E-020

英日機械翻訳のための主辞後置事前並び替えにおける限量詞移動の改善

An Improvement in Quantifiers Movement of Head Finalization Reordering for E-J Machine Translation

谷口 正訓 *1
Masanori TANIGUCHI

丁 塵辰 *1
Chenchen DING

山本 幹雄 *2
Mikio YAMAMOTO

1. はじめに

現状の統計的機械翻訳手法は、フレーズの翻訳などの局所的な翻訳は高性能であるが、語の移動など大局的な調整にまだ改良の余地がある。例えば、英仏翻訳など語族的に近い言語間の翻訳に比べて、英日翻訳などの語の移動が大きい言語間の翻訳性能は低い。これを改善するため、原言語を予め目的言語と同じ語順となるように並び替えを行ったあと統計的機械翻訳を行うという事前並び替え手法が研究されている。本論文では、事前並び替え手法のうち、英日向けでシンプルかつ高性能な Head Finalization[2] に注目し、この手法で失敗しやすい限量詞の並び替えを改善する手法を検討する。翻訳実験を行ったところ、BLEU は 0.32% 上昇し、有意水準 5% で有意差が認められた。

2. 先行研究

2.1.Head Finalization

Head Finalization[2] は日本語の主辞後置性を利用して英文を日本語と同様の語順に並び替える手法である。具体的には、主辞駆動句構造文法による構文解析で得られた情報を元に構文木を作成し、全てのノードに対し英語の主辞 (head) となるノードを後ろ移動するという操作を行うことにより並び替えを実現する。構文解析結果から作成した構文木の例を図 1(左) に示す。破線の方向に存在するノードが主辞ノードとなる。上から 2 段目 VP ノードの子ノードのように主辞が前方に位置する場合は主辞ノードを後ろに移動する。このような処理を全ノードに対して行うことにより、図 1(右) のように英文を主辞後置言語である日本語と同様の語順に並び替えることができる。この並び替えに格助詞相当語 (_va1、_va2) の追加などの細かいルールを加えてさらに日本語の語順に近づける操作を行う。

2.2.Head Finalization の問題点

英日で構造的に異なる表現となる語が存在する。そのような語は例えば限量詞の一部に見られ、“no”、“little”、“low” 等がこれに該当する。構造が変化すれば主辞も変化するため、ノード間の主辞関係を用いて並び替えを行う Head Finalization ではこれに対応できない。図 2(左) の場合、“no” の部分が日本語の否定文の構造と異なるため、並び替えを行っても日本語の適切な場所へは位置しない。並び替え後の図 2(右) を見ると、日本語の “形成しない” に対応する “no” と “forms” が大きく離れていることが分かる。そのため、“no” と “forms” が同じフレーズとして扱われず、“no” の翻訳される場所が不適切になる場合や、言語モデルにより日本語として不自然な位置の否定詞が削除される傾向がある。

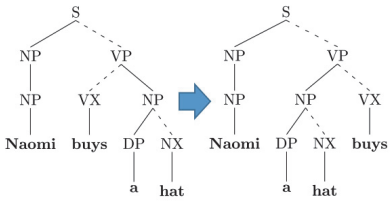

図 1　主辞後置操作の例

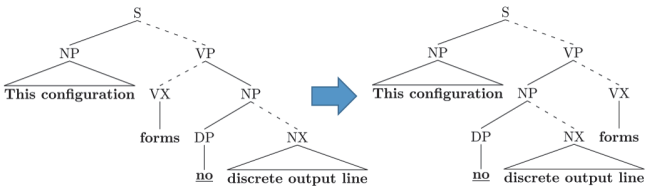

図 2　“no” を含む文の主辞後置例

図 2 の例では対象単語は日本語における接尾語に対応しているが、用法によっては接頭語に対応する場合もある。“a no load induced voltage is 〜”(無負荷誘起電圧が〜) を例として示す。

限量詞は肯定文を否定文に変えるなど文の意味を大きく変更し、翻訳される場所によっても文の意味が大幅に変化する。したがって、対象単語が日本語における接頭語と接尾語のどちらに対応するのかの分類を行った後、適切な位置へ移動させる必要がある。

3. 提案手法

本稿では、英日で構造が異なる表現となる可能性のある限量詞の一部を日本語の適切な位置に移動する方法を提案する。

対象単語の日本語における適切な位置を、英語の構造と同様の場所か、対象単語が日本語において結びつく可能性のある動詞の後の 2 カ所と仮定し、SVM による移動判定を行う。SVM の素性には、対象単語、対象単語の主辞となる語 w_e、対象単語が含まれる節の動詞 w_j、の各ノードの要素を用いる。各ノードの要素には構文解析で得られた品詞等の情報を用いる。例を図 3 に示す。w_e は対象単語の親ノードから主辞をたどり到達した語となり、w_j は対象単語を意味の上で包含する動詞句ノードから主辞をたどり到達した語となる。この場合、対象単語は “no”、w_e は “line”、w_j は “forms” となる。対象

*1 筑波大学大学院システム情報工学研究科 {m.taniguchi,tei}@mibel.cs.tsukuba.ac.jp
*2 筑波大学大学院システム情報工学系 myama@cs.tsukuba.ac.jp

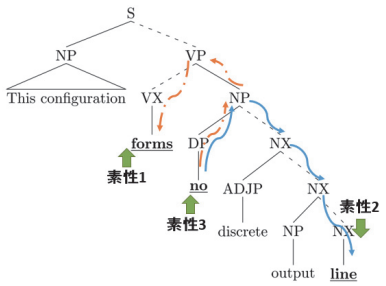

図 3　SVM で用いる素性ノードの例 (対象単語を包含する最近の動詞句から主辞をたどった先が素性ノード 1、対象単語の親ノードから主辞をたどった先が素性ノード 2、対象単語のノードが素性ノード 3 となる)

単語が節内に存在しない場合 w_j は存在しないが、直近の動詞句の主辞となる語、動詞句が存在しなければルートノードの主辞となる語を w_j とおき移動判定により対処を行う。対象単語のリストは人手で作成する。

4. 評価

4.1. 実験条件

Head Finalization によって並び替えられた文と並び替え後に限量詞の移動を行った文で翻訳実験をし比較を行う。

統計的機械翻訳のデコーダに Moses[*3]、単語アライメントに GIZA++[*4]、日本語形態素解析に Mecab[*5]、構文解析に Enju[*6]、移動判定に LIBLINEAR[*7] を用いた。

コーパスは NTCIR7 の特許コーパス [1] を用い、言語モデル作成にトレーニングコーパスの約 180 万文、トレーニングに、トレーニングコーパスの 1 文 40 単語以上のものをカットした約 110 万文、チューニング (MERT) にチューニングコーパスの 915 文、テストにテストコーパスの 1381 文を用いた。

実験には、Moses のフレーズベース翻訳システムを用い、言語モデルには Interpolated Kneser-Ney 5-gram を用いた。翻訳モデルには、grow-diag-final-and を用い、設定は、max-phrase-length=7、ttable-limit=20、stack=100 とした。

SVM の素性に用いる各ノードの要素は、Enju による構文解析で得られた要素とし、ノードの持つ語そのものと、その語の原形である “base”、品詞等を表す “cat”、“xcat”、“pos”、“pred”、“type”、“lexentry” を用いた。

対象単語は英日で構造が異なる表現となる可能性のある語のうち、出現数の多い “no”、“little”、“less”、“never” の 4 語を選択した。対象単語とトレーニングコーパスから作成した教師データのデータ数を表 1 に示す。教師データにおいて、対象単語が “never” の場合、ほぼ全ての文で日本語における接尾語に対応していたため移動判定を行わず全て移動するというルールを加えた。番号を表す “no .” については日本語における動詞

表 1　SVM の教師データ数

対象単語	no	little	less
教師データ数	255	151	204

表 2　翻訳結果の BLEU[単位:%](太字は各手法の最高値を示し、*はブートストラップ・リサンプリング法により Head Finalization の最高値 (35.14) と有意水準 5% で有意差があったものを示す)

dl =	3	6	9	12
Head Finalization	34.37	35.04	**35.14**	35.06
Proposed	34.12	35.27	35.45*	**35.46***

表 3　図 2 の翻訳結果

[HFE]	this configuration _va1 no discrete output line _va2 forms
Head Finalization	この 構成 は 、 離散 的 な 出力 線 を 形成 する
Proposed[HFE]	this configuration _va1 discrete output line _va2 forms no
Proposed	この 構成 は 、 離散 的 な 出力 線 を 形成 し ない

表 4　対象単語が反映されたか否かの目視調査結果

並び替え方法	成功語数	失敗語数
Head Finalization	11	13
Proposed	19	5

と結びつかないため移動判定を行わず全て移動しないというルールを加えた。

4.2. 実験結果

distortion limit(dl) を変更してテストコーパスの 1381 文に対し BLEU を測定した。その結果を表 2 に、翻訳例を表 3 に示す。表 3 の HFE は Head Finalized English の略であり並び替え後の英文を指す。

また、テストコーパス中の対象単語が含まれる文に対して、最も高い BLEU となる dl の場合に対象単語が翻訳結果に反映されているかの調査を目視で行った結果を表 4 に示す。表 4 の成功語は、対象単語が翻訳結果に反映された語とした。

5. おわりに

表 2、表 4 の結果から、変更箇所が少ない割に翻訳精度が向上していることが分かる。現状の課題として、対象単語のリストとその教師データを目視での判断により作成している点があげられる。今後、単語のアライメント情報を用いてこれらの自動化を行う予定である。

参考文献

[1] Fujii, A., Utiyama, M., Yamamoto, M., and Utsuro, T. (2008), Overview of the Patent Translation Task at the NTCIR-7 Workshop, In *Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologie*, 389-400.

[2] Isozaki, H., Sudoh, K., Tsukada, H., Duh, K. (2012) HPSG-Based Preprocessing for English-to-Japanese Translation, *ACM Transactions on Asian Language Processing*, 11(3):8:1–8:16.

*3 http://www.statmt.org/moses/
*4 http://www.statmt.org/moses/giza/GIZA++.html
*5 http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html
*6 http://www.nactem.ac.uk/enju/index.ja.html
*7 http://www.csie.ntu.edu.tw/~cjlin/liblinear/