E-022

単語の分散表現を用いた異言語文間類似度に基づく最適訳選択
Selection of Best Machine Translation Outputs based on Cross-Lingual Similarity using Word Embeddings

藤川 寛基† 越前谷 博‡ 荒木 健治†
Hiroki Fujikawa Hiroshi Echizenya Kenji Araki

1. はじめに

昨今の機械翻訳システムの性能はニューラルネットワークの発展によって向上しているが，ニューラルネットワークから得られる結果に対し訳文の自動評価手法は追随できていない．特に小説のように抽象性が高い文書においては，表層的な情報だけでは正しい結果が得られない．また評価に用いる参照訳の質と量が十分でない場合にも精度は落ちてしまう．そこで本稿では意味的な情報を扱うことができる評価手法として，単語の意味を Word2Vec[1]による分散表現を用いて表し，参照訳を用いずに翻訳対象文と複数の訳候補文で異言語文間類似度を計算することで意味の観点から正しい訳を選択することができる新たな手法を提案する．

2. 提案手法概要

図 1 に提案手法の概要を示す．Word2Vec の分散表現を異言語間で比べる手法として Mikolov らの翻訳行列を用いた研究[2]がある．Mikolov らは翻訳行列を用いた線形変換により，異言語間で単語単位の類似度を比較した．本手法では異言語間の単語ベクトルの比較に翻訳行列を用いた．

また，文書の類似度を計算する手法として Earth Mover's Distance(EMD) を用いた Wan らの研究[3]が挙げられる．EMD は類似画像検索の分野で用いられている技術であり，2 つの分布間の距離を定義する．Wan らはそれを基にして文書の類似度を算出した．柳本の研究[4]では，Word2Vec によって得られた分散表現を特徴量として EMD で文書類似度を計算した結果，単語の分散表現による意味的な表現を反映させた文書類似度 が得られたとしている．しかし，EMD による文間の類似度は全ての単語間での類似度を基に求めているため，対応関係の無い単語間の計算を多く含むという問題がある．

したがって，本研究では英文を翻訳対象文，機械翻訳システムで得られた日本語訳を訳候補文とし，それぞれ単語分割し Word2Vec により分散表現を得る．英語単語ベクトルに対して翻訳行列を用いて日本語単語ベクトルの空間にマッピングしたのち，単語アライメントを行い単語の対応関係を獲得し，tf-idf 法による単語重要度を用いて，EMD による異言語文間の類似度に基づき最適訳を選択した． 以下では提案手法について詳細に述べる．

2.1 翻訳行列を用いた単語ベクトルのマッピング

ここでは異言語間の Word2Vec による単語ベクトルを比較するために用いる翻訳行列について説明する．単語ベクトルは単言語コーパスを用いて学習されているので，異言語間での比較ができない．そこで，ある単語ベクトルと同意味で対応する異言語ベクトルへの線形変換をなす行列 W を翻訳行列とよび，それを用いて異言語間の単語ベクトルを比較する．翻訳行列 W は対訳語ペア$\{x_i, z_i\}_{i=1}^{n}$を用いて最小二乗法により近似することで学習することができるため，以下の式(1)により表すことができる．

$$W = arg\min_{W} \sum_{i=1}^{n} ||Wx_i - z_i||^2 \quad (1)$$

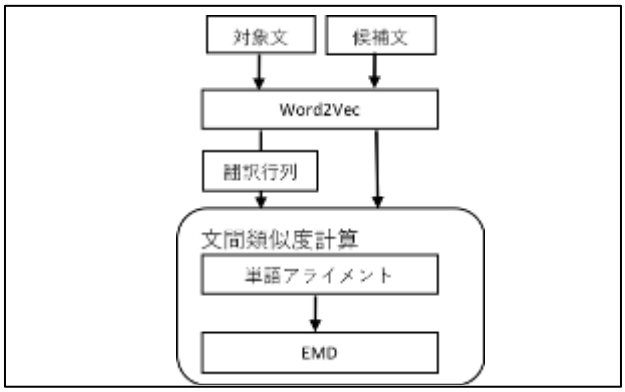

図 1　提案手法の概要

2.2 文間類似度計算

2.2.1 Earth Mover's Distance

本手法では文間の類似度の算出に EMD を用いている．EMD は 2 つの分布間の距離を最適化問題の一つである輸送問題に基づいて算出する手法である．それぞれの分布は特徴量と重みからなるシグネチャの集合から構成されており，m 個の集合からなる分布 P は P={(p_1,w_{p1}),・・・,(p_m,w_{pm})}で表せる．このとき，p_i は特徴量であり，w_{pi} は特徴量に対する重みである．同様に n 個の集合からなる分布 Q を Q={(q_1,w_{q1}),・・・,(q_n,w_{qn})}とした時，分布 P,Q 間で EMD によって最小化する目的関数は以下の式(2)により表せる．

$$WORK = \sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij}f_{ij} \quad (2)$$

ここで d_{ij} は特徴量間の距離であり，f_{ij} は p_i から q_j への輸送量である．d_{ij} は特徴量によって一様に決定できるため，f_{ij} を最小にした時に目的関数は最小となる．したがって，目的関数を最小にする最適な輸送量を f*とすると EMD によって定義される距離の式は以下の式(3)により表せる．

$$EMD(P,Q) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij}f^*_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} f^*_{ij}} \quad (3)$$

† 北海道大学, Hokkaido University
‡ 北海学園大学, Hokkai-Gakuen University

2.2.2 単語アライメント

本手法で用いる単語アライメントについて述べる．EMD は文書間の距離を求める際に総当りで単語間の距離を求めていることから，対応関係のない単語間の距離も含めて文書間の距離を算出している．そこで，単語アライメントによって対象文と訳候補文間での単語の対応関係を獲得し，対応関係の取れている単語のみを用いて類似度を算出した．単語アライメントには両単語ベクトル間のコサイン類似度と DICE 係数を用いてアライメントスコアを定義し，使用した．アライメントスコアがしきい値 t 以上の場合は単語の対応関係があるとみなし，しきい値 t 以下の場合は単語間の距離 d_{ij} を最大値の 1.0 とする．こうすることで，対応関係の存在する単語同士に基づく適切な類似度を求めることが可能となる．計算式を以下の式(4)~(6)に示す．

$$alignment\ score(x,y) = \frac{\cos(x,y) + DICE係数(x,y)}{2.0} \quad (4)$$

$$DICE係数(x,y) = \frac{2 * f_{xy}}{f_x + f_y} \quad (5)$$

$$d_{ij} = \begin{cases} 1.0(t < alignment\ score(x,y)のとき) \\ \cos(x,y)\ (alignment\ score(x,y) < tのとき) \end{cases} \quad (6)$$

ここで f_x は単語 x が対象文全文で出現した回数であり，f_y は単語 y が訳候補文全文で出現した回数である．また，f_{xy} は対応する対象文と候補文の組に単語 x と単語 y が同時に出現する回数である．

アライメントによって得られた式(6)の d_{ij} を式(3)に代入することで EMD の距離が得られる．式(3)で求められた距離を用いて EMD による類似度を以下の式(7)により定義する．(0<EMD<1)

$$類似度 = 1 - EMD(P,Q) \quad (7)$$

3. 性能評価実験

3.1 実験データ

翻訳対象文として川端康成の「古都」を原文とした「The old capital」を用いた．ここで小説を用いたのは，小説の翻訳は意訳を行う必要があるため難しいとされており，Word2Vec を用いた分散表現を評価するのに適当であると考えたためである．また，訳候補文には対象文を一文ずつ「Google 翻訳」[5]「Microsoft Translator」[6]「エキサイト翻訳」[7]の 3 つのサイトで翻訳したものを使用した．Word2Vec の学習コーパスには，英日それぞれの Wikipedia ダンプデータ[8]を用いた．本来学習コーパスは小説にすべきであるが，一貫性のある十分な量のデータを集めることができなかったため，小説を学習コーパスに使用することができなかった．コーパスの容量は，英語版 Wikipedia は約 1.3GB で日本語版 Wikipedia は約 850MB であった．翻訳行列に学習させた対訳語ペアは，英語版 Wikipedia の頻出単語上位 8,000 語をそれぞれ google 翻訳で翻訳し対訳語ペアを獲得した．

また，EMD の計算に用いる特徴量は Word2Vec によって得た単語ベクトルを使用し，重みには対象文全文と 3 つの訳候補文全文それぞれに対して tf-idf 法を文単位で対応させて獲得した単語重要度を使用した．アライメントスコアのしきい値 t は予備実験に基づき得られた値 0.75 を用いた．またアンケートによって人手で 3 つの訳候補文に 5 段階で点数をつけてもらい最適な訳の正解データを作成した．

3.2 実験方法

翻訳対象文の冒頭 100 文に対して，一文ずつ 3 つの候補文との類似度を算出した．得られた 3 つの類似度を比較し候補文の順位付けを行い，同様に人手評価で得られた候補文の評価点からも候補文を順位付けした．システムで得られた順位と人手評価で得られた順位を 1 文毎に比較し，その一致数を評価した．また，提案手法の比較対象としてアライメントを行わない手法による類似度，また文単位の自動評価法である METEOR[9]による評価からも順位付けを行い同様に一致数を評価した．

3.3 実験結果及び考察

表 1 に提案手法で得られた一致数，およびアライメントを行わないで得られた一致数，METEOR による一致数を示す．

表 1:実験結果

	一致数
提案手法	21
アライメントなし	11
METEOR	15

表 1 より，提案手法は単語アライメントなしの手法と自動評価法 METEOR に比べ高い一致数を確信した．これにより提案手法である単語の分散表現を用いた異言語文間類似度が小説の機械翻訳において有効であったと考えられる．METEOR による自動評価は参照訳との表層的な相関で求められる．よって表層的な情報だけでは判断できないような文書においても提案手法は有効であると考えられる．提案手法で獲得できなかった文では，アライメントで対応関係を取る際に 1 対多の対応関係や重複するアライメントにより内容語がアライメントできなかったものが 3 割ほどあった．これは，対応関係のない単語同士でもしきい値を上回ったことでアライメントが行われたためと考えられる．

4. おわりに

本稿では単語の分散表現を用いた異言語文間類似度に基づく最適訳選択手法を提案し，実験によって提案手法の有効性を確認した．今後は，EMD では語順に関係なく文間類似度を算出してしまうため，類似度計算を行う際に語順情報を反映させる方法を導入することや，アライメントの精度を上げることで文選択精度を向上させる予定である．

参考文献

[1] word2vec: Tool for computing continuous distributed representations of words https://code.google.com/p/word2vec/
[2] T Mikolov, QV LE, I Sutskever “Exploiting similarities among languages for machine translation”, arXiv preprint arXiv:1309.4168 (2013).
[3] X Wan, Y Peng “The earth mover’s distance as a semantic measure for document similarity”, Proceedings of the 14th ACM international conference on Information and knowledge management, pp.301~302 (2005)
[4] 柳本 豪一 “単語の分散表現を利用した文書類似度”, The 29th Annual Conference of the Japanese Society for Artificial Intelligence (2015)
[5] Google 翻訳 https://translate.google.co.jp
[6] Microsoft Translator https://translator.microsoft.com/neural
[7] エキサイト翻訳 http://www.excite.co.jp/world/
[8] Wikipedia Downloads https://dumps.wikimedia.org
[9] MDA Lavie “Meteor universal: Language specific translation evaluation for any target language”, ACL2014 (2014)