SIGIR 2019
Paris, France

Evaluating Resource-Lean Cross-Lingual Embedding Models in Unsupervised Retrieval

Robert Litschko*
Goran Glavaš*
Ivan Vulić**
Laura Dietz***

*Data and Web Science Group
University of Mannheim
**Language Technology Lab
University of Cambridge
***University of New Hampshire

Contact: litschko@informatik.uni-mannheim.de

Motivation

Bilingual lexicon induction (BLI) is the standard evaluation task for projection-based CLE-models.

- Does CLIR performance correlate with BLI performance?
- How do CLIR models compare to resource-intensive Machine Translation models?
- How does the CLIR performance of CLE models vary across different language pairs?

CLE Models

Supervised

Canonical Correlation Analysis (CCA) [1]:

- Treat X_S and X_T as different views on the same data points
- Learn the data representations that maximize the correlation between the two views using CCA.

Procrustes Problem (Proc) [3]:

- Treat learning mapping as optimization problem: $W^* = \text{argmin}_{W \in M_d(\mathbb{R})} \|WX_S - X_T\|_F$
- Dictionary of 5k word-pair translations

Procrustes with Bootstrapping (Proc-B):

- Same like Proc, except we start with smaller seed dictionary (1k)
- Dictionary later expanded with bootstrapping

Relaxed Cross-Domain Similarity Local Scaling (RCSLS) [7]:

- Directly optimize for BLI inference metric
- Cross-Domain Similarity Local Scaling (CSLS) between WX_S and X_T
- Cosine similarity normalized with the avg. sim. that each vector has with its cross-lingual nearest neighbors

Unsupervised

Iterative Closest Point Model (ICP) [4]:

- Learn initial projection matrices and word alignments with ICP algorithm
- Each iteration:
  - 1. Fix projections and find the optimal word alignment D
  - 2. Use D to update the projection matrices.
- Expand D with bootstrapping and produce final mapping by solving Procrustes problem

Adversarial Alignment (Muse) [2]:

- Use adversarial learning to learn a projection matrix W, mapping X_S to X_T
- Adversary predicts if embedding comes from WX_S or from X_T.
- Mapper tries to update W to best fool adversary

Heuristic Alignment (VecMap) [5]:

- Assume word translations have similar distributions of similarities with other words from the same language
- Word pairs with closest vectors of monolingual similarity distributions make the initial seed dictionary D

Cross-lingual Embeddings (CLE)

- Cross-lingual Embeddings facilitate cross-lingual NLP and IR
- Start with monolingual word embedding space for source language X_S and target language X_T
- CLE methods map words from source to target language: $X_{CL} = X_S W \cup X_T$, or both languages into a new shared space: $X_{CL} = X_S W_S \cup W_T X_T$
- Resource-Lean CLE → either no, or only weak, bilingual signal (word-translation pairs) used
- **Retrieval models** [6]:
  - Unigram-Language Model + Diritchlet Smoothing (LM-UNI)
  - Google Translate + LM-UNI (MT-IR)
  - (IDF weighted) Bag-of-Word-Embedding-Aggregation (Agg-IDF)
  - Term-by-Term Query Translation + LM-UNI (TbT-QT)
- **CLIR Datasets:**
  - CLEF 2003 [0]: 60 queries, average document collection size: 131K
  - Europarl: Randomly sampled 1K “queries” 100K “documents” for each language
- **BLI dataset:** Most frequent 7K English words automatically translated (2k held out test data)

Document-level CLIR on CLEF (MAP)

Model	CLE Model	DE-FI	DE-IT	DE-RU	EN-DE	EN-FI	EN-IT	EN-RU	FI-IT	FI-RU	AVG
LM-UNI	-	.111	.143	.000	.142	.142	.137	.001	.132	.001	.090
MT-IR	-	.340	.418	.196	.339	.278	.423	.225	.389	.212	.313
Agg-IDF	CCA	.251	.210	.158	.249	.193	.243	.151	.145	.146	.194
	Proc	.255	.212	.152	.261	.200	.240	.152	.149	.146	.196
	Proc-B	.294	.230	.155	.288	.258	.265	.166	.151	.136	.216
	RCSLS	.196	.189	.122	.237	.127	.210	**.133**	.130	.113	.162
	ICP	.252	.170	.167	.230	.230	.231	.119	.117	.124	.182
	Muse	.001	.210	.195	.280	.000	.272	.002	.002	.001	.107
	VecMap	.240	.129	.162	.200	.150	.201	.104	.096	.109	.155
TbT-QT	CCA	.052	.112	.074	.079	.063	.174	.090	.031	.014	.077
	Proc	.061	.098	.058	.081	.048	.181	.069	.044	.021	.073
	Proc-B	.054	.155	.048	.097	.057	.196	.058	.024	.050	.082
	RCSLS	.069	.112	.088	.104	.037	.167	.096	.070	.025	.085
	ICP	.019	.062	.078	.079	.043	.143	.086	.012	.056	.064
	Muse	.000	.131	.111	.102	.001	.196	.001	.004	.001	.061
	VecMap	.204	.166	.080	.205	.087	.237	.117	.140	.115	.150

Bilingual Lexical Induction (MRR)

CLE Model	AVG
CCA	.441
Proc	.447
Proc-B	.422
RCSLS	.481
VecMap	.391
Muse	.211
ICP	.336

Sentence-level CLIR on Europarl (MRR)

Model	CLE Model	DE-FI	DE-IT	EN-DE	EN-FI	EN-IT	FI-IT	AVG
LM-UNI	-	.040	.064	.066	.041	.067	.033	.052
MT-IR	-	.520	.676	.712	.639	.783	.686	.669
Agg-IDF	CCA	.487	.602	.761	.483	.790	.361	.581
	Proc	.497	.614	.766	.481	.791	.371	.587
	Proc-B	.523	.636	.778	.498	.791	.395	.604
	RCSLS	.477	.562	.754	.505	.784	.320	.567
	ICP	.637	.723	.822	.622	.858	.537	.700
	Muse	.020	.630	.764	.009	.774	.010	.368
	VecMap	.590	.599	.741	.551	.789	.442	.619
TbT-QT	CCA	.021	.118	.071	.031	.234	.023	.083
	Proc	.022	.210	.077	.032	.236	.025	.085
	Proc-B	.029	.133	.065	.014	.247	.023	.087
	RCSLS	.025	.140	.140	.044	.282	.048	.113
	ICP	.022	.081	.056	.028	.132	.018	.056
	Muse	.008	.125	.072	.009	.204	.010	.071
	VecMap	.098	.262	.291	.068	.437	.098	.209

Conclusion

- CLIR results **do not follow the trends observed in the BLI** task → Overfitting CLE models to word translation performance may hurt performance in downstream tasks such as CLIR
- MT is a better option for document-level CLIR, Resource-lean CLE models are viable for sentence-level CLIR
- Agg-IDF variants signifcantly outperform corresponding TbT-QT models
- TbT-QT models in many cases perform worse than the LM-UNI baseline

[0] http://catalog.elra.info/product_info.php?products_id=888
[1] Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In Proceedings of EACL. 462–471
[2] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word Translation Without Parallel Data. In ICLR
[3] Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. CoRR, abs/1309.4168 (2013).
[4] Yedid Hoshen and Lior Wolf. 2018. Non-Adversarial Unsupervised Word Translation. In EMNLP. 469–478
[5] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word mbeddings. In ACL. 789–798.
[6] Robert Litschko, Goran Glavaš, Simone Paolo Ponzetto, and Ivan Vulić. 2018. Unsupervised Cross-Lingual Information Retrieval Using Monolingual Data Only.In SIGIR. 1253–1256.
[7] Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in Translation: Learning Bilingual Word Mapping with a Retrieval Criterion. In *EMNLP*. 2979–2984