多言語単語埋め込みのための文脈窓の分析

李 凌寒　　鶴岡 慶雅
東京大学 大学院情報理工学系研究科
{li0123,tsuruoka}@logos.t.u-tokyo.ac.jp

1 はじめに

複数言語で共通の意味空間を学習する多言語単語埋め込み（Cross-lingual Word Embedding）の分野において、独立に学習された複数の単語空間を共通の空間にマップする写像ベース (mapping-based) の手法が盛んに研究されている [1, 8]。写像ベースの手法は、異なる言語の単語埋め込み空間の構造が類似している、もしくは近傍グラフの構造が同型である [10] という強い仮定に基づいている。

単語埋め込み空間の構造は、訓練コーパスにおける単語の共起情報に強く依存している [2, 11]。実際のアルゴリズムでは、文内の単語について文脈窓 (context window) を定義し、その文脈窓にどのような単語が現れるかの情報によって、単語埋め込みが計算される。この文脈窓の選択は、単語埋め込み空間の構造の決定に大きな影響を与える。例えば、前後 1 ～ 2 単語を範囲とする狭い文脈窓で単語埋め込みを学習するとその埋め込み空間は単語の文法的な性質を捉え、広い文脈窓では単語のトピック的な性質を捉える傾向にあることが知られている [5]（表 1）。このように、文脈窓と単語埋め込み空間の構造は密接に関連しているにも関わらず、文脈窓の選択が多言語の単語埋め込み空間の構造的類似性、ひいてはマッピングの質にどのような影響を与えるかを調べた研究は少ない。

本研究では、文脈窓と多言語単語埋め込みの関係について理解を深めるために、異なる文脈窓サイズの単語埋め込みを複数言語で学習し、それらのマッピングの性能を測定する実験を行った。マッピングの性能は近傍から翻訳に当たる単語を取得する bilingual lexicon induction のタスクで評価した。その結果、全体的な傾向として、窓サイズを大きくすればするほど、2 つの埋め込み空間はマップしやすくなることが観察された。

単語	窓サイズ 1	窓サイズ 10
言語学	宗教学	比較言語学
	社会学	類型論
	発生学	記号論
	統計学	言語学者
	音声学	生成文法

表 1: 本研究で得られた日本語単語埋め込み空間における上位近傍単語。小さい窓サイズは機能的な類似性 (「-学」) を捉え、大きな窓サイズはトピック的な類似性を捉えている。

2 実験設定

2.1 単言語単語埋め込みの学習

実験に用いる言語は、言語資源の豊富さと語族の多様性を考慮して、ターゲット言語として英語 (En) 、ソース言語としてフランス語 (Fr)、ドイツ語 (De)、ロシア語 (Ru)、日本語 (Ja) を使用した。単語埋め込みを学習するためのコーパスには the Wikipedia Comparable Corpora[1] を用いた。Comparable コーパスを用いた理由は、各言語である程度のデータ量が確保でき、かつ多言語単語埋め込みが学習しやすい設定にすることで文脈窓の影響を強調できるためである。ソース言語には 100 万文、ターゲット言語には 500 万文用いた。

文脈窓の影響を調査するために、線形文脈窓のサイズを 1, 2, 3, 4, 5, 7, 10, 15, 20 の間で変化させた。線形文脈窓の他に、係り受け関係に基づいた文脈窓 [7] から学習した埋め込みについても調べたが、予備実験の結果、係り受け文脈窓の性能は常に線形窓サイズが小さいものと大きいものの中間に位置し、特徴的な傾向はみられなかった。従って、以下の分析では線形窓についてのみ論じる。

[1] https://linguatools.org/tools/corpora/wikipedia-comparable-corpora/

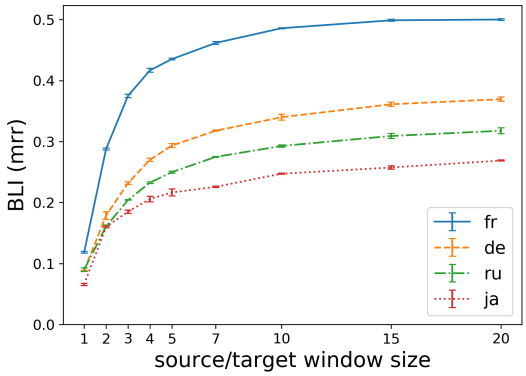

図 1: Comparable な設定における BLI スコア。

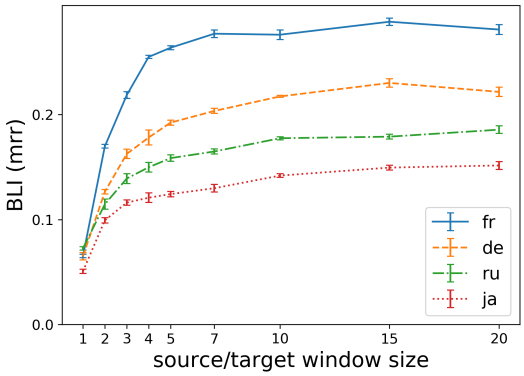

図 2: 異なるドメインの設定における BLI スコア。

単語埋め込みの計算には Skip-gram with Negative Sampling [9] の手法を用いたが、窓サイズの影響について論じる際は、その実装に注意しなくてはならない。オリジナルの C 言語実装 `Word2Vec`[2] や、python 実装である `Gensim`[3] は dynamic window の仕組みを採用しており、各トークンに対する実際の窓サイズは 1 から設定された窓サイズの間から一様にサンプリングされる。また、上に挙げた実装では高頻度のトークンを subsampling によって取り除くことで学習を効率化しているが、この高頻度トークンの除去は単語・文脈語ペアを抽出する前に行われため、実質文脈窓サイズを増やすこととなる（“dirty” subsampling と呼ばれる [6]）。こうした dynamic window と dirty sub-sampling の仕組みは、実質の文脈窓サイズを変化させるため、窓サイズが埋め込みに与える影響を曖昧にしてしまう可能性がある。従って、本実験では `word2vecf`[4] の実装を用いて実験を行う。`word2vecf` は入力として直接、単語・文脈語ペアを与えることが出来るので、固定した窓サイズから単語・文脈語ペアを抽出し、その後に sub-sampling を行った。

2.2 単言語埋め込みのマッピング

単言語埋め込みを学習した後は、写像ベースの手法で 2 言語の埋め込み空間を揃えた。空間を揃えるための線形行列 W は 2 言語の単語埋め込みと単語辞書から計算される。ここでは、単語辞書を $(x_i, y_i)_{i=1}^m$ として、最小化問題 $\arg\min_W \sum_{i=1}^m \|Wx_i - y_i\|^2$ を解く一般的な手法を用いた [8]。このとき、マッピングの品質を上げるために、W には直交行列になるような制約を課し、単語埋め込みにはマッピング前にノルム正規化と中心化の前処理を適用した [1]。

学習と評価に用いた単語辞書は Google Translate [5] から構築した。語彙内の単語を翻訳し、訓練データとして 5,000 対、評価データとして 2,000 対の単語翻訳ペアを各言語対でランダムに抽出した。

プログラムの乱数シードを変え、全ての設定について 3 つずつ多言語埋め込みを学習した。以下の結果では、それらの平均を標準偏差と共に示す。

3 実験結果

文脈窓サイズを変えて得られた多言語単語埋め込みを bilingual lexicon induction (BLI) のタスクで評価する。BLI は評価用辞書内の翻訳単語ペアについて、ソース言語の単語埋め込みから、多言語単埋め込み空間におけるコサイン類似度に基づく近傍探索によってターゲット単語を取得するタスクである。ここでは、評価尺度としては平均逆順位 (mean reciprocal rank; MRR) [6] を用いる。

ソース言語とターゲット言語の窓サイズを同じに揃え、どちらも変化させた時の設定の結果を図 1 に示す。ソース言語とターゲット言語の窓サイズを増やすことにより、BLI のスコアも一貫して向上することが観察される。大きい窓サイズは、単語のトピック的な性質を捉えることを考えると、これはよりトピック的な性質を捉えた単語埋め込みの方が異なる言語間でマップしやすいからだと考えられる。トピックは、言語固有の文法などと異なり、コーパスが comparable である限り言語普遍だと考えられる。従って、トピックをより捉えた単語埋め込みは異なる言語間でマップがしやすいのも妥当なことだと思われる。

[2] `https://code.google.com/archive/p/word2vec/`
[3] `https://radimrehurek.com/gensim/`
[4] `https://bitbucket.org/yoavgo/word2vecf/src/default/`

[5] https://translate.google.com/ (October 2019)
[6] N 単語ペアと、それらの順位 $rank_i$ が与えられたとき、平均逆順位スコアは $\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\text{rank}_i}$ となる。

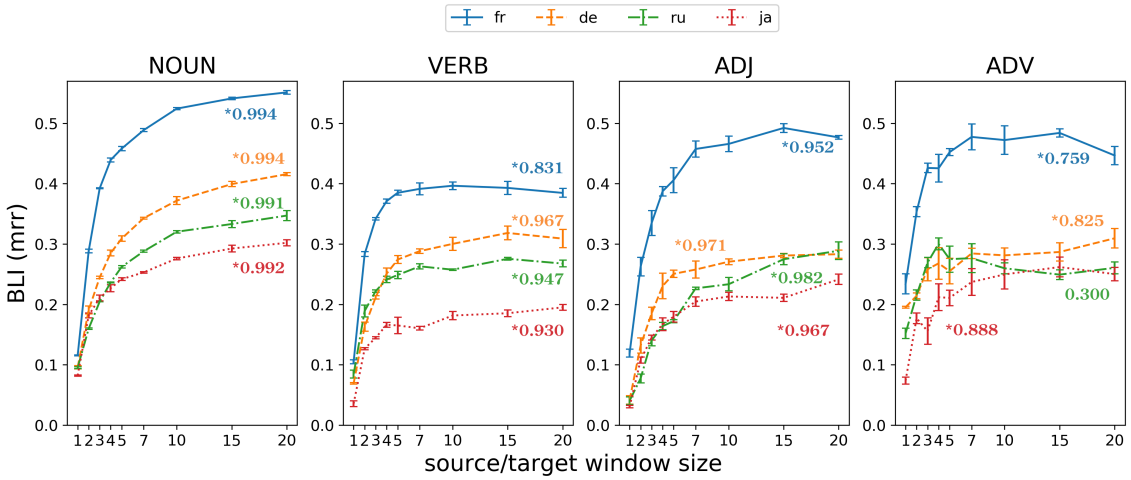

図 3: Comparable な設定における、各品詞の BLI スコア。グラフ上の数値はスコアと窓サイズのスピアマンの順位相関係数。統計的に有意な相関はアスタリスクで示される ($p < 0.05$)。

このトピック的な単語埋め込みはマップがしやすいという仮説は、単語の品詞毎の BLI の結果からも支持されると考えられる。直感的に、名詞は他の品詞に比べてトピックをよく表すことが多いと考えられるため、よりトピック的な性質を持った埋め込みが与えられているはずである。従って、名詞の BLI のスコアは単語埋め込みの窓サイズと特に強い相関を示すことが期待される。図 3 に各品詞のスコアとスピアマンの相関係数を示す[7]。全ての言語において、実際に名詞が相関係数 0.99 以上と最も強い相関を示している。

異なるドメインの設定. これまでの結果は、ソース言語とターゲット言語のコーパスが comparable であるという比較的理想的な条件下で行った実験の結果であった。コーパスが comparable であるとき、2 つのコーパスは同じトピックを持つので、トピック的な単語埋め込みがマップしやすいのは当然だと思われる。この傾向が、異なるドメインのコーパスを用いた時にも見られるかどうかを調べるために、ソース言語のコーパスに異なるドメイン（ニュース）のコーパス[8]を用いた結果を図 2 に示す。

まず、comparable の設定（図 1）に比べて、全体的に 0.1 〜 0.2 ポイント低いスコアを示している。これは、先行研究の、ドメインが一致していることが埋め込み空間の類似性に重要であるという観察 [10] と合致する。

次に、BLI の性能と窓サイズの関係については、窓

[7]各単語の品詞は、和田ら [12] にならい、Brown Corpus におけるその単語の最も頻度の高い品詞を用いた。

[8]https://wortschatz.uni-leipzig.de/en/download

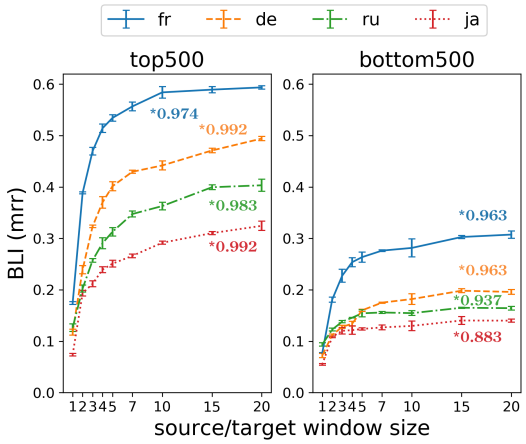

図 4: Comparable な設定における高頻度語と低頻度語の BLI スコア。

サイズを大きくすると BLI のスコアも上がるという、comparable の設定と同じ傾向が観察される。このことは、たとえソース言語とターゲット言語でコーパスのドメインが異なっていても、窓サイズを大きくすることで単語埋め込みはドメイン普遍なトピックを捉えマッピングがしやすくなるということを示唆している。

単語頻度による分析. 窓サイズを広げることでどのような単語のマップがしやすくなるのかについて、さらなる知見を得るために、評価用の単語辞書から頻度が上位 500 の単語と下位 500 の単語を抜き出して、それらについてのスコアを評価した。Comparable の設定における結果を図 4 に示す。

高頻度語 (top500) のスコアが低頻度語 (bottom500) のものより低いのは、既存のマッピングの

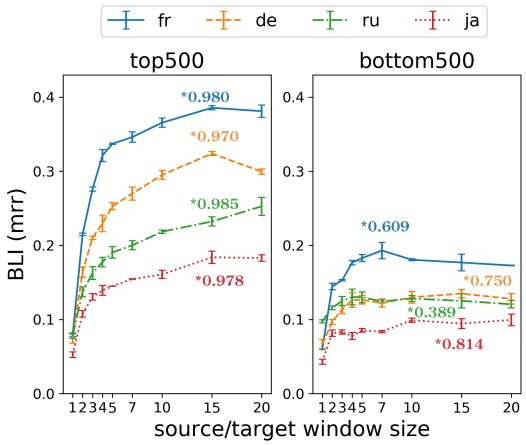

図 5: 異なるドメインの設定における高頻度語と低頻度語の BLI スコア。

手法は低頻度語に弱いという、先行研究の観察と合致する [3, 4]。

窓サイズとの関係については、高頻度語と低頻度語のどちらも大きい窓サイズにするにつれてスコアが上がっているが、日本語 (Ja) とロシア語 (Ru) については上昇の傾向は弱い。一方、ドメインが異なるコーパスでの結果（図 5）からは、低頻度語は、特にフランス語 (Fr) とロシア語 (Ru) で顕著であるが、窓サイズを大きくするとスコアが下がることが観察される。高頻度語は一貫して、窓サイズに比例してスコアが上昇している。これは、ドメイン違いのコーパスでは単語埋め込みを学習する時に、高頻度語は多くの訓練事例（単語・文脈語ペア）と関連づけられることで、大きい窓サイズを広げても意味上関係の無い文脈語（ノイズ）の影響を受けにくく、一方で低頻度語は限られた訓練事例の中で窓サイズを広げてしまうと、ノイズやドメインの違いが増幅され、結果マップがしにくくなるからだと考えられる。

4 おわりに

単語埋め込みを学習する際の文脈窓と、２つの埋め込み空間の構造的類似性との関係は、明らかに大きな影響を与えうることが予想されるにも関わらず、既存研究では十分に調査されていなかった。本研究は、文脈窓サイズと多言語単語埋め込みの BLI の性能との関係についての実験結果を提供し、多言語単語埋め込みの性質について新たな知見を加えた。

参考文献

[1] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of EMNLP*, 2016.

[2] Marco Baroni, Georgiana Dinu, and Germán Kruszewski. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of ACL*, 2014.

[3] Fabienne Braune, Viktor Hangya, Tobias Eder, and Alexander Fraser. Evaluating bilingual word embeddings on the long tail. In *Proceedings of the 2018 Conference of NAACL-HLT*, 2018.

[4] Paula Czarnowska, Sebastian Ruder, Edouard Grave, Ryan Cotterell, and Ann Copestake. Don't Forget the Long Tail! A Comprehensive Analysis of Morphological Generalization in Bilingual Lexicon Induction. In *Proceedings of EMNLP-IJCNLP*, 2019.

[5] Omer Levy and Yoav Goldberg. Dependency-Based Word Embeddings. In *Proceedings of ACL*, 2014.

[6] Omer Levy, Yoav Goldberg, and Ido Dagan. Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Transactions of the Association for Computational Linguistics*, Vol. 3, p. 211–225, 2015.

[7] Bofang Li, Tao Liu, Zhe Zhao, Buzhou Tang, Aleksandr Drozd, Anna Rogers, and Xiaoyong Du. Investigating Different Syntactic Context Types and Context Representations for Learning Word Embeddings. In *Proceedings of EMNLP*, 2017.

[8] Tomas Mikolov, Quoc V Le, and Ilya Sutskever. Exploiting Similarities among Languages for Machine Translation, 2013.

[9] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*. 2013.

[10] Anders Søgaard, Sebastian Ruder, and Ivan Vulić. On the Limitations of Unsupervised Bilingual Dictionary Induction. In *Proceedings of ACL*, 2018.

[11] Peter D. Turney and Patrick Pantel. From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, Vol. 37, No. 1, pp. 141–188, 2010.

[12] Takashi Wada, Tomoharu Iwata, and Yuji Matsumoto. Unsupervised multilingual word embedding with limited resources using neural language models. In *Proceedings of ACL*, 2019.