Zero-Shot Language Transfer for Cross-Lingual Sentence Retrieval Using Bidirectional Attention Model

Goran Glavaš[1] and Ivan Vulić[2]

[1] Data and Web Science Group, University of Mannheim, Germany
goran@informatik.uni-mannheim.de
[2] Language Technology Lab, University of Cambridge, UK
iv250@cam.ac.uk

Abstract. We present a neural architecture for cross-lingual mate sentence retrieval which encodes sentences in a joint multilingual space and learns to distinguish true translation pairs from semantically related sentences across languages. The proposed model combines a recurrent sequence encoder with a bidirectional attention layer and an intra-sentence attention mechanism. This way the final fixed-size sentence representations in each training sentence pair depend on the selection of contextualized token representations from the other sentence. The representations of both sentences are then combined using the bilinear product function to predict the relevance score. We show that, coupled with a shared multilingual word embedding space, the proposed model strongly outperforms unsupervised cross-lingual ranking functions, and that further boosts can be achieved by combining the two approaches. Most importantly, we demonstrate the model's effectiveness in zero-shot language transfer settings: our multilingual framework boosts cross-lingual sentence retrieval performance for unseen language pairs without any training examples. This enables robust cross-lingual sentence retrieval also for pairs of resource-lean languages, without any parallel data.

Keywords: Cross-lingual retrieval · Language transfer · Bidirectional attention model · Sentence retrieval.

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. In general, there are three distinct approaches to CLIR. First, translating queries and/or documents using dictionaries or full-blown machine translation (MT) to the same language enables the use of monolingual retrieval models [31, 17, 23]. Second, the lexical chasm can be crossed by grounding queries and documents in an external multilingual knowledge source (e.g., Wikipedia or BabelNet) [46, 10]. Finally, other systems induce shared cross-lingual semantic spaces (e.g., based on bilingual word embeddings) and represent queries and documents as vectors in the shared space [9, 48, 49].

Each line of work comes with certain drawbacks: 1) robust MT systems require huge amounts of parallel data, while these resources are still scarce for many language pairs and domains; 2) concept coverage in multilingual knowledge bases like BabelNet [29] is still limited for resource-lean languages, and all content not present in a knowledge base is effectively ignored by a CLIR system; 3) CLIR models based on bilingual semantic spaces require parallel or comparable texts to induce such spaces [49, 40].

Due to smaller quantities of text which the ranking functions can exploit, sentence retrieval is traditionally considered more challenging than standard document-level retrieval [28, 32, 21, 23]. Cross-lingual sentence retrieval typically equals to identifying parallel sentences in large text collections: the so-called *mate retrieval* task [30, 37, 42], which benefits the construction of high-quality sentence-aligned data for MT model training [41]. To this end, it is crucial to distinguish exact translation pairs from sentence pairs that are only semantically related. This is why CLIR models that exploit coarse-grained representations, e.g., by inducing latent topics [9, 48] or by aggregating word embeddings [49], are not suitable for modeling such subtle differences in meaning.

In this work, we propose a neural architecture for cross-lingual sentence retrieval which captures fine-grained semantic dependencies between sentences in different languages and distinguishes true sentence translations pairs from related sentences. Its high-level flow is illustrated in Figure 1.

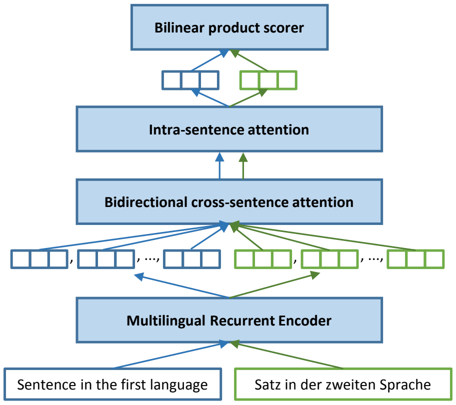
Bilinear product scorer
Intra-sentence attention
Bidirectional cross-sentence attention
Multilingual Recurrent Encoder
Sentence in the first language
Satz in der zweiten Sprache

Fig. 1. High-level overview of the bidirectional cross-lingual attention (BiCLA) model.

First, a joint multilingual word embedding space is coupled with a recurrent encoder that is *shared* between the two languages: this enables contextualization of word representations for word sequences in both languages. Further, selective cross-sentence contextualization is achieved by means of a bidirectional attention mechanism stacked on top of the shared encoder. The attention layer enables the model to assign more weight to relevant information segments from the other sentence. Cross-sentence informed representations are then aggregated into a fixed-size sentence vectors via an intra-sentence attention mechanism. Finally, the model predicts the ranking score for a sentence pair by computing a bilinear product between these fixed-size sentence representations.

We evaluate the model in a mate sentence retrieval task on the Europarl data. We experiment with four languages of varying degrees of similarity, showing that our bidirectional attention model significantly outperforms state-of-the-art unsupervised CLIR models, recently proposed in [19]. Most importantly, we demonstrate the effectiveness of the model in zero-shot language transfer: a model trained on parallel data for one language pair (e.g., German and English) successfully performs CLIR for another language pair (e.g., Czech and Hungarian). Finally, we observe that the proposed bidirectional attention model complements state-of-the-art unsupervised CLIR baselines: we obtain further significant performance gains by ensembling the models.

2 Related Work

Cross-Lingual Information Retrieval. Early CLIR methods combined dictionary-based term-by-term translations of queries to the collection language [4, 34, 17] with monolingual term-based retrieval models [45, 39]. Such models suffer from two main drawbacks: 1) inability to account for in-context meaning of query words and multi-word expressions and 2) inability to capture semantic similarity between queries and documents.[3]

One way to mitigate these issues is to represent documents and queries using concepts from external (multilingual) knowledge resources. Sorg and Cimiano [46] exploit Wikipedia as a multilingual knowledge base and represent documents as vectors where dimensions denote Wikipedia concepts. Franco-Salvador et al. [10] link the document text to concepts in BabelNet [29] and then measure document similarity by comparing BabelNet subgraphs spanned by the linked concepts. These methods, relying on external structured knowledge, are limited by the coverage of the exploited knowledge bases. Another limiting factor is their core dependence on the quality of concept linkers [26, 33], required to associate the concepts from text with knowledge base entries: any piece of text that is not linked to a knowledge base concept is effectively ignored by the model.

Another class of models for cross-lingual text comparison is based on the induction of shared multilingual semantic spaces in which queries and documents in both languages are represented as vectors. These are induced using (Probabilistic) Latent Semantic Analysis (LSA) [9, 35], Latent Dirichlet Allocation [25, 48], or Siamese Neural Networks [51]. In contrast to directly learning bilingual document representations, Vulić and Moens [49] obtain bilingual word embeddings and then compose cross-lingual document and query representations by simply summing the embeddings of their constituent words.

Cross-Lingual Sentence Matching. Approaches to extracting parallel sentences have ranged from rule-based extraction from comparable documents [30, 38, 42], over classifiers trained on sentence-aligned parallel data [27, 41] to cross-lingual sentence retrieval [37]. Supervised approaches typically exploit pretrained SMT models or their components (e.g., word alignment models) to produce features for classification [27, 41]. These are often coupled with a rich set of domain-specific features computed from metadata of the bilingual data at hand [41]: all this limits the portability of such models.

Another related task is cross-lingual semantic similarity of short texts (STS) [1, 7]. The best-performing cross-lingual STS models [6, 14, 47] all employ a similar strategy: they first translate the sentences from one language to the other (resource-rich) language

[3] E.g., a German term *"Hund"* translated as *"dog"* still does not match a term *"canine"* from a relevant document.

(i.e., English) and then apply supervised, feature-rich and language-specific (e.g., they rely on syntactic dependencies and named entity recognizers) regression models. Their dependence on full-blown MT systems and resource-intensive and language-specific features limits their portability to arbitrary (resource-lean) language pairs.

3 Bidirectional Attention CLIR Model

First we describe the induction of a multilingual word vector space and then the components of our bidirectional cross-lingual attention model (BiCLA).

Multilingual Word Vector Space. Multiple methods have recently been proposed for inducing bilingual word vector spaces by learning linear projections from one monolingual space to another [24, 2, 43, 8]. A multilingual vector space for N languages is then induced by simply learning $N-1$ bilingual projections with the same target space (e.g., English). A comparative evaluation by Ruder et al. [40] indicates that all of the above models produce multilingual spaces of similar quality. Due to its large language coverage and accessible implementation, we opt for the model of Smith et al. [43]. They learn the projection by exploiting a set of (10K or less) word translation pairs.

3.1 BiCLA: Cross-Lingual Sentence Retrieval Model

The architecture of the BiCLA model is detailed in Figure 2. We encode the sentences from both languages with the same bidirectional long short-term memory network (Bi-LSTM): word vectors from a pre-trained bilingual space are input to the network. The sentences are then made “aware of each other”: we compute the vector representations of each sentence’s tokens by attending over Bi-LSTM encodings of other sentence’s tokens. Next, we use an intra-sentence attention mechanism to aggregate a fixed-size encoding of each sentence from such cross-sentence contextualized vectors of its tokens. Finally, we compute the relevance score for the sentence pair as the bilinear product between fixed-size representations of the two sentences obtained through intra-sentence attention. In what follows, we describe all components of the BiCLA model.

1. Recurrent Encoder. We encode both sentences using the same shared Bi-LSTM encoder [12].[4] Given an input sequence of T tokens $\{t_i\}_{i=1}^{T}$, the Bi-LSTM layer produces a sequence of T within-sentence *contextualized* token representations $\{h_i = [h_i^f, h_i^b]\}_{i=1}^{T}$, where $h_i^f \in \mathbb{R}^H$ is the i-th token vector produced by the forward-pass LSTM and $h_i^b \in \mathbb{R}^H$ is the i-th token vector produced by the backward-pass LSTM, with H as LSTMs hidden state size. Vector h_i^f contextualizes the i-th token with the meaning of its left context (i.e., preceding tokens), whereas h_i^b makes the representation of the i-th token aware of its right context (following tokens).

2. Bidirectional Cross-Sentence Attention. In neural machine translation [3, 22], the attention mechanism allows to focus more on parts of the source sentence that are most

[4] We also experimented with two different Bi-LSTMs for encoding sentences in two languages, but this exhibited poorer performance.

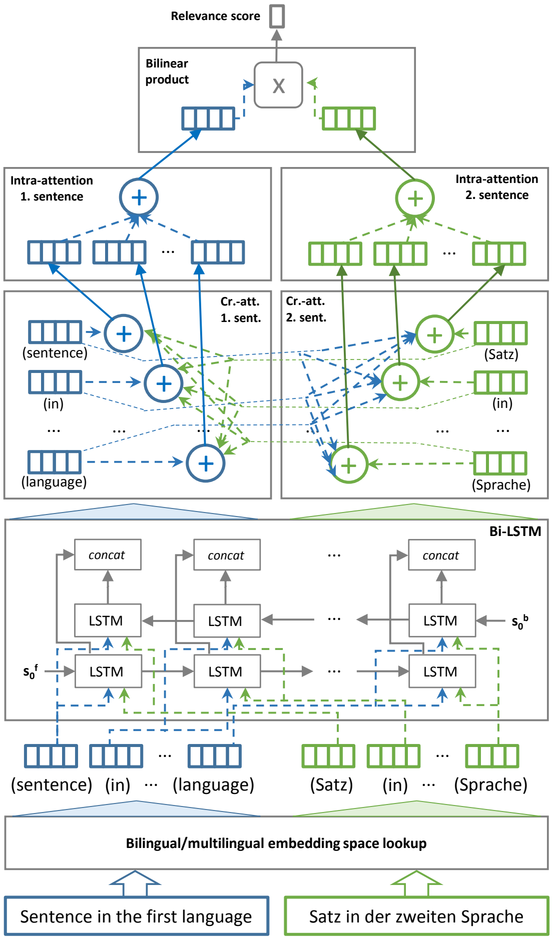
Relevance score
Bilinear product
X
Intra-attention 1. sentence
Intra-attention 2. sentence
Cr.-att. 1. sent.
Cr.-att. 2. sent.
(sentence)
(in)
(language)
(Satz)
(in)
(Sprache)
Bi-LSTM
concat
LSTM
s_0^f
s_0^b
(sentence) (in) ··· (language)
(Satz) (in) ··· (Sprache)
Bilingual/multilingual embedding space lookup
Sentence in the first language
Satz in der zweiten Sprache

Fig. 2. Schema of the BiCLA model.

relevant for translation generation at a concrete position. In sentence retrieval, the goal is to semantically *align* two sentences and determine their semantic compatibility. To this end, we define a bidirectional attention layer that allows to represent tokens of the first sentence by focusing on representations of only relevant tokens from the second sentence, and vice versa. Let $\{h_i^{S_1}\}_{i=1}^{T_1}$ and $\{h_j^{S_2}\}_{j=1}^{T_2}$ be the sequences of token representations of the input sentences produced by Bi-LSTM. The cross-sentence representation $\overline{h}_i^{S_1}$ of the i-th token of the first sequence is then computed as the weighted average of token vectors $\{h_j^{S_2}\}_{j=1}^{T_2}$ of the second sentence, and vice versa:

$$\overline{h}_i^{S_1} = \sum_{j=1}^{T_2} \alpha_{i,j} \cdot h_j^{S_2}; \qquad \overline{h}_j^{S_2} = \sum_{i=1}^{T_1} \beta_{j,i} \cdot h_i^{S_1}. \tag{1}$$

Attention weights $\alpha_{i,j}$ and $\beta_{j,i}$ are obtained by computing the softmax functions over respective raw matching scores $m_{i,j}$ and $n_{j,i}$ which are, in turn, computed on the basis of a bilinear product of token vectors $h_i^{S_1}$ and $h_j^{S_2}$:

$$\alpha_{i,j} = \frac{e^{m_{i,j}}}{\sum_{k=1}^{T_2} e^{m_{i,k}}};\;\; m_{i,j} = \tanh\left(h_i^{S_1} W_{ca}^1 h_j^{S_2} + b_{ca}^1\right); \tag{2}$$

$$\beta_{j,i} = \frac{e^{n_{j,i}}}{\sum_{k=1}^{T_1} e^{n_{j,k}}};\;\; n_{j,i} = \tanh\left(h_j^{S_2} W_{ca}^2 h_i^{S_1} + b_{ca}^2\right). \tag{3}$$

$W_{ca}^1, W_{ca}^2 \in \mathbb{R}^{2H \times 2H}$ and $b_{ca}^1, b_{ca}^2 \in \mathbb{R}$ are attention parameters.[5] Using the cross-attention mechanism we *contextualize* one sentence in terms of the other in a *localized* manner: the vector $\overline{h}_i^{S_1}$ of a first sentence token aggregates information from the semantically most relevant parts of the other sentence, and vice-versa for $\overline{h}_j^{S_2}$.

3. Intra-Sentence Attention. Bi-LSTM contextualizes token representations within the sentence, whereas the cross-attention contextualizes them with respect to the other sentence. We finally produce the task-specific fixed-size sentence representations by aggregating their respective contextualized token vectors. Because not all parts of a sentence are equally contributing to its meaning, we learn how to aggregate the fixed-size sentence representation by means of an intra-sentence attention mechanism. Our intra-sentence attention is a simplified version of the recently introduced self-attention networks [20, 18]. The sentence embeddings e_1 and e_2 are computed as weighted sums of their cross-sentence contextualized token vectors:

$$e_1 = \sum_{i=1}^{T_1} \gamma_i \overline{h}_i^{S_1}; \qquad e_2 = \sum_{j=1}^{T_2} \delta_j \overline{h}_j^{S_2}. \tag{4}$$

The weights γ_i and δ_j are computed as non-linear transformations of dot products between token vectors and intra-sentence attention parameter vectors:

$$\gamma_i = \tanh\left(\overline{h}_i^{S_1} \cdot v_{ia}^1 + b_{ia}^1\right); \delta_j = \tanh\left(\overline{h}_j^{S_2} \cdot v_{ia}^2 + b_{ia}^2\right).$$

We learn the parameters $v_{ia}^1, v_{ia}^2 \in \mathbb{R}^{2H}$ and $b_{ia}^1, b_{ia}^2 \in \mathbb{R}$ during training.[6]

4. Bilinear Scoring. Finally, we can quantify a similarity (i.e., relevance) score for the cross-lingual pair of sentences from the obtained fixed-size representations e_1 and e_2. We combine the vectors e_1 and e_2 into a relevance score r with a bilinear product function, which was previously successfully applied to, e.g., relation prediction for knowledge base completion [44, 50] and predicting semantic matches at a word level [11]:

$$r(S_1, S_2) = \tanh\left(e_1 W_B e_2 + b_B\right), \tag{5}$$

where $W_B \in \mathbb{R}^{2H \times 2H}$ and $b_B \in \mathbb{R}$ are the bilinear product parameters.

[5] Note that by constraining $W_{ca}^1 = W_{ca}^2$ and $b_{ca}^1 = b_{ca}^2$ we enforce a symmetric bidirectional cross-attention mechanism. However, the above asymmetric attention gave better performance.

[6] Again, we could enforce the symmetric intra-sentence attention for both sentences by setting $v_{ia}^1 = v_{ia}^2$ and $b_{ia}^1 = b_{ia}^2$, but doing so resulted in lower performance in our experiments.

5. Objective and Optimization. The model has to assign higher scores $r(S_1, S_2)$ to sentence pairs where S_2 is a complete semantic match (i.e., a translation) of S_1 than to semantically related sentence pairs with only a partial semantic overlap. Therefore, BiCLA relies on a contrastive loss function that maximizes the difference in scores between positive sentence pairs and corresponding negative pairs. Let $\{(S_1^{(i)}, S_2^{(i)})\}_{i=1}^{N}$ be the collection of positive pairs in our training set: these are exact translations. For each source language sentence $S_1^{(i)}$ we create K negative training pairs $\{(S_1^{(i)}, S_2^{(k_j)})\}_{j=1}^{K}$. Half of the these K pairs are created by pairing $S_1^{(i)}$ with $K/2$ randomly selected target language sentences. The remaining pairs are created by coupling $S_1^{(i)}$ with $K/2$ semantically most similar sentences in the other language (excluding the target sentence from the positive example), according to a baseline heuristic function as follows. Let $e(t)$ retrieve the embedding of the term t from the shared bilingual embedding space. We then compute the heuristic similarity between sentences S_1 and S_2 as:

$$sim(S_1, S_2) = \cos\left(\sum_{t_1 \in S_1} e(t_1), \sum_{t_2 \in S_2} e(t_2)\right) \tag{6}$$

By taking the most similar sentences according to the above heuristic, we create – for each positive sentence pair – $K/2$ corresponding negative pairs in which there is at least some semantic overlap between the sentences. The contrastive loss objective for the given training set of translation pairs $\{(S_1^{(i)}, S_2^{(i)})\}_{i=1}^{N}$ is finally defined as follows:

$$J = \sum_{i=1}^{N}\sum_{j=1}^{K} \epsilon - \Big(r\big(S_1^{(i)}, S_2^{(i)}\big) - r\big(S_1^{(i)}, S_2^{(k_j)}\big)\Big). \tag{7}$$

The hyper-parameter ϵ defines the margin between scores of positive and negative pairs. The final objective function J_{MIN} augments the contrastive loss function J with the L_2 regularization of parameters $\mathbf{\Omega}$: $J_{MIN} = J + \lambda\|\mathbf{\Omega}\|_2$, with λ as regularization factor.

4 Evaluation

We first describe the important aspects of the experimental setup – datasets, baselines, and details on model training. We then report and discuss BiCLA performance in 1) standard mate retrieval and 2) zero-shot language-transfer experiments.

4.1 Experimental Setup

Data. We use the parallel Europarl corpus [16][7] in all experiments. Since one of our goals is to examine retrieval performance for languages of varying degree of similarity, we experiment with the Europarl data in English (EN), German (DE), Czech (CS), and Hungarian (HU).[8] The Europarl datasets for all six language-pair combinations (see

[7] http://opus.nlpl.eu/Europarl.php

[8] English, German, and Czech belong to the family of Indo-European languages (EN and DE are representatives of the Germanic branch, and CS is in the Slavic branch), whereas Hungarian belongs to the Uralic language family.

Table 1. Sizes of all datasets used in experiments.

Language pair	Train size (# pairs)	Test size (# pairs)
CS-DE	506,495	1,000
CS-EN	572,889	1,000
CS-HU	543,959	1,000
DE-EN	1,584,202	1,000
HU-DE	501,128	1,000
HU-EN	556,774	1,000

Table 1) were preprocessed by 1) removing stopwords and 2) retaining only sentence pairs in which each sentence has at least three tokens represented in the bilingual embedding space. For each language pair, we use a set of 1000 randomly selected sentence pairs as test data. All the remaining pairs are used for the BiCLA training. The datasets' sizes, in terms of number of sentence pairs, are shown in Table 1.

Multilingual Embedding Space. We use precomputed 300-dimensional monolingual FASTTEXT word embeddings [5][9] for all four languages. We then induce a shared four-lingual embedding space using the lexicon-based projection method with pivoting from [43], outlined in §3.[10]

Baselines. We compare BiCLA with the standard query likelihood model [36], two state-of-the-art unsupervised CLIR models [19], and the reduced architecture without the bidirectional attention:

1) Standard query likelihood retrieval model [36] with Jelinek-Mercer smoothing [13] (**QLM**). The model computes the relevance by multiplying source sentence terms' probabilities under the unigram language model of a target language sentence, smoothed with the their probabilities under the language model of the whole target collection:

$$rel(q,d) = \prod_{t \in q} \mu P(t|M_S) + (1-\mu) P(t|M_{SC}). \tag{8}$$

$P(t|M_S)$ is probability of term t under the local language model of the sentence S, $P(t|M_{SC})$ is the probability of t under the global language model of the target collection SC, and $\mu = 0.95$ is the interpolation coefficient. Designed for monolingual retrieval, QLM's CLIR performance crucially depends on the amount of lexical overlap between languages. We thus employ QLM merely as a "sanity check" baseline.

2) Aggregating word embeddings from the shared embedding space (**AGG**) [49, 19]. AGG computes sentence embeddings by averaging the embeddings of their tokens, obtained from the shared multilingual embedding space. The relevance score is the cosine similarity between aggregated sentence embeddings. We used AGG also as a heuristic for creating negative instances for the contrastive loss (cf. Eq. (6) in §3.1).

3) Term-by-term translation using shared embedding space (**TbT**). Each query token is replaced by the most similar target language token, according to the cosine similarity in

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

[10] `https://tinyurl.com/msrmwee`

the shared space. After the term-by-term translation of the query, we use the monolingual QLM to rank the target sentences. TbT and AGG have recently exhibited state-of-the-art performance on several benchmarks for document-level CLIR [19].

4) BiCLA without the bidirectional cross-sentence attention (**InAtt**). In InAtt, the intra-sentence attention layer is stacked directly on top of representations produced by the Bi-LSTM encoder. The comparison between BiCLA and InAtt directly reveals the contribution that bidirectional cross-sentence attention has on sentence CLIR performance.

Ensemble Models. By design, BiCLA aims to capture semantic similarity stemming from semantic alignments of longer sequences (i.e., phrases, clauses), implicitly capturing semantic compositionality. In contrast, baselines AGG and TbT make simpler similarity assessments: AGG assumes the sentence meaning to be a linear combination of word meanings whereas TbT actually measures the lexical overlap, using the multilingual embedding space as the translation dictionary. Given this complementarity between BiCLA and the baselines, we also evaluate ensemble rankers: they rank target sentences by interpolating between ranks assigned by individual models.

Model Configuration. We train BiCLA in mini-batches of size $N_b = 50$ sentence pairs, each consisting of 10 micro-batches containing one positive sentence pair and $K = 4$ corresponding negative sentence pairs (two created randomly and two using AGG as unsupervised similarity heuristic). We optimize the parameters with the Adam algorithm [15], setting the initial learning rate to 10^{-4}. We tune the hyperparameters on a validation set in a fixed-split cross-validation. We found the following optimal values: BiLSTM state size $H = 100$, regularization factor $\lambda = 10^{-4}$, contrastive margin $\epsilon = 1$. The loss on the validation set was also used as the criterion for early stopping of the training.

4.2 Results and Discussion

First we show the results for the basic mate CLIR evaluation, with train and test set involving the same language pair. We then examine BiCLA's behavior in language transfer settings – a model trained on one language pair is used to perform CLIR for another language pair. In both cases, we evaluate the performance of BiCLA alone and ensembled with the unsupervised CLIR baselines, AGG and TbT.

Task 1: Base Evaluation. We first show the results for the base evaluation task where the train and test set involve the same language pair. We treat each sentence of the source language as a query and the 1,000 test sentences in the target language as a target sentence collection. Performance is reported in terms of the standard mean average precision (MAP) measure. Table 2 summarizes the MAP scores for six language pairs (first language is always the query/source language).

BiCLA strongly outperforms all baselines for all six language pairs.[11] The baselines are able to reduce the gap in performance only for two language pairs: DE-EN and

[11] We tested the significance over 1000 average precision scores obtained for individual queries (which are, in our case, equal to reciprocal rank scores, since there is only one relevant sentence in the other language for each query) using the two-tailed Student's t-test. BiCLA significantly outperforms all baselines with $p < 0.01$.

Table 2. Cross-lingual mate retrieval performance.

Model	DE-EN	CS-EN	HU-EN	CS-DE	HU-DE	CS-HU
QLM	.303	.121	.141	.064	.054	.083
AGG	.390	.547	.372	.374	.356	.378
TbT	.490	.563	.357	.228	.142	.124
InAtt	.506	.597	.462	.495	.422	.404
BiCLA	**.604**	**.665**	**.569**	**.562**	**.577**	**.575**

Table 3. Performance of ensemble models.

	Weights	DE-EN	CS-EN	HU-EN	CS-DE	HU-DE	CS-HU
BiCLA + AGG	.5; .5	.662	.802	.622	.666	.597	.675
	.7; .3	.686	.818	.653	.690	.624	.708
	.9; .1	.706	.784	.660	.691	**.647**	**.702**
BiCLA + TbT	.5; .5	.651	.827	.680	.508	.396	.300
	.7; .3	.650	.832	.685	.553	.442	.351
	.9; .1	.649	.802	.687	.618	.561	.473
BiCLA	$.\dot{3}$; $.\dot{3}$; $.\dot{3}$	.656	.846	.651	.599	.449	.401
+ AGG	.6; .2; .2	.685	**.859**	.683	.647	.525	.463
+ TbT	.8; .1; .1	**.708**	.845	**.726**	**.697**	.626	.568

CS-EN. The baseline scores generally tend to decrease as the languages in the pair become more distant. In contrast, BiCLA exhibits fairly stable performance across all language pairs – the performances for pairs of more distant languages (e.g., HU-DE or CS-HU) are on par with the performances for pairs of closer languages (e.g., DE-EN and CS-DE). Full BiCLA outperforms InAtt, a model without the bidirectional attention layer, by a wide margin, confirming our intuition that fine-grained cross-sentential semantic awareness is crucial for better recognition of sentence translation pairs.

We next investigate the extent to which the supervised BiCLA model complements the state-of-the-art unsupervised CLIR baselines, AGG and TbT [19]. To this end, we ensemble the models at the level of the rankings they produce for the queries. We evaluate three different ensembles: (1) BiCLA and AGG, (2) BiCLA and TbT, and (3) BiCLA and both AGG and TbT. For each of the ensembles we show the performance with different weight configurations, i.e., different weight values w_{BiCLA}, w_{AGG}, and w_{TbT} assigned to individual models, BiCLA, AGG. and TbT, respectively. The results of ensemble methods are shown in Table 3. Almost all ensemble models (exception is BiCLA+TbT with large TbT weight for distant languages), exhibit better performance than BiCLA on its own. BiCLA+AGG and BiCLA+AGG-TbT ensembles with larger weights for BiCLA yield performance gains between 10 and 20% MAP with respect to the BiCLA model alone, suggesting that BiCLA indeed complements the best unsupervised CLIR models. Note that the ensembles in which BiCLA gets a larger weight than the unsupervised models exhibit the best performance. For example, a BiCLA+AGG-TbT ensemble with weights $w_{BiCLA} = 0.8$, $w_{AGG} = 0.1$, and $w_{TbT} = 0.1$ significantly outperforms the same ensemble with equal weights (i.e., $w_{BiCLA} = w_{AGG} = w_{TbT} = 0.\dot{3}$), for all language pairs except CS-EN. Big boosts of ensembles with small contributions from unsupervised baselines suggest that (1) in most cases, BiCLA does a much better job than

Table 4. BiCLA performance in language transfer settings (underlined results denote base CLIR results from Table 2, without language transfer. Bold scores indicate BiCLA transfer scores that are above all baseline scores, cf. Table 2).

	DE-EN	CS-EN	HU-EN	CS-DE	HU-DE	CS-HU
DE-EN	.604	**.602**	**.537**	.231	.177	.213
CS-EN	.356	.665	**.421**	**.440**	.307	**.419**
HU-EN	.440	.484	.569	.371	**.414**	.369
CS-DE	.299	.493	.292	.562	**.374**	**.398**
HU-DE	.365	.398	**.448**	**.399**	.577	.329
CS-HU	.360	**.604**	**.432**	**.524**	**.459**	.575

unsupervised baselines and (2) in cases where BiCLA fails, the unsupervised baselines perform very well – even their small contribution significantly improves the ranking.

Task 2: Zero-Shot Language Transfer. We next investigate the predictive capability of BiCLA in transfer learning settings, that is, we test whether a model trained for one language pair may successfully, both on its own and in ensembles with unsupervised models, perform sentence CLIR for another language pair. The language transfer results are shown in Table 4: rows denote the language pair of the train set and columns the language pair of the test set. We analyze the results in view of three types of language transfer: 1) source language transfer (SLT) (same collection language in training and test); 2) target language transfer (TLT) (same query language in training and test), and 3) full language transfer (FLT) (query and collection language in test are both different than in training). BiCLA outperforms all baselines in 7/8 SLT settings, 5/8 TLT settings, and (only) 2/14 FLT experiments. It is not surprising to observe better performance in SLT and TLT settings than in FLT: BiCLA seems to be able to account for the change of one language, but not for both simultaneously. Although BiCLA alone does not outperform the unsupervised baselines in most FLT setting (e.g., DE-EN→CS-HU), we find the SLT and TLT results significant. Drops in performance in some of the SLT and TLT setups, compared to respective basic setups (no language transfer), are almost negligible (e.g., DE-EN→HU-EN has only a 3-point lower MAP compared to the basic HU-EN setup). SLT and TLT seem to work even when we switch between distant languages (e.g., when we replace the query language from DE to HU), which we find encouraging.

Finally, we evaluate the ensemble between BiCLA and AGG (since, on average, AGG exhibits better performance than TbT) in zero-shot language transfer. We assign equal weights to both rankers, i.e., $w_{BiCLA} = w_{AGG} = 0.5$. The results of the zero-shot language transfer ensembles are shown in Table 5. While on its own BiCLA outperforms the unsupervised models in half of the language transfer setups, when ensembled with the AGG baseline, it drastically boosts the CLIR performance on *all* test collections. This also holds for all FLT setups – where BiCLA is trained on a completely different language pair from the language pair of the test collection. For example, a BiCLA model trained on DE-EN, when ensembled with AGG, boosts the CS-HU retrieval by almost 15 MAP points. We hold this to be the most important finding of our work – it implies that we can exploit readily available large parallel corpora for major languages in order

Table 5. Results of language transfer ensembles. Bold scores denote ensembles that surpass the performance of the unsupervised AGG model alone.

	DE-EN	CS-EN	HU-EN	CS-DE	HU-DE	CS-HU
AGG	.390	.547	.372	.374	.356	.378
Ensemble: BiCLA (transfer) + AGG						
DE-EN	–	**.818**	**.634**	**.520**	**.447**	**.518**
CS-EN	**.564**	–	**.550**	**.630**	**.520**	**.624**
HU-EN	**.578**	**.679**	–	**.556**	**.562**	**.537**
CS-DE	**.549**	**.771**	**.529**	–	**.569**	**.594**
HU-DE	**.544**	**.667**	**.562**	**.550**	–	**.499**
CS-HU	**.544**	**.773**	**.534**	**.639**	**.584**	–

to train a model that significantly improves mate retrieval for pairs of under-resourced languages, for which we have no parallel resources.

5 Conclusion

We have presented a novel neural framework for mate sentence retrieval across languages. We introduced the bidirectional cross-lingual attention (BiCLA) model, a multi-layer architecture which learns to encode sentences in a shared cross-lingual space in such a way to recognize true semantic similarity between sentences: this means that BiCLA is able to distinguish true translation pairs from only semantically related sentences with partial semantic overlap. A series of experiments for six language pairs have verified the usefulness of the model – we have shown that BiCLA outperforms unsupervised retrieval baselines, and that further gains, due to the complementarity of the two approaches, can be achieved by combined ensemble methods. Most importantly, we have shown that the multilingual nature of the BiCLA model allows for a zero-shot language transfer for CLIR: a model trained on one pair of languages (e.g., German and English) can be used to improve CLIR for another pair of languages (e.g., Czech and Hungarian). This indicates that we can perform reliable cross-lingual sentence retrieval even for pairs of resource-lean languages, for which we have no parallel corpora.

In future work, we plan to experiment with deeper architectures and more sophisticated attention mechanisms. We will also test the usability of the framework in other related retrieval tasks and evaluate the model on more language pairs. We make the BiCLA model code along with the datasets used in our experiments publicly available at: `https://github.com/codogogo/bicla-clir`.

Acknowledgments

The work described in this paper has been partially supported by the "Eliteprogramm für Postdoktorandinnen und Postdoktoranden" of the Baden Württemberg Stiftung, within the scope of the AGREE (*Algebraic Reasoning over Events from Text and External Knowledge*) grant. We thank the anonymous reviewers for their useful comments.

References

1. Agirre, E., Banea, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Mihalcea, R., Rigau, G., Wiebe, J.: Semeval-2016 Task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In: SemEval. pp. 497–511. ACL (2016)
2. Artetxe, M., Labaka, G., Agirre, E.: Learning bilingual word embeddings with (almost) no bilingual data. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 451–462. Association for Computational Linguistics, Vancouver, Canada (July 2017), `http://aclweb.org/anthology/P17-1042`
3. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: International Conference on Learning Representations (2014)
4. Ballesteros, L., Croft, B.: Dictionary methods for cross-lingual information retrieval. In: International Conference on Database and Expert Systems Applications. pp. 791–801. Springer (1996)
5. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the ACL **5**, 135–146 (2017), `http://arxiv.org/abs/1607.04606`
6. Brychcín, T., Svoboda, L.: UWB at semeval-2016 Task 1: Semantic textual similarity using lexical, syntactic, and semantic information. In: SemEval. pp. 588–594. ACL (2016)
7. Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., Specia, L.: Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). pp. 1–14. Association for Computational Linguistics, Vancouver, Canada (August 2017), `http://www.aclweb.org/anthology/S17-2001`
8. Conneau, A., Lample, G., Ranzato, M., Denoyer, L., Jégou, H.: Word translation without parallel data. arXiv preprint arXiv:1710.04087 (2017)
9. Dumais, S.T., Letsche, T.A., Littman, M.L., Landauer, T.K.: Automatic cross-language retrieval using latent semantic indexing. In: AAAI spring symposium on cross-language text and speech retrieval. vol. 15, p. 21 (1997)
10. Franco-Salvador, M., Rosso, P., Navigli, R.: A knowledge-based representation for cross-language document retrieval and categorization. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics. pp. 414–423 (2014)
11. Glavaš, G., Ponzetto, S.P.: Dual tensor model for detecting asymmetric lexico-semantic relations. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 1757–1767. Association for Computational Linguistics, Copenhagen, Denmark (September 2017), `https://www.aclweb.org/anthology/D17-1185`
12. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
13. Jelinek, F., Mercer, R.: Interpolated estimation of markov source parameters from sparse data. In: Proc. Workshop on Pattern Recognition in Practice, 1980. pp. 381–402 (1980)
14. Jimenez, S.: SERGIOJIMENEZ at semeval-2016 Task 1: Effectively combining paraphrase database, string matching, WordNet, and word embedding for semantic textual similarity. In: SemEval. pp. 749–757. ACL (2016)
15. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization. In: Proceedings of ICLR (Conference Track) (2015), `https://arxiv.org/abs/1412.6980`
16. Koehn, P.: Europarl: A parallel corpus for statistical machine translation. In: Proceedings of the 10th Machine Translation Summit. pp. 79–86 (2005)
17. Levow, G.A., Oard, D.W., Resnik, P.: Dictionary-based techniques for cross-language information retrieval. Information processing & management **41**(3), 523–547 (2005)

18. Lin, Z., Feng, M., Santos, C.N.d., Yu, M., Xiang, B., Zhou, B., Bengio, Y.: A structured self-attentive sentence embedding. In: Proceedings of the International Conference on Learning Representations (2017)
19. Litschko, R., Glavaš, G., Ponzetto, S.P., Vulić, I.: Unsupervised cross-lingual information retrieval using monolingual data only. arXiv preprint arXiv:1805.00879 (2018)
20. Liu, Y., Sun, C., Lin, L., Wang, X.: Learning natural language inference using bidirectional lstm model and inner-attention. arXiv preprint arXiv:1605.09090 (2016)
21. Losada, D.E.: Statistical query expansion for sentence retrieval and its effects on weak and strong queries. Information retrieval **13**(5), 485–506 (2010)
22. Luong, T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 1412–1421. Association for Computational Linguistics, Lisbon, Portugal (September 2015), http://aclweb.org/anthology/D15-1166
23. Martino, G.D.S., Romeo, S., Barrón-Cedeno, A., Joty, S., Marquez, L., Moschitti, A., Nakov, P.: Cross-language question re-ranking. In: Proceedings of SIGIR. pp. 1145–1148 (2017)
24. Mikolov, T., Le, Q.V., Sutskever, I.: Exploiting similarities among languages for machine translation. arXiv preprint arXiv:1309.4168 (2013)
25. Mimno, D., Wallach, H.M., Naradowsky, J., Smith, D.A., McCallum, A.: Polylingual topic models. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2. pp. 880–889. Association for Computational Linguistics (2009)
26. Moro, A., Raganato, A., Navigli, R.: Entity linking meets word sense disambiguation: a unified approach. Transactions of the Association for Computational Linguistics **2**, 231–244 (2014)
27. Munteanu, D.S., Marcu, D.: Improving machine translation performance by exploiting non-parallel corpora. Computational Linguistics **31**(4), 477–504 (2005)
28. Murdock, V., Croft, W.B.: A translation model for sentence retrieval. In: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. pp. 684–691. Association for Computational Linguistics (2005)
29. Navigli, R., Ponzetto, S.P.: Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence **193**, 217–250 (2012)
30. Nie, J.Y., Simard, M., Isabelle, P., Durand, R.: Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. pp. 74–81. ACM (1999)
31. Oard, D.W.: A comparative study of query and document translation for cross-language information retrieval. In: Conference of the Association for Machine Translation in the Americas. pp. 472–483. Springer (1998)
32. Otterbacher, J., Erkan, G., Radev, D.R.: Using random walks for question-focused sentence retrieval. In: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. pp. 915–922. Association for Computational Linguistics (2005)
33. Pappu, A., Blanco, R., Mehdad, Y., Stent, A., Thadani, K.: Lightweight multilingual entity extraction and linking. In: Proceedings of the Tenth ACM International Conference on Web Search and Data Mining. pp. 365–374. ACM (2017)
34. Pirkola, A.: The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. pp. 55–63. ACM (1998)
35. Platt, J.C., Toutanova, K., Yih, W.t.: Translingual document representations from discriminative projections. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. pp. 251–261. Association for Computational Linguistics (2010)

36. Ponte, J.M., Croft, W.B.: A language modeling approach to information retrieval. In: SIGIR. pp. 275–281. ACM (1998)
37. Rauf, S.A., Schwenk, H.: Parallel sentence generation from comparable corpora for improved smt. Machine translation **25**(4), 341–375 (2011)
38. Resnik, P., Smith, N.A.: The web as a parallel corpus. Computational Linguistics **29**(3), 349–380 (2003)
39. Robertson, S., Zaragoza, H., et al.: The probabilistic relevance framework: Bm25 and beyond. Foundations and Trends® in Information Retrieval **3**(4), 333–389 (2009)
40. Ruder, S., Vulić, I., Søgaard, A.: A survey of cross-lingual word embedding models. arXiv preprint arXiv:1706.04902 (2017), `http://arxiv.org/abs/1706.04902`
41. Smith, J.R., Quirk, C., Toutanova, K.: Extracting parallel sentences from comparable corpora using document level alignment. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. pp. 403–411. Association for Computational Linguistics (2010)
42. Smith, J.R., Saint-Amand, H., Plamada, M., Koehn, P., Callison-Burch, C., Lopez, A.: Dirt cheap web-scale parallel text from the common crawl. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). vol. 1, pp. 1374–1383 (2013)
43. Smith, S.L., Turban, D.H., Hamblin, S., Hammerla, N.Y.: Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In: Proceedings of International Conference on Learning Representations (ICLR 2017, Conference Track) (2017)
44. Socher, R., Chen, D., Manning, C.D., Ng, A.: Reasoning with neural tensor networks for knowledge base completion. In: Proceedings of the 2013 Annual Conference on Neural Information Processing Systems. pp. 926–934 (2013)
45. Song, F., Croft, W.B.: A general language model for information retrieval. In: Proceedings of the eighth international conference on Information and knowledge management. pp. 316–321. ACM (1999)
46. Sorg, P., Cimiano, P.: Exploiting wikipedia for cross-lingual and multilingual information retrieval. Data & Knowledge Engineering **74**, 26–45 (2012)
47. Tian, J., Zhou, Z., Lan, M., Wu, Y.: Ecnu at semeval-2017 task 1: Leverage kernel-based traditional nlp features and neural networks to build a universal model for multilingual and cross-lingual semantic textual similarity. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). pp. 191–197 (2017)
48. Vulić, I., De Smet, W., Moens, M.F.: Cross-language information retrieval models based on latent topic models trained with document-aligned comparable corpora. Information Retrieval **16**(3), 331–368 (2013)
49. Vulić, I., Moens, M.F.: Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In: Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval. pp. 363–372. ACM (2015)
50. Yang, B., Yih, W.t., He, X., Gao, J., Deng, L.: Embedding entities and relations for learning and inference in knowledge bases. In: Proceedings of the 2015 International Conference on Learning Representations (2015)
51. Yih, W.t., Toutanova, K., Platt, J.C., Meek, C.: Learning discriminative projections for text similarity measures. In: Proceedings of the Fifteenth Conference on Computational Natural Language Learning. pp. 247–256. Association for Computational Linguistics (2011)