

統計機械翻訳によるやさしい日本語書き換えの性能向上

熊野 正 田中 英輝

NHK 放送技術研究所

{kumano.t-eq, tanaka.h-ja}@nhk.or.jp

1 はじめに

NHK は 2012 年 4 月より、インターネットで国内在住の外国人や子供を対象としたやさしい日本語ニュース「NEWSWEB EASY」を提供している。各記事は、NHK のニュースサイト「NEWSWEB」で公開しているニュースを、まずニュースの構成技術を持つ記者が記事の内容を要約・簡易化した後に、やさしい日本語の知見を持つ日本語教師が語彙・表現を平易化することで作成している [7]。我々は、NEWSWEB EASY 制作業務の効率化を目指して、日本語教師が行うやさしい日本語への書き換えを統計機械翻訳によって実現する研究を進めている [5]。これは NEWSWEB EASY 制作の過程で得られた日本語教師書き換え前後の記事対を収集して、学習データとして用いるものである [4]。

これまでサービスの提供を継続してきたことで、現時点で 15,000 文対程度の学習データを得ることができた。広範なトピックのニュース文入力を十分カバーする書き換え知識を得るにはまだまだ学習データの規模が不足しているが、一般の言語間翻訳タスクと異なり、言語内翻訳タスクでは書き換え方を知らない単語を未知語としてそのまま出力しても意味が通らなくなることはない。入力と十分な書き換えの中間状態として、日本語教師による書き換えの下訳などに利用できる可能性がある。しかし、学習データの不足は単に書き換え知識のカバレッジ不足につながるだけでなく、共起頻度に基づく alignment 推定の精度低下の原因にもなる。この結果、誤った書き換えを引き起こし、下訳としての有用度が大きく低下してしまう。

低頻度の語句は統計的に対応を推定することが難しいが、一方で、それらの語句は固有表現など、やさしい日本語書き換えタスクでは書き換え不要であることも多い。従って、学習データの文対にともに同一の語句が現れているならば、それらは対応している可能性が高く、文対全体の alignment 推定精度を改善する強いヒントとなる。本稿では、この性質に基づいたヒューリスティクスを用いて、文対にともに表れている同じ

名詞列の間に対応制約を与えることで単語 alignment 推定精度を向上する手法を提案する。また、書き換え知識のカバレッジ不足への対処として、入力名詞列を non-terminal に汎化した Hiero 文法を用いる書き換え手法を検討した。前述の単語 alignment 手法との組み合わせによって顕著な書き換え性能向上が得られたので、あわせて報告する。

2 同一名詞列対応制約による 単語 alignment 推定の改良

これまでのやさしい日本語書き換えデータの分析から、体言は他の表現より書き換え時の欠落や湧き出しが起りにくく、かつ書き換えられずにそのまま用いられることが多いことや、特に長い名詞列は固有表現等の特定性の高い表現であり、さらに書き換えられない傾向が強いことなどがわかった。我々はこれを踏まえ、やさしい日本語書き換え文対間の対応に関して以下の仮定を置くことにする。

文対の一方を入力、他方を出力としたとき、

- 出力側長さ 2 以上の名詞列は、同じ名詞列が入力側に 1 つ以上あるならば、必ずそのいずれか 1 つと対応（すなわち、出力側名詞列中の各単語が順に入力側名詞列中の各単語と対応）している。出力側の複数の名詞列が同じ入力側名詞列と対応していてもよい。
- 出力側で重なりのある複数の名詞列に対しては、そのうち 1 つの名詞列のみ、名詞列対応を考える。例えば出力文「… 東京都知事 …」に対して、入力側に名詞列「東京都」と「都知事」がともに存在していても、「東京都」の名詞列対応と「都知事」の名詞列対応は両立しない。
- より長い名詞列の対応先が優先される。例えば出力文「… 東京都知事 …」に対して、入力側に

名詞列「東京都知事」が1個所、「都知事」が2個所（うち1個所は「東京都知事」の一部）ある場合には、「東京都知事」の名詞列対応が優先して決まるので、入力側の「都知事」と出力側の「東京都知事」の外にある「都知事」との対応可能性は考えない。

Gao らは、IBM model の推定時に文中の特定の翻訳先単語の対応先翻訳元単語を外部から強制的に指定することが可能な単語 alignment 推定手法を提案している [3]。我々は、この手法に基づく単語 alignment ツール MGIZA の拡張版 (MGIZA 配布物の中にこの拡張の試験的な実装が含まれている¹) を利用することで、上述の仮定に基づく単語対応制約を与えた単語 alignment 推定を実現する。

2.1 アルゴリズム

前述の仮定 3. に従って、学習データの各文対について、出力側に存在する名詞列の長いものから順に対応先を決定し、最後にすべての名詞列対応を単語対応に分解したものを制約として各文対に与えて、学習データ全体の (制約が与えられていない単語の) alignment を推定すればよい。ただし、ある出力側名詞列に対して入力側に同じ名詞列が複数ある場合や、仮定 2. のように同じ長さで両立しない複数の名詞列対応が考えられる場合には、最適な対応を推定し選択する必要がある。

文対に対するある長さ n の名詞列の対応づけ方に複数の選択肢がある場合、実際に各選択肢を対応制約に用いて単語 alignment を推定し、その結果得られる対応尤度に応じて最適なものを選択することができる。単語 alignment の推定は各文対独立には行うことができないため、学習データ全体に対して同時に、長い名詞列から順に対応づけ方を決定していくことにする。

学習データの i 番目の文対 (s_i, t_i) を考える。名詞列対応は、入力文および出力文での名詞列の区間を特定する [開始位置:終了位置] 情報の対の形式で表現することとし、すでに (s_i, t_i) に対して長さ $n+1$ 以上の名詞列について名詞列対応が決定しているものとして、これを名詞列対応集合 $\hat{A}_i^{\geq n+1}$ とする。この条件の下で、 (s_i, t_i) 中の長さ n の名詞列対応を決定するアルゴリズムを以下に示す。

1. (s_i, t_i) 中に存在する、長さ n の可能なすべての同一名詞列どうしの名詞列対応を列挙する。これ

¹<https://github.com/moses-smt/mgiza/tree/master/experimental/alignment-enabled/>

を名詞列対応集合 A_i^n とする。

2. すでに決定している長さ $n+1$ 以上の名詞列対応と両立 (名詞列対応の出力文側区間の重複がない) し、かつ可能な限り多く名詞列対応を追加できるような、長さ n の名詞列の対応づけ方の候補を列挙する。これを名詞列対応集合族 $\mathcal{A}_i^n = \{A_{i,1}^n, \dots, A_{i,m}^n\}$ と表す。各要素 $A_{i,j}^n$ ($1 \leq j \leq m$) は、 A_i^n のべき集合 (すべての可能な部分集合からなる集合族) のうち、以下の条件を満たす要素である。

- $\hat{A}_i^{\geq n+1} \cup A_{i,j}^n$ のどの2要素も互いに両立 (既決の名詞列対応と両立可能)
- $A \setminus A_{i,j}^n$ のどの要素も、 $A_{i,j}^n$ のいずれかの要素と非両立 (可能な限り多くの名詞列対応を与える)

3. \mathcal{A}_i^n の各要素に対応した、単語 alignment 推定のための文対事例を作成する。 $A_{i,j}^n$ に対応した事例は、以下の要素の組である。

- 事例重み $w_{i,j}^n = 1/m$
- 文対 (s_i, t_i)
- 長さ n 以上の名詞列対応制約 $\hat{A}_i^{\geq n+1} \cup A_{i,j}^n$ (MGIZA 処理時には単語対応制約に分解)

4. 以下を適当な回数反復する。

- (a) 学習データ全体の全文対事例をまとめて拡張版 MGIZA で処理し、各事例の単語 alignment を推定する。また、各事例の対応尤度 $s_{i,j}$ を得る。
- (b) 各事例の事例重みを更新する。

$$w_{i,j}^n \leftarrow (w_{i,j}^n s_{i,j}) / (\sum_{j'} w_{i,j'}^n s_{i,j'})$$

5. \mathcal{A}_i^n の中から長さ n の名詞列の最尤な対応づけ方を選択し、 $\hat{A}_i^{\geq n+1}$ と組み合わせて長さ n 以上の名詞列の名詞列対応 $\hat{A}_i^{\geq n}$ を決定する。

$$\hat{A}_i^{\geq n} \leftarrow \hat{A}_i^{\geq n+1} \cup A_{i,k}^n \quad (k = \arg \max_j w_{i,j}^n)$$

学習データ全体で可能な最大の名詞列長 n から上記のアルゴリズムで処理を始め、 n を1ずつ減じながら $n=2$ で処理を終了する。最後に、 $n=2$ の手順 2 の反復処理の最後で得られた単語 alignment 推定結果を、最終的な推定結果として採用する。一般的な phrase-based 等の翻訳モデル学習と同様に、単語 alignment 推定は、学習データの書き換え元側、書き換え先側をそれぞれ入力と考えたときの推定を個別に行い、後段の処理で対称化を行って利用する。

3 名詞列汎化 Hiero 文法の導入

学習データが不足している場合、phrase-based では翻訳時に有用な長さの phrase を用いることができず、翻訳性能に影響を及ぼすことが多い。hierarchical phrase-based、いわゆる Hiero [2] 文法の導入は、non-terminal を用いて phrase 対の汎化を行うことで、長い phrase のバリエーションを増やし、翻訳性能を向上できる可能性がある。

しかし一方で、構文構造の近い言語間の翻訳ではあまり有効ではないとされる。その理由は、1) 学習データが大量にあるならば、phrase-based でそれなりに性能が得られる、2) phrase-based より phrase 対中の細かな単語 alignment の精度が必要、3) 単語対応に交差が少ない場合、phrase 対中の汎化可能なスパンを制約する手がかりが乏しく、有害な汎化を行った規則を大量に抽出してしまう、などが考えられる。

本タスクの場合も、上記の問題があるため単純に Hiero 文法を導入しても性能向上を期待しにくい。しかし、前節の改良の結果、名詞列については alignment 精度の向上が見込めることから、汎化個所を入力側が名詞列である対に限定することで 3) の問題に対処して、Hiero 文法を導入してみる。

規則の抽出手法は、non-terminal に汎化可能な個所を限定する以外は、通常の Hiero 文法抽出手法と変わらない。また、デコーダも Moses のチャートデコーダをそのまま利用できる。

4 実験

2012年4月～2014年9月の NEWSWEB EASY 制作で得られた日本語教師書き換え前後の記事対より構築した、評価用データセット [4] (訓練データ: 10,769 文対²、開発データ: 723 文対、評価データ: 2,012 文対) を用いて自動書き換え器を試作し、提案手法の有効性を評価した。形態素解析器には MeCab+IPAdic³ を、機械翻訳ツールキットには Moses ver. 3.0 を、単語 alignment ツールには 2 節で述べた拡張版 MGIZA を、言語モデルツールキットには Moses 同梱の KenLM を、それぞれ使用した。

言語モデルには表層 7-gram と品詞 2-gram の factored モデル [1] を採用し、訓練データの書き換え後の全データから学習した。品詞言語モデルは、形態素解

²1 文の形態素数が 100 個を超える文対を分割するなど、若干の人手修正を行ったため、後藤らが報告した文対数と異なる。

³訓練データの書き換え前後の全データに対して人手で形態素正解タグを付与したうえで、書き換え前用/書き換え後用にそれぞれ、配布されているモデルを再学習して用いた。

析で得られた IPA 品詞体系の 1 階層目を用いた。翻訳モデル学習時には、学習データの書き換え元側異なり形態素の各々について、その語 1 語どうしの「文対」を生成し、学習データに追加した [5]。単語 alignment 推定の事前知識として機能する。また、単語 alignment の対称化ヒューリスティクスには grow-diag を用いた。

2 節の名詞列対応制約処理において、手順 2. の尤度推定の反復回数は 3 回とした。3 節の文法抽出処理においては、一般的な Hiero 文法抽出の設定と同様、入力側の non-terminal の接続は認めないこととした。また、汎化可能なスパン長には制限を設けない (連続名詞列の制約で十分であるため)。

4.1 各書き換え器の諸元

以下の 5 種類の書き換え器を作成し、性能を比較する。

dummy ベースライン。入力を書き換えなしで出力。

PB Moses 標準設定による phrase-based システム (distortion-limit は 0 とした)。

PB+NP-Const PB システムに対して、単語 alignment 推定時に 2 節の名詞列対応制約を導入。

NP-Hiero 3 節の名詞列汎化 Hiero 文法を導入した hierarchical phrase-based システム。単語 alignment 推定は通常の MGIZA。

NP-Hiero+NP-Const NPHiero システムに対して、単語 alignment 推定時に 2 節の名詞列対応制約を導入。

4.2 評価方法

データセットの評価データを用い、書き換え結果の BLEU 値による評価を行う。またこれに加えて、人手による「書き換え形態素数」および「形態素書き換え妥当率」の評価を行うこととした。これは、本研究が日本語教師の書き換えの下記の自動生成を目指すものであり、ある書き換えが対象読者にとって十分に適切なものでなくても、日本語教師が意味をとる上で問題ない (妥当な) ものであれば下記として無用でないであろう、という考えに基づく。妥当性評価は一般の日本語話者が行うことができ、専門性を要しないので、一般的な翻訳の人手評価に比べてコストが低い。

表 1: 評価結果

書き換え器	BLEU	書換数	妥当率
dummy	43.04	0	—
PB	47.71	154	66.9%
PB+NP-Const	47.86	144	68.1%
NP-Hiero	47.52	108	77.8%
NP-Hiero+NP-Const	48.07	111	80.2%

人手評価は、BLEU 値評価と同じ評価データの書き換え結果の一部、30 文（入力側 1,086 形態素）に対して、以下の手順で行った。

- 出力側各形態素に対して、入力側の 1 形態素を手手で対応づける。複数の出力側形態素が同じ入力側形態素に対応してもよい。
- 入力側形態素のうち、出力側の同じ形態素と 1 対 1 で対応づけられていないものを、書き換え形態素とする。この総数が書き換え形態素数である。
- 書き換え形態素の各々について、その書き換えが妥当かどうかをチェックする。

$$\frac{\text{（妥当な書き換え形態素の総数）}}{\text{（書き換え形態素数）}} = \text{形態素書き換え妥当率}$$
である。

4.3 結果・考察

4.1 節の各書き換え器による評価データ書き換えの評価結果を表 1 に示す。なお、デコーダの素性重みは、PB および NP-Hiero については開発データを用いて MERT チューニングを行った結果を用い、PB+NP-Const および NP-Hiero+NP-Const についてはそれぞれ PB、NP-Hiero と同じ値を用いた。

2 節の単語 alignment 推定への名詞列対応制約の導入は、phrase-based システムでの利用 (PB+NP-Const) では、BLEU 値、単語書き換え妥当率とも改善は微小だった。一方、3 節の名詞列汎化 Hiero 文法の導入 (NP-Hiero) は、単語書き換え妥当率の改善に効果があったものの、BLEU 値は逆に悪化している。これに対して、両者を同時に導入した場合 (NP-Hiero+NP-Const) は、BLEU 値、単語書き換え妥当率とも一定の改善が見られた。Hiero モデルが phrase-based モデルより単語 alignment の精度に強い影響を受けるという先行報告 [6] と同様に、本稿で提案する単語 alignment 推定の改良は効果は確かにあったが、単語 alignment の細かな誤りの影響を受けにくい phrase-based モデルでは十分な性能改善には至らず、

Hiero モデルとの組み合わせで初めて有意な改善となったと考えられる。

なお、上記いずれの改良も、書き換え形態素妥当率は上昇し、書き換え形態素数は減少する方向に働いている。書き換え数と書き換え妥当率はトレードオフの関係にあるため、これらは総合的に評価する必要がある。評価手法については今後検討したい。

5 おわりに

本稿では、統計機械翻訳を用いたやさしい日本語への自動書き換えタスクにおいて、学習データの規模が小さい場合の書き換え性能を改善するための学習手法を提案した。単語 alignment 手法の改良、および、入力側が名詞列である対のみを non-terminal に汎化した Hiero 文法の利用によって、BLEU 値は 47.71 → 48.07 に、単語書き換えの妥当率は 66.9% → 80.2% に、それぞれ向上した。

書き換え妥当性が高いことは、自動書き換え結果が日本語教師による書き換えの下訳として有用であるために重要であるので、引き続き書き換え性能の改善を進める。また、自動書き換えの利用が NEWSWEB EASY 制作業務の効率化につながるための条件についても、調査を進めていきたい。

参考文献

- [1] Amittai E. Axelrod. Factored language models for statistical machine translation. Master's thesis, University of Edinburgh, 2006.
- [2] David Chiang. Hierarchical phrase-based translation. *Computational Linguistics*, Vol. 33, No. 2, pp. 201–228, 2007.
- [3] Qin Gao, Nguyen Bach, and Stephan Vogel. A semi-supervised word alignment algorithm with partial manual alignments. In *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR*, pp. 1–10, 2010.
- [4] 後藤功雄, 熊野正, 田中英輝. やさしい表現へのニュースの自動変換評価用データセットの構築. 言語処理学会第 21 回年次大会, pp. 151–154, 2015.
- [5] 熊野正, 後藤功雄, 田中英輝. 統計機械翻訳を用いたニュース文のやさしい日本語への自動変換. 2015 年映像情報メディア学会年次大会, 32D–2, 2015.
- [6] 丹生伊左夫, Graham Neubig, 小林和也, Sakriani Sakti, 戸田智基, 中村哲. 構文情報が機械翻訳に及ぼす影響の分析. 情報処理学会研究報告 自然言語処理研究会報告 2013-NL-212(8), pp. 1–7, 2013.
- [7] 田中英輝, 美野秀弥, 越智慎司, 柴田元也. やさしい日本語ニュースの公開実験. NHK 技研 R&D, No. 139, pp. 20–29, 2013.