Where the journey is headed:

Collaboratively constructed multilingual Wiki-based resources

Michael Matuschek and Iryna Gurevych
Technische Universität Darmstadt
Hochschulstraße 10
64289 Darmstadt, Germany
http://www.ukp.tu-darmstadt.de

Abstract

Multilingual lexical-semantic resources play an important role in computational linguistics, e.g. in cross-lingual information retrieval or machine translation. However, multilingual resources with sufficient quality and coverage are rare, as the effort of manually constructing such a resource is substantial. In recent years, the emergence of Web 2.0 has opened new possibilities for handling the effort of constructing large scale lexical-semantic resources. We identified Wiktionary and OmegaWiki as two important multilingual initiatives where a community of users („crowd“) collaboratively edits and refines the lexical information. We argue that collaborative construction is a promising approach to cope with the enormous effort of building such resources. It seems especially appropriate in the multilingual domain as users from all languages and cultures can easily contribute. However, despite their advantages such as open access and coverage of multiple languages, these resources have hardly been systematically investigated until now. Therefore, the goal of our contribution is two-fold: First, we focus on two promising multilingual resources containing lexical-semantic information. To this end, we analyze the way they emerged and characterize the resulting content. Second, we propose how a collaboratively constructed multilingual resource should be designed in order to be maximally useful for text analysis.

Keywords: lexical semantic resources, collaboratively constructed resources, multilingual resources

1. Introduction

Multilingual lexical-semantic resources play an important role in computational linguistics, e.g. in cross-lingual information retrieval (Herbert, Szarvas, & Gurevych, 2011) or machine translation (Okuma, Yamamoto, & Sumita, 2007). However, multilingual resources with sufficient quality and coverage are rare, as the effort of manually constructing a monolingual resource is already substantial, and it is naturally even greater if multiple languages are involved. Thus, past research focused on the construction of monolingual resources and aligning them in some appropriate way (Fellbaum, 2010). However, while monolingual resources such as WordNet (Fellbaum, 1998) or GermaNet (Hamp & Feldweg, 1997) have reached substantial size and acceptance in the research community, the efforts to align them (most prominently in the EuroWordNet project (Vossen, 1998)) struggle from problems such as limited size, insufficient overlap and structural differences between resources in different languages (Miháltz, et al., 2008). Licensing issues also sometimes impair their usefulness.

In recent years, the emergence of Web 2.0 has opened new possibilities for handling the effort of constructing large scale lexical-semantic resources. The most prominent example is Wikipedia[1], which can also be used as a resource in computational linguistics (Zesch, Gurevych, & Mühlhäuser, 2007). In this work, we focus on resources such as dictionaries which explicitly encode linguistic (as opposed to encyclopedic) knowledge, and most importantly multilingual information such as translations. This yields direct exploitation of this knowledge in text analysis. We identified Wiktionary[2] and OmegaWiki [3] as two important multilingual initiatives where a community of users („crowd“) collaboratively edits and refines the lexical information. This information is in turn free to use for everyone. We argue that collaborative construction is the most

[1] http://www.wikipedia.org

[2] http://www.wiktionary.org

[3] http://www.omegawiki.org

promising approach to cope with the enormous effort of building such resources. It seems especially appropriate for a multilingual resource as users speaking any language and from any culture can easily contribute. This is crucial for minor, usually resource-poor languages. However, despite their advantages such as open access and coverage of multiple languages, these resources have hardly been systematically investigated until now. This makes it hard to fully understand their characteristics and in turn exploit them for text analysis purposes. Therefore, the goal of our contribution is two-fold: First, we focus on two promising multilingual resources containing lexical-semantic information. To this end, we analyze the way they emerged and characterize the resulting content. Second, we propose how a collaboratively constructed multilingual resource should be designed in order to be maximally useful for text analysis.

2. Wiktionary

Wiktionary is a side project of Wikipedia. It is also based on the Wiki principle so that users are free to add and edit entries. There are templates which recommend how these entries should be structured, but these can be changed and extended according to the user's needs. This freedom (paired with the popularity of the "mother project" Wikipedia) has led to a substantial number of entries in various languages (see table 1). It has already successfully been used for different purposes (Zesch, Müller, & Gurevych, 2008; Navarro, et al., 2009; Meyer & Gurevych, 2010). It is especially interesting for multilingual applications (Mausam et al., 2008) as many language versions of Wiktionary contain foreign language entries (e.g. a German explanation for an English word) as well as translations and links to the corresponding entry in other languages.

Wiktionary has been primarily designed to be used by human readers, which is problematic for its use as a machine readable resource. The Wiki markup used does not follow common standards for the encoding of resources such as the Lexical Markup Framework (LMF) (Francopoulo, et al., 2009). Thus, the representation of information is not always explicit (Meyer & Gurevych, 2010). E.g., synonymy links are not sense disambiguated, they only point to the lexicon entry. While the disambiguation is usually not a problem for humans, it is a considerable obstacle for machines. Furthermore, the coding of symmetric relations is often incomplete, as all links have to be set manually. Deviations from the templates and encoding mistakes made by editors introduce some errors into the parsing process. The templates and guidelines for entries in different language versions are also different which requires an adaptation of the parser for each language. Even worse, templates may change over time, which again requires adjusting the parser.

	WKT	**OW**
Entries (Total)	14,021,155	442,723
Entries (English)	2,457,506	55,182
Languages	>400	290
Languages with >10.000 entries	54	12
Information storing	Wiki Markup/XML	Relational DB
Inconsistencies & encoding errors	Yes	No

Table 1: Descriptive statistics about Wiktionary (WKT) and OmegaWiki (OW) as of May 2011. Further statistics about these resources are to be found at http://www.ukp.tu-darmstadt.de/data/lexical-resources/

3. OmegaWiki

OmegaWiki relies on a fundamentally different concept of storing the information than Wiktionary. To avoid Wiktionary's problems caused by free editing of entries, OmegaWiki is based on a fixed database structure which users have to stick to[4]. Central to this are language-independent concepts to which lexicalizations of the concepts are attached. This directly yields unambiguous translations (E.g., concept no. 5616 carries the labels "hand", "main", "mano" etc. and also glosses in different languages which describe the concept such as "That part of the fore limb below the forearm or wrist"). Another useful consequence of the design for multilingual applications is that relations are unambiguously defined between concepts regardless of existing lexicalizations, which is useful for tasks such as cross-lingual semantic relatedness (e.g., "dedo" is marked as hypernym of "finger" and "toe" although there exists no corresponding term in English). Exploiting this kind of information is

[4] www.omegawiki.org/Help:OmegaWiki

not as easy in resources like EuroWordNet where concepts are linked across languages but the respective taxonomies are different or even contradictory (Jansen, 2004).

Although the fixed structure of OmegaWiki is proprietary and does not conform to standards such as LMF, it is definitely easier to utilize this resource in text analysis than Wiktionary as the underlying database ensures straightforward structured extraction of the information and less error-prone results, as opposed to Wiktionary. However, it also has some disadvantages due to limited expressiveness. As an example, the coding of grammatical properties is only possible to a small extent. Moreover, an extension of this structure is not easy, as this would require all present and future entries to conform to the new structure, other than for Wiktionary entries, whose structure is more flexible. Consequently, ordinary users are not allowed to extend the structure of OmegaWiki and thus are tied to what has been defined. This lack of flexibility, in combination with the fact that Wiktionary was already quite popular at its creation time, has caused OmegaWiki to remain rather small (see table 1). Nevertheless, it is useful in terms of research because it exemplifies how the process of creating such a resource by ordinary users can be moderated to yield a machine readable result.

4. Conclusions and future work

Our analysis shows that collaborative editing presents a viable solution for creating large-scale multilingual lexical-semantic resources. However, both presented approaches have their disadvantages: While the open approach of Wiktionary has attracted many users, leading to a resource of considerable size and richness, it also leads to difficulties in the exploitation for text analysis. OmegaWiki, on the other hand, does not suffer from this problem, but the self-imposed limitations to maintain integrity also constrain its expressiveness and, along with that, the range of information which can be represented in the resource. Thus, our conclusion is that a collaboratively constructed multilingual lexical-semantic resource has to strike a balance between these two approaches, i.e. moderate the work of the crowd without imposing overly narrow restrictions. This conclusion yields the following requirements for such a resource:

1) Fixed structure: The structure of entries must be unchangeable by users to avoid inconsistencies in encoding, and this structure must be supported by a corresponding database or XML schema which ensures the consistency validation.
2) Elaborate structure: The structure must be elaborate and expressive enough to cater for a wide range of lexical-semantic information in different languages, as found in other machine readable resources. This requires an exhaustive preparatory analysis of existing resources by language experts.
3) Interoperability: The resource should be in a format which is not only machine readable but also compliant to existing standards to allow for easy reuse and integration into applications. Our proposal is to model the resource in LMF, as this is a recognized, expressive standard for lexical resources and allows for easy storage in XML or a database.

As an additional requirement, the interface should be designed in a way that the complexity of the resource does not impair understandability or usability, so that ordinary users are able to contribute easily. A good example how this can be achieved is the WISIGOTH add-on for Wiktionary (Navarro et al., 2009).

Consequently, the current focus of our work is to design a representation format for a multilingual lexical-semantic resource based on LMF which we plan to make available to the public in the near future. As the effort of filling such a resource from scratch is prohibitive, we work on the mapping of the existing resources to the LMF model and the tools to import and merge multilingual data from Wiktionary, OmegaWiki and other resources. We plan to make the mappings to the LMF format available for off-the-shelf usage too, if the licensing of resources permits that. In parallel, we work on an accompanying API and web interface. A primary design goal of the overall project is the interoperability of a wide range of lexical-semantic resources, including the widely used ones such as FrameNet (Baker & Fillmore, 2010) or WordNet, based on the LMF standard and on the sense alignment between resources. To this end, we also investigate how different word senses can be aligned across resources and languages automatically (Niemann & Gurevych, 2011; Meyer & Gurevych, 2011). The resulting meta-resource will be evaluated in multilingual text analysis.

5. Acknowledgements

This work has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant No. I/82806 and by the Hessian research excellence program “Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz” (LOEWE) as part of the research center “Digital Humanities”**.** We would also like to thank Judith Eckle-Kohler and Christian M. Meyer for valuable comments and discussions.

6. References

Baker, C. F., & Fillmore, C. J. (2010). A Frame Approach to Semantic Analysis. In B. Heine, & H. Narrog, Oxford Handbook of Linguistic Analysis. Oxford, UK

Fellbaum, C. (2010). Translating with a Semantic Net. In B. Lewandowska-Tomaszczyk, & M. Thelen, Meaning in Translation (pp. 255 - 265). Lodz.

Fellbaum, C. (1998). WordNet: An Electronic Lexical Database. Cambridge, MA

Francopoulo, G., Bel, N., George, M., Calzolari, N., Monachini, M., Pet, M., Soria, C. (2009). Multilingual Resources for NLP in the Lexical Markup Framework. Language Resources and Evaluation, Vol. 43, No. 1, pp. 57-70.

Hamp, B., & Feldweg, H. (1997). GermaNet - A Lexical-Semantic Net For German. Proceedings of the Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications, pp. 9-15.

Herbert, B., Szarvas, G., & Gurevych, I. (2011). Combining Query Translation Techniques to Improve Cross-Language Information Retrieval. Proceedings of the 33rd European Conference on Information Retrieval (pp. 712-715). Dublin, Republic of Ireland

Jansen, P. (2004, 3). Lexicography in an Interlingual Ontology. Canadian Undergraduate Journal of Cognitive Science, pp. 1-5.

Mausam, Soderland, S., Etzioni, O., Weld, D. S., Skinner, M., & Bilmes, J. (2008). Compiling a Massive, Multilingual Dictionary via Probabilistic Inference. Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1 (pp. 262-270). Suntec, Singapore

Meyer, C. M., & Gurevych, I. (2010). How Web Communities Analyze Human Language: Word Senses in Wiktionary. Proceedings of the Second Web Science Conference. Raleigh, CA.

Meyer, C. M., & Gurevych, I. (2010). Worth its Weight in Gold or Yet Another Resource — A Comparative Study of Wiktionary, OpenThesaurus and GermaNet. Proceedings of the 11th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing) (pp. 38-49). Iaşi, Romania.

Meyer, C. M., & Gurevych, I. (2011). What Psycholinguists Know About Chemistry: Aligning Wiktionary and WordNet for Increased Domain Coverage. Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP), (to appear), November 2011. Chiang Mai, Thailand.

Miháltz, M., Hatvani, C., Kuti, J., Szarvas, G., Csirik, J., Prózeky, G., Váradi, T. (2008). Methods and Results of the Hungarian WordNet Project. Proceedings of the Fourth Global Wordnet Conference. Szeged, Hungary.

Navarro, E., Sajous, F., Gaume, B., Prévot, L., ShuKai, H., Tzu-Yi, K., Magistry, P., Huang, C. R. (2009). Wiktionary and NLP: Improving Synonymy Networks. Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources (pp. 19-27). Suntec, Singapore

Niemann, E. & Gurevych, I. (2011). The People’s Web meets Linguistic Knowledge: Automatic Sense Alignment of Wikipedia and WordNet. Proceedings of the 9th International Conference on Computational Semantics (pp. 205-214). Oxford, UK

Okuma, H., Yamamoto, H., & Sumita, E. (2007). Introducing Translation Dictionary into Phrase-based SMT. Proceedings of MT Summit XI , pp. 361-368.

Vossen, P. (1998). EuroWordNet: A Multilingual Database for Information Retrieval. Norwell, USA

Zesch, T., Gurevych, I., & Mühlhäuser, M. (2007). Analyzing and Accessing Wikipedia as a Lexical Semantic Resource. In G. Narr, Data Structures for Linguistic Resources and Applications (pp. 197-205).

Zesch, T., Müller, C., & Gurevych, I. (2008). Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. Proceedings of the 6th International Conference on Language Resources and Evaluation. Marrakech, Morocco.