ECAI 2020
G.D. Giacomo et al. (Eds.)

doi:10.3233/FAIA200332

A Unified and Unsupervised Framework for Bilingual Phrase Alignment on Specialized Comparable Corpora

Jingshu Liu [1, 2] and **Emmanuel Morin**[1] and **Sebastián Peña Saldarriaga**[2] and **Joseph Lark**[2]

Abstract. Significant advances have been achieved in bilingual word-level alignment, yet the challenge remains for phrase-level alignment. Moreover, the need for parallel data is a critical drawback for the alignment task. In particular, this makes multi-word terms very difficult to align in specialized domains. This work proposes a system that alleviates these two problems: a unified phrase representation model using cross-lingual word embeddings as input, and an unsupervised training algorithm inspired by recent works on neural machine translation. The system consists of a sequence-to-sequence architecture where a short sequence encoder constructs cross-lingual representations of phrases of any length, then an LSTM network decodes them w.r.t their contexts. After training, our encoder provides cross-lingual phrase representations that can be compared without further transformation. Experiments on five specialized domain datasets show that our method obtains state-of-the-art results on the bilingual phrase alignment task, and improves the results of different length phrase alignment by a mean of **8.8** points in MAP.

1 Introduction

We consider the problem of bilingual word alignment (BWA) from non-parallel corpora. We are particularly interested in domain-specific words and expressions extracted from modestly sized corpora. Unlike machine translation, which is a text generation task, bilingual alignment is a vector comparison task where word candidates in target language are ranked according to their similarity w.r.t a given word in source language. As a consequence, the vector representation plays a key role.

Beginning with the seminal works of [12] and [37] based on word co-occurrences for BWA, significant improvements have been recently achieved by neural network based approaches [31, 10, 50, 2], but most work on the subject focus on single words. Alignment of multi-word expressions (MWE) from comparable corpora is much less discussed [38, 34]. More in line with our work, [29] proposes a unified approach able to handle single and multiple word expressions at the same time. Since our goal is to handle both without distinction, hereafter we refer to them as phrases.

The vector representation of phrases is one of the crucial parts of alignment. An intuitive baseline representation can be obtained by adding vectors of separate words in the phrase [32, 33, 18, 29]. Regardless of its simplicity, it still hold comparable results on many vector comparison tasks [29, 7]. Additive approaches will succeed when the meaning of the phrase is yielded by the combination of each component, like in "wind turbine" and "life quality". However, less effective representations will be generated when it comes to domain-specific phrases such as "Savonius rotor"[3] and "her3"[4]. Moreover, for phrases whose semantics are predominantly determined by one of its components like in the phrases "*sneaker* shoe" and "blood *serum*", a uniform weight will be given to each component disregarding its importance in the phrase. Finally, an obvious weak point is that additive models ignore word order, hence "control system" and "system control" will have exactly the same representation.

In this paper we propose a new framework for bilingual phrase alignment inspired by recursive neural networks (denoted by RNN in this paper) and encoder-decoder architectures [4, 46]. The RNN capture component word relations and distribute different weights depending on word semantics but require tree structures and task-oriented labeled data to calculate loss during training [44, 45, 22, 35]. As part of our framework, we propose a phrase encoder that can do without a tree structure.

Since the meaning of domain-specific phrases is highly context related, sequence-to-sequence systems fit naturally our needs. After phrase encoding, we can decode its representation to predict its context, thus establishing a relation between the phrase and its context. Unlike common neural machine translation sequence-to-sequence systems, our model encodes a phrase and decodes it with regard to its context. In order to be able to align phrases in different languages, we make the encoder cross-lingual which means that the input vectors in different languages share the same vector space [29, 2]. We also incorporate a back-translation mechanism [40] of single words during training by using pre-trained bilingual word embeddings (BWE). Other than that, our model relies exclusively on monolingual data, and is trained in an unsupervised manner. After completion of training we obtain a shared cross-lingual phrase encoder that can generate a unified representation of phrases of any length.

We provide four open datasets with 108 phrase pairs in the medical domain and 90 pairs in the renewable energy domain with phrase pairs in different languages: English-Spanish, English-French and English-Chinese. We also evaluate our models on two existing datasets of specialized domains [19, 29]. All the training corpora are non-parallel comparable corpus. It is worth mentioning that the small size of the reference lists can be explained by the small size of the specialized corpora which contain few specialized terms and synonym variants [20]. Our experiments on these datasets show that our method outperforms existing unsupervised methods across multiple language pairs.

The remaining of this paper is structured as follows. In Section 2 we present works related to ours, while Section 3 describes our framework. Section 4 presents experiments and results obtained. Fi-

[1] LS2N - UMR CNRS 6004 - Université de Nantes, France
email: {jingshu.liu,emmanuel.morin}@ls2n.fr
[2] Dictanova, Nantes, France, email: firstname@dictanova.com

[3] A kind of rotor used in wind turbines
[4] A human protein

nally Section 5 concludes the paper and opens new perspectives.

2 Related work

In this section we present three concepts related to our method. First we describe cross-lingual word embeddings which are a prerequisite for our model. Next, recursive neural networks approaches are briefly described, finally we describe several neural machine translation training systems which use little or no cross-lingual resources.

2.1 Cross-lingual word embeddings

Following the success of word embeddings [32] trained on monolingual data, a large proportion of research aimed at mapping word embeddings into a common space for multiple languages. Cross-lingual word embeddings were pioneered by [31] by using a linear transformation matrix. A large number of works tried since then to improve the linear transformation method [26, 1, 29]. [2] compiled a substantial amount of similar works [31, 10, 50, 41, 52, 1, 42] into a multi-step bilingual word embedding framework.

2.2 Short sequence representation

The additive approach remains an effective way to encode multi-word expressions. However, it ignores word order and always distributes uniform weights to components regardless of their importance. We can also pre-train phrase embeddings if we consider them as a single token, but it ignores compositionality and inner component relations of the phrase. Besides, learning phrase embeddings as individual vocabulary entries is extremely memory intensive and will lead to a data sparsity problem. Finally, phrases not seen during training cannot be handled by this approach.

[7] used an LSTM to encode two-word phrases in order to enhance English-Estonian machine translation. Pre-trained language models like ELMo [36] or BERT [8] obtained appealing results on various NLP tasks, however pre-trained models such as the BERT multilingual model cannot distinguish words like “pain” in English and “pain”(bread) in French without context, as in our alignment phase. Furthermore in many real life scenarios, although phrase contexts are available during training, during alignment phrases are the only input available.

Recursive neural networks [14] were proposed for encoding hierarchical structured data, they can be seen as a generalization of recurrent neural networks [9] and naturally handle word order in sequences. To distinguish it from the recurrent neural network, we denote it by RNN while the recurrent neural network is denoted by RecurrentNN in this paper. Figure 1 shows an example with three input tokens.

Given a tree structure, e.g. a parse tree, the network visits each node in topological order, applying transformations to generate further representations from previously computed representations, and finally reaching the root level where one single representation is generated for the whole sequence.

The disadvantage of RNN in our scenario is the need of a tree structure because as said before, we do not have the sentence where a phrase appears during alignment, we could parse the multi-word expressions but it seems unreasonable to apply parsing on single words. We would like to train a model similar to a RNN without the required structure information.

This architecture has been successfully exploited in a variaty of tasks, [43] use an *untied weight* RNN for the constituent parsing where they use different weight matrices depending on the constituent syntactic category, [27] collect the context information by adding an outer representation for each node. Their system is served in a dependency parsing task. Besides, various works [45, 22, 35] apply the RNN to generate sentence level representation for the sentiment analysis using some labeled data.

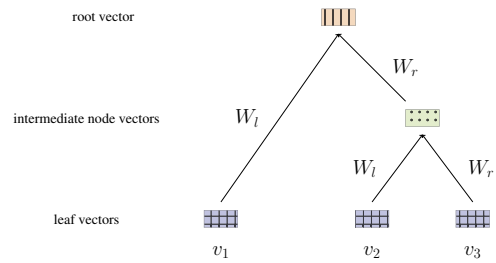

Figure 1. Diagram of a recursive neural network.

2.3 Sequence-to-sequence models in neural machine translation

To train our network, we use the widely exploited sequence-to-sequence model in neural machine translation (NMT) [4, 6, 13, 48]. Although there are many different models, they all implement an encoder-decoder architecture optionally combined with an attention mechanism [4, 30] to tackle long sequences. This type of model has become the main trend in recent years producing the current state-of-the-art results. It takes advantage of longer context information and continuous representations and can be easily trained in an end-to-end system.

In [6], a model to learn representations of variable-length sequences was proposed, however this approach requires parallel phrase pairs for training. Therefore we looked at NMT models making use of monolingual corpora to enhance translation in low-resource scenarios. When no parallel data exists between source and target languages, several works proposed the use of a pivot language [11, 39, 5] acting as a bridge between source and target. Following the same idea, [23] proposed a multilingual NMT model which creates an implicit bridge between language pairs for which no parallel data is used for training. Whether explicit- or implicitly, all these works still require the use of parallel corpora between the pivot language and other languages.

More interestingly for our work, a few researches have been recently conducted on training NMT models with monolingual corpora only [25, 3, 51]. They all use pre-trained cross-lingual word embeddings as input. Then a shared encoder is involved to encode different noised sequences in the source and the target languages. The decoder decodes the encoded vector to reconstruct its original sequence. This strategy is called *denoising* [49] with the objective to minimize the following cross-entropy loss:

$$\begin{aligned} &\mathcal{L}_{denoising}(\theta_{enc}, \theta_{dec}) = \\ &- \mathbb{E}_{x \in D_l} H(x, dec_{\rightarrow l}(enc(\mathcal{N}(x)))) \end{aligned} \tag{1}$$

where θ_{enc} and θ_{dec} respectively means the parameters in the encoder and the decoder, $x \in D_l$ is a sampled sequence from the monolingual data and $dec_{\rightarrow l}(enc(\mathcal{N}(x)))$ represents a reconstructed sequence from the noised version of the original sequence x.

These works exploit the back-translation approach [40] to build the link between the two languages by alternatively applying the source-to-target model to source sentences to generate inputs for training the target-to-source model (and vice versa):

$$\begin{aligned} &\mathcal{L}_{backtranslation}(\theta_{enc}, \theta_{dec}) = \\ &- \mathbb{E}_{x \in D_{l1}} H(x, dec_{\rightarrow l1}(enc(y))), \\ &y = transl(x) = dec_{\rightarrow l2}(enc(x)) \end{aligned} \tag{2}$$

where D_{l1} is the source language corpora, $dec_{\rightarrow l1}$ means that the decoder will decoder the sequence in $l1$ language (or $l2$ resp.). Suppose y is the translation of $x \in D_{l1}$, then $dec_{\rightarrow l1}(enc(y))$ represents the reconstructed source sentence from the synthetic translation.

Also pertaining to our work, [51] introduce a semi-shared encoder to retain specific properties of each language, and directional self-attention to model word order. Based on the works discussed in this section, we propose an encoder-decoder network with a novel encoder architecture and a new unsupervised training objective, more details will be presented in the next section.

3 Proposed method

We present our unified and unsupervised framework in this section, first we introduce the tree-free phrase encoder which is a short sequence encoder in Section 3.1. Then Section 3.2 describes the global architecture of our system and how we train it without parallel data.

3.1 Tree-free phrase encoder

Recall that our objective is to encode multi-word or single-word phrases without tree structures, because we want our framework to be robust enough without too much constraint. Extra information such as tree structures is not always available. To this end we propose a network that can be seen as a short sequence encoder similar to a recursive neural network with three levels. In the first level we split the semantics of each word by a linear transformation into two parts: the right side and the left side. Then we associate these nodes by concatenation, the left side is supposed to be associated with the right side of the previous token and vice versa. In fact, by doing this we recreate a pseudo-tree structure where each word is directly associated with its left and right neighbours. The second level is composed of a fully connected layer that maps the input vectors to output vectors in a specified dimension. Finally the third level consists in summing all intermediate level nodes and outputting a single fixed-length vector. The sum operation is motivated by the additive characteristics mentioned in [32]. Instead of summing directly over the input, we sum over the combination of each pair to capture word inner relationship.

Figure 2 shows the schema of the proposed network. The input of the network is a sequence of word vectors for a phrase of any length, and the output is a fixed length vector which can be considered as the representation of the whole sequence.

We use pre-trained cross-lingual embeddings as the input vector sequence $[v_1, v_2, v_3, ..., v_n]$ with $v_i \in \mathbb{R}^d$, the output vector $v_o \in \mathbb{R}^p$ is calculated as follows:

$$\begin{aligned} v_{i,l} &= \tanh(W_l v_i + b_l) \\ v_{i-1,r} &= \tanh(W_r v_{i-1} + b_r) \\ v_{inter,i} &= \tanh(U[v_{i-1,r}; v_{i,l}] + b) \\ v_o &= \sum_{i=1}^{n} v_{inter,i} \end{aligned} \tag{3}$$

where $W_l \in \mathbb{R}^{d \times d}$ and $W_r \in \mathbb{R}^{d \times d}$ denote respectively the left and the right weight matrix in the linear transformation of the semantic association, $b_l \in \mathbb{R}^d$ and $b_r \in \mathbb{R}^d$ are the corresponding bias vectors, $U \in \mathbb{R}^{p \times d}$ and $b \in \mathbb{R}^p$ are the parameters in the fully connected layer with d the input dimension and p the output vector dimension.

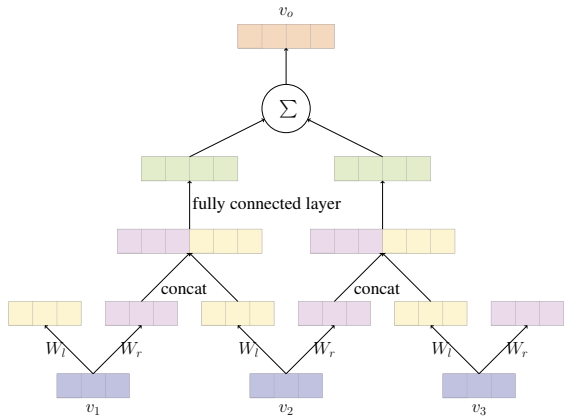

Figure 2. Illustration of the tree-free phrase encoder.

Consequently, our phrase encoder produces vector representations that are word order sensitive and that can distribute different weights for the different phrase components without using structured input. In addition, the output dimension can be different from the input dimension unlike in the original RNN.

3.2 System overview

The general encoder-decoder architecture of our method is shown in Figure 3. Since the input sequence is always a phrase, usually much shorter than a sentence, we did not use attention which is intended to capture long-range dependencies. The network tries to predict the sentence containing the input phrase from its encoded vector.

For a phrase x in language D_{l1}, we first use the shared tree-free phrase encoder, then the system can be trained in two subnetworks. The first networks decodes the encoded output v_0 of x w.r.t D_{l1}. The second network is applied if x is a single-word phrase, it decodes the encoded vector of $BWE(x) \in D_{l2}$ w.r.t D_{l2}. We alternatively iterate through all the phrases in the two languages.

As illustrated in Figure 3, we input synthetic translations of single words to the second subnetwork. Hence, we create a link between the two languages in the absence of parallel bilingual data. This can be seen as a pseudo back-translation mechanism based on bilingual word embeddings [2, 29] in place of decoding it into the target translation.

The language decoders consist of a 2-layer LSTM and a fully connected layer on top of it. The goal of the decoder is to reconstruct the *wrapped sentence* which contains the current input phrase. We name this process *context prediction*. The intuition behind context prediction is based on the distributional hypothesis [17], i.e., *words in similar contexts tend to have similar meanings*. Predicting the context from a central language unit has already been applied in Skip-gram which predicts the context independently. Using a generator to generate the context can be seen as a conditional Skip-gram which predicts

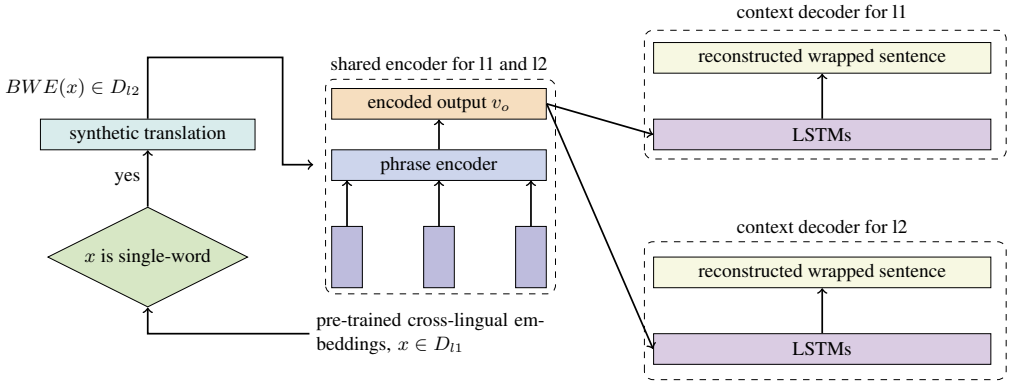

Figure 3. Overview of the training architecture. The objective of the decoder is to reconstruct a *wrapped sentence* containing x (as described in Section 3.2).

a continuous sequence where each predicted token is related. In addition, in [7], instead of an end-to-end system, the authors first learn all the phrase embeddings by Skip-gram considering them as a single word, and then learn the composition function by a regression model which predicts the pre-trained phrase embeddings from its composing word embeddings. However they limit the phrase length to 2, while we would like to propose a unified end-to-end framework which is able to learn the phrase composition of variable length and the mapping simultaneously. Overall, the system uses three key concepts:

Wrapped sentences. Like in NMT, we use special tokens to mark the start and the end of a sentence. But, apart from the standard special tokens, we insert a universal ⟨PHR⟩ token to the left and right border of each phrase in a sentence. This allows the system to recognize the phrase when decoding and strengthen links between languages.

Shared phrase encoder. The system treats input phrases in different languages via the universal encoder detailed in 3.1. Works using a similar idea are [21, 28] and, [3]. As the input embeddings are already mapped to a common space, the representation generated by the shared encoder is also a cross-lingual vector. After the training, we use exclusively the shared encoder (TF-RNN) to generate cross-lingual phrase representations, which is essential for our final task: bilingual phrase alignment.

Pseudo back-translation. Since we do not have cross-lingual data, a direct link between a phrase in language $l1$ and one in language $l2$ is not feasible. However, synthetic translations of single words can be easily obtained using bilingual word embeddings. By using translated single-word phrases to train our model, we create stronger links between the two languages. This can be viewed as pseudo *back-translation* as we generate synthetic translations by BWE while in NMT systems the translation is generated by the corresponding decoder [40, 3].

Therefore, the system potentially has four objective loss functions when we alternatively iterate all phrases in the two languages $l1$ and $l2$:

$$\mathcal{L}_{cp\ l1\to l1}(\theta_{enc}, \theta_{dec_{\to l1}}) = - \mathbb{E}_{x\in D_{l1}} H(ws(x), dec_{\to l1}(enc(x))), \tag{4}$$

$$\mathcal{L}_{cp\ l2\to l1}(\theta_{enc}, \theta_{dec_{\to l1}}) = - \mathbb{E}_{x\in D_{l1}} H(ws(x), dec_{\to l1}(enc(\text{BWE}(x)))), \tag{5}$$

$$\mathcal{L}_{cp\ l2\to l2}(\theta_{enc}, \theta_{dec_{\to l2}}) = - \mathbb{E}_{x\in D_{l2}} H(ws(x), dec_{\to l2}(enc(x))), \tag{6}$$

$$\mathcal{L}_{cp\ l1\to l2}(\theta_{enc}, \theta_{dec_{\to l2}}) = - \mathbb{E}_{x\in D_{l2}} H(ws(x), dec_{\to l2}(enc(\text{BWE}(x)))) \tag{7}$$

where $\mathcal{L}_{cp\ lp\to lq}$ means the *context prediction* loss from an encoded phrase in language lp to the context of language lq, $dec_{\to l}(enc(x))$ is the reconstructed version of the wrapped sentence, $ws(x)$ denotes the real wrapped sentence containing the phrase x and $\text{BWE}(x)$ is the translated single-word phrase for x using bilingual word embedding. The alignment process for the reference phase consists in a forward pass of the trained shared encoder for the source phrase and all the candidate phrases in target language, then candidates are ranked using cosine similarity.

4 Experiments

We evaluate the proposed method on four publicly available datasets across three language pairs: English-French, English-Spanish and English-Chinese. We first cover the datasets and the experiment settings, then we present the baseline method along with the results of our experiments.

4.1 Resource and data

Two corpora from specialized domains were used in our experiments: *breast cancer* (BC) and *wind energy* (WE), these are comparable corpora meaning that texts are not translations even if they share the same domain. The BC corpus has English and Spanish texts, while the WE corpus has English, French, Spanish and Chinese texts.

We use the *IXA pipes* library[5] to tokenize and lemmatize French and Spanish corpora. It is worth noting that the WE Chinese corpus is already pre-segmented. Then we use the *Stanford CoreNLP* library[6] pos-tagger for all languages, then pos-tags are used to extract phrases of a maximal length of 7 tokens using PKE[7]. After hapax filtering, each corpus contains roughly 6,000 phrases.

BC corpus was crawled from a scientific website[8]. The corpus is based on open access articles in English and Spanish related to breast

[5] http://ixa2.si.ehu.es/ixa-pipes/
[6] https://stanfordnlp.github.io/
[7] https://github.com/boudinfl/pke
[8] https://www.sciencedirect.com

cancer and related pathologies (e.g. ovarian cancer). The English corpus has 26,716 sentences and the Spanish one has 62,804 sentences. The gold standard was constructed based on the MeSH 2018 thesaurus[9] and contains 108 phrase pairs. We made the English-Spanish breast cancer corpus and all the reference lists publicly available.[10]

WE corpora are available for download from the authors' academic website[11]. The English, French, Spanish and Chinese corpora contain 13,338, 33,887, 29,083 and 17,932 sentences respectively. [19] proposed a reference list consisting of 139 single words for the English-French corpus, while [29] provided a gold standard with 73 multi-word phrases for the same corpus. Based on the reference list of [29], we propose a new gold standard including also single words. Moreover, we extended this gold standard to other languages while ensuring that all reference lists share the same 90 English phrases to be aligned. Finally, reference lists were obtained for three languages pairs: English-French, English-Spanish and English-Chinese. For the sake of comparability, we report results on the datasets previously published in [29] and [19]. Again, on the domain specialized corpora, it is difficult to build a large reference list of the non general meaning phrases because domain specific terms represent a small subset of all the n-grams [20, 29].

4.2 Experiment settings

We implement the bilingual word embedding framework mentioned in Section 2.1 using *deeplearning4j 1.0.0-beta3*[12]. We also use this library to train domain-specific 100-dimensional word embeddings using the Skip-gram model, with 15 negative samples and a window size of 5. Since our corpora are fairly small, we concatenate these embeddings with the 300-dimensional *fastText* vectors pre-trained on the *wikipedia* [15][13], resulting in 400-dimensional vectors. This technique follows the idea discussed in [20] and, [29]. Next we apply the bilingual word embedding framework so all word embeddings at the input level in each experiment are mapped to a common space. For each language pair, the seed lexicon is selected by a frequency threshold of 50, obtaining around 2,000 word pairs. We use unit length normalization, mean centering, matrix whitening, re-weighting and the de-whitening to generate cross-lingual word embeddings. Since our goal is to evaluate the contributions of our system, we will not measure the impact of different pre-trained embeddings and focus those achieving state-of-the-art results to date.

The dimension of the encoded vector (v_o in Figure 2) for the shared encoder is set to 500. This is also the hidden size for the LSTM decoders. A max length of 100 tokens is applied to discard the long sentences so that the training is quicker and more stable. The model is trained by a minibatch of 20. We run our experiments for a maximum of 200 epochs with an early-stop condition of three consecutive loss increases. Each model takes about 2 days to train on a single Geforce 1080 Ti GPU with Pytorch 1.0 and Cuda 10 on Ubuntu 16.04.

4.3 Reference methods

We implement two baseline methods to compare with our approach. The first one is a traditional co-occurrence based approach while the second used bilingual word embeddings and vector sum to generate phrase representations. We also compare the result of our phrase encoder to the results of existing neural networks used as encoders.

Co-occurrence based approach The compositional approach [16, 47, 38] is a quick and direct method to align multi-word expressions. It is basically a dictionary look-up approach which translates each word via a dictionary and sort all candidates by frequency. [34] proposed a co-occurrence based approach called *compositional approach with context based method* (CMCBP) to tackle the problem of out of dictionary words. However, this approach can only align phrases of the same length, so we compare only a subset of multi-word phrase pairs.

Addition based BWE approach. Using the addition to generate multi-word expressions is originally mentioned in [32]. [29] make a deeper review of this method and apply it to bilingual phrase alignment. By summing up cross-lingual word vectors in a phrase to represent the whole sequence, we can directly align each phrase with this approach. We also implement this approach for the sake of comparison, this method is fully comparable with our system.

Context prediction with baseline phrase encoders. We implement four baseline phrase encoders based on regular neural networks which do not require structured input: **RecurrentNN** (referred to as Rec. below), **CNN LSTM** and a **Transformer encoder** [48] (referred to as TXM below). LSTM is reported to obtain the best results in [7]. To be comparable with the other architectures, we use a 4 headed transformer cell with 4 hidden layers. The output dimension of the transformer encoder is the same as the input word embedding dimension, 400. The other encoders have the same output dimension of 500 and the CNN has a kernel size of 2 and a zero-padding so that even single-word phrases can be encoded. It should be worth mentioning that we have tested both the max and the average pooling for the CNN and TXM encoders because they both output a sequence of representations without a pooling layer. We only report the best results in the next section.

4.4 Results and discussion

Table 1 shows the overall results on all test phrases. Since the distributional approach [34] does not include the alignment of variable length phrases, we ignore the corresponding results in the table.

It is also worth noting that we do not present the results of the distributional approach on the English-Chinese corpus because we do not have enough resources to build the co-occurrence matrix as in the other language pairs[14].

It is clearly shown that the proposed method has a better overall performance. Especially when it comes to different length phrase alignment, the new approach improves significantly the results with an average score of 8.8 points. This proves that the proposed method is able to produce high-quality alignment for phrases of variable length. Keep in mind that the different length distribution represents a small proportion of all test phrases except for the English-Chinese corpus, so the overall score would be furthermore improved if we has uniform distribution over all kinds of alignment. The second best method on overall results is the addition approach, previously reported to obtain decent results [32, 7]. However, we observe that between linguistically distant language pairs (English-Chinese), all

[9] https://meshb.nlm.nih.gov/search
[10] https://bitbucket.org/stevall/phrase-dataset
[11] https://www.ls2n.fr/corpus-comparable-multilingue
[12] https://deeplearning4j.org/
[13] https://github.com/facebookresearch/fastText/
[13] This is the optimal application order reported by [2].

[14] The distributional approach requires a high coverage bilingual dictionary, furthermore if the dictionary does not use the same Chinese word segmentation approach as the WE corpus, it is even harder to find words in it.

Table 1. Overall MAP % for all phrase alignment. *sw*, *n2n* and *p2q* respectively mean single-word to single-word, same length multi-word and variable length phrase alignment.

Dataset		Method		Encoder				Our
Corpus	Phrases	CMCBP	ADD	Rec.	CNN	LSTM	TXM	method
BC en-es	sw (72)	35.72	47.46	46.71	45.12	46.25	43.37	**47.76**
	n2n (21)	68.73	81.10	28.52	62.10	50.05	59.26	**86.11**
	p2q (9)	-	42.18	1.11	10.65	7.04	4.49	**49.11**
	all (108)	-	52.85	36.78	43.04	43.72	43.22	**55.40**
WE en-fr	sw (15)	65.56	78.25	77.22	78.33	79.36	**85.56**	79.44
	n2n (61)	42.09	57.37	6.16	40.84	18.64	41.82	**62.19**
	p2q (14)	-	15.83	<0.5	10.07	9.09	12.35	**37.95**
	all (90)	-	55.77	17.25	43.33	27.42	44.53	**62.10**
WE en-es	sw (15)	63.35	77.92	**88.89**	75.78	87.18	84.44	87.62
	n2n (61)	35.94	**62.68**	7.31	40.33	23.07	44.68	61.35
	p2q (14)	-	43.28	<0.5	28.57	17.86	37.20	**46.21**
	all (90)	-	62.20	19.77	44.41	32.94	50.14	**63.38**
WE en-zh	sw (17)	-	53.43	70.26	**76.47**	71.43	65.92	66.50
	n2n (47)	-	23.34	17.53	16.55	**25.24**	18.86	23.01
	p2q (26)	-	4.97	5.13	7.60	2.37	5.80	**12.32**
	all (90)	-	22.67	23.91	25.28	27.36	23.98	**28.13**
WE en-fr	n2n (40)	67.32	78.36	46.07	68.51	44.82	48.47	**88.01**
	p2q (33)	-	34.38	2.38	20.01	7.93	28.25	**41.83**
Liu2018	all (73)	-	58.48	26.06	46.59	28.13	39.33	**67.13**

encoder-decoder systems outperform the addition based approach. The CNN has some interesting results in same length alignment and the LSTM is powerful at short phrase alignment but unlike in [7], it falls much behind on other types. This difference may be explained by the fact that they limit the alignment to two-word phrases. The Transformer encoder does not obtain better results than the addition based approach nor much better results than the other encoders. First, the addition is still more adaptive and effective for short sequence comparison between linguistically close language pairs [20, 29, 7]. Second, as we set a maximum epoch of 200, we think that the transformer encoder may not be converged after 200 epochs because it has a much bigger parameter-sample ratio than the other encoders. Finally, Transformer architectures are basically multi-head self-attentions which are designed for capturing the relations in long sequences while we encode mostly short sequences. On the English-Spanish and English-Chinese *wind energy* corpus, the addition based approach slightly outperforms our approach by 1.33 and 0.33 points for the same length alignment but falls much behind our proposal in other types of alignment.

The relative poor results on the English-Chinese corpus may be due to the segmentation of Chinese words. More concretely, as the input vectors for the Chinese sequences are in word-level, many words in our gold standard are not segmented in the same way as in the given corpus which is already pre-segmented. We would like to incorporate character-level embeddings in our future works as this particularly makes sense in Chinese. Concerning the single-word alignment on BC, 25 among the 72 single words are in fact acronyms which are particularly difficult to align. This would explain why the single-word alignment has much poorer results than other distributions. Besides, the proposed method obtains strong results for single-word alignment, we believe this happens because the system sees more single-word alignment samples generated by the pseudo back-translation during training.

In order to show that the proposed method can still hold a reasonable performance on single words, we present in Table 2 the results for single words compared to state-of-the-art work on bilingual word embedding [2], including the 139 English-French single word dataset of [19] (suffixed -HM in the Table 2). In order to be comparable, we only test on single-word phrases and the candidate list is limited to all single words in the corpus vocabulary.

We can see that in general, compared to [2], the proposed approach does not degrade much the results except for the English-Chinese words. In addition to that, we succeed to hold better results with regard to the original transformation matrix method [31] with only one exception on the English-French *wind energy* dataset. This shows that our approach is not biased by the compositionality of the multi-word expressions.

Table 2. MAP (%) for bilingual alignment of single words only.

	BC	WE			
Method	**en-es**	**en-fr**	**en-es**	**en-zh**	**en-fr-HM**
Mikolov et al. [31]	39.96	91.33	87.27	45.88	79.47
Artetxe et al. [2]	**49.13**	**95.56**	**90.39**	**73.52**	**84.01**
Our method	45.96	89.44	88.89	58.75	82.23

4.5 Qualitative analysis

For a better understanding of how the proposed method succeed or fail to align different types of phrases, we analyzed some of the alignments proposed by our system.

Table 3 shows examples extracted from top 2 nearest candidates to the source phrase in column 2. Again we see that the proposed method is capable of generating better results over different types

of alignment. In the first example, with the proposed approach, the source phrase *breast cancer* is aligned to *cáncer de mama* (lit. "cancer of breast") which is the expected phrase in Spanish and is far more idiomatic than *cáncer mamario* (lit. "cancer mammary") obtained by the addition approach. In line 7 we see that the perfect translation for *wind vane* is found by our proposal: 风向标, while the additive approach finds 偏航␣电机 (lit. "yaw electric machine"). Besides, examples in lines 3, 5, 6, 7 and 8 are all composed of phrases of variable length, the corresponding reference phrase can be found in the fourth column. Interestingly, we find that the proposed system find paraphrases referring to fairly domain-specific phrases like *blade tip* which is aligned to *côté supérieur de la pale* (lit. "side top of the blade"). This is also the case for *Darrieus rotor* aligned to *rotor vertical*, which is remarkable since the Darrieus rotor is a kind of vertical rotor.

Table 3. Phrase alignment examples within top 2 candidates. ("␣" is the segmentation point for Chinese words)

Dataset	Source	Addition	Our method
BC	breast cancer	cáncer mamario	cáncer de mama
en-es	cell death	muerte celular	muerte
WE	blade tip	angle des pales	côté supérieur de la pale
en-fr	Darrieus rotor	rotor tripale	rotor vertical
WE	airflow	freno aerodinámico	flujo de aire
en-es	wind power plant	electricidad del viento	planta eólica
WE	wind vane	偏航␣电机	风向标
en-zh	electricity power	电力	电力

Though the proposed method performs generally well on phrases, we observe that it emphasizes occasionally too much the syntactic head in a multi-word phrase. For instance, in the second example, *cell death* is aligned to *muerte* ("death"), while the addition based approach succeeds to align it to *muerte celular* (lit. "death cellular") which is the reference phrase in Spanish. Undoubtedly, *death* is the syntactic head for the noun phrase *cell death*, it is clear that the proposed method puts more weight on the syntactic information rather than the compositional property for this phrase. This also explains why we do not obtain better results on equal-length phrase alignment on the English-Spanish and English-Chinese *wind energy* corpora (Table 1). This bias could be due to the increased amount of single-word phrase samples of the pseudo back-translation reinforced learning. This suggests that we could improve the system by adding synthetic translations for multi-word phrases while training.

5 Conclusion

This work proposes an unsupervised bilingual alignment framework for phrases of variable length. We implement an adapted encoder-decoder system that uses pre-trained cross-lingual embeddings as input. The system does not require parallel data but instead includes a shared encoder and a pseudo back-translation mechanism. Our experiments show that our proposal improves significantly the results of different length phrase alignment compared to existing methods (+8.8 in MAP) while holding comparable results on equal length phrases. It should be emphasized that this work focuses on the phrase alignment task, one may argue that the output of this task is actually a cross-lingual phrase table for statistical machine translation (SMT). Indeed, we could have applied our output to the SMT task however this does not directly evaluate our phrase encoder for the alignment. Nonetheless, for our future work we would like to evaluate our system in downstream tasks such as SMT.

Despite of the strong empirical performance, one aspect of our method that we identified as sub-optimal is that the pseudo back-translation is only used for single words, therefore we would like to explore strategies for generating synthetic translations of multi-word phrases. In addition, incorporating character-level input vectors may allow us to extract more versatile features to further improve the performance, particularly in Chinese. Finally we look forward to incorporate recently released cross-lingual pre-trained models [24].

REFERENCES

[1] Mikel Artetxe, Gorka Labaka, and Eneko Agirre, 'Learning principled bilingual mappings of word embeddings while preserving monolingual invariance', in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP'16)*, pp. 2289–2294, (2016).

[2] Mikel Artetxe, Gorka Labaka, and Eneko Agirre, 'Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations', in *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI'18)*, pp. 5012–5019, (2018).

[3] Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho, 'Unsupervised neural machine translation', in *Proceedings of the 6th International Conference on Learning Representations (ICLR'18)*, (2018).

[4] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, 'Neural machine translation by jointly learning to align and translate', *CoRR*, **abs/1409.0473**, (2014).

[5] Yun Chen, Yang Liu, Yong Cheng, and Victor O.K. Li, 'A teacher-student framework for zero-resource neural machine translation', in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL'17)*, pp. 1925–1935, (2017).

[6] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, 'Learning phrase representations using rnn encoder–decoder for statistical machine translation', in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*, pp. 1724–1734, (2014).

[7] Maksym Del, Andre Tättar, and Mark Fishel, 'Phrase-based unsupervised machine translation with compositional phrase embeddings', in *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pp. 361–367, (2018).

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, 'Bert: Pre-training of deep bidirectional transformers for language understanding', *CoRR*, **abs/1810.04805**, (2018).

[9] Jeffrey L. Elman, 'Finding structure in time', *Cognitive Science*, **14**(2), 179–211, (1990).

[10] Manaal Faruqui and Chris Dyer, 'Improving vector space word representations using multilingual correlation', in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL'14)*, pp. 462–471, (2014).

[11] Orhan Firat, Baskaran Sankaran, Yaser Al-Onaizan, Fatos T. Yarman Vural, and Kyunghyun Cho, 'Zero-resource translation with multi-lingual neural machine translation', in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP'16)*, pp. 268–277, (2016).

[12] Pascale Fung, 'Compiling bilingual lexicon entries from a non-parallel english-chinese corpus', in *Proceedings of the 3rd Annual Workshop on Very Large Corpora (VLC'95)*, pp. 173–183, (1995).

[13] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin, 'Convolutional sequence to sequence learning', in *Proceedings of the 34th International Conference on Machine Learning (ICML'17)*, pp. 1243–1252, (2017).

[14] C. Goller and A. Küchler, 'Learning task-dependent distributed representations by backpropagation through structure', in *Proceedings of International Conference on Neural Networks*, pp. 347–352, (1996).

[15] Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov, 'Learning word vectors for 157 languages', in *Proceedings of 11th edition of the Language Resources and Evaluation Conference (LREC'18)*, pp. 3483–3487, (2018).

[16] Gregory Grefenstette, 'The world wide web as a resource for example-based machine translation tasks', in *Proceedings of the ASLIB Conference on Translating and the Computer 21*, (1999).

[17] Zellig Harris, 'Distributional structure', *Word*, **10**(2–3), 146–162, (1954).

[18] Amir Hazem and Béatrice Daille, 'Word embedding approach for synonym extraction of multi-word terms', in *Proceedings of the 11th edition of the Language Resources and Evaluation Conference (LREC'18)*, pp. 297–303, (2018).

[19] Amir Hazem and Emmanuel Morin, 'Efficient data selection for bilingual terminology extraction from comparable corpora', in *Proceedings of the 26th International Conference on Computational Linguistics (COLING'16)*, pp. 3401–3411, (2016).

[20] Amir Hazem and Emmanuel Morin, 'Bilingual word embeddings for bilingual terminology extraction from specialized comparable corpora', in *Proceedings of the 8th International Joint Conference on Natural Language Processing (IJCNLP'17)*, pp. 685–693, Taipei, Taiwan, (2017).

[21] Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma, 'Dual learning for machine translation', in *Advances in Neural Information Processing Systems 29 (NIPS'16)*, pp. 820–828, (2016).

[22] Ozan Irsoy and Claire Cardie, 'Deep recursive neural networks for compositionality in language', in *Advances in Neural Information Processing Systems 27 (NIPS'14)*, pp. 2096–2104, (2014).

[23] Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean, 'Google's multilingual neural machine translation system: Enabling zero-shot translation', *Transactions of the ACL*, **5**, 339–351, (2017).

[24] Guillaume Lample and Alexis Conneau, 'Cross-lingual language model pretraining', *arXiv preprint arXiv:1901.07291*, (2019).

[25] Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato, 'Unsupervised machine translation using monolingual corpora only', in *Proceedings of the 6th International Conference on Learning Representations (ICLR'18)*, (2018).

[26] Angeliki Lazaridou, Georgiana Dinu, and Marco Baroni, 'Hubness and pollution: Delving into cross-space mapping for zero-shot learning', in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP'15)*, pp. 270–280, (2015).

[27] Phong Le and Willem Zuidema, 'The inside-outside recursive neural network model for dependency parsing', in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*, pp. 729–739, Doha, Qatar, (2014).

[28] Jason Lee, Kyunghyun Cho, and Thomas Hofmann, 'Fully character-level neural machine translation without explicit segmentation', *Transactions of the ACL*, **5**, 365–378, (2017).

[29] Jingshu Liu, Emmanuel Morin, and Sebastián Peña Saldarriaga, 'Towards a unified framework for bilingual terminology extraction of single-word and multi-word terms', in *Proceedings of the 27th International Conference on Computational Linguistics (COLING'18)*, pp. 2855–2866, (2018).

[30] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning, 'Effective approaches to attention-based neural machine translation', in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP'15)*, pp. 1412–1421, (2015).

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

[32] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean, 'Distributed representations of words and phrases and their compositionality', in *Advances Neural Information Processing Systems 26 (NIPS'13)*, pp. 3111–3119, (2013).

[33] Jeff Mitchell and Mirella Lapata, 'Language models based on semantic composition', in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP'09)*, pp. 430–439, (2009).

[34] Emmanuel Morin and Béatrice Daille, 'Revising the compositional method for terminology acquisition from comparable corpora', in *Proceedings of the 24rd International Conference on Computational Linguistics (COLING'12)*, pp. 1797–1810, (2012).

[35] Romain Paulus, Richard Socher, and Christopher D Manning, 'Global belief recursive neural networks', in *Advances in Neural Information Processing Systems 27 (NIPS'14)*, pp. 2888–2896, (2014).

[36] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer, 'Deep contextualized word representations', in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL'18)*, pp. 2227–2237, (2018).

[37] Reinhard Rapp, 'Automatic identification of word translations from unrelated english and german corpora', in *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, pp. 519–526, (1999).

[38] Xavier Robitaille, Yasuhiro Sasaki, Masatsugu Tonoike, Satoshi Sato, and Takehito Utsuro, 'Compiling french-japanese terminologies from the web', in *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL'06)*, pp. 225–232, (2006).

[39] Amrita Saha, Mitesh M. Khapra, Sarath Chandar, Janarthanan Rajendran, and Kyunghyun Cho, 'A correlational encoder decoder architecture for pivot based sequence generation', in *Proceedings of the 26th International Conference on Computational Linguistics (COLING'16)*, pp. 109–118, (2016).

[40] Rico Sennrich, Barry Haddow, and Alexandra Birch, 'Improving neural machine translation models with monolingual data', in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL'16)*, pp. 86–96, (2016).

[41] Yutaro Shigeto, Ikumi Suzuki, Kazuo Hara, Masashi Shimbo, and Yuji Matsumoto, 'Ridge regression, hubness, and zero-shot learning', in *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD'15)*, pp. 135–151, (2015).

[42] Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla, 'Offline bilingual word vectors, orthogonal transformations and the inverted softmax', in *Proceedings of the 5th International Conference on Learning Representations (ICLR'17)*, (2017).

[43] Richard Socher, John Bauer, Christopher D. Manning, and Ng Andrew Y., 'Parsing with compositional vector grammars', in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL'13)*, pp. 455–465, Sofia, Bulgaria, (2013).

[44] Richard Socher, Christopher D. Manning, and Andrew Y. Ng, 'Learning continuous phrase representations and syntactic parsing with recursive neural networks', in *Proceedings of the NIPS-2010 Deep Learning and Unsupervised Feature Learning Workshop*, pp. 1–9, (2010).

[45] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts, 'Recursive deep models for semantic compositionality over a sentiment treebank', in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP'13)*, pp. 1631–1642, (2013).

[46] Ilya Sutskever, Oriol Vinyals, and Quoc V Le, 'Sequence to sequence learning with neural networks', in *Advances in Neural Information Processing Systems 27 (NIPS'14)*, pp. 3104–3112, (2014).

[47] Takaaki Tanaka, 'Measuring the similarity between compound nouns in different languages using non-parallel corpora', in *Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*, pp. 1–7, (2002).

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin, 'Attention is all you need', in *Advances in Neural Information Processing Systems 30 (NIPS'17)*, pp. 5998–6008, (2017).

[49] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol, 'Extracting and composing robust features with denoising autoencoders', in *Proceedings of the 25th International Conference on Machine Learning (ICML'08)*, pp. 1096–1103, (2008).

[50] Chao Xing, Dong Wang, Chao Liu, and Yiye Lin, '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 (NAACL'15)*, pp. 1006–1011, (2015).

[51] Zhen Yang, Wei Chen, Feng Wang, and Bo Xu, 'Unsupervised neural machine translation with weight sharing', in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL'18)*, pp. 46–55, (2018).

[52] Yuan Zhang, David Gaddy, Regina Barzilay, and Tommi Jaakkola, 'Ten pairs to tag – multilingual pos tagging via coarse mapping between embeddings', in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL'16)*, pp. 1307–1317, (2016).