The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

単語の分散表現を用いた Earth Mover's Distance と文ベクトルによる対訳コーパスの自動生成

Automatic Generation of Bilingual Corpus by Earth Mover's Distance and Sentence Vectors Using Word Embedding

田上 諒[*1] Ryo Tanoue
越前谷 博[*1] Hiroshi Echizen'ya
荒木 健治[*2] Kenji Araki

[*1] 北海学園大学大学院工学研究科
Graduate School of Engineering, Hokkai-Gakuen University

[*2] 北海道大学大学院情報科学研究科
Graduate School of Information Science and Technology, Hokkaido University

We propose a new method to automatically generate bilingual corpus using word embedding. The bilingual corpus is effective as the language resource for natural language processing. However, the generation of bilingual corpus by human needs high cost. In this paper, we propose a new method to automatically generate bilingual corpus using word embedding. Our proposed method uses two similarities: One is the similarity based on Earth Mover's Distance (EMD); another is the similarity based on the sentence vector. Through the evaluation experiments, we confirmed that the weighted mean between the similarity based on EMD and the similarity based on the sentence vector is effective to extract the correct bilingual sentence pair.

1. はじめに

対訳コーパスは，言語学や外国語教育などの言語分析の観点で大きな意義を持つことはもちろん，工学的な観点においても不可欠な言語リソースである．特に近年ではニューラル機械翻訳の進展に伴い，より大規模な対訳コーパスが求められている．しかし，人手により対訳コーパスを収集することはコストが大きな問題となる．そこで対訳コーパスを自動生成するための研究[内山 03]が従来より盛んに行われている．

対訳コーパスを自動生成する手法には，単語の文脈情報を利用した手法[Fung 98]や翻訳用辞書と単語の出現回数を利用した手法[Tamura 12]，そして，潜在的意味解析を利用した手法[Preiss 12, 江里口 14]などが挙げられる．しかし，これらの手法は単語の意味を十分に考慮したものとはなっていない．

そこで，本研究では，単語の分散表現に基づき対訳コーパスを自動生成する新たな手法を提案する．提案手法では，異言語間の文同士の対応関係を単語の分散表現に基づく類似度により決定する．その際，2 つの分布間の距離である Earth Mover's Distance(EMD)と文ベクトルに基づく距離の 2 つの距離を用いて類似度を求める．EMD [Yossi 00, 柳本 07]を用いるのは単語の分散表現を特徴量とすることで，単語の意味を考慮した文間の類似度を得ることができるためである．提案手法では対訳辞書などの高品質な対訳知識を用いることなく，単語の意味を考慮した対訳コーパスの自動生成が可能である．

性能評価実験では，日英 Wikipedia 京都関連文書対訳コーパス [1] にある 100 の対訳文を英語文と日本文に分離したうえで，提案手法により正しい対訳文をどの程度抽出できるかを評価した．その結果，EMD による類似度と文ベクトルによる類似度の加重平均を用いた提案手法に基づくシステムでは EMD による類似度のみと文ベクトルに基づく類似度のみのシステムよりも多くの正しい対訳文が得られた．

2. 提案手法

2.1 提案手法の概要

言語が異なる文を直接比較することはできないため，一方の言語の文の単語分散表現をもう一方の言語の文の単語分散表現に翻訳行列を用いてマッピングする．そして，2 つの分布間の距離を EMD を用いて求める．さらに，対応関係にある文間においては，構成単語数も近いとの仮定に基づき，単語の分散表現の平均同士でベクトル間類似度を求める．本稿ではこの単語の分散表現の平均を文ベクトルと呼ぶ．そして EMD と文ベクトルの 2 つの類似度の加重平均を最終的な類似度とする．

図1に英語から日本語へ向けた類似度計算の概要を示す．

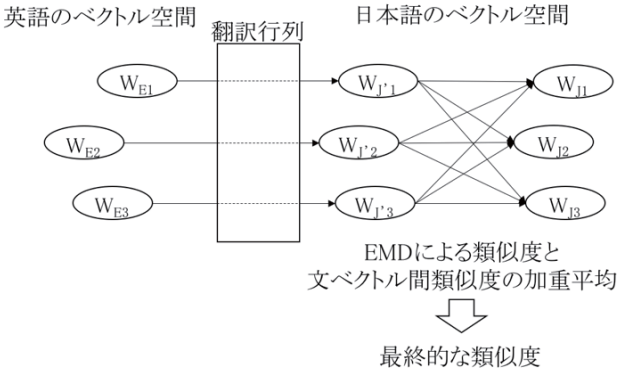

図1 英日方向における異言語文間の類似度計算

連絡先：田上 諒，北海学園大学大学院 工学研究科電子情報生命工学専攻，〒 064-0926 札幌市中央区南 26 条西 11-1-1，6717101r@hgu.jp

[1] https://alaginrc.nict.go.jp/WikiCorpus/

2.2 翻訳行列の生成

翻訳行列はある言語の単語ベクトル空間から他の言語の単語ベクトル空間へ任意の単語を写像するために用いる. 翻訳行列Wは訳語ペアの分散表現の集合$\{x_i, z_i\}_{i=1}^{n}$を用いて, 次の式(1)より得る.

$$\min_{W} \sum_{i=1}^{n} \|Wxi - zi\|^2 \quad (1)$$

2.3 Earth Mover's Distance による類似度計算

ある言語を他の言語のベクトル空間へマッピングした後, 同じベクトル空間上で EMD より類似度を求める. EMD は特徴量, 重み, 距離式を定義することで 2 つの分布間の距離を得ることができる. 2 分布P, Qの特徴量をそれぞれ$\{(p_1, w_{p_1}), \cdots, (p_m, w_{p_m})\}$と$\{(q_1, w_q), \cdots, (q_m, w_{q_m})\}$とする. 分布$P$は$m$個の特徴量で表現されており, p_iは特徴量, w_{pi}はその特徴量に対する重みである. 分布Qも同様である. また, p_iとq_jの距離をd_{ij}とし, 輸送量をf_{ij}と定義する. ${f_{ij}}^*$を最小化された輸送量とすると EMD は以下の式(2)より得られる.

$$\mathrm{EMD}(P, Q) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij}{f_{ij}}^*}{\sum_{i=1}^{m}\sum_{j=1}^{n} {f_{ij}}^*} \quad (2)$$

提案手法では特徴量として単語の分散表現, 重みとして$tfidf$[徳永 99], 距離式にはコサインを用いる. $tfidf$は任意のドキュメントにおける出現単語の重要度を表す尺度である. また, 距離dを計算する際に, 翻訳行列によりマッピングされた単語ベクトルと最も近い単語を近似単語とし, それらが他言語の文に存在するか否かの結果を反映させる. 近似単語が存在する場合には, 単語間の対応関係の信頼性は高いと考えられる. 距離dは以下の式(3)より得る.

$$d = \begin{cases} 1.0 - \cos(\text{近似単語あり}) \\ 1.0 - cos^{2.0}(\text{近似単語なし}) \end{cases} \quad (3)$$

式(3)より, 近似単語が存在しない場合には, コサインを 2 乗することで, 近似単語が存在する場合よりも距離が大きくなるように制御する. さらに$tfidf$の利用においては文中の各単語の$tfidf$の合計が 1.0 になるように正規化することで, EMD の値を 0.0 から 1.0 の範囲に収める. また, 値が大きいほど類似度が大きくなるように, 1.0 から EMD の値を引く.

2.4 文ベクトルによる類似度計算

文の長さが大きく異なる 2 文間において, 短い文の単語の多くが長い文の単語と意味に近い場合でも, 文としての類似度は低いと考えられる. そこで提案手法では, 文の長さの違いをより反映した類似度として, 文を構成する単語の分散表現の平均を文ベクトルとして文ベクトル間のコサインを求める. そして, EMD による類似度と文ベクトルによる類似度との間の加重平均を求め, それを最終的なスコアとする. 加重平均を用いるのは抽出精度に対して EMD による類似度と文ベクトルによる類似度の影響が等しいとは限らないためである.

2.5 対応関係にある文の決定

得られた類似度に基づき対応関係にある文同士を決定する. 英日方向の場合, 1 つの英語文とすべての日本語文との類似度を求め, 類似度の高い順にソートする. その結果, 上位 1 位にランキングされた日本文を基準として, 今度はその日本語文と類似度が最も高い英語文を求める. その結果, 双方向で類似度が最も高い場合, 対応関係にあるとして, 抽出する. 図 2 にその具体例を示す. 図 2 では英日方向で E1 と J1, 日英方向でも J1 と E1 が対応したと見なされたため, (E1:J1)が対応関係にある文として決定される.

2.2 以降の処理をさらに日英方向でも行う. そして, 双方向で一致した組み合わせを最終的な対応関係にある文同士とする. 例えば, 図 2 で使用された英語文と日本文において, 日英方向で(J1: E1)が得られた場合は(E1:J1)と一致するため, E1 と J1 は対応関係にある文として決定される.

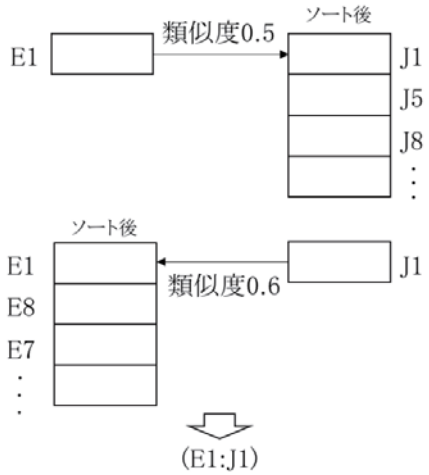

図 2 英日方向における対応関係にある文の決定

3. 性能評価実験

3.1 実験方法

性能評価実験を行うための実験データには Wikipedia 日英京都関連文書対訳コーパスに含まれているカテゴリ「仏教」から対訳文 100 文をランダムに抽出した. そして, 対訳文を英語文と日本語文に分離し, 提案手法により正しい対訳文をどの程度得られるのかを評価した.

提案手法を用いる際, 単語の分散表現は英語と日本語の Wikipedia ダンプデータから word2vec[Tomas 13 (a)]を用いて得た. ただし, サブサンプリングの閾値パラメタを 0.001 に設定することで, 機能語のような高頻度単語の影響を抑制する. 翻訳行列については, 文献[Tomas 13 (b), 石渡 16]により, 英日方向, 日英方向共に原言語の次元数は 800, マッピング先の目標言語の次元数は 200 で学習を行なった. また, 学習データに用いた対訳語ペアについては Wikipedia ダンプデータにおいて出現頻度が上位の単語と Google 翻訳を用いて得られた訳語のペアを用いた. 本研究では, 対訳語 10,000 ペアを翻訳行列の学習データとして用いた.

また, 提案手法の有効性を検証するために, EMD のみのシステムと文ベクトルのみのシステム, そして, 提案手法のシステムを用いた. ただし, EMD のみのシステムにおいては近似単語を用いていない.

3.2 実験結果

3.1 で述べたシステムの実験結果を表1に示す. 抽出精度は英日 100 文ずつから正しい対訳文を抽出できた割合を示す. また, 表 1 の提案手法は, EMD による類似度と文ベクトルによる類似度の重みが等しい相加平均によるシステムと文ベクトルによる類似度の重みを EMD による類似度に対して 2 倍とした加重平均によるシステムの結果とした.

表1 実験結果

手法	抽出精度
EMDのみ	19%
文ベクトルのみ	44%
提案手法(相加平均)	36%
提案手法(加重平均)	45%

3.3 考察

表1より提案手法における加重平均が最も高い抽出精度を示した. 加重平均では文ベクトルによる類似度の重みが EMD による類似度の重みの 2 倍であるため, 文ベクトルが効果的な役割を果たしていることがわかる. 文ベクトルのみの抽出精度が44%であったことからも文ベクトルの効果は大きいと考えられる. 提案手法の加重平均と文ベクトルのみで抽出された対訳文を比べると, 提案手法により新たに得られた対訳文は 16, 変化のなかった対訳文は 29 であった. 提案手法により新たに得られた対訳文の一つは英語文「by intellectual understanding, it is impossible to clap with one hand and make a sound.」と日本文「知的な理解では片手では拍手はできず音はしない。」からなる対訳文であった. 提案手法では, この対訳文においては英日方向, 日英方向のいずれも類似度が最も高かったため, 正しい対訳文として抽出された. それに対して, 文ベクトルのみでは, 英日方向では 100 番中 6 番に類似度が高く, 日英方向では 13 番目に類似度が高かったため, 正しい対訳文として抽出されなかった. したがって, 抽出精度としては提案手法の加重平均は文ベクトルに対してわずかに上回ったという結果ではあるが, 得られた対訳文の 3 割が異なるため, さらなる精度向上が期待できる.

4. おわりに

本稿では, 単語の分散表現を用いた対訳コーパスの自動生成のための新たな手法を提案した. 提案手法により EMD による類似度と文ベクトルによる類似度を組み合わせることの有効性を確認した. しかし, 抽出精度はまだ不十分である. 今後は, 2 つの類似度のそれぞれの利点をさらに引き出すことで抽出精度の向上を図る予定である.

参考文献

[内山 03] 内山将夫, 谷村緑: パラレルコーパスの自動生成技術, 情報通信研究機構季報. 53(3), 23-28 ,情報通信研究機構 (2007).

[Fung 98] Fung P. and Yee L. Y.: An IR approach for translating new words from nonparallel, comparable texts, COLING '98 Proceedings of the 17th international conference on Computational linguistics - Volume 1, 414-420 (1998).

[Tamura 12] Tamura A., Watanabe T. and Sumita E.: Bilingual lexicon extraction from comparable corpora using label propagation, EMNLP-CoNLL '12 Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 24-36 (2012).

[Preiss 12] Preiss J.: Identifying comparable corpora using LDA, NAACL HLT '12 Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 558-562 (2012).

[江里口 14] 江里口瑛子, 小林一郎: 潜在情報を利用したパラレルコーパス生成, 第 28 回人口知能学会全国大会論文集, 3I4-1, 人工知能学会 (2014).

[Yossi 00] Yossi Rubner, Carlo Tomasi, Leonidas J. Guibas: The Earth Mover's Distance as a Metric for Image Retrieval, International Journal of Computer Vision November 2000 Volume 40 Issue 2, 99–121 (2000).

[柳本 07] 柳本豪一, 大松繁: Earth Mover’s Distance を用いたテキスト分類, 第 21 回人口知能学会全国大会論文集, 1G3-4,人工知能学会 (2007).

[徳永 99] 徳永健伸:情報検索と言語処理, 東京大学出版会 (1999)

[Tomas 13 (a)] Tomas M., Kai C., Greg C. and Jeffrey D.: Efficient estimation of word representations in vector space, arXiv:1301.33781v3 (2013)

[Tomas 13 (b)] Tomas M., Quoc V. Le, and Ilya S.: Exploiting similarities among languages for machine translation, arXiv:1309.4168v1 (2013).

[石渡 16] 石渡祥之佑, 鍛治 伸裕, 吉永 直樹, 豊田 正史, 喜連川 優: 文脈語間の対訳関係を用いた単語の意味ベクトルの翻訳, 人工知能学会誌, 32(1), 20-29, 人工知能学会 (2016).