5B-06

日英機械翻訳の前処理としての語順整序

大木 俊樹† 横井 健†
東京都立産業技術高等専門学校†

1. はじめに

近年、翻訳業務における機械翻訳の重要性が高まっている。翻訳業務の従事者 34 人を対象に実施したアンケート[1]では 4 割以上にあたる 14 人が「機械翻訳をほぼ毎日利用している」と答えている。しかし機械翻訳の翻訳品質については、「満足している」と答えた人数が 5 人、「不満である」の人数が 12 人であり、翻訳品質の改善が機械翻訳の課題であるといえる。

本研究では、日英翻訳における機械翻訳の品質に影響を与える要素として原文の語順に着目する。日本語は英語などの言語に比べて語順の自由度が高いため、日英翻訳では原文の語順によって翻訳結果が変化し、翻訳品質が低下することがある。

そこで、日英機械翻訳の前処理として語順整序を行ない、翻訳品質を向上することを目指す。

2. 提案手法

本研究では、あらかじめ学習データとして日本語のテキストを用意し、その語順に基づいて被翻訳文の語順整序を行う。学習データには、一般的な語順の文章であると考えられる新聞や論文を利用する。

語順の分析と入れ替えは図 1 のような係り受け木構造に基づいて行なう。この係り受け木は、文全体の述語を根として修飾関係を表現したもので、同じ親要素を持つ子要素は入れ換えが可能であるとする。ただし、図 2 のように同じ文節に連用修飾と連体修飾が係っている場合、それらは入れ換えが不可である。

語順は、以下の二つの仮定に基づいて決定する。

(A) 各文節の順番は基本的に、文節の機能語（格助詞や接続詞など）の種類によって決まる。
(B) 多くの文節が係っている文節は手前に位置することが多い。

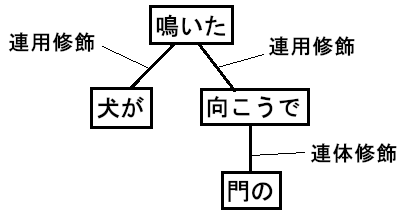

図１：「犬が門の向こうで鳴いた」という文の係り受け木

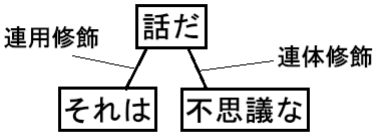

図２：「それは不思議な話だ」という文の係り受け木

(A)の仮定より、機能語の順番を調べれば各文節の順番を定めることが可能である。そのため、学習データに現れた機能語の組み合わせすべてについて、どちらの機能語が手前に位置しやすいかの優先度を算出する。

機能語w_iの機能語w_jに対する優先度p_{ij}は(1)式で定義される。ここで$S_{w_iw_j}$は学習データの部分集合であり、w_iがw_jより手前にありそれらの文節が入れ替え可能な文の集合である。また$len(s,w_i)$は文sの中でw_iを含む文節に係る文節の数であり、(B)の仮定に基づいて、文節数の比を優先度とする。

$$p_{ij} = \sum_{s \in S_{w_iw_j}} \frac{len(s,w_j)}{len(s,w_i)} \quad (1)$$

このようにして算出した優先度に基づいて、被翻訳文の語順整序を行なう。被翻訳文sの中に入れ換え可能な二つの文節があり、一つ目の文節の機能語をw_i、二つ目の文節の機能語をw_jとすると、(2)式が成立したときにそれらの文節を入れ換える。両辺に文節数lenを掛けているのは、(B)の仮定に基づいて、多くの文節が係っている文節ほど手前に位置しやすくするためである。

$$p_{ij} \times len(s,w_i) < p_{ji} \times len(s,w_j) \quad (2)$$

Word Order Rearrangement for the Preprocessing of the Japanese to English Machine Translation
†Toshiki Ohki, Takeru Yokoi,
Tokyo Metropolitan College of Industrial Technology

3. 検証実験

3.1. 実験方法

本システムによって翻訳精度が向上するかどうかの検証のため、実験を行なった。

日本語文の語順の学習データには CD-毎日新聞データ集 08 版を使用した。

係り受け解析のツールには CaboCha[2]を使用した。

対訳コーパスは Wikipedia 日英京都関連文書対訳コーパス[3]を用い、歴史カテゴリからランダムに抽出した記事 200 件(7,262 文)を使用した。

機械翻訳システムは Microsoft Translator を使用した。対訳コーパスの日本語文について、語順整序を行なう前と後の文章をそれぞれ Microsoft Translator で翻訳し、機械翻訳自動評価尺度 RIBES[4]のスコアを求めて比較した。また被翻訳文の全体についてだけでなく、1 文ずつに対しても RIBES のスコアを求めて、スコアが向上したかどうかを分類した。

3.2. 実験結果

実験を行なったところ、語順整序を行なう前と後のスコアは表 1 のようになった。

結果、語順整序を行なうことで RIBES のスコアは僅かに低下しており、翻訳精度が向上していないことが分かった。

1 文ずつ RIBES のスコアを求めて分類した結果は表 2 のようになった。また、語順の入れ替えが発生した文の中から、スコアが向上した例を図 3 に示す。

表 2 から分かるように、被翻訳文の半分以上の文において語順の変化は発生していない。これらの文は文字数が少ないことから、語順を入れ換えられる文構造が少ないものと思われる。そのため、文字数が少ない文の翻訳精度を語順整序で向上させることは難しいと考えられる。

また、語順の変化によって RIBES のスコアが向上している文もあるが、逆に低下している文のほうが多いため、それが全体としてスコアを低下させる要因になっていると思われる。

表 1: 文章全体のスコアの比較

	RIBES スコア
語順整序前	0.426
語順整序後	0.414

表 2: 1 文ずつのスコアの比較

	結果の内訳	平均文字数
語順変化なし	4,048 文	33.3 文字
RIBES スコア向上	1,095 文	62.6 文字
RIBES スコア低下	2,119 文	56.4 文字

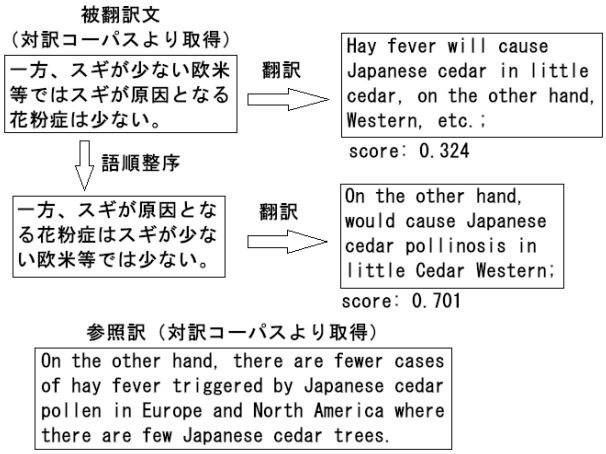

図 3: RIBES スコアが向上した例

スコアが向上した文は、低下した文に比べて文字数が多い。これは、長い文章は係り受け構造が複雑になることが多く、正しく翻訳できない変則的な語順が発生しやすいためであると考えられる。

このことから、本手法は長い文の翻訳前処理に適した手法であると考えられる。

4. おわりに

本研究では、機械翻訳の精度を向上するための語順整序の手法を提案した。

提案手法は、事前に用意した日本語のテキストを学習データとして語順を解析し、その結果に基づいて被翻訳文の語順を入れ換える。この時、各文節の機能語と、その文節を修飾する文節の数によって、語順を入れ換えるかどうかを決定する。

検証のために実験を行なったところ、RIBES のスコアが向上した文もあったが、全体としてはスコアを向上させることはできなかった。また、短い文では語順の変化が発生しづらく、翻訳精度の向上が困難であることが分かった。

参 考 文 献

[1] 長瀬友樹,小谷克則,工藤竜広,佐久間みゆき,秋葉泰弘. ″実務翻訳における機械翻訳の利用に関する調査報告″. 言語処理学会第 20 回年次大会発表論文集. 2014, pp.610-613

[2] 工藤拓,松本裕治. ″チャンキングの段階適用による日本語係り受け解析″. 情報処理学会論文誌. 2002, Vol.43, No.6, pp.1834-1842

[3] 国立研究開発法人情報通信研究機構. ″Wikipedia 日英京都関連文書対訳コーパス″. https://alaginrc.nict.go.jp/WikiCorpus/

[4] 平尾努,磯崎秀樹,須藤克仁,Duh Kevin, 塚田元,永田昌明. ″語順の相関に基づく機械翻訳の自動評価法″. 自然言語処理. 2014, Vol.21, No.3, pp.421-444