# Revisiting the Context Window for Cross-lingual Word Embeddings

**Ryokan Ri and Yoshimasa Tsuruoka**
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan
{li0123,tsuruoka}@logos.t.u-tokyo.ac.jp

## Abstract

Existing approaches to mapping-based cross-lingual word embeddings are based on the assumption that the source and target embedding spaces are structurally similar. The structures of embedding spaces largely depend on the co-occurrence statistics of each word, which the choice of context window determines. Despite this obvious connection between the context window and mapping-based cross-lingual embeddings, their relationship has been underexplored in prior work. In this work, we provide a thorough evaluation, in various languages, domains, and tasks, of bilingual embeddings trained with different context windows. The highlight of our findings is that increasing the size of both the source and target window sizes improves the performance of bilingual lexicon induction, especially the performance on frequent nouns.

## 1 Introduction

Cross-lingual word embeddings can capture word semantics invariant among multiple languages, and facilitate cross-lingual transfer for low-resource languages (Ruder et al., 2019). Recent research has focused on *mapping-based* methods, which find a linear transformation from the source to target embedding spaces (Mikolov et al., 2013b; Artetxe et al., 2016; Lample et al., 2018). Learning a linear transformation is based on a strong assumption that the two embedding spaces are structurally similar or isometric.

The structure of word embeddings heavily depends on the co-occurrence information of words (Turney and Pantel, 2010; Baroni et al., 2014), *i.e.*, word embeddings are computed by counting other words that appear in a specific context window of each word. The choice of context window changes the co-occurrence statistics of words and thus is crucial to determine the structure of an embedding space. For example, it has been known that an embedding space trained with a smaller linear window captures functional similarities, while a larger window captures topical similarities (Levy and Goldberg, 2014a). Despite this important relationship between the choice of context window and the structure of embedding space, how the choice of context window affects the structural similarity of two embedding spaces has not been fully explored yet.

In this paper, we attempt to deepen the understanding of cross-lingual word embeddings from the perspective of the choice of the context window through carefully designed experiments. We experiment with a variety of settings, with different domains and languages. We train monolingual word embeddings varying the context window sizes, align them with a mapping-based method, and then evaluate them with both intrinsic and downstream cross-lingual transfer tasks. Our research questions and the summary of the findings are as follows:

**RQ1: What kind of context windows produces a better alignment of two embedding spaces?** Our result shows that increasing the window sizes of both the source and target embeddings improves the accuracy of bilingual dictionary induction consistently regardless of the domains of the source and target corpora. Our fine-grained analysis reveals that frequent nouns receive the most benefit from larger context sizes.

**RQ2. In downstream cross-lingual transfer, do the context windows that perform well on the source language also perform well on the target languages?** No. We find that even when some context window performs well on the source language task, that is often not the best choice for the target language. The general tendency is that broader context windows produce better performance for the target languages.



*Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 995–1005
July 5 - 10, 2020. 

## 2 Background and Related Work

### 2.1 Context Window of Word Embeddings

Word embeddings are computed from the co-occurrence information of words, *i.e.*, context words that appear around a given word. The embedding algorithm used in this work is the skip-gram with negative sampling (Mikolov et al., 2013c). In the skip-gram model, each word $w$ in the vocabulary $W$ is associated with a word vector $v_w$ and a context vector $c_w$.[1] The objective is to maximize the dot-product $v_{w_t} \cdot c_{w_c}$ for the observed word-context pairs $(w_t, w_c)$, and to minimize the dot-product for negative examples.

The most common type of context is a linear window. When the window size is set to $k$, the context words of a target word $w_t$ in a sentence $[w_1, w_2, ..., w_t, ...w_L]$ are $[w_{t-k}, ..., w_{t-1}, w_{t+1}, ..., w_{t+k}]$. The choice of context is crucial to the resulting embeddings as it will change the co-occurrence statistics associated with each target word. Table 1 demonstrates the effect of the context window size on the nearest neighbor structure of embedding space; with a small window size, the resulting embeddings capture functional similarity, while with a larger window size, the embeddings capture topical similarities.

Among the other types of context windows that have been explored by researchers are linear windows enriched with positional information (Levy and Goldberg, 2014b; Ling et al., 2015a; Li et al., 2017), syntactically informed context windows based on dependency trees (Levy and Goldberg, 2014a; Li et al., 2017), and one that dynamically weights the surrounding words with the attention mechanism (Ling et al., 2015b). In this paper, we mainly discuss the most common linear window and investigate how the choice of the window size affects the isomorphism of two embedding spaces and the performance of cross-lingual transfer.

### 2.2 Cross-lingual Word Embeddings

Cross-lingual word embeddings aim to learn a shared semantic space in multiple languages. One promising solution is to jointly train the source and target embedding, so-called *joint methods*, by exploiting cross-lingual supervision signals in the form of word dictionaries (Duong et al., 2016), parallel corpora (Gouws et al., 2015; Luong et al., 2015), document-aligned corpora (Vulic and Moens, 2016).

| Query word | window size 1 | window size 10 |
|---|---|---|
| words | phrases | word |
| | loanwords | phrases |
| | morphemes | phrase |
| | verses | ungrammatical |
| | phonemes | homographs |
| typological | synchronic | totemism |
| | mechanistic | typology |
| | numerological | categorizations |
| | architectonic | dialectology |
| | dialectical | fusional |

Table 1: The top-5 nearest neighbors in English embedding spaces trained with different context windows in our experiment. The smaller window size captures functional similarities (*-s, -cal, -ic*), while the larger captures topical similarities.

Another line of research is off-line mapping-based approaches (Ruder et al., 2019), where monolingual embeddings are independently trained in multiple languages, and a post-hoc alignment matrix is learned to align the embedding spaces with a seed word dictionary (Mikolov et al., 2013b; Xing et al., 2015; Artetxe et al., 2016), with only a little supervision such as identical strings or numerals (Artetxe et al., 2017; Smith et al., 2017), or even in a completely unsupervised manner (Lample et al., 2018; Artetxe et al., 2018). Mapping-based approaches have recently been popularized by their cheaper computational cost compared to joint approaches, as they can make use of pre-trained monolingual word embeddings.

The assumption behind the mapping-based methods is the isomorphism of monolingual embedding spaces, *i.e.*, the embedding spaces are structurally similar, or the nearest neighbor graphs from the different languages are approximately isomorphic (Søgaard et al., 2018). Considering that the structures of the monolingual embedding spaces are closely related to the choice of the context window, it is natural to expect that the context window has a considerable impact on the performance of mapping-based bilingual word embeddings.

However, most existing work has not provided empirical results on the effect of the context window on cross-lingual embeddings, as their focus is

[1] Conceptually, the word and context vocabularies are regarded as separated, but for simplicity, we assume that they share the vocabulary.

on how to learn a mapping between the two embedding spaces. In order to shed light on the effect of the context window on cross-lingual embeddings, we trained cross-lingual embeddings with different context windows, and carefully analyzed the implications of their varying performance on both intrinsic and extrinsic tasks.

## 3 Experimental Design

### 3.1 Training Monolingual Embeddings

The experiment is designed to deal with multiple settings to fully understand the effect of the context window.

**Languages.** As the target language, we choose English (En) because of its richness of resources, and as the source languages, we choose French (Fr), German (De), Russian (Ru), Japanese (Ja), taking into account the typological variety and availability of evaluation resource.

Note that the language pairs analyzed in this paper are limited to those including English, and there is a possibility that some results may not generalize to other language pairs.

**Corpus for Training Word Embeddings.** To train the monolingual embeddings, we use the Wikipedia Comparable Corpora[2]. We choose comparable corpora for the main analysis in order to accentuate the effect of context window by setting an ideal situation for training cross-lingual embeddings.

We also experiment with different domain settings, where we use corpora from the news domain[3] for the source languages, because the isomorphism assumption is shown to be very sensitive to the domains of the source and target corpora (Søgaard et al., 2018). We refer to those results when we are interested in whether the same trend with respect to context window can be observed in the different domain settings.

For the size of the data, to simulate the setting of transferring from a low-resource language to a high-resource language, we use 5M sentences for the target language (English), and 1M sentences for the source languages.[4]

**Context Window.** Since we want to measure the effect of the context window size, we vary the window size among 1, 2, 3, 4, 5, 7, 10, 15, and 20.

Besides the linear window, we also experimented with the unbound dependency context (Li et al., 2017), where we extract context words that are the head, modifiers, and siblings in a dependency tree. Our initial motivation was that, while the linear context is directly affected by different word orders, the dependency context can mitigate the effect of language differences, and thus may produce better cross-lingual embeddings. However, the performance of the dependency context turned out to be always in the middle between smaller and larger linear windows, and we found nothing notable. Therefore, the following analysis only focuses on the results of the linear context window.

**Implementation of Word2Vec.** Note that some common existing implementations of the skip-gram may obfuscate the effect of the window size. The original C implementation of `word2vec` and its python implementation `Gensim`[5] adopt a dynamic window mechanism where the window size is uniformly sampled between 1 and the specified window size for each target word (Mikolov et al., 2013a). Also, those implementations remove frequent tokens by subsampling *before* extracting word-context pairs (so-called "dirty" subsampling) (Levy et al., 2015), which enlarges the context size in effect. Our experiment is based on `word2vecf`,[6] which takes arbitrary word-context pairs as input. We extract word-context pairs from a fixed window size and afterward perform subsampling.

We train 300-dimensional embeddings. For details on the hyperparameters, we refer the readers to Appendix A.

### 3.2 Aligning Monolingual Embeddings

After training monolingual embeddings in the source and target languages, we align them with a mapping-based algorithm. To induce a alignment matrix $W$ for the source and target embeddings $x, y$, we use a simple supervised method of solving the Procrustes problem $\arg\min_W \sum_{i=1}^{m} \|Wx_i - y_i\|^2$, with a training word dictionary $(x_i, y_i)_{i=1}^{m}$ (Mikolov et al.,

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

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

[4] We also experimented with very low-resource settings, where the source corpus size is set to 100K, but the results showed similar trends to the 1M setting, and thus we only include the result of the 1M settings in this paper.

[5] https://radimrehurek.com/gensim/

[6] https://bitbucket.org/yoavgo/word2vecf/src/default/

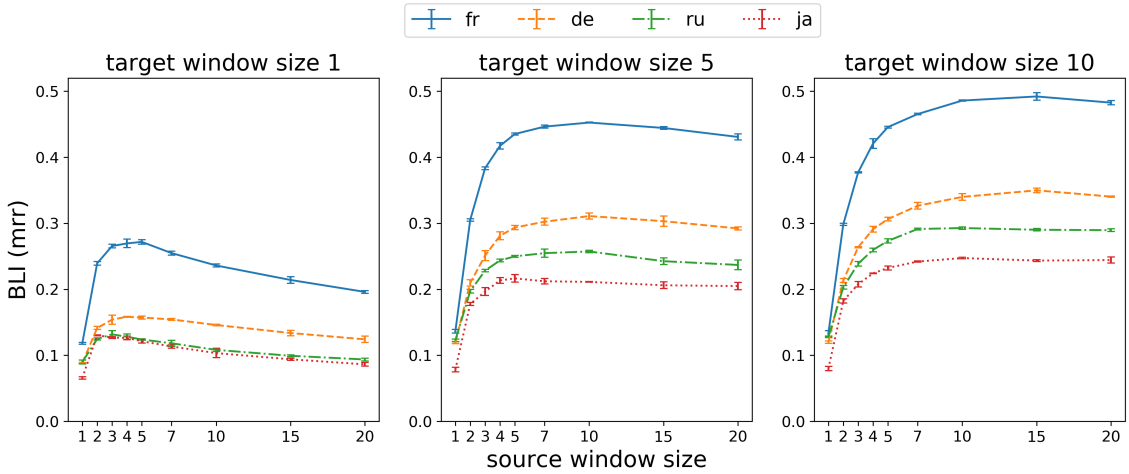


Figure 1: BLI performance in the comparable setting. The target window size is fixed and the source window size is varied.

2013b), with the orthogonality constraint on $W$, length normalization and mean-centering as preprocessing for the source and target embeddings (Artetxe et al., 2016).

The word dictionaries are automatically created by using Google Translate. [7] We translate all words in our English vocabulary into the source languages and filter out words that do not exist in the source vocabularies. We also perform this process in the opposite direction (translated from the source languages into English), and take the union of the two corresponding dictionaries. We then randomly select 5K tuples for training and 2K for testing. Although using word dictionaries automatically derived from a system is currently a common practice in this field, it should be acknowledged that this may sometimes pose problems: the generated dictionaries are noisy, and the definition of word translation is unclear (*e.g.,* how do we handle polysemy?). It can hinder valid comparisons between systems or detailed analysis of them, and should be addressed in future research.

For each setting, we train three pairs of aligned embeddings with different random seeds in the monolingual embedding training, as training word embeddings is known to be unstable and different runs result in different nearest neighbors (Wendlandt et al., 2018). The following results are presented with their averages and standard deviations.

[7] https://translate.google.com/ (October 2019)

## 4 Bilingual Lexicon Induction

We first evaluate the learned bilingual embeddings with bilingual lexicon induction (BLI). The task is to retrieve the target translations with source words by searching for nearest neighbors with cosine similarity in the bilingual embedding space. The evaluation metric used in prior work is usually top-k precision, but here we use a more informative measure, mean reciprocal rank (MRR) as recommended by Glavaš et al. (2019).

**Fixed Target Context Window Settings.** First, we consider the settings where the target context size is fixed, and the source context size is configurable. This setting assumes common situations where the embedding of the target language is available in the form of pre-trained embeddings.

Figure 1 shows the result of the four languages. Firstly, we observe that too small windows (1 to 3) for source embeddings do not yield good performance, probably because the model failed to train accurate word embedding models with insufficient training word-context pairs that the small windows capture.

At first, this result may seem to contradict with the result from Søgaard et al. (2018). They trained English and Spanish embeddings with `fasttext` (Bojanowski et al., 2017) and the window size of 2, and then aligned them with an unsupervised mapping algorithm (Lample et al., 2018). When they changed the window size of the Spanish embedding to 10, they only observed a very slight drop on top-1 precision (from 81.89 to 81.28). We suspect that the discrepancy with our result is due to the different settings. First of

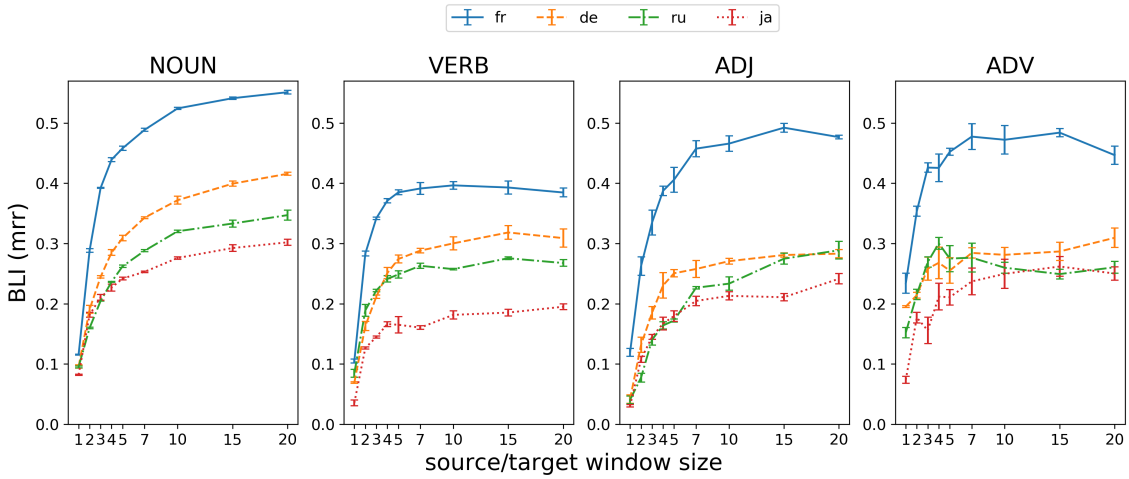


Figure 2: BLI performance for each PoS in the comparable setting.

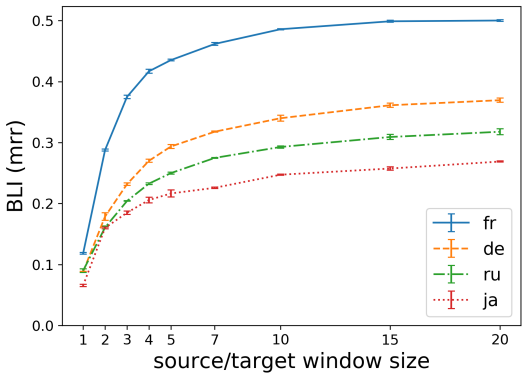


Figure 3: BLI performance in the comparable setting.

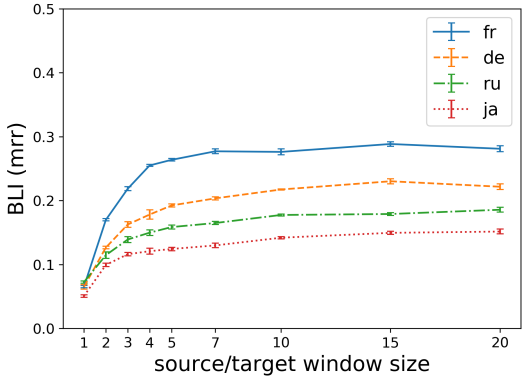


Figure 4: BLI performance in the different domain setting.

all, `fasttext` adopts a dynamic window mechanism, which may obfuscate the difference in the context window. Also, they trained embeddings with full Wikipedia articles, which is an order of magnitude larger than ours; the `fasttext` algorithm, which takes into account the character n-gram information of words, can exploit a non-trivial amount of subword overlap between the quite similar languages.

Overall, we observe that the best context window size for the source embeddings increases as the target context size increases, and increasing the context sizes of both the source and target embedding seems beneficial to the BLI performance.

**Configurable Source/Target Context Window Settings.** Hereafter, we present the results where both the source and target sizes are configurable and set to the same. Figure 3 summarizes the result of the same domain setting.

As we expected from the observation of the settings where the target window size is fixed, the performance consistently improves as the source and target context sizes increase. Given that the larger context windows tend to capture topical similarities of words, we hypothesize that the more topical the embeddings are, the easier they are to be aligned. Topics are invariant across different languages to some extent as long as the corpora are comparable. It is natural to think that topic-oriented embeddings capture language-agnostic semantics of words and thus are easier to be aligned among different languages.

This hypothesis can be further supported by looking at the metrics of each part-of-speech (PoS). Intuitively, nouns tend to be more representative of topics than other PoS, and thus are expected to show a high correlation with the window size. Figure 2 shows the scores for each PoS. [8] In all languages, nouns and adjectives show stronger (almost perfect) correlation than verbs and adverbs.

[8] We assigned to each word its most frequent PoS tag in the Brown Corpus (Kucera and Francis, 1967), following Wada et al. (2019).

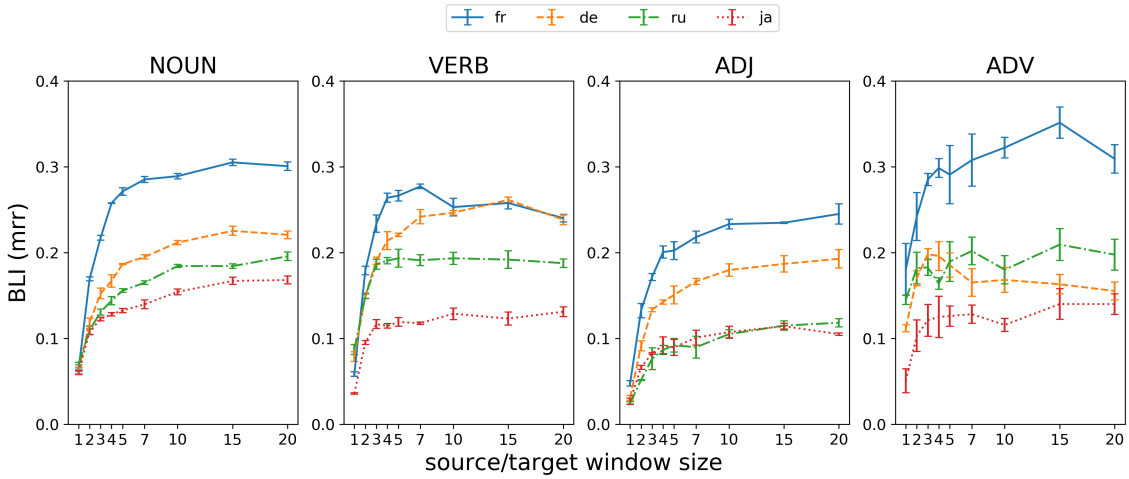


Figure 5: BLI performance for each PoS in the different domain setting.

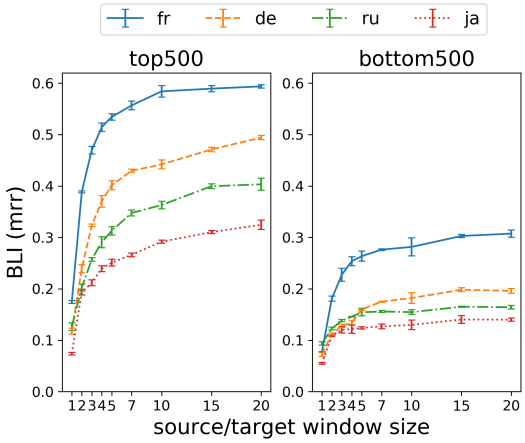


Figure 6: BLI performance with the top 500 frequent and rare words in the comparable setting.

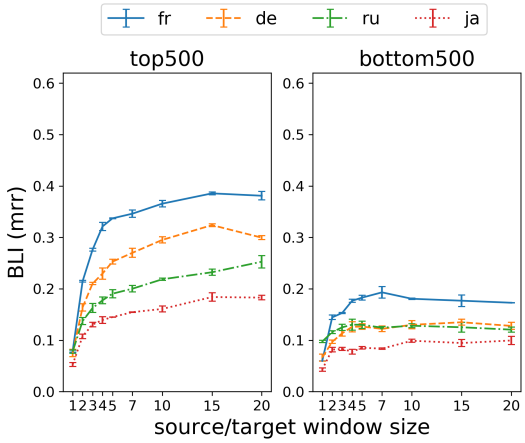


Figure 7: BLI performance on the top 500 frequent and rare words in the different domain setting.

**Different-domain Settings.** The results so far are obtained in the settings where the source and target corpora are comparable. When the corpora are comparable, it is natural that topical embeddings are easier to be aligned as comparable corpora share their topics. In order to see if the observations from the comparable settings hold true for different-domain settings, we also present the result from the different-domain (news) source corpora in Figure 4.

Firstly, compared to the same-domain settings (Figure 3), the scores are lower by around 0.1 to 0.2 points across the languages and context windows, even with the same amount of training data. This result confirms previous findings showing that domain consistency is important to the isomorphism assumption (Søgaard et al., 2018).

As to the relation between the BLI performance and the context window, we observe a similar trend to the comparable settings: increasing the context window size basically improves the performance. Figure 5 summarizes the results for each PoS. The performance on nouns and adjectives still accounts for much of the correlation with the window size. This suggests that even when the source and target domains are different, some domain-invariant topics are captured by larger-context embeddings for nouns and adjectives.

**Frequency Analysis.** To further gain insight into what kind of words receive the benefit of larger context windows, we analyze the effect of word frequency. We extract the top and bottom 500 frequent words[9] from the test vocabularies and evaluate the performance on them respectively.

The results of the comparable setting in each language are shown in Figure 6.

[9]The frequencies were calculated from our subset of the English Wikipedia corpus.

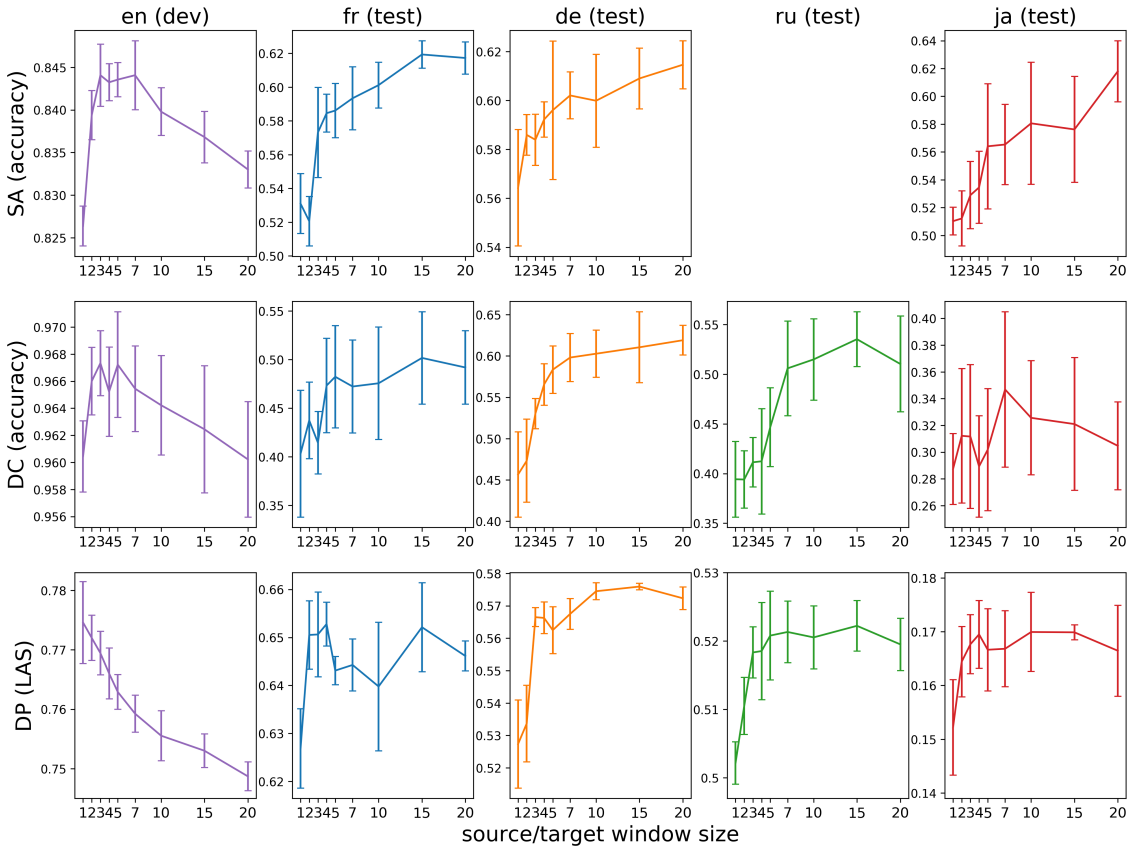


Figure 8: Downstream evaluations in the comparable settings. SA: sentiment analysis; DC: document classification; DP: dependency parsing. The window sizes of both the source and target embeddings are varied.

The scores for the frequent words (top500) are notably higher than the rare words (bottom500). This confirms previous empirical results that existing mapping-based methods perform significantly worse for rare words (Braune et al., 2018; Czarnowska et al., 2019).

With respect to the relation with the context size, both frequent and rare words benefit from larger window sizes, although the gain in the rare words is less obvious in some languages (Ja and Ru).

In the different domain settings, as shown in Figure 7, the rare words, in turn, suffer from larger window sizes, especially for Fr and Ru, but the performance on frequent words still improves as the context window increases.

We conjecture that when training a skip-gram model, frequent words observe many context words, and that would mitigate the effect of irrelevant words (noise) caused by a larger window size and result in high-quality topical embeddings; however, rare words have to rely on a limited number of context words, and larger windows just amplify the noise and domain difference to result in an inaccurate alignment of them.

## 5 Downstream Tasks

Although BLI is a common evaluation method for bilingual embeddings, good performance on BLI does not necessarily generalize to downstream tasks (Glavaš et al., 2019). To further gain insight into the effect of the context size on bilingual embeddings, we evaluate the embeddings with three downstream tasks: 1) sentiment analysis; 2) document classification; 3) dependency parsing. Here, we briefly describe the dataset and model used for each task.

**Sentiment Analysis (SA).** We use the Webis-CLS-10 corpus[10] (Prettenhofer and Stein, 2010), which is comprised of Amazon product reviews in the four languages: English, German, French, and Japanese (no Russian data available). We cast sentiment analysis as a binary classification task, where we label reviews with the scores of 1 or 2 as `negative` and reviews with 4 or 5 as `positive`. For the model, we employ a simple CNN encoder followed by a multi-layer perceptrons classifier.

[10] https://webis.de/data/webis-cls-10.html

**Document Classification (DC).** MLDoc[11] (Schwenk and Li, 2018) is compiled from the Reuters corpus for eight languages including all the languages used in this paper. The task is a four-way classification of the news article topics: `Corporate/Industrial`, `Economics`, `Government/Social`, and `Markets`. We use the same model architecture as sentiment analysis.

**Dependency Parsing (DP).** We train deep biaffine parsers (Dozat and Manning, 2017) with the UD English EWT dataset[12] (Silveira et al., 2014). We use the PUD treebanks[13] as test data.

The hyperparameters used in this experiment are shown in Appendix B.

**Evaluation Setup.** We evaluate in a cross-lingual transfer setup how well the bilingual embeddings trained with different context windows transfer lexical knowledge across languages. Here, we focus on the settings where both the source and target context sizes are varied.

For each task, we train models with our pre-trained English embeddings. We do not update the parameters of the embedding during training. Then, we evaluate the model with the test data in other languages available in the dataset. At test time, we feed the model with the word embeddings of the test language aligned to the training English embeddings.

We train nine models in total for each setting with different random seeds and English embeddings, and we present their average scores and standard deviations.

**Result and Discussion.** The results from all the three tasks are presented in Figure 8. For sentiment analysis and document classification, we observe a similar trend where the best window size is around 3 to 5 for the source English task, but for the test languages, larger context windows achieve better results. The only deviation is the Japanese document classification, where the score does not show a significant correlation. We attribute this to low-quality alignments due to the large typological difference between English and Japanese.

For dependency parsing, embeddings with smaller context windows perform better in the source English task, which is consistent with the observation that smaller context windows tend to produce syntax-oriented embeddings (Levy and Goldberg, 2014a). However, the performance of the small-window embeddings does not transfer to the test languages. The best context window for the English development data (the size of 1) performs the worst for all the test languages, and the transferred accuracy seems to benefit from larger context sizes, although it does not always correlate with the window size. This observation highlights the difficulty of transferring syntactic knowledge across languages. Word embeddings trained with small windows capture more grammatical aspects of words in each language, which, as different languages have different grammars, makes the source and target embedding spaces so different that it is difficult to align them.

In summary, a general trend we observe here is that good context windows in the source language task do not necessarily produce good transferrable bilingual embeddings. In practice, it seems better to choose a context window that aligns the source and target well, rather than using the window size that just performs the best for the source language.

## 6 Conclusion and Future Work

Despite their obvious connection, the relation between the choice of context window and the structural similarity of two embedding spaces has not been fully investigated in prior work. In this study, we have offered the first thorough empirical results on the relation between the context window size and bilingual embeddings, and shed new light on the property of bilingual embeddings. In summary, we have shown that:

- larger context windows for both the source and target facilitate the alignment of words, especially nouns.
- for cross-lingual transfer, the best context window for the source task is often not the best for test languages. Especially for dependency parsing, the smallest context size produces the best result for the source task, but performs the worst for test languages.

We hope that our study will provide insights into ways to improve cross-lingual embeddings by not only mapping methods but also the properties of monolingual embedding spaces.

[11] https://github.com/facebookresearch/MLDoc

[12] https://universaldependencies.org/treebanks/en_ewt/index.html

[13] https://universaldependencies.org/conll17/

## Acknowledgement

We thank the anonymous reviewers for their valuable comments and suggestions. This work was supported by JST CREST Grant Number JPMJCR1513, Japan.

## References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2294, Austin, Texas. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Vancouver, Canada. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia. Association for Computational Linguistics.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247, Baltimore, Maryland. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Fabienne Braune, Viktor Hangya, Tobias Eder, and Alexander Fraser. 2018. Evaluating bilingual word embeddings on the long tail. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 188–193, New Orleans, Louisiana. Association for Computational Linguistics.

Paula Czarnowska, Sebastian Ruder, Edouard Grave, Ryan Cotterell, and Ann Copestake. 2019. Don't Forget the Long Tail! A Comprehensive Analysis of Morphological Generalization in Bilingual Lexicon Induction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 974–983, Hong Kong, China. Association for Computational Linguistics.

Timothy Dozat and Christopher D Manning. 2017. Deep Biafine Attention for Neural Dependency Parsing. In *Proceedings of the International Conference on Learning Representations*.

Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. 2016. Learning Crosslingual Word Embeddings without Bilingual Corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1285–1295, Austin, Texas. Association for Computational Linguistics.

Goran Glavaš, Robert Litschko, Sebastian Ruder, and Ivan Vulić. 2019. How to (Properly) Evaluate Cross-Lingual Word Embeddings: On Strong Baselines, Comparative Analyses, and Some Misconceptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 710–721, Florence, Italy. Association for Computational Linguistics.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. BilBOWA: Fast Bilingual Distributed Representations without Word Alignments. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 748–756, Lille, France.

Henry. Kucera and W. Nelson. Francis. 1967. *Computational analysis of present-day American English*. Brown University Press.

Guillaume Lample, Alexis Conneau, Marc Aurelio Ranzato, Ludovic Denoyer, and Herve Jegou. 2018. Word Translation without Parallel Data. In *Proceedings of the International Conference on Learning Representations*.

Omer Levy and Yoav Goldberg. 2014a. Dependency-Based Word Embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308, Baltimore, Maryland. Association for Computational Linguistics.

Omer Levy and Yoav Goldberg. 2014b. Linguistic Regularities in Sparse and Explicit Word Representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 171–180, Ann Arbor, Michigan. Association for Computational Linguistics.

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

Bofang Li, Tao Liu, Zhe Zhao, Buzhou Tang, Aleksandr Drozd, Anna Rogers, and Xiaoyong Du. 2017. Investigating Different Syntactic Context Types and Context Representations for Learning Word Embeddings. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2421–2431, Copenhagen, Denmark. Association for Computational Linguistics.

Wang Ling, Chris Dyer, Alan W. Black, and Isabel Trancoso. 2015a. Two/Too Simple Adaptations of Word2Vec for Syntax Problems. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1299–1304, Denver, Colorado. Association for Computational Linguistics.

Wang Ling, Yulia Tsvetkov, Silvio Amir, Ramón Fermandez, Chris Dyer, Alan W Black, Isabel Trancoso, and Chu-Cheng Lin. 2015b. Not All Contexts Are Created Equal: Better Word Representations with Variable Attention. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1367–1372, Lisbon, Portugal. Association for Computational Linguistics.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Bilingual Word Representations with Monolingual Quality in Mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 151–159, Denver, Colorado. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient Estimation of Word Representations in Vector Space. In *Proceedings of the International Conference on Learning Representations*.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013b. Exploiting Similarities among Languages for Machine Translation. *Computing Research Repository*, arXiv:1309.4168.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013c. 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*, pages 3111–3119. Curran Associates, Inc.

Peter Prettenhofer and Benno Stein. 2010. Cross-Language Text Classification Using Structural Correspondence Learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1118–1127, Uppsala, Sweden. Association for Computational Linguistics.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A Survey of Cross-lingual Word Embedding Models. *Journal of Artificial Intelligence Research*, 65(1):569–630.

Holger Schwenk and Xian Li. 2018. A Corpus for Multilingual Document Classification in Eight Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Chris Manning. 2014. A Gold Standard Dependency Corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2897–2904, Reykjavik, Iceland. European Language Resources Association (ELRA).

Samuel L Smith, David H P Turban, Steven Hamblin, and Nils Y Hammerla. 2017. Offline Bilingual Word Vectors, Orthogonal Transformations and the Inverted Softmax. In *Proceedings of the International Conference on Learning Representations*, pages 1–10.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the Limitations of Unsupervised Bilingual Dictionary Induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788, Melbourne, Australia. Association for Computational Linguistics.

Peter D. Turney and Patrick Pantel. 2010. From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188.

Ivan Vulic and Marie-Francine Moens. 2016. Bilingual Distributed Word Representations from Document-aligned Comparable Data. *Journal of Artificial Intelligence Research*, 55(1):953–994.

Takashi Wada, Tomoharu Iwata, and Yuji Matsumoto. 2019. Unsupervised multilingual word embedding with limited resources using neural language models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3113–3124, Florence, Italy. Association for Computational Linguistics.

Laura Wendlandt, Jonathan K. Kummerfeld, and Rada Mihalcea. 2018. Factors influencing the surprising instability of word embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2092–2102, New Orleans, Louisiana. Association for Computational Linguistics.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized Word Embedding and Orthogonal Transform for Bilingual Word Translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011, Denver, Colorado. Association for Computational Linguistics.

## A The hyperparameters for training monolingual word embeddings

| hyperparameter | Source embeddings (1M sentences) | Target embeddings (5M sentences) |
|---|---|---|
| embedding size | 300 | |
| number of negative samples | 15 | |
| alpha (learning rate) | 0.025 (linearly decayed during training) | |
| minimum word count | 10 | 15 |
| number of iterations | 10 | 5 |

## B The hyperparameters for downstream tasks

### B.1 Document Classification and Sentiment Analysis

| hyperparameters | | |
|---|---|---|
| CNN Classifier | number of filters | 100 |
| | ngram_filter_sizes | 2, 3, 4, 5 |
| | MLP hidden size | 64 |
| Training | optimizer | Adam |
| | learning rate | 0.001 |
| | lr scheduler | halved each time the dev score stops improving |
| | patience | 3 |
| | batch size | 64 |

### B.2 Dependency Parsing

| hyperparameters | | |
|---|---|---|
| Graph-based Parser | LSTM hidden size | 200 |
| | LSTM number of layers | 3 |
| | tag representation dim | 100 |
| | arc representation dim | 500 |
| | pos tag embedding dim | 50 |
| Training | optimizer | Adam |
| | learning rate | 0.001 |
| | lr scheduler | halved each time the dev score stops improving |
| | patience | 3 |
| | batch size | 32 |