整数計画法を用いたフレーズ対応最適化による翻訳システムの改良

システム情報工学研究科 1 年 200820634 越川 満
指導教員 山本 幹雄
2008 年 10 月 2 日

1 はじめに

現在ウェブ上には、多様な言語で表現された膨大な量のテキスト情報が存在し、これを翻訳する手段の一つとして機械翻訳への需要が高まっている。機械翻訳システムを実現する方法は様々であるが、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翻訳規則を自動学習することにより翻訳システムを構築する統計的手法 [1] [2] は近年の著しい性能向上から注目されている。しかし、その翻訳精度は人手により翻訳規則を記述するルールベース手法には未だ及ばず、更なる翻訳精度の改善が必要とされる。

統計的機械翻訳が大幅に翻訳精度を向上した背景には、フレーズ翻訳モデル [4] [5] の導入がある。フレーズベース翻訳システムでは、フレーズと呼ばれる連続する単語列を翻訳の最小単位として扱う。これにより局所的な語順変化や複数単語の翻訳がモデル化できるようになったが、その反面入力文のフレーズ区切りや翻訳前後での対応関係を考慮しなければならないため探索空間が非常に広くなってしまう。探索は入力文に対する翻訳としての確率により各候補を順位付けて行われるが、探索空間を狭めるための様々な近似やヒューリスティックスを用いるため一般に統計的機械翻訳システムは確率最大の翻訳結果を出力するとは言えない。

一方で、翻訳システムの出力として選ばれなかった翻訳候補中には、より翻訳精度を向上させるものが含まれていることが知られている [6]。これらの候補の中には翻訳システムが探索に近似を用いているために、確率が不当に低く計算されているものが含まれる。そこでより適切なフレーズ区切り・対応を適用する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、与えられた対訳フレーズ集合の中で最適なフレーズの対応付け (phrase alignment) を行う手法の一つとして、整数計画法を用いた手法 [7] が提案されている。この手法を用いてフレーズ区切り・対応を改善することが考えられるが、この手法では現在のフレーズベース翻訳システムで広く用いられている語順変化を表す歪みモデル [8] を取り入れることが出来ない。そこで本研究では歪みモデルをフレーズアライナ (phrase aligner) に適用するため、フレーズアライナの新しい定式化を提案し、これを用いて統計的機械翻訳システムの翻訳候補を再順位付け (reranking) することで翻訳精度の向上を目指す。

2 統計的機械翻訳

2.1 統計的機械翻訳

機械翻訳システムは、翻訳元の言語で表現された文 $\boldsymbol{f}$ を、同一の意味をもつ翻訳先言語の文 $\boldsymbol{e}$ へと変換することを目的とする。ここで翻訳元の言語を原言語、翻訳先の言語を目的言語という。統計的機械翻訳システムは、与えられた入力文に対する翻訳候補としてあらゆる目的言語文を考慮し、最も確率の高い目的言語文を翻訳結果として出力する。すなわち、以下の式に基づいて翻訳候補 $\hat{\boldsymbol{e}}$ は決定される [3]。

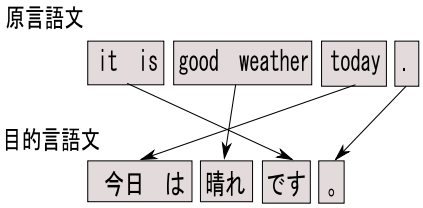

図 1 フレーズベース翻訳

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

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指標である。また $\arg_{\boldsymbol{e}} \max_{\boldsymbol{c},\boldsymbol{e}}$ はデコーダと呼ばれ、与えられた $\boldsymbol{f}$ に対し、その翻訳として最も確率の高い $\boldsymbol{e}$ を探索する。$\boldsymbol{c}$ は $\boldsymbol{f}$、$\boldsymbol{e}$ 間での訳語の対応関係を表す。

2.2 フレーズ翻訳モデル

フレーズ翻訳モデル [4][5] では、翻訳の最小単位としてフレーズを用いる。統計的機械翻訳におけるフレーズとは連続した 1 単語以上の単語列を指す。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。フレーズベースシステムでは以下の近似をおく。

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

ここで、$\bar{f}$、$\bar{e}$ は各々原言語フレーズ、目的言語フレーズを、$\boldsymbol{c} = c_1, c_2, \cdots, c_I$ はフレーズ区切りおよび原言語－目的言語間でのフレーズ同士の対応関係を表す。$P(\bar{f}_i|\bar{e}_{c_i},c_i)$ はフレーズ翻訳確率であり、対訳コーパス中での原言語フレーズおよび目的言語フレーズの共起確率から推定される [4]。$P(\bar{e}_{c_i},c_i|\boldsymbol{e})$ は統計的機械翻訳における歪みモデルであり、翻訳前後での語順変化に対する確率を与える。歪みモデルとしては、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広く用いられている [4][8]。

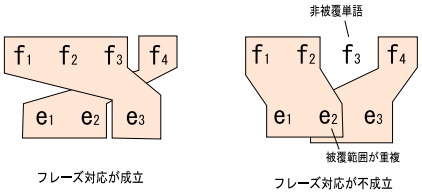

図 2　フレーズ対応の成立と不成立

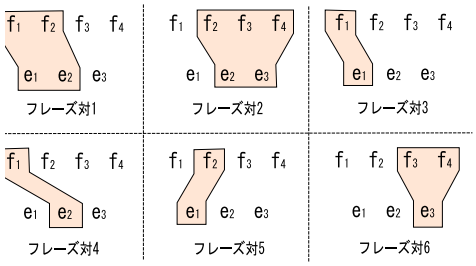

図 3　フレーズ対

3　フレーズアライナ

本節では、本研究で提案する手法の中核をなすフレーズ対応の改善と密接に関連する DeNero らの研究 [7] について説明する。この手法では整数計画法を用いて対訳文へのフレーズ対応付けを定式化している。

3.1　フレーズ対応取得問題

対訳文 $(\boldsymbol{f}, \boldsymbol{e})$ およびフレーズ対集合が与えられたとき、対訳文の各単語を一度ずつ被覆するようなフレーズ対の組み合わせが存在するとき、そのフレーズ対集合および適用位置を対訳文に対するフレーズ対応という [7]。フレーズ対応が成立している場合とそうでない場合の例を図 2 に示す。図中、f_i, e_i はそれぞれ原言語単語、目的言語単語を表す。図 2 左側はフレーズ対応が成立しているが、図 2 右側は重複して被覆されている単語や被覆されていない単語が見られるためフレーズ対応は成立していない。一般に、ある対訳文に対するフレーズ対応は、フレーズ区切り・対応付けの多様性から一意には定まらず複数の候補が存在する。フレーズ対応取得問題とは、対訳文およびフレーズテーブルと呼ばれるフレーズ対とその翻訳確率 $\langle \bar{f}, \bar{e}, P(\bar{f}|\bar{e}) \rangle$ の集合が与えられたとき、フレーズ対応候補の中で最も確率の高いすなわち最適なフレーズ区切り・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 [7]。フレーズアライナは以下の式で定義される [12]。

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

ここで、$c_1^I = c_1, c_2, \cdots, c_I$ であり $\bar{f}_1^I$, $\bar{e}_1^I$ についても同様である。

3.2　定式化

DeNero らの手法ではフレーズアライナを以下のように定式化する。なお、簡単化のため対訳文に対する最適フレーズ区切り・対応対応を行う際、フレーズ翻訳確率のみを考慮する [7]。フレーズテーブル中の各翻訳対 $\langle \bar{f}_k, e_{c_k}^- \rangle$ 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、目的言語文中で被覆する単語位置を 1 で表す 2 値行列 F および E を導入する。例えば対訳文 $\langle \boldsymbol{f} = f_1, f_2, f_3, f_4,\ \boldsymbol{e} = e_1, e_2, e_3 \rangle$ に対して図 3 のようなフレーズ対が適用可能なとき、F, E は次のようになる。

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

これらの被覆行列は行が各言語の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。このときフレーズ対応取得問題は以下のように定式化される [7]。

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

ここで p_k は $P(\bar{f}_k | e_{c_k}^-, c_k)$ 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

4　提案手法

本研究ではフレーズアライナを利用して、デコーダにより求められた翻訳候補を reranking する手法を提案する。また reranking の際に翻訳確率だけでなく現在フレーズベース翻訳システムで広く用いられている歪み確率を考慮したフレーズ対応最適化が可能となるよう、フレーズアライナの新たな定式化を提案する。

4.1　reranking の流れ

本節ではフレーズベース翻訳システムにより得られた翻訳候補のフレーズ対応を改善し、その reranking を行うまでの流れを説明する。まず、フレーズベースデコーダを用いて翻訳候補上位 N 個を求める。次に、各翻訳候補と原言語文のペアに対し、フレーズアライナを用いてフレーズ対応を改善する。最後に、最適化されたフレーズ対応をもとに各翻訳候補を再順位付けし、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。

4.2　フレーズアライナの新しい定式化

DeNero らのフレーズアライナの定式化では、各フレーズの使用有無を表す 0-1 変数を用いているため、フレーズ対同士の位置関係を表す歪みモデルを一次式として取り入れるのは難しい。そこで本研究では、歪みモデルを自然に組込むことができる、フレーズアライナの新たな定式化を提案する。

まず対訳文に対して適用可能なフレーズ対を有向グラフ上に表すことを考える。3.1 節で挙げた例を有向グラフとして表したものを図 5 に示す。図 5 において、各フレーズは枠で表されており、各フレーズの位置・大きさは被覆する単語位置に対応する。また原言語側、目的言語側で同じフレーズ番号を持つものはフレーズ対をなしている。対訳文のフレーズ対応は原言語側、目的言語側ともに開始ノード s から終端

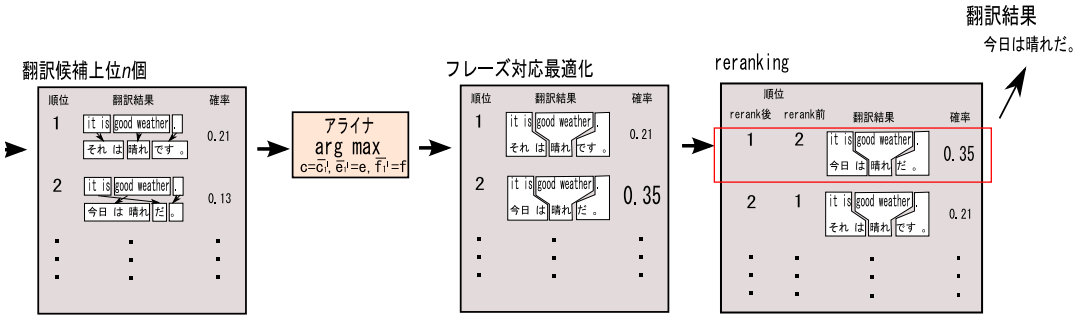

図 4　フレーズアライナを用いた翻訳候補の reranking の流れ

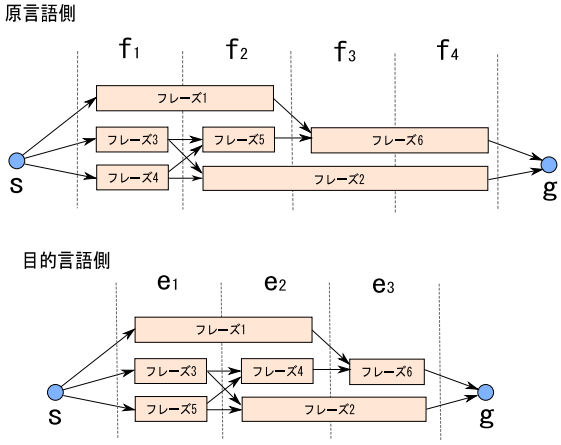

図 5　有向グラフ上に表したフレーズ同士の関係

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 4, 5, 6 を通るパスはこの対訳文のフレーズ対応の候補の一つである。フレーズ対応をなすフレーズ対集合はパスをなす枝の集合から導くことが出来る。このモデル化のもとでは、歪みコストを目的言語側の枝に対する重みとして取り入れることができる。

以上のことからフレーズアライナを定式化すると次のようになる。原言語側有向グラフ中の枝 e に対し、e 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_e を導入する。また同様に目的言語側に対しても仮変数 z_e を導入する。このとき、フレーズアライナは以下のように定式化される [13]。

$$
\begin{aligned}
&maximize && \sum_{k \in K} x_k \log p_k + \sum_{e \in E} z_e \log d_e \\
&subject\ to && M\boldsymbol{y} = \boldsymbol{b} \\
& && N\boldsymbol{y} = \boldsymbol{x} \\
& && M'\boldsymbol{z} = \boldsymbol{b}' \\
& && N'\boldsymbol{z} = \boldsymbol{x} \\
& && x_k \in \{0,1\} \quad (\forall k \in K) \\
& && y_e \in \{0,1\}, z_e \in \{0,1\} \quad (\forall e \in E)
\end{aligned}
\tag{6}
$$

ここで $M\boldsymbol{y} = \boldsymbol{b}, M'\boldsymbol{z} = \boldsymbol{b}'$ は原言語・目的言語側でノード s からノード g へのパスとなっているための条件であり流量保存則と呼ばれる。N, N' は原言語側、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 E はすべての枝の集合を、d_e は目的言語側の各枝に割り当てられる歪み確率を表す。

表 1　NTCIR-7 対訳コーパス詳細

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)	899	29,674	3,867
テストセット (日)	899	31,848	3,696

表 2　評価実験条件

項目	条件
デコーダ	Moses (08/02/20 release)
ビーム幅	200
beam-threshold	1e-5
ttable-limit	20
歪み距離制限	なし
数理計画問題の Solver	CPLEX 11.0

5　評価実験

5.1　実験条件

実験には、NTCIR-7 特許文翻訳タスク [10] で配布される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 セットを用いた。またテストセットには同テストセットを用いた。コーパスの詳細を表 1 に示す。翻訳精度の評価には BLEU[11] を用いた。BLEU は正解文と翻訳結果の一致率から算出される翻訳精度の指標であり、100% に近いほど精度は高い。デコーダパラメータ等の実験条件を表 2 に示す。本実験でのベースラインは、オープンソースのフレーズベース統計的機械翻訳システム Moses デコーダ [9] の翻訳結果とする。またフレーズアライナによる reranking 対象は Moses デコーダの翻訳候補上位 100 個とし、提案手法を用いて reranking を行う。

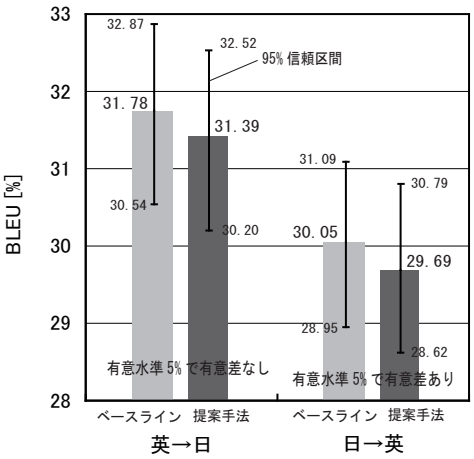

図 6　翻訳実験結果

5.2 実験結果

ベースラインおよび提案手法によるテストセット翻訳結果に対して求めた BLEU 値を図 6 に示す。図 6 より、提案手法はベースラインを改善することはできなかったと言える。この原因としては、翻訳精度の評価基準が BLEU であるのに対して、各翻訳候補の順位付けには確率が用いられており、確率が最大の候補を選んだとしても BLEU を向上させるとは限らないことが考えられる。また確率計算など、実装に不備がなかったか最適化後のフレーズ対応を元に検証したい。

6 まとめ

本稿では、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アライナを利用してフレーズ対応最適化を行うことにより、各候補の確率をより正確に計算しなおし reranking を行う手法について提案した。また、それに付随して歪みモデルを取り入れることのできるフレーズアライナの新たな定式化を提案した。翻訳実験の結果、提案手法は Moses と比べて翻訳精度を向上しないことがわかった。今後は実験に不備がなかったか検証をし、実験結果の裏づけを行いたい。また、提案したフレーズアライナの定式化は DeNero らの定式化に比べて変数が多い。そこで、原言語側の制約条件を DeNero らの制約式で置き換えることにより、変数を削減し高速化を行う予定である。

参考文献

[1] W.Weaver. Translation. *Machine Translation of Languages: fourteen essays*. pp.15-23. Technology press of MIT. Wieley and sons. New York. 1949.

[2] Peter F. Brown, John Cocke, Stephan A. Della Pietra, Vincent J. Pietra, Fredelik Jelinek, John D. Lafferty, Robert L. Mercer and Paul S. Roossin. A Statistical Approach to Machine Translation. In *Computational Linguistics*, Vol.16, No.2, pp.79-85. 1990.

[3] P.F.Brown, J.Cocke, S.A.Della Pietra, V.J.Della Pietra, and R.L.Mercer, "The mathmatics of statistical machine translation: Parameter estimation", *Computational Linguistics*, vol.19, no.2, pp.263-311, 1993

[4] Philipp Koehn, Franz J. Och and Daniel Marcu. 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. Edmonton, Canada. 2003.

[5] Franz J. Och and Hermann Ney. The alignment template approach to statistical machine translation. In *Computational Linguistics*, Vol.30, No.4, pp.417-449. 2004.

[6] Saša Hasan, Richard Zens and Hermann Ney. Are Very Large N-Best Lists Useful for SMT? In Proceedings of Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers, pp.57-60. Rochester, New York, USA. Association for Computational Linguistics. 2007.

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

[8] Christoph Tillmann and Tong Zhang, 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. Ann Arbor, Michigan, USA. Association for Computational Linguistics. 2005.

[9] P.Koehn, H.Hoang, A.Birch, C.Callson-Burch, M.Federico, N.Bertoldi, B.Cowan, W.Shen, C.Moran, R.Zens, C.Dyer, O.Bojar, A.Constantin, and E.Herbst, "Moses: Open Source Toolkit for Statistical Machine Translation", In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pp.177-180, Prague, Czech Republic, June, 2007.

[10] A.Fujii, T.Utsuro, M.Yamamoto and M.Utiyama, "The definition of the patent translation task at NTCIR-7", NTCIR-7, 2007 (to appear).

[11] Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu. 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. Morristown, NJ, USA. Association for Computational Linguistics. 2001.

[12] 越川満. 統計的機械翻訳モデルを用いたフレーズアライナ. 筑波大学第三学群情報学類卒業論文. 2008.

[13] 松井知己. personal communication, 2008.