Investigating Cross-lingual Alignment Methods for Contextualized Embeddings with Token-Level Evaluation

Qianchu Liu, Diana McCarthy, Ivan Vulić, Anna Korhonen
Language Technology Lab, University of Cambridge
English Faculty Building, 9 West Road, Cambridge CB3 9DA, United Kingdom
{ql261,iv250,alk23}@cam.ac.uk, diana@dianamccarthy.co.uk

Abstract

In this paper, we present a thorough investigation on methods that align pre-trained contextualized embeddings into shared cross-lingual context-aware embedding space, providing strong reference benchmarks for future context-aware crosslingual models. We propose a novel and challenging task, Bilingual Token-level Sense Retrieval (BTSR). It specifically evaluates the accurate alignment of words with the same meaning in cross-lingual non-parallel contexts, currently not evaluated by existing tasks such as Bilingual Contextual Word Similarity and Sentence Retrieval. We show how the proposed BTSR task highlights the merits of different alignment methods. In particular, we find that using context average type-level alignment is effective in transferring monolingual contextualized embeddings cross-lingually especially in non-parallel contexts, and at the same time improves the monolingual space. Furthermore, aligning independently trained models yields better performance than aligning multilingual embeddings with shared vocabulary.

1 Introduction

Contextualized embeddings have been shown to achieve superior performance compared to static word embeddings in English (Peters et al., 2018; Devlin et al., 2019). Despite recent efforts to better understand their multilingual variants (Pires et al., 2019), leveraging these available pretrained contextualized embeddings to learn cross-lingual contextualized embeddings is still an under-explored area: past cross-lingual embedding alignment methods have mainly focused on static embeddings (Ruder et al., 2019). In this paper, we introduce a first study that investigates and compares different ways of aligning the pretrained contextualized embeddings. In particular, we make the comparisons focused on the following properties: (1) aligning contextualized embeddings at the level of word tokens versus word types; (2) different training signals: static dictionaries, word alignment, or sentence alignment from parallel data; and (3) aligning different model variants: aligning from independently trained models versus aligning embeddings from a multilingual model with shared vocabulary.

We evaluate the methods on a variety of context-aware tasks. Besides two previously established evaluation tasks (1) Bilingual Contextual Word Similarity (Chi and Chen, 2018) and (2) Sentence Retrieval (Conneau et al., 2017), we introduce a new task: Bilingual Token-level Sense Retrieval (BTSR). It is more challenging than the alternatives as it requires the accurate cross-lingual retrieval of contextualized words on the token level which are disambiguated both in the source and the target language using non-parallel contexts. We provide BTSR task data and run evaluations on two language pairs: English–Chinese (EN–ZH) and English–Spanish (EN–ES). The data and guidelines can be found at: https://github.com/qianchu/BTSR

Our main findings are as follows. (1) Using the average of the contextualized word representations as type-level anchors is effective and robust for aligning pre-trained contextualized embeddings cross-lingually, and can also improve the monolingual contextualized space as it brings the largest gains in English context-aware evaluation compared to results from aligning on other levels. (2) Using a dictionary with a few thousand entries is able to yield performance comparable to leveraging training signals from parallel corpora. (3) Aligning independently trained models performs better than aligning embeddings from a multilingual model trained with shared vocabulary.

2 Related Work

Cross-lingual Word Embeddings. We conduct our experiments using a popular projection-based approach that learns an orthogonal mapping between pretrained embeddings (Xing et al., 2015; Artetxe et al., 2016). The orthogonality of the mapping is crucial as it preserves monolingual invariance and is empirically proven to be more robust (Smith et al., 2017; Xing et al., 2015). This projection-based method can be applied post-hoc on pretrained monolingual embeddings with an exact analytical solution. Moreover, its performance is often competitive to that of jointly trained cross-lingual models using additional bilingual signals in the form of parallel or comparable corpora (Ruder et al., 2019; Glavaš et al., 2019).

However, projection-based cross-lingual embeddings are still predominantly concerned with static word embeddings (Glavaš et al., 2019; Vulić et al., 2019; Mohiuddin and Joty, 2019). Learning cross-lingual contextualized embeddings is still a large unexplored area with only two concurrent papers at the moment. First, Aldarmaki and Diab (2019) adopt the same projection-based approach as our paper to align contextualized embeddings on the token-level using parallel data. They find that context-aware mapping using parallel data outperforms context-independent mappings from static dictionaries on a parallel Sentence Retrieval task. Second, Schuster et al. (2019) introduce anchor embeddings as the average of contextualized embeddings of a word to perform alignment for contextualized models, and show its effectiveness in cross-lingual dependency parsing. These two studies are not directly comparable, whereas our paper provides a comprehensive and systematic comparison of various methods for learning cross-lingual contextualized embeddings and introduces a new and more challenging evaluation task.

Evaluation of (Contextualized) Cross-lingual Embeddings. The traditional task to evaluate cross-lingual embeddings is Bilingual Dictionary Induction (BDI) (Vulić and Moens, 2013; Mikolov et al., 2013a; Gouws et al., 2015): given a source query word, the task is to retrieve the translation word in the target language. The test words in BDI are out-of-context and polysemy cannot be addressed properly. The same issue is found in another relevant lexical task, Cross-lingual Semantic Similarity. (Camacho-Collados et al., 2017).

The only context-aware dataset for evaluating cross-lingual embeddings on the word level is Bilingual Contextual Word Similarity (BCWS) (Chi and Chen, 2018). It challenges a system to predict similarity scores between cross-lingual word pairs with sentential context provided in both languages. However, BCWS does not explicitly test for the retrieval of meaning-equivalent cross-lingual contextualized embeddings, which is explicitly tested in our test. Also, BCWS is only available for one language pair: English-Chinese.

Another task used for evaluating contextualized embeddings is Sentence Retrieval (Aldarmaki and Diab, 2019): given a query source sentence, the task is to retrieve the corresponding parallel sentence in the target language. Sentences can be represented as averages of contextualized embeddings of their constituent words. As the task does not explicitly evaluate at the word level, even if a system cannot accurately capture polysemy, it can rely on other words in the sentence to retrieve the correct parallel sentence. Therefore, Sentence Retrieval may lead to superficially high scores.

Cross-lingual Word Sense Disambiguation. Our new task is also related to Cross-lingual Word Sense Disambiguation (Lefever and Hoste, 2009): given a source language word in context, a system needs to provide the correct sense labels as clustered translation words in a number of target languages. Another related task is Cross-lingual Lexical Substitution (Sinha et al., 2009): the model must provide plausible target language translations for the source language lexical item in the source language context. In contrast, our BTSR task: (1) directly evaluates token-level word representations without the need to predict sense labels from a sense inventory and (2) it contextualizes both the source query and the target candidates ensuring full sense disambiguation. The core differences between the three tasks are illustrated in the following examples below:

(1) Cross-lingual Word Sense Disambigution:
source query: the national [coach] of the Irish teams ...
answer: allenatore (Italian); Fußbaltrainer; Nationaltrainer; Trainer (German); entrenador(Spanish) ...

(2) Cross-lingual Lexical Substitution :
source query: She looked as [severely] as she could muster at Draco.
answer: rigurosamente, seriamente

(3) BTSR:
source query: The reflections included in this document are linked to discussions with many colleagues and friends, in the present [tense].

answer: Scott Peterson metió la pata elfondo y usó el [tiempo] pasado mientras afirmaba que su esposa asesinada estaba viva , lanzando una búsqueda (...)

3 Methods

3.1 Monolingual Contextualized Embeddings

Compared to static word embeddings (Mikolov et al., 2013b; Bojanowski et al., 2017), more recent contextualized embeddings provide dynamic representations for a word in context as hidden layers in a deep neural network. They are typically obtained by unsupervised pretraining based on language modeling objectives (Devlin et al., 2019; Yang et al., 2019). The underlying contextualized method in our study is the pretrained $BERT_{base}$ cased model[1] (Devlin et al., 2019). BERT is trained using a transformer architecture (Vaswani et al., 2017) with masked language modelling (MLM) and next sentence prediction (NSP) tasks. MLM predicts the vocabulary id of a randomly masked word in a sentence based on the word's context. NSP trains text-pair representations to predict whether the text-pair contains consecutive sentences from a monolingual corpus.[2]

We work with two BERT variants. First, we explore aligning independently trained BERT models, that is, models with separate model parameters for each language. For English and Chinese, we align independently trained Chinese and English monolingual models. For Spanish and English, since there is no pretrained BERT Spanish model, we take the Spanish embeddings from the BERT multilingual model and align it with the monolingual English model. We take this alignment as an approximation to aligning two independently trained models. We have also experimented with directly aligning embeddings obtained from the BERT multilingual model, which is a joint model trained with the same model parameters with shared subword vocabulary (Devlin et al., 2019). This means that identical words in two different languages will obtain the same embeddings.

3.2 Orthogonal Mapping and MIM

Given a dictionary with item pairs from source and target languages (s_i, t_i), and matrices S and T that contain the vector representations corresponding to the item pairs in the columns, we follow the standard practice (Glavaš et al., 2019) to find an orthogonal alignment matrix W that minimizes the distance between the transformed matrix WS and T. For improved performance, following Artetxe et al. (2016), we normalize and mean center the embeddings in S and T. The mapping is as follows:

$$W = \arg\min_{W} \|WS - T\|^2 \quad s.t. \quad W^T W = I. \tag{1}$$

The closed-form solution can be found by solving the orthogonal Procrustes problem (Schönemann, 1966) as follows:

$$TS^T = U\Sigma V^T; W = UV^T \tag{2}$$

We also optionally apply a post-processing *Meeting-in-the-Middle* (MIM) technique, recently proposed by Doval et al. (2018). It first calculates the average of each dictionary item representation in a pair after the orthogonal mapping: we denote the matrix U as the matrix where each column is such an average vector. Then, it finds a linear mapping M from both the source language (denoted as M_s) and the target language (M_t) after the previous step of orthogonal mapping to minimize the distance to U via a closed-form solution. Equation (3) formulates how to find M_s, and we do the same from target to source.

$$M_s = \arg\min_{M_s} \|M_s WS - U\|^2 \tag{3}$$

We apply the orthogonal mapping and MIM both on static embeddings (for baselines) and contextualized embeddings. For mapping the contextualized embeddings, we either extract type-level embeddings from the contextualized models to serve as anchors for the alignment using static dictionaries, or we use parallel sentences as dictionary items to directly align contextualized word representations on the token level. We discuss this in what follows.

3.3 Alignment Levels

We explore aligning contextualized models on two levels: *type-level* and *token-level*. Type-level word representation refers to static word representation that assigns one fixed embedding to a word. All the traditional word embedding models (e.g., skip-gram, CBOW, fastText) provide such embeddings, and cross-lingual alignment is typically applied on

[1] To produce the contextualized representation for a word in context, we average the 12 hidden layers of the word's subword representations in BERT and then average the subword representations as input for the cross-lingual alignment. We leave other ways to extract the representations for future work.

[2] We have also experimented with ELMo in lieu of BERT (Peters et al., 2018; Che et al., 2018). However, as we reach similar conclusions in terms of relative performance, while BERT-based cross-lingual embeddings outperform their ELMo-based counterparts in absolute terms, we do not report ELMo's results for brevity. It should be noted that these pretrained models used different training data.

these type-level embeddings (Ruder et al., 2019). Token-level word representation refers to dynamic representations for words *in context*, i.e., contextualized word representations.

Contextualized models such as BERT provide token-level embeddings by default: a natural way to align these embeddings is token-level alignment. This has been proposed concurrently to our work by Aldarmaki and Diab (2019). This method requires token-level training data , e.g., from a word-aligned parallel corpus.

As an alternative, we obtain static type-level representations in the same space as our contextualized embeddings and use these type-level representations as anchors to learn the crosslingual mapping. The type-level anchors can be seen as taking a representative sample of the infinite space of the contextualized embeddings. The mapping learned via the anchors will hopefully be generalizable to align the dynamic token-level contextualized embeddings as well. The advantage of this approach is that we can align the contextualized embeddings with a standard dictionary now that we have one representation per word.

We experiment with two different kinds of anchor type-level embeddings: iso_type and avg_type. The iso_type refers to type-level embeddings that are produced by simply inputting the word in isolation to the contextualized model. Avg_type embeddings are obtained by taking the average of the contextualized representations of a word.[3] The context-average avg_type embeddings has been proposed recently by Schuster et al. (2019). In this work, we provide a systematic comparison of embeddings aligned on the token level, and on the two kinds of type-level alignments.

3.4 Alignment Training Signal

We explore a number of different supervision signals for learning the alignment between monolingual embeddings. First, we evaluate traditional methods that exploit word-level training signals (Ruder et al., 2019). We use (1) a static manually created (i.e., external) dictionary to obtain the alignment, and (2) we rely on word alignments from a parallel corpus as the source of the training signal. For word alignments, we either treat them as a large dictionary to perform type-level alignment or we additionally leverage the context in the aligned sentences to extract a dynamic contextualized dictionary to perform token-level alignment.

We also exploit the training signal coming from the aligned parallel sentences alone without word alignments. We first create sentence representations by averaging type-level or token-level embeddings, and then align the parallel sentence representations from source to target language.

The configurations for learning cross-lingual contextualized word embeddings explored in this work are summarized in Table 1, and we rely on the configuration labels from the table throughout the paper. Type-level configurations which ignore context are treated as baselines.

4 Bilingual Token-level Sense Retrieval Task (BTSR)

Task Description. In §2, we already discussed the main properties of the two other tasks that can be used to evaluate cross-lingual context-aware embeddings: BCWS and parallel Sentence Retrieval. In short, BCWS only measures similarity between cross-lingual word pairs in context, and it does not evaluate the translation capacity of different methods. The Sentence Retrieval task does not evaluate on the word level and can be solved by relying on the context alone.

To bridge this gap in evaluation, we introduce a new task: Bilingual Token-level Sense Retrieval (BTSR). It tests for the retrieval of meaning-equivalent cross-lingual contextualized word embeddings relying on non-parallel context information. Our task can be seen as a contextualized variant of the BDI task. Its comparison to the traditional BDI task is provided in Table 2.

In what follows, we define the BTSR task formally and provide details on how the task data is created. To build a representative sample of contextualized words in the source and target languages, we collect translation pairs and contextualize the word pairs into token-level representations. Then we manually check a sample of the contextualized word pairs to ensure correspondence of sense on the token-level. To understand the effect of the size of the search space, we experiment with 20k and 200k candidates respectively.

Formal Definition. In BTSR, we define S : $s^1_{tk,1}, s^1_{tk,2}, s^2_{tk,1}, \dots, s^n_{tk,m}$ as a set of queries from the source language. A query $s^i_{tk,j}$ is a token-level contextualized representation of the ith source

[3] In practice, we take 1000 random samples for a word from the training data of the parallel corpora used in our experiments.

Component	Options	Label
Alignment Signal	Word alignment from parallel data	wa
	Sentence alignment from parallel data	sa
	MUSE training dictionary	dict
Alignment Level	Token-level alignment	token
	Type-level alignment from context average	avg_type
	Type-level alignment from inputting the word in isolation	iso_type
	Type-level alignment in static embeddings (eg. Fasttext)	type
Models	monolingual English BERT model	mono_en
	monolingual Chinese BERT model	mono_zh
	BERT multilingual English model	multi_en
	BERT multilingual Spanish model	multi_es
	Fasttext baseline	fasttext
Alignment techniques	the original orthogonal linear transformation	orig
	post-processing linear transformation after the orthogonal transformation	mim
Evaluation level	Evaluated on token-level representations	[token]
	Evaluated on type-level representations	[type]

Table 1: Different components used for the model configurations in our evaluation.

BDI		BTSR	
uniform	制服	..[uniforms] were black...	他的[制服].. (His [uniform]..)
subdue	制服	..mosquito was [subdued]..	..[制服]刺客.. (...[subdue] the assassin...)
uniform	一致	the [uniform] convergence of the regular solution	...[一致] 漸近穩定...定理 (the theorem of [uniform] asymptotic stability...)

Table 2: BTSR: examples and a comparison with traditional (non-contextualized) BDI.

word that corresponds to the word's jth sense. Similarly, we define $T : t^1_{tk,1}, t^1_{tk,2}, \dots, t^p_{tk,q}$ as a set of candidates in the target language where each candidate is a contextualized token-level word that represents a specific sense of a word in the target language. For each query s_{tk}, the task is to find a target contextualized token-level word t_{tk} that has the same word sense as in the query. $Sim(s_{tk}, t_{tk})$ is a function that computes the similarity of s_{tk} and t_{tk}. In our experiments, we use cosine similarity. Using $Sim(s_{tk}, t_{tk})$, for each query, we retrieve $t_{tk,i^1}, \dots, t_{tk,i^K}$: the top K most similar token-level contextualized words from the target set T in the cross-lingual space as the nearest neighbours. We report *Precision@K*, i.e. precision of finding the gold t_{tk} in the top K retrieved candidates.

Collecting Translation Pairs. We select a representative set of query words from WordNet (Miller, 1998) (one unique word per WordNet synset). For each source word, we retrieve its WordNet senses and the corresponding translations in the target language from Multilingual WordNet (Bond and Foster, 2013). As WordNet senses are too fine-grained, we collapse senses into clusters if they contain the same translation for the source word. For example, "uniform" has five WordNet senses which are translated into four distinct Chinese words: 制服(the clothes worn by a particular group), 一致(the translation of two senses: consistent and undifferentiated)[4], 不變(unchanged) and 相同(the same) . We take these four Chinese words to form four translation pairs with "uniform".

Word Pair Contextualization. For each word in a word pair, we "contextualize" the word by selecting a sentence in which the word appears, and ensure that the resulting contextualized word can be translated into the other word. Therefore, if a polysemous word occurs in multiple word pairs with distinct translations, it will be accompanied with different contexts that correspond to each translation. We achieve this by selecting a pair of parallel sentences in which the source word and the target word from the word pair are aligned after we run word alignment. The context in the source language in this parallel sentence pair is used to "contextualize" the source word. When we select context for the target word, we choose a different parallel sentence in which the two words in the pair are aligned. Therefore, the final contexts for the source and target word in the word pair are indeed non-parallel.

The use of non-parallel contexts here is crucial because when we perform the token retrieval task, parallel contexts can be superficially retrieved by simply matching the contexts rather than repre-

[4] Notice the senses are different thus contexts are needed to find the pair corresponding to the same meaning.

senting the words in context appropriately. We empirically verified that a simplistic context average baseline outperforms contextualized word embeddings in a variant of our task which relies on parallel contexts.

We set aside 1M parallel sentences from the UMCorpus (Tian et al., 2014) (EN–ZH) and the WMT13 news dataset (Bojar et al., 2013) (EN–ES) for extracting the sentence contexts. We end up with 14,604 distinct word pairs with contexts extracted for EN–ZH, and 9,623 pairs for EN–ES.

Creation of Test Data. As the contexts are non-parallel in a word pair, we need to check if the contextualized words in a word pair genuinely represent the same meaning. We manually checked a sample of the word pairs extracted in the previous step to produce the final test set for BTSR. To produce the sample, we selected the translation pairs that satisfy any of the following constraints: 1) target or source word belongs to the top 250 frequent words in each language, 2) target or source word belongs to the top 250 most ambiguous words in each language. We take the number of sense clusters as introduced above as a measure of ambiguity for each word.

The first author then provided an initial manual annotation of the samples for both EN–ES and EN–ZH on whether the contextualized words in a pair correspond to the same meaning. The samples from the two language pairs were subsequently annotated by one native Chinese speaker and one native Spanish speaker respectively. The final agreement rate calculated as pairwise inter-annotator agreement on a binary choice[5] for EN–ZH is 94.5%, and 94.7% for EN–ES. Finally, we take the subsets where all annotators agree as the test sets for EN–ZH (1,181 pairs) and EN–ZH (994 pairs).

Target Candidates. We treat the token-level representations of the target words from all words pairs in the contextualization process described above as our candidate space. To make the target candidate space more representative of the language, we supplement the space with words outside of the WordNet inventory from monolingual Wikipedia dumps in the target language. For each of these words, we randomly select a sentence in which it occurs to contextualize the word into a token-level target candidate. We experiment with 20k target candidates and 200k target candidates.

5 Experiments

Training Setup. To test the effects of corpora size on the induction of the cross-lingual alignment, we vary the size of the parallel corpus from 100 up to 200k parallel sentences in the UMCorpus and the WMT13 corpus. Word alignment was produced by IBM Model 2 using Fastalign (Dyer et al., 2013). We also induce cross-lingual alignments relying on static dictionaries provided by MUSE (Conneau et al., 2017). BERT variants (see §3.1) are taken from Devlin et al. (2019). For comparison with BERT, we also run fasttext (Bojanowski et al., 2017) to produce baseline static embeddings using the same training Wikipedia corpora for English, Chinese and Spanish.

5.1 Bilingual Contextual Word Similarity

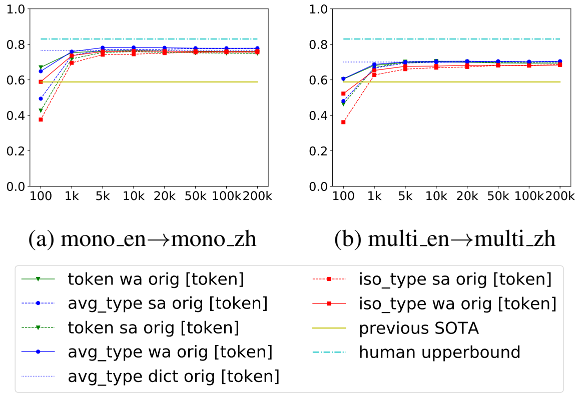
We first evaluate the models on two previous evaluation tasks: BCWS and Sentence Retrieval. For both tasks, we compute cosine similarity to measure the distance between representations. For BCWS, we evaluate embedding distance against human annotations via Spearman correlation. Results on the BCWS task for EN–ZH are shown in Figure 1. The main finding is that all cross-lingual contextualized embeddings in our comparison surpass the previous state-of-the-art (SOTA) based on a cross-lingual multi-sense model (Chi and Chen, 2018) as soon as they are fed 5K or more parallel sentences. Note that the previous SOTA model was trained on the full EN–ZH parallel corpus of around 2M sentences. Although BERT was pretrained on a corpus comprising 3.3B words , it is reasonable to assume that it is easier to procure abundant monolingual data than parallel data. Therefore, aligning pretrained monolingual embeddings using only a small amount of parallel data rather than training on a large parallel corpus is a more favorable choice.

Alignment based on independent monolingual models (mono_en→mono_zh) is particularly effective, achieving human-level performance. While different methods achieve comparable results, avg_type consistently takes the lead.

5.2 Sentence Retrieval

For the Sentence Retrieval task, we compute cosine similarity between the query sentence representation and sentence representations in the tar-

[5] For each language pair, it is calculated as the percentage of token pairs marked correct by both annotators (the first author and one native speaker of the language) divided by the number of all the token pairs.



(a) mono_en→mono_zh (b) multi_en→multi_zh

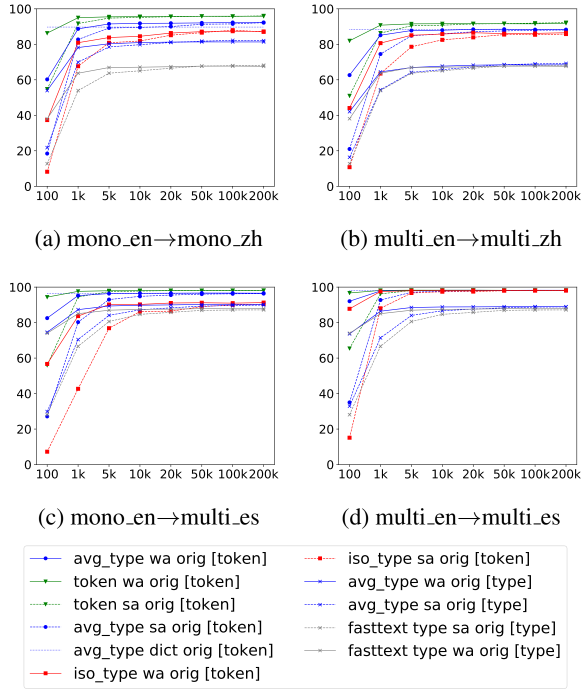
Figure 1: BCWS (Spearman's ρ). The horizontal axis indicates the number of parallel sentences used for learning the alignment transformation. Please refer to Table 1 for understanding the method acronyms in the legend. For example, 'token wa orig [token]' refers to token-level orthogonal mapping trained with word alignment and it is evaluated on token-level data.

get language in the test set of UMcorpus (English-Chinese) and WMT13 corpus (English-Spanish). Precision results for finding the parallel sentence in the top 5 candidates are reported in Figure 2. We find that evaluating with contextualized embeddings on the token-level (all the *[token]* lines) performs consistently better than type embedding baselines. Among the different ways to transfer the contextualized embeddings, aligning directly on the token level with parallel data outperforms aligning via type-level anchoring. Concerning the alignment training signal, sentence alignment starts low but is able to yield comparable results with word alignment after 50K sentences. For the EN–ZH Sentence Retrieval, aligning independently trained BERT models outperforms aligning embeddings with shared vocabulary. For the EN–ES Sentence Retrieval task, aligning from both independent models and from shared embeddings achieves ceiling performance.

5.3 Bilingual Token-level Sense Retrieval

We report *Precision@5* scores for 20k target words in Figure 3. We also report the results from aligning using 200k parallel sentences on BTSR with 200k target words and applying the additional MIM technique in Table 3.

Baselines. We evaluate four baselines that help us better understand the models' performance in this task. For *BL(word)* methods, we discard the contexts and use only the query and target word's type representations. Therefore, polysemous words in the dataset will have only one static representation. We implement both a fasttext baseline and a context-average type embedding baseline for each contextualized model. We also provide baselines which use context but ignore the word in focus (*BL(context)*). These baselines take an average of the context embeddings both at the token level and at the type level of the contextualized models. Instead of finding the best translation word in context, these baselines retrieve the target sentence with the best translation of the source context.[6] Finally, we evaluate a simple baseline that combines both word and context as an average of the two representations. Context representation here is the average of the context embeddings. Both word and context embeddings here are calculated using the avg_type embeddings.



(a) mono_en→mono_zh (b) multi_en→multi_zh

(c) mono_en→multi_es (d) multi_en→multi_es

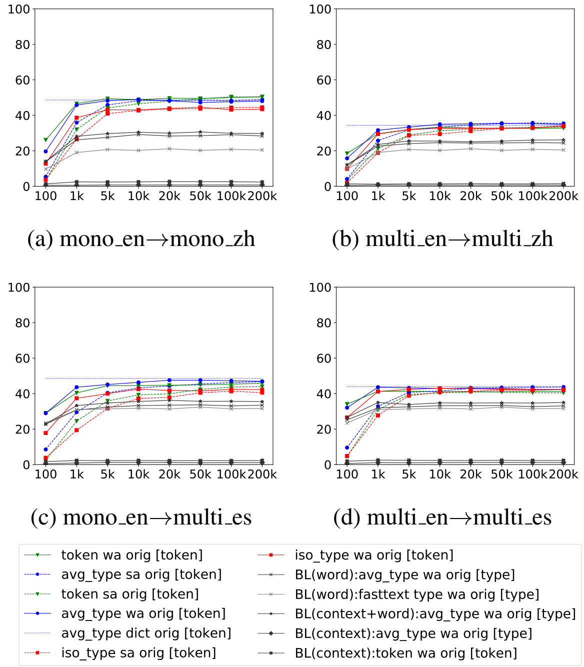
Figure 2: Results on the Sentence Retrieval task from the testset of UMcorpus and WMT13 corpus; the scores are *Precision@5* (%). The horizontal axis indicates the number of parallel sentences used for learning the alignment transformation. Please refer to Table 1 for understanding the method acronyms.

Discussion. The low performance of all the baselines suggest that the proposed task is more challenging than the alternatives: it can not be easily

[6] On our trivial parallel variant of the task, this context baseline gives the best performance.

	token		avg_type		iso_type	
	wa	sa	wa	sa	wa	sa
mono_en→mono_zh	30.84	28.87	**32.04**	31.7	25.43	26.46
+ mim	29.98	30.15	**34.79**	34.45	26.37	27.15
multi_en→multi_zh	17.14	16.97	19.9	**20.84**	16.8	18.17
+ mim	15.93	16.62	21.62	**21.79**	14.81	14.9
mono_en→multi_es	33.47	30.15	**34.37**	33.37	29.25	28.44
+mim	32.46	30.55	**35.38**	33.57	27.34	25.43
multi_en→multi_es	27.14	25.43	**29.35**	29.25	27.04	26.33
+mim	28.44	27.94	**31.86**	31.76	26.73	25.03

Table 3: BTSR results with 200k candidates; alignment learned from 200k parallel sentences. Please refer to Table 1 for the explanation of the acronyms.



(a) mono_en→mono_zh

(b) multi_en→multi_zh

(c) mono_en→multi_es

(d) multi_en→multi_es

Figure 3: EN–ZH and EN–ES BTSR results; *Precision@5* (%). The horizontal axis indicates the number of parallel sentences used for learning the alignment transformation. Please refer to Table 1 for understanding the method acronyms.

tackled by looking at word in isolation (i.e., at type-level representations) or the context alone, or a simple combination of context and the query word.

Regarding the alignment level, compared to the Sentence Retrieval task, the benefit of dynamic token-level alignment from parallel corpora now disappears. Aligning the contextualized embeddings via context-average anchor type embeddings, i.e. avg_type alignment, (which consistently outperform iso_type embeddings) is the best model in most cases, or yields comparable performance with token-level alignment. Their advantage becomes more pronounced in the experiments with 200K target candidates, see Table 3. We suspect that this method is particularly robust when generalizing to words in non-parallel contexts: we find the same pattern in the BCWS task which is also constructed with nonparallel sentences.

Applying MIM brings consistent improvement for the best (avg_type) alignment method. Such improvements for the other methods are less stable. This suggests MIM is only effective when the alignment methods already learn a high-quality cross-lingual space before applying MIM.

As for training signals, relying only on a small dictionary (5K word pairs) yields comparable results with the methods that are trained on large amounts of parallel data. This suggests that a small seed dictionary may be enough to transfer the contextualized embeddings cross-lingually and be able to disambiguate words in context cross-lingually.

When comparing model variants, we see an advantage of aligning independent models over aligning shared models as we increase the training data. This advantage becomes more obvious with 200K target candidates, see Table 3. For EN–ES results in Figure 3, we observe that all alignment methods which use the shared model (i.e., multi_en→multi_es) start higher than results from aligning independently trained mono_en→multi_es. With the 'avg_type wa orig' method for example, aligning mono_en→multi_es starts at 29.04(%) whereas multi_en→multi_es starts at 34.07(%) given 100 parallel sentences. This is intuitive as English and Spanish share a larger portion of their vocabulary compared to English and Chinese: this gives the multilingual model a head start, but it is quickly surpassed by aligning from independently-trained models, especially via the avg_type alignment, as we increase training data.

In sum, we show that (1) BTSR is a challenging task; (2) unlike in Sentence Retrieval, context

	English original		token		avg_type		iso_type	
	mono_en	multi_en	wa mim	sa mim	wa mim	sa mim	wa mim	sa mim
mono_en→mono_zh	76.37	-	76.9	77.98	78.16	**78.28**	73.82	74.37
mono_en→multi_es	76.37	-	75.89	76.76	**77.2**	76.83	73.12	72.05
multi_en→multi_zh	-	72.6	73.56	**75.31**	74.89	75.1	68.55	68.07
multi_en→multi_es	-	72.6	72.3	73.78	**74.1**	73.72	68.43	66.99

Table 4: Evaluating alignment methods and model variants on the monolingual SCWS dataset which measures word similarity in context (in English). Spearman's ρ ($\times$ 100%). Previous best reported score is 69.3 (Neelakantan et al., 2014). Please refer to Table 1 for the explanations of the acronyms.

average type-level alignment performs the best in our task and in the BCWS task where the contexts are non-parallel, and can be further improved with the MIM technique. (3) Using a small dictionary is sufficient to transfer the contextualized embeddings via type-level alignment. (4) Aligning from a shared model gives a head start when two languages contain some shared vocabulary, but aligning from independently trained monolingual embeddings is able to achieve better performance given sufficient training data (5) Overall, increasing the search space from 20K to 200K target words results in a decrease of 10% in precision in BTSR, but the relative performance of different methods is more consistent and more pronounced.

Monolingual Contextual Evaluation. We also examine whether the cross-lingual alignment with MIM post-processing can improve the monolingual contextualized embeddings by evaluating the EN models on the Stanford Contextualized Word Similarity Task which measures similarity of word pairs with context in English. We evaluate the alignments learned from using 200K parallel sentences. The results are in Table 4. It seems that aligning independently trained models, which have better monolingual performance, outperforms aligning from shared models as found in BTSR. Also, we see consistent improvement over the original monolingual space after MIM, especially with avg_type alignment level. This indicates that the avg_type alignment level is effective not only in transferring the contextualized embeddings to the target language, but it can also improve the context-aware monolingual space.

We also observe that the EN contextualized models in their original space (both mono_en and multi_en) outperform SOTA (69.3%), a multi-sense static embedding model (Neelakantan et al., 2014). This indicates that the present contextualized embeddings are already capturing context effect including sense-level information without explicitly assigning embeddings to discrete sense categories.

6 Conclusion

We have conducted novel comparisons and analyses of various alignment methods for aligning contextualized embeddings cross-lingually. We have also introduced a novel task, Bilingual Token-level Sense Retrieval, which directly evaluates the retrieval of meaning-equivalent cross-lingual contextualized embeddings. The proposed task is challenging and enables a finer-grained analysis of different cross-lingual alignment methods. We have found that using context-average type-level alignment (avg_type) is effective and robust in transferring monolingual contextualized embeddings cross-lingually and at the same time improves the monolingual space. Using a small static dictionary as the alignment signal provides comparable results to word alignment methods relying on parallel corpora. We have also found that aligning independently trained monolingual embeddings yields better performance than aligning embeddings from a shared model. As our paper focuses only on the projection-based alignment methods, future work may explore other ways to learn the cross-lingual contextualized embeddings, e.g., based on joint training (Mulcaire et al., 2019).

Acknowledgments

We thank the anonymous reviewers for their helpful feedback on this work. We acknowledge Peterhouse College at University of Cambridge for funding Qianchu Liu's PhD research. The work was also supported by the ERC Consolidator Grant LEXICAL: Lexical Acquisition Across Languages (no 648909) awarded to Anna Korhonen. We also appreciate many helpful discussions and feedback from our colleagues in the Language Technology Lab, in particular Victor Prokhorov, Yi Zhu, Edoardo Ponti, Dr Haim Dubossarsky, and Dr Ehsan Shareghi.

References

Hanan Aldarmaki and Mona Diab. 2019. Context-aware cross-lingual mapping. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3906–3911, Minneapolis, Minnesota. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2294, Austin, Texas. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria. Association for Computational Linguistics.

Francis Bond and Ryan Foster. 2013. Linking and extending an Open Multilingual WordNet. In *Proceedings of ACL*, pages 1352–1362.

Jose Camacho-Collados, Mohammad Taher Pilehvar, Nigel Collier, and Roberto Navigli. 2017. SemEval-2017 task 2: Multilingual and cross-lingual semantic word similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 15–26, Vancouver, Canada. Association for Computational Linguistics.

Wanxiang Che, Yijia Liu, Yuxuan Wang, Bo Zheng, and Ting Liu. 2018. Towards better UD parsing: Deep contextualized word embeddings, ensemble, and treebank concatenation. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 55–64, Brussels, Belgium. Association for Computational Linguistics.

Ta-Chung Chi and Yun-Nung Chen. 2018. CLUSE: Cross-lingual unsupervised sense embeddings. In *Proceedings of EMNLP*, pages 271–281.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yerai Doval, Jose Camacho-Collados, Luis Espinosa Anke, and Steven Schockaert. 2018. Improving cross-lingual word embeddings by meeting in the middle. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 294–304, Brussels, Belgium. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Goran Glavaš, Robert Litschko, Sebastian Ruder, and Ivan Vulić. 2019. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In *Proceedings of ACL*, pages 710–721.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. BilBOWA: Fast bilingual distributed representations without word alignments. In *Proceedings of ICML*, pages 748–756.

Els Lefever and Veronique Hoste. 2009. SemEval-2010 task 3: Cross-lingual word sense disambiguation. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, pages 82–87, Boulder, Colorado. Association for Computational Linguistics.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013a. Exploiting similarities among languages for machine translation. *CoRR, abs/1309.4168*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Proceedings of NeurIPS*, pages 3111–3119.

George Miller. 1998. *WordNet: An electronic lexical database*. MIT press.

Tasnim Mohiuddin and Shafiq Joty. 2019. Revisiting adversarial autoencoder for unsupervised word translation with cycle consistency and improved training. In *Proceedings of NAACL-HLT*, pages 3857–3867.

Phoebe Mulcaire, Jungo Kasai, and Noah A. Smith. 2019. Polyglot contextual representations improve crosslingual transfer. In *Proceedings of NAACL-HLT*, pages 3912–3918.

Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. 2014. Efficient non-parametric estimation of multiple embeddings per word in vector space. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1059–1069, Doha, Qatar. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of ACL*, pages 4996–5001.

Sebastian Ruder, Anders Søgaard, and Ivan Vulić. 2019. A survey of cross-lingual embedding models. *Journal of Artificial Intelligence Research*, 65:569–630.

Peter H Schönemann. 1966. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. In *Proceedigns of NAACL-HLT*, pages 1599–1613.

Ravi Sinha, Diana McCarthy, and Rada Mihalcea. 2009. SemEval-2010 task 2: Cross-lingual lexical substitution. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, pages 76–81, Boulder, Colorado. Association for Computational Linguistics.

Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *arXiv preprint arXiv:1702.03859*.

Liang Tian, Derek F. Wong, Lidia S. Chao, Paulo Quaresma, Francisco Oliveira, Yi Lu, Shuo Li, Yiming Wang, and Longyue Wang. 2014. UM-corpus: A large English-Chinese parallel corpus for statistical machine translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1837–1842, Reykjavik, Iceland. European Language Resources Association (ELRA).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of NIPS*, pages 5998–6008.

Ivan Vulić, Goran Glavaš, Roi Reichart, and Anna Korhonen. 2019. Do we really need fully unsupervised cross-lingual embeddings? In *Proceedings of EMNLP*.

Ivan Vulić and Marie-Francine Moens. 2013. Cross-lingual semantic similarity of words as the similarity of their semantic word responses. In *Proceedings of NAACL-HLT*, pages 106–116.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011, Denver, Colorado. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237.