複数の事前並べ替え候補を用いた句に基づく統計的機械翻訳

小田 悠介 † 工藤 拓 ‡ 中川 哲治 ‡ 渡辺 太郎 ‡

† 奈良先端科学技術大学院大学 情報科学研究科
‡ グーグル株式会社

oda.yusuke.on9@is.naist.jp, {taku, tnaka, tarow}@google.com

1 はじめに

句に基づく統計的機械翻訳 (PBMT) [1] では原言語と目的言語の語順の違いをどのように扱うかが問題となる. 並べ替えモデル [2, 3, 4] は訳出時に句同士の位置関係を考慮する手法だが, 原言語の大域的な情報を取り入れるのが難しく, 結果として統語的に正しくない語順を生成してしまう可能性がある. また翻訳器は事実上任意の順序で句の連結を考慮する必要があり, 訳出処理の複雑さを増加させる要因にもなっている. 一方, 事前並べ替え [5, 6, 7, 8] は訳出処理の前に原言語の文を目的言語の語順へと変換する手法であり, 原言語について大域的に妥当な並べ替え結果を生成しやすいと考えられる. しかし PBMT は単一の入力単語列を前提としており, 事前並べ替え後の語順に誤りがある場合の対処が難しい. 事前並べ替え候補が複数得られる場合は単純に全ての候補を翻訳する手法も考えられるが [9], より多くの処理時間を必要とするために実用面で問題がある. このため従来の PBMT では並べ替えモデルと事前並べ替えを併用することが多いが, これらは互いに性質の異なる手法であり, 組み合わせた際の影響を推測することが困難となる.

本研究では, 単一の文に対する複数の事前並べ替え候補を用いた訳出手法を提案する. 具体的には, 複数の事前並べ替え候補を単一のグラフ構造として表現し, 一度の処理で全ての事前並べ替え候補を考慮した翻訳文を生成する手法について述べる. 翻訳器の入力として並べ替えに関するグラフを扱う手法は他にも提案されているが [10, 11], これらは主に事前並べ替え手法自身がグラフを生成する手法である. 提案手法はこれらとは異なり, 既存の事前並べ替え手法が出力する単語の並べ替え候補のみを用いてグラフ構造を生成する. このため任意の事前並べ替え手法を前処理として用いることができ, より汎用性の高い手法であると考えられる. また提案手法は訳出時に考慮する並べ替えを制限することに相当し, 従来の並べ替えモデルに基づく手法よりも訳出処理の簡略化が可能である.

実験により, 複数の言語から英語への翻訳において提案手法により翻訳精度が向上することを示す. また訳出に必要な計算時間が従来の並べ替えモデルに基づく手法と比較して短い点, 訳出の計算量と翻訳精度の向上に明確な関係がある点も示す.

2 事前並べ替え候補のグラフ表現

本節では複数の事前並べ替え候補を単一のデータ構造として表現する手法を述べる. まず, 入力となる原言語の単語列 $S = [s_0, s_1, \cdots, s_{I-1}]$ に対して, 事前並べ替えにより複数の並べ替え候補 $\mathcal{A} = \{A_1, A_2, \cdots, A_N\}, A_n = [a_0^n, a_1^n, \cdots, a_{I-1}^n]$ とその信頼度 $\mathcal{C} = \{C_1, C_2, \cdots, C_N\} \in \mathbb{R}^N$ が得られているものとする. ここで I は原言語の単語数, N は事前並べ替え候補の数である. 各 $a_i^n \in \{0, 1, \cdots, I-1\}$ は事前並べ替え候補で i 番目に現れる単語の入力単語列での位置とする. 事前並べ替え候補の単語列を前から順に翻訳するものと仮定すると, 訳出過程で現れる「翻訳済み単語」(カバレッジ) の集合が事前並べ替え候補から $\{\}, \{a_0^n\}, \{a_0^n, a_1^n\}, \cdots$ のように明示的に得られる. これらの翻訳済み単語の集合を頂点, 次に翻訳される単語を辺として, 単一の事前並べ替え候補を図 1(a) に示す有向非巡回グラフとして表現する. 以後はこのグラフを指して「並べ替えグラフ」と呼ぶこととする.

並べ替えグラフは各事前並べ替え候補について個別に得られるが, 本研究では複数の並べ替えグラフを統合し, 少なくとも全ての事前並べ替え候補を含むような単一のグラフを生成する. 具体的には, 異なる並べ替えグラフ中に同じ翻訳済み単語の集合が存在する場合, これらを同一と見なし, 並べ替えグラフの頂点を統合する. 例えば図 1(b) では, 新たな並べ替えグラフのうち{}, {5}, {0-5}, {0-6}が共通しており, これらに対応する頂点を統合している. この操作を全ての事前並べ替え候補に対して適用することで, 最終的に図 1(c) のような並べ替えグラフが生成される. また訳出時に事前並べ替えの信頼性を考慮できるよう, 各辺には事前並べ替え候補の信頼度に基づいてスコアを付与する. 本研究ではこのスコアとして, 事前実験で最も高い翻訳精度となった各辺の信頼度の最大値と単語数の比を採用した. 図 1 の各 c_n は C_n/I を表す.

このようにして生成された並べ替えグラフは統合された全ての事前並べ替え候補を含むことが保証されるが, 元の集合には含まれないような事前並べ替え候補も暗黙的に含むことになり, 実際には並べ替えグラフを用いることでより多くの事前並べ替え候補を考慮することとなる.

3 訳出処理

並べ替えグラフを用いた訳出では, 並べ替えモデルを適用せず, 単純にグラフ構造に従って動的計画法を実行することで, 従来の訳出よりも効率の良い探索を行う. 具体的には, 図 1 の左端から右端へ向けて, 各頂点における部分的な翻訳候補を順に確定してゆくこととなる. 並べ替えモデルを考慮しなくてもよい点については次章において実験結果とともに示す.

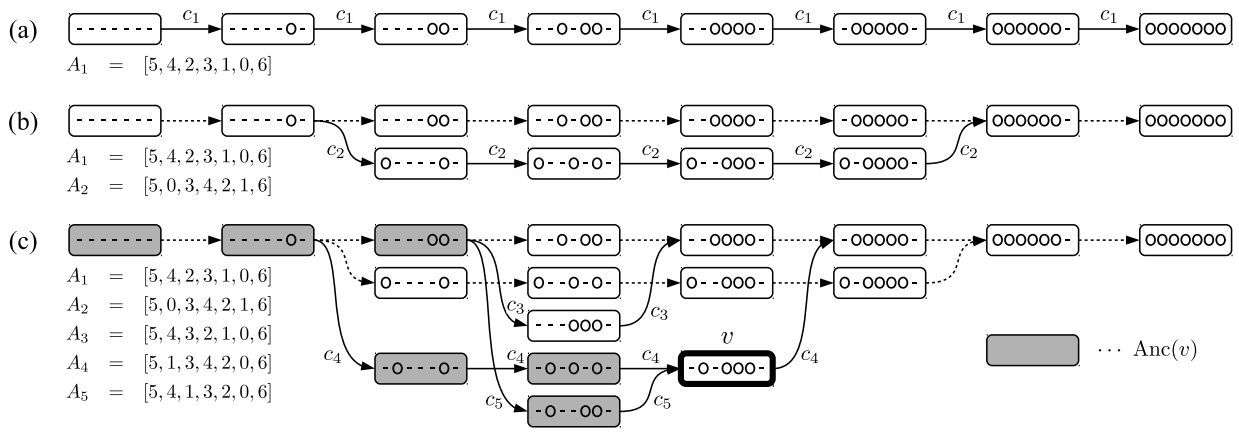


図 1: 事前並べ替え候補集合からの並べ替えグラフの構築

Algorithm 1 並べ替えグラフに基づく訳出生成

```
G ← Reordering Graph
v_L ← Leftmost(G)
H[v_L] ← {""}
for v ← TopologicalSort(Nodes(G)\{v_L}) do
    H[v] ← { }
    for v' ← Anc(v) do
        for h' ← H[v'] do
            for φ ← PhrasePairs(v', v) do
                H[v] ← H[v] ∪ {Concat(h', φ)}
            end for
        end for
    end for
    H[v] ← Prune(H[v])
end for
return BestResult(H[Rightmost(G)])
```

今, 並べ替えグラフ上のある頂点, 例えば図 1(c) の太枠で示した頂点 v に注目する. v に到達可能な祖先の頂点の集合 Anc(v) は並べ替えグラフを逆に辿ることで得られ, 図 1(c) の色付きの頂点が該当する. v に対する翻訳候補を得るには, 各 $v' \in$ Anc(v) について v' に対する翻訳候補と v' から v へのフレーズペアの組み合わせを列挙し, スコアの高い候補を選択すればよい. Algorithm 1 にスコア計算を考慮しない場合の擬似コードを示す.

ここで Leftmost(G), Rightmost(G) は並べ替えグラフの左端および右端の頂点を表し，それぞれ翻訳開始前，および全ての単語が翻訳された状態に相当する. また PhrasePairs(v', v) は v' から v へ至る経路に現れる原言語の単語列からフレーズペアを得る操作を表す. 並べ替えグラフの形状によっては元の事前並べ替え候補集合には存在しないような経路が出現する可能性もあるが, このような経路を全て列挙するのは計算量の面で問題がある. このため, 本研究では元の事前並べ替え候補集合に存在する経路のみを列挙の対象とした.

実際の訳出処理では Algorithm 1 に加えて, 各翻訳候補に対して対数線形モデルに基づくスコアの付与を行い, これに基づいて候補の枝刈りを行う. 並べ替えグラフの辺が持つスコアは対数線形モデルの素性の一つとして探索時に考慮される.

4 実験

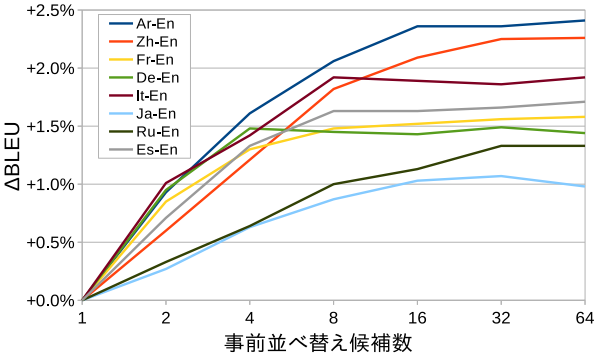
4.1 実験設定

英語を目的言語とした翻訳に提案手法を適用し, 単一の事前並べ替え候補のみを用いた従来の PBMT と翻訳精度の比較を行った. 原言語は英語に対して異なる種類の語順を持つ言語とし, アラビア語, 中国語, フランス語, ドイツ語, イタリア語, 日本語, ロシア語, スペイン語を選択した. 訓練データには Web から自動的に収集した対訳文を使用し, 開発, テストデータには Web から収集した英語 3000 文, 5000 文を人手翻訳したものを使用した. 以下に示す各手法の最適化は全て開発データで行い，翻訳精度の比較はテストデータで行った. 事前並べ替え手法としては Bracketing Transduction Grammar に基づく手法 [8] を採用し, ビーム探索による N-best 候補を $\mathcal{A}$, 対応するスコアを $\mathcal{C}$ とした.

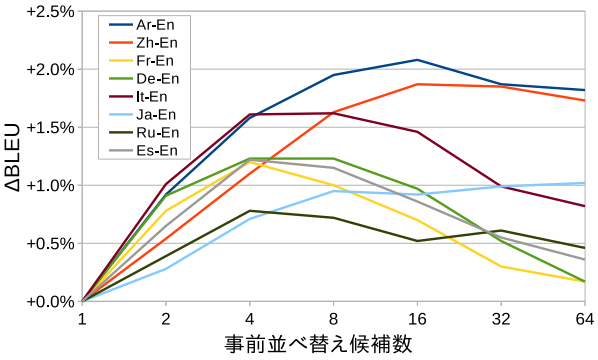
既存手法として, 訳出時の並べ替えモデルを含む state-of-the-art の PBMT システムを用意した. 入力には最も信頼度の高い事前並べ替え候補を使用し, 各言語対について訳出時の並べ替え制限を 0 から 6 単語の間で変化させた各モデルを作成し, 最も BLEU [12] の高いものを選択した. また, 事前並べ替え候補を統合せず従来の PBMT で個別に翻訳した場合として, 事前並べ替えの信頼度 C と訳出の信頼度 D を式 (1) で線形補間し, 最もスコアの高い翻訳候補を選択する再ランキング法についても同様に実験を行った. 再ランキング法は従来の PBMT を事前並べ替え候補の数だけ実行するため, 実行時間はこれに比例して長く必要となる. このため, 実際のシステムにおける再ランキング法の使用は想定していない.

$$Score := \lambda \cdot C + (1 - \lambda) \cdot D \tag{1}$$

提案手法, 再ランキング法については適用する事前並べ替え候補数を 1, 2, 4, 8, 16, 32, 64 個と変化させ, それぞれ個別に最適化を行い, 最も BLEU の高いモデルについて従来の PBMT との精度の比較を行った. 再ランキング法については並べ替え制限が上述の最良設定の場合, および 0 単語の場合について実験を行った. また提案手法については各言語対ごとに翻訳



(a) 並べ替えの信頼度を考慮



(b) 並べ替えの信頼度を無視

図 2: 事前並べ替え候補数による BLEU の変化

結果 400 文をランダムに抽出し, PBMT の翻訳結果と併せて 0(最低) から 6(最高) の 7 段階による主観評価を行った.

各手法の対数線形モデルの最適化には MERT [13] を用いた. 再ランキング法の λ は 0 から 1 まで 0.01 ごとに個別に評価し, 最も翻訳精度の高い値を選択した. またフレーズペアは各言語対ごとに同じデータ, 英語の言語モデルは全ての実験で同じデータとなっている.

4.2 実験結果と考察

表 1 に, 従来の PBMT, 提案手法, 再ランキング法についての BLEU を示す. また PBMT では採用した並べ替え制限 (DL) , 提案手法については主観評価による平均値の差分と両側 t 検定による p 値, および実際の翻訳結果が PBMT から変化していた文の割合 (CR) も示す. なお $\emptyset$ は微小値を表す.

まず, おおよその言語対について提案手法と再ランキング法の BLEU の傾向は一致しており, いずれも従来の PBMT と比較して翻訳精度の向上が見られる. このため, より多くの事前並べ替え候補を考慮することは翻訳精度の向上に直接貢献するものと考えられる. また再ランキング法の並べ替え制限の有無による BLEU を比較すると, 並べ替えを行わない場合の方が BLEU が向上している言語対もあり, 事前並べ替え候補を複数用いる場合に並べ替えモデルを考慮することが必ずしも有効であるとは限らないことが分かる. この結果より, 提案手法の枠組みでも並べ替えモデルを考慮する必要はないと考えられる.

主観評価では全ての言語対で提案手法が良い翻訳を生成していることが分かり, 特にアラビア語と中国語以外では 95%以上の有意水準で統計的に有意な差が認められる. また翻訳結果の異なり率を見ると, アラビア語, 中国語, 日本語で半数以上の文が PBMT と異なる翻訳結果を生成していることが分かる. これらの言語は英語とは語順の異なる傾向が強く, 事前並べ替え結果のばらつきが最終的な翻訳に与える影響が大きい. アラビア語と中国語で主観評価の p 値が相対的に高いのはこの傾向を反映しているものと考えられる. 対照的に, 日本語では翻訳結果が大きく異なるにも関わらず有意に良い翻訳を生成している. 日本語は語順の曖昧性が強い言語であり, 英語のような語順が比較的厳密に定まる言語への事前並べ替えは本質的に難しい. このため, 複数の事前並べ替え候補を考慮することが特に有効に作用したものと考えられる. その他の英語と語順の比較的近い言語については翻訳異なり率が小さく, また統計的に有意な翻訳精度の向上が見られる. このため, 提案手法が事前並べ替えの誤りを選択的に改善しているものと考えられる.

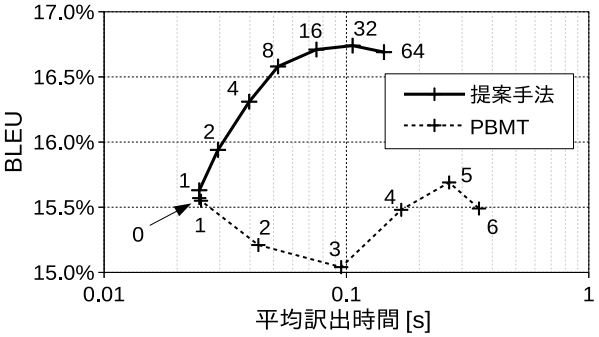


図 3: 日英翻訳における訳出時間と BLEU の関係

図 2 に, 並べ替えの信頼度を考慮および無視した場合について, 提案手法で用いる事前並べ替え候補数による BLEU の変化を示す. 縦軸の基準は候補数が 1 の場合の BLEU であり, 並べ替え制限を 0 単語とした場合の従来の PBMT とほぼ一致する. また言語対ごとに実際の BLEU の範囲が異なることに注意する. 図 2(a) より, いずれの言語対においても事前並べ替え候補数を増加させることで翻訳精度が向上していることが分かる. また図 2(a) では全ての言語対で翻訳精度の飽和は見られるものの低下は見られず, 対照的に (b) では事前並べ替え候補数の増加に従い翻訳精度の低下が見られる. これは妥当でない並べ替えを多数考慮した結果, 訳出のスコアが偶然高くなるような翻訳文を生成してしまっているためと考えられ, 事前並べ替えの信頼度を素性として導入することが翻訳精度を保証するのに有効であることが分かる.

図 3 に, 日英翻訳における提案手法と PBMT による平均の訳出時間と BLEU の関係を示す. 図中の線に付随する数値は, 提案手法では使用した事前並べ替え候補数, PBMT では並べ替え制限の単語数を表す. 提案手法は考慮する事前並べ替え候補数が増加するに従って訳出に多くの時間が必要となるが, その増加量は PBMT の並べ替え制限による変化に比べて小さいことが分かる. また PBMT では訳出時間と翻訳精度に明確な関係が見られない一方, 提案手法では訳出時間の増加に見合う翻訳精度の向上が得られていることも確認できる.

表 1: 各手法における翻訳結果の評価

言語対	PBMT		提案手法				再ランキング法 BLEU%	
	BLEU%	DL	BLEU%	主観評価 Δ	p	CR%	DL>0	DL=0
Ar-En	36.47	6	36.99 (+0.52)	+0.066	0.063	64.55	37.26 (+0.79)	36.88 (+0.41)
Zh-En	29.93	6	31.14 (+1.21)	+0.096	0.070	78.44	30.96 (+1.03)	31.35 (+1.42)
Fr-En	33.19	5	34.03 (+0.84)	+0.157	∅	27.56	33.87 (+0.68)	33.93 (+0.74)
De-En	30.45	6	31.05 (+0.60)	+0.111	∅	30.10	31.53 (+1.08)	31.27 (+0.82)
It-En	37.59	5	38.22 (+0.63)	+0.074	0.001	25.66	38.62 (+1.03)	38.31 (+0.72)
Ja-En	15.66	5	16.68 (+1.02)	+0.238	∅	73.35	17.00 (+1.34)	16.53 (+0.87)
Ru-En	25.58	6	25.79 (+0.21)	+0.057	0.017	29.74	26.12 (+0.54)	25.54 (−0.04)
Es-En	34.41	2	36.11 (+1.70)	+0.133	∅	40.44	36.06 (+1.65)	36.36 (+1.95)

表 2: 日英翻訳における PBMT と提案手法の翻訳例と主観評価スコア

種類	文	評価
原文	では, この問題をどうやって解決するつもりですか.	
PBMT	So, are you going to solve **how** this problem.	1
提案手法	So, **how** do you intend to solve this problem.	6
原文	私の車は, 私を含む全員がシートベルトを着用するまで駆動しません.	
PBMT	My car, **everyone including the I does not drive up** to wear a seat belt.	1
提案手法	My car **does not drive until everyone, including me** to wear a seat belt.	5
原文	技術革新により, 情報と画像をカードの表面に印刷できます.	
PBMT	By technological innovation, you can print the **information and images** on the card surface.	6
提案手法	By technological innovation, you can print the **image information** on the surface of the card.	4

表 2 に日英翻訳における提案手法の翻訳例を挙げる. 最初の 2 例は提案手法が効果的に作用した例である. 従来の PBMT では語順に起因する訳出の失敗が見られるが, 提案手法ではより正しい語順となっていることが分かる. 最後の例は提案手法が失敗した例である. この例のように, 並列構造のような並べ替えの難しいパターンで提案手法を適用すると, 言語モデルなどの強力な素性が並べ替えの信頼度を上回り, これに引きずられて誤った並べ替えを選択してしまう可能性がある. このような誤りをどのように抑制するかが今後の課題となる.

5 おわりに

本研究では句に基づく統計的機械翻訳の新たな訳出手法として, 複数の事前並べ替え候補を単一のグラフ構造として表現し, このグラフ上を探索することで高精度かつ高速な訳出を行うアルゴリズムを示した. 実験により, 提案手法が従来の PBMT と比較してより高い翻訳精度を示し, 訳出処理の計算速度についても同等以上の性能であることを示した.

本研究では単一の事前並べ替え手法のみについて実験を行ったが, 並べ替えの信頼度を何らかの方法で計算可能であれば, 他の事前並べ替え手法についても本手法をそのまま適用可能である. 今後は他の事前並べ替え手法についても実験を行い, それぞれの手法の特徴がどのように影響するのかを検証したい. また並べ替えグラフに複数の素性を導入することで, 異なる事前並べ替え手法を組み合わせることが可能であると考えられる. これについても今後の研究課題としたい.

謝辞

本研究の一部は日本学術振興会特別研究員奨励費 (15J10649) の助成を受けたものである.

参考文献

[1] Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *Proc. NAACL-HLT*, pp. 48–54, 2003.

[2] Philipp Koehn, Amittai Axelrod, Alexandra Birch, Chris Callison-Burch, Miles Osborne, David Talbot, and Michael White. Edinburgh system description for the 2005 iwslt speech translation evaluation. In *Proc. IWSLT*, pp. 68–75, 2005.

[3] Richard Zens and Hermann Ney. Discriminative reordering models for statistical machine translation. In *Proc. WMT*, pp. 55–63, 2006.

[4] Michel Galley and Christopher D. Manning. A simple and effective hierarchical phrase reordering model. In *Proc. EMNLP*, pp. 848–856, 2008.

[5] Fei Xia and Michael McCord. Improving a statistical mt system with automatically learned rewrite patterns. In *Proc. COLING*, pp. 508–514, 2004.

[6] Hideki Isozaki, Katsuhito Sudoh, Hajime Tsukada, and Kevin Duh. Head finalization: A simple reordering rule for sov languages. In *Proc. WMT-MetricsMATR*, pp. 244–251, 2010.

[7] Graham Neubig, Taro Watanabe, and Shinsuke Mori. Inducing a discriminative parser to optimize machine translation reordering. In *Proc. EMNLP-CoNLL*, pp. 843–853, 2012.

[8] Tetsuji Nakagawa. Efficient top-down btg parsing for machine translation preordering. In *Proc. ACL-IJCNLP*, pp. 208–218, 2015.

[9] Chi-Ho Li, Minghui Li, Dongdong Zhang, Mu Li, Ming Zhou, and Yi Guan. A probabilistic approach to syntax-based reordering for statistical machine translation. In *Proc. ACL*, pp. 720–727, 2007.

[10] Jan Niehues and Muntsin Kolss. A POS-based model for long-range reorderings in SMT. In *Proc. WMT*, pp. 206–214, 2009.

[11] Teresa Herrmann, Jan Niehues, and Alex Waibel. Combining word reordering methods on different linguistic abstraction levels for statistical machine translation. In *Proc. SSST*, pp. 39–47, 2013.

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

[13] Franz Josef Och. Minimum error rate training in statistical machine translation. In *Proc. ACL*, pp. 160–167.