Using Wiktionary to Create Specialized Lexical Resources and Datasets

Lenka Bajčetić and Thierry Declerck

Austrian Centre for Digital Humanities and Cultural Heritage; DFKI GmbH, Multilinguality and Language Technology Lab, Saarland University

Abstract

This poster presents our work which utilizes Wiktionary data for creating specialized lexical datasets which can be used for enriching other lexical (semantic) resources or for generating datasets that can be used for evaluating or improving NLP tasks, like word sense disambiguation, word-in-context challenges, or sense linking across lexicons and dictionaries. We have focused on Wiktionary data about pronunciation information in English, and grammatical number and grammatical gender in German.

Introduction

Wiktionary is a freely available web-based multilingual dictionary for over 170 languages, supported by the Wikimedia Foundation. The fact that Wiktionary is built by a collaborative effort means that the coverage and variety of lexical information is much larger than any single curated resource, but there are inevitably some variations in data organization and not all language versions follow the same structure for encoding their data.

We focus on **linguistic properties of lexical entries that are often omitted from other resources**, especially from lexical semantics resources, and can be a source of **sense ambiguity**: pronunciation, grammatical number, or grammatical gender.

Approach

Our approach consists of:

- automatically extracting the relevant linguistic data associated with lexical entries of different languages
- storing it in easy-to process formats
- using this data to enrich other lexical resources or providing for lexical datasets to support the evaluation of NLP tasks

English pronunciation dataset

We automatically extracted pronunciation information for **72.067 entries**, out of a total of 887.259 entries, from the English Wiktionary XML dump. Using this data, we have recently generated a lexical dataset containing all the entries of Wiktionary including pronunciation information, together with their definitions, senses, and example usages.

In cases when there is no pronunciation ambiguity, the pronunciation information can be easily automatically integrated to other lexical resources, which are lacking such information. We are happy to report that **the new version of the Open English WordNet includes now over 35,000 entries that are equipped with pronunciation information**.

Heteronyms

Heteronyms are words that have the same spelling, but different pronunciations associated with different senses. The goal of focusing on heteronyms was twofold. First, we considered them an interesting case of ambiguity, and we wanted to see in what way we could tackle sense linking across lexicons and dictionaries when dealing with heteronyms. Hence, one of the goals of our work in this case was to produce a special **gold standard dataset for the task of heteronym sense linking**. This dataset lists a sample of the heteronym entries extracted from Wiktionary, together with definitions and usage examples, so that the entry and its pronunciation are accompanied by textual sequences that can be used for training systems for WSD. Secondly, we have also noticed that even the widely-used lexical resources do not have the infrastructure to output several pronunciations for a single word even when it has multiple senses. It is important to rectify this in order to have the resources capable of capturing all the nuances of a language. As a result of our work, **the Open English WordNet has modified its schema to accept (multiple) pronunciations**.

German dataset for ambiguity in grammatical gender and number

In this part of our work, we are focusing on German nouns having a different meaning related to their grammatical gender or number.

German has three grammatical genders: masculine, feminine, and neutral. The German edition of Wiktionary includes 1,164 nouns having two genders, and 44 having 3 genders. When the grammatical gender of a word changes, so does the flexion of its article and adjectives in the sentences in which they occur, giving thus a (only partial) clue to gender detection, and therefore possibly also for meaning disambiguation. Grammatical gender ambiguity occurs when the ambiguity of a word is related to its possible grammatical genders, each carrying at least one different meaning.

Similarly, ambiguity in grammatical number occurs if a singular word form has a plural form with it own meaning, which is not merely expressing a quantification of the singular word form. Grammatical number ambiguity in German often occurs when two words that look like they are forming a singular-plural pair have different meanings. In fact the two words are not forming such a pair, but are examples of a singulare tantum and a plurale tantum, as can be seen with “Kost” (diet or meal) versus “Kosten” (costs, expenses).

In the end, **we created two German datasets which include those grammatical gender and number ambiguities**. The resulting two datasets consist of the lexical entries associated with the definitions and example sentences that are associated with the different senses listed for the entries. We expect those datasets to be helpful for training WSD tasks on German text, and also for supporting MT, as we noticed problems with the machine translation of German words carrying a gender ambiguity, as it seems that popular MT systems do not take into account enough context for translating such words.

Conclusion and future enhancement

We presented ongoing work in generating specialized datasets from Wiktionary lexical resources. A result was the possibility to add pronunciation information to the Open English WordNet, with a focus on heteronyms. Our approach consists in generating a dataset containing examples sentences and definitions associated to the heteronyms in order to support the automated sense linking of the heteronyms to a WordNet lexical resource.

Another aspect of our work consists in extracting from the German edition of Wiktionary entries that are ambiguous with respect to number and gender information, together with their sense specific definitions and usage examples. As we discovered that a relevant number of Wiktionary multiword entries are missing pronunciation information, we are working on generating such pronunciation information, combining it from the existing pronunciation of their subcomponents. In this way, we can contribute to the enrichment of Wiktionary itself. Completing this missing pronunciation would certainly prove helpful for text-to-speech and speech-to-text tasks. We will also extend the work to other languages and specific linguistic phenomena.

Acknowledgements

This paper is based upon work from the COST Action NexusLinguarum – European network for Web-centered linguistic data science (CA18209), supported by COST (European Cooperation in Science and Technology). It is also supported by the Horizon 2020 research and innovation programme with the projects Pret-a-LLOD (grant agreement no. 825182) and ELEXIS (grant agreeement no. 731015). We would like to thank Annegret Janzso for her very valuable work on the German data. We also thank the anonymous reviewers for their helpful comment