OMWEdit - The Integrated Open Multilingual Wordnet Editing System

Luís Morgado da Costa and Francis Bond
Linguistics and Mulitingual Studies
Nanyang Technological University
Singapore
{luis.passos.morgado@gmail.com, bond@ieee.org}

Abstract

Wordnets play a central role in many natural language processing tasks. This paper introduces a multilingual editing system for the Open Multilingual Wordnet (OMW: Bond and Foster, 2013). Wordnet development, like most lexicographic tasks, is slow and expensive. Moving away from the original Princeton Wordnet (Fellbaum, 1998) development workflow, wordnet creation and expansion has increasingly been shifting towards an automated and/or interactive system facilitated task. In the particular case of human edition/expansion of wordnets, a few systems have been developed to aid the lexicographers' work. Unfortunately, most of these tools have either restricted licenses, or have been designed with a particular language in mind. We present a web-based system that is capable of multilingual browsing and editing for any of the hundreds of languages made available by the OMW. All tools and guidelines are freely available under an open license.

1 Introduction

Lexical semantic resources, such as wordnets (WNs), play a central role in many Natural Language Processing (NLP) tasks. Word Sense Disambiguation (WSD), for example, relies heavily on existing lexical semantic resources. Likewise, many other unsolved problems of NLP (e.g. Machine Translation, Q&A Systems) rely on WSD and, consequently, indirectly, rely also on the existence of resources like WNs.

This explains why substantial resources have been employed in the development of high quality lexical semantic resources. The Princeton Wordnet (PWN: Fellbaum, 1998) pioneered the development of such a resource for English. Following its steps, many other projects followed PWN into building similar resources for different languages.

The lexicographic work-flow for these early projects included hand-typing linked complex data structures in electronic text files. The result was a huge net of concepts, senses and definitions linked through a variety of relations. This kind of work is ultimately very time consuming and prone to mistakes. The direct manipulation of text files makes it extremely easy to unintentionally violate the data syntax. In recent times, the creation and expansion of these resources has been increasingly shifting into an automated and/or interactive system facilitated task. Simple and intuitive user interfaces can help to both speed-up and to immediately check for inconsistencies in the data (e.g. relatedness to nonexistent keys, reduplicated information, typos or omission of minimal required information). Using modern relational databases and web-serviced interactive platforms has also allowed for remote and parallel collaboration, as well as effective journaling systems.

As the coverage of dictionaries is never complete, WNs are in constant development. Even should the main lexicon of a language be described, as languages evolve, new words and senses appear, while old senses fade away. For this reason, maintaining a WN is a demanding task that should be facilitated in every possible way.

In this paper we present a web-based system designed to exploit the OMW multilingual structure, allowing a multilingual editing environment (e.g. allow multilingual lexicographers to edit multiple languages at the same time), to allow remote parallel access to the editing environment, requiring minimal to no technical knowledge from the lexicographers side to install/run the editing interface, and to facilitate the management overhead of mantaining a WN.[1]

The system has been tested by the developers

[1] `http://compling.hss.ntu.edu.sg/omw/`

Proceedings of ACL-IJCNLP 2015 System Demonstrations, pages 73–78,
Beijing, China, July 26-31, 2015.

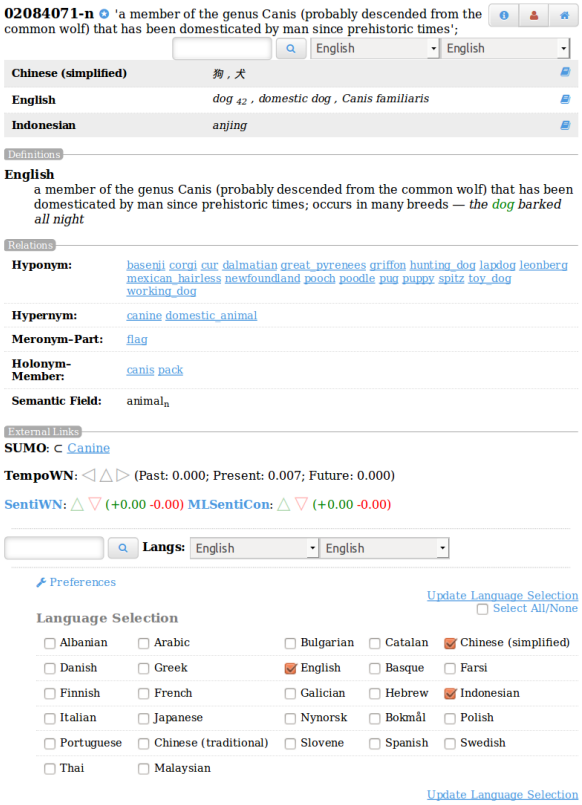

Figure 1: OMW Browsing / Language Selection

and ten annotators (linguistics students) for over 10 months, who made feature requests and gave feedback during the development. The lexicographic work was done in parallel with the semantic annotation of a portion of the NTU Multilingual Corpus (NTU-MC: Tan and Bond, 2012) in (Mandarin) Chinese, English, Indonesian and Japanese (Bond et al., 2015).

The remainder of this paper is arranged as follows. In Section 2 we discuss related work, including the OMW and other similar tools available. The main functionality of our system are described in Section 3. Section 4 will summarize and point to our current and future work.

2 Related Work

2.1 The Open Multilingual Wordnet (OMW)

The OMW is a combination of open-licenced wordnets, along with data extracted from Wiktionary and the Unicode Common Locale Data Repository. In total, OMW has over 2 million senses for over 100 thousand concepts, linking over 1.4 million words in hundreds of languages (Bond and Foster, 2013). It is used as a source of data for projects such as BabelNet (Navigli and Ponzetto, 2012) and Google translate. OMW uses the basic structure of the Princeton Wordnet (PWN: Fellbaum, 1998) to pivot other languages around PWN3.0 synset IDs. Even though it is a single resource, data from each language and project is available separately, respecting their individual licenses. Figure 1 shows the language selection menu that allows browsing this resource as a monolingual or a multilingual resource. The OMW is also fully integrated with the tools currently in use for the development of the NTU Multilingual Corpus (Tan and Bond, 2012). Even though the specifics of this integration go beyond the scope of this paper, it is important to note that most of the development that tested this tool was according to the needs of the semantic annotation of the NTU-MC.

2.2 Other Available Systems

Building and expanding large lexical semantic resources is not an easy task. More importantly, many realized early on that building a WN is not a simple translation task (e.g., Vossen, 1998a). Being able to modify its hierarchy and creating new concepts is important when expressing individual languages semantic hierarchies. Still, due to the lack of resources, many WN projects bootstrap themselves by translating the PWN. However, as individual projects grow, they tend to move away from the inherited English concept hierarchy. This is the moment when systems to support easy manipulation and expansion of their WNs are needed.

Among the available systems we can find VisDic (Horák et al., 2004) (later developed into DEBVisDic, Horák et al., 2006), used for the development of BalkaNet (Tufis et al., 2004), plWordNetApp (succeeded by WordNetLoom (Piasecki et al., 2013)) used for the construction of the Polish Wordnet (Derwojedowa et al., 2008), GernEdiT, the GermaNet editing tool (Henrich and Hinrichs, 2010), KUI (Sornlertlamvanich et al., 2008, used in the Asian Wordnet) and Polaris (Louw, 1998) used in the development of EuroWordnet (Vossen, 1998b).

Out of the above mentioned, we excluded Polaris as not being released. Even though GernEdiT seemed well designed, it was mainly developed for a monolingual environment and a restrictive license (i.e. it does not allow redistribution or commercial use). WordNetLoom, the successor of plWordNetApp, develops an interesting editing mode based directly on the WN hierarchy graph, but the fact that it was not offered as a web-service limited our interest. VisDic was originally devel-

oped in response to Polaris commercial licence, but the direct manipulation of XML limited its usefulness when compared to the original work-flow of direct manipulation of text files. DEBVisDic was later developed on top of VisDic, enhacing it many ways. It can be served as a web application and it supports the development of multiple link wordnets. Unfortunately, while experimenting with it, we found that its installation and user experience was not intuitive. Its development and usability is strongly dependent on Mozilla's Firefox, making any further development less appealing. And, most importantly, its license also restricts use of the tool to noncommercial, nonprofit internal research purposes only. KUI was open source, but only contained functionality for adding lemmas, not making changes to the wordnet structure. We decided we had enough motivation to start the development of a new tool.

3 System Overview and Architecture

OMWEdit follows a simple yet powerful web-based architecture. It is built on an SQLite database, allowing fast querying and reliable storage. It is fully tested for Firefox, Chrome and Safari browsers. Its main focus is on semi-automation and consistency checking of the WN development work, supporting the lexicographer's work. In this section we discuss the OMWEdit's main functionality.

3.1 Browsing and Authentication

The OMW can be browsed either monolingually or multilingually. Figure 1 shows how languages can be filtered through the navigation interface. Filtering languages is an important feature for both browsing and editing since many concepts have data for over 100 languages. This amount of information can be overwhelming, especially within the edition interface. The OMW interface also integrates an authentication menu. As guests, users are free to browse through the resource. Once logged in (provided they are given access), a user can access the editing mode. All changes committed are immediately available for further browsing, editing and usage by linked tools (i.e. the OMW is currently linked to a set of corpus annotation tools).

3.2 Creating and Editing Concepts

The lexicographic work centers around editing existing concepts and adding new concepts, senses or relations to the WN. For this reason, our system has been optimized for these two tasks.

Our system integrates the lexical, concept and relation levels in a single semi-automated process. Most of the above mentioned systems sustain a separate development between lexical entries and concepts (e.g. in order to be linked to a concept, a lexical unit has to be previously created as a separate entity). Contrary to this practice, the OMWEdit has a fully automated lexical management — e.g. the creation, linking, unlinking, correction and deletion of lexical entries is fully automated behind the concept creation/edition screen. In order to add a lemma to a concept, for example, a lexicographer has simply to type the word form in the appropriate field of creation/editing concept view. The system then verifies if a new lexical entry needs to be created. In the event that the lexical entry already exists, its ID is automatically fetched and bound to the concept. Otherwise, a new lexical entry is created and linked to the concept ID. Likewise, correcting an existing lexical entry within a concept will trigger a similar process. The system checks if a lexical entry that matches the corrected version already exists, or if needs to be created. The link between the previously corrected lexical unit is dropped and a new link is created for the newly corrected form. Lexical entries that are not linked to any concept are periodically removed from the database.

Similar processes are put in practice for the main components of a concept. We currently allow to edit/add lemmas, definitions, examples and synset relations. The web interface was designed to be intuitive and as automated as possible, in order to shield the lexicographer's work to include checking. The editing interfaces include quick guidelines that summarize the workflow of that particular task. Typos are avoided by either checking inputs with regular expressions or through the use of closed selection lists (see Figure 2). The inputs are processed for further consistency before being written in the database (e.g. white-space and punctuation stripping).

Fields such as definitions, examples and lemmas are associated with languages. Most of our lexicographers are, at least, bilingual. Having the possibility of creating multilingual concepts is a single submission is, therefore, most efficient. The languages displayed and available when creating/editing a concept are constrained by the se-

Figure 2: Adding a new concept

lected browsing languages, as seen in Figure 1. It is especially important to be able to constrain the languages in the editing mode, since too much information quickly becomes hard to manage.

The creation of a new synset has been optimized to fall under one of three categories. Named Entities have a quick creation setting where only minimal information is required (not shown) – as this system knows where to place them in the hierarchy. The creation of new concepts can also be done from scratch (Figure 2), or through the selection of a linked concept. In this case, the information available for the linked concept is displayed and some information is copied across into the creation form for further edition.

The tool has a link to a multiple lexicon search, where the lexicographers can simultaneously query multiple lexicons for different languages (e.g. wiktionary, JMDict for Japanese, CC-edict for Chinese and so on). This makes it easy to check the meanings of words without relying too much on a single source.

Other consistency checks are enforced by the system. For example, when creating new entries, the minimal information required to constitute a concept (name, pos, definition in English, link to existing synset, at least one lemma) is enforced by the interface, making it impossible to unwittingly create an ill-formed entry.

3.3 Journaling and Reporting System

Although the wordnets are stored in a relational database, they can easily be exported as other standard formats such as Wordnet LMF (Vossen et al., 2013) or plain text triples.

The WN development is done directly into this database. All editable tables are associated with triggers that record every change committed to the database, along with the authorship and the timestamp of the event. By keeping the metadata in a separate table, we ensure that the wordnet itself does not get unwieldy. Both manual and scripted manipulation of a data is dated and signed. Individual lexicographers have to go through an online login system to be able to see the editing interface. The authorship of scripted manipulation of data is often the name of the script — this allows us to keep track of what was changed when. The ease of manipulation of the data by scripts is important to the development — it is easy to develop new data as a separate project and import it when it is ready.

This careful database journaling keeps a tractable history of all the changes in the OWN. This information allows a better management of the lexicographers' workflow, and also a better control of the quality of data that is inserted. Management of the lexicographic work is facilitated by a reporting interface that displays the rate/progress of each contributor's work (Figure 3). This interface allows the coordinators to adjudicate and revert any changes they deem fit, to assert the work pace of individual contributors and also to judge the quality of the lexicographer's work. This interface also provides some consistency checks to the quality of the information contained in the synset, for example lack of senses or definitions, and how often it appears in the corpus.

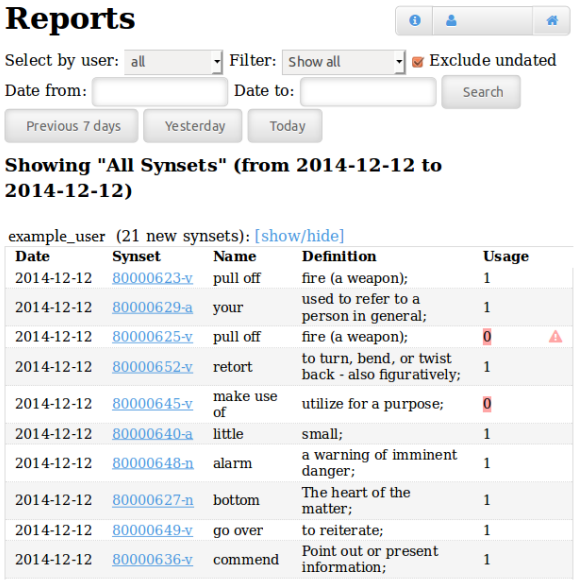

Figure 3: Reporting Interface (extract of list)

4 Summary and Future Work

We have described the architecture and main functionality of the OMWEdit. Considering its short development history, our system has proved itself an increasingly stable and useful tool for the expansion of a few major Wordnet projects. Our web-based system architecture has proved itself ideal for a medium to large scale lexicographic team, regardless of the location of each member. During the development of this system, we were able to confirm an increase in the speed of the lexicographer's workflow. The managing overhead, such as maintaining consistency and quality of the introduced changes has also become a much easier task to accomplish.

Nevertheless, we are aware that the nature of this kind of system is always open ended, and we hope to keep supporting and developing it further. We are aware of some shortcomings and have a list of ongoing planned development of future implementation. This list includes (but is not restricted to):

- the ability to change and/or add lexical relations and verb frames
- the ability to easily comment on entries and watch entries for changes
- the ability to express all relations (both lexical and between concepts) by language — allowing to move away from only using the hierarchy given by the PWN
- the ability to seed a new concept by copying a similar concept (with all its internal structure and relations)
- the ability to do a live check for similarity scores in definitions, accounting for probable matching/mergeable concepts
- further development of the reporting interface
- the development of a graphic component to help visualizing the best placement of a new concept in the hierarchy
- Also, considering our multilingual context, to further develop our support for multilingual users by translating the user interface.

Even though our system was developed with the goal of expanding and correcting wordnets, we believe that our system can also be used to help create new wordnets that use the PWN hierarchy as their backbone. Though the hierarchical relations are still currently imposed by the PWN, we have abolished the limitation to a fixed concept inventory by allowing the creation of new concepts.

Although the tool is far from perfect, we encourage existing and new projects to use the OMW and OMWEdit as a platform to for their WN development. Furthermore, we intend to feedback the changes committed to the individual wordnet projects: the Princeton Wordnet (Fellbaum, 1998), the Japanese Wordnet (Isahara et al., 2008), the Wordnet Bahasa (Nurril Hirfana *et al.* 2011) and the Chinese Open Wordnet (Wang and Bond, 2013), respectively, so that changes committed to the OMW can be incorporated to the original WN projects.

Acknowledgments

This research was supported in part by the MOE Tier 2 grant *That's what you meant: a Rich Representation for Manipulation of Meaning* (MOE ARC41/13) and Fuji-Xerox Co. Ltd (Joint Research on *Multilingual Semantic Analysis*). We would also like to thank our lexicographers for their feedback during the system's development.

References

Francis Bond, Luís Morgado da Costa, and Tuấn Anh Lê. 2015. IMI — a multilingual semantic annotation environment. In *ACL-2015 System Demonstrations*. (this volume).

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *51st Annual Meeting of the Association for Computational Linguistics: ACL-2013*, pages 1352–1362. Sofia. URL `http://aclweb.org/anthology/P13-1133`.

Magdalena Derwojedowa, Maciej Piasecki, Stanisław Szpakowicz, Magdalena Zawisławska, and Bartosz Broda. 2008. Words, concepts and relations in the construction of polish wordnet. In *Proceedings of the Global WordNet Conference, Seged, Hungary*, pages 162–177.

Christine Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Verena Henrich and Erhard W Hinrichs. 2010. Gernedit-the germanet editing tool. In *ACL (System Demonstrations)*, pages 19–24.

Aleš Horák, Karel Pala, Adam Rambousek, Martin Povolnỳ, et al. 2006. Debvisdic–first version of new client-server wordnet browsing and editing tool. In *Proceedings of the Third International Wordnet Conference (GWC-06), Jeju Island, Korea*.

Aleš Horák, Pavel Smrž, et al. 2004. Visdic–wordnet browsing and editing tool. In *Proceedings of the Second International WordNet Conference–GWC*, pages 136–141.

Hitoshi Isahara, Francis Bond, Kiyotaka Uchimoto, Masao Utiyama, and Kyoko Kanzaki. 2008. Development of the Japanese WordNet. In *Sixth International conference on Language Resources and Evaluation (LREC 2008)*. Marrakech.

Michael Louw. 1998. Polaris user's guide. *EuroWordNet (LE-4003), Deliverable D023D024*.

Nurril Hirfana Mohamed Noor, Suerya Sapuan, and Francis Bond. 2011. Creating the open Wordnet Bahasa. In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation (PACLIC 25)*, pages 258–267. Singapore.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Maciej Piasecki, Michał Marcińczuk, Radosław Ramocki, and Marek Maziarz. 2013. Wordnetloom: a wordnet development system integrating form-based and graph-based perspectives. *International Journal of Data Mining, Modelling and Management*, 5(3):210–232.

Virach Sornlertlamvanich, Thatsanee Charoenporn, Kergrit Robkop, and Hitoshi Isahara. 2008. KUI: Self-organizing multi-lingual wordnet construction tool. In Attila Tanács, Dóra Csendes, Veronika Vincze, Christiane Fellbaum, and Piek Vossen, editors, *4th Global Wordnet Conference: GWC-2008*, pages 417–427. Szeged, Hungary.

Liling Tan and Francis Bond. 2012. Building and annotating the linguistically diverse NTU-MC (NTU-multilingual corpus). *International Journal of Asian Language Processing*, 22(4):161–174.

Dan Tufis, Dan Cristea, and Sofia Stamou. 2004. Balkanet: Aims, methods, results and perspectives. a general overview. *Romanian Journal of Information science and technology*, 7(1-2):9–43.

Piek Vossen, editor. 1998a. *Euro WordNet*. Kluwer.

Piek Vossen. 1998b. *A multilingual database with lexical semantic networks*. Springer.

Piek Vossen, Claudia Soria, and Monica Monachini. 2013. LMF - lexical markup framework. In Gil Francopoulo, editor, *LMF - Lexical Markup Framework*, chapter 4. ISTE Ltd + John Wiley & sons, Inc.

Shan Wang and Francis Bond. 2013. Building the Chinese Open Wordnet (COW): Starting from core synsets. In *Proceedings of the 11th Workshop on Asian Language Resources, a Workshop at IJCNLP-2013*, pages 10–18. Nagoya.