MUNI FI

Evaluation of the Cross-lingual Embedding Models from the Lexicographic Perspective

Michaela Denisová, Pavel Rychlý

449884@mail.muni.cz, pary@fi.muni.cz

Natural Language Processing Centre

Faculty of Informatics, Masaryk University

June 27-29, 2023

Content

1. Word Vectors
2. Cross-lingual Embedding Models
3. Cross-lingual Embedding Models and Lexicography
4. Current Drawbacks
5. Aims
6. Experimental Setup
7. Evaluation Datasets
8. Parameters:
   - Vocabulary
   - Inflected Word Forms
   - Part of Speech
   - Senses
9. Conclusion

Word Vectors

- Mathematical representation of the word
- Computed from the corpus where the algorithm computes how often the words occur next to each other
- Occurrences as represented by an array of numbers
- Monolingual word vectors are close to each other

Cross-lingual Embedding Models

- Bilingual or multilingual vector representations of words that are projected into shared space
- Similar words obtain similar vectors

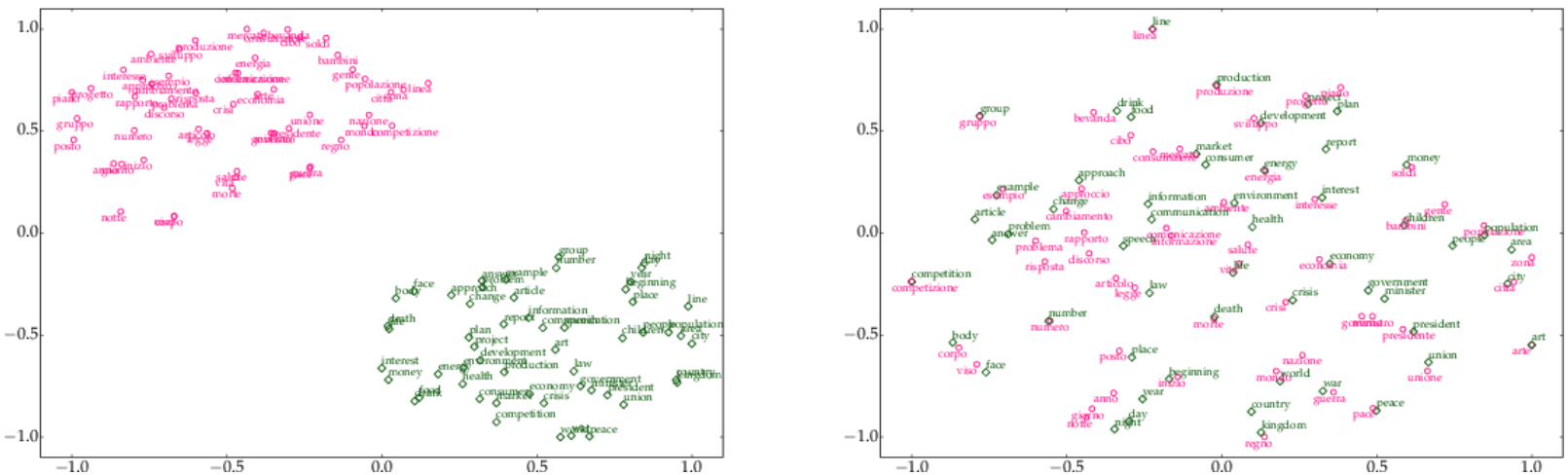

Figure: Monolingual vs. joint cross-lingual space [16]

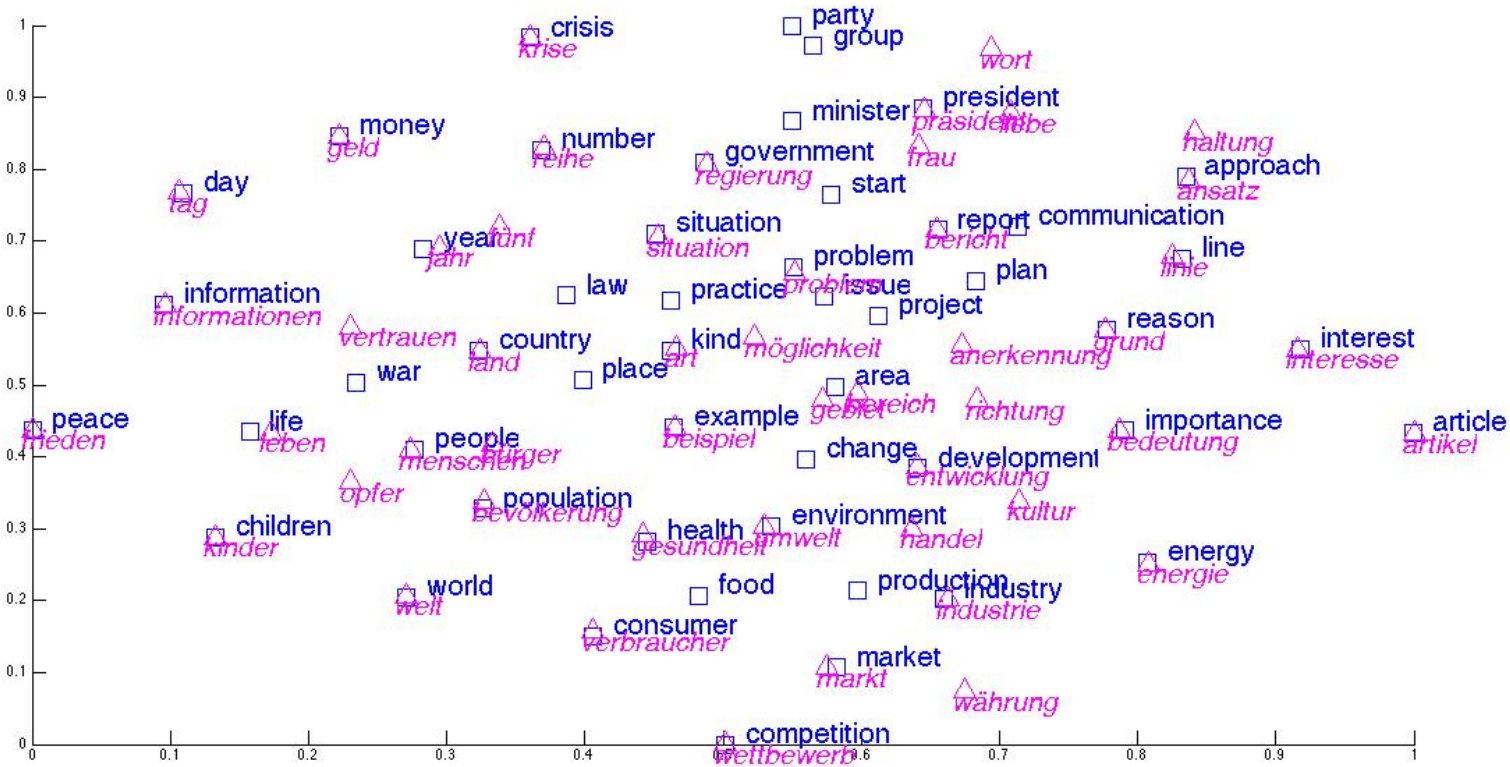

Figure: Cross-lingual space English-German [15]

Cross-lingual Embedding Models and Lexicography

- Importance of the cross-lingual embedding models in lexicography:
  - Connect meanings across the languages
  - Translation equivalents' extraction
  - Availability for small/ rare language pairs
  - Balanced texts in genres
- The bilingual lexicon induction task

Current Drawbacks

- Inconsistent evaluation and metrics (e.g., each paper uses a different evaluation dataset or metric)
- Erroneous evaluation datasets [12, 7] (e.g., automatically compiled evaluation datasets with occurring mistakes and irrelevant words such as proper names)
- Lacking lexicographic perspective (e.g., papers concentrate mainly on the technical side of the problem)

Aims

- Evaluation of three benchmark models (MUSE [5, 14]; VecMap [4, 3, 2, 1]; FastText [11])
- On three diverse language pairs (ET-SK; CZ-SK; EN-KO)
- Crucial parameters for the evaluation datasets:
  - Unifying
  - Reproducible
  - High-quality
- Better evaluation = better training
- Link the NLP and lexicography for cross-lingual embedding models

Experimental Setup

- Training involves:
  - Monolingual word embeddings: FastText [8]; SketchEngine [9]
  - The level of supervision: supervised, semi-supervised, unsupervised
  - Seed lexicons: word-to-word datasets:
    - ET-SK: Pivot dictionary [6] (e.g., *lõpmatu - nekonečný*; *kannatlik - trpezlivý*)
    - CZ-SK: Manual, consisting of identically spelt words (e.g., *krása - krása*; *modrou - modrou*)
    - EN-KO: MUSE dataset (e.g., *and - 그리고*; *with - 함께*; *earth - earth*) [5]
- Metrics (%):
  - Precision (P@k)
  - Recall

Evaluation Datasets

- ET-SK: Pivot dictionary [6](e.g., *aeglaselt* - *pomaly*; *üllatama* - *prekvapiť*)
- CZ-SK: Manual, consisting of different words (e.g., *želva* - *korytnačka*, *turtle*; *březen* - *marec*)
- EN-KO: MUSE [5] (e.g., *abdullah* - *압둘라*; *ibrahim* - *이브라힘*) SketchEngine [13] (e.g., *cake* - *케이크*; *woman* - *대상*)

Vocabulary

- Monolingual word embeddings influence the resulting quality
- Type and size of the vocabulary
- Out-of-the-vocabulary words:
  - Multi-word expressions (*take off*, *office supplies*; de *Grundschule* - en *primary school*, *elementary school*)
  - Low-frequency words or words with 0 occurrences in the corpus
  - Words left out during training
- FastText: between 300K-600K words
- SketchEngine: between 800K-6 mil. words

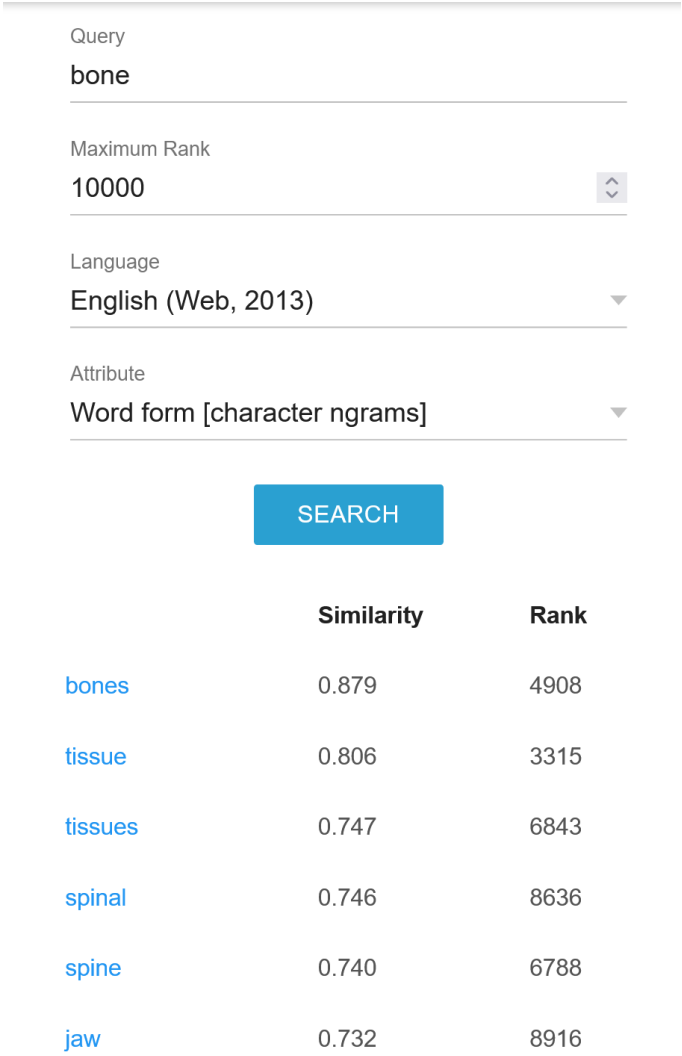

Figure: Search for the word *bone* with a word rank of 10,000.

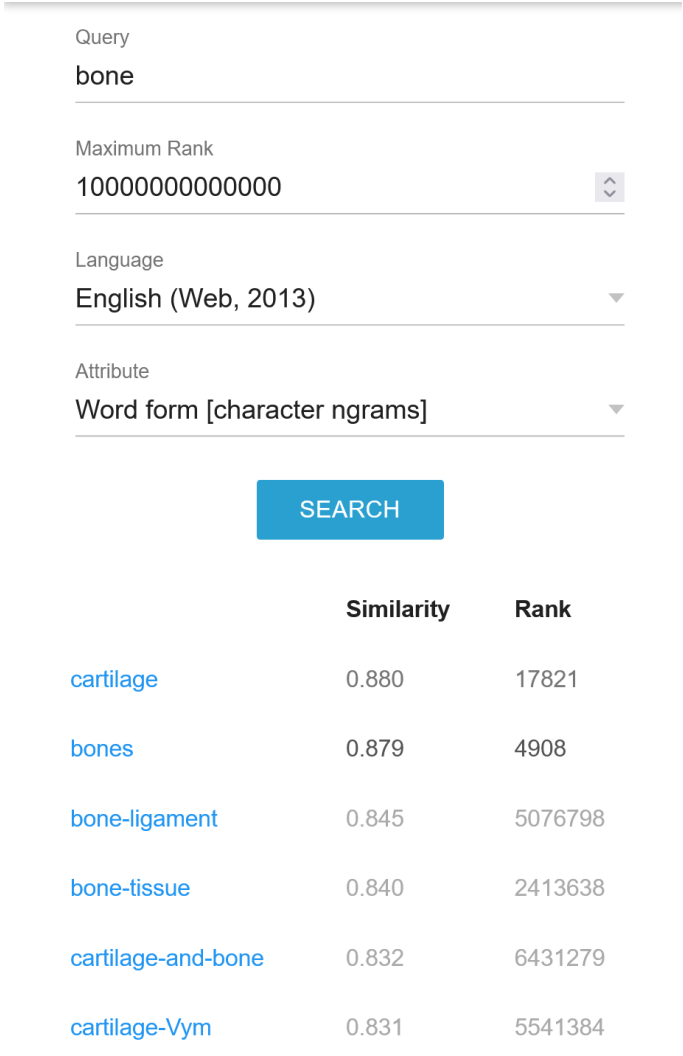

Figure: Search for the word *bone* with a word rank of 10,000,000,000,000.

FastText/ SketchEngine (%)	50K loaded			300-400K loaded		
	ET-SK	CZ-SK	EN-KO	ET-SK	CZ-SK	EN-KO
MUSE-S	19.33 20.00	57.84 70.94	39.97 31.01	27.86 42.40	68.73 78.95	29.98 34.14
MUSE-I	19.26 19.40	57.91 **71.00**	39.00 28.41	25.80 38.93	68.73 79.02	23.17 29.65
MUSE-U	19.80 18.80	58.58 **71.00**	36.46 26.14	24.46 34.80	69.13 79.02	24.58 25.33
VecMap-S	20.73 20.33	58.24 70.67	**50.51** 32.52	34.93 **51.86**	69.73 79.02	49.00 35.44
VecMap-I	21.00 19.20	59.05 **71.00**	**41.59** 28.63	34.73 46.00	71.87 **80.09**	33.98 29.87
VecMap-U	**21.20** 18.86	58.98 70.67	36.35 21.93	33.53 44.80	71.94 **80.09**	29.76 12.42
FastText	20.60 21.06	57.51 70.54	50.40 26.90	31.93 49.33	67.93 78.28	**51.91** 37.60

Figure: The recall of models before and after changing the parameter for loaded embeddings.

Inflected Word Forms

- Words occur in context; they are not necessarily in their basic form
- For example:
  - *tund* (*hour*): *hodiny*, *hodinu*, *hodín*, *hodina*
- Do we want the basic form only or the most frequent form according to the corpus?
- Solutions:
  - Include all word forms in the evaluation dataset
  - Lemmatise: Majka[1]

[1] https://nlp.fi.muni.cz/czech-morphology-analyser/

FT MEs	ET		CS	
	NON-L	LEM	NON-L	LEM
MUSE-S	27.86	29.40	68.73	70.80
MUSE-I	25.80	27.86	68.73	70.94
MUSE-U	24.56	28.80	69.13	71.20
VM-S	34.93	35.93	69.73	71.74
VM-I	34.73	36.06	71.87	72.94
VM-U	33.53	35.20	71.94	73.01
FT	31.93	32.06	67.93	70.14

Table: The results before (NON-L) and after lemmatisation (LEM) of the results from the models trained with FastText monolingual word embeddings (FT MEs)

Part of Speech

- Criticism of datasets containing a large number of proper nouns [12]
- Datasets with similar and even POS distribution [10]
- Not all POS are relevant to reflect the model's performance (e.g., pronouns, articles, conjunction, prepositions)
- Tagging the evaluation datasets with part-of-speech tags

Estonian-Slovak

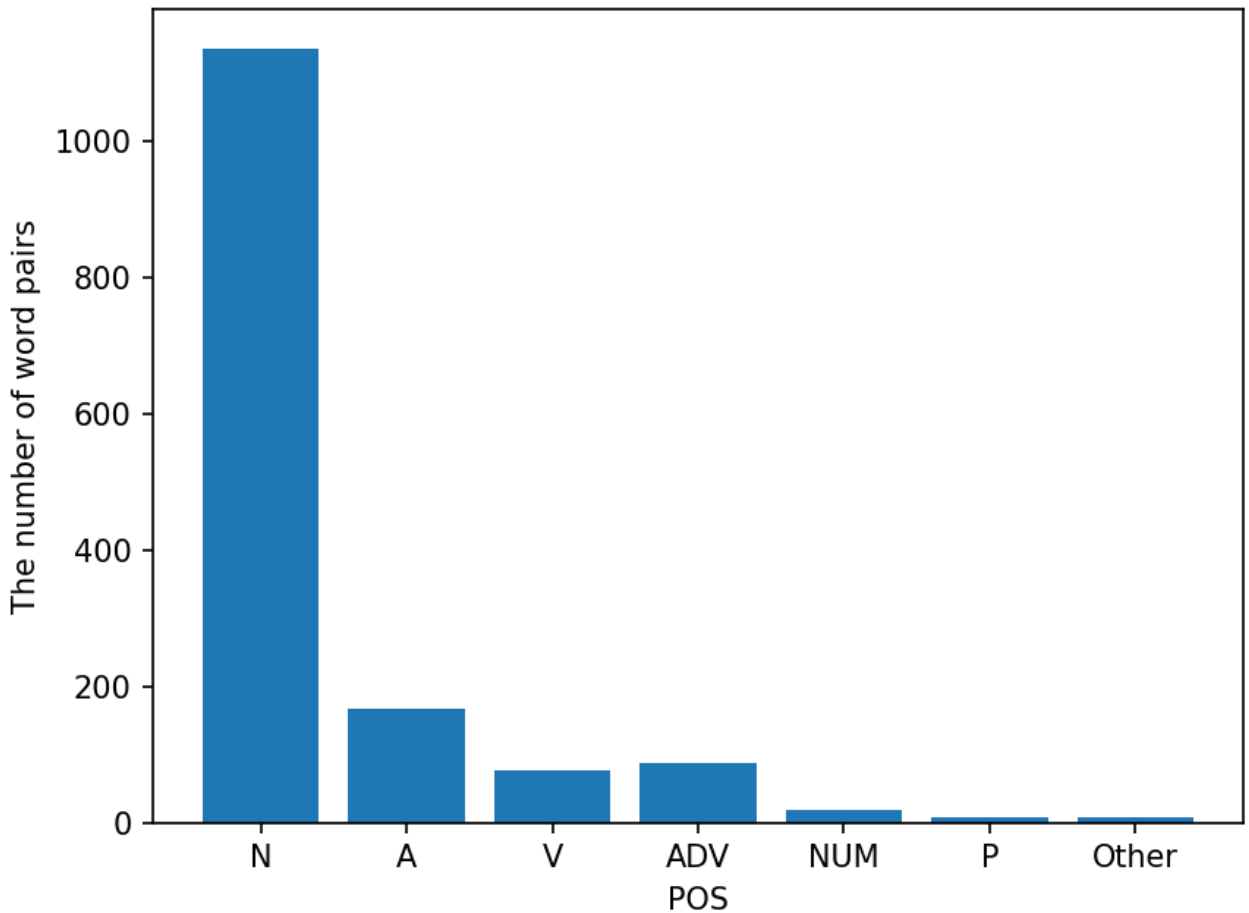

Figure: The POS distribution of the Estonian-Slovak evaluation dataset.

Czech-Slovak

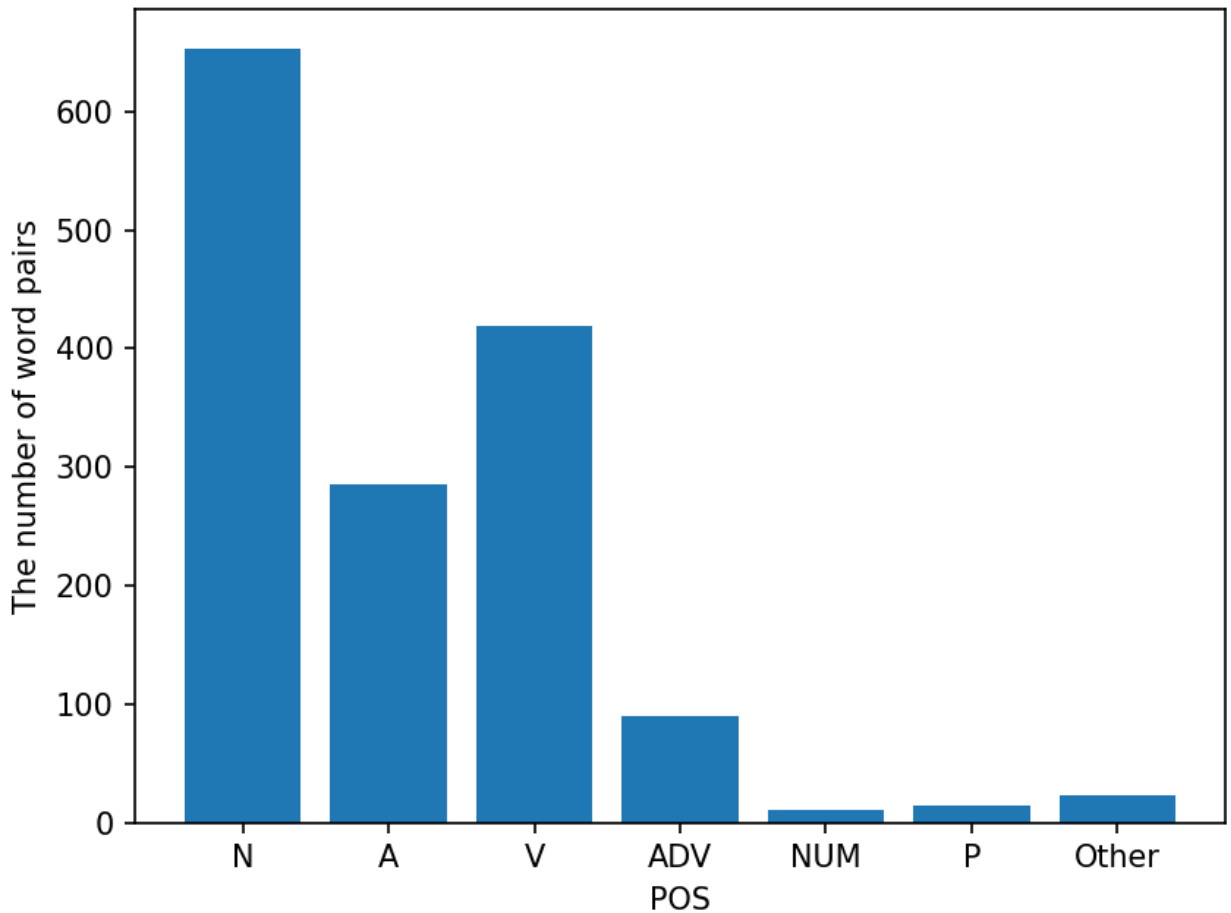

Figure: The POS distribution of the Czech-Slovak evaluation dataset.

English-Korean

MUSE [5] and SketchEngine [13] datasets

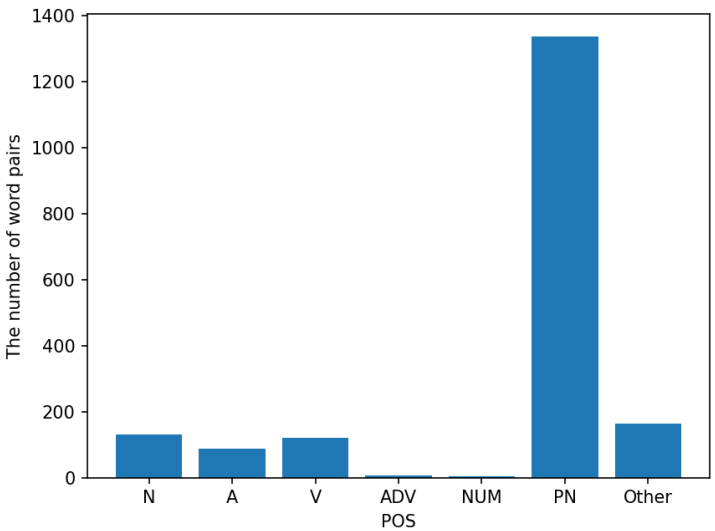

Figure: The POS distribution of the English-Korean evaluation dataset from MUSE.

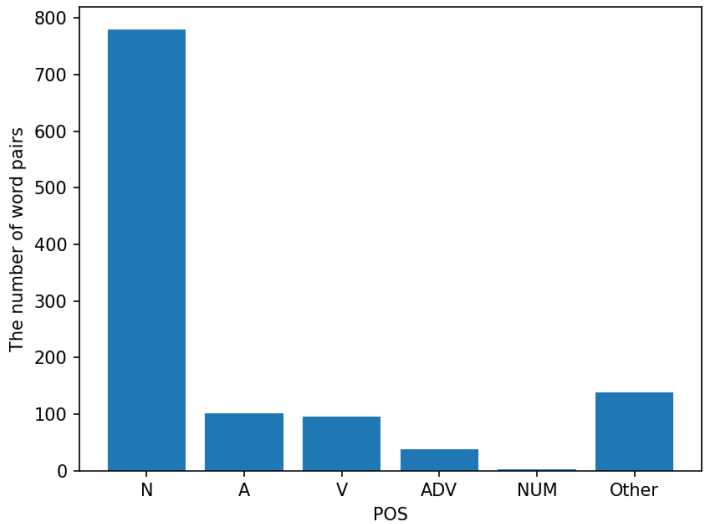

Figure: The POS distribution of the English-Korean evaluation dataset from SketchEngine.

(%)	MUSE dataset		SketchEngine dataset	
	FastText MEs	SketchEngine MEs	FastText MEs	SketchEngine MEs
MUSE-S	29.98	34.14	20.58	31.83
MUSE-I	23.17	29.65	19.89	24.56
MUSE-U	24.58	25.33	19.37	23.18
VecMap-S	49.00	35.44	**29.41**	36.24
VecMap-I	33.98	29.87	23.44	25.25
VecMap-U	29.76	12.42	22.31	15.22
FastText	**51.91**	**37.60**	28.37	**37.80**

Figure: The results comparison for the models trained on English-Korean after changing the POS distribution.

VecMap-S result's comparison

- MUSE dataset:
  - More correct proper nouns: *Abdullah*, *Alexandra*, *Cameroon*, *Helsinki*
  - More correct international words: *alias*, *android*, *idol*
  - Incorrect caused mainly by mistakes in the evaluation dataset: *android-android*; *Yemen-South Yemen*
- SketchEngine dataset:
  - Words for foods, animals, numbers: *coffee-tea*, *fifty-fourteen*
  - More verbs

VecMap-S (%) FT/SE	ET-SK	CZ-SK	EN-KO
Nouns	31.89/ 48.54	76.87/ 86.21	48.46/ **50.00**
Adjectives	48.21/ 63.09	73.07/ 77.97	47.19/ 43.82
Verbs	35.52/ 64.47	63.48/ 67.30	35.00/ 21.66
Adverbs	41.37/ 56.32	70.00/ 81.11	**75.00**/ 37.50
Numerals	61.11/ 77.77	**81.81**/ 81.81	25.00/ 25.00
P/ PN	**62.50**/ 75.00	80.00/ **100**	51.00/ 34.25
Others	25.00/ 37.50	69.56/ **100**	43.55/ 39.26

Figure: The recall for each POS in VecMap-S.

Senses

- How many senses to include?
- *band* - music group, piece of cloth, range of values
- How are precision and recall changing when top 1, 5, or 10 target words are considered
- Higher recall - lower precision and reversely
- *kokkulepe* - *agreement*, *contract*, *lease*, *deal*:
  - *zmluva*
  - *zmluve*
  - *zmluvám*
  - *zmluvná*
  - *dohody*
  - *zmluve*
  - *dohodu*
  - *zmluvy*
- Our goal: material for lexicographers or students

Conclusion

- Parameters: selected vocabulary, word forms, POS distribution, and precision vs recall
- Set the goal beforehand
- Monolingual word embeddings influence the quality of the resulting model (FastText vs SketchEngine)
- Supervised mode is better when training distant language pairs; identical or unsupervised fits for close ones
- Cross-lingual embedding models: data supplement, low-resource or rare language pairs, technical dictionaries

Bibliography I

[1] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. “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)*. Association for Computational Linguistics, 2018, pp. 789–798.

[2] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. “Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations”. In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*. 2018, pp. 5012–5019.

Bibliography II

[3] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. “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)*. Association for Computational Linguistics, 2017, pp. 451–462.

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

[5] Alexis Conneau et al. “Word Translation Without Parallel Data”. In: *ArXiv* abs/1710.04087 (2017).

Bibliography III

[6] Michaela Denisová. “Compiling an Estonian-Slovak Dictionary with English as a Binder”. In: *Proceedings of the eLex 2021 conference*. Lexical Computing CZ, s.r.o., 2021, pp. 107–120.

[7] Michaela Denisová and Pavel Rychlý. “When Word Pairs Matter: Analysis of the English-Slovak Evaluation Dataset”. In: *Recent Advances in Slavonic Natural Language Processing (RASLAN 2021)*. Brno: Tribun EU, 2021, pp. 141–149.

[8] Edouard Grave, Armand Joulin, and Quentin Berthet. “Unsupervised Alignment of Embeddings with Wasserstein Procrustes”. In: *International Conference on Artificial Intelligence and Statistics*. 2018.

Bibliography IV

[9] Ondřej Herman. “Precomputed Word Embeddings for 15+ Languages”. In: *RASLAN 2021 Recent Advances in Slavonic Natural Language Processing* (2021), pp. 41–46.

[10] Mike Izbicki. “Aligning Word Vectors on Low-Resource Languages with Wiktionary”. In: *Proceedings of the Fifth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2022)*. Association for Computational Linguistics, 2022, pp. 107–117. URL: https://aclanthology.org/2022.loresmt-1.14.

[11] Armand Joulin et al. “Loss in Translation: Learning Bilingual Word Mapping with a Retrieval Criterion”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2018, pp. 2979–2984.

Bibliography V

[12] Yova Kementchedjhieva, Mareike Hartmann, and Anders Søgaard. “Lost in Evaluation: Misleading Benchmarks for Bilingual Dictionary 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)*. Association for Computational Linguistics, 2019, pp. 3336–3341.

[13] Vojtěch Kovář, Vít Baisa, and Miloš Jakubíček. “Sketch Engine for Bilingual Lexicography”. In: *International Journal of Lexicography* 29.3 (2016), pp. 339–352.

[14] Guillaume Lample, Ludovic Denoyer, and Marc’Aurelio Ranzato. “Unsupervised Machine Translation Using Monolingual Corpora Only”. In: *ArXiv* abs/1711.00043 (2017).

Bibliography VI

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

[16] Sebastian Ruder, Ivan Vulić, and Anders Søgaard. “A Survey of Cross-lingual Word Embedding Models”. In: *The Journal of Artificial Intelligence Research* 65 (2019), pp. 569–631.

Thank You for Your Attention!

MUNI
FACULTY
OF INFORMATICS