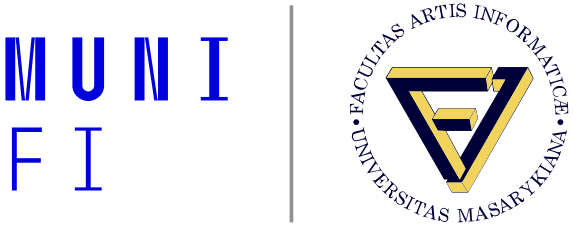

When Word Pairs Matter

Analysis of the English-Slovak Evaluation Dataset

Michaela Denisová and Pavel Rychlý
449884@mail.muni.cz
pary@fi.muni.cz

Faculty of Informatics, Masaryk University

December 11, 2021

Content

Introduction

- Reliability of the evaluation datasets
- English-Slovak dataset with 2,739 word pairs used to evaluate MUSE [5]
- Aim of the project

Related work

- MUSE
  - Open-source library[1] with pre-trained multilingual word embeddings and available evaluation and training datasets
- VecMap [4][3][2][1]
  - Open-source cross-lingual embedding model[2]
  - 4 types of training: supervised, semi-supervised, identical and unsupervised
  - Strong supervision with 5,000 word-pair dataset using FastText English and Slovak monolingual word embeddings [7]

[1] https://github.com/facebookresearch/MUSE
[2] https://github.com/artetxem/vecmap

Analysis of the Dataset

Three aspects:

1. Frequency
2. Category
3. Similarity

Frequency

- Frequencies for each word pair from English-Slovak parallel corpus OPUS2 [8] via SketchEngine API [6]
- Corpus size: 8,000,000 sentences, 8,000 documents

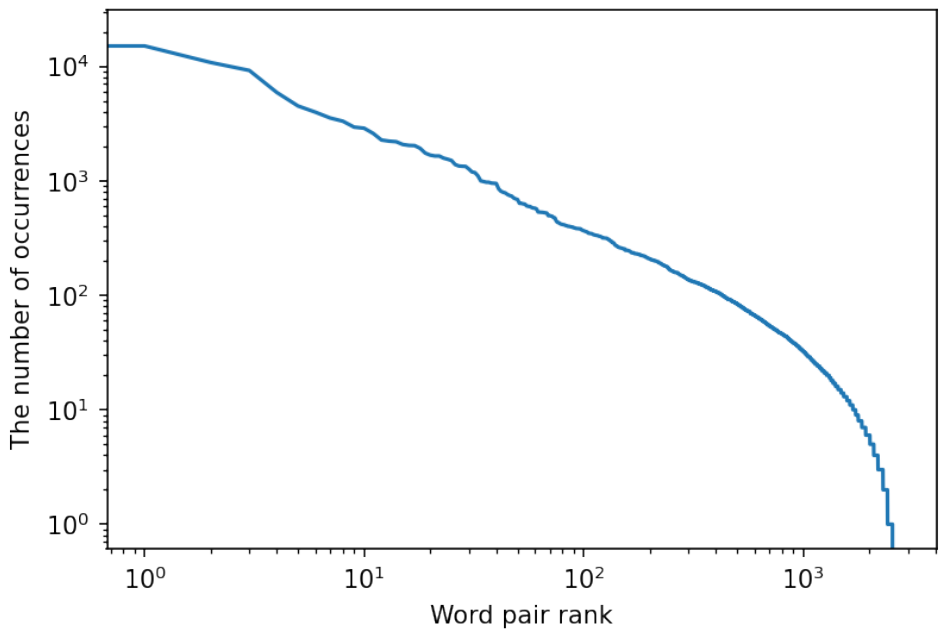

Figure: Frequency distribution of each word pair in the parallel English-Slovak corpus OPUS2 represented by logarithm of Zipf's curve

Category

- Manually annotated with categories from A to J
  - A: Correct translation
  - B - J: Incorrect translation, major or minor mistake

Category	Description	Weight	Example
A	correct translation	1	'*admit*' : '*priznať*'
B	inflected word form	0.80	'*advocacy*' : '*obhajobu*'
C	different part of speech	0.30	'*darkness*' : '*temné*'
D	translated as same non-Slovak word, abbreviations	0.20	'*bbc*' : '*bbc*'
E	proper names	0.20	'*bruno*' : '*bruno*'
F	synonym or incorrect translation	0.10	'*intensity*' : '*svietivosť*'
G	incomplete word pair	0.20	'*brigadier*' : '*brigádny*' (generál)
H	non-existing English word	0.10	'*wwe*' : '*mozeme*'
I	interjection	0.80	'*boom*' : '*bum*'
J	missing diacritics	0.60	'*joy*' : '*radosť*'

Table: Categories, their description, weights and an example of a word pair from the respective category

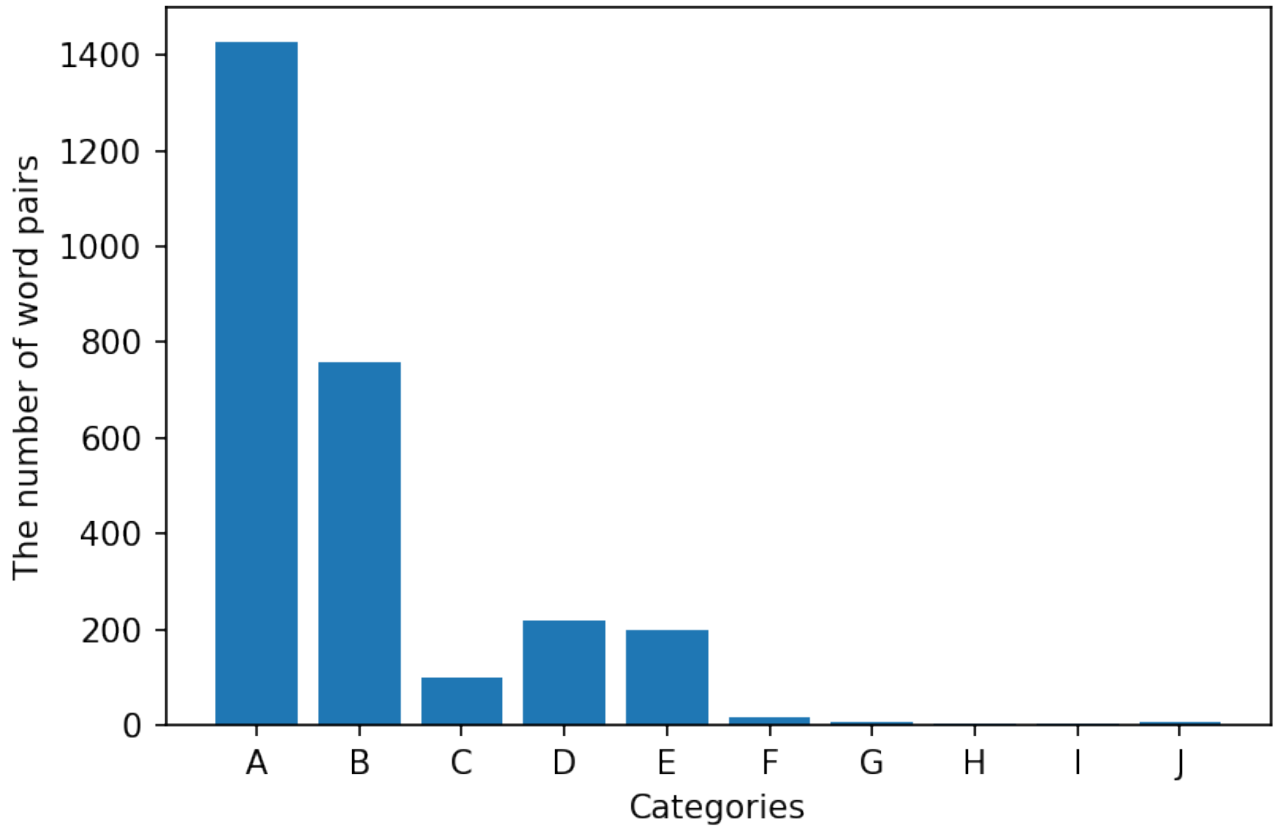

Figure: The number of word pairs in each category

Similarity

- Proposal of Slovak translation to every English headword in category from B to J
- Cosine similarity measurement

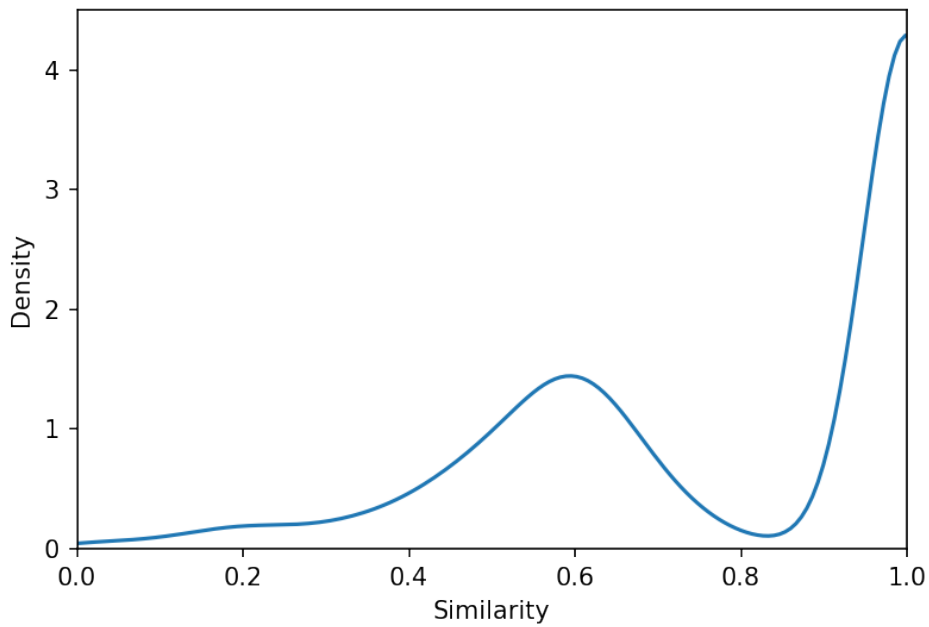

Figure: Cosine similarity distribution from 0 to 1.

Assigning weights

- Aims:
  - Reflect word pairs relevance
  - Not to penalize model for word pairs with lower weight and increase the accuracy for the word pairs with higher weight
- Re-scaling numbers to the range from 0 to 1
- Using only scaled frequencies as weights

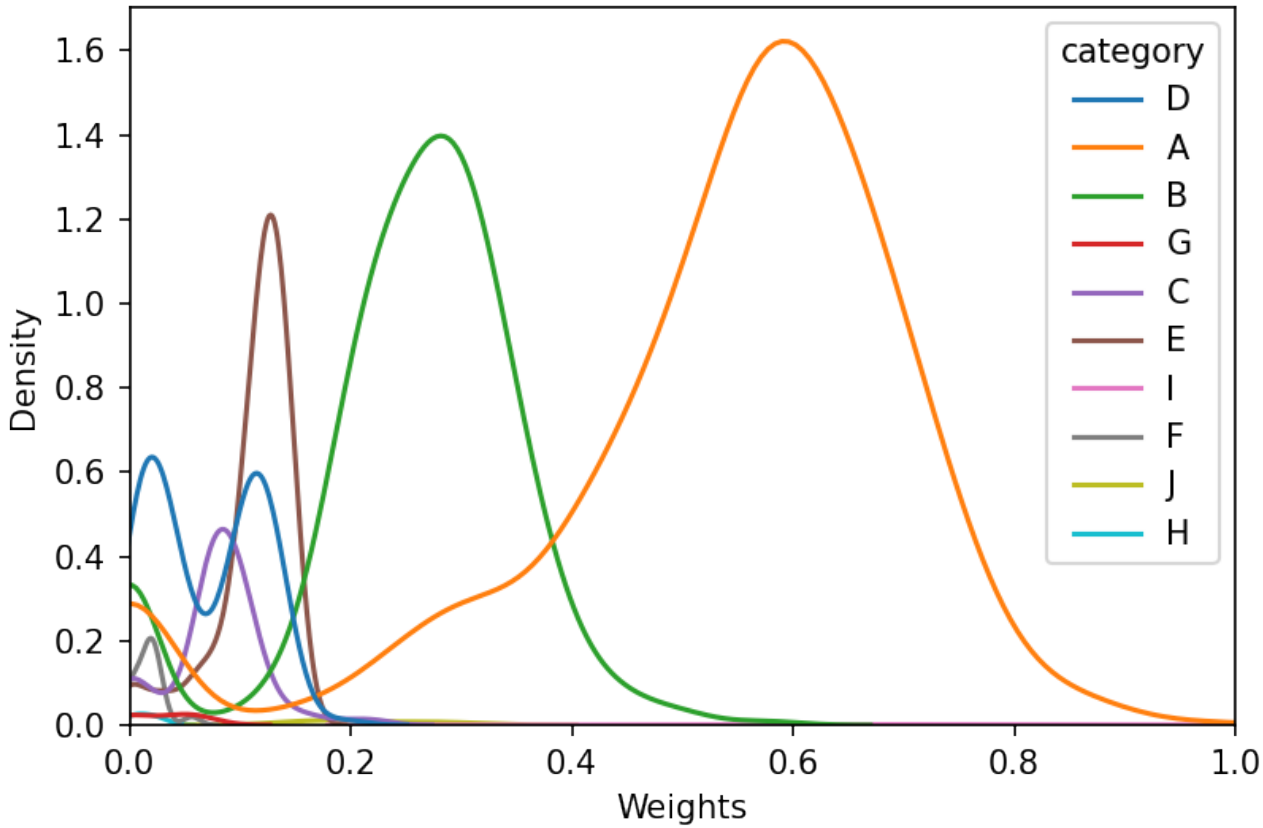

Figure: Histograms of weights distribution in each category

Evaluation

- MUSE and VecMap
- Extracting nearest neighbors of every English headword and comparing them to the original evaluation dataset

	Without Weights	With Weights	Scaled frequencies
MUSE	30.41	**34.60**	32.82
VecMap	38.15	48.43	**54.74**

Table: The performance of MUSE and VecMap models before and after applying weights

- MUSE: 294 word pairs
- VecMap: 506 word pairs
- Match: 539 word pairs

EN	SK	Frequency	Weight	MUSE	VecMap
decrease	zníženie	**274**	**0.8709**	Yes	No
estonia	estónsko	42	0.7592	Yes	No
luxembourg	luxembursko	39	0.7555	Yes	No
euro	eurá	188	0.3957	Yes	No
vii	vii	254	0.1733	Yes	No
carefully	starostlivo	101	0.8115	No	Yes
decrease	pokles	253	0.8663	No	Yes
infection	infekcia	283	**0.8730**	No	Yes
hey	hej	1349	0.7728	No	Yes
tel	tel	**2384**	0.2000	No	Yes

Table: Comparison of the word pairs with the highest frequency (in hits per million) and/or highest weight that were found either by MUSE or VecMap model

Conclusion

- More accurate picture of the models
- VecMap outperforms MUSE
- Quality of the word pairs in the evaluation dataset plays an important role when evaluating models

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)*. 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)*. 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*. 2016, pp. 2289–2294.

[5] Alexis Conneau et al. “Word Translation Without Parallel Data”. In: *arXiv preprint arXiv:1710.04087* (2017). url: `https://arxiv.org/abs/1710.04087`.

Bibliography III

[6] Adam Kilgarriff et al. “The Sketch Engine: ten years on”. In: *Lexicography* (2014), pp. 7–36. url: http://dx.doi.org/10.1007/s40607-014-0009-9.

[7] Tomas Mikolov et al. “Advances in Pre-Training Distributed Word Representations”. In: *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*. 2018. url: https://arxiv.org/abs/1712.09405.

[8] Jörg Tiedemann. “Parallel Data, Tools and Interfaces in OPUS”. In: *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Ed. by Nicoletta Calzolari (Conference Chair) et al. Istanbul, Turkey: European Language Resources Association (ELRA), May 2012. isbn: 978-2-9517408-7-7. url: https://aclanthology.org/L12-1246/.

Summary

- Reliability of the evaluation datasets
- MUSE and VecMap models
- 3 aspects
  - Frequency
  - Category
  - Similarity
- Evaluating two models with and without weights
- Word pairs are an important factor when evaluating

Thank You for Your Attention!

MUNI
FACULTY
OF INFORMATICS