Results of the Translation Inference Across Dictionaries 2019 Shared Task

Jorge Gracia[1], Besim Kabashi[2,3], Ilan Kernerman[4], Marta Lanau-Coronas[1], and Dorielle Lonke[4]

[1] Aragon Institute of Engineering Research (I3A), University of Zaragoza, Spain
{jogracia,mlanau}@unizar.es
[2] Ludwig-Maximilian University of Munich, Germany
[3] Friedrich-Alexander University of Erlangen-Nuremberg, Germany
besim.kabashi@fau.de
[4] K Dictionaries, Tel Aviv, Israel
{ilan,dorielle}@kdictionaries.com

Abstract. The objective of the Translation Inference Across Dictionaries (TIAD) shared task is to explore and compare methods and techniques that infer translations indirectly between language pairs, based on other bilingual/multilingual lexicographic resources. In its second, 2019, edition the participating systems were asked to generate new translations automatically among three languages - English, French, Portuguese - based on known indirect translations contained in the Apertium RDF graph. The evaluation of the results was carried out by the organisers against manually compiled language pairs of K Dictionaries. This paper gives an overall description of the shard task, the evaluation data and methodology, and the systems' results.

Keywords: TIAD · Apertium RDF · translation inference · lexicographic data

1 Introduction

A number of methods and techniques have been explored in the past aimed at automatically generating new bilingual and multilingual dictionaries based on existing ones. For instance, given a bilingual dictionary containing translations from one language L1 to another language L2, and another dictionary with translations from L2 to L3, a new set of translations from L1 to L3 is produced. The intermediate language (L2 in this example) is called pivot language, and it is possible to use multiple pivots for this purpose. When using intermediate languages, it is necessary to discriminate wrong inferred translations caused by translation ambiguities. The method proposed by Tanaka and Umemura [13] in 1994, called One Time Inverse Consultation (OTIC), identified incorrect translations when constructing bilingual dictionaries intermediated by a third language. This was a pioneering work in this field and it still constitutes a baseline that is hard to beat, as we will see in this paper. The OTIC method has been further adapted

and evolved in the literature, for instance by Lim et al. [6], who grounded on it for their method for multilingual lexicons creation. From a different perspective, other works were proposed that relied on cycles and graph exploration to validate indirectly inferred translations, such as the SenseUniformPaths algorithm by Mousam et al. [7], the CQC algorithm by Flati et al. [2] or the exploration based on cycle density by Villegas et al. [15].

However, previous work on the topic of automatic bilingual/multilingual dictionary generation was usually conducted on different types of datasets and evaluated in different ways, applying various algorithms that are often not comparable. In this context, the objective of the Translation Inference Across Dictionaries (TIAD) shared task is to support a coherent experiment framework that enables reliable validation of results and solid comparison of the processes used. This initiative also aims to enhance further research on the topic of inferring translations across languages. In this paper, we give an overall description of the shard task, the evaluation data and methodology, and the systems results of TIAD 2019.

The remainder of this paper is organised as follows. In Section 2, an overall description of the shared task is given. Section 3 describes the evaluation data and Section 4 explains the evaluation process. In Section 5 the systems results are reported, and conclusions are summarised in Section 6.

2 Shared task description

The objective of TIAD shared task was to explore and compare methods and techniques that infer translations indirectly between language pairs, based on other bilingual resources. Such techniques would help in auto-generating new bilingual and multilingual dictionaries based on existing ones.

In this second edition, the participating systems were asked to generate new translations automatically among three languages: English, French, and Portuguese, based on known translations contained in the Apertium RDF graph[5]. As these languages (EN, FR, PT) are not directly connected in this graph, no translations can be obtained directly among them there. Based on the available RDF data, the participants had 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.

Participants could also make use of other freely available sources of background knowledge (e.g. lexical linked open data and parallel corpora) to improve performance, as long as no direct translation among the studied language pairs were available. Beyond performance, participants were encouraged to consider also the following issues in particular:

1. The role of the language family with respect to the newly generated pairs
2. The asymmetry of pairs, and how translation direction affects the results
3. The behavior of different parts of speech among different languages

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

4. The role that the number of pivots plays in the process

The evaluation of the results was carried out by the organisers against manually compiled pairs of K Dictionaries, extracted from its Global Series[6], which were not accessible to the participants.

3 Evaluation data

In this section we briefly describe the input data source that has been proposed in the shared task as a source of known translations, i.e., Apertium RDF, as well as the data used as golden standard, from K Dictionaries.

3.1 Source data

As mentioned above, the shared task relies on known translations contained in Apertium RDF, which were used to infer new ones. Apertium RDF is the linked data counterpart of the Apertium dictionary data. Apertium [3] is a free open-source machine translation platform. The system was initially created by Universitat d'Alacant and it is released under the terms of the GNU General Public License. In its core, Apertium relies on a set of bilingual dictionaries, developed by a community of contributors, which covers more than 40 languages pairs. Apertium RDF [5] is the result of publishing 22 Apertium bilingual dictionaries as linked data on the Web. The result groups the data of the (originally disparate) Apertium bilingual dictionaries in the same graph, interconnected through the common lexical entries of the monolingual lexicons that they share.

In its fist version, Apertium RDF was modelled using the *lemon* model [8] jointly with its translation module [12]. Each Apertium bilingual dictionary was converted into three different objects in RDF: source lexicon, target lexicon, and translation set. As a result, two independent monolingual lexicons were published as linked data on the Web per dictionary, along with a set of translations that connects them. Notice that the naming rule used to build the identifiers (URIs) of the lexical entries allows to reuse the same URI per lexical entry across all the dictionaries, thus explicitly connecting them. For instance the same URI is used for the English word *bench* as a noun: `http://linguistic.linkeddata.es/id/apertium/lexiconEN/bench-n-en` throughout the Apertium RDF graph, no matter if it comes from, e.g., the EN-ES dictionary or the CA-EN. More details about the generation of Apertium RDF based on the Apertium data can be found at [5].

Figure 1 illustrates the Apertium RDF unified graph. The nodes in the figure are the languages and the edges are the translation sets between them. All the generated information is accessible on the Web both for humans (via a Web in-

[6] https://www.lexicala.com/

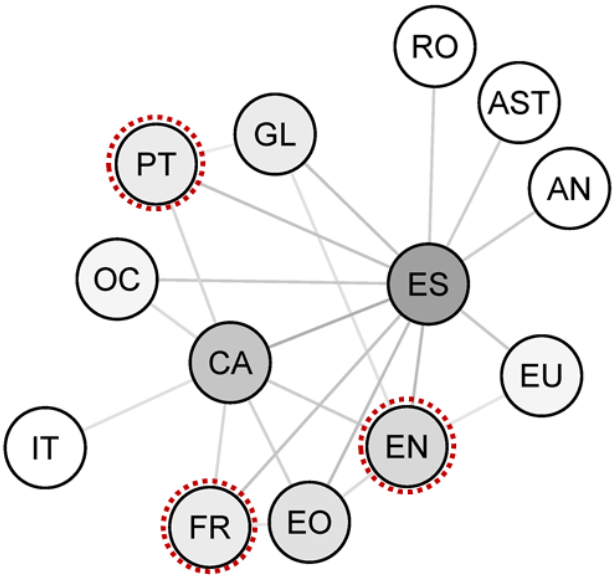

Fig. 1. The Apertium RDF graph. The nodes in the figure represent the monolingual lexicons and the edges are the translation sets between them. The darker the colour, the more connections a node has. We have highlighted the three languages of this evaluation campaign: PT, FR, and EN.

terface[7]) and software agents (with SPARQL[8]). All the datasets are documented in Datahub[9].

There were several ways in which the evaluation data was available to the participants: though the data dumps available in Datahub, through the SPARQL endpoint[10], and in a ZIP file in tab separated values (TSV) format[11]. More details on how to access the data are available in the TIAD 2019 website[12].

3.2 Gold standard

The evaluation of the results was carried out by the organisers against manually compiled language pairs of K Dictionaries, extracted from its Global series, particularly the following pairs: BR-EN, EN-BR, FR-EN, EN-FR, FR-PT, PT-FR. The translation pairs extracted from these dictionaries served as a golden

[7] `http://linguistic.linkeddata.es/apertium/`
[8] http://linguistic.linkeddata.es/apertium/sparql-editor/
[9] `https://datahub.ckan.io/dataset?q=apertium+rdf`
[10] See an example query at `https://tiad2019.unizar.es/docs/ApertiumRDF_ExampleQuery_10.txt`
[11] `https://tiad2019.unizar.es/data/TranslationSetsApertiumRDF.zip`
[12] See the "how to get the data source" section at `https://tiad2019.unizar.es/task.html`

standard and remained blind to the participants. Notice that the Brazilian Portuguese variant was used for the translations to/from English (whereas the European Portuguese variant was used with French), which might introduce a bias; however its influence should be equivalent to every participant system thus still allowing for a valid comparison.

Given the fact that the coverage of KD is not the same as Apertium, we took the subset of KD that is covered by Apertium to build the gold standard and allow comparisons, i.e., those KD translations for which the source and target terms are present in both Apertium RDF source and target lexicons. This is shown graphically in Figure 2 for the FR-PT pair.

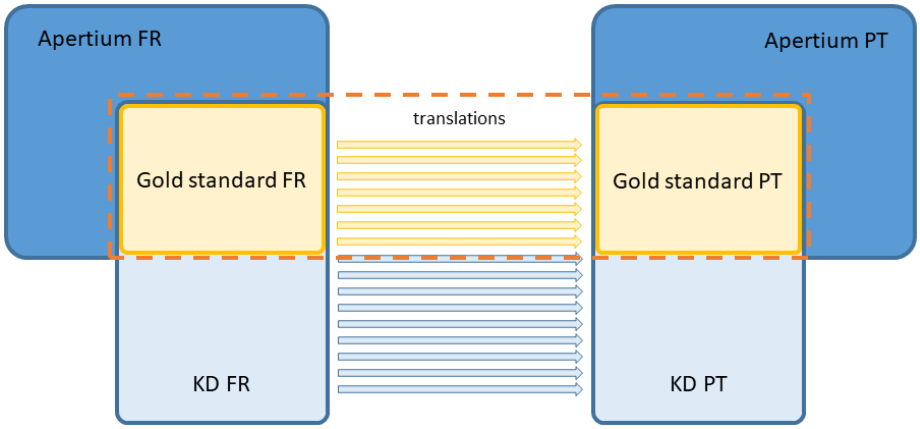

Fig. 2. Gold standard construction for the FR-PT pair. The translations in the dashed area in the middle of the figure constitute the gold standard, selected amongst all the KD translations (for FR-PT) for which both source and target lexical entries are present in their respective Apertium RDF lexicons.

Table 1 shows the size (in number of translations) of the different language pairs in the gold standard.

Table 1. Number of translations per language pair in the gold standard.

Lang. pair	Size
EN-FR	14,512
EN-PT	12,811
FR-EN	20,800
FR-PT	10,791
PT-EN	17,498
PT-FR	10,808

4 Evaluation methodology

The participants run their systems locally, using the Apertium RDF data as known translations, to infer new translations among the three studied languages: FR, EN, PT. Once the output data (inferred translations) were obtained, they loaded the results into a file per language pair in TSV format, containing the following information per row (tab separated):

"source written representation"
"target written representation"
"part of speech"
"confidence score"

The confidence score takes float values between 0 and 1 and is a measure of the confidence that the translation holds between the source and target written representations. If a system does not compute confidence scores, this value had to be put to 1.

4.1 Evaluation process

The organisers compared the obtained results with the gold standard automatically. This process was followed for each system results file and per language pair:

1. Remove duplicated translations (some systems produced duplicated rows, i.e., identical source and target words, POS and confidence degree).
2. Filter out translations for which the source entry is not present in the golden standard (otherwise we cannot assess whether the translation is correct or not). We call *systemGS* the subset of translations that passed this filter, and *GS* the whole set of gold standard translations, in the given language pair.
3. Translations with confidence degree under a given threshold were removed from *systemGS*. In principle, the used threshold is the one reported by participants as the optimal one during the training/preparation phase.
4. Compute the coverage of the system with respect to the gold standard, i.e., how many gold standard entries in the source language were effectively translated by the system (no matter if they were correct or wrong ones).
5. Compute precision as P =(#correct translations in systemGS) / |systemGS |
6. Compute recall as R =(#correct translations in systemGS) / |GS |
7. Compute F-measure as $F = 2 * P * R/(P + R)$

4.2 Baselines

We have run the above evaluation process with results obtained with two baselines, to be compared with the participating systems results:

Baseline 1 - Word2Vec. The method uses Word2Vec [11] to transform the graph into a vector space. A graph edge is interpreted as a sentence and the nodes are word forms with their POS tag. Word2Vec iterates multiple times over the graph and learns multilingual embeddings (without additional data). We used the Gensim[13] Word2Vec implementation. For a given input word, we calculated a distance based on the cosine similarity of a word to every other word with the target-POS tag in the target language. The square of the distance from source to target word is interpreted as the confidence degree. For the first word the minimum distance is 0.6^2, for the others it is 0.8^2. Therefore multiple results are only in the output if the confidence is not extremely weak. In our evaluation, we applied an arbitrary threshold of 0.5 to the confidence degree.

Baseline 2 - OTIC. In short, the idea of the One Time Inverse Consultation (OTIC) method [13] is to explore, for a given word, the possible candidate translations that can be obtained through intermediate translations in the pivot language. Then, a 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. In our evaluation, we have applied the OTIC method using Spanish as pivot language, and using an arbitrary threshold of 0.5.

5 Results

In this section we review the participating systems in TIAD 2019 and their evaluation results.

5.1 Participating systems

Four teams participated in the shared task. Unlike the first TIAD edition [10], all of them were able to complete the evaluation. The participants contributed with eleven system results. One team (Frankfurt) submitted the results of a single system, while the other three run the experiment on several systems or variations of the same system. Table 2 lists the participant teams and systems.

The first team, García et al. from Universidade da Coruña, developed four systems [4]: three transitive systems differing only in the pivot language used, and a fourth 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. The second team, Torregrosa et al. from National University of Ireland Galway, presented three methods [14] based on graph analysis and neural machine that did not make use of parallel data. The third contribution, by John P. McCrae, also from National University of Ireland Galway [9] applied explicit topic modelling over comparable corpora to the task

[13] `https://radimrehurek.com/gensim/`

Table 2. Participant systems.

Team	System	Comment
García et al. (Univ. da Coruña) [4]	LyS	Using the third language of the shared task as pivot (e.g., PT is pivot in an EN-FR translation)
	LyS_EN	English as pivot language
	LyS_CA	Catalan as pivot language
	LyS_DT	No pivot language
Torregrosa et al. (National University of Ireland Galway) [14]	UNLP-4CYCLE	Cycle based approach
	UNLP-GRAPH	Graph based approach
	UNLP-NMT-3PATH	Neural Machine Translation and Path based approach
	UNLP-NMT-4CYCLE	Neural Machine Translation and Cycle based approach
McCrae (National University of Ireland Galway) [9]	ONETA-ES	Spanish as pivot language
	ONETA-CA	Catalan as pivot language
Donandt and Chiarcos (Goethe Universität at Frankfurt) [1]	FRANKFURT	Multilingual word embeddings

of inferring translation candidates. In particular, he used the Orthonormal Explicit Topic Analysis (ONETA) model. Finally, the fourth team, Donandt and Chiarcos from Goethe-Universität at Frankfurt, constructed a multi-lingual word embedding space by projecting new languages in the feature space of a language for which a pre-trained embedding model exists [1]. They used the similarity of the word embeddings to predict candidate translations.

5.2 Evaluation results

The complete evaluation results per system and per language pair are accessible in the TIAD 2019 website[14]. In order to give an overview of the results, we include here Table 3, which shows the averaged results, evaluated by using the confidence threshold that every participant reported as optimal according to their internal tests. In addition, we evaluated the systems results with other thresholds in the range [0,1]. The results are plotted in Figure 3.

5.3 Discussion

As can be seen in Table 3, the two baselines obtained better results than the participating systems in terms of F-measure, which gives an idea of the difficulty of the task. Strictly speaking, these are not baselines as they are conceived in other shared tasks, meaning naive approaches with a straightforward implementation, but state-of-the-art methods to solve the task.

[14] See `https://tiad2019.unizar.es/results.html` under the section "Evaluation results".

Table 3. Averaged system results, ordered by F-measure in descending order.

System	Precision	Recall	F-measure	Coverage
BASELINE(OTIC)	0.64	0.26	0.37	0.45
BASELINE(Word2Vec)	0.66	0.24	0.35	0.51
FRANKFURT	0.64	0.22	0.32	0.43
LyS-DT	0.36	0.31	0.32	0.64
LyS-ES	0.33	0.3	0.31	0.64
LyS-CA	0.31	0.29	0.29	0.64
LyS	0.32	0.28	0.29	0.64
UNLP-NMT-3PATH	0.66	0.13	0.21	0.25
UNLP-GRAPH	0.76	0.1	0.18	0.2
UNLP-NMT-4CYCLE	0.58	0.11	0.18	0.25
ONETA-ES	0.81	0.1	0.17	0.17
ONETA-CA	0.83	0.08	0.14	0.13
UNLP-4CYCLE	0.75	0.07	0.11	0.13

Some of the participating systems kept a good balance between precision and recall (FRANKFURT, LyS-DT) while some promoted precision at the cost of recall (ONETA, UNLP), and others obtained very good recall and coverage at the cost of precision (LyS, LyS-ES, LyS-CA). Interestingly, the OTIC method, based on purely graph exploration and dated back to 1994, outperformed more contemporary methods based on word embeddings and distributional semantics. We argue, however, that OTIC is not upper bound and that there is still much room for improvement for such recent methods, that could benefit from a different selection of training data and dictionary-related features.

Notice that the precision values shown in Table 3 are conservative since there is a small but undefined number of false negatives (correct translations that are not present in the gold standard) that can be found in the results. Some examples, from the EN→FR set of translations:

"wizard"→"sorcier" noun 0.81 [BASELINE Word2Vec]
"abandon"→"quitter" verb 0.99 [FRANKFURT]
"dump"→"vider" verb 0.71 [LyS-CA]
"ban"→"prohibition" noun 0.31 [ONETA-CA]
"portion"→"ration" noun 0.4 [UNLP-GRAPH]

6 Conclusions

In this paper we have given an overview of the 2nd Translation Inference Across Dictionaries (TIAD) shared task, and a description of the results obtained by the 11 participating systems and two baselines. In this edition, the participating systems were asked to generate new translations automatically among English, French, Portuguese, based on known indirect translations contained in the Apertium RDF graph. The evaluation of the results was carried out by the organisers against manually compiled pairs of K Dictionaries.

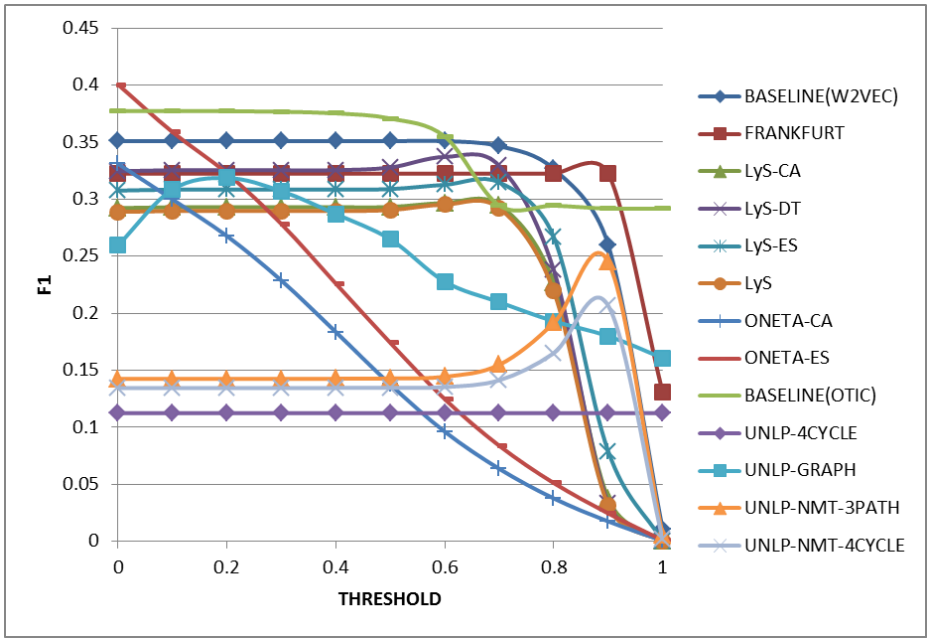

Fig. 3. Averaged system results (F-measure) with variable threshold.

The results are promising and illustrate the difficulty of the tasks, showing that there is still much room for research and improvement in the area of translation inference across dictionaries.

7 Acknowledgements

We would like to thank Michael Ruppert (University of Erlangen-Nuremberg) for his assistance with the Word2Vec baseline. This work has been supported by the European Union's Horizon 2020 research and innovation programme through the projects Lynx (grant agreement No 780602), Elexis (grant agreement No 731015) and Prêt-à-LLOD (grant agreement No 825182). It has been also partially supported by the Spanish National projects TIN2016-78011-C4-3-R (AEI/ FEDER, UE) and DGA/FEDER.

References

1. Donandt, K., Chiarcos, C.: Translation inference through multi-lingual word embedding similarity. In: Proc. of TIAD-2019 Shared Task Translation Inference Across Dictionaries, at 2nd Language Data and Knowledge (LDK) conference. CEUR-WS (May 2019)

2. Flati, T., Navigli, R.: The CQC Algorithm: Cycling in Graphs to Semantically Enrich and Enhance a Bilingual Dictionary (Extended Abstract). In: Proce. of the 23th International Joint Conference on Artificial Intelligence. pp. 3151–3155. IJCAI '13, AAAI Press (2013)
3. Forcada, M.L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J.A., Sánchez-Martínez, F., Ramírez-Sánchez, G., Tyers, F.: Apertium: a free/open-source platform for rule-based machine translation. Machine Translation **25**(2), 127–144 (2011)
4. García, M., García-Salido, M., Alonso, M.A.: Exploring cross-lingual word embeddings for the inference of bilingual dictionaries. In: Proc. of TIAD-2019 Shared Task Translation Inference Across Dictionaries, at 2nd Language Data and Knowledge (LDK) conference. CEUR-WS (May 2019)
5. Gracia, J., Villegas, M., Gómez-Pérez, A., Bel, N.: The apertium bilingual dictionaries on the web of data. Semantic Web **9**(2), 231–240 (2018)
6. Lim, L.T., Ranaivo-Malançon, B., Tang, E.K.: Low Cost Construction of a Multilingual Lexicon from Bilingual Lists. Polibits **43**, 45–51 (2011)
7. Mausam, Soderland, S., Etzioni, O., Weld, D.S., Skinner, M., Bilmes, J.: Compiling a Massive, Multilingual Dictionary via Probabilistic Inference. In: Proc. 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: Volume 1 - Volume 1. pp. 262–270. ACL '09, Association for Computational Linguistics, Stroudsburg, PA, USA (2009)
8. McCrae, J., Aguado-de Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D., Wunner, T.: Interchanging lexical resources on the Semantic Web. Language Resources and Evaluation **46**, 701–719 (2012)
9. McCrae, J.P.: Tiad shared task 2019: Orthonormal explicit topic analysis for translation inference across dictionaries. In: Proc. of TIAD-2019 Shared Task Translation Inference Across Dictionaries, at 2nd Language Data and Knowledge (LDK) conference. CEUR-WS (May 2019)
10. McCrae, J.P., Bond, F., Buitelaar, P., Cimiano, P., Declerck, T., Gracia, J., Kernerman, I., Montiel-Ponsoda, E., Ordan, N., Piasecki, M. (eds.): Proc.of the LDK 2017 Workshops: 1st Workshop on the OntoLex Model (OntoLex-2017), Shared Task on Translation Inference Across Dictionaries & Challenges for Wordnets co-located with 1st Conference on Language, Data and Knowledge (LDK 2017). CEUR Press, Galway (Ireland) (2017)
11. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space. In: Proc. of International Conference on Learning Representations (ICLR) (2013)
12. Montiel-Ponsoda, E., Gracia, J., Aguado-De-Cea, G., Gómez-Pérez, A.: Representing translations on the semantic Web. In: Proc. of the 2nd International Workshop on the Multilingual Semantic Web (MSW) at ISWC '11. vol. 775. CEUR Press (2011)
13. Tanaka, K., Umemura, K.: Construction of a Bilingual Dictionary Intermediated by a Third Language. In: COLING. pp. 297–303 (1994)
14. Torregrosa, D., Arcan, M., Ahmadi, S., McCrae, J.P.: Tiad 2019 shared task: Leveraging knowledge graphs with neural machine translation for automatic multilingual dictionary generation. In: Proc. of TIAD-2019 Shared Task Translation Inference Across Dictionaries, at 2nd Language Data and Knowledge (LDK) conference. CEUR-WS (May 2019)

15. Villegas, M., Melero, M., Bel, N., Gracia, J., Bel, N.: Leveraging RDF Graphs for Crossing Multiple Bilingual Dictionaries. In: Proc. of 10th Language Resources and Evaluation Conference (LREC'16) Portorož (Slovenia). pp. 868–876. European Language Resources Association (ELRA), Paris, France (may 2016)