Exploring cross-lingual word embeddings for the inference of bilingual dictionaries*

Marcos Garcia[1,3][0000−0002−6557−0210]
Marcos García-Salido[1,3][0000−0002−5667−2119]
Miguel A. Alonso[2,3][0000−0001−9254−4934]

[1] Universidade da Coruña, Grupo LyS, Departamento de Letras
Campus da Zapateira, 15071 A Coruña, Spain
marcos.garcia.gonzalez@udc.gal, http://www.grupolys.org/~marcos/
marcos.garcias@udc.gal
[2] Universidade da Coruña, Grupo LyS, Departamento de Computación
Campus de Elviña, 15071 A Coruña, Spain
miguel.alonso@udc.es, http://www.grupolys.org/~alonso/
[3] Universidade da Coruña, CITIC, Campus de Elviña, 15071 A Coruña, Spain

Abstract. We describe four systems to generate automatically bilingual dictionaries based on existing ones: three transitive systems differing only in the pivot language used, and a system based on a different approach which only needs monolingual corpora in both the source and target languages. All four methods make use of cross-lingual word embeddings trained on monolingual corpora, and then mapped into a shared vector space. Experimental results confirm that our strategy has a good coverage and recall, achieving a performance comparable to to the best submitted systems on the TIAD 2019 gold standard set among the teams participating at the TIAD shared task.

Keywords: Bilingual dictonaries · cross-lingual word embeddings · distributional semantics.

1 Introduction

Most research and development in natural language processing (NLP) and text mining was initially conducted for English. In the last decades, there was a surge in the application of NLP to other European languages as well as to the most-spoken languages worldwide (e.g., Chinese [19], Arabic [8]). Thus, in present day, the tendency is not to focus on monolingual texts, but to try to develop

* Research supported by a 2017 Leonardo Grant for Researchers and Cultural Creators (BBVA Foundation), by Ministerio de Economía, Industria y Competitividad (Grant TIN2017-85160-C2-1-R), and by Xunta de Galicia (Grant ED431B-2017/01). Marcos Garcia has been funded by a Juan de la Cierva grant (IJCI-2016-29598), and Marcos García-Salido by a post-doctoral grant from Xunta de Galicia (ED481D 2017/009). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

techniques and tools that can analyze texts written in different languages. In this context, the availability of bilingual lexicons is a necessity, not only for machine translation but for multilingual text mining tasks [17], such as sentiment analysis [18] or information extraction and summarization [12]. However, it is not possible to get a manually constructed bilingual lexicon for any arbitrary pair of languages. Hence the interest in defining models and techniques to construct bilingual language resources from those available for other pairs of languages.

For this purpose, most methods rely on comparable corpora [15], taking a measure of common words in their contexts as first clue to the similarity between a word and its translation. However, in this type of resources the position and the frequency of the source and target words are not comparable, and the translation of a word might not exist in a given pair of comparable documents [4]. Taking into account these limitations, several techniques have been tried, ranging from adaptations of traditional Information Retrieval metrics [5] to unsupervised multilingual word embeddings [9, 2]. In the case of rare words (terms that appear from 1 to 5 times in a collection of documents), it is even possible to use a vector representation of contexts and a classifier trained on one pair of languages to extract translations for another pair of languages with a reasonable performance [13], but this technique does not scale to the full range of words.

In [14], Rapp proposes a methodology for extracting bilingual lexicons from comparable corpora which is based on aligning comparable documents, using multiple pivot languages and considering word senses rather than words to solve ambiguities. Usually, taking into account word senses requires the existence of a previously compiled dictionary with all the possible senses for each word, in the style of WordNet, but since the construction of this kind of dictionaries is expensive even when reusing existing resources [3], they will not be available for most pairs of languages. An alternative consists in projecting words into a high-dimensional concept space: each word is converted into a vector of real numbers such that words having similar vectors are supposed to correspond to similar meanings. This, in brief, is the foundation of our proposal for using word embeddings on comparable corpora for the construction of bilingual dictionaries.

2 Description of the LyS systems

The objective of TIAD shared task is to generate automatically bilingual dictionaries based on existing ones. In particular, participants were required to generate new translations automatically between English (EN), French (FR) and Portuguese (PT) based on known translations contained in the Apertium RDF graph[4]. As these three languages are not directly connected in this graph, no translations can be obtained directly among them. The use of other freely available sources of background knowledge to improve performance is allowed, as long as no direct translation among the target language pairs is applied.

We presented four similar systems to the TIAD 2019 shared task: three transitive systems differing only in the pivot language used, and a system based on

[4] http://linguistic.linkeddata.es/apertium/

Table 1. Size of the source corpora (in number of tokens) and of the vocabularies (in number of *lemma_PoSTag* entries) for each language. English has an additional model (*EN10*) with a smaller vocabulary

Language	Corpus	Vocabulary
English	1,616,172,368	2,464,876
English EN10	1,616,172,368	1,497,155
Portuguese	299,788,709	482,728
French	817,787,536	1,097,192
Spanish	608,818,563	759,421
Catalan	219,973,695	409,002

a different approach which only needs monolingual corpora in both the source and target languages. Every system makes use of cross-lingual word embeddings trained on monolingual corpora, and then mapped into a shared vector space using VecMap [1]. Apart from the three source–target languages of the shared task (EN, PT, and FR), we also evaluated two additional pivot languages: Spanish (ES), and Catalan (CA).

2.1 Data processing

In order to obtain similar models, and also to cover a general vocabulary in each case, we decided to employ the different editions of Wikipedia as corpora for learning the word embeddings. For English, Portuguese, and Spanish, we used the Wikipedia dumps from January 2018, while for French and Catalan we have downloaded the January 2019 data.

Since we work with dictionary data, we have pre-processed each corpus to obtain morphosyntactic information as well as to reduce the vocabulary size of each model. Thus, we can obtain the PoS (Part of Speech) tag of each word and avoid the extraction of inflected forms that do not appear in dictionaries. This process was carried out using LinguaKit [7, 6] for Spanish, Portuguese, and English, and FreeLing [11] for Catalan and French. These processed data were then converted into *lemma_PoSTag* corpora to train distributional models with this lexical and morphosyntactic information.

After that, we used these modified corpora to train distributional semantics models using *word2vec* [10]. We took advantage of the *gensim* implementation [16] to train skip-gram models with 300 dimensions using windows of 5 tokens and a frequency threshold of 5. Table 1 contains the size of each Wikipedia (in number of tokens) as well as the number of *lemma_PoSTag* entries of the *word2vec* vocabularies. It is worth mentioning that the large size of the English vocabulary may reduce the computational efficiency of our approach when combined with other large models such as the French one, so we decided to train an additional English model (*EN10*) selecting only those *lemma_PoStag* elements with 10 or more occurrences in the corpus.

Once we built the monolingual models for each language, we mapped them into shared bilingual spaces using VecMap [1], a framework to learn cross-

lingual embeddings mappings. We applied the semi-supervised approach to each language pair, thus obtaining bilingual word-embeddings models. For EN/FR, FR/PT, and PT/EN, we only used 100 digits in *lemma_PoSTag* format (e.g., 0_NUM, 1_NUM, ..., 99_NUM) as seed dictionary, but we did not take advantage of any bilingual resource. For the other pairs, we used these 100 numbers as well as 300 randomly selected words (100 adjectives, 100 nouns, and 100 verbs) which were automatically translated and then reviewed by the authors.

2.2 Algorithm

With a view to exploring solely the performance of cross-lingual word embeddings in this task, it must be noted that our approach only uses the first and last columns of the input data (the source lemma and its PoS tag), so we do not utilize information such as the sense of each entry or its translation in other languages. Besides, the strategy was principally designed to obtain translations of single lexical words, so multiword expressions (specially compound proper nouns and non-compositional expressions such as idioms) are not well covered by the algorithm. These are, however, interesting cases for further research.

To translate between a source (*src*) and a target (*tgt*) language, our algorithm uses a pivot language (*pvt*) relying on the referred cross-lingual word embeddings and computing the similarity by means of the cosine distance between the two word vectors. We take the following steps:

1. For each single word (*lemma_PoSTag*, except proper nouns) in the input dictionary (e.g., EN-GL), we first check whether it appears in our source vocabulary with the same PoS tag. In this case, we select the two most similar entries with the same PoS tag in the pivot vocabulary. Then, for each of these words we get the two closest entries in the target model (also with the same PoS tag). If any of the words appears in the models with the same morphosyntactic category, the most similar entry (adjective, noun, verb, or adverb) is selected. At the end of this process, we have 0 to 4 candidate translations for each input word, together with a confidence value (the cosine distance between the words in the *pvt* and *tgt* models). If no translation is found, the algorithm applies a default rule which uses the input lemma as the translation with a 0 confidence value.
2. For single proper nouns (composed by just one word), we first check whether they appear in the pivot language (both in upper-case and lower-case) with a similarity greater than 0.5. In this case, we use the same procedure between *pvt* and *tgt*, selecting the target entry with a confidence value of 1. If the input proper noun does not appear in the pivot or target languages, we simply select the most similar proper noun (from the closest 50 words). If no proper noun is found, we apply the default rule.
3. To translate multiword expressions (MWEs) we apply a basic approach which uses a list of MWEs in the target language extracted from the input dictionaries (*mwe-list*). Thus, for each input MWE, we select its two longest words (*src_1* and *src_2*), and applying a similar strategy to that of

single words, we select two candidate translations for each one: *w1* and *w1b* for *src_1*, and *w2* and *w2b* for *src_2*. Then, using the *mwe-list* in the target language, we select the first MWE which contain both *w1* and *w2*. Otherwise, we look for expressions containing other combinations of the candidate translations. If no translation is found, we also apply the default rule.

The output of this process contains, for each input entry, 1 to 4 candidate translations with their confidence value. In a post-processing step we select the first candidate as the target translation for *lemma_PoSTag* pairs with only one sense in each input dictionary. However, as our method entirely relies on the input word and PoS tag (and not on its translation or its semantic information), it produces the same output for homonyms and the different senses of polysemous words. To alleviate this issue, the post-processing step respectively selects the third, second, and fourth candidates (if any) as the translations of other entries of the same word form.

Using different pivot languages, we applied this algorithm in four runs:

- **LyS**: this run uses, for each translation direction, the third language of the shared task as pivot. Thus, for EN→FR, Portuguese was the pivot language, while for FR→PT and PT→EN, English and French were the pivot languages, respectively.
- **LyS-CA**: in this system, Catalan was used as the pivot language for each translation.
- **LyS-ES**: this approach is identical to the previous one, but with Spanish (instead of Catalan) as pivot.
- **LyS-DT**: this last strategy does not use a pivot language, but only the source and target cross-lingual word embeddings models. As mentioned, and to follow the guidelines of the shared task, the mapping of these models was carried out without any bilingual information.

Except for the EN↔FR translations (where we used the *EN10* model), the full English model was used in all the other directions (see Table 1).

3 Results

Evaluation of the results was carried out by the organisers against a gold-standard set built by extracting translations from manually compiled pairs of K Dictionaries[5] (KD). As the coverage of KD is not the same as Apertium, the subset covered by Apertium was taken to build the gold standard, i.e., those KD translations for which the source and target terms are present in both Apertium RDF source and target lexicons. The gold standard set was not available to participants, so that we could not carry out a systematic error analysis which may be useful for further research.

Due to a misunderstanding, our team, like other participants, understood that evaluation would be performed in the direction EN→FR, FR→PT and

[5] https://www.lexicala.com/resources#dictionaries

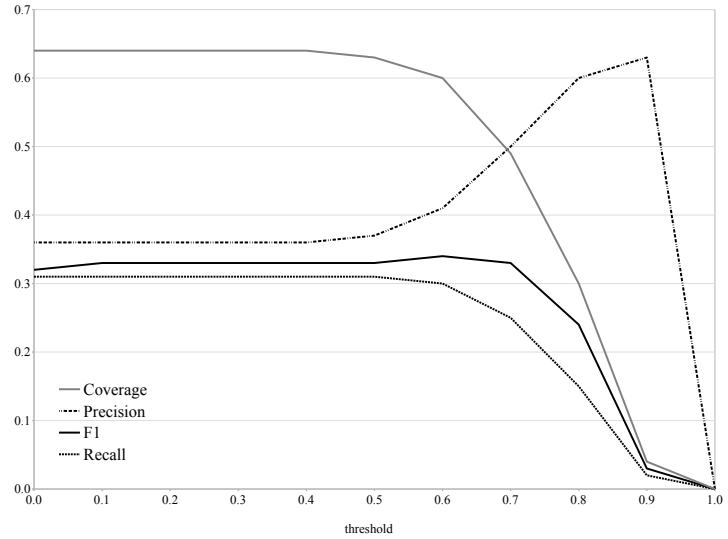

Fig. 1. Performance variation according to the chosen threshold for LyS-DT.

PT→EN. As a result, translations in the opposite direction were not sent for evaluation.

When compared to the other submissions of the shared task, the first overview of the general results confirm that our approach has a good coverage and recall (LyS-DT achieved the best numbers in this respect), even if the precision is lower than that of the other participants. In this regard, it must be mentioned that we decided to submit every output of the systems (even those with 0 confidence), so that a better analysis of the results should take into account different confidence thresholds. Interestingly, no submission could beat the baselines proposed by the organisers. The results of our four runs had the same tendencies in each of the three translations, being LyS-DT the best system followed by LyS-ES, LyS-CA, and LyS, respectively. Thus, we can infer that using a pivot language worsens the translation performance. The differences among each system, however, were not considerable, and there is no evidence that language family or corpus size have had a decisive impact, so that only a careful error analysis could shed some light into these results.

With regard to each specific translation pair, LyS-DT was the best submitted system in the EN→FR direction (0.34 F1), and the second one in FR→PT (0.36 F1) and in PT→EN (0.27 F1), tied and below the Frankfurt team, respectively. In general, the results of our other systems were slightly below these values.

We carried out a brief analysis of the LyS-DT submission, aimed at knowing a little better the output of the system. First, we observed the impact of the confidence threshold in the final results. Then, we analyzed the number of

translations of our system, paying special attention to those entries for which our system were not designed: multiword expressions and proper nouns.

Figure 1 shows a graph for LyS-DT submission with the evolution of the four metrics used to measure the performance according to the threshold of the degree of confidence for translations. As can be seen, the results remain stable up to a value of 0.5. It is worth noting that the F1 corresponding to a threshold 0 is slightly lower than the F1 between 0.1 and 0.4, so we could have climbed to the first position in the final ranking of participants teams by simply considering as official the results corresponding to a threshold 0.1. From a threshold of 0.5 on there is a rise in accuracy at the cost of a drop in coverage and recall, with F1 also dropping since the increase in accuracy is not enough to compensate for the drop in recall. Maximum precision is obtained between thresholds 0.8 and 0.9, but, while for 0.8 the recall has only dropped by half the value for 0.5, from that point on it descends with a steeper slope.

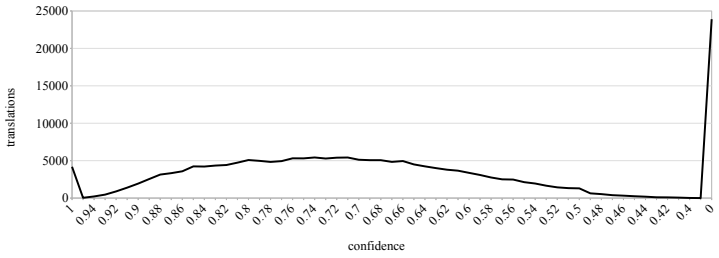

Fig. 2. Total number of extractions versus confidence value of the LyS-DT system.

Figure 2 shows the number of translations produced by LyS-DT with each confidence value. As this figure reveals, most translations have confidence values between ≈0.6-0.9. Besides, it is striking that from a total of 191, 373 entries there were 4, 184 with confidence of 1, and 23, 902 with 0 (which basically means that no translation was found). Among the first group, it is worth noting that all translations with confidence 1 were proper nouns, because they were found in both pivot and target languages with a high similarity value. However, it is likely that many of these equivalents are incorrect (see Figure 1), since our Wikipedia-based monolingual models contain proper nouns in different languages.

With regard to the entries with 0 confidence, Table 2 contains the number of non translated elements discriminated by single or multiword expressions as well as by PoS tags. For single words, most cases were proper nouns, followed by common nouns, adverbs, adjectives and verbs. In this respect, it was expected that our approach could not find suitable translations for many proper nouns. Also, several adverbs, adjectives, and nouns could also be wrongly lemmatized and subsequently not found in our models. However, a brief look at these data also revealed other issues: on one hand, the input dictionaries contain wrong *lemma_PoStag* pairs (e.g., actual proper nouns such as BBC, ETA or AT&T

Table 2. Number of entries without translation (confidence equal to 0) of the LyS-DT submission.

PoS tag	**Single words**	**Multiwords**
Proper noun	3,770	3,292
Common noun	3,357	3,063
Adverb	2,353	4,408
Adjective	1,439	216
Verb	170	1,834
Total	11,089	12,813

and several adjectives wrongly labeled as nouns). On the other hand, in some dictionaries there were typos and words in other languages (e.g., the Portuguese inputs include Spanish words like *garantía*, *asesor*, *cazador*, or *ateo*, instead of their Portuguese equivalents *garantia*, *assessor*, *caçador*, and *ateu*). Therefore, they were not present in our vocabularies. Apart from these issues, several other mistranslations of LyS-DT were technical words from specific domains (e.g., Medicine and Biology), which perhaps have low frequency in the Wikipedia corpus.

Finally, out of 13,157 MWEs of our runs, LyS-DT could only translate 344, proving that our basic MWE approach was not enough to obtain suitable translations. A simple pre-processing of proper nouns (joint as single tokens) could improve both their precision and recall. For the other cases, specially for non-compositional expressions, more complex strategies need to be designed.

4 Conclusions and future work

We consider that our participation in the TIAD shared task has been fruitful, as one of our systems has obtained the first position among the participant teams (tied with Frankfurt), while the other three results we sent were placed from third to fifth position. However, our systems failed to beat the "baselines" established by the organisers. In this regard, we must clarify that these are not really baselines but quite sophisticated state-of-the-art methods (one based on multilingual word embeddings and the other one based on the degree of overlap with the translations obtained by means of a pivot language).

Besides, it is worth mentioning that our approach only makes use of the lemma and PoS tag of each entry, so different senses of a word are simply inferred by their distributional contexts. In this respect, it could be interesting to make use of contextual information for each input word, in order to automatically select a specific sense. Also, the use of contextualized distributional models could be an interesting topic for further research.

Finally, in future work we plan to improve our method by dealing with constructions that in the current version are not processed properly. In particular, we intend to design different strategies, both compositional and non-compositional, for the processing of compound proper nouns as well as for obtaining better vector representations of multiword expressions.

References

1. Artetxe, M., Labaka, G., Agirre, E.: 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 (Volume 1: Long Papers). pp. 789–798. Association for Computational Linguistics, Melbourne, Australia (Jul 2018), https://www.aclweb.org/anthology/P18-1073
2. Chen, X., Cardie, C.: Unsupervised multilingual word embeddings. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 261–270. Association for Computational Linguistics, Brussels, Belgium (2018), https://www.aclweb.org/anthology/D18-1024
3. Fiser, D., Sagot, B.: Constructing a poor man's wordnet in a resource-rich world. Language Resources and Evaluation **49**(3), 601–635 (Sep 2015)
4. Fung, P.: A statistical view on bilingual lexicon extraction: From parallel corpora to non-parallel corpora. In: Proceedings of the Third Conference of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup. pp. 1–17. AMTA '98, Springer-Verlag, London, UK (1998), http://dl.acm.org/citation.cfm?id=648179.749226
5. Fung, P., Yee, L.Y.: An IR approach for translating new words from nonparallel, comparable texts. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1. pp. 414–420. Association for Computational Linguistics, Montreal, Quebec, Canada (Aug 1998). https://doi.org/10.3115/980845.980916
6. Gamallo, P., Garcia, M., Pineiro, C., Martinez-Castaño, R., Pichel, J.C.: LinguaKit: a Big Data-based multilingual tool for linguistic analysis and information extraction. In: 2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS). pp. 239–244. IEEE (2018)
7. Garcia, M., Gamallo, P.: Yet Another Suite of Multilingual NLP Tools. In: José-Luis Sierra-Rodríguez and José Paulo Leal and Alberto Simões (ed.) Languages, Applications and Technologies. Communications in Computer and Information Science. pp. 65–75. International Symposium on Languages, Applications and Technologies (SLATE 2015), Springer (2015)
8. Habash, N.: Introduction to Arabic Natural Language Processing. Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers (2010), http://dx.doi.org/10.2200/S00277ED1V01Y201008HLT010
9. Madhyastha, P.S., España-Bonet, C.: Learning bilingual projections of embeddings for vocabulary expansion in machine translation. In: Proceedings of the 2nd Workshop on Representation Learning for NLP. pp. 139–145. Association for Computational Linguistics, Vancouver, Canada (Aug 2017). https://doi.org/10.18653/v1/W17-2617
10. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: Workshop Proceedings of the International Conference on Learning Representations (ICLR) 2013 (2013), arXiv preprint arXiv:1301.3781
11. Padró, L., Stanilovsky, E.: Freeling 3.0: Towards wider multilinguality. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012). pp. 2473–2479. European Language Resources Association (ELRA), Istanbul, Turkey (May 2012), http://www.lrec-conf.org/proceedings/lrec2012/pdf/430_Paper.pdf

12. Poibeau, T., Saggion, H., Piskorski, J., Yangarber, R.: Multi-source, Multilingual Information Extraction and Summarization. Springer Publishing Company, Incorporated (2012)
13. Prochasson, E., Fung, P.: Rare word translation extraction from aligned comparable documents. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. pp. 1327–1335. Association for Computational Linguistics, Portland, Oregon, USA (Jun 2011), https://www.aclweb.org/anthology/P11-1133
14. Rapp, R.: A methodology for bilingual lexicon extraction from comparable corpora. In: Proceedings of the Fourth Workshop on Hybrid Approaches to Translation (HyTra). pp. 46–55. Association for Computational Linguistics, Beijing (Jul 2015). https://doi.org/10.18653/v1/W15-4108
15. Rapp, R., Xu, V., Zock, M., Sharoff, S., Forsyth, R., Babych, B., Chu, C., Nakazawa, T., Kurohashi, S.: New areas of application of comparable corpora. In: Skadiņa, I., Gaizauskas, R., Babych, B., Ljubesic, N., Tufis, D., Vasiljevs, A. (eds.) Using Comparable Corpora for Under-Resourced Areas of Machine Translation, pp. 255–290. Springer, Cham (2019)
16. Řehůřek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. pp. 45–50. European Language Resources Association (ELRA), Valletta, Malta (May 2010), http://is.muni.cz/publication/884893/en
17. Steinberger, R.: A survey of methods to easy the development of highly multilingual text mining applications. Language Resources and Evaluation **46**, 155–176 (2012)
18. Vilares, D., Gómez-Rodríguez, C., Alonso, M.A.: Universal, unsupervised (rule-based), uncovered sentiment analysis. Knowledge-Based Systems **118**, 45–55 (2017)
19. Wong, K.F., Li, W., Xu, R., sheng Zhang, Z.: Introduction to Chinese Natural Language Processing, Synthesis Lectures on Human Language Technologies, vol. 4. Morgan & Claypool Publishers, San Rafael, CA (2010)