# 統計的機械翻訳におけるフレーズ対応最適化を用いた翻訳候補のリランキング

システム情報工学研究科 2 年 200820634 越川 満

指導教員 山本 幹雄

2008 年 5 月 14 日

## 1 はじめに

インターネットの普及以来、ウェブ上に存在する情報は増加の一途を辿っている。ウェブ上の情報を最大限に活用するには、様々な言語で書かれたテキスト情報を翻訳する必要があるが、人手によりすべてのウェブページのテキストを翻訳することは非現実的である。そこで、これを翻訳する手段の一つとして機械翻訳が需要されている。

機械翻訳手法のひとつである統計的機械翻訳では、対訳コーパスと呼ばれる同一の意味をもつ異なる言語の文対の集合から確率的翻訳規則を自動学習し、確率的翻訳規則をもとに各翻訳候補を順位付け、最も確率の高い候補を出力する。現在、統計翻訳で最も広く用いられているフレーズベース翻訳手法は、フレーズと呼ばれる連続する 1 単語以上の単語列を翻訳の最小単位として扱う [Koehn et.al., 2003]。そのため、フレーズ内部に収まるような局所的な語順並び替えがモデル化可能であり、近年の著しい性能向上に大きく寄与している。しかし、入力文に対するフレーズ区切りや翻訳前後のフレーズ同士の対応関係の組み合わせ数は膨大であり、探索空間を狭めるために様々な近似やヒューリスティックス用いざるを得ない。そのため、統計翻訳システムは一般に確率最大の翻訳結果を出力するとは言えない。

一方で翻訳結果として採択されなかった第 2 位以下の候補には、より翻訳精度の高いものが含まれることが知られており、これらの候補に対して確率を厳密に最大化するフレーズ区切り・対応を求め直すことで翻訳候補の順位付けを改善し、翻訳精度を向上させることができると思われる。対訳文に対して確率最大すなわち最適なフレーズ対応 (phrase alignment) を求める手法としては、整数計画法を用いた手法 [DeNero and Klein, 2008] が提案されている。しかし、この手法では現在広く用いられている語順並び替え (リオーダリング) モデル [Koehn et.al., 2003][Tillmann and Zhang, 2005] が考慮されていないという問題点がある。そこで本稿では、リオーダリングモデルを組込み可能な新たなフレーズアライナ (phrase aligner) の定式化を提案し、これを統計的機械翻訳システムの翻訳候補のリランキングに応用することで翻訳精度が向上することを実験的に示す。

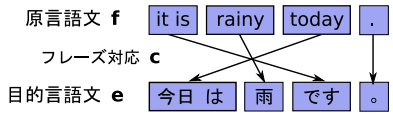


図 1 フレーズベース翻訳

## 2 フレーズアライナ

### 2.1 フレーズベース統計的機械翻訳

統計的機械翻訳システムは、入力された原言語文 $\boldsymbol{f}$ に対する翻訳候補としてあらゆる目的言語文 $\boldsymbol{e}$ を考慮し、最も確率の高い目的言語文 $\hat{\boldsymbol{e}}$ を翻訳結果として出力する。ここで、原言語および目的言語はそれぞれ翻訳元の言語、翻訳先の言語を意味する。すなわち、統計翻訳では以下の式に基づいて翻訳候補 $\hat{\boldsymbol{e}}$ が決定される [Koehn et.al., 2003]。

$$\begin{aligned}\hat{\boldsymbol{e}} &= \arg\max_{\boldsymbol{e}} \sum_{\boldsymbol{c}} P(\boldsymbol{f},\boldsymbol{c}|\boldsymbol{e})P(\boldsymbol{e}) \\ &\approx \arg_{\boldsymbol{e}} \max_{\boldsymbol{e}} \max_{\boldsymbol{c}} P(\boldsymbol{f},\boldsymbol{c}|\boldsymbol{e})P(\boldsymbol{e}) \\ &\approx \arg_{\boldsymbol{e}} \max'_{\boldsymbol{e},\boldsymbol{c}} P(\boldsymbol{f},\boldsymbol{c}|\boldsymbol{e})P(\boldsymbol{e}) \qquad (1)\end{aligned}$$

式 (1) において、$P(\boldsymbol{e})$ は言語モデルと呼ばれ、$\boldsymbol{e}$ の目的言語文らしさを表す。また、$\boldsymbol{c}$ は $\boldsymbol{f}$、$\boldsymbol{e}$ 間での対訳関係を表す。$\arg_{\boldsymbol{e}} \max'_{\boldsymbol{e},\boldsymbol{c}}$ はデコーダと呼ばれ、与えられた $\boldsymbol{f}$ の翻訳として最も確率の高い $\boldsymbol{e}$ をヒューリスティック探索により近似探索する。ここで $\max'$ は $\max$ の近似解を表す。

フレーズベース手法では、翻訳の最小単位として連続した 1 単語以上の単語列であるフレーズを用いる [Koehn et.al., 2003]。フレーズベース翻訳の例を図 1 に示す。図 1 において枠で囲われた単語ないしは単語列がフレーズである。フレーズベース翻訳では以下の近似をおく。

$$P(\boldsymbol{f},\boldsymbol{c}|\boldsymbol{e}) \approx \left(\prod_{i=1}^{I} P(\bar{f}_i|\bar{e}_{c_i})\right) \cdot P(\boldsymbol{c}|\boldsymbol{e}) \qquad (2)$$

ここで、$\bar{f}$、$\bar{e}$ はそれぞれ原言語フレーズ、目的言語フレーズを、$\boldsymbol{c} = c_1, c_2, \cdots, c_I$ は原言語–目的言語間でのフレーズ対応関係を表す。$c_i$ は原言語側で $i$ 番目のフレーズ $\bar{f}_i$ が対応する目的言語側フレーズの番号である。すなわち $\bar{f}_i$ は目的言語側で $c_i$ 番目のフレーズ $\bar{e}_{c_i}$ に対応する。$P(\bar{f}_i|\bar{e}_{c_i})$

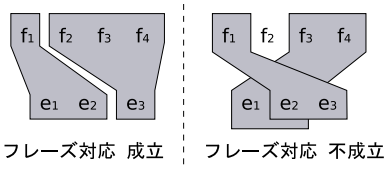


図 2　フレーズ対応の成立と不成立

はフレーズ翻訳確率、$P(\boldsymbol{c}|\boldsymbol{e})$ はリオーダリングモデルとそれぞれ呼ばれる。リオーダリングモデルは翻訳前後での語順変化に対する確率を与えるモデルであり、目的言語側で隣り合うフレーズ同士の原言語側における位置関係を確率モデル化したものが、現在最も広く用いられている [Koehn et.al., 2003][Tillmann and Zhang, 2005]。

### 2.2　フレーズ対応問題

対訳関係にあるフレーズのペア (フレーズ対) の集合および対訳文が与えられたとき、対訳文の各単語を一度ずつ被覆するフレーズ対の組合せをその対訳文に対するフレーズ対応という [DeNero and Klein, 2008]。フレーズ対応が成立している場合と成立しない場合の例を図 2 に示す。図中、$f_1 \sim f_4$ と $e_1 \sim e_3$ とは、各々原言語文、目的言語文を構成する単語列を表し、対訳文をなしている。また色のついた枠はフレーズ対を表す。図 2 の左側の例はフレーズ対応の定義を満たしているのに対し、右側の例は $f_2$ が被覆されておらず、$e_2$ が二重に被覆されているためフレーズ対応は成立していない。

フレーズ対応問題とは、対訳文 $\langle \boldsymbol{f}, \boldsymbol{e} \rangle$、およびフレーズ対とその翻訳確率 $\left\langle \bar{f}, \bar{e}, P(\bar{f}|\bar{e}) \right\rangle$ の集合であるフレーズテーブルが与えられたとき、確率最大すなわち最適なフレーズ対応を求める問題であり、これを実現するシステムをフレーズアライナと呼ぶ [DeNero and Klein, 2008]。フレーズアライナは以下の式で定義される。

$$\left\langle \hat{\bar{f}}_1^I, \hat{\bar{e}}_1^I, \hat{\boldsymbol{c}} \right\rangle = \arg \max_{\bar{e}_1^I = \boldsymbol{e}, \bar{f}_1^I = \boldsymbol{f}, \boldsymbol{c} = c_1^I} P\left(\bar{f}_1^I | \bar{e}_1^I, \boldsymbol{c}\right) P\left(\bar{e}_1^I, \boldsymbol{c} | \boldsymbol{e}\right) \tag{3}$$

ここで、$c_1^I = c_1, c_2, \cdots, c_I$ であり $\bar{f}_1^I$, $\bar{e}_1^I$ についても同様に定義される。式 (3) の右辺第一項および第二項は、それぞれ式 (2) のものと対応している。

### 2.3　整数計画問題としての定式化

フレーズテーブル中のフレーズ対 $\left\langle \bar{f}_k, \bar{e}_{c_k} \right\rangle$ に対してその使用の有無を表す 2 値変数 $x_k \in \{0, 1\}$、および各フレーズ対が原言語文で被覆する単語位置を 1、それ以外を 0 で表す 2 値行列 $F$ を導入する [DeNero and Klein, 2008]。なお、$F$ は行が原言語文に含まれる単語で、列が各フレーズ対でインデクシングされている。目的言語側についても $F$ と同様に行列 $E$ を定義する。

対訳文 $\langle \boldsymbol{f} = f_1, f_2, f_3, f_4,\ \boldsymbol{e} = e_1, e_2, e_3 \rangle$ に対して図 3 のようなフレーズ対が適用可能な場合を考える。図 3 において四角い枠がフレーズを表し、線で結ばれたフレーズ同士がフレーズ対である。このとき、$F, E$ は式 (4) のように

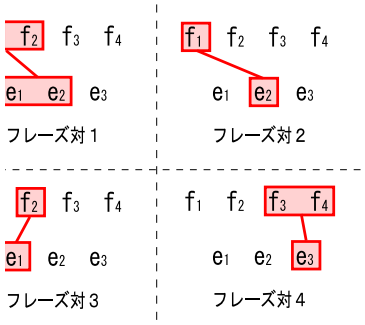


図 3　フレーズ対集合

なる。例えば、フレーズ対 1 は原言語側で $f_1, f_2$ を被覆するため、行列 $F$ の 1 行 1 列および 2 行 1 列成分が 1 となっている。

$$F = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}, E = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{4}$$

簡単化のため、式 (3) の右辺第二項すなわちリオーダリング確率を 1 と置き、フレーズ翻訳確率のみを考慮すると仮定すれば、フレーズ対応問題は次のように表される [DeNero and Klein, 2008]。

$$\begin{array}{ll} \text{maximize} & \displaystyle\sum_{k \in K} x_k \log p_k \\ \\ \text{subject to} & F\boldsymbol{x} = \mathbf{1}, \\ & E\boldsymbol{x} = \mathbf{1}, \\ & x_k \in \{0, 1\} \quad (\forall k \in K)\,. \end{array} \tag{5}$$

ここで $p_k$ は $P(\bar{f}_k|\bar{e}_{c_k})$ の略記であり、$K$ は対訳文に適用可能なフレーズ対の集合を表す。また $\mathbf{1} = (1 \cdots 1)^T$ である。

## 3　提案手法

### 3.1　フレーズアライナによる翻訳候補のリランキング

本研究では、フレーズアライナを利用した翻訳候補のリランキング (再順位付け) 法を提案する。リランキングの手順としては、まずフレーズベースデコーダを用いて翻訳候補上位 $n$ 個を求め、次に、各翻訳候補と原言語文のペアを対訳文とみなし、フレーズアライナにより最適なフレーズ対応を付与する。そして、最適化されたフレーズ対応をもとに各翻訳候補をリランキングし、最も確率の高い候補を翻訳システムの出力とする。以上を図示すると、図 4 のようになる。

### 3.2　有向グラフのパスとしてのフレーズ対応

[DeNero and Klein, 2008] の定式化では各フレーズ対に対して変数を置くため、フレーズ対同士の位置関係を表すリオーダリングモデルを一次式として目的関数に組込むのは難しい。一般に非線形問題は、線形問題に比べて計算コストが非常に高くなるという問題がある。そこで本節で

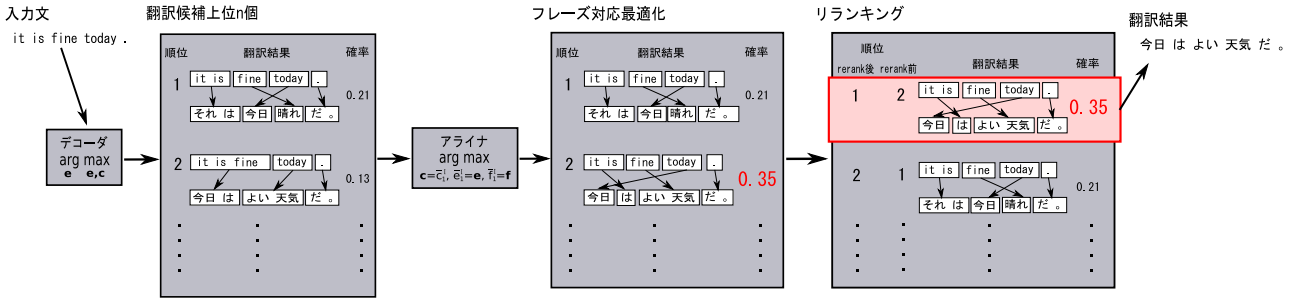


図 4　フレーズアライナによる翻訳候補のリランキングの流れ

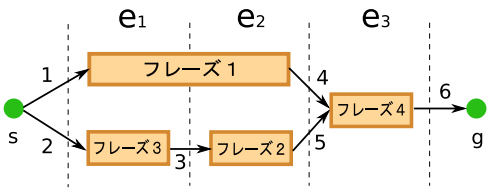


図 5　有向グラフ上に表したフレーズ同士の関係

は、上記の問題を解決するフレーズ対応問題の新たな定式化を提案する。

本稿で提案する定式化では、対訳文に適用可能なフレーズ対を目的言語側について有向グラフ上に表すことを考える。図 3 の各フレーズ対を有向グラフ上に表すと図 5 のようになる。図 5 において $e_1$〜$e_3$ は目的言語文中の単語である。フレーズは四角い枠で表されており、各フレーズの位置と大きさは被覆する単語位置に対応する。またフレーズの番号は図 3 のフレーズ対の番号に対応している。フレーズ同士を結ぶ有向枝に割り振られた数字は枝番号 $a$ を表す。フレーズ対応は、原言語側について式 (5) の原言語側制約式 $F\boldsymbol{x} = \mathbf{1}$ を満たし、かつ目的言語側グラフにおいて開始ノード $s$ から終端ノード $g$ へのパスとなっている必要がある。例えば図 5 中、フレーズ 2, 3, 4 を通るパスは対訳文に対するフレーズ対応候補の一つである。フレーズ対応に含まれるフレーズ対は、目的言語側グラフのパスに含まれる枝からも導くことが出来る。このモデル化のもとでは、リオーダリング確率を目的言語側グラフの枝に対する重みとして組込むことができる。

[DeNero and Klein, 2008] の定式化で用いられている記号に加えて、目的言語側有向グラフ中の枝 $a$ に対し、$a$ がパスに含まれる場合は 1、そうでない場合は 0 をとる仮変数 $y_a$ を新たに導入する。このとき、フレーズ対応問題は以下のように定式化される。

$$
\begin{aligned}
\text{maximize} \quad & \sum_{k \in K} x_k \log p_k + \sum_{a \in A} y_a \log d_a \\
\text{subject to} \quad & F\boldsymbol{x} = \mathbf{1}, \\
& M\boldsymbol{y} = \boldsymbol{b}, \\
& N\boldsymbol{y} = \boldsymbol{x}, \\
& x_k \in \{0,1\} \quad (\forall k \in K), \\
& y_a \in \{0,1\} \quad (\forall a \in A).
\end{aligned}
\tag{6}
$$

表 1　コーパス詳細

| データセット | 文数 | 単語数 | 語彙サイズ |
|---|---|---|---|
| 学習セット (英) | 1,798,571 | 59,974,173 | 142,435 |
| 学習セット (日) | 1,798,571 | 64,184,179 | 121,652 |
| dev. セット (英) | 915 | 30,028 | 3,986 |
| dev. セット (日) | 915 | 32,427 | 3,653 |
| テストセット (英) | 1,381 | 45,334 | 4,116 |
| テストセット (日) | 1,381 | 48,737 | 3,882 |

ここで $M\boldsymbol{y} = \boldsymbol{b}$ は目的言語側グラフでノード $s$ からノード $g$ へのパスとなるための条件であり流量保存則と呼ばれる。$N$ は目的言語側パスに含まれる枝と各フレーズ対との関係を与える行列である。また $A$ はすべての枝の集合を、$d_a$ は目的言語側の各枝に割り当てられるリオーダリング確率を表す。

図 5 について、流量保存則を書き下すと式 (7) のようになる。左辺第一項は行列 $M$ であり、各行は $s$、フレーズ 1、... 、フレーズ 4、$g$ のノードに対応する。例えば式 (7) の 5 行目は、フレーズ対 4 のノードについての流量保存則に対応し、$y_4 + y_5 = y_6$ を表す。

$$
\begin{pmatrix}
-1 & -1 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & -1 & 0 & 0 \\
0 & 1 & -1 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & -1 & 0 \\
0 & 0 & 0 & 1 & 1 & -1 \\
0 & 0 & 0 & 0 & 0 & 1
\end{pmatrix}
\begin{pmatrix}
y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6
\end{pmatrix}
=
\begin{pmatrix}
-1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1
\end{pmatrix}
\tag{7}
$$

## 4　評価実験

### 4.1　実験条件

実験には、NTCIR-7 特許翻訳タスク [Fujii et.al., 2008] で配布された英日対訳コーパスを用いた。フレーズテーブルの学習には同学習セットを、デコーダの各素性に対する重み学習には development セットを使用した。テストセットは NTCIR-7 フォーマルランで配布されたテストセットとした。コーパスの詳細を表 1 に示す。翻訳方向は日英とし、ベースラインにはオープンソースのフレーズベース統

表 2　評価実験条件

| 項目 | 条件 |
|---|---|
| デコーダ | Moses |
| ビーム幅 (翻訳候補数) | 10, 20, 50, 100, 200, 500, 1,000 |
| ttable-limit | 20 |
| 数理計画問題の Solver | CPLEX 11.0 |

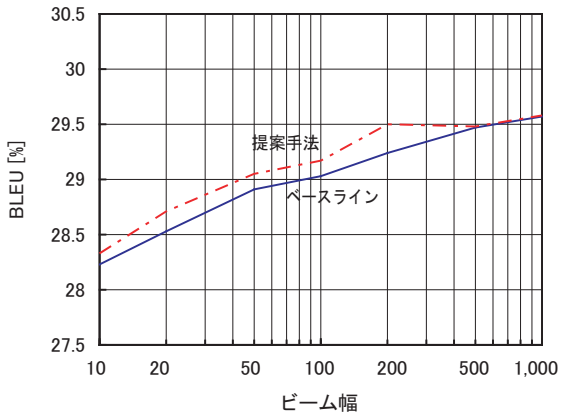


図 6　提案手法とベースラインのビーム幅に対する翻訳精度

計的機械翻訳システムである Moses デコーダ[*1]を用いた。デコーダパラメータ等の実験条件を表 2 に示す。表 2 において、ttable-limit は原言語フレーズひとつ当たりの翻訳先の目的言語フレーズ数である。ビーム幅はベースライン翻訳システム Moses の翻訳精度に対するパラメータであり、ビーム幅が大きいほどデコーダの探索精度は高くなる。また、デコーダが出力する翻訳候補数 $n$ はビーム幅と同値とした。翻訳精度の評価には BLEU[Papineni et.al.,2001] を用いた。BLEU は正解翻訳例との一致率から計算される評価指標であり、100% に近いほどシステムの翻訳精度が高いことを表す。

### 4.2　実験結果

ビーム幅を変化させたときのベースラインシステム Moses および提案手法による翻訳結果に対する BLEU を図 6 に示す。提案手法とベースラインとを比較すると提案手法は、常にベースラインより BLEU 値が高くなっている。有意水準 5% として有意差検定を行ったところ、ビーム幅 200 以下の全測定点において提案手法が Moses を有意に上回っているという結果が得られた。しかし、ビーム幅が 500〜1,000 のとき提案手法による改善はほとんど見られなかった。これはビーム幅を大きくとることでベースラインシステム Moses の探索精度が向上し、提案手法による改善の余地が小さくなってしまうためであると考えられる。

[*1] http://www.statmt.org/moses/

## 5　おわりに

本稿では、リオーダリングモデルを考慮したフレーズ対応問題の新しい定式化を提案した。また、それを用いてフレーズベース翻訳システムの翻訳候補に対しフレーズ対応最適化を行い、$P(\boldsymbol{f}, \boldsymbol{c}|\boldsymbol{e})$ を厳密に最大化することによってリランキングを行う手法について検討した。提案手法は NTCIR-7 特許翻訳タスクのデータセットを用いた評価実験において、常に Moses が翻訳候補に与えたスコアを改善し、BLEU を向上させることが確認できた。しかし、提案手法によるスコア改善はそれほど大きくなく、従来から用いられているヒューリスティック探索であっても、フレーズ対応についてはかなり高い精度の解が得られていると言える。今後は、翻訳時に最適化手法を取り入れる方法について検討していきたい。

## 参考文献

[Koehn et.al., 2003] P.Koehn, F.J.Och and D.Marcu. 2003. “Statistical phrase-based translation.” In *NAACL ’03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pp.48–54.

[DeNero and Klein, 2008] J.DeNero and D.Klein. 2008. “The Complexity of Phrase Alignment Problems.” In *Proceedings of ACL-08: HLT, Short Papers*, pp.25–28.

[Fujii et.al., 2008] A.Fujii, M.Utiyama, M.Yamamoto, and T.Utsuro. 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 Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access*, pp.389–400.

[Tillmann and Zhang, 2005] Christoph Tillmann and Tong Zhang, 2005. “A Localized Prediction Model for Statistical Machine Translation.” In *ACL ’03: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pp.557–564.

[Papineni et.al.,2001] Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu. 2001. “BLEU: a method for automatic evaluation of machine translation.” In *ACL ’02: Proceedings of the 40th Anuual Meeting on Association for Computational Linguistics*, pp.311-318.