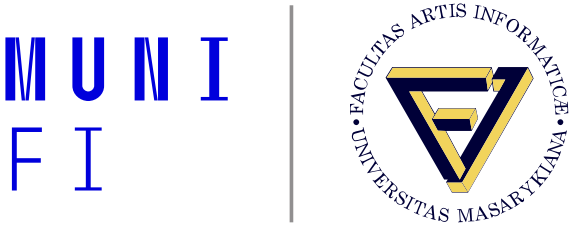


Comparison of the Parallel and Comparable Data-based methods

Michaela Denisová
449884@mail.muni.cz

Faculty of Informatics, Masaryk University

December 10, 2022

Introduction

- Parallel data- and comparable data-based methods for the bilingual dictionaries
- Can comparable data compete with standard, widely used parallel data?
- Estonian-Slovak language

Parallel Data

- Rich context information
- Lower the human intuition
- Not always available, unbalanced texts
- Parallel Corpus
  - EUR-Lex[5][1], 300,000,000 tokens
  - Statistics-based method, logDice score[14, 10]
  - Manually evaluated 1,000 word pairs
- Pivot Dictionary
  - Two dictionaries that have one common language
  - Estonian-Slovak dictionary compiled from English-Estonian, and English-Slovak dictionaries[7]
  - Manually evaluated 1,000 word pairs

[1] https://eur-lex.europa.eu/homepage.html

Comparable Data

- Quickly developing
- Available for any language
- No context, no multi-word expressions
- Cross-lingual Embedding Models
  - Align two monolingual embeddings into one joint space[13]
  - Supervised, semi-supervised, unsupervised
  - Three SotA models: MUSE[6, 11][2], VecMap[1, 3, 2, 4][3], FastText[9][4]

[2] https://github.com/facebookresearch/MUSE
[3] https://github.com/artetxem/vecmap
[4] https://fasttext.cc/

Setup

- FastText[12] and SketchEngine[8][5] monolingual word embeddings
- Supervised, unsupervised, identical strings

[5] https://embeddings.sketchengine.eu/

Evaluation

- Parallel Corpora
  - Randomly sampled 1,000 word pairs
  - Two constrains: frequency above 1,000, logDice score above 10
  - numbers ('558' : '558'), incorrect proper names ('engström' : 'alfonsi'), website links ('vormis' : 'http://eur-lex.europa.eu/en/index.htm'), different language words ('nuts' : 'nuts')
- Cross-lingual Embedding Models
  - 100 Estonian words from the Basic Estonian Dictionary[6]
  - 70 high-frequency, 30 low-frequency
  - Manually evaluated at P@10, P@5, P@1

[6] http://www.eki.ee/dict/psv/

Table 1: Manual controls' labels and examples.

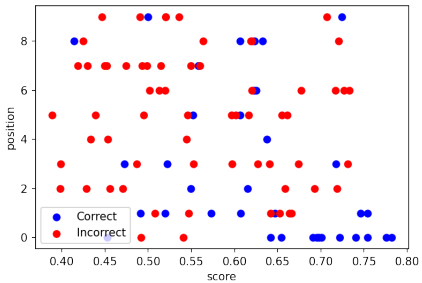
Correctness	Label	Example
YES	meanings match	*'ammu'* : *'dávno'*(long time ago)
YES	inflected word form	*'demokraatia'* : *'demokracii'* (democracy)
YES	adjective in different grade	*'õnnelik'* : *'najšťastnejší'* (the happiest, correct: happy)
YES	near equivalent or synonym	*'sõitma'* : *'šoférovať'* (travel/drive)
YES	additional word needed	*'kontoritarbed'* : *'kancelárske'* (office, correct: office supplies)
NO	different part of speech	*'sõbralik'* : *'priateľstvo'* (friendship, correct: friendly)
NO	antonym	*'kiire'* : *'pomalé'* (slow, correct: fast)
NO	number	*'kell'* : *'17.00'* (clock)
NO	shortcut	*'kilo'* : *'kg'*
NO	symbol	*'kell'* : *'+'*
NO	meanings do not match	*'linn'* : *'radnica'* (city hall, correct: city)

Results

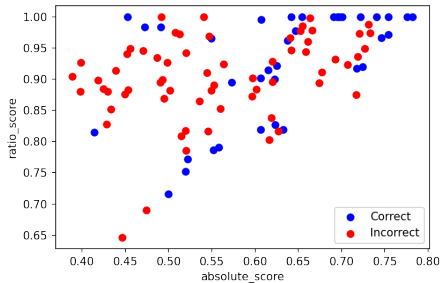
Table 2: The comparison of the precision P@10, P@5, and P@1 when separating words into high- and low-frequency words.

(high-/low-frequency)	P@10	P@5	P@1
FastText			
MUSE-S (%)	26.86/12.12	40/8	58.57/30
MUSE-I (%)	28.76/7.4	25/16.66	45.71/13.33
MUSE-U (%)	27.27/30.43	43.05/32.14	60/26.66
VECMAP-S (%)	43.66/17.24	45.2/29.62	74.28/43.33
VECMAP-I (%)	28.04/22.22	46.05/37.5	60/33.33
VECMAP-U (%)	28.57/20	36.23/41.93	61.42/ 30
FastText (%)	35.21/17.24	33.8/27.58	72.85/40
SketchEngine			
MUSE-S (%)	47.05/34.37	39.39/41.17	70/66.66
MUSE-I (%)	28.35/27.27	38.15/29.16	68.57/56.66
MUSE-U (%)	30.12/23.52	48.64/38.46	71.42/53.33
VECMAP-S (%)	39.70/12.5	48.57/46.66	74.28/66.66
VECMAP-I (%)	38.88/17.85	47.14/23.33	75.71/63.33
VECMAP-U (%)	40/20	39.72/40.74	77.14/60
FastText (%)	13.88/21.42	45.45/47.05	77.14/56.66

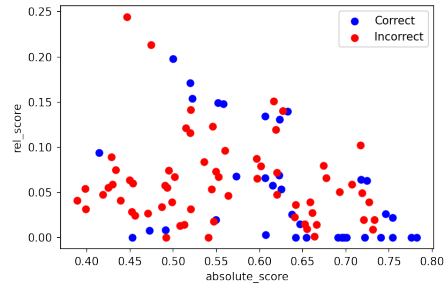
Limit



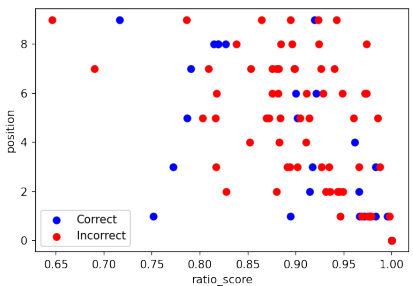
(a) Abs./Pos.



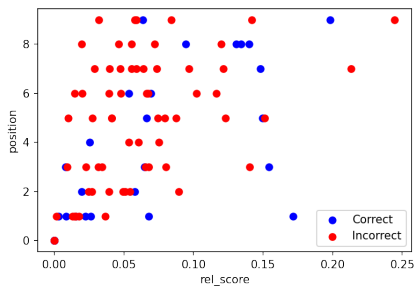
(b) Abs./Ratio



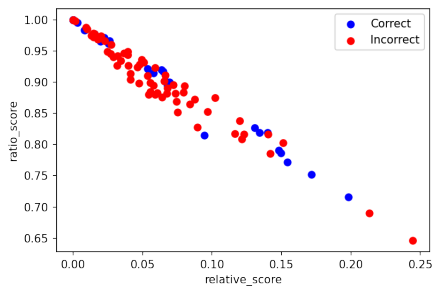
(c) Abs./Differ.



(d) Ratio/Pos.



(e) Differ./Pos.



(f) Differ./Ratio

Figure: Various graphs for correct and incorrect translation equivalents extracted from VecMap trained in a supervised mode with FastText embeddings

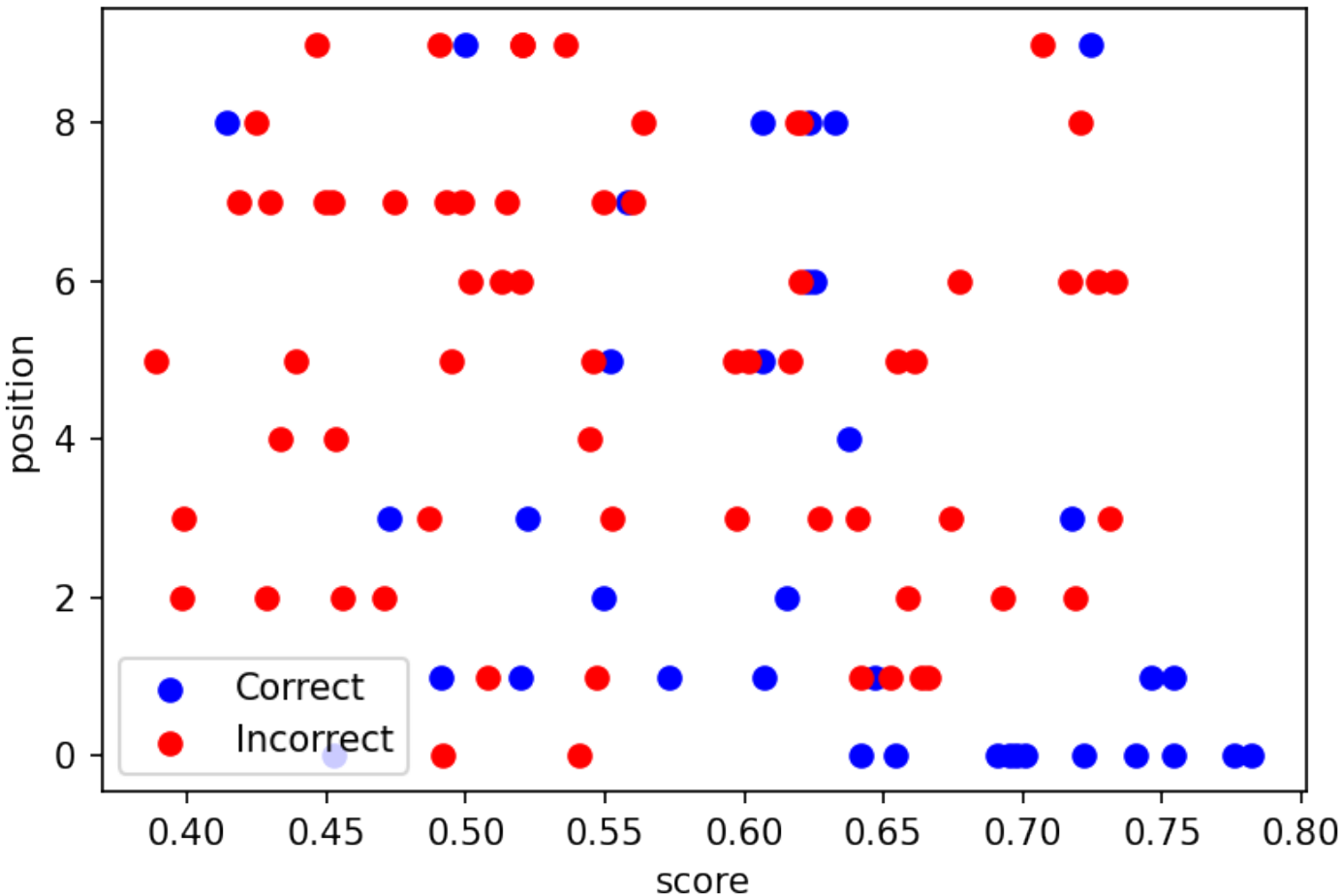


Figure: Absolute score vs. Position

$$limit = [0.45 - 0.65] + (position * 0.01)$$

Table 3: The precision P@10, P@5, and P@1 of comparable data-based models (MUSE, VECMAP, FastText) before and after applying a limit for the extraction of the translation equivalents compared to the performance of the parallel data-based methods.

	P@10/ Limit	P@5/ Limit	P@1
Comparable data			
FastText			
MUSE-S (%)	22/ 36.84	32/ 45.94	50
MUSE-I (%)	23/ 35.71	23/ 35.29	36
MUSE-U (%)	28/ 34.48	40/ 46.05	50
VECMAP-S (%)	36/ 45.2	41/ 46.06	65
VECMAP-I (%)	27/ 31.57	44/ 50.60	52
VECMAP-U (%)	26/ 32.81	38/ 42.35	52
FastText (%)	30/ 40.9	32/ 48.83	63
SketchEngine			
MUSE-S (%)	43/ 47.16	40/ 49.12	69
MUSE-I (%)	28/ 33.33	36/ 54.54	65
MUSE-U (%)	29/ 33.33	46/ 52.72	66
VECMAP-S (%)	31/ 35.36	48/ 52.17	72
VECMAP-I (%)	33/ 37.5	40/ 47.76	72
VECMAP-U (%)	33/ 33.76	40/ 58.33	72
FastText (%)	16/ 21.81	46/ 52.56	71
Parallel data			
Pivot dictionary (%)	40	-	-
Parallel corpus (%)	16.1	-	-

Table 4: Comparison of the word pairs that were found or were not found by MUSE either trained with FastText (MUSE-S-F) or SketchEngine (MUSE-S-S).

	ET	SK	Pos.	Score	Rank	Correct
MUSE-S-F	*laupäev*	*piatok*	1	0.621176	34506	**No**
	laupäev	*sviatok*	8	0.578794	34506	**No**
	suhkur	*cukry*	1	0.962491	28078	**Yes**
	samuti	*rovnako*	5	0.500055	108	**Yes**
MUSE-S-S	*laupäev*	*nedeľu*	4	0.756580	14506	**Yes**
	laupäev	*Nedeľa*	8	0.733683	14506	**Yes**
	laupäev	*víkend*	5	0.756557	14506	**No**
	suhkur	*škrob*	6	0.802334	7490	**No**
	samuti	*rovnako*	4	0.792155	190	**Yes**

Table 5: Comparison of the word pairs that were found or were not found by FastText either trained with FastText (FastText-F) or SketchEngine (FastText-S).

	ET	SK	Pos.	Score	Rank	Correct
FastText-F	*laul*	*pesnička*	0	0.253858	752	**Yes**
	lõplik	*konečné*	4	0.214931	5056	**Yes**
	õnnelik	*milovaný*	1	0.172903	9829	**No**
	sõitma	*cestovať*	0	0.264190	13698	**Yes**
	sõitma	*jazda*	2	0.180665	13698	**No**
FastText-S	*laul*	*hymna*	7	0.573021	4021	**No**
	lõplik	*presný*	8	0.522405	6906	**No**
	õnnelik	*šťastné*	6	0.517561	2381	**Yes**
	sõitma	*viezť*	3	0.628647	2734	**Yes**

Conclusion

- Comparable data outperformed parallel data in some cases
- Good supplement data for low-resource languages or rare language pairs
- Active research area
- Algorithm for automatic limit computing

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.

Bibliography III

[5] Vıt Baisa et al. “European Union Language Resources in Sketch Engine”. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*. Portorož, Slovenia: European Language Resources Association (ELRA), May 2016, pp. 2799–2803.

[6] Alexis Conneau et al. “Word Translation Without Parallel Data”. In: *arXiv preprint arXiv:1710.04087* (2017).

[7] Michaela Denisova. “Compiling an Estonian-Slovak Dictionary with English as a Binder”. In: *Electronic lexicography in the 21st century. Proceedings of the eLex 2021 conference*. 2021, pp. 107–120.

Bibliography IV

[8] Miloš Jakubíček et al. “The TenTen Corpus Family”. In: *7th International Corpus Linguistics Conference CL 2013*. Lancaster, 2013, pp. 125–127. url: http://ucrel.lancs.ac.uk/cl2013/.

[9] 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*. 2018.

[10] Vojtěch Kovář, Vít Baisa, and Miloš Jakubíček. “Sketch Engine for Bilingual Lexicography”. In: *International Journal of Lexicography* 29.3 (July 2016), pp. 339–352. issn: 0950-3846.

Bibliography V

[11] Guillaume Lample et al. “Unsupervised Machine Translation Using Monolingual Corpora Only”. In: *arXiv preprint arXiv:1711.00043* (2017).

[12] 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.

[13] Sebastian Ruder, Ivan Vulić, and Anders Søgaard. “A Survey of Cross-Lingual Word Embedding Models”. In: *J. Artif. Int. Res.* 65.1 (May 2019), pp. 569–630. issn: 1076-9757.

[14] P. Rychlý. “A Lexicographer-Friendly Association Score”. In: *RASLAN*. 2008.

Thank You for Your Attention!

MUNI
FACULTY
OF INFORMATICS