統計的機械翻訳は意訳しにくいのか？

竹元 勇太，山本 和英

長岡技術科学大学 電気系

E-mail：{takemoto,ykaz}@nlp.nagaokaut.ac.jp

1 はじめに

統計的機械翻訳において、翻訳モデル(機械翻訳知識)を対訳コーパスから自動構築する際、コーパス中の翻訳の多様性が高い場合や単語アライメントが推定しにくい場合、翻訳精度が低くなってしまうという問題がある。この問題に対して今村ら[1]は、対訳文の直訳性を利用して、対訳コーパスから翻訳モデルを構築しやすい対訳文だけを用いることで解決しようとしている。その結果、直訳性の高い対訳文は翻訳モデルを効率よく構築でき、反対に直訳性の低い(意訳性の高い)対訳文は翻訳モデルを構築しにくいことが確認されている。しかし、統計的機械翻訳は直訳しやすく、そして意訳しにくいという傾向があるかどうかまでは示されていない。

もし、そのような傾向が統計的機械翻訳にあるならば、目的言語に対して意訳となる文を入力とするべきではない。また、意訳をさせる場合は、水田ら[4]のように、意訳しやすい機械翻訳を行うなど、意訳に特化した手法を用いる必要があると考える。

そこで本研究では、統計的機械翻訳は直訳しやすく、意訳しにくい傾向があるのか確認をすることに焦点を当てて研究を行った。

2 直訳性の測定方法

対訳文の直訳性を測定する方法として、今村らが考案した対訳対応率(*Translation Correspondence Rate* ； *TCR*)がある。*TCR*は対訳文の単語対応率から直訳性を測定するものである。そこで本研究では直訳性を測定するための指標として *TCR*を使用することにした。そして、*TCR*値の高い対訳文を直訳的対訳文、*TCR*値の低い対訳文を意訳的対訳文と呼ぶことにする。

*TCR*は対訳文の対訳対応率から直訳性を測定するため、対訳辞書や統計的単語アライメントが必要となる。以下に*TCR*の式を示す。

$$TCR = \frac{2L}{Ts + Tt} \quad (1)$$

*Ts*は対訳辞書中の見出しに含まれる原言語の単語数で、*Tt*は対訳辞書中の訳語に含まれる目的言語の単語数である。*L*は対訳文の単語対の内、対訳辞書中に対訳として含まれる数である。図１に今村らから引用した *TCR*の計算例を示す。図１の原文と翻訳文１は直訳の関係で、原文と翻訳文２は意訳の関係である。*TCR*はこのように直訳性の判定が可能である。

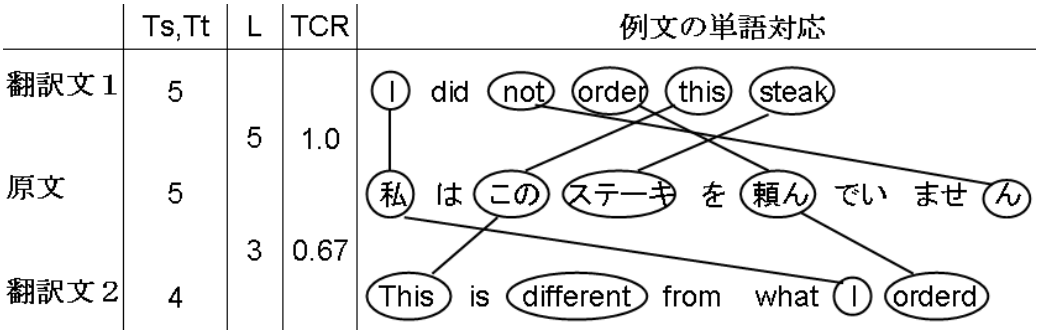

図１　*TCR*の計算例
（丸囲み単語の個数が *Ts*及び *Tt*、直線の数が *L*に値する）

3 翻訳実験

本研究では、*TCR*を用いて直訳的対訳文のテストセットと意訳的対訳文のテストセットを対訳コーパスから抽出した。そして、日英機械翻訳での精度を比較することによって、統計的機械翻訳は直訳しやすく、意訳しにくいかを確認する。また、本研究では、アライメント推定ツールに *GIZA++*[1]、言語モデル構築ツールに *IRSTLM*[3]、デコーダに *Moses*[5]を使用した。

3.1 実験データ

対訳コーパス

本研究では、言語資源として日英対訳コーパス[7]の内、374,085対訳を使用した。対訳コーパスには前処理として以下の処理を行っている。

原言語(日本語)コーパス

日本語の形態素解析器 *ChaSen*[4]を使用して、分ち書きにし、原形化を行った。また、数字やアルファベットは分ち書きをせず、半角に統一した。

目的言語(英語)コーパス

英語の形態素解析器 *TreeTager*[2] を使用して、分ち書き及び原形化を行った。また、大文字は小文字に、全角は半角に変換した。

対訳辞書

対訳辞書には以下の 2 種類を使用した。

A：*GIZA++*を使用して、全対訳コーパス(374,085 対訳)から単語アライメントを推定して構築した対訳辞書(748,258 対訳)。

B：英辞郎の対訳辞書(1,473,940 対訳)[6] の内、1 単語の対訳となっているものだけで構成した対訳辞書(153,067 対訳)。

評価用データ

評価用データの詳細は以下の各実験で説明する。

学習データ

学習データには対訳コーパスから評価用データを抜いたものを使用した。その他の詳細は以下の各実験で説明する。

言語モデル構築用データ

言語モデル構築用のデータは対訳コーパスの目的言語を使用した。そして、構築する際は評価用データの目的言語の文と完全一致するものは除いている。言語モデルの構築には *IRSTLM* を使用しており、構築する際の単語数(*N-gram*)は猪澤ら[3] の研究結果を参考に 5-gram とした。

3.2 対訳辞書 A を使用した実験

評価用データの作成

始めに、対訳辞書 A を使用して全対訳文(374,085)に *TCR* 値を付与する。その後、*TCR* 値の上位 30,000 対訳と *TCR* 値の下位 30,000 対訳、そして対訳コーパスからランダムに、それぞれから 1,500 対訳を抽出し、500 対訳をチューニング用の開発データに、1,000 対訳をテストセットに使用した。

翻訳モデルの構築

翻訳モデルは以下の 3 種類を構築しており、全て評価用データを除いたもの(369,608 対訳)から構築している。以後、各翻訳モデルを以下の名称で示す。

直訳モデル

369,608 対訳の内、*TCR* 値の高い方から 184,804 対訳を使用して構築したもの。

意訳モデル

369,608 対訳の内、*TCR* 値の低い方から 184,804 対訳を使用して構築したもの。

ランダムモデル

369,608 対訳の内、ランダムで 184,804 対訳を使用して構築したもの。

各翻訳モデルに対して 3 種類のテストセットで評価実験を行った。本研究では、評価方法として *BLEU*[5] を用い、使用した参照訳は 1 原文あたり 1 つで、4-gram まで使用した。

表 1　翻訳モデルの違いによる各テストセットの評価結果(BLEU)

翻訳モデル	テストセット		
	直訳	意訳	ランダム
直訳	0.297	0.087	0.257
意訳	0.201	0.125	0.226
ランダム	0.270	0.099	0.229
平均	0.256	0.104	0.237

この実験結果から、どの翻訳モデルを使用した場合でも、翻訳精度(*BLEU* 値)は意訳的対訳文のテストセットが最も低く、最大でも 0.125 であった。この値は直訳的対訳文やランダムのテストセットの *BLEU* 値と比べると約 1/2 程度と非常に低い。また、直訳的対訳文のテストセットはランダムと比べて平均して高い *BLEU* 値を出している。これらのことから、統計的機械翻訳は直訳しやすく、特に意訳しにくい傾向があると考える。

次に、翻訳モデルの違いによる傾向を見てみる。ランダムのテストセットの評価結果から、最も *BLEU* 値が高いのは直訳モデルを使用した場合である。これは、今村らの実験結果からも同じ傾向が表れていることから、統計的機械翻訳では直訳的対訳文で翻訳モデルを構築することで、翻訳精度を向上させることができると考える。

意訳的対訳文のテストセットでは意訳モデルが最も高い *BLEU* 値を出している。このことから、統計的機械翻訳は意訳しにくいという傾向はあるが、意訳的対訳文で翻訳モデルを構築することによって、意訳する傾向が高くなることがわかった。また同様に、直訳的対訳文のテストセットは直訳モデルがもっとも高い *BLEU* 値を出していることから、直訳的対訳文で学習することによって、より直訳しやすくなると考える。

3.3 対訳辞書 B を使用した実験

3.2 節の実験結果で統計的機械翻訳は意訳しにくいという傾向が表れていたが、それは *GIZA++* の単語アライメント推定の傾向が表れているだけの可能性もあると考える。そのため、推定したものではなく、正確な対訳辞書 B(英辞郎の 1 単語対訳データ)を使用することによって、その可能性の真偽を確かめるための実験を行った。

評価用データの作成

対訳辞書 B は対訳辞書 A に比べ対訳数が約 1/5 と圧倒的に少ない。そのため、*TCR* 値の信頼度を上げるために、以下の条件式を満たす対訳文だけを使用することにした。

$$\frac{Ts+Tt}{Ws+Wt} \geq 0.9 \qquad (2)$$

Ws は原言語の単語数、*Wt* は目的言語の単語数である。この条件を満たす対訳文は 26,095 対訳であった。ここで、*TCR* 値を変化させた時の *BLEU* 値

の変化も見るため、抽出した対訳文(26,095 対訳)を分割した。

抽出した対訳文を *TCR* 値で昇順にソートし、上から 2,610 対訳ごとに分割した。そして、分割された 2,610 対訳ごとから開発データ 500 対訳とテストセット 1,000 対訳を抽出した。

翻訳モデルの構築

翻訳モデルは、対訳コーパスから全開発データ(5,000 対訳)と全テストセット(10,000 対訳)を除いた 359,431 対訳から構築した。

評価結果を図 2 に示す。図の各点の *TCR* 値は分割された 2,610 対訳ごとの *TCR* 値の平均である。

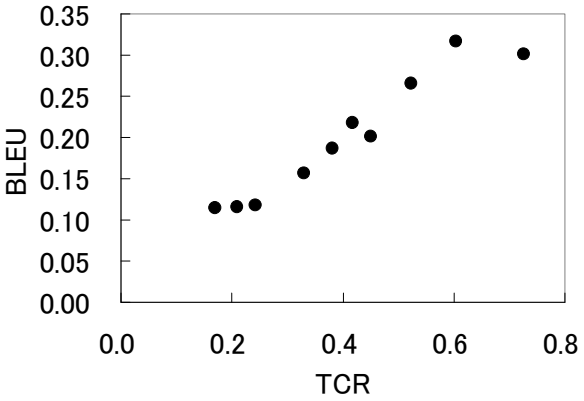

図 2　*TCR* 値を変化させた時の翻訳精度(*BLEU*)

このグラフから、*GIZA++*の精度や特徴などに関係ない対訳辞書 B を使用した場合であっても、*TCR* 値の低い(意訳性の高い)テストセットは *BLEU* 値が低いという結果となった。また、*TCR* 値と *BLEU* 値には正の相関があることも確認された。

3.4 翻訳モデルの構築方法についての実験

3.2 節の実験結果から、統計的機械翻訳では直訳的対訳文で翻訳モデルを構築すれば、より直訳しやすくなり、意訳的対訳文で構築すれば、より意訳しやすくなるという傾向があることが確認された。このことから、対訳コーパスを直訳的対訳文と意訳的対訳文の 2 つに分割して翻訳モデルを構築することによって、翻訳の直訳性と意訳性が向上し、同時に翻訳精度も向上すると考える。そこで、以下 4 種類の翻訳モデル構築方法を比較して確認をすることにした。

評価用データは 3.2 節と同様で、学習用の対訳コーパスは 3.2 節の 369,608 対訳の内、ランダムに抽出した 360,000 対訳を使用した。

全体モデル

全対訳文 360,000 対訳から構築したもの。

直訳モデルと意訳モデル

始めに、全対訳文に *TCR* 値を付与し、降順にソートした。そして、*TCR* 値の高い方から 30,000 対訳ずつに分割点を付けた。その後、分割点の上部の対訳文を直訳モデル構築用データ、下部を意訳モデル構築用データとした。つまり、直訳モデルと意訳モデルそれぞれに 11 個の翻訳モデルを構築した。

合成モデル

分割点で区切られたデータでそれぞれ構築した直訳モデルと意訳モデルを組み合わせた翻訳モデル。学習量としては全体モデルと常に同じ 360,000 対訳である。合成モデルの個数は全部で 11 個である。

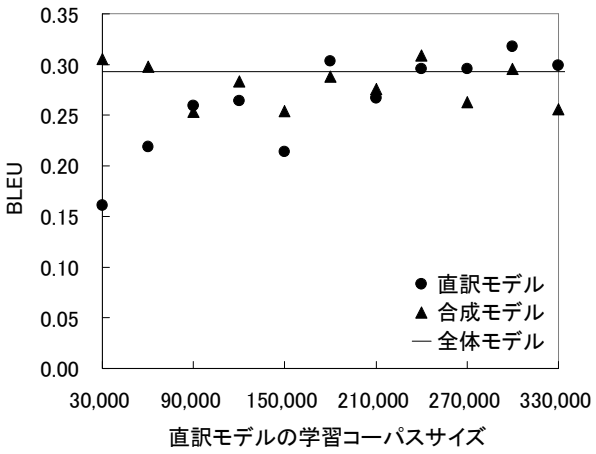

図 3 直訳的対訳文のテストセットで評価したコーパスサイズによる翻訳精度の変化

図 3 から、直訳モデルは学習コーパスサイズの量と比例して、上下しながらも *BLEU* 値が上昇した。そして、直訳モデルは学習コーパスサイズが 300,000 対訳の時、*BLEU* 値が 0.318 と最大になり、学習コーパス全部を使用した全体モデルや合成モデルより翻訳精度が高くなった。

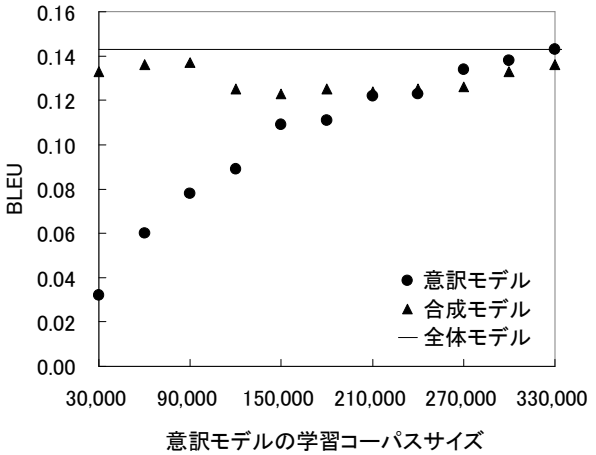

図 4 意訳的対訳文のテストセットで評価したコーパスサイズによる翻訳精度の変化

図 4 から、意訳モデルはコーパスサイズが 330,000 対訳の時が 0.143 と最も高く、全体モデルと同じ *BLEU* 値であった。このように、意訳モデルは図 3 の直訳モデルと比べて全体モデルを上回るまでの翻訳精度を出すことはできなかった。そのため、意訳モデルを用いて意訳に特化させようとするよりは、対訳コーパス全てを使用した方が高い翻訳精度を出せると考える。

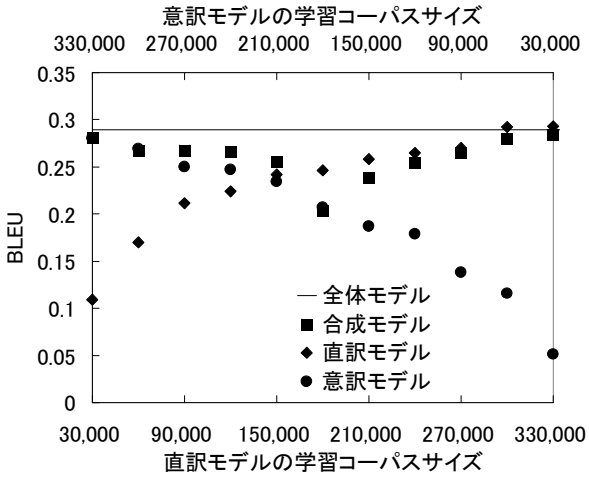

図 5 ランダムのテストセットで評価した
コーパスサイズによる翻訳精度の変化

図 5 から、合成モデルは意訳モデルと直訳モデルのコーパスサイズが同等の 180,000 対訳の時、最も *BLEU* 値が低くなり、0.203 であった。そして、180,000 対訳を境に、直訳モデルと意訳モデル構築に使用する対訳コーパスの量が多くなるにつれて、*BLEU* 値も上昇した。この傾向は、図 3 と図 4 にも見られる。

図 5 において、合成モデルと直訳モデルを比較すると、直訳モデルの学習コーパスサイズが 180,000 以上では、常に直訳モデルの方が合成モデルより高い *BLEU* 値を出している。

合成モデルは対訳コーパス(360,000 対訳)を 2 つに分割し、直訳モデルと意訳モデルを構築した後、2 つのモデルを混同させたモデルである。そのため、同じフレーズに異なる翻訳確率が付与される場合がある。また、フレーズに付与される翻訳確率は、学習コーパスサイズに比例する。そのため、直訳モデルの学習コーパスサーズが大きくなったとしても、意訳モデルの学習コーパスサイズが小さくなることで、合成モデルの翻訳確率の精度が減少したと考えられる。

4 おわりに

本研究では、統計的機械翻訳は意訳しにくい傾向があるのか確認をするために、2 種類の辞書を使用して翻訳実験を行った。その結果、統計的機械翻訳は直訳しやすく、意訳しにくいという傾向があることがわかった。

また、直訳的対訳文で構築した翻訳モデルと対訳コーパス全てを使用した翻訳モデルを比較した。その結果、直訳的対訳文のテストセットを使用した評価において、高い翻訳精度を出すことができた。このことから、直訳性の高い対訳文に制限して翻訳モデルを構築することによって、より直訳しやすくなることが確認された。

しかし、意訳的対訳文で構築した翻訳モデルと対訳コーパス全てを使用した翻訳モデルを比較した場合、意訳的対訳文のテストセットを使用した評価において、全体モデルを上回って意訳しやすくなることはなかった。

結論として、統計的機械翻訳を行う場合、目的言語に対して意訳となる文を入力とするべきではないことが確認された。

参考文献

1] 今村賢治, 隅田英一郎, 松本裕治. 直訳性を利用した機械翻訳知識の自動構築. 自然言語処理, Vol.11, No.2, pp.85-99, 2004.

2] 今村賢治, 隅田英一郎. 直訳性に着目した対訳コーパスフィルタリング, FIT 2002, E-52, Vol.2, pp.185-186, September 2002.

3] 猪澤雅史, 村上仁一, 徳久雅人, 池原悟. 統計翻訳における単文・重文複文の翻訳精度の評価, 言語処理学会第 14 回年次大会, pp.869-872, 2008.

4] 水田理夫, 徳久雅人, 村上仁一, 池原悟. 重文・複文文型パターン辞書による意訳の可能性, 電子情報通信学会ソサイエティ大会発表, 基礎・境界, 言語の意味と思考過程, AS-5-4, 2007.

5] 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 of the Association for Computational Linguistic (ACL)*, pp.311-318, 2002.

6] Philipp Koehn, Franz J. Och, and Daniel Marcu. Statistical phrase-based translation. In Proceedings of HLT-NAACL 2003, pp. 127–133, 2003.

使用した言語資源及びツール

1) アライメント推定ツール GIZA++, Franz Josef Och and Hermann Ney, Ver.1.0.2 , http://www.fjoch.com/GIZA++.html
2) 英語形態素解析器 TreeTagger, Ver.3.2, The University of Stuttgart, http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
3) 言語モデル構築ツール IRSTLM, Ver.5.20.0 http://sourceforge.net/projects/irstlm/
4) 日本語形態素解析器 ChaSen, Ver.2.4.2, 奈良先端科学技術学院大学 松本研究室, http://ChaSen.naist.jp/hiki/ChaSen/
5) デコーダ Moses, Philipp Koehn, http://www.statmt.org/moses/index.php?n=Main.HomePage/
6) 日英対訳辞書 英辞郎, Ver.54
7) 日英対訳コーパス, CREST「セマンティックタイポロジーによる言語の等価交換と生成技術」プロジェクト.