SIGIR 2018
Ann Arbor, MI, US

Robert Litschko*
Goran Glavaš*
Simone Paolo Ponzetto*
Ivan Vulić**

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

Unsupervised Cross-Lingual Information Retrieval using Monolingual Data Only

Contact: litschko@informatik.uni-mannheim.de

Recepient of Student Travel Grant

MOTIVATION

- Fully unsupervised Cross-lingual Information Retrieval (CLIR) model
- Baseline: Unigram Language Model (LM-UNI)
- Models exploiting induced cross-lingual word embedding spaces:
  - (IDF weighted) Bag-of-Word-Embedding-Aggregation (BWE-Agg-{Add, IDF})
  - Term-by-Term Query Translation + LM-UNI (TbT-QT)
  - Ensemble of BWE-Agg-IDF and TbT-QT ($\lambda = 0.7$ -> emphasize TbT-QT model)

Cross-lingual Embedding Space from ...

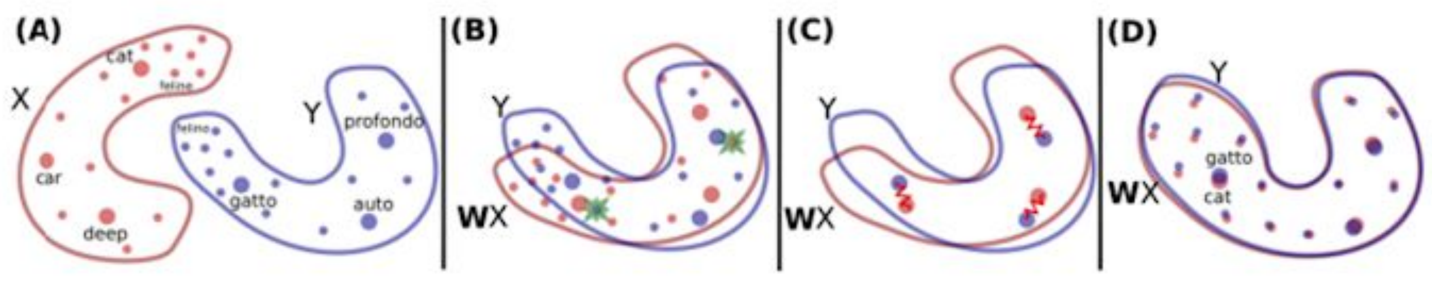
... No Bilingual Supervision

- Uses adversarial learning to learn a projection matrix mapping one embedding space to another.
- Uses projection matrix W to find mutual nearest neighbors between two vocabularies
- The automatically obtained word-translation pairs become synthetic training set for refined projection.

Optimal translation matrix $W^* = \text{argmin}_{W\in M_d(\mathbb{R})}\|WX - Y\|_F$

Discriminator objective $\mathcal{L}_D(\theta_D|W) = -\frac{1}{n}\sum_{i=1}^{n} \log P_{\theta_D}(\text{source} = 1|Wx_i) - \frac{1}{m}\sum_{i=1}^{m} \log P_{\theta_D}(\text{source} = 0|y_i)$

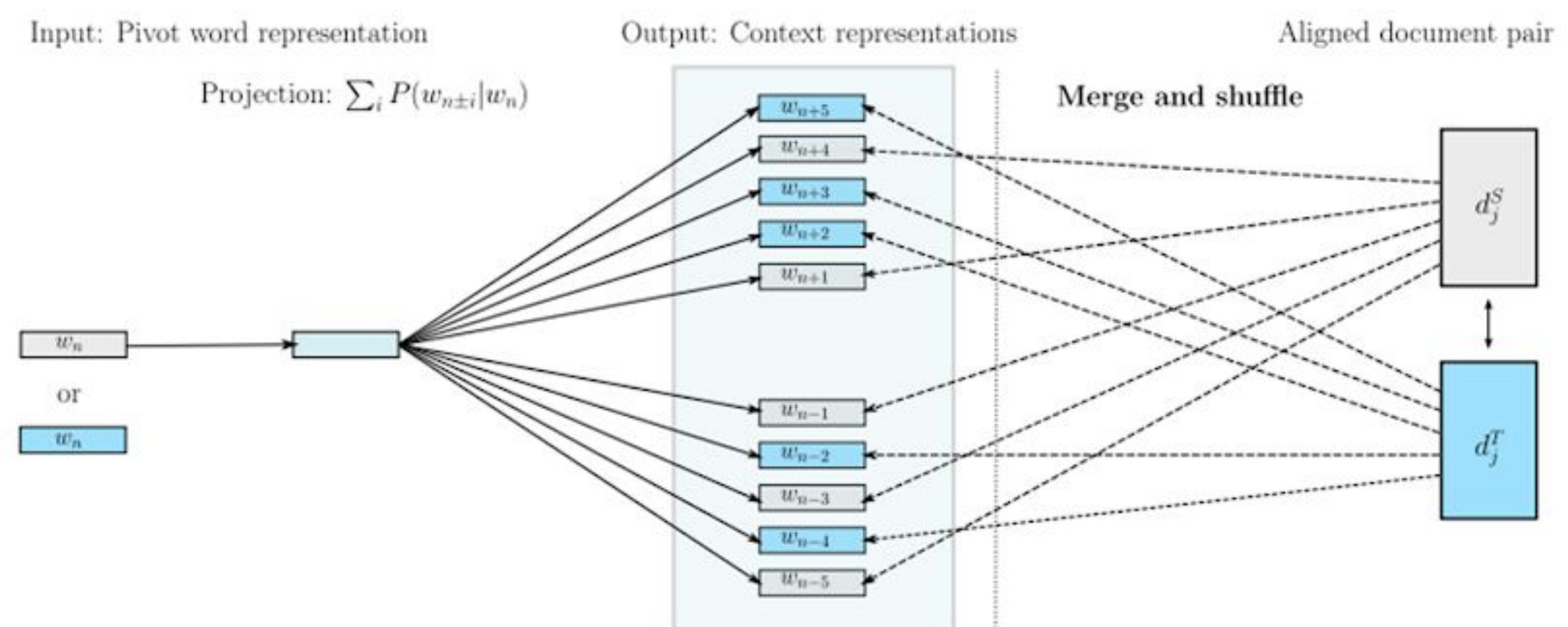
Mapping objective $\mathcal{L}_W(W|\theta_D) = -\frac{1}{n}\sum_{i=1}^{n} \log P_{\theta_D}(\text{source} = 0|Wx_i) - \frac{1}{m}\sum_{i=1}^{m} \log P_{\theta_D}(\text{source} = 1|y_i)$



Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word Translation Without Parallel Data. In ICLR

... Comparable Documents

- Exploits large document-aligned comparable corpora (Wikipedia)
- Creates merged corpus of bilingual pseudo-documents by intertwining pairs of documents.
- Applies standard monolingual Skip-Gram model with negative sampling on merged corpus.



Ivan Vulić and Sien Moens. 2015. Monolingual and Cross-lingual Information Retrieval Models Based on (Bilingual) Word Embeddings. In SIGIR. 363–372

... Word Translation Pairs

- Learns linear projection matrices for two monolingual word embedding spaces into shared embedding space
- Mapping is computed with SVD from similarity matrix (which needs word-alignments)

Samuel L. Smith, David H.P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Ofine Bilingual Word Vectors, Orthogonal Transformations and the Inverted Softmax. In *ICLR*

RESULTS

- Evaluation on CLEF 2000-2003 ad-hoc retrieval Test Suite* (**M**ean **A**verage **P**recision)
- The TbT-QT model based on CL-UPSUP embeddings almost always performs best
- CLIR models based on CL-WT embeddings outperform models based on CL-CD embeddings on average
- Ensembles outperform best-performing individual models by wide margin
- Proximity of language plays a role only to a certain extent

		EN --> NL			EN --> IT			EN --> FI	
CL Embs	**Model**	**2001**	**2002**	**2003**	**2001**	**2002**	**2003**	**2002**	**2003**
-	LM-UNI	.199	.196	.136	.085	.167	.137	.111	.142
CL-CD	BWE-Agg-Add	.111	.138	.137	.087	.114	.147	.026	.084
	BWE-Agg-IDF	.144	.203	.189	.127	.157	.188	.082	.125
	TbT-QT	.125	.196	.120	.106	.148	.143	.176	.140
	Ensemble ($\lambda = 0.5$)	.145	.216	.174	.120	.183	.216	.179	.189
	Ensemble ($\lambda = 0.7$)	.142	.216	.180	.127	.180	.207	.183	.197
CL-WT	BWE-Agg-Add	.149	.168	.203	.138	.155	.236	.078	.217
	BWE-Agg-IDF	.185	.196	.243	.169	.166	.248	.086	.204
	TbT-QT	.159	.164	.176	.129	.150	.218	.095	.095
	Ensemble ($\lambda = 0.5$)	.202	.198	.280	.187	.168	.228	.117	.190
	Ensemble ($\lambda = 0.7$)	.202	.198	.263	.181	.171	.230	.120	.164
CL-UNSUP	BWE-Agg-Add	.125	.153	.198	.119	.126	.213	.078	.239
	BWE-Agg-IDF	.172	.204	.250	.157	.161	.253	.102	.223
	TbT-QT	.229	.257	.299	.232	.257	.345	.145	.243
	Ensemble ($\lambda = 0.5$)	.258	.300	.330	.225	.248	.325	.154	.307
	Ensemble ($\lambda = 0.7$)	.259	.303	.336	.236	.253	.347	.151	.307

*http://catalog.elra.info/product_info.php?products_id=888