New features in DEBVisDic for WordNet Visualization and User Feedback

Adam Rambousek, Aleš Horák, David Klement, Jiří Kletečka

Natural Language Processing Centre
Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
deb@aurora.fi.muni.cz
http://deb.fi.muni.cz

Abstract. This paper presents two new features that help with wordnet management and presentation in DEBVisDic. The first is the new interface to gather user feedback about errors in wordnet and streamlined management of revisions approval and possible updates to the wordnet database. The second feature is the new visualization interface, providing both textual and graphical representation of wordnet data, with emphasis on user-friendly and responsive design. New visualization interface will be included in the DEBVisDic editor and also published as a stand-alone web application.

Key words: DEB platform, crowdsourcing, wordnet, visualization, error checking

1 Introduction

Projects to build a large ontology or semantic network usually do not include capacities needed for long-term management and updates. Although there are various automatic ontology consistency checkers [1,2,3], manual reviews are always the most reliable method for database fixes and updates. However the process is time consuming and cannot be completely finished by a small group of linguists. General audience often discovers mistakes in the published version, it is thus very useful to take user feedback into consideration.

In the following text, we present a new interface for error checking and reporting in wordnets. The interface is developed with the idea of crowdsourcing by wide public which should speed up the process of errors disovery and correction. The tool is developed withing the DEB (Dictionary Editor and Browser [4,5]) framework and connected to the backend database of DEBVisDic used for developing a number of national wordnets.

For users, simple and user-friendly method for discovering wordnet data is an important aspect. After reviewing existing visualization tools, we have decided to develop new text and graph based interface within the DEBVisDic tool, which we believe offers a best solution to combined user needs in the network exploration process.

Aleš Horák, Pavel Rychlý, Adam Rambousek (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2017, pp. 3–10, 2017.

2 Wordnet Development and Issues

Issues in wordnet may be divided to two main categories:

- surface errors – issues with synset description, e.g. spelling errors in literals or definitions,
- structural errors – issues with semantic relations, appropriate literal selection, varying subtrees depth and granularity, or orphaned synsets.

Two general methodologies defined during the EuroWordNet project [6] are usually used to build new wordnets:

- *Expand model* – with this approach, Princeton WordNet (or its part) is translated to a new language, keeping the semantic relations mostly intact. Some projects translated the synsets semi-automatically, which may introduce surface errors if the results are not verified properly.
- *Merge model* – new wordnet is created either from scratch, or based on existing dictionary, which does not contain semantic relations and entries are not grouped to synsets. Wordnets utilizing this method contain more structural errors.

Many of the errors may be prevented during the wordnet development phase. Important part is to design and follow detailed guidelines [7,2]. Software tools may help significantly. Wordnet editing software should check for a range of errors, from spellchecking to semantic relations completeness [8]. Some projects also use periodical heuristic testing to check recently added or updated synsets [9].

3 Crowdsourcing in Linguistics

In linguistics and NLP research, crowdsourcing is usually used to manually annotate large datasets with semantic or syntactic information [10], word sense disambiguation [11], or to evaluate the results of automatic tools [12], but may even help to detect epidemics outbreak [13].

The results of crowdsourcing experiments in NLP research were evaluated multiple times, concluding that combining annotation by several "unskilled" annotators may result in cheaper and faster annotation. Study by [14] concluded that on average 4 non-expert annotations achieve the equivalent inter-annotator agreement as a single expert. Another experiment [15] evaluated machine translation using crowdsourcing and concluded that combination of many non-expert evaluations provides equivalent quality as experts.

In the field of lexicography, Wiktionary,[1] a sister project of Wikipedia, is one of the most prominent crowdsourced resource. The goal of Wiktionary is to create a freely available “dictionary of all words in all languages” [16] edited by volunteers. Several analysis [17,18,19] found Wiktionary to be a useful linguistic resource, however, entry quality varies from well-crafted to unreliable.

[1] `http://www.wiktionary.org/`

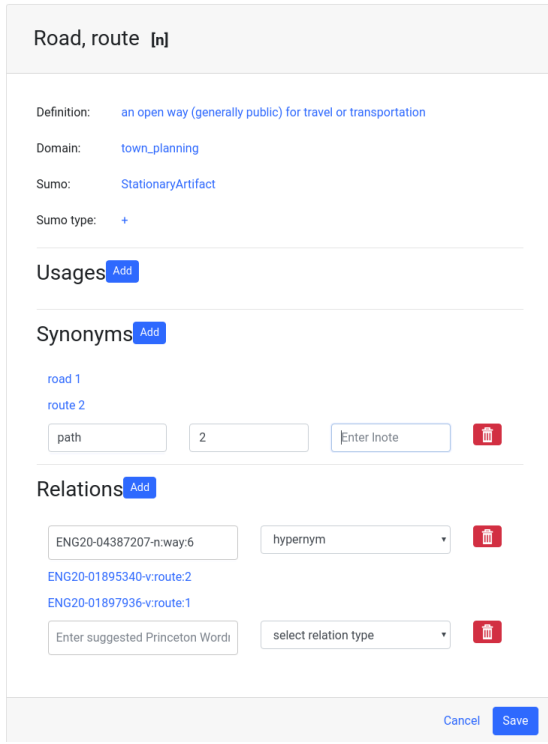

Fig. 1. User feedback form to provide synset data suggestions.

4 Crowdsourcing Tool and Review Process

Czech WordNet (CzWN) was published as a part of the EuroWordNet and Balkanet projects [6,20] and since then CzWN was mostly just maintained. However, there are several versions with various amount of edits, together with version semi-automatically extended using English-Czech translation dictionary [21]. NLP Centre (the CzWN developer) is currently running a project to integrate all updates to Czech WordNet and publish new Open Czech WordNet linked to Collaborative Interlingual Index [22].

Czech WordNet was developed using the expand model, translating the English wordnet synsets. Most notable example of errors caused by this approach are the synsets containing words that are not exactly synonyms, or only rare in the Czech language, but present in the Czech WordNet because of the translation from English. For example, the English synset *cabriolet:1, cab:2* has the equivalent Czech synset *kabriolet:2, dvoukolový jednospřežní povoz:1, koňská drožka:1* (cabriolet, two-wheeled one horse cart, horse-drawn carriage). Although the translation is correct, this sense of *kabriolet* in Czech is very archaic, in current language the only sense used in spoken language is the convertible car. Another problem is

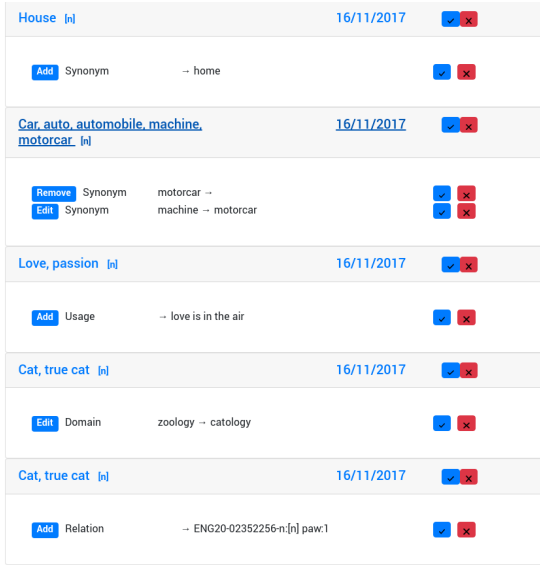

Fig. 2. Administrator view of suggested changes.

the inclusion of multiword expressions in the synset, which may be justified in some cases, but these are not fixed lexical units in the Czech language.

However, during the integration we will not have enough resources and lexicographers to check all synsets and relations in the Czech WordNet. We have developed new software tool that allows any wordnet user to report issues they spot in the data. Although we are testing the tool on the Czech WordNet, it is language-independent and available for all wordnets developed using the DEBVisDic editor.

The tool is not directly integrated into DEBVisDic editor, but rather uses the DEBVisDic server API to access wordnet data. On the other hand, all available synset representation (editor, simplified browser, API call) will enable users to easily move to the error reporting application. Users are presented with the data from the synset they were browsing and may update any data value – change existing value, add a new one if some part of synset is missing, or remove an unwanted value. See Figure 1 for an example of the user feedback form. Updates are stored in a separate database as suggestions. Each value (e.g. gloss or relation) is stored as a single suggestion.

Any member of the editing team with access permissions to the given wordnet may browse all user suggestions (or filter them by reporting user, information type, or review status). The editor may approve or reject any single suggestion, or approve/reject all suggestions for any synset at once. Of course, it is also possible to approve/reject all suggestions based on the selected filter. Before deciding, the editor may compare user feedback with previously

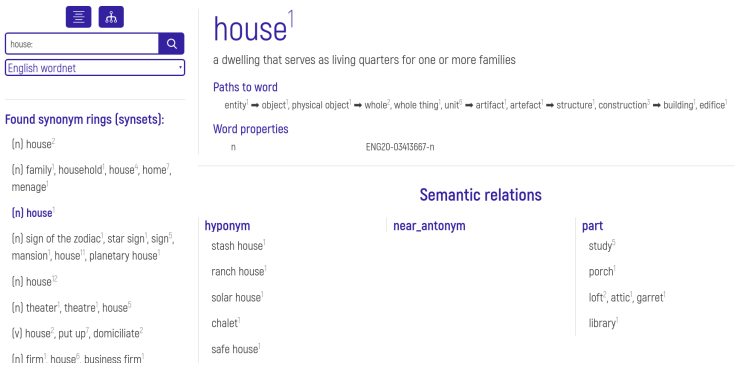

Fig. 3. DEBVisDic synset information in the text mode.

approved or rejected updates for the selected synset. See Figure 2 for preview of editor's interface.

All approved suggestions are immediately transferred to the development version of the wordnet database and presented to users. When a user's feeedback is rejected by the editor, the information is kept in the database and future users trying to suggest the same update are notified about the previous refusal.

5 New DEBVisDic Visualization Interface

The goal of the new DEBVisDic wordnet interface is to facilitate visualization of wordnets in both textual and graphical forms to the widest possible audience. The tool aims to provide access to wordnets in a platform independent way so that user is not bound to use only e.g. desktop computer or a mobile phone to access the network data. This was achieved by developing the tool as a web application which allows user to utilize it on any device that is equipped by a web browser. For good usability on any device, the tool needs to be responsive and to adapt itself to any reasonable size of a user's screen. The responsive design goes in hand with the other goal of the tool which is a visually appealing and modern style. This is essential when using the application for educational purposes as wordnets offer a rich basis for language trainings of both children and adults considering the correspondence of the semantic network structure with the presumed human brain organization. Another goal achieved by developing the tool as a web application was its broad accessibility. With the only requirement of having a web browser installed, the tool can be accessed on almost any device and there is no need for complicate preparation of e.g. school computers for the usage in classes.

From the technical point of view, the new DEBVisDic interface is an HTML document partly generated on the client side by JavaScript according to the data sent by the server in the JSON format. The look of the document is defined by

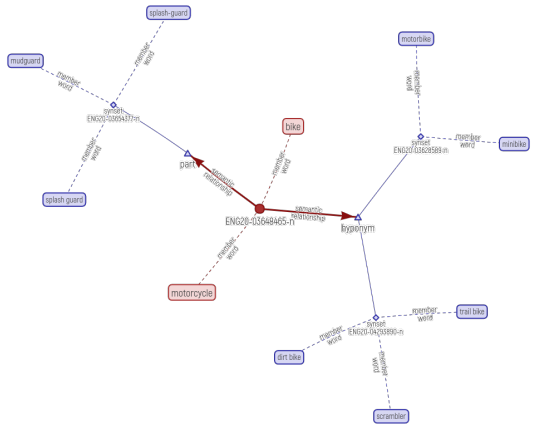

Fig. 4. Graph representation of synset relations in DEBVisDic.

CSS, the content being a result of what was sent by the server. The server sends synsets which contain a searched word or have an ID identical to the one that was requested by the client. The server response contains all the synsets that comply with the searched term and the interface displays by default the first one that was in data and the rest is offered to the user to choose in a sidebar menu on the left (or on the top if the screen is too small for the sidebar). The right (or bottom) part of the page then contains either textual or graph representation of the presented synset. The two views can be switched by two buttons on the top of the sidebar.

In the text mode, the interface displays all available information i.e. the synonymic set contained in the synset (the literals), the hypero-hyponymic path leading to the synset and last but not least the synset semantic relations. These are displayed as a column for each relation with the connected synsets. See Figure 3 for an example of the synset information in text mode.

The graph mode is an alternative representation of relations to which the synset belongs. The central node of the graph is the displayed synset with edges leading to its literals and then to nodes representing all connected semantic relations. An example of the graph-based synset visualization can be found in Figure 4.

6 Conclusion

We have presented a new tool for crowdsourcing reporting of wordnet errors. The process workflow takes into account all the needed phases of lexical database updates and enhancement. After thorough public testing with the

Czech Wordnet, we plan to release the tool for all wordnets developed within the DEB platform.

To present wordnet data in a visually attractive and understandable way for users, we have developed new DEBVisDic Visualization Interface providing both textual and graphic mode for synset preview. This interface is both integrated with the DEBVisDic editor, and as a standalone web application.

Acknowledgements. This work was partly supported by the Ministry of Education of CR within the LINDAT-Clarin project LM2015071 and by the project of specific research *Čeština v jednotě synchronie a diachronie* (Czech language in unity of synchrony and diachrony; project no. MUNI/A/0915/2016).

References

1. Alvez, J., Atserias, J., Carrera, J., Climent, S., Laparra, E., Oliver, A., Rigau, G.: Complete and consistent annotation of wordnet using the top concept ontology. In: Proceedings of LREC 2008. (2008)
2. Tufiş, D., Cristea, D.: Methodological issues in building the romanian wordnet and consistency checks in balkanet. In: Proceedings of LREC2002 Workshop on Wordnet Structures and Standardisation. (2002) 35–41
3. Rath, H.H.: Technical issues on topic maps. In: Proceedings of Metastructures 99 Conference, GCA (1999)
4. Rambousek, A., Horák, A.: DEBVisDic: Instant WordNet Building. In: Proceedings of the Eighth Global WordNet Conference, GWC 2016. (2016) 25–29
5. Horák, A., Vossen, P., Rambousek, A.: A Distributed Database System for Developing Ontological and Lexical Resources in Harmony. In: Lecture Notes in Computer Science: Computational Linguistics and Intelligent Text Processing, Haifa, Israel, Springer-Verlag (2008) 1–15
6. Vossen, P., ed.: EuroWordNet: a multilingual database with lexical semantic networks for European Languages. Kluwer (1998)
7. Pociello, E., Agirre, E., Aldezabal, I.: Methodology and construction of the basque wordnet. Language Resources and Evaluation **45**(2) (May 2011) 121–142
8. Horák, A., Pala, K., Rambousek, A., Povolný, M.: DEBVisDic – First Version of New Client-Server Wordnet Browsing and Editing Tool. In: Proceedings of the Third International WordNet Conference - GWC 2006, Jeju, South Korea, Masaryk University, Brno (2006) 325–328
9. Čapek, T.: SENEQA - System for Quality Testing of Wordnet Data. In Fellbaum, C., Vossen, P., eds.: 6th International Global Wordnet Conference Proceedings, Matsue, Japan, Toyohashi University of Technology (2012) 400–404
10. Grác, M.: Rapid Development of Language Resources. PhD thesis, Faculty of Informatics, Masaryk University (2013)
11. Rumshisky, A.: Crowdsourcing word sense definition. In: Proceedings of the 5th Linguistic Annotation Workshop, Association for Computational Linguistics (2011) 74–81
12. Nevěřilová, Z.: Paraphrase and Textual Entailment Generation in Czech. PhD thesis, Faculty of Informatics, Masaryk University (2014)
13. Munro, R., Gunasekara, L., Nevins, S., Polepeddi, L., Rosen, E.: Tracking Epidemics with Natural Language Processing and Crowdsourcing. In: AAAI Spring Symposium: Wisdom of the Crowd. Volume SS-12-06., AAAI (2012)

14. Snow, R., O'Connor, B., Jurafsky, D., Ng, A.Y.: Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In: Proceedings of the conference on empirical methods in natural language processing, Association for Computational Linguistics (2008) 254–263
15. Callison-Burch, C.: Fast, cheap, and creative: evaluating translation quality using Amazon's Mechanical Turk. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1, Association for Computational Linguistics (2009) 286–295
16. Wikipedia: Wiktionary — Wikipedia, The Free Encyclopedia (2017) [Online; accessed 2-October-2017].
17. Hanks, P.: Corpus evidence and electronic lexicography. In Granger, S., Paquot, M., eds.: Electronic Lexicography. Oxford University Press, Oxford (2012) 57–82
18. Meyer, C.M., Gurevych, I.: Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography. In Granger, S., Paquot, M., eds.: Electronic Lexicography. Oxford University Press, Oxford (2012) 259–291
19. Fuertes-Olivera, P.A.: The Function Theory of Lexicography and Electronic Dictionaries: WIKTIONARY as a Prototype of Collective Free Multiple-Language Internet Dictionary. In Bergenholtz, H., Nielsen, S., Tarp, S., eds.: Lexicography at a Crossroads. Peter Lang, Bern (2009) 103–120
20. Christodoulakis, D.: Balkanet Final Report, University of Patras, DBLAB (2004) No. IST-2000-29388.
21. Blahuš, M., Pala, K.: Extending Czech WordNet using a bilingual dictionary. In Fellbaum, C., Vossen, P., eds.: 6th International Global Wordnet Conference Proceedings, Matsue, Japan, Toyohashi University of Technology (2012) 50–55
22. Bond, F., Vossen, P., McCrae, J.P., Fellbaum, C.: CILI: the Collaborative Interlingual Index. In Barbu Mititelu, V., Forascu, C., Fellbaum, C., Vossen, P., eds.: Proceedings of the Eighth Global WordNet Conference, Romanian Academy (2016) 50–57