Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 3457–3460
Marseille, 20-25 June 2022

Using Wiktionary to Create Specialized Lexical Resources and Datasets

Lenka Bajčetić[1], **Thierry Declerck**[2]
[1] Austrian Centre for Digital Humanities and Cultural Heritage, Austrian Academy of Sciences, Vienna, Austria
[2] DFKI GmbH, Multilinguality and Language Technology Lab, Saarland University Campus D3 2, Germany
[1]Lenka.Bajcetic@oeaw.ac.at, [2]declerck@dfki.de

Abstract

This paper describes an approach aiming at utilizing 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.

Keywords: Wiktionary, ambiguities, pronunciation,grammatical number, grammatical gender

1. Introduction

In this paper, we investigate the use of Wiktionary (Wikimedia, 2021b) for building (lexical) datasets that can support the improvement and/or evaluation of challenging NLP tasks, like Word Sense Disambiguation (WSD), Word-in-Context challenges (WiC), Sense Linking (SL), or Machine Translation (MT). Extracting lexical information from Wiktionary can also be used for enriching other lexical resources.

Wiktionary is a freely available web-based multilingual dictionary for over 170 languages, supported by the Wikimedia Foundation (Wikimedia, 2021a). Wiktionary provides detailed information on lexical entries, which may include inflectional and derivational information, definitions, examples of usage, pronunciation, etymology, translations, to name just the more relevant for our experiments.

Like other Wikimedia (Wikimedia, 2021a) supported initiatives, Wiktionary is a collaborative project. This implies that there might be inaccuracies in the resource, but the editing system is helping in mitigating this risk. 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, while Wiktionary is integrating information from expert-based dictionary resources, when their licensing conditions allow it.

In addition to its availability, Wiktionary is easily accessible and well-structured. Since Wiktionary grows 'organically' as a product of collaborative contributions, there are inevitably some variations in data organization and not all language versions follow the same structure for encoding their data.

Wiktionary provides a range of data-dumps in XML, which we access using Python scripts. Each lexical entry has its own Wiktionary page and in the XML dump its own <page>-tag, individual words can be found within the <title>-tag, and parts containing the lexical information are available within the <text>-tag. But the information within the <text>-tag is encoded by using a wiki mark-up language. The structure of information at this last tag, however, is not the same throughout all languages, so that each language specific XML dump file has to be accessed individually.

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. Our approach consists of automatically extracting the relevant linguistic data associated with lexical entries of different languages, storing it in easy-to process formats, and finally use this data to enrich other lexical resources, such as adding pronunciation information to the Open English WordNet (McCrae et al., 2020) or providing for lexical datasets to support the evaluation of NLP tasks, like word sense disambiguation, senses linking across lexicons and dictionaries, or machine translation.

2. Extracting an English Pronunciation Dataset from Wiktionary

As we noticed that many lexical-semantics resources are lacking information about pronunciation, and knowing that this information can be relevant for distinguishing between senses of a word, we have in a first step focused on extracting pronunciation information from the English edition of Wiktionary, aiming at enriching with it the Open English WordNet (McCrae et al., 2020), which was lacking this type of information.

We automatically extracted pronunciation information for 72.067 entries, out of a total of 887.259 entries, from the English Wiktionary XML dump.[1] Preliminary work on the base of which we were building the current datasets is detailed in (Declerck and Bajčetić, 2021). Pronunciation information in the English edition of Wiktionary is attached mostly only to lemma forms, and is often not provided for multiword entries. Still, the word forms which do have pronunciation information have provided us with a large enough

[1]We worked on a Wiktionary dump of July 2021.

dataset of English words with pronunciations, structured around the etymologies of the word senses: In Wiktionary, the various senses of an entry, and the associated lexical information, are included in an "etymology" element. 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.

However, there are a number of words that carry more than one pronunciation, being within or across part-of-speech categories. For those cases, we are particularly interested in dealing with so-called heteronyms,[2] which 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 (Bajčetić et al., 2021). 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 word (in written and spoken data) sense disambiguation.[3]

Besides the gold standard dataset, we recently generated a lexical dataset containing all the entries of Wiktionary including pronunciation information, together with their definition(s) (the sense(s)) and examples usage(s). This lexical dataset will be made available in the same GitHub space containing the gold standard dataset.

Second, we have also noticed that even the widely-used lexical resources, like BabelNet,[4] 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.

We can report that, as a result of our work, the Open English WordNet has modified its schema to accept (multiple) pronunciations. Additionally, the new version of the Open English WordNet, which was released in November 2021,[5] includes now over 35,000 entries that are equipped with pronunciation information.

[2]Wiktionary encodes 930 English heteronyms, out of which ca 450 were identified as heteronyms within one PoS category, mainly nouns.

[3]The heteronym gold standard and the work done on heteronym sense linking, can be found here: https://github.com/acdh-oeaw/heteronym_sl.

[4]BabelNet displays the correct IPA transcription for the word "lead" (in the "metal" sense, but the wrong audio signal. See https://babelnet.org/synset?id=bn%3A00006915n&orig=lead&lang=EN

[5]The latest version of Open English WordNet can be found here:https://github.com/globalwordnet/english-wordnet

3. Generation of a German lexical Dataset for Cases of Ambiguity in Grammatical Gender and Number

We extended our "quest" for specific types of lexical ambiguities described in Wiktionary and extracted cases of grammatical gender and grammatical number ambiguities for German. The German edition of Wiktionary includes 1,164 nouns having two genders, and 44 having 3 genders. The English edition of Wiktionary lists 1,269 "German nouns with multiple genders".[6]

There are many cases in which the fact of having more than one grammatical gender does not have impact on the meaning (or sense) of the noun. One reason for this are regional differences, like for example for the word "Butter" (*butter*), which is used as masculine in certain German-speaking regions, else bearing the feminine gender, or, less frequently, neutral gender. Some brand names used in everyday context, like 'Nutella", can be used with the three genders, depending on the region. The same is valid for certain loan words, like "Joghurt". It would be interesting to keep track of this national or regional differences, when translating from another language into a national or regional version of German. But for now we are focusing on German nouns having a different meaning related to their grammatical gender or number, similar to the heteronyms we discussed in Section 2.

German has three different grammatical genders: masculine, feminine, and neutral. When the grammatical gender of a word changes, so does the flexion of its article and adjectives in the sentence, which are a (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.

A German example of grammatical gender ambiguity is given with the lexical entry "Plastik", which can bear the sense of an object of art (*sculpture*), when used with the feminine grammatical gender, or the chemical sense (*plastic*), when used with the masculine or neutral grammatical gender. This example is interesting, as the second sense is given only for the use of the word in singular. We have thus also a case of number ambiguity, as the plural use of the word in a text should not be associated with the chemical sense. This is not trivial, as contextual markers for Gender, like determiners, can have exactly the same surface realisation, depending on the grammatical case in which they occur.

[6]See respectively https://de.wiktionary.org/w/index.php?title=Kategorie:Substantiv_zwei_Genera_(Deutsch), https://de.wiktionary.org/w/index.php?title=Kategorie:Substantiv_drei_Genera_(Deutsch) and https://en.wiktionary.org/wiki/Category:German_nouns_with_multiple_genders.

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. An example in English is given by the word *glasses*, which can be either the regular plural form of the word *glass* or, used only in plural form, mean *Spectacles, frames bearing two lenses worn in front of the eyes*.[7] While both meanings of *glasses* appear to be etymologically related, they do mean different things. A similar example is given by the pair silk/silks.[8]

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*).

We extracted from the German Wiktionary two datasets including those grammatical gender and number ambiguities and made them availabe on Github.[9] As for the work dealing with the extraction of pronunciation information, 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 this dataset 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.

4. Related Work

It has been shown that the access and use of Wiktionary data can be helpful in a series of applications. (Kirov et al., 2016), for example, describes work to extract and standardize the data in Wiktionary and to make it available for a range of NLP applications, while the authors focus on extracting and normalizing a huge number of inflectional paradigms across a large selection of languages. This effort contributed to the creation of the UniMorph data (`http://unimorph.org/`). The UniMorph project was focusing on (scraping) the HTML representation of Wiktionary (mostly the English version, but also looking at other language editions). (Metheniti and Neumann, 2020) describe a related approach, but making use of a combination of the HTML pages and the underlying XML dump of the English edition of Wiktionary, which is covering also 4,293 other languages, but some of them with a very low number of entries.

(Hanoka and Sagot, 2014) details how lexical data included in Wiktionary can be used for supporting the creation of a multilingual translation graph, extracting information from 21 language specific Wiktionaries (and from OPUS parallel data).

(Sajous et al., 2020) describes a more general approach as ours, but limited to English and with a focus on building a machine-readable dictionary, while we are also aiming at generating corpora from Wiktionary data, with the aim of supporting specific tasks, like detecting ambiguities in text.

(Hellmann et al., 2012) and (Sérasset, 2015) present approaches consisting in transforming multilingual Wiktionaries into a linked data compliant format. While we are more concerned in generating corpora from Wiktionary or extending the coverage of existing lexical resources with elements taken from Wiktionary, it will be interesting to see if we can achieve similar results extracting information from such linked data compliant versions of Wiktionary, as this framework is offering a consistent representation of lexical information across the various languages, harmonizing thus the data included in the original Wiktionaries.

Wiktionary is often used as a source for various text-to-speech or speech-to-text models. For instance, the work of (Schlippe et al., 2010) developed a system which automatically extracts phonetic notations in IPA from Wiktionary to use for automatic speech recognition. This work also assess the quality of pronunciation information in Wiktionary for four languages (English, French, German, and Spanish) and come to satisfying results, especially in the case of French, when it comes to the evaluation of the coverage and also to the impact on automatic speech recognition (ASR) systems,

A more recent example is the work of (Peters et al., 2017) which is aimed at improving grapheme-to-phoneme conversion by utilizing Wiktionary. Grapheme-to-phoneme is necessary for text-to-speech and automatic speech recognition systems

Besides text-to-speech, there are various other applications which rely on extracting pronunciation information from Wiktionary. A very recent tool is WikiPron (Lee et al., 2020). WikiPron is an open-source command-line tool for extracting pronunciation data from Wiktionary. It stores the extracted pronunciation information database, which at the date of the publication contains 1.7 million pronunciations from 165 languages. Results of our work could be included directly or via Wiktionary updates into this database.

In general, dataset extraction or generation is rather done on the data available on Wikipedia, and we do not find a lot of works dealing with (lexical) dataset generation using Wiktionary, where we can use the definitions and the examples sentences associated with an entry.

5. Conclusion and Future work

We presented ongoing work in generating specialized datasets from Wiktionary lexical resources. A result

[7] Example taken from `https://en.wiktionary.org/wiki/glasses`

[8] Those cases of grammatical number ambiguities in English are discussed in depth in (Gromann and Declerck, 2019)

[9] `https://github.com/Declerck/LexicalDatasetforGenderNumberAmbiguities`.

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.

6. 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). The article is also supported by the Horizon 2020 research and innovation programme with the projects Prêt-à-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 would also like to thank the anonymous reviewers for their helpful comments.

7. Bibliographical References

Bajčetić, L., Declerck, T., and McCrae, J. (2021). Heteronym sense linking. In *Proceedings of the eLex Conference on 'post-editing lexicography'*. eLex.

Declerck, T. and Bajčetić, L. (2021). Towards the addition of pronunciation information to lexical semantic resources. In *Proceedings of the 11th Global Wordnet Conference*, pages 284–291, University of South Africa (UNISA), January. Global Wordnet Association.

Gromann, D. and Declerck, T. (2019). Towards the detection and formal representation of semantic shifts in inflectional morphology. In Maria Eskevich, et al., editors, *2nd Conference on Language, Data and Knowledge (LDK)*, volume 70 of *OpenAccess Series in Informatics (OASIcs)*, pages 21:1–21:15. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 5.

Hanoka, V. and Sagot, B. (2014). An open-source heavily multilingual translation graph extracted from wiktionaries and parallel corpora. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3179–3186, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Hellmann, S., Brekle, J., and Auer, S. (2012). Leveraging the crowdsourcing of lexical resources for bootstrapping a linguistic data cloud. In *JIST*.

Kirov, C., Sylak-Glassman, J., Que, R., and Yarowsky, D. (2016). Very-large scale parsing and normalization of wiktionary morphological paradigms. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Lee, J. L., Ashby, L. F., Garza, M. E., Lee-Sikka, Y., Miller, S., Wong, A., McCarthy, A. D., and Gorman, K. (2020). Massively multilingual pronunciation modeling with WikiPron. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4223–4228, Marseille, France, May. European Language Resources Association.

McCrae, J. P., Rademaker, A., Rudnicka, E., and Bond, F. (2020). English WordNet 2020: Improving and extending a WordNet for English using an open-source methodology. In *Proceedings of the LREC 2020 Workshop on Multimodal Wordnets (MMW2020)*, pages 14–19, Marseille, France, May. The European Language Resources Association (ELRA).

Metheniti, E. and Neumann, G. (2020). Wikinflection corpus: A (better) multilingual, morpheme-annotated inflectional corpus. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*. LREC.

Peters, B., Dehdari, J., and van Genabith, J. (2017). Massively multilingual neural grapheme-to-phoneme conversion. *CoRR*, abs/1708.01464.

Sajous, F., Calderone, B., and Hathout, N. (2020). ENGLAWI: From human- to machine-readable Wiktionary. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3016–3026, Marseille, France, May. European Language Resources Association.

Schlippe, T., Ochs, S., and Schultz, T. (2010). Wiktionary as a source for automatic pronunciation extraction. In *11th Annual Conference of the International Speech Communication Association, Makuhari, Japan*. Interspeech 2010.

Sérasset, G. (2015). Dbnary: Wiktionary as a lemon-based multilingual lexical resource in RDF. *Semantic Web*, 6(4):355–361.

Wikimedia. (2021a). Wikimedia foundation. `https://wikimediafoundation.org/`, 01.

Wikimedia. (2021b). Wiktionary. `https://www.wiktionary.org/`, 07.