Data Augmentation with Unsupervised Machine Translation Improves the Structural Similarity of Cross-lingual Word Embeddings

Sosuke Nishikawa, Ryokan Ri and Yoshimasa Tsuruoka
The University of Tokyo, Japan

Outline

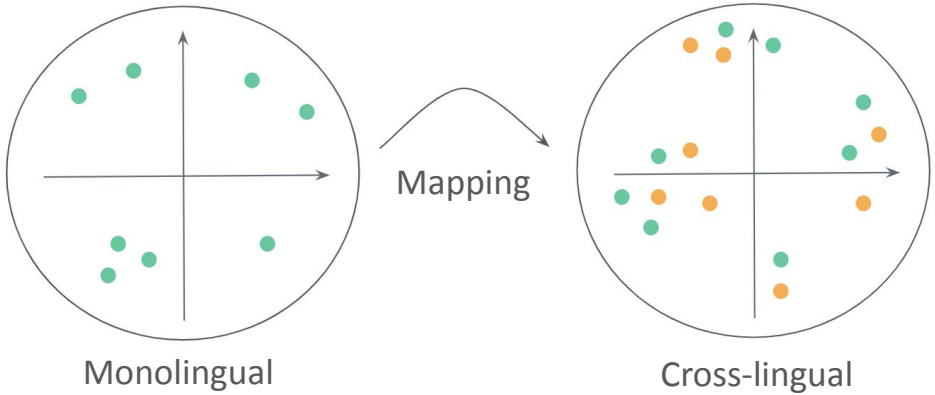
- Background & Motivation
- Proposed method
- Experiments
- Analysis
- Conclusion

Outline

- **Background & Motivation**
- Proposed method
- Experiments
- Analysis
- Conclusion

Cross-lingual Word Embedding

Unsupervised Cross-lingual word embedding (CLWE) methods learn a semantic space shared between languages without any cross-lingual supervision



Isomorphism assumption:

"The two embeddings are structurally similar"

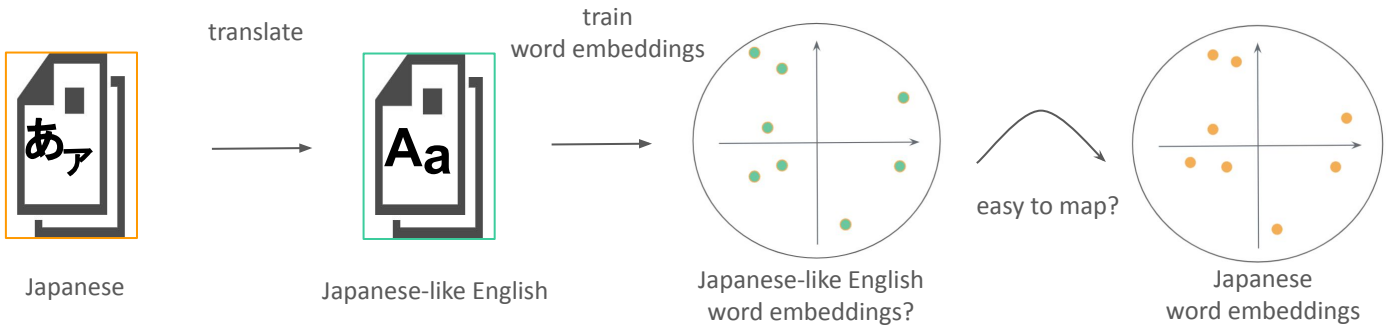
Limitation of Mapping Methods

Isomorphism assumption does not hold true when the two corpora are from **different domains** or the two languages are **typologically very different** [Søgaard et al., 2018]

➢ needs for improving the structural similarity of the two word embeddings before mapping

Hypothesis

Word embeddings trained using translated sentences have a similar structure to the word embedding space in the original language?



Unsupervised Machine Translation

Unsupervised machine translation (UMT) is a machine translation system which does not require any translation resources [Artetxe et al., 2018]

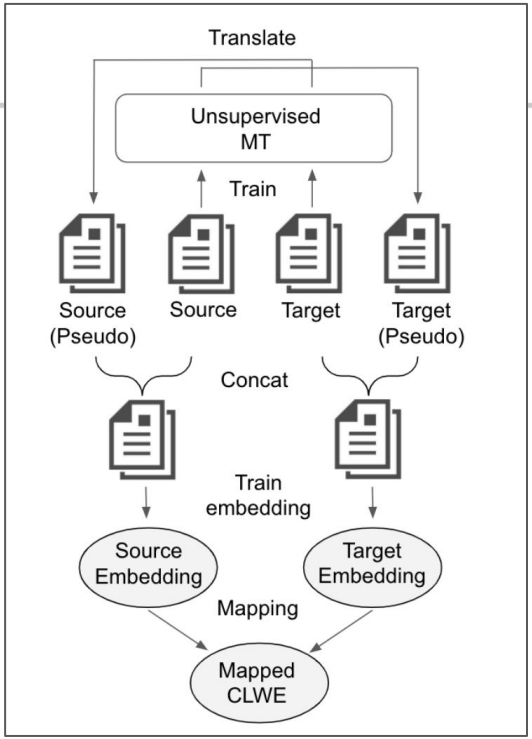
Learn word embeddings using translated sentences by UMT to improve structural similarity

Outline

- Background & Motivation
- **Proposed method**
- Experiments
- Analysis
- Conclusion

Method

- train UMT using the source/target training corpora and translate them
- learn word embeddings independently using machine-translated corpora (pseudo corpora)
- map them to a shared CLWE space



Our framework for training CLWEs

Outline

- Background & Motivation
- Proposed method
- **Experiments**
- Analysis
- Conclusion

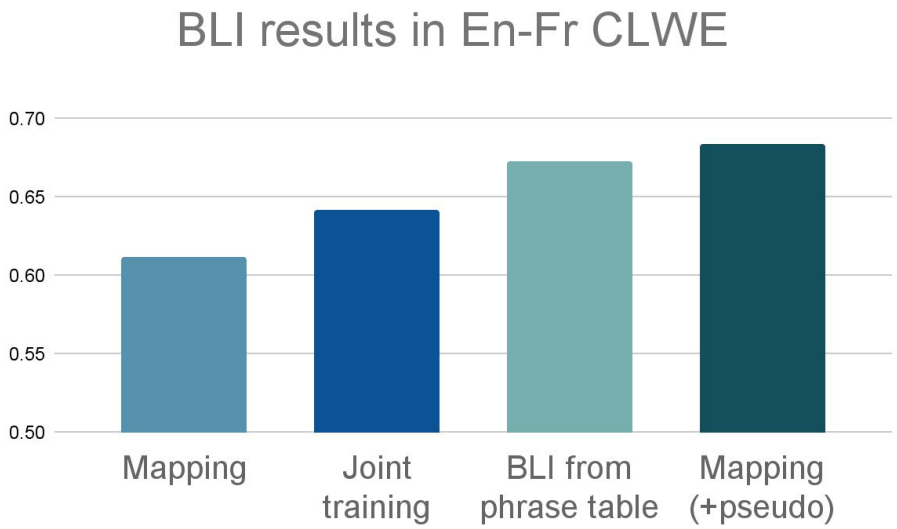
Experimental Settings

Evaluation task:	Bilingual lexicon induction (BLI)
Language pairs:	English-French, English-German, English-Japanese
Data:	10M sentences from Wikipedia dumps for each language
UMT system:	Phrase-based statistical UMT [Lample et al., 2018]
Word embedding method:	fastText [Bojanowski et al., 2017]
Word mapping method:	VecMap [Artetxe et al., 2018]

Baseline Methods

- CLWE using training corpora only (Mapping)
- BLI using a phrase table built with synthetic parallel corpora from UMT (BLI from phrase table) [Artetxe et al., 2019]
- CLWE trained using a bilingual skip-gram algorithm with a synthetic parallel corpus from UMT (Joint-training) [Marie and Fujita, 2019]

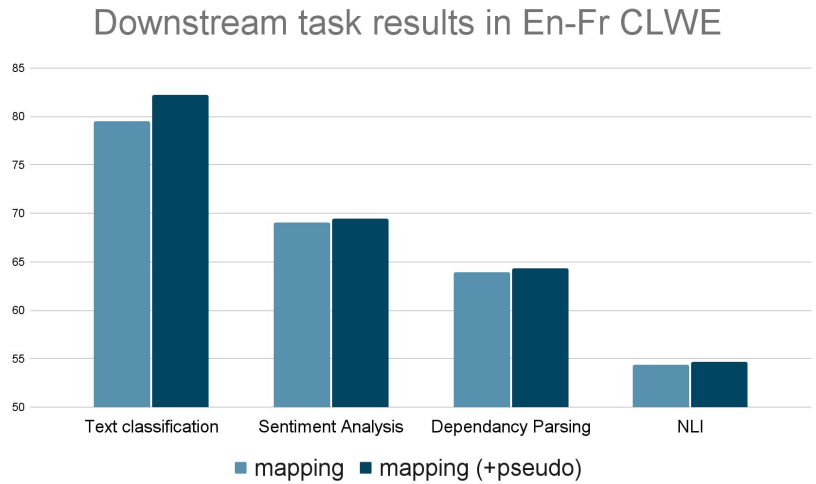
BLI Results



Mapping (+pseudo) clearly outperformed other alternative approaches

Downstream tasks

Evaluate our method in four downstream tasks



Mapping (+pseudo) consistently outperformed baseline mapping mehod

➢ Why did this method work well?

Outline

- Background & Motivation
- Proposed method
- Experiments
- **Analysis**
- Conclusion

Why did this method work?

It is not simply because of data augmentation but because:

- It bridges the domain gap between texts in two languages
- It mitigates linguistic differences between texts in two languages

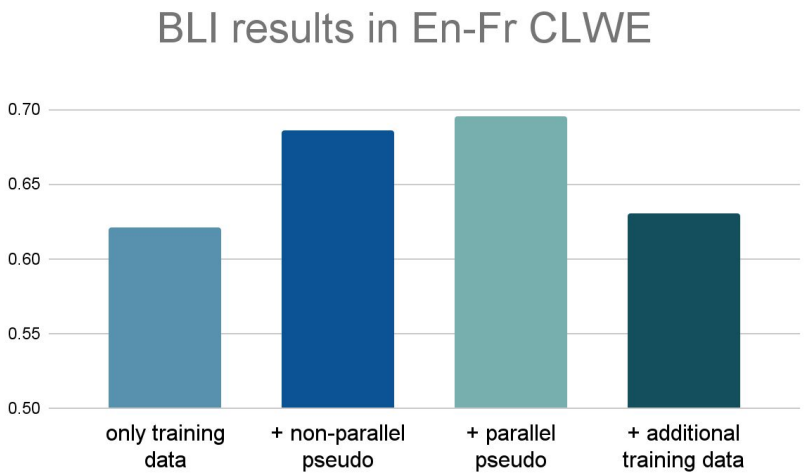
Why did this method work?

It is not simply because of data augmentation but because:

- **It bridges the domain gap between texts in two languages**
- It mitigates linguistic differences between texts in two languages

Bridging domain gap

Compare the extensions with a non-parallel pseudo, parallel pseudo and training data



Extension by parallel-pseudo corpora yielded the best BLI score

➢ Parallel-pseudo corpus makes the domains similar, and thus improves cross-lingual mapping

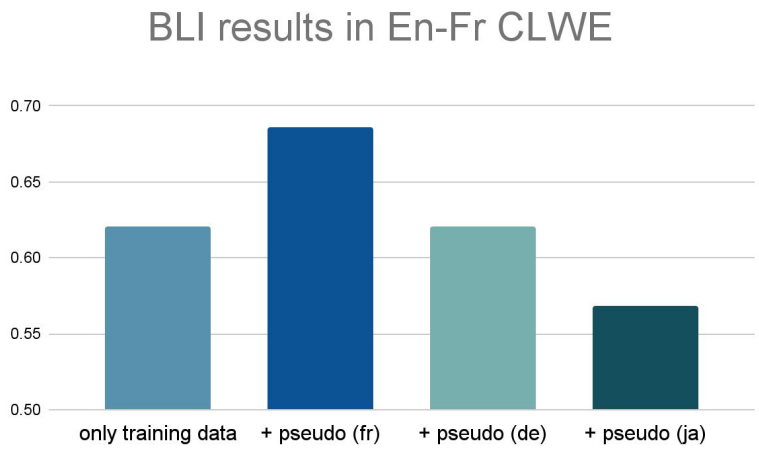
Why did this method work?

It is not simply because of data augmentation but because:

- It bridges the domain gap between texts in two languages
- **It mitigates linguistic differences between texts in two languages**

Effect of Language Pairs

Compare with pseudo corpora from non-target languages



Extension from the corresponding language (i.e. Fr) corpora specifically improved the BLI score

➢ The pseudo corpus makes documents linguistically similar,
and thus improves cross-lingual mapping

Outline

- Background & Motivation
- Proposed method
- Experiments
- Analysis
- **Conclusion**

Conclusion

- Proposed a method to learn word embeddings using translated sentences from UMT to improve the mapping for CLWEs.
- The proposed method outperformed the existing methods in the BLI task and Downstream tasks.
- The proposed method works by bridging the domain gap and mitigating linguistic differences between languages.

Thank you for listening