Dependency Based Bilingual word Embeddings without word alignment

1st Taghreed Alqaisi
department of computer science
University of York
York, UK
ta808@york.ac.uk
Taibah University
Madinah, KSA
taqesi@taibahu.edu.sa

2nd Alexandros Komninos
department of computer science
University of York
York, UK
alexandros.komninos@york.ac.uk

2nd Simon O'Keefe
department of computer science
University of York
York, UK
simon.okeefe@york.ac.uk

***Abstract*—In this work, we trained different bilingual word embeddings models without word alignments (BilBOWA) using linear Bag-of-words contexts and dependency-based contexts. BilBOWA embedding models learn distributed representations of words by jointly optimizing a monolingual and a bilingual objective. Including dependency features in the monolingual objective, improves the accuracy of learning bilingual word embeddings up to 6% points in English-Spanish (En-Es) and up to 2.5% points in English-German (En-De) language pairs in word translation task compared to the baseline model. However, using these dependency features in both monolingual and bilingual objectives does not lead to any improvement in the En-Es language pair and only shows minor improvement for En-De. Moreover, our results provide evidence that using dependency features in bilingual word embeddings has a different effect based on syntactic and sentence structure similarity of the language pair.**

***Index Terms*—word embeddings, bilingual word embeddings, dependency context, syntax features.**

I. Introduction

Word embedding has shown a positive effect on various natural language processing (NLP) tasks due to its ability to distribute word embeddings into a low dimensional continuous vector space, according to the syntactic and semantic similarities between these words. [12] presents a successful bag of words-based word embedding method that improves many NLP applications in monolingual scenarios, including language modelling ([9], [10], [13]), machine translation ([14], [15], [16]), named entity recognition [17], document classification, sentiment analysis [18], [19] and [20] and parsing [32].

In cross-lingual scenarios, many works have been introduced for bi/cross-lingual word embedding. Bilingual word embedding methods aim to drive similar words into a shared vector space of different languages. These introduced methods can be classified into three categories based on how the parallel corpus is used with different alignment levels:

- A word aligned dictionary [21]–[24].
- Phrase/Sentence-aligned parallel corpus [25], [26].
- Word and sentence level alignment datasets [25], [26].
- None aligned comparable datasets [28].

[21] extends the skip-gram model [12] to learn an efficient bilingual word embedding. While [22] introduces a bilingually-constrained phrase embeddings (BRAE) model that learns source-target phrase embeddings by minimising the semantic distance between translation equivalents and maximising the semantic distance between non-translation equivalents. Then [23] extends the BRAE model by introducing a "bilingual correspondence recursive autoencoder" (BCorrRAE) model by incorporating a word alignment that learns better bilingual phrase embeddings by capturing different levels of their semantic relations. An attention-based method has been introduced by [24]. It introduces a Bidimensional attention-based Recursive AutoEncoder (BattRAE) model that learns bilingual phrase embeddings by integrating source-target interactions at different levels of granularity.

With sentence level alignment, recently, models such as the BilBOWA model [25] and the Transgram method [26] have been introduced to learn and align word embeddings without word alignment. Moreover, [27] proposes a Bilingual paRAgraph VEctors (BRAVE) model that learns bilingual embeddings from either a sentence-aligned parallel corpus or label-aligned non-parallel document corpus. While a multilingual (two or more languages) word embeddings model that uses document-aligned comparable data has been proposed by [28].

[29] utilises bilingual word embeddings with syntactic dependency (DepBiWE). In this model, they extract context from dependency parsed trees to be used jointly with Bag-of-words context to learn bilingual word embeddings.

As obtaining word alignment is an expensive process in terms of time and data, we propose a bilingual model which is an extension to the BilBOWA model. The main difference between the two models lies in integrating dependency context (Dep-BilBOWA). The BilBOWA model is trained by jointly optimising a monolingual objective for each language and a bilingual objective that aligns the representations of the two languages. The skip-gram objective with negative sampling is used as the monolingual objective and the bilingual objective minimises the Euclidean distance of the Bag-of-words

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

representation between the two languages in the embedding space. We propose two methods to add syntactic information to BilBOWA model. Using a dependency based skip-gram model for the monolingual objective while keeping the bilingual objective the same (MonoDep-BilBOWA), or extending the Bag-of-words representation with dependency features for the bilingual objective (BiMonoDep-BilBOWA).

The main contribution in this paper is to consider different syntactic structures in learning bilingual word representations without word alignment. In this work, we show that one of the proposed models, namely MonoDep-BilBOWA model, learns better bilingual word embeddings using Bag-of-words and dependency contexts.

In this paper, we extend the BilBOWA model by integrating dependency features in both monolingual and bilingual objectives to investigate their effects on learning bilingual word embeddings on the cross-lingual dictionary induction (CLDI) task.

In Section II, we give an overview of some related recent work on dependency-based word embeddings. Section III describes the proposed models. The next section is the experimental section that contains the training dataset, preprocessing settings and training hyper-parameters for each trained model. This is followed by the evaluation section which explains the evaluation method and presents the results. We then discuss the trained models evaluation results in more details. Finally, we draw final conclusions in Section VII.

II. Related Work

A. Monolingual dependency-based Word Embeddings

Since the success of Bag-of-words context for learning word embedding models, a few dependency-based word embedding models have been introduced in the literature. The research shows that syntax-based embeddings have different properties to word similarity evaluations, as they are known to capture better functional properties of words compared to their window based counterparts embedding models.

Recently, a few researchers have proposed dependency-based word embedding models that integrate dependency contexts to capture syntactic features from the sentence to train skip-gram model variations [7], [6], [5]. [7] modified the skip-gram model by replacing the linear Bag-of-words context with context features from a word's neighbourhood in a dependency graph as shown in Figure 1. While [6] propose another variation of dependency-based skip-gram word embedding model. They extend the notion of token co-occurrence in a dependency neighbourhood to include additional pairs compared to the model of [7]. In addition, they show that the dependency features can be used in various sentence representations to improve performance in several sentence classifications tasks. Also, [5] introduce a multi-order dependency-based context into the skip-gram model with adaptive dependency weights.

B. Bilingual Dependency-based Word Embeddings

In terms of the learning process, bilingual word embeddings have been classified into three categories namely, monolingual mapping, cross-lingual training and joint optimisation approaches. In monolingual mapping, after learning word representations separately for each language, the model learns a transformation matrix to map the word representation from one language to the word representation from the other, using word translation pairs [4]. Parallel corpus models require either word-level [3] or sentence-level alignments [2], [1] [25]. These models aim to have the same word/sentence representations for equivalence translations.

Finally, in the joint optimisation method, the monolingual and cross-lingual objectives are optimised jointly to enforce bilingual constraints [25], [26]. [25] proposes a bilingual Bag-of-words without word alignment model (BilBOWA) that uses a skip-gram model as the monolingual objective. It jointly learns the bilingual embeddings by minimising the distance between aligned sentences, by assuming that each word in the source sentence is aligned to all words in the target sentence. The model can utilize large amounts of monolingual data along with a few translation pairs of sentences. The model shows success in the English-Spanish (En-Es) translation task and the English-German (En-De) languages pair in document classification task.

Recently, [29] proposes a first model that learns bilingual word embeddings using syntactic dependencies. Their model learns the bilingual word embeddings using both dependency context and Bag-of-words context. As with the Bag-of-words method, word order has been ignored in cross-lingual scenarios as it can produce context words that are not related to the target words. [29] obtains the dependency contexts of aligned words to capture the syntactic information among languages.

III. Models

Recently, the use of bilingual/cross-lingual word embeddings has attracted many researchers' attention due to the importance of learning word representations that capture the relations among languages [25]. The BilBOWA model is a simple, efficient model to learn bilingual distributed word representations without word alignment [25]. Therefore, in this paper, we proposed dependency-based bilingual word embeddings models that extend the BilBOWA model to incorporate sentences' syntactic information.

A dependency representation of a sentence is a directed graph with one node per word and type labelled edges representing the syntactic relations between nodes. We use Universal Dependencies (UD) [33] as the syntactic relation types. The UD types are specifically designed to be consistent among different languages, making them suitable for multilingual syntactic analysis. The dependency features are extracted from the parse tree to implement DepBilBOWA models using different settings – modelling dependency features at monolingual objective (MonoDep-BilBOWA), and modelling dependency features at both monolingual and bilingual objectives (BiMonoDep-BilBOWA).

A. Bilingual Word Embeddings without Word Alignment (BilBOWA) (Baseline model)

In this work, we train BilBOWA models[1] for En-Es and En-De language pairs as a baseline. Using a sentence-level aligned corpus, the Baseline-BilBOWA model assumes that each word in the source language sentence is aligned to every word in the target language sentence and vice versa. (This feature is an advantage of this model as the word alignment process is very time consuming). In the BilBOWA model, both monolingual and bilingual objective functions are learnt jointly.

- Model 1: BilBOWA (Baseline model)
  - Monolingual Features
    The BilBOWA model learns monolingual word representations using a skip-gram model with the negative sampling approach by [12]. The skip-gram model learns distributed representations of words by estimating the conditional probability of a target word w occurring in the context of word c. The (target, context) pairs are determined by a context definition function, which is typically a predefined window around each target word. To avoid the computational cost of estimating a categorical distribution over all possible words, the objective is converted to a binary classification problem. The target word is assigned a positive label and a small number of sampled words are used as the negative samples. The skip-gram with negative sampling training objective for a single sample is given in [12] as:

$$\log \sigma(v_w^{'T} u_{cp}) + \sum_{i=ng}^{NG} E_{w_i \sim P_n(w)} [\log \sigma(-v_w^{'T} u_{cn})] \quad (1)$$

    where $v_w \in \mathbb{R}^k$ denotes the target word representation, u_{cp}, $u_{cn} \in \mathbb{R}^k$ represent positive and negative context word representations respectively, NG is the number of negative samples and σ is the sigmoid logistic function. The objective is averaged over each word instance in the corpus and maximized by stochastic gradient ascent. The skip-gram model maintains two different representations of each word: v to be used as the target word and u to be used a context word. The sampling distribution $P_n(w)$ is the unigram distribution of words estimated by their frequency in the training corpus, raised to the power of 3/4. skip-gram also sub-samples training instances based on the frequency of the target word, i.e. instances of frequent words have a higher probability of being skipped during training, which results in better representations for rare words as their contribution to the objective increases.
    This method allows the model to learn high-quality monolingual features as well as speeding up the computation process [11], [25].

  - Bag-of-words Bilingual/Cross-lingual Features
    The bilingual word embeddings are learnt by minimising the distance between source and target sentence representations in each aligned sentence pair. In other words, the model minimises the mean square error loss between sentence representation pairs, where sentence representations are computed as the mean of their word embeddings.
    [25] defines the bilingual objective as:

$$\Omega = ||\frac{1}{m}\sum_{i=1}^{m} v_i - \frac{1}{n}\sum_{j=1}^{n} v_j||^2 \quad (2)$$

    where m and n are the number of words in the source and target language, and v_i and v_j denote the word representations for each language respectively. While this objective can be trivially minimised by setting all the vectors equal to zero, when used along with the monolingual objective it acts as a regularizer that forces the word representations of the two languages to share a common aligned space, where translation word pairs are close.

B. Dependency Based Bilingual Word Embeddings without Word Alignment(Dep-BilBOWA)

As a main contribution in this work, we propose two different dependency-based BilBOWA models that learn word representations by updating the shared embeddings jointly for both monolingual and bilingual objectives using additional features, namely dependency context features. As it has been mentioned above, the BilBOWA model uses a skip-gram model to learn monolingual relations between words in the same language. In this paper, we follow the work of [6], which extends the use of the skip-gram model to integrate dependency contexts with Bag-of-words contexts, as explained below.

- Model 2: Monolingual Dependency-Based model (MonoDep-BilBOWA)
  At the monolingual level, dependency-based skip-gram embedding models learn representations by extracting (target, context) token pairs from dependency graphs instead of word sequences. To encode the graph's structure, they use two types of tokens: words and dependency features. Words correspond to nodes of the dependency graph and dependency features are composite features representing a node and an incident edge as a unit. We denote dependency features as a concatenated string of the edge type and word. The direction of the edge is encoded by adding a `^-1` to the edge type if it is an outgoing edge. Dependency-based skip-gram models jointly learn distributed representations of both token types using the same objective as skip-gram, but change the context definition that determines co-occurring tokens from a window to a node neighbourhood.
  The extended dependency-based skip-gram [6] defines context as the (target, context) token pairs that can be extracted within the one-hop neighbourhood of a

[1] https://github.com/gouwsmeister/bilbowa

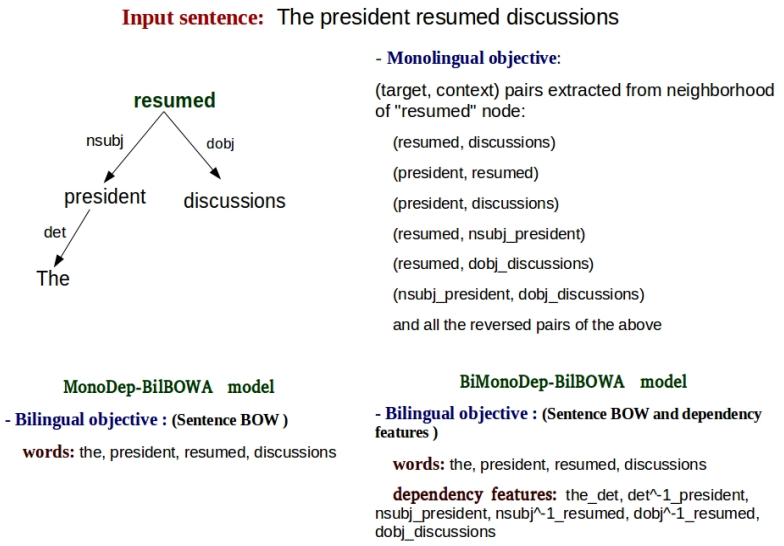

Fig. 1. Model 2 and Model 3 input features example

dependency graph node. In particular, pair extraction is performed by visiting each node in the dependency graph and constructing one bag with the neighbouring words and one bag with the dependency features formed by the neighbouring nodes and their edges. The centre node is added to both bags. The (target, context) pairs are then all the ordered pairs of tokens that can be formed within each of the two bags. In this model, the bilingual objective remains the same as the baseline model (Bag-of-words sentence representations), as is shown in Fig. 1.

- Model 3: Bi/Mono-lingual Dependency-Based model (BiMonoDep-BilBOWA)
  In addition to the dependency-based monolingual objective, and similar to the baseline, the dependency-based bilingual objective minimises the loss between sentence representation pairs. The Bag-of-words representation for sentences is modified to include syntactic information by adding dependency features extracted from the sentence's dependency graph. The sentence's distributed representation is then formed by the mean of embeddings of all the sentence tokens (words and dependency features) in the bag. As the number of dependency features (twice the number of edges in the graph) is larger than the number of words in the sentence, a weighting scheme can be applied to balance their contribution in the representation [6]. Alternatively, we can represent each sentence with two separate feature bags, one for each token type, and form two aligned representations for each parallel sentence pair (For example, See Fig. 1).

IV. Implementation

We trained three different versions of the BilBOWA model for En-Es and En-De: Baseline-BilBOWA, MonoDep-BilBOWA and BiMonoDep-BilBOWA models. We used the Euoropal parallel corpus v7 for monolingual training, and the News Commentary v8 parallel corpus (that has been provided for statistical machine translation tasks [31]) to train with the bilingual objective. For more details, see Table I. We use the same code as [26] to train the models. Our implementation is based on the observation that the extended dependency skip-gram can be trained as a window-based skip-gram by an appropriate transformation of the input. For each context neighbourhood in the corpus we create two auxiliary sentences, one with the word context features and one with the dependency context features. Each sentence consists of all the tokens in the target word's neighbourhood in any order. Setting

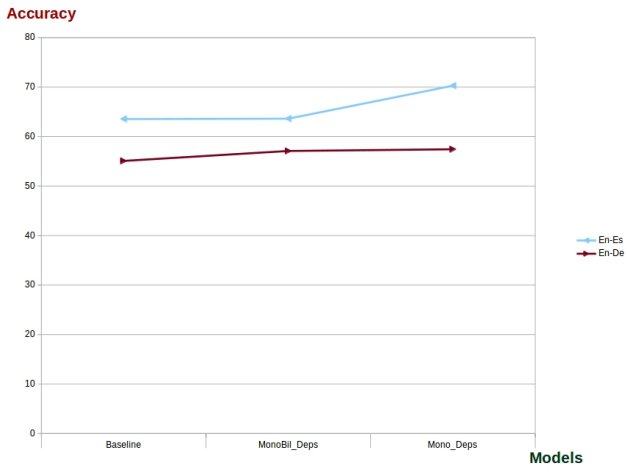

Fig. 2. Experiment results

TABLE I
TOKENISED AND CLEANED DATASETS

Language pair	Europal v7			News		
En-Es	**Sentences**	**tokens**	**MonoDep tokens**	**Sentences**	**tokens**	**BiDep tokens**
en	1916071	51520106	301933100	132571	3280918	9406630
es	1916071	53804104	316022276	132571	3737853	10737580
En-De						
en	1879003	50896257	297861530	176850	4492424	13123596
de	1879003	48458495	283234958	176850	4547691	13289413

TABLE II
PRECISION AT K ON WORD-LEVEL TRANSLATION TASK

En-Es	**k=**1	**k=**3	**k=**5
Baseline	63.54	76.42	78.74
MonoDep-BilBOWA	**70.28**	**82.3**	**84.38**
BiMonoDep-BilBOWA	63.62	75.4	80.22
En-De	**k=**1	**k=**3	**k=**5
Baseline	55.08	68.89	72.5
MonoDep-BilBOWA	**57.44**	**70.62**	**73.82**
BiMonoDep-BilBOW	57.09	70.14	73.26

the window size larger than the length of the longest auxiliary sentence (or equivalently larger than the maximum degree of the dependency graphs in the corpus) results in creating all the positive pairs defined by the extended dependency skip-gram model. We note that no undesired pairs are created by having a large window because windows do not go across line breaks. We can create a Bag-of-words sentence representation with dependency features for the bilingual objective by including all the dependency context features to the Bag-of-words representation of the sentence. To implement the weighting scheme of [6] where word and dependency tokens are given equal weight, we instead form two aligned sentences per original sentence pair, one for each type of token. The models were trained with 200 dimensional word embeddings, with window size 35, and 15 negative samples for 5 epochs using stochastic gradient decent.

A. Datasets and Preprocessing

In all our experiments, the datasets used have been tokenised, lower-cased and the empty lines have been removed. For the other models, a dependency parser has been used to parse the Europarl v7 and News Commentary v6 parallel corpus. Then, we extracted the dependency contexts from the parsed datasets, to be used for monolingual and bilingual training. For parsing, we used a neural network based model for joint part-of-speech (POS) tagging and dependency parsing, introduced by [30] [2]. This model is an extension of the BIST graph-based dependency parser discussed in [8]. They incorporating BiLSTM-based tagging to predict POS tags for the parser automatically. We parsed the En-Es and En-De Europarl datasets to be used in the monolingual objective to train MonoDep-BilBOWA. For BiMonoDep-BilBOWA, Europarl and News Commentary datasets for the same languages pairs have been cleaned and preprocessed to train this model with monolingual and bilingual objectives respectively. After preprocessing and parsing the datasets, the number of features (tokens) have increased dramatically as shown in Table I. The increase happens due to multiple dependency features being extracted for each word.

[2]https://github.com/datquocnguyen/jPTDP

V. EVALUATION

In a similarly way to [25], the trained bilingual word embeddings have been evaluated on the Cross Language Dictionary Induction (CLDI) task, which is a word translation task. The exact setting was first introduced by [11]. To perform this evaluation, firstly, we created two testing dataset pairs, for En-Es and En-De language pairs. We extracted the most frequent 4,000 words from the Europarl En-Es and En-De datasets. Then a dictionary was created for each language pair by translating the extracted words using the Google translator. After having these translation pairs (w_{l1}, w_{l2}), we calculate the precision at k for word translation by finding w_{l2} in the nearest top-k neighbours (1,3 and 5) to w_{l1} in the embedding space. We computed the mean precision from 10 runs, each time randomly selecting 500 source words and their k nearest neighbours. The results from our experiments are shown in Table II.

VI. RESULTS AND DISCUSSION

In our experiments, comparing the three different trained models with different dependency settings allows us to investigate the effect of utilising dependency context features on the process of learning bilingual word embeddings at monolingual and bilingual objectives.

The experiments conducted show that incorporating dependency-based features at the monolingual level has a positive effect on the learning process. These dependency contexts lead to better learning of bilingual word embeddings in the CLDI task compared to the baseline BilBOWA model.

In contrast, the BiMonoDep-BilBOWA model, that uses dependency features with monolingual and bilingual objectives, has not improved the learning process and produces similar results to the BilBOWA baseline model using En-Es language pair (See Table II).

Using different language pairs (En-Es and En-De) with different levels of language differentiation, our experiments show that the language pair with similar sentence structure (En-De) learns better bilingual word embeddings using dependency features at the bilingual level, and the accuracy increased in the CLDI task by more than 2.5% points compared to the baseline, as shown in Fig. 2.

VII. Conclusion

We compare three different BilBLOWA models using different contextual features: no dependency features, dependency features at monolingual level and dependency features at both mono/bilingual levels. Our results show that dependency word embeddings at the monolingual level leads to learn better bilingual word embeddings which improves the performance of word translation task in both language pairs: En-Es and En-De compared to the baseline model. However, these features show moderate improvement in the learning process of the BiMonoDep-BilBOWA model on En-De language pair and has shown almost no impact on En-Es language pair.

References

[1] S. Lauly, A. Boulanger, and H. Larochelle, "Learning multilingual word representations using a bag-of-words autoencoder," CoRR, 2014.

[2] K. M. Hermann and P. Blunsom," The role of syntax in vector space models of compositional semantics," In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 894–904, Sofia, Bulgaria. Association for Computational Linguistics, 2013.

[3] M. Xiao and Y. Guo, "Distributed word representation learning for cross-lingual dependency parsing", In CoNLL, pp. 119–129, 2014.

[4] S. Ruder." A survey of cross-lingual embedding models," CoRR, 2017.

[5] C. Li, J. Li, Y. Song and Z. Lin ,"Training and Evaluating Improved Dependency-Based Word Embeddings,"The Thirty-Second AAAI Conference on Artificial Intelligence ,2018,

[6] A. Komninos and S. Manandhar, "Dependency Based Embeddings for Sentence Classification Tasks," Proceedings of NAACL-HLT 2016, pp. 1490–1500,Association for Computational Linguistics, San Diego, California, 2016.

[7] O. Levy and Y. Goldberg, "Dependency-Based Word Embeddings," Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pp. 302–308, Baltimore, Maryland, USA, 2014.

[8] E. Kiperwasser and Y. Goldberg, "Simple and Accurate Dependency Parsing Using Bidirectional LSTM Feature Representations".Transactions of ACL4:313–327, 2016.

[9] T. Mikolov, M. Karafit, L. Burget, J. Cernock, and S. Khudanpur, " Recurrent neural network based language model," In Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTER-SPEECH, volume 2, pages 1045–1048, 2010.

[10] T. Mikolov and G. Zweig. "Context dependent recurrent neural network language model," IEEE Workshop on Spoken Language Technology, SLT 2012 - Proceedings,2012.

[11] T. Mikolov, Q.c V. Le, and I. Sutskever, "Exploiting similarities among languages for machine translation," CoRR, 2013a.

[12] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean," Distributed representations of words and phrases and their compositionality,". In Proceedings of the 26th International Conference on Neural Information Processing Systems, volume 2 of NIPS'13, pages 3111–3119, USA. Curran Associates Inc, 2013b.

[13] Y. Shi, W. Zhang, J. Liu, and M T. Johnson "RNN language model with word clustering and class-based output layer," EURASIP Journal on Audio, Speech, and Music Processing, 2013(1):22.

[14] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. 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), pages 1724–1734, Doha, Qatar. Association for Computational Linguistics, 2014.

[15] D. Bahdanau, K. Cho, and Y. Bengio, " Neural machine translation by jointly learning to align and translate," In 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.

[16] T. Luong, H. Pham, and C. D. Manning, " Effective approaches to attention-based neural machine translation," In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics, 2015b.

[17] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," CoRR, abs/1603.01360, 2016.

[18] C. dos Santos and M. Gatti, " Deep convolutional neural networks for sentiment analysis of short texts," In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 69–78, Dublin, Ireland. Dublin City University and Association for Computational Linguistics, 2014.

[19] Y. Kim. 2014, " Convolutional neural networks for sentence classification," In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) , page 17461751, Doha, Qatar,2014.

[20] A. Severyn and A. Moschitti ," Twitter sentiment analysis with deep convolutional neural networks," In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '15, pages 959–962, New York, NY, USA. ACM, 2015.

[21] T. Luong, H. Pham, and C. D. Manning, " Bilingual word representations with monolingual quality in mind," In Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing, pages 151–159, Denver, Colorado. Association for Computational Linguistics, 2015a.

[22] J. Zhang, S. Liu, M. Li, M. Zhou, and C. Zong, " Bilingually-constrained phrase embeddings for machine translation," In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 111–121, Baltimore, Maryland. Association for Computational Linguistics, 2014.

[23] J. Su, D. Xiong, Bi. Zhang, Y. Liu, J. Yao, and M. Zhang, " Bilingual correspondence recursive autoencoder for statistical machine translation," In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1248–1258, Lisbon, Portugal. Association for Computational Linguistics, 2015.

[24] B. Zhang, D. Xiong, and J. Su, " BattRAE: Bidimensional attention-based recursive autoencoders for learning bilingual phrase embeddings," CoRR, abs/1605.07874, 2016.

[25] S. Gouws, Y. Bengio, and G. Corrado, " Bilbowa: Fast bilingual distributed representations without word alignments," In Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37, ICML'15, pages 748–756. JMLR.org, 2015.

[26] J. Coulmance, J. Marty, G. Wenzek, and A. Benhalloum, " Transgram, fast cross-lingual word-embeddings," In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1109–1113, Lisbon, Portugal. Association for Computational Linguistics, 2015.

[27] A. Mogadala and A. Rettinger, " Bilingual word embeddings from parallel and non-parallel corpora for cross-language text classification," In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 692–702, San Diego, California. Association for Computational Linguistics, 2016.

[28] I. Vulic and M-F. Moens, " Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction," In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL2015), page 719725, 2015.

[29] L. Xu, W. Ouyang, X. Ren, Y. Wangand and L. Jiang, "Enhancing Semantic Representations of Bilingual Word Embeddings with Syntactic Dependencies," Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18),pp. 4518-4524, 2018.

[30] D. Q. Nguyen and K. Verspoor, " An improved neural network model for joint POS tagging and dependency parsing," Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from RAw Text to Universal Dependencies, pp. 81-91, Brussels, Belgium, October 31-November 1, 2018.

[31] P. Koehn, " Europal: A Parallel Corpus for Statistical Machine Translation," MT Summit, 2005.

[32] R. Socher, J. Bauer, C. D Manning and A. Y Ng, " Parsing with compositional vector grammers," In ACL (1), pp. 455-465, 2013.

[33] M. C. De Marneffe, T. Dozat, N. Silveira, K. Haverinen, F. Ginter, J. Nivre and C. D. Manning, "Universal Stanford dependencies: A cross-linguistic typology," In LREC, pp. 4585-4592, 2014.