Don’t Forget Cheap Training Signals Before Building Unsupervised Bilingual Word Embeddings

Silvia Severini, Viktor Hangya, Masoud Jalili Sabet, Alexander Fraser, Hinrich Schütze

LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

BUCC@LREC 2022

Outline

1. Introduction
2. Background
3. Contribution
4. Approach
5. Evaluation
6. Conclusion

Outline

Introduction

- **Bilingual Word Embeddings** (BWEs) can be built effectively even for low-resource settings

* (Zhang et al., 2017; Lample et al., 2018; Artetxe et al., 2018; Alvarez-Melis and Jaakkola, 2018; Chen and Cardie, 2018; Hoshen and Wolf, 2018; Mohiuddin and Joty, 2019; Alaux et al., 2019; Dou et al., 2020; Grave et al., 2019; Li et al., 2020).

Introduction

- **Bilingual Word Embeddings** (BWEs) can be built effectively even for low-resource settings
- Various unsupervised methods have been proposed relying on the assumption that embedding spaces are isomorphic*

* (Zhang et al., 2017; Lample et al., 2018; Artetxe et al., 2018; Alvarez-Melis and Jaakkola, 2018; Chen and Cardie, 2018; Hoshen and Wolf, 2018; Mohiuddin and Joty, 2019; Alaux et al., 2019; Dou et al., 2020; Grave et al., 2019; Li et al., 2020).

Introduction

- **Bilingual Word Embeddings** (BWEs) can be built effectively even for low-resource settings
- Various unsupervised methods have been proposed relying on the assumption that embedding spaces are isomorphic*

…but

* (Zhang et al., 2017; Lample et al., 2018; Artetxe et al., 2018; Alvarez-Melis and Jaakkola, 2018; Chen and Cardie, 2018; Hoshen and Wolf, 2018; Mohiuddin and Joty, 2019; Alaux et al., 2019; Dou et al., 2020; Grave et al., 2019; Li et al., 2020).

Introduction

- **Bilingual Word Embeddings** (BWEs) can be built effectively even for low-resource settings
- Various unsupervised methods have been proposed relying on the assumption that embedding spaces are isomorphic*

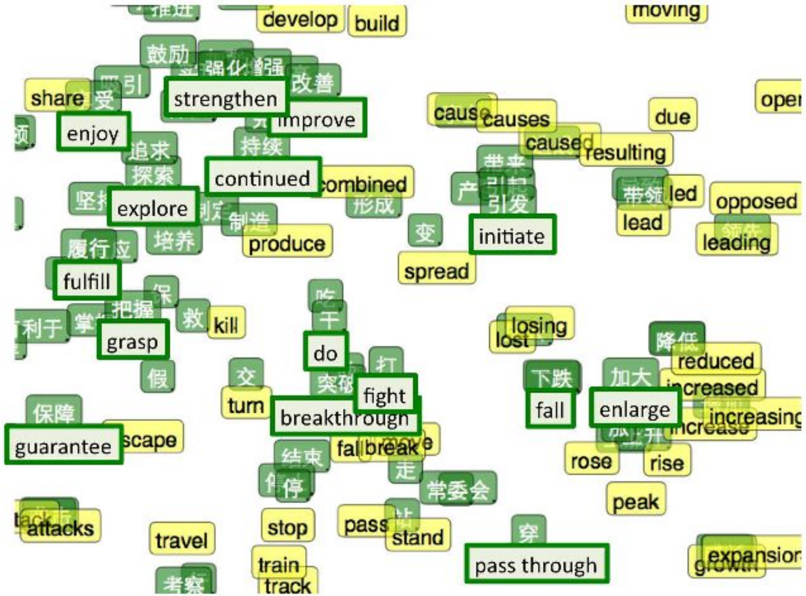
…but

- Many methods fail for distant language pairs (Vulic et al. (2019))
- They don't compare with straightforward baselines

* (Zhang et al., 2017; Lample et al., 2018; Artetxe et al., 2018; Alvarez-Melis and Jaakkola, 2018; Chen and Cardie, 2018; Hoshen and Wolf, 2018; Mohiuddin and Joty, 2019; Alaux et al., 2019; Dou et al., 2020; Grave et al., 2019; Li et al., 2020).

Questions

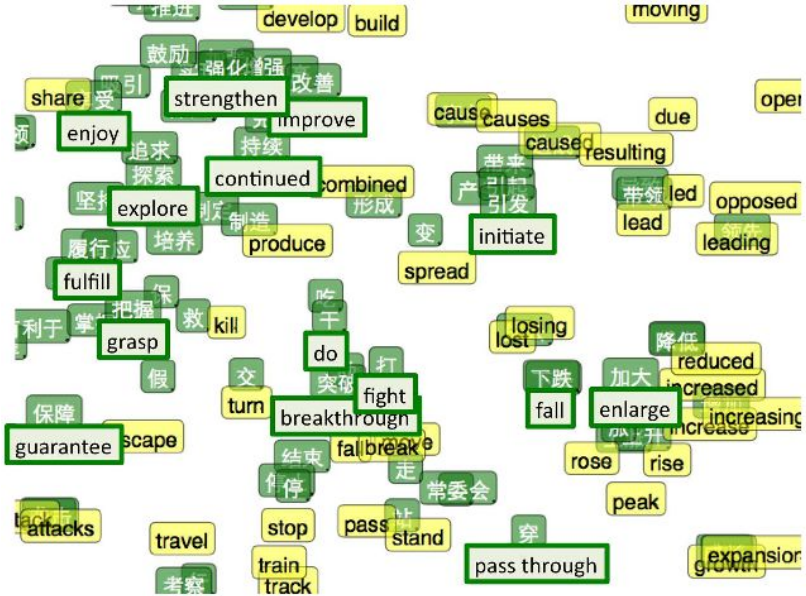
- Do we really need **unsupervised** approaches for building Bilingual Word Embeddings?



Questions

- Do we really need **unsupervised** approaches for building Bilingual Word Embeddings?

- If yes, aren't we missing any **baselines**?



Outline

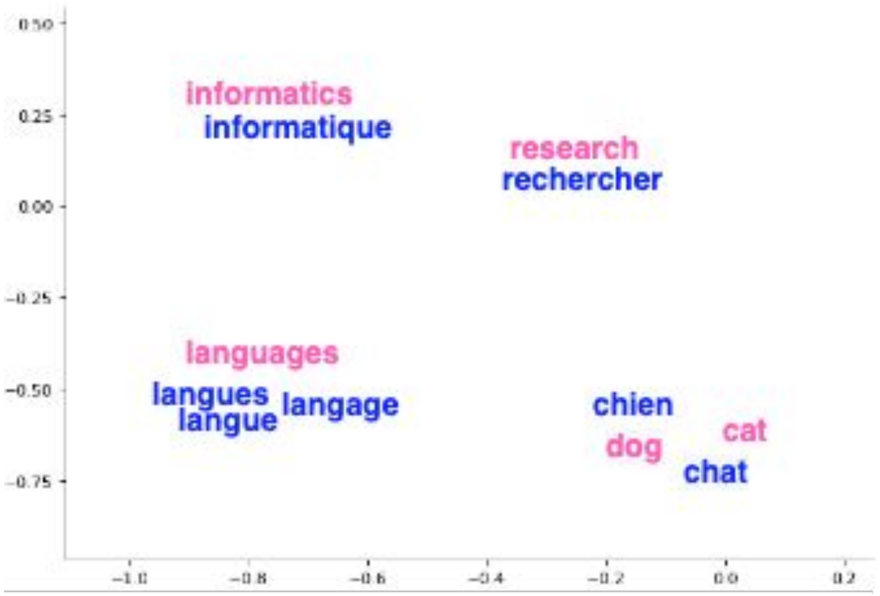
1. Introduction
2. Background
3. Contribution
4. Approach
5. Evaluation
6. Conclusion

Outline

1. Introduction
2. Background
3. Contribution
4. Approach
5. Evaluation
6. Conclusion

Bilingual Word Embeddings

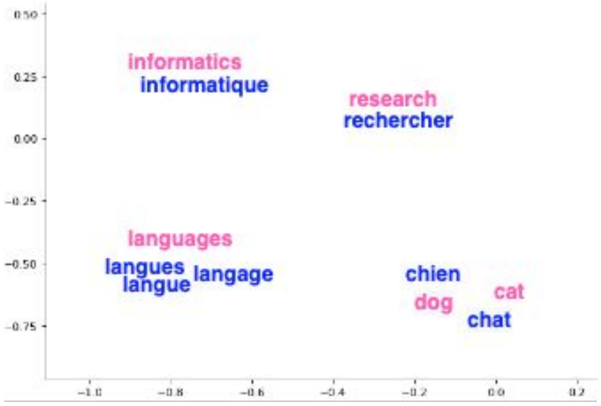
- They represent lexicons of different languages in a shared embedding space



Bilingual Word Embeddings

- They represent lexicons of different languages in a shared embedding space

- They are essential for supporting semantic and knowledge transfers in a variety of **cross-lingual** NLP tasks (Machine translation, Bilingual NER, …)

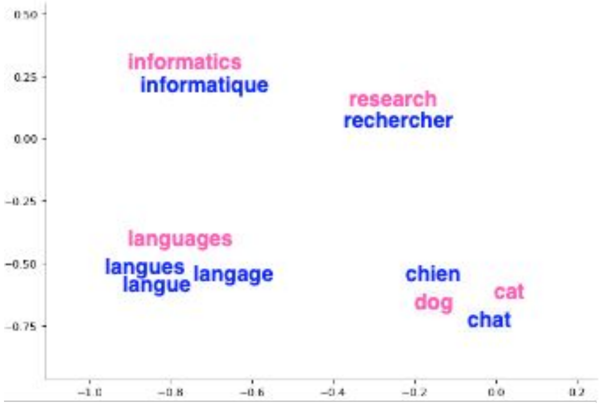


Bilingual Word Embeddings

- They represent lexicons of different languages in a shared embedding space

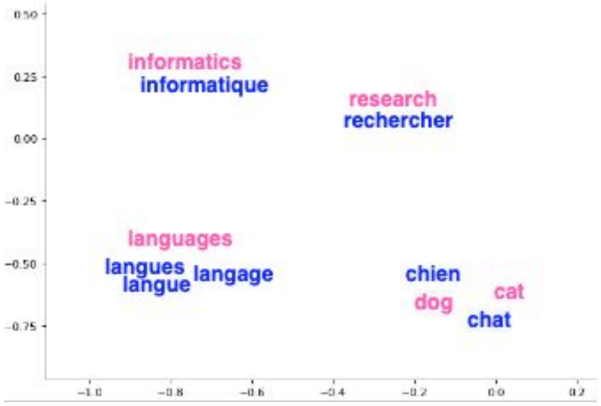
- They are essential for supporting semantic and knowledge transfers in a variety of **cross-lingual** NLP tasks (Machine translation, Bilingual NER, …)

- They can be built effectively even when only a **small** seed lexicon is available



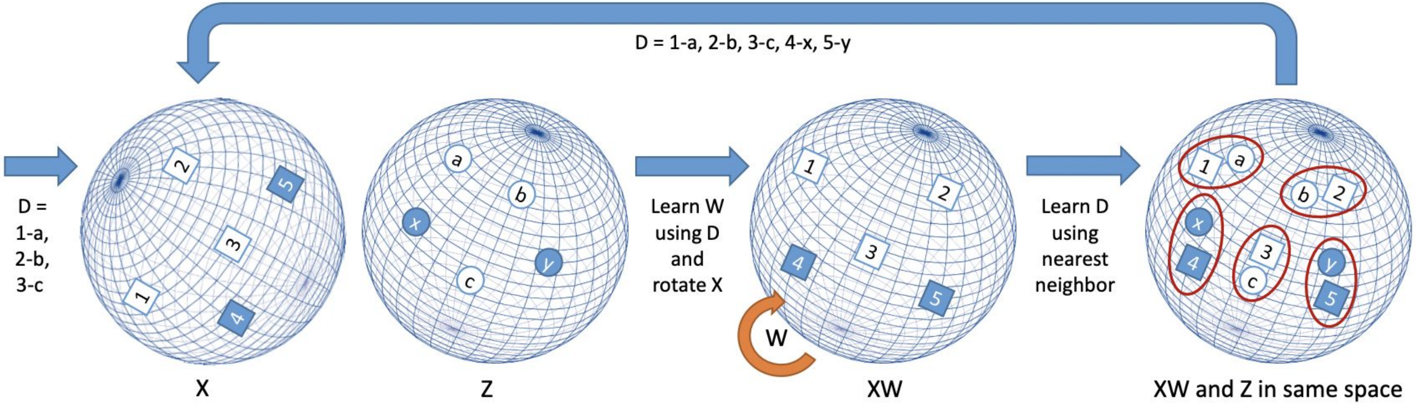
Bilingual Word Embeddings

- They represent lexicons of different languages in a shared embedding space
- They are essential for supporting semantic and knowledge transfers in a variety of **cross-lingual** NLP tasks (Machine translation, Bilingual NER, …)
- They can be built effectively even when only a **small** seed lexicon is available
- They work even for low-resource language not covered by PLMs



Semi-supervised mapping

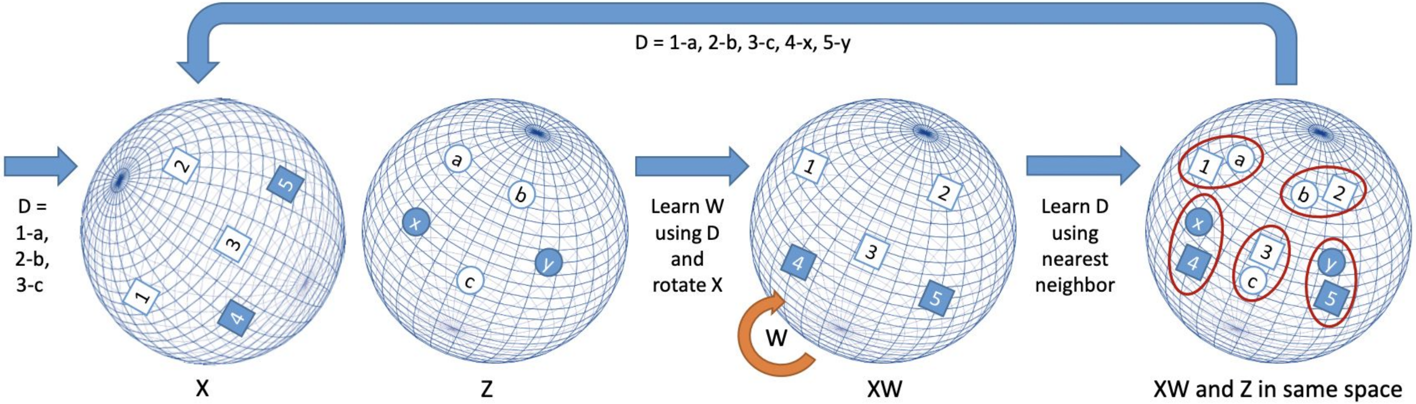
- VecMap: build BWE from noisy lexicon and monolingual embeddings



Artetxe, Mikel, Gorka Labaka, and Eneko Agirre. "Learning principled bilingual mappings of word embeddings while preserving monolingual invariance." *EMNLP,* 2016.

Semi-supervised mapping

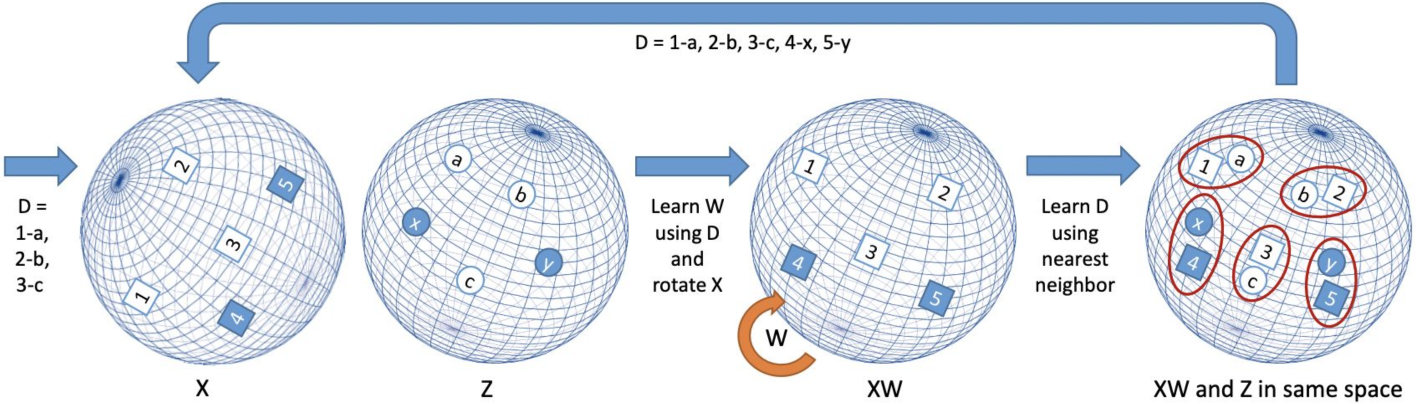
- VecMap: build BWE from noisy lexicon and monolingual embeddings
- VecMap iterates over two steps: embedding mapping and dictionary induction.



Artetxe, Mikel, Gorka Labaka, and Eneko Agirre. "Learning principled bilingual mappings of word embeddings while preserving monolingual invariance." *EMNLP,* 2016.

Semi-supervised mapping

- VecMap: build BWE from noisy lexicon and monolingual embeddings
- VecMap iterates over two steps: embedding mapping and dictionary induction.
- Semi-supervised approach performs well with small and **noisy seed lexicons** by iteratively refining them.



Artetxe, Mikel, Gorka Labaka, and Eneko Agirre. "Learning principled bilingual mappings of word embeddings while preserving monolingual invariance." *EMNLP,* 2016.

Outline

1. Introduction
2. Background
3. Contribution
4. Approach
5. Evaluation
6. Conclusion

Outline

1. Introduction
2. Background
3. Contribution
4. Approach
5. Evaluation
6. Conclusion

Contribution

- We test identical word pairs on multiple language pairs with **distinct scripts**, including pairs using **distinct numerals**

Contribution

- We test identical word pairs on multiple language pairs with **distinct scripts**, including pairs using **distinct numerals**

- We propose to strengthen identical pairs by extending them with further easily accessible pairs based on **romanization** and edit distance

Contribution

- We test identical word pairs on multiple language pairs with **distinct scripts**, including pairs using **distinct numerals**
- We propose to strengthen identical pairs by extending them with further easily accessible pairs based on **romanization** and edit distance
- We focus on distant language pairs having distinct scripts for many of which unsupervised approaches have failed or had very poor performance so far

Contribution

- We test identical word pairs on multiple language pairs with **distinct scripts**, including pairs using **distinct numerals**.
- We propose to strengthen identical pairs by extending them with further easily accessible pairs based on **romanization** and edit distance
- We focus on distant language pairs having distinct scripts for many of which unsupervised approaches have failed or had very poor performance so far
- Our work calls into question, at least for **BDI**, the strong trend toward unsupervised approaches in recent literature

Outline

1. Introduction
2. Background
3. Contribution
4. Approach
5. Evaluation
6. Conclusion

Outline

1. Introduction
2. Background
3. Contribution
4. **Approach**
5. Evaluation
6. Conclusion

Unsupervised pair extraction

- Extract seed lexicon for mapping approaches

Unsupervised pair extraction

- Extract seed lexicon for mapping approaches
- No need for labeled data -> Applicable to a wide range of languages

Unsupervised pair extraction

- Extract seed lexicon for mapping approaches
- No need for labeled data -> Applicable to a wide range of languages
- Two approaches:
  a. **ID**: Identical pair approach for different scripts
  b. **ID++** : Unsupervised romanization-based augmentation

ID: Identical pairs for distinct scripts

- Available in **large** quantities:
  - even for distinct scripts and with different numerals

Lang	ID	Lang	ID	Lang	ID
ko-th*	17K	ko-he*	11K	he-th*	15K
en-zh*	62K	en-bn*	31K	en-ar*	19K
en-th	46K	en-hi*	30K	en-ru	18K
en-ja	43K	en-ta*	23K	en-he*	17K
en-el	35K	en-kn*	21K	en-ko*	15K
en-fa*	32K				

Life - ಜೀವನ
Language - ಸಮ್ಮೇಳನ
Conference - ಭಾಷೆ

Søgaard, A., Ruder, S., and Vulic´, I. (2018). On the limitations of unsupervised bilingual dictionary induction.

ID: Identical pairs for distinct scripts

- Available in **large** quantities:
  - even for distinct scripts and with different numerals
- Examples:
  - Punctuation marks and digits
  - Non-transliterated named entities written in the Latin script
  - English words (assumingly words of a title) which were not translated in the non-English languages

Lang	ID	Lang	ID	Lang	ID
ko-th*	17K	ko-he*	11K	he-th*	15K
en-zh*	62K	en-bn*	31K	en-ar*	19K
en-th	46K	en-hi*	30K	en-ru	18K
en-ja	43K	en-ta*	23K	en-he*	17K
en-el	35K	en-kn*	21K	en-ko*	15K
en-fa*	32K				

Søgaard, A., Ruder, S., and Vulic´, I. (2018). On the limitations of unsupervised bilingual dictionary induction.

Rom : Unsupervised pair extraction

- Exploit the concept of **transliteration** and orthographic similarity to find a cheap signal between languages

Rom : Unsupervised pair extraction

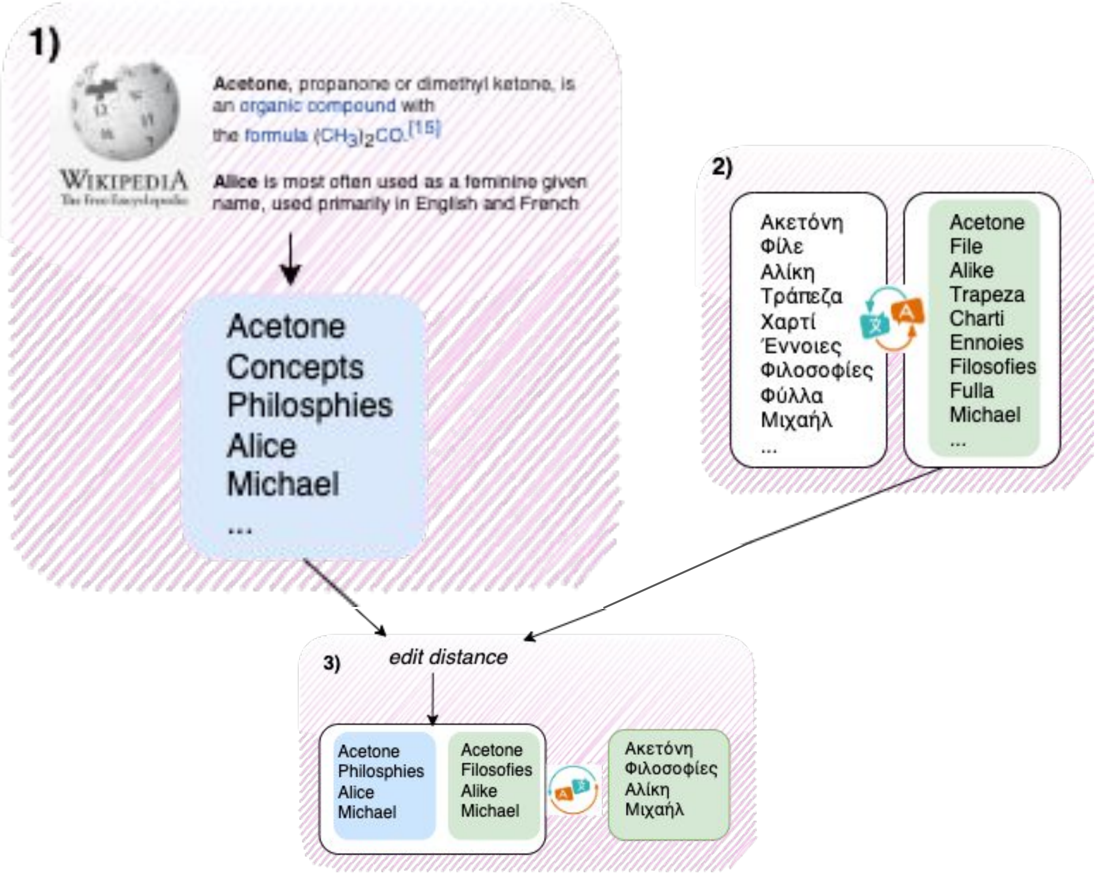
- Exploit the concept of **transliteration** and orthographic similarity to find a cheap signal between languages

"Transliteration is a type of conversion of a text from one script to another that involves swapping letters."

Greek	***English***	***Translit***
Ελληνική Δημοκρατία	*Hellenic Republic*	*Ellēnikē Dēmokratia*
Ελευθερία	*Freedom*	*Eleutheria*

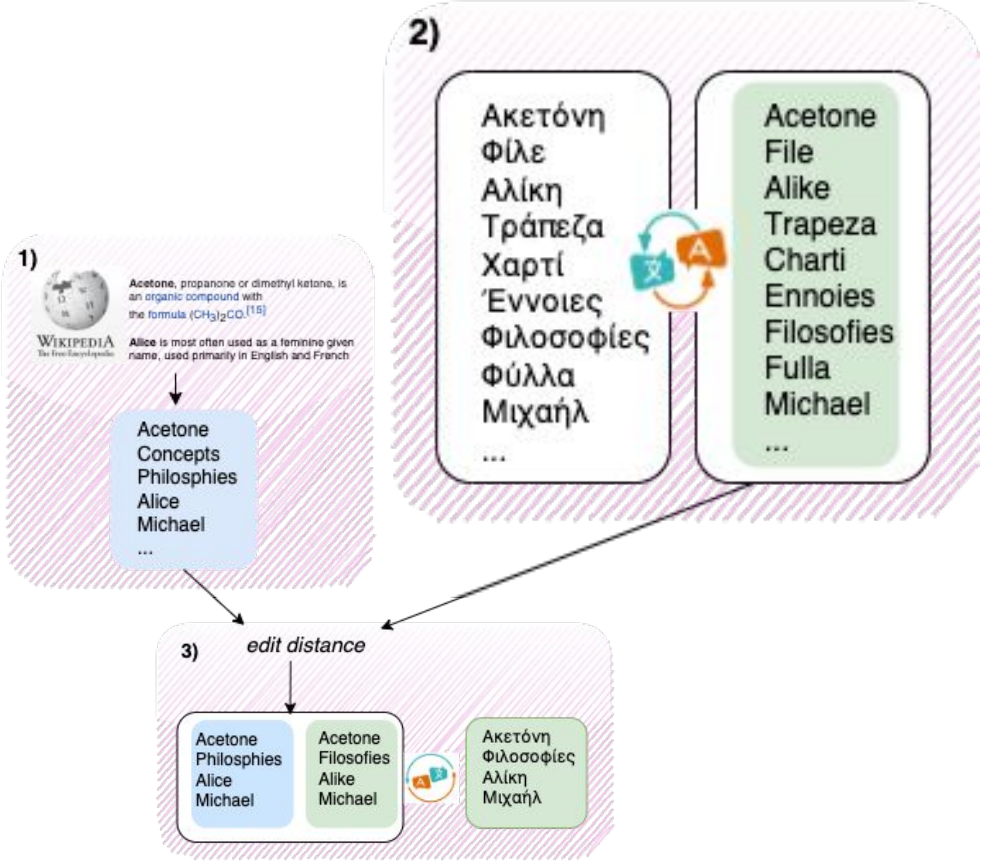
Rom : Unsupervised pair extraction

1) Source candidate extraction



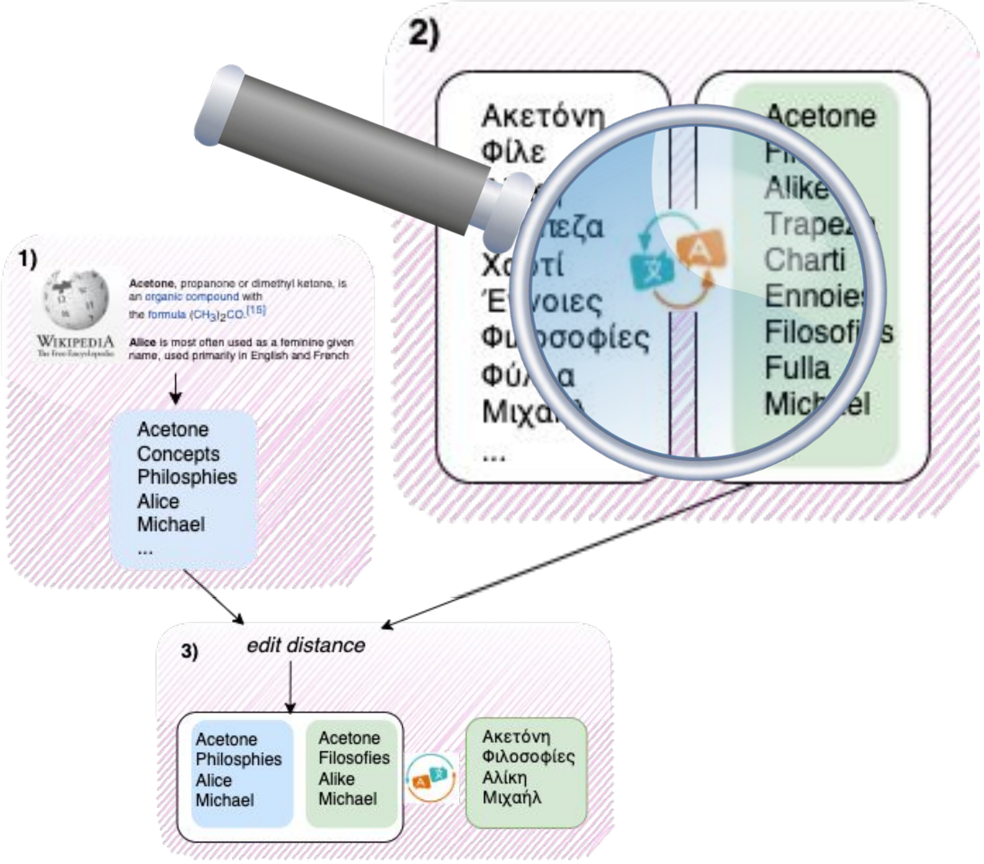
Rom : Unsupervised pair extraction

1) Source candidate extraction
2) Target candidate extraction



Rom : Unsupervised pair extraction

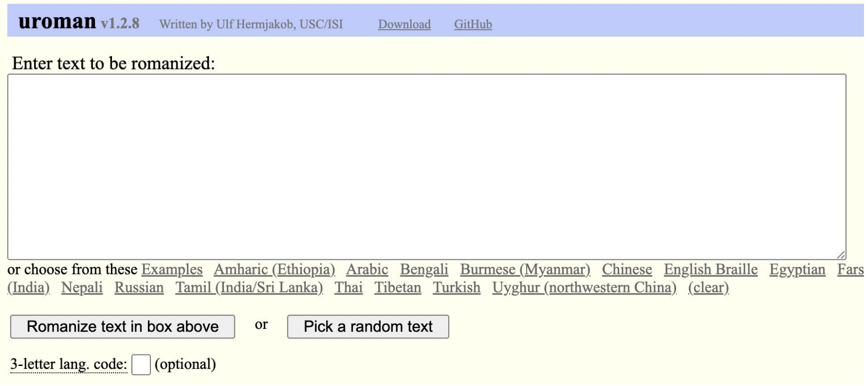
1) Source candidate extraction
2) Target candidate extraction



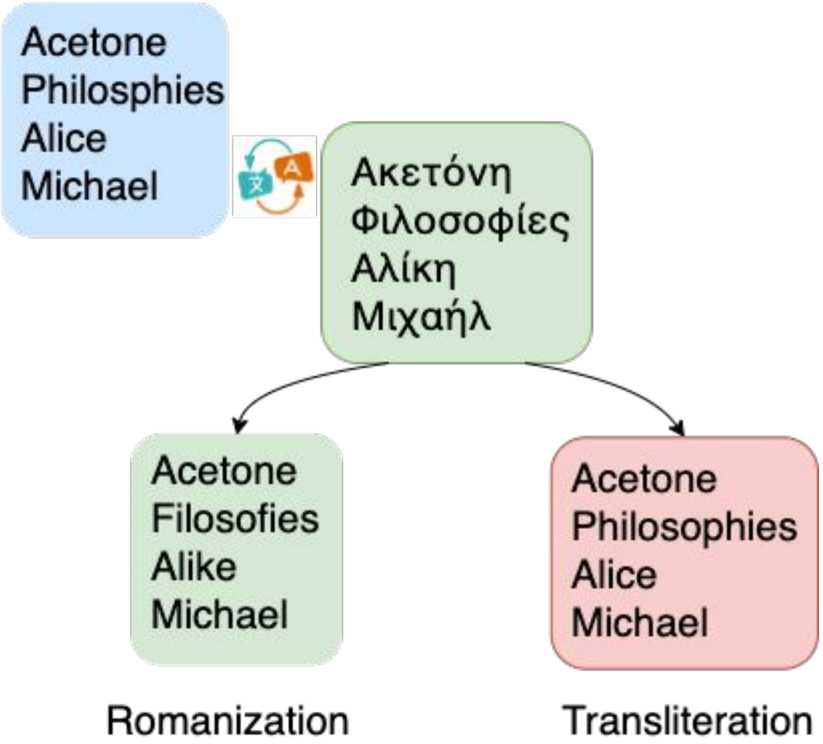
Romanization vs Transliteration

- Uroman romanizer:

"*uroman* is a *universal romanizer*. It converts text in any script to the Latin alphabet."

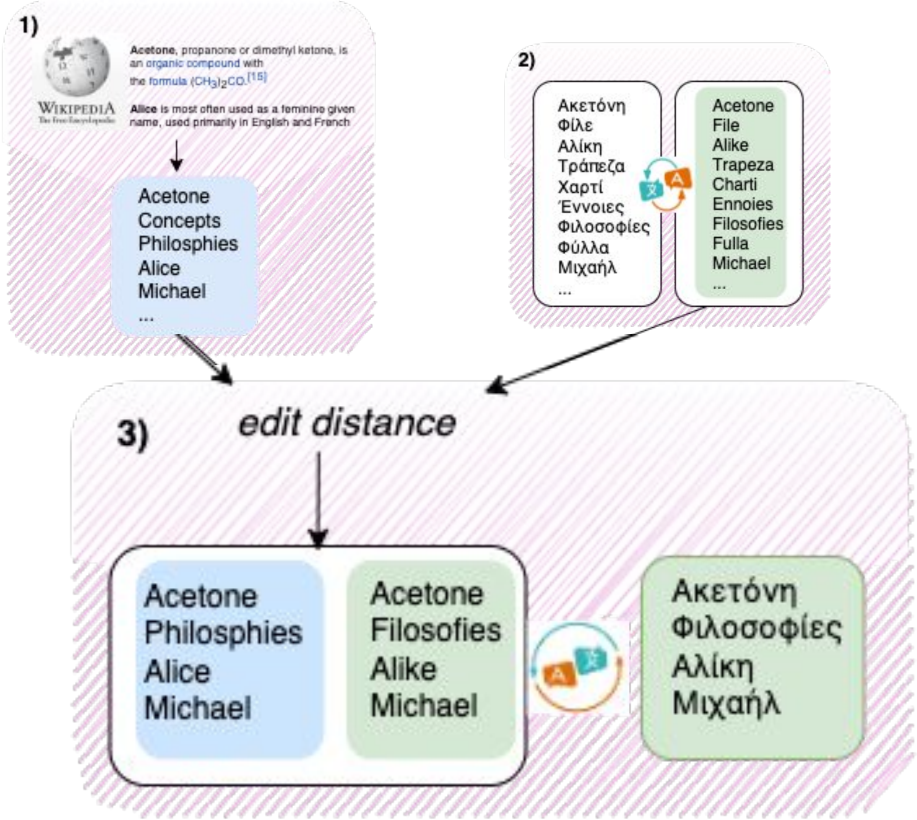


https://github.com/isi-nlp/uroman



Rom : Unsupervised pair extraction

1) Source candidate extraction
2) Target candidate extraction
3) Candidate matching



OOVs analysis

	MUSE	ID	Romanized
en-th	6,799	46,653	10,721 / 53,804
en-ja	7,135	43,556	11,488 / 118,626
en-kn	1,552	21,090	12,888 / 59,207
en-ta	8,091	23,538	5,987 / 120,836
en-zh	8,728	62,289	6,360 / 41,829
en-ar	11,571	19,275	4,773 / 61,031
en-hi	8,704	30,502	16,180 / 73,553
en-ru	10,887	18,663	9,913 / 301,698
en-el	10,662	35,270	20,740 / 150,472
en-fa	8,869	32,866	10,226 / 85,210
en-he	9,634	17,012	4,005 / 40,258
en-bn	8,467	31,954	10,721 / 53,804
en-ko	7,999	15,518	9956 / 134156

OOVs analysis

	MUSE	ID	Romanized
en-th	6,799	46,653	10,721 / 53,804
en-ja	7,135	43,556	11,488 / 118,626
en-kn	1,552	21,090	12,888 / 59,207
en-ta	8,091	23,538	5,987 / 120,836
en-zh	8,728	62,289	6,360 / 41,829
en-ar	11,571	19,275	4,773 / 61,031
en-hi	8,704	30,502	16,180 / 73,553
en-ru	10,887	18,663	9,913 / 301,698
en-el	10,662	35,270	20,740 / 150,472
en-fa	8,869	32,866	10,226 / 85,210
en-he	9,634	17,012	4,005 / 40,258
en-bn	8,467	31,954	10,721 / 53,804
en-ko	7,999	15,518	9956 / 134156

Outline

1. Introduction
2. Background
3. Contribution
4. Approach
5. Evaluation
6. Conclusion

Outline

1. Introduction
2. Background
3. Contribution
4. Approach
5. Evaluation
6. Conclusion

Evaluation

- **Bilingual Dictionary Induction** task:
  - **Goal**: generate translations in the target language of the source word in the source language.
  - Given a BWEs representing two words in two languages, create n-best list by taking the top n words with the closest representranslation to the source word according to the cosine distance

Evaluation

- **Bilingual Dictionary Induction** task:
  - **Goal**: generate translations in the target language of the source word in the source language.
  - Given a BWEs representing two words in two languages, create n-best list by taking the top n words with the closest representranslation to the source word according to the cosine distance

- acc@1 scores calculated by the MUSE evaluation tool

Results - low-resource

	en-th	en-ja	en-kn	en-ta	en-zh
Unsupervised					
Artetxe et al. (2018)	0.00	0.96	0.00	0.07	0.07
Grave et al. (2019)	0.00	0.48	0.00	0.07	0.00
Mohiuddin and Joty (2019)	0.00	0.00	0.00	0.00◇	0.00
Semi-supervised (Artetxe et al., 2018)					
ID	<u>24.40</u>	48.87	22.03	17.93	<u>37.00</u>
Rom.	23.33	48.46	22.90	18.00	0.27
ID++	23.47	<u>49.14</u>	<u>24.23</u>	18.20	35.00
MUSE	24.33	48.73	23.78	<u>18.80</u>	36.53

Results - low-resource

	en-th	en-ja	en-kn	en-ta	en-zh
Unsupervised					
Artetxe et al. (2018)	0.00	0.96	0.00	0.07	0.07
Grave et al. (2019)	0.00	0.48	0.00	0.07	0.00
Mohiuddin and Joty (2019)	0.00	0.00	0.00	0.00$^{\diamond}$	0.00
Semi-supervised (Artetxe et al., 2018)					
ID	<u>24.40</u>	48.87	22.03	17.93	<u>37.00</u>
Rom.	23.33	48.46	22.90	18.00	0.27
ID++	23.47	<u>49.14</u>	<u>24.23</u>	18.20	35.00
MUSE	24.33	48.73	23.78	<u>18.80</u>	36.53

Results - low-resource

	en-th	en-ja	en-kn	en-ta	en-zh
Unsupervised					
Artetxe et al. (2018)	0.00	0.96	0.00	0.07	0.07
Grave et al. (2019)	0.00	0.48	0.00	0.07	0.00
Mohiuddin and Joty (2019)	0.00	0.00	0.00	$0.00^{\diamond}$	0.00
Semi-supervised (Artetxe et al., 2018)					
ID	<u>24.40</u>	48.87	22.03	17.93	<u>37.00</u>
Rom.	23.33	48.46	22.90	18.00	0.27
ID++	23.47	<u>49.14</u>	<u>24.23</u>	18.20	35.00
MUSE	24.33	48.73	23.78	<u>18.80</u>	36.53

Results - low-resource

	en-th	en-ja	en-kn	en-ta	en-zh
Unsupervised					
Artetxe et al. (2018)	0.00	0.96	0.00	0.07	0.07
Grave et al. (2019)	0.00	0.48	0.00	0.07	0.00
Mohiuddin and Joty (2019)	0.00	0.00	0.00	$0.00^{\diamond}$	0.00
Semi-supervised (Artetxe et al., 2018)					
ID	24.40	48.87	22.03	17.93	37.00
Rom.	23.33	48.46	22.90	18.00	0.27
ID++	23.47	49.14	24.23	18.20	35.00
MUSE	24.33	48.73	23.78	18.80	36.53

Results - low-resource

	en-th	en-ja	en-kn	en-ta	en-zh
Unsupervised					
Artetxe et al. (2018)	0.00	0.96	0.00	0.07	0.07
Grave et al. (2019)	0.00	0.48	0.00	0.07	0.00
Mohiuddin and Joty (2019)	0.00	0.00	0.00	$0.00^{\diamond}$	0.00
Semi-supervised (Artetxe et al., 2018)					
ID	24.40	48.87	22.03	17.93	37.00
Rom.	23.33	48.46	22.90	18.00	0.27
ID++	23.47	49.14	24.23	18.20	35.00
MUSE	24.33	48.73	23.78	18.80	36.53

Results - high-resource

	Unsup.	ID	Rom.	ID++	MUSE
en-ar	36.30	40.27	39.33	40.20	39.87
en-hi	40.20	40.47	39.60	40.20	40.33
en-ru	44.80	49.13	48.87	49.53	48.80
en-el	47.90	47.87	48.00	48.27	48.00
en-fa	36.70	37.67	36.80	37.67	38.00
en-he	44.60	44.47	44.53	44.67	45.00
en-bn	18.20	19.87	19.80	20.13	21.60
en-ko	19.80	27.92	28.40	28.81	28.94

Results - high-resource

	Unsup.	ID	Rom.	ID++	MUSE
en-ar	36.30	40.27	39.33	40.20	39.87
en-hi	40.20	40.47	39.60	40.20	40.33
en-ru	44.80	49.13	48.87	49.53	48.80
en-el	47.90	47.87	48.00	48.27	48.00
en-fa	36.70	37.67	36.80	37.67	38.00
en-he	44.60	44.47	44.53	44.67	45.00
en-bn	18.20	19.87	19.80	20.13	21.60
en-ko	19.80	27.92	28.40	28.81	28.94

MUSE without Proper Nouns

			Baselines		Our		
			Unsup	Semi-sup. MUSE	Semi-supervised ID	Rom.	ID++
1	en-th	→	0.00	**27.21**	**27.13**	26.35	26.11
		←	0.00	18.93	19.83	18.25	19.83
2	en-ja	→	0.71	**46.15**	45.04	46.31	**46.39**
		←	0.56	**39.14**	38.86	**40.73**	39.52
3	en-kn	→	0.00	**23.78***	22.03	22.90	**24.23**
		←	0.00	**41.25***	**43.04**	42.50	41.79
4	en-ta	→	0.08	**20.12**	19.35	18.97	**19.43**
		←	0.08	**24.60**	24.60	23.71	**25.00**
5	en-zh	→	0.07	**37.34**	**38.14**	0.07	35.74
		←	0.00	**32.48**	**34.83**	0.00	32.48

Kementchedjhieva, Y., Hartmann, M., and Søgaard, A. (2019). Lost in evaluation: Misleading benchmarks for bilingual dictionary induction. *arXiv preprint arXiv:1909.05708*.

Non-English centric evaluation

- PanLex dictionaries

	Unsup.	ID	Rom.	ID++	PanLex
th-ko	0.00	2.81	<u>3.37</u>	3.09	2.95
th-he	0.00	<u>9.75</u>	0.00	8.86	10.13
ko-th	0.00	<u>15.90</u>	14.23	15.26	14.36
ko-he	14.62	15.68	<u>16.08</u>	16.00	15.11
he-th	0.00	16.42	0.00	<u>16.54</u>	17.90
he-ko	14.30	<u>15.39</u>	15.15	15.09	16.06

Outline

1. Introduction
2. Background
3. Contribution
4. Approach
5. Evaluation
6. Conclusion

Outline

1. Introduction
2. Background
3. Contribution
4. Approach
5. Evaluation
6. Conclusion

Conclusion

- We exploited identical pairs that **surprisingly** appear in corpora of **distinct scripts**

Conclusion

- We exploited identical pairs that **surprisingly** appear in corpora of **distinct scripts**

- We combined them with a simple method to extract the initial hypothesis set via **romanization** and edit distance

Conclusion

- We exploited identical pairs that **surprisingly** appear in corpora of **distinct scripts**
- We combined them with a simple method to extract the initial hypothesis set via **romanization** and edit distance
- With both approaches, we obtained results that are competitive with high-quality dictionaries

Conclusion

- We exploited identical pairs that **surprisingly** appear in corpora of **distinct scripts**
- We combined them with a simple method to extract the initial hypothesis set via **romanization** and edit distance
- With both approaches, we obtained results that are competitive with high-quality dictionaries
- Without using explicit cross-lingual signal, we outperformed previous unsupervised work

Conclusion

- We exploited identical pairs that **surprisingly** appear in corpora of **distinct scripts**
- We combined them with a simple method to extract the initial hypothesis set via **romanization** and edit distance
- With both approaches, we obtained results that are competitive with high-quality dictionaries
- Without using explicit cross-lingual signal, we outperformed previous unsupervised work
- We question unsupervised approaches, and show that cheap cross-lingual signals should always be considered for building BWEs, even for distant languages.

References (selected)

https://dumps.wikimedia.org/ (01.04.2020)

https://github.com/isi-nlp/uroman

Artetxe, M., Labaka, G., and Agirre, E. (2017). 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), pages 451–462.

Lample, G., Conneau, A., Ranzato, M., Denoyer, L., and Jégou, H. (2018). Word translation without par- allel data. In International Conference on Learning Representations.

Artetxe, M., Labaka, G., and Agirre, E. (2018). 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), pages 789–798.

Grave, E., Joulin, A., and Berthet, Q. (2019). Unsupervised alignment of embeddings with wasserstein procrustes. In The 22nd International Conference on Artificial Intelligence and Statistics, pages 1880– 1890. PMLR.

Mohiuddin, T. and Joty, S. (2019). Revisiting adversarial autoencoder for unsupervised word translation with cycle consistency and improved training. In Proceedings of NAACL-HLT, pages 3857–3867.

Baldwin, T., Pool, J., and Colowick, S. (2010). Panlex and lextract: Translating all words of all languages of the world. In Coling 2010: Demonstrations, pages 37–40.

Kamholz, D., Pool, J., and Colowick, S. (2014). Panlex: Building a resource for panlingual lexical translation. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pages 3145–3150.

Vulic´, I., Glavaš, G., Reichart, R., and Korhonen, A. (2019). Do we really need fully unsupervised cross- lingual embeddings? 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), pages 4398–4409.

Thank you!

Silvia Severini
Oettingenstraße 67 - 80538 Munich - Germany
silvia@cis.uni-muenchen.de - https://silviaseverini.github.io/

Future work

- Extend this work to LMs:
  - our approach would be applicable to this paper that uses identical words to improve the cross-lingual alignment in multilingual LMs:
    "UNKs Everywhere: Adapting Multilingual Language Models to New Scripts"