UWN: A Large Multilingual Lexical Knowledge Base

Gerard de Melo
ICSI Berkeley
demelo@icsi.berkeley.edu

Gerhard Weikum
Max Planck Institute for Informatics
weikum@mpi-inf.mpg.de

Abstract

We present UWN, a large multilingual lexical knowledge base that describes the meanings and relationships of words in over 200 languages. This paper explains how link prediction, information integration and taxonomy induction methods have been used to build UWN based on WordNet and extend it with millions of named entities from Wikipedia. We additionally introduce extensions to cover lexical relationships, frame-semantic knowledge, and language data. An online interface provides human access to the data, while a software API enables applications to look up over 16 million words and names.

1 Introduction

Semantic knowledge about words and named entities is a fundamental building block both in various forms of language technology as well as in end-user applications. Examples of the latter include word processor thesauri, online dictionaries, question answering, and mobile services. Finding semantically related words is vital for query expansion in information retrieval (Gong et al., 2005), database schema matching (Madhavan et al., 2001), sentiment analysis (Godbole et al., 2007), and ontology mapping (Jean-Mary and Kabuka, 2008). Further uses of lexical knowledge include data cleaning (Kedad and Métais, 2002), visual object recognition (Marszałek and Schmid, 2007), and biomedical data analysis (Rubin and others, 2006).

Many of these applications have used English-language resources like WordNet (Fellbaum, 1998). However, a more multilingual resource equipped with an easy-to-use API would not only enable us to perform all of the aforementioned tasks in additional languages, but also to explore cross-lingual applications like cross-lingual IR (Etzioni et al., 2007) and machine translation (Chatterjee et al., 2005).

This paper describes a new API that makes lexical knowledge about millions of items in over 200 languages available to applications, and a corresponding online user interface for users to explore the data. We first describe link prediction techniques used to create the multilingual core of the knowledge base with word sense information (Section 2). We then outline techniques used to incorporate named entities and specialized concepts (Section 3) and other types of knowledge (Section 4). Finally, we describe how the information is made accessible via a user interface (Section 5) and a software API (Section 6).

2 The UWN Core

UWN (de Melo and Weikum, 2009) is based on WordNet (Fellbaum, 1998), the most popular lexical knowledge base for the English language. WordNet enumerates the senses of a word, providing a short description text (gloss) and synonyms for each meaning. Additionally, it describes relationships between senses, e.g. via the hyponymy/hypernymy relation that holds when one term like '*publication*' is a generalization of another term like '*journal*'.

This model can be generalized by allowing words in multiple languages to be associated with a meaning (without, of course, demanding every meaning be lexicalized in every language). In order to accomplish this at a large scale, we automatically link

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 151–156,
Jeju, Republic of Korea, 8-14 July 2012.

terms in different languages to the meanings already defined in WordNet. This transforms WordNet into a multilingual lexical knowledge base that covers not only English terms but hundreds of thousands of terms from many different languages.

Unfortunately, a straightforward translation runs into major difficulties because of homonyms and synonyms. For example, a word like '*bat*' has 10 senses in the English WordNet, but a German translation like '*Fledermaus*' (the animal) only applies to a small subset of those senses (cf. Figure 1). This challenge can be approached by disambiguating using machine learning techniques.

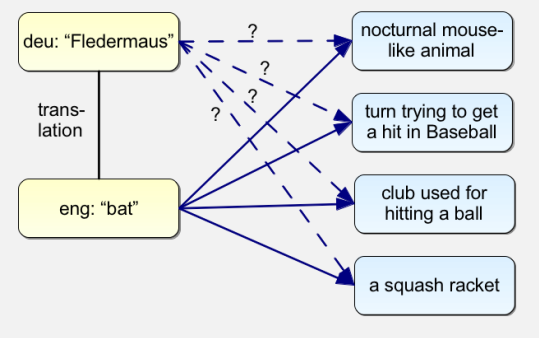

Figure 1: Word sense ambiguity

Knowledge Extraction An initial input knowledge base graph G_0 is constructed by extracting information from existing wordnets, translation dictionaries including Wiktionary (*http://www.wiktionary.org*), multilingual thesauri and ontologies, and parallel corpora. Additional heuristics are applied to increase the density of the graph and merge near-duplicate statements.

Link Prediction A sequence of knowledge graphs G_i are iteratively derived by assessing paths from a new term x to an existing WordNet sense z via some English translation y covered by WordNet. For instance, the German '*Fledermaus*' has '*bat*' as a translation and hence initially is tentatively linked to all senses of '*bat*' with a confidence of 0. In each iteration, the confidence values are then updated to reflect how likely it seems that those links are correct. The confidences are predicted using RBF-kernel SVM models that are learnt from a training set of labelled links between non-English words and senses. The feature space is constructed using a series of graph-based statistical scores that represent properties of the previous graph G_{i-1} and additionally make use of measures of semantic relatedness and corpus frequencies. The most salient features $x_i(x, z)$ are of the form:

$$\sum_{y \in \Gamma(x, G_{i-1})} \phi(x, y)\ \mathrm{sim}^*_x(y, z) \tag{1}$$

$$\sum_{y \in \Gamma(x, G_{i-1})} \frac{\phi(x, y)\ \mathrm{sim}^*_x(y, z)}{\mathrm{sim}^*_x(y, z)\ + \mathrm{dissim}_x(y, z)} \tag{2}$$

The formulae consider the out-neighbourhood $y \in \Gamma(x, G_{i-1})$ of x, i.e. its translations, and then observe how strongly each y is tied to z. The function sim^* computes the maximal similarity between any sense of y and the current sense z. The dissim function computes the sum of dissimilarities between senses of y and z, essentially quantifying how many alternatives there are to z. Additional weighting functions ϕ, γ are used to bias scores towards senses that have an acceptable part-of-speech and senses that are more frequent in the SemCor corpus.

Relying on multiple iterations allows us to draw on multilingual evidence for greater precision and recall. For instance, after linking the German '*Fledermaus*' to the animal sense of '*bat*', we may be able to infer the same for the Turkish translation '*yarasa*'.

Results We have successfully applied these techniques to automatically create UWN, a large-scale multilingual wordnet. Evaluating random samples of term-sense links, we find (with Wilson-score intervals at $\alpha = 0.05$) that for French the precision is $89.2\% \pm 3.4\%$ (311 samples), for German $85.9\% \pm 3.8\%$ (321 samples), and for Mandarin Chinese $90.5\% \pm 3.3\%$ (300 samples). The overall number of new term-sense links is 1,595,763, for 822,212 terms in over 200 languages. These figures can be grown further if the input is extended by tapping on additional sources of translations.

3 MENTA: Named Entities and Specialized Concepts

The UWN Core is extended by incorporating large amounts of named entities and language- and domain-specific concepts from Wikipedia (de Melo and Weikum, 2010a). In the process, we also obtain

human-readable glosses in many languages, links to images, and other valuable information. These additions are not simply added as a separate knowledge base, but fully connected and integrated with the core. In particular, we create a mapping between Wikipedia and WordNet in order to merge equivalent entries and we use taxonomy construction methods in order to attach all new named entities to their most likely classes, e.g. '*Haight-Ashbury*' is linked to a WordNet sense of the word '*neighborhood*'.

Information Integration Supervised link prediction, similar to the method presented in Section 2, is used in order to attach Wikipedia articles to semantically equivalent WordNet entries, while also exploiting gloss similarity as an additional feature. Additionally, we connect articles from different multilingual Wikipedia editions via their cross-lingual interwiki links, as well as categories with equivalent articles and article redirects with redirect targets.

We then consider connected components of directly or transitively linked items. In the ideal case, such a connected component consists of a number of items all describing the same concept or entity, including articles from different versions of Wikipedia and perhaps also categories or WordNet senses.

Unfortunately, in many cases one obtains connected components that are unlikely to be correct, because multiple articles from the same Wikipedia edition or multiple incompatible WordNet senses are included in the same component. This can be due to incorrect links produced by the supervised link prediction, but often even the original links from Wikipedia are not consistent.

In order to obtain more consistent connected components, we use combinatorial optimization methods to delete certain links. In particular, for each connected component to be analysed, an Integer Linear Program formalizes the objective of minimizing the costs for deleted edges and the costs for ignoring soft constraints. The basic aim is that of deleting as few edges as possible while simultaneously ensuring that the graph becomes as consistent as possible. In some cases, there is overwhelming evidence indicating that two slightly different articles should be grouped together, while in other cases there might be little evidence for the correctness of an edge and so it can easily be deleted with low cost.

While obtaining an exact solution is NP-hard and APX-hard, we can solve the corresponding Linear Program using a fast LP solver like CPLEX and subsequently apply region growing techniques to obtain a solution with a logarithmic approximation guarantee (de Melo and Weikum, 2010b).

The clean connected components resulting from this process can then be merged to form aggregate entities. For instance, given WordNet's standard sense for '*fog*', water vapor, we can check which other items are in the connected component and transfer all information to the WordNet entry. By extracting snippets of text from the beginning of Wikipedia articles, we can add new gloss descriptions for fog in Arabic, Asturian, Bengali, and many other languages. We can also attach pictures showing fog to the WordNet word sense.

Taxonomy Induction The above process connects articles to their counterparts in WordNet. In the next step, we ensure that articles without any direct counterpart are linked to WordNet as well, by means of taxonomic hypernymy/instance links (de Melo and Weikum, 2010a).

We generate individual hypotheses about likely parents of entities. For instance, articles are connected to their Wikipedia categories (if these are not assessed to be mere topic descriptors) and categories are linked to parent categories, etc. In order to link categories to possible parent hypernyms in WordNet, we adapt the approach proposed for YAGO (Suchanek et al., 2007) of determining the headword of the category name and disambiguating it.

Since we are dealing with a multilingual scenario that draws on articles from different multilingual Wikipedia editions that all need to be connected to WordNet, we apply an algorithm that jointly looks at an entity and all of its parent candidates (not just from an individual article, but all articles in the same connected component) as well as superordinate parent candidates (parents of parents, etc.), as depicted in Figure 2. We then construct a Markov chain based on this graph of parents that also incorporates the possibility of random jumps from any parent back to the current entity under consideration. The stationary probability of this Markov chain, which can be obtained using random walk methods, provides us a ranking of the most likely parents.

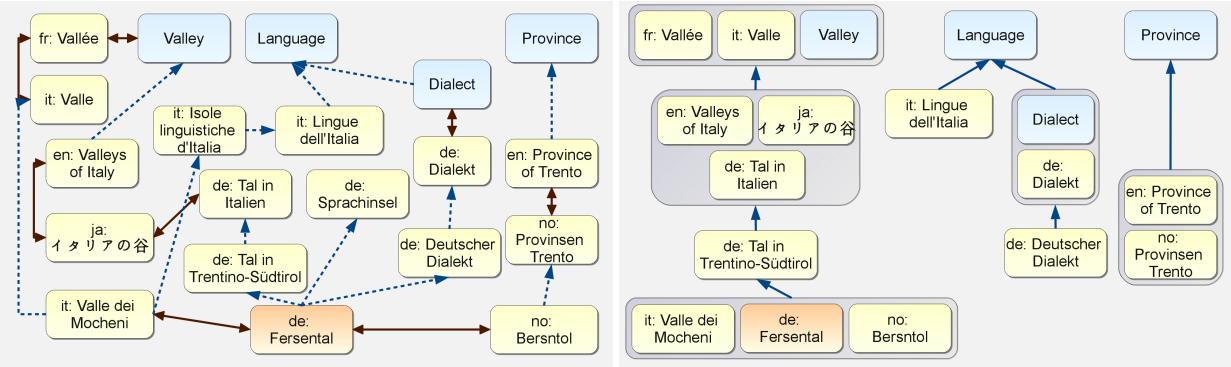

Figure 2: Noisy initial edges (left) and cleaned, integrated output (right), shown in a simplified form

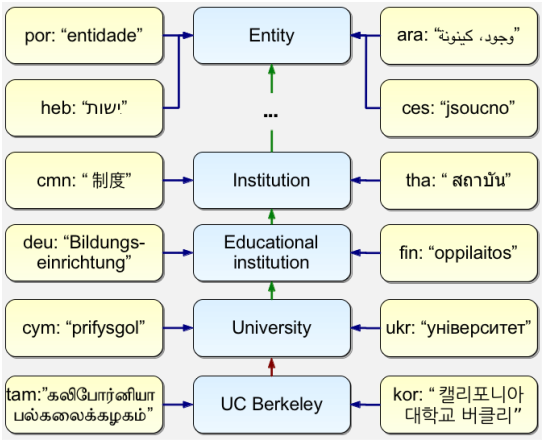

Figure 3: UWN with named entities

Results Overall, we obtain a knowledge base with 5.4 million concepts or entities and 16.7 million words or names associated with them from over 200 languages. Over 2 million named entities come only from non-English Wikipedia editions, but their taxonomic links to WordNet still have an accuracy around 90%. An example excerpt is shown in Figure 3, with named entities connected to higher-level classes in UWN, all with multilingual labels.

4 Other Extensions

Word Relationships Another plugin provides word relationships and properties mined from Wiktionary. These include derivational and etymological word relationships (e.g. that '*grotesque*' comes from the Italian '*grotta*': grotto, artificial cave), alternative spellings (e.g. '*encyclopædia*' for '*encyclopedia*'), common misspellings (e.g. '*minuscule*' for '*minuscule*'), pronunciation information (e.g. how to pronounce '*nuclear*'), and so on.

Frame-Semantic Knowledge Frame semantics is a cognitively motivated theory that describes words in terms of the cognitive frames or scenarios that they evoke and the corresponding participants involved in them. For a given frame, FrameNet provides definitions, involved participants, associated words, and relationships. For instance, the `Commerce_goods-transfer` frame normally involves a seller and a buyer, among other things, and different words like '*buy*' and '*sell*' can be chosen to describe the same event.

Such detailed knowledge about scenarios is largely complementary in nature to the sense relationships that WordNet provides. For instance, WordNet emphasizes the opposite meaning of the words '*happy*' and '*unhappy*', while frame semantics instead emphasizes the cognitive relatedness of words like '*happy*', '*unhappy*', '*astonished*', and '*amusement*', and explains that typical participants include an experiencer who experiences the emotions and external stimuli that evoke them. There have been individual systems that made use of both forms of knowledge (Shi and Mihalcea, 2005; Coppola and others, 2009), but due to their very different nature, there is currently no simple way to accomplish this feat. Our system addresses this by seamlessly integrating frame semantic knowledge into the system. We draw on FrameNet (Baker et al., 1998), the most well-known computational instantiation of frame semantics. While the FrameNet project is generally well-known, its use in practical applica-

tions has been limited due to the lack of easy-to-use APIs and because FrameNet alone does not cover as many words as WordNet. Our API simultaneously provides access to both sources.

Language information For a given language, this extension provides information such as relevant writing systems, geographical regions, identification codes, and names in many different languages. These are all integrated into WordNet's hypernym hierarchy, i.e. from language families like the Sinitic languages one may move down to macrolanguages like Chinese, and then to more specific forms like Mandarin Chinese, dialect groups like Ji-Lu Mandarin, or even dialects of particular cities.

The information is obtained from ISO standards, the Unicode CLDR as well as Wikipedia and then integrated with WordNet using the information integration strategies described above (de Melo and Weikum, 2008). Additionally, information about writing systems is taken from the Unicode CLDR and information about individual characters is obtained from the Unicode, Unihan, and Hanzi Data databases. For instance, the Chinese character '娴' is connected to its radical component '女' and to its pronunciation component '闲'.

5 Integrated Query Interface and Wiki

We have developed an online interface that provides access to our data to interested researchers (*yago-knowledge.org/uwn/*), as shown in Figure 4.

Interactive online interfaces offer new ways of interacting with lexical knowledge that are not possible with traditional print dictionaries. For example, a user wishing to find a Spanish word for the concept of persuading someone not to believe something might look up the word '*persuasion*' and then navigate to its antonym '*dissuasion*' to find the Spanish translation. A non-native speaker of English looking up the word '*tercel*' might find it helpful to see pictures available for the related terms '*hawk*' or '*falcon*' – a Google Image search for '*tercel*' merely delivers images of Toyota Tercel cars.

While there have been other multilingual interfaces to WordNet-style lexical knowledge in the past (Pianta et al., 2002; Atserias and others, 2004), these provide less than 10 languages as of 2012. The most similar resource is BabelNet (Navigli and Ponzetto, 2010), which contains multilingual synsets but does not connect named entities from Wikipedia to them in a multilingual taxonomy.

Figure 4: Part of Online Interface

6 Integrated API

Our goal is to make the knowledge that we have derived available for use in applications. To this end, we have developed a fully downloadable API that can easily be used in several different programming languages. While there are many existing APIs for WordNet and other lexical resources (e.g. (Judea et al., 2011; Gurevych and others, 2012)), these don't provide a comparable degree of integrated multilingual and taxonomic information.

Interface The API can be used by initializing an accessor object and possibly specifying the list of plugins to be loaded. Depending on the particular application, one may choose only Princeton WordNet and the UWN Core, or one may want to include named entities from Wikipedia and frame-semantic knowledge derived from FrameNet, for instance. The accessor provides a simple graph-based lookup API as well as some convenience methods for common types of queries.

An additional higher-level API module implements several measures of semantic relatedness. It also provides a simple word sense disambiguation method that, given a tokenized text with part-of-

speech and lemma annotations, selects likely word senses by choosing the senses (with matching part-of-speech) that are most similar to words in the context. Note that these modules go beyond existing APIs because they operate on words in many different languages and semantic similarity can even be assessed across languages.

Data Structures Under the hood, each plugin relies on a disk-based associative array to store the knowledge base as a labelled multi-graph. The outgoing labelled edges of an entity are saved on disk in a serialized form, including relation names and relation weights. An index structure allows determining the position of such records on disk.

Internally, this index structure is implemented as a linearly-probed hash table that is also stored externally. Note that such a structure is very efficient in this scenario, because the index is used as a read-only data store by the API. Once an index has been created, write operations are no longer performed, so B+ trees and similar disk-based balanced tree indices commonly used in relational database management systems are not needed. The advantage is that this enables faster lookups, because retrieval operations normally require only two disk reads per plugin, one to access a block in the index table, and another to access a block of actual data.

7 Conclusion

UWN is an important new multilingual lexical resource that is now freely available to the community. It has been constructed using sophisticated knowledge extraction, link prediction, information integration, and taxonomy induction methods. Apart from an online querying and browsing interface, we have also implemented an API that facilitates the use of the knowledge base in applications.

References

Jordi Atserias et al. 2004. The MEANING multilingual central repository. In *Proc. GWC 2004*.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proc. COLING-ACL 1998*.

Niladri Chatterjee, Shailly Goyal, and Anjali Naithani. 2005. Resolving pattern ambiguity for English to Hindi machine translation using WordNet. In *Proc. Workshop Translation Techn. at RANLP 2005*.

Bonaventura Coppola et al. 2009. Frame detection over the Semantic Web. In *Proc. ESWC*.

Gerard de Melo and Gerhard Weikum. 2008. Language as a foundation of the Semantic Web. In *Proc. ISWC*.

Gerard de Melo and Gerhard Weikum. 2009. Towards a universal wordnet by learning from combined evidence. In *Proc. CIKM 2009*.

Gerard de Melo and Gerhard Weikum. 2010a. MENTA: Inducing multilingual taxonomies from Wikipedia. In *Proc. CIKM 2010*.

Gerard de Melo and Gerhard Weikum. 2010b. Untangling the cross-lingual link structure of Wikipedia. In *Proc. ACL 2010*.

Oren Etzioni, Kobi Reiter, Stephen Soderland, and Marcus Sammer. 2007. Lexical translation with application to image search on the Web. In *Proc. MT Summit*.

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

Namrata Godbole, Manjunath Srinivasaiah, and Steven Skiena. 2007. Large-scale sentiment analysis for news and blogs. In *Proc. ICWSM*.

Zhiguo Gong, Chan Wa Cheang, and Leong Hou U. 2005. Web query expansion by WordNet. In *Proc. DEXA 2005*.

Iryna Gurevych et al. 2012. Uby: A large-scale unified lexical-semantic resource based on LMF. In *Proc. EACL 2012*.

Yves R. Jean-Mary and Mansur R. Kabuka. 2008. ASMOV: Results for OAEI 2008. In *Proc. OM 2008*.

Alex Judea, Vivi Nastase, and Michael Strube. 2011. WikiNetTk – A tool kit for embedding world knowledge in NLP applications. In *Proc. IJCNLP 2011*.

Zoubida Kedad and Elisabeth Métais. 2002. Ontology-based data cleaning. In *Proc. NLDB 2002*.

Jayant Madhavan, P. Bernstein, and E. Rahm. 2001. Generic schema matching with Cupid. In *Proc. VLDB*.

Marcin Marszałek and C. Schmid. 2007. Semantic hierarchies for visual object recognition. In *Proc. CVPR*.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proc. ACL 2010*.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. MultiWordNet: Developing an aligned multilingual database. In *Proc. GWC*.

Daniel L. Rubin et al. 2006. National Center for Biomedical Ontology. *OMICS*, 10(2):185–98.

Lei Shi and Rada Mihalcea. 2005. Putting the pieces together: Combining FrameNet, VerbNet, and WordNet for robust semantic parsing. In *Proc. CICLing*.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. YAGO: A core of semantic knowledge. In *Proc. WWW 2007*.