Does Size Matter?

Comparing Evaluation Dataset Size for the Bilingual Lexicon Induction

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

Natural Language Processing Centre

Faculty of Informatics, Masaryk University

December 8, 2023

Introduction

- eLex 2023 extension
- Evaluation of cross-lingual embedding models with datasets of various sizes
- Can we evaluate the model with fewer word pairs with the same precision while minimising the time and effort?
- Estonian-Slovak, Czech-Slovak, and English-Korean language

Cross-lingual Embedding Models

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

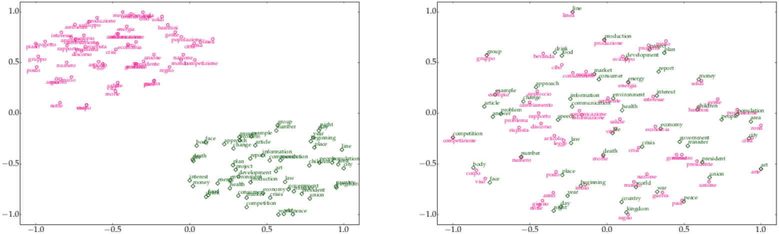

Figure: Monolingual vs. joint cross-lingual space [8]

- Bilingual lexicon induction task

Baselines

- Muse [3] - supervised (Muse-S), unsupervised (Muse-U), identical string relying (Muse-I)
- VecMap [2, 1] - supervised (VM-S), unsupervised (VM-U), identical string relying (VM-I)
- RCLS [7] - supervised only
- Trained on two MWEs trained on two types of data: Wikipedia (FastText) [5] and web corpora (SketchEngine) [6]

Evaluation Datasets

- Estonian-Slovak [4], Czech-Slovak (manual, *želva* - *korytnačka*), English-Korean [3]
- Make each group of word pairs as similar as possible
- Random sampling
- 200, 500, 1.5K, and 3K source words

	et	sk
0	abivalmis	užitočný
1	abivalmis	nápomocný
2	komplekt	súprava
3	geograaf	geograf
4	rohkem	viac

Evaluation

- P@1

$$P = TruePositives/(TruePositives + FalseNegatives)$$

- Output = [tgw0, tgw,1, tgw2, ...]

Results

Estonian-Slovak

et-sk (%)	FastText				SketchEngine			
	200	500	1.5K	3K	200	500	1.5K	3K
MUSE-S	17.34	18.93	21.37	**23.18**	26.53	27.02	32.30	**36.14**
MUSE-I	10.20	13.40	15.30	**16.65**	27.55	24.68	28.82	**32.03**
MUSE-U	11.73	12.76	13.52	**15.64**	20.40	20.85	23.80	**27.14**
VM-S	19.89	25.53	26.88	**30.72**	28.57	28.72	34.81	**38.85**
VM-I	17.34	18.29	22.18	**24.60**	21.93	22.76	26.63	**30.15**
VM-U	15.30	16.17	19.67	**21.72**	22.95	22.12	26.63	**29.80**
RCLS	16.83	19.78	22.99	**27.05**	27.55	26.59	34.73	**38.28**

Table 1: The results for the Estonian-Slovak language combination.

Results

Czech-Slovak

cs-sk (%)	FastText				SketchEngine			
	200	500	1.5K	3K	200	500	1.5K	3K
MUSE-S	58.08	62.10	64.99	**68.72**	62.26	65.89	71.50	**75.72**
MUSE-I	59.59	61.68	64.92	**68.93**	61.61	65.68	70.97	**75.48**
MUSE-U	60.60	62.31	65.51	**69.25**	61.00	65.68	70.97	**75.44**
VM-S	59.09	60.63	66.41	**69.13**	62.62	65.47	71.50	**75.84**
VM-I	59.09	64.42	68.66	**72.10**	61.61	65.89	71.42	**75.52**
VM-U	59.09	64.21	68.58	**72.10**	61.61	65.89	71.50	**75.60**
RCLS	57.57	61.05	64.32	**68.04**	64.14	67.36	72.70	**76.48**

Table 2: The results for the Czech-Slovak language combination.

Results

English-Korean

en-ko (%)	FastText				SketchEngine			
	200	500	1.5K	3K	200	500	1.5K	3K
MUSE-S	13.91	13.57	**17.44**	15.91	16.49	19.82	**21.23**	19.00
MUSE-I	11.34	14.22	**17.16**	15.80	10.30	**15.51**	14.64	13.90
MUSE-U	10.30	11.42	**13.94**	12.78	12.37	**13.36**	12.05	11.63
VM-S	29.38	29.52	**35.31**	33.80	21.13	20.90	**23.75**	21.58
VM-I	20.61	17.67	**21.72**	19.03	13.91	15.30	**15.41**	13.43
VM-U	12.37	14.22	**16.53**	14.51	**6.70**	5.81	6.51	5.63
RCLS	30.92	27.80	**34.40**	32.54	21.13	20.90	**22.91**	20.25

Table 3: The results for the English-Korean language combination.

Results

VM-S

VM-S	ET-SK		CS-SK		EN-KO	
	FastText	SketchEngine	FastText	SketchEngine	FastText	SketchEngine
I.	26.73	29.34	64.64	70.50	28.33	17.45
II.	21.42	25.10	61.89	68.00	28.45	17.78
III.	26.30	33.04	60.45	67.28	31.12	18.67
IV.	22.82	30.65	58.10	63.36	27.55	20.00
V.	22.73	30.31	57.74	64.22	29.35	19.07
VI.	21.61	29.55	53.52	57.64	30.06	18.88

Table 4: The results of 3K-headword evaluation datasets split into groups of 500.

Error Analysis

- Performed manual check
- Large gaps between the results

VM-S	ET-SK		CS-SK	
	FastText	SketchEngine	FastText	SketchEngine
3K	45.41	56.51	86.57	94.41
1.5K	**47.61**	59.67	**87.28**	94.83
500	46.59	58.51	86.31	**94.94**
200	46.93	**60.71**	86.86	94.44

Table 5: Manual error analysis of the results of the model VM-S for Estonian-Slovak and Czech-Slovak.

Error Analysis

- *ajajärk* (*time period*, *era*, *epoch*) - *obdobie* (VM-S), *doba* (evaluation dataset).
- *puhuma* (*to blow*) had multiple target words such as, *pofúkať*, *fúkať*, *trúbiť*, *vanúť*, *zaviať*, *viať*
- Uneven distribution of OOV words

Comparison of the two MWEs

Type	SRC	ED	FT	SkeEng	Description
A	Clara	클라라	클라라	외가에서	proper names
	Emma	엠마	엠마	희진	
	Erik	에릭	에릭	동료인	
	Phillip	필립	필립	옹은	
B	vms	vms	vms	램도	same word with same word
	pgm	pgm	pgm	변환하고	
C	hnědá	hnedá (brown)	žltohnedá	hnedá	precise translations
D	stýskat	cniet' (to miss)	-	cniet'	low-frequency words, slang
	chasník	mládenec (young man)	-	chasník	
	emps	mamka	-	mamka	

Table 6: The differences between the models trained with FastText and SketchEngine MWEs (examples from EN-KO, ET-SK, and CS-SK trained with the VM-S model).
(FT = FastText; SkeEng = SketchEngine)

Conclusion

- Random splitting of datasets does not ensure an equal underlying distribution within all the datasets
- Result strongly depends on the appropriate vocabulary choice rather than on the size of the dataset
- Smaller datasets work - focus on the quality of the chosen vocabulary for the evaluation dataset

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.

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

Bibliography II

[4] 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.

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

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

Bibliography III

[7] 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.

[8] 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