パラレルテキストの自動生成に基づく越日統計的機械翻訳

Le Tuan Anh　　秋葉　友良
豊橋技術科学大学
{toxuann, akiba}@cl.ics.tut.ac.jp

1. はじめに

機械翻訳の分野では、短時間低コストでシステム開発が可能な統計的機械翻訳が有望視されている。この手法は、大規模なパラレル・テキストに基づいて翻訳に必要な情報を統計的に取得し、それを用いて新規入力文の翻訳を行う。統計的機械翻訳を実装するためには、翻訳を行う言語対の大規模なパラレル・テキストが必要となる。しかし、言語資源が限られている言語を一方に持つ言語対では、大規模なパラレル・テキストは入手困難である。パラレル・テキストが不足する言語対で、統計的機械翻訳を行う手法として、中間言語を用いる研究が多い。Wangら[1]は中国 - 日本語の単語対訳のエラーを改良するために、英語 - 日本語と英語 - 中国語の2つ大規模なコーパスを用いた。Utiyamaら[2]は英語を中間言語として、2つの翻訳テーブルを作成してから、それらを統合した翻訳テーブルを作成し、統計的機械翻訳を行う手法を提案した。Gisperら[3]は英語 - カタロニア語のコーパスを作成するために、統計的機械翻訳で、英語 - スペイン語コーパスのスペイン語文をカタロニア語に翻訳し、パラレル・テキストを自動生成する。これらの研究は、2対の大規模なパラレル・テキストを用いる手法である。しかし、本稿で対象とするベトナム語に対しては、利用可能な大規模なベトナム語と他言語の間のパラレル・テキストは存在しない。Nguyenら[4]はベトナム語－英語の統計的機械翻訳を構築しているが、利用したパラレル・テキストは小規模なものである。本稿では、英語を中間言語として、ベトナム語と英語の類似性を利用し、英語 - 日本語のコーパスから、新たなベトナム語 - 日本語の対訳コーパスを作成し、統計的機械翻訳を実現する手法を提案する。本手法の利点は英語と目的言語の大規模なパラレルテキストが入手できれば、自動的にベトナム語と目的言語のパラレル・テキストが得られる点にある。

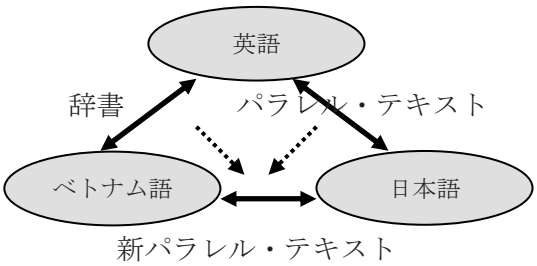

図 1：パラレル・テキストの自動生成

2. ベトナム語の特徴

本稿の手法ではベトナム語の 2 つの特徴を利用した。ベトナム文の例を図２に示す。

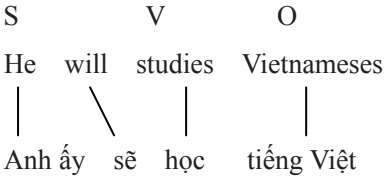

図 2：ベトナム文の例

第１の特徴は文法である。ベトナム語の語順は、SVOの形式で、英語とよく似ている点である*。そして、ベトナム語は、名詞の単数・複数や男女名詞などによる語形が変化するような規則がない。更に、時制の違いによっても、動詞や形容詞は活用しない。その代わりに、時制を現す

* 疑問文でも語順が変化しないという英語と異なる点もある。

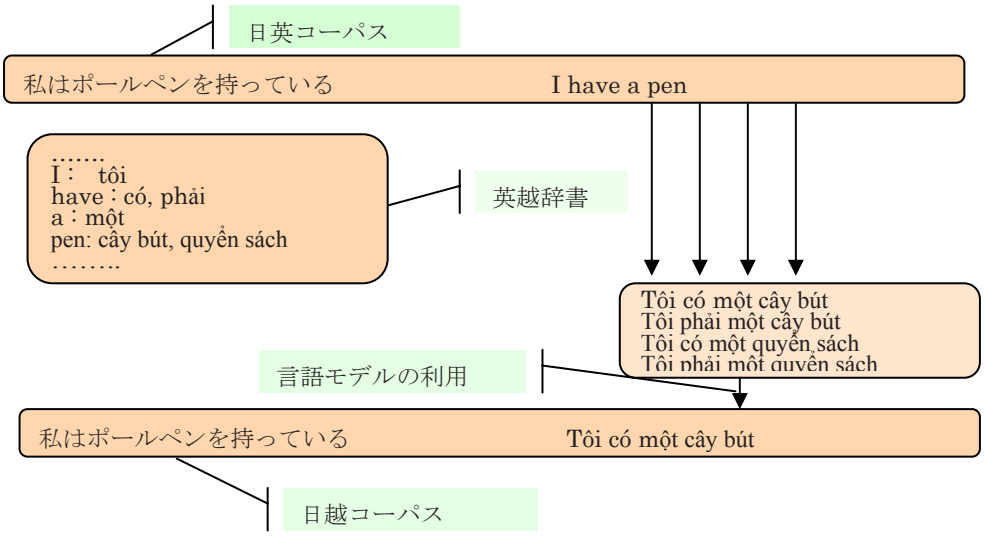

図３パラレルテキストの自動生成の例

特定の単語を付け加える。

第２の特徴は語彙である。ベトナム語は英語のように、文中でスペースを用いるが、スペースの役割は、単語の区切りではなく、音節の区切りである。ベトナム語では、多音節の単語もあるため、単語中でもスペースが表れる場合がある。

3. 中間言語を用いたコーパスの作成

ベトナム語と英語の類似性を利用して、新たなベトナム語 - 日本語パラレル・テキストの作成方法を考える。パラレル・テキストを作成するために、英語の単語・フレーズからベトナム語の単語・フレーズへの対応関係を表す辞書（単語・フレーズ対訳辞書）と英語 - 日本語のパラレル・テキストを用いる。作成方法の例を図３で示す。作成の手順を以下に述べる。

① 英語 - 日本語のパラレル・テキストから対訳ペアを１つ取り出し、そのうちの英語文を取り出す。

② その英文に対して、POS タガー[10]を適用して、英単語を原型に変換する。

③ 英語 - ベトナム語の単語・フレーズ対辞書を参照しながら、辞書項目に合致する英単語・フレーズを対応するベトナム語の単語・フレーズに置き換える。英文全体に対し、置換を繰り返し、全ての英単語がベトナム単語に置き換えられた文だけを選択する。

④ 対訳辞書には１つの英単語・フレーズに対し、２つ以上のベトナム語単語・フレーズが対応する可能性があること、置き換えの適用順によって、選択されるベトナム語単語・フレーズが変わることから、2 つ以上のベトナム文候補が得られる場合がある。この場合、考えられるベトナム文中で最も良い文を選択するために、ベトナム語言語モデルを用いて、最も確率の高い文を選択する。

⑤ 選択したベトナム文と元の対訳ペア中の日本語文を組み合わせて、ベトナム文 - 日本語文の対訳ペアを作成する。

以上の操作を全ての対訳ペアに対して繰り返すことにより、ベトナム語 - 日本語の対訳コーパスが作成できる。

4. データ

提案手法で利用した言語資源について述べる。

☆ 英語 - ベトナム語の対訳辞書

英語 - ベトナム語の機械可読辞書とベトナ

ム語 - 英語の機械可読辞書[1]から、単語・熟語の項目を抽出して、単語・フレーズ対訳辞書を作成した。対訳辞書は約 8.8 万英語項目を含んでいる。また、同じ機械可読辞書から、5 節で述べるベースライン・システムに用いるベトナム語 - 英語間のパラレル・テキストのため、例文対訳も抽出した。約 5.1 万文ペアが収集された。

☆ ベトナム語の単言語コーパス

ベトナム語の言語モデルを作成するために、ベトナム語のコーパスをオンライン新聞サイトtuoitre[2]から収集した。収集した新聞記事は約 19 万記事で、約 300 万文を含んでいる。

☆ 英語 - 日本語のコーパス

英語 - 日本語のコーパスは機械可読辞書[3]から収得した。収集した文は約 9 万文である。

5. 評価実験

提案した方法を評価するために、翻訳実験を行った。

ベースラインシステムとしては、英語を介して、ベトナム語 - 英語と英語 - 日本語のフレーズ・ベース統計的機械翻訳を続けて適用する手法を実装した。ベトナム語から英語への統計的機械翻訳には翻訳モデルの学習データとして、4 節で述べた 5．1 万対のベトナム語 - 英語パラレル・テキストを、英語の言語モデルは Phraoh[7]パーケージに付属する Europarl の言語モデルを用いた。英語から日本語への統計的機械翻訳には、翻訳モデルの学習データとして 4 節で説明した 9 万文の英和パラレル・テキストを、また言語モデルの学習データには新聞記事の 20 万文を用いた。

提案方法は節 3 で説明した手順で、新しいベトナム語 - 日本語パラレル・テキストを作成した後、統計的機械翻訳を行った。日本語の言語モデルにはベースラインと同じ、20 万文の新聞記事を用いた。

5.1 パラレル・テキストの作成

3 節で述べた方法で、ベトナム語 - 日本語のパラレル・テキストを作成した。ただし、3 節で述べた作成手順のステップ 3 において、置き換える単語・フレーズを選択するために、言語モデルを用いる代わりに、辞書項目との最長一致で、置き換えるべき単語・フレーズを選択する方法を用いた。4 節で述べた 8．8 万項目の英語 - ベトナム語対訳辞書と約 9 万文の英日パラレル・テキストを用いた。対訳辞書に項目がない英単語を含む文が存在するため、全ての英文が必ずベトナム文に変更できるとは限らない。抽出したベトナム語 - 日本語のパラレル・テキストは約 3 万文である。

5.2 翻訳の評価

作成したパラレル・テキストを利用して、

単語ベース統計的機械翻訳とフレーズベース統計的機械翻訳を構築した。利用したツール・モデルを表 1 で示す。

表 1 利用するツール

	翻訳モデル	言語モデル	デコーダ
単語ベース[5]	IBM モデル GIZA++ [11]	3-gram 言語モデル CMUCambridge Toolkit[8]	ISI-rewrite decoder
フレーズベース[6]	フレーズベース翻訳モデル GIZA++,Thot	3-gram 言語モデル SRILM Toolkit	Pharaoh decoder [7]

評価のため、ベトナム語 - 日本語の 100 文ペアを人手で作成した。ベトナム語側を入力として、翻訳された日本語文と人で作成した日本語文の比較を行った。評価尺度には BLEU[9]を用いた。結

[1] Free Vietnamese dictionary project
http://www.informatik.uni-leipzig.de/~duc/Dict/
[2] Tuoi tre online
http://www.tuoitre.com.vn
[3] 英和活用大辞典

果を表２に示す。

表２実験結果の BLEU スコア

ベースライン	単語ベース 統計的機械翻訳	フレーズベース 統計的機械翻訳
0.0021	0.0246	0.0542

BLEU スコアの比較より、提案した方法はベースラインに比べて、良い結果が得られた。単語ベース統計的機械翻訳の BLEU スコアはベースライン・の BLEU スコアの１０倍であり、フレーズ・ベース統計的機械翻訳の BLEU スコアはベースラインの２５倍であった。

6. 終わりに

本稿では、言語資源が限られた言語に対し、言語資源が豊富な中間言語との類似性を利用することで、新しいパラレル・テキストを自動的に作成し、統計的機械翻訳を行う手法を提案した。ベトナム語 - 日本語を対象言語とし、評価実験を行った。実験により、提案方法は中間言語を介して、2 回翻訳するベースライン手法より、良い性能を示すことがわかった。今後の課題として、より良いパラレル・テキストを作成するために、形態変換や構文変形などのような前処理を行うことが考えられる。

参考文献

[1] Haifeng Wang, Hua Wu, and Zhanyi Liu.Word alignment for languages with scarce resources using bilingual corpora of other language pairs. In Proc. of COLING/ACL 2006 Main Conference Poster Sessions, 2006.page 874-881

[2] Masao Utiyama and Hitoshi Isahara. A comparison of pivot methods for phrasebased statistical machine translation. In HLT-NAACL, 2007 page 484-491

[3] Adri´a de Gispert and Jos´e B. Mari no. Catalan-English statistical machine translation without parallel corpus: Bridging through Spanish. In Proc. of LREC 5th Workshop on Strategies for developing Machine Translation for Minority Languages,2006. page 65-68

[4] Thai Phuong Nguyen and Akira Shimazu. 2006. Improving Phrase-Based SMT with Morpho-Syntactic Analysis and Transformation.In Proc. of AMTA 2006, pages 138–147.

[5] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation.Computational Linguistics, 19(2): 263-311.

[6] Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation.In HLT-NAACL, 2003.page 48-54

[7]Philipp Koehn. Pharaoh: a beam search decoder for phrase-based statistical machine translation models. In Proc of AMTA, 2004. page 115-124

[8]Philip Clarson, Ronald Rosenfeld. Statistical language modeling using the CMU-cambridge toolkit.In Proc of Eurospeech, 1997.page 2707-2710

[9] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation.In ACL, 2002.page 311-318

[10] H.Schmid TreeTagger – a language independent part-of-speech tagger. http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreetagger.html

[11] F.J.Och GIZA++:Training of statistical translation model. http://www-i6.informatik.rwth-aachen.de/Colleagues/och/software/GIZA++.html