

Wikipediaにおける複数言語空間の相関について検討

ダオ ヴァン トゥアン 佐藤 浩
防衛大学校 知能情報研究室

1. 背景

近年、自然言語処理において、MikolovらのWord2Vec[2]に代表される、単語の分散表現を使って単語間の意味関係を表す研究が盛んとなっている。

現在、多くの研究では単一の言語データが中心であるが、複数言語のデータで単語空間を作ることは少ない。本研究では多数の言語データを持つWikipediaを用いて複数言語空間を生成することで各言語間の関係を示す。

2. 関連研究

多言語を扱う領域に、機械翻訳があるが、Melvinら[1]の研究では、中間言語を経由する、翻訳システムを提案している(図1)。

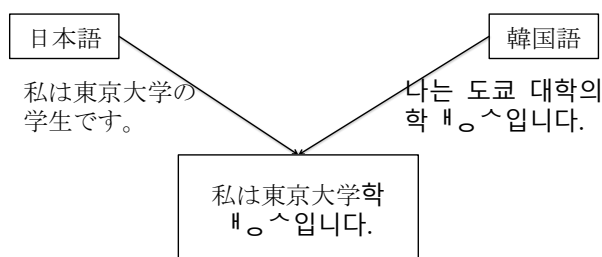


図1. 文書の組み合わせ

この方法は英語を介すものに比べ、翻訳時間を半分に減らせるものの、精度に関しては、従来の多言語モデルより翻訳品質が低い。

本研究では、単語ベクトル化手法 Word2Vec[2, 3] に複数言語データを同時に処理させることで、多言語間の相関を検討する。Word2Vec には、CBOW と Skip-gram の2つのモデルがあるが、本研究では、Skip-gram モデルを用いて実験を行う。

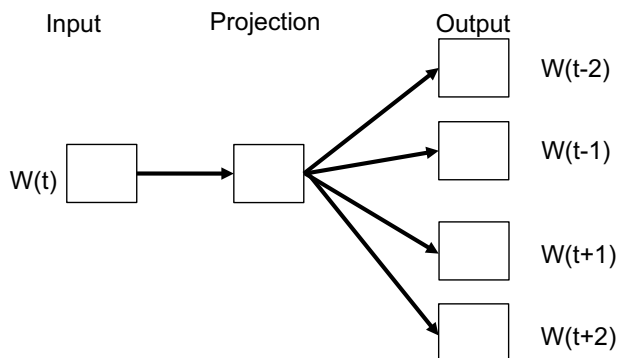


図2 skip-gramモデルのアーキテクチャ

Skip-gram モデルの式は以下に示す。

$$\frac{1}{T} \sum_{t=1}^T \sum_{-c \leq j \leq c, j \neq 0} \log p(w_{t+j} | w_t)$$

ここで、Tは学習データの単語数、cは文脈のサイズを示す。

3. 提案手法

複数言語を組み合わせた空間の作成を2つの方法で実施した：

【複数言語空間1】英語-日本語-ベトナム語の順で、全記事を結合した文書を学習する。

【複数言語空間2】各言語データの中から共通する記事を抽出したものを結合した文書を学習する。

言語データとして、Wikipediaダンプデータから、英語版(14.5GB)、日本語版(2.54GB)、およびベトナム語版(500MB)を用いた。また、Word2vecのパラメータは、先行研究[4]において良いとされたものを用いた (Min-count: 1, Window: 2, Size: 50)。

単語ベクトル空間は、学習データの質によって大きく反映される。

英語、ベトナム語などの場合ではストップワードといった言語処理における必要のない単語が単語ベクトル空間に影響を与えることがある。日本語の場合漢字、ひらがな、カタカナも存在する。本研究においては、前処理として、ストップワードを取り除き、必要単語として名詞、動詞、形容詞のみを扱うとする。

4. 実験

【実験1】空間ごとのベクトル距離の違い Word2vecの有名な事例である：

“king - man + woman = queen”

よりvector(king), vector(man), vector(woman)とvector(queen)のユークリッド距離関係を測った。各言語における結果を表1に示す。

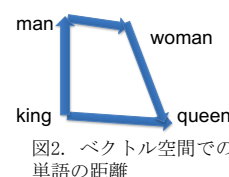


図2. ベクトル空間での単語の距離

表1. 単語距離の比較

	英語空間	日本語空間	ベトナム語空間	複数言語空間1	複数言語空間2
king-man	2.71	2.80	3.01	2.91	2.83
king-queen	2.58	2.72	2.85	2.82	2.64
man-woman	1.54	1.62	1.70	1.68	1.58
woman-queen	2.89	2.92	3.23	3.02	2.94

表1より、各言語の空間において、distance(king-queen)/distance(man-woman)が同じ値をとる。複数言語空間1, 2においても同様な結果が得られた。

【実験2】

言語の組み合わせの仕方が空間にどんな影響を与えるかを検討する。単語の関連度を示すコサイン類似度を比較した。

表2. 単語の関連性の比較

	複数言語空間1	複数言語空間2
(king - man + woman) vs (queen)	0.81	0.84
(王様 - 男 + 女) vs (女王)	0.76	0.78
(vua - dân ông + dân bà) vs (hoàng hậu)	0.80	0.81
(japan - tokyo + vietnam) vs (hanoi)	0.89	0.93
(日本 - 東京 + ベトナム) vs (ハノイ)	0.88	0.91
(nhật bản - tokyo + việt nam) vs (hà nội)	0.91	0.92

表2より、どちらの空間でも結果はほぼ等しい。複数言語空間2記事を選択することによって学習時間を短縮できた。

5. まとめと今後の課題

本研究では Word2vec を用いて、複数言語が混在するベクトル空間を構築した。記事ことを組み合わせた結果は、高い単語の関連性を確認でき、学習時間も短縮できた。今後の課題として、複数言語空間による、越日翻訳システムの実現を目指す。

6. 参考文献

- [1] Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat “Google’s Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation”, arXiv:1611.04558v2 [cs. CL] 21 Aug 2017.
- [2] Tomas Mikolov, Kai Chen, Grag Corrado, Jeffery Dean, “Efficient Estimation of Word Representations in Vector Space” Cornell University Library arXiv.org, arXiv:1301.3781v3[cs. CL], 2013.
- [3] Tomas Mikolov, T., Sutskever, I., Chen, K., Corrado G. and Dean, J.: “Distributed representations of words and phrases and their compositionality, Advances in Neural Information Processing Systems”, pp.3111-3119, 2013.
- [4] Dao V. Tuan, Hiroshi Sato “A proposal for search assistance method by Word2vec”, The 10th Vietnam-Japan Scientific Exchange Meeting VJSE September 2017.