統計的機械翻訳のための単語列ラベリングに基づく語順推定モデル

Distortion Model Based on Word Sequence Labeling for Statistical Machine Translation

日本放送協会　放送技術研究所ヒューマンインターフェース研究部専任研究員　**後藤　功雄**

PROFILE　1997年NHK入局。2008年より情報通信研究機構に出向。2013年NHKに復帰。2014年京都大学大学院博士課程修了。博士（情報学）。自然言語処理の研究に従事。

1　はじめに

筆者らは統計的機械翻訳（SMT）において語順推定を改良する研究を情報通信研究機構にて実施した[1, 2]。本稿では、この研究成果について紹介する。SMTでは、訳語選択と語順推定という2つの処理が必要である。特に日英など語順が大きく異なる言語間では、語順の推定が難しい課題である。この課題を解決するために、これまでに語彙化語順推定モデル[3]、並べ替え制約[4]、プレオーダリング[5]、階層フレーズベースSMT[6]、構文ベースSMT[7]などが提案されている。一般的に、原言語の構文構造は長距離の語順推定に有用である。しかし、多くの言語では高性能な構文解析器が利用できない。そこで、構文解析器を必要としない翻訳手法も必要である。フレーズベースSMT[8]は、構文解析器を必要としないSMT手法で広く利用されているものの1つである。我々は、フレーズベースSMTで利用する新しい語順推定モデルを提案する。

2　フレーズベースSMTでの語順推定モデル

フレーズベースSMT[8]では、目的言語文を文頭から文末へ連続的に生成していく。そのため語順推定モデルの役割は、翻訳処理中において最後に生成した目的言語の表現の右隣に生成するべき目的言語表現に対応する原言語文中の表現の位置を推定することである。例を図1に示す。図1では、「彼 は」のみが翻訳されている状態を表している。次に生成されるべき語は“bought”であるので、次に翻訳されるべきである原言語文中の語は「買った」である。そのため、語順推定モデルは、「買った」を含むフレーズを次に翻訳するべきフレーズの位置として推定する必要がある。

図1　日英翻訳の語順推定の例。四角の枠はフレーズを表す。

ここで、語順推定の課題をより正確に説明するためにcurrent position（CP）およびsubsequent position（SP）という2つの用語を導入する。CPは、生成中の目的言語文中でアラインメントされている右端の語に対応する原言語文中の位置である。SPは、次に生成する目的言語のフレーズ中でアラインメントされた左端の語に対応する原言語文中の位置である。フレーズベースSMTによる翻訳処理での語順推定モデルの役割は、原言語文中において最後に翻訳した位置（CP）が分かっている状態で次に翻訳する位置（SP）[1]を推定することである[2]。

SPの推定は難しい課題である。図2にCPとSPの例を示す。上付き数字は原言語の単語位置を示す。

図2において、(a)ではSPは8である。しかし、(b)ではCPの語は同じであるがSPは異なる。これらの例

1　正解訳は複数あり得るため、SPは1カ所だけとは限らない。
2　この定義は既存手法[8]の定義とは少し違いがある。既存手法では、フレーズ内の単語対応を考慮せずに、CPは最後に翻訳したフレーズの右端位置、SPは次に翻訳するフレーズの左端位置としている。

図２　日英翻訳での CP と SP の例。CP はボールド、SP はイタリックで表されている。×は SP 以外の SP 候補を表している。

から距離は SP を決める本質的な要因ではないことが分かる。また、SP を推定するのに CP の語とその文脈だけでは十分ではないことが分かる。そのため、CP の語だけでなく SP 候補の語も同時に考慮する必要がある。

図２において、(c)と(d)では CP の語は同じであり、どちらにも SP 候補として「借りた」と「買った」がある。(c)では「借りた」が SP であるが、(d)では「買った」が SP である。この違いの原因の１つは SP 候補間の相対的な語順の違いである。そのため、SP 候補の語だけでなく、SP 候補間の相対的な語順も重要である。

図２において、(d)と(e)では CP の語は同じであり、SP 候補の「買った」と「借りた」の語順も同じである。しかし、(d)と(e)では SP の語が異なる。この要因は CP や SP 候補の語の文脈の違いである。そのため、近い SP 候補を選択すれば良いわけではなく、CP や SP 候補の文脈を考慮する必要がある。

すなわち、SP を推定するためには、CP の語、SP 候補の語、SP 候補間の相対的な語順、CP および SP 候補の文脈を考慮する必要がある。

フレーズベース SMT の語順推定モデルはこれまでに、MSD 語彙化語順推定モデル[3, 9, 10] や、CP もしくは SP 候補の文脈を考慮する語順推定モデル[11, 12, 13] が提案されている。これらのモデルはいずれも SP を CP からの相対位置（距離と方向）で特定し、相対位置の確率を計算する。相対位置はいくつかに分類されて、各分類（クラス）の確率を計算する。この場合、同じクラスに含まれる位置が同じ確率になってしまうという問題がある。この問題を軽減するために分類を細かくすると、各クラスに該当する学習データの量が少なくなって多くのクラスでフレーズや語の頻度が０になり、統計量の信頼性が低下してしまうという問題が起きてしまう。

3　提案手法

我々は、既存手法のように相対位置の確率を計算するというアプローチではなく、各 SP 候補を特徴付ける特徴量に基づいて各 SP 候補が SP である確率を直接計算するというアプローチを提案する。これによって、前記の問題を起こさずに各位置の確率を計算することができる。SP 候補 j が SP である確率は、

$$P(X=j|i, S) \quad (1)$$

と表すことができる。ここで、i は CP、S は原言語文、X は SP の確率変数である。

次に(1)式を実現する２つのモデルを説明する。

3.1　ペアモデル

ペアモデルは、(1)式を実現する手法で、メインの提案手法のベースとなるモデルである。我々はこのモデルを最大エントロピー法で構築する。s を原言語の単語、$s_1^n=s_1 s_2 \cdots s_n$ を原言語文とする。文頭に BOS、文末に EOS の記号を追加する。すると原言語文 S は s_0^{n+1}（s_0＝BOS、s_{n+1}＝EOS）で表される。ペアモデルは次式により確率を計算する。

$$P(X=j|i, S)=\frac{1}{Z_i}\exp(\mathbf{w}^{\mathrm{T}}\mathbf{f}(i, j, S, o, d)) \quad (2)$$

ここで、

$$o=\begin{cases}0 & (i<j)\\ 1 & (i>j)\end{cases},\quad d=\begin{cases}0 & (|j-i|=1)\\ 1 & (2\leqq|j-i|\leqq 5)\\ 2 & (6\leqq|j-i|)\end{cases}$$

$$Z_i=\sum_{j\in\{1\leqq j\leqq n+1\wedge j\neq i\}}\exp(\mathbf{w}^{\mathrm{T}}\mathbf{f}(i, j, S, o, d))$$

であり、$\mathbf{f}(\cdot)$ はバイナリ素性関数を要素とするベクトルで、$\mathbf{w}$ はそれに対応する重みパラメータのベクトルである。Z_i は正規化項、o は i から j への方向、d は距

表１　素性テンプレート

テンプレート
$\langle o\rangle, \langle o, s_{i+p}\rangle^{1}, \langle o, s_{j+p}\rangle^{1}, \langle o, t_i\rangle, \langle o, t_j\rangle, \langle o, d\rangle,$ $\langle o, s_{i+p}, s_{j+q}\rangle^{2}, \langle o, t_i, t_j\rangle, \langle t, t_{i-1}, t_i, t_j\rangle, \langle o, t_i, t_{i+1}, t_j\rangle,$ $\langle o, t_i, t_{j-1}, t_j\rangle, \langle o, t_i, t_j, t_{i+1}\rangle, \langle o, s_i, t_i, t_j\rangle, \langle o, s_j, t_i, t_j\rangle$

[1] $p \in \{p \mid -2 \leqq p \leqq 2\}$
[2] $(p, q) \in \{(p, q) \mid -2 \leqq p \leqq 2 \wedge -2 \leqq q \leqq 2 \wedge (|p| \leqq 1 \vee |q| \leqq 1)\}$

離クラスである。バイナリ素性関数は素性に一致した場合に１、それ以外で０を返す。表１に素性を構築する際に用いる素性テンプレートを示す。t_i は s_i の品詞を表す。

(2)式で i、j、S は素性関数で利用されている。これによって、CP の語、SP 候補の語、CP および SP の文脈を同時に考慮して確率を計算することができる。

3.2　系列モデル

ペアモデルでは SP 候補間の相対的な語順を考慮できないという問題がある。この問題を解決するモデルを提案する。このモデルがメインの提案手法であり、系列モデルと呼ぶ。

図２の(c)と(d)で、「借りた」と「買った」はどちらも原言語文に出現している。ペアモデルは距離の違いによる影響を距離クラスである d のみで考慮している。そのため、もしこれらの語が同じ距離クラスに出現した場合は、これらの語の CP からの距離の違いを扱うことができない。このようなケースは、学習時では適切に分離できない学習データとなり、翻訳時では適切に推定することが困難となる。

系列モデルは相対的な語順を考慮することで、この問題を解決する。これは、CP から各 SP 候補（SPC）の範囲を表すラベル系列を識別することで行う。各ラベル系列は１つの SPC に対応しているので、SP に対応するラベル系列を特定することができれば、SP を特定することができる。ラベル系列は C、I、S という３種類のラベルを用いて CP から SPC の範囲を表す。ラベル C は CP を表し、ラベル S は SPC を表し、ラベル I は CP と SPC の間の位置を表す。図２(c)に対するラベル系列の例を図３に示す。SPC は各ラベル系列の ID として用いる。

ラベル系列は相対的な語順を扱うことができる。例えば、図３のラベル系列 ID が 10 のラベル系列は、「買った[10]」の SPC より CP に近い位置に「借りた」が存在

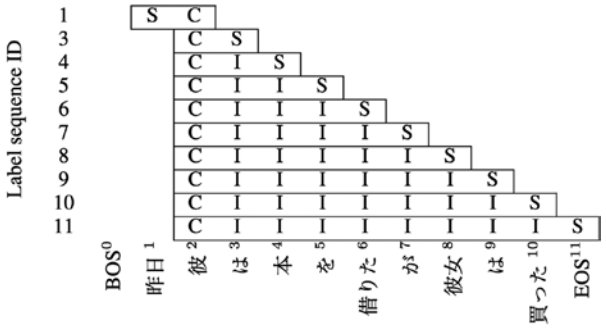

図３　ラベル系列の例。図２(c)の場合。四角の枠内のラベルが系列を表す。

することを把握することができる。なぜなら、「借りた[6]」にはラベル I が対応し、一方「買った[10]」にはラベル S が対応している。そして、ラベル I はラベル S より CP に相対的に近いと定義されているためである。ラベル系列と入力文の単語を用いることで、系列モデルは「買った[10]」の SPC より CP に近い位置に「借りた」が存在することを確率の計算に反映させることができる。

我々は、CRF[14] を基にした系列ラベリング技術を用いて、SP に対応するラベル系列を識別する。我々のタスクと CRF のタスクとは２つの違いがある。違いの１つは、CRF は全てのラベル候補からラベル系列を識別するが、我々のラベル系列は、CP はラベル C、SPC はラベル S、それらの間はラベル I と制約を与えている。もう１つの違いは、CRF は同じ系列データに対するラベル系列を識別することを目的としているが、我々は長さの異なる系列に対するラベル系列を識別する。すなわち、CP と SPC の外側にはラベルを付与しない。この違いは CRF のタスクにおいて、CP と SPC の外側に別のラベル（例えばラベル E）を付与するが、ラベル E に関する素性が全く定義されていない場合に相当する。本稿では、異なる長さの単語列に対応するラベル系列を識別する手法を *partial CRF* と呼ぶ。

我々は partial CRF を用いて系列モデルを構築する。系列モデルは、ラベルと系列を利用するようにペアモデルを拡張することで得られる。ラベルを利用するため

に、図3のようなCPからSPCの範囲を特定するラベル系列を各SPCに対して想定する。そして素性にラベルの情報を加えるために、表1の全ての素性テンプレートに素性テンプレート$\langle l_i, l_j\rangle$を追加する。ここでl_iはiの位置に対応するラベルを表す。例えば、表1の素性テンプレート$\langle o, s_{i+1}, s_j\rangle$に$\langle l_i, l_j\rangle$を追加することで、素性テンプレートは$\langle o, s_{i+1}, s_j, l_i, l_j\rangle$となる。さらに系列を利用するように拡張する。ペアモデルでは、素性テンプレートを適用する際に位置ペア(i, j)を用いた。それに対して系列モデルでは、素性テンプレートを適用する際に位置ペア(i, k), $k\in\{k|i<k\leqq j\vee j\leqq k<i\}$、および$(k, j)$, $k\in\{k|i\leqq k<j\vee j<k\leqq i\}$を用いる。この際、表1の素性テンプレートでは、2つの位置を表す際にiとjを用いているが、素性テンプレートを系列モデルに適用する際には、これらの2つの位置を表す変数のどちらかにkを用いる。例えば、位置ペア(i, k)の場合、iとjを用いた素性テンプレート$\langle o, s_{i+1}, s_j, l_i, l_j\rangle$を適用する際は、$\langle o, s_{i+1}, s_k, l_i, l_k\rangle$とする。

系列モデルはSPCのjがSPである確率を次式で計算する。

$$P(X=j|i, S) = \frac{1}{Z_i}\exp\Big(\sum_{k\in M\cup\{j\}}\mathbf{w}^{\mathrm{T}}\mathbf{f}(i, k, S, o, d, l_i, l_k) + \sum_{k\in M\cup\{i\}}\mathbf{w}^{\mathrm{T}}\mathbf{f}(k, j, S, o, d, l_k, l_j)\Big) \quad (3)$$

ここで、

$$M = \begin{cases}\{m|i<m<j\} & (i<j)\\ \{m|j<m<i\} & (i>j)\end{cases},$$

$$Z_i = \sum_{j\in\{1\leqq j\leqq n+1\wedge j\neq i\}}\exp\Big(\sum_{k\in M\cup\{j\}}\mathbf{w}^{\mathrm{T}}\mathbf{f}(i, k, S, o, d, l_i, l_k) + \sum_{k\in M\cup\{i\}}\mathbf{w}^{\mathrm{T}}\mathbf{f}(k, j, S, o, d, l_k, l_j)\Big)$$

である。jはSPCの他にラベル系列IDを表しているため、(3)式はラベル系列を識別していると見なすことができる。

(3)式の$\exp(\cdot)$内の最初の項は位置iと系列内の他の位置との組およびそれらの文脈を利用し、2番目の項は位置jと系列内の他の位置との組およびそれらの文脈を利用する。

長さの異なる系列を識別するようにモデルを設計することで、系列モデルは距離の影響を自然に扱うことができる。なぜなら、長い系列には多くのラベルがあるため含まれる素性が多く、短い系列にはラベルが少ないため含まれる素性が少ない。この素性の量のバイアスが距離の影響を学習する際に重要な手がかりとなるためである。

3.3 訓練データ

提案手法の語順推定モデルの学習には教師有りの訓練データを用いる。このデータはパラレルコーパスと言語間の単語アラインメントから構築する。図4に訓練データの例を示す。対訳文の目的言語側でアラインメントされている単語を左から右にたどっていくと、それらの単語に対応する原言語単語の順番(原言語側の矢印)が教師データとなる。原言語文と原言語側の矢印で表される情報を訓練データとして用いる。

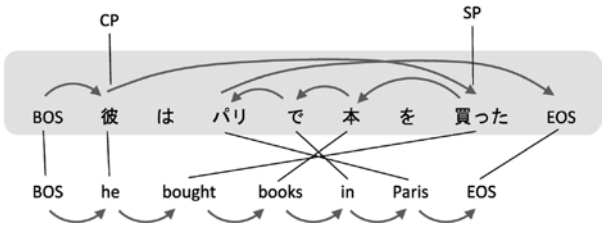

図4 日英翻訳での教師有り訓練データの例。日英の単語間の線は単語アラインメントを表す。

4 実験

4.1 実験設定

NTCIR-9特許機械翻訳タスク[15]のデータを用いて日英(JE)、中英(CE)の特許翻訳の実験を行った。提案手法で用いたフレーズベースSMTの翻訳システムにはMoses[8]と同等のシステムを用いた。単語アラインメントにはGIZA++を用いた。言語モデルは、対訳データの目的言語側のデータで学習し、5-gramを用いた。翻訳モデルは40単語以下の対訳データで学習した。紙面の都合上、詳細な実験設定[1]の説明は省略する。

下記の5つの語順推定モデルをSMTシステムの素性として用いた場合を比較して、提案手法の有効性を評価した。

・線形歪みコストモデル[8](LINEAR)

・LINEARとMSD語彙化語順推定モデル[9](LINEAR+LEX)
・9クラス分類に基づく語順推定モデル[12](9-CLASS)
・提案手法のペアモデル（PAIR）
・提案手法の系列モデル（SEQUENCE）

提案手法のPAIRとSEQUENCEの構築には、計算量を削減するために、翻訳モデルを構築する際に用いたデータのうち20万文のデータを用いて学習した。モデルパラメータの推定にはL-BFGS[16]を用いた。

9-CLASSは、提案手法のモデルで用いたデータと同じ20万文を用いて構築した。モデルパラメータの推定にはL-BFGSを用いた。

MSD語彙化語順推定モデルは、翻訳モデルを構築する際に用いたデータを全て用いて構築した。

線形歪みコストモデルは、距離[3]のみに応じてコストが決まるモデルである。

フレーズベースSMTのdistortion limitには10, 20, 30, ∞の4種類の値を用いた。本稿では、開発データを用いてシステム毎にdistortion limitを選択した。

また、階層フレーズベースSMT（HIER）の翻訳も行った。このシステムにはMosesを用いた。max-chart-spanは∞に設定した。

4.2 評価結果

評価は、BLEU-4（case insensitive）で行った。図5にJEの結果を、図6にCEの結果を示す。提案手法のSEQUENCEの結果は、比較したフレーズベースSMTの語順推定モデルの結果よりも評価が高かった。これにより提案手法の系列モデルの有効性が確認された。SEQUENCEの結果はPAIRの結果より評価が高かった。これにより、SEQUENCEで系列を扱うことで候補間の相対的な語順を考慮できることなどが有効であることが確認された。また、階層フレーズベースSMT（HIER）との比較でもSEQUENCEはJEで値が高く、CEでも同等以上の値であった。

提案手法の系列モデルが距離の影響をどれだけ適切に学習できているか調べるために、距離と平均確率との関係を調べた。具体的には、日英翻訳において、distortion=$j-i-1$と次の3種類の確率との関係を調べた。(1)日本語テスト文におけるSEQUENCEの平均確率。(2)日本語テスト文におけるPAIRの平均確率。(3)訓練データにおける最尤推定の確率（CORPUS）。結果を図7に示す。図7より、距離クラスが同じdistortion（5から20は同じ距離クラス素性）では、PAIRはほぼ同じ平均確率であるのに対し、SEQUENCEはdistortionが大きくなると平均確率が低くなっている。この傾向はCORPUSと同じである。これにより、SEQUENCEが距離による影響をPAIRに比べて適切に学習できていることが確認された。

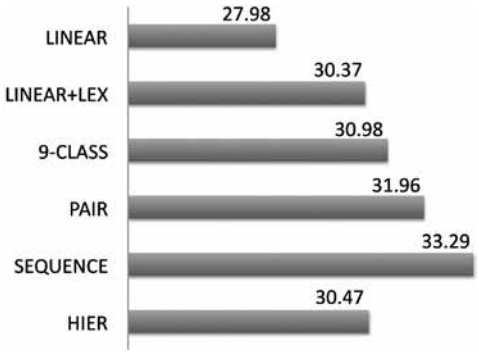

図5　日英翻訳の評価結果

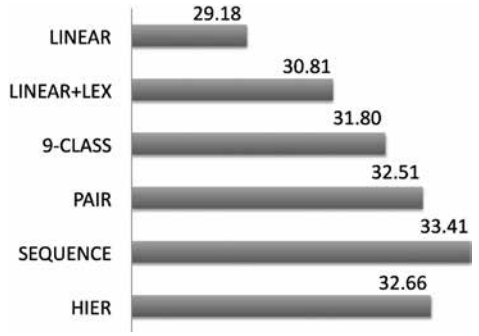

図6　中英翻訳の評価結果

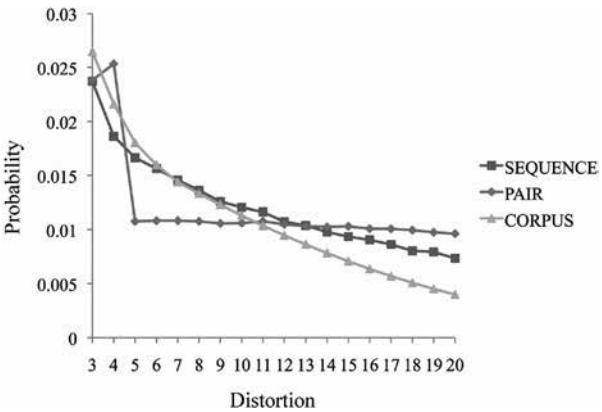

図7　日英翻訳でのdistortionと確率との関係

3　より正確には、次節で説明するdistortionである。

5 まとめ

フレーズベース SMT のための新しい語順推定モデルを提案した。提案した系列モデルは 1 つの確率モデルのみで構成され、候補間の相対的な語順や距離の影響を考慮することができる。日英、中英の特許翻訳の実験で有効性を確認した。

参考文献

[1] Isao Goto, Masao Utiyama, Eiichiro Sumita, Akihiro Tamura, and Sadao Kurohashi. Distortion Model Based on Word Sequence Labeling for Statistical Machine Translation. ACM Transactions on Asian Language Information Processing 13, 1, Article 2, 2014.

[2] Isao Goto, Masao Utiyama, Eiichiro Sumita, Akihiro Tamura, and Sadao Kurohashi. Distortion Model Considering Rich Context for Statistical Machine Translation. In Proceedings of ACL, pp. 155–165, 2013.

[3] Christoph Tillman. A Unigram Orientation Model for Statistical Machine Translation. In Proceedings of HLT-NAACL, pp. 101–104, 2004.

[4] Richard Zens et al. Reordering Constraints for Phrase-Based Statistical Machine Translation. In Proceedings of Coling, pp. 205–211. 2004.

[5] Fei Xia and Michael McCord. Improving a Statistical MT System with Automatically Learned Rewrite Patterns. In Proceedings of Coling, pp. 508–514, 2004.

[6] David Chiang. Hierarchical Phrase-Based Translation. Computational Linguistics 33, 2, pp. 201–228, 2007.

[7] Kenji Yamada and Kevin Knight. A Syntax-based Statistical Translation Model. In Proceedings of ACL, pp. 523–530, 2001.

[8] Philipp Koehn et al. Moses: Open Source Toolkit for Statistical Machine Translation. In Proceedings of ACL, pp. 177–180. 2007.

[9] Philipp Koehn et al. Edinburgh System Description for the 2005 IWSLT Speech Translation Evaluation. In Proceedings of IWSLT, 2005.

[10] Michel Galley and Christopher D. Manning. A Simple and Effective Hierarchical Phrase Reordering Model. In Proceedings of EMNLP, pp. 848–856. 2008.

[11] Richard Zens and Hermann Ney. Discriminative Reordering Models for Statistical Machine Translation. In Proceedings of WMT, pp. 55–63, 2006.

[12] Spence Green et al. Improved Models of Distortion Cost for Statistical Machine Translation. In Proceedings of HLT-NAACL, pp. 867–875. 2010.

[13] Minwei Feng et al. Advancements in Reordering Models for Statistical Machine Translation. In Proceedings of ACL, pp. 322–332. 2013.

[14] John D. Lafferty et al. Conditional Random Fields : Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceedings of ICML, pp. 282–289, 2001.

[15] Isao Goto et al. Overview of the Patent Machine Translation Task at the NTCIR-9 Workshop. In Proceedings of NTCIR-9, pp. 559–578, 2011.

[16] D.C. Liu and J. Nocedal. On the Limited Memory Method for Large Scale Optimization. Mathematical Programming B 45, 3, pp. 503–528. 1989.