# 統計的機械翻訳における機能語を利用した階層的広範囲フレーズ並び替えルール

丁 塵辰 † 乾 孝司 ‡ 山本 幹雄 ‡
† 筑波大学 システム情報工学研究科
‡ 筑波大学 システム情報系情報工学域
{tei@mibel., inui@, myama@}cs.tsukuba.ac.jp

## 1 はじめに

機械翻訳においては、原言語から目的言語への単語・フレーズを変換することに加え、並び方も決めなければならない。しかし、対象言語の構文構造の相違の程度により、語順調整の困難さは大きく変化する。

図1と図2は単語単語対応テーブルであり、それぞれ仏英と日英の対訳文対の例である。

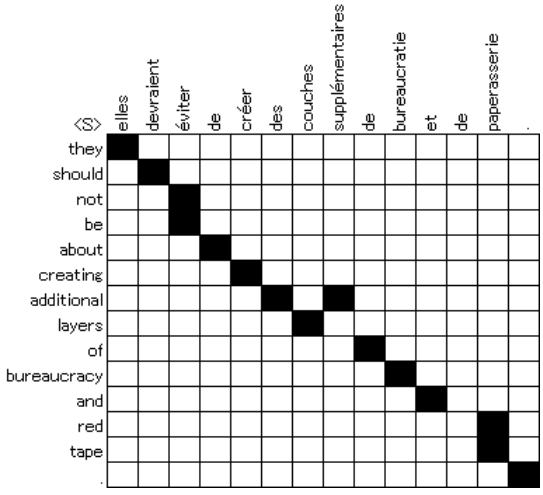


図 1: 仏英単語単語対応テーブルの一例

図1（仏英）の例を見ると、対応している単語（黒いマス）がほぼ主対角線に並んでいることから、仏英の間では単語・フレーズの並び方の一致性が高いことが分かる。それに対して図2の日英の単語対応テーブルにおいては、黒いマスが対角線に直交して散らばっており、日英間の語順の大きな違いが示唆される。

図1と図2はあくまで一例であるが、一般的に、仏英方向の翻訳において文頭から単語・フレーズごとに変換し、そのまま並べた後に局所的な調整を行えばおおよそ翻訳できる場合が多い。しかしながら、日英翻訳の場合はフレーズの並び方が大きく変化するため、強力な語順並び替えモデルで対応しなければ、意味が通る翻訳文が得られにくい。

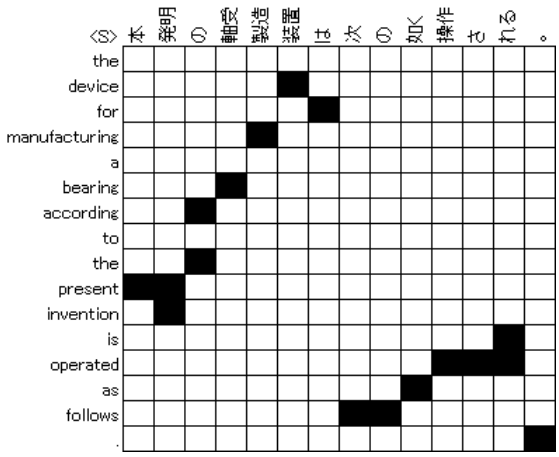


図 2: 日英単語単語対応テーブルの一例

現在のフレーズ翻訳システムは一般的に Lexical Orientation Model[Tillmann, 2004] と呼ばれるモデルを利用し、語順調整を行う。Lexical Orientation Model はフレーズ対の局所的な交換をする／しないといった傾向をモデル化するものであり、広範囲語順調整はできない。

一方、Chiang[Chiang, 2007] によって提案された階層フレーズモデルは、単語列の文脈を利用しながら、比較的広範囲の移動を正確に行える。しかし、従来の Chiang の抽出法で抽出された対訳階層フレーズ対の集合は非常に大きく、有用なルールとそうではないルールが混ざっている。このため、単純なヒューリスティックスで大量の無用なルールを削除すると、一部の有用なルールも同時に削除してしまう可能性が高い。

例えば、階層フレーズモデルの抽出時に「イニシャルフレーズの長さ」に上限を設定するとルール集合をコンパクトにできるが、広範囲の語順調整ルールが抽出できない。さらに、「非終端記号」の数を 2 に設定することで、3 個以上のフレーズを同時に移動することが不可能となってしまう。

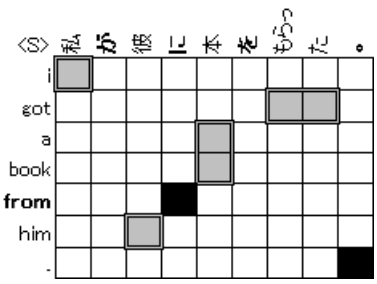


図 3: 構文構造の変換の例

図 3 は文構造レベルの語順調整の例である。二重線枠で囲まれる四つのフレーズ対はそれぞれ文の主語、間接目的語、直接目的語と述語部分であり、それらの移動を同時にコントロールすることは従来の階層フレーズモデルでは処理できない。つまり図 3 から

- $X \rightarrow < X_1$ が $X_2$ に $X_3$ を $X_4$ ;
  $X_1\ X_4\ X_3\ from\ X_2 >$

というような翻訳ルールが抽出できない。これに対して、本論文では文構造を正確にとらえる広範囲フレーズ並び替えルールを検討する。次節で抽出法を述べるが、提案の抽出法はイニシャルフレーズの長さと非終端記号数を明示的に制限しないことで、従来法より広範囲の対訳ルールが抽出できる。一方、これらのルールは特定の小さい単語集合のみに語彙化されるため、従来法よりルール数が大幅に削減される。

## 2 広範囲フレーズ並び替えルール

最初に原言語側文構造を示唆できる単語（例えば、機能語）からなる単語集合 $\mathcal{I}$ を決定する。この集合をイニシャル単語集合と呼ぶ。例えば、助詞、助動詞、接続詞、句読点といった単語である。

$f_1^J = f_1, f_2, \cdots, f_J$ と $e_1^I = e_1, e_2, \cdots, e_I$ の単語対応 $\mathcal{A}$ を次のように定義する。

$$\mathcal{A} = \{(i,j) \mid e_i\ is\ aligned\ to\ f_j\} \tag{1}$$

イニシャル単語集合 $\mathcal{I}$ を用いて、$\mathcal{A}$ をさらに $\mathcal{R}$ と $\mathcal{O}$ の二つの集合に分割する。

$$\mathcal{R} = \{(i,j) \mid (i,j) \in \mathcal{A}, and\ f_j \in \mathcal{I}\} \tag{2}$$

$$\mathcal{O} = \{(i,j) \mid (i,j) \in \mathcal{A}, and\ f_j \notin \mathcal{I}\} \tag{3}$$

単語対応付きの対訳文対に対して、Koehn らの方法 [Koehn et al., 2003] で抽出できるすべての対訳フレーズの集合を $\mathcal{P}$ で表記すると、集合 $\mathcal{R}$ と $\mathcal{O}$ に基づき、$\mathcal{P}$ のサブ集合 $\mathcal{P}_o$ が以下のように決定できる。ここで、対訳フレーズは図 3 のような単語対応テーブルの中の大きな矩形で表現できる。また、「$(i,j)\ covered\ by\ p$」はフレーズ対 $p$ の矩形が $(i,j)$ を含むことを意味する。

$$\mathcal{P}_o = \{p \mid p \in \mathcal{P},\ and\ \forall (i,j)\ covered\ by\ p : (i,j) \in \mathcal{A} \rightarrow (i,j) \in \mathcal{O}\} \tag{4}$$

集合 $\mathcal{P}_o$ は $\mathcal{O}$ に属するマスのみを覆う対訳フレーズ対の集合である。次に $\mathcal{P}_o$ の「最大フレーズ対」で構成される $\mathcal{P}_o$ のサブ集合 $\mathcal{M}$ を以下のように定義する。

$$\mathcal{M} = \{p \mid p \in \mathcal{P}_o,\ and\ \exists (i,j)\ covered\ by\ p : (i,j)\ is\ covered\ by\ q \in \mathcal{P}_o \rightarrow q = p\} \tag{5}$$

最終的に、$\mathcal{M}$ の中の対訳フレーズ対を一つの変数どうしの対とみなす。図 4 の例を変換した単語アライメントテーブルは図 5 のようになる。$\mathcal{M}$ 中のフレーズ対が変数 $X$ どうしの対となっている。

図 5 のような非終端記号を含む単語単語対応テーブルを用いて、従来の対訳フレーズ抽出法を応用すれば非終端記号を含む対訳フレーズ対（つまり、階層的な対訳フレーズ対）が抽出できる。

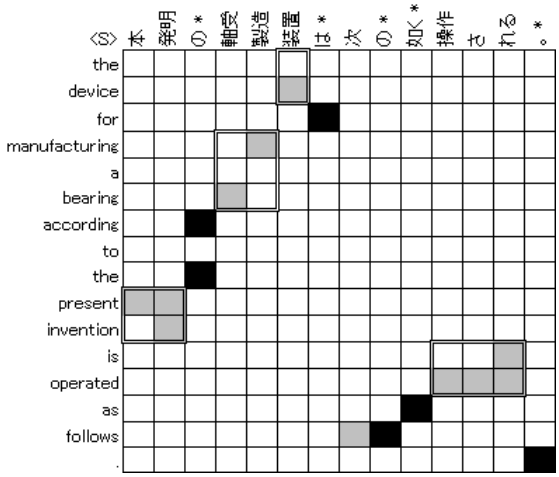


図 4: $\mathcal{R}$(黒マス)、$\mathcal{O}$(灰色マス)、$\mathcal{M}$(二重線枠) の例

## 3 ベースルールセット

上記で得られる広範囲フレーズ並び替えルール（以下 S と略記）を提案モデルのコアルールセットとする。

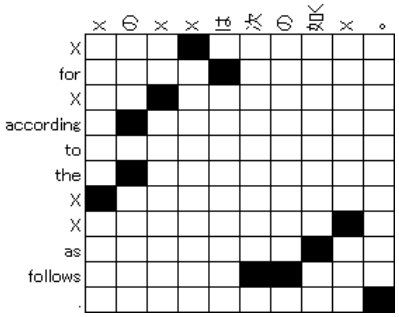


図 5: 単語対応テーブルの収縮

ただし、局所的な並び替えルールも必要なので、単純なフレーズ並び替えルールセット（以下 R と略記）と Chiang の方法で抽出した階層フレーズから特定のパターンのルールセット（以下 G と略記）をベースルールセットとして用いる。

R は、以下の形のルールで構成される。

- $X \rightarrow <$ ア , $A >$
- $X \rightarrow <$ ア $X_1$ , $A\ X_1 >$
- $X \rightarrow <$ ア $X_1$ , $X_1\ A >$
- $X \rightarrow <$ $X_1$ ア , $A\ X_1 >$
- $X \rightarrow <$ $X_1$ ア , $X_1\ A >$

G は、以下の形のルールで構成される。

- $X \rightarrow <$ ア $X_1$ イ , $A\ X_1 >$
- $X \rightarrow <$ ア $X_1$ イ , $X_1\ A >$
- $X \rightarrow <$ ア $X_1$ , $A\ X_1\ B >$
- $X \rightarrow <$ $X_1$ ア , $A\ X_1\ B >$
- $X \rightarrow <$ ア $X_1$ イ , $A\ X_1\ B >$

ここで「ア」と「イ」は原言語側のフレーズ、「A」と「B」は目的言語側のフレーズを意味している。

## 4 実験結果

表 1 は従来の階層フレーズモデル（$Hiero$）と各ルールセットで構成する提案モデルの実験結果である。トレーニングデータは NTCIR7 対訳コーパスからランダムに抽出された 10 万文対であり、NTCIR7 の dev と fmlrun をそれぞれデベロップメントとテストセットとして用いた。デコーディングの際、階層ルールの解析スパンの上限を 20 又は 30 と設定した。

翻訳精度において、$S$+$R$ はほぼ従来の階層フレーズモデルと同じ精度に達している[1]。$S$+$G$+$R$ はさらに向上し、従来の階層フレーズモデルの性能を上回っている[2]ことが分かる。

[1]統計的有意差なし。

表 1: 従来の階層フレーズモデル及び各提案モデルのテストセットでの BLEU 値

| $span$ | $Hiero$ | $R$ | $G$+$R$ | $S$+$R$ | $S$+$G$+$R$ |
|---|---|---|---|---|---|
| 20 | 28.61 | 27.36 | 28.05 | 28.47 | **28.79** |
| 30 | 28.49 | 27.37 | 28.08 | 28.22 | **28.91** |

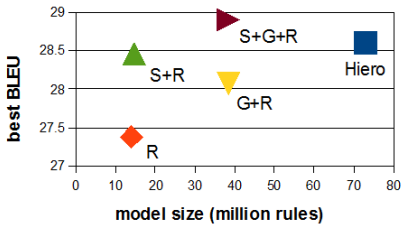


図 6: 翻訳精度とモデルサイズ

図 6 の横軸はルール数（百万）であり、縦軸は各モデルの最高の BLEU 値である。図 6 から、$S$+$R$ は従来の階層フレーズの約五分の一のサイズであり、$S$+$G$+$R$ はその約半分のサイズであることが分かる。$S$ は翻訳精度を効率よく改善できることが分かる。

表 2 は翻訳例であり、図 7 と図 8 はそれぞれ従来の階層フレーズモデルと $S$+$G$+$R$ の導出木である。この例では、広範囲フレーズ並び替えルールにより、3 個以上のフレーズの移動を制御でき、構文構造として正しい出力文が得られていることが分かる。

## 5 終わりに

本研究は日英翻訳のような広範囲語順調整が必要とされる翻訳タスクに着目し、階層的広範囲フレーズ並び替えルールを提案した。これらのルールは機能語で語彙化され、非終端記号数やイニシャルフレーズの長さの制限がない。長いイニシャルフレーズからも抽出できるため、複数のフレーズ（3 つ以上）に対して同時に比較的長距離の移動を可能とする。NTCIR-7 対訳コーパス上の日英翻訳実験結果により、提案手法は、従来の階層フレーズモデルの五分の一のサイズでほぼ同じ性能、半分のサイズで性能がやや上回ることを確認した。今後の課題は NTCIR8 や NTCIR9 などの大きな対訳コーパスで実験を行うことである。

[2]統計的有意差は $span = 30$ のときのみあり。

表 2: 翻訳例

| | |
|---|---|
| *Source* | また 、 同時に 強 誘電 体 キャパシタ １ ２ から BL ２ へ 電荷 が 移動する 。 |
| *Reference* | at the same time , electric charges transfer from the ferroelectric capacitor 12 to bl2 . |
| *Hiero* | at the same time , bl2 to charge is moved from the ferroelectric capacitor 12 . |
| *S+G+R* | at the same time , the charge moves from the ferroelectric capacitor 12 to bl2 . |

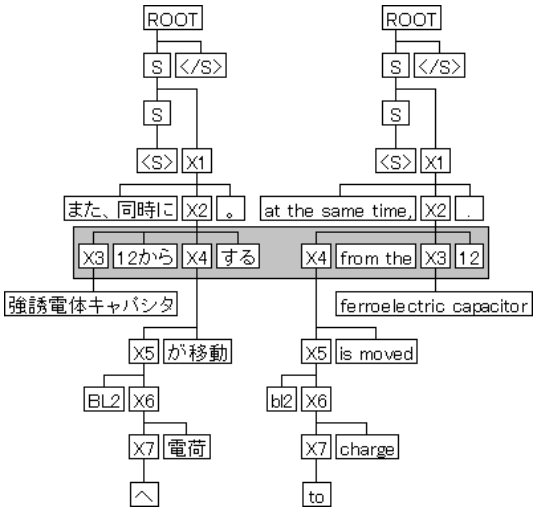


図 7: 従来の階層フレーズモデルの導出木

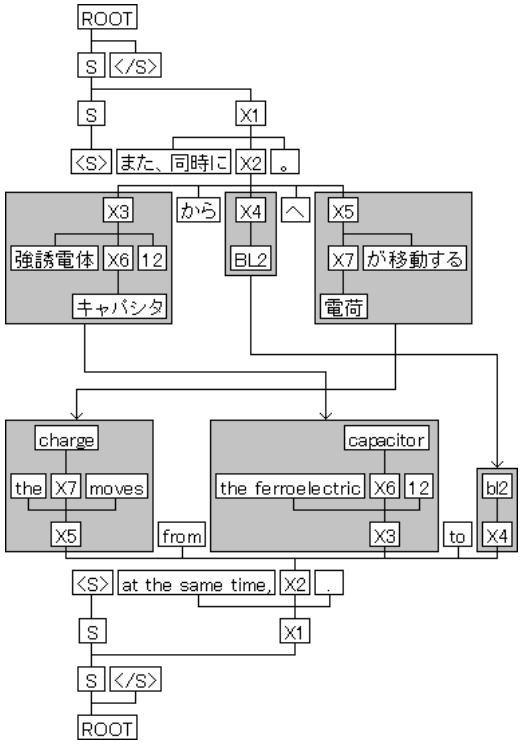


図 8: *S+G+R* の導出木

## 参考文献


[Brown et al., 1993] Brown, P. F., Della Pietra, V. J., Della Pietra, S. A., and Mercer, R. L., The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, Vol.19, No.2, pp. 263-311, 1993.

[Och and Ney, 2003] Och, F. J., and Ney, H., A systematic comparison of various statistical alignment models *Computational Linguistics*, Vol.29, No.1, pp. 19-51, 2003.

[Koehn et al., 2003] Koehn, P., Och, F. J., and Marcu, D., Statistical phrase based translation. In *Proceedings of HLT-NAACL*, pp. 48-54, Edmonton, May-June, 2003.

[Tillmann, 2004] Tillmann, C., A unigram orientation model for statistical machine translation. In *Proceedings of HLT-NAACL*, 2004.

[Galley and Manning, 2008] Galley, M., and Manning, C. D., A simple and effective hierarchical phrase reordering model. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 848-856, Honolulu, October, 2008.

[Chiang, 2007] Chiang, D., Hierarchical phrase-based translation. *Computational Linguistics*, Vol.33, No.2, pp. 201-228, 2007.

[Papineni et al., 2002] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J., BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 311-318, Philadelphia, July, 2002.

[Och and Ney, 2002] Och, F. J., and Ney, H., Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 295-302, Philadelphia, July, 2002.

[Koehn, 2011] Koehn, P., Moses statistical machine translation system user manual and code guide. September, 2011.