Proceedings of the 13th Workshop on Building and Using Comparable Corpora, pages 56–60
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

TALN/LS2N Participation at the BUCC Shared Task: Bilingual Dictionary Induction from Comparable Corpora

Martin Laville, Amir Hazem and Emmanuel Morin
LS2N, UMR CNRS 6004, Université de Nantes, France
firstname.lastname@ls2n.fr

Abstract

This paper describes the TALN/LS2N system participation at the Building and Using Comparable Corpora (BUCC) shared task. We first introduce three strategies: (i) a word embedding approach based on fastText embeddings; (ii) a concatenation approach using both character Skip-gram and character CBOW models, and finally (iii) a cognates matching approach based on an exact match string similarity. Then, we present the applied strategy for the shared task which consists in the combination of the embeddings concatenation and the cognates matching approaches. The covered languages are French, English, German, Russian and Spanish. Overall, our system mixing embeddings concatenation and perfect cognates matching obtained the best results while compared to individual strategies, except for English-Russian and Russian-English language pairs for which the concatenation approach was preferred.

Keywords: Bilingual lexicon induction, Comparable corpora, Cognates, Word embeddings

1. Introduction

Cross-lingual word embeddings learning has triggered great attention in the recent years and several bilingual supervised (Mikolov et al., 2013; Xing et al., 2015; Artetxe et al., 2018a) and unsupervised (Artetxe et al., 2018b; Conneau et al., 2017) alignment methods have been proposed so far. Also, multilingual alignment approaches which consists in mapping several languages in one common space via a pivot language (Smith et al., 2017) or by training all language pairs simultaneously (Chen and Cardie, 2018; Wada et al., 2019; Taitelbaum et al., 2019b; Taitelbaum et al., 2019a; Alaux et al., 2018) are attracting a great attention.

Among possible downstream applications of cross-lingual embedding models: Bilingual Lexicon Induction (BLI) which consists in the identification of translation pairs based on a comparable corpus. The BUCC shared task offers the first evaluation framework on BLI from comparable corpora. It covers six languages (English, French, German, Russian, Spanish and Chinese) and two corpora (Wikipedia and WaCKy). We describe in this paper our participation at the BLI shared task. We start by evaluating the cross-lingual word embedding mapping approach (VecMap) (Artetxe et al., 2018a) using fastText embeddings. Then, we present an extension of VecMap approach that uses the concatenation of two mapped embedding models (Hazem and Morin, 2018). Finally, we present a cognates matching approach, merely an exact match string similarity.

Based on the obtained results of the studied approaches, we derive our proposed system –Mix (Conc + Dist)– which combines the outputs of the embeddings concatenation and the cognates matching approaches. Overall, the obtained results on the validation data sets are in favor of our system for all language pairs except for English-Russian and Russian-English pairs, where the cognates matching approach obviously showed very weak results and for which the concatenation approach was preferred.

In the following, Section 2 describes the shared task data sets, Section 3 presents the tested approaches and the chosen strategy. The results are given in Section 4, Section 5 discusses the quality of the seed lexicons, and finally, Section 6 concludes our work.

2. BLI Shared Task

The topic of the shared task is bilingual lexicon induction from comparable corpora. Its aim is to extract for each given source word, its target translations. The quality of the extracted lexicons is measured in terms of F1-score. To allow a deeper results analysis, the evaluation is conducted on three test sets corresponding to frequency ranges of the source language word: high (the frequency is among the 5000 most frequent words), mid (words ranking between 5001 and 20000) and low (words ranking between 20001 to 50000).

2.1. Tracks

The BLI shared task is composed of two tracks that is: (i) the closed task and (ii) the open task. In the closed task, only the data sets provided by the organizers can be used, while in the open track, external data as well as other language pairs evaluation are allowed. In this paper, only the closed track is addressed.

2.2. Data Sets

Two comparable corpora are provided: Wikipedia and WaCKy corpora (Baroni et al., 2009). Following the recommendations of the organizers, Table 1 illustrates the language pairs and their corresponding corpora that we address in the closed track.

Language	*de*	*es*	*fr*	*ru*
en	WaCKy	Wikipedia	WaCKy	WaCKy
de	-	-	WaCKy	-

Table 1: Corpus used for every language pair

Our training seed lexicons are from Conneau et al. (2017), for the validation results, we split these lists $80/20$.

3. Approach

In this section, we present the three tested strategies as well as the chosen system to address the BLI shared task.

3.1. Word Embeddings and Mapping

To extract bilingual lexicons from comparable corpora, a well-known word embedding approach that maps source words in a target space has been introduced (Mikolov et al., 2013) and several mapping improvements have been proposed (Xing et al., 2015; Artetxe et al., 2018a). The basic idea is to learn an efficient transfer matrix that preserves translation pairs proximity of a seed lexicon. After the mapping step, a similarity measure is used to rank the translation candidates.

To apply the mapping approach, several embedding models can be used such as Skip-gram and CBOW (Mikolov et al., 2013), Glove (Pennington et al., 2014), character Skip-gram (Bojanowski et al., 2016), etc. In our approach, we used fastText (Bojanowski et al., 2016) as our word embeddings representations. We trained character Skip-gram and CBOW models, using the same parameters as the given pre-trained embeddings for both methods: minCount: 30; dim: 300; ws (context window): 7; epochs: 10; neg (number of negatives sampled): 10. For the English-Spanish pair, our embeddings were trained on Wikipedia. For all the other language pairs, the embedding models were trained on their corresponding WaCKy corpora.

After training our embeddings, we used the VecMap tool from Artetxe et al. (2018a) to project by pairs every source embeddings space in its corresponding target space (i.e. Skip-gram English mapped with Skip-gram Spanish or CBOW French mapped with CBOW German). We used the supervised method and split the training seed lexicon $80/20$ for training and validation. For the submitted results, we took the whole seed lexicon as training for the mapping.

Once our embeddings were projected in the same space, we compared every source word of our reference lists to every target word of the vocabulary with a similarity measure. We used the CSLS (Conneau et al., 2017), which is based on the cosine similarity but reduces the similarity for word vectors in dense areas and increases it for isolated ones:

$$CSLS(x_s, y_t) = 2cos(x_s, y_t) - knn(x_s) - knn(y_t) \quad (1)$$

where x_s (y_t) is the vector from source (target) space and $knn(x_s)$ ($knn(y_t)$) is the mean of the cosine of its k-nearest neighbors in the target (source) space.

This similarity measure allows us to order the target words from the most to the less likely to be the translation, but as there is multiple words as valid translations, we can not just keep the first word of each ranking. We used two criteria to select the candidates from the embeddings approach: i) a maximal number of candidates that we want to keep for each source word and ii) a minimal CSLS value to validate the candidates. We present the different values that we used for every language pair in Table 2. These values were fixed empirically on the validation set.

Language pair	*Cand.* $\leq$	*Sim.* $\geq$
en-es	4	0.1
es-en	2	0.08
en-de	5	0.06
de-en	5	0.04
en-fr	3	0.08
fr-en	2	0.04
en-ru	4	0.05
ru-en	2	0.03
de-fr	2	0.08
fr-de	2	0.06

Table 2: Parameters for selection of candidates for every language pair

3.2. Embeddings Concatenation

In order to take advantage of several embedding models, Hazem and Morin (2018) proposed an extension of the mapping approach by applying the concatenation or addition of two embedding models before performing the mapping approach. In our case, and for each language, we applied the concatenation of character CBOW and character Skip-gram models for each word. Starting from the mapped 300 dimensional embeddings from the previous step, we obtained a concatenated embedding vector of 600 dimensions for each source and target words.

3.3. Perfect Cognates

A careful analysis of the training reference lists revealed that many translation pairs were graphically identical, especially for the low frequency lists. While some of these words are perfect cognates, a part of them are inconsistencies (i.e. the English to French translation pair *someone - someone*). We give more details of these problems in Section 5. To take this into consideration, we selected as valid candidates for every source word its perfect cognates if present in the target vocabulary. We added the constraint that each translation word pairs must have a distribution with a proportional factor of n. Given a source word w_s and its corresponding translation w_t, and given the frequency of w_s ($freq(w_s)$), respectively the frequency of w_t ($freq(w_t)$). The constraint is represented as:

$$\frac{1}{n} \leq \frac{freq(w_s)}{freq(w_t)} \leq n \quad (2)$$

where n was fixed empirically to 100.

3.4. Mixing the Candidates

To improve performance, combining several approaches is often performed. As will be shown in Table 3 of the results Section, the embeddings approach performs better on high frequency pairs while the perfect cognates method shows good results on lower range pairs. Hence, we naturally combined the extracted candidates of both strategies to provide one final mixed list, without taking into account the previous limit of the number of candidates. This mixing approach also noted –Mix (Conc + Dist)–, corresponds to our participating system to the BLI shared task. One exception however, concerns English and Russian languages for which we applied the concatenation approach only.

	en-es				*es-en*			
Frequency	high	mid	low	all	high	mid	low	all
Skip-gram	60.1	62.8	57.2	60.4	62.5	64.2	65.9	63.9
CBOW	57.1	56.8	54.1	56.4	59.7	60.2	56.2	59.0
Concatenation	60.9	**64.5**	62.8	62.4	62.6	65.5	65.3	64.3
Perfect Cognates	23.3	37.5	63.3	38.3	22.8	37.8	65.4	40.9
Mix (Conc + Dist)	**61.0**	61.8	**74.4**	**64.3**	**63.5**	**68.6**	**79.1**	**69.5**
	en-de				*de-en*			
Frequency	high	mid	low	all	high	mid	low	all
Skip-gram	47.6	43.6	29.8	43.4	50.6	47.6	33.7	45.8
CBOW	43.4	41.4	23.0	39.6	45.5	43.9	31.6	41.8
Concatenation	47.9	45.2	30.8	44.3	50.8	50.0	34.0	46.7
Perfect Cognates	21.1	35.6	67.8	37.2	24.1	35.7	69.9	41.2
Mix (Conc + Dist)	**50.9**	**55.0**	**71.8**	**56.4**	**57.2**	**62.3**	**72.9**	**63.1**
	en-fr				*fr-en*			
Frequency	high	mid	low	all	high	mid	low	all
Skip-gram	56.5	45.7	31.8	48.0	60.2	49.1	30.3	49.7
CBOW	51.4	42.0	31.1	44.1	58.5	48.7	29.4	48.4
Concatenation	57.8	45.8	34.6	49.3	62.8	55.4	36.2	54.0
Perfect Cognates	27.2	42.7	74.6	45.6	32.5	51.9	75.0	52.0
Mix (Conc + Dist)	**60.6**	**60.4**	**80.3**	**65.2**	**66.5**	**68.1**	**78.5**	**70.4**
	en-ru				*ru-en*			
Frequency	high	mid	low	all	high	mid	low	all
Skip-gram	41.3	31.7	13.2	34.0	53.8	40.6	20.7	41.9
CBOW	40.6	28.2	13.7	32.8	49.5	39.5	19.1	39.3
Concatenation	**42.6**	**32.6**	14.4	**35.3**	**55.5**	**44.3**	**22.8**	**44.4**
Perfect Cognates	7.4	6.6	13.2	8.6	0.0	0.0	0.0	0.0
Mix (Conc + Dist)	42.3	29.9	**21.0**	34.5	-	-	-	-
	de-fr				*fr-de*			
Frequency	high	mid	low	all	high	mid	low	all
Skip-gram	58.3	41.9	17.4	43.1	56.2	44.0	12.3	42.4
CBOW	52.7	32.7	14.4	36.6	51.2	39.9	11.7	38.5
Concatenation	60.2	44.2	17.9	44.6	56.8	46.9	14.9	44.2
Perfect Cognates	43.4	72.2	82.9	67.4	41.5	68.3	86.9	67.4
Mix (Conc + Dist)	**67.9**	**78.8**	**85.5**	**77.0**	**62.9**	**74.7**	**87.7**	**74.0**

Table 3: F1-score for our different approaches and language pairs

4. Results

Table 3 presents the obtained results (F1-score) of the individual strategies: (i) the mapping approach (Skip-gram and CBOW); (ii) the concatenation approach (Concatenation); (iii) the perfect cognates approach; and our proposed system (iv) Mix (Conc + Dist), on the validation sets for all language pairs.

We notice that mixing the candidates from the concatenated embeddings method and the perfect cognates extraction (Mix (Conc + Dist)) obtains the best results in almost every configuration, except one from English to Spanish and, obviously, the two pairs containing Russian, due to the different alphabets between English and Russian. Nevertheless, the English to Russian pair has a F1-score superior to zero, meaning that some Russian words are not written in Cyrillic, questioning the consistency of the lists.

The better results of the mixed method indicate a good complementarity of both approaches, which is confirmed by the trends regarding the frequency lists. We observe that the embeddings approach performs better on high frequency pairs and then degrades as the frequency decreases. Conversely, for the perfect cognates approach, the results are very high for the low frequency pairs and degrades for translation pairs of higher frequencies. The decline of results for perfect cognates is mostly due to the fact that high frequency words tend to have more translations than low ones (see Table 4) and the perfect cognates can at most predict one translation per source word.

The numbers illustrated in Table 4 corresponds to the validation lists, and not to the whole dictionaries.

As additional information, not shown in Table 3, it is to note that the perfect cognates method has a high precision for most language pairs, and it finds usually for more than half of the source words a perfect cognate in the target vocabulary. And thus, the results in F1-score are particularly high for the German-French pair in both directions as only few source words have more than one translation on the reference lists (1.03 target words per source words).

Finally, we note that the embeddings approach for the English-Spanish pair in both directions presents way better

Language pair	high	mid	low	all
en-es	2.34	1.58	1.10	1.67
en-de	2.83	1.81	1.14	1.93
fr-en	1.64	1.42	1.15	1.40
de-fr	1.08	1.02	1.00	1.03

Table 4: Ratio of target words per source words for the validation lists for some language pair on different lists

results than other language pairs (10 to almost 30 points). Unlike other pairs trained on WaCKy, this pair is the only one trained on Wikipedia, contradicting the idea that "the WaCKy corpora seem somewhat better suited for the dictionary induction task than Wikipedia". To verify this statement, we used pre-trained word embeddings from Grave et al. (2018) to check if the corpus was really the main problem. And actually, using the pre-trained embeddings on Wikipedia or Common Crawl led to much better results than the results obtained using the WaCKy corpora, reaching about the same F1-score as the English-Spanish language pair.

Our final results for the shared task were reported from the mixed approach for all language pairs but the two with Russian, for which we only took the results from the concatenation approach.

5. Seed Lexicon Analysis

As mentioned in the shared task, we report here the problems found in the seed lexicon.

We first noticed the presence of graphically identical pairs on the English-Russian pair, whereas the two languages have a different alphabet. This results are visible in Table 3 at the Perfect Cognates corresponding list. These instances are only present on the English to Russian language pair, suggesting a better control has been done for the source part of the lists.

A brief inspection of the lists makes us notice the presence of multiple words not belonging to the language of interest (i.e. on the French part of the English to French seed lexicon: *grammy, gov, god, northwest, phoenix* and many others) and we suggest the usage of monolingual dictionary to get rid of them. We even find pairs with none of the words belonging to one of the two languages (in the German to French seed lexicon the pair *times - times*, which should be *zeit - temps* if we translate it from English, or *ram - ram* instead of *ramm - bélier*).

We also observe many proper names and while some of them can be interesting to translate, most of them are graphically identical words (*jura, edward, lille...* on French to German or *calais, guanajuato...* on English to French), and we question the utility of translating such words, especially when some of them are not correctly presented (the German to French seed lexicon proposes a *mans - mans* pair, and we assume this is an incomplete form of the city "Le Mans" in France).

Focusing on the French part of some lists, we notice inconsistency with the use of diacritics (i.e. é, è...), the word *events* in English has four proposed translations in French, each being a variation of accents: *évènements, evénements, evenements,* and *événements*. While in French, both *é* or *è* are accepted for the second *e*, the first one should always be an *é*. The English word *development* being another example with *developpement* and *développement* while only the latter should be a correct translation.

Still on the French part, we notice that the inflectional morphology also suffers from incoherence. In the German to French pair, *allein* is only translated with its masculine (*seul*) and feminine (*seule*) and not its plural forms (*seuls* and *seules*), but *ausgebildet* translations are only *formés* and *formé*, forgetting the feminine forms. We add that in the English to French pair *christian* being translated to *chrétiens, chrétienne, chrétien* (and *christian*, which can only be a proper name in French) instead of *chrétiens* (and *chrétiennes* which is not even here) being the translation of *christians*.

Finally, some conjugation omissions are observed, for the English word *believe* for instance, the proposed translations are *croyez, croire, croient,* and *crois* but not *croyons* and we later have *believed* with only *croyait* as translation.

All these inconsistencies open important questions about the evaluation process and suggest a careful handcrafted validation which will undoubtedly strengthen the BLI shared task.

6. Conclusion

We presented in this paper the participation of the TALN/LS2N team at the BUCC shared task. We used concatenation of classic embeddings models (character Skip-gram and character CBOW) from fastText to get our first results. Graphical proximity of many translation pairs led us to strengthen our system based on a perfect cognates strategy. This latter tend to beat embedding methods on some language pairs. As both methods were effective in different frequency ranges, we combined them to pump up our results on all the language pairs except the two containing Russian. We add that the Wikipedia corpora seem to be more suited for our approach for bilingual lexicon induction than the WaCKy corpora, contradicting the initial claim of the organizers. Finally, we noted and reported multiple problems on the training seed lexicons, the most visible one being the presence of graphically identical pairs on the English-Russian pair, whereas the two languages have a different alphabet. Also, the presence of multiple words not belonging to the language of interest and many proper names, with many of them being graphically identical, making the utility of these pairs questionable. At last, some inconsistencies are present (at least for the French part of these lists) with the inflectional morphology, and with the verb conjugation.

Acknowledgments

We would like to thank the organizers for this exciting challenge. This research has received funding from the French National Research Agency under grant ANR-17-CE23-0001 ADDICTE (Distributional analysis in specialized domain) as well as the Canadian Institute for Data Valorisation (IVADO).

References

Alaux, J., Grave, E., Cuturi, M., and Joulin, A. (2018). Unsupervised hyperalignment for multilingual word embeddings. *CoRR*, abs/1811.01124.

Artetxe, M., Labaka, G., and Agirre, E. (2018a). Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5012–5019, New Orleans, LA, USA.

Artetxe, M., Labaka, G., and Agirre, E. (2018b). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL'18)*, pages 789–798, Melbourne, Australia.

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching word vectors with subword information. *CoRR*, abs/1607.04606.

Chen, X. and Cardie, C. (2018). Unsupervised multilingual word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 261–270, Brussels, Belgium.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jégou, H. (2017). Word translation without parallel data. *CoRR*, abs/1710.04087.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC'18)*, Miyazaki, Japan.

Hazem, A. and Morin, E. (2018). Leveraging meta-embeddings for bilingual lexicon extraction from specialized comparable corpora. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 937–949. Association for Computational Linguistics.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, et al., editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar.

Smith, S. L., Turban, D. H. P., Hamblin, S., and Hammerla, N. Y. (2017). Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *CoRR*, abs/1702.03859.

Taitelbaum, H., Chechik, G., and Goldberger, J. (2019a). A multi-pairwise extension of Procrustes analysis for multilingual word translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP'19)*, pages 3560–3565, Hong Kong, China.

Taitelbaum, H., Chechik, G., and Goldberger, J. (2019b). Multilingual word translation using auxiliary languages. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP'19)*, pages 1330–1335, Hong Kong, China.

Wada, T., Iwata, T., and Matsumoto, Y. (2019). Unsupervised multilingual word embedding with limited resources using neural language models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL'19)*, pages 3113–3124, Florence, Italy.

Xing, C., Wang, D., Liu, C., and Lin, Y. (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 (NAACL-HLT'15)*, pages 1006–1011, Denver, CO, USA.