shows statistics related to word tokens/types in the 100GB collection and ten years of "Mainichi Shimbun" newspaper articles in 1991–2000. We shall use the term "word token" to refer to occurrences of words, and the term "word type" to refer to vocabulary items. Roughly, the size of the 100G collection ("Web") is ten times that of ten years of newspaper articles ("News"), which was one of the largest Japanese corpora available for the purpose of research and development in language modeling. In other words, the Web is a vital, as yet untapped, corpus for language modeling.

Table 1: The number of words in source corpora for language modeling.

	Web (100GB)	News (10 years)
# of Word types	2.57M	0.32M
# of Word tokens	2.44G	0.26G

System Description

Overview

Figure 2 depicts the overall design of our speech-driven text retrieval system, which consists of speech recognition and text retrieval modules.

In the off-line process, a target IR collection is used to produce a language model, so that user speech related to the collection can be recognized with a high accuracy. However, an acoustic model was produced independent of the target collection.

In the on-line process, given an information need spoken by a user (i.e., a spoken query), the speech recognition module uses acoustic and language models to generate a transcription of the user speech. Then, the text retrieval module searches the target IR collection for documents relevant to the transcription, and outputs a specific number of top-ranked documents according to the degree of relevance in descending order. In the following two sections, we explain speech recognition and text retrieval modules, respectively.

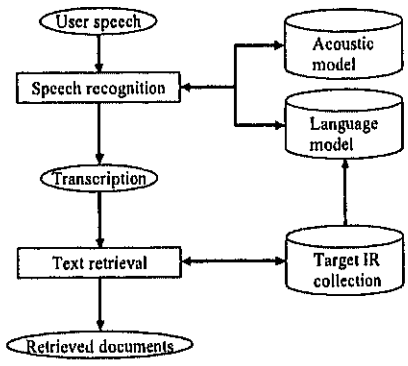

Figure 2: The overview of our speech-driven text retrieval system.

Speech Recognition

The speech recognition module generates word sequence W, given phone sequence X. In a stochastic speech recognition framework (Bahl, Jelinek, & Mercer 1983), the task is to select the W maximizing $P(W|X)$, which is transformed as in Equation (1) through the Bayesian theorem.

$$\arg\max_W P(W|X) = \arg\max_W P(X|W) \cdot P(W) \quad (1)$$

Here, $P(X|W)$ models a probability that word sequence W is transformed into phone sequence X, and $P(W)$ models a probability that W is linguistically acceptable. These factors are called acoustic and language models, respectively.

We used the Japanese dictation toolkit (Kawahara *et al.* 2000)[3], which includes the Julius decoder and acoustic/language models. Julius performs a two-pass (forward-backward) search using word-based forward bigrams and backward trigrams.

The acoustic model was produced from the ASJ speech database (Itou *et al.* 1998), which contains approximately 20,000 sentences uttered by 132 speakers including the both gender groups. A 16-mixture Gaussian distribution triphone Hidden Markov Model, in which states are clustered into 2,000 groups by a state-tying method, is used. The language model is a word-based trigram model produced from 60,000 high frequency words in ten years of Mainichi Shimbun newspaper articles.

This toolkit also includes development softwares so that acoustic and language models can be produced and replaced depending on the application. While we used the acoustic model provided in the toolkit, we used new language models produced from the 100GB collections, that is, the Web20K and Web60K models.

Text Retrieval

The retrieval module is based on an existing retrieval method (Robertson & Walker 1994), which computes the relevance score between the transcribed query and each document in the collection. The relevance score for document d is computed by Equation (2).

$$\sum_t f_{t,q} \cdot \frac{(K+1) \cdot f_{t,d}}{K \cdot \{(1-b) + \frac{dl_d}{b \cdot avgdl}\} + f_{t,d}} \cdot \log \frac{N - n_t + 0.5}{n_t + 0.5} \quad (2)$$

Here, $f_{t,q}$ and $f_{t,d}$ denote the frequency that term t appears in query q and document d, respectively. N and n_t denote the total number of documents in the collection and the number of documents containing term t, respectively. dl_d denotes the length of document d, and $avgdl$ denotes the average length of documents in the collection. We empirically set $K = 2.0$ and $b = 0.8$, respectively.

Given transcriptions (i.e., speech recognition results for spoken queries), the retrieval module searches a target IR collection for relevant documents and sorts them according to the score in descending order.

We used content words, such as nouns, extracted from documents as index terms, and performed word-based indexing. We used the ChaSen morphological analyzer to extract content words. We also extracted terms from transcribed queries using the same method. We used words and bi-words (i.e., word-based bigrams) as index terms.

We used the same retrieval module to participate in other text retrieval workshops, such as NTCIR-2. However, the 10GB/100GB Web collections were different from existing Japanese test collections in the following two perspectives.

First, the Web collections are much larger than existing test collections. For example, the file size of the NTCIR-2 Japanese collection including 736,166 technical abstracts is approximately 900MB (NII 2001). Thus, tricks were needed to index larger document collections. Specifically, files of more than 2GB size were problematic for file systems and tools in existing operating systems.

To resolve this problem, we divided the 100GB collection into 20 smaller sub-collections so that each file size did not exceed 2GB, and indexed the 20 files independently. Given queries, we retrieved documents using the 20 indexes and sorted documents according to the relevance score. The relevance score of a document was computed with respect to the sub-collection from which the document was retrieved.

Second, target documents are Web pages, in which HTML (Hyper Text Markup Language) tags provide the textual information with a certain structure. However, the use of HTML tags are usually different depending on the author. Thus, we discarded HTML tags in documents, and indexed only textual contents. Additionally, we did not use hyperlink information among Web pages for retrieval purposes.

Experimentation

Evaluating Text-to-Text Retrieval

In the Web retrieval main task, different types of text retrieval were performed. The first type was "Topic Retrieval" resembling the TREC ad hoc retrieval. The second type was "Similarity Retrieval," in which documents were used as queries instead of keywords and phrases. The third type was "Target Retrieval," in which systems with a high precision were highly valued. This feature provided a salient contrast to the first two retrieval types, in which both recall and precision were equally used as evaluation measures.

Although the produced spoken queries can be used for the first and third task types, we focused solely on the Topic Retrieval for the sake of simplicity. In addition, our previous experiments (Fujii, Itou, & Ishikawa 2002a; 2002b; Itou, Fujii, & Ishikawa 2001), in which the IREX[4] and NTCIR[5] collections were used, were also a type of Target Retrieval. We used the 47 topics for the Topic Retrieval task to retrieve 1,000 top documents, and used the TREC evaluation software to calculate mean average precision (MAP) values (i.e., non-interpolated average precision values, averaged over the 47 topics).

Relevance assessment was performed based on four ranks of relevance, that is, highly relevant, relevant, partially relevant and irrelevant. In addition, unlike conventional retrieval tasks, documents hyperlinked from retrieved documents were optionally used for relevance assessment. To

[3] http://winnie.kuis.kyoto-u.ac.jp/dictation/

[4] http://cs.nyu.edu/cs/projects/proteus/irex/index-e.html

[5] http://research.nii.ac.jp/ntcir/index-en.html

sum up, the following four assessment types were available to calculate the MAP values:

- (highly) relevant documents were regarded as correct answers, and hyperlink information was NOT used (RC),
- (highly) relevant documents were regarded as correct answers, and hyperlink information was used (RL),
- partially relevant documents were also regarded as correct answers, and hyperlink information was NOT used (PC),
- partially relevant documents were also regarded as correct answers, and hyperlink information was used (PL).

In the formal run for the main task, we submitted results obtained with different methods for the 10GB and 100GB collections, respectively. First, we used title (<TITLE>) and description (<DESC>) fields independently as queries. Second, we used as index terms either only words or a combination of words and bi-words. As a result, we investigated the MAP values for 32 cases as shown in Table 2.

By looking at Table 2, there was no significant difference among the four methods in performance. However, by comparing two indexing methods, the use of both words and bi-words generally improved the MAP values of that obtained with only words, irrespective of the collection size, topic field used, and relevance assessment type.

Evaluating Speech-Driven Retrieval

The purpose of experiments for speech-driven retrieval was two-fold. First, we investigated the extent to which a language model produced based on a target document collection contributes to improve the performance. Second, we investigated the impact of the vocabulary size for speech recognition to speech-driven retrieval. Thus, we compared the performance of the following four retrieval methods:

- text-to-text retrieval, which used written queries, and can be seen as the perfect speech-driven text retrieval ("Text"),
- speech-driven text retrieval, in which the Web60K model was used ("Web60K"),
- speech-driven text retrieval, in which a language model produced from 60,000 high frequency words in ten years of Mainichi Shimbun newspaper articles was used ("News60K"),
- speech-driven text retrieval, in which the Web20K model was used ("Web20K").

For text-to-text retrieval, we used descriptions (<DESC>) as queries, because the spoken queries used for speech-driven retrieval methods were descriptions dictated by speakers. In addition, we used both bi-words and words for indexing, because the experimental results in Table 2 showed that the use of bi-words for indexing improved the performance of that obtained with only words.

For speech-driven text retrieval methods, queries dictated by the ten speakers were used independently, and the final result was obtained by averaging results for all speakers. Although the Julius decoder used in the speech recognition module generated more than one transcription candidate (hypothesis) for a single speech, we used only the one with the greatest probability score.

All language models were produced by means of the same softwares, but were different in terms of the vocabulary size and source documents.

Table 3 shows the MAP values with respect to the four relevance assessment types and the word error rate in speech recognition, for different retrieval methods targeting the 10GB and 100GB collections.

As with existing experiments for speech recognition, word error rate (WER) is the ratio between the number of word errors (i.e., deletion, insertion, and substitution) and the total number of words. In addition, we investigated error rate with respect to query terms (i.e., keywords used for retrieval), which we shall call term error rate (TER). It should be noted that unlike MAP, smaller values of WER and TER are obtained with better methods.

Table 3 also shows test-set out-of-vocabulary rate (OOV), which is the ratio between the number of words not included in the speech recognition dictionary and the total number of words in spoken queries. In addition, the column of "Time" denotes CPU time (sec.) required for speech recognition per query, for which we used a PC with two CPUs (AMD Athlon MP 1900+) and a memory size of 3GB.

Suggestions which can be derived from the results in Table 3 are as follows.

Looking at columns of WER and TER, News60K and Web20K were comparable in the speech recognition performance, but Web60K outperformed both cases. However, difference of News60K and Web20K in OOV did not affect WER and TER. In addition, TER was greater than WER, because in computing TER, functional words, which are generally recognized with a high accuracy, were excluded.

While the MAP values of News60K and Web20K were also comparable, the MAP values of Web60K, which were roughly 60-70% of those obtained with Text, were greater than those for News60K and Web20K, irrespective of the relevance assessment type. These results were observable for both the 10GB and 100GB collections.

The only difference between News60K and Web60K was the source corpus for language modeling in speech recognition, and therefore we can conclude that the use of target collections to produce a language model was effective for speech-driven retrieval. In addition, by comparing the MAP values of Web20K and Web60K, we can conclude that the vocabulary size for speech recognition was also influential for the performance of speech-driven retrieval.

CPU time for speech recognition did not significantly differ depending on the language model, despite the fact that the number of words and N-gram tuples in Web60K was larger than those in News60K and Web20K. In other words, Web60K did not decrease the time efficiency of News60K and Web20K, which is crucial for read-world usage. At the same time, response time also depends on various factors, such as the hardware and decoder program used, we do not pretend to draw any premature conclusions regarding the time efficiency.

We analyzed speech recognition errors, focusing mainly on those attributed to the out-of-vocabulary problem. Ta-

Table 2: MAP values for different text-to-text retrieval methods targeting the 10GB and 100GB collections.

Field	Index	MAP (10GB)				MAP (100GB)			
		RC	RL	PC	PL	RC	RL	PC	PL
<DESC>	word & bi-word	.1470	.1286	.1612	.1476	.0855	.0982	.1257	.1274
<DESC>	word	.1389	.1187	.1563	.1374	.0843	.0928	.1184	.1201
<TITLE>	word & bi-word	.1493	.1227	.1523	.1407	.0815	.0981	.1346	.1358
<TITLE>	word	.1402	.1150	.1437	.1323	.0808	.0938	.1280	.1299

Table 3: Experimental results for different retrieval methods targeting the 10GB and 100GB collections (OOV: test-set out-of-vocabulary rate, WER: word error rate, TER: term error rate, MAP: mean average precision).

Method	OOV	WER	TER	Time (sec.)	MAP (10GB)				MAP (100GB)			
					RC	RL	PC	PL	RC	RL	PC	PL
Text	—	—	—	—	.1470	.1286	.1612	.1476	.0855	.0982	.1257	.1274
Web60K	.0073	.1311	.2162	7.2	.0966	.0916	.0973	.1013	.0542	.0628	.0766	.0809
News60K	.0157	.1806	.2991	7.0	.0701	.0681	.0790	.0779	.0341	.0404	.0503	.0535
Web20K	.0423	.1642	.2757	6.7	.0616	.0628	.0571	.0653	.0315	.0378	.0456	.0485

ble 4 shows the ratio of the number of out-of-vocabulary words to the total number of misrecognized words (or terms) in transcriptions. However, it should be noted that the actual ratio of errors due to the OOV problem can potentially be higher than those figures, because non-OOV words collocating with OOV words are often misrecognized. Remaining reasons of speech recognition errors are associated with insufficient N-gram statistics and the acoustic model.

Table 4: The ratio of the number of OOV words/terms to the total number of misrecognized words/terms.

	Word	Term
Web60K	.0704	.1838
News60K	.0966	.2143
Web20K	.2855	.5049

As can be predicted, the ratio of OOV words (terms) in Web20K was much higher than those in Web60K and News60K. However, by comparing News60K and Web20K, WER and TER of News60K in Table 3 were higher than those of Web20K. This suggests that insufficient N-gram statistics were more problematic in News60K, when compared with Web20K.

Although we used only the top-ranked transcription hypotheses as queries, certain words can potentially be correctly transcribed in lower-ranked hypotheses. Thus, to investigate the effect of multiple hypotheses, we varied the number of hypotheses used as queries and evaluated its effect on the MAP value. Table 5 shows the result, in which we used the Web60K model for speech recognition and targeted the 100G collection. In the case of $H = 1$, the MAP values are the same as those in Table 3. According to this table, the MAP values marginally decreased when we increased the number of hypotheses used as queries, irrespective of the relevance assessment type.

Table 5: MAP values of the Web60K speech-driven retrieval method with different numbers of hypotheses (H), targeting the 100G collection.

	RC	RL	PC	PL
$H = 1$	.0542	.0628	.0766	.0809
$H = 3$	.0527	.0608	.0755	.0794
$H = 5$	.0529	.0609	.0754	.0794

Conclusion

In the NTCIR-3 Web retrieval task, we organized the speech-driven retrieval subtask and produced 105 spoken queries dictated by ten speakers. We also produced word-based trigram language models using approximately 10M documents in the 100GB collection used for the main task. We used those queries and language models to evaluate the performance of our speech-driven retrieval system. Experimental results showed that a) the use of target documents for language modeling and b) enhancement of the vocabulary size in speech recognition were effective to improve the system performance. As with the collection for the main task, all speech data and language models produced for this subtask are available to the public. Future work will include experiments using spontaneous spoken queries.

Acknowledgments

The authors would like to thank the organizers of the NTCIR-3 Web retrieval task for their support to the speech-driven retrieval subtask.

References

Bahl, L. R.; Jelinek, F.; and Mercer, R. L. 1983. A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 5(2):179–190.

Barnett, J.; Anderson, S.; Broglio, J.; Singh, M.; Hudson, R.; and Kuo, S. W. 1997. Experiments in spoken queries for document retrieval. In *Proceedings of Eurospeech97*, 1323–1326.

Crestani, F. 2000. Word recognition errors and relevance feedback in spoken query processing. In *Proceedings of the Fourth International Conference on Flexible Query Answering Systems*, 267–281.

Eguchi, K.; Oyama, K.; Kuriyama, K.; and Kando, N. 2002. The Web retrieval task and its evaluation in the third NTCIR workshop. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 375–376.

Fujii, A.; Itou, K.; and Ishikawa, T. 2002a. A method for open-vocabulary speech-driven text retrieval. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, 188–195.

Fujii, A.; Itou, K.; and Ishikawa, T. 2002b. Speech-driven text retrieval: Using target IR collections for statistical language model adaptation in speech recognition. In Coden, A. R.; Brown, E. W.; and Srinivasan, S., eds., *Information Retrieval Techniques for Speech Applications (LNCS 2273)*. Springer. 94–104.

Garofolo, J. S.; Voorhees, E. M.; Stanford, V. M.; and Jones, K. S. 1997. TREC-6 1997 spoken document retrieval track overview and results. In *Proceedings of the 6th Text REtrieval Conference*, 83–91.

Itou, K.; Yamamoto, M.; Takeda, K.; Takezawa, T.; Matsuoka, T.; Kobayashi, T.; Shikano, K.; and Itahashi, S. 1998. The design of the newspaper-based Japanese large vocabulary continuous speech recognition corpus. In *Proceedings of the 5th International Conference on Spoken Language Processing*, 3261–3264.

Itou, K.; Yamamoto, M.; Takeda, K.; Takezawa, T.; Matsuoka, T.; Kobayashi, T.; and Shikano, K. 1999. JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research. *Journal of Acoustic Society of Japan* 20(3):199–206.

Itou, K.; Fujii, A.; and Ishikawa, T. 2001. Language modeling for multi-domain speech-driven text retrieval. In *IEEE Automatic Speech Recognition and Understanding Workshop*.

Johnson, S.; Jourlin, P.; Moore, G.; Jones, K. S.; and Woodland, P. 1999. The Cambridge University spoken document retrieval system. In *Proceedings of ICASSP'99*, 49–52.

Jones, G.; Foote, J.; Jones, K. S.; and Young, S. 1996. Retrieving spoken documents by combining multiple index sources. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 30–38.

Kawahara, T.; Lee, A.; Kobayashi, T.; Takeda, K.; Minematsu, N.; Sagayama, S.; Itou, K.; Ito, A.; Yamamoto, M.; Yamada, A.; Utsuro, T.; and Shikano, K. 2000. Free software toolkit for Japanese large vocabulary continuous speech recognition. In *Proceedings of the 6th International Conference on Spoken Language Processing*, 476–479.

Kupiec, J.; Kimber, D.; and Balasubramanian, V. 1994. Speech-based retrieval using semantic co-occurrence filtering. In *Proceedings of the ARPA Human Language Technology Workshop*, 373–377.

National Institute of Informatics. 2001. *Proceedings of the 2nd NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization.*

Paul, D. B., and Baker, J. M. 1992. The design for the Wall Street Journal-based CSR corpus. In *Proceedings of DARPA Speech & Natural Language Workshop*, 357–362.

Robertson, S., and Walker, S. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 232–241.

Sheridan, P.; Wechsler, M.; and Schäuble, P. 1997. Cross-language speech retrieval: Establishing a baseline performance. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 99–108.

Singhal, A., and Pereira, F. 1999. Document expansion for speech retrieval. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 34–41.

Srinivasan, S., and Petkovic, D. 2000. Phonetic confusion matrix based spoken document retrieval. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 81–87.

Steeneken, H. J. M., and van Leeuwen, D. A. 1995. Multilingual assessment of speaker independent large vocabulary speech-recognition systems: The SQALE-project. In *Proceedings of Eurospeech95*, 1271–1274.

Voorhees, E. M. 1998. Variations in relevance judgments and the measurement of retrieval effectiveness. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 315–323.

Wechsler, M.; Munteanu, E.; and Schäuble, P. 1998. New techniques for open-vocabulary spoken document retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 20–27.

Whittaker, S.; Hirschberg, J.; Choi, J.; Hindle, D.; Pereira, F.; and Singhal, A. 1999. SCAN: Designing and evaluating user interfaces to support retrieval from speech archives. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 26–33.

Young, S. 1996. A review of large-vocabulary continuous-speech recognition. *IEEE Signal Processing Magazine* 45–57.

A System for On-demand Video Lectures

Atsushi Fujii[†,†††] **Katunobu Itou**[††,†††] **Tetsuya Ishikawa**[†]
[†] Institute of Library and Information Science
University of Tsukuba
1-2 Kasuga, Tsukuba, 305-8550, Japan
[††] National Institute of Advanced Industrial Science and Technology
1-1-1 Chuuou Daini Umezono, Tsukuba, 305-8568, Japan
[†††] CREST, Japan Science and Technology Corporation
fujii@slis.tsukuba.ac.jp

Abstract

We propose a lecture-on-demand system, which searches lecture videos for segments relevant to user information needs. We utilize the benefits of textbooks and audio/video data corresponding to a single lecture. Our system extracts the audio track from a target lecture video, generates a transcription by large vocabulary continuous speech recognition, and produces a textual index. Users can selectively view specific video segments by submitting textual queries associated with the textbook for the target lecture. Experimental results showed that by adapting speech recognition to the lecture topic, the recognition accuracy increased and the retrieval accuracy was comparable with that obtained by human transcriptions. Our system is implemented as a client-server system over the Web to facilitate e-education.

Introduction

Given the growing number of multimedia contents available via the World Wide Web, CD-ROMs, and DVDs, information technologies across speech, image, and text processing have of late become crucial. Among various types of contents, lectures (audio/video) are very typical and valuable multimedia contents, in which speeches (i.e., oral presentations) are usually organized based on textual materials, such as resumes, slides, and textbooks. In lecture videos, image information, such as flip charts, is often additionally used. In other words, a single lecture consists of different types of compatible multimedia contents.

However, since a single lecture often includes multiple stories and takes long time, it is useful to selectively obtain specific segments (passages) so that audience can satisfy their information needs with a minimal cost. To resolve this problem, in this paper we propose a lecture-on-demand system, which retrieves relevant video/audio passages in response to user queries. For this purpose, we utilize the benefits of different media types to improve retrieval performance.

On the one hand, textual contents are advantageous in the sense that users can view/scan the entire contents quickly and easily identify relevant passages using layout information (e.g., text structures based on sections and paragraphs). In other words, textual contents can be used for random-access purposes.

On the other hand, speech contents are fundamentally used for sequential-access purposes. Thus, it is difficult to identify relevant passages unless target video/audio data include additional annotations, such as indexes. Even if target data are indexed, users are not necessarily able to come up with effective queries. To resolve this problem, textbooks are desirable materials, from which users can extract effective keywords and phrases.

However, while textbooks are usually concise, speeches are relatively redundant and thus are easy to understand more than textbooks, specifically in the case where additional image information is provided.

In view of the above discussion, we model our lecture-on-demand (LOD) system as follows. A user selects text segments (i.e., keywords, phrases, sentences, and paragraphs) relevant to their information needs, from a textbook for a target lecture. By using selected segments, a textual query is automatically generated. In other words, queries can be formulated even if users cannot come up with effective keywords and phrases. Users can also submit additional keywords as queries, if necessary. Video passages relevant to a given query are retrieved and presented to the user.

To retrieve video passages in response to textual queries, we extract the audio track from a lecture video, generate a transcription by means of large vocabulary continuous speech recognition, and produce a textual index, prior to the system usage.

Our on-demand system should not be confused with video-on-demand (VOD) systems, which search video archives for specific videos in response to user requests. While in VOD systems, minimal unit for retrieval is the entire program, in our system, retrieval units are passages smaller than the entire program.

System Description

Overview

Figure 1 depicts the overall design of our lecture-on-demand system, in which left/right-hand regions correspond to the on-line and off-line processes, respectively. Although our system is currently implemented for Japanese, our methodology is fundamentally language-independent. For the pur-

pose of research and development, we tentatively target lecture programs on TV for which textbooks are published. We explain the basis of our system using this figure.

In the off-line process, given the video data of a target lecture, audio data are extracted and segmented into more than one passage. Then, speech recognition transcribes each passage. Finally, transcribed passages are indexed as performed in conventional text retrieval systems, so that each passage can be retrieved efficiently in response to textual queries.

To adapt speech recognition to a specific lecturer, we perform unsupervised speaker adaptation using an initial speech recognition result (i.e., a transcription).

To adapt speech recognition to a specific topic, we perform language model adaptation, for which we search a general corpus for documents relevant to the textbook related to a target lecture. Then, retrieved documents (i.e., a topic-specific corpus) are used to produce a word-based N-gram language model.

We also perform image analysis to extract textual contents (e.g., keywords and phrases) in flip charts. These contents are also used later to improve our language model.

In the on-line process, a user can view specific video passages by submitting any textual queries, i.e., keywords, phrases, sentences, and paragraphs, extracted from the textbook. Any queries not in the textbook can also additionally be used. The current implementation is based on a client-server system over the Web. While both the off-line and on-line processes are performed on servers, users can utilize our system with Web browsers on their own PCs.

Figure 2 depicts a prototype interface of our LOD system, in which a lecture associated with "nonlinear multivariate analysis" is given. In this interface, an electronic version of a textbook is displayed in the left-hand side, and a lecture video is played in the right-hand side. In addition, users can submit any textual queries to the box in the bottom of the interface. The operation is similar to that for existing Web search engines.

In Figure 3, a text paragraph related to "discriminant analysis" is copied and pasted into the query input box. It should be noted that unlike conventional keyword-based retrieval systems, in which users usually submit a small number of keywords, in our system users can easily submit longer queries using textbooks. In the case where submitted keywords are misrecognized in transcriptions, the retrieval accuracy decreases. However, longer queries are relatively robust for speech recognition errors, because the effect of misrecognized words are overshadowed by a large number of words correctly recognized.

Figure 4 depicts retrieval results, in which top-ranked transcribed passages for the query in Figure 3 are listed according to the degree of relevance. Users can select (click) transcriptions to play the corresponding video passage. We explain each module in the following three sections.

Passage Segmentation

The basis of passage segmentation is to divide the entire video data for a single lecture into more than one minimal unit to be retrieved. We shall call those units passages.

For this purpose, both speech and image data can be promising clues. For example, Hamada et al. (2000) performed a structural analysis on cooking TV programs by means of speech/image/text processing. However, in lecture TV programs, it is often the case that a lecturer sitting still is mainly focused and a small number of flip charts are occasionally used. In such cases, image data is less informative. Thus, we tentatively use only speech data for the passage segmentation process.

However, unlike the case where a target speech (e.g., a news program) consists of multiple different topics (Allan 2002; Takao, Ogata, & Ariki 2000), topic segmentation for lectures is relatively difficult, because a single lecture consists of sub-topics closely related to one another, and thus topic boundaries are not necessarily clear.

Existing methods to segment written texts (e.g., one proposed by Hearst (1997)) rely only on lexical information in texts, and thus are not robust against errors in automatic transcriptions. Additionally, in our LOD system, segmentation can potentially vary depending on the user query. Thus, it is difficult to predetermine a desirable segmentation in the off-line process.

In view of the above problems, we first extract the audio track from a target video, and perform a simple pause-based segmentation method to obtain minimal speech units, such as sentences and long phrases. In other words, speech units are continuous audio segments that do not include pauses longer than a certain threshold. Finally, we generate variable-length passages from one or more speech units. To put it more precisely, we combine N speech units into a single passage, with N ranging from 1 to 5 in the current implementation.

Figure 5 shows an example of variable-length passages, in which any sequences of speech units that are 1-5 in length are identified as passages.

Speech Recognition

The speech recognition module generates word sequence W, given phone sequence X. In a stochastic speech recognition framework (Bahl, Jelinek, & Mercer 1983), the task is to select the W maximizing $P(W|X)$, which is transformed as in Equation (1) through the Bayesian theorem.

$$\arg\max_W P(W|X) = \arg\max_W P(X|W) \cdot P(W) \qquad (1)$$

$P(X|W)$ models a probability that word sequence W is transformed into phone sequence X, and $P(W)$ models a probability that W is linguistically acceptable. These factors are called acoustic and language models, respectively.

We use the Japanese dictation toolkit (Kawahara *et al.* 2000)[1], which includes the Julius decoder and acoustic/language models. Julius performs a two-pass (forward-backward) search using word-based forward bigrams and backward trigrams.

The acoustic model was produced from the ASJ speech database (Itou *et al.* 1998), which contains approximately 20,000 sentences uttered by 132 speakers including the both

[1] http://winnie.kuis.kyoto-u.ac.jp/dictation/

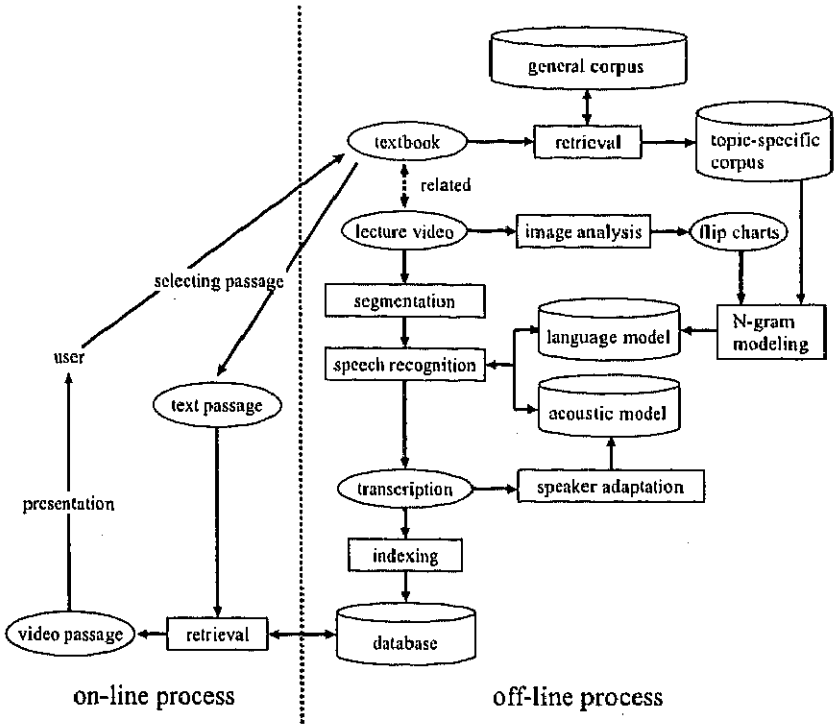

Figure 1: The overview of our lecture-on-demand system.

gender groups. A 16-mixture Gaussian distribution triphone Hidden Markov Model, in which states are clustered into 2,000 groups by a state-tying method, is used. We adapt the provided acoustic model by means of an MLLR-based unsupervised speaker adaptation method (Leggetter & Woodland 1995), for which in practice we use the HTK toolkit[2].

Existing methods to adapt language models can be classified into two fundamental categories. In the first category, the *integration* approach, general and topic-specific (domain-specific) corpora are integrated to produce a topic-specific language model (Auzanne *et al.* 2000; Seymore & Rosenfeld 1997). Since the sizes of those corpora are different, N-gram statistics are calculated by the weighted average of statistics extracted independently from those corpora. However, it is difficult to determine the optimal weight depending on the topic.

In the second category, the *selection* approach, a topic-specific subset is selected from a general corpus and is used to produce a language model. This approach is effective if general corpora contain documents associated with target topics, but N-gram statistics in those documents are overshadowed by other documents in resultant language models.

We followed the selection approach, because the 10M Web page corpus (Eguchi *et al.* 2002)[3] containing mainly Japanese pages associated with various topics was available to the public. The quality of the selection approach is dependent of the method to select topic-specific subsets. An existing method (Chen *et al.* 2001) uses hypotheses in the initial speech recognition phase as queries to retrieve topic-specific documents from a general corpus. However, errors in the initial hypotheses potentially decrease the retrieval accuracy. Instead, we use textbooks related to target lectures as queries to improve the retrieval accuracy and consequently the quality of language model adaptation.

[2]http://htk.eng.cam.ac.uk/

[3]http://research.nii.ac.jp/ntcir/index-en.html

Retrieval

Given transcribed passages and textual queries, the basis of the retrieval module is the same as conventional text retrieval. We use an existing probabilistic text retrieval method (Robertson & Walker 1994) to compute the relevance score between the query and each passage in the database. The relevance score for passage p is computed by Equation (2).

$$\sum_t f_{t,q} \cdot \frac{(K+1) \cdot f_{t,p}}{K \cdot \{(1-b) + \frac{dl_p}{b \cdot avgdl}\} + f_{t,p}} \cdot \log \frac{N - n_t + 0.5}{n_t + 0.5} \quad (2)$$

Here, $f_{t,q}$ and $f_{t,p}$ denote the frequency that term t appears in query q and passage p, respectively. N and n_t denote the total number of passages in the database and the number of passages containing term t, respectively. dl_p denotes the length of passage p, and $avgdl$ denotes the average length of

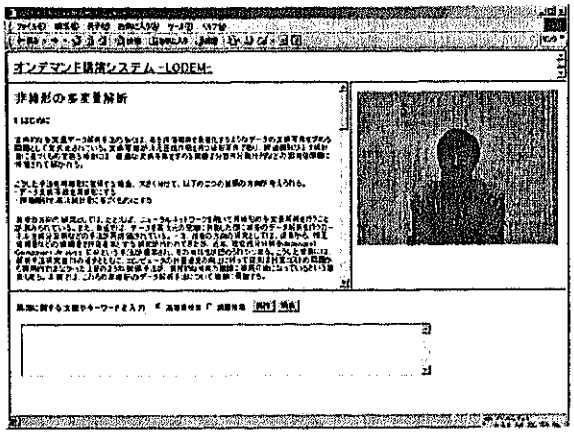

Figure 2: The interface of our LOD system over the Web.

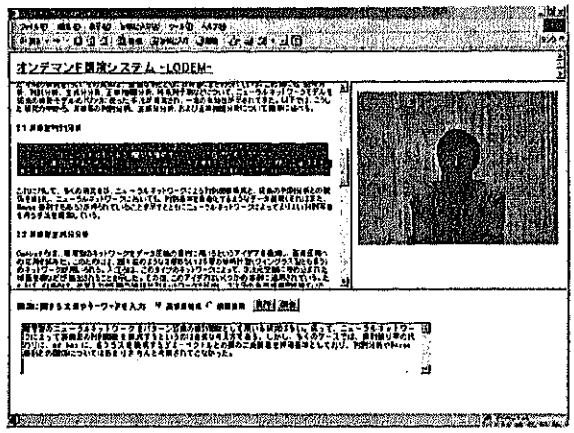

Figure 3: An example scenario, in which a paragraph in the textbook is copied and pasted into the query input box.

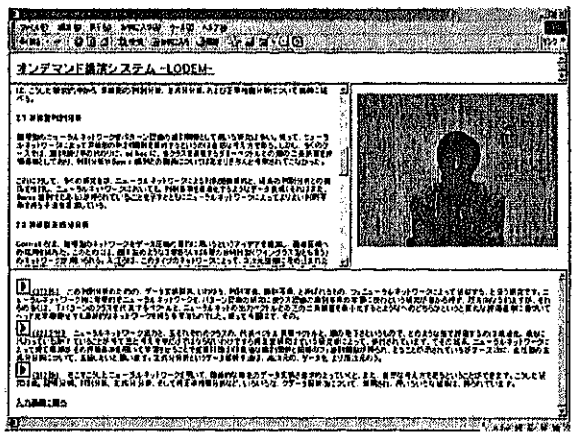

Figure 4: Example retrieved transcriptions.

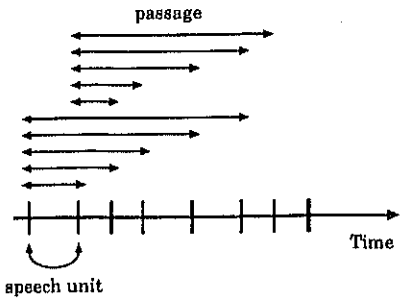

Figure 5: An example of the passage segmentation process.

passages in the database. We empirically set $K = 2.0$ and $b = 0.8$, respectively.

It should be noted that in Equation (2), the score is normalized with the length of each passage. Thus, longer passages, which potentially include more index terms, are not necessarily assigned with a higher score. This property is important, because variable-length passages are considerably different in terms of length.

We use content words, such as nouns, extracted from transcribed passages as index terms, and perform word-based indexing. We use the ChaSen morphological analyzer[4] to extract content words. We also extract terms from queries using the same method.

However, retrieved passages are not disjoint, because top-ranked passages often overlap with one another in terms of the temporal axis. It is redundant to simply list the top-ranked retrieved passages as they are. Thus, we reorganize those overlapped passages into a single passage. In Figure 6, which uses the same basic notation as Figure 5, illustrates an example scenario. In this figure, top-ranked passages are organized into three groups.

The relevance score for a group (a new passage) is computed by averaging scores for all passages belonging to the group. New passages are sorted according to the degree of relevance and are presented to users as the final result.

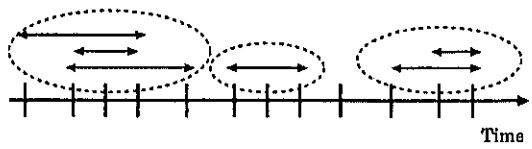

Figure 6: An example of grouping retrieved passages.

Experimentation

Methodology

To evaluate the performance of our LOD system, we produced a test collection (a kind of benchmark data set) and performed experiments partially resembling one performed

[4]http://chasen.aist-nara.ac.jp/

in the TREC spoken document retrieval (SDR) track (Garofolo *et al.* 1997).

Two lecture programs on TV, for which printed textbooks were also published, were videotaped in DV and were used as target lectures. Both lectures were manually transcribed and sentence boundaries with temporal information (i.e., correct speech units) were also manually identified. The textbooks for the two target lectures were read by an OCR software and were manually revised. The accuracy of the OCR software was roughly 97% on a word-by-word basis.

For both lectures, each paragraph in the corresponding textbook was used as a query independently. For each query, a human assessor (a graduate student other than authors of this paper) identified one or more relevant sentences in the human transcription.

Table 1 shows details of our test collection, in which lectures #1 and #2 were associated with the criminal law and histories of ancient Greece, respectively. Each lecture was 45 minutes long. In this table, we shall use the term "word token" to refer to occurrences of words, and the term "word type" to refer to vocabulary items. The column "# of Fillers" denoting the number of interjections in speech partially shows the fluency of each lecturer.

Table 1: Details of our test collection used for experiments.

ID	#1	#2
Topic	Law	History
# of Word tokens in lecture	6917	8092
# of Word types in lecture	1029	1219
# of Fillers in lecture	3	953
# of Sentences in lecture	181	191
# of Queries	25	13
Avg. # of relevant sentences per query	7.6	6.8
Avg. length of queries (Avg. # of words)	154	247

By using our test collection, we evaluated the accuracy of speech recognition and passage retrieval. It may be argued that passage segmentation should also be evaluated. However, to evaluate the extent to which the accuracy of passage segmentation affects the entire system performance, relevance assessment for passage retrieval has to be performed for multiple segmentations, which is expensive.

For both target lectures, our system used the sentence boundaries in human transcriptions to identify speech units, and performed speech recognition. We also used human transcriptions as perfect speech recognition results and investigated the extent to which speech recognition errors affect the retrieval accuracy. Our system retrieved top-ranked passages in response to each query. It should be noted that passages here are those grouped based on the temporal axis, which should not be confused with those obtained in the passage segmentation method.

For lecture #1, we adapted the acoustic model to the lecturer by means of the MLLR-based method. However, for lecture #2 we did not perform acoustic model adaptation, because the speech data contained constant background noise and the sound quality was not good enough to adapt the acoustic model. For both lectures #1 and #2, we did not use flip chart information obtained by means of image analysis.

Results

To evaluate the accuracy of speech recognition, we used word error rate (WER), which is the ratio between the number of word errors (i.e., deletion, insertion, and substitution) and the total number of words. We also used test-set out-of-vocabulary rate (OOV) and trigram test-set perplexity (PP) to evaluate the extent to which our language model was adapted to target topics.

We used human transcriptions as test set data. For example, OOV is the ratio between the number of word tokens not contained in the language model for speech recognition and the total number of word tokens in the transcription. It should be noted that smaller values of OOV, PP, and WER are obtained with better methods.

The final outputs (i.e., retrieved passages) were evaluated based on recall and precision, averaged over all queries. Recall (R) is the ratio between the number of correct speech units retrieved by our system and the total number of correct speech units for the query in question. Precision (P) is the ratio between the number of correct speech units retrieved by our system and the total number of speech units retrieved by our system. To summarize recall and precision into a single measure, we used F-measure (F), which is calculated by Equation (3).

$$\frac{(\beta^2 + 1) \cdot \mathrm{R} \cdot \mathrm{P}}{\beta^2 \cdot \mathrm{R} + \mathrm{P}} \tag{3}$$

Here, β is a parametric constant used to control the preference between recall and precision. In our case, recall and precision were equally important, and thus we set $\beta = 1$.

Table 2 shows the accuracy of speech recognition (WER) and passage retrieval (R, P, and F), for each lecture. In this table, the columns "HUM" and "ASR" correspond to the results obtained with human transcriptions and automatic speech recognition, respectively. The results for ASR are also divided into those obtained with/without language model adaptation (LA).

To adapt language models, we used the textbook corresponding to a target lecture and searched the 10M Web page corpus for 2,000 relevant pages, which were used as a source corpus. In the case where the language model adaptation was not performed, all 10M Web pages were used as a source corpus. In either case, 20,000 high frequency words were selected from a source corpus to produce word-based trigram language model. We used the ChaSen morphological analyzer to extract words (morphemes) from source corpora, because Japanese sentences lack lexical segmentation.

In passage retrieval, we regarded the top N passages as the final outputs. In Table 2, the value of N ranges from 1 to 3. As the value of N increases, the recall improves, but potentially sacrificing the precision.

Discussion for Speech Recognition

By comparing the results of ASR with/without LA in Table 2, OOV, PP, and WER decreased by adapting language

models to target topics, irrespective of the lecture. This suggests that our language model adaptation method was effective to improve the quality of speech recognition.

The values of OOV, PP, WER for lecture #2 were generally greater than those for lecture #1. One possible rationale is that the lecturer of #1 spoke more fluently and the number of erroneous utterances were smaller, when compared with the lecturer of #2. This tendency was partially observable in the column "# of Fillers in lecture" of Table 1. Additionally, the acoustic model was not adapted to the lecturer of #2, because the sound quality of the speech data for lecture #2 was not good enough to perform acoustic model adaptation.

Discussion for Passage Retrieval

By comparing the results of ASR with/without LA in Table 2, recall, precision, and F-measure increased by adapting language models to the topic of lecture #2, irrespective of the number of passages retrieved. This suggests that our language model adaptation method was effective to improve the retrieval accuracy.

For lecture #1, the retrieval accuracy did not significantly differ whether or not we adapted the language model to the topic. One possible rationale is that WER of lecture #1 without language model adaptation (20.9%) was small enough to obtain the retrieval accuracy comparable with text retrieval (Jourlin *et al.* 2000). In fact, the difference between HUM and ASR was marginal in terms of the retrieval accuracy. Thus, the effect of the language model adaptation method was overshadowed in passage retrieval.

Surprisingly, for lecture #2, recall, precision, and F-measure of ASR with LA were better than those of HUM except for the case of $N = 3$. In other words, the automatic transcription was more effective than the human transcription for passage retrieval purposes.

One possible rationale is associated with Japanese variants (i.e., more than one spelling form corresponding to the same word), such as "*girisha*/*girishia* (Greece)." Since the language model was adapted by means of the textbook corresponding to a target lecture, the spelling in automatic transcriptions systematically resembled one in queries extracted from textbooks. In contrast, it is difficult to standardize the spelling in human transcriptions. Thus, relevant passages in automatic transcriptions were retrieved more likely than passages in human transcriptions.

For all cases, recall was better than precision. This is attributed to our retrieval method. Since passages (one or more sentences) obtained by the initial phase were grouped into a single passage based on the temporal axis, irrelevant sentences were often contained in the retrieval results.

The retrieval accuracy for lecture #1 was generally higher than those for lecture #2. While the story of lecture #1 was organized based primarily on the textbook, the story of lecture #2 was relatively independent of the contents in the textbook. This suggests that the performance of our LOD system is dependent of the organization of target lectures.

At the same time, since our test collection includes only two lectures, experiments using larger test collections associated with various topics should be further explored.

Table 2: Experimental results for speech recognition (OOV: test-set out-of-vocabulary rate, PP: trigram test-set perplexity, WER: word error rate) and passage retrieval (N: # of passages retrieved, R: recall, P: precision, F: F-measure).

ID		#1			#2		
			ASR			ASR	
		HUM	w/o LA	w/ LA	HUM	w/o LA	w/ LA
OOV		—	.0444	.0203	—	.0729	.0821
PP		—	48.91	43.27	—	122.1	96.69
WER		—	.2088	.1335	—	.5161	.4232
N=1	R	.6947	.7263	.7316	.4494	.2584	.5506
	P	.5344	.5476	.5187	.3774	.3194	.3858
	F	.6041	.6244	.6070	.4103	.2857	.4537
N=2	R	.8474	.8579	.8316	.6629	.3596	.6742
	P	.4411	.4478	.4580	.3010	.2105	.3141
	F	.5802	.5884	.5907	.4140	.2656	.4286
N=3	R	.8789	.8684	.8737	.7640	.4382	.7079
	P	.4103	.4054	.4010	.2688	.1625	.2520
	F	.5595	.5528	.5497	.3977	.2371	.3717

Related Work

Informedia (Hauptmann & Witbrock 1997) retrieves video passages from TV news programs in response to textual queries, for which users have to type the entire queries. This feature is problematic in the case where users have difficulty formulating effective queries. However, in our case, users can utilize segments of the textbook associated with a lecture as queries even if they cannot come up with effective keywords and phrases.

Hamada et al. (2000) performed structural analysis on cooking TV programs by means of speech/image/text processing, in which the textbook for a program was additionally used. However, while they focused mainly on analyzing video data, we intended to retrieve video passages.

Unlike our study in this paper, in the above two cases no quantitative experimental results were shown with respect to the accuracy of retrieving video data. Thus, it is difficult to compare the performance of our system with those for those existing systems.

Spoken document retrieval (SDR), in which textual queries are used to search speech archives for relevant information, is primarily related to our research. Initiated partially by the TREC-6 SDR track (Garofolo *et al.* 1997), various SDR methods targeting broadcast news have been proposed (Johnson *et al.* 1999; Jones *et al.* 1996; Sheridan, Wechsler, & Schäuble 1997). State-of-the-art SDR methods, with WER being approximately 20%, are comparable with text retrieval methods in performance (Jourlin *et al.* 2000), and thus are already practical.

However, as shown in Table 2 (lecture #2), the speech recognition accuracy for lectures was not necessarily high when compared with broadcast news. While the TREC conference concluded that SDR in English was a solved problem, SDR for lectures remains unsolved and should be further explored, specifically for languages other than English.

Segmenting lectures into passages is associated with

the Topic Detection and Tracking (TDT) evaluation workshop (Allan 2002), in which one task is to segment a single broadcast news stream into topically coherent stories. However, in the case of lectures, stories in a single lecture are closely related to one another, and therefore topic segmentation is more difficult than that for broadcast news programs.

Our research is also associated with speech summarization (Hori & Furui 2000), because a specific number of passages extracted from the entire speech data are organized so that users can understand important contents with a minimal cost. However, unlike existing methods for generic summaries, our method is classified as a query-biased (user-focused) summarization (Mani & Bloedorn 1998; Tombros & Sanderson 1998), in which different summaries are generated depending on the user information needs.

Finally, our research is crucial for e-education purposes, in which educational contents, such as lecture video/audio data are provided in real-time over computer networks. For example, in the WIDE University, School of Internet[5], lecture video data manually synchronized with presentation slides are available to the public over the Web.

Jones and Edens (2002) proposed a system to automatically synchronize an audio track with presentation slides, which is expected to reduce a cost required for manual indexing. Their system is similar to our system, because textual materials (slides and textbooks) are used to identify corresponding passages in a presentation. However, while their system was mainly intended to match transcriptions with slides, we also addressed problems in adapting language models for speech recognition, and showed its effectiveness by means of experiments.

Conclusion

Reflecting the rapid growth in the utilization of multimedia contents, information technologies across speech, image, and text processing are crucial. Among various content types, in this paper we focused video data of lectures organized based on textbooks, and proposed a system for on-demand lectures, in which users can formulate textual queries using the textbook for a target lecture to retrieve specific video passages.

To retrieve video passages in response to textual queries, we extract the audio track from a lecture video, generate a transcription by large vocabulary continuous speech recognition, and produce a textual index, prior to the system usage. The current system is implemented as a server-client system over the Web to facilitate e-education.

We also evaluated the performance of our system by means of experiments, for which two TV lecture programs were used. The experimental results showed that the accuracy of speech recognition varied depending on the domain and presentation style of lectures. However, the accuracy of speech recognition and passage retrieval was improved by adapting language models to the topic of a target lecture. In addition, even if the word error rate was approximately 40%, the accuracy of retrieval was comparable with that obtained by human transcriptions.

[5]http://www.soi.wide.ad.jp/

Future work will include improvement of each component in our system and extensive experiments using larger test collections related to various domains.

References

Allan, J., ed. 2002. *Topic Detection and Tracking: Event-based News Organization.* Kluwer Academic Publishers.

Auzanne, C.; Garofolo, J. S.; Fiscus, J. G.; and Fisher, W. M. 2000. Automatic language model adaptation for spoken document retrieval. In *Proceedings of RIAO 2000 Conference on Content-Based Multimedia Information Access.*

Bahl, L. R.; Jelinek, F.; and Mercer, R. L. 1983. A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 5(2):179–190.

Chen, L.; Gauvain, J.-L.; Lamel, L.; Adda, G.; and Adda, M. 2001. Language model adaptation for broadcast news transcription. In *Proceedings of ISCA Workshop on Adaptation Methods For Speech Recognition.*

Eguchi, K.; Oyama, K.; Kuriyama, K.; and Kando, N. 2002. The Web retrieval task and its evaluation in the third NTCIR workshop. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 375–376.

Garofolo, J. S.; Voorhees, E. M.; Stanford, V. M.; and Jones, K. S. 1997. TREC-6 1997 spoken document retrieval track overview and results. In *Proceedings of the 6th Text REtrieval Conference*, 83–91.

Hamada, R.; Ide, I.; Sakai, S.; and Tanaka, H. 2000. Associating cooking video with related textbook. In *Proceedings of the International Workshop on Multimedia Information Retrieval*, 237–241.

Hauptmann, A. G., and Witbrock, M. J. 1997. Informedia: News-on-demand multimedia information acquisition and retrieval. In Maybury, M. T., ed., *Intelligent Multimedia Information Retrieval.* AAAI Press/MIT Press. 215–239.

Hearst, M. A. 1997. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics* 23(1):33–64.

Hori, C., and Furui, S. 2000. Automatic speech summarization based on word significance and linguistic likelihood. In *Proceedings of ICASSP2000*, 1579–1582.

Itou, K.; Yamamoto, M.; Takeda, K.; Takezawa, T.; Matsuoka, T.; Kobayashi, T.; Shikano, K.; and Itahashi, S. 1998. The design of the newspaper-based Japanese large vocabulary continuous speech recognition corpus. In *Proceedings of the 5th International Conference on Spoken Language Processing*, 3261–3264.

Johnson, S.; Jourlin, P.; Moore, G.; Jones, K. S.; and Woodland, P. 1999. The Cambridge University spoken document retrieval system. In *Proceedings of ICASSP '99*, 49–52.

Jones, G. J. F., and Edens, R. J. 2002. Automated alignment and annotation of audio-visual presentations. In *Pro-*

ceedings of the 6th European Conference on Development for Digital Libraries, 276–291.

Jones, G.; Foote, J.; Jones, K. S.; and Young, S. 1996. Retrieving spoken documents by combining multiple index sources. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 30–38.

Jourlin, P.; Johnson, S. E.; Jones, K. S.; and Woodland, P. C. 2000. Spoken document representations for probabilistic retrieval. *Speech Communication* 32:21–36.

Kawahara, T.; Lee, A.; Kobayashi, T.; Takeda, K.; Minematsu, N.; Sagayama, S.; Itou, K.; Ito, A.; Yamamoto, M.; Yamada, A.; Utsuro, T.; and Shikano, K. 2000. Free software toolkit for Japanese large vocabulary continuous speech recognition. In *Proceedings of the 6th International Conference on Spoken Language Processing*, 476–479.

Leggetter, C., and Woodland, P. 1995. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language* 9:171–185.

Mani, I., and Bloedorn, E. 1998. Machine learning of generic and user-focused summarization. In *Proceedings of AAAI/IAAI-98*, 821–826.

Robertson, S., and Walker, S. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 232–241.

Seymore, K., and Rosenfeld, R. 1997. Using story topics for language model adaptation. In *Proceedings of Eurospeech97*, 1987–1990.

Sheridan, P.; Wechsler, M.; and Schäuble, P. 1997. Cross-language speech retrieval: Establishing a baseline performance. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 99–108.

Takao, S.; Ogata, J.; and Ariki, Y. 2000. Topic segmentation of news speech using word similarity. In *Proceedings of the Eighth ACM International Conference on Multimedia*, 442–444.

Tombros, A., and Sanderson, M. 1998. Advantages of query biased summaries in information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2–10.

SELECTIVE BACK-OFF SMOOTHING FOR INCORPORATING GRAMMATICAL CONSTRAINTS INTO THE N-GRAM LANGUAGE MODEL

Tomoyosi AKIBA[†], *Katunobu ITOU*[†], *Atsushi FUJII*[‡], *Tetsuya ISHIKAWA*[‡]

[†] National Institute of Advanced Industrial Science and Technology (AIST)
1-1-1 Umezono, Tsukuba, 305-8568, JAPAN, E-mail: t-akiba@aist.go.jp
[‡] University of Library and Information Science
1-2 Kasuga, Tsukuba, 305-8550, JAPAN

ABSTRACT

Spoken queries submitted to question answering systems usually consist of query contents (e.g. about newspaper articles) and frozen patterns (e.g. WH-words), which can be modeled with N-gram models and grammar-based models, respectively. We propose a method to integrate those different types of models into a single N-gram model. We represent the two types of language models in a single word network. However, common smoothing methods, which are effective for N-gram models, decrease grammatical constraints for frozen patterns. For this problem, we propose a selective back-off smoothing method, which controls a degree to which smoothing is applied depending the network fragment. Additionally, resulting models are compatible with the conventional back-off N-gram models, and thus existing N-gram decoders can easily be used. We show the effectiveness of our method by way of experiments.

1. INTRODUCTION

The N-gram model has been used successfully as a language model for large vocabulary continuous speech recognition (LVCSR) systems. The N-gram model is simple but robust enough to model all word sequences in the vocabulary. However, it needs a large training corpus and such a corpus cannot be easily constructed unless there already exists a large text corpus based on, for example, newspaper articles.

On the other hand, the grammar-based model is used as a language model for tasks involving a relatively small vocabulary. This model does not need a training corpus because it takes advantage of linguistic knowledge. It can model correlations more distant than is possible with the N-gram model, which can model only local relations between words.

Thus, some spoken sentences can be modeled more suitably by the N-gram models and some more suitably by the grammar-based model. This is also true from an intra-sentence perspective – some parts of sentence are best modeled by N-gram model and some parts are best modeled by a grammar-based model.

For example, question answering systems receive queries that often consist of a part that conveys various query contents about, for example, newspaper articles, and a part that represents a frozen pattern for query sentences. The first part seems to be best dealt with by using an N-gram model trained with a corpus of newspaper articles, and the second part is best dealt with by using the grammar-based model.

The second and third authors are also members of CREST, Japan Science and Technology Corporation.

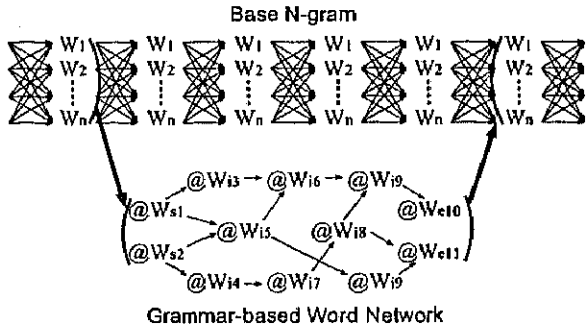

Fig. 1. Integrated Word Network

Another example is phrases that cannot be predicted from past training data, such as those concerning telephone numbers, ID numbers, a date, a time for a reservation, etc. Recognizing such a phrase is often necessary to achieve the given task. Those phrases seems to be best dealt with by using the grammar-based model while the other parts of the sentence is dealt with by using an N-gram model.

In this paper, we will explain how these two types of models can be integrated into a single N-gram model. The key idea is that we assume the grammar is described in regular language [1], so we should be able to represent it in an N-gram which is also equal to regular language. The problem is that the grammatical constraint is incompatible with smoothing, because the grammar tries to assign a zero probability to non-grammatical connections of the words while smoothing tries to assign a non-zero probability to avoid the zero frequency problem. To solve this inconsistency, we have developed a selective back-off smoothing method.

2. INTEGRATION OF THE CONVENTIONAL N-GRAM AND GRAMMAR-BASED MODELS

The sentences modeled by a conventional N-gram model can be expressed as a fully connected word network, in which a word can be followed by any word in the vocabulary of the model. On the other hand, the sentences modeled by a grammar-based model can be expressed as a partially connected word network, in which a word can be followed by only specific words according to the grammar. Integra-

[1] This is a reasonable assumption because any language of finite length is known to be included in regular language. Moreover, an algorithm exists to approximate CFGs into finite state automata [1].

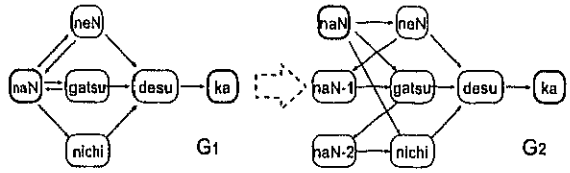

Fig. 2. Word Network

tion of the conventional N-gram model (we will refer to it as "base N-gram") and the grammar-based model can be achieved by joining the two different types of word networks so that any word in the base N-gram might be followed by the beginning words of the grammar and the ending words of the grammar might be followed by any word of the base N-gram (Fig. 1).

In the integrated word network, the N-gram probabilities are assigned so that the N-gram probability that runs along the directed arcs of the network might have a positive value and the other probabilities might be zero. With such a probability distribution, the integrated model is able to convey the two different characters of the conventional N-gram model and the grammar-based model at the same time.

To obtain such a distribution, we must give the N-gram counts along the arcs of the network, then calculate the probabilities. The next section of this paper will show how to construct the word network for the grammar and how to obtain the N-gram counts on the network. In section 4, we will introduce the selective back-off smoothing method that is used calculate the probabilities while preserving the characteristics of both the N-gram and the grammar.

3. GRAMMAR-BASED LANGUAGE MODELING

3.1. Word Network

Though there are many way to describe regular language, we will use a word network that consists of nodes corresponding to words and directed arcs corresponding to word-to-word transitions.

Such a network can be easily obtained for given example sentences. For example, we can obtain the network at the left of Fig.2 from the following three Japanese sentences that denote questions about the date,

naN / neN / desu / ka

naN / neN / naN / gatsu / desu / ka

naN / gatsu / naN / nichi / desu / ka

using the words "naN" (what), "neN" (year), "gatsu" (month), "nichi" (day), "desu" (is) and "ka" (*interrogative*). From the three sentences, we can get neighboring word pairs as follows.

$$\begin{aligned} A = \; & \{(\text{naN, neN})(\text{naN, gatsu})(\text{naN, nichi}) \\ & (\text{neN, naN})(\text{gatsu, naN})(\text{neN, desu}) \\ & (\text{gatsu, desu})(\text{nichi, desu})(\text{desu, ka})\} \end{aligned}$$

If we assume word-to-word transition is possible if the word pair appear in A, the word network G_1 is defined by 4-tuples (W_A, W_B, W_E, A) where W_a, W_b and W_e are, respectively, the set of all words, the set of beginning words, the set of ending words of the network. In this case,

$$\begin{aligned} W_A &= \{\text{naN neN gatsu nichi desu ka}\} \\ W_B &= \{\text{naN}\} \\ W_E &= \{\text{ka}\} \end{aligned}$$

Unfortunately, the simple network G_1 fails to model the Japanese query sentences about the date correctly, because the sentences below that can be modeled by G_1 are never used in practice.

* naN / neN / naN / neN / desu / ka

(the word "neN (year)" is repeated)

* naN / gatsu / naN / neN / desu / ka

(a more specific word "gatsu (month)" comes before a less specific word "neN (year)")

To exclude such ill-formed sentences, the word network should be improved by separating the word nodes according to their context (as shown at the right of Fig. 2). With the new symbol assigned to the separated nodes, the improved word network G_2 is describe by (W'_A, W_B, W_E, A') where:

$$\begin{aligned} W'_A &= W_A \cup \{\text{naN-1 naN-2}\} \\ A' &= \{(\text{naN, neN})(\text{naN, gatsu})(\text{naN, nichi})(\text{neN, naN-1}) \\ &\quad (\text{naN-1, gatsu})(\text{gatsu, naN-2})(\text{naN-2, nichi}) \\ &\quad (\text{neN, desu})(\text{gatsu, desu})(\text{nichi, desu})(\text{desu, ka})\} \end{aligned}$$

In this way, introducing the new nodes allow word networks to express any long distance dependence between words.

To integrate the word network with the base N-gram, two requirements must be met when constructing the word network. First, the vocabulary for the network must be distinguished from that for the base N-gram. This can be easily achieved by applying different word symbols to the same words in the network and the base N-gram. Namely, we gave "@" as the prefix for word symbols in the network. Formally, the vocabulary of the integrated model consists of the exclusive word sets W_U for the base N-gram and W_A for the word network, i.e. $W_U \cap W_A = \phi$.

Second, the beginning words must not arrive other than at the beginning of the network and the ending words must not arrive other than at the ending of the network. In other words, the word set W_A consists of exclusive word sets W_B, W_I and W_E (i.e. $W_A = W_B \cup W_I \cup W_E$), that respectively correspond to beginning words, intermediate words, and ending words, i.e. $W_B \cap W_I = \phi \wedge W_I \cap W_E = \phi \wedge W_B \cap W_E = \phi$.

3.2. Calculating Probability with Extra N-gram Counts

To give the N-gram probabilities of the words in the word network, we need the N-gram counts both on the word network and on the bridge connecting the network with the base N-gram.

The conventional N-gram counts are consistent and complete. To calculate the N-gram probability, only the longest N-gram counts $C(w_{i-N+1}^{i})$ (where w_{i-N+1}^{i} denotes the word sequence $w_{i-N+1} \cdots w_{i-1} w_i$) are needed because the shorter N-gram counts are obtained from them by recursively summing them up giving all the last (or first) words. Consequently, an N-gram count $C(w_{i-N+1}^{i})$ is used to calculate all the probabilities that predict the words

$w_{i-N+1} \cdots w_i$. In other words, raising an N-gram count $C(w_{i-N+1}^i)$ results in raising all the probabilities used to predict the words $w_{i-N+1} \cdots w_i$.

If we provide an arbitrary number of extra N-gram counts of an arbitrary length, the consistency and the completeness of the conventional N-gram counts are broken. Furthermore, we want to give the extra N-gram counts $C(w_{i-N+1}^i)$ only to raise the probability of the last word w_i. Thus, we need to redefine the usage of the N-gram counts as follows.

- The N-gram counts $C_1(w_i)$, $C_2(w_{i-1}^i)$, $\cdots$, $C_N(w_{i-N+1}^i)$ are given separately according to their length $1 \cdots N$.
- Each N-gram count $C_n(w_{i-n+1}^i)$ is used only to calculate the probability used to predict the last word w_i.

For example, the probability estimated through the maximum likelihood method can be calculated from the conventional N-gram counts as $P_{ML}(w_{i-n+1}^i) = C_n(w_{i-n+1}^i)/C_{n-1}(w_{i-n+1}^{i-1})$ which relies on consistency among the counts. On the other hand, with the extra N-gram counts, the probability must be calculated as $P_{ML}(w_{i-n+1}^i) = C_n(w_{i-n+1}^i)/\sum_{w_i} C_n(w_{i-n+1}^i)$

3.3. Providing N-gram Counts onto the Word Network

To provide N-gram counts to the constructed word network, we used the existing counts on the base N-gram, though several approaches are applicable that include giving equal counts at each branch of the network. To do so, the words in the network must be of the same unit with the base N-gram[2].

The N-gram counts related to the network can then be obtained from the corresponding counts in the base N-gram. We obtain the N-gram count $C_n(@w_{i-n+1}^i)$ in the network by copying the count $C_n(w_{i-n+1}^i)$ found in the base N-gram. We also need the counts corresponding to the bridge between the network and the base N-gram. Such counts can also be obtained by copying the corresponding counts in the base N-gram. Namely, both $C_n(w_{i-n+1}^{i-k}@w_{i-k+1}^i)$ and $C_n(@w_{i-n+1}^{i-k}w_{i-k+1}^i)$ are obtained from $C_n(w_{i-n+1}^{i-k}w_{i-k+1}^i)$. Among these, we obtain $C_n(w_{i-n+1}^{i-1}@w_i)$ by multiplying $C_n(w_{i-n+1}^{i-1}w_i)$ by the weight γ. γ is introduced so that we can assign a higher probability to a sentence that traces the network and γ is assumed to be given a value not less than 1 (Fig. 3).

4. SELECTIVE BACK-OFF SMOOTHING

The basic formula for back-off smoothing is

$$P(w_i|w_{i-n+1}^{i-1}) = \begin{cases} d_{w_{i-n+1}^i} P_{ML}(w_i|w_{i-n+1}^{i-1}) & \\ \qquad \cdots C_n(w_{i-n+1}^i) > 0 & \\ \alpha_{n-1}(w_{i-n+1}^{i-1}) P(w_i|w_{i-n+2}^{i-1}) & \\ \qquad \cdots C_n(w_{i-n+1}^i) = 0 & \end{cases} \quad (1)$$

[2]The definition of the word unit is important and should be done beforehand to model Japanese because words are not explicitly separated by spaces in Japanese sentences.

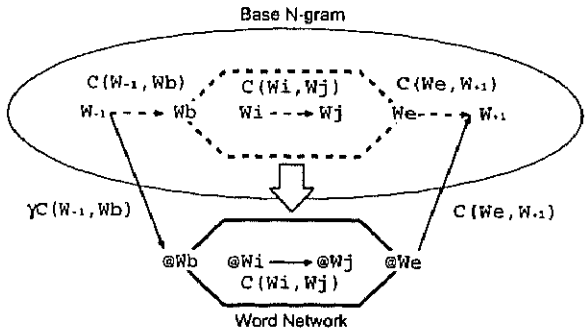

Fig. 3. Copying the N-gram counts (in case of bi-gram)

where d is the discount coefficient, P_{ML} is the probability obtained by maximum likelihood estimation, and α is the normalized function chosen to equalize the total probabilities to one.

If we apply the smoothing equally to all of the words in the N-gram model, the resulting model allows any word-to-word transition and loses the grammatical constraint. On the other hand, if we do not apply the smoothing, the grammatical constraint is preserved, but the model suffers from the data sparseness problem.

Our solution is to apply the two smoothing methods selectively according to the words of the N-gram model (Fig. 4). The probabilities used to predict the words that belong to the base N-gram and the beginning of the word network ($w_i \in W_U \cup W_B$) are calculated using the ordinary back-off smoothing method (Equation 1)[3]. The probabilities used to predict the other words that belong to the word network, excluding the beginnings ($w_i \in W_I \cup W_E$), are calculated so as not to "back-off" to the uni-gram probabilities. In other words, the probabilities are calculated by the ordinary back-off smoothing for the length $n > 2$ but using the following formula for the length $n = 2$.

$$P(w_i|w_{i-1}) = \begin{cases} d'_{w_{i-1}^i} P_{ML}(w_i|w_{i-1}) & C_2(w_{i-1}^i) > 0 \\ 0 & C_2(w_{i-1}^i) = 0 \end{cases} \quad (2)$$

As a result, we get

$$P(w_i) = 0 \quad \text{if } w_i \in W_I \cup W_E \quad (3)$$

$$\alpha_1(w_{i-1}) = 0 \quad \text{if } w_{i-1} \in W_B \cup W_I \quad (4)$$

The model to which the selective back-off smoothing is applied has two significant features.

- The probability $P(w_i|w_{i-N+1}^{i-1})$, where $w_i \in W_I \cup W_E$ and the word sequence w_{i-N+1}^i is not allowed by the grammar, becomes zero.

  In this case, the N-gram counts $C_n(w_{i-n+1}^i)$ for $n > 1$ are zero. Thus the probability is "back-off"ed to the uni-gram:

  $$P(w_i|w_{i-N+1}^{i-1}) = \alpha_{N-1} \cdots \alpha_2 \cdot \alpha_1(w_{i-1}) P(w_i) = 0$$

[3]More precisely, the uni-gram discounting must be applied to only the word set $W_U \cup W_B$ because the uni-gram probabilities on $W_I \cup W_E$ are set to 0 by Equation 2.

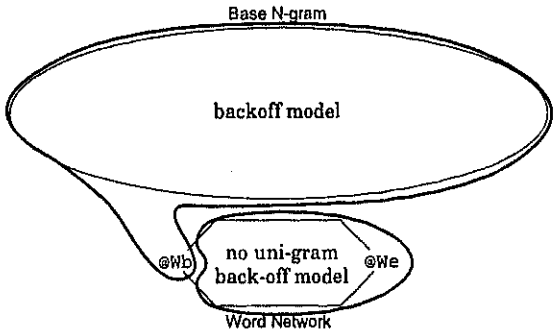

Fig. 4. Selective Back-off Smoothing

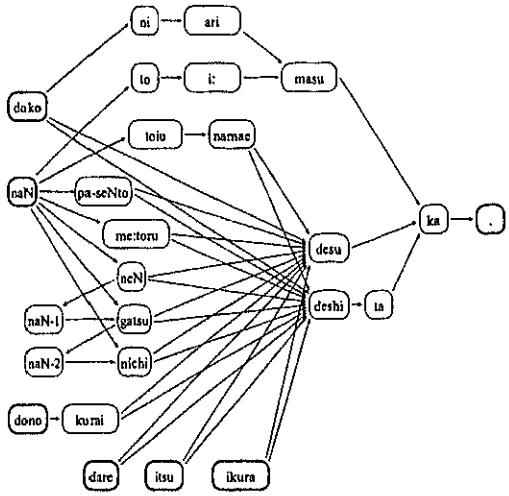

Fig. 5. Word network used for the experiment

because $P(w_i) = 0$ from Equation 3.

Especially, the probabilities used to predict the word $w \in W_I \cup W_E$ from the words (in the base N-gram) $w \in W_U$ are always zero.

- The probabilities used to predict the word (in the base N-gram) $w \in W_U$ from the words (in the network grammar, but excepting the endings) $w \in W_B \cup W_I$ are always zero.

  Also in this case, the N-gram counts $C_n(w_{i-n+1}^{i})$ for $n > 1$ are zero. Thus, the probability is "back-off"ed to the uni-gram:

  $$P(w_i|w_{i-N+1}^{i-1}) = \alpha_{N-1} \cdots \alpha_2 \cdot \alpha_1(w_{i-1})P(w_i) = 0$$

  because the context word w_{i-1} is in $W_B \cup W_I$ and we get $\alpha(w_{i-1}) = 0$ from Equation 4.

The resulting integrated model is compatible with conventional back-off N-gram models so that it can work as the language model for existing LVCSR systems.

5. EXPERIMENTAL RESULTS

We developed a word network for the Japanese frozen patterns used for Question Answering [3] (Fig.5). We also prepared a base N-gram of 20,000 words that were obtained from newspaper articles collected over 111 months. We then integrated these two models according to our method explained above (referred to as the *base+net* model). To enable comparison, we made the N-gram model from only the newspaper articles by using the conventional method (referred to as the *base* model). We used the Witten-Bell discounting method [2] for smoothing in both models.

Table 1. Results ($\gamma = 2$)

target (# of sent.)	language model	WER (2-gram)	WER (3-gram)
NP (100)	*base*	20.6	12.3
	base+net	21.6	12.3
QA (50)	*base*	30.6	20.9
	base+net	26.6	18.4
QA' (36)	*base*	32.8	21.4
	base+net	25.5	16.7

We prepared 100 sentences from the newspaper articles (referred to as *NP*) and 50 query sentences for the QA system (referred to as *QA*), and these were recorded for four speakers (two men and two women). Though the word network was relatively small and had only 33 nodes (31 words) in the network, the 36 of 50 query has the frozen patterns written by the network (referred as *QA'*).

The existing N-gram decoder [4] was used for the recognition experiments. The language model weight and the insertion penalty were set to the best values for the newspaper (*base*) model. The results are shown in Table 1.

We found that the integrated model significantly reduced the word error rate (WER) for the QA queries (*QA*) while scarcely increased the WER for the newspaper articles (*NP*). Moreover, when the word network covered the frozen patterns of the queries (*QA'*), it reduced the WER further.

6. CONCLUSION

Our selective back-off smoothing method was developed to enable the incorporation of grammatical constraints into the conventional N-gram model. The resulting integrated model is compatible with the conventional back-off N-gram model so that it can work as the language model for an existing N-gram decoder. We applied our model to model the frozen patterns used in queries for question answering systems. Results showed that our model significantly reduced the WER for queries for QA systems while scarcely increased the WER for the newspaper articles that had been modeled by the base N-gram before integration.

7. REFERENCES

[1] F.C.N.Pereira and R.R.Wright, "Finite-state approximation of phrase-structure grammars", In Proc. of ACL 1991, pp.246-255, 1991.

[2] P. Placeway, R. Schwartz, P. Fung and L. Nguyen, "The Estimation of Powerful Lnaguage Models from Small and Large Corpora", In Proc. of ICASSP, Vol.II, pp.33-36, 1993.

[3] E. Voorhees, D. Tice, "The TREC-8 Question Answering Track Evaluation", In Proc. of The Eighth Text Retrieval Conference, pp.83-106, 1999.

[4] A. Lee, T. Kawahara, K. Shikano, "Julius — an Open Source Real-Time Large Vocabulary Recognition Engine", In Proc. of Eurospeech, pp.1691-1694, 2001

A Method for Open-Vocabulary Speech-Driven Text Retrieval

Atsushi Fujii*
University of Library and Information Science
1-2 Kasuga, Tsukuba
305-8550, Japan
fujii@ulis.ac.jp

Katunobu Itou
National Institute of Advanced Industrial Science and Technology
1-1-1 Chuuou Daini Umezono
Tsukuba, 305-8568, Japan
itou@ni.aist.go.jp

Tetsuya Ishikawa
University of Library and Information Science
1-2 Kasuga, Tsukuba
305-8550, Japan
ishikawa@ulis.ac.jp

Abstract

While recent retrieval techniques do not limit the number of index terms, out-of-vocabulary (OOV) words are crucial in speech recognition. Aiming at retrieving information with spoken queries, we fill the gap between speech recognition and text retrieval in terms of the vocabulary size. Given a spoken query, we generate a transcription and detect OOV words through speech recognition. We then correspond detected OOV words to terms indexed in a target collection to complete the transcription, and search the collection for documents relevant to the completed transcription. We show the effectiveness of our method by way of experiments.

1 Introduction

Automatic speech recognition, which decodes human voice to generate transcriptions, has of late become a practical technology. It is feasible that speech recognition is used in real-world human language applications, such as information retrieval.

Initiated partially by TREC-6, various methods have been proposed for "spoken document retrieval (SDR)," in which written queries are used to search speech archives for relevant information (Garofolo et al., 1997). State-of-the-art SDR methods, where speech recognition error rate is 20-30%, are comparable with text retrieval methods in performance (Jourlin et al., 2000), and thus are already practical. Possible rationales include that recognition errors are overshadowed by a large number of words correctly transcribed in target documents.

However, "speech-driven retrieval," where spoken queries are used to retrieve (textual) information, has not fully been explored, although it is related to numerous keyboard-less applications, such as telephone-based retrieval, car navigation systems, and user-friendly interfaces.

Unlike spoken document retrieval, speech-driven retrieval is still a challenging task, because recognition errors in short queries considerably decrease retrieval accuracy. A number of references addressing this issue can be found in past research literature.

Barnett et al. (1997) and Crestani (2000) independently performed comparative experiments related to speech-driven retrieval, where the DRAGON speech recognition system was used as an input interface for the INQUERY text retrieval system. They used as test queries 35 topics in the TREC collection, dictated by a single male speaker. However, these cases focused on improving text retrieval methods and did not address problems in improving speech recognition. As a result, errors in recognizing spoken queries (error rate was approximately 30%) considerably decreased the retrieval accuracy.

Although we showed that the use of target document collections in producing language models for speech recognition significantly improved the performance of speech-driven retrieval (Fujii et al., 2002; Itou et al., 2001), a number of issues still remain open questions.

* The first and second authors are also members of CREST, Japan Science and Technology Corporation.

Section 2 clarifies problems addressed in this paper. Section 3 overviews our speech-driven text retrieval system. Sections 4-6 elaborate on our methodology. Section 7 describes comparative experiments, in which an existing IR test collection was used to evaluate the effectiveness of our method. Section 8 discusses related research literature.

2 Problem Statement

One major problem in speech-driven retrieval is related to out-of-vocabulary (OOV) words.

On the one hand, recent IR systems do not limit the vocabulary size (i.e., the number of index terms), and can be seen as open-vocabulary systems, which allow users to input any keywords contained in a target collection. It is often the case that a couple of million terms are indexed for a single IR system.

On the other hand, state-of-the-art speech recognition systems still need to limit the vocabulary size (i.e., the number of words in a dictionary), due to problems in estimating statistical language models (Young, 1996) and constraints associated with hardware, such as memories. In addition, computation time is crucial for a real-time usage, including speech-driven retrieval. In view of these problems, for many languages the vocabulary size is limited to a couple of ten thousands (Itou et al., 1999; Paul and Baker, 1992; Steeneken and van Leeuwen, 1995), which is incomparably smaller than the size of indexes for practical IR systems.

In addition, high-frequency words, such as functional words and common nouns, are usually included in dictionaries and recognized with a high accuracy. However, those words are not necessarily useful for retrieval. On the contrary, low-frequency words appearing in specific documents are often effective query terms.

To sum up, the OOV problem is inherent in speech-driven retrieval, and we need to fill the gap between speech recognition and text retrieval in terms of the vocabulary size. In this paper, we propose a method to resolve this problem aiming at open-vocabulary speech-driven retrieval.

3 System Overview

Figure 1 depicts the overall design of our speech-driven text retrieval system, which consists of speech recognition, text retrieval and query completion modules. Although our system is currently implemented for Japanese, our methodology is language-independent. We explain the retrieval process based on this figure.

Given a query spoken by a user, the speech recognition module uses a dictionary and acoustic/language models to generate a transcription of the user speech. During this process, OOV words, which are not listed in the dictionary, are also detected. For this purpose, our language model includes both words and syllables so that OOV words are transcribed as sequences of syllables.

For example, in the case where "*kankitsu* (citrus)" is not listed in the dictionary, this word should be transcribed as /ka N ki tsu/. However, it is possible that this word is mistakenly transcribed, such as /ka N ke tsu/ and /ka N ke tsu ke ko/.

To improve the quality of our system, these syllable sequences have to be transcribed as *words*, which is one of the central issues in this paper. In the case of speech-driven retrieval, where users usually have specific information needs, it is feasible that users utter contents related to a target collection. In other words, there is a great possibility that detected OOV words can be identified as index terms that are phonetically identical or similar.

However, since a) a single sound can potentially correspond to more than one word (i.e., homonyms) and b) searching the entire collection for phonetically identical/similar terms is prohibitive, we need an efficient disambiguation method. Specifically, in the case of Japanese, the homonym problem is multiply crucial because words consist of different character types, i.e., "*kanji*," "*katakana*," "*hiragana*," alphabets and other characters like numerals[1].

To resolve this problem, we use a two-stage retrieval method. In the first stage, we delete OOV words from the transcription, and perform text retrieval using remaining words, to obtain a specific number of top-ranked documents according to the degree of relevance. Even if speech recognition is not perfect, these documents are potentially associated with the user speech more than the entire col-

[1] In Japanese, *kanji* (or Chinese character) is the idiogram, and *katakana* and *hiragana* are phonograms.

lection. Thus, we search only these documents for index terms corresponding to detected OOV words.

Then, in the second stage, we replace detected OOV words with identified index terms so as to complete the transcription, and re-perform text retrieval to obtain final outputs. However, we do not re-perform speech recognition in the second stage.

In the above example, let us assume that the user also utters words related to "*kankitsu* (citrus)," such as "*orenji* (orange)" and "*remon* (lemon)," and that these words are correctly recognized as words. In this case, it is possible that retrieved documents contain the word "*kankitsu* (citrus)." Thus, we replace the syllable sequence /ka N ke tsu/ in the query with "*kankitsu*," which is additionally used as a query term in the second stage.

It may be argued that our method resembles the notion of pseudo-relevance feedback (or local feedback) for IR, where documents obtained in the first stage are used to expand query terms, and final outputs are refined in the second stage (Kwok and Chan, 1998). However, while relevance feedback is used to improve only the retrieval accuracy, our method improves the speech recognition and retrieval accuracy.

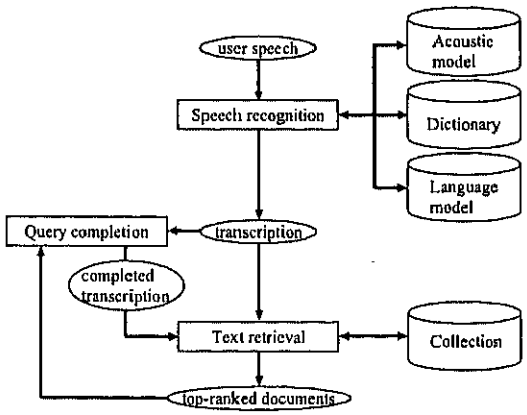

Figure 1: The overall design of our speech-driven text retrieval system.

4 Speech Recognition

The speech recognition module generates word sequence W, given phone sequence X. In a stochastic speech recognition framework (Bahl et al., 1983), the task is to select the W maximizing $P(W|X)$, which is transformed as in Equation (1) through the Bayesian theorem.

$$\arg\max_W P(W|X) = \arg\max_W P(X|W) \cdot P(W) \quad (1)$$

Here, $P(X|W)$ models a probability that word sequence W is transformed into phone sequence X, and $P(W)$ models a probability that W is linguistically acceptable. These factors are usually called acoustic and language models, respectively.

For the speech recognition module, we use the Japanese dictation toolkit (Kawahara et al., 2000)[2], which includes the "Julius" recognition engine and acoustic/language models. The acoustic model was produced by way of the ASJ speech database (ASJ-JNAS) (Itou et al., 1998; Itou et al., 1999), which contains approximately 20,000 sentences uttered by 132 speakers including the both gender groups.

This toolkit also includes development softwares so that acoustic and language models can be produced and replaced depending on the application. While we use the acoustic model provided in the toolkit, we use a new language model including both words and syllables. For this purpose, we used the "ChaSen" morphological analyzer[3] to extract words from ten years worth of "Mainichi Shimbun" newspaper articles (1991-2000).

Then, we selected 20,000 high-frequency words to produce a dictionary. At the same time, we segmented remaining lower-frequency words into syllables based on the Japanese phonogram system. The resultant number of syllable types was approximately 700. Finally, we produced a word/syllable-based trigram language model. In other words, OOV words were modeled as sequences of syllables. Thus, by using our language model, OOV words can easily be detected.

In spoken document retrieval, an open-vocabulary method, which combines recognition methods for words and syllables in target speech documents, was also proposed (Wechsler et al., 1998). However, this method requires an additional computation for recognizing syllables, and thus is expensive. In contrast, since our language model is a regular statistical N-gram model, we can use the same speech recognition framework as in Equation (1).

[2] http://winnie.kuis.kyoto-u.ac.jp/dictation
[3] http://chasen.aist-nara.ac.jp

5 Text Retrieval

The text retrieval module is based on the "Okapi" probabilistic retrieval method (Robertson and Walker, 1994), which is used to compute the relevance score between the transcribed query and each document in a target collection. To produce an inverted file (i.e., an index), we use ChaSen to extract content words from documents as terms, and perform a word-based indexing. We also extract terms from transcribed queries using the same method.

6 Query Completion

6.1 Overview

As explained in Section 3, the basis of the query completion module is to correspond OOV words detected by speech recognition (Section 4) to index terms used for text retrieval (Section 5). However, to identify corresponding index terms efficiently, we limit the number of documents in the first stage retrieval. In principle, terms that are indexed in top-ranked documents (those retrieved in the first stage) and have the same sound with detected OOV words can be corresponding terms.

However, a single sound often corresponds to multiple words. In addition, since speech recognition on a syllable-by-syllable basis is not perfect, it is possible that OOV words are incorrectly transcribed. For example, in some cases the Japanese word "*kankitsu* (citrus)" is transcribed as /ka N ke tsu/. Thus, we also need to consider index terms that are phonetically *similar* to OOV words. To sum up, we need a disambiguation method to select appropriate corresponding terms, out of a number of candidates.

6.2 Formalization

Intuitively, it is feasible that appropriate terms:

- have identical/similar sound with OOV words detected in spoken queries,
- frequently appear in a top-ranked document set,
- and appear in higher-ranked documents.

From the viewpoint of probability theory, possible representations for the above three properties include Equation (2), where each property corresponds to different parameters. Our task is to select the t maximizing the value computed by this equation as the corresponding term for OOV word w.

$$\sum_{d \in D_q} P(w|t) \cdot P(t|d) \cdot P(d|q) \qquad (2)$$

Here, D_q is the top-ranked document set retrieved in the first stage, given query q. $P(w|t)$ is a probability that index term t can be replaced with detected OOV word w, in terms of phonetics. $P(t|d)$ is the relative frequency of term t in document d. $P(d|q)$ is a probability that document d is relevant to query q, which is associated with the score formalized in the Okapi method.

However, from the viewpoint of empiricism, Equation (2) is not necessarily effective. First, it is not easy to estimate $P(w|t)$ based on the probability theory. Second, the probability score computed by the Okapi method is an approximation focused mainly on *relative* superiority among retrieved documents, and thus it is difficult to estimate $P(d|q)$ in a rigorous manner. Finally, it is also difficult to determine the degree to which each parameter influences in the final probability score.

In view of these problems, through preliminary experiments we approximated Equation (2) and formalized a method to compute the degree (not the probability) to which given index term t corresponds to OOV word w.

First, we estimate $P(w|t)$ by the ratio between the number of syllables commonly included in both w and t and the total number of syllables in w. We use a DP matching method to identify the number of cases related to deletion, insertion, and substitution in w, on a syllable-by-syllable basis.

Second, $P(w|t)$ should be more influential than $P(t|d)$ and $P(d|q)$ in Equation (2), although the last two parameters are effective in the case where a large number of candidates phonetically similar to w are obtained. To decrease the effect of $P(t|d)$ and $P(d|q)$, we tentatively use logarithms of these parameters. In addition, we use the score computed by the Okapi method as $P(d|q)$.

According to the above approximation, we compute the score of t as in Equation (3).

$$\sum_{d \in D_q} P(w|t) \cdot \log(P(t|d) \cdot P(d|q)) \qquad (3)$$

It should be noted that Equation (3) is independent of the indexing method used, and therefore t can be any sequences of characters contained in D_q. In other words, any types of indexing methods (e.g., word-based and phrase-based indexing methods) can be used in our framework.

6.3 Implementation

Since computation time is crucial for a real-time usage, we preprocess documents in a target collection so as to identify candidate terms efficiently. This process is similar to the indexing process performed in the text retrieval module.

In the case of text retrieval, index terms are organized in an inverted file so that documents including terms that *exactly* match with query keywords can be retrieved efficiently.

However, in the case of query completion, terms that are included in top-ranked documents need to be retrieved. In addition, to minimize a score computation (for example, DP matching is time-consuming), it is desirable to delete terms that are associated with a diminished phonetic similarity value, $P(w|t)$, prior to the computation of Equation (3). In other words, an index file for query completion has to be organized so that a *partial* matching method can be used. For example, /ka N ki tsu/ has to be retrieved efficiently in response to /ka N ke tsu/.

Thus, we implemented a forward/backward partial-matching method, in which entries can be retrieved by any substrings from the first/last characters. In addition, we index words and word-based bigrams, because preliminary experiments showed that OOV words detected by our speech recognition module are usually single words or short phrases, such as "*ozon-houru* (ozone hole)."

7 Experimentation

7.1 Methodology

To evaluate the performance of our speech-driven retrieval system, we used the IREX collection[4]. This test collection, which resembles one used in the TREC ad hoc retrieval track, includes 30 Japanese topics (information need) and relevance assessment (correct judgement) for each topic, along with target documents. The target documents are 211,853 articles collected from two years worth of "Mainichi Shimbun" newspaper (1994-1995).

Each topic consists of the ID, description and narrative. While descriptions are short phrases related to the topic, narratives consist of one or more sentences describing the topic. Figure 2 shows an example topic in the SGML form (translated into English by one of the organizers of the IREX workshop).

However, since the IREX collection does not contain spoken queries, we asked four speakers (two males/females) to dictate the narrative field. Thus, we produced four different sets of 30 spoken queries. By using those queries, we compared the following different methods:

1. text-to-text retrieval, which used written narratives as queries, and can be seen as a perfect speech-driven text retrieval,

2. speech-driven text retrieval, in which only words listed in the dictionary were modeled in the language model (in other words, the OOV word detection and query completion modules were not used),

3. speech-driven text retrieval, in which OOV words detected in spoken queries were simply deleted (in other words, the query completion module was not used),

4. speech-driven text retrieval, in which our method proposed in Section 3 was used.

In cases of methods 2-4, queries dictated by four speakers were used independently. Thus, in practice we compared 13 different retrieval results. In addition, for methods 2-4, ten years worth of *Mainichi Shimbun* Japanese newspaper articles (1991-2000) were used to produce language models. However, while method 2 used only 20,000 high-frequency words for language modeling, methods 3 and 4 also used syllables extracted from lower-frequency words (see Section 4).

Following the IREX workshop, each method retrieved 300 top documents in response to each query, and non-interpolated average precision values were used to evaluate each method.

[4] http://cs.nyu.edu/cs/projects/proteus/irex/index-e.html

```
<TOPIC><TOPIC-ID>1001</TOPIC-ID>
<DESCRIPTION>Corporate merging</DESCRIPTION>
<NARRATIVE>The article describes a corporate merging and in the article, the
name of companies have to be identifiable. Information including the field
and the purpose of the merging have to be identifiable. Corporate merging
includes corporate acquisition, corporate unifications and corporate buy-
ing.</NARRATIVE></TOPIC>
```

Figure 2: An English translation for an example topic in the IREX collection.

7.2 Results

First, we evaluated the performance of detecting OOV words. In the 30 queries used for our evaluation, 14 word *tokens* (13 word *types*) were OOV words unlisted in the dictionary for speech recognition. Table 1 shows the results on a speaker-by-speaker basis, where "#Detected" and "#Correct" denote the total number of OOV words detected by our method and the number of OOV words correctly detected, respectively. In addition, "#Completed" denotes the number of detected OOV words that were corresponded to correct index terms in 300 top documents.

It should be noted that "#Completed" was greater than "#Correct" because our method often mistakenly detected words in the dictionary as OOV words, but completed them with index terms correctly. We estimated recall and precision for detecting OOV words, and accuracy for query completion, as in Equation (4).

$$\begin{aligned} recall &= \frac{\#Correct}{14} \\ precision &= \frac{\#Correct}{\#Detect} \\ accuracy &= \frac{\#Completed}{\#Detect} \end{aligned} \tag{4}$$

Looking at Table 1, one can see that recall was generally greater than precision. In other words, our method tended to detect as many OOV words as possible. In addition, accuracy of query completion was relatively low.

Figure 3 shows example words in spoken queries, detected as OOV words and correctly completed with index terms. In this figure, OOV words are transcribed with syllables, where " : " denotes a long vowel. Hyphens are inserted between Japanese words, which inherently lack lexical segmentation.

Second, to evaluate the effectiveness of our query completion method more carefully, we compared retrieval accuracy for methods 1-4 (see Section 7.1). Table 2 shows average precision values, averaged over the 30 queries, for each method[5]. The average precision values of our method (i.e., method 4) was approximately 87% of that for text-to-text retrieval.

By comparing methods 2-4, one can see that our method improved average precision values of the other methods irrespective of the speaker. To put it more precisely, by comparing methods 3 and 4, one can see the effectiveness of the query completion method. In addition, by comparing methods 2 and 4, one can see that a combination of the OOV word detection and query completion methods was effective.

It may be argued that the improvement was relatively small. However, since the number of OOV words inherent in 30 queries was only 14, the effect of our method was overshadowed by a large number of other words. In fact, the number of words used as query terms for our method, averaged over the four speakers, was 421. Since existing test collections for IR research were not produced to explore the OOV problem, it is difficult to derive conclusions that are statistically valid. Experiments using larger-scale test collections where the OOV problem is more crucial need to be further explored.

Finally, we investigated the time efficiency of our method, and found that CPU time required for the query completion process per detected OOV word was 3.5 seconds (AMD Athlon MP 1900+). However, an additional CPU time for detecting OOV words, which can be performed in a conventional speech recognition process, was not crucial.

[5]Average precision is often used to evaluate IR systems, which should not be confused with evaluation measures in Equation (4).

Table 1: Results for detecting and completing OOV words.

Speaker	#Detected	#Correct	#Completed	Recall	Precision	Accuracy
Female #1	51	9	18	0.643	0.176	0.353
Female #2	56	10	18	0.714	0.179	0.321
Male #1	33	9	12	0.643	0.273	0.364
Male #2	37	12	16	0.857	0.324	0.432
Total	176	40	64	0.714	0.226	0.362

OOV words	Index terms (syllables)	English gloss
/gu re : pu ra chi na ga no/	*gureepu-furuutsu* /gu re : pu fu ru : tsu/	grapefruit
/ya yo i chi ta/	*Yayoi-jidai* /ya yo i ji da i/	the *Yayoi* period
/ni ku ku ra i su/	*nikku-puraisu* /ni q ku pu ra i su/	Nick Price
/be N pi/	*benpi* /be N pi/	constipation

Figure 3: Example words detected as OOV words and completed correctly by our method.

7.3 Analyzing Errors

We manually analyzed seven cases where the average precision value of our method was significantly lower than that obtained with method 2 (the total number of cases was the product of numbers of queries and speakers).

Among these seven cases, in five cases our query completion method selected incorrect index terms, although correct index terms were included in top-ranked documents obtained with the first stage. For example, in the case of the query 1021 dictated by a female speaker, the word "*seido* (institution)" was mistakenly transcribed as /se N do/. As a result, the word "*sendo* (freshness)," which is associated with the same syllable sequences, was selected as the index term. The word "*seido* (institution)" was the third candidate based on the score computed by Equation (3). To reduce these errors, we need to enhance the score computation.

In another case, our speech recognition module did not correctly recognize words in the dictionary, and decreased the retrieval accuracy.

In the final case, a fragment of a narrative sentence consisting of ten words was detected as a single OOV word. As a result, our method, which can complete up to two word sequences, mistakenly processed that word, and decreased the retrieval accuracy. However, this case was exceptional. In most cases, functional words, which were recognized with a high accuracy, segmented OOV words into shorter fragments.

Table 2: Non-interpolated average precision values, averaged over 30 queries, for different methods.

Speaker\Method	1	2	3	4
Female #1	–	0.2831	0.2834	0.3195
Female #2	–	0.2745	0.2443	0.2846
Male #1	–	0.3005	0.2987	0.3179
Male #2	–	0.2787	0.2675	0.2957
Total	0.3486	0.2842	0.2734	0.3044

8 Related Work

The method proposed by Kupiec et al. (1994) and our method are similar in the sense that both methods use target collections as language models for speech recognition to realize open-vocabulary speech-driven retrieval.

Kupiec et al's method, which is based on word recognition and accepts only short queries, derives multiple transcription candidates (i.e., possible word combinations), and searches a target collection for the most plausible word combination. However, in the case of longer queries, the number of candidates increases, and thus the searching cost is prohibitive. This is a reason why operational speech recognition systems have to limit the vocabulary size.

In contrast, our method, which is based on a recent *continuous* speech recognition framework, can accept longer sentences. Additionally, our method uses a two-stage retrieval principle to limit a search space in a target collection, and disambiguates only detected OOV words. Thus, the computation cost can be minimized.

9 Conclusion

To facilitate retrieving information by spoken queries, the out-of-vocabulary problem in speech recognition needs to be resolved. In our proposed method, out-of-vocabulary words in a query are detected by speech recognition, and completed with terms indexed for text retrieval, so as to improve the recognition accuracy. In addition, the completed query is used to improve the retrieval accuracy. We showed the effectiveness of our method by using dictated queries in the IREX collection. Future work would include experiments using larger-scale test collections in various domains.

References

Lalit. R. Bahl, Frederick Jelinek, and Robert L. Mercer. 1983. A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(2):179–190.

J. Barnett, S. Anderson, J. Broglio, M. Singh, R. Hudson, and S. W. Kuo. 1997. Experiments in spoken queries for document retrieval. In *Proceedings of Eurospeech97*, pages 1323–1326.

Fabio Crestani. 2000. Word recognition errors and relevance feedback in spoken query processing. In *Proceedings of the Fourth International Conference on Flexible Query Answering Systems*, pages 267–281.

Atsushi Fujii, Katunobu Itou, and Tetsuya Ishikawa. 2002. Speech-driven text retrieval: Using target IR collections for statistical language model adaptation in speech recognition. In Anni R. Coden, Eric W. Brown, and Savitha Srinivasan, editors, *Information Retrieval Techniques for Speech Applications (LNCS 2273)*, pages 94–104. Springer.

John S. Garofolo, Ellen M. Voorhees, Vincent M. Stanford, and Karen Sparck Jones. 1997. TREC-6 1997 spoken document retrieval track overview and results. In *Proceedings of the 6th Text REtrieval Conference*, pages 83–91.

K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi. 1998. The design of the newspaper-based Japanese large vocabulary continuous speech recognition corpus. In *Proceedings of the 5th International Conference on Spoken Language Processing*, pages 3261–3264.

Katunobu Itou, Mikio Yamamoto, Kazuya Takeda, Toshiyuki Takezawa, Tatsuo Matsuoka, Tetsunori Kobayashi, and Kiyohiro Shikano. 1999. JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research. *Journal of Acoustic Society of Japan*, 20(3):199–206.

Katunobu Itou, Atsushi Fujii, and Tetsuya Ishikawa. 2001. Language modeling for multi-domain speech-driven text retrieval. In *IEEE Automatic Speech Recognition and Understanding Workshop*.

Pierre Jourlin, Sue E. Johnson, Karen Spärck Jones, and Philip C. Woodland. 2000. Spoken document representations for probabilistic retrieval. *Speech Communication*, 32:21–36.

T. Kawahara, A. Lee, T. Kobayashi, K. Takeda, N. Minematsu, S. Sagayama, K. Itou, A. Ito, M. Yamamoto, A. Yamada, T. Utsuro, and K. Shikano. 2000. Free software toolkit for Japanese large vocabulary continuous speech recognition. In *Proceedings of the 6th International Conference on Spoken Language Processing*, pages 476–479.

Julian Kupiec, Don Kimber, and Vijay Balasubramanian. 1994. Speech-based retrieval using semantic co-occurrence filtering. In *Proceedings of the ARPA Human Language Technology Workshop*, pages 373–377.

K.L. Kwok and M. Chan. 1998. Improving two-stage ad-hoc retrieval for short queries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 250–256.

Douglas B. Paul and Janet M. Baker. 1992. The design for the Wall Street Journal-based CSR corpus. In *Proceedings of DARPA Speech & Natural Language Workshop*, pages 357–362.

S.E. Robertson and S. Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 232–241.

Herman J. M. Steeneken and David A. van Leeuwen. 1995. Multi-lingual assessment of speaker independent large vocabulary speech-recognition systems: The SQALE-project. In *Proceedings of Eurospeech95*, pages 1271–1274.

Martin Wechsler, Eugen Munteanu, and Peter Schäuble. 1998. New techniques for open-vocabulary spoken document retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 20–27.

Steve Young. 1996. A review of large-vocabulary continuous-speech recognition. *IEEE Signal Processing Magazine*, pages 45–57, September.

LANGUAGE MODELING FOR MULTI-DOMAIN SPEECH-DRIVEN TEXT RETRIEVAL

Katunobu Itou[1], *Atsushi Fujii*[2], *Tetsuya Ishikawa*[2]

[1] National Institute of Advanced Industrial Science and Technology
1-1-1 Chuuou Daini Umezono, Tsukuba, 305-8568, Japan, E-mail: itou@ni.aist.go.jp
[2] University of Library and Information Science
1-2 Kasuga, Tsukuba, 305-8550, Japan, E-mail: {fujii,ishikawa}@ulis.ac.jp

ABSTRACT

We report experimental results associated with speech-driven text retrieval, which facilitates retrieving information in multiple domains with spoken queries. Since users speak contents related to a target collection, we produce language models used for speech recognition based on the target collection, so as to improve both the recognition and retrieval accuracy. Experiments using existing test collections combined with dictated queries showed the effectiveness of our method.

1. INTRODUCTION

Automatic speech recognition, which decodes human voice to generate transcriptions, has of late become a practical technology. It is feasible that speech recognition is used in real world computer-based applications, specifically, those associated with human language. In fact, a number of speech-based methods have been explored in the information retrieval (IR) community, which can be classified into the following two fundamental categories:

- spoken document retrieval, in which written queries are used to search speech (e.g., broadcast news audio) archives for relevant speech information [1].
- speech-driven retrieval, in which spoken queries are used to retrieve relevant textual information [2, 3].

Initiated partially by the TREC-6 spoken document retrieval (SDR) track [1], various methods have been proposed for spoken document retrieval. However, a relatively small number of methods have been explored for speech-driven text retrieval, although they are associated with numerous keyboard-less retrieval applications, such as telephone-based retrieval, car navigation systems, and user-friendly interfaces.

Barnett et al. [2] performed comparative experiments related to speech-driven retrieval, where the DRAGON speech recognition system was used as an input interface for the INQUERY text retrieval system. They used as test inputs 35 queries collected from the TREC topics and dictated by a single male speaker. Crestani [3] also used the above 35 queries and showed that conventional relevance feedback techniques marginally improved the accuracy for speech-driven text retrieval.

These above cases focused solely on improving text retrieval methods and did not address problems of improving speech recognition accuracy. In fact, an existing speech recognition system was used with no enhancement. In other words, speech recognition and text retrieval modules were fundamentally independent and were simply connected by way of an input/output protocol.

However, since most speech recognition systems are trained based on specific domains, the accuracy of speech recognition across domains is not satisfactory. Thus, as can easily be predicted, in cases of Barnett et al. [2] and Crestani [3], a speech recognition error rate was relatively high and considerably decreased the retrieval accuracy. Additionally, speech recognition with a high accuracy is crucial for interactive retrieval, such as dialog-based retrieval.

Motivated by these problems, in this paper we integrate (not simply connect) speech recognition and text retrieval to improve both recognition and retrieval accuracy in the context of speech-driven text retrieval.

Unlike general-purpose speech recognition aimed to decode any spontaneous speech, in the case of speech-driven text retrieval, users usually speak contents associated with a target collection, from which documents relevant to their information need are retrieved. In a stochastic speech recognition framework, the accuracy depends primarily on acoustic and language models [4]. While acoustic models are related to phonetic properties, language models, which represent linguistic contents to be spoken, are related to target collections. Thus, it is intuitively feasible that language models have to be produced based on target collections.

To sum up, our belief is that by adapting a language model based on a target IR collection, we can improve the speech recognition and text retrieval accuracy, simultaneously.

Section 2 describes our speech-driven text retrieval system, which is currently implemented for Japanese. Section 3 elaborates on comparative experiments, in which IR test collections in different domains are used to evaluate the effectiveness of our system.

2. SYSTEM DESCRIPTION

2.1. Overview

Figure 1 depicts the overall design of our speech-driven text retrieval system, which consists of speech recognition and text retrieval modules. In the following sections, we explain two modules in Figure 1, respectively.

The first and second authors are also members of CREST, Japan Science and Technology Corporation.

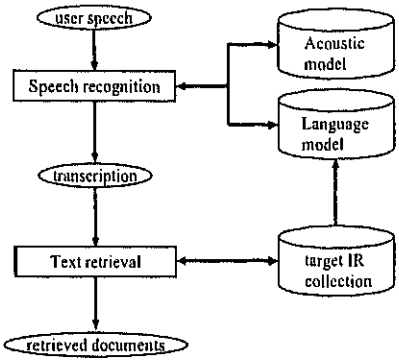

Fig. 1. The design of our speech-driven text retrieval system.

2.2. Speech Recognition

For the speech recognition module, we use the Japanese dictation toolkit [5][1], which includes the "Julius" recognition engine and acoustic/language models. Julius performs a two-pass (forward-backward) search using word-based forward bigrams and backward trigrams on the respective passes.

The acoustic model was produced by way of the ASJ speech databases of phonetically balanced sentences (ASJ-PB) and newspaper articles texts (ASJ-JNAS) [6], which contain approximately 20,000 sentences uttered by 132 speakers including the both gender groups. We used a 16-mixture Gaussian distribution triphone Hidden Markov Model, where states were clustered into 2,000 groups by a state-tying method.

This toolkit also includes development softwares, so that acoustic and language models can be produced and replaced depending on the application. While we use the acoustic model provided in the toolkit, we use new language models produced by way of source documents (i.e., target IR collections).

2.3. Text Retrieval

The text retrieval module is based on the "Okapi" method [7], which computes the relevance score between the transcribed query and each document in the collection, based on the distribution of index terms, and sorts retrieved documents according to the score in descending order.

We use content words extracted from documents as index terms, and perform a word-based indexing. For this purpose, we use the ChaSen morphological analyzer [8] to extract content words. We extract terms from transcribed queries using the same method.

3. EXPERIMENTATION

3.1. Test Collections

To investigate the performance of our multi-domain speech-driven retrieval system, we used two different types of Japanese IR test (benchmark) collections: the NTCIR and IREX collections. Both collections, which resemble one used in the TREC ad hoc retrieval track, include topics (information need) and relevance assessment (correct judgement) for each topic, along with target documents. However, these collections are associated with different domain, respectively.

The NTCIR collection [9][2] includes 736,166 abstracts collected from technical papers published by 65 Japanese associations for various fields. On the other hand, the IREX collection [10][3] includes 211,853 articles collected from two years worth of "Mainichi Shimbun" newspaper articles[4].

The NTCIR and IREX collections include 132 and 30 Japanese topics, respectively, for a sample of which English translations are also provided. Figures 2 and 3 show example topics in each collection, which consist of different fields (for example, descriptions and narratives) tagged in an SGML form.

Since both collections do not contain spoken queries, we asked four speakers (two males/females) to dictate topics. For this purpose, we selectively used a specific field, so as to simulate a realistic speech-driven retrieval.

In the case of the NTCIR topics, titles are not informative for the retrieval. On the other hand, narratives, which usually consist of several sentences, are too long to speak. Thus, only descriptions, which consist of a single phrase and sentence, were dictated by each speaker, so as to produce four different sets of 132 spoken queries. However, in the case of the IREX topics, since descriptions are not informative for the retrieval, only narratives were dictated by each speaker, to produce four different sets of 30 spoken queries.

3.2. Comparative Evaluation

We compared the performance of the following retrieval methods:

- text-to-text retrieval, which used written queries, and can be seen as the perfect speech-driven text retrieval,
- speech-driven text retrieval, in which a language model produced based on the NTCIR collection was used,
- speech-driven text retrieval, in which a language model produced based on the IREX collection was used.

In cases of speech-driven text retrieval methods, queries dictated by four speakers were used independently, and the final result was obtained by averaging results for different speakers.

Although the Julius decoder outputs more than one transcription candidates for a single speech, we used only the one with the greatest probability score. The results did not significantly change depending on whether or not we used lower-ranked transcriptions as queries.

The only difference in producing two different language models (i.e., those based on the NTCIR and IREX collections) is the source documents. In other words, both language models were of the same vocabulary size (20,000), and were produced by way of the same softwares.

Table 1 shows statistics related to word tokens/types in two different collections for language modeling, where the line "Coverage" denotes the ratio of word tokens contained in the resultant language model. Most of word tokens were covered irrespective of the collection.

[1] http://winnie.kuis.kyoto-u.ac.jp/dictation/

[2] http://research.nii.ac.jp/ntcadm/index-en.html

[3] http://cs.nyu.edu/cs/projects/proteus/irex/index-e.html

[4] In practice, the IREX collection provides only article IDs, which corresponds to articles in Mainichi Shimbun newspaper CD-ROM'94-'95. Participants must get a copy of the CD-ROMs themselves.

```
<TOPIC q=0123>
<TITLE>Biofilms</TITLE>
<DESCRIPTION>Are there any documents about the biofilms produced by some microorganisms in
which chronic diseases are mentioned?</DESCRIPTION>
<NARRATIVE>Biofilms are thought to occur when microorganisms grow in microcolonies embedded
in the adherent gel surface on tunica mucosa, and teeth, or on catheters, prosthetic valves,
and other artifacts. A relevant document will report on any studies into the relationship
between biofilms produced by some microorganisms and chronic diseases. Documents that in-
clude reports on biofilms produced by non-medical microorganisms that do not cause infectious
diseases are not relevant.</NARRATIVE>
</TOPIC>
```

Fig. 2. An English translation for an example topic in the NTCIR collection.

```
<TOPIC>
<TOPIC-ID>1001</TOPIC-ID>
<DESCRIPTION>Corporate merging</DESCRIPTION>
<NARRATIVE>The article describes a corporate merging and in the article, the name of compa-
nies have to be identifiable. Information including the field and the purpose of the merging
have to be identifiable. Corporate merging includes corporate acquisition, corporate unifi-
cations and corporate buying.</NARRATIVE>
</TOPIC>
```

Fig. 3. An English translation for an example topic in the IREX collection.

Table 1. Statistics related to source words for language modeling.

	NTCIR	IREX
# of Types	454K	179K
# of Tokens	175M	53M
Coverage	97.9%	96.5%

Each method retrieved 1,000 top documents, and the TREC evaluation software was used to calculate non-interpolated average precision values and plot recall-precision curves.

Table 2 shows the non-interpolated average precision values (AP) and word error rate in speech recognition, for different retrieval methods. As with existing experiments for speech recognition, word error rate (WER) is the ratio between the number of word errors (i.e., deletion, insertion, and substitution) and the total number of words. In addition, we investigated error rate with respect to query terms (i.e., keywords used for retrieval), which we shall call "term error rate (TER)". Table 2 also shows trigram test-set perplexity (PP) and test-set out-of-vocabulary rate (OOV).

It should noted that for all the evaluation measures in Table 2 excepting average precision, smaller values are generally obtained with better methods. Suggestions which can be derived from these results are as follows.

First, by comparing results of different language models, one can see that the performance was significantly improved with a language model produced from the target collection, which was observable irrespective of the domain. Thus, producing language models based on target collections was quite effective for speech-driven text retrieval.

Second, while in the case of the NTCIR collection, the average precision for speech-driven retrieval was approximately 77% of that obtained with text-to-text retrieval, in the case of the IREX collection, the average precision for speech-driven retrieval was quite comparable that obtained with text-to-text retrieval.

Third, TER was generally higher than WER irrespective of the speaker. In other words, speech recognition for content words was more difficult than functional words, which were not contained in query terms.

Finally, we investigated the trade-off between recall and precision. Figures 4 and 5 show recall-precision curves of different retrieval methods, for the NTCIR and IREX collections, respectively. In these figures, the relative superiority for precision values due to different language models in Table 2 was also observable, regardless of the recall.

4. CONCLUSION

Aiming at speech-driven text retrieval with a high accuracy, we proposed a method to integrate speech recognition and text retrieval methods, in which target text collections are used to produce statistical language models for speech recognition. We also showed the effectiveness of our method by way of experiments, where dictated information needs in the NTCIR/IREX collections were used as queries to retrieve documents in different domains.

Acknowledgments

The authors would like to thank the National Institute of Informatics for their support with the NTCIR collection and the IREX committee for their support with the IREX collection.

Table 2. Results for different retrieval methods targeting the NTCIR/IREX collections (AP: average precision, WER: word error rate, TER: term error rate, PP: trigram test-set perplexity, OOV: test-set Out-of-Vocabulary rate).

	NTCIR					IREX				
Language Model	AP	WER	TER	PP	OOV	AP	WER	TER	PP	OOV
Text	0.337	—	—	—	—	0.367	—	—	—	—
NTCIR	0.261	18.6%	23.6%	60	4.2%	0.166	31.1%	41.0%	138	6.1%
IREX	0.111	41.4%	54.6%	195	9.4%	0.334	19.5%	22.9%	108	1.4%

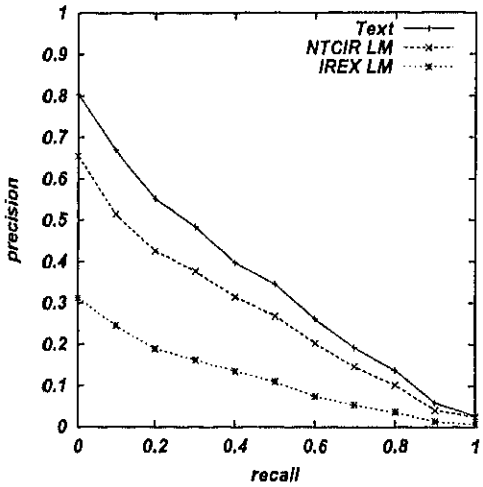

Fig. 4. Recall-precision curves for different methods targeting the NTCIR collection.

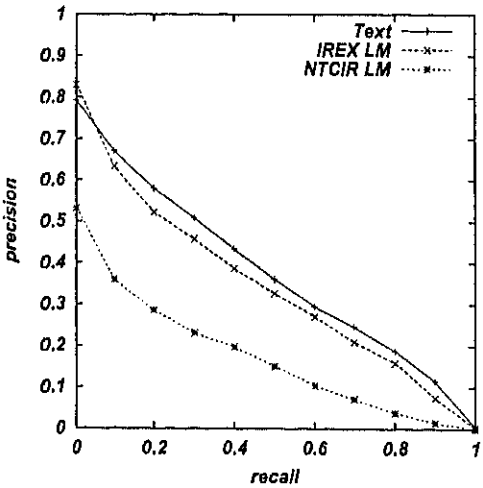

Fig. 5. Recall-precision curves for different methods targeting the IREX collection.

5. REFERENCES

[1] John S. Garofolo, Ellen M. Voorhees, Vincent M. Stanford, and Karen Sparck Jones, "TREC-6 1997 spoken document retrieval track overview and results," in *Proceedings of the 6th Text REtrieval Conference*, 1997, pp. 83–91.

[2] J. Barnett, S. Anderson, J. Broglio, M. Singh, R. Hudson, and S. W. Kuo, "Experiments in spoken queries for document retrieval," in *Proceedings of Eurospeech97*, 1997, pp. 1323–1326.

[3] Fabio Crestani, "Word recognition errors and relevance feedback in spoken query processing," in *Proceedings of the Fourth International Conference on Flexible Query Answering Systems*, 2000, pp. 267–281.

[4] Lalit. R. Bahl, Frederick Jelinek, and Robert L. Mercer, "A maximum linklihood approach to continuous speech recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 5, no. 2, pp. 179–190, 1983.

[5] T. Kawahara, A. Lee, T. Kobayashi, K. Takeda, N. Minematsu, S. Sagayama, K. Itou, A. Ito, M. Yamamoto, A. Yamada, T. Utsuro, and K. Shikano, "Free software toolkit for Japanese large vocabulary continuous speech recognition," in *Proceedings of the 6th International Conference on Spoken Language Processing*, 2000, pp. 476–479.

[6] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "The design of the newspaper-based Japanese large vocabulary continuous speech recognition corpus," in *ICSLP-98*, 1998, pp. 3261–3264.

[7] S. E. Robertson and S. Walker, "Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval," in *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1994, pp. 232–241.

[8] Yuji Matsumoto, Akira Kitauchi, Tatsuo Yamashita, Yoshitaka Hirano, Hiroshi Matsuda, and Masayuki Asahara, "Japanese morphological analysis system ChaSen version 2.0 manual 2nd edition," Tech. Rep. NAIST-IS-TR99009, NAIST, 1999.

[9] National Institute of Informatics, *Proceedings of the 2nd NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization*, 2001.

[10] Satoshi Sekine and Hitoshi Isahara, "IREX: IR and IE evaluation project in Japanese," in *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, 2000, pp. 1475–1480.

講演音声の認識における言語モデルのタスク適応*

△坂本尚美 (図書館情報大), △藤井敦, △石川徹也 (筑波大),
○秋葉友良, 伊藤克亘 (産総研)

1 まえがき

講演のビデオデータを対象にして、要求に応じた内容を視聴するオンデマンド講演 (Lecture-On-Demand: LOD) システムの研究開発を行っている [6]。本システムを用いると、教科書や予稿などの講演資料を閲覧しながら、関心があるビデオ内容 (音声と画像) を選択的に視聴する事が可能である。本稿では、本システムにおいて講演音声を音声認識する際に使用する言語モデルのタスク適応手法を提案し、評価する。

2 提案する言語モデルのタスク適応手法

2.1 関連文書の検索

本研究では、講演ビデオデータとして「放送大学」を対象とする。LOD システムにおいて、放送大学の教科書は、視聴者 (ユーザ) が講義内容を検索するための手掛かりとして利用される。他方において、教科書を認識対象のタスク適応テキストコーパスとしても利用する点に提案手法の特長がある。

提案手法の概要を図 1 に示す。一般コーパスとして、WWW から収集した 10M ページコーパス [7](WEB コーパス) を用いる。認識する講義に該当する教科書を検索質問として類似文書検索 [1] を行い、WEB コーパスからタスク関連文書 (WEB ページ) を取得する。取得した文書集合と教科書をタスク適応コーパスとする。また認識用辞書の語彙は、タスク適応コーパスの高頻度語と、教科書に出現する全単語から、重複を除いて選択する。

2.2 学習コーパスの正規化

WEB コーパスは、多種多様な内容を含むコーパスであり、また言語表現にも統制がない。そこで、形態素解析後のコーパスに対し、「異表記統一」と「複合語分割」の 2 種類の正規化処理を行う。

「異表記統一」では、表記の揺れをまとめ、一つの辞書エントリになるように統一する。例えば、「申し込み」「申込み」「申込」「申しこみ」のような表記の違いを検出し、最頻出の代表表記に統一する。

「複合語分割」では、形態素解析の結果として得られた複合語を、それを構成する単語へと分割する。形態素解析システムに登録されている複合語には一貫性がなく、言語モデルの辞書エントリとしてそのまま登録するには弊害があると考えたためである。

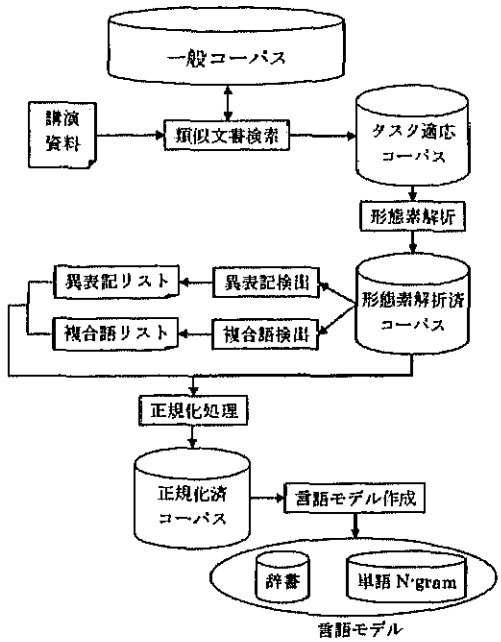

図 1: 言語モデル作成の概要

3 評価実験

放送大学の講義を音声認識し、タスク適応した言語モデルの評価を行った。6 種類の講義を選択し、それぞれ 1 回分の講義 45 分を対象とした。

WEB コーパス全体 (10M) から、高頻度語 2 万語の tri-gram 言語モデルを作成しベースラインとした。次に提案手法により、関連文書数を上位 1,000 件 (1K)、2,000 件 (2K)、10,000 件 (10K) 検索してタスク適応コーパスを抽出し、2 万語+α(教科書にのみ含まれる単語による増加分) の tri-gram 言語モデルを作成した。

音声認識デコーダには JULIUS[8] を使用した。音響モデルは、教師無し MLLR 法で話者適応した。

各言語モデルに対する、作成に使用した文書 (WEB ページ) 数、文書集合のサイズ (のべ語数)、認識対象となる講義音声の書き起こしテキストに対する未知語率 (OOV)、テストセット・パープレキシティ(Perplexity)、音声認識実験による単語誤り率 (WER)、を表 1 に示す。

「現代モード論」を除くすべての講義で認識率の改善が見られ、言語モデル適応の効果を確認する事ができた。しかし、個々の講義でタスク適応の効果が大きく異なる。各講義音声の言語的特徴を明らかにし、各特徴に応じた適応手法を適用することで、認識率をさらに改善できる可能性がある。

本稿で検討した範囲 (1,000〜10,000 件) での関連文

* "Adapting Language Models in Recognizing Lecture Speech" by Naomi SAKAMOTO (University of Library and Information Science), Atsushi FUJII, Tetsuya ISHIKAWA (University of Tsukuba), Tomoyosi AKIBA, and Katunobu ITOU (National Institute of Advanced Industrial Science and Technology)

表 1: 実験結果

講義名	検索文書数	サイズ (語)	OOV(%)	Perprexity	WER(%)
法と裁判	10M	25,837,418	4.44	48.9	20.9
	1K	4,970,889	1.89	46.6	13.9
	2K	8,735,505	2.03	43.3	13.3
古典古代の歴史	10M	25,837,418	7.29	122.1	51.6
	1K	3,457,168	8.27	96.7	42.3
	2K	6,873,183	8.21	91.7	44.1
太陽系の科学	10M	25,837,418	5.11	163.3	63.7
	1K	1,219,166	5.27	130.0	48.2
	2K	2,341,807	5.13	125.5	48.1
現代モード論	10M	25,837,418	3.82	83.6	62.4
	1K	5,661,675	7.18	171.0	64.7
	2K	10,532,078	7.31	157.0	64.3
	10K	41,353,533	7.49	127.9	63.9
食物とからだ	10M	25,837,418	5.28	89.3	48.9
	1K	3,967,952	4.61	146.9	44.0
	2K	5,544,780	4.48	127.0	43.2
	10K	17,024,551	4.10	107.5	41.6
家族法	10M	25,837,418	3.92	136.3	60.4
	1K	3,923,692	4.84	186.4	56.5
	2K	7,576,822	4.69	178.2	56.5
	10K	35,657,505	4.91	131.9	54.3

書数の増加は、総じてテストセットパープレキシティを引き下げ、「古典古代の歴史」を除いて単語誤り率を改善している。文書数と言語モデル改善の関係を調べるには、さらなる調査が必要である。

「法と裁判」と「食物とからだ」を除き、類似文書検索による言語モデルは未知語率を増加させた。特に「現代モード論」は、未知語率の大きな増加が単語誤り率の悪化の原因となったと考えられる。未知語率の増加を抑えるための改善手法としては、WEB コーパス全体と検索された文書による部分的なコーパスを重み付きで混合する手法 [5] が有効であろう。

4 関連研究

西村ら [10] は、本研究と同様に放送大学を対象としている。しかし、本研究との差異は講義冒頭 60 文のみを認識している点にある。講義冒頭は、挨拶や概要説明などを丁寧に発話する傾向にあるため、一般に認識率は高い。言語モデルは、講義書き起こしコーパスを利用した講義調発話スタイルの獲得や、間投詞などの不要語の処理を行っているが、専門用語などへのタスク適応は行っていない。

山本ら [11] は、放送大学を対象とし、教科書の索引に現れる専門用語を言語モデルの未知語カテゴリとして登録した言語モデルを用いている。Niesler ら [4] や南條ら [9] は、教科書などの講演資料を用いない手法を示している。一般コーパスから作成した言語モデルによる認識結果を利用し、タスク適応コーパスを一般コーパスから選択している。

一般コーパスから適応するタスクに関連する部分集合を抽出し適応コーパスとして利用する手法 [2, 3] が提案されている。関連文選択の尺度としては、初期適応コーパスに対するパープレキシティ減少の基準 [2] や情報検索の分野で広く利用されている TF.IDF[3] があり、本手法では TF.IDF を用いた。

5 まとめ

講演音声の音声認識において、講演資料を用いた類似文書検索によって言語モデルをタスクに適応させる手法を提案した。6 分野に関する講義を対象にした実験から、本手法によって音声認識率を改善できることが分かった。

参考文献

[1] A. Fujii and K. Itou. Evaluating speech-driven IR in the NTCIR-3 Web retrieval task. In *Proceedings of the Third NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering*, 2003.

[2] D. Klakow. Selecting articles from the language model training corpus. In *Proceedings of International Conference on Acoustics Speech and Signal Processing*, Vol. 3, pp. 1695–1698, Istanbul, Turkey, June 2000.

[3] M. Mahajan, D. Beeferman, and X. D. Huang. Improved topic-dependent language modeling using information retrieval techniques. In *Proceedings of International Conference on Acoustics Speech and Signal Processing*, Phoenix, Arizona, March 1999.

[4] T. Niesler and D. Willett. Unsupervised language model adaptation for lecture speech transcription. In *Proceedings of International Conference on Spoken Language Processing*, Vol. 2, pp. 1413–1416, Denver, Colorado, Sept. 2002.

[5] 伊藤, 好田. N-gram 出現回数の混合によるタスク適応の性能解析. 信学論, J83-D-II(11):2418–2427, 2000.

[6] 伊藤, 藤井, 石川. 音声文書検索を用いたオンデマンド講義システム. 情報処理学会研究報告 SLP-39, pp. 165–170, 2001.

[7] 伊藤, 秋葉, 藤井. WWW は大語彙連続音声認識の学習データとして使えるか? 日本音響学会秋季研究発表会講演論文集, pp. 131–132, 2002.

[8] 鹿野, 伊藤, 河原, 武田, 山本 (編). 音声認識システム. オーム社, 2001.

[9] 南條, 河原. 講演音声認識における言語モデル適応の検討. 日本音響学会秋期研究発表会講演資料, pp. 185–186, 2002.

[10] 西村, 伊東. 講義コーパスを用いた自由発話の大語彙連続音声認識. 信学論, J83-D-II(11):2473–2480, 2000.

[11] 山本, 緒方, 有木. トピックセグメンテーションに基づく講義ビデオの構造化とそれを用いた学習支援システムの検討. 日本音響学会秋期研究発表会講演資料, pp. 207–208, 2002.

音声入力型情報検索のための自由発話収録*

○秋葉友良, 伊藤克亘 (産総研), △藤井敦 (筑波大学)

1 まえがき

近年の大語彙音声認識技術の発展に伴い、様々なアプリケーションに音声入力手段を利用する可能性が広がっている。中でも、音声入力による情報検索 [2, 5] は、キーボードリテラシーがないユーザ、非晴眼者、携帯端末ユーザなどを支援する手段として有効である。

従来の音声入力型情報検索では、キーワードリストや整った検索文をユーザに発声してもらうといった、キーボード入力を音声入力に置換しただけの方法であった。しかし、キーボード入力に比べた音声入力の特長は、思考の即時表現可能性や、表現に要する省労力性にあり、これらを活かした方法が望まれる。

キーワードリストや整った文を入力とする従来の情報検索は、はっきりと言語化された情報要求 (Taylor の 4 階層 [6] での「調整済みの要求 (compromised need)」や「形式化された要求 (formalized need)」) を扱うものであった。音声入力の特長を活用することで、ユーザが十分に言語化できていない漠然とした情報要求 (Taylor の「意識された要求 (conscious need)」) を持つ場合でも、とりあえず話し始める事で情報検索するという新たな利用方法が期待できる。

以上の背景から、我々は発話形態の制限を設けない自由発話音声入力による音声入力情報検索を目指し、システムを開発中である。本稿では、自由発話による検索要求のテストコレクションを収集するための手法を提案し、提案手法による音声収録実験を通して得られた自由発話音声データの分析結果を報告する。

2 収録実験

2.1 方針

情報検索のテストコレクションを構築するには、あらかじめ注意深く設計した検索課題に対して発話を収集することが望ましい。一方、検索者の自由発話を引き出すためには、検索要求の言語表現や表現スタイルをできるだけ制限しないように注意する必要がある。

そこで、検索課題を直接表した検索質問文の代わりに、検索要求の背景の詳しい説明を記した文章を被験者に提示した。これによって、被験者に適合文書集合をイメージさせ、検索質問を間接的に想像させる。また、検索質問を文字列としてそのまま記憶し発声するのを避けるため、課題を記したカードを提示後、発声開始まで一定時間カードを隠した。さらに、被験者の自由な表現をなるべく妨げないようにするため、発声に要する時間には制限を設けないようにした。質問は何回発話しても途中で考える時間を挟んでも良く、なるべく多くの情報を与えた方が検索には有利であると

<TOPIC>
<NUM>0031</NUM>
<DESC>情報処理や IT と言った分野の資格試験にはどのようなものがあるのか知りたい</DESC>
<NARR><BACK>近年 IT 革命や情報化が叫ばれ、これに対応するべく情報処理技術者の需要が高まっている。それに伴い情報処理・IT 分野の資格も広く注目を集めている。そこで、それらの情報処理・IT 分野の資格を取りたいと思うがどのような資格があるのか？どのような資格試験が存在し、それらがどのようなものなのか知りたい。</BACK><RELE>IT・情報処理分野の資格試験に対する情報を提供するものを適合とする。資格試験のための専門学校などの紹介は不適合であるが、その中で試験に関しての情報を提供している場合は適合とする。</RELE></NARR>
</TOPIC>

図 1: NTCIR-3 WEB 検索課題の例

説明し、もう十分と判断した場合に最後にキーワード「検索開始」を発声するように求めた。

検索課題には、NTCIR-3 の WEB 検索テストコレクション [3] を利用した。本コレクションには、検索課題を簡潔に記述した Description の他に、課題の背景を詳しく記した Narrative が与えられている。検索課題の一例を図 1 に示す。

2.2 手順

全 13 課題を以下の手順で収録した。

1. 実験の内容を示したカードを被験者に見せ、実験手順を理解させる。不明な点は質問に答えた。
2. 課題 1〜6 までを続けて実行。1 課題の実行手順は以下の通り
   (i) 課題を記したカードをテーブルに置く。カードには、WEB 検索課題の Description(図 1 の <DESC>) と Narrative(図 1 の<NARR>) が記されている。
   (ii) 30 秒間、課題を読んで内容を理解させる。
   (iii) カードを隠し、30 秒間待つ。
   (iv) 録音開始。キューの合図 (手でジェスチャー) で、発話を開始させる。
   (v) 被験者の「検索開始」の発話後、録音終了。
3. 続いて、課題 7〜12 を、2 つの課題を同時に (7 と 8,9 と 10,11 と 12 の組み合わせで) 行った。課題 1〜6 との違いは以下のとおり。
   - カードは左右に並べて 2 枚同時に提示する。
   - カードを読む時間は 1 分。
   - 発話は、左のカード、右のカードの順に。
   - 2 課題連続で発声、録音。
4. 最後に自由課題を行う。何でも良いから、自分で興味を持つ検索課題を考えさせる。考える時間は十分に与え、準備ができ次第すぐに録音開始。

3 収録データの評価

大学 (院) 生男女各 2 名を被験者として、各人 13 件の発話収録実験を行った。実験用に選択した 12 課題と NTCIR-3 の WEB 検索課題番号 (ntcir#) との対応を表 1 に示す。収録した音声データ (自由課題を除く)

* "Collecting Spontaneous Speech Towards Speech-Driven Information Retrieval" by Tomoyosi AKIBA, Katunobu ITOU, (National Institute of Advanced Industrial Science and Technology), and Atsushi FUJII (University of Tsukuba)

表 1: 未知語率 (上段) と単語誤り率 (下段) (%)

	課題 1	2	3	4	5	6	7	8	9	10	11	12	全体	13
ntcir#	0008	0026	0023	0031	0027	0011	0021	0033	0036	0020	0046	0022	-	(自由)
A(女性)	4.6	4.5	9.6	9.1	7.7	9.8	8.9	8.0	10.7	17.0	9.7	7.4	9.0	8.6
	65.1	45.5	53.1	56.7	69.5	52.3	49.4	52.2	54.1	63.4	55.2	51.6	56.1	68.4
B(女性)	1.5	0.0	1.5	0.0	0.0	0.0	0.0	0.0	0.0	5.9	0.0	0.0	0.5	1.2
	40.9	24.4	59.4	30.0	35.0	24.1	44.0	16.0	54.5	39.4	34.0	18.8	34.4	44.6
C(男性)	3.9	0.0	0.0	0.0	0.0	1.6	0.0	0.0	0.0	4.4	0.0	0.0	1.2	1.5
	44.7	25.7	32.6	30.2	29.5	30.0	42.5	44.0	31.2	47.8	11.4	22.4	34.3	35.4
D(男性)	3.0	3.1	2.9	3.0	7.1	6.2	5.6	2.1	2.7	10.9	13.0	5.7	5.7	2.6
	75.0	48.4	73.5	58.5	78.0	61.9	85.7	52.2	50.0	74.1	44.4	97.1	65.5	60.5
全体	3.4	2.2	5.0	4.2	4.6	5.4	5.2	4.2	4.7	10.8	6.5	4.2	5.1	4.9
	54.2	36.7	53.9	46.4	56.4	42.9	54.0	44.3	49.5	58.3	43.7	47.1	48.8	56.1

(被験者 A)

情報処理やＩＴの産業に関係する、[えっと] 資格試験について、[えと] どういった種類があるのか知りたい。[えっと] その ((資格、)) 資格試験を受けるための、そういった、((こう))((え)) 講座を開いている教室、とかではなくて、[えっとー◎、] どういった種類の資格試験があって、((そ)) その一◎資格試験、の◎[えっとー◎、]((具体)) 具体的な [えっとー◎] その試験内容や、開催時期、[えっと] 受験料、[えっとー◎、] その資格を取ることで◎、[えっとー◎] どういった、どういった、え、((知識、)) 知識が、あ、あどういったことに有利に、働くのか、[えっとー◎、] そういった [えっと] 資格、試験の種類、について知りたい。検索開始。

(被験者 B)

最近、ＩＴ革命とか情報処理とかいう言葉をよく ((聞き、)) 耳にします。そのようなことに興味があるので、ＩＴ、ＩＴや情報技術に関連する、資格試験の情報、を知りたいです。どのような試験があって、それはどのような、内容で、どんな資格なのかという情報を、教えて下さい。資格を取るためだけの学校とかの情報はいりません。検索開始。

図 2: 収録した自由発話音声の書き起こし

は、1 課題平均入力時間 45.3 秒、最大入力時間 98.4 秒、最短入力時間 14.0 秒であった。課題 1〜6 と 7〜12 で異なる収録方法を試したが、収録した発話に大きな差は認められなかった。図 1 の課題に対する自由発話音声の書き起こしテキストを図 2 に示す。

収録した音声データを入力とした情報検索実験を行った。音声認識では、特に自由発話認識用の工夫は行わず、ディクテーション用の認識システムをそのまま用いた。音声認識デコーダには JULIUS[7]、言語モデルには WWW から収集した大規模コーパスから作成した 6 万語 tri-gram[4]、音響モデルは 2000 状態の状態共有 tri-phone を用いた。情報検索には、筑波大学で開発したシステム [4] を用いた。

言語モデルに対する収録した発話の未知語率と、音声認識の単語誤り率 (WER) を表 1 に示す。課題と被験者毎に単語誤り率を比較すると、課題による差に比べ被験者による差が大きい。被験者 A と D の発話は B と C に比べ未知語が多く、その結果認識率が悪化したと考えられる。この 2 グループの発話を調べると、図 2 の話者二人の書き起こしのように、発話スタイルが大きく異なっており、発話の自由度の差によるものと考えられる。

また、情報検索の精度 (平均適合率) を調べた。実験に用いた検索課題 (Description) をそのまま検索システムへの入力とした場合、Description の読み上げ音声[1] を入力とした場合、収録した自由発話音声の書き起こしを入力とした場合、自由発話音声を入力とした場合、について比較した。その結果、それぞれ 0.13, 0.069, 0.055, 0.031 であった。テキスト入力に比べ、読み上げ、自由発話となるに従い、平均適合率が約 50%ずつ性能低下した。また、自由発話の書き起こし (認識率 100%に相当) を入力とした場合、テキスト入力の半分以下の性能となった[2] 。検索に関係のないノイズを排除したり、間接的な表現 (例えば、「〜ではない文書」などの否定の表現) を扱うなど、検索手法の改善が必要である。

4 まとめ

情報検索課題の背景情報を用いて、自由発話音声による情報要求を収録する方法を提案した。また、提案手法を用いて発話収録実験を行った。収録した音声を分析した結果、提案手法により自由発話検索要求の収集が可能であることを確認できた。今後、本手法によって収集したデータを用いて、システムの開発を進めたい。また、音声入力質問応答システム [1] の自由発話入力化も検討課題である。

参考文献

[1] T. Akiba, K. Itou, A. Fujii, and T. Ishikawa. Selective back-off smoothing for incorporating grammatical constraints into the n-gram language model. In *Proceedings of International Conference on Spoken Language Processing*, Vol. 2, pp. 881–884, Denver, Colorado, Sept. 2002.

[2] J. Barnett, S. Anderson, J. Broglio, M. Singh, R. Hudson, and S. W. Kuo. Experiments in spoken queries for document retrieval. In *Proceedings of European Conference on Speech Communication and Technology*, pp. 1323–1326, Rhodes, Greece, Sept. 1997.

[3] K. Eguchi, K. Oyama, K. Kuriyama, and N. Kando. The Web retrieval task and its evaluation in the third NTCIR workshop. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 375–376, 2002.

[4] A. Fujii and K. Itou. Evaluating speech-driven IR in the NTCIR-3 Web retrieval task. In *Proceedings of the Third NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering*, 2003.

[5] A. Fujii, K. Itou, and T. Ishikawa. Speech-driven text retrieval: Using target IR collections for statistical language model adaptation in speech recognition. In A. R. Coden, E. W. Brown, and S. Srinivasan eds., *Information Retrieval Techniques for Speech Applications (LNCS 2273)*, pp. 94–104. Springer, 2002.

[6] R. S. Taylor. The process of asking questions. *American Documentation*, 13(4):391–396, 1962.

[7] 鹿野, 伊藤, 河原, 武田, 山本 (編) . 音声認識システム. オーム社, 2001.

[1] NTCIR-3 WEB 検索コレクション [3] から 4 人 (実験の被験者とは異なる) の音声を選んで使用した。

[2] ただし、自由発話入力の性能が悪い理由の一つは、適合文書の判定に行う pooling に参加していないことにある。

社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

[招待論文] 音声による言語バリアフリーな多言語情報アクセス

藤井 敦†,††

† 筑波大学 図書館情報学系 〒305-8550 茨城県つくば市春日 1-2
†† 科学技術振興事業団 CREST
E-mail: †fujii@slis.tsukuba.ac.jp

あらまし ワールドワイドウェブの発展によってオンライン情報が急増した反面，様々な障害によって，利用できる情報の量や質に個人差が生じるといった情報格差問題が深刻化している．本稿では，言葉に起因する障害に焦点を当てて，それを克服する情報アクセスシステムの研究事例やプロジェクトについて紹介する．
キーワード バリアフリー，情報検索，音声認識，機械翻訳，未知語問題，テストコレクション

Speech-driven Language-barrier-free Multi-lingual Information Access

Atsushi FUJII†,††

† Institute of Library and Information Science, University of Tsukuba Kasuga 1-2, Tsukuba, Ibaraki, 305-8550 Japan
†† CREST, Japan Science and Technology Corporation
E-mail: †fujii@slis.tsukuba.ac.jp

Abstract While the number of machine readable information accessible via the World Wide Web is growing, the digital divide problem caused by various barriers has of late become crucial. This paper describes research examples and projects of information access aimed to overcome barriers associated with human languages.
Key words barrier free, information retrieval, speech recognition, machine translation, out-of-vocabulary problems, test collections

1. まえがき

インターネットやワールドワイドウェブに代表される情報基盤技術の急速な発展と普及によって，自宅・オフィスの端末や携帯電話を使って，世界中の情報に瞬時にアクセスが可能になった．ウェブ上の検索システムは，ウェブが流行り始めた当初に比べれば使いやすくなり，情報要求に関連するページがすぐに見つかることも多い．しかし他方で，コンピュータに対する習熟度や外国語の運用能力によって，利用できる情報の量や質に個人差が生じるといった，いわゆる「情報格差」や「digital divide」が深刻な国際社会問題になっている．

全ての人々によるアクセスを志向した概念として「バリアフリー」や「ユニバーサルアクセス」が存在する．これらの概念を広義に解釈すれば，技術者/一般人の区別なく誰もが科学技術の恩恵を受けることができる社会の仕組み，身体障害者でも利用できるインフラ整備，地域間の経済格差に起因する問題の解決など様々な要因を含んでいる．しかし，本稿では情報アクセスにおけるバリアフリー化に限定し，とりわけ言語の運用能力に起因するバリアとその克服方法に焦点を当てながら，音声言語処理技術がどのように貢献するのかついて論述する．

以下，2. で既存の検索システムが抱える様々な障害（バリア）について説明し，3. でそれらのバリアを克服するためのシステム設計について考察する．4. で筆者らが研究開発した音声による多言語情報検索システムについて説明し，5. で当該問題に関連する研究プロジェクトを紹介する．

2. 検索システムが抱えるバリア

既存の検索システムの多くは，検索対象文書と同じ言語でキーワードや検索式をキーボード入力することが前提である．一般的に，効果的なキーワードや複雑な論理式を構成することで，ユーザの要求を満たす情報を取得することが可能になるという利点がある．しかし見方を変えれば，有効なキーワードや論理式を構成もしくは入力できない状況では，必要な情報を取得することが困難になるという問題が生じる．

すなわち，キーボード入力できない状況に置かれたユーザに対するバリアがある．キーボードリテラシーがないユーザ，非晴眼者，携帯端末ユーザにとっては，キーボードよりも音声入力の方がより自然なインタフェースである．

次に，情報要求を具体的に表現できないユーザに対するバリアがある．キーボードは簡潔に表現されたキーワードや論理式を入力することに適した手段である．しかし他方で，漠然・曖昧とした情報要求をキーボードで入力するのは困難を伴うため，音声入力の方がより自然なインタフェースである．

近年は，古典的な情報検索（information retrieval）よりも高度な情報アクセスを指向した技術として，質問応答（question answering）が実用的なアプリケーションになりつつある．質問応答システムは，キーワードよりも長く，かつ論理式よりも日常言語に近い質問文を許容することができるため，キーボード入力よりも音声入力に馴染みやすい性質を持っている．

以上の議論から，音声入力インタフェースの意義とは，キーボードの代替手段だけにとどまらないことが分かる．情報要求が漠然としている場合に，ユーザの発想支援を促すためのツールとしての効用も兼ね備えている．

既存の検索システムが抱えるその他のバリアとして，外国語運用能力に起因する言葉の壁がある．日本語のページはウェブ全体の10%にも満たないと言われている．ウェブ全体の70%を占めると言われる英語ページに英語圏以外のユーザが手軽にアクセスするためには，母国語で表現した検索要求を使って外国語を用意に取得できなければならない．画像などの検索はともかく，言語コンテンツ（テキスト情報）を検索する場合，この問題は顕著になる．さらに，検索された情報が外国語のままでは利用が難しいため，ユーザ言語に翻訳し，内容に基づいて分類するなどの閲覧支援も必要である．Altavista(注1)は比較的早い段階から入出力インタフェースに機械翻訳を導入しており，上記問題の解消に努めている．

3. バリアフリーシステムの実現における諸問題

既存の検索システムが抱える言語に起因する障害（すなわち，言語バリア）は，以下の4点に整理することができる．

(1) 音声入力が困難
(2) キーワード化できない漠然とした要求の入力が困難
(3) 外国語情報の検索が困難
(4) 検索された外国語情報の閲覧が困難

ここで，ひとまず(2)のバリアに関する議論を後回しにすると，残り3つのバリアを克服するためには，検索エンジンの入出力インタフェースとして音声認識と機械翻訳を接続するという安易な解決方法がある．しかし，このような単純な方法では，音声認識，機械翻訳，情報検索などの個別処理の誤りが累積して，システム全体の有用性は著しく低下することが容易に想像できる．もっとも，個別の技術について今後も研究を進めれば，いずれは実用レベルに到達することが期待できる．

しかし，個別の技術を高度化するだけでは解決が難しい，より本質的な問題がある．それは，音声認識，機械翻訳，情報検索で扱える語彙のサイズや内容が異なるという点である．すなわち，個別の処理を接続する段階で「未知語問題」が発生する．

近年の情報検索システムは，古典的な統制語彙型システムとは異なり，検索対象文書中の任意の語による検索を可能とする．索引のサイズが数千万～億の単位に達することは珍しくない．助詞などの機能語などは不要語として索引から除外されることがあるものの，これらが検索キーワードとして利用されることは稀であるため，事実上，語彙制限はないと考えてよい．

それに対して，近年の音声認識システムでは語彙サイズ（辞書登録語数）が制限される．ハードウェアに関する制約や統計モデルの学習効率が主な原因であるため[24]，登録語数を増やすという単純な方法では解決が困難である．そのため，多くの言語において語彙サイズは高々数万語に制限されており[15],[19],[22]，実用的な検索システムの索引サイズに比べると極端に小さい．また，統計的音声認識では，機能語などの高頻出語は高精度で認識されるのに対して，情報検索では特定の文書にしか出現しない低頻度語ほど効果的な索引語になりやすい．そこで，ユーザ発話中の効果的な検索キーワードほど誤認識されやすくなる．

機械翻訳の標準的な語彙サイズは明らかではないものの，比較的規模が大きい翻訳辞書でも登録語数は百万語程度であり，音声認識と情報検索の中間規模に位置する．すなわち，音声認識 → 機械翻訳 → 情報検索という具合いに後段の処理に進むほど扱うことができる語彙サイズが大規模化する．それにも拘らず，既存のシステムを組み合わせると，前段の処理において使用できる語彙が制限されてしまう．

音声認識と機械翻訳を統合する研究として音声翻訳（speech-to-speech translation）がある．しかし，既存の音声翻訳（例えば，Verbmobilプロジェクト[23]）においては，扱える語彙サイズは音声認識できるサイズ（数万語）に制限されているため，ウェブ検索などの大規模な情報検索のための入力インタフェースとしては十分でない．

ここで，上述した(2)のバリアに話を戻すと，これも未知語問題の一種として捉えることができる．ユーザは自分に馴染みのある平易な言葉によって漠然とした要求を言語化し，発話するかもしれない．しかし，発話内容が機械翻訳しやすい表現であったり，検索に有効なキーワードであるとは限らない．こうして考えると，未知語問題とは，ユーザとシステムの間にも生じることが分かる．各モジュールにおいて処理しやすい言葉に変換することで，異種技術のシームレスな統合が必要である．

以上をまとめると，言語バリアフリーな情報アクセスシステムにおいて「未知語問題」は本質的に不可避である．基盤技術（音声認識，機械翻訳，情報検索）を個別に高度化するに止まらず，異種技術を横断した総括的な解決策が必要である．

4. 多言語情報検索エージェント MIRAGE

4.1 システム概要

2.と3.で提起した問題を解消するために，筆者らは多言語情報検索エージェント MIRAGE（Multi-lingual Information Retrieval AGEnt）を研究開発した[7],[8],[10]．MIRAGEは，日本語の音声入力によってウェブや各種コレクションから日本語と英語の文書情報を検索することができる．また，閲覧支援のために，検索結果に含まれる英語情報を日本語に翻訳し，内容に基づいて分類表示することが可能である．

(注1)：http://www.altavista.com/

図1にMIRAGEのシステム構成を示す．破線矢印と実線矢印はそれぞれオフライン処理とオンライン処理に対応する．用途に応じて，各モジュールは適宜スキップすることが可能である．ここでは，全てのモジュールを利用する状況を用いてMIRAGEの動作を説明する．まずユーザが音声によって漠然とした情報要求を入力すると，音声認識によって転記（transcription）が自動生成される．次に，キーワード生成によって転記された情報要求が具体的な検索キーワードに変換される．さらに，キーワード翻訳によって日本語キーワードを英語に翻訳し，日本語と英語の情報を同時に検索する．検索結果に含まれる英語情報は文書翻訳によって日本語に翻訳され，最後に文書分類によって内容に基づくグループに分割される．

図1では「電子メールに感染するものについて知りたい」という発話から「マクロウイルス」「macro virus」という日英キーワードを生成し，検索された文書を「病気に関するウイルス」と「コンピュータウイルス」という観点で分類した例を示している．以下の節で，個別のモジュールについて説明する．

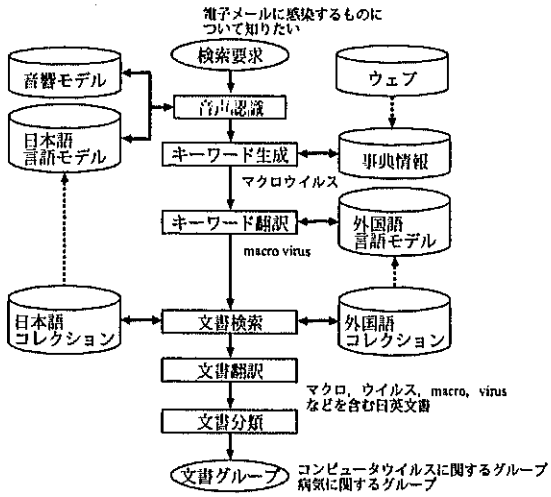

図1　MIRAGEのシステム構成
Fig. 1　The overall design of MIRAGE.

4.2 音声認識

統計的な音声認識システム[2]は主に音響モデルと言語モデルで構成され，両者は音声認識精度に強く影響する．音響モデルは音響的な特性に関するモデルであり，検索対象文書とは独立な要素である．

それに対して，言語モデルは音声認識候補の言語的な妥当性を定量化するためのモデルである．しかし，あらゆる言語現象全てをモデル化することは不可能であるため，一般的には与えられた学習用コーパスに出現する言語現象に特化したモデルを作成する．本システムのユーザは検索対象に関連する内容を発話する可能性が高い．そこで，検索対象文書に基づいて言語モデルを作成すれば，音声認識の精度向上が期待できる．その結果，ユーザの発話が正しく認識されるので，テキスト入力に近い検索精度を実現することが可能になる[11], [14].

本システムでは，日本語ディクテーションツールキット[27][(注2)]で提供されている音声認識エンジン（デコーダ）と音響モデル[15]を利用している．本システムには，言語モデルの作成を工夫することで，以下の2つの特長がある．

まず，ユーザ発話中に含まれる未知語を検出し，検索用の索引語によって自動的に正しい語に補完する機能を持つ[10], [26]．簡単に説明すると，学習コーパスから抽出した高頻度語を単語として辞書に登録し，それ以外の語は音節に分割する．そして，単語と音節を併用したNグラムモデルを作成する．その結果，辞書未登録の語は，単語ではなく音節（カタカナ）列として検出される．音節の数は限られているので（我々の実験では異なりで700）辞書登録語数を増やすことなく，認識できる語数を増やすことができる．また，音節の新しい組み合わせによって，学習コーパスに存在しない語を合成することも可能になる．また，当該言語モデルは通常の統計的Nグラムなので，既存のデコーダを拡張せずに利用できる．

「オレンジやグレープフルーツなどの柑橘系果物の輸入に関する記事」という発話を例にとると，

> オレンジや /グ レ ー プ フ ラ チ ナ ガ ノ/ などの /カン キ ツ ケ イ/ 果物の輸入に関する記事

のように「グレープフルーツ」や「柑橘系」が未知語として検出される（ここでは未知語部分をスラッシュで括っている）．なお「柑橘系」のように未知語箇所の検出と音韻列の特定に成功する場合や「グレープフルーツ」のように音韻列の特定は不完全でも未知語箇所の検出に成功する場合がある．

さらに，検出されたカタカナ語と音韻的な類似度が高い語を検索対象の索引から探索し，正しい語（「グレープフルーツ」や「柑橘系」）に置換する．以上の処理によって，音声認識と情報検索における未知語問題を解消することが可能になる．

もう一つの特長は，質問応答システムに入力される質問文に対応している点にある．あらゆる種類の質問文をモデル化することは困難である．しかし，既存の質問応答システムでは5W1H的要素に関する質問文を想定している．そこで，質問文の多くは，検索トピックに関するキーワードと質問文固有の有限個の定型表現（「～はどこにありますか?」など）に分解することができる．そこで，単語Nグラム（キーワードに対応）と人手で記述した文法（定型表現に対応）を統合した言語モデルを作成し，認識精度の向上を実現している[1], [33].

以上の点から，MIRAGEは，音声認識と文書検索を単に接続したシステム[3], [4]や単語認識による検索[18]とは異なる．

4.3 キーワード生成

ユーザが具体的・効果的なキーワードを思い付かない問題（ユーザとシステムの間に生じる未知語問題）は，情報検索の分野で古くから認知されており，シソーラス中の同義語や初期検索文書中の索引語を用いた検索要求の拡張（query expansion）が研究されている．しかし，ユーザのあらゆる要求に対応するためには大規模なシソーラスが必要であり，人手による構築は高価である．また，初期検索文書に基づく手法では，初期検索

(注2)：http://www.lang.astem.or.jp/CSRC/

の精度がシステム全体の性能を左右する．

本システムでは，既存の手法とは異なるアプローチをとっている．筆者らは，ウェブに新規情報や専門情報が数多く含まれている点に着目し，用語辞典・百科事典的な知識の構築を行っている [6], [8], [28], [31] ．概説すると，ウェブページ中の HTML タグのレイアウトや言語表現に基づいて，特定の用語について説明している段落を抽出し，説明内容の専門分野に基づいて分類することで，人手で編纂される事典に近い知識情報を自動構築する．現在，見出し語数は 20 万語に到達し，人手による精度評価や情報処理技術者試験問題の自動回答などの応用評価において概ね良好な結果を得ている．

MIRAGE では，この事典情報を一種のシソーラスとして利用し，漠然とした情報要求を具体的なキーワードに語彙化する．直感的に言うと，事典情報を逆引きすることで，自然言語文から見出し語への変換を行う．具体的には，用語説明を個別の文書と見なして索引付けし，既存の文書検索手法によって，ユーザの情報要求に適合する用語説明を検索する．その結果，例えば「電子メールに感染するもの」から「マクロウイルス」を生成したり，また「大画面の薄型ディスプレイ」から「PDP（プラズマディスプレイパネル）」を生成することが可能になる．複数のキーワード候補が得られた場合は，検索対象コレクションにおける TD·IDF 値などに基づいて，検索に有効な候補から順番に提示することもできる．

4.4 キーワード翻訳

通常の機械翻訳では，文や文章を対象にして，人間が理解しやすい自然な訳を生成することに主眼が置かれている．それに対して，検索キーワードの翻訳では比較的短い語やフレーズが対象であるため，問題が矮小化されている印象があるかもしれない．しかし，ユーザが検索したくなるようなキーワードに迅速に対応するために新語や専門用語を高精度で翻訳しなければならないという問題がある．事実，3. でも議論したように，情報検索システムで扱うことができる索引サイズは機械翻訳で扱える語彙サイズよりも一般的に大きい．そこで，機械翻訳と情報検索における未知語問題を解消する必要がある．

翻訳用辞書における未登録語として，既存の語を組み合わせた複合語の専門用語（「アドレス空間制御ブロック」など）やカタカナ表記された外来語（「コラボレーション」など）が多い．本システムにおける翻訳モジュールの特長は，複合語と外来語の翻訳を高精度で行う点にある [7], [30].

翻訳対象の検索キーワードが辞書に登録されていない場合は，辞書に登録されている細かな単位（単語や短い複合語）に分割し，辞書引きを行いながら翻訳を行う．ただし，細かな単位で翻訳すると，訳語候補の組み合わせによって可能な訳語数が急激に増加する．そこで，検索対象コレクションから作成した言語モデル（単語 N グラム）に基づく統計的手法によって訳語曖昧性を解消する．

翻訳辞書に未登録のカタカナ語は「翻字（transliteration）」と呼ばれる処理によって音韻的に類似する英語へ変換する．具体的には，文字列（カタカナ列とアルファベット列）を対応させた翻字辞書を自動的に構築しておき，文字列単位で辞書を引きながら翻訳を行う．翻字においても多数の訳語候補が生成されるため，検索対象コレクションから作成した言語モデルに基づいて尤もらしい訳語を選択する．

当該技術はテキスト入力の多言語検索や言語横断検索 [29] でも有効であり，筆者らの翻訳エンジンは多言語特許検索システム [13] において既に実用化されている．

4.5 その他のモジュール

文書検索のために確率型の検索手法 [20] を実装し，形態素単位で索引付けを行っている．また，論文抄録，新聞記事，特許情報，ウェブページなど（主に日本語と英語）を対象に研究を進めている．文書翻訳には PAT-Transer(注3)を，文書分類には汎用連想検索エンジン GETA(注4)を用いている．

5. 関連する研究プロジェクト

5.1 諸外国における動向

近年の音声認識技術は，認識精度，処理速度，ツールとして使いやすさなどの観点から実用的になってきた．そこで，情報検索などのアプリケーションに音声認識を導入することは比較的容易である．これらは目的に応じて以下の 2 つに大別できる．

- 音声データの検索

放送音声データなどを検索対象とする．入力手段は問わないものの，テキスト入力が中心である．

- 音声による検索

検索要求を音声入力によって行う．検索対象の形式は問わないものの，テキストが中心である．

音声データの検索は，米国の TREC(注5)における Spoken Document Retrieval（SDR）トラック [12] で放送音声データを対象にしたテストコレクションが整備されたことを背景にして，盛んに研究が行われている [16], [17], [21]．また，欧州の CLEF(注6)では多言語音声データ検索が行われている．他方において，音声による検索はカーナビゲーションシステムやコールセンターのようにキーボード入力を前提としないバリアフリーなアプリケーションを支える重要な基盤技術であるにも拘らず，音声データ検索に比べて研究事例は少ない．TREC や CREF では音声による検索はタスクとして実施されていない．

5.2 国内における動向

日本において，TREC や CLEF と主旨を同じくする評価ワークショップとして NTCIR(注7)がある．実験用のデータセット（テストコレクション）と実験結果を評価するための標準化された手順が用意され，参加チームが開発したシステムの性能を比較評価する．NTCIR プロジェクトは 1998 年から開始され，2002 年 10 月に第 3 回目の NTCIR-3 ワークショップが開催された．NTCIR-3 では，言語横断検索，特許検索，質問応答，テキスト自動要約，Web 検索の 5 つのタスクが実施され，9ケ国から大学や企業など合計 65 チームが参加した．

(注3)：http://www.nova.co.jp/
(注4)：http://geta.ex.nii.ac.jp/
(注5)：http://trec.nist.gov/
(注6)：http://clef.iei.pi.cnr.it/
(注7)：http://research.nii.ac.jp/ntcir/

Web 検索タスク [5] では，テキスト入力による検索をメインタスクとして実施しつつ，外部オーガナイザ（参加チームの有志）による自由タスクの提案が認められていた．筆者らは，この機会を利用して音声入力型 Web 検索タスク [9], [25](注8)を提案し，音声入力による検索用テストコレクションを構築した．

5.3 Web 検索メインタスク

Web 検索タスクの目的は，タグとリンク構造を持った Web 文書の検索に関する研究の推進である．文書コレクションとして，参加者が扱いやすい規模（10GB 程度）と比較的現実的な規模（100GB 程度）の 2 つが用意された．

検索課題（トピック）の例を図 2 に示す．検索課題は SGML 形式で記述されており，課題 ID（<NUM>），中心的な主題を表わす 1〜3 語程度のキーワード（<TITLE>），検索要求を記述した DESCRIPTION（<DESC>），背景，検索目的，判定基準，用語の定義などを詳説した NARRATIVE（<NARR>），課題作成者によって定義された同義語，関連語，上位語を示す CONCEPTS（<CONC>），適合文書 ID の例（<RDOC>），課題作成者のプロフィール（<USER>）からなる．

さらに，各検索課題に対する正解（適合性）判定もコレクションに含まれている．そこで，当該コレクションを再利用して様々な検索手法を比較評価することができる．

5.4 音声入力型 Web 検索タスク

音声入力型 Web 検索タスクは，再利用可能なテストコレクションと関連ツールを，情報検索分野の研究者にも音声処理分野の研究者にも広く利用できるような形で整備することを目的とした．タスク設定として，読み上げ音声発話入力による検索を想定した．参加形態として，音声認識から情報検索まで行う全参加と，音声認識のみ行う部分参加を設定した．参加チームに提供したデータとツールは以下の通りである．

- 読み上げ音声発話データ

Web 検索メインタスクの<DESC>を読み上げたものを音声による検索要求として使用した（男女各 5 名）．そこで，メインタスクのテキスト入力に対する適合性判定をそのまま利用できる．

- 音声認識ツールキット

NTCIR 参加者の多くは情報検索分野や自然言語処理分野の研究者であり，音声認識システムの開発や扱いには不慣れであることを考慮して，無償の音声認識ツールキット [27] で利用できる言語モデルと辞書を作成した．

- 音声認識用言語モデル

100GB コレクションから高頻度語に制限した単語トライグラムおよびバイグラムを作成した．語彙サイズは，2 万語と 6 万語の 2 種類を用意した．

- 検索システム

音声認識のみ行うチームのために，10GB/100GB コレクションを検索できるシステムに対するアカウントを用意した．

音声入力型 Web 検索タスクには，オーガナイザーをメンバーに含む「図情大-産総研チーム」が全参加した．また「豊橋技科大チーム」が音声認識のみに参加し，提供された検索システム

(注8)：http://research.nii.ac.jp/ntcir/workshop/web/sdwr_index.html

```
<TOPIC><NUM>0008</NUM>
<TITLE CASE="b">サルサ，学ぶ，方法</TITLE>
<DESC>サルサを踊れるようになる方法が知りたい</DESC>
<NARR><BACK>最近はやっているサルサという踊りを学ぶためにどうすればよいか具体的な方法が知りたい。例えば教室に通うという場合には、その場所や授業形態など、具体的な内容を必要とする。</BACK><RELE>具体的な方法の表記のない、流行であることのみを扱った文書は不適合とする。</RELE></NARR>
<CONC>サルサ，習う，方法，場所，カリキュラム</CONC>
<RDOC>NW011992774，NW011992731，NW011992734</RDOC>
<USER>大学院修士 1 年，女性，検索歴 2.5 年</USER></TOPIC>
```

図 2 Web 検索課題の例

Fig. 2 A fragment of the Web IR collection topic.

(5.4 参照) を利用して実験を行った．概して，検索対象コレクションから作成した言語モデルの有効性が示され，また，語彙サイズが音声認識，検索精度ともに影響を与えることも分かった．音声入力による検索精度はテキスト入力（音声認識誤りなしの入力）の 60%程度であった(注9)．

5.5 今後の展望

NTCIR-3 での経験を踏まえて，以下の点について検討する必要がある．まず，読み上げ音声から自由発話（spontaneous speech）への移行が必要である．筆者らの予備調査について別稿 [32] で報告する予定である．また，質問応答や言語横断検索などの他タスクとの連携によって，理想に近い情報アクセスの実現を目指す．そのためには，異なる分野の研究者が共用できるデータやツールの整備が必要である．テキスト入力を対象にした情報検索や自然言語処理では，入力が誤りを含まないことが前提である．音声認識などによる誤りを含んだ入力に対する頑健さについて今後さらに研究する必要がある（Google(注10)の綴り誤り修正機能は分かりやすい例である）．モジュールが増えることで評価が煩雑になるため，グラスボックス評価の枠組みについても検討が必要である．

6. むすび

音声入力による漠然とした要求を使って多言語情報検索し，閲覧する情報アクセスについて概説した．システム実現における諸問題について議論し，筆者らが開発したシステムを紹介した．また，国内における研究動向として，NTCIR-3 音声入力型 Web 検索タスクを紹介し，今後の展望について検討した．

謝辞

本稿で紹介した研究成果は，伊藤克亘氏，秋葉友良氏（産業技術総合研究所），石川徹也教授（筑波大学）との共同研究によるものです．音声入力型 Web 検索タスクを実施する機会を与えて下さった NTCIR-3 Web 検索タスクのオーガナイザーに感謝致します．

(注9)：最終的な評価結果については現在分析中であり，NTCIR-3 ワークショップ予稿集で最終報告する予定である [9]．

(注10)：http://www.google.com/

文　献

[1] Tomoyosi Akiba, Katunobu Itou, Atsushi Fujii, and Tetsuya Ishikawa. Selective back-off smoothing for incorporating grammatical constraints into the n-gram language model. In *Proceedings of the 7th International Conference on Spoken Language Processing*, pp. 881–884, 2002.

[2] Lalit. R. Bahl, Frederick Jelinek, and Robert L. Mercer. A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 5, No. 2, pp. 179–190, 1983.

[3] J. Barnett, S. Anderson, J. Broglio, M. Singh, R. Hudson, and S. W. Kuo. Experiments in spoken queries for document retrieval. In *Proceedings of Eurospeech97*, pp. 1323–1326, 1997.

[4] Fabio Crestani. Word recognition errors and relevance feedback in spoken query processing. In *Proceedings of the Fourth International Conference on Flexible Query Answering Systems*, pp. 267–281, 2000.

[5] Koji Eguchi, Keizo Oyama, Kazuko Kuriyama, and Noriko Kando. The Web retrieval task and its evaluation in the third NTCIR workshop. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 375–376, 2002.

[6] Atsushi Fujii and Tetsuya Ishikawa. Utilizing the World Wide Web as an encyclopedia: Extracting term descriptions from semi-structured texts. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pp. 488–495, 2000.

[7] Atsushi Fujii and Tetsuya Ishikawa. Japanese/English cross-language information retrieval: Exploration of query translation and transliteration. *Computers and the Humanities*, Vol. 35, No. 4, pp. 389–420, 2001.

[8] Atsushi Fujii and Tetsuya Ishikawa. Organizing encyclopedic knowledge based on the Web and its application to question answering. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pp. 196–203, 2001.

[9] Atsushi Fujii and Katunobu Itou. Evaluating speech-driven IR in the NTCIR-3 Web retrieval task. In *Proceedings of the 3rd NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering*, 2003. (To appear).

[10] Atsushi Fujii, Katunobu Itou, and Tetsuya Ishikawa. A method for open-vocabulary speech-driven text retrieval. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pp. 188–195, 2002.

[11] Atsushi Fujii, Katunobu Itou, and Tetsuya Ishikawa. Speech-driven text retrieval: Using target IR collections for statistical language model adaptation in speech recognition. In Anni R. Coden, Eric W. Brown, and Savitha Srinivasan, editors, *Information Retrieval Techniques for Speech Applications (LNCS 2273)*, pp. 94–104. Springer, 2002.

[12] John S. Garofolo, Ellen M. Voorhees, Vincent M. Stanford, and Karen Sparck Jones. TREC-6 1997 spoken document retrieval track overview and results. In *Proceedings of the 6th Text REtrieval Conference*, pp. 83–91, 1997.

[13] Shigeto Higuchi, Masatoshi Fukui, Atsushi Fujii, and Tetsuya Ishikawa. PRIME: A system for multi-lingual patent retrieval. In *Proceedings of MT Summit VIII*, pp. 163–167, 2001.

[14] Katunobu Itou, Atsushi Fujii, and Tetsuya Ishikawa. Language modeling for multi-domain speech-driven text retrieval. In *IEEE Automatic Speech Recognition and Understanding Workshop*, 2001.

[15] Katunobu Itou, Mikio Yamamoto, Kazuya Takeda, Toshiyuki Takezawa, Tatsuo Matsuoka, Tetsunori Kobayashi, and Kiyohiro Shikano. JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research. *Journal of Acoustic Society of Japan*, Vol. 20, No. 3, pp. 199–206, 1999.

[16] S.E. Johnson, P. Jourlin, G.L. Moore, K. Spärck Jones, and P.C. Woodland. The Cambridge University spoken document retrieval system. In *Proceedings of ICASSP'99*, pp. 49–52, 1999.

[17] G.J.F. Jones, J.T. Foote, K. Spärck Jones, and S.J. Young. Retrieving spoken documents by combining multiple index sources. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 30–38, 1996.

[18] Julian Kupiec, Don Kimber, and Vijay Balasubramanian. Speech-based retrieval using semantic co-occurrence filtering. In *Proceedings of the ARPA Human Language Technology Workshop*, pp. 373–377, 1994.

[19] Douglas B. Paul and Janet M. Baker. The design for the Wall Street Journal-based CSR corpus. In *Proceedings of DARPA Speech & Natural Language Workshop*, pp. 357–362, 1992.

[20] S.E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 232–241, 1994.

[21] Páraic Sheridan, Martin Wechsler, and Peter Schäuble. Cross-language speech retrieval: Establishing a baseline performance. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 99–108, 1997.

[22] Herman J. M. Steeneken and David A. van Leeuwen. Multilingual assessment of speaker independent large vocabulary speech-recognition systems: The SQALE-project. In *Proceedings of Eurospeech95*, pp. 1271–1274, 1995.

[23] Wolfgang Wahlster. Mobile speech-to-speech translation of spontaneous dialogs: An overview of the final Verbmobil system. In Wolfgang Wahlster, editor, *Verbmobil: Foundations of Speech-to-Speech Translation*, pp. 3–21. Springer, 2000.

[24] Steve Young. A review of large-vocabulary continuous-speech recognition. *IEEE Signal Processing Magazine*, pp. 45–57, September 1996.

[25] 伊藤克亘, 藤井敦. NTCIR-3 ワークショップにおける音声入力型ウェブ検索タスク. 情報処理学会研究報告 2002-NL-43, pp. 25–32, 2002.

[26] 伊藤克亘, 田中和世. 被覆率を重視した大語彙連続音声認識用統計的言語モデル. 日本音響学会講演論文集, pp. 65–66, March 1999.

[27] 鹿野清宏, 伊藤克亘, 河原達也, 武田一哉, 山本幹雄（編）. 音声認識システム. オーム社, 2001.

[28] 藤井敦, 伊藤克亘, 石川徹也. WWW は百科事典として使えるか? –大規模コーパスの構築–. 情報処理学会研究報告 2002-NL-149, pp. 7–14, 2002.

[29] 藤井敦. 言語横断検索への入門 –翻訳と検索の統合が生み出す新たな可能性–. 情報処理, Vol. 42, No. 3, pp. 327–329, 2000.

[30] 藤井敦, 石川徹也. 技術文書を対象とした言語横断情報検索のための複合語翻訳. 情報処理学会論文誌, Vol. 41, No. 4, pp. 1038–1045, 2000.

[31] 藤井敦, 石川徹也. World Wide Web を用いた事典知識情報の抽出と組織化. 電子情報通信学会論文誌, Vol. J85-D-II, No. 2, pp. 300–307, 2002.

[32] 秋葉友良, 伊藤克亘, 藤井敦. 音声入力型情報検索のための自由発話収録. 日本音響学会講演論文集, Mar. 2003. （掲載予定）.

[33] 秋葉友良, 伊藤克亘, 藤井敦, 石川徹也. 部分 N-gram 頻度情報を利用した質問応答定型表現への言語モデル適応. 情報処理学会研究報告 2002-SLP-42, pp. 31–38, 2002.

NTCIR-3 ワークショップにおける 音声入力型ウェブ検索タスク

伊藤克亘 ‡, 藤井敦 †‡
産業技術総合研究所, † 筑波大学 図書館情報学系
‡ 科学技術振興事業団 CREST
E-mail: itou@ni.aist.go.jp, fujii@slis.tsukuba.ac.jp

近年、音声認識技術の性能が向上し、実際の応用まで視野に入れた研究を推進する必要性が向上している。情報検索は歴史も長く、主張な情報処理アプリケーションの一つであるため、音声認識の応用の有力な候補として期待される。そこで、我々は、NTCIR-3 ワークショップ Web 検索タスクにおいて音声入力サブタスクを実施した。タスクで用いるテストコレクションとして、Web 検索メインタスクの検索要求を 10 名の話者により読み上げたデータを整備した。このコレクションを音声認識システムと情報検索システムから構成される音声入力型情報検索システムで評価した。情報検索システムは確率モデルに基づくシステムであり、各ページの言語的な内容だけを用いており、HTML タグやハイパーリンク情報は用いていない。実験の結果、a) 音声認識用言語モデルとして対象文書を利用すること。b) 音声認識の語彙サイズを拡張すること。がシステムの性能向上に役立つことが明らかになった。

Speech-Driven Web Retrieval Task in the NTCIR-3 Workshop

ITOU Katunobu‡, FUJII Atsushi†‡
National Institute of Advanced Industrial Science and Technology,
†Institute of Library and Infromation Science, University of Tsukuba,
‡CREST, Japan Science and Technology Corporation
E-mail: itou@ni.aist.go.jp, fujii@slis.tsukuba.ac.jp

Speech recognition has of late become a practical technology for real world applications. For the purpose of research and development in speech-driven retrieval, which facilitates retrieving information with spoken queries, we organized the speech-driven retrieval subtask in the NTCIR-3 Web retrieval task. Search topics for the Web retrieval main task were dictated by ten speakers and recorded as collections of spoken queries. We used those queries to evaluate the performance of our speech-driven retrieval system, where speech recognition and text retrieval modules were integrated. In the text retrieval module, which is based on a probabilistic model, indexed only textual contents in documents (Web pages), but did not use HTML tags and hyperlink information in documents. Experimental results showed that a) the use of target documents for language modeling and b) enhancement of the vocabulary size in speech recognition were effective to improve the system performance.

1 はじめに

近年の音声認識技術は、ある程度内容が整理されている発話に対しては実用的な認識精度を達成できる。また、ハードウェア技術の発展にも支えられ、パソコン上で動作する商用/無償の音声認識ソフトウェアが存在する。そこで、既存のアプリケーションに音声認識を導入することは比較的容易になっている。

とりわけ、情報検索システムは歴史が長く主要な情報処理アプリケーションの一つであるため、音声認識を採り入れた研究も近年数多く行われている。これらは目的に応じて以下の2つに大別できる。

- 音声データの検索 [1, 2, 3, 4]
  放送音声データなどを検索対象とする。入力手段は問わないものの、テキスト入力が中心である。
- 音声による検索 [5, 6, 7, 8, 9]
  検索要求を音声入力によって行う。検索対象の形式は問わないものの、テキストが中心である。

これらは検索対象と検索要求のどちらを音声データと捉えるかが異なる。さらに、両者を統合すれば、音声入力による音声データ検索を実現することも可能である。しかし、そのような研究事例はあまり存在しない。

音声データの検索は、TREC-6 の Spoken Document Retrieval (SDR) トラック [10] で放送音声データを対象にしたテストコレクションが整備されたことを背景にして、盛んに研究が行われている。

他方において、音声による検索はカーナビゲーションシステムやコールセンターのようにキーボード入力を前提としない（バリアフリーな）アプリケーションを支える重要な基盤技術であるにも拘らず、音声データ検索に比べて研究事例は極端に少ない。

Barnett ら [5] は、既存の音声認識システム（語彙サイズ 20,000）をテキスト検索システム INQUERY の入力として利用して、音声による検索の評価実験を行った。具体的には、TREC コレクションに含まれる検索課題 35 件に対する単一話者の読み上げ音声をテスト入力として利用し、検索実験を行った。

Crestani [6] も上記 35 件の読み上げ検索課題を用いた実験を行い（通常のテキスト検索で用いられる）適合性フィードバックによって検索精度が向上することを示している。しかし、どちらの実験においても既存の音声認識システムを改良せずに利用しているため、単語誤り率は 30%以上と比較的高い。

Kupiec ら [11] は数個のキーワードからなる検索要求を単語認識器を用いて認識し、複数の組み合わせを候補とし、対象となるコレクションを用いてもっともらしい組み合わせを決定する手法を提案している。この手法は、対象となるコレクションによく出現するキーワードの組み合わせをもっともらしいとする手法であるが、検索要求が長くなると、候補の数が増加し、探索コストが増加するという問題がある。したがって、この手法は文章の検索要求を連続音声認識を用いて認識する場合に適用することは困難である。

これらの問題点に対して、われわれは音声認識の精度向上にも焦点を当て、音声認識とテキスト検索の有機的な統合を目指してシステムの研究開発を行った [7, 8, 9]。音声認識システムの多くは、単語辞書や言語モデルを切替えることで、目的に応じた使い分けが可能である。そこで、検索対象テキストに基づいて言語モデルを作成し、利用することは自然な発想である。

確率モデルに基づく音声認識システムは主に音響モデルと言語モデルで構成され、両者は音声認識精度に強く影響する。音響モデルは音響的な特性に関するモデルであり、検索対象テキストとは独立な要素である。

それに対して、言語モデルは音声認識候補の言語的な妥当性を定量化するためのモデルである。しかし、あらゆる言語現象全てをモデル化することは不可能であるため、一般的には与えられた学習用コーパスに出現する言語現象に特化したモデルを作成する。

情報検索システムのユーザは検索対象に関連する内容を発話する可能性が高い。そこで、検索対象のテキストに基づいて言語モデルを作成すれば、音声認識の精度向上が期待できる。その結果、ユーザの発話が正しく認識されるので、テキスト入力に近い検索精度を実現することが可能になる。

高精度の音声認識は、対話による検索を円滑に進めたり、自分の発話通りに検索が実行されるという安心感をユーザに与える上でも重要である。現在の音声認識システムでは、システム語彙のサイズを限定するのが一般的であるが、情報検索の場合、キーワードが未知語となり認識されなかったり、別のキーワードに誤認識されると致命的である。したがって、未知語の扱いも重要となる [7]。

NTCIR-3 Web 検索タスクにおいて、メインタスクでは現在主流であるテキストによる検索要求を用いた検索の評価が対象である。しかし、メインタスクに加え、外部オーガナイザによる自由な研究課題

を展開するための自由タスクの提案が勧められている。そこで、我々は、音声入力型情報検索タスクを提案し、音声入力型ウェブ検索用のテストコレクションを構築した。

2章では、NTCIR ワークショップの概要について説明し、3章では、我々が実施した音声入力型検索タスクの概要について説明し、4章では、それを用いた音声入力ウェブ検索の評価結果について述べる。5章では、音声処理と情報検索分野の今後のテストコレクションの方向について議論する。

2 NTCIR ワークショップ

2.1 概要

NTCIR ワークショップは、情報検索とテキスト要約・情報抽出などのテキスト処理技術の研究をより発展させることを目的とした評価会議である。1998年より開始され、以下の3点を目的としている。

1. 繰り返し利用できる大規模な実験用データセットと、システム間の比較を可能にする共通の評価枠組みを提供することによって、情報検索やテキスト処理技術と関連領域の研究の一層の発展を図る。
2. システム間の比較、研究上のアイディアの交換などを行う研究者フォーラムを作る。
3. 情報検索やテキスト処理技術の評価手法および繰り返し実験に利用可能な大規模データセット構築法について研究を行う。

「評価会議」は、通常、実験用のデータセットと実験結果を評価するための統一された手順が用意される。評価会議に参加するグループは、それぞれ、NTCIR 企画グループが用意したデータを用い、さまざまなアプローチで研究と実験を行う。

これまで、日本語のテキストによる学術抄録および新聞記事の情報検索を中心に、言語横断検索(日英中)、テキスト要約に関する評価テストを行なってきており、これまで8ヶ国から40を越える研究グループが参加してきている。

また、NTCIR では、「テストコレクション」と呼ばれる実験用データセットを整備・公開している。テストコレクションとは、情報検索分野で、情報検索システムの検索性能評価に用いるデータで以下の要素からなる。

1. 検索対象文書
2. 利用者の検索要求を記述した「検索課題」
3. 検索課題を満たす「正解文書の網羅的なリスト」

テストコレクションは、情報検索システムの研究・開発、評価に必要不可欠であるとされているものである。

2001 年 8 月から 2002 年 10 月に行なわれた NTCIR-3 では、特許検索と Web 検索、質問応答に関するタスクが新たに取り入れられた。

2.2 NTCIR-3 ワークショップ Web 検索タスク

NTCIR-3 ワークショップ Web 検索タスクの目的は、タグとリンク構造を持った Web 文書の検索に関する研究である。文書集合の規模は、参加者が扱いやすい規模 (10GB 程度) と、ある程度現実に近い規模 (100GB 程度) が設定されている。

タスクには、検索課題検索、類書検索、ターゲット検索がある。検索課題検索は、TREC の Ad Hoc 型検索に相当する。各検索課題ごとに適合度順の検索結果の上位 1000 ページを順位付きで提出する。各参加チームにより提出された上位文書を集めて正解候補とし，人間の判定者が、4 段階適合性判定 (高度に適合、適合、部分的適合、不適合) を行なう[1]。類書検索は、キーワードや文章による検索要求の変わりに、文書を検索要求として用いる。ターゲット検索は、精度重視のタスクである。

それぞれの検索課題は、SGML 形式になっており、課題 ID (<NUM>)、もっとも中心的な主題を表わす 1 ～ 3 語からなる <TITLE>、もっとも基本的な検索要求の記述である DESCRIPTION (<DESC>)、背景、検索目的、判定基準、用語の定義などの詳しい説明である NARRATIVE (<NARR>)、検索課題作成者によって定義された同義語、関連語、上位語を示す CONCEPTS (<CONC>)、3 つの適合文書の例 (文書 ID) を示す <RDOC>、検索課題作成者の属性を示す USER ATTRIBUTES (<USER>) からなる。例を図 1 に示す。

[1] このように参加チームの出力の中からだけ正解を探す手法をプーリングと呼ぶ。

<TOPIC>
<NUM>0043</NUM>
<TITLE CASE="c" RELAT="1-3">シフォンケーキ，作り方，菓子</TITLE>
<DESC>「シフォンケーキ」の作り方が書かれている文書を探したい</DESC>
<NARR><RELE>適合文書は「シフォンケーキ」の作り方（材料や分量）が説明されており、色々な「シフォンケーキ」のバリエーションが紹介されているもの。</RELE></NARR>
<CONC>シフォンケーキ，菓子，作り方，製菓，バリエーション</CONC>
<RDOC>NW013569355，NW011761975，NW009137107</RDOC>
<USER>大学院修士2年，女性，検索歴4年</USER>
</TOPIC>

図 1. NTCIR コレクション検索課題の例（課題番号 0043）

3 NTCIR-3 音声入力 Web 検索タスク

3.1 背景と動機

音声入力を持つシステムの評価方法について考えてみる。システムの評価には大きく分けて「実世界における試験運用」と「研究室における実験」がある。前者は、入出力インタフェースの設計や応答時間などの運用上の問題まで考慮して評価を行う必要がある。それに対して、後者は比較的限定された設定のもとで行う評価である。すなわち、あらかじめ用意された検索要求に対してシステムが出力した検索結果を何らかの尺度によって評価する。

NTCIR ワークショップは後者の評価を行なうワークショップであり、テストコレクションを構築し、評価用ベンチマークとして利用する手法をとっている。大規模なテストコレクションの作成には膨大なコストを要するものの、一旦作ってしまえばシステムの性能評価を繰り返すことが容易になる。そこで、ユーザとなる被験者がいなくても、様々な手法を比較評価しながらシステムを改善できる。

Barnett ら [5] は、TREC の検索課題 35 件を話者に読み上げてもらい、音声による検索要求データを作成した。そこで、当該データを TREC のテキストコレクションおよび適合性判定と併用することで、音声入力によるテキスト検索の精度を定量的に評価することが可能である。

TREC のように一般公開されているテストコレクションを拡張すれば、研究者間でデータを共有することが容易になり、当該分野の発展に貢献できる。しかし、Barnett ら [5] が作成したデータに関しては、単一話者による読み上げであること以外の詳細（検索課題中のどの項目をどのような手順で収録したのかなど）は不明である。

それに対して、我々ではテストコレクションとして望ましいユーザ発話の特性について 3.2 節で検討し、実世界の検索要求にできるだけ近づけること目指す。

3.2 ユーザ発話に関する特性

3.2.1 発話の長さ

情報検索を広義に解釈すれば、ユーザが抱えている問題を解決できる情報を見つけ出すことである。すなわち、言語化されていない直観的な情報要求（visceral need [12]）を検索質問（query）に言語化して検索を行い、検索された内容を理解する一連の処理である。

しかし、多くの研究では情報検索は狭義に解釈され、具体化された検索質問をシステムに対する直接の入力として扱う。例えば、キーワードは極端に具体化された検索質問の形態である。

音声入力型システムの場合は、キーボード入力では躊躇するような曖昧または冗長な要求を発話することは比較的容易である。そこで、キーワードのような簡潔な語から説明的な文まで一律に認識できる必要がある。

他方において、ユーザが一度に無理なく自然に発話できる長さ（単語数や時間）には限度があるため、極端に長い検索要求を利用することは現実的ではない。

3.2.2 内容（語彙）の多様性

テキスト入力の検索においては、検索要求の内容に関する多様性が研究の焦点になることは少ない。しかし、音声入力型の検索においては、音声認識できる内容や語彙の多様性がシステムの適用範囲に深く関わる。

3.2.3 発話スタイル

音声認識の評価においては、読み上げ音声（read speech）、自発音声（spontaneous speech）、会話音声（conversational speech）等の発話スタイルを区別することが重要である。

検索システムの入力としては、読み上げ/自発音声を区別する必要がある。Barnett ら [5] が行ったように、既存の検索課題をそのまま発話すれば読み上げ音声である。他方において、話者が検索課題を理解し、課題達成に必要な検索要求を自分で考えて発話すれば、自発音声に近くなる。

3.2.4 話者の特性

対象を特定話者に限定するのか、不特定話者を想定するのかを考慮して、目的に応じた検索要求を作成する必要がある。また、話者の性別や年齢などについても検討する必要がある。

なお、以上 4 項目のうち、最初の 3 項目は音声認識とテキスト検索に共通の観点であり、残りの 1 項目は音声認識に固有の観点である。

3.3 フィージビリティスタディ

上記の議論に基づき、日本語テキスト検索用テストコレクションに音声による検索要求を加える形で音声入力検索のテストコレクションを試作した。現在一般に入手可能な日本語テキスト検索用テストコレクションのうち、NTCIR [13][2] と IREX [14][3] の日本語検索コレクションを対象にした。どちらのコレクションも米国の TREC[4] と同じようなシステム参加型のワークショップを通して構築された。

NTCIR が技術文書（論文抄録、科研費成果報告書概要）で構成されているのに対して、IREX は新聞記事（毎日新聞 1994-1995 年版）で構成されている[5]。

音声による検索要求としては、文単位の検索要求を採用した。具体的には、NTCIR では DESCRIPTION のみ、IREX では NARRATIVE のみを検索要求として利用した。すなわち、システムの入力としては、専門用語と一般語が混在する文単位の発話

[2]http://research.nii.ac.jp/ntcir/index-ja.html
[3]http://cs.nyu.edu/cs/projects/proteus/irex/
[4]http://trec.nist.gov/
[5]IREX コレクションには毎日新聞の記事 ID のみが含まれており、記事データそのものは含まれていない。

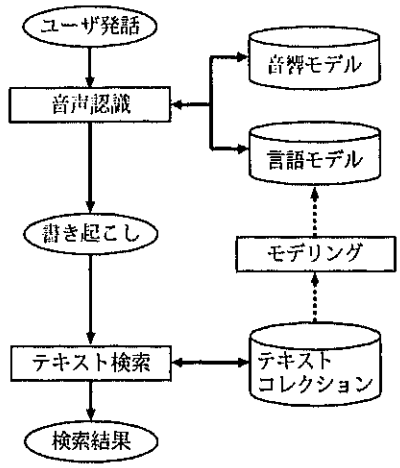

図 2. 音声入力型テキスト検索システムの構成

を想定する。この文単位の検索要求を男女それぞれ 2 名合計 4 名が読み上げる形で発話を収録した。読み上げの場合、発話の自然性はないが、反面、同一条件で多数の話者の発話を評価できる利点がある。

このテストコレクションで想定した音声入力による情報検索システムの構成を図 2 に示す。まず、オフライン処理 (破線矢印) によって、検索対象となるテキストコレクションから音声認識用の言語モデルを作成する。オンライン処理では、ユーザが検索要求を発話すると、音響モデルと言語モデルを用いて音声認識が行われ、書き起こしが生成される。次に、書き起こされた検索要求を用いてテキスト検索を実行し、検索結果を出力する。

このシステムで、テキスト検索以外の部分は誰でも共通に利用できるツールを用いて評価を行なったところ、以下の知見が得られた。

- 平均適合率[6]では、音声入力の場合にテキスト入力の場合より 4 ないし 6 ポイント程度低下する。
- 音声検索要求の認識については全体の単語誤り率より検索キーワードの誤り率が悪い。

この知見から、我々が提案する音声入力型テストコレクションの構築方法で一定の評価が可能なこと、音声入力型特有の解決すべき問題が存在することが明らかになった。

我々は、この知見に基づき、NTCIR-3 Web 検索音声入力タスク用のテストコレクションを作成した。以下にテストコレクションの詳細を述べる。

[6]あらかじめ決めた再現率の点について適合率を平均する。ここでは、再現率が (0.0,0.1,...,1.0) の 11 点を使って計算している。

3.4 音声入力型ウェブ検索タスク

音声入力型ウェブ検索タスクは、再利用可能なテストコレクションと関連ツールを、情報検索分野の研究者にも音声処理分野の研究者にも広く利用できるような形で整備することを目的とした。

3.3 章のフィージビリティスタディの結果から、Web 検索のメインタスクの検索要求を読み上げたものを音声による検索要求として整備することにした。したがって、正解判定は、メインタスクのテキスト入力の判定結果がそのまま利用できる。ただし、メインタスクの正解判定のプーリングには音声入力型のシステムは参加していないため、正解文書の数は目減りする点は注意を要する。

また、NTCIR ワークショップ参加者のほとんどは情報検索分野、または自然言語処理分野の研究者であり、音声認識システムの開発や扱いには不慣れであることを考慮して、フリーの音声認識ツールキット [15] で利用できる音声認識用の言語モデルと辞書を Web 検索の対象文書から作成し、標準モデルとしてテストコレクションに加えた。

音声検索要求としては、男女それぞれ 5 名の話者に 105 検索課題を静かなオフィス環境で読み上げさせ、接話マイクを用いて収録した。さらに、ATR 音韻バランス文セットの A セットを参照用に収録した[7]。

言語モデルはフィージビリティスタディの結果を受け、対象文書だけから作成した。10GB コレクションは、100GB コレクションの一部なので、今回のテストコレクションでは、100GB 文書だけから言語モデルを作成し、10GB コレクションの評価にも用いた。

言語モデルとしては、ごく標準的なものとして高頻度語に制限した単語トライグラムおよびバイグラムを作成した。語彙サイズは、2 万 (Web20K) と 6 万 (Web60K) とした。言語モデルの作成方法は、連続音声認識ツールキットでの作成方法 [16] に準じているが、Web 文書の場合、英語の文書なども混ざっているので、段落が ASCII 文字だけからなる場合などを排除する前処理を追加した。

平滑化には、バックオフスムージング (Witten-Bell ディスカウント) を用いた。カットオフは、Web20K のときに 20、Web60K のとき 10 とした。バイグラムは前向きのもの、トライグラムは前向きのものと逆向きのものの両方を ARPA 形式で作成した。また、認識用の辞書は、高頻度語から形態素解析で未知語となって読みが付与されなかった語を除いて HTK 形式で作成した。

図 1 に言語モデルの学習データに関する統計を示す。比較のため毎日新聞 10 年分 (1991 年から 2000 年) の統計も示す。テストコレクションだけから作成したといっても、新聞 10 年分の約 10 倍の分量があるため、言語モデルの学習には十分であるといえる。

表 1. 言語モデル学習データの統計

	Web (100GB)	News (10 years)
# of Word types	2.57M	0.32M
# of Word tokens	2.44G	0.26G

4 評価

4.1 概要

我々が本テストコレクションで想定している音声入力型情報検索システム (図 2) は、音声認識システムと情報検索システムからなる。

本章では、評価に用いたシステムに関して、4.2 章で音声認識システムについて述べ、4.3 章で情報検索システムについて述べる。さらに 4.4 章で、実験結果について述べる。

4.2 音声認識システム

音声認識システムとしては、日本語ディクテーションツールキット [15] を用いた。2000 年版に収録されたデコーダと音響モデル (16 混合分布 2000 状態のもの) をそのまま用いた。

言語モデルは、検索対象のコレクションから作成したもの (Web20K, Web60K) 以外に比較のために、ツールキット 2000 年版に収録されている毎日新聞 1991 年から 2000 年の 10 年分で作成された語彙サイズが 6 万語のモデルを用いた。以下このモデルを News60K とよぶ。

4.3 情報検索システム

テキスト検索には統計的手法 [17] を用いた。本手法は近年のいくつかの評価実験によって比較的高い検索精度を実現することが示されている。

検索要求が与えられると、索引語の頻度分布に基づいてコレクション中の各テキストに対する適合度

[7] 話者適応などに利用することを考慮した

表 2. 100GB コレクションの検索結果の比較 (OOV: 未知語率, WER: 単語誤り率, TER: キーワード誤り率)

					平均適合率 (100GB)			
Method	OOV	WER	TER	Time (sec.)	RC	RL	PC	PL
Text	—	—	—	—	.0855	.0982	.1257	.1274
Web60K	.0192	.1754	.3452	7.2	.0474	.0552	.0676	.0717
News60K	.0260	.2260	.4527	7.0	.0309	.0369	.0450	.0484
Web20K	.0856	.2208	.4335	6.7	.0281	.0339	.0410	.0438

を計算し、適合度が高いテキストから優先的に出力する。テキスト i の適合度は式 (1) によって計算される。

$$\sum_t \left(\frac{TF_{t,i}}{\frac{DL_i}{avglen} + TF_{t,i}} \cdot \log \frac{N}{DF_t} \right) \qquad (1)$$

ここで、t は検索要求 (本システムでは、ユーザ発話の書き起こしに相当する) に含まれる索引語である。$TF_{t,i}$ はテキスト i における索引語 t の出現頻度である。DF_t は対象コレクションにおいて索引語 t を含むテキストの数であり、N はコレクション中のテキスト総数である。DL_i はテキスト i の文書長 (バイト数) であり、$avglen$ はコレクション中の全テキストに関する平均長である。

適合度を適切に計算するためには、オフラインでの索引語抽出 (索引付け) が必要である。そこで「茶筌」を用いて単語分割、品詞付与を行う。さらに、品詞情報に基づいて内容語 (主に名詞) を抽出し、単語単位で索引付けを行って転置ファイルを作成する。オンライン処理では、書き起こされた検索要求に対しても同様の処理で索引語を抽出し、検索に利用する。

4.4 実験

Web 検索メインタスクの検索課題検索の検索課題のうちフォーマルランで正解判定結果が公表された 47 検索課題に関して評価を行なった。結果を図 2 に示す。検索性能は平均適合率で評価した。音声認識に関しては、単語誤り率以外に、検索に使われる内容語だけで評価したキーワード誤り率でも評価した。また単語誤り率 0% の場合として、テキスト入力の場合の評価結果を比較した。

正解判定では、論文や新聞などのテキスト検索とは違って、ハイパーテキストである特徴を生かした評価方法も取り入れられていてハイパーリンク情報の有無を考慮した評価が行える。表 2 では以下の 4 つの基準が用いられている。

RC (高) 適合、ハイパーリンク情報を用いない。
RL (高) 適合、ハイパーリンク情報を用いる。
PC 部分適合、ハイパーリンク情報を用いない。
PL 部分適合、ハイパーリンク情報を用いる。

我々の評価では、ハイパーリンク情報の有無で顕著な差は見られなかった。

音声認識まで含めた評価結果としては、News60K と比べ、Web60K の方が認識性能もよく、検索まで含めた評価結果もよいことから、対象文書から言語モデルを作成することの有効性が示された。

また、Web20K よりも Web60K の方が性能がよいことから、語彙サイズを大きくすることが有効であることもわかる。

全データの平均では、単語誤り率 が 20% 前後であるのに対して、キーワード誤り率は 40% 前後であり、検索用キーワード、すなわち名詞などの内容語の認識性能の向上の必要性が示唆されている。

今後は、キーワード誤り率や単語誤り率と平均適合率の関係など実験結果のより細かい分析を行なう予定である。

5 NTCIR ワークショップにおける音声技術の今後

残念ながら、今回の音声入力型タスクは多くの参加者を集めるには至らなかった。そのためタスクとして不十分な点もあるだろうし、テストコレクションに関しても不十分な点があるだろう。音声入力型タスクの今後についてはより参加者を増やすためにも、以下にあげる点を検討していきたいと考えている。

- <TITLE>に対応する単語発声を検索要求として追加する。
- 音声検索要求をより自由/自発的なものにする。
- NTCIR には Web 検索タスクの他に質問応答タスクも行われている。音声入力型タスクの適用可能性を検討する。
- 情報検索システム、音声認識システム、それぞれ単独の研究を行なっている研究グループが参

加しやすくなるようなツールの整備。

さらに、NTCIR ワークショップにおいて、音声処理技術と情報検索技術の融合を考えるとすれば、欧米の動向なども考慮すると、以下の点を検討する時機ではないだろうか。

- テキスト要約のタスクを、音声文書要約タスクに発展させる。
- 欧米では、音声文書検索のコレクションの整備段階を終え、ビデオ文書検索のコレクションの整備段階に移行している。この点を踏まえ、ビデオ文書に関する、より先進的なタスクの設定。

6 まとめ

NTCIR-3 ワークショップの Web 検索タスクにおいて、我々が実施した音声入力タスクについてその概要と評価結果を述べた。音声認識の評価という観点から見ると、情報検索という応用の評価まで一貫して行えることで、未知語の影響、認識誤りの影響、などについてより現実的な評価が可能になることがわかった。

情報検索も音声認識もそれぞれが複雑で大規模なシステムではあるが、NTCIR ワークショップのように再利用可能なコレクションを整備し、それらのコレクションを通じて、多様な分野の研究者が交流することで、より高度な研究が可能になることを期待する。

参考文献

[1] S. Johnson, P. Jourlin, G. Moore, K. S. Jones, and P. Woodland. The Cambridge University spoken document retrieval system. In *Proceedings of ICASSP'99*, pages 49–52, 1999.

[2] G. Jones, J. Foote, K. S. Jones, and S. Young. Retrieving spoken documents by combining multiple index sources. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 30–38, 1996.

[3] A. Singhal and F. Pereira. Document expansion for speech retrieval. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 34–41, 1999.

[4] M. Wechsler, E. Munteanu, and P. Schäuble. New techniques for open-vocabulary spoken document retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 20–27, 1998.

[5] Barnett, J., Anderson, S., Broglio, J., Singh, M., Hudson, R. and Kuo, S. W.: Experiments in Spoken Queries for Document Retrieval, *Proceedings of Eurospeech97*, pp. 1323–1326 (1997).

[6] F. Crestani. Word recognition errors and relevance feedback in spoken query processing. In *Proceedings of the Fourth International Conference on Flexible Query Answering Systems*, pages 267–281, 2000.

[7] A. Fujii, K. Itou, and T. Ishikawa. A method for open-vocabulary speech-driven text retrieval. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 188–195, 2002.

[8] A. Fujii, K. Itou, and T. Ishikawa. Speech-driven text retrieval: Using target IR collections for statistical language model adaptation in speech recognition. In A. R. Coden, E. W. Brown, and S. Srinivasan, editors, *Information Retrieval Techniques for Speech Applications (LNCS 2273)*, pages 94–104. Springer, 2002.

[9] K. Itou, A. Fujii, and T. Ishikawa. Language modeling for multi-domain speech-driven text retrieval. In *IEEE Automatic Speech Recognition and Understanding Workshop*, 2001.

[10] J. S. Garofolo, E. M. Voorhees, V. M. Stanford, and K. S. Jones. TREC-6 1997 spoken document retrieval track overview and results. In *Proceedings of the 6th Text REtrieval Conference*, pages 83–91, 1997.

[11] J. Kupiec, D. Kimber, and V. Balasubramanian. Speech-based retrieval using semantic co-occurrence filtering. In *Proceedings of the ARPA Human Language Technology Workshop*, pages 373–377, 1994.

[12] Taylor, R. S.: The Process of Asking Questions, *American Documentation*, Vol. 13, No. 4, pp. 391–396 (1962).

[13] National Institute of Informatics: *Proceedings of the 2nd NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization* (2001).

[14] Sekine, S. and Isahara, H.: IREX: IR and IE Evaluation project in Japanese, *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, pp. 1475–1480 (2000).

[15] T. Kawahara, A. Lee, T. Kobayashi, K. Takeda, N. Minematsu, S. Sagayama, K. Itou, A. Ito, M. Yamamoto, A. Yamada, T. Utsuro, and K. Shikano. Free software toolkit for Japanese large vocabulary continuous speech recognition. In *Proceedings of the 6th International Conference on Spoken Language Processing*, pages 476–479, 2000.

[16] 鹿野清宏, 伊藤克亘, 河原達也, 武田一哉, 山本幹雄 (編): 音声認識システム, オーム社 (2001).

[17] S. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 232–241, 1994.

WWW は大語彙連続音声認識の学習データとして使えるか? *

○伊藤克亘 †, 秋葉友良 (産総研) △藤井敦 †(図書館情報大) (†JST CREST)

1 はじめに

単語を単位とする N-gram モデルは、モデルの単純さから大語彙連続音声認識でも広く用いられている。しかし、最低でも数万語の規模のトライグラムを構築しようとすると、それなりの性能を示すためには、数百万から数千万の単語からなる電子化されたコーパスが必要であった。

日本語に関しては、大語彙連続音声認識の研究が本格化した 1990 年代後半に、それだけの規模を持ち、研究者が容易に入手できる電子化テキストコーパスは、新聞データくらいしかなく、音声認識用言語モデルとして適切なコーパスを質や量といった面から議論することも難しい状態であった。

しかし、ここ数年の WWW の成長で、今や、日本語のコンテンツに限っても、6000 万ページ近くに達する[1]といわれており、新聞数百年分ともいえる電子化テキストが手に入れられる状態となっているともいわれている。

また、WWW は、企業サイトや個人サイト、学校サイト、自治体サイトと提供母体が多様であり、内容も千差万別で非常に幅が広い。これらの特徴から、用途によってはバランスコーパスと同様に使えるという報告もある [1]。また、出現する語の種類も、従来の単一の辞書や事典の見出し語を越える見込みも得られている [2]。

そこで、本研究では、WWW を日本語の大語彙連続音声認識用の汎用言語モデルの学習用コーパスとして利用できるかどうかを実際に大量の Web ページからトライグラムモデルを構築して検討する。

2 大語彙連続音声認識のための言語モデル学習用コーパス

汎用の大語彙連続音声認識のための言語モデル学習用コーパスとして WWW 上の文書はどのような可能性があるか、ここで考えてみる。

日本語の大語彙連続音声認識用のコーパスとして公表されているものとしては、新聞がその中心であった [3, 4]。しかし、新聞は、紙面の制約などや客観性などの制限があるため、文体や表記が厳しく統制されている。また、報道される内容も非常に多くの読者を意識したものである。これらの特徴から、日本語全体から見ると、かなり偏った存在であるといえる。

特に音声認識の場合は認識対象は主として話し言葉であるため、話し言葉への対応を強化するために、パソコン通信の電子会議室のデータを利用した例などがあった [4]。さらに、放送大学の講義を書き起こしたコーパスも構築されている [5]。

話し言葉独特の表現も音声認識用言語モデルにとっては重要な課題ではある。しかし、いい淀みやいい直しなど、本来、テキストだけでは表現しにくく、それゆえに言語モデル単独で実現するのが適当でないと考えられる現象を除くと、99% 程度カバーするような語彙サイズで新聞からトライグラムを作成すると、その頻度は別にすれば多くの表現はカバーされるようになる。

むしろ、新聞では使われない専門用語の方が問題になる点も多く、これらの内容語は、音声認識の後段の応用で重要な役割を担うため、認識誤りを起さないように、強化する必要がある。そういった専門用語を含むコーパスとしてサーチエンジンで専門用語を検索した結果を利用する例 [6] や、検索した結果を利用して言語モデルを補強する例がある [7]。しかし、サーチエンジンは検索アルゴリズムやページ収集アルゴリズムが全て公開されているわけではなく、さらに日本語の場合、サーチエンジンが利用する形態素解析システムの問題もあり、サーチエンジンが出力する文書数などの情報がトライグラムモデルの構築にどのように影響を及ぼすかはよくわかっていない。

WWW 上の文書の特徴を、これらの視点からまとめ直してみると、まず、量が非常に多いことがあげられる。次に、日記など、話し言葉ふうの文書もかなり多い。また、専門用語を扱う文書も多く、専門用語の幅も広い。これらの特徴がどのように言語モデルに生かせるかを考えるためには、むしろ、まず、基礎的なデータとして WWW 上の文書から N-gram モデルを構築してみることの方が、WWW の潜在能力を測れるのではないかと考え、本稿では、なるべく多くの web ページを用いて N-gram モデルを構築し評価する。

3 実験

学習データの統計を表 1 に示す。全体では、新聞 10 年分のほぼ 10 倍と非常に膨大な学習データが得られた[2]。M は 100 万、G は 10 億を示す。1 で述べたように、日本語のページが 6000 万ページ程度だとすると、ページ数で見ると全体の約 20% 程度を使ったといえる。

評価用データとしては、放送大学のテレビ講義 (法律関係) 1 回分 (45 分) と、NTCIR の web 検索タス

---

* Can the WWW be used as a corpus for language modeling of large vocaburaly speech recognition by ITOU, K. (AIST, JST CREST) et. al.

[1] 平成 12 年 第 6 回 WWW コンテンツ統計調査 (郵政省郵政研究所) より

[2] 異り形態素数は、chasen が未知語として出力する文字列を全て形態素として計量しているため、実際の形態素の種類に比べるとかなり多めになっている。

表 1 学習データ

ページ数	形態素数 (のべ)	形態素数 (異り)
11.0 M	2.44G	2.57M
(新聞 10 年分)	0.26G	0.32M

ク [8] の検索要求を男女各 5 名が読み上げたものを用いた。

言語モデルの学習は、文献 [9] で述べた方法と同じであるが、web は新聞と違い日本語以外のコンテンツが多く含まれるため、文献 [6] で述べられているような選別をする必要がある。今回は、ページ数が多く重い処理ができなかったため、行ごとにアスキー文字だけの行のみを省く処理のみを行なった。したがって、かなり日本語でないデータが残っており、高頻度 2 万語にも、英単語が含まれていた。

認識実験としては、まず、新聞の場合と比較するために、語彙サイズをそれぞれ 2 万語, 6 万語と変化させてトライグラムモデルを作成した[3]。結果を表 2 に示す。表では、WER は単語誤り率、OOV は未知語率、#trigram はトライグラムの種類の数を示す。

表 2 新聞モデルと web モデルの比較

	web		新聞	
	2 万	6 万	2 万	6 万
放送大学 (WER %)	29.8	25.5	29.9	30.7
(OOV%)	3.9	0.8	2.7	0.9
NTCIR (WER %)	22.1	17.5	27.7	22.6
(OOV%)	8.6	1.9	6.8	2.6
#trigram(M)	8.5	16.4	13.8	15.9

次に学習データの影響を見るために、web 6 万語モデルにおいて学習データの量を変化させて言語モデルを作成し、放送大学データで評価した。結果を表 3 に示す[4]。

表 3 学習データ量を変化させた場合の比較

	1/10	1/5	1/2	全体
ページ数 (M)	1.0	2.2	5.5	11.1
形態素数 (のべ)(G)	0.23	0.49	1.24	2.44
WER(%)	28.7	27.6	27.2	25.5
#trigram(M)	3.6	7.3	8.9	16.4

最後にカットオフの値の影響を見るために、web 6 万語モデルにおいてカットオフの値を変化させて言語モデルを作成し、放送大学データで評価した。結果を表 4 に示す。

表 4 カットオフを変化させた場合の比較

カットオフ	20	10	5
WER(%)	26.6	25.5	25.9
#trigram(M)	9.0	16.4	29.5

ごくわずかの評価データの結果ではあるが、新聞モデルよりも良好な認識性能が得られていることがわかる。これらの認識結果の誤りを検討したが、改善された表現や単語に傾向などは特になかった。一般的に学習データを増やしたときに見られる未知語率の改善や、出現する三つ組の増加が、学習された言語モデルの性能向上に貢献していると考えられる。

web モデルの 6 万語のカットオフ 5 のときには、出現する三つ組の数が増大しているのにもかかわらず、単語誤り率が向上していない。これは、前述したように、日本語でないデータをはじめとしたデータを排除しきれておらず、それらの割合がかなり増えていると考えられ、前処理に課題があることを示している。

4 おわりに

日本語の web ページの約 1/5 にあたるデータから実際にトライグラムモデルを構築し、音声認識実験により評価を行なった。その結果、従来、よく利用されていた新聞記事 10 年分よりも単語誤り率を向上させることができた。

より多様な評価データでの評価を行なう必要があるが、「WWW は大語彙連続音声認識の学習データとして使えるか?」という問いの答えは、「使える」といってよいのではないだろうか。

[3] カットオフは新聞モデルは 1、web モデルは 2 万のときは 20、6 万のときは 10 である

[4] カットオフは、1/10 と 1/5 のときが 5 で 1/2 と全体のときが 10 である。また語彙は全て共通とした。

参考文献

[1] Frank Keller, Maria Lapata, and Olga Ourioupina. Using the web to overcome data sparseness. In *Proc. of EMNLP*, pp. 230–237. ACL, July 2002.

[2] 藤井敦, 伊藤克亘, 石川徹也. WWW は百科事典として使えるか? – 大規模コーパスの構築 –. 情報処理学会研究報告 自然言語処理, Vol. 2002-NL-149, pp. 7–14, May 2002.

[3] 河原達也, 李晃伸, 小林哲則, 武田一哉, 峯松信明, 伊藤克亘, 伊藤彰則, 山本幹雄, 山田篤, 宇津呂武仁, 鹿野清宏. 日本語ディクテーション基本ソフトウェア (97 年度版). 日本音響学会誌, Vol. 55, No. 3, pp. 175–180, Mar. 1999.

[4] 西村雅史, 伊東伸泰, 山崎一孝. 単語を認識単位とした日本語の大語彙連続音声認識. 情報処理学会論文誌, Vol. 40, No. 4, pp. 1395–1403, Apr. 1999.

[5] 伊東伸泰, 西村雅史. 口語体言語モデルのためのコーパス. 情報処理学会研究会資料 自然言語処理, Vol. 99-NL-134, pp. 9–14, 1999.

[6] 西村竜一, 長友健太郎, 小松久美子, 黒田由香, 李晃伸, 猿渡洋, 鹿野清宏. Web からの音声認識用言語モデル自動生成ツールの開発. 情報処理学会研究報告 音声言語情報処理, Vol. 2001-SLP-35, pp. 43–48, 2001.

[7] Xiaojin Zhu and Ronald Rosenfeld. Improving trigram language modeling with the world wide web. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, 2001.

[8] 大山敬三, 神門典子, 江口浩二, 栗山和子. Web 検索チャレンジの課題 – NTCIR ワークショップ 3 の新タスク–. 情報処理学会研究報告 情報学基礎, Vol. 2001-FI-63, pp. 41–48, 2001.

[9] 伊藤克亘, 山田篤, 天白成一, 山本俊一郎, 踊堂憲道, 宇津呂武仁, 山本幹雄, 鹿野清宏. 日本語ディクテーションのための言語資源・ツールの整備. 情報処理学会音声言語情報処理研究会資料, Vol. 99, No. 26-5, pp. 31–38, May 1999.

単語連接を重視したN-gram言語モデル平滑化手法の検討*

○秋葉友良, 伊藤克亘 (産業技術総合研究所), △藤井敦 (図書館情報大学)

1 はじめに

N-gram言語モデルでは、学習データに現れない単語列を扱うため、種々の平滑化 (スムージング) 手法が適用される。バックオフ・スムージングでは、高次のN-gramが存在しない場合、低次のN-gramで代用する。

tri-gramモデルの場合、tri-gramエントリが学習データに現れない場合、bi-gramを用いる。さらに、bi-gramがない場合、uni-gramを用いる。このうちuni-gramは、単に単語の出現頻度だけを反映し、平滑化手法にとってゼロ頻度を避けるファクタとして機能する一方、単語連鎖に関する情報を全く持たない。そのため、単語連鎖を表現したN-gramモデルにとって、必ずしも適切なファクタとはなっていない。例えば、日本語の場合、出現頻度の高い格助詞「の」などは、文脈に関係なく高確率で予測されることになる。そのため、従来のN-gramモデルを用いた大語彙音声認識では、このような高頻度語を誤認識する場合が多く見られる。

本稿では、十分なn-gram($n > 1$)があれば、uni-gramの影響を無効化あるいは軽減することで言語モデルの性能を改善できるのではないかという直感に基づき、uni-gramを極力用いない種々の平滑化手法の検討を行なう。そして、各手法を認識実験により比較した結果を報告する。

2 単語連接を重視したバックオフ平滑化手法

バックオフ・スムージングの一般式は次のように表される。

$$P(w_i|w_{i-n+1}^{i-1}) = \begin{cases} d_{w_{i-n+1}^{i}} P_{ML}(w_i|w_{i-n+1}^{i-1}) & C(w_{i-n+1}^{i}) > 0 \\ \alpha(w_{i-n+1}^{i-1}) P(w_i|w_{i-n+2}^{i-1}) & C(w_{i-n+1}^{i}) = 0 \end{cases} \quad (1)$$

ここでd, P_{ML}, αは、それぞれ、ディスカウント係数、最尤推定によるN-gram確率、確率の総和を1とするための正規化係数である。

2.1 品詞を用いた選択的なスムージング

品詞によってuni-gramを用いないようにスムージングを行う。前単語との接続が限定される品詞を人手で選別し、予測する単語の品詞によって、uni-gramへバックオフする/しないを選択的に切替える [1, 8] ことにする。

uni-gramを用いない品詞集合をAとする。また、単語wの品詞を$pos(w)$とする。式1において、$n = 2$の場合に以下の式を用いる。[1]

$$P(w_i|w_{i-1}) = \begin{cases} d_{w_{i-1}^{i}} P_{ML}(w_i|w_{i-1}) & C(w_{i-1}w_i) > 0 \\ 0 & C(w_{i-1}w_i) = 0 \\ & \wedge\, pos(w_i) \in A \\ \alpha(w_{i-1}) P(w_i) & \text{otherwise} \end{cases} \quad (2)$$

2.2 Kneser-Neyスムージング

KneserとNey[3]は、絶対法 [4] を拡張した平滑化手法を示している (Kneser-Neyスムージングと呼ぶ)。純粋に平滑化手法として他の手法と比べた場合でも、英語に適用した例で優れた性能を示すことが報告されている [2]。Kneser-Neyスムージングでは、高次のN-gram確率が利用できない (信頼できない) 場合に使用する低次の確率として、最尤推定による確率$P_{ML}(w_i|w_{i-n+1}^{i-1})$の代わりに次の値$P_{KN}(w_i|w_{i-n+1}^{i-1})$を用いる。

$$P_{KN}(w_i|w_{i-n+1}) = \frac{|\{w_{i-n}|C(w_{i-n}^{i}) > 0\}|}{\sum_{w_i} |\{w_{i-n}|C(w_{i-n}^{i}) > 0\}|} \quad (3)$$

例えば、英語において"San Francisco"を多く含む学習データを用いる場合、"Francisco"のuni-gram確率が高くなり、その結果学習データに現れない"on Francisco"に対しても高い確率値が与えられる。しかし、"Francisco"は"San Francisco"の文脈でしか使われないので、uni-gramを用いるのは多くの場合不適切である。Kneser-Neyスムージングでは、uni-gram確率を文脈語の異なり語数に比例するファクタで置き換え、この問題を回避する。

Kneser-Neyスムージングでは、長さNのN-gramについて、$n < N$である全てのnに対し$P_{KN}(w_i|w_{i-n+1})$を用いる。同時に本稿では、uni-gramを置き換えるファクタとしてのみP_{KN}を用いるモデルも検討する。すなわち、式1において、$n = 2$

---

* "Using Word Connetion in N-gram Language Model Smoothing Methods" by Tomoyosi AKIBA, Katunobu ITOU (National Institute of Advanced Industrial Science and Technology), Atsushi FUJII (University of Library and Information Science)

[1] Aの決め方によっては、ある単語コンテキストw_iについて、bi-gramの存在しない、$C(w_iw_j) = 0$となる単語w_jの品詞が、すべてAに含まれる可能性がある。この場合、正規化係数αの操作でコンテキストw_iの確率の総和を1にすることができず、ディスカウント係数dの再調整 (あるいは$d = 1$とする) が必要となる。本稿で扱う範囲では、$|A|$を十分小さく取ったため、上記のような場合は当てはまらなかった。

表 1: uni-gram を用いない品詞

大分類	小分類
名詞	非自立-一般, 非自立-副詞可能, 非自立-助動詞語幹, 非自立-形容動詞語幹, 特殊, 特殊-助動詞語幹, 動詞非自立的, 引用文字列
動詞	非自立
形容詞	非自立
助詞	接続助詞, 間投助詞, 終助詞, 副助詞／並立助詞／終助詞, 連体化, 副詞化, 特殊
助動詞	*

の場合に以下の式を用いる。

$$P(w_i|w_{i-1}) = \begin{cases} d_{w_{i-1}^i} P_{ML}(w_i|w_{i-1}) \\ \qquad C(w_{i-1}w_i) > 0 \\ \alpha(w_{i-1}) \frac{|\{w_{i-1}|C(w_{i-1}w_i)>0\}|}{\sum_{w_i}|\{w_{i-1}|C(w_{i-1}w_i)>0\}|} \\ \qquad C(w_{i-1}w_i) = 0 \end{cases} \quad (4)$$

3 実験

WWW 上のテキストから獲得した [7] 単語数 2 万語 (20k) の頻度情報について、品詞を用いた選択的スムージング (*SU*)、Kneser-Ney スムージング (*KN*)、uni-gram へのバックオフのみ Kneser-Ney スムージング (*KN1*)、の各手法を用いて言語モデルを作成した。*SU* では、形態素解析システム Chasen の品詞体系を用い、表 1 の品詞集合を用いた。この品詞集合に対応する単語は 2 万語中 1075 語であった。

上記平滑化手法と同時に、ディスカウント係数を求めるために、Witten-Bell 法 [5](*WB*) と絶対法 [4](*ABS*) を併用した。ただし、*KN* と *KN1* については、本来の定義通り絶対法とのみ組み合わせた。

bi-gram および tri-gram のカットオフは、5-5, 10-10、また bi-gram 以上の HIT 率を上げるために 1-10, 1-40、の 4 通りで行った。

評価データには、新聞記事 100 文を、男性 2 人女性 2 人によって読み上げた音声データを用いた。デコーダには大語彙音声認識デコーダ julius[6] のバージョン 3.2 を使用した[2]。音響モデルには 2000 状態 16 混合性別非依存 triphone を用いた。言語モデル重みは julius の標準設定から変更を行っていない。

実験結果を表 2 に示す。提案法の中では、*SU* と Witten-Bell 法の組み合わせ、および *KN1* で認識率の改善が認められた[3]。

[2] uni-gram を 0 とする *SU* を考慮して、uni-gram factoring は行わない設定を用いた。また、Kneser-Ney では、LR モデルと RL モデルに不整合が生じる (uni-gram 確率が異なる。また、一方には存在するが他方には存在しない bi-gram エントリが生じる) ため、両者を同一のデータ構造で管理する julius ではそのまま扱うことができない。この点を修正し、LR モデル RL モデルを個別のデータ構造に保持するように変更を行った。

[3] *KN* の結果が良くないのは、デコーダの性質 (LR モデルと RL モデルを同時に用いる) に起因する可能性がある。

表 2: 認識実験結果

LM	disc.	WER (%)			
(cutoffs)		5-5	10-10	1-10	1-40
BASE	*WB*	17.7	17.6	17.6	18.2
BASE	*ABS*	17.9	17.9	18.5	18.7
SU	*WB*	17.3	17.7	17.5	18.1
SU	*ABS*	18.0	18.1	18.4	18.9
KN	*ABS*	18.8	19.2	20.1	19.6
KN1	*ABS*	17.2	17.2	18.2	18.2

bi-gram のカットオフを小さくとったモデル (1-10, 1-40) は、全般的に認識率の改善は認められなかった上、絶対法との組合わせで逆に悪化している。形態素解析誤りによるノイズの影響が考えられる。

ディスカウント手法との組合わせでは、絶対法に比べ Witten-Bell 法が全般的に良い結果を示した。Witten-Bell 法は後続する異なり単語数を用いたディスカウント手法であり、やはり単語連接を考慮した手法であるのは興味深い結果である。

4 結論

単語の連接を重視した平滑化手法として、品詞により uni-gram を用いない手法 (*SU*)、Kneser-Ney スムージング (*KN*)、および uni-gram のみ Kneser-Ney スムージングを用いる手法 (*KN1*) を検討した。認識実験の結果から、uni-gram の影響を押さえて単語連接を重視した手法を用いることにより、言語モデルの性能を改善できることを確認した。単語連接の情報は、学習データや、未知語の存在に大きく影響を受けると考えられるため、今後は学習データ量や言語モデルの語彙サイズとの関係を調査する予定である。

参考文献

[1] T. Akiba, K. Itou, A. Fujii, and T. Ishikawa. Selective back-off smoothing for incorporating grammatical constraints into the N-gram language model. In *Proceedings of ICSLP*, 2002.

[2] J. T. Goodman. A bit of progress in language modeling. *Computer Speech and Language*, 15(4):403–434, Oct 2001.

[3] R. Kneser and H. Ney. Improved backing-off for m-gram language modeling. In *Proceedings of ICASSP*, Vol. 1, pp. 181–184, 1995.

[4] H. Ney, U. Essen, and R. Kneser. On structuring probabilistic dependences in stochastic language modelling. *Computer Speech and Language*, 8(1):1–38, 1994.

[5] P. Placeway, R. Schwartz, P. Fung, and L. Nguyen. The estimation of powerful language models from small and large corpora. In *Proceedings of ICASSP*, Vol. 2, pp. 33–36, 1993.

[6] 鹿野, 伊藤, 河原, 武田, 山本 (編) . 音声認識システム. オーム社, 2001.

[7] 伊藤, 秋葉, 藤井. WWW は大語彙連続音声認識の学習データとして使えるか? 日本音響学会秋季研究発表会講演論文集, Sep 2002.

[8] 秋葉, 伊藤, 藤井, 石川. 部分 N-gram 頻度情報を利用した質問応答定型表現への言語モデル適応. 情報処理学会研究報告 2002-SLP-42, pp. 31–38, 2002.

部分N-gram頻度情報を利用した質問応答定型表現への言語モデル適応

秋葉友良† 伊藤克亘†* 藤井敦‡* 石川徹也‡

† 産業技術総合研究所 ‡ 図書館情報大学 * 科学技術振興事業団 CREST

概要 音声入力に対応した質問応答システムのための言語モデルを作成する手法について述べる。検索対象となる新聞記事から作成したN-gramをベースに、人手で与えた質問文定型表現を用いて適応化する2つの手法を提案する。一つは、ベースとなる新聞記事N-gramモデル中の質問文定型表現に対応するN-gram頻度を、部分的なN-gram頻度で強調する手法である。もう一つは、定型表現を記述文法で表し、新聞記事N-gramと統合する手法である。認識実験を行い、N-gram頻度を重み付きで混合する従来法とくらべ、どちらの手法も単語誤り率を減少させることが示された。特に、前者の手法が認識率と頑健性の面でより良い結果を示した。

Using Extra N-gram Counts for Statistical Language Model Adaptation in Speech-Driven Question Answering

Tomoyosi AKIBA† Katunobu ITOU†* Atsushi FUJII‡* Tetsuya ISHIKAWA‡

† National Institute of Advanced Industrial Science and Technology (AIST)

‡ University of Library and Information Science

*CREST, Japan Science and Technology Corporation

Abstract Aiming at speech-driven question answering, we propose two methods to produce statistical language models for recognizing spoken questions with a high accuracy. Both methods use a target collection (i.e., a document set from which answers are derived) to extract N-grams, and adapt them to the question answering task by way of frozen patterns typically used in interrogative questions. The first method magnifies N-gram counts corresponding to the frozen patterns in the original N-gram. The second method combines N-grams extracted from the collection and grammars associated with frozen patterns, to produce a single N-gram model. Our experiments showed that the two proposed methods outperformed a conventional language model adaptation method in terms of the recognition accuracy, and that the first method was more accurate and robust than the second method.

1 はじめに

質問応答 (QA) は、1999年のTREC-8[8]にタスクとして採択されて以来、次世代の情報検索技術を目指した評価タスクとして注目されている。従来の情報検索タスクも音声入力に対応するように拡張されてきたが[4, 12]、質問応答では入力が質問文という話し言葉に近い表現が使用されることから、より音声入力に適したタスクであると考えられる。我々はこのような、音声入力を前提とした質問応答システムを開発中である。本稿では、音声認識部で利用する、質問文に対応した言語モデルを構築する手法について述べる。

質問応答システムへの入力は、検索のトピックに関する表現、質問文の文末に現れる定型的な表現、の2つの部分から構成される言語表現となる。例えば、次のような質問文が想定される。

「1976年に火星に軟着陸した探査機は何という名前でしたか」

前半の「1976年に火星に軟着陸した探査機は」の部分は、検索のトピックに関する部分で、新聞記事・辞典など質問応答の検索対象となる文書からそのまま学習した言語モデルで対応できる。本稿では、新聞記事を対象とした質問応答[1]を想定し、以降新聞記事モデルと呼ぶことにする。一方、後半の「何という名前でしたか」の部分は質問文に典型的に現れるパターンであり、新聞記事では稀な表現となるので、新聞記事モデルだけで扱うのでは不十分である。

言語モデルのタスク適応については、一般的な大量の学習データと、対象タスク用の比較的少量の学習データから、対象タスク用の言語モデルを作成する方法が一般的である。この枠組みに従えば、QA質問

文用の言語モデル作成には、新聞記事モデルをベースに、定型表現で構成された学習データを用いてタスク適応する方法が考えられる。しかし、本稿で対象とするQA質問文は、以下にあげる特徴を持つと仮定することで、単に漠然と2つのモデルを混合するよりは、より適切な手法が適用できると考えられる。(1) 前半部分だけなら、適応なしの新聞記事モデルで扱える。(2) 後半部分の語彙は新聞記事モデルでカバーできる程度の一般性を持つ。すなわち「語彙」ではなく「言い回し」を獲得したい。(3) 後半部分は「定型的」で、人手で列挙可能な程度の多様度を持つ。(4) 一文は性質の異なる前半と後半の2つの部分から構成される。よって、両者の接続部分のモデル化が重要となる。特に、2つの部分を独立したテキストから学習し、混合したのでは接続部分をうまく学習できないと考えられる。

本稿では、このような特徴を持つQA質問文を扱う2種類の言語モデル適応手法を提案する。一つは、定型表現の部分に対応する部分的なN-gram頻度情報を割増しする手法である(3節)。もう一つは、人手で記述した文法で表される制約をN-gram言語モデルに導入する手法である(4節)。

2 N-gram頻度の混合によるタスク適応

言語モデルのタスク適応については、一般的な大量の学習テキストデータと、対象タスク用の比較的少量の学習データから、対象タスク用の言語モデルを獲得する方法が知られている。多くの適応化手法では少量の学習データが利用可能であることを仮定している。しかし、少量のテキストデータでも獲得のコストは決して小さくない。代替手法として、対象タスクの文法を記述してテキストデータを自動生成する方法[5]、対象タスクの典型的な例文を用いる方法[11]が提案されている。提案法では、1節で述べた適応タスクの特徴(3)「人手で列挙可能な程度の多様度を持つ」を考慮して、定型表現部の言語表現を人手で(例文か文法で)与えることにする。

適応化手法としては、学習データのN-gram頻度を重み付きで足し合わせる方法が知られている[3, 9]。この手法の概念図を図1に示す。ここで混合するN-gram頻度は、共にテキストデータから直接獲得した頻度である。そのため、それぞれのN-gram頻度、および混合後のN-gram頻度は、次のような性質を持つ。

長さに関する整合性 任意の長さのN-gram頻度は、より長いN-gram頻度から一意に計算できる。

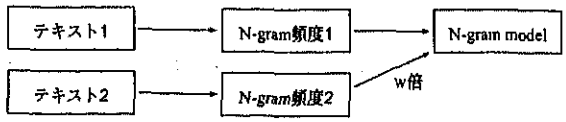

図1: N-gram頻度の混合によるタスク適応

(例) tri-gram頻度$C_{(3)}$が与えられているとする。この時、bi-gram w_pw_q の頻度は、以下のいずれかの方法で求められる。

$$C_{(2)}(w_pw_q) = \sum_{w_i} C_{(3)}(w_iw_pw_q) = \sum_{w_i} C_{(3)}(w_pw_qw_i)$$

同様に、1-gram頻度$C_{(1)}$は次のように求められる。

$$C_{(1)}(w_p) = \sum_{w_i} C_{(2)}(w_iw_p) = \sum_{w_i} C_{(2)}(w_pw_i)$$

この性質は、単語の連続としてのテキストの性質が反映された結果である。

このN-gram頻度の混合による適応化手法を、本稿のQA質問文タスクに適用することを考える。前半を新聞記事から、後半を定型表現を記した例文から学習できる。しかし、タスクの特徴(4)「一文は性質の異なる前半と後半の2つの部分から構成される」を考慮すると、新聞記事と定型表現のどちらの学習データにも接続部分に対するテキストを含んでいない。そのため、接続部分をうまく学習できないと考えられる。

3 部分N-gram頻度情報を用いたタスク適応

タスクの特徴(2)「後半部分の語彙は新聞記事モデルでカバーできる程度の一般性を持つ」を考慮すると、定型表現に対応する語彙はすでに新聞記事モデルに含まれていると仮定できる。また、定型表現に対応する単語の連接情報(N-gram頻度)もある程度含まれていると考えられる。そこで、新聞記事モデルに含まれる定型表現に関するN-gram頻度をそのまま活用してタスク適応することを考える。

提案する手法の概念図を図2に示す。新聞記事から作成したベースとなるN-gram頻度情報に対し、**部分的なN-gram頻度情報**を与える。ここで「部分的」とは、テキストデータから作成した整合的なN-gram頻度**ではない**ことを意味する。何らかの知識源によって、任意のN-gram頻度を直接与える。その結果、N-gram

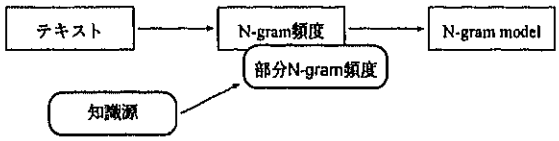

図 2: 部分 N-gram 頻度を用いたタスク適応

頻度は、2 節で述べた「長さに関する整合性」が成り立たない。

3.1 部分 N-gram 頻度情報を用いた確率計算

整合性のない N-gram 頻度を扱う場合、次に述べるような注意が必要となる。まず、より短い N-gram 頻度は、より長い N-gram 頻度から計算することができない。したがって、N-gram 頻度情報は、長さ n 毎に保持する必要がある。以下では、長さ n 毎に N-gram 頻度を $\mathbf{C}_n$ と記す。

スムージング・モデルの計算式にも注意が必要である。バックオフ・スムージングの一般式は次のように表される。

$$P(w_i|w_{i-n+1}^{i-1}) = \begin{cases} d_{w_{i-n+1}^i} P_{ML}(w_i|w_{i-n+1}^{i-1}) \\ \qquad \cdots C(w_{i-n+1}^i) > 0 \\ \alpha(w_{i-n+1}^{i-1}) P(w_i|w_{i-n+2}^{i-1}) \\ \qquad \cdots C(w_{i-n+1}^i) = 0 \end{cases} \tag{1}$$

ここで d, P_{ML}, α は、それぞれ、ディスカウント係数、最尤推定による N-gram 確率、確率の総和を 1 とするための正規化係数である。このうち、α は他の値から自動的に求まるので、d と P_{ML} について計算式を再検討する。

最尤推定による N-gram 確率は、通常次の式で求めることができる。

$$P_{ML}(w_{i-n+1}^i) = \frac{C(w_{i-n+1}^i)}{C(w_{i-n+1}^{i-1})}$$

しかし、部分的な N-gram 頻度を用いる場合、$\mathbf{C}_n$ と $\mathbf{C}_{n-1}$ の不整合から、両者を同時に用いることはできず、$\mathbf{C}_n$ だけから次のように計算する必要がある。

$$P_{ML}(w_{i-n+1}^i) = \frac{\mathbf{C}_n(w_{i-n+1}^i)}{\sum_{w_i} \mathbf{C}_n(w_{i-n+1}^i)}$$

同様に、ディスカウント係数 d の計算にも注意が必要である。例えば、Witten-Bell 法 [7] では、次の式を用いる。

$$d_{\mathrm{WB},w_{i-n+1}^i} = \frac{C(w_{i-n+1}^{i-1})}{C(w_{i-n+1}^{i-1}) + r(w_{i-n+1}^{i-1})}$$

ここで $r(w_{i-n+1}^{i-1})$ は、文脈 w_{i-n+1}^{i-1} の次に出現する異なり単語数である。部分的な N-gram 頻度を用いる場合は、長さ毎に r の情報源を特定する必要がある。

$$d_{\mathrm{WB},w_{i-n+1}^i} = \frac{\sum_{w_i} \mathbf{C}_n(w_{i-n+1}^i)}{\{\sum_{w_i} \mathbf{C}_n(w_{i-n+1}^i)\} + \mathbf{r}_n(w_{i-n+1}^{i-1})}$$

ここで $\mathbf{r}_n(w_{i-n+1}^{i-1})$ は、N-gram 頻度情報 $\mathbf{C}_n(w_{i-n+1}^i)$ から求めた、文脈 w_{i-n+1}^{i-1} の次に出現する異なり単語数である。

以上のように、長さ n の確率値は、$\mathbf{C}_n$ だけを使って計算するように注意する。

3.2 部分 N-gram 頻度情報による定型表現の強調

前節に述べた確率値計算に注意すれば、任意の部分的な N-gram 頻度を導入できる。特に、頻度情報は長さ毎に与えることができる。このことを利用して、定型表現を含む文だけが相対的に確率値が高くなるように、部分的な N-gram 頻度を与える。

定型表現を表す単語列 $\hat{w}_p^q = \hat{w}_p \hat{w}_{p+1} \cdots \hat{w}_{q-1} \hat{w}_q$ と、その左文脈となる単語列 $w_{p-N+1}^{p-1} = w_{p-N+1} \cdots w_{p-1}$ を考える。この単語列 $\hat{w}_p^q$ を含む文集合が共通して持つ、以下の N-gram 頻度を γ 倍して強調する (図 3)。

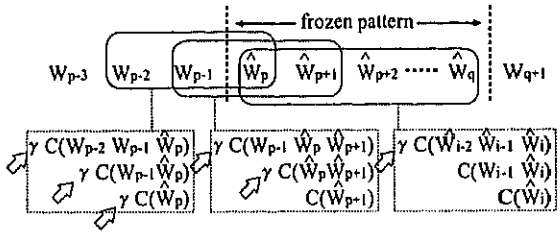

図 3: 部分 N-gram 頻度による定型表現の強調

1. 定型表現内部の単語列に対して、最も長い N-gram 頻度だけを強調する。

$$\mathbf{C}_N(\hat{w}_{i-N+1}^i) = \gamma C(\hat{w}_{i-N+1}^i) \tag{2}$$

例えば tri-gram モデルを構築する場合、tri-gram 頻度だけを γ 倍する。

$$\mathbf{C}_3(\hat{w}_{i-2}\hat{w}_{i-1}\hat{w}_i) = \gamma C(\hat{w}_{i-2}\hat{w}_{i-1}\hat{w}_i)$$

2. 定型表現の接頭単語列に対して、接頭単語列長以上の長さのN-gram頻度を強調する。接頭単語帳をkとすると、すべての文脈単語列w_{p-n+k}^{p-1}について、

$$\mathrm{C}_n(w_{p-n+k}^{p-1}\hat{w}_p^{p+k-1}) = \gamma C(w_{p-n+k}^{p-1}\hat{w}_p^{p+k-1}) \quad (3)$$

tri-gramモデルの場合、すべての文脈単語列$w_{p-2}w_{p-1}$について、次のように各長さの頻度を強調する。

$$\begin{aligned}
\mathrm{C}_3(w_{p-1}\hat{w}_p\hat{w}_{p+1}) &= \gamma C(w_{p-1}\hat{w}_p\hat{w}_{p+1}) \\
\mathrm{C}_2(\hat{w}_p\hat{w}_{p+1}) &= \gamma C(\hat{w}_p\hat{w}_{p+1}) \\
\mathrm{C}_3(w_{p-2}w_{p-1}\hat{w}_p) &= \gamma C(w_{p-2}w_{p-1}\hat{w}_p) \\
\mathrm{C}_2(w_{p-1}\hat{w}_p) &= \gamma C(w_{p-1}\hat{w}_p) \\
\mathrm{C}_1(\hat{w}_p) &= \gamma C(\hat{w}_p)
\end{aligned}$$

3. それ以外のN-gram頻度情報は、元のN-gram頻度と同じとする。$n = 1 \cdots N$について、

$$\mathrm{C}_n(w_{i-n+1}^{i}) = C(w_{i-n+1}^{i}) \quad (4)$$

4 N-gramモデルへの文法的制約の導入

QA質問文への言語モデル適応のもう一つの手法を示す。定型表現を記述文法で表し、新聞記事などから学習したN-gramモデル(以降、ベースN-gramと呼ぶ)と統合する手法である[2, 10]。

N-gramでモデル化される単語列は、全ての単語が互いに連接可能な単語ネットワークとして表現することができる。一方、記述文法(正規言語[1])で表される単語列は、文法によって部分的な単語連接だけを許した単語ネットワークで表現される。ベースN-gramの任意の位置から記述文法の先頭単語へ、記述文法の末尾単語からベースN-gramの任意の位置へ、両ネットワークを結合することによって、両モデルを統合し、一つの単語ネットワークで表すことができる(図4)。

統合した単語ネットワークにおいて、連接しない(弧の存在しない)単語列に対し確率値が0となるように、

[1]本手法で扱える記述文法は、(N-gramの表現力の制約の故に)正規言語までである。自然言語の記述として広く用いられている文脈自由言語は、そのまま埋め込むことはできない。しかし、有限長の文は必ず正規言語で表現できること、文脈自由文法を正規文法に近似するアルゴリズムが知られている[6]こと、などから実用上ほとんど問題はない。

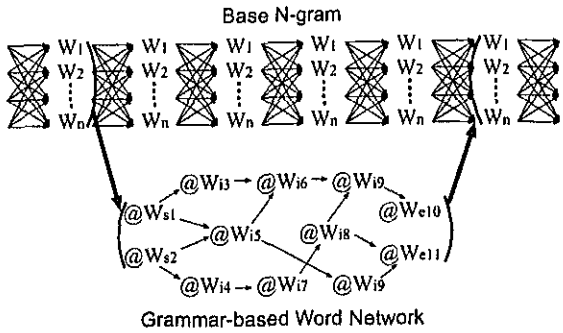

図4: 単語ネットワークの統合

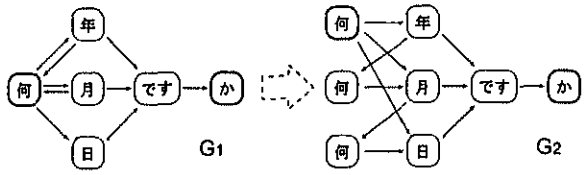

図5: 単語ネットワーク

N-gram確率を割り当てれば、汎用N-gramと記述文法の両方の性質を同時に保持するモデルを獲得できる。

4.1 単語ネットワーク

正規言語を、単語を頂点とし可能な単語連接を有効弧で表した単語ネットワークで表現することを考える。このような単語ネットワークは、例文の集合から簡単に獲得可能である。例えば、年月日を尋ねる発話を表した以下の例文から文法を獲得することを考える。

何/年/です/か　　何/年/何/月/です/か

何/月/何/日/です/か

この3文から獲得できる連接可能な単語対は以下の通りである。

A = { (何, 年) (何, 月) (何, 日) (年, 何) (月, 何) (年, です) (月, です) (日, です) (です, か) }

この単語対だけが連接可能であると考えると、文法(G_1)は4つ組(W_a, W_s, W_f, A)で表現できる。ここで、W_a,W_s,W_fは、それぞれ、全単語集合、先頭単語集合、末尾単語集合であり、

W_a = { 何 年 月 日 です か },

W_s = { 何 }, W_f = { か }

となる。G_1のグラフ表現を図5左に示す。

文法G_1は「何年何年ですか」「何月何年ですか」「何年何日ですか」のような、意図されない言語表現まで

受理してしまう。そこで、文法作成者の持つ言語知識を利用して、好ましくない表現を排除し、図5右のような文法 G_2 に修正することを考える。新たに導入したノード(文脈)毎に、新たな単語記号を導入して、次のような文法 (W'_a, W_s, W_f, A') として表現する。

$W'_a = W_a \cup \{$ 何1 何2$\}$

$A' = \{$(何, 年) (何, 月) (何, 日) (年, 何1) (何1, 月) (月, 何2) (何2, 日) (年, です) (月, です) (日, です) (です, か) $\}$

文法 G_2 は、「何年何月ですか」「何月何日ですか」のような、作成者の意図する表現だけを受理し、それ以外を排除する。このように、単語ネットワーク(正規言語)では、人の持つ言語知識を利用して、N-gramでは獲得不可能な、単語間の長距離の依存関係も表現することが可能である。

ここで、ベースN-gramと統合するために、単語ネットワークで表された文法は以下の2つの条件を満たす必要がある。第1に、ベースN-gramと文法の語彙は区別されている必要がある。これは、文法の語彙に特別な単語記号を割り当てれば良い。本稿では、単語記号の先頭に "@" をつけて表すことにする。ベースN-gramの語彙を W_U、文法の語彙を W_A とすると、$W_U \cap W_A = \phi$ である。第2に、文法の先頭単語は、先頭以外の個所に現れてはいけない。同様に、文法の末尾単語は、末尾以外に現れてはいけない。すなわち、文法の語彙 W_A は、互いに共通部分のない、先頭単語集合 W_B、中間単語集合 W_I、末尾単語集合 W_E から構成されるとする $(W_A = W_B \cup W_I \cup W_E \wedge W_B \cap W_I = \phi \wedge W_I \cap W_E = \phi \wedge W_B \cap W_E = \phi)$。

4.2 文法部へのN-gram頻度割り当て

文法部の単語ネットワークにも確率値を割り当てるために、頻度情報を与える。ここで与える頻度も「部分的」であるため、3節で述べた計算方法に従うことに注意する。頻度の与え方は様々な方法が考えられる。例えば、最も簡単な方法として、全ての分岐で等しい頻度を与える方法が考えられる。ここでは、タスクの特徴(2)から、ベースN-gramの頻度情報を利用する方法を採用した。文法の語彙をベースN-gramと同じ単位で構成すれば、文法内部の単語列 $@w^i_{i-n+1}$ に対して、対応するベースN-gramの単語列 w^i_{i-n+1} が必ず存在することを利用する方法である。

文法内部の単語列 $@w^i_{i-n+1}$ に対するN-gram頻度 $\mathbf{C}_n(@w^i_{i-n+1})$ を、ベースN-gramでの単語頻度 $\mathbf{C}_n(w^i_{i-n+1})$ を利用し、そのまま値をコピーして与える。また、ベースN-gramから文法へ遷移する部分の頻度 $C_n(w^{i-k}_{i-n+1}@w^i_{i-k+1})$ と、文法からベースN-gramへ遷移する部分の頻度 $C_n(@w^{i-k}_{i-n+1}w^i_{i-k+1})$ も、対応するベースN-gramでの頻度 $C_n(w^{i-k}_{i-n+1}w^i_{i-k+1})$ を利用する。このうち前者については、頻度を γ 倍して強調する。これは、文法でモデル化する定型表現を優先的に扱うためである(図6)。

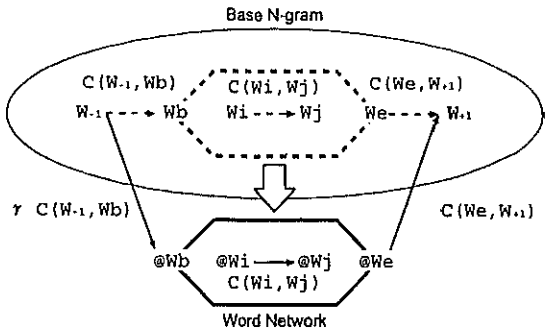

図6: 頻度情報のコピー (bi-gramの場合)

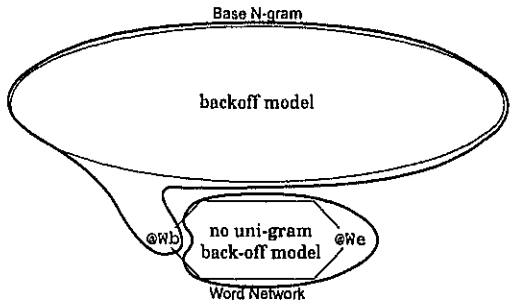

図7: Selective Back-off Smoothing

4.3 選択的なバックオフ・スムージング

獲得した単語ネットワークと頻度情報から、N-gram確率を計算する。この時問題となるのは、スムージング手法と文法的制約は、両立できないということである。バックオフ・スムージングでは、高次のN-gramが存在しない場合、低次のN-gramで補間する。ネットワーク全体を等しくスムージングすると、文法によって記述された二値的な単語連接の制約がuni-gramで補間され、結局全ての単語間の連接を許すモデルとなってしまう。一方、全くスムージングを行わないモデルを作成することも出来るが、その場合文法部の二値的制約は獲得されるが、N-gram部にゼロ頻度問題が生じ、精度が落ちてしまう。

そこで、ネットワークの部分によってスムージング手法を切り替える選択的なバックオフ・スムージングを用いる (図 7)。ベース N-gram の単語と文法の先頭単語 ($w_i \in W_U \cup W_B$) を予測する確率値は、式 (1) で表される通常のバックオフ・スムージングで計算する。ただし、uni-gram 確率値は $W_U \cup W_B$ を全単語集合とみなして計算する (これは、次に述べるように、$W_I \cup W_E$ について uni-gram 確率を 0 とするからである)。残りの単語、すなわち文法の中間単語と末尾単語 ($w_i \in W_I \cup W_E$) を予測する確率値は、uni-gram へバックオフしないように計算する。すなわち、長さ $n > 2$ については、式 (1) をそのまま使用するが、$n = 2$ については次の計算式を用いる。

$$P(w_i|w_{i-1}) = \begin{cases} d'_{w_{i-1}^i} P_{ML}(w_i|w_{i-1}) & \mathrm{C}_2(w_{i-1}^i) > 0 \\ 0 & \mathrm{C}_2(w_{i-1}^i) = 0 \end{cases} \quad (5)$$

この計算によって、次の関係が常に成立する。

$$P(w_i) = 0 \quad \text{if} \quad w_i \in W_I \cup W_E \quad (6)$$

$$\alpha_1(w_{i-1}) = 0 \quad \text{if} \quad w_{i-1} \in W_B \cup W_I \quad (7)$$

選択的なバックオフスムージングで獲得した N-gram モデル $P(w_i|w_{i-N+1}^{i-1})$ は、以下のような性質を示す。

- 文法部の中間単語および末尾単語 ($W_I \cup W_E$) を予測する確率値 $P(w_i|w_{i-N+1}^{i-1})$ は、単語列 w_{i-N+1}^i が文法により許されていない場合、必ず 0 となる。

文法で許されない単語列には頻度情報が与えられない。よって、$n > 1$ の N-gram 頻度 $\mathrm{C}_n(w_{i-n+1}^i)$ は 0 となり、$P(w_i|w_{i-N+1}^{i-1})$ は、uni-gram までバックオフされる。また、$w_i \in W_I \cup W_E$ なので、式 (6) より $P(w_i) = 0$。よって、

$$\begin{aligned} P(w_i|w_{i-N+1}^{i-1}) &= \alpha_{N-1} \cdots \alpha_2 \cdot \alpha_1(w_{i-1}) P(w_i) \\ &= 0 \end{aligned}$$

特に、ベース N-gram の単語 $w \in W_U$ から $w_i \in W_I \cup W_E$ を予測する確率は必ず 0 となり、ベース N-gram から文法の途中への単語への遷移は生じないことがわかる。

- 文法部の開始単語および中間単語 ($W_B \cup W_I$) からベース N-gram の単語 (W_U) を予測する確率は、必ず 0 となる。

文法の途中からベース N-gram の単語への N-gram 頻度は与えられていない。よって、この場合も $n > 1$ の N-gram 頻度 $\mathrm{C}_n(w_{i-n+1}^i)$ は 0 となり、確率値計算は uni-gram までバックオフする。この時、$w_{i-1} \in W_B \cup W_I$ なので、式 (7) より $\alpha(w_{i-1}) = 0$。よって、

$$\begin{aligned} P(w_i|w_{i-N+1}^{i-1}) &= \alpha_{N-1} \cdots \alpha_2 \cdot \alpha_1(w_{i-1}) P(w_i) \\ &= 0 \end{aligned}$$

以上により、獲得したモデルは文法の二値的な性質と N-gram の汎用性を兼ね備えていることが分る。また、文法で記述した言語表現に対し、uni-gram へバックオフしないで (discount なしで) 計算すること、γ で重み付けすること、から相対的に高い確率値を与える。これにより、文法で表現した表現を強調して扱うタスクに適応した言語モデルを得ることができる。また、従来のバックオフ言語モデルと互換性があり、言語モデルを置換えるだけで既存の音声認識デコーダでそのまま利用可能という、既存のシステムとの互換性の面でも優れた特徴を持つ。

5 実験

新聞記事 111 か月分から 2 万語の N-gram 頻度情報を抽出した。これをベースに、本稿で述べた種々の適応化手法の比較を行った。スムージング手法は、すべて Witten-Bell 法 [7] を用いた。適応タスクの学習データとして、QA 質問文の定型表現を受理する文法 (単語ネットワーク) (図 8) を作成した。また、この単語ネットワークから全文生成を行い、定型表現のパターンの集合を作成した。表記違いの単語を含めると、172 パターンが得られた。

まず、新聞記事だけから bi-gram および tri-gram を作成した (*BASE* と記す)。従来法として、N-gram 頻度の混合による手法 (2 節) で適応モデルを作成した。定型表現のパターン集合を適応化テキストデータとみなして N-gram 頻度を抽出し、重み w をかけて新聞記事 N-gram 頻度と混合、bi-gram および tri-gram を作成した (*MIX* と記す)。提案法の 1 として、3 節で述べた手法でモデルを作成した。定型表現のパターン集合を用い、新聞記事モデルの N-gram 頻度中の定型表現を (重み γ で) 強調し、bi-gram および tri-gram を作成した (*EMP* と記す)。提案法の 2 として、作成した文法を、4 節で述べた手法を用いて新聞記事 N-gram と (重み γ で) 統合、bi-gram および tri-gram を作成した (*NET* と記す)。

評価データには、新聞記事 100 文 (*NP*) と QA タスク用質問文 50 文 [13](*QA*) を、男性 2 人女性 2 人によって読み上げた音声データを用いた。作成したネッ

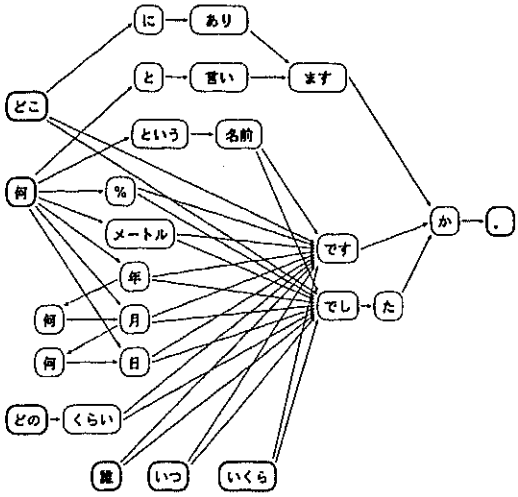

図 8: QA タスク定型表現の文法

表 1: 認識実験結果

target (# of sent.)	language model	WER (2-gram)	WER (3-gram)
NP (100)	*BASE*	18.0	10.7
	MIX	18.0	10.6
	EMP	18.0	10.3
	NET	18.8	10.4
QA (50)	*BASE*	26.3	16.9
	MIX	21.2	15.4
	EMP	19.6	13.8
	NET	23.4	14.7
QA' (36)	*BASE*	28.1	17.2
	MIX	21.8	15.2
	EMP	20.4	13.3
	NET	20.8	13.6

トワーク文法は、29 単語と比較的小規模のものであるが、質問文のうち 72%の 36 文 (*QA'*) が、この文法のモデル化する表現を含んでいた。

デコーダには大語彙音声認識デコーダ julius[14] のバージョン 3.2 を使用し[2]、音響モデルには 2000 状態 16 混合性別非依存 triphone を、言語モデル重みは新聞記事 N-gram(*BASE*) での最適値を用いた。bi-gram の比較には第一パスの結果を、tri-gram の比較には第一パスと第二パスの結果を用いた。

実験結果を表 1、図 9、図 10 に示す。

表 1 は、各手法で重みパラメータを調節して最も良い結果を示したものである。どの適応化手法もベースラインの新聞記事モデル (*BASE*) に比べて、単語誤り率 (WER) を改善している。また、従来法 (*MIX*) と比べ、提案法 (*EMP*,*NET*) は、おおむねより良い結果を示していることがわかる。

図 9、図 10 に、各手法における重みパラメータと単語誤り率との関係を示す。*NET* モデルは、文法で扱える表現だけの発話 (*QA'*) に対しては、重み γ の影響なく WER を引き下げるが、文法から少しはずれた発話を含む場合 (*QA*)、重みを上げることで WER が悪化する。文法の表現とは大きく異なる発話 (*NP*) の場合は、この傾向は認められない。これは、類似した発話を無理矢理文法で扱える表現で誤認識しまうことによる弊害と思われる[3]本手法を用いる場合、漏れのない文法記述が必要となることを示唆している。一方、*EMP* モデルではこのような傾向は認められず、より頑健な手法と考えられる。また認識率の面でも、*NET* よりも良い結果を示した。

6 まとめ

音声入力に対応した質問応答システムの言語モデルを獲得するため、検索対象となる新聞記事から作成した N-gram をベースに、人手で与えた質問文定型表現を用いて適応化する 2 つの手法を提案した。認識実験の結果、N-gram 頻度を重み付きで混合する従来法とくらべ、どちらの手法も単語誤り率を減少させることが示された。特に、前者の手法が認識率と頑健性の面でより良い結果を示した。提案法は、既存の N-gram 言語モデルを、比較的多様度が小さい (人手で記述できる程度の多様性を持つ) 表現に対応するための適応化手法として、他の分野にも適用可能であろう。

参考文献

[1] NTCIR workshop3 質問応答タスク. http://www.nlp.cs.ritsumei.ac.jp/qac, 2001.

[2] T. Akiba, K. Itou, A. Fujii, and T. Ishikawa. Selective back-off smoothing for incorporating grammatical constraints into the n-gram language model. In *Proceedings of International Conference on Spoken Language Processing*, 2002. (to appear).

[2]提案法で用いた部分的な N-gram 頻度から計算した N-gram 確率表は、LR モデルと RL モデルで不整合が生じる (一方には存在するが他方には存在しない N-gram エントリが生じる)。そのため、両者を同一のデータ構造で管理する julius ではそのまま扱うことができない。この点を修正し、LR モデル RL モデルを個別のデータ構造に保持するように変更を行った。

[3]特に、実験に使用したデコーダ (julius) では、第一パスで認識中の仮説を最尤近似するため、影響が大きく現れたと考えられる。

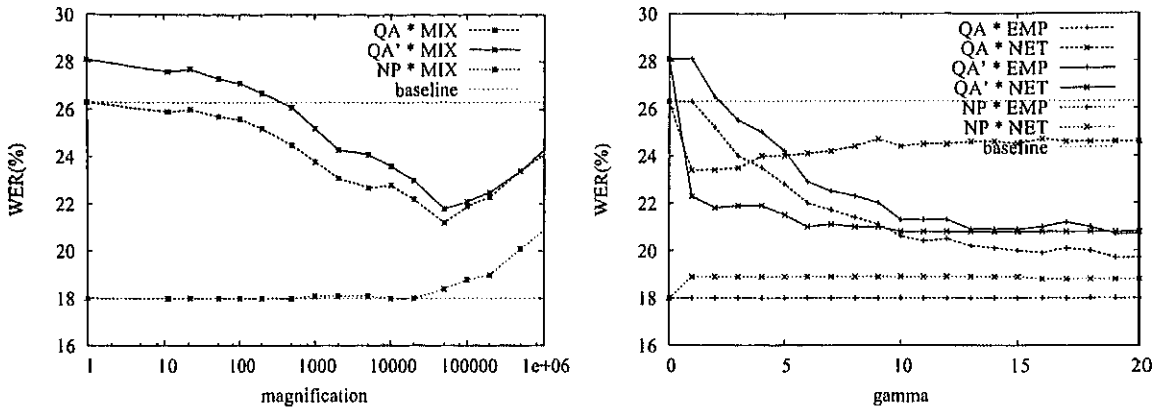

図 9: bi-gram 単語誤り率

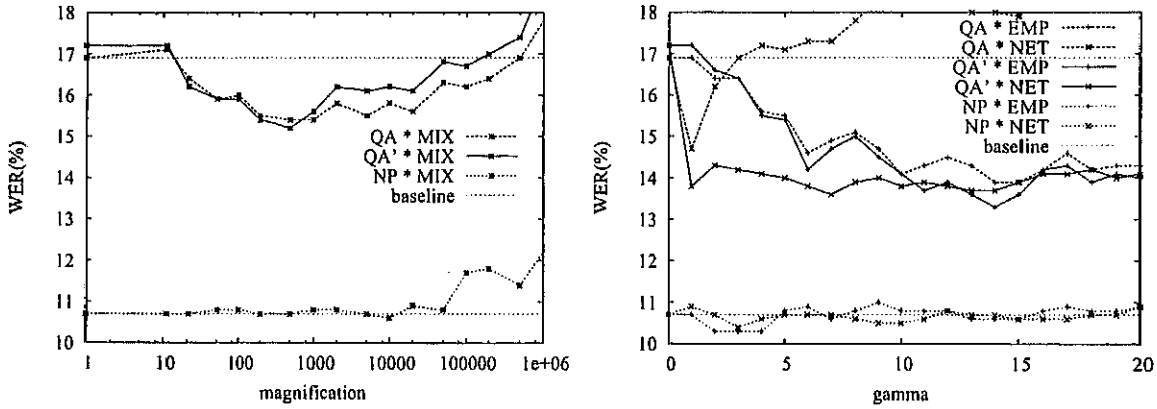

図 10: tri-gram 単語誤り率

[3] M. Federico. Bayesian estimation methods for n-gram language model adaptation. In *Proceedings of International Conference on Spoken Language Processing*, pp. 240–243, 1996.

[4] A. Fujii, K. Itou, and T. Ishikawa. Speech-driven text retrieval: Using target IR collections for statistical language model adaptation in speech recognition. In A. R. Coden, E. W. Brown, and S. Srinivasan eds., *Information Retrieval Techniques for Speech Applications (LNCS 2273)*, pp. 94–104. Springer, 2002.

[5] L. Galescu, E. Ringger, and J. Allen. Rapid language model development for new task domains. In *Proceedings of International Conference on Language Resources and Evaluation*, pp. 807–812, 1998.

[6] F. C. N. Pereira and R. R. Wright. Finite-state approximation of phrase-structure grammars. In *Proceedings of Annual Meeting of the Association for Computational Linguistics*, pp. 246–255, 1991.

[7] P. Placeway, R. Schwartz, P. Fung, and L. Nguyen. The estimation of powerful language models from small and large corpora. In *Proceedings of International Conference on Acoustics Speech and Signal Processing*, Vol. 2, pp. 33–36, 1993.

[8] E. Voorhees and D. Tice. The TREC-8 question answering track evaluation. In *Proceedings of the 8th Text Retrieval Conference*, pp. 83–106, 1999.

[9] 伊藤, 好田. N-gram 出現回数の混合によるタスク適応の性能解析. 信学論, J83-D-II(11):2418–2427, 2000.

[10] 秋葉, 伊藤, 藤井, 石川. 音声入力による質問応答システムのための音声認識用言語モデルの検討. 言語処理学会第 8 回年次大会発表論文集, pp. 244–247, 2002.

[11] 岡登, 石井, 花沢. タスクの例文を用いた自由発話音声認識のための言語モデルの構築. 日本音響学会秋季研究発表会講演論文集, pp. 73–74, Oct 2001.

[12] 伊藤, 秋葉, 藤井, 石川. 音声入力型テキスト検索システムのための音声認識. 日本音響学会秋季研究発表会講演論文集, pp. 193–194, Oct 2001.

[13] 佐々木, 磯崎, 平, 廣田, 賀沢, 中島. 質問応答システムの比較と評価. 信学技法 NLC-24, pp. 17–24, 2000.

[14] 鹿野, 伊藤, 河原, 武田, 山本（編）. 音声認識システム. オーム社, 2001.

音声文書検索の応用によるオンデマンド講演システム

藤井 敦 †,†††　伊藤克亘 ††,†††　石川徹也 †
† 図書館情報大学
†† 産業技術総合研究所
††† 科学技術振興事業団 CREST
fujii@ulis.ac.jp

1 はじめに

近年，マルチメディア情報の普及や情報通信のブロードバンド化に伴って，多種多様な情報を誰もが容易にオンラインで受信や発信できるようになった．このような現状では，大量な情報の中から必要な情報だけをいつでもどこでも手軽に活用できる基盤技術が重要である．

現在普及しているマルチメディア情報として，テキスト，音声，画像がある．これらが有機的に混在し，しかも繰返し利用したくなるコンテンツとして「講演」が挙げられる．講演は，予稿やスライドなどの資料を併用しながら，対面で話すのが一般的である．また，出版された教科書に基づいた講義を放映するテレビ番組もある．

我々は，講演のビデオデータを対象にして，要求に応じた内容を視聴するオンデマンドシステムの研究開発を行っている [4, 8]．本システムを用いると，教科書や予稿などの講演資料テキストを閲覧しながら，関心があるビデオ内容（音声と画像）を選択的に視聴することが可能になる．

2 資料と講演の相違点

一つの内容を伝達するために，資料（書き言葉）と発話による講演（話し言葉）という2つの異なる手段が存在する．もしも，どちらか一つの手段で講演が十分に成立するならば，本研究で提案するシステムの意義は希薄である．しかし，資料と講演は，一方が存在すれば他方は必要がないというような排他的な関係ではない．そこで，両者の相違点について考察することは，本研究の意義を明確にするために有効である．

資料は，冊子体や電子版などの形態に依らず，章立てのような文書構造や文字種などの表層情報を手掛かりにして「斜めに読む」ことができる．また，何度も繰り返し読み返すことができる．そこで，全体の概要を把握したり，関心のある箇所を高速に探索することが容易である．すなわち，ランダムアクセスに適している．

それに対して，講演は逐次アクセスが原則であり，ランダムアクセスには適さない．実際の講演では，内容が資料のページやスライドの単位で構造化されていても，資料のように戻ったり飛ばしたりして聞くことはできない．録画された講演ビデオの場合は，意味内容に基づいて索引付けを行わない限り，構造を手掛かりに早送りや巻戻しによって必要な箇所を特定することは困難である．音声認識（ディクテーション）によって発話内容をテキストに変換しても，書き起こしされた話し言葉を読むことは，資料を読む場合に比べて負担が大きい．

他方において，情報の量は講演の方が相対的に多い．資料は講演内容に対してページ数が制限されることがある．これは主に印刷コスト等の理由によるものなので電子版には該当しない．しかし，現状では冊子体と電子版の資料が等価な形で存在することが多いので，必然的に電子版も字数の制限を受ける．それに比べると，講演は発表内容に適した長さであることが多い．

また，資料を読むよりも講演を聞く方が分かりやすいことがある．発話内容が適度に冗長であったり，会話的な表現も使用される．熟練した講演者ならば，聴衆の反応に応じて難易度や説明の仕方を動的に調整することもある．講演時には，資料を執筆した時よりも新しくかつ正確な情報が補足されることもある．

以上の考察から，講演ビデオデータを高度に "視聴" するためのある一つのモデルが成立する．まず，資料を読んで内容の概要を把握し，興味のある箇所を素早く特定する．そして，特定した箇所に関する内容だけを講演ビデオから選択的に視聴し，理解を深める．その結果，講演ビデオを全て見なくても，最小限のコストで必要な情報を取得することが可能になる．

3 システム構成

3.1 概要

本研究で提案するオンデマンド講演システム（Lecture-On-DEMand system: LODEM）の構成を図1に示す．システムで利用するコンテンツは，同一の講演に関する資料テキストとビデオデータである．ユーザが指定した章や節などに基づいて資料テキストから対象範囲を抽出し，それに関連するビデオトラックの頭出しを行い，再生する．また，ユーザが思いついたキーワード，フレーズ，文など資料テキストに書かれていない任意のテキスト情報も入力することができる．すなわち，システムの

オンライン処理は，資料テキストから抽出した一定の範囲やユーザが思いついたキーワードを検索質問（query）として用いて，講演ビデオデータから関連するビデオ内容を検索する処理に相当する．

現在は，検索処理だけをサーバで行い，クライアントPC上のウェブブラウザで資料の閲覧，検索質問の入力，ビデオの視聴を行うことができる．

そのためには，通常の情報検索と同様に，講演ビデオデータに対する索引付けが必要である．具体的には，講演ビデオから音声データを抽出し，発話単位に基づいて講演ビデオを意味のある一定のまとまりに分割する．さらに音声認識によって音声データの書き起こしを生成して講演データベースを作成する．ここでは，分割されたそれぞれの部分を「パッセージ」と呼ぶことにする．

当該データベースは，テキスト情報をキーにして，関連するビデオパッセージを検索することができるように編成されている．現在，索引付けは全て自動で行われる．

本システムの検索処理は，音声文書検索（Spoken Document Retrieval: SDR）[1]の一種である．英語を対象にしたSDRの研究では，音声認識の単語誤り率が30%程度あっても，テキスト検索と同等の検索精度を達成できることが知られている[2]．音声認識の精度は対象とする音声データの品質や講演内容によって変化するものの，現在の音声認識技術で本システムを実用化できる可能性は十分にある．

また，本研究は講演の特徴に着目し，通常の検索システムにはない機能を実現した．講演資料は簡潔に書かれるため，資料中では一つにまとまっている内容が，実際の講演では離散した箇所で説明されることがある．そこで，単一の検索質問に対して，関連する説明を尤度が高い順に複数出力する方が網羅性の点では好ましい．

しかし，同じような説明を何度も出力することは冗長であり，非効率的である．そこで，以前出力した内容との重複を避けながらビデオパッセージを検索する機能を実装した（3.4節参照）．この機能によって，最小限の情報でユーザの情報要求を満足することが可能になる．

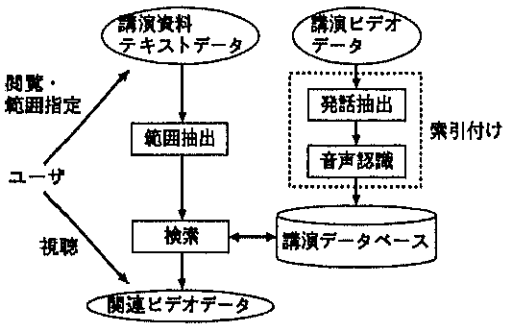

図 1: オンデマンド講演システムLODEMの構成

3.2 講演ビデオデータの索引付け

講演ビデオデータの索引付けは以下の手順からなる．バッチ処理で動作するツール群によって，全ての処理は自動化されている．

1. ビデオデータから音声データを抽出する．
2. 音声データをパッセージに分割する．
3. 音声認識によって各パッセージに対する書き起こしを生成する．
4. 書き起こしに対して，通常のテキスト検索と同様の索引付けを行う．

手順1は，対象とするビデオデータの規格（フォーマット）が分かっていれば技術的な問題はない．

手順2では，書き言葉の段落に相当するような意味的・論理的なまとまりに音声データを分割することが理想的である．しかし，話し言葉に対しては，文を認定することすら難しく，また人間が書き起こしを作成しても，意味のある単位に分割することは難しいのが現状である．

本システムでは音声認識システムを用いて自動的に書き起こしを作成しているため，話題の転換に用いられる接続表現が正しく認識されない可能性がある．また，音声認識用の統計的言語モデルを書き言葉の学習データから作成しているため，話し言葉特有の表現に対する認識率が低い．

以上の問題点を考慮して，現在はパワーなどの物理的な尺度で音声を分解し，無音区間を区切りとして発話を抽出する．さらに，複数の発話をまとめて一つのパッセージを構成する．

手順3では，連続音声認識コンソーシアムのディクテーションソフトウェア[7]で提供されている音声認識エンジン（デコーダ）と音響モデルを利用する．また，新聞記事や論文抄録などから学習した言語モデルを独自に作成して[5]，対象に応じて適宜使い分けている．

手順4では形態素解析システム「茶筌」[1]を用いて書き起こしを単語に分割し，品詞情報に基づいて名詞を索引語として抽出する．また，カタカナ語や新語は未知語と認定されることが多いので，未知語も索引語として抽出する．抽出した索引語を用いて転置ファイルを編成し，テキストによる検索質問を用いた検索を可能にする．

3.3 関連説明の検索

3.2節で説明した処理によって，講演ビデオデータ（書き起こし）を複数のパッセージに分割することができる．そこで，各パッセージを異なる文書と見なせば，情報検

[1] http://chasen.aist-nara.ac.jp/

索の分野で提案された各種の手法を用いて，検索質問に関連するパッセージを効率的に特定できる．

関連度の計算には確率型の手法 [3] を用いた．これは，近年の情報検索手法の中でも比較的高い検索精度を実現することで知られている．具体的には，文書 d の関連度スコアを式 (1) によって計算し，スコアが高い順番に文書を出力する．

$$\sum_t \left(\frac{TF_{t,d}}{\frac{DL_d}{avglen} + TF_{t,d}} \cdot \log \frac{N}{DF_t} \right) \qquad (1)$$

ここで，$TF_{t,d}$ は索引語 t が文書 d に出現する頻度である．DF_t は t を含む文書数であり，N は総文書数である．DL_d は文書 d の長さ（バイト数）であり，$avglen$ は平均文書長である．

検索質問からは，書き起こしパッセージの場合と同じ方法（3.2 節の手順 4）で索引語を抽出する．

3.4 効用最大化に基づく再帰的検索

資料中では一つの章や節としてまとまっている内容が，実際の講演では複数の箇所で分散して説明されることがある．そこで，単一の説明だけを出力するのではなく，3.3 節の式 (1) で計算されるスコアが高い順に複数のパッセージを出力する必要がある．

しかし，ユーザが十分理解したにも拘わらず，同じようなビデオ内容を再生することは効果的ではない．一度検索（再生）した内容は次回以降なるべく検索しないようにすれば，少ない情報によって効用を最大化することが期待できる．これは，ユーザがある文書を閲覧したときに，ユーザが抱える情報要求のうち，まだ満足していない部分を特定し，その部分に対して効果的な別の文書を提示する問題と捉えることができる．

松村ら [6] は，ユーザが文書 d を閲覧しても依然として満たされない情報要求を，検索質問と文書 d との差分ベクトルとして表現し，次回以降の検索に利用する手法を提案した．ここで，検索質問と各文書があらかじめ索引語のベクトルとして表現されていることが前提である．

我々は，この手法を応用し，検索質問ベクトルと一度検索されたパッセージに対応するベクトルとの差分を取りながら再帰的に検索を繰り返すことで，冗長な内容の再生を回避する．また，検索を繰り返すたびに，検索質問ベクトルが縮退するので，式 (1) で計算されるスコアは次第に低下する．これはユーザの要求が次第に満たされていく過程をモデル化している．そこで，スコアに対する閾値を設定し，スコアが閾値を下回った時点でユーザの要求が十分満たされたと判断して処理を終了する．

現在は，スコアの閾値に対して決定的な値や一般的な範囲は分かっていない．テストデータを用いて実験を繰り返しながら，経験的に設定する必要がある．

4 実行例と考察

3 章で説明したシステムを実装し，以下に示す複数のコンテンツを対象に試験運用を行った．

- 教科書が市販され，その内容に基づく講義が放映されているテレビ番組（45 分間）
- 学会でのチュートリアル講演（予稿あり）を再現したビデオ（30 分間）

前者では，CATV から受信したビデオデータを DV に録画し，後者では，DVCAM を用いて聴衆がいない状態でスタジオで撮影した．

構築したシステムは，ウェブブラウザで検索および視聴できるように実装した．検索処理はサーバで実行される．しかし，現状ではビデオデータのサイズが大きくサーバから短時間でダウンロードすることが困難である．そこで，ビデオデータはクライアントの PC 上に保存している．この点は今後改善する必要がある．

システムインタフェースの外観を図 2 に示す．この図では，学会チュートリアル講演を視聴している．

ユーザは画面左の資料を読みながら，画面下の入力ボックスに講演資料からコピーした内容や任意のキーワードを入力することができる．検索を実行すると，検索されたパッセージの書き起こしが複数提示され，ユーザが選択した内容が画面右に再生される．

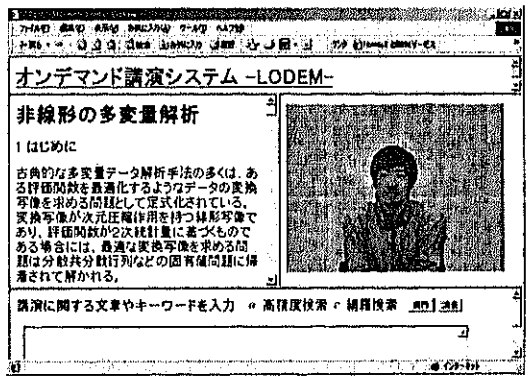

図 2: オンデマンド講演システムのインタフェース

以下，実行例に利用したコンテンツを対象にいくつかの考察を行った．

まず，話し言葉と書き言葉の違いについて調べるために，講演資料テキストと講演発話の文字数を比較した．ここで，発話の文字数は，実際には音声認識によって生成された書き起こしに含まれる文字数である．音声認識には認識誤りがあるものの，文字数に関しては正解とそれほどの差はない点に注意を要する．結果を以下に示す．

	講演資料	書き起こし
テレビ番組	3,489	11,038
学会講演	3,342	6,017

この結果から分かるように，話し言葉は，書き言葉に比べると 2〜3 倍程度に冗長であった．話し言葉と書き言葉の違いは，講演のスタイルなどに依存するため一般化することは困難であるものの，今回対象にしたコンテンツでは，書き言葉固有の表現として以下のようなものが顕著に現れた．

- 話の進行に必要なメタ表現（あらすじなど）
- 具体例（例えば，教科書では「公害事件」と表記されている箇所に対して「水俣病」の例が挙げられている）
- 聞き手に語りかけるような表現（「皆さんご存知のように」など）
- 丁寧な表現（「ですます調」など）

話し言葉と書き言葉の違いに関して，テレビ番組の例を示す．太字が話し言葉のみに現われた表現である．

> 刑法とは，**皆さんご存知のように**どのような行為が犯罪となり，その行為にどのような刑罰が科されるかを定めた法律のことを指します．**六法を開くとどの六法にも刑法と名前のついた法律，すなわち刑法典がのっています．**

DV で録画されたデータの分割は，パワーに基づいて 400ms 程度の無音区間を検出したところで音声データを区切り，それを一つの発話とした．さらに，発話 3 つで一つのパッセージを構成した．その結果，45 分の講義が 34 のパッセージに分割された．ここで，パッセージあたり最長で 582 文字，最短で 96 文字，平均 312 文字が含まれていた．

さらに，テレビ番組の冒頭 10 分間に対して音声認識精度を評価した．論文抄録や新聞記事から作成した言語モデルを使い分けた結果，単語誤り率（Word Error Rate: WER）は 20〜30%という結果が得られ，音声文書検索には十分な認識精度であることが分かった．

今回対象としたテレビ番組は，法律や裁判に関する講義だった．教科書から「人の健康に係る公害犯罪の処罰に関する法律」という一部を抜き出して検索質問とした場合の検索結果（書き起こし）を以下示す．

> また，1960 年代には，熊本の水俣病事件を，はじめとする．公害事件が多発しました．今の安全率で得るのは，公害事件の，民事裁判に関する映像です．このような発光が事件の多発を契機として，1970 年には向上などから，ヒトの健康がする物質を輩出して，ヒト腺眼，身体に実験を調査する行為を処罰する．**人・高にかかる血行が犯罪の処罰に関する法律**で稼いでされています．

当該検索結果において「人の健康に係る公害犯罪の処罰に関する法律」は最後の太字部分である．この部分は正しく音声認識されていないものの，周辺の文脈に現れる語によって正しく検索されている．このように，検索質問や検索対象パッセージをある程度の長さにすることで，音声認識誤りに対しても頑健な検索が可能となった．

5 おわりに

同一の内容に関するテキスト情報とビデオ（音声・画像）情報を併用し，テキスト入力によってビデオ内容を選択的に視聴するオンデマンドシステムを実現した．現在までにテレビの講義番組，学会チュートリアルなどを対象に実験を行った．その他，新聞とニュース番組，レシピのウェブページと料理番組のようなマルチメディアコンテンツの利用が考えられる．現状では画像はユーザが講演内容を理解する助けにはなるものの，処理の対象にはなっていない．今後は，画像解析を応用したビデオパッセージ分割などについて研究を行う予定である．また，評価方法などについても今後検討する必要がある．

参考文献

[1] John S. Garofolo, Ellen M. Voorhees, Vincent M. Stanford, and Karen Sparck Jones. TREC-6 1997 spoken document retrieval track overview and results. In *Proceedings of the 6th Text REtrieval Conference*, pp. 83–91, 1997.

[2] Pierre Jourlin, Sue E. Johnson, Karen Spärck Jones, and Philip C. Woodland. Spoken document representations for probabilistic retrieval. *Speech Communication*, Vol. 32, pp. 21–36, 2000.

[3] S.E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 232–241, 1994.

[4] 伊藤克亘, 藤井敦, 石川徹也. 音声文書検索を用いたオンデマンド講義システム. 情報処理学会研究報告 2001-SLP-39, pp. 165–170, 2001.

[5] 伊藤克亘, 秋葉友良, 藤井敦, 石川徹也. 音声入力型テキスト検索システムのための音声認識. 日本音響学会講演論文集, pp. 193–194, Oct. 2001.

[6] 松村真宏, 大澤幸生, 谷内田正彦. AAS: 文書の組み合わせによってユーザの興味を満足する検索システム. 人工知能学会誌, Vol. 14, No. 6, pp. 1177–1185, 1999.

[7] 鹿野清宏, 伊藤克亘, 河原達也, 武田一哉, 山本幹雄（編）. 音声認識システム. オーム社, 2001.

[8] 藤井敦, 伊藤克亘, 秋葉友良, 石川徹也. 音声言語データの構造化に基づく講演発表の自動要約. ワークショップ「話し言葉の科学と工学」, pp. 173–177, 2001.

音声入力による質問応答システムのための音声認識用言語モデルの検討

秋葉友良[†]　伊藤克亘[†*]　藤井敦[†*]　石川徹也[‡]
[†]産業技術総合研究所　[‡]図書館情報大学　[*]科学技術振興事業団 CREST
e-mail: t-akiba@aist.go.jp

1 はじめに

質問応答 (QA) は、1999 年の TREC-8 にタスクとして採択されて以来、次世代の情報検索技術を目指した評価タスクとして注目されている。従来の情報検索タスクも音声入力に対応するように拡張されてきたが [4]、質問応答では入力が質問文というより話し言葉に近い表現が使用されることから、より音声入力に適したタスクであると考えられる。我々はこのような、音声入力を前提とした質問応答システムを開発中である。

質問応答システムへの入力となる検索者の発話は、質問文という定型的な表現となる一方、QA の検索対象に関する多様な表現が使用される。そのため音声認識部では、これら性質の異なる 2 種の表現を同時に扱う言語モデルが必要となる。前者の定型的な表現には、記述文法を使い人手で特定タスクの言語モデルを作成することが考えられる [1]。データベースなどを対象としたタスクの限定された対話システムなどでは有効な方法であるが、QA タスクでは質問の対象が広範囲にわたるためすべてを人手で用意するのは現実的ではない。

一方、後者の検索対象に関する表現には、QA の知識源となる辞典や新聞記事などを利用して、そこから学習した N-gram 言語モデルを使うのが現実的である。しかし、これらを学習データとして使うだけでは、定型的な質問表現を扱うには不十分である。質問文のための学習データを別途用意する必要があるが、収集のためのコストの面を考えると、これは容易ではない。この問題は、N-gram モデルのタスク適応の問題として捉えられてきており、比較的少数のタスク特有の学習データを使って、元のモデルをタスク用に再訓練する手法が提案されている ([6] など)。

これら 2 つのアプローチの利点を共に生かす方法として、N-gram と記述文法を何らかの方法で統合する方法が考えられる。鹿島ら [5] は、タスクの言語表現をモデル化した記述文法の間を、類似タスクから学習した bi-gram で補間する手法を提案している。

ところで、N-gram モデルは (確率付きの) 有限状態オートマトンと等価であり、正規言語の表現力を持つ。したがって、N-gram モデルの枠組の中で、正規文法[1] (ネットワーク文法) を表現することは可能なはずである。そこで本稿では、汎用の N-gram モデルの中に、ネットワーク文法の持つ二値的制約を埋め込む手法を提案する。本手法により、N-gram による統計的知識と、記述文法による人手による知識を同時に一つのモデルに表すことが可能となり、広範囲の言語表現を N-gram でカバーしつつ、ある特定のタスクに用いる表現を記述文法で重点的に扱う言語モデルを獲得することが可能になる。また、提案法の言語モデルは従来の N-gram 表現形式で記述可能であり、言語モデルを置換えるだけで既存の音声認識デコーダでそのまま利用可能という、既存のシステムとの互換性の面でも優れた特徴を持つ。

[1] 本手法で扱える記述文法は、(N-gram の表現力の制約の故に) 正規文法までである。自然言語の記述として広く用いられている文脈自由文法は、そのまま埋め込むことはできない。しかし、有限長の文は必ず正規言語で表現できること、文脈自由文法を正規文法に近似するアルゴリズムが知られている [2] こと、などから実用上ほとんど問題はない。

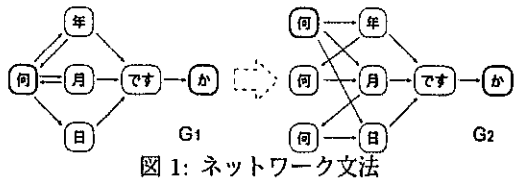

図 1: ネットワーク文法

2 ネットワーク文法の bi-gram モデル表現

単語 bi-gram は、単語を頂点とし、全ての単語間の有向弧に確率が付与された、重み付き (ループのある) 完全有向グラフと見ることができる。この時、有向弧の bi-gram 確率が 0 である場合、その単語連続があり得ないことを表すため、弧が存在しないことと等価である。したがって、単語を頂点として表現した任意のネットワーク文法から、有向弧が存在する場合は 0 でない確率値を、有向弧が存在しない場合は確率 0 を割り当て、単語 bi-gram で表現することが可能となる。もしネットワーク文法中で、ある単語から入出力する有向弧を、文脈に応じて変えたいのであれば、文脈の数だけ同じ単語を表す頂点を複製して表現すればよい。

このような、単語 bi-gram で表現されたネットワーク文法は、例文の集合から簡単に獲得可能である。例えば、年月日を尋ねる発話を表した以下の例文から文法を獲得することを考える。

何/年/です/か　　何/年/何/月/です/か
何/月/何/日/です/か

この 3 文から獲得できる連接可能な単語対は以下の通りである。

A = { (何, 年) (何, 月) (何, 日) (年, 何) (月, 何) (年, です) (月, です) (日, です) (です, か) }

この単語対だけが連接可能であると考えると、ネットワーク文法 (G_1) は 4 つ組 (W_a, W_s, W_f, A) で表現できる。ここで、W_a,W_s,W_f は、それぞれ、全単語集合、開始単語集合、終了単語集合であり、

W_a = { 何 年 月 日 です か },
W_s = { 何 }, W_f = { か }

となる。G_1 のグラフ表現を図 1 左に示す。

この時、ネットワーク文法は、以下の制約を満たす bi-gram として表現できる。

$$P(w_j|w_i) = \begin{cases} p(>0) & \text{if } (w_i, w_j) \in A \\ 0 & \text{otherwise} \end{cases}$$
$$\text{where } \textstyle\sum_{w_j \in W_a} P(w_j|w_i) = 1$$

文法 G_1 は「何年何年ですか」「何月何年ですか」「何年何日ですか」のような、意図されない言語表現までモデル化してしまう。そこで、文法作成者の持つ言語知識を利用して、好ましくない表現を排除し、図 1 右のようなネットワーク文法 G_2 に修正することを考える。新たに導入したノード (文脈) 毎

に、新たな単語記号を導入して、次のような文法 (W'_a, W_s, W_f, A') として表現する。

$W'_a = W_a \cup \{$ 何1 何2$\}$

$A' = \{$(何, 年) (何, 月) (何, 日) (年, 何1) (何1, 月) (月, 何2) (何2, 日) (年, です) (月, です) (日, です) (です, か) $\}$

文法 G_2 は、「何年何月ですか」「何月何日ですか」のような、作成者の意図する表現だけを受理し、それ以外を排除する。このように、ネットワーク文法では、人の持つ言語知識を利用して、N-gram では獲得不可能な、単語間の長距離の依存関係も表現することが可能である。

3 ネットワーク文法と汎用N-gramの統合

前節で述べた方法によって作成したネットワーク文法を、新聞記事などから学習した既存の N-gram モデルと統合する。N-gram がモデル化する単語列の中に、ネットワーク文法によって表される単語列が過不足無く現れる (ネットワーク文法開始単語から終了単語までのパスが存在するような単語列がある) 場合には、その単語列を優先するように、モデルを作成する。このような統合によって、N-gram により汎用の単語列が認識されるのと同時に、タスクに特有の表現を優先して認識することが可能になる。例として、QA タスクの入力には、答えを得るための次のような質問文が想定される。

1976 年に火星に軟着陸した探査機は何という名前でしたか

この入力中、文末の「何という名前でしたか」の部分は質問文に典型的に現れるパターンであり、ネットワーク文法でモデル化する。一方、「1976 年に火星に軟着陸した探査機は」の部分は汎用の N-gram モデルで扱う。

ここで問題となるのは、性質の異なる 2 種のモデルをどのように統合して 1 つのモデルで表現するかということである。ネットワーク文法は、単語連接が二値的で、陽に指定した単語連接以外は許さない。また、ネットワークの開始から終了までの単語列が過不足無く現れている必要がある。一方、N-gram モデルでは、全ての単語が互いに連接可能とし、連接の強弱は確率値で表現するように作成するのが普通である。

既存の N-gram とネットワーク文法を混合して、単純に N-gram モデルを作成すると、全ての単語が連接可能なモデルとなる。これは、N-gram モデルの学習には、スムージングが行われるためである。バックオフ・スムージングでは、高次の N-gram が存在しない場合、低次の N-gram で補間される。bi-gram で二値的制約を表現しても uni-gram で補間されるため、結局すべての単語とネットワーク内単語との間で連接可能となってしまう。一方、全くスムージングを行わないモデルを作成することも出来るが、その場合ネットワーク部の二値的制約は獲得されるが、N-gram 部にゼロ頻度問題が生じ、精度が落ちてしまう。

提案法の基本的アイデアは、統合モデルの部分に応じてスムージング方法を切り替えて学習を行うというものである。二値的な連接制約が必要な部分はバックオフスムージングを行わずに学習を行い、その他の部分はその部分の単語集合だけを元にバックオフスムージングを行う。以下では、既存の bi-gram とネットワーク文法から、統合 bi-gram を作成する場合について説明する。tri-gram の場合も、同様の方法が適用可能である。

3.1 ネットワーク文法の作成

まず、既存 N-gram の語彙を用いてネットワーク文法を作成する[2]。2 節で述べたように、ネットワーク文法は例文から獲得することもできるが、より良いモデルを作成するためには人手によるチューニングを行うのが好ましい。

開始単語と終了単語は、他のネットワーク内単語と区別して、それぞれ開始時、終了時のみ到達可能とするようにネットワークを構成する[3]。すなわちネットワーク内単語集合 W_G は、互いに共通部分のない、開始単語集合 W_{Gs}、終了単語集合 W_{Gf}、中間に現れる単語集合 W_{Gm} から構成されるとする。また、ネットワーク文法の単語は、元の N-gram の同じ単語と区別するために、別の単語記号を割り当てる。本稿では、N-gram 中の単語 w に対して、ネットワーク文法の単語を、記号 "@" を付けて @w と表すことにする。

作成したネットワーク文法を元の N-gram とマージし、統合モデルとする。統合モデルの語彙 W_A は、元の N-gram からの単語 $w \in W_U$ と、ネットワーク文法からの単語 $W_G (= W_{Gs} \cup W_{Gm} \cup W_{Gf})$ から構成される。$(W_U \cap W_G = \phi)$

3.2 単語列頻度の付与

統合モデルに、bi-gram 確率を付与するために、各単語対 (w_i, w_j) の頻度 $C(w_i, w_j)$ と単単語頻度 $C(w_j)$ を与える[4]。頻度 C は、元の bi-gram モデルの単語列頻度 $C_0(w_i, w_j), C_0(w_j)$ を利用して獲得する (図 2)。

N-gram 内単語列の頻度

元の単語列の、bi-gram および uni-gram 頻度をそのまま使う。

$$C(w_i, w_j) = C_0(w_i, w_j) \quad w_i, w_j \in W_U$$
$$C(w_j) = C_0(w_j) \quad w_j \in W_U$$

ネットワーク文法内の単語列頻度

ネットワーク文法の有向弧の存在する単語対 $(@w_i, @w_j) \in A$ に対し、対応する N-gram モデル中の単語対 (w_i, w_j) の頻度を与える。

$$C(@w_i, @w_j) = C_0(w_i, w_j) \quad (@w_i, @w_j) \in A$$

ただし、$C_0(w_i, w_j) = 0$ となるような単語対 $(@w_i, @w_j)$ が存在する可能性があるため、ディスカウンティングを行う必要がある[5]。一方、開始単語 $@w_s$ を除いて単単語頻度を 0 とする。

$$C(@w_i) = 0 \quad @w_i \in W_{Gf} \cup W_{Gm}$$

---

[2] 例えば、日本語ディクテーション基本ソフトウェア [8] の言語モデルは、2 万語と 6 万語のものがあり、特殊な固有名詞を使わない限り、ネットワーク文法を記述するには十分だと考えられる。もし N-gram 語彙に含まれない単語を使用する場合は未知語に対応づけることも可能である。また、N-gram の語彙を利用するのは、評価のために単語の単位を合わせるためでもある。

[3] 開始単語と終了単語のために、他の単語と区別する特別な単語記号を用意すればよい。

[4] 普通、N-gram 確率学習では、高次の単語列頻度と低次の単語列頻度が無矛盾なので、高次の単語列頻度だけを与える。一方、提案法での N-gram 確率学習では、各長さ毎に頻度を用意することに注意されたい。

[5] 利用するディスカウント方法は種々考えられるが、本稿では単純なラプラス法 (全ての頻度に 1 を加える) を用いた。

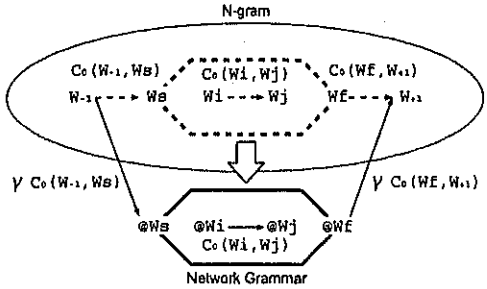

図 2: 統合モデルへの単語列頻度の付与

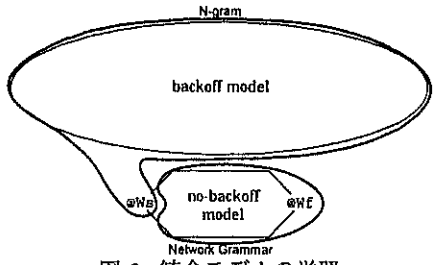

図 3: 統合モデルの学習

N-gram とネットワーク文法を結ぶ単語列頻度

ネットワーク文法の開始単語 $@w_s \in W_{Gs}$ に対し、対応する単語 w_s の頻度を用いて、次のような頻度を与える。

$$C(w_i, @w_s) = \gamma C_0(w_i, w_s)$$
$$C(@w_s) = \gamma C_0(w_s)$$

γ は、ネットワーク文法の表す単語列の、対応する N-gram モデルでの単語列に対する相対的な優先度を表す。1 以上の値を与え、大きな値を与えるほど、ネットワーク文法の表現が優先的に考慮される。後で述べるように、$\gamma = 1$ としても、ネットワーク文法の表現は優先的に扱われるようになる。

ネットワーク文法の終了単語 $@w_f \in W_{Gf}$ から N-gram への単語列頻度も、同様に与える。ただし、$@w_f$ の単単語頻度は 0 とする。

$$C(@w_f w_j) = \gamma C_0(w_f w_j)$$
$$C(@w_f) = 0$$

3.3 モデルの学習

与えた頻度からモデルを学習する[6]。その際、予測する単語によってスムージング方法を切り替えて計算する (図 3)。一般に、バックオフスムージングが行われた bi-gram は、次の式で表される。

$$P(w_j|w_i) = \begin{cases} d_{C(w_i,w_j)} P_{ML}(w_j|w_i) & C(w_i,w_j) > 0 \\ \alpha(w_i) P(w_j) & C(w_i,w_j) = 0 \end{cases}$$

ここで、$d_{C(w_i,w_j)}$、$\alpha(w_i)$、$P_{ML}(w_j|w_i)$ は、それぞれディスカウント係数、uni-gram への配分を正規化する関数、最尤推定による確率を表す。

まず、統合モデルの語彙 W_A のうち、開始単語を除くネットワーク文法中の単語 $@w_j \in W_{Gm} \cup W_{Gf}$ を予測する確率 $P(@w_j|@w_i)$ を計算する。この時、コンテキストを構成する単語は $@w_i \in W_{Gs} \cup W_{Gm}$ である。このモデルは、バックオフを行なわないで求める。すなわち、bi-gram 頻度 $C(@w_i, @w_j)$ が存在するものに対してのみ bi-gram 確率を割り当て、$\alpha(@w_i) = 0$ とする。$W_{Gm} \cup W_{Gf}$ に再配分される uni-gram 確率は 0 となる。

次に、N-gram 内の単語とネットワーク開始単語 $w_j \in W_U \cup W_{Gs}$ を予測する確率 $P(w_j|w_i)$ を計算する。コンテキストを構成する単語は、$w_i \in W_U \cup W_{Gf}$ である。このモデルは、普通にバックオフスムージングを行なって求める。その際、$W_{Gm} \cup W_{Gf}$ にはすでに uni-gram 確率 0 が割り当てられていることに注意し、残りの単語集合 $W_U \cup W_{Gs}$ だけを使って再配分を行なう。

3.4 モデルの性質

以上の方法で作成した統合言語モデルは、例えば ARPA 形式などの、従来の N-gram モデルと同じフォーマットで表現できる。そのため、N-gram を言語モデルとして利用する既存の音声認識デコーダでそのまま利用できる。そして以下に示すように、N-gram とネットワーク文法の両方の性質を併せ持ったモデルとして利用できる。

・N-gram からネットワーク文法内部の単語を予測する確率は必ず 0 となる。 開始単語を除く Network 内単語 $@w_j \in W_{Gm} \cup W_{Gf}$ の予測モデルは、bi-gram のみが用いられ、uni-gram は 0 となる。Network 内単語で、かつ bi-gram 確率の存在する (有向弧の存在する) 単語からのみ予測可能となる。N-gram 内単語 $w_i \in W_U$ からの N-gram 確率は、

$$P(@w_j|w_i) = \alpha(w_i) P(@w_j) = \alpha(w_i) \cdot 0 = 0$$

となり、w_i から $@w_j$ への遷移は生じない。

・ネットワーク文法内部から N-gram 単語を予測する確率は必ず 0 となる。 終了単語を除くネットワーク内単語 $@w_i \in W_{Gs} \cup W_{Gm}$ に対し、$\alpha(@w_i) = 0$ となる。ネットワーク内単語 $@w_i$ から N-gram 内単語 w_j への頻度 $C(@w_i, w_j)$ は必ず 0 なので、

$$P(w_j|@w_i) = \alpha(@w_i) P(w_j) = 0 \cdot P(w_j) = 0$$

したがって、$@w_i$ から w_j への遷移は生じない。

・ネットワーク文法の開始単語/終了単語では N-gram との連接が可能。 N-gram 内単語からネットワーク開始単語、およびネットワーク終了単語から N-gram 内単語は、通常の N-gram モデルと同様のスムージングされた確率値が割り当てられる。したがって、全ての N-gram 内単語からネットワーク文法開始単語への遷移、文法終了単語から全ての N-gram 内単語への遷移が可能である。

・N-gram 内の単語列だけから成る文に割り当てられる確率値の順序関係は保存される。 N-gram 内単語列に与える頻度は、元の N-gram モデル学習用の頻度と同じである。

・同じ単語列では、ネットワーク文法を通る単語列が優先される。 ネットワーク文法内単語列を含む文には、それに対応する (同じ音素列を持つ) N-gram 内単語だけで構成された文が必ず存在する。既存の認識デコーダは、文の確率をパスの最大確率で近似する方法 (ビタビ・アルゴリズム) が普通であるので、認識時には両者の確率値を比較し、高い方が採用されることになる。両者の bi-gram 確率は同じ頻度か

[6] 提案法では、bi-gram 頻度と uni-gram 頻度を個別に与えており、その間に整合性がないので、確率計算において uni-gram 頻度 $C(w_j)$ と bi-gram のコンテキスト頻度 $C_c(w_i) = \sum_{w_j} C(w_i, w_j)$ を区別して用いる。例えば、P_{ML} は、bi-gram 頻度とそのコンテキスト頻度を用いて、$P_{ML}(w_j|w_i) = C(w_i, w_j)/C_c(w_i)$ と計算する。

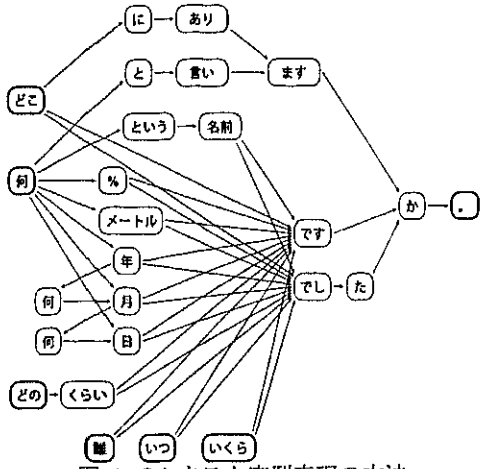

図 4: QA タスク定型表現の文法

ら学習されているが、ネットワーク文法内単語列は、連接しない単語 (特にネットワーク外の単語) を予測するための確率配分がないこと、バックオフを行わないで学習したモデルであること、から相対的に高い確率値が割り当てられることになる。また、ネットワーク文法開始単語への単語列頻度を γ で調節することで、N-gram に対する優先性をコントロールすることができる。

4 実験

QA タスクの質問文を想定した定型表現を受理可能なネットワーク文法 (図 4) を作成し、新聞記事 111 か月分から学習した 2 万語 bi-gram および tri-gram と統合、ネットワーク文法統合モデル (net) を作成した。今回は、既存の N-gram モデル作成ツール [3] を用いて、近似的に作成した[7]。γ は 2 とした。また比較のため、新聞記事のみから学習した N-gram モデル (base) を作成した。スムージング手法は、共に Witten-Bell 法を用いた。

評価データには、新聞記事 100 文 (NP) と QA タスク用質問文 50 文 [7](QA) を、男性 2 人女性 2 人によって読み上げた音声データを用いた。作成したネットワーク文法は、29 単語と比較的小規模のものであるが、質問文のうち 72%の 36 文 (QA') が、この文法のモデル化する表現を含んでいた。

デコーダには大語彙音声認識デコーダ julius[8] のバージョン 3.2 を使用し、音響モデルには 2000 状態 16 混合性別非依存 triphone を、言語モデル重みは新聞記事 N-gram での最適値を用いた。探索アルゴリズムの変更は行っていない[8]。

[7] 以下の手順で作成した。(1)palmkit[3] の idngram2lm を修正。頻度読み込み部を修正して、N-gram 頻度を長さ N 毎に別々に読み込めるように変更。同時に、各頻度のコンテキスト頻度を求め、確率計算に使用するように変更。(2) 単語集合を $W_U \cup W_{Gs} \cup W_{Gf}$ と $W_{Gs} \cup W_{Gm} \cup W_{Gf}$ に分割して、前者からスムージングモデルの ARPA 形式を、後者からスムージングしないモデルの ARPA 形式を作成。(3) 結合部の確率値に注意しながら両モデルをマージ。(4) 言語モデル評価ツール (evallm) を用いて、正しいモデルとなっていることを確認。

[8] ネットワーク文法内部の単語と N-gram モデル内の単語の間での N-gram 確率は必ず 0 となり、この間の単語間遷移は起こり得ない。この単語間遷移の探索を抑制するようにデコーダを修正することで、より効率的な探索を行うことができるようになると考えられる。

表 1: 実験結果

n-gram	評価データ	言語モデル	COR	ACC
2	NP (100)	base	81.9	79.4
		net	81.9	79.4
	QA (50)	base	72.9	69.4
		net	74.2	70.4
	QA'(36)	base	71.3	67.2
		net	73.8	69.6
3	NP (100)	base	90.4	87.8
		net	90.0	87.8
	QA (50)	base	85.0	80.3
		net	85.1	80.5
	QA'(36)	base	84.4	79.9
		net	85.2	81.0

COR=単語正解率 (%), ACC=単語正解精度 (%)

実験結果を表 1 に示す。新聞記事の認識精度を下げること無く、質問文の精度が向上することを確認した。bi-gram にくらべ、tri-gram の場合にはあまり効果が現れなかった。既存の実装方法との整合性に問題が生じている可能性も考えられ、今後調査が必要である。また、今回の提案法の tri-gram との統合モデルでは、ネットワーク文法部は bi-gram のままであったが、tri-gram を用いることも理論上可能であり、今後検討していきたい。

5 まとめ

音声入力による質問応答システムのために、定型的な質問文表現と多様な検索対象に関する言語表現を、同時にモデル化した言語モデルを提案した。提案法は、N-gram 言語モデル表現を利用しつつ、その内部に記述文法で表現した二値的な制約を統合する手法である。N-gram では不可能であった単語間の長距離の依存関係も、記述文法を使って表現可能である。提案法で作成した言語モデルは、従来の N-gram モデルと互換性があり、既存の大語彙音声認識デコーダでそのまま利用できるという特徴を持つ。提案法により、新聞記事から作成した言語モデルと、質問文を表現したネットワーク文法を統合した言語モデルを作成し、既存の音声認識デコーダで認識実験を行った。元の汎用 N-gram 言語モデルが対象とする新聞記事の読み上げに対する認識精度を下げることなく、記述文法で扱うタスク特有の質問文に対し認識精度が向上することを確認した。

参考文献

[1] T. Akiba and K. Itou. Semi-Automatic Language Model Acquisition without Large Corpora. In Proc. of ICSLP-2000, vol.4, pp.49-52, 2000.

[2] F.C.N.Pereira and R.R.Wright. Finite-state approximation of phrase-structure grammars. In Proc. of ACL 1991, pp.246-255, 1991.

[3] 伊藤彰則, 好田正紀. 単語およびクラス n-gram 作成のためのツールキット. 信学技報, SP2000-106, pp.67-72, 2000.

[4] 伊藤克亘, 秋葉友良, 藤井敦, 石川徹也. 音声入力型テキスト検索システムのための音声認識. 日本音響学会講演論文集, pp.193-194, Oct. 2001.

[5] 鹿島博晶, 河原達也. 複合的言語制約に基づくキーフレーズスポッティングによる対話音声理解. 信学技報, SP2000-106, pp.115-120, 2000.

[6] 駒谷和範, 河原達也, 浦田陽司, 黒橋禎夫, Pascale Fung. 柔軟な言語モデルとマッチングを用いた音声によるレストラン探索システム. 信学技報, SP2001-113, pp.67-72, 2001.

[7] 佐々木 裕, 磯崎 秀樹, 平博順, 廣田 啓一, 賀沢秀人, 平尾努, 中島浩之, 加藤恒昭. 質問応答システムの比較と評価. 信学技報, NLC2000-24, pp.17-24, 2000.

[8] 鹿野清宏, 伊藤克亘, 河原達也, 武田一哉, 山本幹雄 (編). 音声認識システム. オーム社, 2001.

ユーザ発話中の未知語を自動補完する音声入力型検索システム

藤井 敦 [†,†††]　伊藤克亘 [††,†††]　石川徹也 [†]
[†] 図書館情報大学
[††] 産業技術総合研究所
[†††] 科学技術振興事業団 CREST
fujii@ulis.ac.jp

1 はじめに

近年の音声認識技術は，ある程度内容が整理されている発話に対しては実用的な認識精度を達成できるようになっており，様々な応用が考えられる．情報検索の分野では音声認識を採り入れた研究も数多く行われている．これらの研究は目的に応じて「音声データの検索」と「音声による検索」の2つに大別される．前者は，TRECのSpoken Document Retrieval (SDR) トラック [4] で放送音声データを対象にしたテストコレクションが整備されていることを背景にして盛んに研究が行われ，既に実用レベルに達している [7]．

それに対して，音声による検索はカーナビゲーションシステムやコールセンターのようにキーボード入力を前提としないアプリケーションを支える重要な基盤技術であるにも拘らず，音声データ検索に比べて研究事例は少ない．また，従来の研究では既存の音声認識とテキスト検索システムが単純に接続されているだけであり，音声認識誤りによって検索精度が顕著に低下する [1, 2]．これに対して，筆者らは検索対象のコレクションを用いて音声認識用の言語モデルを作成し，音声認識と検索精度の両方を向上させる手法を提案した [3, 5]．

しかし，音声入力型の検索システムでは，未知語（システム辞書未登録語）の問題がある．近年の情報検索システムは，古典的な統制語彙型システムとは異なり，検索対象テキスト中の任意の語による検索を可能とする．索引のサイズが数100万のオーダーに達することは珍しくない．機能語などは不要語として索引から除外されるものの，これらが検索キーワードとして利用されることは稀であるため，事実上，語彙制限はないと考えてよい．

他方において，近年の音声認識システムでは語彙サイズ（辞書登録語数）が制限される．これはハードウェアに関する制約や統計モデルの学習効率が主な原因であるため [13]，登録語数を増やすという単純な方法では解決が困難である．多くの言語において，語彙サイズは高々数万語に制限されており [6, 9, 11]，実用的な検索システムの索引サイズに比べると極端に小さい．

また，統計的な音声認識では，機能語などの高頻出語ほど高い精度で認識されるのに対して，情報検索では特定の文書にしか出現しない低頻度語ほど効果的な索引語になりやすい．すなわち，ユーザ発話中の効果的な検索キーワードほど誤認識されやすいという矛盾が生じる．

以上まとめると，音声入力型の検索システムにおいて「未知語問題」は本質的に不可避であり，何らかの積極的な解決策が必要である．本研究では，音声認識でカバーできない単語を検索用の索引語によって自動的に補完する手法を提案する．また，評価実験によって実装したシステムの有効性を示す．

2 システム概要

本研究で提案する音声入力型テキスト検索システムの構成を図1に示す．本システムは，音声認識，テキスト検索，未知語補完の3つのモジュールで構成されている．現在は日本語を対象に実装されているものの，本研究で提案する手法は言語の種類を問わない．以下，図1に基づいて本システムの処理について説明する．

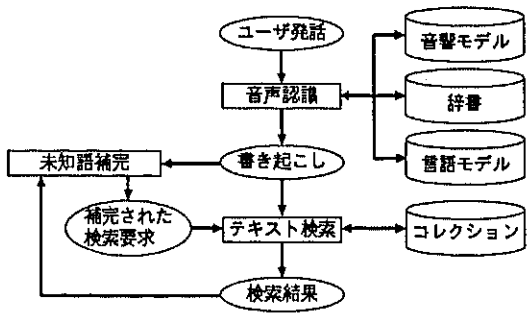

図 1: 音声入力型検索システムの構成

まず，ユーザが検索要求を発話すると，音声認識部が辞書，音響モデル，言語モデルを用いてユーザ発話の書き起こしを生成する．本システムでは，日本語ディクテーションツールキット [15] で提供されている音声認識エンジン（デコーダ）と音響モデルを利用した．しかし，ユーザ発話中に含まれる未知語を検出するために，辞書と言語モデルは独自に作成して利用した [14]．

具体的には，毎日新聞CD-ROM 10年分（1991–2000）の記事を「茶筌」[1]で形態素解析し，高頻度語20,000語

[1] http://chasen.aist-nara.ac.jp/

を抽出して辞書を構成した．通常は，辞書中の単語 N グラムなどによって言語モデルを作成する．しかし，これでは辞書未登録語は認識できない．そこで，辞書に登録されなかった約 30 万語（異なり数）を音節単位に分割し，単語と音節を併用してトライグラムを作成した．音節は異なりで 700 件あった．

すなわち，本システムの言語モデルにおいて，辞書未登録語は音節の組合せとしてモデル化されている．その結果，辞書未登録語は単語としては認識されないものの，音節単位でカタカナ列として書き起こしされる．また，当該言語モデルは通常の統計的 N グラムなので，既存のデコーダを拡張せずに利用できる．そこで，音韻系列の認識を別途必要とする手法 [12] とは異なる．

「オレンジやグレープフルーツなどの柑橘系果物の輸入に関する記事」という発話を例にとると，

> オレンジや /グ レープ ラチナガ ノ/ などの
> /カン キ ツ ケ イ/ 果物の輸入に関する記事

のように「グレープフルーツ」や「柑橘系」が未知語として検出される（ここでは未知語部分をスラッシュで括っている）．なお「柑橘系」のように未知語箇所の検出と音韻列の特定に成功する場合や「グレープフルーツ」のように音韻列の特定は不完全でも未知語箇所の検出に成功する場合がある．いずれの場合も，未知語に対する正しい語を推定することが出来れば，音声認識精度が向上し，結果として検索精度も向上する．

本システムのユーザは，検索対象のテキストコレクションから何らかの情報を引き出したいという意図を持って発話を行う．言い替えれば，ユーザの発話はコレクション中の情報に関連したものである可能性が高い．そこで，上記「グレープラチナガノ」や「カンキツケイ」に対応する正しい語がコレクション中に含まれていると考えることは自然な発想である．

直観的には，検索対象コレクションの索引語から，検出された未知語と音韻的に等価な語もしくは類似する語を探索してユーザ発話中の未知語を補完すればよい．しかし，音韻的に「類似する」語の探索（すなわち，音韻列の部分一致による探索）を大規模な索引に対して行うことは効率が悪く，実時間処理には耐えない．

そこで，まず，ユーザ発話中で単語として認識された部分だけを用いて初期検索を実行し，ユーザの検索要求に関連する文書を選択的に取得する．テキスト検索には確率型の「Okapi 法」[10] を用いた．当該手法は，与えられた検索要求に対するスコアを各文書に対して計算し，スコアが高い順番に文書を出力する．本システムでは，対象テキストを「茶筌」で形態素解析して名詞を索引語として抽出し，単語単位で索引付けを行って転置ファイルを事前に作成する．

次に，初期検索で得られた文書から，検出された未知語に対応する語を探索し，未知語と置き換えることで検索要求を補完する．具体的な方法については 3 章で説明する．最後に，補完された検索要求を用いて再検索を行い，最終的な検索結果が得られる．

上記の手法は，初期検索の結果を用いて最終的な検索精度を向上させるという点において，情報検索で用いられる検索要求の拡張（query expansion）やローカルフィードバックに類似している [8]．しかし，これらは検索精度を向上させることに主眼が置かれ，ユーザが意図しない索引語を追加する可能性がある．それに対して本手法は，ユーザの発話を正しく認識することを目的としている点が異なる．これは「自分が発話（意図）した通りに検索が行われている」という安心感をユーザに与える上で重要である（残念ながら，このような観点は情報検索の研究ではあまり考慮されていない）．

3 未知語の自動補完

3.1 方法論

本システムの特長は，音声認識で検出された未知語の音韻系列を，初期検索で取得された上位文書中の索引語に対応付けることによって単語として正しく認識する点にある．この処理を「未知語の補完」と呼ぶことにする．

同音意義語のために，一つの音韻系列が複数の単語に対応することがある（例えば「河川」と「架線」）．また，未知語の音韻系列は誤って検出されることがあるため，補完対象の音韻系列一つに対して，音韻的に類似する複数の索引語を考慮する必要がある．すなわち，未知語の自動補完では，複数の候補から適切な索引語を選択するための曖昧性解消が必要である．

そこで，選択されるべき索引語が満たす条件について検討し，以下に示す 3 つの基準を設定した．

- 補完対象の未知語との音韻的な類似度が高い（完全一致すれば類似度は最大となる）．
- 上位文書における出現頻度が高い．
- より上位の文書に出現する．

これらを確率論的な枠組で定式化すると，未知語補完は，式 (1) で計算されるスコアを最大化する t を選択することに相当する．

$$\sum_{d \in D_q} P(w|t) \cdot P(t|d) \cdot P(d|q) \qquad (1)$$

ここで，D_q は検索要求 q によって初期検索された上位文書の集合である．$P(w|t)$ は t が音韻的に w と等価である確率，$P(t|d)$ は上位文書の一つ d から索引語を無作為に選んだ場合に，それが t である確率，$P(d|q)$ は検索要求 q によって文書 d が検索される確率である．これらのパラメタは，上記 3 つの基準にそれぞれ対応している．

しかし，実際には $P(w|t)$ や $P(d|q)$ の確率値を正確に推定することは難しい．また，音韻的な類似度（上記，第1の基準）が他の基準よりもかなり強い制約になることが経験的に分かっている．そこで，予備実験の結果に基づいて，式 (1) を式 (2) のように近似する．

$$\sum_{d \in D_q} P(w|t) \cdot \log(P(t|d) \cdot P(d|q)) \qquad (2)$$

ここで，$P(w|t)$ は t と w が共有する音韻数と w に含まれる音韻総数の比率によって計算する．具体的には，DP マッチングによって t と w を音韻単位で比較し，両者に共通して含まれる音韻列を特定する．$P(t|d)$ は d における t の相対頻度で計算する．$P(d|q)$ として Okapi 法で計算される文書 d のスコアで代用する．また，$P(t|d)$ と $P(d|q)$ の log を用いることで，これら2つの影響力が相対的に小さくなるように制御している．

以上の方法は，索引付けの手法に依存しない点に注意が必要である．言い替えれば，索引語 t の単位として，文字，単語，複合語など文書中に現れる任意の文字列を対象とすることができる．

3.2 実装

初期検索によって文書数を制限しても，索引語数は膨大なものになる場合がある．特に，DP マッチングによる音韻単位の比較は実時間応答を低下させる要因となる．また，上位文書中の索引語の多くは，補完対象の未知語と音韻的に全く類似しないため，これらのノイズを早期に排除できれば，計算効率の向上が期待できる．

通常のテキスト検索に用いられる索引（本システムでは転置ファイル）は，入力されたキーワードとの完全一致によって，該当する項目を効率良く検索できる．しかし，未知語補完用の索引では，入力された音韻列に対して，部分一致を許容しながら，ある程度類似した項目だけを効率良く特定できなければならない．

本システムで用いる未知語検出の傾向を調査した結果，検出された未知語と，それに対応する正しい索引語は，前方もしくは後方で一致していることが多く，両端が一致せずに語中のみが一致することは少ない．そこで，未知語補完用の索引を以下の手順で事前に作成した．

まず，コレクション中の全文書を「茶筌」で形態素解析し，単語表記とカナ表記を抽出する．次に，カナ表記を規則によって音韻系列に変換する（規則数 143）．最後に，音韻系列の前方と後方から任意長の部分列を抽出して，前方/ 後方部分一致探索が可能な索引を編成する．

このとき，単語一つと単語バイグラムを併用して索引を作成することで「弥生/時代」や「オゾン/ホール」のように2単語で構成される複合語にも対応した．

原理的には，3単語以上で構成される長い複合語も扱うことができる．しかし，未知語の長さに比例して探索時間がかかるため，現在は2単語までとしている．また，現状の音声認識では機能語のような高頻度語は既知語として正しく認識されやすいため，長い単語列（例えば「情報検索の応用分野」）が一つにまとまった未知語として検出されることは稀である．

4 評価実験

本研究で実装したシステムを評価するために，IREX の日本語検索コレクション[2]を用いて実験を行った．当コレクションは，毎日新聞 1994–1995 年（記事総数 211,853 件）を対象にした検索課題 30 件と各課題に対する正解記事 ID で構成されている．検索課題の例を以下に示す．

```
<TOPIC><TOPIC-ID>1010</TOPIC-ID>
<DESCRIPTION>柑橘類の輸入</DESCRIPTION>
<NARRATIVE>オレンジ、レモン、グレープフルーツなどの柑橘系果物の日本への輸入の記事。政府の市場解放や輸入による日本生産地の影響、値段への影響や消費者の反応などの記事を含む。</NARRATIVE></TOPIC>
```

さらに，4名の話者（男女各2名）に<NARRATIVE>フィールドを読み上げてもらい，合計 120 件の音声発話データを作成して実験に利用した．初期検索，再検索ともに上位 300 件を出力した．

まず，未知語の検出と補完に関する評価を行った．30 件の検索要求（<NARRATIVE>のみ）に含まれる単語は，のべ数で約 400 語あり，14 単語（異なりで 13 単語）が音声認識用辞書に登録されていなかった．

未知語検出の再現率と精度はそれぞれ 71.4%と 22.6%であった．本システムは未知語を網羅的に特定する傾向があることが分かる．さらに，未知語の補完精度を調べた結果，36.2%であった．ここでは，辞書登録語が未知語として誤検出されても，補完処理によって正しい索引語に対応付けられた場合は正解と判定した．正しく補完された未知語と索引語の例を以下に示す．

グレープラチナガノ / グレープフルーツ
ヤヨイチタ / 弥生時代
ニククライス / ニックプライス
ベンピ / 便秘

次に，検索精度への影響を調べるために，以下の異なる検索手法（システム）を比較した．

1. テキスト入力型検索システム
2. 高頻度語 20,000 語のみを含む言語モデルを音声認識に使用した音声入力型検索システム
3. 本システム（検出した未知語は補完しない）
4. 本システム（未知語の検出・補完を併用）

[2] http://cs.nyu.edu/cs/projects/proteus/irex/

システム4が本研究で提案するシステムに相当する．システム2は未知語音節をモデル化していないため，未知語の検出と補完を行わない点を除けば，本システムと同じである．各システムの平均適合率（%）を以下に示す．

話者＼システム	1	2	3	4
男性#1	–	30.1	25.8	31.8
男性#2	–	27.9	26.8	29.6
女性#1	–	28.3	28.3	32.0
女性#2	–	27.5	24.4	28.5
平均	35.0	28.4	26.3	30.4

本システムの精度はテキスト検索には及ばないものの，約87%を再現している．また，全ての話者に対して，それ以外の音声入力型システム（2と3）の検索精度を向上させた．システム3と4を比較することで未知語補完の効果が分かり，システム2と4を比較することで，未知語の検出と補完を併用した提案手法の有効性が分かる．

しかし，精度の向上はそれほど大きくなかった．今回の実験では未知語が本質的に少なかったため，全体的な差異が大きくならなかった．また，未知語を人工的に作るような不自然な実験設定は避けた．未知語問題がより深刻な対象（例えば，技術文書やウェブページ）について，今後さらなる評価実験を行う予定である．

システム2に比べて，本システムの精度が顕著に低下した課題を分析した結果，初期検索の上位文書に正しい索引語が含まれているにも拘らず，式(2)のスコアで適切に選択されなかった事例が大半を占めた．例えば「制度」が未知語検出によって「センド」と誤認識されたために「鮮度」のように音韻的に等価な別の語が選択されてしまった．文書中の索引語頻度や文書順位などとのバランスについて今後検討が必要である．また，未知語音節をモデル化したために，辞書登録語を誤認識し，検索精度が低下した事例が若干あった．

最後に，オンライン処理のCPU時間を測定した．未知語の検出は通常の統計的音声認識の枠組内で行われるため，それに伴う付加的なCPU時間は発生しない．補完に要したCPU時間は未知語あたり平均3.5秒だった（AMD Athlon MP 1900+）．依然として改善の余地はあるものの，ほぼ実時間で動作すると考えてよい．

5 おわりに

音声入力型の検索システムでは，音声認識と検索における語彙サイズの不整合は不可避である．本研究は単語と音節を併用した言語モデルによってユーザ発話中の未知語を検出し，検索対象文書中の索引語によって適切に補完する手法を提案した．新聞記事を対象にした実験の結果，本手法は実時間で動作し，既存の手法を上回る検索精度を実現することができた．論文のような技術文書やウェブページの検索では，未知語問題がより深刻になる．今後は，これらを対象に研究を行う予定である．

参考文献

[1] J. Barnett, S. Anderson, J. Broglio, M. Singh, R. Hudson, and S. W. Kuo. Experiments in spoken queries for document retrieval. In *Proceedings of Eurospeech97*, pp. 1323–1326, 1997.

[2] Fabio Crestani. Word recognition errors and relevance feedback in spoken query processing. In *Proceedings of the Fourth International Conference on Flexible Query Answering Systems*, pp. 267–281, 2000.

[3] Atsushi Fujii, Katunobu Itou, and Tetsuya Ishikawa. Speech-driven text retrieval: Using target IR collections for statistical language model adaptation in speech recognition. In *ACM SIGIR'01 Workshop on Information Retrieval Techniques for Speech Applications*, 2001.

[4] John S. Garofolo, Ellen M. Voorhees, Vincent M. Stanford, and Karen Sparck Jones. TREC-6 1997 spoken document retrieval track overview and results. In *Proceedings of the 6th Text REtrieval Conference*, pp. 83–91, 1997.

[5] Katunobu Itou, Atsushi Fujii, and Tetsuya Ishikawa. Language modeling for multi-domain speech-driven text retrieval. In *IEEE Automatic Speech Recognition and Understanding Workshop*, 2001.

[6] Katunobu Itou, Mikio Yamamoto, Kazuya Takeda, Toshiyuki Takezawa, Tatsuo Matsuoka, Tetsunori Kobayashi, and Kiyohiro Shikano. JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research. *Journal of Acoustic Society of Japan*, Vol. 20, No. 3, pp. 199–206, 1999.

[7] Pierre Jourlin, Sue E. Johnson, Karen Spärck Jones, and Philip C. Woodland. Spoken document representations for probabilistic retrieval. *Speech Communication*, Vol. 32, pp. 21–36, 2000.

[8] K.L. Kwok and M. Chan. Improving two-stage ad-hoc retrieval for short queries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 250–256, 1998.

[9] Douglas B. Paul and Janet M. Baker. The design for the Wall Street Journal-based CSR corpus. In *Proceedings of DARPA Speech & Natural Language Workshop*, pp. 357–362, 1992.

[10] S.E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 232–241, 1994.

[11] Herman J. M. Steeneken and David A. van Leeuwen. Multilingual assessment of speaker independent large vocabulary speech-recognition systems: The SQALE-project. In *Proceedings of Eurospeech95*, pp. 1271–1274, 1995.

[12] Martin Wechsler, Eugen Munteanu, and Peter Schäuble. New techniques for open-vocabulary spoken document retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 20–27, 1998.

[13] Steve Young. A review of large-vocabulary continuous-speech recoguition. *IEEE Signal Processing Magazine*, pp. 45–57, September 1996.

[14] 伊藤克亘, 田中和世. 被覆率を重視した大語彙連続音声認識用統計的言語モデル. 日本音響学会講演論文集, pp. 65–66, March 1999.

[15] 鹿野清宏, 伊藤克亘, 河原達也, 武田一哉, 山本幹雄（編）. 音声認識システム. オーム社, 2001.

未知語検出を用いた語彙統制のない音声検索*

○伊藤克亘 †(産総研) △藤井敦 † △石川徹也 (図書館情報大) (†JST CREST)

1 はじめに

大語彙連続音声認識技術は、ここ数年広く利用されるようになってきた。しかし、現在の技術では、まだ様々な限界が残されている。例えば、言語モデルに関しては、確率モデルの構築やメモリなどハードウェアの制約から、システム語彙の大きさを数万程度に制限せざるを得ない。(本稿ではこのような構成の音声認識システムを語彙統制型と呼ぶ。) 語彙統制型システムでは、未知語の問題を避けられない。

音声認識を何らかのシステムの一部として利用した場合に、未知語がシステム全体にどのような影響を及ぼすかは、応用によって異なる。したがって、応用によっては、未知語処理の有無が重要になる場合もあると考えられる。

本稿では、音声入力型のテキスト検索システムに、音声認識部の未知語処理を導入することで、音声認識システムの語彙統制によってテキスト検索に悪影響を及ぼさない手法を提案する。

2 語彙統制が音声入力型テキスト検索に与える悪影響

全文検索型のテキスト検索においては、全ての語を検索用に索引付けするわけではなく、検索要求として入力されやすい語や効率的な検索を可能にする語のみを索引語として利用する。日本語のテキスト検索においては、索引語は名詞などの内容語に限定されることが多い [1] [1]。また、関連性のスコアを計算するときには、索引語によってその重要度は異なり、一般的な (出現頻度の非常に高い) 語よりは、ある話題に特有の語の方が重要度が高くなる。つまり、この応用においては、助詞の認識に関する正確さはさほど重要ではなく、自立語の名詞の部分が正しく認識されることが重要となる。

この全文テキスト検索の入力に、制限語彙型の音声認識システムを利用した場合について考えてみる。制限語彙型の音声認識システムは、高頻度語を中心にモデル化される。したがって、助詞などの機能語や非常によく出現する自立語の部分のモデル化の精度は高いが、一般的には余り出現しない専門用語などは、未知語になってしまう可能性もある。

こういった未知語の問題は、学習データを増やし、辞書の登録語数も多くするなどの方法で、ある程度回避することができる。しかし、重要な索引語が未知語となる可能性は残り、そのことにより致命的な誤りを起こすこともありえる。

したがって、テキスト検索への入力としての音声認識の場合には、上記のような間接的な対策ではなく、未知語を適切に扱う手法が必要となる。

テキスト検索の場合、重要な索引語が未知語となる可能性もあるため、音声認識の段階で、未知語を検出するだけでなく、その未知語がどのような語であるか推定できることが望ましい。

本研究では、われわれが既に提案している未知語処理の手法 [2] を用いて未知語を検出し、音節系列を推定する。その結果とテキスト検索を用いて未知語を推定し、索引語として用いることによって、語彙統制のない音声入力型の検索を実現する。

3 サブワード併用型単語 N-gram

本稿で提案する手法で用いる言語モデルを、サブワード併用型単語 N-gram とよぶ。このモデルでは、高頻度語による語彙統制型単語 N-gram ではシステム語彙とならない低頻度の語を、サブワード (音節など) に分解する。システム語彙にサブワードのエントリも加えることにより、語彙の制限を取り払う。このようにモデル化することにより、システム語彙の部分は単語 (形態素) の系列が得られ、未知語の部分は音節の系列が得られる。

この手法では、全体の認識率で比較した場合には、未知語処理を全くおこなわない場合とほぼ同等であった。しかし、未知語処理をしない場合と比較すると、未知語である内容語の認識に成功する一方で、未知語周辺の助詞の認識に失敗するという傾向にあった。この特徴は、前節の議論をふまえると全文テキスト検索と親和性が高いと考えられる。

本研究では、毎日新聞 10 年分のデータを用いて、語彙サイズ 20000 語、サブワードとして音節を用いた音節併用型単語 trigram/bigram を作成した。

4 テキスト検索を用いた未知語の推定

テキスト検索部は、まず、音声認識結果から未知語として検出された語を除外して検索を行う。正しく認識された語によって検索要求に関連する文書が絞られるため、上位の検索文書には未知語部分に相当する語が含まれている可能性が高い。そこで、検索文書中の語から未知語に音韻的に類似する語を DP マッチングによって探索し、未知語部分を補完して認識結

---

* Using OOV Estimation for Open-Vocabulary Speech IR by ITOU, K. (AIST, JST CREST) et. al.

[1] 本稿で利用したデータで計量したところ、対象文書・検索要求どちらにおいても、内容語は全体の 50〜60% 程度の割合であった。

果を修正する。

このため、索引付け処理において、通常の索引に加えて、文書中の語を音素によって探索できるように索引ファイルを拡張しておく。最後に、補完された認識結果を検索要求として再検索を行い、最終的な検索結果が得られる。

以下、実験に用いたIREX日本語検索コレクションの検索課題[3]を用いて具体例を示す。IREXは新聞記事（毎日新聞1994-1995年版）を検索対象にしており、検索課題30件が含まれている。図1に検索課題の例を示す。

```
<TOPIC>
<TOPIC-ID>1010</TOPIC-ID>
<DESCRIPTION>柑橘類の輸入</DESCRIPTION>
<NARRATIVE>オレンジ、レモン、グレープフルーツなどの柑橘系果物の日本への輸入の記事。政府の市場解放や輸入による日本生産地の影響、値段への影響や消費者の反応などの記事を含む。</NARRATIVE>
</TOPIC>
```

図 1　IREXコレクション検索課題の例（課題番号1010）

この課題では、「グレープフルーツ」と「柑橘」の2語が未知語である。図1の<NARRATIVE>に対して、未知語モデルなしで認識した結果、未知語モデルを用いて認識した結果、提案手法によって補完した結果をそれぞれ示す。未知語として特定された箇所は<U>でタグ付けされている。

オレンジ、レモン、グレープラスなの、完結型果物の、日本への輸入の記事。政府の市場開放や、輸入により日本生産地の影響を、二大への影響や、少数の反応などの記事を含む。

オレンジ、レモン、<U>グレープラチナガノ</U>、<U>カンケツケコン</U>だものの、日本への輸入の記事。政府の市場開放や、輸入による日本<U>セイサンチ</U>の影響を、<U>ゲダン</U>への影響や、少数の反応などの記事を含む。

オレンジ、レモン、<U>グレープフルーツ</U>、<U>かんきつ類</U>だものの、日本への輸入の記事。政府の市場開放や、輸入による日本<U>生産</U>の影響を、<U>値段</U>への影響や、少数の反応などの記事を含む。

「かんきつ類」や「生産」は、音声認識の結果としては正しくないものの、検索キーワードとしては有効である。上記の検索課題の場合、未知語モデルを用いない音声認識結果では検索精度（平均適合率）が0.104であったのに対して、補完後の認識結果を用いると0.208に向上した。

5　実験

IREX検索課題中の<NARRATIVE>を男女2名ずつ合計4名の話者に読み上げてもらい、音声による検索要求データを作成し、利用した[4]。これらの検索要求には、今回利用した語彙を用いた場合には、14個の未知語が含まれる。収録環境は録音スタジオで、各発話は卓上型マイクを用いてDATに収録した。全検索課題を通しで収録し、読み間違えた場合には、正しく読めるまで同じ検索要求を続けて発話してもらった。検索文書数はIREXの指針に従って300件としている。

本実験の評価尺度には、音声認識率と検索精度の2つがある。今回は紙面の都合上、検索精度に焦点を当てて説明する。話者ごとの未知語検出数と平均適合率を表1に示す。

表 1　平均適合率の比較

話者 ID	1	2	3	4
未知語検出数	51	56	33	37
未知語モデルなし	0.283	0.275	0.287	0.279
未知語モデル+補完	0.316	0.271	0.266	0.286

話者1と4の検索精度が向上し、本手法の有効性を示す複数の検索課題が存在した。話者2と3では、一部の検索課題において検索精度が顕著に低下して、全体の検索精度を下げる結果となった。誤りには、未知語モデルを用いたことによる誤認識と、索引語を補完する段階での誤りがあった。後者は本研究固有の先駆的な処理であるため未解決な問題が多い。複合語の未知語への対応や統計情報の利用などについて今後検討する必要がある。

6　おわりに

本稿では、語彙統制のない音声入力型のテキスト検索システムの構築手法を提案した。この手法では、音声認識エンジンは従来のままで、未知語部分を音節に分解した言語モデルを用いることで、未知語を検出しどのような音節列かを推定する。検索部では、推定結果と検索結果を比較することで文脈を考慮した未知語の推定を行なうことも可能である。

今後は、より詳細なモデルを用いた未知語の推定精度の向上などを検討する予定である。

参考文献

[1] 徳永健伸. 情報検索と言語処理. 東京大学出版会, 1999.

[2] 伊藤克亘, 田中和世. 被覆率を重視した大語彙連続音声認識用統計的言語モデル. 日本音響学会講演論文集, March 1999.

[3] Satoshi Sekine and Hitoshi Isahara. IREX: IR and IE evaluation project in japanese. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, pp. 1475–1480, 2000.

[4] Katunobu Itou, Atsushi Fujii, and Tetsuya Ishikawa. Language modeling for multi-domain speech-driven text retrieval. *IEEE Automatic Speech Recognition and Understanding Workshop*, 2001.

社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

音声文書検索を用いたオンデマンド講義システム

伊藤 克亘† 藤井 敦†† 石川 徹也††

† 産業技術総合研究所 情報処理研究部門
〒 305-8568 つくば市梅園 1-1-1
†† 図書館情報大学
〒 305-8550 つくば市春日 1-2

E-mail: †itou@ni.aist.go.jp, ††{fujii,ishikawa}@ulis.ac.jp

あらまし 本研究は、講演をオンデマンドで視聴することを目的とし、音声データを媒介として講演資料テキストと講演ビデオデータを統合して利用するモデルを提案し、そのモデルに基づきシステムを実装した。講演のビデオデータの音声トラックから、音声認識システムによって書き起こしを作成し、講演データベースを構築する。ユーザは、講演資料テキストのある範囲や、それに類する文章やキーワードを用いてそのデータベースを検索すると、システムは、検索結果に基づいて関連するビデオデータを再生する。このことにより、視聴したい部分をテキストから高速に検索し、その部分に対応したビデオデータの一部を視聴することが可能になる。

キーワード 音声文書検索、音声認識、情報検索、オンデマンド講義システム、インターネット教育

A Lecture-On-demand System using Spoken Document Retrieval

Katunobu ITOU†, Atsushi FUJII††, and Tetsuya ISHIKAWA††

† National Institute of Advanced Industrial Science and Technology
1-1-1 Chuuou Daini, Umezono, Tsukuba, 305-8568, Japan
†† University of Library and Information Science
1-2, Kasuga, Tsukuba, 305-8550, Japan

E-mail: †itou@ni.aist.go.jp, ††{fujii,ishikawa}@ulis.ac.jp

Abstract This research proposes a lecture-on-demand system, which retrieves video data in response to users' information needs. For this purpose, we utilize texts and audio/video data for a single lecture. Our system extracts audio tracks from lecture video data, transcribes them by way of a large vocabulary continuous speech recognition system, and produces a lecture database. Users can selectively browse a specific video track by submitting textual queries, i.e., keywords, sentences and paragraphs associated with the text for the target lecture. Our current system is implemented over the Web, so as to facilitate e-education.

Key words Spoken Document Retrieval, Speech Recognition, Information Retrieval, Lecure-on-demand System, e-education

1. はじめに

近年、情報通信技術の発達にともなって、大量の電子化されたマルチメディア情報を誰もが容易に受発信できるようになった。このような現状では、必要な情報をいつでもどこでも手軽に活用できる技術が重要となる。

現在、広く使われているマルチメディア情報としては、テキスト、音声、画像があげられる。これらの情報を併用し、何度も利用したくなるコンテンツの一つとして、講演があげられる。講演は、予稿やスライドなどの資料を併用し、対面で話すのが普通である。また、テキストを用意し、それに基づいた講演を放映する形のテレビ番組もたくさん放映されている。

我々は、そのような講演ビデオデータを対象にして、要求に応じた内容を視聴するオンデマンドシステムを実現した。このシステムでは、教科書や予稿などの、講演資料テキストを閲覧しながら、ビデオデータ(音声と画像)の対応する内容を選択的に視聴することが可能である。

このシステムでは、講演のビデオデータの音声トラックから、音声認識システムによって書き起こしを作成し、講演データベースを構築する。そのデータベースに対して、ユーザが、講演資料テキストのある範囲や、それに類する文章やキーワードを用いて検索し、検索結果に基づいてビデオデータを再生する。

2. 講演データの諸相

講演では、一般にあらかじめテキスト資料が用意され、補助資料を用いながら口頭での発表が行なわれる。予稿などの資料をテキスト情報、口頭での発表を音声情報、講演者の様子などを映像情報ととらえると、講演は、互いに関連し補完しあうマルチメディア情報から成り立っていると見なせる。

しかし、これらの情報のうち、講演の場以外では、一般に予稿などのテキスト情報しか利用できない。いつでも、どこでも、手軽に、講演情報を利用するためには、講演の場以外でも、テキスト情報以外の情報も利用できるようにしなければならない。

そこで、以下、講演を構成する情報の相異点とどのように利用できればよいかを考察する。

教科書や予稿といった講演を構成するテキスト情報をここでは、講演資料テキストと総称する。講演資料テキストは、冊子体や電子版といった形態にかかわらず、ある程度の範囲を一度に見わたせ、さらに、章立てのような文書構造や文字種などの表層情報を手がかりにして、いわゆる「斜め読み」ができる。また、何度も繰り返し読み返すことができる。そこで、全体の概要を把握したり、関心のある箇所を高速に探索することが比較的容易である。すなわち、ランダムアクセスに適している。

それに対して、講演情報は、局所的な逐次アクセスが原則であり、ランダムアクセスには適さない。実際の講演では、スライドなどの構造化されたビジュアル情報が併用されたり、話の内容自体が構造化されていることも多いが、それらも、講演者が意図した順序でしかアクセスできず、論文のように戻ったり進んだり、繰り返したりはできない。講演を録画/録音した場合には、早送りや巻戻しは可能であるが、利用者が意図したところで止めたりすることはあらかじめインデックスが用意されていない限り、困難である。講演データの書き起こしが利用可能であっても、そのままでは、講演資料テキストに比べるとかなり読みにくい(注1)。

他方、情報の量は、講演の方が相対的に多い傾向にある。予稿論文は発表内容に対してページ数が厳しく制限されることがある。これは、主として印刷コストやレイアウト上の理由によるものなので、電子版の論文には該当しない。しかし、現状では、電子版の論文が冊子版の形態の域を出ないものであることが多く、電子版も字数の制限を受けることが多い。

また、講演の方が、わかりやすいことが多い。発話内容が適度に冗長であったり、会話的な表現も使用される。熟練した講演者であれば、聴衆にあわせた例を挿入したり、その場の反応に応じて説明を調整することもある。あらかじめ書かれた講演資料テキストに対して講演する場合には、テキスト執筆当時よりも新しくかつ正確な情報が補足されることもある。

これらを整理すると、講演を視聴するための、ある一つのモデルが成立する。すなわち、講演資料テキストを利用して、講演の中の重要部分を探し、その部分だけを講演の中から選択して視聴し、より理解を深める。この結果、講演の一部を重点的に視聴することが可能になり、不必要な部分を何度も見ることを避けられる。

このモデルを可能にするために、音声認識技術と

(注1):いうまでもないことであるが、講演録のようなものがテキストの形で利用できることもあるが、多くの場合、読みやすいように書き言葉ふうに書き変えられていることが多い。

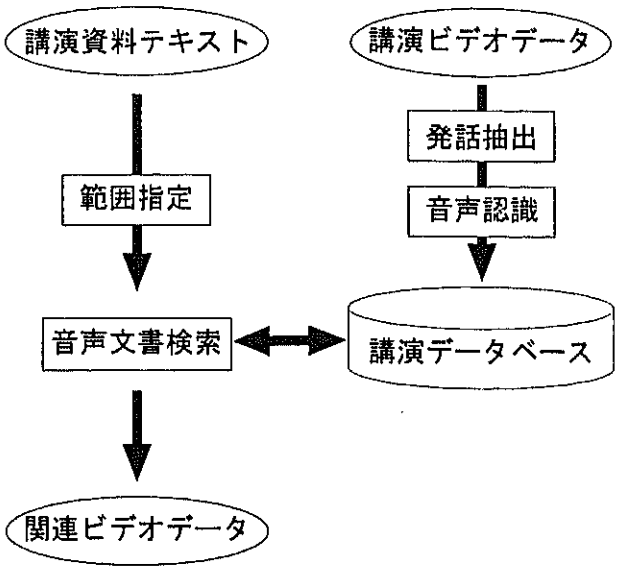

図 1　システム構成図
Fig. 1　The Architecture of the Lecture-On-demand System

情報検索技術を統合して用いる。具体的には、講演データの音声情報を媒介として、講演資料テキストと講演ビデオデータを自動的に対応させる。以下、このモデルに基づいて構築したシステムについて述べる。

3.　オンデマンド講義システム

3.1　システム概要

システムの概要を図 1 に示す。

このシステムでは、教科書や予稿などの講演資料テキストと講演ビデオデータを前提としている。ユーザが、講演資料テキストの一部分を指定したり、調べたい内容を表現したキーワードや文章等を入力すると、システムは、それに関連する講演部分 (本システムでは、パッセージとよぶ) を出力する。

いいかえれば、システムのオンライン処理は、論文から抽出した範囲を問い合わせ (query) と見なし、講演から関連パッセージを検索することである。そこで、問い合わせとパッセージにおける語の頻度分布に基づいて関連度を計算し、関連度の高いパッセージを選択する。

そのためには、一般の情報検索システムにおいて、あらかじめ索引処理を行うように、講演に対して何らかの加工を施して、検索を可能にしておくことが必要となる。具体的には、本システムでは、音声文書検索技術を適用している。あらかじめ、講演データの音声データを取り出して、そのデータを音声認識して書き起こしを作成し、それらをある程度のまとまりごとにパッセージとして整理して、データベース化する。各パッセージは、開始時刻と終了時刻を持つため、その情報を利用して、ビデオデータを再生することで、問い合わせに関連するビデオデータを視聴することができる。

講演資料テキスト中では、簡潔に一つにまとまっている内容が、講演では、いくつかに分かれて点在したり、繰り返し説明されたりすることがある。そこで、単一の問い合わせに対して、関連する説明を尤度順に複数出力することが、網羅性という点では好ましい。しかし、他方、講演の内容や問い合わせの内容によっては、同じような説明を何度も出すことは冗長であり、内容を検索する点から見ると、非効率なこともある。そこで、既に出力した内容との重複を避けることで、最小限の閲覧によって利用者の情報要求を満足するための工夫も行なっている [1]。

3.2　講演データベースの構築

講演データベースを構築する際に問題となるのは、書き起こしを作るための音声認識の精度と、パッセージをどのような単位とするかという二点である。

書き起こしの精度については、対象とする音声データの品質や、講義の内容によって変化するが、おおむね 50% 以上の単語認識率は期待できる。この程度の認識率であれば音声文書検索に十分である [2] と思われる。

パッセージをどのような単位にするかに関しては、非常に難しい問題がある。理想的には、書き言葉の段落のように意味のあるまとまりに分割できることが望ましいが、話し言葉については、文の認定すら難しいのが現状であり、人間が書き起こしても、意味のある単位に分割のは難しい。

さらに、本システムでは、音声認識システムを用いて書き起こしを作成しているため、認識誤りも生じる可能性があり、話し言葉の話題の転換に用いられるような接続表現なども認識誤りの可能性があること、そもそも、言語モデルを書き言葉の学習データから作成しているため話し言葉特有の表現に関しては、特に認識率が低い可能性もある。

したがって、便宜上、パワーなど物理的な尺度で音声を分解し、その単位を必要に応じて、複数個まとめてパッセージとすることで対処する。

3.3　関連パッセージの検索

前節で述べた処理で作成したパッセージを、それぞれ異なる文書と見なすと、情報検索の分野で提案されている各種の手法を用いて、問い合わせに関連するパッセージを効率的に特定できる。

情報検索で典型的に使われる手法は、問い合わせと文書のそれぞれを索引語の統計頻度に関するベクトル (索引語の頻度分布) で表現する。そして、ベクトルの類似度に基づいて問い合わせと文書の関連度を定量化し、関連度が高い文書から優先的に出力する [3], [4]。一般には、特定の文書だけに表われる索引語や、問い合わせと対象文書の両方に共通して多く現われる索引語が重視される。

我々は、形態素解析システム「茶筌」[5] を用いて問い合わせとパッセージを単語に分割し、品詞情報に基づいて名詞を索引語として抽出する。また、カタカナ語や新語は未知語として解析されることが多いので、未知語も索引語として抽出する。

関連度の計算には、確率型の手法 [6] を用いた。これは、近年の情報検索手法の中でも比較的高い検索精度を実現することで知られている。具体的には、文書 i の関連度スコアを式 1 によってもとめる。

$$\sum_t \left(\frac{TF_{t,i}}{\frac{DL_i}{\text{avglen}} + TF_{t,i}} \cdot \log \frac{N}{DF_t} \right) \qquad (1)$$

ここで、$TF_{t,i}$ は索引語 t が文書 i に出現する頻度である。DF_t は t を含む文書数である、N は総文書数である。DL_i は文書 i の長さ (バイト数) であり、avglen は平均文書長である。

3.4 効用最大化に基づく再帰的検索

教科書や予稿では、一つの章や節としてまとまっている内容が、実際の講演では、複数の箇所で分散して説明されることがある。そこで、単一の説明だけを出力するのではなく、スコア順に複数の説明を出力する必要がある。

しかし、他方、十分に理解できる説明を出力したにもかかわらず、同様な説明を何度も出力し続けることは、効果的ではない。一度提示した内容は、それ以降、なるべく出力しないようにすれば、ユーザは少ない労力によって効用を最大化することが期待できる。

これを情報検索の観点からとらえると、利用者がある文書を閲覧した場合に、利用者の情報要求のうち、まだ満足していない部分を特定し、その部分に対して、効果的な別の文書を提示する問題におきかえることができる。

松村ら [7] は、利用者が文章 i を閲覧しても依然として満たされない検索要求を、問い合わせと文章 i との差分ベクトルとして表現し、それ以降の検索に利用する手法を提案した。ここでは、問い合わせと各文書があらかじめ索引語のベクトルとして表現されていることが前提である。

	講演資料テキスト	書き起こし
テレビ講義	3489	11038
学会講演	3342	6017

表 1 文字量の比較
Table 1 A Comparison of The Number of Characters

我々は、この手法を応用し、問い合わせベクトルと一度提示された説明に対応するベクトルとの差分を取りながら再帰的に検索を繰り返すことで、冗長な出力を回避する。

また、検索を繰り返すたびに、問い合わせベクトルが縮退するので、スコアは次第に低下する。これは利用者の要求が次第に満たされていく過程をモデル化している。そこで、スコアに対するしきい値を設定し、スコアがその値を下回った時点で利用者の要求が十分に満たされたと判断して処理を終了する。

現状では、スコアのしきい値や、問い合わせベクトルの縮退方法に関して、決定的な値や方法は分かっていない。様々な検索要求や対象データで調査する必要がある。

4. システム実装例

前章で説明したシステムを、1) 教科書が市販されているテレビ放映されている講義 (45 分間) と、2) 学会におけるチュートリアル講演の再現ビデオ (30 分間、予稿あり) の二種類のコンテンツを用いて構築した。1) では、CATV から受信したビデオデータを DV に録画し、2) では、DVCAM を用いて聴衆がいない状態で撮影した。

構築したシステムは、Web ブラウザで検索および視聴できるように実装した。検索部分などは全てサーバとして実装されているが、現状では、ビデオデータだけはサーバから短時間でダウンロードするにはサイズが大きいため、クライアントのマシン上に保存している。

構築したシステムインタフェースの外観を図 2 に示す。

まず、話し言葉と書き言葉の違いについて調べるために、講演資料テキストと書き起こしの文字数を計量した(注2)。結果を表 1 に示す。

このように、話し言葉は、書き言葉に比べると、2

(注2)：書き起こしは、音声認識により得られたものを利用している。音声認識結果は認識誤りはあるものの、文字数に関しては、正解とそれほどの差はない。

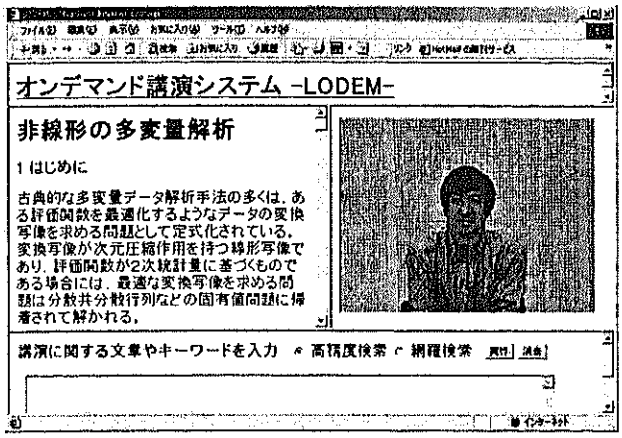

図 2 Web オンデマンド講義システム: LODEM
Fig. 2 LODEM: the Web Lecture-On-demand System

倍から 3 倍程度に冗長であることがわかる。話し言葉と書き言葉の違いは、講演のスタイルなどに依存するため、一概に説明することは困難であるが、今回用いたデータでは、書き言葉には現われないが話し言葉に現われる表現として、以下のような表現が目についた。

（1） 話の進行のためのメタな表現 (あらすじなど)

（2） 具体的な例 (教科書では「公害事件」となっているものが「水俣病」となるなど)

（3） 聞き手に話しかけるような表現 (「皆さんごぞんじのように」など)

話し言葉と書き言葉の違いに関して、テレビ講義での例を示す。太字の部分が話し言葉にのみ現われた表現である。

> 刑法とは、**皆さんご存知のように**どのような行為が犯罪となり、その行為にどのような刑罰が科されるかを定めた法律のことを指します。**六法を開くとどの六法にも刑法と名前のついた法律、すなわち刑法典がのっています。**刑法典は…

DV で録画されたデータの分割は、パワーに基づいて、400ms 程度の無音区間を検出したところで音声データを区切り、それを一つの発話とし、発話三つで一つのパッセージとした。このようにすると、45 分の講義が、34 のパッセージに分割された。このとき、最長で 582 文字、最短で 96 文字、平均 312 文字だった。

音声データの認識性能をはかるために、テレビ講義の一部 (10 分間分) で、連続音声認識コンソーシアムのディクテーションソフトウェア [8] を用い、さらに、いくつかの言語モデル [9] を用いて予備調査を行なった。この言語モデルは、新聞および論文から構築したものではあるが、単語誤り率 (Word Error Rate: WER) で、20% から 30% という結果が得られた。この精度なら、音声文書検索に十分である [2] と思われる。

(検索要求) 公害犯罪を処罰するために制定された「人の健康に係る公害犯罪の処罰に関する法律」
(検索結果) また、1960 年代には、熊本の水俣病事件を、はじめとする。公害事件が多発しました。今の安全率で得るのは、公害事件の、民事裁判に関する映像です。このような発光が事件の多発を契機として、1970 年には向上などから、ヒトの健康がする物質を輩出して、ヒト脳喉、身体に実験を調査する行為を処罰する。人・高にかかる血行が犯罪の処罰に関する法律で稼いでされています。

図 3 検索要求と検索結果の例
Fig. 3 An Example of a Query and an Output

このようにして構築した講義データベースを対象に、実際の教科書の一部を選択して検索要求として、動作確認を行なった。図 3 に検索要求とそれに対する出力を示す。

5. おわりに

本研究では、音声認識技術と情報検索技術を用いて、講演に用いられるマルチメディアデータである、テキスト、音声、ビデオデータを組み合わせて利用するモデルを提案した。本モデルは、講演資料テキストは、構造や表層情報や媒体の特性から、興味ある場所を高速に探索することに適した媒体、講演音声/ビデオデータは、テキスト中の簡潔な表現の理解を助けるための媒体、という仮説に基づいている。これらの媒体を自由に行き来しながら、効率的に講演視聴を可能にするためのモデルである。

さらに、このモデルに基づいて、計算機上のシステムとして実装した。このシステムでは、利用者が興味のある講義資料の一部分を指定したり、調べたい内容を表現したキーワードや文章等を入力すると、システムはそれに対するパッセージを検索し、その部分のビデオデータを再生する。

今後の研究課題としては、システムの要素技術や全体に関する客観的な評価が必要であるが、現状では、評価方法の確立自体が課題である。

謝　　辞

本研究の一部は、科学技術振興事業団 CREST によって行なわれました。

文　　献

[1] 藤井敦, 伊藤克亘, 秋葉友良, 石川徹也. 音声言語データの構造化に基づく講演発表の自動要約. ワークショップ「話し言葉の科学と工学」, pp. 173–177, 3 2001.

[2] Pierre Jourlin, Sue E. Johnson, Karen Sparck-Jones, and Philip C. Woodland. Spoken document representations for probabilistic retrieval. *SPEECH COMMUNICATION*, Vol. 32, No. 1-2, pp. 21–36, 2000.

[3] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval.* Addison Wesley, 1999.

[4] Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill, 1983.

[5] 松本裕治, 北内啓, 山下達雄, 平野善隆, 松田寛, 浅原正幸. 日本語形態素解析システム『茶筌』version 2.0 使用説明書. Technical Report NAIST-IS-TR99012, 奈良先端科学技術大学院大学, 1999.

[6] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 232–241, 1994.

[7] 松村真宏, 大澤幸生, 谷内田正彦. AAS: 文書の組み合わせによってユーザの興味を満足する検索システム. 人工知能学会誌, Vol. 14, No. 6, pp. 1177–1185, 1999.

[8] 鹿野清宏, 伊藤克亘, 河原達也, 武田一哉, 山本幹雄(編). 音声認識システム. オーム社, 2001.

[9] 伊藤克亘, 秋葉友良, 藤井敦, 石川徹也. 音声入力型テキスト検索システムのための音声認識. 日本音響学会講演論文集, 10 2001.

音声入力型テキスト検索システムのための音声認識*

○伊藤克亘 † 秋葉友良 (産総研) △藤井敦 † △石川徹也 (図書館情報大) (†JST CREST)

1 はじめに

音声認識技術は、ここ数年広く利用されるようになってきた。次の段階としては、音声認識以外のアプリケーションと組み合わせていかに高度なアプリケーションを創出していくか、また、それらをどのようにして評価していくかが重要である。

まず考えられるのは、既存のアプリケーションに音声認識を導入することである。とりわけ、情報検索システムは、歴史が長く、主要な情報処理アプリケーションであり、インターネットの普及でさらに利用の範囲が広がっているため、音声認識を導入した研究も近年数多くおこなわれている。これらは、目的に応じて、(1) 音声データの検索 (検索要求の入力は、キーボードが中心である) (2) 音声による検索 (検索要求を音声入力によっておこなう。検索対象はテキストが中心である) に大別できる。

音声データの検索は、TREC の Spoken Document Retrieval (SDR) トラックで放送音声データを対象としたテストコレクションが整備されている [1] ことなどを背景に、盛んに研究が行なわれている。

音声による検索に関しては、重要な技術であるにもかかわらず、評価セットなどの整備も遅れており、まだまだ研究が盛んだとはいえない。

Barnett らは、WSJ のテキストから構築した 20K 語彙のバイグラムを用いたディクテーションシステムを、テキスト検索システム INQUERY の入力として利用した実験を行なっている [2]。評価セットとしては、TREC の検索課題 35 件に対する単一話者の読み上げ音声を入力として利用し、TREC コレクションの検索実験を行なっている。Crestani も同じ読み上げ課題を用いた実験を行なっている [3]。しかし、どちらの実験においても、そもそも音声認識システムの性能が単語誤り率 (Word Error Rate: WER) 30% 以上と、十分でないという問題がある。

現行の音声認識システムでは、音響モデルも言語モデルも確率モデルが使われることが多いため、あるアプリケーションへの音声認識の導入には、音響モデルや言語モデルを学習するためのデータの質や量も重要な問題であり、評価の際には、それらの点についても論じられるべきである。

情報検索では、検索対象のテキストが検索システムにとって、既知であることが多い。したがって、その検索対象のテキストを用いて、言語モデルを学習することは自然であると考えられる。これらの言語モデルを用いることで、どの程度、音声認識システムの性能が変化するかを評価した。

2 実験

2.1 評価データ

現在、一般に入手可能な日本語テキスト検索用テストコレクションには、NTCIR[4, 5](論文の抄録) IREX[6], BMIR-J2 (新聞記事) などがある。これらのコレクションは、検索対象の大規模なテキストデータとそれらに対する検索課題からなる。例を図 1 に示す。

```
<TOPIC q=0101> <TITLE>B型肝炎</TITLE>
<DESCRIPTION>遺伝子工学的手法によるB型肝炎ワクチンの開発について論じている文献</DESCRIPTION>
<NARRATIVE>肝炎などのウイルス性疾患に対する安全かつ有効な予防法の確立は21世紀に向けての医療分野での重要な課題である。そのため、(中略) しかし、遺伝子工学的手法に触れていない論文は不可。また、B型肝炎以外のワクチンも不可。</NARRATIVE>
<CONCEPT>a. B型肝炎, b. 遺伝子工学的手法, c. ワクチン, 予防接種</CONCEPT>
<FIELD>7. 医学・歯学</FIELD> </TOPIC>
```

図 1 検索課題の例 (NTCIR)

これらの項目のうち、<TITLE> <DESCRIPTION> <NARRATIVE> が実際の検索要求として想定されているが、今回は、このうち、NTCIR の<DESCRIPTION> の部分と IREX の <NARRATIVE>の部分を一文づつに区切ったものを読み上げたデータを評価用入力として用いることにした[1]。発話数は、NTCIR-1(予備版、公式版) NTCIR-2 をあわせて 132 件、IREX は 72 件だった。全ての課題を、男女それぞれ 2 名分、録音スタジオで、NTCIR は卓上マイクを用い、IREX は接話型マイクを用いて DAT に収録した。

2.2 音声検索システム

実験に用いた音声検索システムは、音声認識部とテキスト検索部からなる [8]。音声認識部には、日本語ディクテーション基本ソフトウェア [7] を用いた。ただし、言語モデルは、毎日新聞 1991 年から 2000 年の 10 年分から学習した MNP モデルと、NTCIR-2 の文書コレクションから学習した NTCIR モデル、両方の文書をあわせたデータから学習した混合モデルの三つを新たに構築して用いた。比較のため、語彙サイズは同一 (20K) とした。学習データの量、各言

---

* A Speech Recognition System for a Speech-Driven Text Retrieval System by ITOU, K. (AIST, JST CREST) et. al.

[1] IREX の検索要求は、NTCIR のものに比べるとかなり短いので、IREX に関しては<NARRATIVE>を用いた。

表 1 言語モデルの構成および認識実験結果

		NTCIR	MNP	混合
総形態素数		175M	262M	
異なり語数		454K	315K	652K
被覆率 (%)	NTCIR	97.9	88.3	96.8
	MNP	88.3	96.5	95.8
未知語率 (%)	NTCIR	4.2	8.7	4.3
	IREX	6.1	0.9	1.2
PP	NTCIR	60	159	59
	IREX	138	96	105
WER	NTCIR	0.186	0.414	0.214
	IREX	0.329	0.166	0.206

語モデルの学習データに対する被覆率、評価データに対する未知語率、評価データに対するテストセットパープレキシティ(PP) を表 1 に示す。

言語モデルは、高頻度語上位 20000 語で制限してバイグラム、逆向きトライグラムを作成した。カットオフは、それぞれ 1 とした。(混合モデルは 2 とした。) 上位 20000 位の出現回数が、MNP は 502、NTCIR は 166 だったので、混合モデルでは、NTCIR の頻度に 3.02(= 502/166) を乗じて足し合わせて作成した。

2.3 実験方法

二つのコレクション (NTCIR と IREX) を読み上げたデータを、二つの言語モデルを用いて認識実験し、WER (= $\frac{\text{削除+挿入+置換}}{\text{全語数}}$) で評価した。実験結果は表 1 に示した通りである。

いずれの課題の場合も、検索対象のテキストから作成したモデルとそうでないモデルでは、誤り率が 2 倍程度の差があることがわかる。また、混合モデルは、どちらの課題に対しても検索対象のテキストから作成したモデルに近い認識率を示しており、安定した性能であることがわかる。また、これらの傾向は、表 1 の未知語率、テストセットパープレキシティにも現われている。

NTCIR の課題については、検索実験も行なった [8]。入力として音声入力を用いる他は、NTCIR ワークショップの評価と同様の方法で、平均適合率 (AP) を求めたところ、表 2 のようになった。

表 2 検索実験結果

	テキスト	NTCIR	MNP	混合
平均適合率	0.337	0.261	0.111	0.231

検索結果でも、認識率同様に NTCIR モデルの効果があらわれていることがわかる。

NTCIR の課題に対しては、専門用語を多く含む発話に関して NTCIR モデルを用いた効果が顕著であった。たとえば、検索課題 0027「シソーラスを用いたテキストの検索課題について」の NTCIR モデルと MNP モデルの AP と認識結果の例を表 3 に示す。これを見ると、「シソーラス」のように専門性の高い用語の認識誤りは、検索性能にも強く影響することがわかる。

表 3 認識結果と平均適合率の例

モデル	認識結果	AP
NTCIR	シソーラスを用いた、テキストの連想検索について	0.687
MNP	首相ら総務した。テキストの連鎖を検索について	0.007

一方、IREX の課題に対しては、スポーツや芸能、事件といった、論文で取り上げられないものに関して、新聞モデルを用いた効果が顕著であった。

3 おわりに

検索対象文書から言語モデルを作成することで、検索課題に対する認識性能を向上することができることを示した。

また、既存のテキスト検索テストコレクションの検索要求を読み上げた評価用音声コーパスを構築した。

今後は、話者を男女各 5 名程度に増やす予定である。また、IREX に関する検索実験の評価も行なう予定である。

謝辞

NTCIR コレクションは国立情報学研究所の許諾を得て使用させていただきました。関和広氏、伊堂史織氏 (図書館情報大) には実験用データの実験用データの整備を補助していただきました。

参考文献

[1] John S. Garofolo, Ellen M. Voorhees, Vincent M. Stanford, and Karen Sparck Jones. TREC-6 1997 spoken document retrieval track overview and results. In *Proceedings of the 6th Text Retrieval Conference*, pp. 83–91, 1997.

[2] J. Barnett, S. Anderson, J. Broglio, S. Singh, R. Hudson, and S. W. Kuo. Experiments in spoken queries for document retrieval. In *Proc. Eurospeech 97*, pp. 1323–1326. ESCA, 1997.

[3] F. Crestani. Word recognition errors and relevance feedback in spoken query processing. In *Proceedings of the Fourth International Conference on Flexible Query Answering Systems*, pp. 267–281, 2000.

[4] NACSIS. *Proceedings of the 1st NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, 1999.

[5] NII. *Proceedings of the 2nd NTCIR Workshop Meting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization*, 2001.

[6] Satoshi Sekine and Hitoshi Isahara. IREX: IR and IE evaluation project in japanese. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, pp. 1475–1480, 2000.

[7] 鹿野清宏, 伊藤克亘, 河原達也, 武田一哉, 山本幹雄 (編). 音声認識システム. オーム社, 2001.

[8] 藤井敦, 伊藤克亘, 秋葉友良, 石川徹也. 音声入力型文書検索システムの開発とテストコレクションの構築. 情処学報, Vol. 2001-FI-63, pp. 65–72, July 2001.

音声入力型文書検索システムの開発とテストコレクションの構築

藤井 敦[†,†††] 伊藤 克亘[††,†††] 秋葉 友良[††] 石川 徹也[†]

[†] 図書館情報大学
〒 305-8550 つくば市春日 1-2
[††] 産業技術総合研究所
〒 305-8568 つくば市梅園 1-1-1 中央第 2
[†††] 科学技術振興事業団 CREST

E-mail: fujii@ulis.ac.jp

あらまし 本研究は，音声発話によって必要な情報を検索するためのシステムに関する研究開発を行った．従来の関連研究では，テキスト検索精度の向上が主要なテーマであり，音声認識精度の向上は対象とされてこなかった．本研究で提案するシステムの特長は，検索対象テキストを用いて音声認識用の言語モデルを作成し，利用する点にある．ユーザの発話は検索対象テキストに関連するものが中心なので，本システムは比較的高精度の音声認識を実現できる．さらに，システムを評価するためのテストコレクションを構築するために，既存のテキスト検索用コレクションを利用し，検索要求に関する音声発話データを収録した．当該テストコレクションを用いた実験によって本システムの有効性を確認することができた．

キーワード テキスト検索，音声認識，言語モデル，NTCIR コレクション，読み上げ音声

A Speech-Driven Text Retrieval System and its Evaluation Using a Test Collection

Atsushi Fujii[†,†††], Katunobu Itou[††,†††], Tomoyosi Akiba[††], Tetsuya Ishikawa[†]

[†] University of Library and Information Science
1-2 Kasuga, Tsukuba, 305-8550
[††] National Institute of Advanced Industrial Science and Technology
1-1-1 Chuuou Daini Umezono, Tsukuba, 305-8568, Japan
[†††] CREST, Japan Science and Technology Corporation

E-mail: fujii@ulis.ac.jp

Abstract To facilitate retrieving information with spoken queries, we propose a speech-driven text retrieval system. In past research, no attempt has been made to improve speech recognition in the context of speech-driven retrieval. In our system, a language model used for speech recognition is produced based on a target text collection, so that user queries associated with the collection can be recognized with a high accuracy. We also produced a test collection to evaluate our system, for which we recorded dictated queries in an existing collection for text retrieval. We show the effectiveness of our system by way of experiments using this collection.

Keywords text retrieval, speech recognition, language models, the NTCIR collection, read speech

1 はじめに

近年の音声認識技術は，内容がある程度整理されている発話に対しては実用的な認識精度を達成できる．また，ハードウェア技術の発展にも支えられ，パソコン上で動作する商用/無償の音声認識ソフトウェアが存在する．そこで，既存のアプリケーションに音声認識を導入することは比較的容易になっており，その需要は今後ますます増加するだろう．

とりわけ，情報検索システムは歴史が長く主要な情報処理アプリケーションの一つであるため，音声認識を採り入れた研究も近年数多く行われている．これらは目的に応じて以下の2つに大別できる．

- 音声データの検索 [5, 12, 13, 14, 16, 17]
  放送音声データなどを対象にした検索である．入力手段は問わないものの，テキスト（キーボード）入力が中心である．
- 音声による検索 [2, 3]
  検索要求を音声入力によって行う．検索対象の形式は問わないものの，テキストが中心である．

すなわち，これらは検索対象と検索要求のどちらを音声データと捉えるかが異なる．さらに，両者を統合すれば，音声入力による音声データ検索を実現することも可能である．しかし，現在そのような研究事例はあまり存在しない．

音声データの検索は，TRECのSpoken Document Retrieval (SDR) トラック [4] で放送音声データを対象にしたテストコレクションが整備されていることを背景にして，盛んに研究が行われている．

他方において，音声による検索はカーナビゲーションシステムやコールセンターのようにキーボード入力を前提としない（バリアフリーな）アプリケーションを支える重要な基盤技術であるにも拘らず，音声データ検索に比べて研究事例は極端に少ない．

そこで，本研究は音声入力によるテキスト検索に焦点を当て，システムの研究開発を行った（2章）．さらに，テストコレクションを構築・整備し（3章），本システムの評価に利用した（4章）．

2 本研究で提案する検索システム

2.1 背景と動機

音声による検索に関する従来のシステムでは，概して，音声認識とテキスト検索は完全に独立したモジュールとして存在し，単に入出力インタフェースで接続されているだけである．また，検索精度の向上に焦点が当てられ，音声認識精度の向上は研究対象となっていないことが多い．

Barnettら [2] は，既存の音声認識システム（語彙サイズ20,000）をテキスト検索システムINQUERYの入力として利用して，音声による検索の評価実験を行った．具体的には，TRECの検索課題35件（101-135）に対する単一話者の読み上げ音声をテスト入力として利用し，TRECコレクションの検索実験を行った．Crestani [3] も上記35件の読み上げ検索課題を用いた実験を行い（通常のテキスト検索で用いられる）適合性フィードバックによって検索精度が向上することを示している．しかし，どちらの実験においても既存の音声認識システムを改良せずに利用しているため，単語誤り率は比較的高い（30%以上）．

それに対して，本研究では音声認識の精度向上にも焦点を当て，音声認識とテキスト検索の有機的な統合を目指してシステムの研究開発を行った．音声認識システムの多くは，内部データ（単語辞書など）を切替えることで，目的に応じた使い分けが可能である．そこで，検索対象テキストに基づいて内部データを作成し，利用することは自然な発想だろう．

統計的な音声認識システム [1] は主に音響モデルと言語モデルで構成され，両者は音声認識精度に強く影響する．音響モデルは音響的な特性に関するモデルであり，検索対象テキストとは独立な要素である．

言語モデルは音声認識結果（候補）の言語的妥当性を定量化するためのモデルである．しかし，あらゆる言語現象全てをモデル化することは不可能であるため，一般的には，与えられた学習用コーパスに出現する言語現象に特化したモデルを作成する．

本研究で提案するシステムにおいては，ユーザの発話は検索対象テキストに関連する内容である可能性が高い．そこで，検索対象テキストに基づいて言語モデルを作成すれば，音声認識の精度向上が期待できる．その結果，ユーザの発話が正しく認識されるので，テキスト入力に近い検索精度を実現することが可能になる．

音声認識の精度を高めることは，インタラクティブ検索を円滑に進めたり，発話通りの要求に基づいて検索が行われている安心感をユーザに与える上でも重要である．

2.2 システム構成

本研究で提案する検索システムの構成を図1に示す．本システムの特長は，検索テキストに基づいて音声認識精度を高めることで，音声認識とテキスト検

索の有機的な統合を実現する点にある．そこで，まずオフライン処理（破線矢印）によって，検索対象となるテキストコレクションから音声認識用の言語モデルを作成する．

オンライン処理では，ユーザが検索要求を発話すると，音響モデルと言語モデルを用いて音声認識が行われ，書き起こしが生成される．実際には，複数の書き起こし候補が生成され，尤度を最大化する候補が選択される．ここで，言語モデルはテキストコレクションに基づいて作成されているので，コレクション中のテキストに言語的に類似する書き起こしが優先的に選択される点に注意を要する．

次に，書き起こされた検索要求を用いてテキスト検索を実行し，検索結果をユーザに提示する．検索結果閲覧を支援するためには効果的な提示手法が必要である．しかし，出力インタフェースについては今回の研究では対象外とし，今後検討を行う．

なお，本システムは現在は日本語を対象に実装されているものの，原理的には対象言語を問わない．

以下，2.3，2.4節で音声認識とテキスト検索についてそれぞれ説明する．

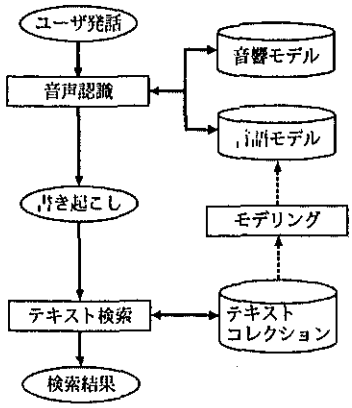

図 1: 音声入力型テキスト検索システムの構成

2.3 音声認識

音声認識には連続音声認識コンソーシアムの日本語ディクテーション基本ソフトウェアを用いた [6, 19] [1]．本ソフトウェアは2万語規模の単語辞書を用いて，ほぼ実時間に近い動作で90%の認識精度を実現できる．

音響モデルと認識エンジン（デコーダー）は，本ツールキット付属のものを変更せずに利用する．

他方において，統計的言語モデル（単語 N グラム）は検索対象のテキストコレクションに基づいて作成する．本ソフトウェアに付属されている関連ツール群や一般に利用可能な形態素解析システム「茶筌」[20] を併用することで，様々な対象に対して比較的容易に言語モデルを作成できる．すなわち，対象テキストから不要部分を削除するなどの前処理を行い「茶筌」を用いて形態素に分割し，読みを考慮した高頻度語制限モデルを作成する [18]．

[1] http://www.lang.astem.or.jp/CSRC/

2.4 テキスト検索

テキスト検索には確率的手法 [9] を用いた．本手法は，近年のいくつかの評価実験によって比較的高い検索精度を実現することが示されている．

検索要求が与えられると，索引語の頻度分布に基づいてコレクション中の各テキストに対する適合度を計算し，適合度が高いテキストから優先的に出力する．テキスト i の適合度は式 (1) によって計算する．

$$\sum_t \left(\frac{TF_{t,i}}{\frac{DL_i}{avglen} + TF_{t,i}} \cdot \log \frac{N}{DF_t} \right) \qquad (1)$$

ここで，t は検索要求（本システムでは，ユーザ発話の書き起こしに相当する）に含まれる索引語である．$TF_{t,i}$ はテキスト i における索引語 t の出現頻度である．DF_t は対象コレクションにおいて索引語 t を含むテキストの数であり，N はコレクション中のテキスト総数である．DL_i はテキスト i の文書長（バイト数）であり，$avglen$ はコレクション中の全テキストに関する平均長である．

適合度を適切に計算するためには，オフラインでの索引語抽出（索引付け）が必要である．そこで「茶筌」を用いて単語分割，品詞付与を行う．さらに，品詞情報に基づいて内容語（主に名詞）を抽出し，単語単位で索引付けを行って転置ファイルを作成する．オンライン処理では，書き起こされた検索要求に対しても同様の処理で索引語を抽出し，検索に利用する．

3 テストコレクションの構築

3.1 背景と動機

研究開発した検索システムの性能を定量的に評価することは，問題点の分析や改善を行うために重要である．システムの評価には大きく分けて「実世界における試験運用」と「研究室における実験」がある．

前者は，入出力インタフェースのデザインや応答時間などの運用上の問題まで考慮して評価を行う必要がある．

それに対して，後者は比較的限定された設定のもとで行う評価である．すなわち，あらかじめ用意さ

れた検索要求に対してシステムが出力した検索結果を何らかの尺度によって評価する．そのためには，テストコレクションを構築し，評価用ベンチマークとして利用する手法が効果的である．

大規模なテストコレクションの作成には膨大なコストを要するものの，一旦作ってしまえばシステムの性能評価を繰り返すことが容易になる．そこで，被験者をその都度雇わなくても，様々な手法を比較評価しながらシステムを改善できる．

そこで，本研究は「研究室での実験」に焦点を当て，音声入力型テキスト検索システムのためのテストコレクション構築を目的とした．

Barnett ら [2] は，TREC の検索課題 35 件を話者に読み上げてもらい，音声による検索要求データを作成した．そこで，当該データを TREC のテキストコレクションおよび適合性判定と併用することで，音声入力によるテキスト検索の精度を定量的に評価することが可能である．

また，一般公開されているテキスト検索コレクションを利用すれば，関連データを研究者間で共有することも容易になり，当該分野の発展に貢献できる．

ただし，研究室における実験の場合であっても，テスト入力として利用する検索要求は実世界の検索要求にできるだけ近づけることが好ましい．そのためには，ユーザ発話の特性について（少なくとも）次節に示す 5 つの観点から検討する必要がある．

3.2 ユーザ発話に関する特性

内容の詳細度　情報検索を広義に解釈すれば，ユーザが抱えている問題を解決できる情報を見つけ出すことである．すなわち，言語化されていない直観的な情報要求（visceral need [15]）を検索質問（query）に言語化してから検索を行い，検索したテキスト内容を理解する一連の処理である．

しかし，多くの研究では情報検索は狭義に解釈され，言語化された検索質問をシステムに対する直接の入力として扱う．例えば，キーワードは極端に具体化された検索質問の形態である．

しかし，音声入力型システムの場合は，キーボード入力では躊躇するような曖昧な要求でも，発話することは容易かもしれない．そこで，通常のテキスト入力検索に比べて，相対的に詳細度の低い検索要求まで考慮する必要がある．すなわち，音声認識の観点から見れば，短い単語から長い文まで一律に認識できる必要がある．

発話の長さ　ユーザが一度に無理なく自然に発話できる長さ（単語数や時間）には限度があるため，極端に長い検索要求を利用することは現実的ではない．

これはテキスト入力型の検索システムについても当てはまる．しかし，テキスト入力の場合は，類似文書検索のように他の場所からダウンロード（コピー）した長いテキストも入力できるため，検索要求の長さに関する制約は比較的緩やかである．

内容（語彙）の多様性　適用範囲の広いシステムを目的とする場合，多様な検索要求を評価に用いることが好ましい．情報検索の研究分野では，経験的に数十以上の検索要求を利用することが推奨されている．音声認識の観点からは，検索要求に含まれる語彙の多様性に関する評価と捉えることができる．

発話スタイル　音声認識の評価では，読み上げ音声（read speech），自発音声（spontaneous speech），会話音声（conversational speech）等の発話スタイルを区別することが重要である．

検索システムの入力として利用する場合は，読み上げ/自発音声を区別する必要がある．Barnett ら [2] が行ったように，既存の検索課題をそのまま発話すれば読み上げ音声である．他方において，話者が検索課題を理解し，課題達成に必要な検索要求を自分で考えて発話すれば，自発音声に近くなる．

話者の特性　対象を特定話者に限定するのか，不特定話者を想定するのかを考慮して，目的に応じた検索要求を作成する必要がある．また，性別や年齢などの多様性についても検討する必要がある．

以上 5 項目のうち，最初の 3 項目は音声認識とテキスト検索に共通の観点であり，残りの 2 項目は音声認識に固有の観点である点に注意が必要である．

3.3 対象にしたテキストコレクション

現在一般に入手可能な日本語テキスト検索用テストコレクションのうち，代表的なものを以下に挙げる．ただし，各コレクションの詳細は割愛する．

- NTCIR [7, 8][2]
- IREX [11][3]
- BMIR-J2 [10][4]

IREX と BMIR-J2 が毎日新聞記事を検索対象テキストとしているのに対して，NTCIR は技術文書（論文抄録，科研費成果報告書概要）を対象にしている点

[2] http://research.nii.ac.jp/ntcir/index-ja.html
[3] http://cs.nyu.edu/cs/projects/proteus/irex/
[4] http://www.ulis.ac.jp/~ishikawa/bmir-j2/

が異なる．いずれのコレクションも本研究の目的に適しているものの，今回はNTCIRを対象にした．

広義には，NTCIRコレクションには多言語検索や自動要約用のコレクションも含まれる．しかし，今回は日本語検索コレクションのみを対象にした．そこで，以下では，NTCIRコレクションを日本語検索コレクションと同義で扱う．さらに，日本語検索コレクションには，NTCIR-1（予備版，公式版），NTCIR-2（公式版）の3種類の異なるコレクションが存在する．

それぞれの検索課題数およびテキスト数を表1に示す．ここで，NTCIR-1の予備・公式版に含まれるテキストは同一であり，NTCIR-2公式版テキストはNTCIR-1を完全に包含している．ただし，複数のコレクションに重複して含まれる検索課題はない．

すなわち，合計132件の検索課題を用いて音声発話による検索要求データを作成した．

表1: NTCIR日本語検索コレクションの構成

コレクション名	検索課題数（課題番号）	テキスト数
NTCIR-1予備版	30（0001-0030）	332,918
NTCIR-1公式版	53（0031-0083）	332,918
NTCIR-2公式版	49（0101-0149）	736,166

3.4 音声検索要求データの作成

図2にNTCIRコレクションに含まれる検索課題の例を示す．各検索課題はSGML形式で記述された複数の項目で構成される．これらのうち，検索システムの評価に使用することが推奨されている項目は，主に<TITLE>，<DESCRIPTION>，<NARRATIVE>の3つである．これらは，いずれも検索要求を表現するものの，後半ほど詳細であり長い．

我々は3.2節の議論に基づいて，音声による検索要求を作成した．

まず，内容の詳細度と発話の長さの観点から，<DESCRIPTION>のみを利用した．すなわち，システムの入力としては，専門用語と一般語が混在する文単位の発話を想定する．

内容（語彙）の多様性については，NTCIR検索課題の範囲に事実上制限される．しかし，情報処理，医学などの多分野に渡っている点に注意を要する．

発話スタイルに関しては，<DESCRIPTION>の読み上げを行った．今回のテストコレクション構築はパイロット的な側面もあるため，比較的音声認識を行いやすい対象を選択した．今後は対象の難易度を段階的に上げていき，最終的には検索課題に基づいた自発音声を収録することも検討している．

話者の特性については，男女2名ずつ合計4名（いずれも大学院生，年齢は20代）の話者に読み上げを依頼した．最終的には，合計10名を目標に収録を行う予定である．

収録環境は録音スタジオで，各発話は卓上型マイクを用いてDATに収録した．132件の<DESCRIPTION>は全て通して収録した（途中で一度休憩を入れた）．ただし，132件をランダムな順序で記述した原稿を渡した．すなわち，疲れや慣れなどの身体的・心理的な影響によって，コレクションごとの発話状態がばらつかないようにした．読み間違えた場合には，正しく読めるまで同一検索課題を続けて発話してもらった．

なお，Barnettら[2]が作成したTREC検索課題の音声発話については，単一話者による読み上げであること以外の詳細（検索課題中のどの項目をどのような手順で収録したのかなど）は不明である．

4 実験

4.1 方法

本研究の検索システム（2章）をテストコレクション（3章）を用いて評価した．評価方法は，システムの入力として音声発話データを利用する点を除けば，NTCIRワークショップにおける評価と同じである．

すなわち，各検索要求に適合するテキスト上位1,000件を出力し，再現率-適合率曲線と補間なし平均適合率に基づいて，以下に示す異なるシステム（手法）の比較評価を行った．

1. テキスト入力型のシステム
   音声認識を100%正しく行うシステム，すなわち理想的なシステムと見なすことができる．
2. 音声入力型システム
   本研究で提案するシステムであり，NTCIR-2公式版コレクションに基づいて作成された言語モデル（NTCIRモデル）を用いて音声認識を行う．NTCIR-2公式版テキストはNTCIR-1予備・公式版を包含しているので，検索対象によらずに全て当該モデルを使用した．
3. 音声入力型システム
   本評価実験におけるベースラインとして利用する．日本語ディクテーション基本ソフトウェアに付属されている言語モデルを用いて音声認識を行う．当該モデルは，毎日新聞75か月分に基づいて作成された（新聞モデル）．

上記システムのうち，2と3における違いは，言語モデルに利用するデータ（コーパス）だけに起因す

<TOPIC q=0101>
<TITLE>B型肝炎</TITLE>
<DESCRIPTION>遺伝子工学的手法によるB型肝炎ワクチンの開発について論じている文献</DESCRIPTION>
<NARRATIVE>肝炎などのウイルス性疾患に対する安全かつ有効な予防法の確立は21世紀に向けての医療分野での重要な課題である。そのため、遺伝子工学的手法によるB型肝炎ワクチンの開発について論じていれば検索要求を満たす。開発されたB型肝炎ワクチンの物理化学的特性を論じているものやその免疫力増強に有効な免疫アジュバントについて論じているものも検索要求を満たす。しかし、遺伝子工学的手法に触れていない論文は不可。また、B型肝炎以外のワクチンも不可。</NARRATIVE>
<CONCEPT>a. B型肝炎，b. 遺伝子工学的手法，c. ワクチン，予防接種</CONCEPT>
<FIELD>7. 医学・歯学</FIELD>
</TOPIC>

図 2: NTCIR コレクション検索課題の例（NTCIR-2 公式版，課題番号 0101）

る．各言語モデルに使用した学習データの内訳を表2に示す．いずれの場合も同一のツール群を用いて，語彙サイズ 20,000 の単語トライグラムを作成した．

表 2: 言語モデル作成に使用したデータの比較

	NTCIR-2	新聞記事
総形態素数	178M	117M
異なり語数	459K	253K

システム 2 と 3 では話者 4 名の発話を個別に使用したので，事実上，9 種類の異なる検索結果を比較評価した．

NTCIR コレクションの適合判定には，適合（S,A），部分適合（B），不適合（C）のレベルがある．本実験では「適合」のみを正解と見なした．

4.2 結果と考察

まず，各システムの検索精度を概観するために，再現率-適合率（recall-precision）曲線をコレクションごとに図3〜5に示す．グラフが右上に位置するほど検索精度が高い．

これらの図より，話者によらずに，本システムの検索精度は新聞モデルを用いたシステムの検索精度を大幅に改善していることが分かる．言い替えれば，本研究の趣旨すなわち検索対象テキストに基づく言語モデルの利用が妥当であることが分かった．

本システムの検索精度はテキスト入力型システムの精度には及ばないものの，NTCIR-1 予備版においては，かなり接近していることが分かる．

次に，検索精度と音声認識精度の関係を分析するために，補間なし平均適合率と音声認識誤り率を表3に示す．誤り率としては単語誤り率（word error rate: WER）が一般的である．これは，式 (2) を用いて，正解発話に対する削除・挿入・置換の割合を単語単位で計算する尺度である．

$$\frac{\text{削除} + \text{挿入} + \text{置換}}{\text{全語数}} \qquad (2)$$

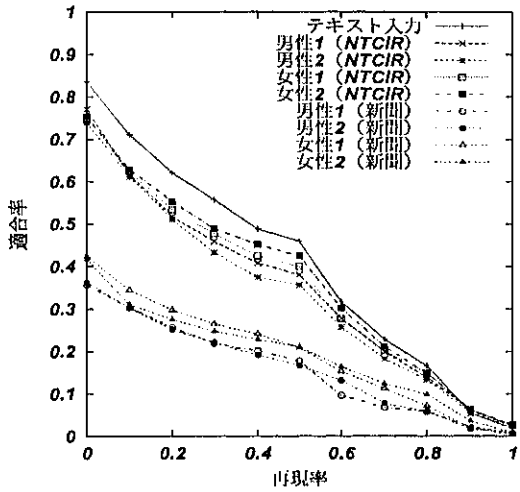

図 3: 再現率-適合率曲線（NTCIR-1 予備版）

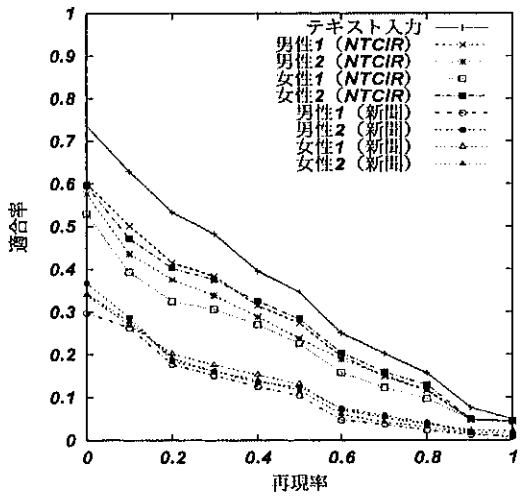

図 4: 再現率-適合率曲線（NTCIR-1 公式版）

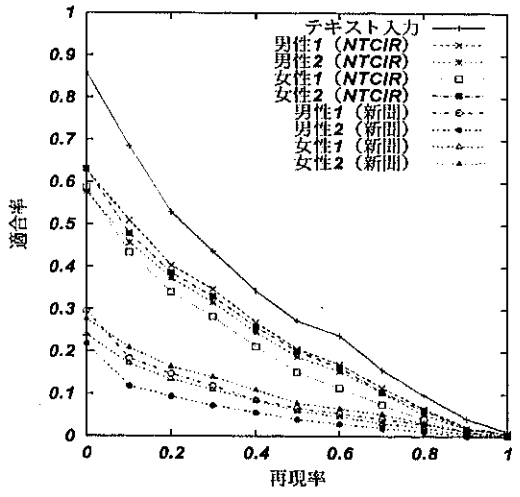

図 5: 再現率-適合率曲線（NTCIR-2 公式版）

しかし，実際には音声認識結果中の単語全てが検索に利用される訳ではない．そこで，検索に利用された検索キーワードの単位でも誤り率（keyword error rate: KER）を計算した．

概して，NTCIR モデルは新聞モデルの単語・キーワード誤り率を大幅に削減し，音声認識，検索精度（平均適合率）の両面から効果的であることが分かる．この傾向は検索課題に専門用語を多く含む場合に，より顕著であった．

例えば，検索課題 0027「シソーラスを用いたテキストの連想検索について」の認識結果と平均適合率を言語モデルごとに比較した結果を表 4 に示す．これは「シソーラス」や「連想検索」などの専門性の高い単語の認識誤りが平均適合率に強く影響する典型的な例である．

しかし，検索課題 0119「日本人の生活価値観の変化」のような一般語だけで構成される発話は，どちらのモデルでも話者によらず正しく認識された．

他方において，NTCIR-1 予備版以外は，単語誤り率よりもキーワード誤り率の方が高いことも分かった．NTCIR モデルを用いると，検索対象テキスト（すなわち，技術文書）に含まれない表現に対する認識精度が低い．

例えば，NTCIR 検索課題には「～の文献が欲しい」という表現が頻出する．しかし「欲しい」という語が NTCIR モデルでは未知語となるため「～の文献が星」のように誤認識され，その結果「星」が名詞性のキーワードとして抽出されてしまった．

また，トライグラムモデルでは語の共起が考慮されるため，発話中の一部が誤認識されると周辺語も連鎖的に誤認識される可能性がある．そこで，今後は検索対象テキストと検索要求との表現の不整合について対処する必要がある．

5 おわりに

音声認識とテキスト検索を統合し，音声発話によって高精度のテキスト検索を実現するシステムを提案した．本システムの特長は，検索対象テキストに基づいて音声認識用の言語モデルを作成することで音声認識精度を高める点にある．

さらに，システム評価用のベンチマークを構築するために，既存のテキスト検索用テストコレクションを用いて検索要求に対する音声発話データを作成した．

当該テストコレクションを用いた評価実験の結果，本システムの検索精度はテキスト入力型システムの精度に比べて若干劣るものの，既存の音声認識システムを拡張せずに利用した場合に比べてはるかに高いことが示された．今後は，音声発話データの規模拡張など行いながら，システムの性能をさらに向上させる予定である．

謝辞

NTCIR コレクションは国立情報学研究所の許諾を得て使用させて頂きました．関和広氏，伊堂史織氏（図書館情報大学）には実験用データの整備を補助して頂きました．本研究の一部は日本学術振興会科学研究費補助金（課題番号 12680406）の助成による．

参考文献

[1] Lalit. R. Bahl, Frederick Jelinek, and Robert L. Mercer. A maximum linklihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 5, No. 2, pp. 179–190, 1983.

[2] J. Barnett, S. Anderson, J. Broglio, M. Singh, R. Hudson, and S. W. Kuo. Experiments in spoken queries for document retrieval. In *Proceedings of Eurospeech97*, pp. 1323–1326, 1997.

[3] Fabio Crestani. Word recognition errors and relevance feedback in spoken query processing. In *Proceedings of the Fourth International Conference on Flexible Query Answering Systems*, pp. 267–281, 2000.

[4] John S. Garofolo, Ellen M. Voorhees, Vincent M. Stanford, and Karen Sparck Jones. TREC-6 1997 spoken document retrieval track overview and results. In *Proceedings of the 6th Text REtrieval Conference*, pp. 83–91, 1997.

[5] S.E. Johnson, P. Jourlin, G.L. Moore, K. Spärck Jones, and P.C. Woodland. The Cambridge University spoken document retrieval system. In *Proceedings of ICASSP'99*, pp. 49–52, 1999.

表 3: 検索実験結果（AP: 平均適合率, WER: 全単語誤り率, KER: キーワード誤り率）

入力方法	NTCIR-1 予備版			NTCIR-1 公式版			NTCIR-2 公式版		
	AP	WER	KER	AP	WER	KER	AP	WER	KER
テキスト	0.3877	—	—	0.3320	—	—	0.3118	—	—
男性 1（NTCIR）	0.3301	0.2123	0.2041	0.2609	0.1598	0.2120	0.2320	0.1605	0.2482
男性 2（NTCIR）	0.3145	0.2945	0.2789	0.2379	0.2228	0.2753	0.2119	0.2297	0.2956
女性 1（NTCIR）	0.3388	0.2055	0.2245	0.2116	0.1719	0.2690	0.1853	0.1841	0.2847
女性 2（NTCIR）	0.3507	0.1678	0.1429	0.2617	0.1380	0.2089	0.2213	0.1635	0.2555
男性 1（新聞）	0.1504	0.4658	0.5510	0.1030	0.3668	0.5759	0.0847	0.3918	0.5876
男性 2（新聞）	0.1536	0.6986	0.7278	0.1219	0.4529	0.6456	0.0512	0.5110	0.6606
女性 1（新聞）	0.1820	0.5514	0.6190	0.1213	0.3850	0.5854	0.0727	0.4021	0.5620
女性 2（新聞）	0.1803	0.4760	0.5374	0.1138	0.3341	0.5348	0.0941	0.3697	0.5803

表 4: 検索課題 0027（シソーラスを用いたテキストの連想検索について）の音声認識結果と平均適合率

入力	NTCIR モデル		新聞モデル	
男性 1	シソーラスを用いたテキストの連想検索について。	(0.6872)	市長ら屑を用いたテキストの連想を原作について。	(0.0109)
男性 2	シソーラスを用いたテキストの連鎖を検索について。	(0.2336)	起訴をら三落ちたテキストの連鎖を検索について。	(0.0073)
女性 1	シソーラスを用いた。エキスの連想検索について。	(0.6644)	小さな村落ちたが、って聞いても連想検索について。	(0.2491)
女性 2	シソーラスを用いた、テキストの連想検索について。	(0.6872)	市長ら数を用いた。テキストの連投検索について。	(0.0119)

[6] T. Kawahara, A. Lee, T. Kobayashi, K. Takeda, N. Minematsu, S. Sagayama, K. Itou, A. Ito, M. Yamamoto, A. Yamada, T. Utsuro, and K. Shikano. Free software toolkit for Japanese large vocabulary continuous speech recognition. In *Proceedings of the 6th International Conference on Spoken Language Processing*, pp. 476–479, 2000.

[7] National Center for Science Information Systems. *Proceedings of the 1st NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, 1999.

[8] National Institute of Informatics. *Proceedings of the 2nd NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization*, 2001.

[9] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 232–241, 1994.

[10] Tetsuya Sakai, Tsuyoshi Kitani, Yasushi Ogawa, Tetsuya Ishikawa, Haruo Kimoto, Ikuo Keshi, Jun Toyoura, Toshikazu Fukushima, Kunio Matsui, Yoshihiro Ueda, Takenobu Tokunaga, Hiroshi Tsuruoka, Hidekazu Nakawatase, Teru Agata, and Noriko Kando. BMIR-J2: A test collection for evaluation of japanese information retrieval systems. *ACM SIGIR FORUM*, Vol. 33, No. 1, pp. 13–17, 1999.

[11] Satoshi Sekine and Hitoshi Isahara. IREX: IR and IE evaluation project in Japanese. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, pp. 1475–1480, 2000.

[12] Páraic Sheridan, Martin Wechsler, and Peter Schäuble. Cross-language speech retrieval: Establishing a baseline performance. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 99–108, 1997.

[13] Amit Singhal and Fernando Pereira. Document expansion for speech retrieval. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 34–41, 1999.

[14] Savitha Srinivasan and Dragutin Petkovic. Phonetic confusion matrix based spoken document retrieval. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 81–87, 2000.

[15] Roberto S. Taylor. The process of asking questions. *American Documentation*, Vol. 13, No. 4, pp. 391–396, 1962.

[16] Martin Wechsler, Eugen Munteanu, and Peter Schäuble. New techniques for open-vocabulary spoken document retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 20–27, 1998.

[17] Steve Whittaker, Julia Hirschberg, John Choi, Don Hindle, Fernando Pereira, and Amit Singhal. SCAN: Designingand evaluating user interfaces to support retrieval from speech archives. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 26–33, 1999.

[18] 伊藤克亘, 山田篤, 天白成一, 山本俊一郎, 鵜寺義道, 宇津呂武仁, 山本幹雄, 鹿野清宏. 日本語ディクテーションのための言語資源・ツールの整備. 情報処理学会研究報告 99-SLP-26-5, 1999.

[19] 鹿野清宏, 伊藤克亘, 河原達也, 武田一哉, 山本幹雄（編）. 音声認識システム. オーム社, 2001.

[20] 松本裕治, 北内啓, 山下達雄, 平野善隆, 松田寛, 浅原正幸. 日本語形態素解析システム『茶筌』version 2.0 使用説明書. Technical Report NAIST-IS-TR99012, 奈良先端科学技術大学院大学, 1999.

音声言語データの構造化に基づく講演発表の自動要約

藤井 敦[†,†††] 伊藤 克亘[††,†††] 秋葉 友良[††] 石川 徹也[†]

[†] 図書館情報大学
〒 305-8550 つくば市春日 1-2
[††] 電子技術総合研究所
〒 305-8568 つくば市梅園 1-1-4
[†††] 科学技術振興事業団 CREST

E-mail: fujii@ulis.ac.jp

あらまし 本研究は，講演音声データの重要箇所を選択的に聴講することを目的とし，音声要約に関するモデルを提案して，計算機上のシステムとして実装した．本システムの特長は，講演とそれに対応する予稿論文の利点を併用する点にある．予稿論文は文書構造や表層情報に基づいて興味がある箇所を高速に探索する目的に適しており，講演は論文中で簡潔にまとめられている箇所について詳細に理解する目的に適している．本システムは，利用者が指定した論文中の章や節などの範囲に対応する説明内容を講演データから検索し，冗長性を排除しながら簡潔に提示する．開放的融合研究「話し言葉工学」プロジェクトの講演コーパスを用いた予備実験の結果，本システムの妥当性に関する見通しを得ることができた．

キーワード 音声要約，書き言葉と話し言葉の対応付け，話し言葉工学

A Structure-based Method for Speech Summarization

Atsushi Fujii[†,†††], Katunobu Itou[††,†††], Tomoyosi Akiba[††], Tetsuya Ishikawa[†]

[†] University of Library and Information Science
1-2 Kasuga, Tsukuba, 305-8550
[††] Electrotechnical Laboratory
1-1-4 Umezono, Tsukuba, 305-8568
[†††] CREST, Japan Science and Technology Corporation

E-mail: fujii@ulis.ac.jp

Abstract This paper proposes a model and system for speech summarization, aimed at selectively listening to specific contents in the entire lecture speech data. Our system uses both a target speech and its corresponding paper. Papers are used to identify contents where users are interested, based on structure/surface information. On the other hand, speech is effective to deeply understand specific contents. Thus, given a specific region in papers, such as a chapter and section, identified by the user, our system searches the lecture data for corresponding contents, and presents them based on a utility maximization method. To evaluate our system, we used a speech corpus produced by the spontaneous speech engineering project. Our preliminary experiments showed that our system was practical.

Keywords speech summarization, aligning text and speech, spontaneous speech engineering

1 はじめに

近年，情報通信技術が急速に発達し，大量の電子化情報を誰もが容易に受発信できるようになった．このような現状は，学会の運営形態にも影響を及ぼしている．論文誌や予稿集などの従来の冊子体刊行物は，オンライン/CD-ROM 版などの電子媒体に着実に移行しつつある．これらは主にテキスト，すなわち「書き言葉」である．

しかし，研究成果を公開するための手段には，口頭発表などの，いわゆる「話し言葉」もある．書き言葉と話し言葉は，一方が存在すれば他方は必要がないというような排他的な関係ではない．このことは，研究会などで予稿集と講演が両立している事実からも分かる．そこで，話し言葉の電子化を推進し，高度利用に関する研究を行うことは学術的・社会的に大きな意義がある．

我々は，開放的融合研究「話し言葉工学」プロジェクトで作成された日本語話し言葉コーパスの学会講演データ [5] を題材に，話し言葉情報の高度利用に関する萌芽的研究を行った．具体的には，口頭発表された内容の要点だけを掻い摘んで "聴講" することを目的とし，講演データの要約に関する手法を提案し，予備実験を行った．

以下，2 章で本研究の背景について方法論の観点から議論し，3 章で本研究の方法論および実装したシステムについて詳説する．そして，4 章でシステム評価に関する予備実験について説明する．

2 方法論に関する検討

2.1 先行研究

自然言語処理の研究分野では，書き言葉を対象とした要約に関する様々な手法が提案されている [2, 6]．これらの手法を応用し，さらに音声認識の誤り率を考慮した話し言葉の要約手法もある [9]．いずれの場合も，元の情報（テキストあるいは音声データ）内の相対的な重要度に基づいて，重要箇所を抽出する処理が中核をなす．

それに対して，我々は講演に対応する予稿論文を一種の外部情報として利用することを試みた．もしも予稿論文と講演が完全に等価な情報ならば，電子化された予稿論文を要約すれば十分なので，我々のアプローチは無意味である．しかし，1 章でも述べたように，予稿論文と講演は現実に両立しており，それ相応の理由があると考える．そこで，以下，人間の利用者にとっての予稿論文と講演の相違点について考察する．

2.2 予稿論文と講演の相違に関する考察

予稿論文は形態（冊子体/電子版）に拘わらず，章立てのような文書構造や文字種などの表層情報を手掛かりにして，いわゆる「斜めに読む」ことができる．また，何度も繰り返し読み返すことができる．そこで，全体の概要を把握したり，関心のある箇所を高速に探索することが比較的容易である．すなわち，ランダムアクセスに適している．

それに対して，講演は逐次アクセスが原則であり，ランダムアクセスには適さない．実際の講演では，内容がスライド単位などで構造化されていたとしても，論文のように戻ったり飛ばしたりして聞くことはできない．録音された講演の場合は，スライドの区切りに関する情報を付与しない限り，構造を手掛かりに早送りや巻戻しによって重要箇所を特定することは困難である．音声認識（ディクテーション）によって講演をテキストに正しく変換できたとしても，書き起こしされた話し言葉を読むことは，予稿論文を読む場合に比べて非常に負担が大きい．

他方において，情報の量は講演の方が相対的に多い．予稿論文は発表内容に対してページ数が厳しく制限されることがある．これは主に印刷コスト等の理由によるものなので電子版の論文には該当しない．しかし，現状では冊子体と電子版の論文が等価な形で共存するので，必然的に電子版も字数の制限を受ける．それに対して，講演は（予稿論文に比べれば）発表内容に適した長さであることが多い．

また，講演の方が理解しやすいことがある．発話内容が適度に冗長であったり，会話的な表現も使用される．熟練した講演者ならば，聴衆の反応に応じて難易度や説明の仕方を動的に調整することもある．あらかじめ書かれた予稿論文に対して講演する場合には（そうでない場合もある），論文執筆当時よりも新しくかつ正確な情報が補足されることもある．

2.3 本研究の方法論

2.2 節での議論から，他者の研究成果を電子媒体を介して "聴講" するためのある一つのモデルが成立する．すなわち，まず，予稿論文を読んで発表の概要を把握し，興味のある箇所を素早く特定する．そし

て，特定した箇所に関する説明だけを講演の中から選択的に聴講し，理解を深める．その結果，講演を全て聞かなくても，最小限のコストで必要な情報を取得できる．

仮に，論文の冒頭から最後まで順番に関連する講演内容を聴講しても，講演内の不要な部分や過度に冗長な部分が排除されるので，講演を最初から最後まで聞くよりは効率的であることが期待できる．

実際の（電子媒体でない）講演の場合は，講演を聞きながら予稿集を読むことが前提である．我々の聴講モデルは，予稿論文と講演の両方が電子化されたときに，はじめて成り立つ点に注意を要する．

以上は，構造がない（または欠落した）講演情報を，それに対応する予稿論文の文書構造に基づいて要約する計算モデルと見なすこともできる．我々は「話し言葉工学」プロジェクトの学会講演データを用いて，予稿論文の構造に基づく講演要約の実現に向けて研究を行った．

当該データには，4 件の学会講演に対して，波形データ，人手による書き起こし，書き起こしの形態素解析結果が収録されている．我々は，人手による書き起こしを（疑似的な）講演データとして利用した．やがては，波形データを音声認識システムによって自動的に書き起こし，講演として利用することが可能になるだろう．

収録された講演 4 件のうち，3 件に対応する予稿論文を独自に入手できたので，これら 3 件を研究対象とした．

3 予稿論文に基づく講演要約

3.1 システム概要

本研究で提案する講演要約システムの概要を図 1 に示す．システムの入力は，予稿論文とそれに対応する講演である．論文からは，利用者が指定した章や節などに基づいて対象範囲を抽出し，それに関連する講演部分（関連説明）を出力する．現在は，予稿論文を人手で SGML 形式に編集しており，章や節の文書構造に基づいて特定範囲を容易に抽出できるようにしている．

言い替えれば，システムのオンライン処理は，論文から抽出した範囲を問い合わせ（query）と見なして，講演から関連説明を検索することである．そこで，問い合わせと説明における語の頻度分布に基づいて関連度を計算し，関連度が高い説明を選択する．

そのためには，検索処理において，あらかじめ（オフラインで）索引付けを行うように，講演に対して何らかの加工を施して，検索効率を高めることが好ましい．具体的には，講演を意味のあるまとまり，すなわち説明ごとに分割してデータベース化する．

論文中では一つにまとまっている内容が，講演では離散した箇所で説明されることがある．そこで，単一の問い合わせに対して，関連する説明を尤度が高い順に複数出力する方が網羅性の点では好ましい．

しかし他方において，同じような説明を何度も出力することは冗長であり，非効率的である．そこで，以前出力した内容との重複を避けながら，最小限の情報によって利用者の情報要求を満足するための工夫を行う．

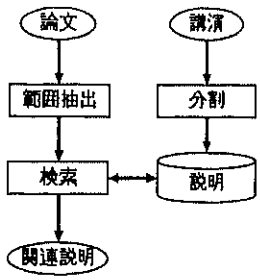

図 1: 講演要約システムの概要

3.2 講演データの分割

講演を意味のあるまとまりに分割することは容易ではない．2.2 節で議論したように，我々の研究は，講演は論文と同じようには構造化されていないことを前提としている．そこで，論文が持つ構造とは別の観点，すなわち話し言葉特有の性質に基づいて講演を分割する必要がある．

講演は自然な発話（spontaneous speech）に比べれば整っているものの，つなぎ語（filler）や言い直しなどがあり，話し言葉に近い特性を持つ [8]．事実，本研究で用いた講演データでも話し言葉特有の言語現象が観測された．

話し言葉に特有の言語現象のうち，つなぎ語は話題の転換に用いられることがあるため，講演の分割に有効な手掛かりになる．我々は，講演データを分析し，以下のつなぎ語によって講演を分割した．

あのー，あー，えー，まー，そのー，えーっと

これら全てが常に話題の展開に利用される訳ではない．また，文の切れ目だけではなく，文中に現れる

こともある．しかし，以上のつなぎ語によって講演を分割し，近接する断片を連結することで，ある程度意味のあるまとまりを構成できることが分かった．現在は，連続する3つの断片を連結して一つの説明としている．

3.3 関連説明の検索

3.2節で説明した処理によって，講演を複数の説明に分割することができる．そこで，各説明を異なる文書と見なせば，情報検索の分野で提案された各種手法を用いて，問い合わせに関連する説明を効率的に特定できる．

情報検索で典型的に使われる手法は，問い合わせと文書のそれぞれを索引語の統計頻度に関するベクトル（索引語の頻度分布）で表現する．そして，ベクトルの類似度に基づいて問い合わせと文書の関連度を定量化し，関連度が高い文書から優先的に出力する [1, 4]．一般には，特定の文書だけに現れる索引語や，問い合わせと対象文書の両方に共通して多く現れる索引語が重視される．

我々は，日本語形態素解析システム「茶筌」[7] を用いて問い合わせと説明を単語に分割し，品詞情報に基づいて名詞を索引語として抽出する．また，カタカナ語や新語は未知語と認定されることが多いので，未知語も索引語として抽出する．

関連度の計算には確率型の手法 [3] を用いた．これは，近年の情報検索手法の中でも比較的高い検索精度を実現することで知られている．具体的には，文書 i の関連度スコアを式 (1) によって計算する．

$$\sum_t \left(\frac{TF_{t,i}}{\frac{DL_i}{avglen} + TF_{t,i}} \cdot \log \frac{N}{DF_t} \right) \qquad (1)$$

ここで，$TF_{t,i}$ は索引語 t が文書 i に出現する頻度である．DF_t は t を含む文書数であり，N は総文書数である．DL_i は文書 i の長さ（文字数）であり，$avglen$ は平均文書長である．

3.4 効用最大化に基づく再帰的検索

論文中では一つの章や節としてまとまっている内容が，実際の講演では複数の箇所で分散して説明されることがある．そこで，単一の説明だけを出力するのではなく，3.3節の式 (1) で計算されるスコアが高い順に複数の説明を出力する必要がある．

しかし他方において，十分説明したにも拘わらず説明を出力し続けたり，同じような説明を何度も出力することは効果的ではない．一度提示した（利用者が聞いた）内容は次回以降なるべく検索しないようにすれば，少ない情報によって効用を最大化することが期待できる．

これを情報検索の観点から捉えると，利用者がある文書を閲覧した場合に，利用者の情報要求のうち，まだ満足していない部分を特定し，その部分に対して効果的な別の文書を提示する問題に置き換えることができる．

松村ら [7] は，利用者が文書 i を閲覧しても依然として満たされない情報要求を，問い合わせと文書 i との差分ベクトルとして表現し，次回以降の検索に利用する手法を提案した．ここで，問い合わせと各文書があらかじめ索引語のベクトルとして表現されていることが前提である（3.3節）．

我々は，この手法を応用し，問い合わせベクトルと一度検索された説明に対応するベクトルとの差分を取りながら再帰的に検索を繰り返すことで，冗長な説明を回避する．

また，検索を繰り返すたびに，問い合わせベクトルが縮退するので，式 (1) で計算されるスコアは次第に低下する．これは利用者の要求が次第に満たされていく過程をモデル化している．そこで，スコアに対する閾値を設定し，スコアが閾値を下回った時点で利用者の要求が十分満たされたと判断して処理を終了する．

現在は，スコアの閾値に対して決定的な値や一般的な範囲は分かっていない．データを用いて実験を繰り返しながら，経験的に設定する必要がある．

4 実験

3章で説明した講演要約システムの性能を評価するための観点として，講演分割の精度，問い合わせに対する検索精度，再帰的検索の効果，要約率などが挙げられる．しかし，現状では，これらのほとんど（要約率以外）が，客観的に評価することが困難である．また，講演データが3件しかないため，実験で得られた知見をどこまで一般化できるかについて議論することも難しい．そこで，講演データ3件に対して講演要約システムを実行し，結果を主観的に分析した．

3件のデータうち，1つは他の2つに比べると予

稿論文と講演の説明順序や内容がよく対応していた．この1件に対して，本システムは比較的質の高い要約を生成した．予稿論文全体に対して（最初から最後まで連続して）関連説明を検索したところ，元の講演データが文字単位で62%，説明単位で69%に要約された．他の2つの講演に関する分析は今後も継続して行う予定である．

明らかになった問題として，予稿論文と講演における表記の差異がある．まず，カタカナの異表記がある．これには，講演者（筆者）が論文と講演で意図的に異なる表記を使う場合もあれば，講演者は論文に書いた通りに発話しているつもりでも，書き起こしの段階で異表記になる場合も考えられる．

英数字の異表記（「T/ティー」,「3/三」など）もあったものの，これに対しては機械的な対処が可能である．しかし，カタカナ異表記は，今後解決しなければならない大きな問題である．

5 おわりに

本研究では，電子化された講演とそれに対応する予稿論文を用いて，他者の研究成果を効率的に聴講・理解するためのモデルを提案した．本モデルは，

- 予稿論文は，文書構造や表層情報に基づいて興味がある箇所を高速に探索することに適した媒体，
- 講演は，論文中で簡潔にまとめられている箇所について詳細に理解するための媒体，

という2つの仮説に基づいている．言い替えれば，予稿論文を外部情報として利用し，講演を要約するためのモデルである．

また，本モデルを計算機上のシステムとして実装した．まず，利用者が予稿論文中で特に関心がある箇所を特定し，その箇所に対してシステムが講演データから関連する説明を検索し，提示する．提示の段階では，単に複数の説明を列挙するのではなく，一度提示した説明との重複を避けながら最小限の説明によって効用を最大化するための手法を導入した．

「話し言葉工学」プロジェクトの学会講演データ3件を用いて実験を行った結果，予稿論文と講演の説明順序や内容が類似している場合は，本システムは比較的良質の要約を生成することが分かった．今後の研究課題として，カタカナや英数字の異表記に対処してシステムを改善する必要がある．また，より多くのデータに対して客観的にシステムの性能評価を行う必要がある．

謝辞

本研究の一部は日本学術振興会科学研究費補助金（基盤研究C，課題番号12680406）の助成による．

参考文献

[1] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.

[2] Inderjeet Mani and Mark T. Maybury, editors. *Advances in Automatic Text Summarization*. MIT Press, 1999.

[3] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 232–241, 1994.

[4] Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.

[5] 小磯花絵, 土屋菜穂子, 間淵洋子, 斉藤美紀, 籠宮隆之, 菊池英明, 前川喜久雄．「日本語話し言葉コーパス」の書き起こし基準について．電子情報通信学会技術研究報告 SP2000-97, pp. 55–60, 2000.

[6] 奥村学, 難波英嗣．テキスト自動要約技術の現状と課題．言語処理学会第4回年次大会ワークショップ論文集, pp. 80–87, 1998.

[7] 松村真宏, 大澤幸生, 谷内田正彦．AAS: 文書の組み合わせによってユーザの興味を満足する検索システム．人工知能学会誌, Vol. 14, No. 6, pp. 1177–1185, 1999.

[8] 河原達也．話し言葉音声認識の概観．電子情報通信学会技術研究報告 SP2000-95, pp. 1–5, 2000.

[9] 堀智織, 古井貞熙．係り受けSCFGに基づく音声自動要約法の改善．電子情報通信学会技術研究報告 SP2000-116, pp. 127–132, 2000.

[10] 松本裕治, 北内啓, 山下達雄, 平野善隆, 松田寛, 浅原正幸．日本語形態素解析システム『茶筌』version 2.0 使用説明書．Technical Report NAIST-IS-TR99012, 奈良先端科学技術大学院大学, 1999.