Proceedings of the Globalex Workshop on Linked Lexicography @LREC2022, pages 26–31
Marseille, 20 June 2022

Creating Bilingual Dictionaries from Existing ones by Means of Pivot-Oriented Translation Inference and Logistic Regression

Yves Bestgen
Laboratoire d'analyse statistique des textes
Université catholique de Louvain
Place Cardinal Mercier, 10 1348 Louvain-la-Neuve, Belgium
yves.bestgen@uclouvain.be

Abstract

To produce new bilingual dictionaries from existing ones, an important task in the field of translation, a system based on a very classical supervised learning technique, with no other knowledge than the available bilingual dictionaries, is proposed. It performed very well in the Translation Inference Across Dictionaries (TIAD) shared task on the combined 2021 and 2022 editions. An analysis of the pros and cons suggests a series of avenues to further improve its effectiveness.

Keywords: Apertium RDF graph, transitivity, supervised learning

1. Introduction

Despite recent advances in neural machine translation, bilingual dictionaries remain useful resources for both language learning and human post-editing of automatic translation as well as for language technologies (Goel et al., 2022). Unfortunately, such dictionaries are never completely up-to-date because languages evolve and speakers create new words. Moreover, some languages have far fewer bilingual dictionaries than others. Being able to automatically produce new bilingual dictionaries from existing ones has thus been an active area of research for the last 30 years (Tanaka-Ishii and Umemura, 1994; Mausam et al., 2009; Goel et al., 2022).

It is in this context that the Translation Inference Across Dictionaries (TIAD) shared task was created (Alper, 2017; Gracia et al., 2019; Gracia et al., 2021). In its 2021 and 2022 editions, it proposes to the participating teams to create automatically bilingual dictionaries between English, French and Portuguese, based on the many other bilingual dictionaries connected in the Apertium RDF graph (Gracia et al., 2018). This report presents the participation of SATLab to the fifth edition of this task proposed as part of the GLOBALEX 2022 workshop at LREC 2022.

This task has several characteristics that make it particularly complex. First of all, if the Apertium RDF graph is large since it contains 51 bilingual dictionaries[1] covering 42 languages, only three languages are present in many bilingual dictionaries (twelve for Spanish and ten for Catalan and English) whereas twenty languages are present in only one dictionary. Moreover, Apertium is largely focused on Spanish languages (Aragonese, Asturian, Basque, Catalan, Galician and Spanish) and even more on Catalan and Spanish since these languages are present in 21 dictionaries out of 51.

Secondly and most importantly, the evaluation of the effectiveness of the systems is not carried out on materials similar to that of the learning phase, but on the basis of manually compiled pairs of *K Dictionaries* (`https://lexicala.com/lexical-data/#dictionaries`) and other resources. The organizers provide a sample of this gold standard, but its use for optimizing systems is not easy as it is small (only 80 instances for one of the pairs of languages). On the other hand, it is far from clear that optimizing the system by means of a cross-validation procedure on the learning materials could be useful for the test materials. Consequently, the goal of the SATLab was to develop without optimization a system that potentially works and see what result it gets. If they are good, it will be interesting to look for an evaluation situation in which an optimization is easy to achieve.

To try to reach this goal, I chose to convert the problem into a supervised learning task, handled without external resources or complex learning procedures, an approach I have already used, sometimes successfully, to solve other NLP problems (Bestgen, 2021a; Bestgen, 2021b). The chance of success was not a priori zero since an approach of this type has already been recently used in this context (McCrae and Arcan, 2020; Ahmadi et al., 2021) and has produced interesting results, even if they were outperformed by more complex systems. The approach developed by these authors was graph-based and was a step towards more complex techniques such as Neural Machine Translation and cross-lingual word embedding mapping techniques. For my part, and even if the two approaches are similar, I started without preconceived ideas by considering the problem as a statistical data mining situation, based on computational procedures that rely on the sole notion of transitivity.

[1]The numbers given here refer to the CSV version of Apertium, provided by the task organizers, which was used in this study and contains two less dictionaries than the RDF version.

2. Approach

The objective is to arrive at potential translations of words for each of which a series of features will be used to decide whether these translations are assumed to be correct or not. These lists of translations or inferred dictionaries were obtained for all language pairs for which the gold standard is available in Apertium and for the six test language pairs. The approach developed is based on the following steps.

2.1. Data Processing

Reading the data. All CSV files were read and only the following three variables were kept: the word in the source language and in the target language and the grammatical category. These files were duplicated by reversing the source and target languages. All the dictionaries where one of the two languages is present in only one dictionary were deleted recursively.

Path search. For each dictionary, all paths, however long, from the source language to the target language through the bilingual dictionaries available in Apertium were identified. The only limiting condition applied is that the path cannot include the same language twice. As an example, 241 paths were found to go from FR[2] to PT and vice versa, and 146 to go from EN to PT. For a large number of language pairs, only 40 paths, including the direct path, were found (e.g., AN>ES, CA>ES or IT>SC). Some paths between the source and target languages pass through eight intermediate languages such as this one from EN to PT via EO>FR>OC>CA>SC>IT>ES>GL.

Producing bilingual dictionaries by inference. The paths identified in the previous step are used to produce bilingual dictionaries by inference, i.e. on the basis of at least one intermediate language, using transitivity. A more formal description of this approach under the name of *pivot-oriented translation inference* is given in Torregrosa et al. (2019). Starting from the source dictionary, the procedure is to use each intermediate dictionary as a pivot to the next one until the target dictionary is reached. At the end of the procedure, we obtain *for each path* an inferred bilingual dictionary which contains the source and target words, the grammatical category and the number of intermediate languages used. For each of these quadruplets, the following numerical data are computed in the dictionary in question:

- #Source: the number of occurrences of the source word.
- #Target: the number of occurrences of the target word.
- #Pair: the number of occurrences of the pair of words. It is indeed possible to reach the same pair by passing through different intermediate words.
- #SourceInPair: the number of different pairs containing the source word.
- #TargetInPair: the number of different pairs containing the target word.
- Source Ratio: #Pair divided by #Source.
- Target Ratio: #Pair divided by #Target.

Pooling of all bilingual dictionaries for a language pair. The average values of all the indices from the previous step is computed for each quadruplet. Two new indices are added: N, the frequency of each quadruplet, as well as the Total N, the frequency of the triplet composed of the source word, the target word, and the grammatical class. The first of these two values is therefore the number of paths of a given length that led to this triplet and the second is the total number of "paths" that led to this triplet. These two values are divided by the total number of paths that go from the source language to the target one. Finally, it is added whether the translation is correct or not according to the Apertium dictionary for this language pair. All these operations were also performed on the six test language pairs, but of course the gold standard, according to the Apertium dictionaries, is not added since it is unknown.

Preparing the data for supervised learning procedure. For each number of intermediate languages, ten features are encoded for each pair of translated words: the nine already described and the number of intermediate languages. So there can be from 10 to 80 features for each pair of words. To these, the size of the smallest path found that leads to this translation is added.

The values of each feature are then normalized by a MinMax transformation slightly modified compared to the classical formula:

$$MinMax = \frac{Feature_i\_score - min}{max - min} + 0.01 \quad (1)$$

The value of 0.01 is added to distinguish the minimum value of a feature with the value of 0, which codes the absence of a feature.

2.2. Supervised Learning Procedure

The supervised learning procedure used is the L1-regularized logistic regression as implemented in the LIBLinear package (Fan et al., 2008), The two parameters to optimize are the regularization parameter C and -wi which allows to adjust this parameter C for the two categories. After a few trials, the regularization parameter C was set to 80 and w1 to 2. The bias (B) was set to 1.

2.3. First Evaluation

In order to determine if the proposed approach had a chance to be sufficiently efficient, it was applied to the prediction of the EN to ES pair by means of a cross-validation procedure with 80% of the instances for training and the rest for testing. To also have an

[2]The languages are indicated by means of ISO 639-1 codes (see `https://en.wikipedia.org/wiki/List_of_ISO_639-1_codes`)

Type	Max Nbr	P	R	F1
L	5	0.8223	0.7139	0.7643
T	5	0.8184	0.7115	0.7612
L	4	0.8213	0.7157	0.7649
T	4	0.8176	0.7121	0.7612
L	3	0.8218	0.7114	0.7626
T	3	0.8166	0.7133	0.7614
L	2	0.8251	0.7094	0.7629
T	2	0.8205	0.7104	0.7615
L	1	0.8085	0.6808	0.7391
T	1	0.8105	0.6819	0.7406

Table 1: Results for the cross-validation analyses on EN to ES translations

Type	Max Nbr	P	R	F1
L	5	0.8225	0.7137	0.7643
T	5	0.8425	0.6224	0.7159
L	4	0.8213	0.7125	0.7631
T	4	0.8384	0.6223	0.7144
L	3	0.8209	0.7120	0.7626
T	3	0.8519	0.6118	0.7122
L	2	0.8242	0.7095	0.7626
T	2	0.8572	0.6210	0.7202
L	1	0.8086	0.6810	0.7393
T	1	0.8620	0.6120	0.7158

Table 2: Results by learning using EN to ES translations and predicting for FR to ES translations

evaluation situation that resembles the test situation in which it is not possible to learn and test on the same pair of languages, the system was also evaluated using a semi-external validation procedure, by learning on the EN to ES pair and testing on the FR to ES pair. The measures of effectiveness were precision (P), recall (R) and F1-score (F1) for the predicted translations according to whether they are present in the gold standard or not. In other words, the translations inferred but rejected by the logistic regression were not included in the calculation of the system's efficiency. In this evaluation, the maximum path size was manipulated. The results are presented in Tables 1 and 2.

These tables suggest that the performances are not too bad, but they do not seem exceptional either. They also indicate a total absence of overfit in CV and a relatively limited loss in semi-external validation. The impact of the number of intermediate languages is very small in both cross- and semi-external validation, except when the prediction is based on a single intermediate language. It was therefore decided to use this approach for the shared task by setting the number of intermediate languages at maximum three.

D1	D2	C	D1	D2	C	D1	D2	C
0	1	.18	4	0	.71	2	3	.89
1	0	.18	1	2	.81	3	2	.89
0	2	.36	2	1	.81	2	4	.90
1	1	.36	1	3	.82	4	2	.90
2	0	.36	3	1	.82	3	3	.93
0	3	.54	1	4	.83	3	4	.96
3	0	.54	4	1	.83	4	3	.96
0	4	.71	2	2	.83	4	4	1.0

Table 3: Computation of the Confidence score (C) according to the number of semi-external learning sets which lead to the translation for each direction (D1 and D2).

System	P	R	F1
PivotAlign-R	0.71	0.58	0.64
PivotAlign-F	0.81	0.51	0.62
SATLab	0.86	0.48	0.62
ACDcat	0.75	0.53	0.61
TUANWEsg	0.81	0.47	0.59
TUANWEcb	0.81	0.47	0.59
ULD_graphSVR	0.70	0.49	0.57
fastRP	0.85	0.28	0.42
PivotAlign-P	0.86	0.24	0.37
Baseline W2V	0.69	0.23	0.33

Table 4: Official results for 2021 et 2022 editions

2.4. System Submitted for the Shared Task

The system used for the official task has some specificities compared to the one described above. It should be noted that no further evaluation attempts were made since, as explained in the introduction, there is no guarantee that an Apertium-based optimization would be informative for the official test set.

First, several semi-external learning sets were arbitrarily selected for each target language:

- For FR>EN and PT>EN: CA>EN, ES>EN, EU>EN and EO>EN
- For EN>FR and PT>FR: CA>FR, ES>FR, OC>FR and EO>FR
- For EN>PT and FR>PT: CA>PT, ES>PT and GL>PT

Then the predictions in both directions for the same pair of test languages were combined to obtain the same inferred bilingual dictionary. Finally, the final decision for each pair was based on the number of models that predict this translation for each of the two directions as shown in Table 3. The threshold used for the official submission was set to 0.80.

3. Results

3.1. Results on the Official Test Set

The SATLab submitted only one system, as the official challenge website (https://tiad2022.unizar.es) did not indicate that more than one system could be submitted.

As the 2022 edition of the TIAD challenge is identical to the 2021 edition, the organizers have released, in addition to the results for 2022, the combined results for these two editions. Table 4 presents the results of the ten best submissions in this combined ranking. The official measure of the challenge is the F1-score.

The SATLab ranked third, close to the top two submissions of the first team in 2021 (Steingrimsson et al., 2021). Compared to the system also based on pivot-oriented translation inference and supervised learning (ULD_graphSVR in Ahmadi et al., 2021), the SATLab gets 5 more F1-points.

Since the systems submitted a confidence score for each proposed translation, the results reported in Table 4 were obtained by dichotomizing the scores, using the threshold value proposed by the teams. The organizers also provided an analysis of the performance of the systems when varying this threshold. As shown in Table 5, the SATLab scores best for the majority of thresholds, but the differences between the best systems are small and it is unlikely that an analysis using confidence intervals (Bestgen, 2022), unfortunately not possible here because the complete data are not available, would report important or statistically significant differences.

Finally, Table 6, provided by the organizers, presents the SATLab results separately for the six test language pairs. One can observe very strong variations according to the pair and the direction since the maximum difference between two F1-scores is 0.22. It would be really interesting to try to understand the origin of such differences, but it seems impossible without having access to the test set.

3.2. Additional Evaluation on the Learning Materials

As the test materials is not available, it is interesting to evaluate the proposed system in different (semi-) external learning configurations using Apertium. Table 7 presents the main results of these analyses.

The first section answers the question whether the same pair of languages used for learning (here, EN>ES) produces equivalent results for different test materials. The answer is very clearly negative, the difference between the test on AN>ES and on EO>ES being almost 0.40 of F1-score.

In the second section, the same semi-external evaluation procedure is used, but the language to be predicted is no longer ES. The results are overall worse than with ES, suggesting that this language is probably easier to predict.

In the last section, a completely external evaluation procedure is used since the four languages are different from each other. In two out of three cases, very poor performance is observed. These results suggest that it is desirable to learn by the semi-external procedure and therefore that one language should be the same for learning and testing.

4. Conclusion

The proposed system for the TIAD 2022 task scored well above my expectations, but it is important to note that many systems get very close scores. This system employs no knowledge other than the training set and is based on a very classical supervised learning technique. This submission has only scratched the surface of this interesting task. Indeed, there are still a number of options to try which are as many possibilities for future work. The main avenues seem to be the following:

- Optimizing features. It is very likely that some features are not very useful, but it is also far from obvious that all the features are computed in an optimal way. For example, it is questionable whether calculating the mean of #Pair is really preferable to taking the sum, since the values of this variable are almost always equal to 1.
- Reducing the number of paths. The results presented in Tables 2 and 3, as well as other analyses not reported here, suggest that limiting the number of intermediate languages to two might be beneficial.
- Evaluating other cases of semi-external validation. The proposed system relies on the presence of the same target language in the learning sets and in the test sets (e.g., ES>**EN** and FR>**EN**). It would be desirable to also evaluate models in which the source language is identical (e.g., **FR**>ES and **FR**>EN). It would also be interesting to see if using more than 7 or 8 models would improve the results.
- Finally, it would be interesting to compare the models of the logistic regression for different pairs of test languages to determine if the features are used in a similar way.

However, it is far from obvious that optimizing on the learning materials is relevant for the test materials, which is understandably not available. In this regard it would be interesting to find out if it is possible to put the task on a competition site by evaluating on one part of the data during development and on another part during the official test phase, possibly even on different test language pairs. I think this would make the task more attractive, but more importantly it would allow the development of better systems.

5. Acknowledgements

The author is a Research Associate of the Fonds de la Recherche Scientifique (FRS-FNRS).

Threshold	SATLab	PivotAlign-R	PivotAlign-F	ACDcat	TUANWEcb	ULD Gr SVR
0.0	**0.65**	0.64	0.62	0.61	0.59	0.60
0.1	**0.65**	0.64	0.62	0.61	0.59	0.59
0.2	0.63	**0.64**	0.62	0.61	0.59	0.59
0.3	0.62	**0.63**	0.62	0.61	0.59	0.58
0.4	**0.62**	0.61	0.60	0.61	0.59	0.57
0.5	**0.62**	0.58	0.58	0.61	0.59	0.57
0.6	**0.62**	0.54	0.53	**0.62**	0.60	0.57
0.7	**0.62**	0.49	0.47	**0.62**	0.60	0.55
0.8	**0.62**	0.41	0.38	**0.62**	0.59	0.54
0.9	0.47	0.30	0.25	**0.61**	0.57	0.51
1.0	0.31	0.12	0.07	**0.59**	0.14	0.25

Table 5: Results for the threshold analysis (best F1-scores are bolded)

Test	P	R	F1
EN>PT	0.85	0.41	0.55
EN>FR	0.82	0.40	0.54
FR>EN	0.83	0.47	0.60
FR>PT	0.89	0.53	0.67
PT>EN	0.86	0.42	0.57
PT>FR	0.90	0.66	0.76

Table 6: SATLab results for the six test language pairs

Learn	Test	P	R	F1
EN>ES	RO>ES	0.8818	0.7155	0.7900
EN>ES	AN>ES	0.9487	0.8332	0.8872
EN>ES	EO>ES	0.6346	0.4007	0.4912
SC>CA	EN>CA	0.7042	0.5427	0.6130
FR>CA	EN>CA	0.7823	0.5795	0.6658
ES>EU	EN>EU	0.8186	0.5341	0.6464
CA>SC	IT>SC	0.7087	0.4265	0.5325
EN>EU	FR>ES	0.5688	0.7549	0.6488
ES>CA	EN>EU	0.9564	0.0787	0.1455
ES>CA	IT>SC	0.8022	0.1934	0.3116

Table 7: Results of the post-hoc analyses on Apertium

6. Bibliographical References

Ahmadi, S., Ojha, A. K., Banerjee, S., and McCrae, J. P. (2021). NUIG at TIAD 2021: Cross-lingual word embeddings for translation inference. In *Proceedings of the Workshops and Tutorials held at LDK 2021*, CEUR Workshop Proceedings. CEUR-WS.org.

Alper, M. (2017). Auto-generating bilingual dictionaries: Results of the TIAD-2017 shared task baseline algorithm. In *Proceedings of the LDK 2017 Workshops*, CEUR Workshop Proceedings, pages 85–93. CEUR-WS.org.

Bestgen, Y. (2021a). LAST at CMCL 2021 shared task: Predicting gaze data during reading with a gradient boosting decision tree approach. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 90–96, Online, June. Association for Computational Linguistics.

Bestgen, Y. (2021b). A simple language-agnostic yet strong baseline system for hate speech and offensive content identification. In *Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation*, CEUR Workshop Proceedings. CEUR-WS.org.

Bestgen, Y. (2022). Please, don't forget the difference and the confidence interval when seeking for the state-of-the-art status. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, Marseille, France. European Language Resources Association (ELRA).

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

Goel, S., Gracia, J., and Forcada, M. L. (2022). Bilingual dictionary generation and enrichment via graph exploration. *Semantic Web – Interoperability, Usability, Applicability*.

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

Gracia, J., Kabashi, B., Kernerman, I., Lanau-Coronas, M., and Lonke, D. (2019). Results of the translation inference across dictionaries 2019 shared task. In *Proceedings of TIAD-2019 Shared Task – Translation Inference Across Dictionaries*, CEUR Workshop Proceedings, pages 1–12. CEUR-WS.org.

Gracia, J., Kabashi, B., and Kernerman, I. (2021). Results of the translation inference across dictionaries 2021 shared task. In *Proceedings of the Workshops and Tutorials held at LDK 2021*, CEUR Workshop Proceedings. CEUR-WS.org.

Mausam, Soderland, S., Etzioni, O., Weld, D., Skinner, M., and Bilmes, J. (2009). Compiling a massive, multilingual dictionary via probabilistic inference. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 262–270, Suntec, Singapore, August. Association for Computational Linguistics.

McCrae, J. P. and Arcan, M. (2020). NUIG at TIAD: Combining unsupervised NLP and graph metrics for translation inference. In *Proceedings of the 2020 Globalex Workshop on Linked Lexicography*, pages 92–97.

Steingrimsson, S., Loftsson, H., and Wa, A. (2021). PivotAlign: Leveraging high-precision word alignments for bilingual dictionary inference. In *Proceedings of the Workshops and Tutorials held at LDK 2021*, CEUR Workshop Proceedings. CEUR-WS.org.

Tanaka-Ishii, K. and Umemura, K. (1994). Construction of a bilingual dictionary intermediated by a third language. In *Proceedings of The 5th International Conference on Computational Linguistics (COLING'94)*, pages 297–303.

Torregrosa, D., Arcan, M., Ahmadi, S., and McCrae, J. P. (2019). TIAD 2019 shared task: Leveraging knowledge graphs with neural machine translation for automatic multilingual dictionary generation. In *Proceedings of TIAD-2019 Shared Task – Translation Inference Across Dictionaries*, CEUR Workshop Proceedings, pages 24–31. CEUR-WS.org.