句に基づく統計翻訳における語句の並べ替えパターンの分析

大橋一輝 [1]　山本和英 [1]　齋藤邦子 [2]　永田昌明 [2]

長岡技術科学大学 [1]
{ohashi, ykaz}@nlp.nagaokaut.ac.jp

NTT サイバースペース研究所 [2]
{saito.kuniko, nagata.masaaki}@lab.ntt.co.jp

1　はじめに

近年、対訳コーパスから翻訳モデルを学習することで統計的に翻訳を行う統計翻訳についての研究が盛んに行われている。その中でも句に基づく翻訳モデル [1] は、語に基づく翻訳モデルに比べて、文脈に基づく訳語選択や局所的な語の並べ替えを表現する能力が高いため、現在の統計翻訳モデルの主流になっている。しかし、句に基づく翻訳モデルは、翻訳する句の相対的な位置のみに依存しているため、大局的な語句の並び替えを表現する能力が低い。よって、日本語と英語のような語順が大きく異なる言語間の翻訳では翻訳精度が劣るという問題点がある。

本稿では、英日翻訳において、どのような語句の並べ替えパターンがどれぐらい存在するかについて統計的に分析した。どの語句からどのような順番で翻訳されていくのかを見ることで、より精度の高い語句の並べ替えモデルを構築できるのはないかと考えている。

分析の方法としては、まず、英日対訳コーパスから句に基づく翻訳モデルを構築する。そして、この翻訳モデルの確率を用いて、ビタビアルゴリズムによる英日対訳コーパスの句対応付けをする。原言語の文と目的言語の文との句対応を用いて、並べ替えについての分析を行った。さらに、句対応の中で IBM 制約および ITG 制約 [3] を満たす文の割合を調べた。前者は原言語を翻訳する順番に関する制約、後者はふたつの句の両言語における相対的な位置に関する制約である。両者が、離れた言語間の翻訳において有効であるかどうかを評価した。

その結果、英日翻訳における語句の並び替えパターンはある程度品詞に依存していることを確認した。英語側の句の先頭が名詞の場合では並べ替えが発生しにくく、動詞や前置詞の場合では並べ替えが発生しやすかった。この分析の上で、句の並び替えをうまく扱えるような歪みモデルを構築したい。さらに、ふたつの制約についてはどちらも有効であることを確認した。

2　句に基づく翻訳モデル

我々は、句に基づく翻訳モデルとして [1] を使用する。本節では、このモデルについて説明する。

統計翻訳では、原言語 f が目的言語 e へ翻訳される確率 $p(e|f)$ を最大とする目的言語の文 $\hat{e}$ を求める。これはベイズの法則により $p(f|e)p(e)$ を最大化すればよい。

$$\hat{e} = \arg\max_e p(e|f) = \arg\max_e p(f|e)p(e)$$

ここで、p(e) を言語モデル、$p(f|e)$ を翻訳モデルと呼ぶ。翻訳モデルは次式で表される。

$$p(\bar{f}_1^I|\bar{e}_1^I) = \prod_{i=1}^{I} \phi(\bar{f}_i|\bar{e}_i) d(a_i - b_{i-1}) \quad (1)$$

ここで、$\phi(\bar{f}_i|\bar{e}_i)$ を翻訳確率、$d(a_i - b_{i-1})$ を歪み確率と呼ぶ。翻訳モデルはこれら二つの確率を考慮する。式中の I は原言語 f の形態素の連なり、$\bar{f}_1^I$ はこれを句に分割したもの、$\bar{f}_i$ は分割したそれぞれの句、$\bar{e}_i$ は $\bar{f}_i$ に対応した句、a_i は新たに翻訳する句の左端の位置、b_{i-1} は直前に翻訳した句の右端の位置とする。

すなわち、翻訳モデルの確率 $p(f|e)$ は、文を構成する句すべての翻訳確率と歪み確率の積となる。翻訳確率は、次式による相対確率で算出する。

$$\phi(\bar{f}|\bar{e}) = \frac{count(\bar{f}, \bar{e})}{\sum_f count(\bar{f}, \bar{e})} \quad (2)$$

歪み確率は、次式によって算出する。

$$d(a_i - b_{i-1}) = \alpha^{|a_i - b_{i-1} - 1|}$$

これは、翻訳する原言語の句の位置のずれに依存するモデルである。この例を図 1 に示す。

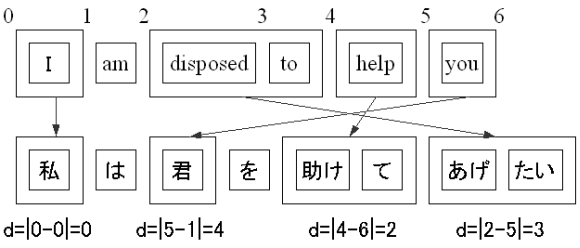

図 1　歪み確率のパラメータ

例えば、英語の「disposed to」の場合を考える。a_i は新たに翻訳する句の左端の位置であり、「disposed to」の左端は 2 となる。そして、b_{i-1} は直前に翻訳した句の右端の位置であり、「help」の右端は 5 である。よって、$d = |2-5| = 3$ となる。

句の相対的な位置のみを考慮しているため、英日翻訳で動詞が文頭近くから文末まで移動する、というような並べ替えをすることができない。語順が大きく異なる言語間では、句についての情報をもっと考慮する必要があると考えられる。

3　語句の並べ替えパターン

語句の並べ替えパターンを考えるために、図 2 を示す。

図では英日の対訳文の句対応付けを示している。ここで、図の英語文を翻訳して日本語文を頭から生成していくことを考える。「言語」は「Language」から生成する。「は」は英語側に表現がないので挿入する。次の「コミュニケーションの」は英語の文末の「of communication」から生成する。

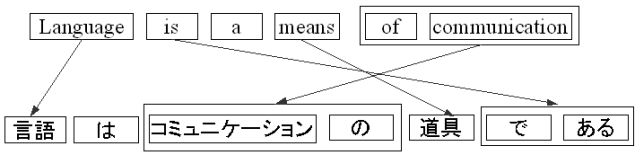

図 2　考慮する並べ替え

さらに、「道具」を「means」、「である」を「is」から生成して終了する。

以上のように、日本語の文を先頭から順に生成しようとすると、語順の違いによって英語の生成元の単語の位置が異なってくる。この違いを、次のようなパターンとして考える。

(1) 直前に翻訳した原言語の句と今回翻訳した原言語の句が正順に連なっている (正順)。

(2) 直前に翻訳した原言語の句と今回翻訳した原言語の句が逆順に連なっている (逆順)。

(3) 直前に翻訳した原言語の句と今回翻訳した原言語の句が正順で間隙がある (正順間隙あり)。

(4) 直前に翻訳した原言語の句と今回翻訳した原言語の句が逆順で間隙がある (逆順間隙あり)。

図 2 でこのパターンの例を示す。最初に翻訳される「Language」は、文頭であり (1) とする。次の「of communication」は、直前に翻訳した「Language」とは正順で且つ離れているため (3) である。「means」は「of communication」とは逆順で連なっているため (2) である。「is」は「means」とは逆順で且つ離れているため (4) である。

それから、文献 [1] で用いるモデルには無いが、語の挿入および削除をパターンとして考える。これは、対訳文の一方では表現されるが他方では表現されない語句の存在に対応するためである。原言語の単語が目的言語では表現されない場合を削除、目的言語の単語に対応する原言語の単語が存在しない場合を挿入とする。図 2 では、「a」が削除であり、「は」が挿入である。

以上の点について考えることにより、言語間の語順の違いがどのように並び替えを強制しているのかについて分析を行う。

4　分析方法

まず、GIZA++[2] を用いて英日対訳コーパスの単語対応付けを双方向に行い、式 (2) の相対頻度を用いて句翻訳モデルを構築する。次に、構築したモデルに入力として対訳コーパスを与え、確率最大となるビタビ対応を算出する。以上で得られた句の対応付けを用いて先述のパターンについて調べる。

4.1　句翻訳モデルの構築

原言語から目的言語、目的言語から原言語の両方向について、GIZA++ を用いて対訳コーパスの単語対応を求める。ここで、2 つの対応付けの積集合 (intersection) と和集合 (union) を考える。 両方向で一致している積集合の方がより信頼できる対応付けとなる。この例を図 3 に示す。

積集合の要素を句の中心とし、その近傍の積集合および和集合の要素をひとつの句だと見なすことで句の範囲を大きくしていく。近傍とは、図 4 に示すように、ある単語対応の左下と右上を除いた周囲 6 箇所とする。これは、右上

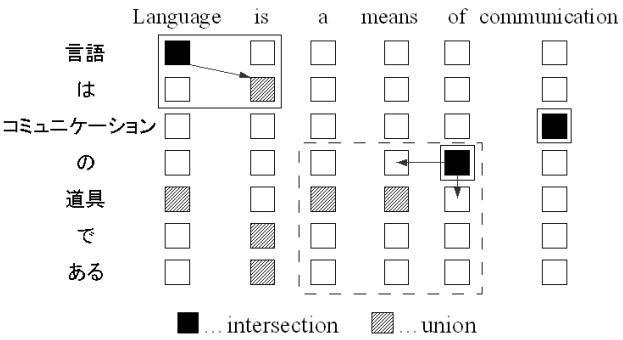

図 3　句の抽出

の対応は、中心の対応に対する位置の違いが原言語で＋１、目的言語で－１と大きいためで、左下も原言語で－１、目的言語で＋１となり同様である。図 3 では、一番左上の積集合の要素から右下の和集合の要素へと句を拡大している。句の大きさに制限は設けておらず、近傍がある限り句を広げていく。

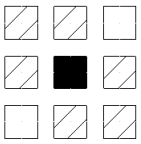

図 4　単語対応の近傍

積集合および和集合以外に句の手がかりは存在しない。しかし、句を広げきった後にも句を構成できない単語が存在する場合がある。そのときは、句が無矛盾 (consistent) であることを条件として、句へと吸収させる。無矛盾とは、原言語と目的言語の対応する句に含まれる単語が、お互いの句の単語だけに対応付けられており、句の外側の単語には対応付けられていない状態を指す。吸収の手順としては、ある句が上記条件に当てはまるかどうかを調べ、可能であれば単語ひとつだけを吸収させる、という処理をすべての句に対して繰り返す。

図 3 では、句を構成していない英語の「a」、「means」、日本語の「道具」「で」「ある」を句に取り込んでいく。すべての単語が句を構成したら終了である。

本稿では以上のように句の抽出を行った。ここで、両方向で対応付けが存在しない語句は、相手側の言語で表現されない語句ではないかと考えられる。このような語句を扱うための機構も必要だと考えている。

それから、可能な限りのすべての句を組み合わせとして抽出する処理も可能であるが、本稿ではこれを行わない。これは、クローズドテストの際に支障が出るためである。クローズドテストではすでに学習した文に対して翻訳を行うため、原言語の文全体と目的言語の文全体を記憶しておくことで、簡単に翻訳ができてしまう。句の組み合わせを学習すると、複数の句をつなげた翻訳が可能になるために、語句の並び替えを分析するという目的に反する。オープンテストに関しては後述する。

以上によって翻訳となる対の句を求める。そして、式 (2) の相対頻度により翻訳確率を求める。

4.2 ビタビ対応

構築した翻訳モデルを用いて、目的言語の文頭から文末方向へ一単語ずつ翻訳句の解析していくビタビアルゴリズムによって、対訳コーパスのビタビ対応を求める。

まず、翻訳モデルの翻訳候補の中で、翻訳となりうる候補をすべて抽出する。翻訳入力文と翻訳出力文が対訳コーパスから与えられているので、考慮すべき候補は限られている。

次に、目的言語の文頭から処理を始め、現在の解析位置までの候補と、その解析位置の右側の単語を含む翻訳候補の組み合わせを考える。英日翻訳における例を図 5 に示す。

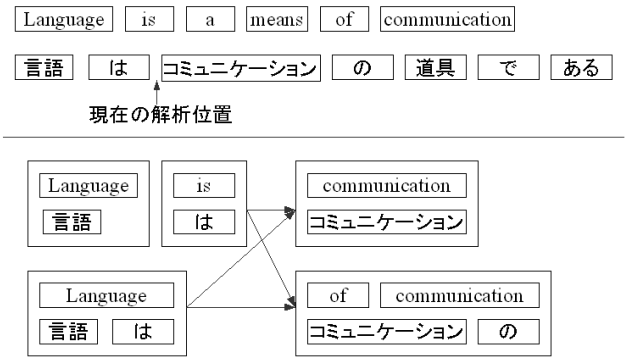

図 5　ビタビアルゴリズム

文頭から現在の解析位置までの左側は、日本語「言語 は」をどう句にするのか、その句はどの英単語に対応するのか、という候補。そして、現在の解析位置から右側の日本語にも同様に候補がある。この左側と右側の組み合わせを新たなる候補として生成する。

しかし、これらの組み合わせは可能な場合と不可能な場合がある。文頭から現在の解析位置の左側までの目的言語の句によってカバーされる原言語の集合と、その解析位置の翻訳候補の目的言語の句によってカバーされる原言語の集合が重なってはならない。原言語の同じ単語を二度も翻訳してしまうような組み合わせは成り立たないためである。

対応付けの確率としては、句翻訳モデルの翻訳確率のみを考える。ただし、解析中の解析位置における翻訳候補が存在しない場合もありうる。このときは、一つの単語だけから構成される目的言語の句を挿入し、ペナルティを与える。

最後に、探索が目的言語の文末に達したら、原言語の単語がすべて翻訳されているかどうかを調べる。翻訳されていない原言語の単語がある場合は、それぞれを一つの単語から構成される句と考え、この句を削除するコストを与える。

以上によって対応付けの確率を求めることができ、すべての候補の中で確率最大の対応付けがビタビ対応となる。この対応付けについて、3 節で述べたパターンの数がどの程度存在するのかを調べる。

5 実験

本稿では、英日翻訳についての実験を行った。英日の対訳コーパスとして、次に挙げる計 78 万文を用いた。

- 日英新聞記事対応付けデータ (JENAAD) 15 万文 [6]
- 辞書例文 55 万分
- 技術文献 8 万文

これらのうち 500 文をテストデータとして分けておく。他をすべて学習して翻訳モデルを作り、テストデータの 500 文の対応付けをした。しかし、3483 単語中に 1183 語 (34%) もの挿入が発生してしまい、評価ができなかった。これは、英語と日本語が両方とも決定されているためである。テスト文の両言語の句がそのままの形でモデルに存在しなければ翻訳することができない。よって、以降では 500 文のクローズドテストに対する評価のみを述べる。

5.1 語句の並び替えパターンの数

並び替えパターンの全体数は、英日翻訳において、正順が 515、正順間隙ありが 355、逆順が 448、逆順間隙ありが 87 であった。[1] の翻訳モデルでは語順に関して句の相対的な位置のみを考慮しているが、逆順や間隙のある並び替えが合わせて全体の 6 割を占めていることを考えると、もっと句の情報を考慮するモデルが必要であると言える。

5.2 品詞による語句の並べ替え

語句の並べ替えが発生するのは双方の言語の語順の違いが原因であり、この違いは品詞に依存していると考えられる。そこで、英日翻訳における語句の並べ替えパターンに対する英語句の品詞の割合、および、この英語句に対応する日本語句の品詞の割合について調べた結果が表 1 および表 2 である。英語側の句は先頭の語の品詞、日本語側の句は末尾の語の品詞を用いた。なお、品詞のタグ付けには、日本語は茶筌 [5]、英語は MXPOST[4] を使用し、茶筌の品詞は第 2 階層までを使用した。

表 1　品詞と語句の並べ替えパターン (英語句)

	正順	正順間隙あり
1	冠詞 (24%)	冠詞 (23%)
2	代名詞 (18%)	名詞 (19%)
3	名詞 (17%)	前置詞 (12%)
4	動詞 (12%)	代名詞 (11%)
	逆順	逆順間隙あり
1	動詞 (23%)	動詞 (20%)
2	前置詞 (19%)	代名詞 (18%)
3	限定詞 (15%)	名詞 (16%)
4	名詞 (13%)	限定詞 (9%)

表 2　品詞と語句の並べ替えパターン (日本語句)

	正順	正順間隙あり
1	助詞-係助詞 (20%)	名詞-一般 (23%)
2	名詞-一般 (15%)	助詞-格助詞 (16%)
3	動詞-自立 (9%)	助詞-連体化 (10%)
4	助詞-格助詞 (9%)	助動詞 (7%)
	逆順	逆順間隙あり
1	助動詞 (26%)	助動詞 (29%)
2	動詞-自立 (17%)	助詞-格助詞 (13%)
3	助詞-格助詞 (13%)	動詞-自立 (11%)
4	名詞-一般 (10%)	助詞-連体化 (8%)

英日翻訳における英語句の正順では、冠詞、代名詞、名詞をまとめると 60%くらいが名詞であった。正順間隙ありでは、正順と似ているが、前置詞が上位に入ってきている。英語は前置詞を置いて後ろから修飾するのに対し、日本語は前から修飾していくことが表れていると考えられる。逆

順では動詞、前置詞が上位になっている。これは、英語と日本語の動詞の位置の違いが原因だと推測できる。英語は主語の次に動詞が来るのに対し、日本語では動詞が末尾に来る。英語と日本語の動詞の位置が逆順であるため、これがそのまま反映されているのである。逆順間隙ありは 逆順に似ているが、代名詞が上位になっている。対応付けを見てみると、日本語文の主語が省略されている文に多く見られた。日本語の文末を生成するとき、最後に英語の文頭に戻って動詞を含む句を翻訳しているため、そのときの主語が句の先頭の代名詞となっている。そして以上の英語句に対応する日本語句は、英語句が正順のときは係助詞や名詞が多く、英語句が逆順のときは助動詞や動詞が増えている。語順の品詞への依存が表れていると言える。

5.3 IBM 制約と ITG 制約

語順の制約として IBM 制約と ITG 制約 [3] がある。IBM 制約は、翻訳は常に言語の文頭から文末へ行われていくと仮定する。原言語の文の先頭からまだ翻訳されていない語句 n 個のみを翻訳対象とし、それ以降の語句は考慮しないという制約である。本稿では、対応付けした 500 文の中で、単語数 n による IBM 制約を満たすものがどの程度あるのかを調べた。その結果を表 6 に示す。

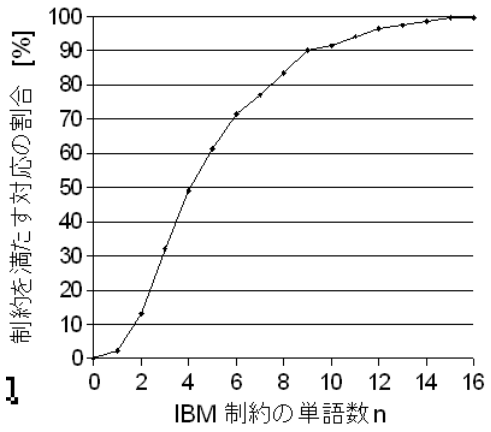

図 6　IBM 制約を満たす対応の割合

単語数 7、すなわちテスト文の平均単語数で 8 割弱の対応を満たしている。計算量の削減をどの程度できるのかは調べていないが、ある程度は有効に機能するのではないかと考えられる。

次に、ITG 制約は、ふたつの句が正順もしくは逆順でつながっている状態を文全体で保つ、という制約である。この制約を満たす対応の割合についても同様に調べた。結果、500 文のすべてが制約を満たしており、ITG 制約は英日翻訳において非常に精度の高い制約であると言える。

6　考察

統計的機械翻訳のひとつの特徴として、扱う言語対に依存せずに翻訳ができるという点がある。しかし、英日翻訳では語句の並べ替えパターンに正順や逆順が入り混じり複雑であることから、語順の大きく異なる言語間では、言語に依存した文法的な知識を扱う必要が出てくると考えられる。

本稿のオープンテストは、翻訳モデルの学習した句が不足していて評価することができなかった。その原因のひとつとして、日本語の活用や表記揺れをすべてコーパスから学習することは難しい点が挙げられる。このような部分は自動生成したり、知識として持っておくなどの処理が必要である。

それから句の抽出についてであるが、本稿で示したような、品詞に依存して語句の並べ替えを表現できるモデルを構築したとする。句の先頭や末尾の品詞に依存して並べ替えるのだとすると、句の境界の選び方によって品詞が変わってくるため、高い精度で句を抽出すべきである。モデルにだけでなく、句の抽出においても何らかの文法的な知識を用いる必要があるのではないかと考えている。

最後に、有効性を示した IBM 制約と ITG 制約について、本稿で実験に用いたのは平均単語数 7 語の非常に短いテストデータである。文が長くなるにしたがって語句の並べ替えは複雑になるため、ある程度の単語にそろえたテストデータを用いて実験することが必要である。

7　おわりに

本稿では、文中における句の構文的な役割を、英語では句の先頭の単語の品詞で代表させ、日本語では句の末尾の単語の品詞で代表させることにより、正順・逆順という観点からみた英日翻訳における句の並び替えの傾向を、ある程度うまく説明できることが分かった。今後は、この分析をふまえ、句に基づく翻訳モデルにおいて、句に依存した歪み確率のモデルを検討したい。

参考文献

[1] Philipp Koehn, Franz J. Och, and Daniel Marcu. Statistical phrase-based translation. In Marti Hearst and Mari Ostendorf, editors, *HLT-NAACL 2003: Main Proceedings*, pp. 127–133, Edmonton, Alberta, Canada, May 27 - June 1 2003. Association for Computational Linguistics.

[2] Franz Josef Och and hermann Ney. A systematic comparison of various statistical alignment models. In *Computational Linguistics.*, Vol. 29. No. 1.

[3] T. Watanabe R. Zens, H. Ney and E. Sumita. Reordering constraints for phrase-based statistical machine translation. In *Proceedings of the 20th International Conference on Computational Linguistics (CoLing 2004)*, pp. pp. 205–211, Geneva, Switzerland, August 2004.

[4] Adwait Ratnaparkhi. Mxpost(maximum entropy pos tagger), ver.1.0. http://www.cis.upenn.edu/~adwait/statnlp.html, 1997.

[5] 奈良先端科学技術大学院大学松本研究室. 形態素解析器「茶筌」, ver.2.3.2. http://chasen.aist-nara.ac.jp/, 2003.

[6] 独立行政法人通信総合研究所. 日英新聞記事対応付けデータ (jenaad). http://www2.nict.go.jp/jt/a132/members/mutiyama/jea/index-ja.html.