様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２４年　５月　25 日現在

機関番号：12102
研究種目：基盤研究（Ｂ）
研究期間：2009～2011
課題番号：21300048
研究課題名（和文）
単語・フレーズ・言語モデルを統合したフレーズ並び替えモデルに基づく統計的機械翻訳
研究課題名（英文）
Phrase reordering models integrating words, phrases and language models for statistical machine translation
研究代表者
山本　幹雄（YAMAMOTO MIKIO）
筑波大学・システム情報系・教授
研究者番号：40210562

研究成果の概要（和文）：
本研究では､高精度かつ長距離のフレーズ並び替えを可能とするルールの抽出手法を開発した。抽出されたフレーズ並び替えルールの特徴は、フレーズの並び替えに重要な働きをする機能語（助詞など）の対訳関係を中心に語彙化されている点である。これにより、翻訳対象である文構造を的確に捉えながらフレーズの並び替えが可能となる。日英の翻訳実験において、提案ルールによって翻訳性能を改善できることを明らかにした。

研究成果の概要（英文）：
In this project, we developed an extraction method for accurate and long-distance phrase reordering rules that are lexicalized by function words such as prepositions. Since patterns of function words determine the structure of a sentence, the rules can correctly reorder phrases on the basis of the structure. In Japanese-English translation experiments, we improved the performance of our translation system using the proposed rule set.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2009 年度	7,400,000	2,220,000	9,620,000
2010 年度	3,400,000	1,020,000	4,420,000
2011 年度	3,400,000	1,020,000	4,420,000
年度			
年度			
総　計	14,200,000	4,260,000	18,460,000

研究分野：総合領域
科研費の分科・細目：情報学、知能情報学
キーワード：機械翻訳, 統計的機械翻訳, 翻訳モデル, 並び替えモデル, 階層フレーズ

１．研究開始当初の背景

十数年前、将来的にはおそらく WEB で使われる言語は英語が支配的になるだろうと言われていたが、現実は逆となり年々英語ページの割合は減少している。例えば、2002 年の時点で英語ページは WEB 全体の約 2/3 を占めていたが、最近のデータでは英語ページの割合は約 1/3 程度まで減少している。主要な原因としては多数の新興国の情報インフラが急速に発展し WEB で情報発信を自国語ではじめていることが挙げられる。その他にも様々な理由があり現実として WEB で使われる言語は確

実に多様化してきている。このような言語の多様化は、世界規模で WEB 上の情報が爆発的に増大している中、広い範囲で情報収集を行う場合の大きな障壁となりえる。このため、異なる言語間の翻訳を自動的に行う機械翻訳技術は年々重要となっている。

ここ 10 年における「句」を翻訳単位とした統計的機械翻訳技術の発展は目覚しく、類似した言語間の翻訳（例えばフランス語から英語）では、数十年に渡って改良されてきた従来の商用機械翻訳システムよりも良質の翻訳結果を生成できるとの報告があるほどである。

しかし、構造的に近い関係にある言語間の翻訳で高性能を発揮する統計的手法も、日本語-英語間のように構造的に遠い言語間の翻訳でうまく行っているとは言いがたく、従来の規則に基づく機械翻訳手法のレベルには及んでいない。例えば、国立情報学研究所が主催する NTCIR-7 の特許翻訳タスクでは世界から集まった15チームが特許文書の日英翻訳性能を競ったが、人手による詳細な翻訳品質評価では５段階評価で約 0.5ー１段階程度ルールベースの手法が統計的手法よりも勝っていることが明らかになった。翻訳結果の分析により、フレーズに基づく統計的機械翻訳はフレーズ内のレベルでは極めて正確に翻訳することが可能であるが、翻訳されたフレーズを並び替えるところでの誤りが多いことが知られていた。

２．研究の目的

フレーズの並び替えをモデル化するには主に次の３つのアプローチがある。(1)隣接フレーズモデル：目的言語文における二つの隣り合ったフレーズが原言語文において位置が逆か、そのままであるか、あるいは遠く離れているのかによってモデル化する方法。(2)構文モデル：原言語文と目的言語文の構文対応を自動的に推定し、構文木の上で翻訳の際にどのような位置交換が生じるかをモデル化する方法。(3)言語モデル：言語モデルは目的言語文の言語らしさを評価するモデルであるが、これは間接的にフレーズの並び替えを評価している。

これら３つの比較研究を行い優劣を見極めた上で、これまでにない高性能なフレーズ並び替えモデルを開発し、日本語-英語間の翻訳でも高精度なフレーズの並び替えを達成することを目的とする。

３．研究の方法

以下のような方法で3年間､研究を行った。

(1)「比較研究」（１年目：H21 年度）

主要なフレーズ並び替えモデル３つをそれぞれ高度化するとともに、組み合わせにおける性能比較を行った。比較の結果、隣接フレーズモデルと言語モデルの組み合わせよりも、構文モデルと言語モデルの組み合わせの性能が高いことが分かった。これは、隣接フレーズモデルの制約が比較的弱いために、言語モデルと競合してしまう場合があるためである。一方、構文モデルと言語モデルの組み合わせでは、構文モデルによる複数の並び替えの可能性の中から言語モデルによって一つに絞り込まれており、うまく協調動作していた。これらの実験結果と考察により、構文モデルと言語モデルの組み合わせで新モデルを検討することとした。

(2)「新モデルの開発」(2 年目：H22 年度)

構文モデルの詳しい解析によって、特に助詞などの機能語を含むルールがフレーズの並び替えに有効であることが分かった。そこで、機能語を明示的に指定し、機能語を中心としてかつ長距離のフレーズ並び替えをモデル化するルールを抽出する方法を考案した。詳しくは次節「4.研究成果」で述べる。

(3)「新システムの開発」(3 年目：H23 年度)

実際の翻訳ができるように(2)で開発したフレーズ並び替えモデルをデコーダ（統計的機械翻訳システム）へ組み込む方法を検討し、実現した。評価実験を行い、その結果を検討しながらさらなるフレーズ並び替えモデルの改良を行った。

４．研究成果

最終的な成果として提案した 2 つのフレー

ズ並び替えモデルについて説明し、日英翻訳実験による評価結果について述べる。

(1) 広範囲語順調整ルール

本節で述べる並べ替えルールは隣接フレーズ並び替えモデルで用いられる方法を構文モデルに埋め込んだルールとみなすことができる。図 1 のような単語対応がある学習データについて具体的に説明する。図 1 は縦軸が英文、横軸がその英文を翻訳した日本語文であるとし、黒い箱がそれぞれの言語の単語が対応していることを意味する。以下の説明では、すべて〈ア;A〉という単語対（日本語の「ア」という単語と、英語の「A」という単語が対となっている）の前後のフレーズに関する並べ替えルールを抽出する場合である。

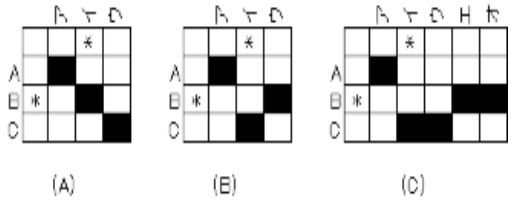

図 1　単語対応例

まず図 1(A)の場合、〈ア;A〉という対訳単語対（フレーズ対でもよい）の右上と左下のマスが白く(対応単語対がないことを意味し、*でマークしてある)、かつ右下のマスが黒い（対応単語対が存在）ことより、〈ア X; A X〉という階層フレーズ対が抽出される。このルールにおける X は任意のフレーズに置換可能であり、日本語の「ア」という単語の右側のフレーズは、英語の「A」という単語の右側に来ることを意味している。この手法は単に隣接のマス（単語対）を考慮するため、単語に基づく手法と呼ぶ。一方、図 1(B)の場合、〈ア;A〉の周辺のマス全部が白いことから、上述の単語に基づく抽出法は失敗する。実際には〈ア;A〉の右下に〈イ ウ; B C〉というフレーズ対が存在している。ここから、〈ア X; A X〉を抽出するのがフレーズに基づく手法である。つまり、単に隣接のマスではなく、隣接のフレーズ対を探すというのがフレーズに基づく手法の長所であり、単語に基づく手法より広い範囲の語順調整情報を取得できる。図 1(C)はより大きなフレーズが右下に存在する場合である。

このように、非常に大きなフレーズが隣接する場合、フレーズの位置関係を階層フレーズルールに取り込むことが困難であった。本研究では、フレーズ対の長さ制限なしで大きなフレーズ対をすべて認識することにより、長距離のフレーズ移動を可能とする階層フレーズルールを抽出する方法を開発した。

(2) 機能語パターンによる並び替えルール

本手法は本研究の中心的な成果であるため、提案するルールをやや詳しく述べる。

まず、原言語側の単語集合$\mathcal{F}$を決定する。この集合をイニシャル単語集合と呼び、文構造を示唆できる機能語のみを含むものとする。本節では説明の都合上、助詞、助動詞、接続詞、句読点という 4 種類の日本語品詞を持つ単語とする。対訳文対 $f_1^I = f_1, f_2, ..., f_I$ と $e_1^J = e_1, e_2, ..., e_J$ のすべて可能な単語インデックスペア集合を以下の$\mathcal{S}$で記す。

$$\mathcal{S} = \{(i,j) \mid 1 \le i \le I, 1 \le j \le J\}$$

f_1^I と e_1^J の単語アラインメントで$\mathcal{S}$の一つサブ集合$\mathcal{A}$が決定される。つまり、

$$\mathcal{A} = \{(i,j) \mid (i,j) \in \mathcal{S}, and\ e_i\ is\ aligned\ to\ f_j\}$$

である。さらに、イニシャル単語集合$\mathcal{F}$を用いて、$\mathcal{A}$をさらに$\mathcal{R}$と$\mathcal{O}$の二つ集合に分割する。

$$\mathcal{R} = \{(i,j) \mid (i,j) \in \mathcal{A}, and\ f_j \in \mathcal{F}\}$$
$$\mathcal{O} = \{(i,j) \mid (i,j) \in \mathcal{A}, and\ f_j \notin \mathcal{F}\}$$

で定義すると、明らかに$\mathcal{R} \cup \mathcal{O} = \mathcal{A}$と$\mathcal{R} \cap \mathcal{O} = \varnothing$を満たす。図 3 は図 2 の単語対応表において集合$\mathcal{R}$と$\mathcal{O}$の例である。黒いマスは$\mathcal{R}$に属する単語ペアであり、灰色のマスは$\mathcal{O}$に属する単語ペアである。$\mathcal{F}$に属する日本語側単語は*で表記する。

単語対応付きの対訳文対に対して、抽出できるすべての対訳フレーズの集合を$\mathcal{P}$で表記すると、集合$\mathcal{R}$と$\mathcal{O}$に基づき、$\mathcal{P}$のサブ集合$\mathcal{P}_o$が以下のように決定できる。

$$\mathcal{P}_o = \{p \mid p \in \mathcal{P}, and\ \forall (i,j)\ covered\ by\ p : (i,j) \in \mathcal{A} \rightarrow (i,j) \in \mathcal{O}\}$$

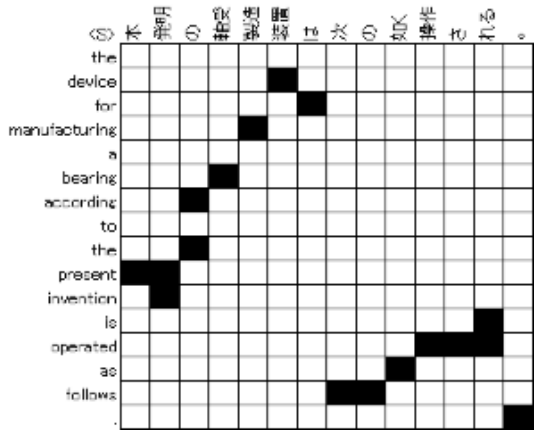

図 2　日英の単語対応例

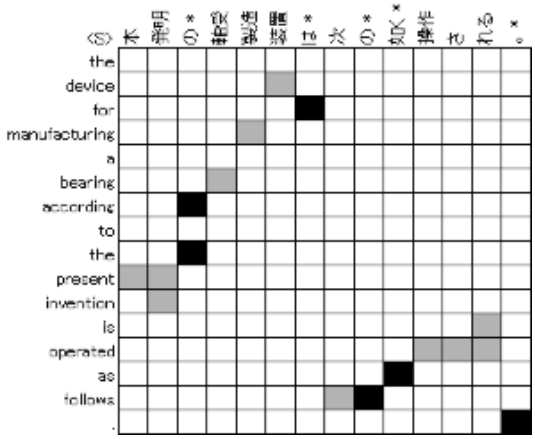

図 3　$\mathcal{R}$ と $\mathcal{O}$ の例

つまり、集合 $\mathcal{P}_o$ は $\mathcal{O}$ に属するマスのみを覆う対訳フレーズ対の集合である。

次に、 $\mathcal{P}_o$ のサブ集合となる $\mathcal{P}_o$ の中の「最大フレーズ対」で構成される集合 $\mathcal{M}$ を以下のように定義する。

$$\mathcal{M} = \{p \mid p \in \mathcal{P}_o,\ and\ \exists(i,j)\ covered\ by\ p : (i,j)\ is\ covered\ by\ q \in \mathcal{P}_o \rightarrow q = p\}$$

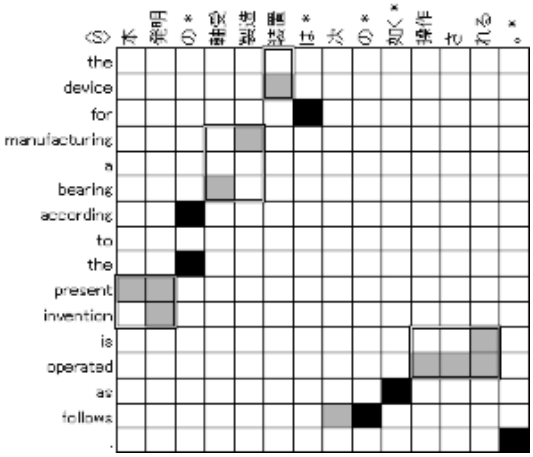

図 4　$\mathcal{M}$ の例

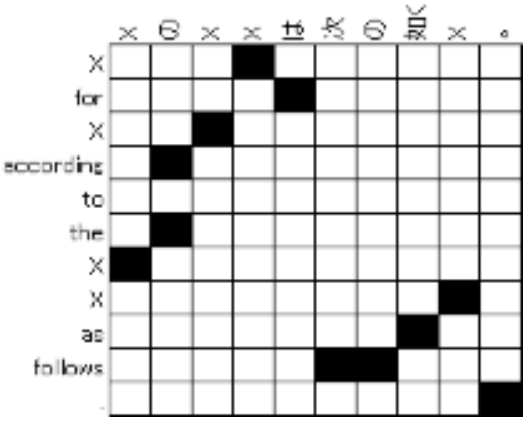

図 5　縮小した単語対応

図 4 は図 3 の $\mathcal{R}$ と $\mathcal{O}$ から得られる $\mathcal{M}$ を示している。二重線枠で囲まれる対訳フレーズ対は $\mathcal{M}$ に属するフレーズ対である。ただし、$\mathcal{P}_o$ が常にすべての $\mathcal{O}$ に属するマスを覆うことはないというところには注意。

最終的に、$\mathcal{M}$ の中の対訳フレーズ対をそれぞれ一つの黒いマスに収縮させる（フレーズを一つ単語のように見なす）。図 4 の例に対して、その収縮されたアライメントテーブルは図 5 のようになる。収縮処理から得られる黒いマスはその両辺に共に非終端記号 X を置き、対応される非終端記号という意味を表す。図 5 というような非終端記号を含む単語アラインメントテーブルを用いて、従来の対訳フレーズ抽出法を応用すれば非終端記号を含む対訳フレーズ対（つまり、階層的な対訳フレーズ対）が抽出できる。ただし、抽出において、幾つかのヒューリスティックスを施す必要がある。

図 5 の収縮した単語対応表より例えば、以下の階層フレーズ対を抽出できる。X1ーX6 の記号は任意のフレーズに置換可能な変数であり、原言語と目的言語中で同じ変数が対訳フレーズのそれぞれの言語における対応する位置を表す。

〈X1 の X2; X2 according to the X1〉
〈次の如く X1 。; X1 as follows .〉
〈X1 および X2　の X3 に関する
　　X4 を X5 する X6;
a X6 for X5 the X1 and X4 concerning X3
　　of the X2〉

3 つ目のルールは非常に広範囲をカバーしており、かつ特許文（今回の翻訳対象分野）においてよく現れる文パターンである。

(3) 日英翻訳実験（評価実験）

①実験の概要

これまで述べてきた提案階層フレーズルール集合を基本的な翻訳ルールと組み合わせた場合の性能評価を行うとともに、従来法との比較を行った。分野として特許文書の日英翻訳において実験を行い、客観評価指標の一つである BLEU（値が大きいほど高性能）を用いて性能を評価した。

②実験データ

学習用データとして、NTCIR-7 特許翻訳タスクで用いられた日英特許対訳データベースからランダムに抽出された 10 万文を用いた。デベロップメントデータとテストセットは、同じく NTCIR-7 の dev データと formal run データを用いた。

③比較ルール集合

提案手法による階層フレーズルールは広範囲の並び替えを目的としており、基本的な近距離の並び替えルールと組み合わせる必要がある。組み合わせのための基本ルールとして、Chiang の基本的な階層フレーズの中から以下の形に限定したルール集合を G と呼ぶ。

X → 〈ア X イ, A X〉
X → 〈ア X イ, X A〉
X → 〈ア X, A X〉
X → 〈X ア, A X〉
X → 〈ア X イ, A X B〉

ここで「ア」と「イ」は原言語側のフレーズ、「A」と「B」は目的言語側のフレーズを意味している。X は任意のフレーズである。

上で述べた提案ルールの(1)を R、(2)を S と呼び、G と組み合わせて評価する。ただし、R は以下のような形に限定した。

X → 〈ア, A〉
X → 〈ア X, A X〉
X → 〈ア X, X A〉
X → 〈X ア, X A〉
X ← 〈X ア, A X〉

④実験結果

表 1 は従来の階層フレーズモデル(Hiero と表記）と各ルールセットで構成する提案モデルの性能評価結果である。

表 1　日英特許翻訳の性能

システム	Hiero	R	G+R	S+R	S+G+R
BLEU(%)	28.49	27.37	28.08	28.22	28.91

翻訳精度において、S+R はほぼ従来の階層フレーズモデルと同じ精度に達している。S+G+R はさらに向上し、従来の階層フレーズモデルの性能を上回っていることが分かる。

図 6 は、翻訳性能(縦軸, BLEU(%))とルール数（横軸, Million rules）を対比させた図である。図 6 より、S+R は従来の階層フレーズの約五分の一のサイズであり、S+G+R は約半分のサイズであることが分かる。提案ルール集合である S は少数のルールによって精度を効率よく改善できることが分かる。

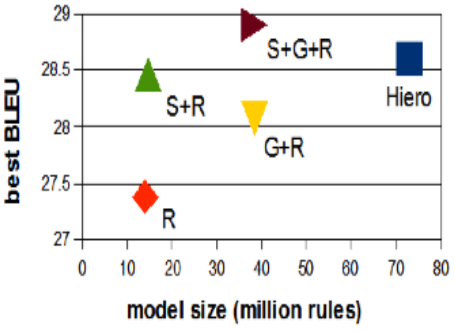

図 6 翻訳性能とルール数の関係

表 2 は翻訳例であり、図 7 と図 8 はそれぞれ従来の階層フレーズモデル、提案手法である S+G+R ルール集合を使った場合の導出木である。表 2 の原言語は入力文としての日本語、正解は人間が与えた正解、Hiero は従来法システムの翻訳結果、S+G+R は提案法による翻訳結果である。この例では、広範囲フレーズ並び替えルールにより、3 個以上のフレーズの移動を正しく制御しており、構文構造として正しい出力文が得られていることが分かる。

表２　翻訳例

原言語	また、同時に強誘電体キャパシタ１２からＢＬ２へ電荷が移動する。
正解	At the same time , electric charges transfer from the ferroelectric capacitor 12 to BL2.
Hiero	At the same time , BL2 to charge is moved from the ferroelectric capacitor 12.
S+G+R	At the same time , the charge moves from the ferroelectric capacitor 12 to BL2.

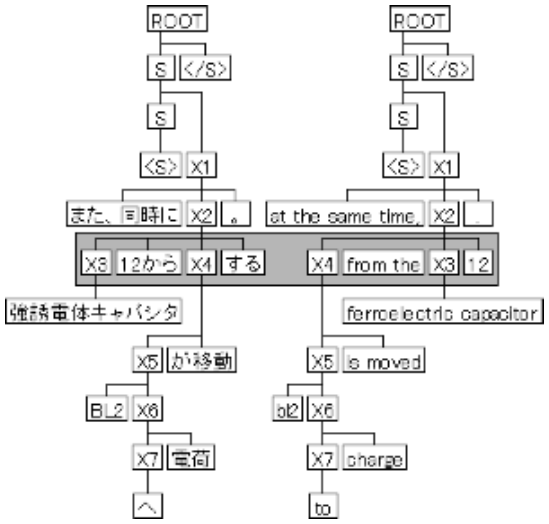

図７ 従来法(Hiero)による翻訳時の導出木

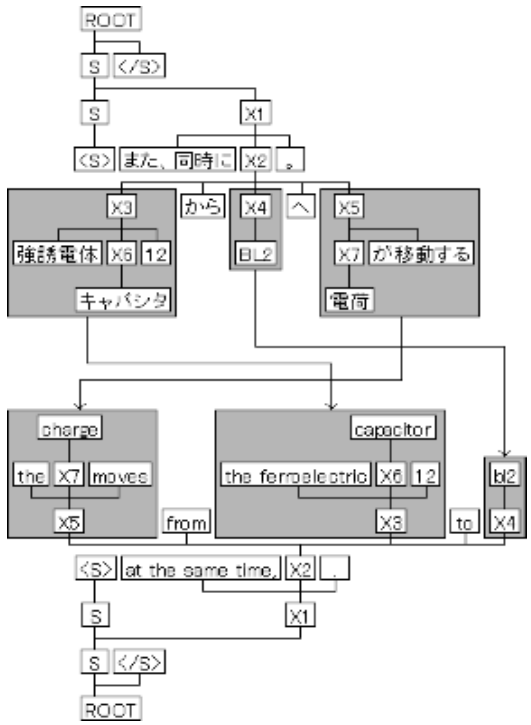

図８　提案法(S+G+R)による翻訳時の導出木

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計７件）

① Chenchen Ding, Takashi Inui and Mikio Yamamoto. Long-distance hierarchical structure transformation rules utilizing function words. In Proc. of the IWSLT 2011, pp.159-166. 査読あり. Dec. 9, 2011(San Francisco). http://www.mt-archive.info/IWSLT-2011-TOC.htm

② 越川満，内山将夫，梅谷俊治，松井知己，山本幹雄．統計的機械翻訳におけるフレーズ対応最適化を利用したN-best翻訳候補のリランキング．情報処理学会論文誌，Vol.51，pp.1443-1451．査読あり．2010.

〔学会発表〕（計10件）

① 丁塵辰，乾孝司，山本幹雄．統計的機械翻訳における機能語を利用した階層的広範囲フレーズ並び替えルール．言語処理学会第18回年次大会，2012.3.15(広島市立大学・広島県).

② 安田隆浩，越川満，乾孝司，山本幹雄．Khafra：語順並べ替えモデルに対応した動的計画法に基づくSMTデコーダ．言語処理学会第１６回年次大会．2010.3.10(東京大学・東京都).

６．研究組織
(1)研究代表者
山本　幹雄（YAMAMOTO MIKIO）
筑波大学・システム情報系・教授
研究者番号：40210562

(2)研究分担者
乾　孝司（INUI TAKASHI）
筑波大学・システム情報系・助教
研究者番号：60394031

(3)研究協力者
丁 塵辰（DING CHENCHEN）
筑波大学大学院・システム情報工学研究科・コンピュータサイエンス専攻