Proceedings of the Globalex Workshop on Linked Lexicography, pages 106–110
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Graph Exploration and Cross-lingual Word Embeddings for Translation Inference Across Dictionaries

Marta Lanau-Coronas, Jorge Gracia
Aragon Institute of Engineering Research (I3A), University of Zaragoza, Spain
{mlanau, jogracia}@unizar.es

Abstract

This paper describes the participation of two different approaches in the 3rd Translation Inference Across Dictionaries (TIAD 2020) shared task. The aim of the task is to automatically generate new bilingual dictionaries from existing ones. To that end, we essayed two different types of techniques: based on graph exploration on the one hand and, on the other hand, based on cross-lingual word embeddings. The task evaluation results show that graph exploration is very effective, accomplishing relatively high precision and recall values in comparison with the other participating systems, while cross-lingual embeddings reaches high precision but smaller recall.

Keywords: Translation inference, Graph exploration, Cross-lingual word embeddings

1. Introduction

The fact that the open-source Apertium[1] bilingual dictionaries (Forcada et al., 2011) have been converted into RDF and published on the Web following Linked Data principles (Gracia et al., 2018) allows for a large variety of exploration opportunities. Nowadays, the Apertium RDF Graph[2] contains information from 22 bilingual dictionaries. However, as can be seen in Figure 1, where languages are represented as nodes and the edges symbolise the translation sets between them, not all the languages are connected to each other. In this context, the objective of the Translation Inference Across Dictionaries (TIAD) shared task[3] is to automatically generate new bilingual dictionaries based on known translations contained in this graph.

In particular, in this TIAD edicion (TIAD 2020), the participating systems were asked to generate new translations automatically among three languages, English, French, Portuguese, based on known translations contained in the Apertium RDF graph. As these languages (EN, FR, PT) are not directly connected in such a graph (see Figure 1), no translations can be obtained directly among them in this graph. Based on the available RDF data, the participants were asked to apply their methodologies to derive translations, mediated by any other language in the graph, between the pairs EN/FR, FR/PT and PT/EN. The evaluation of the results was carried out by the organisers against manually compiled pairs of K Dictionaries[4].

We have proposed two different systems for participating in the task.

1. *Cycles-OTIC*. The first one is a hybrid technique based on graph exploration. It includes translations coming from a method that explores the density of cycles in the translations graph (Villegas et al., 2016), combined with the translations obtained by the One Time Inverse Consultation (OTIC) method, which generates translation pairs by means of an intermediate pivot language (Tanaka and Umemura, 1994).

2. *Cross-lingual embeddings*. The second proposed system has a different focus. It does not rely on the graph structure but on the distribution of embeddings across languages. To that end, we reuse the system proposed by Artetxe et al. (Artetxe et al., 2018) to build cross-lingual word embeddings trained with monolingual corpora and mapped afterwards through an intermediate language.

The remainder of this paper is organised as follows. In Section 2 we give an overview of the used techniques. Then, in Section 3 we comment the results obtained in the evaluation and, finally, in Section 4 we present some conclusions and future directions of our research.

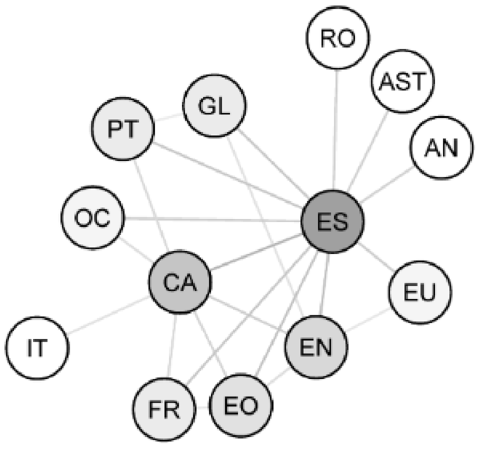

Figure 1: Apertium RDF Graph. It represents how the language pairs are connected by means of bilingual translation sets. The darker the colour, the more connections a node has. [Figure taken from `https://tiad2020.unizar.es/task.html`]

[1] `https://www.apertium.org/`

[2] `http://linguistic.linkeddata.es/apertium/`

[3] `https://tiad2020.unizar.es/`

[4] `https://lexicala.com/resources#dictionaries`

2. Systems overview

As it was stated previously, we developed two different techniques in order to automatically generate new bilingual dictionaries between the language pairs proposed in the task. Following the TIAD rules, the output data of the system was encoded in a TSV (tab separated values) file and had to contain the following information for all the translation pairs: *source and target written representation, part of speech and a confidence score*.

2.1. Cycles-OTIC system

Cycles-OTIC is a hybrid system that combines the translation pairs generated by means of the two graph-based methods described in the following paragraphs. The objective of this collaborative system is to reinforce both techniques and cover translations that can not be reached separately by any of the two methods.

Because of word polysemy, translation cannot be considered as a transitive relation. Specifically, when an intermediate language is used to generate a bilingual dictionary, the ambiguity of words in the pivot language may infer inappropriate equivalences. Avoiding those wrong translations is the main motivation of both methods.

2.1.1. Cycle-based method

The Cycle-based method was proposed by Villegas et al. (2016). The idea was using cycles to identify potential targets that may be a translation of a given word. A cycle can be considered a sequence of nodes that starts and ends in the same node, without repetitions of nodes nor edges. The confidence value of each translation is calculated by means of nodes' degree and graph density. The density is higher when higher is the number of edges in the graph, as can be seen in the Equation 1, where *E* represents the number of edges and *V* the number of vertices (nodes).

$$D = \frac{|E|}{|V| * (|V| - 1)} \tag{1}$$

The confidence score of a potential target is assigned by the density value of the more dense cycle where the source and target words appear. This value can achieve values from 0 to 1 (from completely disconnected to fully connected graph). Table 1 (Villegas et al., 2016) shows an illustrative example of some target candidates obtained in the Apertium RDF graph when translating the English word 'forest', along with the confidence score and the more dense cycle.

2.1.2. OTIC method

The second method utilised in our system was proposed by Tanaka and Umemura (1994) and adapted by Lim et al. (2011) afterwards for the creation of multilingual lexicons. This method is known as One Time Inverse Consultation (OTIC) and its objective is to construct bilingual dictionaries through intermediate translations by a pivot language. The OTIC method, even if relatively old, has proven to be a simple but effective one and a baseline very hard to beat, as it was shown by the previous TIAD edition results (Gracia et al., 2019) and corroborated with the latest TIAD 2020 results (see Table 6).

The OTIC method works as follows. In order to avoiding ambiguities caused by polysemy, for a given word, a confidence score is assigned to each candidate translation based on the degree of overlap between the pivot translations shared by both the source and target words. The higher is the overlap, the higher is the confidence score. The computation of this value is calculated by the Equation 2, where *T1* and *T2* are the number of translations into the pivot language from the source and target words respectively, and *I* the size of the intersection between those translations.

$$score = \frac{2 * I}{T1 + T2} \tag{2}$$

As it was mentioned before, the Apertium RDF Graph has been the source data of the experiments. In order to chose a suitable pivot language for the experiments, we explored the two possible ones: Spanish and Catalan. Table 2 shows a comparison of the size of the translation sets depending on using Spanish or Catalan as intermediate language. It can be observed that the Catalan language is quite unbalanced. For this reason, Spanish has been chosen as pivot language in our experiments[5].

2.1.3. Hybrid Cycles-OTIC method

Both methods have obtained good results in previous experiments (Villegas et al., 2016; Gracia et al., 2019). Our hypothesis is that the addition of the Cycles method should increase the coverage of the OTIC baseline, since there are possibly some translation pairs that cannot be linked through Spanish (our pivot language) but trough other languages in the graph. Additionally, we wanted to measure the benefits of adding the Cycles method in terms of precision and recall.

During development, some experiments with the Apertium RDF Graph were carried out to evaluate the performance of two possible ways of combining both methods: through the union and through the intersection of the translations results provided by both techniques. Some existing Apertium dictionaries were removed from the Apertium RDF graph and used as golden-standard during the development phase, where the explored method had to re-construct the removed Apertium dictionary. Results provided by those experiments showed that whereas the union of the translations sets from the Cycle-based and the OTIC method reached similar o even better results than the OTIC method alone, the results of the translations obtained from the intersection between both methods achieves much worse values of recall, as many correct translations reached by only one method were dismissed. Therefore we opted for the union operation when combining both systems. It was also observed that the hybrid system improved the results of the OTIC method when the pivot language has a small translation set with source and/or target languages.

Thus, the Cycles-OTIC method is simply the result of the union of the sets of translations generated by both methods individually. The translation pairs keep the confidence score obtained by the individual methods. However, when the same translation is provided by the two methods, the

[5] Spanish is also used as pivot language in the baseline evaluation carried out by the organisers, which uses also the same implementation: `https://gitlab.com/sid_unizar/otic`

bois-fr	0.9	[bosque-es, bosc-ca, bois-fr, arbaro-eo, forest-en]
fort-fr	0.9	[bosque-es, fort-fr, bosc-ca, arbaro-eo, forest-en]
bòsc-oc	0.833	[bosque-es, bòsc-oc, bosc-ca, forest-en]
bosque-pt	0.833	[bosque-gl, bosque-pt, bosque-es, forest-en]
floresta-pt	0.7	[fraga-gl, floresta-pt, bosque-gl, bosque-es, forest-en]
selva-es	0.619	[bosque-es, bosc-ca, arbaro-eo, fort-fr, selva-es, baso-eu, forest-en]

Table 1: 'Forest-en' best targets, its scores and cycles (Villegas et al., 2016).

	EN	FR	PT
ES	25,830	21,475	12,054
CA	33,029	6,550	7,111

Table 2: Size of the translation sets (in number of translations) for different intermediate languages (ES, CA).

score assigned is the maximum of the two values. The default threshold proposed for this combined method is 0.5.

2.2. CL-embeddings system

The second system developed makes use of cross-lingual word embeddings and a third intermediate language to generate new dictionaries. The vectors of the three languages (source, pivot and target) were all trained with monolingual corpora on Common Crawl and Wikipedia using fastTest (Grave et al., 2018). Then, they were mapped in pairs into a shared vector space through VecMap (Artetxe et al., 2018), a framework to learn cross-lingual word embedding mappings. The VecMap system allows for either a supervised or an unsupervised mode. In our case, it was supervised since we use the Apertium dictionaries as source of initial mappings between the source and intermediate monolingual embeddings, and also for the intermediate and target vectors. Given a word in the source language contained in the source vector, the algorithm gets the closest word vector in the embedding mapped. It is obtained by means of the cosine similarity metric, which can reach values from 0 to 1. The closer the vector, the higher the cosine metric. Afterwards, the same mechanism is done for getting the closest word in the target language from the one in the pivot language. Finally, the confidence score of the pair generated is computed by the product of both cosine similarity values calculated. The translation only is considered as candidate if the part of speech of source, pivot and target words are the same.

The language used as pivot between source and target were Spanish. In Table 3 can be seen the sizes of the extracts used for doing the initial mappings. These translation sets were obtained from the Apertium RDF Graph excluding those which contain spaces.

EN-ES	FR-ES	PT-ES
21610	18484	11634

Table 3: Size of the translation sets (in number of translations) used for mapping the monolingual vectors.

3. Results and Evaluation

The final evaluation of the results was carried out by the organisers against the test data[6]. These gold-standard consisted of the intersection between manually compiled pairs of K Dictionaries and the entries in Apertium dictionaries. The performance was measured in terms of precision, recall, F-measure and coverage. The official results of our systems with variable threshold are shown in Table 4 and Table 5. It can be seen that in both systems, when threshold gets higher values, precision increases while recall is reduced, as expected.

Threshold	Precision	Recall	F1	Coverage
0.0	0.64	0.47	0.54	0.76
0.1	0.64	0.47	0.54	0.76
0.2	0.64	0.47	0.54	0.76
0.3	0.64	0.47	0.54	0.76
0.4	0.64	0.47	0.54	0.76
0.5	0.65	0.47	0.54	0.75
0.6	0.67	0.45	0.54	0.73
0.7	0.73	0.38	0.49	0.63
0.8	0.74	0.36	0.48	0.60
0.9	0.77	0.31	0.44	0.53
1.0	0.77	0.31	0.44	0.53

Table 4: TIAD results for the Cycles-OTIC system with variable threshold

Threshold	Precision	Recall	F1	Coverage
0.0	0.58	0.33	0.41	0.81
0.1	0.58	0.33	0.41	0.81
0.2	0.58	0.33	0.41	0.81
0.3	0.59	0.33	0.41	0.81
0.4	0.59	0.33	0.42	0.79
0.5	0.62	0.32	0.42	0.73
0.6	0.68	0.29	0.40	0.60
0.7	0.75	0.20	0.31	0.38
0.8	0.79	0.07	0.13	0.12
0.9	0.40	0	0	0
1.0	0	0	0	0

Table 5: TIAD results for the CL-embeddings system with variable threshold

[6] Notice that one of the co-authors is co-organiser of TIAD. However, the test data was also treated as blind for the participating systems reported in this paper, to allow a fair comparison

System	Precision	Recall	F1	Coverage
Baseline-OTIC	0,70	0,47	0,56	0,70
Cycles-OTIC	**0,64**	**0,47**	**0,54**	**0,76**
NUIG	0,77	0,35	0,49	0,54
Multi-StrategyI+II+III+IV	0,61	0,33	0,43	0,63
Multi-StrategyI+II+III	0,62	0,33	0,43	0,63
CL-embeddings	**0,62**	**0,32**	**0,42**	**0,73**
Multi-StrategyI+II	0,65	0,30	0,40	0,59
ACOLIbaseline	0,60	0,28	0,38	0,48
Baseline-Word2Vec	0,30	0,37	0,33	0,68
Multi-StrategyI	0,63	0,22	0,32	0,44
ACOLIwordnet	0,61	0,16	0,25	0,28

Table 6: Averaged results per language pair for every system and ordered by F-measure in descending order.

A graph of the average of F-measure per threshold comparing all systems can be seen in Figure 2. The Cycles-OTIC system achieves the second position in terms of F-measure, although is beaten by the OTIC baseline. The other system, based in cross-lingual word embeddings gets the fifth position. As it is shown in Tables 4 and 5, both systems obtain high precision values, and the graph-based system obtains the highest coverage score among all the participating systems and baselines.

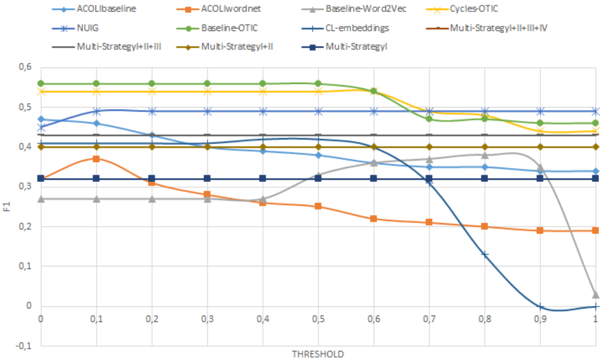

Figure 2: F1 results per different values of threshold for all systems

Discussion. The results prove our hypothesis that the addition of the Cycles method increases the coverage of the OTIC baseline. In particular from 0.70 to 0.76, being the largest value achieved in the shared task. The reason is that the Cycles method helps to discover, through alternative paths, some translation pairs that cannot be discovered through the pivot language. We see, however, that many of these extra translations are not present in the golden standard, since the value of precision drops from 0.70 to 0.64, while recall is preserved (0.47). We will perform a more careful inspection of the validation data results to better understand this effect. Out initial intuition is that the explored languages (PT, EN, FR) are already very well connected through the pivot language (SP), therefore OTIC can be very effective; while the Cycles strategy could play a more important role between other language pairs that are less directly connected in the graph.

As it can be seen in Table 6, the evaluation related to the CL-embeddings method shows that, in average, this technique has the second better value of coverage (0.73), just after the Cycles-OTIC method. The precision achieves also a high value (0.62), but regarding the recall, the value is not so high (0.32). One of the possible reasons behind this is that the embedding-based method only gives one target candidate per source entry (the one with best score). A further research considering different numbers of translations per word will be done in order to optimise recall while minimising the loss in precision.

4. Conclusions

In this paper we have described our participation in the TIAD 2020 shared task with two different techniques: one based on graph exploration and another one based on cross-lingual word embeddings. The official results provided by the organisers demonstrate that the performance of such methods for translation inference across dictionaries are good, specially in terms of precision and coverage. However none of the systems could beat the OTIC baseline in terms of F-measure, although the analysis of the results suggested us some improvements that will be carried out as future steps in this research line.

5. Acknowledgements

This work has been supported by the European Union's Horizon 2020 research and innovation programme through the projects Lynx (grant agreement No 780602) and Prêt-à-LLOD (grant agreement No 825182). It has been also partially supported by the Spanish projects TIN2016-78011-C4-3-R (AEI/FEDER, UE) and DGA/FEDER 2014-2020 "Construyendo Europa desde Aragón".

6. Bibliographical References

Artetxe, M., Labaka, G., and Agirre, E. (2018). 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)*, pages 789–798.

Forcada, M. L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Sánchez-Martínez, F., Ramírez-Sánchez, G., and Tyers, F. M. (2011). Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25(2):127–144.

Gracia, J., Villegas, M., Gomez-Perez, A., and Bel, N. (2018). The apertium bilingual dictionaries on the web of data. *Semantic Web*, 9(2):231–240.

Gracia, J., Kabashi, B., Kernerman, I., Lanau-Coronas, M., and Lonke, D. (2019). Results of the Translation Inference Across Dictionaries 2019 Shared Task. In Jorge Gracia, et al., editors, *Proc. of TIAD-2019 Shared Task – Translation Inference Across Dictionaries co-located with the 2nd Language, Data and Knowledge Conference (LDK 2019)*, pages 1–12. CEUR Press.

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 2018)*.

Lim, L. T., Ranaivo-Malançon, B., and Tang, E. K. (2011). Low cost construction of a multilingual lexicon from bilingual lists. *Polibits*, (43):45–51.

Tanaka, K. and Umemura, K. (1994). Construction of a bilingual dictionary intermediated by a third language. In *Proceedings of the 15th conference on Computational linguistics-Volume 1*, pages 297–303. Association for Computational Linguistics.

Villegas, M., Melero, M., Bel, N., and Gracia, J. (2016). Leveraging rdf graphs for crossing multiple bilingual dictionaries. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 868–876.