Proceedings of TIAD-2017 Shared Task – Translation Inference Across Dictionaries

Language, Data and Knowledge conference
Galway, Ireland, 18 June 2017

Introduction

Various methods and techniques have been explored in the past in the aim of automatically generating new bilingual (and multilingual) dictionaries based on existing ones, so that given L1>L2 and L2>L3 sets, a new L1>L3 dictionary is produced. The intermediate language that is used in the process is called a pivot, and it is possible to use multiple pivots for this purpose. Although considerable work has been done to this end, it was usually conducted on different types of datasets and evaluated in different ways, applying various algorithms that are often not comparable.

The Translation Inference Across Dictionaries shared task (TIAD-2017) was launched with the intention of offering quality lexical resources within the context of a single coherent experiment, which enables reliable validation of results and solid comparison of methods and techniques for auto-generating translation equivalents across languages, as well as stimulating and enhancing further research.

The data for TIAD-2017 was released by K Dictionaries Ltd and extracted from its Global Series, which includes monolingual, bilingual and multilingual lexical sets for 24 languages (Kernerman, 2011). Each language core is compiled independently within a common overall framework and sharing the same technical infrastructure with all other languages. It consists of detailed, well-structured lexicographic information on the core language, which serves as a base for translation to any other language. The language pairs were thus created directly between each other, with no intermediate language bias (e.g. English, in WordNet), and joining the different pairs of each language core forms a multilingual network.

All the evaluation data and results is freely available on the Web (see https://tiad2017.wordpress.com/)

The shared task

The objective of the shared task was to indirectly generate translations for three language pairs, based on the available translations among eight languages altogether, in 14 bilingual dictionaries, involving four possible paths – all from German to Brazilian Portuguese – that feature between 1 to 4 pivot languages.

The test dataset consisted of 100 randomly-selected German dictionary entries with their translations into a second language, and recursively exploring further translations in chained-up dictionaries – including up to 817 entries with 1,532 translation equivalents in the largest language pair that is provided. Besides the headwords and translations, the data includes information about parts of speech, subject domains and synonyms, as well as examples of usage and their translations.

The following language pairs were provided for the four different paths:

1. German>English | English>Portuguese

2. German>Japanese | Japanese>Spanish | Spanish>Portuguese
3. German>Danish | Danish>French | French>Spanish | Spanish>Portuguese
4. German>Dutch | Dutch>Spanish | Spanish>Danish | Danish>French | French>Portuguese

Also included were four Portuguese>German datasets, for closing the loop in each path, to help with the validation of the results.

The three new language pairs that were generated are:

- German>Portuguese
- Danish>Spanish
- Dutch>French

The Portuguese>German sets were provided for the sole purpose of closing the loop as an aid for validating the results, but were not allowed to be reversed to improve results. However, it was allowed to reverse the Spanish>Danish dataset – that was provided as part of path (4) – o helpt with improving the results.

Evaluation of the results of each system was carried out against KD's manually compiled dictionaries for these pairs from the Global Series and other resources, as well as by human translators.

Participants were invited to contribute on either or both of the following tracks:

- Systems that use only the KD data released for the task
- Systems that exploit, in addition to the KD data, other freely available sources of background knowledge (e.g., lexical linked open data and parallel corpora) to improve performance

Beyond performance, participants were encouraged to consider the following issues in particular:

- The role of the language family with respect to the newly generated pairs
- The asymmetry of pairs, and how the translation direction affected the results
- The behavior of different parts of speech among different languages
- The role that the number of pivots played in the process

Results

Four teams participated in the shared task, but one of them was not able to produce results on time for the workshop. All the teams were invited to submit a paper with a technical description of their system and their results. We received three submissions. After the peer reviewing process, one of them was withdrawn and two were accepted for publication. We gave all the participants the opportunity of sharing their ideas and results with an oral presentation at the workshop: long presentations for the two participant teams with accepted papers, and short presentations for the other two. The list of presented works is as follows:

Paper presentations:

- Thomas Proisl, Philipp Heinrich, Stefan Evert and Besim Kabashi. "Translation inference across dictionaries via a combination of graph-based methods and co-occurrence statistics"
- Kathrin Donandt, Christian Chiarcos and Maxim Ionov. "Using Machine Learning for Translation Inference Across Dictionaries"

Short presentations:

- Tom Knorr. "Translation Inference Across Dictionaries TIAD 2017 – Shared Task Data Analysis"
- Uliana Sentsova. "Report on TIAD participation"

Organisation

Organisers

Noam Ordan, K Dictionaries
Jorge Gracia, Ontology Engineering Group, Universidad Politécnica de Madrid
Morris Alper, K Dictionaries
Ilan Kernerman, K Dictionaries

Reviewing committee

Irith Ben-Arroyo Hartman, University of Haifa, Israel
Thierry Declerck, German Research Center for Artificial Intelligence, Germany
Thierry Fontenelle, Translation Center for the Bodies of the EU, Luxembourg
Mikel Forcada, Universidad de Alicante, Spain
Jorge Gracia, Universidad Politécnica de Madrid, Spain
Miloš Jakubíček, Lexical Computing, Czech Republic
Jelena Kallas, Institute of the Estonian Language, Estonia
Ilan Kernerman, K Dictionaries, Israel
Iztok Kosem, Trojina Institute and University of Ljubljana, Slovenia
Nikola Ljubešić, University of Zagreb, Croatia
Shervin Malmasi, Harvard University, USA
John McCrae, National University of Ireland, Galway
Elena Montiel-Ponsoda, Universidad Politécnica de Madrid, Spain
Preslav Nakov, Hamad Bin Khalifa University, Qatar
Noam Ordan, K Dictionaries and The Arab Academic College of Education, Israel
Georg Rehm, German Research Center for Artificial Intelligence, Germany
Victor Rodriguez-Doncel, Universidad Politécnica de Madrid, Spain
Liling Tan, Saarland University / Nanyang Technological University
Carole Tiberius, Institute of Dutch Language, Netherlands
Marta Villegas, Spain
Marcos Zampieri, University of Köln, Germany

Workshop schedule

9:30 – 9:45	Ilan Kernerman. *Translation Inference Across Dictionaries - Introduction*
9:45 – 10:00	Jorge Gracia. *Previous experiences in translation inference: the Apertium RDF case*
10:00 – 10:15	Noam Ordan. *Shared task description*
10:15 – 10:30	Morris Alper & Noam Ordan. *Base line computation and evaluation process*
10:30 – 11:00	coffee break
11:00 – 11:30	Thomas Proisl, Philipp Heinrich, Stefan Evert and Besim Kabashi. *Translation inference across dictionaries via a combination of graph-based methods and co-occurrence statistics*
11:30 – 12:00	Kathrin Donandt, Christian Chiarcos and Maxim Ionov. *Using Machine Learning for Translation Inference Across Dictionaries*
12:00 – 12:15	Tom Knorr. *Translation Inference Across Dictionaries TIAD 2017 – Shared Task Data Analysis*
12:15 – 12:30	Uliana Sentsova. *Report on TIAD participation*