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

Robert Litschko
University of Mannheim
litschko@informatik.uni-mannheim.de

Goran Glavaš
University of Mannheim
goran@informatik.uni-mannheim.de

Ivan Vulić
University of Cambridge
iv250@cam.ac.uk

Laura Dietz
University of New Hampshire
dietz@cs.unh.edu

ABSTRACT

Cross-lingual embeddings (CLE) facilitate cross-lingual natural language processing and information retrieval. Recently, a wide variety of resource-lean projection-based models for inducing CLEs has been introduced, requiring limited or no bilingual supervision. Despite potential usefulness in downstream IR and NLP tasks, these CLE models have almost exclusively been evaluated on word translation tasks. In this work, we provide a comprehensive comparative evaluation of projection-based CLE models for both sentence-level and document-level cross-lingual Information Retrieval (CLIR). We show that in some settings resource-lean CLE-based CLIR models may outperform resource-intensive models using full-blown machine translation (MT). We hope our work serves as a guideline for choosing the right model for CLIR practitioners.

CCS CONCEPTS

• **Information systems → Multilingual and cross-lingual retrieval**; *Retrieval models and ranking*.

KEYWORDS

Cross-Lingual IR, Cross-Lingual Embeddings, CLIR Evaluation

ACM Reference Format:

Robert Litschko, Goran Glavaš, Ivan Vulić, and Laura Dietz. 2019. Evaluating Resource-Lean Cross-Lingual Embedding Models in Unsupervised Retrieval. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19), July 21–25, 2019, Paris, France.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3331184.3331324

1 INTRODUCTION

Distributional word vectors, that is, word embeddings have become ubiquitous in natural language processing (NLP) and information retrieval (IR) [2, 12, 17]. Researchers have soon broadened their work towards cross-lingual word embeddings (CLEs). CLE models represent words from two or more languages with vectors lying in the same *shared cross-lingual vector space*, so that words with similar meanings end up with similar vectors, regardless of their actual language. Due to this trait, CLEs offer support to cross-lingual NLP [5–7, 10, 19, *inter alia*] and IR applications [13, 17].

Earlier models induced CLEs by exploiting bilingual supervision in the form of bilingual corpora, aligned either at the level of documents or at the sentence level (see [15] for a comprehensive overview). Recently, the focus has been put on *projection-based* (also known as *mapping-based* or *offline*) CLE models. These models learn a projection (i.e., a mapping) between two (separately) pre-trained monolingual embedding spaces. The projection-based models are particularly suitable for resource-lean settings as they require only limited word-level bilingual supervision (i.e., dictionaries commonly containing only few thousands word translation pairs) [14, 16] or even no bilingual supervision at all [1, 3, 8]. Despite requiring weaker and cheaper supervision (or no supervision at all), projection-based CLE models still deliver the same end product – a shared cross-lingual word vector space. However, evaluations of recent projection-based CLEs have almost exclusively been limited to testing word translation quality, commonly framed as the bilingual lexicon induction (BLI) task, which can be seen as a type of intrinsic evaluation of CLEs. Supported by the wide usage of cross-lingual embeddings in various tasks, we argue that word translation (i.e., BLI) is not the main reason for inducing CLEs and that BLI evaluations of projection-based CLE models should be coupled with downstream (i.e., extrinsic) evaluations.

In this work, we use CLIR tasks as benchmarks for extrinsic evaluation of projection-based CLE models. We perform a systematic evaluation of a range of, both supervised and unsupervised, projection-based CLE models on both document-level and sentence-level CLIR tasks for a variety of different language pairs. Experimental results of our evaluation study, in which we couple different CLE models with two simple semantically-informed ranking functions [13], provide answers to the following questions: **(1)** Does CLIR performance correlate with word translation performance of CLE models (i.e., is the best-performing CLE model according to BLI performance also the best-performing model in CLIR tasks)? **(2)** How do unsupervised CLE models that do not employ any bilingual signal perform in CLIR tasks in comparison to supervised models using (seed) dictionaries with word translation pairs? **(3)** Can CLIR models relying on resource-lean CLE models outperform corresponding CLIR models relying on resource-demanding MT models? **(4)** How does the CLIR performance of CLE models vary across different language pairs (i.e., pairs of close vs. distant languages)?

SIGIR '19, July 21–25, 2019, Paris, France

ACM ISBN 978-1-4503-6172-9/19/07...$15.00
https://doi.org/10.1145/3331184.3331324

2 RESOURCE-LEAN CLE MODELS

Not requiring aligned multilingual data and by not being tied to any specific embedding model, projection-based CLE models are resource-lean and widely applicable. We formalize the projection-based CLE framework and describe the models in evaluation.

2.1 Projection-Based CLE Framework

We start from two independently pre-trained monolingual word embedding spaces ($\mathbf{X}_{L1}$ and $\mathbf{X}_{L2}$) and seek to learn the projection/mapping function(s) that either project vectors from one monolingual space to the other or vectors from both monolingual spaces to the new shared vector space. The projection(s) are learned using the dictionary of word translations pairs $D = \{w^i_{L1}, w^i_{L2}\}^N_{i=1}$. Supervised models (§2.2) use some readily available external *seed* translation dictionary (usually consisting of few thousand word translation pairs), whereas unsupervised models (§2.3) induce D automatically (typically iteratively through self-learning), assuming that approximate isomorphism holds between two monolingual word embedding spaces. Using the seed dictionary, projection-based CLE models create word-aligned matrices – $\mathbf{X}_S = \{\mathbf{x}^i_{L1}\}^N_{i=1}$ and $\mathbf{X}_T = \{\mathbf{x}^i_{L2}\}^N_{i=1}$ – by looking up vectors for aligned words from D in $\mathbf{X}_{L1}$ and $\mathbf{X}_{L2}$, respectively. In the general framework, a CLE model uses $\mathbf{X}_S$ and $\mathbf{X}_T$ to learn two projection matrices $\mathbf{W}_{L1}$ and $\mathbf{W}_{L2}$, projecting respectively $\mathbf{X}_{L1}$ and $\mathbf{X}_{L2}$ to the shared cross-lingual space $\mathbf{X}_{CL} = \mathbf{X}_{L1}\mathbf{W}_{L1} \cup \mathbf{X}_{L2}\mathbf{W}_{L2}$. In practice, however, many of the models we evaluate learn only a single-direction projection matrix $\mathbf{W}_{L1}$ which projects vectors from $\mathbf{X}_{L1}$ to $\mathbf{X}_{L2}$. This can be seen as a special instantiation of the framework in which $\mathbf{W}_{L2} = I$, i.e., $\mathbf{X}_{CL} = \mathbf{X}_{L1}\mathbf{W}_{L1} \cup \mathbf{X}_{L2}$.

2.2 Supervised Models

We first examine supervised CLE models that require an externally created seed translation dictionary D.

Canonical Correlation Analysis (CCA). Faruqui and Dyer [4] treat $\mathbf{X}_S$ and $\mathbf{X}_T$ as different views on the same data points and apply CCA to learn the data representations that maximize the correlation between the two views. CCA learns both projection matrices $\mathbf{W}_{L1}$ and $\mathbf{W}_{L2}$ and projects both monolingual spaces to the new shared space. CCA is a simple and efficient CLE baseline that has mostly been ignored in recent BLI evaluations.

Euclidean Distance and Procrustes Problem. Mikolov et al. [14] cast the CLE induction as a problem of learning the unidirectional projection $\mathbf{W}_{L1}$ that minimizes Euclidean distance between the projected source language vectors $\mathbf{X}_S$ and their corresponding target language vectors $\mathbf{X}_T$: $\mathbf{W}_{L1} = \arg\min_{\mathbf{W}} \|\mathbf{X}_{L1}\mathbf{W} - \mathbf{X}_{L2}\|$. By constraining $\mathbf{W}_{L1}$ to an orthogonal matrix, this minimization becomes a well-known Procrustes problem [16, 18] which has the following closed-form solution:

$$\mathbf{W}_{L1} = \mathbf{U}\mathbf{V}^\top, \text{ with}$$
$$\mathbf{U}\Sigma\mathbf{V}^\top = SVD(\mathbf{X}_T{\mathbf{X}_S}^\top). \quad (1)$$

We evaluate two supervised models based on the solution on the Procrustes problem. First, we evaluate the Proc model that induces $\mathbf{W}_{L1}$ using a larger translation dictionary (5K word translation pairs). The second model, Proc-B, starts from a significantly smaller translation dictionary (1K word pairs): it first learns two single-directional projections – $\mathbf{W}_{L1}$ which induces the cross-lingual space $\mathbf{X}^1_{CL} = \mathbf{X}_{L1}\mathbf{W}_{L1} \cup \mathbf{X}_{L2}$ and $\mathbf{W}_{L2}$ which induces a different cross-lingual space $\mathbf{X}^2_{CL} = \mathbf{X}_{L2}\mathbf{W}_{L2} \cup \mathbf{X}_{L1}$ – and then augments the translation dictionary D with pairs of words that are cross-lingual nearest neighbours according to both projections (i.e., both in $\mathbf{X}^1_{CL}$ and $\mathbf{X}^2_{CL}$). Finally, Proc-B computes the new projection matrix $\mathbf{W}_{L1}$ by solving the Procrustes problem on the augmented dictionary.

Relaxed Cross-Domain Similarity Local Scaling (RCSLS). Instead of minimizing the Euclidean distance, the model of Joulin et al. [9] learns the projection matrix $\mathbf{W}_{L1}$ by maximizing the ranking-based measure called Cross-Domain Similarity Local Scaling (CSLS) [3] between $\mathbf{X}_S\mathbf{W}_{L1}$ and $\mathbf{X}_T$. CSLS, commonly used for inference in word translation (BLI), is the cosine similarity normalized with the average similarity that each of the vectors has with its cross-lingual nearest neighbours. For the maximization of CSLS to be a convex optimization problem, the constraint that $\mathbf{W}_{L1}$ is orthogonal must be relaxed. By using a BLI inference metric as its learning objective RCSLS is particularly tailored for good BLI performance.

2.3 Unsupervised Models

Unsupervised CLE models automatically induce seed translation dictionaries without any bilingual data. In this evaluation we include models that induce seed dictionaries using different strategies: adversarial learning [3], similarity-based heuristics [1], and principal component analysis (PCA) [8]. After obtaining the seed dictionary, a bootstrapping procedure, similar to the one described for Proc-B, is executed. In the final step, the Procrustes problem is again solved, using the dictionary produced through bootstrapping.

Heuristic Alignment (VecMap). Artetxe et al. [1] induce the initial seed lexicon by comparing monolingual distributions of word similarities, assuming that word translations have similar distributions of similarities with other words from the same language. Word pairs having closest vectors of monolingual similarity distributions make the initial seed dictionary, which is then expanded in a self-learning bootstrapping procedure. VecMap's empirical robustness also crucially depends on a multitude of additional steps: unit length normalization, mean centering, ZCA whitening, cross-correlational re-weighting, de-whitening and dimensionality reduction.

Adversarial Alignment (Muse). Conneau et al. [3] use a Generative Adversarial Network (GAN) architecture that learns a projection $\mathbf{W}_{L1}$ (generator) from $\mathbf{X}_{L1}$ to $\mathbf{X}_{L2}$ until a discriminator (a deep feed-forward network) cannot distinguish whether a vector originally comes from the target space $\mathbf{X}_{L2}$ or has been projected from the source space (i.e., comes from $\mathbf{X}_{L1}\mathbf{W}_{L1}$ produced by the generator). The initial projection is then improved in an iterative bootstrapping procedure (similar to Proc-B and VecMap). Muse strongly relies on isomorphism of monolingual spaces, often leading to poor GAN initialization, particularly for distant languages.

Iterative Closest Point Model (ICP). Hoshen and Wolf [8] induce the small seed dictionary by projecting vectors of N most frequent words from both languages to a lower-dimensional space using PCA. They then search for translation matrices $\mathbf{W}_{L1}$ and $\mathbf{W}_{L2}$ that find the optimal alignment (minimal Euclidean distance) between the two sets of N words in this low-dimensional space.

Since the projection matrices and optimal word alignment are both initially unknown, they learn with the Iterative Closest Point algorithm. In each iteration, ICP first fixes the projections and finds the optimal alignment D and then uses D to update the projection matrices. Next, they employ iterative dictionary bootstrapping and produce the final projection by solving the Procrustes problem.

3 EXPERIMENTAL SETUP

CLIR Models and Baselines. For comparing different CLE methods we adopt two simple retrieval methods from Litschko et al. [13]. The first model (AGG-IDF) embeds queries and documents as IDF-weighted sums of corresponding word embeddings from the CLE space and uses cosine similarity as the ranking function. The second model (TbT-QT) employs a cross-lingual embedding space as the translation dictionary, replacing each query term with its cross-lingual nearest neighbour: such term-by-term query translation reduces the task to monolingual retrieval in which the documents are ranked with the unigram language model (LM-UN) with Dirichlet smoothing. We compare the results of CLE-based models to two baselines: (1) a monolingual LM-UN (i.e., without query translation) as a sanity check baseline;[1] (2) a much stronger baseline (MT-IR) translates the query to the collection language using a full-blown MT model and then performs monolingual retrieval using LM-UN. In contrast to CLE-based CLIR, our MT-IR baseline is more resource-demanding as it requires large sentence-aligned corpora.

Languages, Vectors, and Dictionaries. We experiment with five languages – English (EN), German (DE), Italian (IT), Finnish (FI) and Russian (RU) – from which we create nine language pairs of varying language proximity: EN–{DE, FI, IT, RU}, DE–{FI, IT, RU}, and FI–{IT,RU}. For each langage we use pre-trained 300-dimensional fastText embeddings, trained on respective Wikipedias.[2] We obtained dictionaries for supervised CLE models by translating 7K most frequent English words to the other four languages via Google translate. For each language pair, we split the dictionaries into 5K pairs for training[3] and 2K pairs for BLI evaluation.

CLIR Datasets. We evaluate CLE-based models in both sentence-level and document-level CLIR. For document-level retrieval experiments we use the 2003 portion of the CLEF benchmark,[4] which contains test collections for all nine language pairs listed above. All test collections contain 60 queries and the average document collection size per language is 131K (ranging from 17K documents for RU to 295K for DE). For sentence-level CLIR evaluation, we resort to the parallel Europarl corpus [11]. Since Europarl does not contain Russian translations, we evaluate sentence-level CLIR on the remaining six language pairs. For each language pair we randomly sample 1K "queries" (i.e., source language sentences) and 100K "documents" (i.e., target language sentences). Given a sentence in the source language, an ideal CLIR model would rank its mate sentence (i.e., its translation) in the target language on top (i.e., in this setting there is only one relevant "document" per "query").

[1] Relying on lexical overlap between the query and documents, LM-UNI is bound to perform poorly in CLIR where the query language differs from the collection language.
[2] https://fasttext.cc/docs/en/pretrained-vectors.html
[3] We use all 5K pairs to train all supervised models except Proc-B, for which we use training dictionary of only 1K pairs. This is because we want to evaluate whether the bootstrapping procedure can compensate for less bilingual supervision.
[4] http://catalog.elra.info/product_info.php?products_id=888

4 RESULTS AND DISCUSSION

Word Translation Results. We examine how word translation performance of CLE models relates to their CLIR performance in Table 1. We first intrinsically evaluate BLI performance on 2K test dictionaries, in terms of mean reciprocal rank (MRR). Not surprisingly, the RCSLS model with a BLI-tailored objective exhibits the best word translation performance. Simple projection models – CCA and Proc – also exhibit solid performance and the bootstrapping-based model Proc-B, trained using only 1K pairs, does not lag behind by much. Unsupervised CLE models, among which VecMap [1] performs best, despite recent claims [1, 3], do not match the performance of their supervised competitors.

CLIR Results. Table 2 shows CLIR results at the document level (CLEF dataset; MAP), whereas Table 3 summarizes sentence-level CLIR performance (Europarl dataset; MRR) of CLE-based CLIR models. The scores in the upper half of both tables correspond to the embedding aggregation model (Agg-IDF), whereas we obtained the scores in the lower half with the term-by-term CLE-based query translation model (TbT-QT). In both CLIR evaluations, for all CLE models (except for VecMap on CLEF), Agg-IDF variants significantly outperform corresponding TbT-QT models. This is because (1) for most terms there is more than one suitable translation and the translation retrieved by the CLE model often does not match the one used in the document collection and (2) even the best CLE spaces are not perfect word translators. On the other hand, through aggregating semantic CLEs of words, Agg-IDF avoids direct word translation altogether. TbT-QT models in many cases perform even worse than the LM-UNI baseline, since many queries contain named entities, which get replaced with different entities by the CLE model. Compared to the resource-hungry MT-IR baseline, CLE-based models underperform in document retrieval, but Agg-IDF models are competitive in sentence retrieval: the unsupervised ICP model outperforms MT-IR in sentence-retrieval across the board.

Comparing different CLE models, we observe that these CLIR results do not follow the trends observed in the BLI task. For example, the best-performing CLE model on BLI, RCSLS, yields only mediocre CLIR results. This implies that overfitting CLE models to word translation performance may hurt performance in downstream tasks such as CLIR. Furthermore, the Proc-B model, trained using only 1K word pairs, exhibits better CLIR performance than other supervised models (CCA, Proc, and RCSLS), trained on 5K word pairs. Somewhat suprisingly, in sentence-level CLIR evaluation, the unsupervised ICP outperforms all other CLE models, as well as the resource-intensive MT-IR baseline. In combination with ICP's moderate BLI performance, this suggests that ICP induces CLE spaces in which semantic relatedness (albeit not necessarily semantic similarity) is better captured than with other models.

Overall, we conclude that MT is a better option for document-level CLIR, whereas the resource-lean CLE models offer a competitive and viable solution for sentence-level CLIR.

5 CONCLUSION

We have presented a comprehensive evaluation on the usefulness of resource-lean models for inducing cross-lingual embeddings (CLEs) in cross-lingual retrieval. We have shown that word translation performance, the standard evaluation of resource-lean CLE models,

Table 1: BLI performance of different CLE models.

CLE Model	DE-FI	DE-IT	DE-RU	EN-DE	EN-FI	EN-IT	EN-RU	FI-IT	FI-RU	AVG
CCA	0.353	0.506	0.411	0.542	0.383	0.624	0.454	0.353	0.340	0.441
Proc	0.359	0.510	0.425	0.544	0.396	0.625	0.464	0.355	0.342	0.447
Proc-B	0.354	0.507	0.392	0.521	0.360	0.605	0.419	0.328	0.315	0.422
RCSLS	0.395	0.529	0.458	0.580	0.438	0.652	0.510	0.388	0.376	**0.481**
VecMap	0.302	0.493	0.322	0.521	0.292	0.600	0.323	0.355	0.312	**0.391**
Muse	0.000	0.496	0.272	0.520	0.000	0.608	0.000	0.000	0.001	0.211
ICP	0.251	0.447	0.245	0.486	0.262	0.577	0.259	0.263	0.231	0.336

Table 2: Document-level CLIR results (CLEF).

Model	CLE	DE-FI	DE-IT	DE-RU	EN-DE	EN-FI	EN-IT	EN-RU	FI-IT	FI-RU	AVG
LM-UN	–	.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

Table 3: Sentence-level CLIR results (Europarl).

Model	CLE	DE-FI	DE-IT	EN-DE	EN-FI	EN-IT	FI-IT	AVG
LM-UN	-	.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	.120	.077	.032	.236	.025	.085
	Proc-B	.029	.133	.065	.025	.247	.023	.087
	RCSLS	.025	.140	.140	.044	.282	.049	.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

is a poor predictor of downstream CLIR performance. While fully unsupervised CLE models can outperform MT-based CLIR models in sentence retrieval, they lag behind for document-level CLIR. We hope our findings will guide future research on resource-lean CLIR.

REFERENCES

[1] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *ACL*. 789–798.

[2] Stéphane Clinchant and Florent Perronnin. 2013. Aggregating continuous word embeddings for information retrieval. In *CVSC Workshop*. 100–109.

[3] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *ICLR*.

[4] Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of EACL*. 462–471.

[5] Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2015. Cross-lingual Dependency Parsing Based on Distributed Representations. In *Proceedings of ACL*. 1234–1244.

[6] Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual Models for Compositional Distributed Semantics. In *Proceedings of ACL*. 58–68.

[7] Geert Heyman, Ivan Vulić, and Marie-Francine Moens. 2017. Bilingual Lexicon Induction by Learning to Combine Word-Level and Character-Level Representations. In *EACL*. 1085–1095.

[8] Yedid Hoshen and Lior Wolf. 2018. Non-Adversarial Unsupervised Word Translation. In *EMNLP*. 469–478.

[9] 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.

[10] Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *COLING*. 1459–1474.

[11] Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, Vol. 5. 79–86.

[12] Saar Kuzi, Anna Shtok, and Oren Kurland. 2016. Query expansion using word embeddings. In *CIKM*. 1929–1932.

[13] Robert Litschko, Goran Glavaš, Simone Paolo Ponzetto, and Ivan Vulić. 2018. Unsupervised Cross-Lingual Information Retrieval Using Monolingual Data Only. In *SIGIR*. 1253–1256.

[14] Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *CoRR, abs/1309.4168* (2013).

[15] Sebastian Ruder, Anders Søgaard, and Ivan Vulić. 2018. A survey of cross-lingual embedding models. *arXiv preprint arXiv:1706.04902* (2018). arXiv:1706.04902

[16] Samuel L. Smith, David H.P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *ICLR*.

[17] Ivan Vulić and Marie-Francine Moens. 2015. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *SIGIR*. 363–372.

[18] Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *NAACL*. 1006–1011.

[19] Yuan Zhang, David Gaddy, Regina Barzilay, and Tommi Jaakkola. 2016. Ten Pairs to Tag – Multilingual POS Tagging via Coarse Mapping between Embeddings. In *NAACL*. 1307–1317.