# TIAD 2019 Shared Task: Leveraging Knowledge Graphs with Neural Machine Translation for Automatic Multilingual Dictionary Generation

Daniel Torregrosa, Mihael Arcan, Sina Ahmadi and John P. McCrae

Insight Centre for Data Analytics
Data Science Institute
National University of Ireland Galway
firstname.lastname@insight-centre.org

**Abstract.** This paper describes the different proposed approaches to the TIAD 2019 Shared Task, which consisted in the automatic discovery and generation of dictionaries leveraging multilingual knowledge bases. We present three methods based on graph analysis and neural machine translation and show that we can generate translations without parallel data. [1]

**Keywords:** Neural machine translation · Dictionary generation · Automatic inference

## 1 Introduction

The growing amount of semantically structured monolingual, as well as multilingual resources, such as dictionaries or knowledge graphs (KGs), offers an excellent opportunity to explore, link and to enrich them with possibly missing multilingual knowledge. Since a manual translation of such resources is very time consuming and expensive, this work focuses on the automatic generation of dictionary entries, which is the objective of the Translation Inference Across Dictionaries (TIAD-2019) Shared Task.[2] In this task, the participants should submit dictionaries containing pairs of source and target language words or expressions, the part of speech (POS) of the entry and a confidence score.

In this work, we propose several different approaches for this task:

– a graph-based approach where loops of length four are searched in the Apertium [2] dictionaries in order to discover new translations;
– a path-based graph approach which retrieves the translation candidates based on translation inference using the language paths of the Apertium dictionaries;

[1] The datasets used and the trained NMT models are available at server1.nlp.insight-centre.org/tiad2019/
[2] tiad2019.unizar.es/

– a multi-way neural machine translation (NMT) model trained with multi-parallel English and Spanish, Italian and Portuguese and French and Romanian corpora and tuned with the dictionaries produced in the other approaches. This approach is further filtered using monolingual dictionaries.

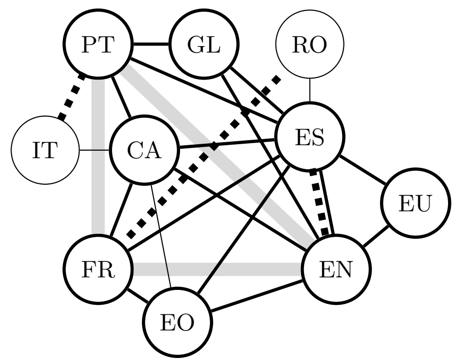


**Fig. 1.** Apertium RDF graph. The graph shows the different available Apertium dictionaries (solid black lines), the new directions targeted in the task (transparent gray lines), and the parallel corpora used by the NMT system (dotted lines).

## 2 System Overview

We developed different approaches to discover dictionary entries: two based on graph traversal heuristics and one using multi-way neural machine translation.

### 2.1 Cycle-based approach

We devised a heuristic that focuses on producing high precision entries, even though the recall might suffer. The model builds a graph with all the bilingual word-POS pairs in all the Apertium dictionaries that can be used as a pivot between Portuguese, French and English, that is, all the language pairs with thick lines in Figure 1. Whenever we find a length 4 cycle in the graph, we connect all the nodes in the cycle. All discovered edges for the respective language pair are used as dictionary. Figure 2 shows an example of discovered translations.

### 2.2 Path-based approach

Similar to the cycle-based method, we use another heuristic technique, which aims to create translation candidates by traversing the paths between the source and the target languages in the Apertium language graph. The candidate translations $T$ are weighted with respect to the path length and the frequency. In this

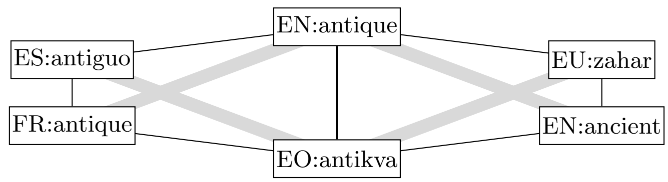


**Fig. 2.** Cycles found in the dictionary (solid lines) and inferred translations (transparent lines). Some of the lines identify possible same-language synonyms (e.g. *ancient* and *antique* in English), while others identify newly discovered possible translations (e.g. *antiguo* in Spanish and *antikva* in Esperanto).

section, *language graph* refers to the Apertium dictionary graph (Figure 1) and *translation graph* refers to a graph where vertices represent a word and edges represent the translations in other languages. Figure 3 illustrates the translation graph of the word spring as a noun in English based on the language path English→Basque→Spanish→French→Esperanto→Catalan→Portuguese.

The basic idea behind pivot-oriented translation inference is the transitivity assumption of translations. If $w_p$, a pivot word in the dictionary $\mathcal{D}_p$, has the translation equivalents $w_i$ and $w_j$ in dictionaries $\mathcal{D}_{p\to 1}$ and $\mathcal{D}_{p\to 2}$ respectively, then $w_i$ and $w_j$ may be equivalents in the $\mathcal{D}_{1\to 2}$ dictionary. Although the pivot-oriented approach can mostly create accurate translations for monosemous words (depending on the lexicon completeness), this oversimplifies for polysemous words leading to incorrect translations [6].

For the current task, we have considered all the simple paths, i.e. paths without any repeating vertex, starting from and ending with the goal languages of the shared task, of which there are 92, 67 and 58 simple paths between Portuguese→French, French→English and English→Portuguese, respectively. As the language graph is undirected, the paths between the vertices are identical in each direction. For instance, the dictionary paths from English to Portuguese are the same as those from Portuguese to English.

In order to describe the likelihood of a translation being correct, we introduce the following weighting factor:

$$w_t = frequency(t) \times \alpha^l \tag{1}$$

where $w_t$ is the weight of the translation candidate $t \in T$, $frequency(t)$ is the number of times translation $t$ is reached, $\alpha \in (0, 1)$ is a penalisation constant and $l$ is the length of the path leading to $w_t$. $\alpha^l$ penalises the translation candidates in a way that paths of lower length and higher frequency get a lower weight. On the other hand, a longer translation path results in a lower weight factor. For instance, in the translation graph in Figure 3, the frequency of the words *primavera*, *font* and *origem* is respectively 2, 1 and 1 and the length of their translation path is 7. For the current task, we set $\alpha = 0.5$ and have included the part-of-speech tags in the inference.

Finally, the weights are normalised such that $\sum_{t\in T} w_t = 1$.

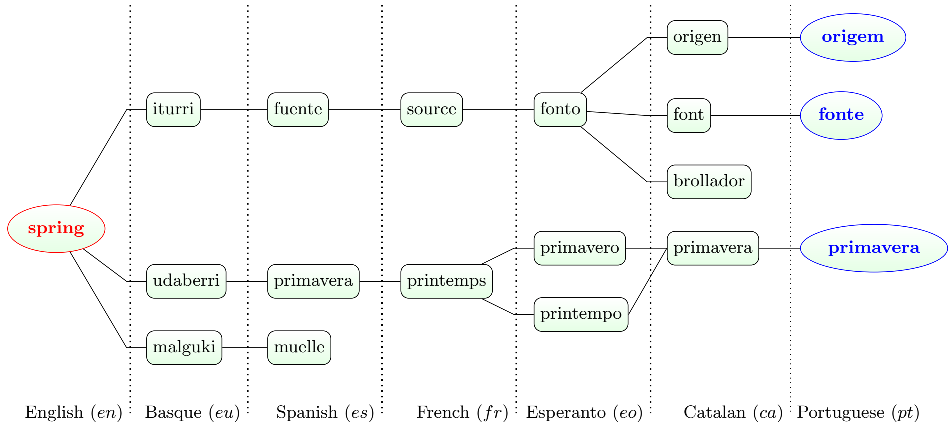


**Fig. 3.** Translation graph of *spring* (noun) (in red) resulting in Portuguese translations (in blue) using the pivot languages.

### 2.3 Multi-way neural machine translation

To perform experiments on NMT models with a minimal set of parallel data, i.e. for less-resourced languages, we trained a multi-source and multi-target NMT model [3, 4] with well-resourced language pairs. In our work, we have chosen the part of the parallel corpora belonging to the Romance language family, i.e. Spanish, Italian, French, Portuguese, Romanian, as well as English. To train the multi-way NMT system, we restricted the language pairs to English-Spanish, French-Romanian and Italian-Portuguese, as shown with a dotted line in Figure 1.

***Continuous training with a discovered dictionary*** To allow the NMT model to align words in the embedding space between the language pairs of the task, we used the trained multi-way model[3] and continued the training of the network based on the output of the approaches presented in Section 2.1 and Section 2.2. Without this procedure, the initial multi-way system could not generate a translation of the targeted language pairs and would instead generate translations into the language paired in the training; e.g. when requesting French to English translations, the system generated Romanian, as the multi-way model only knows how to translate between English and Spanish, between French and Romanian and between Italian and Portuguese.

For the continuous training of the multi-way model, we experimented with two different datasets: each one of the dictionaries generated following the cycle (Section 2.1) and path (Section 2.2) strategies.

***Filtering*** The NMT models were trained without POS information; hence, the NMT model is unable to assign a POS tag to each predicted translation. As the

[3] Trained on English-Spanish, French-Romanian and Italian-Portuguese parallel data

| | EN-FR | FR-EN | EN-PT | PT-EN | FR-PT | PT-FR |
|---|---|---|---|---|---|---|
| Cycle | 7,041 | | 142 | | 100 | |
| Path | 25,594 | 26,492 | 16,273 | 22,195 | 19,079 | 27,678 |
| Wiktionary reference | 57,998 | 58,004 | 46,079 | 46,077 | 46,967 | 46,965 |
| TIAD reference | 14,512 | 20,800 | 12,811 | 17,498 | 10,791 | 10,808 |

**Table 1.** Discovered dictionaries using both the cycle and path approaches, and the dictionary extracted from Wiktionary that will be used as a reference. Unlike the path strategy, that creates different dictionaries for each translation direction, the cycle strategy creates symmetric dictionaries.

TIAD 2019 shared task required POS information, we used a further filtering strategy. We used monolingual Portuguese, French and English dictionaries extracted from Wiktionary, and, for each translation prediction, we examined if the word exists in both dictionaries with the same POS tag.[4] One entry is generated for each shared POS tag.

For example, targeting English to Portuguese, when we request the translation for *snake* the system generates *serpente*; *snake* appears both as a noun and a verb in the English Wiktionary, but *serpente* only appears as a noun in the Portuguese Wiktionary. Hence, we generate an entry in the English-Portuguese dictionary where the noun *snake* has *serpente* as a possible translation. As a side effect this approach removes those words that are incorrectly generated due to the use of subword units, and inflected words (i.e. words not in a canonical form) that are seldom included in dictionaries.

### 2.4 Datasets

For the cycle (Section 2.1) and path (Section 2.2) strategies, we used the Apertium dictionaries[5] shown in Figure 1.

In order to train the multi-way model described in Section 2.3, we used the *Directorate-General for Translation* (DGT) corpus [8]. The English, Spanish, French, Romanian, Italian and Portuguese languages were selected to train the multi-way NMT system. We used the same distribution of the dataset, as described in [1].

### 2.5 Neural Machine Translation Toolkit

We used OpenNMT [5], a generic deep learning framework mainly specialised in sequence-to-sequence models covering a variety of tasks such as machine translation, summarisation, speech processing and question answering as NMT

[4] The POS tags in Wiktionary are different for each language; we mapped each language-specific POS into our generic taxonomy for this task.

[5] Version `cfc5995c5369ddb158cd266fcb8d4b74ed8dbdd0`.

| SYSTEM | PRECISION | RECALL | F1 | COVERAGE |
|---|---|---|---|---|
| BASELINE(OTIC) | 0.64 | 0.26 | $0.37^{*}$ | 0.45 |
| BASELINE(W2VEC) | 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 | 0.30 | 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 | 0.10 | 0.18 | 0.20 |
| **UNLP-NMT-4CYCLE** | 0.58 | 0.11 | 0.18 | 0.25 |
| ONETA-ES | 0.81 | 0.10 | 0.17 | 0.17 |
| ONETA-CA | $0.83^{*}$ | 0.08 | 0.14 | 0.13 |
| **UNLP-4CYCLE** | 0.75 | 0.07 | 0.11 | 0.13 |

**Table 2.** Reported average results for all the submissions. Submissions described in this paper appear in bold. The best result for each metric is marked with $*$.

framework. Due to computational complexity, the vocabulary in NMT models had to be limited. In order to overcome this limitation, we used byte pair encoding (BPE) to generate subword units [7]. BPE is a form of data compression that iteratively replaces the most frequent pair of bytes in a sequence with a single, unused byte.

## 3 Participation in the task

The Translation Inference Across Dictionaries 2019 Shared Task focuses on the automatic generation of dictionary entries in order to enrich knowledge graphs with multilingual knowledge. The task focuses on the generation of French-English, Portuguese-English and French-Portuguese dictionaries. The organisers recommended the use of the Apertium RDF dataset[6] (Figure 1 shows the languages in Apertium RFT used by our submission), although the usage of other resources, other than parallel data between the targeted languages, was permitted.

Evaluation was carried out by the organisers of the task, using manually compiled pairs of *K Dictionaries* as a gold standard. Precision, recall, F-measure and coverage were reported and participants submitted 11 different submissions to the task. The averaged results for all the different models can be seen in Table 2 and the in-depth results for our submitted models can be seen in Table 3.

## 4 Conclusion

Despite the fact that we are using a straightforward approach to extract candidates from Apertium RDF, the precision of our models rank among the highest for the

[6] linguistic.linkeddata.es/apertium/

task. We demonstrate the multi-way NMT approach, which generates translations without requiring any parallel data between the targeted languages. Additionally, when using the path or NMT approaches to generate entries, a threshold can be modified to obtain dictionaries with higher recall and lower precision or vice-versa, what can be used to adapt the method for different use cases.

## Acknowledgment

This publication has emanated from research supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289, co-funded by the European Regional Development Fund, and the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015, ELEXIS - European Lexical Infrastructure.

## References

1. Mihael Arcan, Daniel Torregrosa, Sina Ahmadi, and John P. McCrae. Inferring Translation Candidates for Multilingual Dictionary Generation with Multi-Way Neural Machine Translation. In *Proceedings of the Translation Inference Across Dictionaries (TIAD 2019)*, Leipzig, Germany, 2019.
2. Mikel L Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M Tyers. Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25(2):127–144, 2011.
3. Thanh-Le Ha, Jan Niehues, and Alexander H. Waibel. Toward multilingual neural machine translation with universal encoder and decoder. *CoRR*, abs/1611.04798, 2016.
4. Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017.
5. Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. Opennmt: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada, July 2017. Association for Computational Linguistics.
6. Xabier Saralegi, Iker Manterola, and Inaki San Vicente. 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, 2011.
7. Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 1: Long Papers, pages 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics.
8. Ralf Steinberger, Andreas Eisele, Szymon Klocek, Spyridon Pilos, and Patrick Schlüter. DGT-TM: A freely available Translation Memory in 22 languages. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC12)*, pages 454–459, Istanbul, Turkey, May 2012.

| | UNLP-4CYCLE | | | | UNLP-GRAPH | | | | UNLP-NMT-3PATH | | | | UNLP-NMT-4CYCLE | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| T | P | R | F1 | C | P | R | F1 | C | P | R | F1 | C | P | R | F1 | C |
| 0 | 0.75 | 0.07 | 0.11 | 0.13 | 0.26 | 0.28 | 0.26 | 0.45 | 0.1 | 0.28 | 0.14 | 0.72 | 0.09 | 0.24 | 0.13 | 0.71 |
| 0.1 | 0.75 | 0.07 | 0.11 | 0.13 | 0.4 | 0.26 | 0.31 | 0.45 | 0.1 | 0.28 | 0.14 | 0.72 | 0.09 | 0.24 | 0.13 | 0.71 |
| 0.2 | 0.75 | 0.07 | 0.11 | 0.13 | 0.51 | 0.24 | 0.32 | 0.44 | 0.1 | 0.28 | 0.14 | 0.72 | 0.09 | 0.24 | 0.13 | 0.71 |
| 0.3 | 0.75 | 0.07 | 0.11 | 0.13 | 0.59 | 0.21 | 0.31 | 0.4 | 0.1 | 0.28 | 0.14 | 0.72 | 0.09 | 0.24 | 0.13 | 0.71 |
| 0.4 | 0.75 | 0.07 | 0.11 | 0.13 | 0.65 | 0.19 | 0.29 | 0.36 | 0.1 | 0.28 | 0.14 | 0.72 | 0.09 | 0.24 | 0.13 | 0.71 |
| 0.5 | 0.75 | 0.07 | 0.11 | 0.13 | 0.68 | 0.17 | 0.26 | 0.33 | 0.1 | 0.28 | 0.14 | 0.72 | 0.1 | 0.24 | 0.13 | 0.71 |
| 0.6 | 0.75 | 0.07 | 0.11 | 0.13 | 0.75 | 0.14 | 0.23 | 0.27 | 0.1 | 0.27 | 0.14 | 0.72 | 0.1 | 0.24 | 0.14 | 0.71 |
| 0.7 | 0.75 | 0.07 | 0.11 | 0.13 | 0.76 | 0.12 | 0.21 | 0.24 | 0.12 | 0.26 | 0.15 | 0.71 | 0.11 | 0.23 | 0.14 | 0.7 |
| 0.8 | 0.75 | 0.07 | 0.11 | 0.13 | 0.76 | 0.11 | 0.19 | 0.22 | 0.2 | 0.24 | 0.19 | 0.64 | 0.17 | 0.21 | 0.16 | 0.64 |
| 0.9 | 0.75 | 0.07 | 0.11 | 0.13 | 0.76 | 0.1 | 0.18 | 0.2 | 0.48 | 0.17 | 0.25 | 0.39 | 0.37 | 0.16 | 0.21 | 0.43 |
| 1 | 0.75 | 0.07 | 0.11 | 0.13 | 0.75 | 0.09 | 0.16 | 0.18 | 0.87 | 0 | 0.01 | 0 | 0.89 | 0 | 0 | 0 |

**Table 3.** Results of the evaluation carried out by the organisers of our submissions, with precision (P), recall (R), F-measure (F1), and coverage (C) for all the different thresholds (T).