Combining Word Embeddings with Bilingual Orthography Embeddings for Bilingual Dictionary Induction

Silvia Severini, Viktor Hangya, Alexander Fraser, Hinrich Schütze

Center for Information and Language Processing, University of Munich

{silvia, hangyav, fraser}@cis.uni-muenchen.de

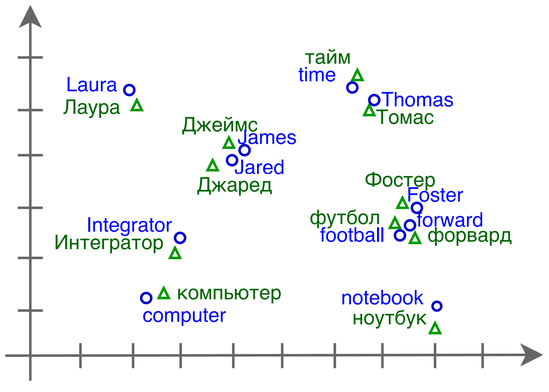
1. Introduction

- Bilingual Dictionary Induction (**BDI**): find target language translations of source language words
- We improve BDI systems in two respects:
  - eliminating the need for language specific orthographic information, such as for Levenshtein distance
  - showing how to decide when to choose transliteration over semantic translation more precisely
- Novel **classification** approach for language pairs with different scripts by combining semantic (**BWEs**) and orthographic information (**BOEs**)
- Novel transliteration system for candidates and BOEs extraction: **seq2seqTr**
- System tested on the English-Russian(En-Ru) data provided in the BUCC 2020 shared task (Rapp et al., 2020).

2. Approach

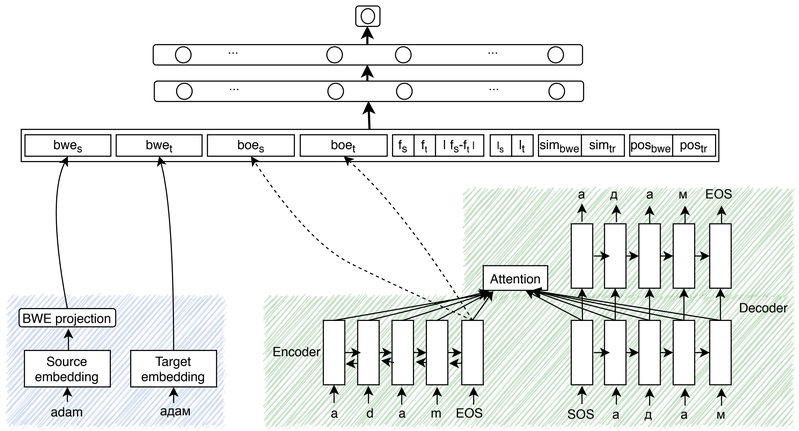
seq2seqTr

- Character-level sequence-to-sequence model with a single-layer encoder and a single-layer decoder
- **Unsupervised**: not relying on transliterations labels
- Trained on the same training dictionary as for building BWEs, but we reduce the number of non-transliteration pairs with our iterative **cleaning process**



BOEs *represent transliteration word pairs with similar vectors. They are the final encoder representations of seq2seqTr (language agnostic encoder with auto-encoder ability).*

Classification model



Green background: seq2seqTr
Blue background: pretrained BWEs
White background: classifier architecture

- BWEs from MWEs, learned with fasttext skipgram (Bojanowski et al., 2017), and aligned with supervised Vecmap on a high-frequency seed dictionary (Artetxe et al. (2018)
- Combines BWEs, BOEs and **additional features**
- Negative sampling: two negative samples for each source word chosen from the candidates lists

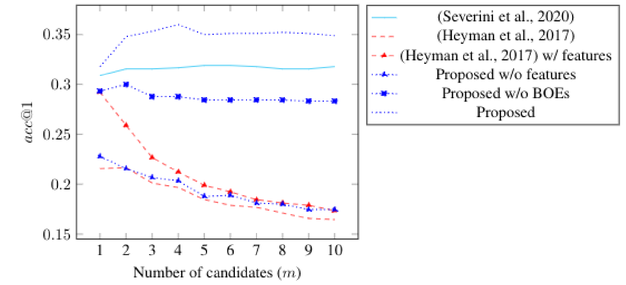
3. Experiments and results

	All	High	Mid	Low
BWEs with CSLS	0.29	0.47	0.29	0.11
Transliteration	0.05	0.05	0.06	0.05
(Severini et al., 2020)	0.33	0.50	**0.33**	0.16
(Heyman et al., 2017)	0.21	0.28	0.22	0.14
(Heyman et al., 2017) w/ features	0.30	0.47	0.29	0.15
Proposed w/o features	0.22	0.33	0.23	0.12
Proposed w/o BOEs	0.31	0.50	0.30	0.12
Proposed	**0.36**	**0.55**	**0.33**	**0.19**

Our approach outperforms all previous approaches both on the joint ("All") and on the separate frequency sets ($acc@1$)

Reranking analysis

- BOEs play a crucial role for the reranking of the candidates
- The added features improve the performance of the models, and help determine the tradeoff between the information from the BWEs and BOEs



$acc@1$ on the development set as a function of the number of candidate words (e.g., 2 means 2 candidates from BWEs and 2 from seq2seqTr.

Transliteration mining with BOEs

- Task where BOEs are the only source of information: transliteration mining on the NEWS 2010 En-Ru test set (Kumaran et al., 2010)
- Score of two words with the cosine similarity of the respective BOEs
- Good performance: BOEs are universal embedding property of representing English and Russian words in a shared space although they use different scripts

	P	R	F
(Jiampojamarn et al., 2010)	88.0	86.9	87.5
(El-Kahky et al., 2011)	92.1	92.5	92.3
(Nabende et al., 2011)	-	-	82.5
(Sajjad et al., 2017)	67.1	97.1	79.4
BOEs	47.0	87.2	61.1
BOEs best	88.8	68.2	77.1

Precision, Recall and F-measure for our BOEs and for state-of-the-art models on transliteration mining. ***sajjad2017statistical*** *and our system are unsupervised while the others are (semi-) supervised.*

4. Conclusion

- We combined semantic (BWEs) and orthographic (BOEs) information for the Bilingual Dictionary Induction task for languages with different scripts
- BOEs are extracted from our seq2seqTr transliteration model
- We improved over the baselines for English-Russian BDI

We gratefully acknowledge funding for this work by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreements № 640550 and № 740516). The work was also supported by the German Research Foundation (DFG; grant FR 2829/4-1).

Presented at The 28th International Conference on Computational Linguistics (COLING)