ベトナム語翻訳への教師なしバイリンガルトークナイザの適用

野村　高広　　塚田　元　　秋葉 友良
豊橋技術科学大学
nomura@nlp.cs.tut.ac.jp tsukada@brain.tut.ac.jp akiba@cs.tut.ac.jp

概要

ベトナム語の翻訳に際して、教師なしのバイリンガルトークナイザの活用結果を報告する。本分割手法は、単語辞書を用いず、対訳の情報を活用するものである。従来の単語辞書を用いる単言語分割手法と比べて、同等の翻訳精度を達成した。

1　序論

統計翻訳は、英語や中国語、アラビア語、ヨーロッパの言語など大量の対訳コーパスが利用可能な言語対でその有用性が示されてきた。一方、多くのアジアの言語については、利用できる対訳コーパスが少なく、統計翻訳を適用しにくい状況にある。ベトナム語はリソースの少ない言語の一つであるが、近年 TED talk の出現により、かなりの量のベトナム語－英語対訳コーパスが利用可能になってきた。これにより、ベトナム語は統計翻訳適用の新たな対象となりつつある。

ベトナム語の分割は英語のそれと異なり、各トークンは必ずしも単語に対応しているわけではない。この特徴は、フレーズアライメントの基となる単語アライメント精度の低下につながると考えられる。そこで、ベトナム語を単語単位に分割しなおすことで、単語アライメント精度を向上させ、翻訳の性能を改善することが期待できる。このようなベトナム語の再分割には、単語辞書を用いる手法（本論文では教師ありトークナイザと呼ぶ）[1] が一般的であるが、統計翻訳の場合、目的言語の単語に合わせた単位に自動的に分割することで、単語辞書を用いた手法を上回る性能向上が期待できる。本論文では、単語辞書を用いない対訳の情報を活用したトークナイザ（教師なしバイリンガルトークナイザと呼ぶ）[2] をベトナム語に適用した検討結果を報告する。ベトナム語-英語の翻訳タスクで、教師なしの手法でありながら教師ありのベトナム語トークナイザと同等の性能を達成することができた。

以下に、論文の構成を示す。2 章ではベトナム語について説明をする。3 章では、分割の手法についての説明をする。そして、4 章では我々のシステムを使った実験の結果を示し、5 章ではこの論文のまとめと今後の予定について述べる。

2　ベトナム語について

ベトナム語と英語のフレーズ対応の例を図 1 に示す。この図が示すように、ベトナム語は英語と同様にスペースで区切られているが、各トークンは単語ではなく、おおむね音節に相当する単位となっている。例えば、図 1 の 2 トークン “kêt́ quá” は、英語の 1 単語 “result” に対応している。ベトナム語の翻訳を考えたときに、ベトナム語のトークンを英語の単語単位に区切ることができれば翻訳性能の改善につながると考えられる。

3　分割手法について

ベトナム語トークナイザのベースラインとして、vnTokenizer[1] を使用した。本トークナイザは単語辞書を使用しているため、教師あり手法呼ぶ。

教師なしバイリンガルトークナイザとして、Tagyoung らが提案した手法 [2] を用いる。本手法は単語辞書を用いる必要がなく、対訳コーパスから得られる統計情報のみ活用して分割を行う。本手法を用いることにより、ベトナム語の分割で英単語との対応しやすさを考慮することができる。本手法は中国語や韓国語のようにスペース区切りされていない言語に適用するために文字単位の処理として提案されたものである。ベトナム語に適用するにあたり、いくつかのベトナム語のトークンを “_” で連結する処理に変えて用いる。

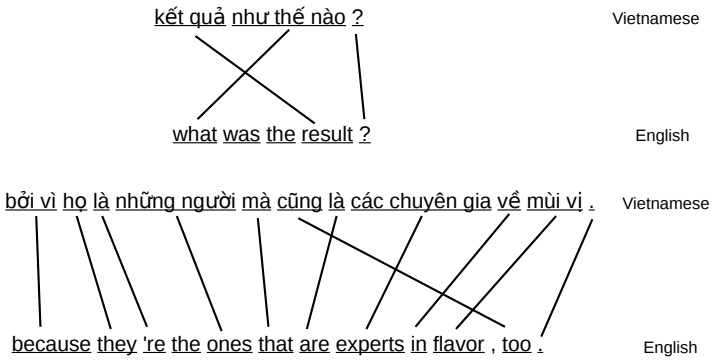

図 1: ベトナム語と英語のフレーズアライメント

3.1 バイリンガルモデル

バイリンガルモデルは以下の式で定義される。入力データは英語のトークン列 e^n とベトナム語のもともとのトークン列 s^m である。

$$P(f, a = k|e) = \frac{\alpha(i)P(f_i|e_k)P(a = k)\beta(j)}{P(\mathbf{s}|\mathbf{e})}$$

ここで、$f = \{s_i s_{i+1} ... s_j\}$ はベトナム語のトークンを i 番目から j 番目までつなげた新たなトークンであり、a は、f を生成するための英単語の位置を示す変数である。ここで、α と β は以下の式で与えられる。

$$\alpha(i) = \sum_{l=1}^{L} \alpha(i-l) \sum_{a} P(a)P(s_{i-l}^{i}|e_a)$$

$$\beta(j) = \sum_{l=1}^{L} \sum_{a} P(a)P(s_{j}^{j+l}|e_a)\beta(j+l)$$

ここで、L は 1 単語あたりの最大の音節数を表す。

このモデルは EM アルゴリズムによって学習される。E ステップとして、それぞれの単語ペアの期待値を計算する。

$$ec(s_i^j, e_k) = \frac{\alpha(i)P(a)P(s_i^j|e_k)\beta(j)}{\alpha(m)}$$

次に、M ステップは単純に正規化を行う。

$$P(f|e) = \frac{ec(f, e)}{\sum_f ec(f, e)}$$

e と **f** の二つの文が与えられたとき、最適な分割はビタビアルゴリズムを使用することによって得ることができる。

$$segments = argmin_{\mathbf{s}} \sum_{i}^{n} \left(-log \sum_{\mathbf{a}} P(s_i|e_a) + \theta \right)$$

ここで、$\mathbf{s} = \{s_1 s_2 ... s_n\}$ はソース文 **f** のセグメント集合で、**a** はソースのセグメントからターゲット単語へのアライメントである。

θ は、トークン数に対するペナルティで、トークン数が少なくなりすぎることを抑制するパラメータである。

3.2 モノリンガルモデル

モノリンガルモデルは以下の式で定義される。

$$P(f) = \sum_{e} P(f|e)P(e)$$

ここで、$P(f|e)$ は 3.1.1 節で計算したバイリンガルモデルの確率である。$P(e)$ は以下の式で計算したモノリンガルモデルの確率である。

$$P(e_i) = \frac{count(e_i)}{\sum_k^K count(e_k)}$$

ここで、$count(e_i)$ は訓練データの英語側の単語 e_i の出現数で、K は、ボキャブラリーのサイズである。

ここでいうモノリンガルモデルは、デコードの際に単言語を用いるという意味であり、ソース側言語の情報のみ用いるわけではない。上記の式が示すように、モデル化に当たってはバイリンガルモデルを活用している。

4 実験

教師なしバイリンガルトークナイザの有効性を検証するために、IWSLT2015 のベトナム語-英語タスクを用いて評価を行った。

4.1 実験条件

今回の実験では、IWSLT2015 の Evaluation Campaign で用いられた TED talk の訓練データと開発データを使用した。もともとのデータでは、各言語は分割されており、文頭の文字は大文字化されている。トレーニングデータ中で出現する回数の多数決により、文頭の単語を小文字または大文字に正規化して用いた。訓練データ中の 80 単語以上の文は捨てて、モデル学習を行っていた。

本実験の開発セットには IWSLT2010 のテストセットを、テストセットには IWSLT2011 と IWSLT2012 のテストセットを用いた。

我々は、翻訳ツールに Moses[3] を使用し、単語対応づけツールに GIZA++[4] を使用した。言語モデルは、kenLM[5] を用いて訓練した。

本実験で使用したシステムは、訓練セットのベトナム語側を教師ありおよび教師無しトークナイザで再分割し、その訓練セットを用いて、フレーズベースの翻訳モデルを学習する。

翻訳実験で用いるフレーズテーブル中の "_" は取り去り、元の表現に戻して翻訳実験を行っている。翻訳に用いる場合、バイリンガルトークナイザは、目的言語の情報が必要であるが、テスト文に対してそれが手に入らない。あらかじめフレーズテーブル中の表現をもとに戻すことによって、この問題に対処している。

θ は、英語とベトナム語のトークン数がほぼ等しくなるように開発セットを用いて設定した。

4.2 実験結果

実験結果を表 1 に示す。

表 1: 実験結果

	test2011	test2012
vnTokenizer	21.07	21.38
unsp-tok(bi)	19.91	19.77
unsp-tok(mono)	20.53	21.40

この表から明らかなように、本実験においては、バイリンガルモデル (unsp-tok(bi)) よりもモノリンガルモデル (unsp-tok(mono)) の方が若干よい翻訳結果が得られた。また、モノリンガルモデル (unsp-tok(mono)) は、単語辞書を用いていないにも関わらず、教師ありの手法 (vnTokenizer) とほぼ同等の翻訳精度を達成した。

5 まとめと今後の課題

ベトナム語の翻訳のためのトークナイザに、単語辞書を用いない手法を適用して、従来の単語辞書を用いる手法と同等の翻訳精度を達成した。今後、最適なパラメータチューニングにより、前者の手法を超えることができるのではないかと考えている。

今回の実験では、バイリンガル情報を用いない純粋なモノリンガルトークナイザとの比較が行えていない。これは今後の課題と考える。

謝辞

本研究のベトナム語の分析にあたっては Doan Thi Thuy Trinh 氏にご協力いただいた。ここに感謝いたします。

参考文献

[1] L. H. Phuong, N. Thi Minh Huyên, A. Roussanaly, and H. T. Vinh, "Language and automata theory and applications," C. Martín-Vide, F. Otto, and H. Fernau, Eds. Berlin, Heidelberg: Springer-Verlag, 2008, ch. A Hybrid Approach to Word Segmentation of Vietnamese Texts, pp. 240–249.

[2] T. Chung and D. Gildea, "Unsupervised tokenization for machine translation," in *In Proc. EMNLP 2009*, 2009.

[3] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open source toolkit for statistical machine translation," in *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ser. ACL '07. Stroudsburg, PA, USA:

Association for Computational Linguistics, 2007, pp. 177–180.

[4] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.

[5] K. Heafield, I. Pouzyrevsky, J. H. Clark, and P. Koehn, "Scalable modified Kneser-Ney language model estimation," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, August 2013, pp. 690–696.