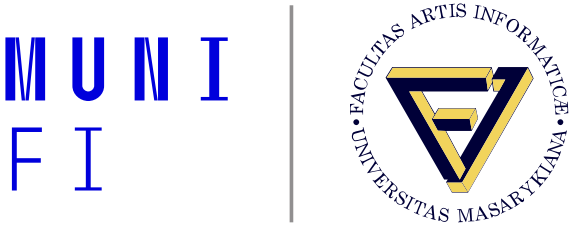


Lost in Morphology: Enhancing Bilingual Lexicon Induction with Lemmatisation

Michaela Denisová, Pavel Rychlý

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

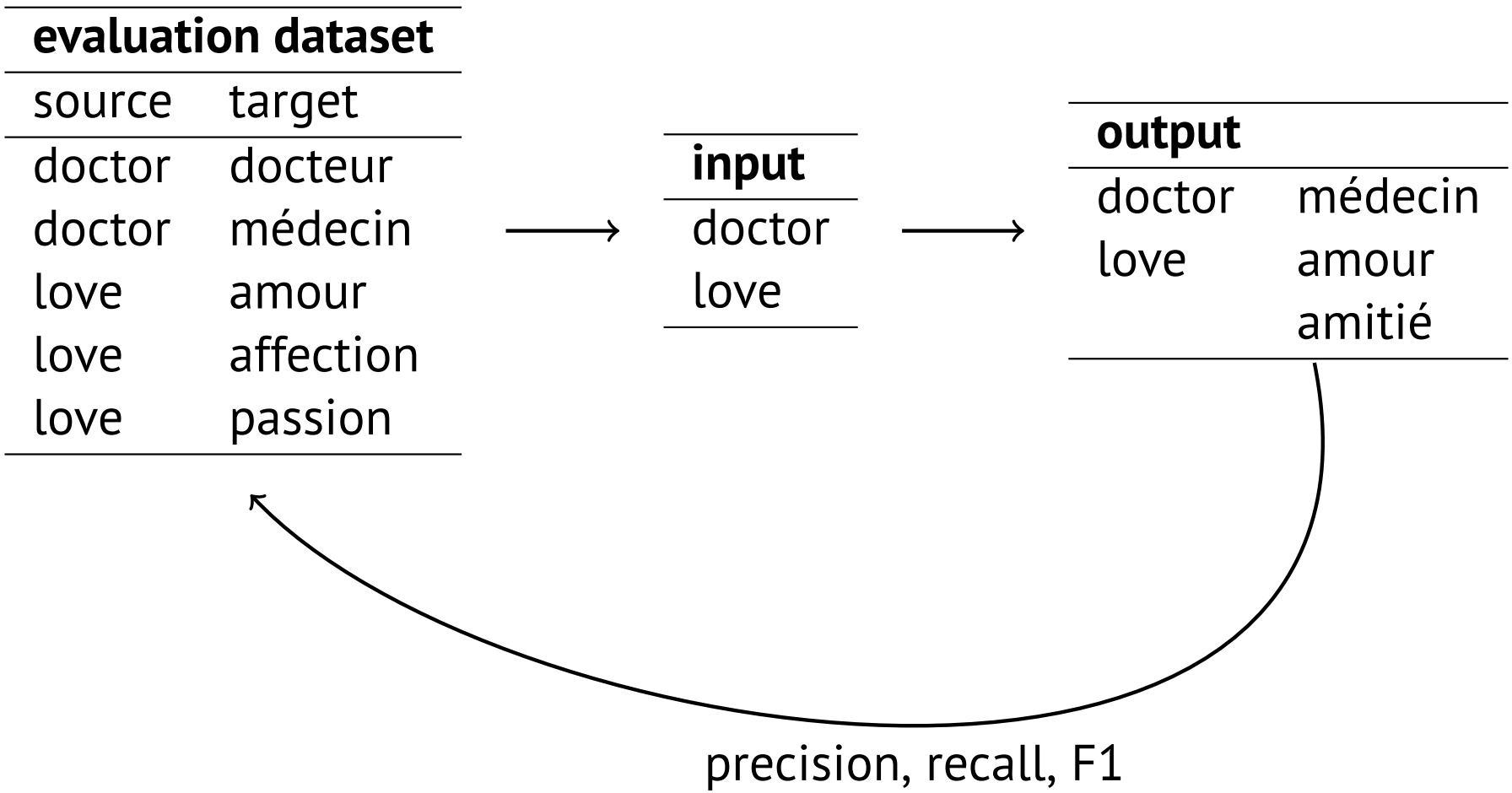
Natural Language Processing Centre

Faculty of Informatics, Masaryk University

December 5-7, 2025

Bilingual Lexicon Induction

- Aims to find the most suitable target word(s) for each source word



Cross-Lingual Word Embeddings

- Bilingual word vector representations, after mapping, similar words obtain similar vectors

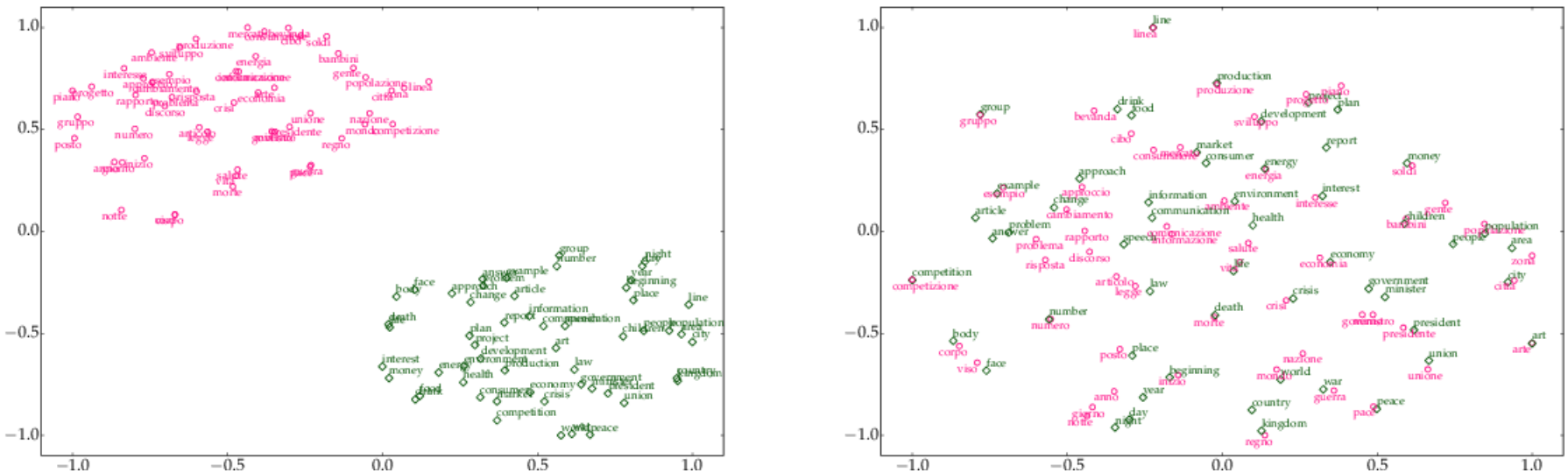
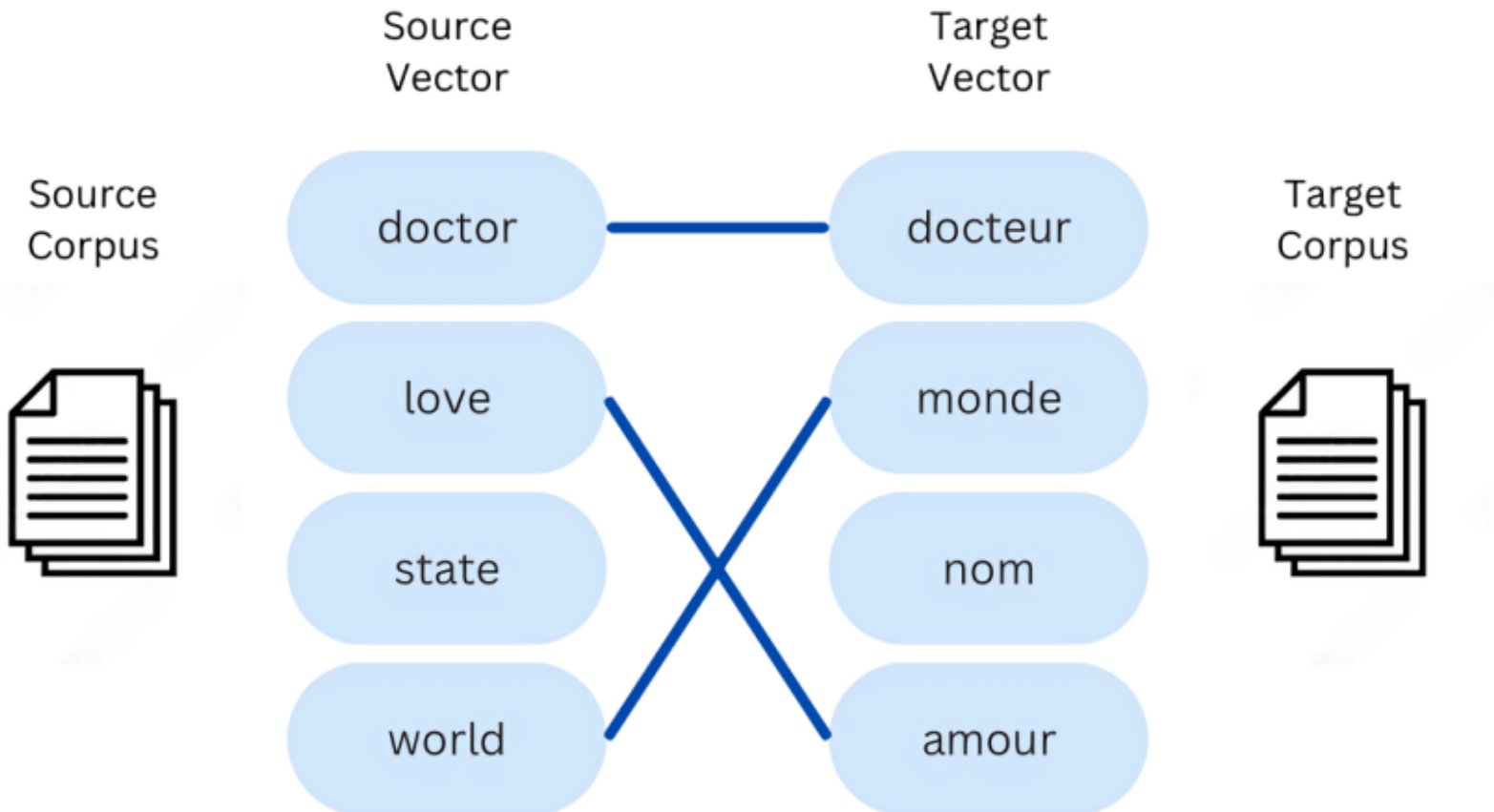


Figure: Monolingual (left) vs. joint cross-lingual (right) space [8]

Mapping


Source
Vector
Target
Vector
Source
Corpus
doctor
docteur
Target
Corpus
love
monde
state
nom
world
amour

Limitations in Morphology

- Word form is often preferred to the lemma
- Morphologically incomplete datasets

Source word	Top-5 nearest neighbours				
frieren	freezing	melt	freezes	thawing	**freeze**
laufen	running	swim	**run**	smoothly	going
schatten	shadows	**shadow**	silhouetted	glare	shadowed

Table: Top-5 target words generated from MUSE pre-trained model

Objectives

- Integrate 2 lemmatisation strategies into the BLI pipeline:
  - Pre-alignment
  - Post-retrieval
- Evaluate across language pairs with different levels of morphological complexity:
  - Italian-English; English-French (low)
  - Estonian-Slovak; Czech-Slovak (medium)
  - German-Hungarian (high)

Experimental Setup

- Sketch Engine monolingual embeddings [5]
- Cross-lingual models: MUSE [4], VecMap [1, 2], RCLS [6]
- Lemmatisers:
  - Pre-alignment: SketchEngine embeddings computed over word forms converted into their lemmas
  - Post-retrieval:
    - Italian-English: Python NLTK [3]
    - English-French: Lefff [9]
    - Estonian-Slovak and Czech-Slovak: Majka
    - German-Hungarian: HuSpaCy [7]

Results I.

P@1

P@1	it-en		en-fr		et-sk		cs-sk		de-hu	
	pre	post	pre	post	pre	post	pre	post	pre	post
MUSE-S	50.41	**62.26**	44.26	**45.76**	36.93	**38.56**	**87.48**	84.91	**27.88**	23.11
MUSE-I	49.58	**61.59**	44.66	**45.76**	37.09	**39.11**	**87.48**	84.91	22.61	**25.87**
MUSE-U	49.27	**62.19**	44.5	**45.02**	35.87	**35.93**	**87.57**	85.26	23.86	**24.37**
VM-S	50.1	**61.29**	**45.94**	45.91	40.74	**42.92**	**87.74**	85.78	**29.14**	28.89
VM-I	46.16	**59.65**	**47.47**	45.98	38.38	**39.74**	**88.27**	85.44	21.6	**23.86**
VM-U	46.57	**59.28**	44.26	**45.61**	**38.61**	38.56	**88.01**	85.44	21.6	**23.61**
RCLS	51.34	**65.02**	45.61	**48.04**	**39.75**	38.2	**87.66**	84.91	**24.87**	22.86

Results II.

R@1

R@1	it-en		en-fr		et-sk		cs-sk		de-hu	
	pre	post	pre	post	pre	post	pre	post	pre	post
MUSE-S	24.55	**42.19**	36.8	**41.4**	**32.33**	28.33	**67.26**	65.06	**12.86**	10.66
MUSE-I	24.15	**41.73**	37.13	**41.4**	**32.46**	28.73	**67.26**	65.06	10.42	**11.93**
MUSE-U	24.0	**42.14**	37.0	**40.73**	**31.4**	26.4	**67.33**	65.33	11.0	**11.23**
VM-S	24.4	**41.53**	38.13	**41.53**	**35.66**	31.53	**67.46**	65.73	**13.44**	13.32
VM-I	22.48	**40.42**	39.46	**41.6**	**33.6**	29.2	**67.86**	65.46	9.96	**11.0**
VM-U	22.68	**40.17**	36.8	**41.26**	**33.8**	28.33	**67.66**	65.46	9.96	**10.89**
RCLS	25.01	**44.06**	41.26	**43.46**	**34.8**	28.06	**67.4**	65.06	**11.47**	10.54

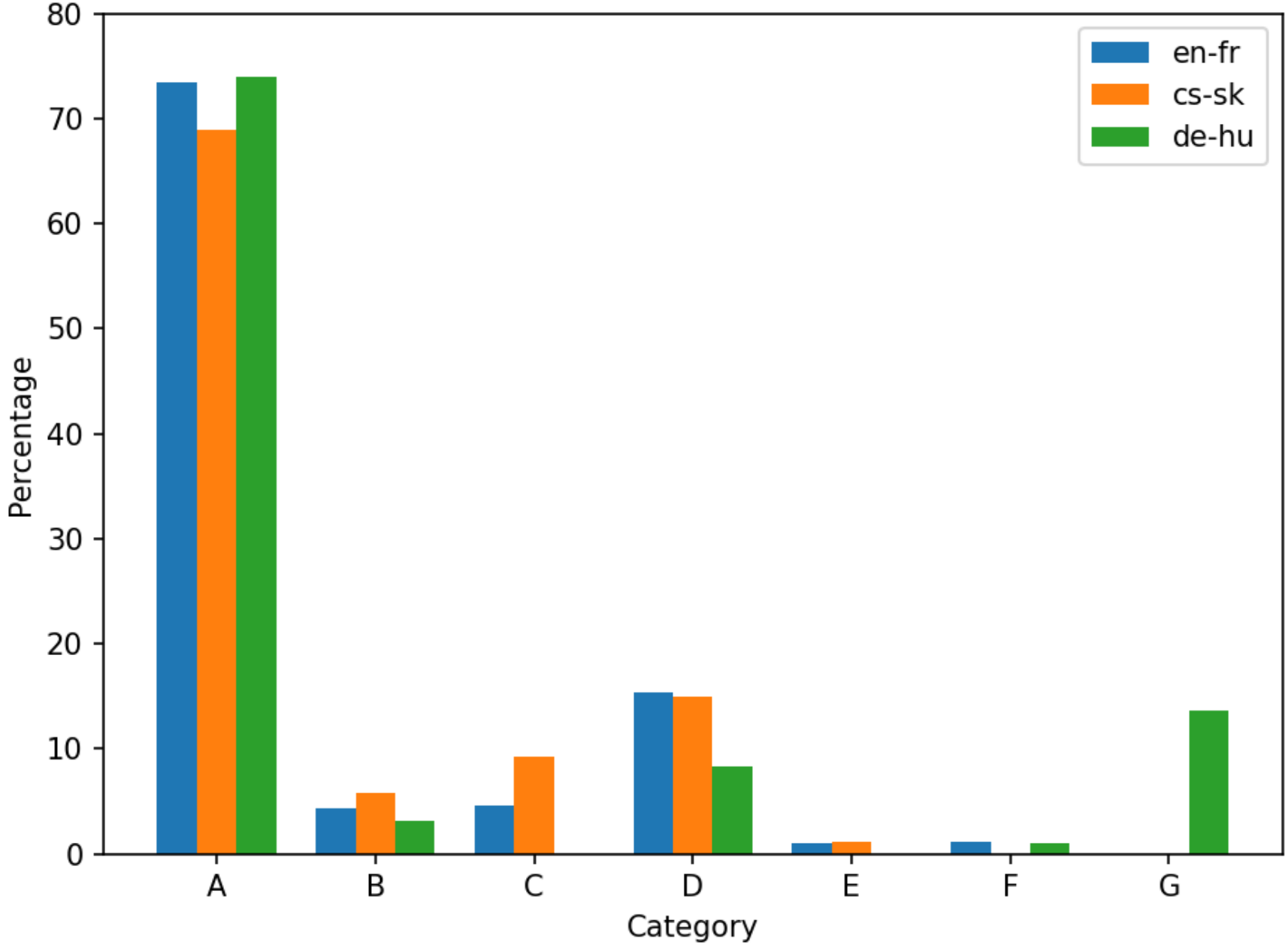
Results III.

P@1	it-en		en-fr		et-sk		cs-sk		de-hu	
	CI	Δ	CI	Δ	CI	Δ	CI	Δ	CI	Δ
MUSE-S	**[-0.074, -0.020]**	-0.046	**[-0.047, -0.006]**	-0.026	**[0.007, 0.054]**	+0.030	**[0.009, 0.047]**	+0.029	**[0.008, 0.089]**	+0.047
MUSE-I	**[-0.068, -0.016]**	-0.041	**[-0.039, -0.003]**	-0.022	[-0.003, 0.051]	+0.023	**[0.009, 0.047]**	+0.029	[-0.058, 0.013]	-0.024
MUSE-U	**[-0.075, -0.023]**	-0.050	**[-0.038, -0.002]**	-0.020	**[0.015, 0.066]**	+0.040	**[0.007, 0.045]**	+0.026	[-0.037, 0.034]	-0.000
VM-S	**[-0.068, -0.014]**	-0.039	**[-0.048, -0.002]**	-0.024	[-0.003, 0.047]	+0.022	**[0.005, 0.039]**	+0.023	[-0.034, 0.042]	+0.003
VM-I	**[-0.087, -0.035]**	-0.060	[-0.026, 0.010]	-0.008	**[0.005, 0.054]**	+0.029	**[0.012, 0.047]**	+0.030	[-0.047, 0.018]	-0.016
VM-U	**[-0.076, -0.025]**	-0.051	**[-0.052, -0.013]**	-0.032	**[0.014, 0.066]**	+0.040	**[0.010, 0.046]**	+0.028	[-0.045, 0.021]	-0.013
RCLS	**[-0.092, -0.039]**	-0.065	**[-0.104, -0.057]**	-0.080	**[0.047, 0.096]**	+0.071	**[0.011, 0.049]**	+0.031	[-0.013, 0.065]	+0.024

Figure: Paired t-tests for P@1 (Δ = A − B) between pre-alignment (A) and post-retrieval (B) lemmatisation approaches. Each Δ is the average over 7 random-seed runs; the 95% confidence interval CI is computed from 1,000 paired bootstrap resamples of those seven differences. Values marked in **bold** are statistically significant.

Error Analysis

Type	SRC	Pre-alignment	Post-retrieval	EN	Explanation
A	strašpytel	strachopud	bojko	coward	synonym
	carpenter	menusier	charpentier	-	
	accidentally	mégarde	accidentellement	-	
	exzellent	elsőrangú	kiváló	excellent	
B	luštěnina	strukovina	-	legume	low frequency
	později	-	neskôr	later	
	wednesday	-	aprèm (eng. *afternoon*)	-	
	zimbabwe	-	-	-	
	zerstörer	-	-	destroyer	
C	fandit	fandiť	fandiť*	cheer on	bad
	nechat	nechať	nenechať (negation)	let	lemmatisation
	atomic	atomique	atomiques	-	
	admission	admissions	admission	-	
D	klokan	kengura	slon (eng. *elephant*)	kangaroo	related words
	patnáctý	pätnásty	dvanásty (eng. *twelfth*)	fifteenth	
	tyler	jesse	dave	-	
	alpha	alpha	béta	-	
E	léto	rok (eng. *year*)	leto	summer	polysemy
	pole	mât (eng. *bar*)	pôle	north pole vs ski pole	
F	advocacy	advocacy	plaidoyer	-	English words
	dream	dreams	yesterday	-	
G	anfangen	kedz	**el**kedz	to start	mixed affixes
	beratung	tanácsadás	tanácsadási	advice	
	lernen	**meg**tanul	tanul	to learn	


80
70
60
50
40
30
20
10
0
Percentage
en-fr
cs-sk
de-hu
A
B
C
D
E
F
G
Category

Summary

- We integrated two lemmatisation strategies into the BLI pipeline to address the issue with mapping word forms
- Pre-alignment - medium complexity
- Post-retrieval - low complexity
- For the most complex language pair, both strategies were equally efficient

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. DOI: 10.18653/v1/P18-1073.

[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. DOI: https://doi.org/10.1609/aaai.v32i1.11992.

Bibliography II

[3] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[4] Alexis Conneau et al. "Word translation without parallel data.". In: *ICLR* (2017).

[5] Ondřej Herman. "Precomputed Word Embeddings for 15+ Languages". In: *Proceedings of the Fifteenth Workshop on Recent Advances in Slavonic Natural Languages Processing (RASLAN)*. 2021, pp. 41–46.

Bibliography III

[6] 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. DOI: 10.18653/v1/D18-1330.

[7] György Orosz et al. “Advancing Hungarian Text Processing with HuSpaCy: Efficient and Accurate NLP Pipelines”. In: *Text, Speech, and Dialogue*. Ed. by Kamil Ekštein, František Pártl, and Miloslav Konopík. Cham: Springer Nature Switzerland, 2023, pp. 58–69. ISBN: 978-3-031-40498-6. DOI: 10.1007/978-3-031-40498-6_6.

Bibliography IV

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

[9] Benoît Sagot. “The Lefff, a Freely Available and Large-coverage Morphological and Syntactic Lexicon for French”. In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA), 2010.

Thank You for Your Attention!

MUNI
FACULTY
OF INFORMATICS