Unsupervised Cross-Lingual Information Retrieval using Monolingual Data Only

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

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

Simone Paolo Ponzetto
University of Mannheim
simone@informatik.uni-mannheim.de

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

ABSTRACT

We propose a fully unsupervised framework for ad-hoc cross-lingual information retrieval (CLIR) which requires no bilingual data at all. The framework leverages shared cross-lingual word embedding spaces in which terms, queries, and documents can be represented, irrespective of their actual language. The shared embedding spaces are induced solely on the basis of monolingual corpora in two languages through an iterative process based on adversarial neural networks. Our experiments on the standard CLEF CLIR collections for three language pairs of varying degrees of language similarity (English-Dutch/Italian/Finnish) demonstrate the usefulness of the proposed fully unsupervised approach. Our CLIR models with unsupervised cross-lingual embeddings outperform baselines that utilize cross-lingual embeddings induced relying on word-level and document-level alignments. We then demonstrate that further improvements can be achieved by unsupervised ensemble CLIR models. We believe that the proposed framework is the first step towards development of effective CLIR models for language pairs and domains where parallel data are scarce or non-existent.

CCS CONCEPTS

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

KEYWORDS

Unsupervised cross-lingual IR, cross-lingual vector spaces

1 INTRODUCTION

Retrieving relevant content across languages (i.e., *cross-lingual information retrieval*, termed CLIR henceforth) requires the ability to bridge the lexical gap between languages [8, 16]. Traditional IR methods based on sparse text representations are not suitable for CLIR, since languages, in general, do not share much of the vocabulary. Even in the monolingual IR, they cannot bridge the lexical gap, being incapable of semantic generalization [6]. A solution is to resort to structured real-valued semantic representations, that is, *text embeddings* [2, 6, 10]: these representations allow to generalize over the vocabularies observed in labelled data, and hence offer additional retrieval evidence and mitigate the ubiquitous problem of data sparsity. Their usefulness has been proven for monolingual [15] and cross-lingual ad-hoc IR models [21].

Besides the embedding-based CLIR paradigms, other approaches to bridging the lexical gap for CLIR exist. **1)** Full-blown Machine Translation (MT) systems are employed to translate either queries or documents [8, 9], but these require huge amounts of parallel data, while such resources are still scarce for many language pairs and domains. **2)** The lexical chasm can be crossed by grounding queries and documents in an external multilingual knowledge source (e.g., Wikipedia or BabelNet) [4, 20]. However, the concept coverage is limited for resource-lean languages, and all content not present in a knowledge base is effectively ignored by a CLIR system.

Bilingual text embeddings, while displaying a wider applicability and versatility than the two other paradigms, still suffer from one important limitation: a *bilingual supervision signal* is required to induce *shared cross-lingual semantic spaces*. This supervision takes form of sentence-aligned parallel data [5], pre-built word translation pairs [11, 19] or document-aligned comparable data [21].[1]

Recently, methods for inducing shared cross-lingual embedding spaces without the need for any bilingual signal (not even word translation pairs) have been proposed [1, 3]. These methods exploit inherent structural similarities of induced monolingual embedding spaces to learn vector space transformations that align the source language space to the target language space, with strong results observed for bilingual lexicon extraction. In this work, we show that these unsupervised cross-lingual word embeddings offer strong support to the construction of fully unsupervised ad-hoc CLIR models. We propose two different CLIR models: **1)** term-by-term translation through the shared cross-lingual space, and **2)** query and document representations as IDF-weighted sums of constituent word vectors. To the best of our knowledge, our CLIR

---

[1]For a complete overview we refer the reader to a recent survey [18].

methodology is the first to allow the construction of CLIR models without any bilingual data and supervision at all, relying solely on monolingual corpora. Experimental evaluation on standard CLEF CLIR data for three different language pairs shows that the proposed fully unsupervised CLIR models outperform competitive baselines and models that exploit word translation pairs or comparable corpora. Our CLIR code and multilingual embedding spaces are publicly available at: https://github.com/rlitschk/UnsupCLIR.

2 METHODOLOGY

The proposed unsupervised CLIR models rely on the existence of a shared cross-lingual word embedding space in which all vocabulary terms of both languages are placed. We first outline three methods for the shared space induction, with a focus on the unsupervised method. We then explain in detail the query and document representations as well as the ranking functions of our CLIR models.

2.1 Cross-Lingual Word Vector Spaces

For our proposed CLIR models, we investigate cross-lingual embedding spaces produced with state-of-the-art representative methods requiring different amount and type of bilingual supervision: **1)** document-aligned comparable data [21], **2)** word translation pairs [19]; and **3)** *no bilingual data at all* [3].

Cross-Lingual Embeddings from Comparable Documents (CL-CD). The BWE Skip-Gram (BWESG) model from Vulić and Moens [21] exploits large document-aligned comparable corpora (e.g., Wikipedia). BWESG first creates a merged corpus of bilingual pseudo-documents by intertwining pairs of available comparable documents. Then it applies a standard monolingual log-linear Skip-Gram model with negative sampling (SGNS) [10] on the merged corpus in which words have bilingual contexts instead of monolingual ones.

Cross-Lingual Embeddings from Word Translation Pairs (CL-WT). This class of models [1, 11, 19] focuses on learning the projections (i.e., mappings) between independently trained monolingual embedding spaces. Let $\{v_{w^i}^S\}_{i=1}^{V_S}, v_{w^i}^S \in \mathbb{R}^{ds}$ be the monolingual word embedding space of the source language L_S with V_S vectors, and $\{v_{w^i}^T\}_{i=1}^{V_T}, v_{w^i}^T \in \mathbb{R}^{dt}$ the monolingual space for the target language L_T containing V_T vectors; ds and dt are the respective space dimensionalities. The models learn a parametrized mapping function $f(v|\theta)$ that projects the source language vectors into the target space: $f(v|\theta) : \mathbb{R}^{ds} \rightarrow \mathbb{R}^{dt}$. The projection parameters θ are learned using the training set of K word translation pairs: $\{w_i^S, w_i^T\}_{i=1}^K$, typically via second-order stochastic optimisation techniques.

According to the comparative evaluation from [18], all projection-based methods for inducing cross-lingual embedding spaces perform similarly. We therefore opt for the recent model of Smith et al. [19] to serve as a baseline, due to its competitive performance, large coverage, and readily available implementation.[2] Technically, the method of Smith et al. [19] learns two projection functions $f_S(v_S|\theta_S)$ and $f_S(v_T|\theta_T)$, projecting the source and target monolingual embedding spaces, respectively, to the new shared space.

Cross-Lingual Embeddings without Bilingual Supervision (CL-UNSUP). Most recently, Conneau et al. [3] have proposed an adversarial learning-based model in order to automatically, in a fully unsupervised fashion, create word translation pairs that can then be used to learn the same projection functions f_S and f_T as in the model of Smith et al. [19]. Let X be the set of all monolingual word embeddings from the source language, and Y the set of all target language embeddings. In the first, adversarial learning step, they jointly learn (1) the projection matrix W that maps one embedding space to the other and (2) the parameters of the discriminator model which, given an embedding vector (either Wx where $x \in X$, or $y \in Y$) needs to predict whether it is an original vector from the target embedding space (y),nor a vector from the source embedding space mapped via projection W to the target embedding space (Wx). The discriminator model is a multi-layer perceptron network. In the second step, the projection matrix W trained with adversarial objective is used to find the mutual nearest neighbors between the two vocabularies – this set of automatically obtained word translation pairs becomes a synthetic training set for the refined projection functions f_S and f_T computed via the SVD-based method similar to the previously described model of Smith et al. [19].

2.2 Unsupervised CLIR Models

With the induced cross-lingual spaces we can directly measure semantic similarity of words from the two languages, but we still need to define how to represent queries and documents. To this end, we outline two models that exploit the induced cross-lingual embedding spaces for CLIR tasks.

BWE aggregation model (BWE-AGG). In the first approach, we derive the cross-lingual embeddings of queries and documents by aggregating the cross-lingual embeddings of their constituent terms. Let $\overrightarrow{t}$ be the embedding of the term t, obtained from the cross-lingual embedding space and let d = $\{t_1, t_2, \ldots, t_{N_d}\}$ be a document from the collection consisting of N_d terms. The embedding of the document d in the shared space can then be computed as:

$$\overrightarrow{d} = \overrightarrow{t1} \circ \overrightarrow{t2} \circ \ldots \circ \overrightarrow{t_{N_d}}$$

where $\circ$ is a semantic composition operator: it aggregates constituent term embeddings into a document embedding.[3] We opt for vector addition as composition for two reasons: 1) word embedding spaces exhibit linear linguistic regularities [12], and 2) addition displays robust performance in compositional and IR tasks [14, 21]. A representation of the query vector $\overrightarrow{q}$ is then the sum of embeddings of constituent terms: $\overrightarrow{q} = \sum_{i=1}^{N_q} \overrightarrow{t_i^q}$. To obtain document representations, we compare two aggregation functions. First, we experiment with a simple non-weighted addition (*BWE-Agg-Add*): $\overrightarrow{d} = \sum_{i=1}^{N_d} \overrightarrow{t_i^d}$. Second, we use weighted addition where each term's embedding is weighted with the term's inverse document frequency (IDF) (*BWE-Agg-IDF*): $\overrightarrow{d} = \sum_{i=1}^{N_d} idf(t_i^d) \cdot \overrightarrow{t_i^d}$. BWE-Agg-IDF relies

---

[2]https://github.com/Babylonpartners/fastText_multilingual

[3]There is a large number of options for the composition operator, ranging from unsupervised operations like addition and element-wise multiplication [14] to complex parametrized (e.g., tensor-based) composition functions [13]. We discard the parametrized composition functions because they require parameter optimization through supervision, and we are interested in *fully unsupervised resource-lean CLIR*.

on the common assumption that not all terms equally contribute to the document meaning: it emphasizes vectors of more document-specific terms.[4] Finally, we compute the relevance score simply as the cosine similarity between query and document embeddings in the shared cross-lingual space: $rel_{Agg}(q, d) = \frac{\vec{q} \cdot \vec{d}}{\|\vec{q}\| \cdot \|\vec{d}\|}$.

Term-by-term query translation model (TbT-QT). Our second CLIR model exploits the cross-lingual word embedding space in a different manner: it performs a term-by-term translation of the query into the language of the document collection relying solely on the shared cross-lingual space. Each source language query term t^q is replaced by the target language term $tr(t^q)$, that is, its cross-lingual nearest neighbour in the embedding space. The cosine similarity is used for computing cross-lingual semantic similarities of terms. In other words, the query $q = \{t_1^q, t_2^q, \ldots, t_{N_q}^q\}$ in L_S is substituted by the query $q' = \{tr(t_1^q), tr(t_2^q), \ldots, tr(t_{N_q}^q)\}$ in L_T.[5]

By effectively transforming a CLIR task into a monolingual IR task, we can apply any of the traditional IR ranking functions designed for sparse text representations. We opt for the ubiquitous query likelihood model [17], smoothing the unigram language model of individual documents with the unigram language model of the entire collection, using the Dirichlet smoothing scheme [23]:

$$rel_{TbT}(q', d) = \prod_{i=1}^{N_{q'}} \lambda \cdot P(t_i^{q'}|d) + (1 - \lambda) \cdot P(t_i^{q'}|D).$$

$P(t_i^{q'}|d)$ is the maximum likelihood estimate (MLE) of $t_i^{q'}$ probability based on the document d, $P(t_i^{q'}|D)$ is the MLE of term's probability based on the target collection D, and $\lambda = N_d / (N_d + \mu)$ determines the ratio between the contributions of the local and global language model, with N_d being the document length and μ the parameter of Dirichlet smoothing (= 1000 [23]). Note that the *TbT-QT* model with unsupervised cross-lingual word embeddings is again a fully unsupervised CLIR framework.

3 EXPERIMENTAL SETUP

Language Pairs and Training Data. We experiment with three language pairs of varying degree of similarity: English (EN) – {Dutch (NL), Italian (IT), Finnish (FI)}.[6] We use precomputed monolingual FASTTEXT vectors [2] (available online)[7] as monolingual word embeddings required by CL-WT and CL-UNSUP embedding models. For the CL-CD embeddings, the BWESG model trains on full document-aligned Wikipedias[8] using SGNS with suggested parameters from prior work [22]: 15 negative samples, global decreasing learning rate is .025, subsampling rate is $1e - 4$, window size is 16.

The CL-WT embeddings of Smith et al. [19] use 10K translation pairs obtained from Google Translate to learn the linear mapping functions. The CL-UNSUP training setup closely follows the default setup of Conneau et al. [3]: we refer the reader to the original

Lang.	2001			2002			2003		
	#doc	#tok	#rel	#doc	#tok	#rel	#doc	#tok	#rel
NL	190K	29.6M	24.5	190K	29.6M	37.2	190K	29.6M	28.2
IT	108K	17.1M	26.5	108K	17.1M	21.9	22.3M	157M	15.9
FI	–	–	–	55K	9.3M	16.7	55K	9.25M	10.7

Table 1: Basic statistics of used CLEF test collections: number of documents (#doc), number of tokens (#tok), and average number of relevant documents per query (#rel).

paper and the model implementation accessible online for more information and technical details.[9]

Test Collections and Queries. We evaluate the models on the standard test collections from the CLEF 2000-2003 ad-hoc retrieval Test Suite.[10] We select all NL, IT, and FI document collections from years 2001-2003[11] and paired them with English queries from the respective year. The statistics for test collections are shown in Table 1. Following a standard practice [7, 21], queries were created by concatenating the *title* and the *description* of each CLEF "topic". The test collections for years 2001-2003 respectively contain 50, 50, and 60 EN queries. Queries and documents were lowercased; stop words, punctuations and one-character words were removed.

Models in Comparison. We evaluate six different CLIR models, obtained by combining each of the three models for inducing cross-lingual word vector spaces – *CL-CD*, *CL-WT*, and *CL-UNSUP* – with each of the two ranking models – *BWE-Agg* and *TbT-QT*. For each cross-lingual vector space, we also evaluate an ensemble ranker that combines the two ranking functions: *BWE-Agg-IDF* and *TbT-QT*. If r_1 is the rank of document d for query q according to the *TbT-QT* model and r_2 is the rank produced by *BWE-Agg-IDF*, the ensemble ranker ranks the documents in the increasing order of the scores $\lambda \cdot r_1 + (1 - \lambda) \cdot r_2$. We evaluate ensembles with values $\lambda = 0.5$, i.e., with equal contributions of both models; and $\lambda = 0.7$, i.e., with more weight allocated to the more powerful *TbT-QT* model (cf. Table 2). Additionally, we evaluate the standard query likelihood model (*LM-UNI*) [17] with Dirichlet smoothing [23] as a direct baseline.[12]

4 RESULTS AND DISCUSSION

We show performance of all models in comparison on all test collections, reported in terms of the standard *mean average precision* (MAP) measure in Table 2.

Unsupervised vs. Supervised CLIR. First, CLIR models based on CL-WT embeddings (the bilingual signal are word translation pairs) outperform models based on CL-CD (requiring document-aligned data) on average. This is an encouraging finding, as word translations pairs are easier to obtain than document-aligned comparable corpora. Most importantly, the unsupervised CL-UNSUP+TbT-QT CLIR model displays peak performance on all but one test collection (EN-FI, 2002). We find this to be a very important result: it

---

[4]Note that with both variants of BWE-Agg, we effectively ignore both query and document terms that are not represented in the cross-lingual embedding space.

[5]If the representation of a query term t_i^q is not present in the cross-lingual embedding space, we retain the query term t_i^q itself. We have also attempted eliminating out-of-vocabulary query terms, but the former consistently leads to better performance.

[6]English and Dutch are Germanic languages, Italian is a Romance language, whereas Finnish is an Uralic language (i.e., not Indo-European)

[7]https://github.com/facebookresearch/fastText

[8]http://linguatools.org/tools/corpora/wikipedia-comparable-corpora/

[9]https://github.com/facebookresearch/MUSE

[10]http://catalog.elra.info/product_info.php?products_id=888

[11]Finnish was included to CLEF evaluation only in 2002 and 2003.

[12]LM-UNI uses the same ranking function as TbT-QT, but without the prior term-by-term query translation via the cross-lingual embedding space. LM-UNI is more suitable for monolingual IR than for CLIR due to limited lexical overlap between languages.

CL Embs	Model	EN→NL			EN→IT			EN→FI	
		2001	2002	2003	2001	2002	2003	2002	2003
–	LM-UNI	.119	.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**

Table 2: CLIR performance on all three test language pairs for all models in comparison (MAP scores reported).

shows that we can perform robust CLIR without any cross-lingual information, that is, by relying purely on monolingual data.

Ensemble CLIR Models. Ensembles generally outperform the best-performing individual CLIR models, and for some test collections (e.g., EN→NL 2002, EN→FI 2003) by a wide margin. For the *CL-CD* and *CL-WT* spaces, we observe similar results for both values of the interpolation factor ($\lambda = 0.5$ and $\lambda = 0.7$). This is not surprising, since the single models *BWE-Agg-IDF* and *TbT-QT* exhibit similar performance for *CL-CD* and *CL-WT*. In contrast, the combined model with $\lambda = 0.7$ (i.e., more weight for the *TbT-QT* ranking) yields larger performance gains for CL-UNSUP spaces, for which the *TbT-QT* model consistently outperforms *BWE-Agg-IDF*.

Language Similarity and Aggregation. The results in Table 2 imply that the proximity of CLIR languages plays a role only to a certain extent. Most models do exhibit lower performance for EN→FI than for the other two language pairs: this is expected since Finnish is lexically and typologically more distant from English than Italian and Dutch. However, even though NL is linguistically closer to EN than IT, for the unsupervised CLIR models we generally observe slightly better performance for EN→IT than for EN→NL. We speculate that this is due to the compounding phenomenon in word formation, which is present in NL, but is not a property of EN and IT. The reported performance on bilingual lexicon extraction (BLE) using cross-lingual embedding spaces is also lower for EN-NL compared to EN-IT (see, e.g., [19]). We observe the same pattern (4-5% lower BLE performance for EN-NL than for EN-IT) with the CL-UNSUP embedding spaces.

The weighted variant of BWE-Agg (BWE-Agg-IDF) outperforms the simpler non-weighted summation model (BWE-Agg-Add) across the board. These results suggest that the common IR assumption about document-specific terms being more important than the terms occurring collection-wide is also valid for constructing dense document representations by summing word embeddings.

5 CONCLUSION

We have presented a fully unsupervised CLIR framework that leverages unsupervised cross-lingual word embeddings induced solely on the basis of monolingual corpora. We have shown the ability of our models to retrieve relevant content cross-lingually without any bilingual data at all, by reporting competitive performance on standard CLEF CLIR evaluation data for three test language pairs. This unsupervised framework holds promise to support and guide the development of effective CLIR models for language pairs and domains where parallel data are scarce or unavailable.

REFERENCES

[1] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *ACL*. 451–462.
[2] Piotr Bojanowski and Edouard Grave et al. 2017. Enriching Word Vectors with Subword Information. *Transactions of the ACL* 5 (2017), 135–146.
[3] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word Translation Without Parallel Data. In *ICLR*.
[4] Marc Franco-Salvador, Paolo Rosso, and Roberto Navigli. 2014. A knowledge-based representation for cross-language document retrieval and categorization. In *EACL*. 414–423.
[5] Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual Models for Compositional Distributed Semantics. In *ACL*. 58–68.
[6] Thomas K. Landauer and Susan T. Dumais. 1997. Solutions to Plato's problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review* 104, 2 (1997), 211–240.
[7] Victor Lavrenko, Martin Choquette, and W. Bruce Croft. 2002. Cross-lingual relevance models. In *SIGIR*. 175–182.
[8] Gina-Anne Levow, Douglas W. Oard, and Philip Resnik. 2005. Dictionary-Based Techniques for Cross-Lingual IR. *IP & M* 41, 3 (2005), 523–547.
[9] Giovanni Da San Martino and Salvatore Romeo et al. 2017. Cross-language question re-ranking. In *SIGIR*. 1145–1148.
[10] Tomas Mikolov and Ilya Sutskever et al. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *NIPS*. 3111–3119.
[11] Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting Similarities among Languages for Machine Translation. *CoRR* abs/1309.4168 (2013).
[12] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic Regularities in Continuous Space Word Representations. In *NAACL-HLT*. 746–751.
[13] Dmitrijs Milajevs and Dimitri Kartsaklis et al. 2014. Evaluating Neural Word Representations in Tensor-Based Compositional Settings. In *EMNLP*. 708–719.
[14] Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *ACL-HLT*. 236–244.
[15] Bhaskar Mitra and Nick Craswell. 2017. Neural Models for Information Retrieval. *CoRR* abs/1705.01509 (2017).
[16] Jian-Yun Nie. 2010. *Cross-Language Information Retrieval*.
[17] Jay M. Ponte and W. Bruce Croft. 1998. A language modeling approach to information retrieval. In *SIGIR*. ACM, 275–281.

[18] Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2017. A Survey of Cross-Lingual Embedding Models. *CoRR* abs/1706.04902 (2017).

[19] 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*.

[20] Philipp Sorg and Philipp Cimiano. 2012. Exploiting Wikipedia for cross-lingual and multilingual information retrieval. *DKE* 74 (2012), 26–45.

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

[22] Ivan Vulić and Sien Moens. 2016. Bilingual Distributed Word Representations from Document-Aligned Comparable Data. *JAIR* 55 (2016), 953–994.

[23] Chengxiang Zhai and John Lafferty. 2004. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems* 22, 2 (2004), 179–214.