E-52

直訳性に着目した対訳コーパスフィルタリング
Bilingual Corpus Filtering Based on Translation Literality

今村 賢治†, 隅田 英一郎†
Kenji Imamura, Eiichiro Sumita

1 はじめに

対訳コーパスの充実に伴い、コーパスから自動学習した翻訳システムも提案されてきている。我々も、大規模対訳（パラレル）コーパスからの翻訳知識自動獲得を前提としたパターンベースの翻訳システムを提案した[1]。これは、対訳文から句レベルの対応を自動的に獲得し、これを変換規則として翻訳を行う、構文トランスファ翻訳方式を採用している。しかし、変換規則の構築過程で、コーパスまたは対訳文に起因する精度低下が観測された。

一般的に、コーパスを用いた言語処理を議論する場合、コーパス量が問題にされることは多いが、コーパスの質に関して議論されることは少ない。制限言語に関する研究では、例外的に表現の質が議論されているが、これは、単言語が対象である。しかし、その研究成果として、使用する語彙や文法を制限することにより、意味解釈の曖昧性が減少するなどの効果が報告されている。対訳コーパスにおいても、表現を制限することにより、曖昧性の減少や機械学習の容易性（小さなコーパスで学習できる）が向上すると期待される。

本稿では、対訳文の質を判断する尺度として、対訳の直訳性を提案する。その上で、より直訳性の高い対訳文に制限して部分コーパスを自動的に作成することにより、小規模コーパスでも大規模コーパスと同等の翻訳システムが構築できることを示す。

2 対訳コーパスにおける翻訳の多様性

まず、翻訳知識自動獲得における対訳コーパスの問題点を、翻訳の多様性という観点から述べる。

言い換え表現 一般的に、原言語1文は、目的言語では複数の文に翻訳することができる。たとえば、英語文「How long does it take to get there?」は、日本語では「そこに行くのにどのくらい時間がかかりますか」「所要時間はどのくらいですか」「どれくらいで着きますか」などに翻訳することができる。これらはすべて正しい訳である。しかし、このような多様性を含むコーパスから翻訳知識を自動学習すると、過剰に知識が生成される。たとえば、[1]で使用した、パターンベース翻訳システムの場合、これらの対訳からはすべて異なる変換規則が作成される。しかし、実際の翻訳処理では1規則だけが必要であるため、不要な規則は曖昧性の増大や翻訳速度の低下を招く。

文脈／状況依存訳の存在 コーパス中には通常、常に正しいとは限らない対訳が存在する。そのうち、顕著なものは、文脈や状況に依存した訳文である。

たとえば、英語を日本語に翻訳する場合、通常定冠詞 '*the*' は訳出されることはないが、文内文脈では解決できない曖昧な表現となる場合、「私の」「その」など、限定表現が付与される場合がある。この限定表現は文脈に依存するので、誤った文脈で使われると、わきだし語など、誤訳の原因となる。

3 直訳性に着目したフィルタリング

上記問題点を解決するため、我々は直訳性を利用した対訳コーパスフィルタリングを提案する。

直訳は、word-to-word translation とも言われ、単語レベルでの翻訳を意味している。また現在のところ、機械翻訳における変換単位は、単語または句が主流を占めており、その最小単位は単語である。つまり、単語の翻訳を要素合成原理に基づいて合成し、文の翻訳を行っているものが多い。したがって、直訳性が高い文は、より機械翻訳に適した対訳と言える。このような対訳を集めて機械翻訳システムを構築することにより、文脈依存訳に起因するわきだし語や過剰な省略を抑えることができる。また、言い換え表現のうち、機械翻訳に適した対訳だけが残るため、曖昧性が減少する。

本稿では、直訳性を以下の式で表し、これを単語対応スコア（WCS）と呼ぶこととする。

$$WCS = \frac{C_s + C_t}{W_s + W_t} \quad (1)$$

ただし、C_s, C_t は、それぞれ原言語、目的言語で対応づけられた単語数、W_s, W_t は原言語、目的言語の単語総数である。つまりこの式は、対応づけられた単語のカバー率を表している。単語同士の対応は、統計的単語アライメント[2]などで提案されているように、コーパスから自動的に取得することが可能である。

単語対応スコアを用いた直訳性の判定例を図1に示す。これは英語「How long does it take to get there?」に対する2つの翻訳文の直訳性を表したものである。翻訳文1は、原文に対して4つの単語対応を持つが、翻訳文2は1つしかない。そのため単語対応スコアは、翻訳文1では $\frac{4+4}{8+11} = 0.42$、翻訳文2では $\frac{1+1}{8+7} = 0.13$ となり、翻訳文1が、より直訳性が高いと判定される。

4 実験

今回、英日翻訳について実験を行い、コーパスフィルタリングの効果を、自動学習された翻訳システムの翻訳品質という観点で評価を行った。

†ATR 音声言語コミュニケーション研究所
*本研究は通信・放送機構の研究委託「大規模コーパスベース音声対話翻訳技術の研究開発」により実施したものである。

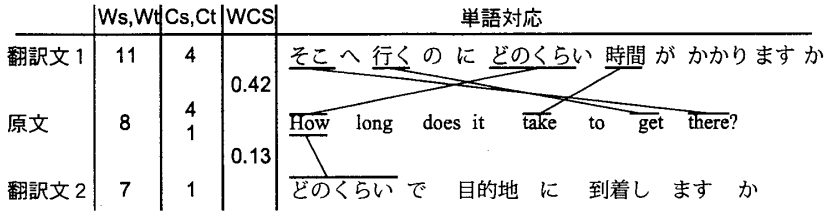

図 1: 単語対応スコアによる直訳性の判定例

4.1 実験条件

対訳コーパス 使用したコーパスは、旅行会話に頻出する基本表現を集めたコーパス [3] である。このうち、原文と翻訳文の対がユニークである 114,535 対訳を使用した。

単語対応抽出方法 コーパス中の共起頻度が 10 回以上の単語同士について、[2] と同様な方法で統計的に単語アライメントを行った。さらに、低頻度語については、シソーラスを参照して、同一グループに属する単語を対応するものとした。

翻訳システム 実験には、構文トランスファ方式のパターンベース翻訳システム HPAT[1] を使用した。

評価方式 3節で述べた単語対応スコアを用いて全対訳をソートし、高単語対応スコア N 対訳、低単語対応スコア N 対訳 (N は可変) で部分コーパスを作成した。そして各部分コーパスから変換規則を作成し、機械翻訳を行った。評価は、テストセット 510 文の翻訳結果に対する一対比較法を用いた。すなわち、コーパス全体を使った機械翻訳文を基本とし、それに対して部分コーパスを用いた機械翻訳文の品質が改善／悪化したのか、あるいは同程度であるのか、日本語ネイティブ話者 1 名による主観評価を行った。したがって、本評価はコーパス全体に対する相対評価である。

4.2 実験結果

各部分コーパスにおける翻訳品質を図 2に示す。なお、図の縦軸は、コーパス全体を用いた機械翻訳文に対する、部分コーパス機械翻訳文の改善率で、次の式で表される。

$$\text{翻訳品質} = \frac{\text{改善文数} - \text{悪化文数}}{\text{テスト文数}} \qquad (2)$$

まず、高 WCS と低 WCS 部分コーパスの翻訳品質を比べると、すべてのコーパスサイズにおいて、高 WCS の方が圧倒的に翻訳品質が勝っている。つまり、単語対応スコアは機械翻訳に適した対訳の選択に寄与している。

高 WCS 部分コーパスのみに着目すると、コーパス全体から 8 万対訳までは、コーパスサイズが減少したにも関わらず、翻訳品質は若干 (約 2%) 向上している。これは、わきだし語など、誤訳の原因となる対訳が削除され、翻訳システムの変換規則の精度が向上したためである。なお実験では、変換規則数はコーパスサイズにほぼ比例していた。

さらにコーパスサイズが減少すると、変換規則のカバレッジが低下するため、翻訳品質も低下する。しかし、

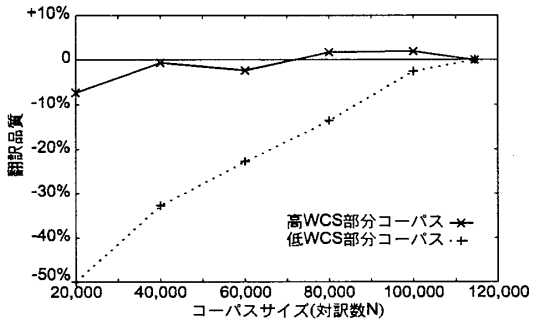

図 2: コーパスサイズによる翻訳品質の変化

4 万対訳まで減少させても、品質はほぼ同等、2 万対訳でも品質低下は 7.3% にとどまった。この数値は翻訳方式に依存すると考えられるが、直訳性に基づくコーパスフィルタリングを行うと、より小規模のコーパスでも大規模コーパスと同等の翻訳システムが構築できることを示している。

5 まとめ

本稿では対訳文の直訳性に基づくコーパスフィルタリングを提案した。単語対応スコアと呼ぶ直訳性評価尺度を定義し、コーパスフィルタリングを実施した。そして、自動学習された機械翻訳の訳文品質として、その効果を評価した。その結果、コーパスサイズを 1/5 以下に削減しても、翻訳品質の低下が 10% 以下にとどまり、本手法は対訳コーパスサイズ削減に有効であることを示した。

参考文献

[1] K. Imamura, "Application of translation knowledge acquired by hierarchical phrase alignment for pattern-based MT," in *Proceedings of TMI-2002*, pp. 74–84, 2002.

[2] I. D. Melamed, "Models of translational equivalence among words," *Computational Linguistics*, vol. 26, pp. 221–249, June 2000.

[3] T. Takezawa, E. Sumita, F. Sugaya, H. Yamamoto, and S. Yamamoto, "Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world," in *Proceedings of LREC 2002*, pp. 147–152, 2002.