構文解析誤りに頑健な日英統計的機械翻訳の事前並べ替え手法

星野 翔 [1,2]　　宮尾 祐介 [1,2]　　須藤 克仁 [3]　　永田 昌明 [3]

[1] 総合研究大学院大学　[2] 国立情報学研究所　[3]NTT コミュニケーション科学基礎研究所

{hoshino,yusuke}@nii.ac.jp {sudoh.katsuhito,nagata.masaaki}@lab.ntt.co.jp

1　はじめに

統計的機械翻訳における統語構造に基づく事前並べ替え手法は、日本語と英語のように、語順の大きく異なる言語対での翻訳精度を改善することができるが、その一方で構文情報を利用するために、構文解析の精度が結果を大きく左右するという欠点がある。

そこで本研究では、構文解析誤りが発生しやすい依存構造を一切使用せず、局所的な、連続する文節間の並列関係という構文情報のみに着目することによって、並列関係を扱いながら構文解析誤りに左右されにくい並べ替え手法を提案する。

提案手法の有効性を調べるため、特許文書翻訳において既存の事前並べ替え手法との翻訳精度を比較したところ、提案手法は既存手法全てを上回り、規則ベースの日英事前並べ替えでは現時点で最高の翻訳精度を示した。

2　関連研究

Komachi et al.[11] や Hoshino et al.[3] など、規則ベースの日英事前並べ替え手法は、述語項構造などの構文情報を利用しており、そのため特に後者は並列句を考慮した複雑な並べ替えを行うことができる。しかしこれらの従来手法には、構文情報の正しさに依拠して並べ替えを行うため、構文解析精度に翻訳精度が左右されてしまうという欠点がある。

一方、Katz-Brown and Collins[7] が提案した 2 つの事前並べ替え手法のうち REV では、文字列を句読点で区切り、区切られたそれぞれの部分文字列について、助詞の「は」以降の語順を逆転させることによって、構文情報を利用せず SVO 語順への並べ替えを達成している。そのため構文解析誤りが並べ替えに影響しない一方、名詞句の語順が逆転してしまう、また並列関係が無視されるなどの問題点がある。

そこで磯崎 [4] は、一部の品詞に対して語順を保存するためのラベル付けを行うことを提案し、REV の問題点のうち大部分を解決したが、並列関係は依然無視されたままである。

なお、構文解析誤りの問題をドメイン適応タスクとして考え、解析器を再訓練する手法 [8] も提案されているが、理想的には、どのようなドメインでも構文解析誤りに左右されにくい頑健な並べ替え手法であることが望ましい。

3　提案手法

関連研究の問題点を踏まえて、Katz-Brown and Collins の REV 手法のように構文解析の誤りに影響されない並べ替えを行いつつ、Hoshino et al. のように並列関係も考慮した複雑な並べ替えを行うために、提案手法では、並列関係を考慮する規則 1、長距離の並べ替えを行う規則 2、局所的な単語の並べ替えを行う規則 3、という 3 つの規則を提案する。このうち規則 1 のみで構文情報を利用することで、規則 2,3 の頑健性を確保している。

これより説明のために、入力文 $input$ は l 個の文節で構成され（$input = c_1...c_l$）、文節 c_x $(1 \leq x \leq l)$ は q 個の単語を持っている（$c = w_1...w_q$）と仮定し、各規則によって左辺を右辺に書き換えていくものとする。また例文中の記号「|」は文節の区切りを表す。

3.1　規則 1: 入力文分割規則

この規則では、構文情報を用いて、入力文を文節に分解するのが目的である。

まず入力文において、並列関係にある連続した文節を 1 つの文節とみなし、入力文を m $(1 \leq m \leq l)$ 個の文節（$input = c_1...c_l \rightarrow c_1...c_m$）に書き換える。この操作により、並列関係にある文節の順序逆転を防ぐことができる。

例えば、入力文「表 1 及び | 図 7 に | 示す」では、

並列関係にある文節「表 1 及び | 図 7 に」が 1 つの文節とみなされ、「表 1 及び図 7 に | 示す」と書き換えられる。

次に、入力文に u 個の句読点があれば、1 句読点を 1 文節とみなし、入力文を $n = m + u$ 個の文節 ($input = c_1...c_m \rightarrow c_1...c_n$) に書き換える。この操作は Katz-Brown and Collins と同じく、特許文書では起こりにくい句読点をまたぐ並べ替えの除外を意図している。ただし、元々並列関係にあり連続していた文節 c_x には以下の例外規則を適用する:

例外 1 文節 c_x においてある句読点の直前が名詞である場合、この文節はその句読点では分割されない。

例外 2 例外 1 が適用され、文節 $c_x = w_1...w_q$ の終端 w_q が「は」または「が」ならば、最後に出現した句読点を 1 文節とみなし、文節 c_x を 3 つの文節に分割する。

例外 3 例外 1,2 が適用されたかを問わず、文節 $c_x = w_1...w_q$ の終端 w_q が句読点であった場合、文節 c_x を句読点以外と句読点の 2 つの文節に分割する。

例えば、入力文「各記号は、| 次のものを | 表している。」は、句読点で分割されて「各記号は | 、| 次のものを | 表している | 。」と書き換えられる。

一方、並列関係にある文節「表 1、表 2」では、例外 1 が適用されるため句読点で分割しない。しかし並列関係にある文節「表 3、表 4 が」のように、「は」または「が」が終端であれば例外 2 が適用され、最後の句読点で「表 3 | 、| 表 4 が」と分割する。

3.2 規則 2: 文節間並べ替え規則

この規則では、Katz-Brown and Collins の REV 手法と同じく、文節の順序を入れ替えることで、SOV 語順を SVO 語順にするような長距離の並べ替えを行う。

まず文節 $c_1...c_n$ から終端が「は」または「が」[*1]である最初の文節 c_i $(1 \leq i \leq n)$ を探す。次に、文節 c_1 から文節 c_i まで、文節 c_{i+1} から文節 c_n までの順序をそれぞれ逆転させる ($c_1...c_n \rightarrow c_i...c_1c_n...c_{i+1}$) ことで、SOV 語順から SVO 語順への並べ替えを行う。最後に、文節 $i = w_1..w_q$ の終端 w_q を残す ($c_i...c_1c_n...c_{i+1} \rightarrow w_1...w_{q-1}c_{i-1}...c_1c_n...c_{i+1}w_q$) ことで、助詞の「は」が文頭へ移動されることを防ぐ。

例えば、入力文「そこで | 各記号は | 次のものを | 表している」は、終端「は」を除いた文節「そこで | 各記号」と文節「次のものを | 表している」が逆転され、「各記号 | そこで | は | 表している | 次のものを」と書き換えられる。

3.3 規則 3: 文節内並べ替え規則

この規則では、Hoshino et al. の手法と同じく、これまで並べ替えられたそれぞれの文節内での単語の順序を並べ替え、局所的にも英語の語順により近づけることを意図している。

まず文節 $c = w_1...w_q$ 内の単語を内容語 $w_1...w_p$ $(0 \leq p \leq q)$ と機能語 (助詞・助動詞) $w_{p+1}...w_q$ に分け、そこから機能語の語順を逆転させて内容語の前に移動 ($c = w_1...w_q \rightarrow w_q...w_{p+1}w_1...w_p$) することで、後置詞句を前置詞句に並べ替える。

例えば、文節「図 3 において」は、機能語「において」が逆転かつ先頭に移動され、「おいてに図 3」と書き換えられる。

3.4 並べ替え結果の比較

表 1 に、文節「表 1 及び」と文節「図 7 に」が並列関係にある例文での、提案手法と既存手法の並べ替え結果とその翻訳例を示している。

Katz-Brown and Collins では並列句「表 1 及び図 7 に」が逆順になっているが、提案手法は並列関係を保つため、このような逆転が発生しない。

また、Hoshino et al. では助詞の「は」が文頭近くに移動されているが、提案手法では文節間の並べ替えで「は」や「が」の並べ替えを行うことで、このような移動を防いでいる。

このように、提案手法は英語により近い語順に日本語を並べ替えることができる。

4 比較実験

提案手法の日英翻訳での効果を確認するため、特許文書において提案手法と既存手法を比較する実験を行った。

この実験では、異なる事前並べ替え手法を共通の統計的機械翻訳システムで翻訳することにより、翻訳結果の優劣を測る。その際に、事前並べ替えを全

[*1] Katz-Brown and Collins では助詞の「は」のみが用いられていたが、提案手法では事前実験によって「は」と「が」両方の使用が最良だと確認した。

日本語入力	ここで、| 表 1 及び | 図 7 に | 示す | 各記号は、| 次のものを | 表している。
Katz-Brown and Collins	でここ | 、| は記号各 | 示す | に 7 図 | 及び 1 表 | 、| いる表して | をもの次の | 。 here , the symbols shown in Table 1 and FIG. 7 , the following construction .
Hoshino et al.	でここ | 、| は各記号 | 示す | に表 1 及び図 7 | 、| いる表して | をもの次の | 。 here , the symbols shown in Table 1 and FIG. 7 shows a following method .
提案手法	でここ | 、| 各記号 | 示す | に表 1 及び図 7 | は | 、| 表している | をもの次の | 。 in this case , the respective symbols shown in Table 1 and FIG. 7 represents the followings .
英語参照文	here , symbols shown in Table 1 and FIG. 7 represent the following items .

表1: 既存手法と提案手法の比較

ベースライン	（前略）| エネルギーが | 240 keV 、ドーズ量が | 4×1012/cm2
提案手法	（前略）| 4×1012/cm2 | 240 keV 、ドーズ量が | エネルギーが

表2: 提案手法の並べ替え失敗例

く用いない設定をベースラインとした。

4.1 実験設定

実験では、対訳データに NTCIR-9 特許機械翻訳テストコレクション[*2]の日英翻訳データ約 320 万文対を用いた。そのうち評価データには 2,000 文対を、開発データには 500 文対を使用し、訓練データは 1 語以上 64 語以下となるようフィルタリングした。対訳データでは、日本語の単語分割および形態素解析に JUMAN 7.0 を、構文解析に KNP 4.1beta を使用した。構文解析に失敗した部分は訓練データから取り除いた。

共通の統計的機械翻訳システムには、SRILM 1.7.0[14] の 6-gram 言語モデル、MGIZA 0.7.3[1]、Moses 1.0[10] をそれぞれ使用した。チューニングには MERT[12] を使用し、デコード時には事前実験で最も良かった distortion limit 10 を設定した。

評価尺度には、BLEU[13]、RIBES[5]、また並べ替えのみの正解率を測るため、Kendall's τ[6] の平均[*3]を用いた。

4.2 実験結果

表 3 に比較実験の結果を示す。提案手法は、BLEU と RIBES の両方の評価尺度で既存手法全てを大幅に上回り、局所的にも大局的にも翻訳精度を著しく改善している。

また、並べ替え後の語順と理想的な語順との順位相関を測る Kendall's τ の平均値においても、提案手法の値が最高となったことから、並べ替え精度でも提案手法が最良であると分かる。一方で、提案手法も $\tau = 1.0$ となる monotonic な語順には程遠く、未だ改良の余地が大きいことが示されている。

事前並べ替え手法	BLEU	RIBES	τ
ベースライン	27.57	68.08	0.3935
Katz-Brown and Collins	29.87	73.10	0.5186
Hoshino et al.	30.56	72.37	0.5829
提案手法	**31.14**	**74.14**	0.6091

表3: 比較実験結果

τ は評価尺度の 1 つである Kendall's τ の平均、太字はブートストラップ・リサンプリング [9] におけるその他全手法との統計的有意性（p<0.01）を表している。

4.3 誤り分析

表 2 に提案手法で正しく並べ替えられなかった例を示す。この例文は、「エネルギーが 240 keV」かつ「ドーズ量が 4×1012/cm2」という is-a 関係の列挙から構成されている。しかし提案手法は、文節内の並べ替えとは独立して文節間の並べ替えを行うため、文節が想定された順序に並べ替えられているにも関わらず、文節内の語順が局所的に逆順になるという問題が発生してしまった。

このような問題を解決するために、追加規則として文節を入れ替えた時の文節内の語順の整合性を確認することも考えられるが、規則 2,3 がより複雑になるにつれて頑健性が確保できなくなる恐れがある。

そこで今後は、提案手法のように規則を人手で与えるのではなく、統計的に並べ替え規則を学習する手法 [15, 2] に取り組んでいきたい。

[*2] http://research.nii.ac.jp/ntcir/permission/ntcir-9/perm-ja-PatentMT.html

[*3] 算出には en-ja.A3.final ではなく ja-en.A3.final を使用した。

5 おわりに

本研究では、構文情報を用いた文節間の並列関係の並べ替えとその他の並べ替えを規則によって切り分けた、構文解析誤りに頑健な事前並べ替え手法を提案した。特許文書において翻訳精度を比較したところ、提案手法は既存手法を全て上回り、これらの規則の有効性が確認できた。

参考文献

[1] Qin Gao and Stephan Vogel. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57, 2008.

[2] Dmitriy Genzel. Automatically learning source-side reordering rules for large scale machine translation. In *Proc. of COLING*, pages 376–384, 2010.

[3] Sho Hoshino, Yusuke Miyao, Katsuhito Sudoh, and Masaaki Nagata. Two-stage pre-ordering for Japanese-to-English statistical machine translation. In *Proc. of IJCNLP*, pages 1062–1066, 2013.

[4] Hideki Isozaki. OkaPU's Japanese-to-English translator for NTCIR-10 PatentMT. In *Proc. of the 10th NTCIR Workshop Meeting*, pages 348–349, 2013.

[5] Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. Automatic evaluation of translation quality for distant language pairs. In *Proc. of EMNLP*, pages 944–952, 2010.

[6] Hideki Isozaki, Katsuhito Sudoh, Hajime Tsukada, and Kevin Duh. Head finalization: A simple reordering rule for SOV languages. In *Proc. of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 244–251, 2010.

[7] Jason Katz-Brown and Michael Collins. Syntactic reordering in preprocessing for Japanese→English translation: MIT system description for NTCIR-7 patent translation task. In *Proc. of the NTCIR-7 Workshop Meeting*, 2008.

[8] Jason Katz-Brown, Slav Petrov, Ryan McDonald, Franz Och, David Talbot, Hiroshi Ichikawa, Masakazu Seno, and Hideto Kazawa. Training a parser for machine translation reordering. In *Proc. of EMNLP*, pages 183–192, 2011.

[9] Philipp Koehn. Statistical significance tests for machine translation evaluation. In Dekang Lin and Dekai Wu, editors, *Proc. of EMNLP*, pages 388–395, 2004.

[10] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *Proc. of ACL Demo and Poster Sessions*, pages 177–180, 2007.

[11] Mamoru Komachi, Yuji Matsumoto, and Masaaki Nagata. Phrase reordering for statistical machine translation based on predicate-argument structure. In *Proc. of IWSLT*, pages 77–82, 2006.

[12] Franz Josef Och. Minimum error rate training in statistical machine translation. In *Proc. of ACL*, pages 160–167, 2003.

[13] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proc. of ACL*, pages 311–318, 2002.

[14] Andreas Stolcke, Jing Zheng, Wen Wang, and Victor Abrash. SRILM at sixteen: Update and outlook. In *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop*, 2011.

[15] Roy Tromble and Jason Eisner. Learning linear ordering problems for better translation. In *Proc. of EMNLP*, pages 1007–1016, 2009.