口語的な表現を含むテキスト情報の理解支援に関する研究

原田智彦

システム情報工学研究科
筑波大学

2015年3月

論文概要

博士 (工学)

口語的な表現を含むテキスト情報の理解支援に関する研究

筑波大学大学院
システム情報工学研究科
原田智彦

近年, インターネットの普及に加え, スマートフォンやタブレットなどのモバイル端末の性能向上や普及とあいまって, メールやソーシャルメディア (マイクロブログ,SNS, 動画投稿) などを主要なコミュニケーションツールとする利用者が急速に増加している. 本研究では, 問い合わせメールやソーシャルメディアなどの口語的な表現を多く含んだユーザーが発信するテキスト情報の正確な内容理解の支援を目的とし, メールとソーシャルメディアデータを扱う二つのビジネス活動に焦点を当て, 統計的な自然言語処理技術を用いた新たな方法を提案し, 評価実験によって提案の方法による効果を確認した.

一つ目は, 問い合わせメールに多く発生する省略の問題に着目し, 特に口語的な表現の多いテキストで見られる「A の B」タイプの名詞句における名詞 B の省略に焦点を当て, トピックモデルを知識として, 省略された語を予測する方法を提案した. 予め候補語集合が与えられた状況下で省略された語を選択する評価実験では, 提案の方法が, 従来方法に対し正解率で 11.34%の改善が見られ,75%を超える高い候補語の選択性能を示すことを確認した. これらの結果から, トピックモデルを利用することにより, 従来の言語モデルである N-gram モデルを補間して候補語の選択性能を向上させる効果があることを確認した.

二つ目は, ソーシャルメディアについて, キーワード検索で収集したツイート集合にキーワードと同名の別の対象がノイズとして含まれてしまう問題に着目し, ツイート中の表層的な情報とユーザー興味情報の両方を利用して, 検索結果から目的のツイートを選り分ける方法を提案した. ツイートの内容が分析対象に関するものかそれ以外かを判定する評価実験では, 提案の方法が, 従来方法に対し正解率で 8.52%の改善と, 見落とし率 (FN 率) で 16.53%の改善効果を確認した. これらの結果から, ツイート中の表層的な情報 (N-gram 言語モデル) とユーザーの潜在的な興味情報 (ユーザーごとのトピックモデル) の両方を利用することにより, ツイート中の表層的な情報だけを利用する従来方法では判定できないケースにおいて, ユーザーの潜在的な興味情報を組み合わせることで判定性を高める効果があることを確認した.

目 次

図目次

表目次

第1章　緒論

近年, インターネットの普及に加え, スマートフォンやタブレットなどのモバイル端末の性能向上や普及とあいまって, ソーシャルメディア (マイクロブログ,SNS, 動画投稿) の利用者が急速に増加している. 例えば, 世界的に展開する最大の SNS サービスを提供している Facebook の利用者は, 既に 10 億人を超えると言われている. 総務省の「情報通信白書」平成 23 年版 [1], 平成 24 年版 [2] では, パソコンや携帯電話に比べて, スマートフォンやタブレット端末の利用者の方が, ソーシャルメディアの利用率が高くなる傾向があり, 身近にいつでもアクセスできるスマートフォン等がさらに普及すれば, ソーシャルメディア利用はさらに広がる可能性があることが示されている.

また, これらソーシャルメディアの多くは,SNS(Facebook や mixi 等), ブログ (Ameba ブログ, Yahoo!ブログ等), LINE, Twitter, 掲示板, ミニブログ (前略プロフィール, リアル等) など文字 (テキスト情報) によるコミュニケーションを主たる目的としたものである. 総務省の「平成 26 年情報通信メディアの利用時間と情報行動に関する調査」[3] によると, コミュケーション手段としての利用時間 (一日平均) は, 平日では電子メール (以下「メール」と略す) が一番長く 26.0 分, 次いでソーシャルメディアが 15.5 分, 休日ではメールとソーシャルメディアが 20.9 分,20.7 分と拮抗しており, 平日と休日の両方で音声によるコミュニケーション手段である携帯電話や固定電話の利用時間を上回っており, テキスト情報を使ったメールやソーシャルメディアが, 現在の主要なコミュニケーション手段になっていることが分かる.

一方で, ソーシャルメディア上でユーザーが書き込むプロフィールやコメント等の構造化されていない非定型のテキスト情報をビジネスに活用する動きが現在既に進んでいる. 近年, これを急速に加速させている背景にビッグデータ活用の拡大がある. ビックデータ活用とは, データの利用者やそれを支援するサービスの提供者それぞれの観点によって捉え方が様々であるが, ここでは, 多種多量な

データを生成・収集・蓄積をリアルタイムで行い, このデータを分析することで未来の予測や異変の察知等を行い, 利用者の個々のニーズに即したサービスの提供, 業務の効率化や新サービスの創出に活かす取り組み [2] とする. ソーシャルメディアには, ユーザーの日々の生活体験, ユーザーが購入した商品やサービスの選択基準や購入後の感想などが書き込まれるため, 企業にとってはこれらユーザー発信のテキスト情報の中から自社のビジネスに役立ちそうな投稿情報を収集・分析することが重要になってきている.

しかし, 通常, メールやソーシャルメディアデータ中のテキスト情報は, 構造化されていないだけでなく, 装飾などもないプレイン・テキスト形式である. また, スマートフォンや携帯電話では, テンキーによる文字入力の不便さや画面の狭小性のために長文を避ける傾向がある. さらに, 若年層の利用者の多さや匿名性の高いサービスも含まれるため, 手紙や報告書などで通常使われるような文語的な文体ではなく, 新語や俗語, 文法誤りの混じった, くだけた表現を含んだ口語に近い文体が多く, 語句の省略も多い. これまでも, 自然言語処理技術を用いて, ウェブサイト上の口コミ情報を対象にしたマーケティング分析が行われてきたが, 従来の自然言語処理技術が新聞等の整った言語表現に対して比較的シンプルな言語解析が行われてきた経緯から, 口語的な言語表現の処理が難しく, 誤りが原因で内容理解に間違いや漏れが発生する課題がある. そこで, 本研究では, メールやソーシャルメディアなどの口語的な表現を多く含んだユーザーが発信するテキスト情報の正確な内容理解の支援を目的とし, メールとソーシャルメディアデータを扱う二つのビジネス活動に焦点を当て, 統計的な自然言語処理技術を用いた新たな方法を提案する.

まず一つ目は, ユーザーから企業などへの問い合わせメールに対する返信を行うカスタマーサポート業務に着目する. 昨今, 企業や自治体, 官公庁などにおいては, 自ら提供する製品やサービスについて, その内容や使い方に関する問い合わせをメールを使って行うことが一般化している. 企業などでは, メールを使って寄せられる問い合わせに対して, 迅速かつ適切に対応を行えば消費者や住民の信頼を得る機会になるが, 反対に適切な対応を怠ると信頼を失いかねない. 実際に, 不適切な対応がもとになり, インターネット上に悪い評判が広がって思わぬ対策が必要になるといった事例も発生しており, ますます慎重な対応が求められてい

る. そのため, 問い合わせメールのテキスト情報からユーザーの意図を正確に読み取ることが重要になっている. しかし, 前述のように, メールのテキスト情報は, 通常, プレーンテキスト形式であり, スマフォなどからの送信が多いことから, 口語的な文体でくだけた表現を多く含んでおり, 語句の省略も多い. もし, 問い合わせメールの対応者が経験のあるベテランの担当者であれば, 背景知識や経験にもとづく知識を手がかりに省略されている語を推論し, 暗黙的に内容を補足しながら意味を解釈することができるが, 配属されたばかりの新人の担当者では背景知識や経験の少ない読み手には, それが難しい. そこで, 本研究では, ユーザーからの問い合わせメール中に多く発生する省略の問題に着目し, 読み手によらない問い合わせテキストの正確な理解のため, 省略された言葉を推定する方法を提案する.

二つ目は, ユーザーが Twitter などのソーシャルメディアへ書き込んだプロフィールやコメントを収集・分析するソーシャルリスニング業務に着目する. ソーシャルリスニングとは, 人々がソーシャルメディア全体で日常的に語っている会話や自然な行動に関する投稿情報を収集・分析し, マーケティングや業務改善に活かす手法である. 前述のように, ビッグデータを収集・分析して社会問題の解決, マーケティング戦略立案や業務改善などのビジネスに活かす取り組みが急速に拡がっており, ソーシャルメディアがその情報源のひとつとして注目されている. 中でも, 代表的なソーシャルメディアである Twitter では, ユーザーは 140 文字以内のツイートと呼ばれるメッセージを使い, 日々の生活体験や思いを投稿できる. 投稿された情報は日常的に人から人へ伝わり, 多くのユーザーによってシェアされる. ツイートには, 購入した商品やサービスの選択基準や購入後の感想などが書き込まれるため, 企業にとっては, Twitter 上の投稿情報からマーケティングや商品開発に役立ちそうな投稿情報を収集・分析することの重要性が増している. しかし, 分析目的で収集したデータ中に, 多くの「ノイズ」がデータとして混入している問題がある. これらのノイズは, 分析の役に立たないだけでなく, 分析結果の精度に影響を与えるため, いかにノイズを除去するかが課題のひとつになっている. しかし, メールの場合と同様に, ツイート中のテキスト情報もプレーンテキスト形式であり, 新語や俗語, 文法誤りの混じった, くだけた表現を含んだ口語的な文体が多く, 140 文字以内という制限から 1 文が短く, 語句の省略も多い. そのため, 従来の方法でシステムによるは解析や処理が難しい. そこで, 本研究では, 分析目的

で収集したツイート中に分析に必要のない同じ名前の別の企業名や商品名が混入する問題(同名他社の問題)に着目し,ツイートの表層に現れる手がかりのみに依らずに,ノイズを含んだ検索結果から目的のツイートを選り分ける方法を提案する.

本論文は6章で構成される. 2章でまず口語的な表現を含むテキスト情報に起因する問題について説明し,その周辺にある関連研究について述べる. そして,本研究で使用する自然言語処理分野における先進的な手法について解説する. 次に3章では前章で述べた課題を予備実験を行い実験結果で確認する. 4章では本研究で提案する問い合わせメール中の省略を推定する方法を説明し,人工データを用いて行った評価実験の結果を示す. 5章では本研究で提案するツイート上の同名他社を判定する方法を説明し,行った評価実験の結果を示す. 最終の6章では結論を述べる.

第2章　口語的な表現を含むテキストの問題

本研究の目的は,メールやソーシャルメディアなどの口語的な表現を多く含んだユーザーが発信するテキスト情報の正確な内容理解を支援することにある. 本章では口語的な表現を含むテキストに起因する問題について説明し,関連研究を交えて解くべき課題と,本研究で使用する自然言語処理によるアプローチについて述べる.

2.1　省略の問題

本研究で扱う一つ目の問題は,ユーザーから企業などへの問い合わせメールに対する返信を行うカスタマーサポート業務に着目したものである.

通常,メールやソーシャルメディアデータ中のテキスト情報は,構造化されていない非定型のプレイン・テキスト形式であり,装飾などもない. また,スマートフォンや携帯電話の利用者が多いことから,テンキーによる文字入力の不便さや画面の狭小性のために長文を避ける傾向がある. さらに,メールやソーシャルメディアの利用者には若年層の利用者が多く,また匿名性の高いサービスも含まれるため,手紙や報告書などで通常使われるような固い表現を用いた書き言葉ではなく,くだけた表現を多く含んだ話し言葉に近い書き言葉が使われる. その結果,語句の省略も多い.

なお,現代の日本語において,書き言葉を「文語」,話し言葉を「口語」と呼ぶ場合があるが,本研究もそれに倣い,同じ書き言葉であっても,書籍や新聞のように固い表現で書かれた文体のものを「文語的」と呼び,一方でメールやソーシャルメディアの書き込みのように話し言葉に近い文体のものを「口語的」と呼ぶこととする.

国立国語研究所では,現代の日本語の書き言葉の全体像を把握するため,書籍,雑誌,新聞,白書,教科書などの11ジャンルから無作為抽出によるサンプリングで,バランス良く集めた約1億語のデータを格納した「現代日本語書き言葉均衡コーパス」(BCCWJ: Balanced Corpus of Contemporary Written Japanese) [4] を構築している. 11ジャンルには,インターネット上で使われる口語に近い書き言葉を収集するため,Yahoo! 知恵袋と Yahoo! ブログも含まれる. Yahoo! 知恵袋とは,質問したい人と回答したい人を結び,知恵と知識を参加者同士で共有することを目的として,2004年4月からヤフー株式会社がサービスを提供している日本国内における代表的な QA サイトである. そこで,本研究においても,口語的な表現を多く含んだ問い合わせテキストのコーパスとして Yahoo! 知恵袋を利用することとした.

Yahoo! 知恵袋のデータは,国立情報学研究所 (NII) が,ヤフー株式会社の契約に基づき研究者に対して提供している「Yahoo! 知恵袋研究機関提供用データ 国立情報学研究所 (NII) 提供版 ver2.1」(以下「Yahoo! 知恵袋データ」と略す) を使用した. Yahoo! 知恵袋データには,期間 2004/4/1 ∼2009/4/7 の質問 16,257,413 件と回答 50,053,894 件が収録されている. 図2.1にはデータ仕様 (レコード中のレイアウト) を,章末の表2.3および表2.4にはカテゴリー別の質問テキストの件数を示す.

カラム	項目	データ
1	質問番号	1424744544
2	カテゴリ名	Windows 全般
3	カテゴリパス	＞インターネット、PC と家電＞パソコン＞ Windows 全般
4	質問タイトル	Hz と Bit について教えてください。
5	質問本文	CPU には、Hz という単位と Bit という単位があります。…
6	質問者 ID	e21c239239317d4c5242919d029e142c3cc47463
7	付随回答の回答数	2
8	質問のステータス	27
9	質問投稿日	2009/3/31 23:59:55
10-25	質問解決日,投票制になった日,役に立つ質問に選択されてるかどうか,質問する際にかける知恵コイン,「BA にふさわしくない」に投票された数,総投票数,画像の枚数,モバイルフラグ,自動カテゴライズ使用可否,補足有無,お礼有無,補足内容,補足日付,お礼内容,お礼日付,お礼アイコン	

図 2.1: Yahoo! 知恵袋データのレイアウト

2.1.1 省略の問題と事例

図 2.2 には, Yahoo! 知恵袋に投稿された質問データの 1 つを例示した. また, 図 2.3 には, 図 2.2 に対して本研究で解決しようとする省略を補完した例を示した.

パソコンに DVD を入れてもディスクを読み取ってもらえなく、ドライブを見ても常に中身は空です。
しかし CD を聴くことは可能です。
こういう場合は修理に出したほうがいいのでしょうか?
すみませんが、ご回答よろしくお願いします。

図 2.2: Yahoo! 知恵袋の質問テキストの例

パソコンのドライブに DVD のディスクを入れても、ディスクのデータを読み取らず、ドライブの中を見ても常に中身は空です。
しかし、CD の音楽を聴くことは可能です。
こういう場合は修理に出したほうがいいのでしょうか？
すみませんが、ご回答よろしくお願いします。

図 2.3: 省略を補完した例

図 2.2 と図 2.3 を比べると, 下線部を補った図 2.2 の方が, 下線部周辺の単語や句の意味的な曖昧性が解消され, 意味をより正確に解釈できる印象を持つかも知れない. 通常, 人は, 与えられた言語表現に対して, 知識を適用し, その意味を解釈している [5] ため, 実は, その印象には個人差があると考えられる. 図 2.2 のようなテキストが与えられると, 読み手に背景知識や経験にもとづく知識があれば, それを手がかりに省略されている語を暗黙的に推論して補完しながら意味を解釈することができるが, 読み手に背景知識や経験の少なければ, それが難しい. 例えば, パソコン (以下「PC」と略す) やインターネットに関する問い合わせに回答する企業などのカスタマーサポート業務で, 図 2.2 のような問い合わせをメールで受

け付けた場合,メール返信作業に対応する担当者がベテランであれば,背景知識や経験を手がかりに省略されている語を暗黙的に推論して補完しながら読み進めることができるが,配属された新人のように背景知識や経験の少ない担当者にはそれが難しく,意図の取り違いや見落としの原因になる. 実際,企業などでは,このスキルのギャップを埋めようと教育コストを投じて対応する一方で,教育した担当者の離職率の高さも課題になっている. そのため,読み手によらない問い合わせテキストの正確な理解を支援することは極めて重要であると考える.

2.1.2 「AのB」タイプの連体修飾関係

次に,図2.2と図2.3の違いに見る現象について,日本語文法による解釈を加え,その周辺にある関連研究と課題について述べる.

最初に,日本語における文の組立ては「主語」,「述語」,「補足語」,「修飾語」の4つの要素からなる.「述語」は文の中心的な要素であり,述語の内容によって文の大枠が決まる.「補足語」は,述語が表す意味を補う働きをし,「修飾語」は与えられた表現に付加的な情報を加え,より精密な記述を与える働きをする文の要素である.「修飾語」はさらに述語を修飾する「連用修飾」と名詞を修飾する「連体修飾」に分かれる [6]. 日本語における名詞の修飾関係は多岐に渡るが,名詞の体言の概念を説明・限定するのが連体修飾である. また,体言の前に置かれて連体修飾の働きをする語が連体修飾語であり,連体修飾語の説明・限定を受ける語が被連体修飾語である [7].

図2.3で補完した下線部の要素は,前に置かれた名詞と接続して「AのB」の形をした連体修飾関係を持つ名詞句の一部である. 特に,図2.2と図2.3を比較すると「AのB」タイプの連体修飾関係における名詞Bの省略が多いことが分かる. 本来「AのB」タイプの連体修飾関係は,名詞Aが名詞Bを意味的に限定する関係で接続する最も一般的な形であり,その形が単純なため,文中に高い頻度で存在する [8]. 3章では,口語的な表現を多く含むYahoo! 知恵袋の質問テキストと,口語的でないテキストを多く含むWikipediaの日本語テキストの省略傾向を比較した予備実験とその結果を示すが,予備実験では口語的な表現を多く含む質問テキストの方が,図2.2と図2.3で示したような「AのB」タイプにおけるBの名詞の省略が多いことを確認している.

このような「A の B」タイプの連体修飾関係に関する研究は, その形が単純で多様な意味構造を持つことから「A の B」の意味解析に関する研究 [8] や「A の B」の言い換えに関する研究 [9] などに重点が置かれ, 本研究のように「A の B」の中の A あるいは B の省略を扱う研究はこれまで十分に行われてこなかったという課題がある.

2.1.3 省略や照応の解析

「A の B」タイプの連体修飾関係に焦点を当てたものではないが, 文中の省略や照応解析に関する研究が, 新聞記事のような文語的な文体を対象として盛んに行われてきた. これら省略や照応解析の周辺にある研究と課題について述べる.

最初に, 照応とは, 文中の語句 (先行詞) と語句 (照応詞) が同じ内容を指すことであり, 省略とはゼロ代名詞による照応を意味する. また, ゼロ代名詞とは, 文中に陽に表現されていないが, 文脈内の他の部分や発話の状況および知識により理解可能な必須格の名詞句を指す [10]. 日本語では文脈から予測できたり, 一度話題になるなどして言語化しなくても聞き手が理解可能な要素は省略されることが多く, 省略される要素も多岐に渡る [6]. このように日本語は省略が多い言語であることから, 照応解析は自然言語処理の重要なタスクの一つになっている.

省略や照応解析の研究の多くは文脈内の表層に現れる手がかりによって解決するものであり, 代表的な研究にセンタリング理論 [11] がある. センタリング理論は,「話題」が文脈内の「センター」に維持され,「センター」が照応や省略の先行詞になりやすいと仮定した考え方である. このセンタリング理論に基づく日本語の照応解析を扱った研究に Walker らの研究 [12] がある. Walker らは先行詞の候補を「主題＞主語＞間接目的語＞直接目的語＞その他」のように顕現性の高いものから並べ, その序列からゼロ代名詞の先行詞を決定する方法を提案している. しかし, センタリング理論では文中に複数の照応詞が存在する場合や前文に先行詞がない場合に適切な解析を行うことができない課題がある. また, 照応解析に関する近年の研究に林部らの研究 [13] がある. 林部らは, センタリング理論に基づく Walker らの方法では, 例えば「X を逮捕した」という文があった場合に, この文のみを手がかりにして同じ文脈内にある「自首した」の主語 (ガ格項) が X であると判定できないケースに着目した. そこで林部らは「格助詞 + 動詞」の

組み合わせを格構造と定義し,あらかじめ収集した格構造同士の間で計算した類似度を用いて,文脈内の前方に出現する格構造の履歴を照合する方法を提案し,従来方法より精度精度が向上することを示した. しかし,この方法でも文脈内に先行詞がない場合に適切な解析を行うことができない. なお,文脈内に明確に表現された先行詞がある照応を「文脈照応」と呼び,文脈内に明確に表現された先行詞がない照応を「外界照応」と呼ぶ.

一方で,文脈内に先行詞が存在しないケースに対応した研究に,清田らの研究[14]がある. 清田らは,換喩表現と換喩を解釈した表現で係り受け関係がずれることが,テキストベースの質問応答システムにおける検索文と対象テキストとのマッチングに影響を与えることから,これを解消するため,自動抽出した換喩表現を用いて係り受け関係のずれを解消する方法を提案した. 清田らの方法は,省略の解決を目的としたものではないが,文脈を超えて省略された語の候補を見つけることができている. なお,換喩とは比喩の一種であり,ある事物をそれと関連する別の事物に置き換えて表現する現象である. 例えば「漱石を読む」について「漱石」は「漱石の小説」を指し,同様に「電源を入れる」においては「電源」は「電源のスイッチ」を指している. 同様の例は,図2.2と図2.3の違いにも見つけることができ,例えば,図2.2「CDを聴く」の「CD」は,図2.3に示したように「CDの音楽」を指す.

換喩は「AのB」タイプの連体修飾関係において被修飾語Bが省略される現象の背景の一つと考えられる. 清田らの方法は,文脈を超えて省略された語の候補を見つけることができるため,本研究の目的と近いが,どの候補が適切かといった優先度について扱っていないという課題がある.

これまで述べたように,省略や照応解析に関する従来研究は,主として文語的な整ったテキストを対象とし,文脈内のどこかに先行詞が存在するケースの解析が中心であり,文脈内に明確な先行詞の存在しないケースを対象外としてきた. しかし,口語的な表現を含むくだけたテキストに対応するためには,文脈全体のどこにも先行詞が存在しない場合の解析に対応する必要がある. そこで,本研究では,この課題に対応し,文や文脈の表層に現れる手がかりのみに依らない解析を行うため,潜在的意味解析の一手法であるトピックモデルを利用する. 次節では,このトピックモデルについて解説する.

2.1.4 トピックモデルと LDA

これまで, 問い合わせメールに含まれる省略の問題をとり上げ, 文中や文脈に先行詞が存在しない場合の解析が必要であることを述べた. 本研究では, この課題に対応するため, 潜在的意味解析の一手法であるトピックモデルを利用する.

トピックモデル [15] は, 大規模かつ不均質な大量のテキストから, 知識を獲得するための統計的モデリング手法として, 近年, 自然言語処理の分野を中心として注目を集めている手法である. これまで, 文の外側にある情報 (大域情報と呼ぶ) を扱うことのできる言語モデルとしては, キャッシュモデルやトリガーモデル [16] が代表的なモデルであったが, これらモデルのように直接単語対単語の関係をモデル化するのではなく, トピックモデルは文書に隠れているトピックと単語との関係をモデル化する.

トピックモデルの基本的な考え方は, 文書中の単語は独立に出現しているのではなく, 潜在的なトピックに基づいて出現するというものである. トピックモデルは, 各文書を複数のトピックの確率分布として, 各トピックを複数の単語の確率分布として表す方法であり, 代表的なものに LDA (Latent Dirichlet Allocation; 潜在的ディリクレ配分法) [17] などがある. 本研究においても, 一つの文書に複数のトピックが存在することを想定し, マルチトピックモデルの一種である LDA を用いたモデル化を行う.

Blei[17] は文書のトピックを表す多項分布の事前分布としてディリクレ事前分布を導入した LDA を提案した. 近年, トピックモデルの有用性が注目されているが, その代表的なものとして LDA がよく機能することが知られている. LDA では一つの文書に対して複数のトピックが存在し, それぞれのトピックがある確率を持って文書上に生起するという考えに基づき, トピックの確率分布を推定する. 図 2.1 に LDA のグラフィカルモデルを示す.

グラフィカルモデルでは, 確率変数およびパラメータは頂点, それらの依存関係は有向辺で表現される. 網掛けの頂点は顕在変数, 他の頂点は潜在変数または未知パラメータを示している. 矩形は角に記された数だけ矩形内の変数の生成が繰り返されることを示している. D は文書数, K はトピック数, N_d は文書 d 内の総単語数を示している. θ と ϕ はそれぞれ文書ごとのトピックの多項分布パラメータとトピックごとの単語の多項分布パラメータである. α と β はそれぞれ θ と ϕ

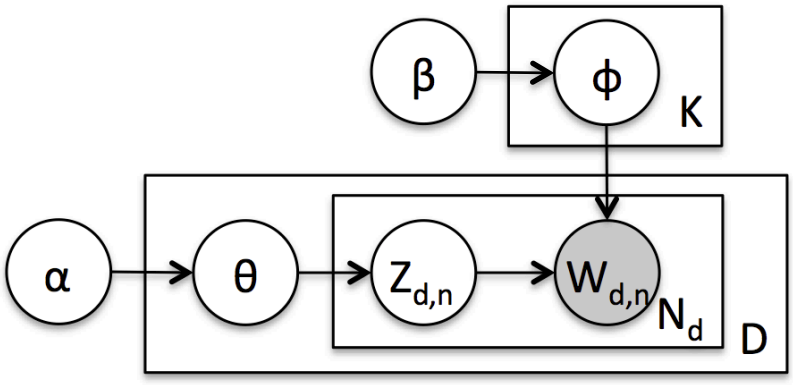

図 2.4: LDA のグラフィカルモデル

のディリクレ分布におけるハイパーパラメータである. このグラフィカルモデルに従った文書 $\boldsymbol{W}$ の生成過程は以下のような手順である.

1. 各トピック $k = 1, \cdots, K$ について:
   (a) ディリクレ分布に従って単語分布 ϕ_k を生成
   $\boldsymbol{\phi}_k \sim Dir(\beta)$
2. 各文書 $d = 1, \cdots, D$ について:
   (a) ディリクレ分布に従ってトピック分布 θ_d を生成
   $\boldsymbol{\theta}_d \sim Dir(\alpha)$
   (b) 文書 d における各単語 $w_{d,n}$ $(n = 1, \cdots, N_d)$ について:
      i. 多項分布に従ってトピックを生成
      $z_{d,n} \sim Multi(\boldsymbol{\theta}_d)$
      ii. 多項分布に従って単語を生成
      $w_{d,n} \sim Multi(\boldsymbol{\phi}_{z_{d,n}})$

なお, $\boldsymbol{\phi}_k$ はトピック k の単語分布, $\boldsymbol{\theta}_d$ は文書 d のトピック分布, $z_{d,n}$ は文書 d の n 番目の単語の潜在トピック, $w_{d,n}$ は文書 d の n 番目の単語を表し, $Dir(\cdot)$ はディリクレ分布, $Multi(\cdot)$ は多項分布を表す. LDA モデルによる全生成確率を式で表

すと式 (2.1) のようになる.

$$P(\boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{\phi}|\alpha, \beta) = \prod_{k=1}^{K} P(\boldsymbol{\phi}_k|\beta) \prod_{d=1}^{D} P(\boldsymbol{\theta}_d|\alpha) \prod_{n=1}^{N} P(z_{d,n}|\boldsymbol{\theta}_d) P(w_{d,n}|\boldsymbol{\phi}_{z_{d,n}}) \quad (2.1)$$

未知パラメータの推定には様々な方法が提案されているが, 本研究では十分な反復回数が得られれば高い精度でモデル推定が行えることで知られる Griffiths ら [18] の Collapsed Gibbs sampling を用いる. LDA のモデルには $\boldsymbol{\theta}$ と $\boldsymbol{\phi}$ が存在するがこれを積分によって消去し, $\boldsymbol{\theta}$ と $\boldsymbol{\phi}$ が現れない形でギプス・サンプリングの更新式を求めるのが Collapsed Gibbs sampling の考え方である. 式 (2.2) に Collapsed Gibbs sampling によるギプス・サンプリングの更新式を示す.

$$P(z_i = k|\boldsymbol{Z}_{\setminus i}, \boldsymbol{W}) \propto \frac{N_{dk\setminus i} + \alpha}{N_{d\setminus i} + \alpha K} \cdot \frac{N_{kw_i\setminus i} + \beta}{N_{k\setminus i} + \beta V} \quad (2.2)$$

ここで, z_i は文書 d における n 番目の単語のトピックを表し, $\boldsymbol{Z}_{\setminus i}$ はトピック集合 $\boldsymbol{Z}$ からトピック z_i を除いたものを表す. また, N_{dk} は文書 d におけるトピック k が割り当てられた単語数で, N_d はこれをすべての k について足し上げたものである. N_{kw} はトピック k における単語 w の出現回数で, N_k はこれをすべての w について足し上げたものである. どちらともに $\setminus i$ は, 文書 d の n 番目の単語を除いた時の数または回数を表す. ギプス・サンプリングによって得られたサンプルから, 各文書のトピック分布 $\boldsymbol{\theta}$ と各トピックの単語分布 $\boldsymbol{\phi}$ の予測分布を計算することができ, 文書 d においてトピック k が生成される確率の推定量 $\hat{\theta}_d^k$, トピック k が選択された場合の単語 w が生成される確率の推定量 $\hat{\phi}_k^w$ は, それぞれ式 (2.3), 式 (2.4) によって求められる.

$$\hat{\theta}_d^k = \frac{N_{dk} + \alpha}{N_d + \alpha K} \quad (2.3)$$

$$\hat{\phi}_k^w = \frac{N_{kw} + \beta}{N_k + \beta V} \quad (2.4)$$

本研究では,メールやソーシャルデータの特性である,文の短さや語の省略の多さに対応し,文や文脈の表層に現れる手がかりのみに依らない解析を行うため,潜在的意味解析の一手法であるトピックモデルの考えを導入し,LDA を利用する.

2.2 同名他社の問題

本研究で扱う二つ目の問題は,ユーザーが Twitter などのソーシャルメディアへ書き込んだプロフィールやコメントを収集・分析するソーシャルリスニング業務に着目したものである.

近年,ビッグデータ活用の急速な拡大とあいまって,その情報源のひとつとしてソーシャルメディアが注目されている. 企業などでは,ソーシャルメディア上でユーザーが書き込むプロフィールやコメント等のテキスト情報をマーケティングや業務改善に活かす取り組みを進めている. 代表的なソーシャルメディアが Twitter である.

Twitter は,ユーザーは 140 文字以内のツイートと呼ぶメッセージを使い,日々様々な意見や感想を投稿できるコミュニケーションツールである. ユーザーは他の複数のユーザーを「フォロー」することで「フォロワー」となり「フォロー」したユーザーのツイートを購読することができる. それにより,ユーザーの投稿したツイートは「フォロワー」関係を持つ複数のユーザーに配信される. また,ユーザーは他のユーザーのツイートを「リツイート」することで,タイムラインと呼ぶ自分のツイートの一覧上に他のユーザーのツイートを掲載したり,他のユーザーのツイートに「返信」することで,他のユーザーと対話的なコミュニケーションを行うこともできる. また,ユーザーは自分のツイートに「ハッシュタグ」を挿入することで話題を明示的にグループ化することができる. Twitter には速報性やリアルタイム性,利用者数,情報量などの顕著な特徴があるが,フォローやリツイート機能などの情報共有のための付加的な仕組みも備えているため,フォロー関係を元にした情報の流れやユーザー間の関係の分析が行えることも特徴である [19]. 一方で,Twitter などのソーシャルメディアデータを対象とした研究や分析には,共通して,検索結果や収集したデータ中に多くの分析の役に立たないデー

タが「ノイズ」として混入しているという問題がある. これらのノイズは, 分析の役に立たないだけでなく, 分析結果の精度に影響を与える.

2.2.1 同名他社の問題と事例

表 2.1 には,2014/1/4∼2014/1/11 の一週間に投稿された日本語ツイートの中から, Twitter API を使って, 試行的にキーワード「アップル」を含むツイートを収集し, 得られた 847 件を内容で分類し, 集計した結果を示した. もし, コンピュータやディジタル家電メーカーの「Apple Inc.」に関するツイートを収集しようとしてキーワード「アップル」で検索を行えば, 図 2.5 に示すような「アップルティー」や「アップルジュース」などのフルーツの意味で使われる「アップル」だけでなく,「アップル」を冠した別の企業名や商品 (以下「同名他社」と呼ぶ) なども混入したデータを収集することになる.

表 2.1: キーワード「アップル」を含むツイートの構成

内容	構成比
アップル (Apple Inc. の意味)	70%
アップルティー	4%
アップルジュース	2%
その他のアップル	24%
計	100%

例えば, 企業の評判分析を行う場合に, このように同じ名前の別の企業名が含まれたツイートを収集してしまうと, 投稿数の集計は不正確なものとなり, 分析精度が低下する要因となる. そのため, 収集したツイートから目的と無関係のツイートを「ノイズ」として除去することは極めて重要である.

しかし, メールの場合と同様に, ツイート中のテキストもプレイン・テキスト形式であり, 新語や俗語, 文法誤りの混じった, くだけた表現を含んだ口語的な文体が多く,140 文字以内という制限から 1 文が短く, 語句の省略も多い. そのため, テキストの中に出現する単語など表層的な情報のみを手がかりにした解析が難し

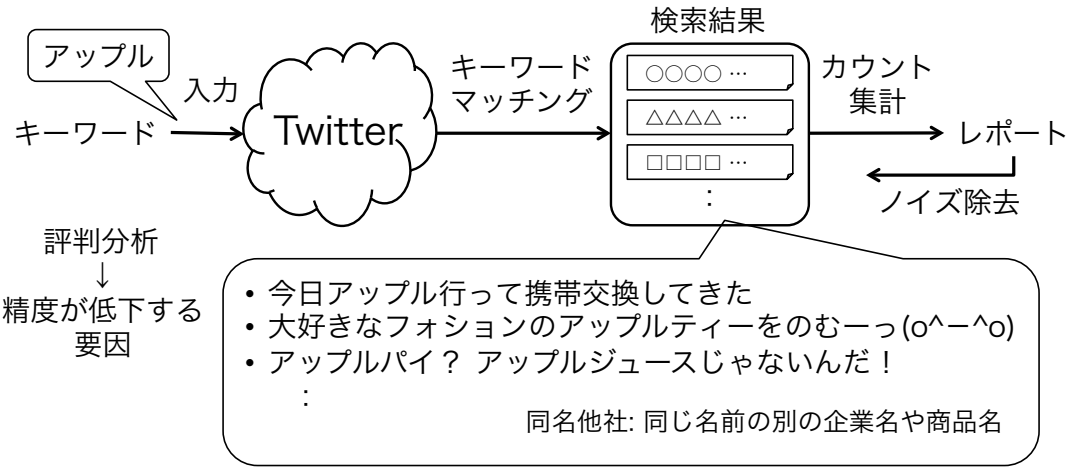

図 2.5: キーワード「アップル」を含むツイート検索のイメージ

いといった課題がある.

2.2.2 Twitter 解析と LDA の適用

Twitter の持つ速報性, リアルタイム性, 利用者数, 情報量などの顕著な特徴や, 情報伝搬・拡散やユーザー間の関係が分析できるトポロジー的な特徴を活用しようと, 多くの研究が行われており, 奥村 [20] の解説などが詳しい.

以下に Twitter の解析技術に関する主な研究テーマを示した. また, 引用した個々の研究に関する要約を奥村の解説に独自の調査を加えた上で章末の表 2.5 ～表 2.10 にまとめた.

「ニュース配信」「リアルタイムイベント検出」

リアルタイム性に注目した研究として, Twitter の情報を元にしたニュース配信の研究 [21] や, 地震や台風などの災害発生検出の研究 [22], インフルエンザなどの病気の流行把握の研究 [23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33] などがある.

「Twitter と実世界の関連性」

Twitter と実世界の関連性に注目し, ツイートの数と株式市場, 原油市場, 選挙結果, 興行収入など実世界で起きる出来事との関連を調査・分析した研究 [34, 35, 36] などがある.

「情報伝播の解析」「デマ拡散の解析」

情報の流れに注目した研究として, 情報伝播の解析 [37] やデマ拡散の解析 [38, 39, 40, 41, 42, 43, 44] の研究などがある.

「東日本大震災後の Twitter の利用動向の分析」

2011 年 3 月 11 日に発生した東日本大震災以降, Twitter の利用動向や交換されているツイートの内容, 情報伝播や拡散の状況などを分析する研究 [45, 46, 47, 48] が盛んに行われている.

「ユーザー属性・影響力推定」「評判・感情分析」「トレンド分析」

マーケティング分析に関連した研究として, キャンペーンやプロモーションのターゲティングに有用なユーザーのデモグラフィック情報 (「性別」「年齢」「居住地」「宗教」「政治的意識」「民族」など) をツイートから抽出するためのユーザー属性推定の研究 [49, 50, 51, 52, 53, 54, 55, 56, 57] や, 他のユーザーの発言に影響を与えるユーザー (「インフルエンサー」や「オーソリティ」と呼ばれる) の発見のためのユーザー影響力推定の研究 [58, 59, 60] などがある.

また, 商品開発などで有用な「ポジティブ」「ネガティブ」「中立」かといったユーザーの評価によってツイートを分類する評判分析 [61, 62, 63, 64, 65] の研究や, 発言に表れるユーザーの感情によってツイートを分類する感情分析 [66] の研究, 流行把握のためのトレンド分析の研究 [67, 68, 69] などがある

「トピック推定」「話題分類」「自動要約」

発言内容に注目した研究として, トピック (=ユーザーの関心) を推定する研究 [70, 71, 72, 73] や, 発言内容に応じてツイートを分類する話題分類の研究 [74, 75], ツイートの集合から発言の要約を行う自動要約の研究 [76, 77, 78] などがある.

「信頼性評価」「スパム・ボット判定」

ツイートの内容が信頼できるかどうかを解析する信頼性評価の研究 [79] や, ツイートがスパムかどうか, 人間かボットかどを判定する研究 [80, 81, 82, 83, 84, 85] などがある.

近年, Twitter に 2.1.4 節で説明したトピックモデルや LDA を適用した研究が数を増しており, 様々なトピックモデルが提案されている. 例えば, Paul と Dredze[25] は, 病気の流行を早期に把握するため, トピックモデルを用いた教師あり機械学習による分類器を使って Twitter の内容が病気に関係するかどうかを判別した. 彼らは疾患をある種のトピックと見立てて, 健康関連のツイートから LDA を用いて推定したトピックモデルを使用している. また, Pennacchiotti と Popescu[54] は, ツイートの情報とソーシャルグラフを用い, 政治的な指向や民族, スターバックスコーヒーへの親近感を推定する手法を提案した. 彼らは LDA を用いて推定したトピックモデルをツイート情報によるユーザー分類モデルに使用している. また, Weng ら [60] は, 影響力の強いユーザーを発見するため, ソーシャルグラフを用い, PageRank アルゴリズムを拡張した TwitterRank を提案している. 彼らは, LDA を用いて推定したトピックごとにユーザーのネットワークを構築し, そのネットワークにランキングアルゴリズムを適用した.

また, トピックモデルや LDA を拡張した研究も多く報告されている. 例えば, Ramage ら [71] は, ハッシュタグなどツイートについている情報を教師情報として利用できるように拡張した Labeled LDA を提案し, 通常の LDA を上回る性能を示した. しかし, ハッシュタグのないツイートなども多く使用できる範囲は限定される. 通常, ツイートは手紙や報告書などに比べて短いため, LDA などの一般的なトピックモデルでは十分に意味を捉えることができないため, LDA をツイッターに対して適用する場合, 1 ツイートを 1 文書とせず, 著者トピックモデル [86] の考えのもとユーザーの全ツイートを 1 文書として扱う方法も用いられている. これに対して, Zhao ら [72] は, 1 ツイートが 1 トピックであるという仮説を元に Twitter-LDA モデルを提案し, ツイートの短さによってトピックモデルが適切に推定できない問題を解消し, 前者のモデルと比べて優れていることを示している. また, 佐々木ら [87] は, ユーザーの興味が日々変化することに対し, Twitter-LDA は従来の LDA と同様にツイートされる時間的な順序を考慮できない点に注目し, 岩田ら [88] が時間発展する購買履歴データのためのトピック追跡モデルとして提案した Topic Tracking Model の機構を Twitter-LDA に加え, ユーザーの興味と話題の時間発展を効率的にモデル化できる方法を提案している. その他, Diao ら [73] は, ユーザーごとのトピック分布に加え, タイムスタンプごとのトピック分布を考

慮した TimeUserLDA を提案し突発的なイベントに関するトピックの抽出を行っている.

本研究では,Twitter データの特性である, 文の短さや語の省略の多さに対応し, 文や文脈の表層に現れる手がかりのみに依らない解析を行うため, 著者トピックモデルの考えに倣いユーザーの過去のツイートで推定したユーザーの興味モデル (トピックモデル) を利用する. その際, 佐々木らが指摘したようにユーザーの興味が日々変化すること (ダイナミクス) を考慮する方法についても検討する.

一方で, 本研究は, ツイート集合に含まれるノイズの問題を扱っているという点では, スパム・ボット判定の研究に近いという見方もできる. しかし, スパム・ボット判定の研究がスパムかどうか [80, 81, 83, 84, 85] や人間かボットか [82] の判定を行うのに対し, 同名他社の問題がいずれも人間が投稿した意味あるメッセージを対象にした判定を行うという点に違いがある.

2.3 まとめ

本章では, 口語的な表現を含むテキストに起因する二つの問題について説明し, 関連研究を交えて解くべき課題と本研究で使用する自然言語処理によるアプローチについて述べた.

本研究テーマと主な関連研究との関係を表 2.2 に示す.

一つ目の問題は, ユーザーから企業などへの問い合わせメールに対する返信を行うカスタマーサポート業務に着目したもので, ユーザーからの問い合わせメール中に多く発生する語の省略が, 背景知識や経験の少ない担当者とっては, 意図の取り違いや見落としの原因になることを説明した. また, 語の省略が「A の B」タイプの連体修飾関係中に多いことを指摘した. なお,「A の B」タイプの連体修飾関係における省略の傾向については, 予備実験によっても確認している. 予備実験の方法や結果については,3 章で説明する.

次に, 周辺研究に関しては,「A の B」タイプの連体修飾関係に関する従来研究は,「A の B」の意味解析や言い換えに関する研究などに重点が置かれ, 本研究のように「A の B」の中の A あるいは B の省略を扱う研究はこれまで十分に行われていないという課題がある. 一方, 省略や照応解析に関する従来研究は, 主として

表 2.2: 本研究のテーマと関連研究の整理

本研究のテーマ	省略の問題			同名他社の問題
研究領域	「AのB」タイプの連体修飾関係	省略や照応の解析	質問応答システム	トピックモデルや LDA
関連研究	「AのB」の意味解析や言い換えに関する研究	文中の省略や照応解析に関する研究	暗喩による係り受け関係のずれを解消する研究	Twitter への適用に関する研究
	森山ら [8], 片岡ら [9] など	Walker ら [12], 林部ら [13] など	清田ら [14]	Steyvers ら [86], Zhao ら [72] など
本研究との関連性	× 口語的な表現を扱っていない ×「AのB」の中の省略の補完については扱っていない	× 新聞記事などのきれいな日本語が対象 × 文脈中に先行詞がない場合に正しく解析できない	× 文脈中に先行詞がなくて対処できる × しかし, 候補語の選択や優先度について扱っていない	○ ユーザーの全ツイートを 1 文書として考える

文語的な整ったテキストを対象としていること, 文脈中のどこかに先行詞が存在するケースの解析が中心であり, 文脈内に明確な先行詞の存在しないケースを対象外としてきたという課題がある. しかし, 口語的な表現を含むくだけたテキストに対応するためには, 文脈全体のどこにも先行詞が存在しない場合の解析に対応する必要がある. ただし, 省略や照応解析を目的とした研究ではないが, 換喩を扱った清田らの方法は, 文脈中に先行詞のないケースに対応する上で参考になる点が多い.

本研究では, 背景知識や経験など読み手のスキルによらない, 問い合わせテキストの正確な理解を支援するため, 省略された言葉を補完する方法を提案する. その際, 文や文脈中に先行詞のない省略に対応するため, トピックモデルを利用し, 大域情報を使うことで, 表層的な手がかりのみに依らない方法を検討する. この後の 4 章では, トピックモデルを利用した, 省略された語を補完する具体的な方法について検討する.

二つ目の問題は, ユーザーが Twitter へ書き込んだプロフィールやコメントを収集・分析するソーシャルリスニング業務に着目したもので, 分析目的で収集したツ

イート中に分析に必要のない同じ名前の別の企業名や商品名 (同名他社) が混入することで, 分析精度を低下する要因になることを説明した. また, 一方で Twitter は,140 文字以内という制限から 1 文が短く, 語句の省略も多いため, テキスト中に出現する単語など表層的な情報のみを手がかりにした解析が難しいことを指摘した.

次に, 周辺研究に関しては, Twitter 解析に関する研究が多く報告されているが, ノイズに関する従来研究は, スパムやボット対策に重点が置かれ, 本研究のように同名他社の問題や多義性に対処する研究は十分に行われていないという課題がある. 一方, 近年, トピックモデルや LDA を適用した研究が数を増しており, Twitter を対象とした様々なトピックモデルが提案されていることを確認した. 中でも,1 文が短いツイートに LDA 適用する場合, 1 ツイートを 1 文書とせず, ユーザーの全ツイートを 1 文書として扱う方法が用いられており, 表層的な手がかりのみに依らない方法を検討する上で参考になる点が多い.

本研究では, ノイズを含んだ検索結果から目的のツイートを選り分ける方法を検討する. その際, 著者トピックモデルの考えに倣い, ユーザーの過去のツイートで推定したユーザーの興味モデル (トピックモデル) を利用することで, ツイート中に現れる表層的な手がかりのみに依らない方法を提案する. この後の 5 章では, 表層的な情報に加え, ユーザーの興味モデルを利用して, 対象のツイートが目的のツイートかどうかを判定する具体的な方法について検討する.

表 2.3: Yahoo! 知恵袋データ中の質問件数 (月別)

	収録期間: 2004 年 4 月 ~2009 年 3 月					
カテゴリー	2004 年	2005 年	2006 年	2007 年	2008 年	2009 年
エンターテインメントと趣味	133,181	272,482	257,003	416,476	1,040,875	428,326
暮らしと生活ガイド	101,089	206,616	255,545	340,745	588,623	188,247
インターネット、PC と家電	111,543	234,562	218,588	350,687	539,732	185,433
健康、美容とファッション	113,766	228,966	209,799	285,594	550,679	171,560
スポーツ、アウトドア、車	76,076	179,398	198,281	269,571	487,855	161,136
教養と学問、サイエンス	94,004	174,159	185,845	237,849	412,874	130,837
生き方と恋愛、人間関係の悩み	63,753	156,153	99,790	211,719	514,483	157,460
Yahoo! JAPAN	167,421	260,043	103,939	145,613	186,662	54,571
子育てと学校	47,027	126,394	96,330	144,246	285,082	97,366
地域、旅行、お出かけ	40,607	92,577	95,668	141,692	276,976	99,522
ニュース、政治、国際情勢	31,367	94,204	82,284	132,335	209,302	58,456
ビジネス、経済とお金	30,650	70,033	66,132	106,450	194,765	64,485
職業とキャリア	26,813	57,623	52,959	93,235	180,774	61,227
その他	31,566	52,089	43,884	64,821	131,379	43,209
マナー、冠婚葬祭	23,684	47,609	47,408	67,364	116,002	35,567
コンピュータテクノロジー	107	1,547	9,571	19,053	37,027	10,503
合計	1,092,654	2,254,455	2,023,026	3,027,450	5,753,090	1,947,905

表 2.4: Yahoo! 知恵袋データ中の質問件数 (合計と構成比)

	収録期間: 2004 年 4 月 ~2009 年 3 月	
カテゴリー	計	構成比
エンターテインメントと趣味	2,548,343	15.83%
暮らしと生活ガイド	1,680,865	10.44%
インターネット、PC と家電	1,640,545	10.19%
健康、美容とファッション	1,560,364	9.69%
スポーツ、アウトドア、車	1,372,317	8.52%
教養と学問、サイエンス	1,235,568	7.68%
生き方と恋愛、人間関係の悩み	1,203,358	7.47%
Yahoo! JAPAN	918,249	5.70%
子育てと学校	796,445	4.95%
地域、旅行、お出かけ	747,042	4.64%
ニュース、政治、国際情勢	607,948	3.78%
ビジネス、経済とお金	532,515	3.31%
職業とキャリア	472,631	2.94%
その他	366,948	2.28%
マナー、冠婚葬祭	337,634	2.10%
コンピュータテクノロジー	77,808	0.48%
合計	16,098,580	100.00%

表 2.5: Twitter を対象とした主な研究 (1/6)

リアルタイムイベント検出	ニュース配信	Sankarana-rayanan ら 2009 [21]	Twitter に基づくニュースの配信システムを構築するため,bag-of-words を素性とした機械学習によりニュースとノイズのフィルタリングを行った
	災害発生の検出	Sakaki ら 2010 [22]	ツイートの内容が地震や台風に関するイベントかどうかを判定し, 地震の震源や台風の進路を推定する手法を提案した
	病気の流行把握	Culotta 2010 [23]	「風邪」や「インフルエンザ」などキーワードとなる単語を経験的選択し, 単語頻度を集計して分類した
		Lampos と Cristianini 2010 [24]	L1 正則化 (LASSO) を用いて単語の次元を圧縮し, 単語頻度の和を集計して分類した
		Paul と Dredze 2011 [25]	疾患をある種のトピックと見立てトピックモデル (LDA) を用いた教師あり機械学習による分類器を使って Twitter の内容が病気に関係するかどうかを判別する手法を提案した
		Aramaki ら 2011 [26]	「風邪」や「インフルエンザ」などキーワードとなる単語を経験的選択し, 単語頻度を集計し,SVM で発言を分類した
		谷田ら 2011 [27]	素性選択の手法 (MMR) を利用し, 単語頻度の線形和を集計して分類した
		Aramaki ら 2012 [28]	「風邪」や「インフルエンザ」などキーワードとなる単語を経験的選択し, 単語頻度と疾患モデルで集計し,SVM で発言を分類した
	株価予測	Bar-Haim ら 2011 [29]	ユーザーが専門家か非専門家を判別し, 専門家のツイート集合のみで株価との関係を学習した分類器を作成した
	その他	Zhao ら 2007 [30]	各文書のトピックの頻度を時間的に分析することでイベントを検出する手法を提案した
		Weng と Lee 2011 [31]	クラスタリングでなく,Wavelet 分析を用いたイベント検出の手法を提案した
		Lee と Sumiya 2011 [32]	ツイートが持つ位置情報に k-means クラスタリングを適用することで, 非日常的な混雑をローカルイベントとして検出する手法を提案した
		渡辺ら 2011 [33]	位置情報のないツイートの発信位置を推定し, ローカルなイベントの検出を可能にする手法を提案した

表 2.6: Twitter を対象とした主な研究 (2/6)

Twitter と実世界の関連性	感情と出来事の関係	Bollen ら 2011 [34]	ツイート中の感情を,POMS(Profile of Mood States) を元にした 6 つの感情 (「緊張」「抑うつ」「怒り」「活気」「疲労」「混乱」) について分析した結果と, 株式市場, 原油市況, 主要な出来事との関係を調査した
	LIWC と選挙結果の関係	Tumasjan ら 2010 [35]	LIWC(Linguistic Inquiry and Word Count) と呼ばれる分析ツールを用いて, 政党や政治家を参照しているツイートを分析し, 参照しているツイート数が選挙結果に反映していることを報告した
	評判と興行収入の関係	Asur と Huberman 2010 [36]	よく語られている映画はよく見られているといった評判情報が映画の興行収入を予測するのに利用できることを示した
情報伝播の解析		Kwak ら 2010 [37]	Twitter のトポロジカルな特徴に注目した情報伝播の解析した
デマ拡散の解析		Ratkiewicz ら 2011 [38]	米国の選挙に関連して, 一般市民を装った特定の候補者を支持や批判, 誹謗中傷, 誤情報の意図的な流布を行っているツイートを検出するシステムを提案した
		Qazvinian ら 2011 [39]	誤情報に関連するツイート群に対し, 誤情報に関して言及の有無で分類し, 言及しているツイートを支持または否定に分類する手法を提案した
		藤川ら 2011 [40]	ツイートに対して疑っているユーザーがどの程度いるか, 根拠付きでデマだと反論されているか等のユーザーの反応を分類することで情報の真偽判断を支援する手法を提案した
		鳥海ら 2012 [41]	ツイートの内容後と「デマ」「噂」「誤報」などの反論を表す語の共起度を調べる手法を提案した
		白井ら 2012 [42]	デマ情報と訂正情報を病気とみなし, 感染症疾患の伝染モデルを拡張することで, デマ情報と訂正情報の拡散をモデル化した
		梅島ら 2012 [43]	「デマ」や「間違い」といった訂正を明示する表現を用いることが訂正ツイートの認識に有用であることを示した
		鍋島ら 2013 [44]	誤情報を訂正する表現に着目し, 誤情報を自動的に収集する手法を提案した

表 2.7: Twitter を対象とした主な研究 (3/6)

東日本大震後の Twitter の利用動向の分析	内容の分析	Doan ら 2011 [45]	大震災後のツイートの中で地震, 津波, 放射能, 心配に関するキーワードが多くつぶやかれたと報告した
		Acar と Mraki 2011 [46]	震災後に Twitter で交換された情報の内容を「警告」「救助要請」「状況報告 (安否, 周りの状況, 心配)」に分類した
	利用状況の分析	Sakaki ら 2011 [47]	地震や計画停電などが発生したときの Twitter の地域別利用状況を分析して報告した
	情報の伝搬・拡散状況の分析	宮部ら 2011 [48]	震災発生後の Twitter の地域別利用動向, 情報の伝搬・拡散状況を分析して報告した
ユーザー属性推定	地理	Cheng ら 2010 [49]	ツイート中の各単語の位置との条件付き確率を元にした確率モデルを用いて, 市レベルでユーザーの位置を推定する手法を提案した
		Eisenstein ら 2010 [50]	特定の位置との結びつきの強い単語が存在するというアイデアを元にし, 潜在トピックと地理的地域を一緒に推論する多レベルの生成モデルを用いた手法を提案した
	性別, 年齢, 宗教, 政治, 民族など	Wing と Jason 2011 [51]	Wikipedia や Twitter の内容を基にその場所を推定する研究を行い,Twitter に適用した場合, 場所推定の平均誤差は 967km であると報告した
		Rao ら 2010 [52]	単語 N-gram, フォロワー数, リツイートの頻度などを素性とした教師あり学習による分類器を用いて, 性別, 年齢, 宗教, 政治的指向の 4 つの属性を分類した
		Burger ら 2011 [53]	単語/文字 N-gram を素性とした教師あり学習による分類器を用いて性別の推定を行い, ツイートの量やプロファイルを記述したメタデータの有無が性能に影響することを報告した
		Pannacchiotti と Popescu 2011 [54]	ツイートの情報と, ユーザー間のリンクで構成されたグラフ中のクラスラベルの分布の情報を利用した分類モデルを提案し, 政治的指向, 民族, スターバックスへの親近感を推定する手法を提案した (ツイートの情報によるユーザー分類に LDA を元にしたトピックモデルを利用)
	その他	Mislove ら 2010 [55]	ソーシャルネットワークをグラフとして用いて, ユーザー属性を推定した

表 2.8: Twitter を対象とした主な研究 (4/6)

		池田ら 2012 [56]	自動構築したキーワードリストを使い, キーワードの出現傾向を SVM で学習した分類器を用いてユーザーごとにプロフィールを推定する手法を提案した
		蔵内ら 2013 [57]	マルコフ確率場を用いてソーシャルグラフ上のユーザー属性をモデル化し, 最適化問題として属性を推定する手法を提案
ユーザー影響力推定		Cha ら 2010 [58]	フォロワー数, リツイート数, 他のユーザーとの会話数などとの関係から, ユーザー影響力を導き出せることを指摘した
		Duan ら 2010 [59]	ツイートの長さ,URL リンクの有無, ユーザー間のフォローする/されるという関係を用いたツイートのランキング手法を提案した
		Weng ら 2010 [60]	ユーザー影響力を推定するため, ソーシャルグラフとして用い, PageRank の拡張である TwitterRank を提案 (LDA を用いてユーザーから推定したトピックごとにユーザーのネットワークを構築した)
意見マイニング	評判分析	Davidov 2010 [61]	特定のハッシュタグや顔文字集合をラベルと対応付けたものを利用し, 教師あり学習による評判分類の手法を提案した
		Barbosa と Feng 2010 [62]	ツイートを対象にした既存の 3 つの評判分類器の出力をラベルとして利用し, 教師あり学習による評判分類の手法を提案した
		Silva ら 2011 [63]	Self-Training により訓練データ自体を最新のものに更新しながら評判分類器を学習する手法を示した
		Brody と Diakopoulos 2011 [64]	単語の長音化が極性 (positive/negative) を含む単語の検出に寄与することを示した
		Wang ら 2011 [65]	ハッシュタグ間の共起関係やハッシュタグ自体の意味 (字義) が有用であることを示し, この 2 つの情報を利用するため, グラフに基づく評判分類モデルを提案した
	感情分析	Go ら 2009 [66]	ツイート内の感情に関する語や顔文字から特徴量を求めることで, ツイートのポジティブ・ネガティブ分類を行った
トレンド分析	バースト検知	Kleinberg 2003 [67]	キーワードの急激な出現頻度の増加 (バースト) により流行の話題を検出できることを示した

表 2.9: Twitter を対象とした主な研究 (5/6)

	蝦名ら 2010 [68]	ドキュメントの発生ごとにバーストを解析し, 短時間に大量のドキュメントが発生した場合でも高速性を保つアルゴリズムを提案した
	Mathioudakis と Koudas 2010 [69]	Twitter のツイートログからバーストキーワードを抽出し, 共起性の高いバーストキーワード同士を集約することにより,Twitter 上のトレンドを提示する手法を提案した
トピック抽出	Ramage ら 2010 [71]	ハッシュタグなどツイートについている情報を教師情報として利用できるように拡張した Labeled LDA を提案し, 通常の LDA を上回る性能を示した
	Pennacchiotti と Gurumurthy 2011 [70]	LDA を用いてユーザーをトピックの混合として表現し, 類似のユーザーを提示する推薦の手法を示した
	Zhao ら 2011 [72]	ユーザーが 1 つのツイートで 1 つのトピックについてのみ言及するという仮定に基づき, 各ユーザーのツイートをまとめて 1 文書とモデル化した Twitter-LDA を提案した
	Diao ら 2012 [73]	Twitter のユーザーごとのトピック分布に加え, タイムスタンプごとのトピック分布を考慮した TimeUserLDA を提案した
	佐々木 ら 2013 [87]	Twitter-LDA が従来の LDA と同様にツイートされる時間的な順序を考慮できない点に注目し, Twitter-LDA に岩田らの Topic Tracking Model の機構を加え, ユーザーの興味と話題の時間発展を効率的にモデル化できる方法を提案した
話題分類	Sriram ら 2010 [74]	ユーザプロフィールに加え,時事イベントや, 感情語などの 8 つの特徴量を用いることで, 高い精度でツイートをニュース, イベント, 意見, 詳細. プライベートメッセージの 5 つに分類できることを示した
	西田ら 2011 [75]	ツイートの圧縮され易さを応用し, 着目する話題に関するツイートか否かを分類する手法を提案した
自動要約	Sharifi ら 2010 [76]	流行っている句を含んだツイート集合から, その句を包含する最頻出の句を抽出することでツイート集合の要約を行う手法を示した
	Takamura ら 2011 [77]	時系列に並んだ文書としてのツイート系列から, 重要なツイートを選択する要約モデルを提案した

表 2.10: Twitter を対象とした主な研究 (6/6)

		Liu ら 2011 [78]	ツイートからリンクされた Web コンテンツも利用し, 整数線形計画によって概念に基づく最適化手法を用いた要約モデルを提案した
信頼性評価		Castillo ら 2011 [79]	ツイートの内容が信頼できるかどうかを判別するため, ツイート関する情報(長さ‘ ! ’ ? ’を含むか, 肯定的/否定的な単語の数,retweet かどうか) やユーザーに関する情報 (年齢, フォロー/フォロワーの数, 過去のツイート数),URL リンクの数) を用いた分類器を決定木学習を用いて構築した
スパム・ボット	スパム	Irani ら 2010 [80]	ツイートがスパムであるか否かについて, ツイートのテキストと, ツイートからリンクされた Web ページの内容を用いて, 機械学習により分類を行った
		Wang 2010 [81]	ユーザーと結びついているユーザー数などを素性にした教師あり学習によって, スパマーを判定する分類器を作成した
	ボット	Chu ら 2010 [82]	ツイートの仕方, ツイートの内容, プロフィールの違いに着目し, 線形判別分析による人間とボットの判別方法を提案した

第3章　口語的な表現を含むテキストの省略傾向

本章では, 前章で説明した口語的な表現を多く含むテキスト情報の課題について確認するため, 口語的な表現を多く含む Yahoo! 知恵袋の質問テキストと, 口語的でないテキストを多く含む Wikipedia の日本語テキストの省略傾向を予備実験によって比較し, それによって確認する.

3.1　予備実験の目的

2 章において, Yahoo! 知恵袋データの質問テキストを例示し, 省略があると見られる部分の周辺に, 前に置かれた名詞と接続して「A の B」の形をした連体修飾関係を持つ名詞句が多く存在すること, また, 省略された語についても「A の B」タイプの連体修飾関係における名詞 B のの省略が多いことを指摘した. この「A の B」タイプの連体修飾関係は, 名詞 A が名詞 B を意味的に限定する関係で接続する単純な形のため, 文中に高い頻度で存在することは先行研究 [8] などが指摘している.

しかし, この傾向が, 口語的な表現を多く含むテキストにより顕著に見られるものかどうかについては, これまで報告がない.

そこで, 本研究の予備実験として, 口語的な表現を多く含む Yahoo! 知恵袋の質問テキストと, 口語的でないテキストを多く含む Wikipedia の日本語テキストの省略の数をカウントし, 省略傾向に違いがあるかどうかを明らかにする. なお, 省略の数のカウントと比較に当たっては,「A の B」タイプの連体修飾関係における「名詞 B の省略」と「名詞 A の省略」の違いも確認する.

3.2 実験方法

予備実験は, 口語的な表現を多く含む Yahoo! 知恵袋の質問テキストと, 口語的でないテキストを多く含む Wikipedia の日本語テキストの比較で行う.

実験データは, 口語的な表現を多く含むデータとして, Yahoo! 知恵袋データ中の質問テキストのうちサブカテゴリーが「インターネットとパソコン」に分類されていたデータ 576,841 件 (表 2.3 および表 2.4 を参照) を使用した. 表 3.1 には, 使用した Yahoo! 知恵袋データの一部を例示した. 一方, 口語的でないテキストを多く含むデータとして, 2012 年 8 月時点で公開されていた Wikipedia の日本語データ 1,763,518 件を使用した. 表 3.2 には, 使用した Wikipedia の日本語データの一部を例示した.

比較は, それぞれのデータに対して, 省略されている語の数をカウントする方法で行うが, そのためには, それぞれのデータから, 省略されている語を見つける方法が必要となる. 本予備実験では, 以下の方法を用いた.

(a) 「A の B」の名詞句中の「名詞 B の省略」をカウントする方法

ステップ 1: 省略前の形をした表現の集合 X を収集する
実験データから「名詞 A–の–名詞 B–格助詞–動詞」のパタンにマッチする表現をすべて収集し, 重複を排除する. これを表現集合 X とする.

ステップ 2: 省略後の形をした表現の集合 Y を収集する
実験データから「(“の” 以外)–名詞–格助詞–動詞」のパタンにマッチする表現をすべて収集し, 重複を排除する. これを表現集合 Y とする.

ステップ 3: 省略表現の候補集合 X’ を人工的に作成する
ステップ 1 で収集した表現集合 X から「の–名詞 B」部分を人工的に除いた「名詞 A–格助詞–動詞」表現を作成し, 重複を排除する. これを省略表現の候補集合 X’ とする.

ステップ 4: 省略表現の候補集合 X’ が省略形かどうかを判定する
省略表現の候補集合 X’ の表現のうち, Y にも存在するものがあれば, これを省略形と見なす.

表 3.1: 使用した Yahoo! 知恵袋データ中の質問テキストの例

質問番号	原文
1023729395	いきなり画像のような画面が出てきてPCが止まります。エラーチェックやデフラグ、メモリの診断をしても原因わからず。ウィルスを調べても検知されず。スパイウェアとかの類もなし。Cドライブを購入時の状態に戻すのが良いかと思うのですが、外付けのハードディスクなどのものがないため必要なデータの一時避難場所が無いため現在できません。初期化以外の方法でこの問題を修正する方法は無いでしょうか？OS は WINDOWS VISTA HOME PREMIUM です。機種は LX55Y/D で改造などはしておりません。以前同文で質問したのですが画像を忘れてたので再度質問します。復元ポイントが適当な場所にないです。
1023732835	WindowsXP のライセンス認証をせず 30 日を過ぎたパソコンがあります。そのままライセンス認証せず WindowsMe の OS を新規でインストールしたいのですが、CD-ROM が起動しません。どうすればインストール出来ますか？ XP はは元々起動に不具合があったパソコンのデータのみ出したく、友人にＯＳを借り違法でインストールしたものです。データのみ取り出してパソコンは使用するつもりではなかったのですが、今回正規でＭｅのＯＳを入れそのパソコンを使うことになりました。ＣＤ－ＲＯＭを入れたのですが、ＸＰが起動し、ライセンス認証をしないとするとまたログインの初期画面に戻ってしまい、CD-ROM が起動しません。セーフモードでタスクマネージャから CD-ROM を起動しセットアップしようとしたのですが、Windows2000 からはインストール出来ないと表示が出ました。どうすればインストールする事が出来ますか？
1023753125	WindowsXP の終了オプションで、”コンピュータの電源を切る ”の「アイコン」画面で終了しますが、いつの間にか ”Windows のシャットダウン ”の「ダイヤログ」画面で終了になりました。元に戻すには？
1023776010	キーボードのことに関して質問なんですが … いつも小文字から大文字にするときシフトを押しながらしていたのですが、多分知らないうちにどこかを触ってしまいいつもの状態と反対になってしまいました。（シフトを押しながら入力すると小文字『a』になってしまう … 逆にシフト押さないで入力すると大文字『A』になるんです … ）説明下手で伝わりにくいかと思いますが、元の状態（シフト押さない状態で小文字、シフト押すと大文字）に戻す方法を知っておられる方がいましたら戻し方教えてください ^^; よろしくお願いします＞＜
1023777925	ウインドウズビスタのデスクトップ画面のごみ箱が行方不明になりました。どうすれば元に戻す事ができますか？
:	

表 3.2: 使用した Wikipedia の日本語データの例

No.	原文
1	日常的な手書きの場合、欧米でアンパサンドは「ε」に縦線を引く単純化されたものが使われることがある。
2	狭義には人間の音声による音声言語を指すが、広義には身振りなど音声以外の要素も含む。また、動物間のコミュニケーションや、コンピュータに指示するための記号体系を指す場合もある。
3	狭義には、人間のコミュニケーション、相互作用を統べる規則の内、声にまつわる部分、あるいはその声の代替としての文字表記などにまつわる部分を指す。手話、トーキングドラムなどの例においても、おおよそ声によるコミュニケーションと対応している。
4	ただし、かつて日本の手話言語学者は手話は音声語とは形態において異なるがゆえに、音声言語学とはまったく異なる言語学用語、文法用語によって研究されるべきであるという立場をとっていた。しかし、近年では手話といえどもれっきとした言語であるゆえに音声語と同様の言語学的手法、用語によって説明できるはずであるという立場が一般的となっている。近年では言語学関連の学会等で音声言語と共に手話言語学者の研究報告がプログラムにのぼることも珍しくない。
5	より広義には視覚言語、身体言語など声によるコミュニケーションに還もとできない場合にも、コミュニケーションを統べる規則があれば、それを言語と呼ぶことがある。
6	自然言語は母語として使用する人々の存在を前提として存在しているため、民族の滅亡や他言語による吸収によって使用されなくなることがある。このような言語は死語と呼ばれ、死語が再び母語として使用されることはヘブライ語の例を除けばほとんどない。
7	言語がいつどのように生まれたのか、生まれたのが地球上の一ヶ所か複数ヶ所かはわかっておらず、複数の説が存在するが、例えばデンマークの言語学者オットー・イェスペルセンは、以下のような説を唱えている。
8	また共通語彙から、言語の分化した年代を割り出す方法も考案されている。
9	最も新しい言語であり、また誕生する瞬間がとらえられた言語としては、ニカラグアの子供達の間で 1970 年代後半に発生した「ニカラグア手話」がある。これは、言語能力は人間に生得のものであるという考えを裏付けるものとなった。
10	こうした困難に際しても、単一の基準を決めて分類していくことは、理屈の上では可能である。しかしあえて単一基準を押し通す言語学者は現実にはいない。ある集団を「言語話者」とするか「方言話者」とするかには、政治的・文化的アイデンティティの問題が深く関係している。どのような基準を設けようと、ある地域で多くの賛成を得られる分類基準は、別の地域で強い反発を受けることになる。そうした反発は誤りだと言うための論拠を言語学はもっていないので、結局は慣習に従って、地域ごとに異なる基準を用いて分類することになる。
:	

(b) 「A の B」の名詞句中の「名詞 A の省略」をカウントする方法

ステップ 1 ～ ステップ 2:
実施後の (a) のデータが利用できるため, 割愛.

ステップ 3: 省略表現の候補集合 X' を人工的に作成する
ステップ 1 で収集した表現集合 X から「名詞 A–の」部分を人工的に除いた「名詞 B–格助詞–動詞」表現を作成し, 重複を排除する. これを省略表現の候補集合 X' とする.

ステップ 4: 省略表現の候補集合 X' が省略形かどうかを判定する
省略表現の候補集合 X' の表現のうち, Y にも存在するものがあれば, これを省略形と見なす.

実験データにステップ 1 を実施したところ, 省略前の形をした「名詞 A–の–名詞 B–格助詞–動詞」のパタンにマッチする事例は, Yahoo! 知恵袋データから 205,788 件の事例 (一部を表 3.3 に例示した), Wikipedia データから 859,707 件の事例 (一部を表 3.4 に例示した) が見つかった. ここから「名詞 A–の–名詞 B–格助詞–動詞」の表現部分を取り出し, 重複を除いて収集したところ, Yahoo! 知恵袋データから 139,786 種類の表現 (出現頻度順で先頭から一部を表 3.5 に例示した), Wikipedia データから 769,034 種類の表現 (出現頻度順で先頭から一部を表 3.6 に例示した) を収集した. これを表現集合 X とする.

同様に, ステップ 2 を実施したところ, 省略後の形をした「(前が "の" 以外)–名詞–格助詞–動詞」のパタンにマッチする事例は, Yahoo! 知恵袋データから 1,228,067 件の事例 (先頭から一部を表 3.7 に例示した), Wikipedia データから 3,538,618 件の事例 (先頭から一部を表 3.8 に例示した) が見つかった. ここから「名詞–格助詞–動詞」の表現部分を取り出し, 重複を除いて収集したところ, Yahoo! 知恵袋データから 289,7116 種類の表現 (省略前の形での出現頻度順で先頭から一部を表 3.9 に例示した), Wikipedia データから 1,578,125 種類の表現 (省略前の形での出現頻度順で先頭から一部を表 3.10 に例示した) を収集した. これを表現集合 Y とする.

ステップ 3 では, 名詞 B の省略については, ステップ 1 で収集した表現集合 X か

表 3.3: 「名詞 A–の–名詞 B–格助詞–動詞」にマッチした事例 (Yahoo! 知恵袋)

質問番号	修飾語	接続助詞	被修飾語	格	用言
113621	Ｔｈｉｎｋｐａｄ	の	キーボード	が	使う
113736	バージョン	の	呼び名	に	ついて
113831	ご存じ	の	方	が	いらっしゃる
103885	プレーヤー	の	残存	が	あらわれる
143904	バッテリーメータ	の	表示設定	を	やる
123967	対応	の	ＬＡＮカード	と	いう
123662	２出力タイプ	の	ＶＧＡ切換器	を	探す
123662	２つ	の	モニタ	を	使う
114081	画面	の	画像	を	換える
124157	パソコン	の	時刻	を	合わせる
124157	１０時間ほど前	の	時刻	に	なる
114191	もと	の	状態	に	もどす
104380	上下	の	間隔	が	開く
104505	４月	の	現在	に	なる
114641	図形	の	円	を	描く
114641	ＧＵＩ	の	プログラム	を	欠かせる
134668	ご存知	の	方	が	おる
144689	アプリ	の	専用	と	する
134703	パソコン	の	進化	に	驚く
134703	ＤＶ	の	ＤＶＤ製作	を	加える
114791	登録	の	仕方	を	教える
124802	パソコン	の	画面	で	見る
134833	プリンタ	の	解像度	に	合わせる
114876	ファイアウォール	の	効果	が	ある
134888	英語	の	サイト	が	立つ
134913	バージョンアップ	の	仕方	が	分かる
114926	３	の	起動ランプ	が	ある
144999	パソコン	の	画面	で	見る
115011	漢字変換候補	の	中	から	選ぶ
115051	下	の	質問	に	ついて
：					

表 3.4: 「名詞 A–の–名詞 B–格助詞–動詞」にマッチした事例 (Wikipedia)

NO.	修飾語	接続助詞	被修飾語	格	用言
2	人間	の	音声	に	よる
2	ため	の	記号体系	を	指す
3	声	の	代替	と	する
6	ヘブライ語	の	例	を	除く
10	多く	の	賛成	を	得る
11	方言	の	区別	に	ついて
11	非標準語	の	関係	に	ある
12	漢語	の	方言	と	する
14	一部	の	国	で	話す
15	大多数	の	日本人	に	向ける
16	語法	の	模倣	を	通ずる
17	偽言語比較論	の	範疇	に	収まる
21	近似	の	鼻母音	に	なる
22	日本語	の	東西	の	違う
23	多く	の	場合	に	おく
23	アクセント	の	型	に	まとめる
24	7	の	文	に	ついて
27	学校文法	の	区分	に	従う
32	文	の	成分	と	する
33	品詞	の	特徴	を	形作る
39	誤解	の	もと	に	なる
42	二人称代名詞	の	使用	が	避ける
48	語彙	の	中核部分	を	占める
49	心地	の	言葉	を	書き表す
49	方言	の	音韻体系	を	記す
54	敬語	の	面	から	言う
57	彼	の	こと	を	話す
57	自分	の	身内	に	対する
65	東西	の	差異	が	取る
67	形式化	の	度合い	を	強める
:					

表 3.5: X に収集した省略前の形をした表現の一部 (Yahoo! 知恵袋)

順位	名詞 A-の-名詞 B-格助詞-動詞	頻度	構成比 %
1	パソコン – の – 電源 – を – 入れる	1,465	0.71
2	ご存知 – の – 方 – が – いらっしゃる	976	0.47
3	パソコン – の – 購入 – を – 考える	699	0.34
4	パソコン – の – 電源 – を – 切る	676	0.33
5	システム – の – 復もと – を – する	670	0.33
6	ノートパソコン – の – 購入 – を – 考える	504	0.24
7	ご存知 – の – 方 – が – いる	484	0.24
8	パソコン – の – 音 – が – 出る	462	0.22
9	パソコン – の – 電源 – が – 入る	327	0.16
10	設定 – の – 仕方 – を – 教える	307	0.15
11	メモリ – の – 増設 – を – する	288	0.14
12	メモリ – の – 増設 – を – 考える	286	0.14
13	メモリ – の – 増設 – に – ついて	222	0.11
14	もと – の – 状態 – に – もどす	220	0.11
15	パソコン – の – 電源 – が – 切れる	218	0.11
16	ご存知 – の – 方 – が – おる	208	0.10
17	パソコン – の – 電源 – が – 落ちる	205	0.10
18	パソコン – の – 電源 – を – 落とす	191	0.09
19	パソコン – の – メモリ – に – ついて	171	0.08
20	もと – の – サイズ – に – もどす	168	0.08
21	ＯＳ – の – 再インストール – を – する	165	0.08
22	パソコン – の – 電源 – を – つける	164	0.08
23	ご存じ – の – 方 – が – いらっしゃる	155	0.08
24	パソコン – の – 中 – に – 入る	140	0.07
25	削除 – の – 仕方 – を – 教える	140	0.07
26	設定 – の – 仕方 – が – わかる	139	0.07
27	中古 – の – パソコン – を – 買う	133	0.06
28	パソコン – の – 中 – に – ある	129	0.06
29	パソコン – の – 電源 – を – いれる	127	0.06
30	故障 – の – 原因 – に – なる	123	0.06
:			

表 3.6: X に収集した省略前の形をした表現の一部 (Wikipedia)

順位	名詞 A–の–名詞 B–格助詞–動詞	頻度	構成比 %
1	２０００年 – の – 国勢調査 – に – よる	1,083	0.13
2	１８歳以上 – の – 女性１００人ごと – に – 対する	750	0.09
3	１８歳以上 – の – 女性１００人 – に – 対する	247	0.03
4	小 – の – 月 – を – 示す	238	0.03
5	もと – の – 姿勢 – に – もどる	217	0.03
6	相対式２面２線 – の – ホーム – を – 持つ	206	0.02
7	独身 – の – 居住者 – が – 住む	178	0.02
8	瀕死 – の – 重傷 – を – 負う	152	0.02
9	２０００年 – の – 国勢調査 – に – おける	149	0.02
10	１８歳未満 – の – 子ども – と – 暮らす	143	0.02
11	多く – の – 観光客 – が – 訪れる	137	0.02
12	島式１面２線 – の – ホーム – を – 持つ	135	0.02
13	近隣 – の – 都市 – と – する	132	0.02
14	後進 – の – 指導 – に – 当たる	131	0.02
15	ソフトウェア – の – 更新 – で – なされる	126	0.01
16	特定 – の – 条件 – を – 満たす	110	0.01
17	一部 – の – 例外 – を – 除く	105	0.01
18	父 – の – 後 – を – 継ぐ	104	0.01
19	単式１面１線 – の – ホーム – を – 持つ	100	0.01
20	偏旁 – の – 意符 – と – する	100	0.01
21	後進 – の – 指導 – に – あたる	99	0.01
22	批判 – の – 対象 – と – なる	96	0.01
23	桜 – の – 名所 – と – する	94	0.01
24	郡 – の – 選挙 – で – 選ぶ	91	0.01
25	絶滅 – の – 危機 – に – 瀕する	90	0.01
26	計２面３線 – の – ホーム – を – 持つ	88	0.01
27	合計２面３線 – の – ホーム – を – 持つ	88	0.01
28	平成 – の – 大合併 – に – よる	88	0.01
29	地域 – の – 領有権 – に – 関する	86	0.01
30	活動 – の – 場 – を – 移す	81	0.01
:			

表 3.7: 「名詞–格助詞–動詞」のパタンにマッチした事例 (Yahoo! 知恵袋)

質問番号	“の” 以外	名詞	格	用言
13153	を	購入しょう	と	思う
13153	、	ハイブリット	と	書く
13153	か	意味	が	ある
122677		バーコード作成ソフト	を	探す
122762		方法	を	教える
133613	と	保護エラー	と	なる
113656		ソフト	を	探す
113656	、	イメージ	が	壊れる
113656	する	機能	に	欠ける
113676	、	パソコン	の	動く
113676	か	方法	が	ある
103725	詳しい	解説	が	載る
133728	、	デメリット	を	教える
113831	、	ＤＯＳソフト	が	動く
113831	見た	こと	が	ある
133723	は	ファイル	の	関連付ける
133723	、	ブラウザ	で	開く
103885	ｒｅａｌｏｎｅ	プレーヤー	を	使う
103885	立ち上げ	画面	に	もどす
103885	この	オンボードメモリ	と	いう
123912	手頃な	価格帯	が	ある
123912	もどる	こと	が	出来る
123912	または	お店	を	知る
123912	る	方	が	いる
123967		ノートパソコン	で	使う
123967	使う	事	の	できる
123967	ので	製品	と	する
123967	ような	気	が	する
113986	手頃な	価格帯	が	ある
113986	もどる	こと	が	出来る
：				

表 3.8: 「名詞–格助詞–動詞」のパタンにマッチした事例 (Wikipedia)

NO.	“の” 以外	名詞	格	用言
1	に	縦線	を	引く
1	れた	もの	が	使う
1	れる	こと	が	ある
2	よる	音声言語	を	指す
3	、	相互作用	を	統べる
3	、	声	に	まつわる
3	まつわる	部分	を	指す
3	おおよそ	声	に	よる
4	は	形態	に	おく
4	、	文法用語	に	よる
4	いう	立場	を	とる
4	は	手話	と	いう
4	、	用語	に	よる
4	が	一般的	と	なる
4	が	プログラム	に	のぼる
5	など	声	に	よる
5	、	コミュニケーション	を	統べる
5	統べる	規則	が	ある
5	を	言語	と	呼ぶ
5	呼ぶ	こと	が	ある
6	は	母語	と	する
6	を	前提	と	する
6	や	他言語	に	よる
6	よる	吸収	に	よる
6	なる	こと	が	ある
6	は	死語	と	呼ぶ
6	再び	母語	と	する
7	ような	説	を	唱える
8	した	年代	を	割り出す
9	する	瞬間	が	とらえる
:				

表 3.9: Y に収集した省略後の形をした表現の一部 (Yahoo! 知恵袋)

順位	名詞-格助詞-動詞	頻度	構成比 %
1	方法 – を – 教える	14,100	1.15
2	こと – が – できる	12,572	1.02
3	こと – が – ある	7,611	0.62
4	パソコン – を – 買う	7,374	0.60
5	方法 – が – ある	6,162	0.50
6	電源 – を – 入れる	5,280	0.43
7	気 – が – する	4,860	0.40
8	気 – に – なる	4,549	0.37
9	こと – が – 出来る	4,505	0.37
10	もと – に – もどす	4,161	0.34
11	修理 – に – 出す	4,109	0.33
12	メッセージ – が – 出る	3,851	0.31
13	音 – が – 出る	3,842	0.31
14	もの – が – ある	3,806	0.31
15	パソコン – を – 使う	3,564	0.29
16	電源 – を – 切る	3,349	0.27
17	こと – に – なる	3,327	0.27
18	エラー – が – 出る	3,298	0.27
19	時間 – が – かかる	3,223	0.26
20	状態 – に – なる	3,218	0.26
21	もと – に – もどる	3,024	0.25
22	パソコン – に – ついて	2,998	0.24
23	音 – が – する	2,962	0.24
24	パソコン – に – 取る	2,928	0.24
25	表示 – が – 出る	2,805	0.23
26	問題 – が – ある	2,805	0.23
27	パソコン – を – 立ち上げる	2,666	0.22
28	方 – が – いる	2,619	0.21
29	の – が – ある	2,505	0.20
30	キー – を – 押す	2,454	0.20
:			

表 3.10: Y に収集した省略後の形をした表現の一部 (Wikipedia)

順位	名詞-格助詞-動詞	頻度	構成比 %
1	こと – が – できる	34,879	0.99
2	こと – に – なる	31,198	0.88
3	こと – が – ある	18,032	0.51
4	こと – に – よる	15,342	0.43
5	こと – と – なる	15,329	0.43
6	中心 – と – する	7,676	0.22
7	こと – が – 出来る	6,001	0.17
8	もの – と – する	5,945	0.17
9	場合 – が – ある	5,794	0.16
10	目的 – と – する	5,636	0.16
11	もの – が – ある	5,422	0.15
12	の – に – 対する	4,733	0.13
13	影響 – を – 与える	4,425	0.13
14	もの – と – なる	3,397	0.10
15	現在 – に – 至る	2,878	0.08
16	こと – を – 知る	2,835	0.08
17	事 – に – なる	2,700	0.08
18	こと – に – する	2,295	0.06
19	対象 – と – する	2,211	0.06
20	影響 – を – 受ける	2,189	0.06
21	結果 – と – する	2,180	0.06
22	手 – に – 入れる	2,161	0.06
23	役割 – を – 果たす	2,096	0.06
24	こと – を – 示す	1,958	0.06
25	日本 – に – おく	1,929	0.05
26	問題 – と – なる	1,827	0.05
27	こと – が – あう	1,775	0.05
28	傾向 – が – ある	1,759	0.05
29	事 – と – なる	1,727	0.05
30	もの – と – 考える	1,698	0.05
:			

ら「の–名詞 B」部分を人工的に除いた表現を作成し, そこから重複を除きいたものを Yahoo! 知恵袋データから 92,270 種類の表現, Wikipedia データから 607,130 種類の表現を収集した. 名詞 A の省略については, 同様に表現集合 X から「名詞 A –の」部分を人工的に除いて作成した表現を重複を除き, Yahoo! 知恵袋データから 92,270 種類の表現, Wikipedia データから 607,130 種類の表現を収集した. これをそれぞれ省略表現の候補集合 X' とする.

最後のステップ 4 では, 名詞 B の省略については, 省略表現 X' のうち, 表現集合 Y の中にも存在するものがあれば「省略形」とみなす方法でしカウントを行い, Yahoo! 知恵袋データから 23,021 種類 (B の省略としたものを表 3.11, B の省略でないものを表 3.12 にそれぞれ省略前の形での出現頻度順で一部を例示した), Wikipedia データから 81,587 種類 (B の省略としたものを表 3.13, B の省略でないものを表 3.14 にそれぞれ省略前の形での出現頻度順で一部を例示した) の省略形を発見した.

同じ方法により, 名詞 A の省略についても, Yahoo! 知恵袋データから 30,785 種類 (A の省略としたものを表 3.15, A の省略でないものを表 3.16 にそれぞれ省略前の形での出現頻度順で一部を例示した), Wikipedia データから 131,569 種類 (A の省略としたものを表 3.17, A の省略でないものを表 3.18 にそれぞれ省略前の形での出現頻度順で一部を例示した) の省略を発見した.

表 3.11: B の省略がある表現の一部 (Yahoo !知恵袋)

順位	B を省略した形	省略前の頻度	省略形の頻度
1	パソコン – に – ついて	3,636	2,998
2	パソコン – を – 入れる	1,657	28
3	もと – に – もどす	944	4,161
4	パソコン – を – 考える	843	11
5	システム – を – する	696	1
6	パソコン – を – 切る	688	77
7	ノートパソコン – に – ついて	660	392
8	パソコン – を – する	627	405
9	パソコン – が – 出る	553	29
10	ノートパソコン – を – 考える	538	18
11	もと – に – もどる	433	3,024
12	ＮＥＣ – を – 使う	410	21
13	パソコン – に – ある	406	311
14	設定 – を – 教える	401	106
15	パソコン – が – 壊れる	384	1,637
16	メモリ – に – ついて	380	350
17	オススメ – を – 教える	379	136
18	パソコン – が – 入る	345	5
19	メモリ – を – する	339	2
20	パソコン – を – 見る	330	208
21	エクセル – に – ついて	319	704
22	ＣＰＵ – に – ついて	317	525
23	オススメ – が – ある	300	163
24	ＨＤＤ – に – ついて	289	344
25	パソコン – と – する	282	140
26	ノートパソコン – が – 壊れる	256	178
27	ファイル – を – する	248	8
28	中古 – を – 買う	247	66
29	パソコン – が – 切れる	241	31
30	画面 – に – ある	235	27
:			

表 3.12: B の省略がない表現の一部 (Yahoo !知恵袋)

順位	B を省略した形	省略前の頻度	省略形の頻度
1	ご存知 – が – いらっしゃる	981	N/A
2	ご存知 – が – いる	492	N/A
3	メモリ – を – 考える	324	N/A
4	ＯＳ – を – する	222	N/A
5	削除 – を – 教える	214	N/A
6	ご存知 – が – おる	211	N/A
7	データ – を – する	200	N/A
8	ご存じ – が – いらっしゃる	157	N/A
9	ＤＶＤ – を – する	148	N/A
10	過去 – を – 見る	137	N/A
11	パソコン – を – 押す	127	N/A
12	変更 – を – 教える	126	N/A
13	皆さん – を – 聞く	120	N/A
14	くらい – が – ある	106	N/A
15	パソコン – を – 上げる	103	N/A
16	くらい – が – かかる	100	N/A
17	ドライブ – を – する	100	N/A
18	程度 – が – ある	100	N/A
19	メモリー – を – する	98	N/A
20	履歴 – を – する	98	N/A
21	ため – を – 教える	97	N/A
22	インストール – を – 教える	91	N/A
23	自分 – に – 入る	82	N/A
24	ハードディスク – を – する	80	N/A
25	とき – を – 教える	79	N/A
26	下記 – を – する	76	N/A
27	方 – を – 聞く	75	N/A
28	ため – が – ある	74	N/A
29	付属 – を – 使う	72	N/A
30	ＮＥＣ – が – 出る	71	N/A
:			

表 3.13: B の省略がある表現の一部 (Wikipedia))

順位	B を省略した形	省略前の頻度	省略形の頻度
1	ため – と – する	1,388	271
2	多く – が – ある	740	6
3	初 – と – なる	733	368
4	２つ – が – ある	578	14
5	もと – に – もどる	564	284
6	一部 – を – 除く	480	517
7	一つ – と – する	382	59
8	日本 – に – おく	364	1,929
9	３つ – が – ある	358	11
10	日本 – と – する	297	17
11	彼 – に – よる	292	198
12	複数 – が – ある	282	1
13	多く – を – 残す	275	1
14	多数 – が – ある	267	11
15	兄弟 – と – する	257	21
16	多く – に – おく	250	5
17	際 – と – する	241	1
18	その他 – と – する	234	2
19	日本 – に – よる	227	137
20	初 – と – する	215	3
21	日本 – に – おける	213	1,355
22	自分 – と – する	210	9
23	父 – を – 継ぐ	210	11
24	ため – と – なる	207	5
25	チーム – と – する	200	143
26	彼 – と – する	199	7
27	複数 – に – よる	196	3
28	４つ – が – ある	193	9
29	ひとつ – と – する	190	14
30	二つ – が – ある	190	5
:			

表 3.14: B の省略がない表現の一部 (Wikipedia)

順位	B を省略した形	省略前の頻度	省略形の頻度
1	２０００年 – に – よる	1,122	N/A
2	１８歳以上 – に – 対する	1,005	N/A
3	多く – が – 訪れる	364	N/A
4	多く – に – よる	313	N/A
5	多く – を – 集める	239	N/A
6	小 – を – 示す	238	N/A
7	多く – を – 持つ	234	N/A
8	間 – と – する	223	N/A
9	１８歳未満 – と – 暮らす	217	N/A
10	相対式２面２線 – を – 持つ	212	N/A
11	初 – を – 果たす	209	N/A
12	多く – を – 出す	207	N/A
13	独身 – が – 住む	182	N/A
14	ため – が – ある	162	N/A
15	ほど – を – 持つ	161	N/A
16	瀕死 – を – 負う	160	N/A
17	多く – が – 集まる	157	N/A
18	ため – を – 行く	149	N/A
19	近隣 – と – する	144	N/A
20	一定 – を – 満たす	140	N/A
21	２度目 – を – 果たす	139	N/A
22	一定 – が – ある	139	N/A
23	後進 – に – 当たる	139	N/A
24	島式１面２線 – を – 持つ	139	N/A
25	多く – を – 生む	136	N/A
26	一定 – を – 得る	133	N/A
27	ため – を – 行う	132	N/A
28	人口 – を – 占める	132	N/A
29	多く – で – 賑わう	131	N/A
30	一定 – を – 持つ	127	N/A
:			

表 3.15: A の省略がある表現の一部 (Yahoo !知恵袋)

順位	A を省略した形	省略前の頻度	省略形の頻度
1	購入 – を – 考える	2,760	578
2	電源 – を – 入れる	1,984	5,280
3	仕方 – を – 教える	1,849	132
4	パソコン – を – 使う	1,529	3,564
5	電源 – を – 切る	1,236	3,349
6	方 – が – いらっしゃる	1,196	1,810
7	パソコン – を – 買う	1,177	7,374
8	方法 – を – 教える	971	14,100
9	仕方 – が – わかる	825	68
10	中 – に – ある	750	136
11	復もと – を – する	730	120
12	状態 – に – もどす	707	230
13	ノートパソコン – を – 使う	633	1,169
14	音 – が – 出る	616	3,842
15	方 – が – いる	602	2,619
16	電源 – が – 入る	585	1,579
17	設定 – を – する	563	1,316
18	画面 – に – なる	554	1,580
19	画面 – が – 出る	547	1,768
20	増設 – を – 考える	496	159
21	中 – に – 入る	473	329
22	増設 – を – する	457	103
23	仕方 – が – 分かる	445	35
24	ボタン – を – 押す	443	1,706
25	増設 – に – ついて	387	44
26	質問 – を – する	386	1,848
27	もの – が – ある	362	3,806
28	パソコン – が – ある	347	805
29	設定 – に – ついて	340	47
30	プロパティ – を – 見る	338	283
:			

表 3.16: A の省略がない表現の一部 (Yahoo! 知恵袋))

順位	A を省略した形	省略前の頻度	省略形の頻度
1	件 – で – 教える	37	N/A
2	画面設定 – を – 開く	28	N/A
3	仕方 – が – 違う	27	N/A
4	前 – に – 座る	26	N/A
5	ほう – に – あう	25	N/A
6	計算式 – に – ついて	21	N/A
7	構文 – が – 間違う	19	N/A
8	種類 – を – 調べる	19	N/A
9	方 – に – あう	19	N/A
10	代わり – に – 使う	18	N/A
11	中 – で – 聞く	18	N/A
12	中 – を – 探す	18	N/A
13	方 – に – 言う	18	N/A
14	選択 – で – 悩む	17	N/A
15	プログラム – に – 移る	16	N/A
16	機密情報 – と – する	16	N/A
17	件 – で – する	16	N/A
18	手 – と – いう	16	N/A
19	中 – から – 消える	15	N/A
20	廃棄 – に – ついて	15	N/A
21	負荷 – が – 高まる	15	N/A
22	方 – に – 行く	15	N/A
23	ほう – に – 行く	14	N/A
24	まま – で – 使う	14	N/A
25	方 – が – 適す	14	N/A
26	ほう – が – 優れる	13	N/A
27	仕方 – が – 載る	13	N/A
28	仕方 – が – 書く	13	N/A
29	中 – に – 組む	13	N/A
30	ダウンロードエラー – と – 出る	12	N/A
:			

表 3.17: A の省略がある表現の一部 (Wikipedia)

順位	A を省略した形	省略前の頻度	省略形の頻度
1	影響 – を – 受ける	2,159	2,189
2	一つ – と – する	1,732	59
3	一環 – と – する	1,639	160
4	対象 – と – なる	1,484	640
5	手 – に – よる	1,290	23
6	国勢調査 – に – よる	1,238	49
7	一部 – と – する	1,040	31
8	一部 – と – なる	969	36
9	子 – と – する	934	74
10	一つ – と – なる	916	26
11	一員 – と – する	884	29
12	一人 – と – する	850	43
13	ホーム – を – 持つ	818	16
14	中心 – と – なる	758	1,651
15	女性１００人ごと – に – 対する	750	719
16	特徴 – と – する	732	1,301
17	対象 – と – する	704	2,211
18	ひとつ – と – する	692	14
19	原因 – と – なる	681	752
20	影響 – に – よる	671	89
21	卵 – を – 産む	646	226
22	こと – を – 指す	637	343
23	中 – に – ある	621	62
24	息子 – と – する	571	70
25	支援 – を – 受ける	535	171
26	長男 – と – する	525	70
27	もの – と – する	517	5,945
28	攻撃 – を – 受ける	514	517
29	舞台 – と – なる	489	532
30	例 – と – する	487	1,604
:			

表 3.18: A の省略がない表現の一部 (Wikipedia)

順位	A を省略させた形	省略前の頻度	省略形の頻度
1	位 – を – 加える	268	N/A
2	居住者 – が – 住む	178	N/A
3	場 – を – 広げる	153	N/A
4	子ども – と – 暮らす	145	N/A
5	勃発 – に – よる	138	N/A
6	更新 – で – なされる	126	N/A
7	教区 – に – 分ける	117	N/A
8	郡区 – に – 分ける	103	N/A
9	領有権 – に – 関する	91	N/A
10	仲介 – に – よる	80	N/A
11	発達 – に – 伴う	78	N/A
12	代 – に – なる	77	N/A
13	もと – で – 学ぶ	74	N/A
14	もと – を – 去る	74	N/A
15	二男 – と – する	74	N/A
16	有無 – に – かかわる	67	N/A
17	矢 – が – 立つ	65	N/A
18	ひとつ – に – 数える	64	N/A
19	作品リスト – を – 示す	62	N/A
20	途 – に – 就く	62	N/A
21	激化 – に – よる	61	N/A
22	もと – に – 送る	60	N/A
23	悪化 – に – 伴う	59	N/A
24	出場 – に – 留まる	59	N/A
25	もと – で – 働く	55	N/A
26	仕方 – に – よる	52	N/A
27	大半 – を – 失う	51	N/A
28	政庁所在地 – と – なる	48	N/A
29	出場 – に – とどまる	47	N/A
30	座 – を – 狙う	46	N/A
:			

説明した方法に沿ったデータ加工の流れを図 3.1 および図 3.2 に示す.

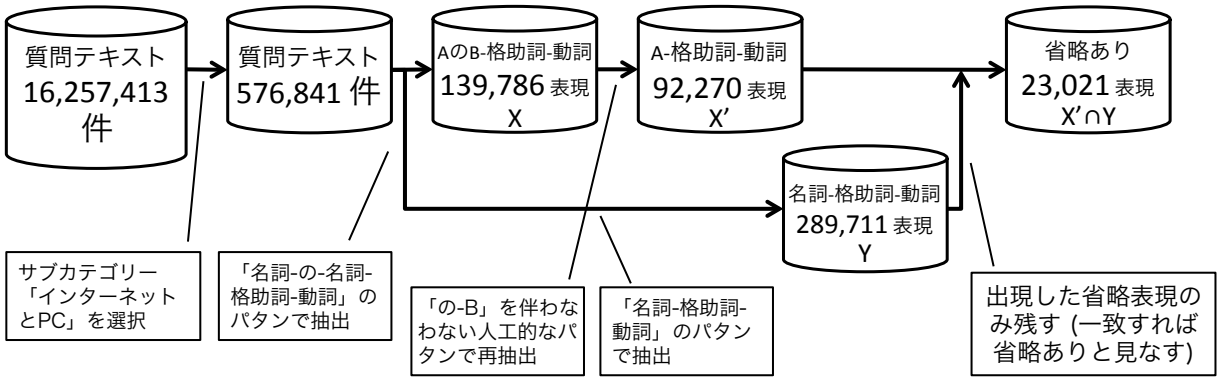

図 3.1: Yahoo! 知恵袋からの名詞 B の省略の収集の流れ

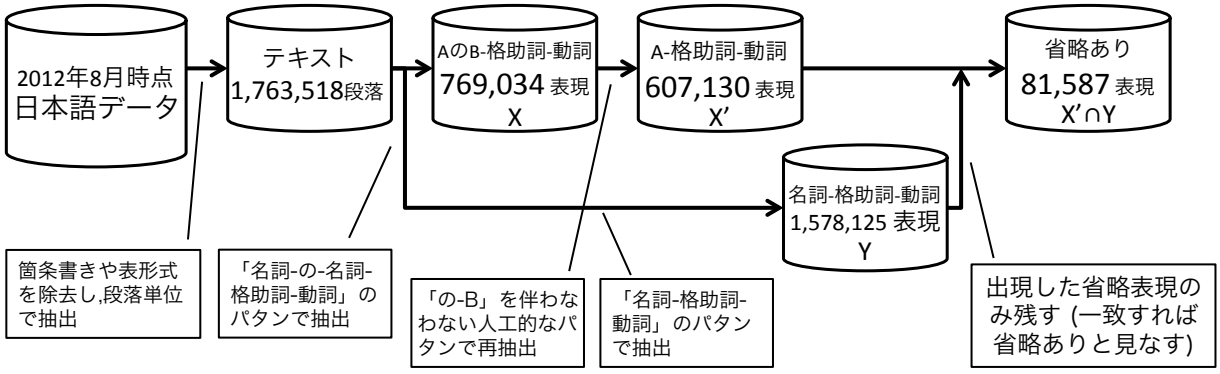

図 3.2: Wikipeda からの名詞 B の省略の収集の流れ

3.3 結果と考察

3.3.1 口語的と非口語的なテキストの比較

口語的な表現を多く含む Yahoo! 知恵袋の質問テキストと, 口語的でないテキストを多く含む Wikipedia の日本語テキストで, 名詞 B の省略と名詞 A の省略の

数を比較した. 結果を表3.19および表3.20に示す.

表3.19は, Yahoo! 知恵袋とWikipediaのそれぞれについて, 名詞Bの省略に関する結果を比較したものである. 表の見方は, 左から, 実験データ中に見つかった省略前の形(名詞A–の–名詞B–格助詞–動詞)をした表現の種類数X, そこから「の–名詞B」を人工的に除いて作成した表現の種類数X', 実験データ中に見つかった省略後の形(名詞–格助詞–動詞)をした表現の種類数Y, Bの省略(X'のうちY中にも存在したもの), Bの省略率(Bの省略をYで割った)の順である.

表3.20は, Yahoo! 知恵袋とWikipediaのそれぞれについて, 名詞Aの省略に関する結果を比較したものである. 表の見方は, 左から, 実験データ中に見つかった省略前の形(名詞A–の–名詞B–格助詞–動詞)をした表現の種類数X, そこから「名詞A–の」を人工的に除いて作成した表現の種類数X', 実験データ中に見つかった省略後の形(名詞–格助詞–動詞)をした表現の種類数Y, Aの省略(X'のうちY中にも存在したもの), Aの省略率(Aの省略をYで割った)の順である.

表3.19: 口語的テキストと非口語的テキストの比較(Bの省略)

データ	省略前の表現集合X	「の–名詞B」を除去X'	省略後の表現集合Y	Bの省略 $X' \cap Y$	not Bの省略 $X' \cap (\overline{X' \cap Y})$	Bの省略率 $(X' \cap Y) / Y$
Y!知恵袋	139,786	92,270	289,711	23,021	69,249	7.95 %
Wikipedia	769,034	607,130	1578,125	81,587	525,543	5.17 %

表3.20: 口語的テキストと非口語的テキストの比較(Aの省略)

データ	省略前の表現集合X	「名詞A–の」を除去X'	省略後の表現集合Y	Aの省略 $Y \cap X'$	not Aの省略 $X' \cap (\overline{X' \cap Y})$	Aの省略率 $(X' \cap Y) / Y$
Y!知恵袋	139,786	67,430	289,711	30,785	36,645	10.63 %
Wikipedia	769,034	407,270	1,578,125	131,569	275,701	8.34 %

表3.19から, Yahoo! 知恵袋におけるBの省略率が7.95%であるのに対して, Wikipediaでは5.17%であることが確認できる. 同様に, 表3.20からYahoo! 知恵袋におけるAの省略率が10.63%であるのに対して, Wikipediaが8.34%であることが確認できる. これら2つの結果は, 名詞Bの省略と名詞Aの省略の両方がとも

に, Wikipedia との比較で Yahoo! 知恵袋の方が省略率が高く, 非口語的なテキストに比べて, 口語的な表現を多く含む質問テキストの省略傾向の高さを示している.

一方, A の B タイプの名詞句では, 元来, 前方にある名詞 A 方が名詞 B より省略され易く, 表 3.19 または表 3.20 の上で, 直接, 名詞 B と名詞 A の省略を比較するのは難しい.

3.3.2 名詞 B の省略と名詞 A の省略の比較

3.3.2 節で述べたように, 表 3.19 と表 3.20 を比較すると, Yahoo! 知恵袋と Wikipedia ともに, B の省略との比較で A の省略の方が省略率が高い.

元々 A の B タイプの名詞句では, 前方にある名詞 A 方が名詞 B より省略され易いため, 表 3.19 または表 3.20 の上で, 名詞 B の省略と名詞 A の省略を比較するのは難しい. そこで, さらに詳しく名詞 B の省略と名詞 A の省略傾向を比較するため, 同じデータを使い, 名詞 A や名詞 B の必須格が省略される割合を比較する追加実験を行った.

本研究では, 笹野ら [89, 90] に倣い, ある名詞にとって必須的な関係をその名詞の必須格と呼ぶことにする. 名詞の必須格の多くは「チケットの値段」などの形で出現する. しかし, 文脈的に明らかな場合は必須格は省略される場合がある. 笹野らは, 各必須格にはどのような用例があるのかという知識を記述した名詞格フレーム辞書をコーパスから自動構築した. 表 3.21 には, 本研究で使用した笹野らの名詞格フレーム辞書から必須格を持つデータの一部を抜粋して示している. 表の左から, 名詞の表記, 格の名前, 用例の頻度, 用例が順に並んでおり, 例えば「価格」であれば「何の価格」かという情報が必須要素となり, その候補として「通販」「土地」などの用例があることが分かる.

実験方法は, まず, 彼らが Web 上の約 16 億文の日本語テキストから自動構築した約 16 万名詞からなる格フレーム辞書を利用し, 予め名詞格フレーム辞書から, 必須格と用例のペア 324,119 件を抽出する. 次に, Yahoo! 知恵袋データ, Wikipedia それぞれについて, 前の実験で「省略形あり」と判断した 名詞 B の省略と名詞 A の省略のそれぞれで必須格と用例のペアに一致する数と一致しない数を集計した.

説明した方法に沿ったデータ加工の流れを図 3.3 および図 3.4 に示す.

表 3.21: 名詞格フレーム辞書中の必須格を持つデータ (一部)

名詞表記	格の名前	用例の頻度	項を構成する 1 つの用例の表記
価格/かかく:名 2	必須格 (属性) 格	18814	通販/つうはん
価格/かかく:名 2	必須格 (属性) 格	3764	各社/かくしゃ
価格/かかく:名 2	必須格 (属性) 格	2682	土地/とち
価格/かかく:名 2	必須格 (属性) 格	2274	驚き/おどろき v
価格/かかく:名 2	必須格 (属性) 格	1767	不動産/ふどうさん
価格/かかく:名 2	必須格 (属性) 格	1430	上記/じょうき
価格/かかく:名 2	必須格 (属性) 格	1390	掲載/けいさい
価格/かかく:名 2	必須格 (属性) 格	1330	資産/しさん
価格/かかく:名 2	必須格 (属性) 格	1267	サービス/さーびす
価格/かかく:名 2	必須格 (属性) 格	927	ガソリン/がそりん
価格/かかく:名 2	必須格 (属性) 格	902	本体/ほんたい
価格/かかく:名 2	必須格 (属性) 格	888	雑誌/ざっし
価格/かかく:名 2	必須格 (属性) 格	819	車/くるま?車/しゃ
価格/かかく:名 2	必須格 (属性) 格	803	税込み/ぜいこみ
価格/かかく:名 2	必須格 (属性) 格	745	住宅/じゅうたく
価格/かかく:名 2	必須格 (属性) 格	730	表示/ひょうじ
価格/かかく:名 2	必須格 (属性) 格	702	込み/こみ v
価格/かかく:名 2	必須格 (属性) 格	677	納得/なっとく
価格/かかく:名 2	必須格 (属性) 格	677	メモリー/めもりー
価格/かかく:名 2	必須格 (属性) 格	667	市場/いちば?市場/しじょう
価格/かかく:名 2	必須格 (属性) 格	662	米/こめ?米/まい
価格/かかく:名 2	必須格 (属性) 格	634	マンション/まんしょん
価格/かかく:名 2	必須格 (属性) 格	628	ソフト/そふと?ソフトだ/そふとだ
価格/かかく:名 2	必須格 (属性) 格	614	下記/かき
価格/かかく:名 2	必須格 (属性) 格	580	野菜/やさい
価格/かかく:名 2	必須格 (属性) 格	576	程度/ていど
価格/かかく:名 2	必須格 (属性) 格	563	くらい/くらい
価格/かかく:名 2	必須格 (属性) 格	560	債券/さいけん
価格/かかく:名 2	必須格 (属性) 格	544	パソコン/ぱそこん
価格/かかく:名 2	必須格 (属性) 格	493	建物/たてもの
価格/かかく:名 2	必須格 (属性) 格	439	金/かね?金/きん
価格/かかく:名 2	必須格 (属性) 格	437	希望/きぼう
価格/かかく:名 2	必須格 (属性) 格	414	本/ほん
価格/かかく:名 2	必須格 (属性) 格	404	記載/きさい
価格/かかく:名 2	必須格 (属性) 格	403	原油/げんゆ
:			

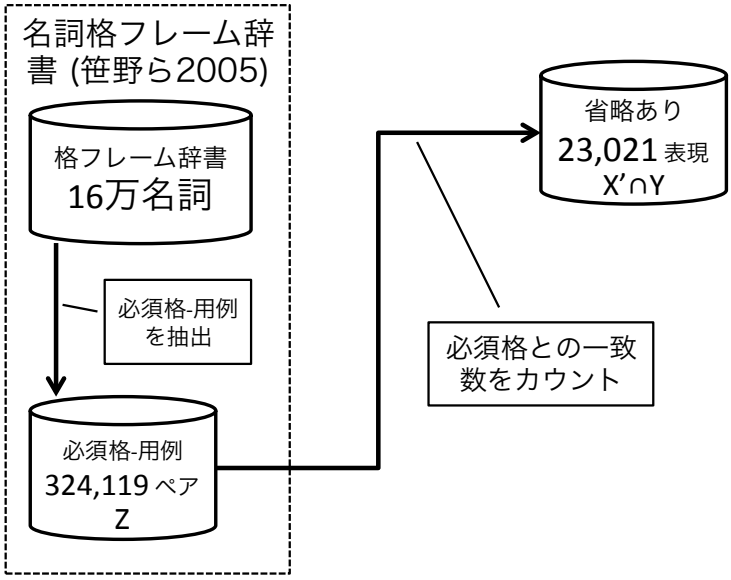

図 3.3: 名詞 B の省略中の必須格の抽出方法 (Yahoo! 知恵袋)

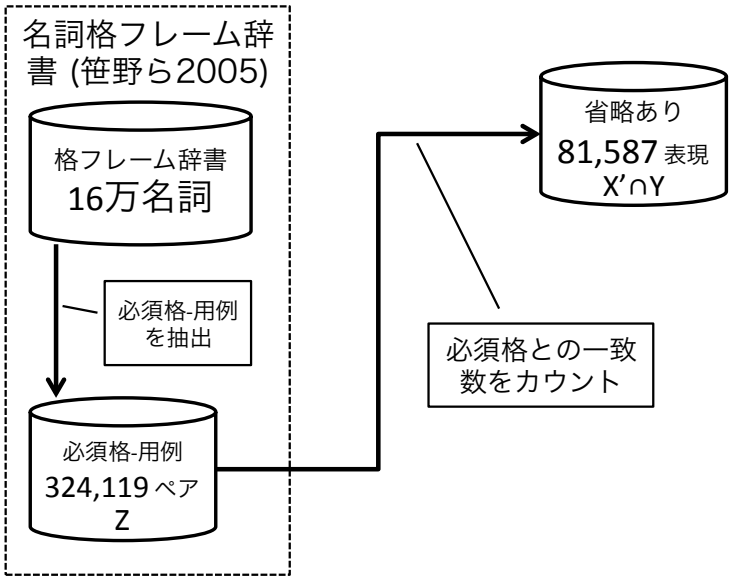

図 3.4: 名詞 B の省略中の必須格の抽出方法 (Wikipedia)

追加実験の結果を表3.22および表3.23に示す.

表 3.22: B の省略と A の省略の比較 (Yahoo! 知恵袋)

省略タイプ	辞書中の必須格 Z	必須格の省略 (X' ∩ Y) ∩ Z	率 %	必須格の省略でない (X' ∩ ($\overline{X' \cap Y}$)) ∩ Z	率 %
B の省略	324,119	1,570	6.82	1,913	2.76
A の省略		1,228	3.99	1,133	3.09

表 3.23: B の省略と A の省略の比較 (Wikipedia)

省略タイプ	辞書中の必須格 Z	必須格の省略 ('X ∩ Y) ∩ Z	率 %	必須格の省略でない ('X ∩ ($\overline{X' \cap Y}$)) ∩ Z	率 %
B の省略	324,119	5,267	6.46	10,047	1.91
A の省略		4,672	3.55	5,407	1.96

表3.22は, Yahoo! 知恵袋データについて, 名詞 B の省略と名詞 A の省略のそれぞれで辞書中の必須格ペアと一致した数と割合, 辞書中の必須格と一致しなかった数と割合を示したものである.

表3.23は, Wikipedia について, 名詞 B の省略と名詞 A の省略のそれぞれで辞書中の必須格ペアと一致した数と割合, 辞書中の必須格と一致しなかった数と割合を示したものである.

表3.22から, Yahoo! 知恵袋では, 名詞 B の省略における必須格の割合が 6.82% であるのに対して, 名詞 A の省略における必須格の割合が 3.99%であることが確認できる. 同様に, 表3.23から, Wikipedia では, 名詞 B の省略における必須格の割合が 6.46%であるのに対して, 名詞 A の省略における必須格の割合が 3.55%であることが確認できる. これら2つの結果は, 名詞 A の省略との比較で名詞 B の省略の方が必須格がより多く含まれていることを示している.

3.4 まとめ

2章において, Yahoo! 知恵袋データの質問テキストを例示し, 省略があると見られる部分の周辺に, 前に置かれた名詞と接続して「A の B」の形をした連体修飾関係を持つ名詞句が多く存在すること, また, 省略された語についても「A の B」タイプの連体修飾関係における名詞 B の省略が多いことを指摘した.

この「A の B」タイプの連体修飾関係は, 名詞 A が名詞 B を意味的に限定する関係で接続する単純な形のため, 文中に高い頻度で存在することは先行研究 [8] などが指摘している. しかし, この傾向が, 口語的な表現を多く含むテキストにより顕著に見られるものかどうかについては, これまで報告がない.

そこで, 本章では, 前章で説明した口語的な表現を多く含むテキスト情報の課題について確認するため, 口語的な表現を多く含む Yahoo! 知恵袋の質問テキストと, 口語的でないテキストを多く含む Wikipedia の日本語テキストの省略傾向を予備実験によって比較し, 確認した.

予備実験の結果から, Yahoo! 知恵袋のような口語的な表現を多く含む質問テキストが,Wikipedia のような口語的でないテキストに比べて, 図 2.2 と図 2.3 で示したような「A の B」のタイプの連体修飾関係における名詞 B の省略が多いことを確認した. また, Yahoo! 知恵袋のような口語的な表現を多く含む質問テキストでは, 名詞 A の省略に比べて名詞 B の省略の方が, 必須格の省略が多いことを確認した.

そのため, 本研究では「A の B」タイプの名詞句における名詞 B の省略に焦点を当て, 省略された名詞 B を補完する方法を検討する. 4 章では, 検討した省略の補完方法と実施した評価実験の結果について述べる.

第4章　文脈情報を利用した省略の補完

本章では, 読み手によらない, 問い合わせテキストの正確な理解を目的とし, 口語的な問い合わせテキストで見られる「A の B」タイプの名詞句にける名詞 B の省略に焦点を当て, トピックモデルを用いて, 省略された名詞 B を補完する方法を提案し, 評価実験によって示された効果について述べる.

4.1　研究の目的

2.1 節では, 口語的な表現を多く含む Yahoo! 知恵袋の質問テキストと, 口語的でないテキストを多く含む Wikipedia の日本語テキストを比較した予備実験から, 口語的な表現を多く含む質問テキストの方が,「A の B」タイプの名詞句における名詞 B の省略が多いことを確認した.

一方で, 2.1 節で述べたように従来の方法では同じ文脈中に先行詞がない場合に適切な解析を行うことが難しい. 一般に, 文脈内に候補を持たない省略語を予測するためには探索範囲を文脈外に広げる必要がある. しかし, 探索範囲を広げると, 仮に解決したい省略の候補集合が予め分かっている場合であっても, 候補語の数が多くなり, 選択はより難しくなる.

そこで, 研究では, 個々の省略語の出現確率を一様ではなく, 文脈に応じて変化すると考え, 言語モデルの一つであるトピックモデルを用いることで大域情報を利用した候補語選択の方法を検討する. また, 予め候補集合が与えられた状況下で, 候補語選択する評価実験を行い, 検討した提案方法の効果を確認する.

4.2　トピックモデルを利用した候補語選択

??節で解説したトピックモデルを用いることで, 省略を解決する候補語の出現確率を一様ではなく, 文脈に応じて変化すると考えることができる. 本研究では,

代表的なトピックモデルである LDA を利用する.

LDA では, 2.1.4 節で示したように, 文書集合を D とし, 各文書は固有のトピック比率 $\boldsymbol{\theta}_d$ を持つと仮定する. また, 文書 d 中の各単語 w は $\boldsymbol{\theta}_d$ に従いトピック k を選択した後, そのトピック k に固有の単語分布 $\boldsymbol{\phi}_k$ に従って生成されたと考える (図 2.4 参照). これにより, 単語 $w(\in V)$ の列によって表現された問い合わせテキストの集合 (大域情報) とトピック数 K を入力として, 各トピック $z_k(k = 1, ..., K)$ における単語 w の確率分布 $P(w|z_k)(w \in V)$ および各文書 d におけるトピック z_k の確率分布 $P(z_k|d)(k = 1, ..., K)$ を推定することができる.

本研究では, 省略を含む問い合わせテキスト d を入力とし, 与えられた I 個の候補語の集合 C の中から候補語 $w_i(w_i \in C, i, ...I)$ を選択する. なお, 候補集合 C の収集には 2.1 節で紹介した清田らの研究で提案されていた方法を用いることとし, 省語の種類や候補集合の収集方法については本研究では扱わない.

候補語の集合 C の中から候補語 $w_i(w_i \in C, i, ...I)$ を選択する方法については, 以下の 2 つの方法を検討した.

方法 1 N-gram モデルによる選択

方法 2 N-gram モデルと LDA モデルの組み合わせによる選択

まず, 方法 1 は, 入力した問い合わせテキスト d の中での候補語 w_i の出現確率 $P_{ngram}(w_i|d)$ を N-gram モデルによって算出し, 最も出現確率の高い候補語 w_i を選択した. N-gram モデルとは, 代表的な言語モデルで「ある時点での単語の生起確率は, その直前の N-1 個の単語にのみ依存する」と仮定し, 単語の生起を N 重マルコフ課程で近似したモデルである.

次に, 方法 2 は, 入力した問い合わせテキスト d の中での候補語 w_i の出現確率 $P_{lda}(w_i|d)$ を LDA モデルを用いて算出し, 方法 1 で算出した $P_{ngram}(w_i|d)$ と式 (4.1) の線形補間により組み合わせてスコア化し, 最もスコアの高い候補語 w_i を選択した. 式 (4.1) の線形補間係数 λ は, 0 から 1 までの 0.05 刻みで変化させた.

$$score_{w_i}(d) = \lambda P_{ngram}(w_i|d) + (1-\lambda)P_{lda}(w_i|d) \tag{4.1}$$

4.3 評価実験

4.3.1 実験方法

評価実験には,2.1 節で紹介した清田らの研究で提案されていた換喩表現ペアの抽出方法を利用した. これは換喩表現および換喩解釈表現の出現パタンに基づいたパタンマッチングを用いる方法である. まず,3.1 節と同じ方法で実験データに対して予め省略された表現を持つ「名詞 A–の–名詞 B–格助詞–動詞」と「名詞–格助詞–動詞」のペアを作成する. 次に, 得られたペアの中から, 一方の表現に存在し他方の表現に存在しない語の集合を抽出し, 省略された語すなわち名詞 B の候補語集合 C とする.

実験は, 評価用データ中に「名詞 A–の–名詞 B–格助詞–動詞」のパタンとマッチする表現が見つかった場合に, マッチした部分を人工的に「名詞 A–格助詞–動詞」に置換し, 置換によって欠落した部分を省略と見立て, 欠落した語を正解として予測する方法で行った.

実験データには,3 章と同じ, 収集期間 2004/4/1∼2009/4/7 の Yahoo! 知恵袋の質問データ 3,206,559 件の中からサブカテゴリーが「インターネットとパソコン」のデータ 576,841 件を使用した. この中から省略された表現を持つ「名詞 A–の–名詞 B–格助詞–動詞」と「名詞–格助詞–動詞」の 4,360 ペアを抽出した. この 4,360 ペアは,「名詞 A–の–名詞 B–格助詞–動詞」のパタンにマッチした 139,786 件と,「名詞–格助詞–動詞」のパタンにマッチした 289,711 件のうち, 低頻度の表現を除くため, それぞれ出現頻度が 3 回以上の表現のみ使用し,「名詞 A–の–名詞 B–格助詞–動詞」から人工的に名詞 B を除いた表現と「名詞–格助詞–動詞」の表現が一致するものから作成した.

なお, 表 4.1 は, 参考として, 抽出した 4,360 ペアを名詞 B を除いた表現である「名詞 A–格助詞–動詞」ごとに集計し, 出現頻度の上位の一部を示したものである. 表の見方は, 左から,「名詞 A –格助詞–動詞」タイプの省略表現, 省略表現に対応する省略前の表現の出現頻度, 出現頻度の構成比および省略表現に対応する名

詞Bの候補数を示した. 同様に, 表4.2は, 参考として, 名詞Bの候補語の一例として表4.1で示した省略表現のうち「元–に–戻す」の持つ39個の候補語について出現頻度の上位の一部を示したものである.

実験は, 576,841件の実験データのうち, 500,000件を100,000件ずつに5分割し, その4つをN-gramモデルおよびLDAモデルの推定用に, 残り1つはランダムに3,000件を選択し評価用として5分割交差検定を行った.

なお, LDAの実装にはGibbsLDA++[91]を用いた. LDAモデルの推定に必要なトピック数Kやディリクレ分布のハイパーパラメータα, βは, 予め実験データの一部を使用して, $K = \{30, 50, 150, 200\}, \alpha = \{0.01, 0.1, 0.5, 1, 1.5\}, \beta = \{0.01, 0.1, 0.5, 1, 1.5\}$の組み合わせでパープレキシティを測定し, 探索的にモデルの精度の良いトピック数$K = 150$とハイパーパラメータ$\alpha = 0.1, \beta = 0.01$に決定した. また, N-gramモデルの実装にはSRILMツールキット[92]を用いた. モデルは5gram確率を用い, パラメータには補間モデルにinterpolate, スムージングにkndiscountを使用した.

表4.3には, 各方式による候補語選択の正解数について, 候補語数の上位の一部を例示した.

4.3.2 結果と考察

評価用データ15,000件(3,000件を5セット)から,「名詞A–の–名詞B–格助詞–動詞」のパタンに一致する表現が1,005件見つかった. 評価実験の結果を表4.4および表4.5に示す.

表4.4は, 1,005件すべてについて, 方法1および2による候補語選択の正解率とベースラインに対する改善率を示したものである.

表4.5は, 1,005件のうち, 候補語が1件の場合を除いた806件について, 正解率とベースラインに対する改善率を示したものである.

なお, ベースラインは, それぞれのパタンについて, 最も高い頻度で名詞Bの位置に出現した候補語を選択した場合の正解率である. また, 方法2の(a)の$\lambda = 0.7$は, すべてのパタンで同じλを使用した場合に最も高い正解率を示した$\lambda = 0.7$の結果であり,(b)のベストλとは, パタンごとに異なるλを使用し, パタンごとで最も高い正解率を得るλを選んだ場合の結果である.

表 4.1: 実験データから抽出した省略表現の一部 (名詞 B を除いた形で集計)

順位	名詞 A –格助詞–動詞	出現頻度	構成比 %	名詞 B の候補数
1	パソコン – に – ついて	2824	6.95	206
2	パソコン – を – 入れる	1618	3.98	13
3	パソコン – を – 考える	796	1.96	11
4	元 – に – 戻す	714	1.76	39
5	パソコン – を – 切る	682	1.68	3
6	ノートパソコン – を – 考える	525	1.29	4
7	パソコン – が – 出る	520	1.28	4
8	パソコン – を – する	465	1.14	42
9	ノートパソコン – に – ついて	413	1.02	40
10	設定 – を – 教える	386	0.95	4
11	ＮＥＣ – を – 使う	353	0.87	13
12	パソコン – が – 入る	332	0.82	2
13	メモリ – に – ついて	328	0.81	20
14	メモリ – を – する	321	0.79	6
15	元 – に – 戻る	316	0.78	18
16	メモリ – を – 考える	313	0.77	3
17	パソコン – に – ある	310	0.76	30
18	パソコン – が – 壊れる	287	0.71	28
19	オススメ – を – 教える	266	0.65	26
20	ＣＰＵ – に – ついて	237	0.58	16
21	パソコン – が – 切れる	221	0.54	2
22	パソコン – を – 見る	220	0.54	28
23	パソコン – が – 落ちる	219	0.54	3
24	オススメ – が – ある	218	0.54	18
25	削除 – を – 教える	212	0.52	2
26	画面 – に – ある	210	0.52	17
27	中古 – を – 買う	209	0.51	8
28	ノートパソコン – が – 壊れる	203	0.50	15
29	エクセル – に – ついて	198	0.49	21
30	ＯＳ – を – する	198	0.49	6
:				

表 4.2: 候補語の一部 (表 4.1 中の “元–に–(名詞 B)–戻す” の場合)

順位	「元–の–(名詞 B)–に–戻す」の候補語	出現頻度	構成比
1	状態	220	30.81
2	サイズ	168	23.53
3	位置	38	5.32
4	画面	34	4.76
5	設定	26	3.64
6	場所	26	3.64
7	色	24	3.36
8	表示	16	2.24
9	メモリ	13	1.82
10	バージョン	12	1.68
11	パソコン	11	1.54
12	ＣＰＵ	10	1.40
13	ローマ字入力	9	1.26
14	アイコン	8	1.12
15	青色	7	0.98
16	メモリー	7	0.98
17	画像	6	0.84
18	ＸＰ	6	0.84
19	名前	5	0.70
20	フォルダ	5	0.70
21	黒	5	0.70
22	フォント	4	0.56
23	ＨＤＤ	4	0.56
24	下	4	0.56
25	小文字	4	0.56
26	壁紙	3	0.42
27	容量	3	0.42
28	ＩＥ６	3	0.42
29	Ｍｅ	3	0.42
30	ＯＳ	3	0.42
:			

表 4.3: パタンごとに集計した候補語選択の正解数 (一部)

				ベースライン (TF)	方法 1 (N-gram)	方法 2 (N-gram と LDA)	
No.	名詞 A–格助詞–動詞	候補数	評価数			(a) $\lambda = 0.7$	(b) ベスト λ
201	パソコン – に – ついて	206	76	5	10	19	21
217	パソコン – を – する	42	17	1	3	2	3
172	ノートパソコン – に – ついて	40	12	5	4	2	4
317	元 – に – 戻す	39	20	9	8	11	11
200	パソコン – に – ある	30	9	5	4	5	6
184	パソコン – が – 壊れる	28	5	0	2	2	2
223	パソコン – を – 見る	28	7	0	1	1	2
100	オススメ – を – 教える	26	6	2	1	2	3
198	パソコン – と – する	23	4	1	1	2	3
236	パソコン – を – 買う	22	13	2	4	5	6
93	エクセル – に – ついて	21	3	1	1	1	2
307	メモリ – に – ついて	20	7	6	6	6	6
39	ＨＤＤ – に – ついて	19	4	1	1	0	2
206	パソコン – に – 関する	19	3	1	1	1	1
99	オススメ – が – ある	18	3	0	1	0	2
225	パソコン – を – 使う	18	2	0	0	0	0
319	もと – に – もどる	18	9	1	6	5	6
353	画面 – に – ある	17	2	1	1	1	1
24	ＣＰＵ – に – ついて	16	9	2	3	3	3
247	ファイル – に – ついて	16	1	0	0	0	1
169	ノートパソコン – が – 壊れる	15	5	0	0	0	2
73	ＸＰ – に – ついて	14	1	0	0	0	1
105	キーボード – に – ある	14	3	1	1	1	1
222	パソコン – を – 教える	14	3	1	0	1	2
249	ファイル – を – する	14	9	4	4	5	5
49	ＮＥＣ – を – 使う	13	4	1	0	1	1
235	パソコン – を – 入れる	13	47	43	43	42	43
237	パソコン – を – 変える	13	5	2	2	2	2
31	ＤＶＤ – に – ついて	12	3	1	1	0	1
144	データ – に – ついて	12	2	0	0	1	1
:							

表 4.4: 候補語選択の正解率と改善率 (候補語の数=1 を含む 1,005 件)

			方法 2 (N-gram と LDA)	
方法	ベースライン (TF)	方法 1 (N-gram)	(a) $\lambda = 0.7$	(b) ベスト λ
正解率%	63.88	65.07	65.97	75.22
改善率%	-	1.19	2.09	11.34

表 4.5: 候補語選択の正解率と改善率 (候補語の数=1 を含まない 806 件)

			方法 2 (N-gram と LDA)	
方法	ベースライン (TF)	方法 1 (N-gram)	(a) $\lambda = 0.7$	(b) ベスト λ
正解率%	54.96	56.45	57.57	69.11
改善率%	-	1.49	2.61	14.14

表 4.4 を見ると, 方法 1 と方法 2 ともにベースラインの正解率 68.88% を上回っている. ベースラインに対する改善率は, 方法 2(b)> 方法 2(a)> 方法 1 の順で高く, 方法 2(b) の正解率は 75.22% で, ベースラインを 11.34% 上回っている. 表 4.5 についても同じ傾向が見られる. 方法 1 と方法 2 ともにベースラインの正解率 54.96% を上回っている. ここでも, ベースラインに対する改善率は, 方法 2(b)> 方法 2(a)> 方法 1 の順で高く, 方法 2(b) の正解率は 69.11% で, ベースラインを 14.14 %上回っている.

表 4.4 と表 4.5 を比較すると, すべての方法において, 候補語の数が 1 の場合を除いた表 4.5 の方が正解率が低いが, その差はベースラインの 8.92% に比べて方法 2(b) の 6.12%の方が小さい.

ベースラインと方法 1 の比較から, 候補語集合から最も出現頻度の高い語を選ぶ方法に比べて, N-gram モデルを使う方法がより高い候補語選択性能を示すことが確認できる.

また, 方法 1 と方法 2 の比較から, LDA モデルを N-gram モデルと組み合わせことで, N-gram モデルで解決できない曖昧性の問題を一部解消し, 候補語選択の性能を改善したと考えられる. それにより, 方法 2 は最も高い 75.22% の正解率を得ている.

次に, 効果とパタンとの関係について詳しく見る. 表 4.6 は, 最も高い候補語選

択性能を示した方法 2(b) の改善効果を示したもである.「名詞–格助詞–動詞」のパタンごとにベースラインと方法 2(b) による正解数を比較し, その差が大きい順の上位 30 件を例示している. 表の見方は, 2 列以降, 左から, パタンに対する候補語の数, 評価用データ中に見つかったパタンの数, ベースラインによる正解数, 方法 2(b) による正解数, 方法 2(b) で採用した λ 値, 改善数 (ベースラインとの差) である.

一見すると, 候補語の数が多ければ改善数も多く見えるが, 改善率で見るとその傾向は顕著ではない. さらに, 候補語の数, 評価数, 正解数, 候補語の数の分散 (表 4.6 上では割愛した), λ 値との間に強い相関は見られなかった. なお, 表 4.6 の 1 行目を外れ値として除いた場合も同様に各項目の間に強い相関は確認できなかった.

4.4 まとめ

2.1 節において, ユーザーから企業などへの問い合わせメールに対する返信を行うカスタマーサポート業務に着目し, ユーザーからの問い合わせメール中に多く発生する語の省略が, 背景知識や経験の少ない担当者とっては, 意図の取り違いや見落としの原因になること, また, 語の省略が「A の B」タイプの連体修飾関係中の名詞 B に多いことを指摘した.

一方で, 2.1 節では, 従来の方法では同じ文脈中に先行詞がない場合に適切な解析を行うことが難しいことも説明した. 一般に, 文脈内に候補を持たない省略語を予測するためには探索範囲を文脈外に広げる必要がある. しかし, 探索範囲を広げると, 仮に解決したい省略の候補集合が予め分かっている場合であっても, 候補語の数が多くなり, 選択はより難しくなる.

そこで, 本章では, 個々の省略語の出現確率を一様ではなく, 文脈に応じて変化すると考え, 言語モデルの一つであるトピックモデルを用いることで大域情報を利用した候補語選択の方法を提案し, 予め候補集合が与えられた状況下で, 候補語選択する評価実験を行い, 検討した提案方法の効果を確認した.

実験の結果, 候補語の集合が与えられるという条件の下で, 提案の方法は従来方法であるベースラインに比べて正解率で 11.34% の改善が見られ, 75% を超える高い候補語の選択性能を示すことを確認し, これらの結果から, トピックモデ

表 4.6: 候補語選択の正解数と改善数 (改善率の上位 30 件)

No.	名詞–格助詞–動詞	候補語数	評価数	TF	方法 2(b)	λ 値	改善数
1	パソコン – に – ついて	206	76	5	21	0.45	16
2	もと – に – もどる	18	9	1	6	1.00	5
3	パソコン – を – 買う	22	13	2	6	0.65	4
4	元 – に – 戻す	39	20	9	11	0.95	2
5	モニター – に – ついて	5	4	2	4	0.20	2
6	パソコン – を – する	42	17	1	3	1.00	2
7	パソコン – と – する	23	4	1	3	0.45	2
8	パソコン – を – 見る	28	7	0	2	0.20	2
9	パソコン – が – 壊れる	28	5	0	2	1.00	2
10	オススメ – が – ある	18	3	0	2	0.50	2
11	ノートパソコン – が – 壊れる	15	5	0	2	0.45	2
12	ＯＳ – を – 使う	2	2	0	2	1.00	2
13	設定 – を – 教える	4	8	6	7	1.00	1
14	パソコン – に – ある	30	9	5	6	0.65	1
15	ファイル – を – する	14	9	4	5	0.75	1
16	メモリ – を – 考える	3	5	4	5	0.85	1
17	オススメ – を – 教える	26	6	2	3	0.50	1
18	ＣＰＵ – に – ついて	16	9	2	3	1.00	1
19	ノートパソコン – を – する	5	3	2	3	1.00	1
20	対処 – を – 教える	2	5	2	3	0.95	1
21	パソコン – を – 打つ	2	3	2	3	1.00	1
22	エクセル – に – ついて	21	3	1	2	0.65	1
23	ＨＤＤ – に – ついて	19	4	1	2	0.35	1
24	パソコン – を – 教える	14	3	1	2	0.60	1
25	パソコン – を – 調べる	10	5	1	2	0.65	1
26	パソコン – に – よる	10	4	1	2	0.70	1
27	ＸＰ – を – 入れる	10	3	1	2	1.00	1
28	インターネット – を – 開く	6	2	1	2	0.95	1
29	パソコン – が – 使える	5	2	1	2	0.70	1
30	ヤフー – を – 開く	3	3	1	2	0.75	1

ル (LDA) を利用するにより, 従来の言語モデルである N-gram モデルを補完して候補語の選択性能を向上させる効果を確認した.

本研究では, λ 値を推定する方法については扱わなかったが, 表 4.4 や表 4.5 で示したように, 方法 2(b) の正解率を得るには, 個々の省略表現に最適な λ 値を決定する方法が必要である. また, トピック数 K やモデルのハイパーパラメータ α, β についても探索的に決定した. これらパラメータの導出や推定方法については, 今後も引き続き取り組む考えである.

第5章　ユーザー興味情報を利用した同名他社の判定

本章では,収集したツイートから目的と無関係のツイートを除去することを目的とし,ユーザー興味モデル(トピックモデル)を用いて,ツイートの表層に現れる手がかりのみに依らずに,ノイズを含んだ検索結果から目的のツイートを選り分ける方法を提案し,評価実験によって示された効果について述べる.

5.1　研究の目的

2.2 節で述べたように,キーワード検索によって目的のツイート集合を収集しようとすると,検索結果にはキーワードと同名の別の対象も含まれる. 例えば,企業の評判分析を行う場合に,このように同じ名前の別の企業名が含まれたツイートを収集してしまうと,投稿数の集計は不正確なものとなり,分析精度が低下する要因となる. そのため,収集したツイートから目的と無関係のツイートを「ノイズ」として除去することは極めて重要である.

一方で,ツイートの多くは文字制限から 1 件当たりの文長が短く,また口語的な崩れた表現を多く含んでいるため,テキスト中に出現する単語など表層的な情報のみによる解析は難しく,課題となっている.

そこで,本研究では,著者トピックモデルの考えに倣い,ユーザーの全ツイートを 1 文書として扱い,ユーザーの興味分布としてモデル化し,目的のツイートの判定に利用する. また,ユーザーの興味と話題のダイナミクスを考慮に入れ,期間の異なるツイートを用いた複数のトピックモデルを予備実験で比較し,良い性能を示した期間のツイートを使ったモデルを選択して用いる.

それにより,ツイート中の表層的な情報からツイートが目的のツイートかどうかを判定するだけでなく,ツイートを投稿したユーザーが目的のツイートを投稿

しやすいかどうかのユーザーの興味情報も利用し,これらノイズを含んだ検索結果から目的のツイートを選り分ける方法を検討し,評価実験によって検討した提案方法の効果を確認する.

5.2 ユーザー興味モデルを利用した方法

2.2節の例で述べたように,キーワード「アップル」で検索を行いコンピュータやディジタル家電メーカーの「Apple Inc.」に関するツイートを収集しようとすると「アップルティー」や「アップルジュース」などのフルーツの「アップル」以外にも,「アップル」を冠した別の企業名などが混入する.

本研究では,ツイート中の情報からツイートごとに目的のツイートかどうかを判定するのでなく,ツイートを投稿したユーザーがコンピュータやディジタル家電メーカーの「Apple Inc.」あるいはフルーツの「アップル」のどちらの発言をしやすいかをユーザーの興味情報を使って判定する.

5.2.1 ユーザー興味モデル

本研究では,ユーザーの興味情報を利用するため,著者トピックモデル[86]の考えのもと,ユーザーが過去に投稿した複数のツイートの集合を1文書として扱い,LDAを適用する.このトピックモデルを本研究では「ユーザー興味モデル」と呼ぶ.まず,このユーザー興味モデルについて,通常のLDAとの対比で説明する.

図5.1のように,通常のLDA(上側の図)では,文書は固有のトピック比率(各トピックの出現確率を表す)を持つと仮定し,各単語は各トピックに固有の単語分布に従って生成されたと考える.これに対し,ユーザー興味モデル(下側の図)では,ユーザーは固有のトピック比率(各トピックに興味を持つ確率を表す)を持つと仮定し,ユーザーがつぶやいた単語は,各トピックに固有の単語分布に従って生成されたと考える.

次に,ユーザー興味モデルと通常のLDAのグラフィカルモデルを図5.1に示す.通常のLDA(左側の図)では,文書集合をDとし,各文書は固有のトピック比率$\boldsymbol{\theta}_d$を持つと仮定する.また,文書d中の各単語wは$\boldsymbol{\theta}_d$に従いトピックkを選択した後,そのトピックkに固有の単語分布$\boldsymbol{\phi}_k$に従って生成されたと考える.これに対

し, ユーザー興味モデル (右側の図) では, ユーザー集合を U とし, 各ユーザーは固有のトピック比率 $\boldsymbol{\theta}_u$(ユーザー u が各トピックに興味を持つ確率を表す) を持つと仮定する (図 5.1). ユーザー u がつぶやいた単語 w は $\boldsymbol{\theta}_u$ に従いトピック k を選択した後, そのトピック k に固有の単語分布 $\boldsymbol{\phi}_k$ に従って生成されたと考える.

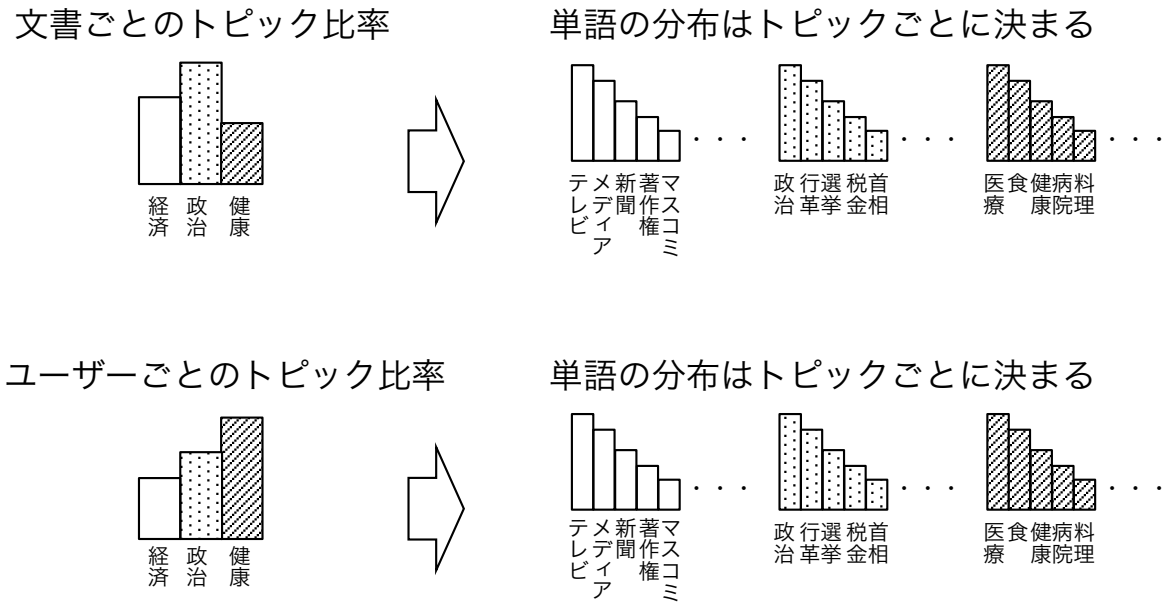

図 5.1: 通常の LDA(上) とユーザー興味モデル (下) のイメージ

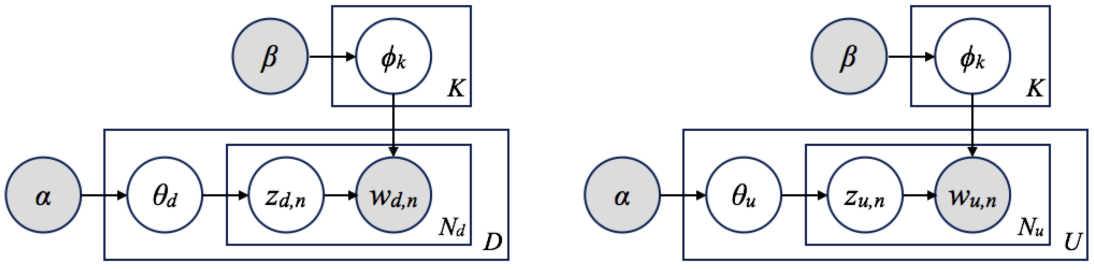

図 5.2: 通常の LDA(左) とユーザー興味モデル (右) のグラフィカルモデル

なお, このグラフィカルモデルに従った, ツイートの生成過程は以下のようになる.

LDA を利用して, ユーザー興味モデルの学習することで, ユーザーがトピック

1. 各トピック $k = 1, \cdots, K$ について:
   (a) ディリクレ分布に従って単語分布 $\boldsymbol{\phi}_k$ を生成
   $\boldsymbol{\phi}_k \sim Dir(\beta)$
2. 各ユーザー $u = 1, \cdots, U$ について:
   (a) ディリクレ分布に従ってトピック分布 $\boldsymbol{\theta}_u$ を生成
   $\boldsymbol{\theta}_u \sim Dir(\alpha)$
   (b) ユーザーの過去ツイート中の各単語 $w_{u,n}$ $(n = 1, \cdots, N_u)$ について:
      i. 多項分布に従ってトピックを生成
      $z_{u,n} \sim Multi(\boldsymbol{\theta}_u)$
      ii. 多項分布に従って単語を生成
      $w_{u,n} \sim Multi(\boldsymbol{\phi}_{z_{u,n}})$

に興味を持つ確率(表5.1)やトピックごとの単語の出現確率(表5.2)を推定することができる. (各表は, トピック数 $K = 150$ とした場合のイメージ)

表5.1: ユーザーがトピックに興味を持つ確率の行列(イメージ)

ユーザーID	トピック1	トピック2	···	トピック150
ユーザーA	0.01	0.05		
ユーザーB	0.05	0.01		
ユーザーC	0.01	0.01		
:				
ユーザーZ				

以降では, このユーザー興味情報を使って, ユーザーが目的のツイートを発言しやすいかどうかを判定する方法について述べる.

5.2.2 同名他社の判定

次に, LDAによって推定したユーザー興味モデルを利用して, 各ツイートが目的のツイートかどうかを判定する. しかしながら, ユーザーが投稿するツイート

表 5.2: トピックごとの単語の出現確率の行列 (イメージ)

トピック#	単語 1	単語 2	···
トピック 1	0.01	0.05	
トピック 2	0.05	0.01	
トピック 3	0.01	0.01	
:			
トピック 150			

の内容は, ユーザーの潜在的な興味情報だけで決まるとは考え難い.

そこで, 本研究では, ツイート中の表層的な情報とユーザーの潜在的な興味情報の両方を組み合わせて利用する. 組み合わせのステップに沿って以下の 3 つの方法を用いる.

方法 1 ツイートの表層的な情報を利用した分類器で判定する

方法 2 ツイートを投稿したユーザーの興味モデルを使った分類器で判定する

方法 3 方法 1 と方法 2 による分類確率を線形補間したスコアで判定する

方法 1 では, まず, 1 ツイートを 1 文書として扱い, 文書中の単語重要度を利用した文書ベクトルを作成する. 次の文書ベクトルを素性とした分類器を作成し, 機械学習によって各ツイートが目的のツイートである確率を見積もる.

方法 2 では, LDA で推定したユーザー興味モデルを利用する. ユーザーごとのトピック構成比を素性とした分類器を作成し, 機械学習によって各ツイートが目的のツイートである確率を見積もる.

方法 3 は, 方法 1 の分類器で見積もった確率 $P_{baseline}$ と方法 2 の分類器で見積もった確率 P_{lda} を式 (5.1) の線形補間で組み合わせスコアを算出する.

$$Score = \lambda P_{baseline} + (1 - \lambda) P_{lda} \tag{5.1}$$

方法 1 と 2 の確率および方法 3 で求めたスコアは 0~1 の間の値を取る. 今回は 2 値の識別のため,「Apple Inc.」に関するツイートである確率およびスコアが 0.5

以上の場合に「Apple Inc.」に関するツイートであると判定する.

5.3 評価実験

5.3.1 実験方法

評価実験は, 2014/1/4 ～ 2014/1/11 に投稿されたキーワード「アップル」を含む日本語のツイートを使用し, 収集したツイートが「Apple Inc.」の製品やサービスに関するものか, フルーツなどそれ以外の「アップル」についてのものかを機械学習を使って判定する方法で行う.

また, 比較のため, 2014/1/4 ～ 2014/1/11 に投稿されたキーワード「キリン」を含む日本語のツイートを使用した. このデータにより, 収集したツイートが「飲料メーカーのキリン」の製品やサービスに関するものか, 動物などそれ以外の「キリン」についてのものかの判定を同じ方法で行う.

なお, ユーザー興味モデルは, ユーザーの興味と話題のダイナミクス (時間的変化) を考慮するため, 予め実施した予備実験で, 投稿期間の異なるデータで学習した 4 種類のモデルの中から, 最も良い性能を示した期間のモデルを選び, 使用する. そのため, 実験は次の 3 ステップで進める.

ステップ 1: 実験データおよびユーザーごとの過去ツイート収集

ステップ 2: ユーザー興味モデルの学習とダイナミクスを考慮したモデル選定

ステップ 3: 選定した LDA モデルを利用したツイートの判定

以下, 各ステップについて説明する.

ステップ 1: 実験データおよびユーザーごとの過去ツイート収集

実験データは, Twitter API を使って収集し, キーワード「アップル」にマッチした 179,079 ツイートから 10,000 ツイートをランダムにサンプルリングして使用した. 次に, 10,000 ツイートの中から PR や bot を除去した上で, 各ツイートの投稿

者について, 過去 1 年を遡ってツイートが収集できた 855 ユーザーによる 904 件を残した.

これに予め人手で「Apple Inc.」に関するツイートかそれ以外かの 2 種類の正解ラベルを付与し, どちらとも区別の付かない 57 件は除外し, ラベルが付与できた 847 件の評価用データを作成した. 評価用データ 847 件から 802 ユーザーを抽出した. 表 5.13 には, 抽出した 847 件の原文の一部を例示し, 表 5.3 にはそれを内容を分類し, 集計した結果を示した. コンピューターやディジタル家電メーカーの「Apple Inc.」に関する製品やサービスに関するツイートの収集を目的とした場合, 目的のツイート (正例) は 70.0% 含まれていた.

表 5.3: キーワード「アップル」で収集した評価用データの構成

内容	件数	構成比 %
アップル (Apple Inc. の意味)	593	70.0
アップルティー	33	3.9
アップルジュース	14	1.7
その他のアップル	207	24.4
計	100	

また, 比較のため, Twitter API を使って収集したキーワード「キリン」にマッチした 101,114 ツイートから 10,000 ツイートをランダムにサンプルリングして使用した. 次に, 10,000 ツイートの中から PR や bot を除去した上で, 各ツイートの投稿者について, 過去 1 年を遡ってツイートが収集できた 769 ユーザーによる 807 件を残した.

これに予め人手で「飲料メーカーのキリン」に関するツイートかそれ以外かの 2 種類の正解ラベルを付与し, どちらとも区別の付かない 40 件は除外し, ラベルが付与できた 767 件のテストデータを作成した. テストデータ 767 件から 730 ユーザーを抽出した. 表 5.6 には抽出した 767 件の原文の一部を例示し, 表 5.5 にはそれを内容ごとに分類し, 集計した結果を示した.「飲料メーカーのキリン」の製品やサービスに関するツイートの収集を目的とした場合, 目的のツイート (正例) は 24.8%含まれており, データ中の偏りの程度については「アップル」の実験

表 5.4: キーワード「アップル」で収集した評価用データの例

No.	原文	ラベル
1	ベビーカーで行列の馬鹿親！林檎信者の馬鹿さ加減は想像以上 RT ASCII.jp：行列 800 人超、アップル福袋「Lucky Bag」2014 前日レポート (2/2) http://t.co/xxxx @xxxx さんから	Apple Inc.
2	バラすのは容易そうだし格好良いな。独自形状な GPU 周りの延命が気になるトコだが。: アップル「Mac Pro」、「修理しやすさ」で高評価–iFixit が分解 http://t.co/xxxx	Apple Inc.
3	でかした！サカナクション。アップルのマーク、アピってましたねぇ。（笑）	Apple Inc.
4	意味不明で笑った。RT @xxxx こういうところに関係改善の道があるのかも。アップル、グーグルに対抗し、日中韓で新しいスマホ OS を開発、来月発表見通し。そう、アジアの中でいがみ合うより、スクラム組んでに米と闘おう。 http://t.co/xxxx …	Apple Inc.
5	いろいろなアクセサリーや周辺機器が入っていて、かなりお得ですね。|大当たりは 16 万超え!?アップル Lucky Bag がスゴイ！ http://t.co/xxxx	Apple Inc.
6	もうすぐ !!!! 緊張してきた !!! #Apple #luckybag2014 #AppleStore 天神 #アップル #福岡 http://t.co/xxxx	Apple Inc.
7	高野さんがアップルティーの茶葉買って来てくれたよ！アップルティー大好きだから素直に嬉しい (*^_^*)！色々プレゼントくれるから京都土産で何かお礼をしよう。	その他
8	【驚愕】アップルラッキーバッグに入っていた T シャツが絶望的にダサい！ パジャマにするしかないレベル http://t.co/xxxx 林檎信者はこれ着て「なんだかんだ言ってもアップルってやっぱセンス良くてオシャレだよね〜」って夢から目を覚まさないで居てて欲しい w	Apple Inc.
9	アップルのタブレットは 20000 円だけど何入ってたんだろう。前の mini 以外ならかなり奮発してるけど	Apple Inc.
10	うちのリーダーがデバッグを開始しました笑 #Apple #luckybag2014 #AppleStore 天神 #アップル #福岡 http://t.co/xxxx	Apple Inc.
11	RT @xxxx: 【明日から】アップルの新社屋設計者ノーマン・フォスターのドキュメンタリー映画がロードショー http://t.co/xxxx http://t.co/xxxx	Apple Inc.
12	コーニングの曲面ガラス―アップルにとって何を意味するか http://t.co/xxxx	Apple Inc.
13	アイフォン様というかアップル様に対しよく分からない敷居の高さを感じる小市民_(:3」∠)_	Apple Inc.
14	@xxxx: 今日、アップルティー 2 口とアルフォートしか食べてない。笑	その他
15	RT @xxxx: 正方形を 31 個見つけたら人材、35 個は天才、40 個以上はアップル社の面接行ったほうがいいらしい（笑） http://t.co/xxxx	Apple Inc.
:		

データと類似するが,正例の数と負例の数の比は「アップル」の実験データと対照的である.「キリン」のケースと比較することで,「アップル」の実験データのように,現実のデータで起こりやすいデータの不均衡による影響を補足的に確認する.

表 5.5: キーワード「キリン」で収集したテストデータの構成

内容	件数	構成比 %
動物	299	39.0
飲料メーカー	190	24.8
ゲーム (モンハン)	134	17.4
ゲーム (その他)	38	5.0
その他のキリン	106	13.8
計	100	

ステップ 2: ユーザー興味モデルの学習とダイナミクスを考慮したモデル選定

LDA の学習データは,実験データ中で過去 1 年を遡ってツイートが収集できた 855 ユーザーの過去 1 年間に投稿された 1,151,739 ツイートを再収集して利用した. ここから直近 1ヶ月分,直近 3ヶ月分,直近 6ヶ月分 1 年分の 4 種類の学習データを用意した. なお,語彙は,形態素解析器 MeCab[93] を利用して一般名詞と固有名詞のみを抽出した. なお,新語や流行語に対応するため,ユーザー辞書には日本語の Wiki タイトルを一般名詞として追加して使用した.

LDA の学習は,期間の異なる 4 種類の学習データに対し,岩田ら [94] に倣い Collapsed ギブスサンプリング [18] を用い,また,ハイパーパラメータ α, β はサンプリングが行われるごとに不動点反復法により推定した. トピック数 K は事前実験による比較検討でパープレキシティ値によるモデルの安定性と処理時間の観点から $K = 150$ に決定した.

LDA モデルの学習で得られた直近 1 ヶ月分,直近 3 ヶ月分,直近 6 ヶ月分,過去 1 年分の 4 種類モデル内のユーザーごとの興味分布を素性に用い,正解を付与した評価用データを使用して,分類器による 4 つのモデルの判定性能を確認する予備実験を行った. 分類器にはデータマイニングソフトウェア WEKA を使用し,機械

表 5.6: キーワード「キリン」で収集したテストデータの例

No.	原文	ラベル
1	今夜のお酒は、キリン 一番搾りです。いただきます☆	飲料メーカー
2	@xxxx あはは。ちなみに私もヴィーナスよりこっちのほうが好きです。あとキリン・シティも大好きです。	飲料メーカー
3	キリンの東京ばな奈いただきます http://t.co/xxxx	その他
4	新年早々キリン一番搾りがうめぇよ !!!!	飲料メーカー
5	どうでしょう新作一挙大放送を観ながら誕生日祝いなう。キリンいっぱい出せるよ w http://t.co/xxxx	その他
6	キリンさんに乗ってみたかった http://t.co/xxxx	その他
7	RT @xxxx: 2014.01.23(木) SOUND CRUE 開場 18:30 開演 19:00 1,500 円 (1 ドリンク別途 500 円) 新年一本目、あけましてライブ!メンバー全員、和装で演奏したいと考えています〜！どうかな〜？ キリン組の出番は最後、22:00…	その他
8	@xxxx キリンのものづくりにかける想いを、ぜひ実感してみてください。	飲料メーカー
9	これは実在するゲームなのか?(笑) 競馬にシマウマとかキリンとか走ってるぞ (笑)http://t.co/xxxx	その他
10	皆さんからたくさんのグランドキリンがっ!落ちついたとこでやってみるよ、あたりますよーに！	飲料メーカー
11	古代ローマ キリン食べていた - Y!ニュース http://t.co/xxxx	その他
12	@xxxx: ハンマー目当てなんだけどキリンのギルクエ持ってないんだよなあ。シャガルより楽なのかな？1805-3103-2390	その他
13	キリン株式会社さんから頂いたビールとつまみで夕飯。	飲料メーカー
14	キリンの睡眠時間は１５分だとか…寿命はどれくらいなのかな？	その他
15	この Blog で加害者がなぜか被害者意識を持っていると不思議がっているが、それは不思議じゃない。http://t.co/xxxx RT @xxxx: 性暴力加害者の心理 - キリンが逆立ちしたピアス http://t.co/xxxx	その他
16	RT @xxxx: 【30,000 本の年賀企画、開始！】年賀メッセージを友だちに贈って、受け取った友だちが抽選にチャレンジ！当選するとグランドキリン ジ・アロマの無料受取チケットをプレゼント。 http://t.co/xxxx	飲料メーカー
17	@xxxx キリン行こーぜー！今年もよろしく！	不明
18 :	古代ローマ キリン食べていた - Y!ニュース http://t.co/xxxx	その他

学習スキーマは事前実験による比較検討で高い性能を示した Sequential Minimal Optimization (SMO) を選択し, 10 分割交差検定によって評価した. なお, その他のオプションについてはデフォルトのままとした. 「アップル」についての予備実験の結果を表 5.7 に示す.

表 5.7: 期間が異なる学習データを用いた 4 モデルの比較 (アップル)

	直近 1ヶ月	直近 3ヶ月	直近 6ヶ月	直近 1 年分
正解率%	75.21	75.95	78.34	76.98
語彙数の平均	505	1,358	2,625	4,899

同様に,「キリン」のケースについて, 同じ作業を行った. LDA の学習データは, 実験データ中で過去 1 年を遡ってツイートが収集できた 769 ユーザーの過去 1 年間に投稿された 1,337,449 ツイートを再収集して利用した. ここから直近 1ヶ月分, 直近 3ヶ月分, 直近 6ヶ月分 1 年分の 4 種類の学習データを用意した.

実験データからの語彙の抽出や LDA の学習, 予備実験の方法およびラメータはすべて「アップル」による実験と同じとした. 「キリン」のケースについての予備実験の結果を表 5.8 に示す.

表 5.8: 期間が異なる学習データを用いた 4 モデルの比較 (キリン)

	直近 1ヶ月	直近 3ヶ月	直近 6ヶ月	直近 1 年分
正解率%	75.79	77.71	79.79	76.66
語彙数の平均	435	1,212	2,356	4,431

一般に, ユーザーの興味と話題のダイナミクスを考慮すると, 新しいツイートのみを用いた方が判定性能は高くなると考えられるが, その傾向を表 5.7 および表 5.8 上の「直近 6ヶ月」と「過去 1 年分」の違いに見ることができる. 一方で,「直近 1ヶ月分」,「直近 3ヶ月分」,「直近 6ヶ月分」の 3 つを比較すると期間が長

くなるにつれて,正解率が改善している. これは,長い期間のツイートを学習することで,ダイナミクスの影響よりも,語彙の増加や興味分布を特徴づけるイベントの発生を捉えていることを示していると考えられる.

参考として,「アップル」による実験で収集した12ヶ月分のツイートで学習したユーザー興味モデル(トピックモデル)を使い1ヶ月分 × 12個のツイートに対して推定した各トピックの出現確率のの変化を時系列でプロットしたものを図5.5に示した. また,このモデルによって抽出されたトピックワードの一部(全150トピックのうちツールの出力順に16トピックを選び,そのトピック中で出現頻度の高いものから10件を表示した)を図5.3に例示した. 図5.5のグラフは左から,全ユーザーの平均,「Apple Inc.」のラベルを付与したユーザーの平均,「その他」のラベルを付与したユーザーの平均に並べている.

同様に,「キリン」のケースについて,12ヶ月分のツイートで学習したユーザー興味モデル(トピックモデル)を使い1ヶ月分 × 12個のツイートに対して推定した各トピックの出現確率のの変化を時系列でプロットしたものを図5.6に示している. また,このモデルによって抽出されたトピックワードの一部(全150トピックのうちツールの出力順に16トピックを選び,そのトピック中で出現頻度の高いものから10件を表示した)を図5.4に例示した. 図5.6のグラフは左から,全ユーザーの平均,「飲料メーカーのキリン」のラベルを付与したユーザーの平均,「その他」のラベルを付与したユーザーの平均に並べている.

図5.5や図5.6のみによって,適切な期間のモデルを選定することは難しいが,例えば「アップル」のケースであれば2013年9月の「新型iPhone」の発売,「キリン」のケースであれば2013年9月のゲームソフト「モンスターハンター4」(ゲーム中で登場するモンスター名に「キリン」がある)の発売などイベントの考慮が必要であることが覗える.

本実験では,「アップル」と「キリン」の両方のケースによる予備実験で示された結果から,直近6ヶ月のLDAモデルを利用することとした.

また,(1)～(2)で説明した方法に沿ったデータ加工の流れを「アップル」と「キリン」のそれぞれについて,図5.7および図5.8に示した.

ステップ3: 選定したLDAモデルを利用したツイートの判定

実験では, 方法 1 でまず, ツイート中の表層的な情報を利用した分類器を使って行う. 1 ツイートを 1 文書とみなして実験データの 847 文書から一般名詞と固有名詞のみを抽出し, TF-IDF を特徴量として文書ベクトルを作成した. なお, 次元数は SVD (特異値分解) を使用して, LDA のトピック数 k と同じ 150 に縮約した. 次に, 文書ベクトルを素性に用いた分類器を作成し, 機械学習によりそのツイートが「Apple Inc.」に関するツイートである確率を見積もった. 機械学習の分類器には前述の予備実験と同様に SMO を利用し, インスタンスごとの確率は WEKA の buildLogisticModels オプションを使い Logistic Model にフィットさせる方法で見積もった. また, 10 分割交差検定は乱数を使用するため,10 種類の異なるシードを使いその結果を平均した.

方法 2 は, 推定した LDA モデルを利用した分類器を使って行う. LDA モデルで推定したユーザーごとのトピック構成比を素性に用いた分類器を作成し, 機械学習によりそのツイートが「Apple Inc.」に関するツイートである確率を見積もった. 素性以外の分類器の使い方やオプションは方法 1 と同じである.

方法 3 は, 方法 1 と方法 2 のそれぞれで見積もったツイートごとの「Apple Inc.」を選択する確率を線形補間したスコアを求めた. なお, 補間係数 λ は $0 \sim 1$ までの 0.05 刻みですべての λ を探索的に試した結果, 最も正解率の高い 0.4 を使用した.

同様に,「キリン」のケースについても同じ作業を行った. 方法 1 でまず, ツイート中の表層的な情報を利用した分類器を使って行う. 1 ツイートを 1 文書とみなして実験データの 767 文書から一般名詞と固有名詞のみを抽出し, TF-IDF を特徴量として文書ベクトルを作成した. なお, 次元数は SVD を使用して, LDA のトピック数 k と同じ 150 に縮約した. 次に, 文書ベクトルを素性に用いた分類器を作成し, 機械学習によりそのツイートが「飲料メーカーのキリン」に関するツイートである確率を見積もった. 機械学習の分類器には SMO を利用し, インスタンスごとの確率は WEKA の buildLogisticModels オプションを使い Logistic Model にフィットさせる方法で見積もった. また, 10 分割交差検定は乱数を使用するため,10 種類の異なるシードを使いその結果を平均した.

方法 2 は, LDA モデルで推定したユーザーごとのトピック構成比を素性に用いた分類器を作成し, 機械学習によりそのツイートが「飲料メーカーのキリン」に関するツイートである確率を見積もった. 素性以外の分類器の使い方やオプショ

ンは方法1と同じである.

方法3は,方法1と方法2のそれぞれで見積もったツイートごとの「飲料メーカーのキリン」を選択する確率を線形補間したスコアを求めた. なお,補間係数λは0 ～ 1までの0.05刻みですべてのλを探索的に試し,最も正解率の高い0.5を使用した.

5.3.2 結果と考察

ツイート中の表層的な情報をのみを利用した方法1をベースラインとして,LDAのみ利用した方法2および線形補間した方法3の各方法で判定した結果のサマリーを表5.9と図5.7に,図5.11の分割表に従って集計したそれぞれの数と求めた正解率,適合率,F値を表5.10に示す.

表5.9: ユーザー興味情報を利用したツイート判定結果のサマリー(アップル)

	方法1(ベースライン)	方法2(LDAのみ)	方法3(線形補間)
正解数	704	664	776
正解率%	83.31	78.58	91.83
FN数	115	64	17
FN率%	19.39	10.79	2.87
FP数	28	117	52
FP率%	11.02	46.43	20.63

表5.10: ユーザー興味情報を利用したツイート判定結果の詳細(アップル)

	TP	TN	FN	FP	正解率	適合率	再現率	F値
方法1	478	226	115	28	0.83	0.94	0.81	0.87
方法2	529	135	64	117	0.79	0.82	0.89	0.85
方法3	576	200	17	52	0.92	0.92	0.97	0.94

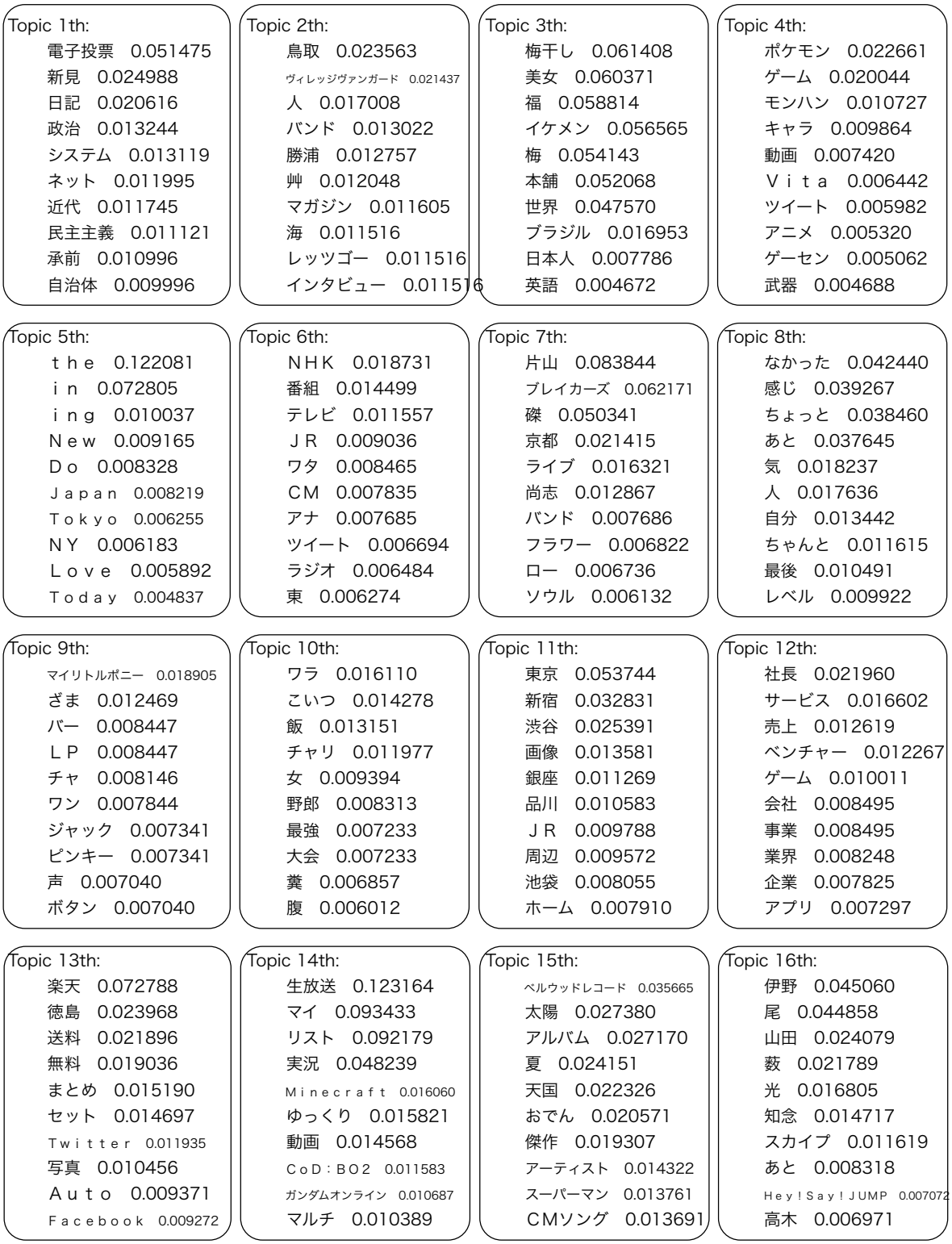

図 5.3: 抽出したトピックワードの例 (アップル)

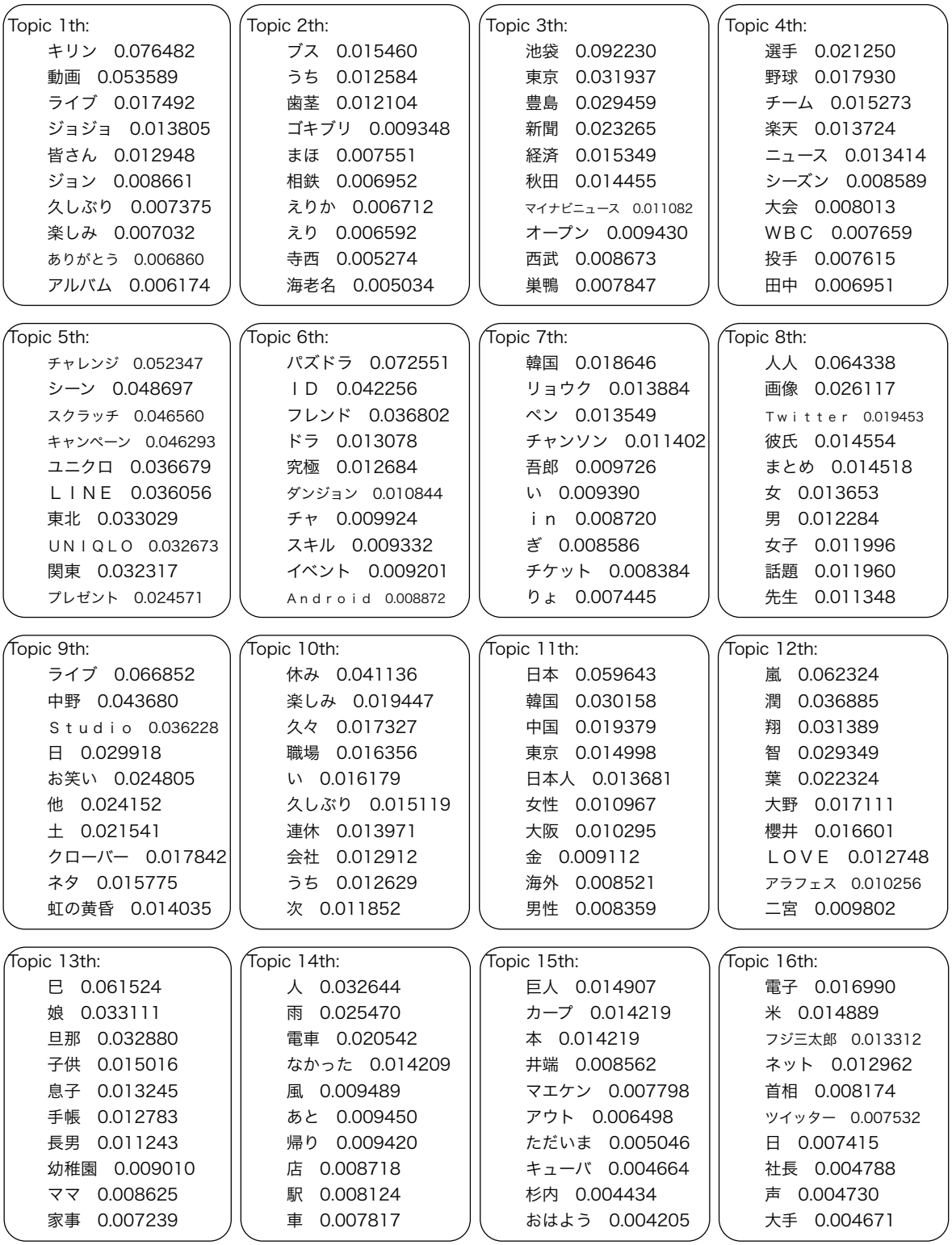

図 5.4: 抽出したトピックワードの例 (キリン)

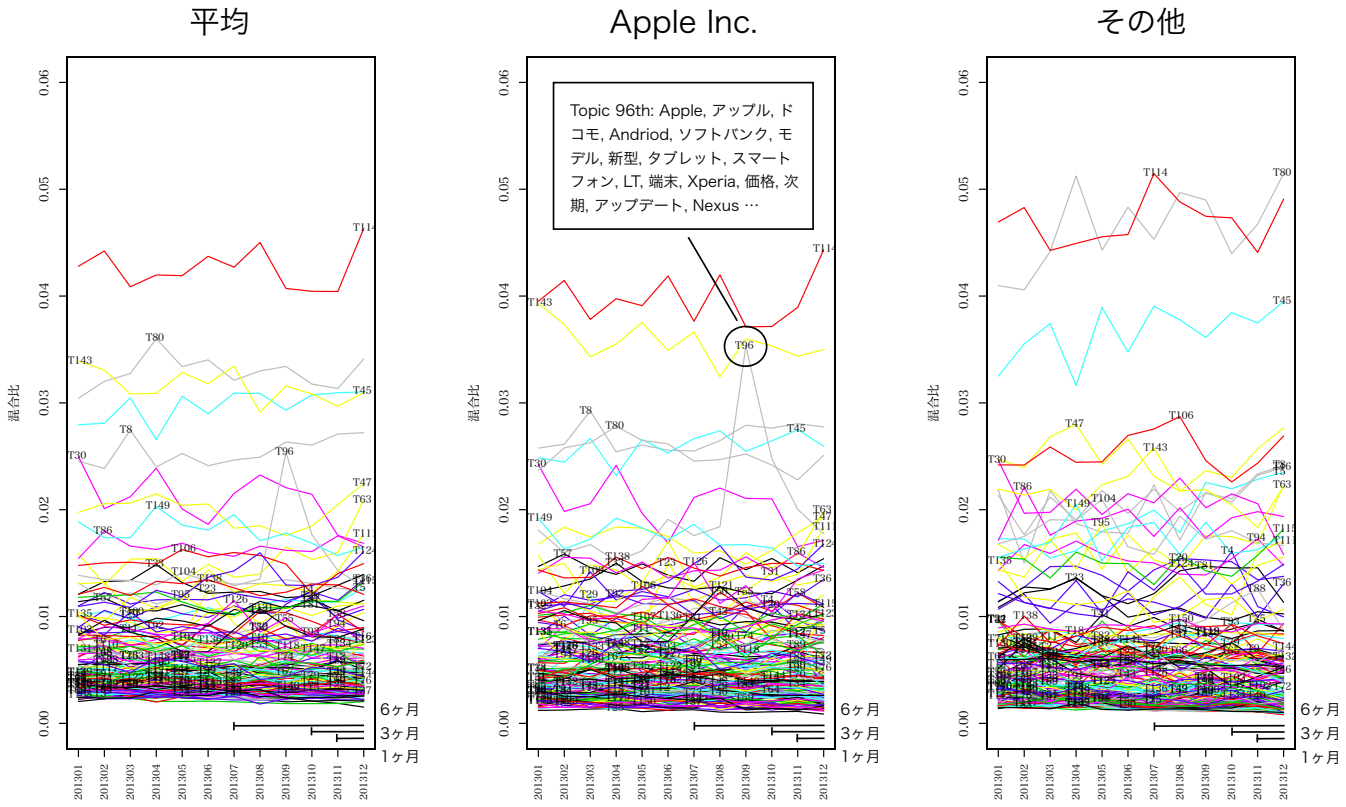

図 5.5: トピックの時系列変化 (アップル)

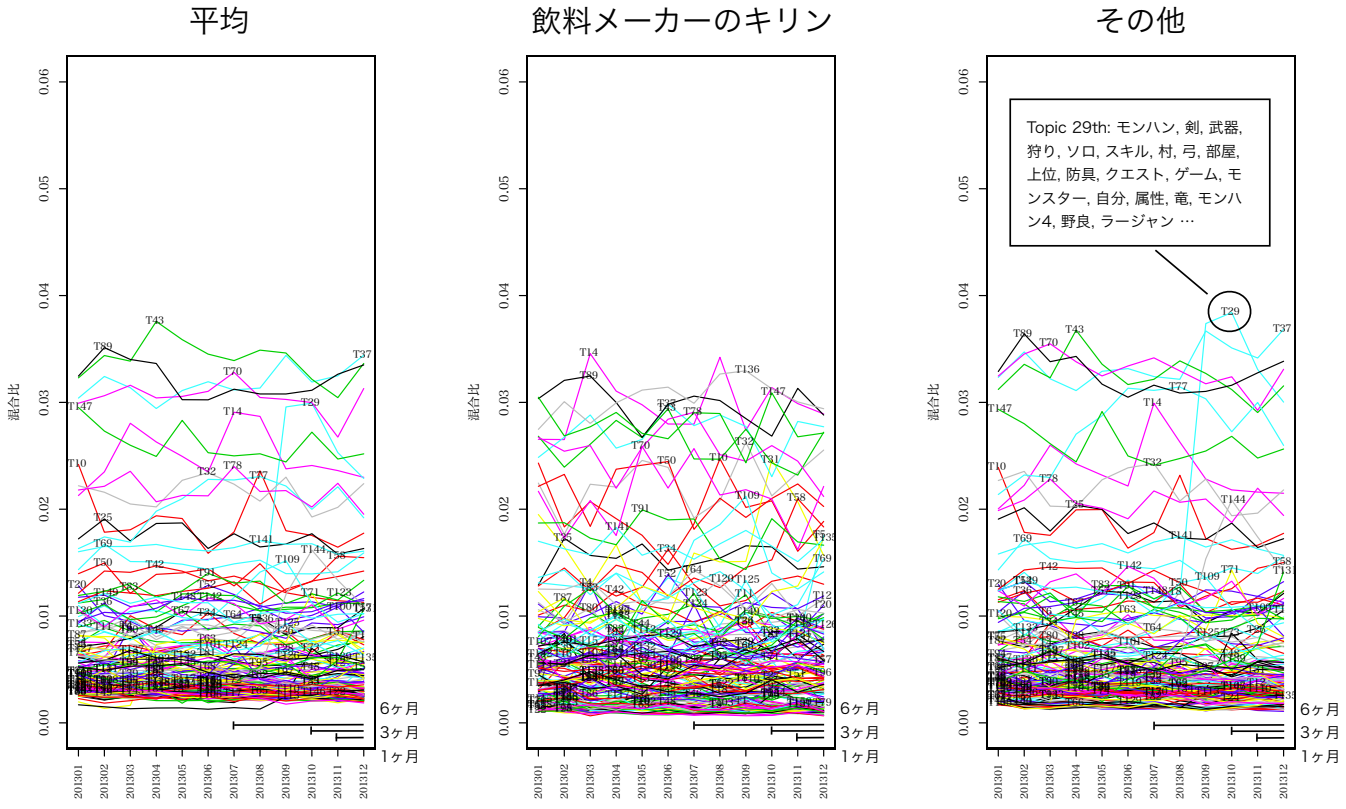

図 5.6: トピックの時系列変化 (キリン)

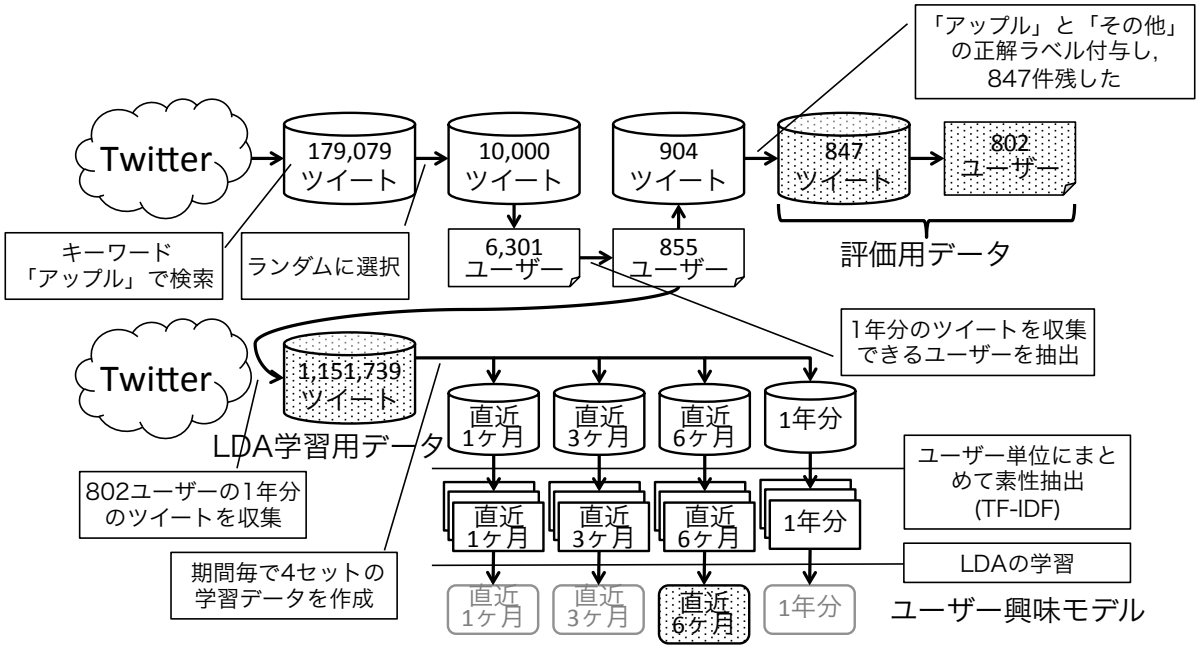

図 5.7: 期間の異なるユーザー興味モデルの学習方法 (アップル)

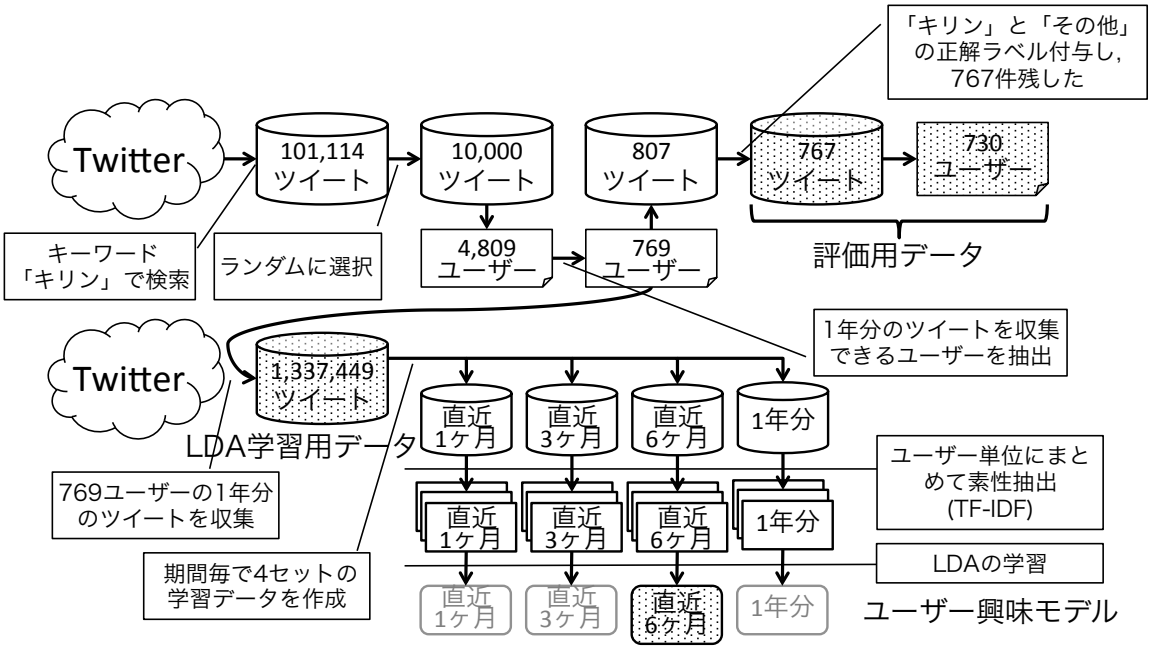

図 5.8: 期間の異なるユーザー興味モデルの学習方法 (キリン)

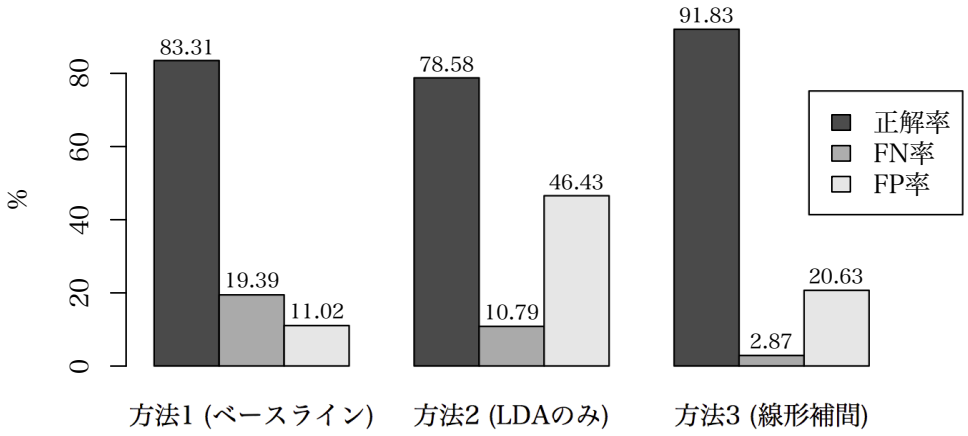

図 5.9: ユーザー興味情報を利用したツイート判定結果のサマリー (アップル)

表 5.11: 分割表の凡例 (「Apple Inc.」とその他に分類する場合)

	Apple Inc. に分類	その他に分類
Apple Inc.	TP (True positive)	FN (False negative)
その他	FP (False positive)	TN (True negative)

表5.9と図5.7の正解率を見ると,方法2はベースラインの方法1に比べて低い. 一方で,方法3はベースラインの方法1に比べて正解率が8.52%改善している. また,表5.10を見ると,「Apple Inc.」として収集したいツイートがその他に分類された見落としの数を示すFN数は,方法3ではベースラインの方法1に比べて98件改善している. 改善率は16.53%である. 一方,「Apple Inc.」以外のツイートが混入した数を示すFP数は,方法3ではベースラインの方法1に比べて24件(9.61%)増加している. これはFN数の改善に比べて小さい.

方法1と方法2の比較から,投稿したユーザーの興味情報のみを利用した方法は,ツイート中の表層的な情報のみを利用した方法に比べて正解率が低いことが確認できる. これは,ユーザーが投稿するツイートの内容が「Apple Inc.」に関するものかそれ以外かをユーザーの潜在的な興味情報だけで判定することができないことを示していると考えられる.

また,方法1と方法3の比較から,ツイート中の表層的な情報とユーザーの潜在的な興味情報の両方を利用する方法は,ツイート中の表層的な情報のみを利用した方法に比べて正解率が高い. これはツイート中の表層的な情報だけでは判定できないケースにおいて,ユーザーの潜在的な興味情報を組み合わせることで判定性を高める効果があることを示していると考えられる.

次に,「キリン」のケースについても同様に,ツイート中の表層的な情報をのみを利用した方法1をベースラインとして,LDAのみ利用した方法2および線形補間した方法3の各方法で判定した結果のサマリーを表5.12と図5.10に,図5.11の分割表に従って集計したそれぞれの数と求めた正解率,適合率,F値を表5.13に示す.

表5.12と図5.10の正解率を見ると,方法2は「アップル」の実験と同様に,ベースラインの方法1に比べて低い. 方法3も「アップル」の実験と同様に,ベースラインの方法1に比べてその差は小さいが正解率が改善している.

また,表5-11を見ると,「飲料メーカーのキリン」以外のツイートが混入した数を示すFP数は,方法3がベースラインの方法1に比べて9件(1.56%)改善している. 一方,「飲料メーカーのキリン」として収集したいツイートがその他に分類された見落としの数を示すFN数は,方法3がベースラインの方法1とほぼ同程度(1件増加)になっている.

表 5.12: ユーザー興味情報を利用したツイート判定結果のサマリー (キリン)

	方法 1 (ベースライン)	方法 2 (LDA のみ)	方法 3 (線形補間)
正解数	669	593	677
正解率%	87.22	77.31	88.27
FN 数	70	119	71
FN 率%	36.84	62.63	37.37
FP 数	28	55	19
FP 率%	4.85	9.53	3.29

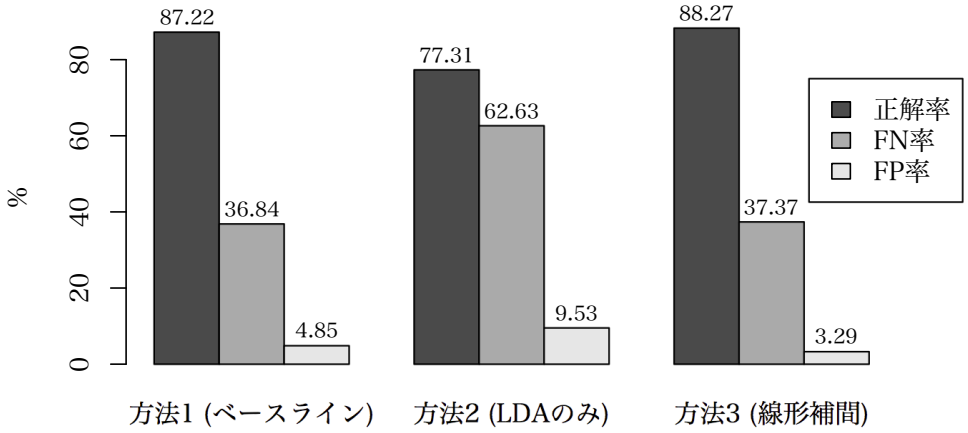

図 5.10: ユーザー興味情報を利用したツイート判定結果のサマリー (キリン)

表 5.13: ユーザー興味情報を利用したツイート判定結果の詳細 (キリン)

	TP	TN	FN	FP	正解率	適合率	再現率	F 値
方法 1	120	549	70	28	0.87	0.81	0.63	0.71
方法 2	71	522	119	55	0.77	0.56	0.37	0.45
方法 3	119	558	71	19	0.88	0.86	0.63	0.73

表 5.14: 分割表の凡例 (「飲料メーカーのキリン」とその他に分類する場合)

	飲料メーカーのキリン に分類	その他に分類
飲料メーカーのキリン	TP (True positive)	FN (False negative)
その他	FP (False positive)	TN (True negative)

これらの結果から,「アップル」の実験と対照的にデータ中に正例の数が少ない「キリン」のケースにおいても,「アップル」の実験結果と類似した傾向が見られ,反作用は確認できなかった. ただし,「アップル」の実験に比べて改善効果が小さいことから,現実のデータに起こり易い不均衡なデータに対する対応が重要であることも示唆している.

次に,「アップル」の実験について,正しく判定できなかったデータについて詳しく見る. 図5.11は,すべての評価用データ847件とベースラインの判定が正解となった141件について,各ツイート中から抽出した一般名詞と固有名詞の単語数を密度推定した結果である. なお,ベースラインで不正解となった141件は,線形補間によって正解に転じた98件(破線)と,不正解の43件(点線)に分けてプロットした.

図5.11から,ベースラインで不正解になった141件の分布は,全ての評価用データに比べて山が左にある. このことから,抽出できた単語数が少ないことが,ベースラインによる予測が不正解になる原因のひとつであることが考えられる.

特に,ベースラインで不正解のうち,線形補間で正解になった43件(点線)については,その傾向がより顕著である. このことは,一方で,ベースラインの手法のように,ツイート中の表層的な情報だけでは判定できないケースでは,ユーザーの潜在的な興味情報を組み合わせることが有効であることを示しており,ユーザーの潜在的な興味情報が判定性を高める効果を示していると考えられる.

5.4 まとめ

2.2節において,ユーザーがTwitterへ書き込んだプロフィールやコメントを収集・分析するソーシャルリスニング業務を着目し,分析目的で収集したツイート中に分析に必要のない同じ名前の別の企業名や商品名(=同名他社)が混入することで,分析精度を低下する要因になるため,収集したツイートから目的と無関係のツイートを「ノイズ」として除去することは極めて重要であることを指摘した.

一方で,2.2節では,ツイートの多くは文字制限から1件当たりの文長が短く,また口語的な崩れた表現を多く含んでいるため,テキスト中に出現する単語など表層的な情報のみによる解析は難しく,課題となっていることも説明した.

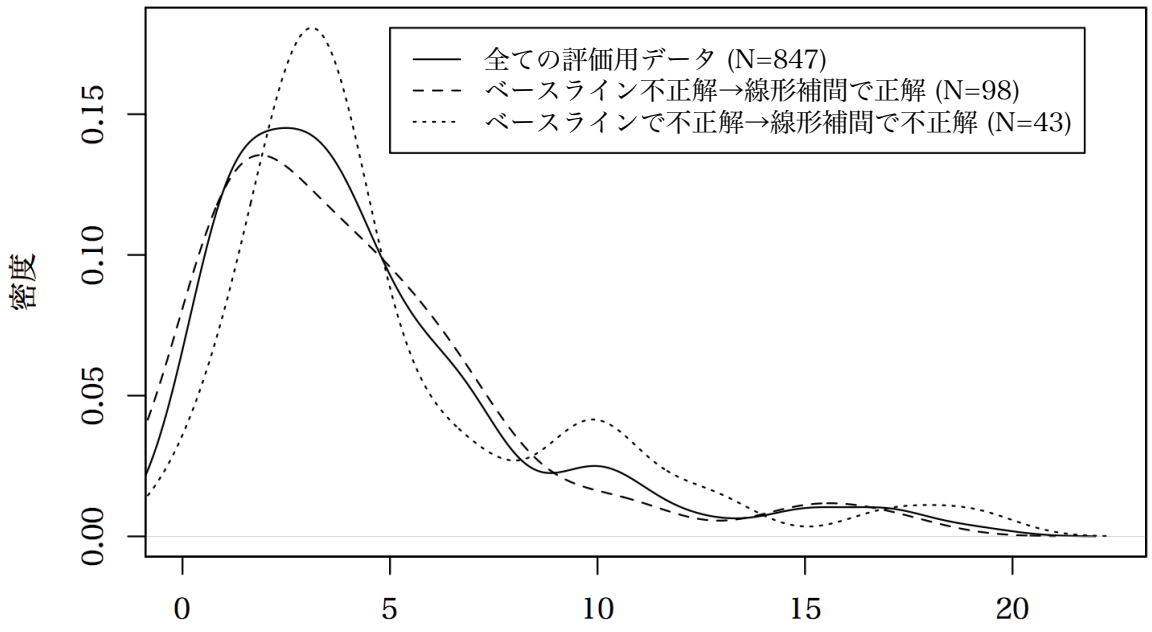

図 5.11: ベースラインで不正解のツイートにおける単語密度 (アップル)

そこで,本章では,ユーザーの興味と話題のダイナミクス(時間的変化)を考慮に入れた,ユーザー興味モデルを利用し,ツイート中の表層的な情報とユーザー興味情報の両方を利用して,検索結果から目的のツイートを選り分ける方法を提案した. また,収集したツイートがコンピュータやディジタル家電メーカーの「Apple Inc.」かそれ以外の「アップル」かを判定する評価実験を行い,検討した提案方法の効果を確認した.

実験の結果,従来の方法に対し8.5%の正解率の改善と,16.5%の見落とし率(FN%)の改善を示し,提案の方法による効果を確認した. これらの結果から,ツイート中の表層的な情報(N-gram 言語モデル)とユーザーの潜在的な興味情報(ユーザーごとのトピックモデル)の両方を利用することにより,ツイート中の表層的な情報だけを利用する従来方法では判定できないケースにおいて,ユーザーの潜在的な興味情報を組み合わせることで判定性を高める効果があることを確認した.

本研究では,提案の方法によって導出した素性が,目的のツイートを判定する性能を持つことを確認することができたが,本章で示したような教師ありの分類器を使用する方法は,判定したい対象ごとに教師データを作成する必要がある. 今後は,導出した素性をクラスタリングなど教師なし学習による分類に適用する方法の検証やダイナミクスの取り入れたモデルによる効果の検証や精度向上について,引き続き取り組む考えである.

第6章　結論

本研究では,問い合わせメールやソーシャルメディアなどの口語的な表現を多く含んだユーザーが発信するテキスト情報の正確な内容理解の支援を目的とし,メールとソーシャルメディアデータを扱う二つのビジネス活動に焦点を当て,統計的な自然言語処理技術を用いた新たな方法を提案し,評価実験によって提案の方法による効果を確認した.

まず,2 章では,口語的な表現を含むテキストに起因する二つの問題について説明し,関連研究を交えて解くべき課題と本研究で使用する自然言語処理によるアプローチについて述べた.

一つ目の問題は,ユーザーから企業などへの問い合わせメールに対する返信を行うカスタマーサポート業務に着目したもので,ユーザーからの問い合わせメール中に多く発生する語の省略が,背景知識や経験の少ない担当者とっては,意図の取り違いや見落としの原因になることを指摘した.

「A の B」タイプの連体修飾関係に関する従来研究は,「A の B」の意味解析や言い換えに関する研究などに重点が置かれ,本研究のように「A の B」の中の A あるいは B の省略を扱う研究はこれまで十分に行われていないという課題がある. 一方,省略や照応解析に関する従来研究は,主として文語的な整ったテキストを対象としていること,文脈中のどこかに先行詞が存在するケースの解析が中心であり,文脈内に明確な先行詞の存在しないケースを対象外としてきたという課題がある. しかし,口語的な表現を含むくだけたテキストに対応するためには,文脈全体のどこにも先行詞が存在しない場合の解析に対応する必要がある.

二つ目の問題は,ユーザーが Twitter へ書き込んだプロフィールやコメントを収集・分析するソーシャルリスニング業務に着目したもので,分析目的で収集したツイート中に分析に必要のない同じ名前の別の企業名や商品名 (同名他社) が混入することで,分析精度を低下する要因になることを指摘した.

Twitter解析に関する研究が多く報告されているが,ノイズに関する従来研究は,スパムやボット対策に重点が置かれ,本研究のように同名他社の問題や多義性に対処する研究は十分に行われていないという課題がある. 一方,近年,トピックモデルやLDAを適用した研究が数を増しており,Twitterを対象とした様々なトピックモデルが提案されていることを確認した. 中でも,1文が短いツイートにLDA適用する場合,1ツイートを1文書とせず,ユーザーの全ツイートを1文書として扱う方法が用いられていることを確認した.

3章では,口語的な表現を多く含むYahoo! 知恵袋の質問テキストと,口語的でないテキストを多く含むWikipediaの日本語テキストの省略傾向を予備実験による比較で確認した. 予備実験の結果から,Yahoo! 知恵袋のような口語的な表現を多く含む質問テキストが,Wikipediaのような口語的でないテキストに比べて,図2.2と図2.3で示したような「AのB」のタイプの連体修飾関係における名詞Bの省略が多いことを確認した. また,Yahoo! 知恵袋のような口語的な表現を多く含む質問テキストでは,名詞Aの省略に比べて名詞Bの省略の方が,必須格の省略が多いことを確認した.

4章では,問い合わせメールに多く発生する省略の問題について,特に口語的な表現の多いテキストで見られる「AのB」タイプの名詞句における名詞Bの省略に焦点を当て,トピックモデルを知識として,省略された語を予測する方法を提案した. 予め候補語集合が与えられた状況下で省略された語を選択する評価実験では,提案の方法が,従来方法に対し正解率で11.34%の改善が見られ,75%を超える高い候補語の選択性能を示すことを確認した. これらの結果から,トピックモデルを利用することにより,従来の言語モデルであるN-gramモデルを補間して候補語の選択性能を向上させる効果があることを確認した.

5章では,ソーシャルメディアについて,キーワード検索で収集したツイート集合にキーワードと同名の別の対象がノイズとして含まれてしまう問題について,ツイート中の表層的な情報とユーザー興味情報の両方を利用して,検索結果から目的のツイートを選り分ける方法を提案した. ツイートの内容が分析対象の「Apple Inc.」に関するものかそれ以外かを判定する評価実験では,提案の方法が,従来方法に対し正解率で8.52%の改善と,見落とし率(FN率)で16.53%の改善効果

を確認した. これらの結果から, ツイート中の表層的な情報 (N-gram 言語モデル) とユーザーの潜在的な興味情報 (ユーザーごとのトピックモデル) の両方を利用することにより, ツイート中の表層的な情報だけを利用する従来方法では判定できないケースにおいて, ユーザーの潜在的な興味情報を組み合わせることで判定性を高める効果があることを確認した.

一方, 4 章の提案方法については, 線形補間係数を推定する方法については扱わなかったが, 高い精度を得るには, 個々の省略表現に最適なパラメータ値を決定する方法が必要である. また, トピック数やモデルのハイパーパラメータについても探索的に決定したが, これらパラメータの導出や推定方法についも検討が必要である. 同様に, 5 章の提案方法についても, 実用性を考慮すると教師なし学習による分類に適用する必要がある. また, ダイナミクスの取り入れたモデルによる効果についても検証が必要である.

これらの課題については, 今後も引き続き取り組む考えである.

謝辞

博士課程の在学中, 公私にわたって大変お世話になった筑波大学大学院システム情報工学研究科の津田和彦教授に深く感謝いたします. また, 本論文の執筆に当たり, 津田研究室の方々には, 日頃より研究の進め方についての貴重な示唆やご意見を頂戴いたしました. 深く感謝いたします.

論文審査を快くお引き受けいただき, 的確なアドバイスを頂戴した筑波大学大学院システム情報工学研究科の吉田健一教授, 佐藤美佳教授, 倉橋節也准教授, 近畿大学理工学部情報学科の溝渕昭二准教授に深く感謝いたします. また, 論文の作成に当たって, 貴重なご助言をいただいた帝京大学文学部社会学科の藤田昌克准教授, 株式会社 KDDI 研究所の鈴木信雄氏に深く感謝いたします.

発表会の場などで, 様々なアドバイスやコメントを下さった筑波大学大学院システム情報工学研究科の教員の方々に, 深く感謝いたします.

最後に, 筆者が在籍した NTT データ先端技術株式会社ソリューション事業部テクノマークグループならびに株式会社 NTT データ技術開発本部サービスイノベーションセンタ知的判断チームの諸先輩方や同僚の方々には, 本研究の意義を理解し, 温かく見守り励ましていただきました. 深く感謝いたします.

ここに記して, 以上の方々に深く感謝いたします.

なお, 本研究の実施に当たっては, ヤフー株式会社が国立情報学研究所に提供した「Yahoo!知恵袋データ (第 2 版)」を利用しました. ここに記して, 感謝いたします.

参考文献

[1] 総務省. 平成23年版 情報通信白書, 2011.

[2] 総務省. 平成24年版 情報通信白書, 2012.

[3] 総務省. 平成25年 情報通信メディアの利用時間と情報行動に関する調査, 2013.

[4] 前川喜久雄. KOTONOHA『現代日本語書き言葉均衡コーパス』の開発(＜特集＞資料研究の現在). 日本語の研究, Vol. 4, No. 1, pp. 82–95, 2008.

[5] 阿部純一. 人間の言語情報処理: 言語理解の認知科学. サイエンス社, 1994.

[6] 益岡隆志, 田窪行則. 基礎日本語文法. くろしお出版, 1992.

[7] 阿辺川武, 奥村学. 日本語連体修飾節と被修飾名詞間の関係の解析. 自然言語処理 = Journal of natural language processing, Vol. 12, No. 1, pp. 107–123, 2005.

[8] 森山健太, 但馬康宏, 藤本浩司, 小谷善行. 枝分かれ同時確率モデルを用いた「A の B」の意味分類(語彙・意味). 情報処理学会研究報告. 自然言語処理研究会報告, Vol. 2008, No. 4, pp. 101–106, 2008.

[9] 片岡明, 増山繁, 山本和英. 要約のための連体修飾節の“A の B”への言い換え. 情報処理学会研究報告. 自然言語処理研究会報告, Vol. 99, No. 73, pp. 37–44, 1999.

[10] 吉本啓. 日本語のゼロ代名詞. 言語研究, Vol. 1987, No. 91, pp. 128–129, 1987.

[11] B. J. GROSZ. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, Vol. 21, No. 2, pp. 203–225, 1995.

[12] Marilyn Walker, Sharon Cote, and Masayo Iida. Japanese discourse and the process of centering. *Computational Linguistics*, Vol. 20, No. 2, pp. 193–233, 1994.

[13] 林部祐太, 小町守, 松本裕治. 文脈情報と格構造の類似度を用いた日本語文間述語項構造解析. 情報処理学会研究報告. 自然言語処理研究会報告, Vol. 2011, No. 10, pp. 1–8, 2011.

[14] 清田陽司, 黒橋禎夫, 木戸冬子. 自動抽出した換喩表現を用いた係り受け関係のずれの解消. 自然言語処理 = Journal of natural language processing, Vol. 11, No. 4, pp. 127–145, 2004.

[15] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 50–57. ACM, 1999.

[16] Frederick Jelinek. *Statistical methods for speech recognition*. MIT press, 1997.

[17] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, Vol. 3, pp. 993–1022, 2003.

[18] T. L. Griffiths. Finding scientific topics. *Proc. National Academy of Sciences*, Vol. 101, pp. 5228–5235, 2004.

[19] 奥村学. マイクロブログマイニングの現在 (第3回集合知シンポジウム). 電子情報通信学会技術研究報告. NLC, 言語理解とコミュニケーション, Vol. 111, No. 427, pp. 19–24, 2012.

[20] 奥村学. ソーシャルメディアを対象としたテキストマイニング. 電子情報通信学会 Fundamentals Review, Vol. 6, No. 4, pp. 285–293, 2013.

[21] Jagan Sankaranarayanan, Hanan Samet, Benjamin E Teitler, Michael D Lieberman, and Jon Sperling. Twitterstand: news in tweets. In *Proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems*, pp. 42–51. ACM, 2009.

[22] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pp. 851–860. ACM, 2010.

[23] Aron Culotta. Detecting influenza outbreaks by analyzing twitter messages. *arXiv preprint arXiv:1007.4748*, 2010.

[24] Vasileios Lampos and Nello Cristianini. Tracking the flu pandemic by monitoring the social web. In *Cognitive Information Processing (CIP), 2010 2nd International Workshop on*, pp. 411–416. IEEE, 2010.

[25] Michael J Paul and Mark Dredze. You are what you tweet: Analyzing twitter for public health. In *ICWSM*, pp. 265–272, 2011.

[26] Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. Twitter catches the flu: detecting influenza epidemics using twitter. In *Proceedings of the Conference on Empirical*

Methods in Natural Language Processing, pp. 1568–1576. Association for Computational Linguistics, 2011.

[27] 谷田和章, 荒牧英治, 佐藤一誠, 吉田稔, 中川裕志. Twitter による風邪流行の推測. 2011.

[28] 荒牧英治, 増川佐知子, 森田瑞樹. 文章分類と疾患モデルの融合によるソーシャルメディアからの感染症把握. 自然言語処理= Journal of natural language processing, Vol. 19, No. 5, pp. 419–435, 2012.

[29] Roy Bar-Haim, Elad Dinur, Ronen Feldman, Moshe Fresko, and Guy Goldstein. Identifying and following expert investors in stock microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1310–1319. Association for Computational Linguistics, 2011.

[30] Qiankun Zhao, Prasenjit Mitra, and Bi Chen. Temporal and information flow based event detection from social text streams. In *AAAI*, Vol. 7, pp. 1501–1506, 2007.

[31] Jianshu Weng and Bu-Sung Lee. Event detection in twitter. *ICWSM*, Vol. 11, pp. 401–408, 2011.

[32] Ryong Lee and Kazutoshi Sumiya. Measuring geographical regularities of crowd behaviors for twitter-based geo-social event detection. In *Proceedings of the 2nd ACM SIGSPATIAL international workshop on location based social networks*, pp. 1–10. ACM, 2010.

[33] 渡辺一史, 大知正直, 岡部誠, 尾内理紀夫. Twitter を用いた実世界ローカルイベント検出. 第 4 回楽天研究開発シンポジウム予稿集, 2011.

[34] Johan Bollen, Huina Mao, and Alberto Pepe. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *ICWSM*, 2011.

[35] Andranik Tumasjan, Timm Oliver Sprenger, Philipp G Sandner, and Isabell M Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM*, Vol. 10, pp. 178–185, 2010.

[36] Sitaram Asur and Bernardo A Huberman. Predicting the future with social media. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, Vol. 1, pp. 492–499. IEEE, 2010.

[37] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pp. 591–600. ACM, 2010.

[38] Jacob Ratkiewicz, Michael Conover, Mark Meiss, Bruno Gonçalves, Snehal Patil, Alessandro Flammini, and Filippo Menczer. Truthy: mapping the spread of astroturf in microblog streams. In *Proceedings of the 20th international conference companion on World wide web*, pp. 249–252. ACM, 2011.

[39] Vahed Qazvinian, Emily Rosengren, Dragomir R Radev, and Qiaozhu Mei. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1589–1599. Association for Computational Linguistics, 2011.

[40] 藤川智英, 鍜治伸裕, 吉永直樹, 喜連川優. マイクロブログ上の流言に対するユーザの態度の分類. 言語処理学会第 18 回年次大会, 2012.

[41] 鳥海不二夫, 篠田孝祐, 兼山元太. ソーシャルメディアを用いたデマ判定システムの判定精度評価. 情報処理学会デジタルプラクティス, Vol. 3, No. 3, pp. 201–208, 2012.

[42] 白井嵩士, 榊剛史, 鳥海不二夫, 篠田孝祐, 風間一洋, 野田五十樹, 沼尾正行, 栗原聡. Twitter におけるデマツイートの拡散モデルの構築とデマ拡散防止モデルの推定. 第 26 回人工知能学会, 2012.

[43] 梅島彩奈, 宮部真衣, 灘本明代, 荒牧英治. マイクロブログにおける流言マーカー自動抽出のための特徴分析. 言語処理学会第 18 回年次大会, 2012.

[44] 鍋島啓太, 渡邉研斗, 水野淳太, 岡崎直観, 乾健太郎. 訂正パターンに基づく誤情報の収集と拡散状況の分析. 自然言語処理, Vol. 20, No. 3, 2013.

[45] Son Doan, Bao-Khanh Ho Vo, and Nigel Collier. An analysis of twitter messages in the 2011 tohoku earthquake. In *Electronic Healthcare*, pp. 58–66. Springer, 2012.

[46] Adam Acar and Yuya Muraki. Twitter for crisis communication: lessons learned from japan's tsunami disaster. *International Journal of Web Based Communities*, Vol. 7, No. 3, pp. 392–402, 2011.

[47] Takeshi Sakaki, Fujio Toriumi, and Yutaka Matsuo. Tweet trend analysis in an emergency situation. In *Proceedings of the Special Workshop on Internet and Disasters*, p. 3. ACM, 2011.

[48] 宮部真衣, 荒牧英治, 三浦麻子. 東日本大震災における twitter の利用傾向の分析. 情報処理学会研究報告. EIP,[電子化知的財産・社会基盤], Vol. 2011, No. 17, pp. 1–7, 2011.

[49] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pp. 759–768. ACM, 2010.

[50] Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 1277–1287. Association for Computational Linguistics, 2010.

[51] Benjamin P Wing and Jason Baldridge. Simple supervised document geolocation with geodesic grids. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pp. 955–964. Association for Computational Linguistics, 2011.

[52] Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pp. 37–44. ACM, 2010.

[53] John D Burger, John Henderson, George Kim, and Guido Zarrella. Discriminating gender on twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1301–1309. Association for Computational Linguistics, 2011.

[54] Marco Pennacchiotti and Ana-Maria Popescu. Democrats, republicans and starbucks afficionados: user classification in twitter. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 430–438. ACM, 2011.

[55] Alan Mislove, Bimal Viswanath, Krishna P Gummadi, and Peter Druschel. You are who you know: inferring user profiles in online social networks. In *Proceedings of the third ACM international conference on Web search and data mining*, pp. 251–260. ACM, 2010.

[56] 池田和史, 服部元, 松本一則. マーケット分析のための twitter 投稿者プロフィール推定手法 (コンシューマ・デバイス & システム vol. 2 no. 1). 情報処理学会論文誌 論文誌トランザクション, Vol. 2011, No. 2, pp. 82–93, 2012.

[57] 蔵内雄貴, 内山俊郎, 内山匡. マルコフ確率場を用いたソーシャルネットワークからのユーザ属性推定. 電子情報通信学会論文誌 D, Vol. 96, No. 6, pp. 1503–1512, 2013.

[58] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and P Krishna Gummadi. Measuring user influence in twitter: The million follower fallacy. *ICWSM*, Vol. 10, pp. 10–17, 2010.

[59] Yajuan Duan, Long Jiang, Tao Qin, Ming Zhou, and Heung-Yeung Shum. An empirical study on learning to rank of tweets. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pp. 295–303. Association for Computational Linguistics, 2010.

[60] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining*, pp. 261–270. ACM, 2010.

[61] Dmitry Davidov, Oren Tsur, and Ari Rappoport. Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pp. 241–249. Association for Computational Linguistics, 2010.

[62] Luciano Barbosa and Junlan Feng. Robust sentiment detection on twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pp. 36–44. Association for Computational Linguistics, 2010.

[63] Ismael Santana Silva, Janaína Gomide, Adriano Veloso, Wagner Meira Jr, and Renato Ferreira. Effective sentiment stream analysis with self-augmenting training and demand-driven projection. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pp. 475–484. ACM, 2011.

[64] Samuel Brody and Nicholas Diakopoulos. Cooooooooooooooollllllllllllll!!!!!!!!!!!!!!: using word lengthening to detect sentiment in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 562–570. Association for Computational Linguistics, 2011.

[65] Xiaolong Wang, Furu Wei, Xiaohua Liu, Ming Zhou, and Ming Zhang. Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pp. 1031–1040. ACM, 2011.

[66] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, pp. 1–12, 2009.

[67] Jon Kleinberg. Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, Vol. 7, No. 4, pp. 373–397, 2003.

[68] 蝦名亮平, 中村健二, 小柳滋. リアルタイムバースト検出手法の提案. 日本データベース学会論文誌, Vol. 9, No. 2, pp. 1–6, 2010.

[69] Michael Mathioudakis and Nick Koudas. Twittermonitor: trend detection over the twitter stream. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, pp. 1155–1158. ACM, 2010.

[70] Marco Pennacchiotti and Siva Gurumurthy. Investigating topic models for social media user recommendation. In *Proceedings of the 20th international conference companion on World wide web*, pp. 101–102. ACM, 2011.

[71] Daniel Ramage, Susan T Dumais, and Daniel J Liebling. Characterizing microblogs with topic models. *ICWSM*, Vol. 10, pp. 1–1, 2010.

[72] Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. Comparing twitter and traditional media using topic models. In *Advances in Information Retrieval*, pp. 338–349. Springer, 2011.

[73] Qiming Diao, Jing Jiang, Feida Zhu, and Ee-Peng Lim. Finding bursty topics from microblogs. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pp. 536–544. Association for Computational Linguistics, 2012.

[74] Bharath Sriram, Dave Fuhry, Engin Demir, Hakan Ferhatosmanoglu, and Murat Demirbas. Short text classification in twitter to improve information filtering. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pp. 841–842. ACM, 2010.

[75] 西田京介, 坂野遼平, 藤村考, 星出高秀. データ圧縮による twitter のツイート話題分類. In *DEIM Forum 2011*, pp. A1–6, 2011.

[76] Beaux Sharifi, Mark-Anthony Hutton, and Jugal Kalita. Summarizing microblogs automatically. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 685–688. Association for Computational Linguistics, 2010.

[77] Hiroya Takamura, Hikaru Yokono, and Manabu Okumura. Summarizing a document stream. In *Advances in Information Retrieval*, pp. 177–188. Springer, 2011.

[78] Fei Liu, Yang Liu, and Fuliang Weng. Why is sxsw trending?: exploring multiple text sources for twitter topic summarization. In *Proceedings of the Workshop on Languages in Social Media*, pp. 66–75. Association for Computational Linguistics, 2011.

[79] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pp. 675–684. ACM, 2011.

[80] Danesh Irani, Steve Webb, Calton Pu, and Kang Li. Study of trend-stuffing on twitter through text classification. In *Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*, 2010.

[81] Alex Hai Wang. Don't follow me: Spam detection in twitter. In *Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference on*, pp. 1–10. IEEE, 2010.

[82] Zi Chu, Steven Gianvecchio, Haining Wang, and Sushil Jajodia. Who is tweeting on twitter: human, bot, or cyborg? In *Proceedings of the 26th annual computer security applications conference*, pp. 21–30. ACM, 2010.

[83] Fabrıcio Benevenuto, Gabriel Magno, Tiago Rodrigues, and Virgılio Almeida. Detecting spammers on twitter. In *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*, Vol. 6, p. 12, 2010.

[84] Daniel Gayo-Avello and David J Brenes. Overcoming spammers in twitter œ a tale of five algorithms. 2010.

[85] Chris Grier, Kurt Thomas, Vern Paxson, and Michael Zhang. @ spam: the underground on 140 characters or less. In *Proceedings of the 17th ACM conference on Computer and communications security*, pp. 27–37. ACM, 2010.

[86] Mark Steyvers, Padhraic Smyth, Michal Rosen-Zvi, and Thomas Griffiths. Probabilistic author-topic models for information discovery. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 306–315. ACM, 2004.

[87] 佐々木謙太朗, 吉川大弘, 古橋武. Twitter におけるユーザの興味と話題の時間発展を考慮したオンライン学習可能なトピックモデルの提案. 情報処理学会研究報告. MPS, 数理モデル化と問題解決研究報告, Vol. 2013, No. 3, pp. 1–6, 2013.

[88] 岩田具治, 渡部晋治, 山田武士, 上田修功. 購買行動解析のためのトピック追跡モデル (人工知能, データマイニング). 電子情報通信学会論文誌. D, 情報・システム, Vol. 93, No. 6, pp. 978–987, 2010.

[89] 笹野遼平, 河原大輔, 黒橋禎夫. 名詞格フレーム辞書の自動構築とそれを用いた名詞句の関係解析. 自然言語処理 = Journal of natural language processing, Vol. 12, No. 3, pp. 129–144, 2005.

[90] Ryohei Sasano and Sadao Kurohashi. A probabilistic model for associative anaphora resolution. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*, pp. 1455–1464. Association for Computational Linguistics, 2009.

[91] Xuan-Hieu Phan and Cam-Tu Nguyen. GibbsLDA++: A C/C++ implementation of latent Dirichlet allocation (LDA), 2007.

[92] Andreas Stolcke, et al. SRILM-an extensible language modeling toolkit. In *INTERSPEECH*, 2002.

[93] Taku Kudo. Mecab: Yet another part-of-speech and morphological analyzer. *http://mecab. sourceforge. net/*, 2005.

[94] 岩田具治. 潜在トピックモデルを用いたデータマイニング. 第 1 回 Latent Dynamics Workshop, 2010/6, 2010.

関連業績リスト

参考論文

・公表済み論文

[1] Tomohiko Harada, Nobuo Suzuki and Kazuhiko Tsuda, “The Prediction of Ellipses Using Topic Model for Japanese Colloquial Inquiry Text,” *Procedia Computer Science*, Vol.22, pp.1311-1318, Procedia, 2013.

[2] Tomohiko Harada, Nobuo Suzuki and Kazuhiko Tsuda, “Japanese Ellipsis Resolution in ‘A NO B’ Noun Phrases for Colloquial Inquiry Text Using Latent Topic Models,” *In: Signal-Image Technology & Internet-Based Systems (SITIS), International Conference on. IEEE*, pp.901-908, IEEE, 2013.

[3] Tomohiko Harada and Kazuhiko Tsuda, “Classifying homographs in Japanese social media texts using a user interest model,” *Procedia Computer Science*, Vol.35, pp.929-936, Procedia, 2014.

・採録決定論文

[1] Tomohiko Harada, Nobuo Suzuki, Yoshikatsu Fujita and Kazuhiko Tsuda, “The Estimate Method of the Omission of Japanese Inquiry Texts using an LDA Algorithm,” *International Journal of Computer Applications in Technologies*, Vol. x, No. x, pp.xxxxxx.

その他の論文

・査読のない発表論文

[1] 原田智彦, 津田和彦,「トピック情報を用いた口語的記述の省略語推定」, 第 19 会年次大会発表論文集, pp.944-947, 言語処理学会, 2013.

[2] 原田智彦, 津田和彦,「トピックモデルを利用したソーシャルテキスト上の同名他者推定」, 第 27 回全国大会論文集, pp.1-2, 人工知能学会, 2014.