Proceedings of the Globalex Workshop on Linked Lexicography, pages 111–115
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Multi-Strategy system for translation inference across dictionaries

Lacramioara Dranca
Centro Universitario de la Defensa
Ctra. de Huesca, Zaragoza, España
licri@unizar.es

Abstract

This paper describes four different strategies proposed to the TIAD 2020 Shared Task for automatic translation inference across dictionaries. The proposed strategies are based on the analysis of Apertium RDF graph, taking advantage of characteristics such as translation using multiple paths, synonyms and similarities between lexical entries from different lexicons and cardinality of possible translations through the graph. The four strategies were trained and validated on the Apertium RDF $EN \leftrightarrow ES$ dictionary, showing promising results. Finally, the strategies, applied together, obtained an F-measure of 0.43 in the task of inferring the dictionaries proposed in the shared task, ranking thus third with respect to the other new systems presented to the TIAD 2020 Shared Task. No system presented to the shared task exceeded the baseline proposed by the TIAD organizers.

Keywords: Dictionary generation, Automatic inference translation, Graph based heuristics

1. Introduction

The TIAD (Gracia and Kabashi, 2020) shared task is aimed at exploring methods and techniques that infer translations indirectly between language pairs, based on other bilingual resources.

The organizers provide Apertium RDF (Gracia et al., 2014), a set of 22 Apertium bilingual dictionaries, published as linked data on the Web. The Apertium RDF groups the bilingual dictionaries in the same graph, interconnected through the common lexical entries of the monolingual lexicons that they share.

Although the Apertium RDF graph contains multiple connections that represent translations, not all the Apertium RDF lexicons are interconnected. The challenge of the task is to automatically infer translations between English and French lexicons, French and Portuguese lexicons, and Portuguese and English lexicons, respectively, based on the existing bilingual dictionaries from Apertium RDF. Additionally, there is also possible to make use of other freely available sources of background knowledge to improve performance, as long as no direct translation among the target language pairs is applied.

The automatically inferenced translation methods could reduce the costs of constructing bilingual dictionaries. Nevertheless, despite the advantages that the automatic translation inference across dictionaries might have, this task is still challenging (Gracia et al., 2019).

Translation inference across dictionaries based on current methods such as word embeddings (Donandt and Chiarcos, 2019; Garcia et al., 2019) still obtains lower results than more traditional heuristics (Tanaka and Umemura, 1994). Some graph traversal heuristics for this shared task have been proposed previously in (Torregrosa et al., 2019). The hypothesis of this work is that graph-based heuristics may still have potential for improving results. The aim of this work is to try to take full advantage of the potential of translation inference heuristics, based on the Apertium RDF graph, with the benefit of obtaining possibly more interpretable methods.

2. Materials and methods

The Apertium RDF (Gracia et al., 2014) is used to develop the proposed translation heuristics proposed in this paper. The Figure 1 shows the Apertium RDF graph available for the TIAD shared task. The graph contains 13 lexicons, the solid lines show the available translations, the dashed line between English (EN) and Spanish (ES) lexicons is the available translation set that is used in this work for training and validation of the translation strategies proposed in this paper. The dotted lines show the translations aimed to infer with the TIAD shared task and are used for testing the strategies.

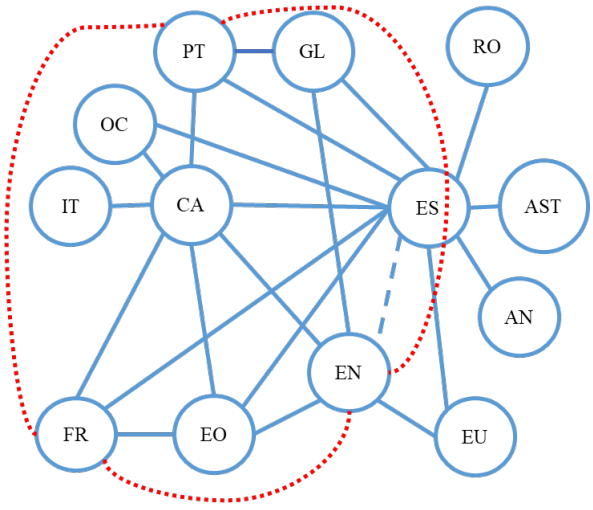

Figure 1: Apertium RDF Graph

As explained by (Saralegi et al., 2011) inferring an A-B bilingual dictionary by merging A-P and P-B dictionaries, using P as pivot lexicon, often produce wrong translations due to polysemous pivot words. To avoid this problem, four translation heuristics or strategies are proposed here, in order to infer translations from a lexicon A to a lexicon B. These strategies are presented below.

2.1. Strategy I

A natural way to address the problem, when multiple paths are available between lexicons A and B, is to validate a translation from $a \in A$ to $b \in B$ if there are multiple paths from a to b across different pivot lexicons. We consider the translation $T = a \leftrightarrow b$ as correct if:

$$b \in translation_{A \leftrightarrow P \leftrightarrow B}(a) \cap translation_{A \leftrightarrow P' \leftrightarrow B}(a) \tag{1}$$

where

$$translation_{A \leftrightarrow P \leftrightarrow B}(a) = translation_{P \leftrightarrow B}(translation_{A \leftrightarrow P}(a)) \tag{2}$$

The strategy requires the existence of two different paths from word a to word b, each path crossing a different pivot lexicon (P and P'), in order to consider the translation as correct. Figure 2 illustrates this strategy (solid lines show existing translations, dashed line shows a new inferred translation). Notice that a, p, b, p' form a 4-cycle graph, a heuristic already used by (Torregrosa et al., 2019).

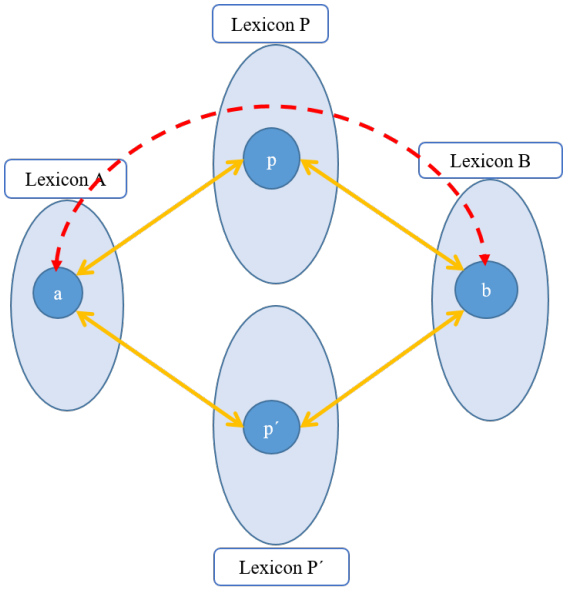

Figure 2: Translation inference across multiple paths

2.2. Strategy II

As mentioned before, the main problem of a direct translation through a pivot lexicon is polysemous words. The polysemy of pivot words implies several meanings for the same word. However, considering the available dictionaries as complete, if the cardinality of a translation through a pivot lexicon is one in both directions, then we assume that it is less likely the translation to be influenced by polysemous words in the pivot lexicon. For these situations we have considered the translation as correct, as expressed in the following equation:

$$\begin{aligned} &a \in A, b \in B \\ &(translation_{A \leftrightarrow P \leftrightarrow B}(a) = \{b\}) \\ &\wedge (translation_{B \leftrightarrow P \leftrightarrow A}(b) = \{a\}) \\ &\Rightarrow \exists T = a \leftrightarrow b \end{aligned} \tag{3}$$

2.3. Strategy III

This strategy attempts to exploit the similarities between different lexicons. A lexical similarity measure s is defined for $a \in A, b \in B$ as follows:

$$s(a,b) = \frac{2 * levenshtein(a,b)}{length(a) + length(b)} \tag{4}$$

This similarity measure is based on the levenstein distance and the leght of the compared words. Notice that for $a = b$ then $s(a,b) = 0$.

Before calculating the lexical similarity between two words, the special characters, typical of each lexicon, have been replaced by the most similar characters from the English alphabet.

For the inference of translations based on lexical similarity for $a \in A, b \in B$ we have considered three settings as follows.

The equation 5 exploits the end-to-end lexical similarity across a path with P as pivot lexicon:

$$\begin{aligned} &(b \in translation_{A \leftrightarrow P \leftrightarrow B}(a)) \wedge (s(a,b) < t_1) \\ &\Rightarrow \exists T = a \leftrightarrow b \end{aligned} \tag{5}$$

The equation 6 exploits the overall lexical similarity across a path with P as pivot lexicon:

$$\begin{aligned} &(p \in translation_{A \leftrightarrow P}(a)) \wedge (b \in translation_{P \leftrightarrow B}(p)) \\ &\wedge (s(a,p) + s(p,b) < t_2) \\ &\Rightarrow \exists T = a \leftrightarrow b \end{aligned} \tag{6}$$

The equation 7 exploits the lexical similarity between translations of the same word $a \in A$ to different lexicons:

$$\begin{aligned} &(b \in translation_{A \leftrightarrow P \leftrightarrow B}(a)) \\ &\wedge (p' \in translation_{A \leftrightarrow P'}(a)) \\ &\wedge (s(b,p') < t_3) \\ &\Rightarrow \exists T = a \leftrightarrow b \end{aligned} \tag{7}$$

The three equations have a corresponding threshold that has been adjusted during training phase. In this work, t_1 and t_3 have been set to 0.17 and t_2 to 0.5.

2.4. Strategy IV

This strategy attempts to exploit the existence of synonymous words in a lexicon, words that might have the same translation to another lexicon. As with the previous strategy 2.3, three settings have been considered.

The first approach is shown in Figure 3 (solid lines show existing translations, dashed line shows new inferred translation)

The equivalent equation is shown below. For $a \in A, b \in B$,

$$\begin{aligned} &\{p_k, p_l\} \in translation_{A \leftrightarrow P}(a)) \\ &\wedge \{p_k, p_l\} \in translation_{B \leftrightarrow P}(b)) \\ &\Rightarrow \exists T = a \leftrightarrow b \end{aligned} \tag{8}$$

where p_k anf p_l might be synonymous words in lexicon P, as reported also in (Torregrosa et al., 2019).

The second approach related to synonymous words is shown in Figure 4 (solid lines show existing translations, dashed lines show new inferred translations)

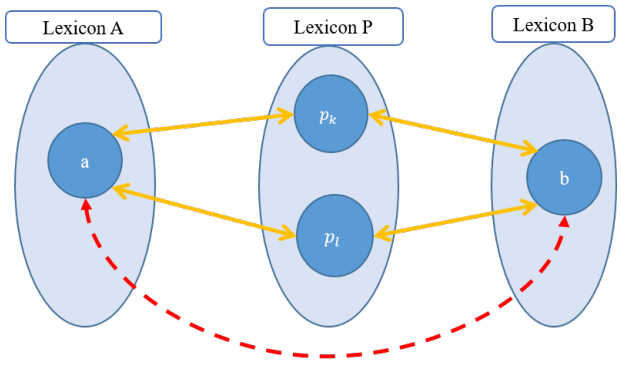

Figure 3: Synonymous strategy across one path

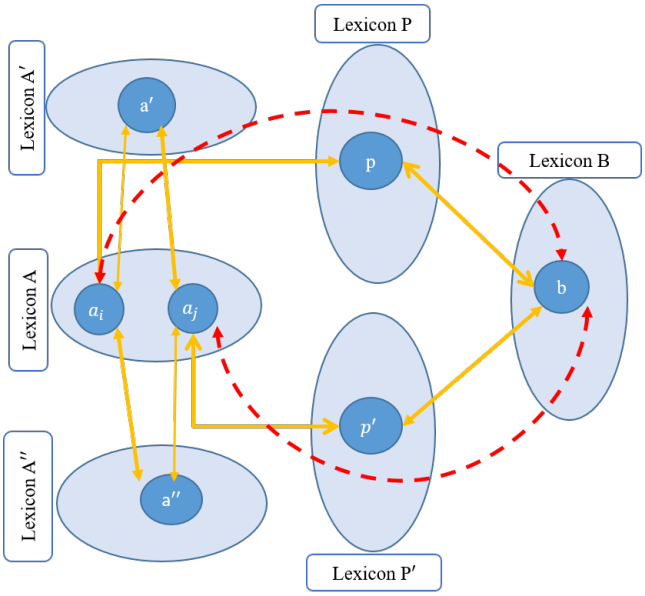

Figure 4: Synonymous strategy across two paths

The equivalent equation is shown below. For $a_i, a_j \in A, b \in B$,

$$\begin{aligned}
&(a_i \in translation_{A \leftrightarrow P \leftrightarrow B}(b)) \\
&\wedge (a_j \in translation_{A \leftrightarrow P' \leftrightarrow B}(b)) \\
&\wedge (\exists a' \in A') \wedge (\exists a'' \in A'') \\
&\wedge \{a_i, a_j\} \in translation_{A' \leftrightarrow A}(a')) \\
&\wedge \{a_i, a_j\} \in translation_{A'' \leftrightarrow A}(a'')) \\
&\Rightarrow (\exists T_1 = a_i \leftrightarrow b) \wedge (\exists T_2 = a_j \leftrightarrow b)
\end{aligned} \tag{9}$$

where a_i and a_j are considered as synonymous words in lexicon A.

This configuration assumes a graph cycle of $length = 7$ $words$ across 6 lexicons.

The third approach related to synonymous words is shown in Figure 5 (solid lines show existing translations, dotted line show an inferred translation in a previous step of the algorithm, dashed line shows a new inferred translation)

The equivalent equation is shown below.

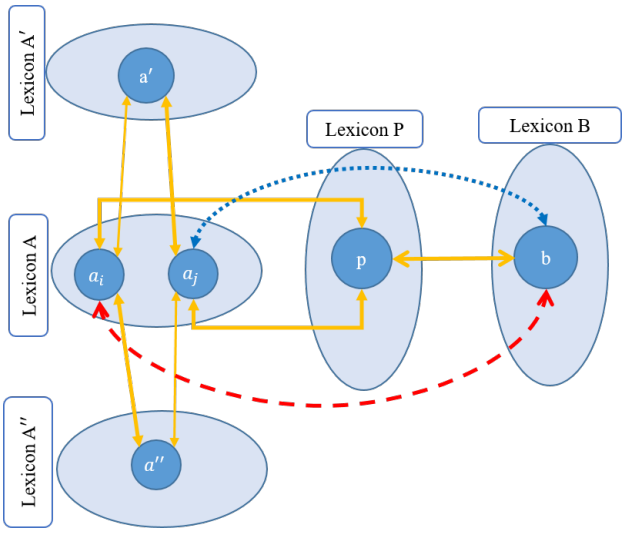

Figure 5: Synonymous strategy across one paths, with previously inferred translations

For $a_i, a_j \in A, b \in B$,

$$\begin{aligned}
&(a_i \in translation_{A \leftrightarrow P \leftrightarrow B}(b)) \\
&\wedge (a_j \in translation_{A \leftrightarrow P \leftrightarrow B}(b)) \\
&\wedge (\exists T_1 = a_j \leftrightarrow b) \\
&\wedge (\exists a' \in A') \wedge (\exists a'' \in A'') \\
&\wedge \{a_i, a_j\} \in translation_{A' \leftrightarrow A}(a')) \\
&\wedge \{a_i, a_j\} \in translation_{A'' \leftrightarrow A}(a'')) \\
&\Rightarrow (\exists T_2 = a_i \leftrightarrow b)
\end{aligned} \tag{10}$$

where a_i and a_j are synonymous words in lexicon A.

3. Results and discussion

The strategies have been evaluated on the available $EN \leftrightarrow ES$ Apertium dictionary and tested by TIAD organizers on $EN \leftrightarrow FR$, $PT \leftrightarrow EN$ and $FR \leftrightarrow PT$ dictionaries. The obtained results are shown below.

3.1. Validation of $EN \leftrightarrow ES$ inferred translations

The validation results of the different translation inference strategies can be seen in Table 1. Proper nouns were not considered for training and validation tasks.

As may be noticed in Table 1, individually, the best strategy for the $EN \leftrightarrow ES$ translations case, considering its F-measure, is strategy II. However, the lexical similarity, the basis of strategy III, is close to strategy II and largely overlaps its correct translations. In the case of Strategy I, it seems important the correct selection of the paths to use, that is, the lexical pivots. Different combinations of two lexical pivots may obtain high degrees of precision (see results for strategy I'), depending on the pivot lexicons used. The path $EN \leftrightarrow EO \leftrightarrow ES$ proven lower precision for both strategies I and IV.

The combination of strategies II, III, IV, without considering Strategy I, also produces competitive results.

Strategy	Precision	Recall	F measure
I	0.87	0.39	0.54
I'	0.94	0.29	0.44
II	0.85	0.48	0.62
III	0.88	0.46	0.60
IV'	0.81	0.15	0.25
$I + II$	0.83	0.63	0.72
$I + II + III$	0.81	0.66	**0.73**
$I + II + III + IV'$	0.80	0.67	**0.73**
$II + III$	0.83	0.59	0.69
$II + III + IV'$	0.81	0.65	0.72

Table 1: Validation results of $EN \leftrightarrow ES$ translations

I' path $EN \leftrightarrow EO \leftrightarrow ES$ has not been considered
IV' path $EN \leftrightarrow EO \leftrightarrow ES$ has not been considered

3.2. Test of TIAD inferred translations

Several systems have been presented to TIAD 2020 shared task. The average results of those systems can be seen in Table 2 (in bold letters the strategies proposed in this paper).

System	P	R	F
Baseline-OTIC	0.70	0.47	0.56
Ciclos-OTIC	0.64	0.47	0.54
NUIG	0.77	0.35	0.49
Multi-StrategyI+II+III+IV	**0.61**	**0.33**	**0.43**
Multi-StrategyI+II+III	**0.62**	**0.33**	**0.43**
CL-embeddings	0.62	0.32	0.42
Multi-StrategyI+II	**0.65**	**0.30**	**0.40**
ACOLIbaseline	0.60	0.28	0.38
Baseline-Word2Vec	0.30	0.37	0.33
Multi-StrategyI	**0.63**	**0.22**	**0.32**
ACOLIwordnet	0.61	0.16	0.25

Table 2: TIAD shared task - average systems results

P stands for Precision
R stands for Recall
F stands for F-measure

As it can be seen, the four strategies proposed in this paper, applied together, obtain a medium result, ranking third with respect to the other new systems presented to the shared task and below the Baseline-OTIC of the task. (results ordered by F-measure - F columns in Table 2). It should be noted that the OTIC method proposed as baseline by the TIAD organizers continues to be the method with the best results, despite being a traditional method (Tanaka and Umemura, 1994) that only use the Apertium RDF graph.

The results for the three dictionaries that we have inferred with the strategies presented in this work can be seen in Table 3 for $EN \rightarrow FR$ translations, in Table 4 for $PT \rightarrow EN$ translations and in Table 5 for $FR \rightarrow PT$ translations, respectively.

As may be observed, the precision of this proposal is superior to the Baseline-OTIC only in the case of the dictionary

System	P	R	F
Ciclos-OTIC	0.57	0.44	0.50
Baseline-OTIC	0.64	0.38	0.48
NUIG	0.68	0.31	0.43
CL-embeddings	0.52	0.35	0.42
Multi-StrategyI+II+III+IV	**0.52**	**0.34**	**0.41**
Multi-StrategyI+II+III	**0.52**	**0.34**	**0.41**
Multi-StrategyI+II	**0.53**	**0.31**	**0.39**
Multi-StrategyI	**0.53**	**0.28**	**0.37**
ACOLIbaseline	0.48	0.24	0.32
Baseline-Word2Vec	0.23	0.39	0.29
ACOLIwordnet	0.54	0.13	0.21

Table 3: Systems results for $EN \rightarrow FR$

System	P	R	F
Ciclos-OTIC	0.68	0.43	0.53
Baseline-OTIC	0.71	0.40	0.51
Multi-StrategyI+II+III+IV	**0.74**	**0.32**	**0.45**
Multi-StrategyI+II+III	**0.76**	**0.31**	**0.44**
CL-embeddings	0.80	0.28	0.41
Multi-StrategyI+II	**0.8**	**0.27**	**0.4**
ACOLIbaseline	0.66	0.26	0.38
Baseline-Word2Vec	0.37	0.33	0.35
Multi-StrategyI	**0.74**	**0.17**	**0.28**
ACOLIwordnet	0.67	0.16	0.25
NUIG	-	-	-

Table 4: Systems results for $PT \rightarrow EN$

System	P	R	F
Baseline-OTIC	0.74	0.54	0.62
Ciclos-OTIC	0.67	0.55	0.6
NUIG	0.84	0.40	0.54
Multi-StrategyI+II+III+IV	**0.58**	**0.34**	**0.43**
Multi-StrategyI+II+III	**0.59**	**0.34**	**0.43**
CL-embeddings	0.55	0.34	0.42
Multi-StrategyI+II	**0.62**	**0.31**	**0.41**
ACOLIbaseline	0.63	0.27	0.38
Multi-StrategyI	**0.61**	**0.21**	**0.31**
Baseline-Word2Vec	0.27	0.34	0.30
ACOLIwordnet	0.62	0.15	0.24

Table 5: Systems results for $FR \rightarrow PT$

$PT \rightarrow EN$. This may be due to the fact that the strategies developed here have been trained on the $EN \leftrightarrow ES$ dictionary and the Portuguese lexicon might have more similarities to the Spanish lexicon and the pivot lexicons used. Perhaps, having used other dictionaries for training, either as an alternative or in addition to the $EN \leftrightarrow ES$ dictionary, could have improved the results of the approach used in this work. The worst precision of this strategy is obtained for the $EN \rightarrow FR$ dictionary, that might prove that using EO lexicon as pivot was not a correct approach.

From the results obtained in the shared task, it can be seen that Strategy I is the least stable, with disparate results, de-

pending on the dictionary to be inferred. This was also observed during the training and validation phases. This strategy is highly dependent on the pivot lexicons used (see results in Table 1). Those good results for Strategy I in the vaidation phase, as concerns the precision, have been a mirage. This may be because some of the pivot lexicons used for the shared task might be related to each other and might share polysemous cases.

Strategy II has better precision than Strategy I in the shared task and it seems a good strategy to maintain in a future work.

Regarding Strategy III, lexical similarity could be useful to improve results in dictionary inference. It improves de F-measure results in the three inferred dictionaries for the shared task, with respect to StrategyI+II. It could be a good method to improve results of other approaches. This method still has room for improvement, using, for example, other similarity measures and optimization techniques to set the thresholds of the method.

Strategy IV, based on the use of synonyms, has proven to be a valid strategy, although it improves the final results very little, at least when used in conjunction with Strategy I. In the validation phase, Strategy IV, used in conjunction with strategies II and III, has shown more significant improvements in the results with respect to the StrategyII + III configuration. Nevertheless, the second setting of Strategy IV (see Figure 4) might have similar drawbacks as Strategy I. It would have been interesting to test a Multi-StrategyII+III+IV in the TIAD shared task, as the results in the validation phase were promising.

4. Conclusion

In this paper four strategy for translation inference across dictionaries have been proposed. The strategies are based on translation using multiple paths, the use of synonyms and similarities between lexical entries from different lexicons and cardinality of possible translations through the graph. The strategies have been trained and validated on the Aperium RDF graph using the dictionary $EN \leftrightarrow ES$, showing promising results. The four proposed strategies, applied together, obtained an F-measure of 0.43 in the task of inferring the proposed dictionaries for the TIAD 2020 Shared Task, thus ranking third with respect to the new systems presented to the shared task. Among the four strategies, the strategy based on lexical similarity stands out. It is a strategy that could enhance other systems and that still has room for improvement.

5. Acknowledgements

This work has been supported by the European Union's Horizon 2020 research and innovation programme through the project Prêt-à-LLOD (grant agreement No 825182). It has been also partially supported by the Spanish National projects TIN2016-78011-C4-2-R (AEI/FEDER, UE) and DGA/FEDER 2014-2020 "Construyendo Europa desde Aragón".

6. Bibliographical References

Donandt, K. and Chiarcos, C. (2019). Translation inference through multi-lingual word embedding similarity. In *Proceedings of TIAD-2019 Shared Task – Translation Inference Across Dictionaries)*, Leipzif, Germany, May.

Garcia, M., Garcıa-Salido, M., and Alonso, M. A. (2019). Exploring cross-lingual word embeddings for the inference of bilingual dictionaries. In *Proceedings of TIAD-2019 Shared Task – Translation Inference Across Dictionaries*, Leipzif, Germany, May.

Gracia, J. and Kabashi, B. (2020). 3rd Translation Inference Across Dictionaries shared task. In conjunction with the GLOBALEX 2020 at LREC2020. Marseille, France, May. URL: `https://tiad2020.unizar.es/`.

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*, Leipzif, Germany, May.

Saralegi, X., Manterola, I., and Vicente, I. S. (2011). Analyzing methods for improving precision of pivot based bilingual dictionaries. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 846–856. Association for Computational Linguistics.

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.

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. May.

7. Language Resource References

Jorge Gracia and Esther Lozano and Julia Bosque-Gil. (2014). *Apertium RDF*. URL: `http://linguistic.linkeddata.es/apertium/`.