

Towards Universal Multilingual Knowledge Bases

Gerard de Melo

Max Planck Institute for Informatics
Saarbrücken, Germany
demelo@mpi-inf.mpg.de

Gerhard Weikum

Max Planck Institute for Informatics
Saarbrücken, Germany
weikum@mpi-inf.mpg.de

Abstract

Lexical, ontological, as well as encyclopedic knowledge is increasingly being encoded in machine-readable form. This paper deals with knowledge representation in multilingual settings. It begins by proposing a generic graph-based knowledge base framework, and then, in three case studies, explains how pre-existing knowledge can be cast into this framework. The first case study involves enriching WordNet with information about human languages and their relationships. The second study shows how machine learning techniques can be used to bootstrap a large-scale multilingual version of WordNet where semantic relationships between terms in many languages are captured. The final study examines how information can be extracted from Wiktionary to produce a lexical network of etymological and derivational relationships between words.

1 Introduction

Knowledge of various sorts, including lexical, ontological, and encyclopedic knowledge, is increasingly being captured in machine-readable form. When multiple human languages are involved, additional challenges need to be addressed. For instance, it is not evident how one best represents languages and their relationships, or how related words from different languages may be connected. This paper proposes a generic framework for representing multilingual knowledge in terms of semantic entity-relationship graphs in the spirit of WordNet (Fellbaum, 1998), a well-known monolingual lexical database. It presents three multilingual lexical knowledge bases that exemplify how one can accommodate pre-existing knowledge within the framework using automatic or semi-automatic techniques and simultaneously addresses the following three questions:

1. How can relationships between languages be captured?
2. How can semantic relationships between words in different languages be captured?

3. How can superficial (etymological, derivational) relationships between words in different languages be captured?

The rest of this paper is organized as follows. Section 2 defines the basic framework and discusses approaches to model terms (words and expressions), word senses, and languages. Section 3 introduces the first case study where WordNet is enriched with domain-specific knowledge about human languages and their relationships, addressing the first question. Section 4 describes a large-scale extension of WordNet to cover not just English words but over 800,000 terms from many different languages, which aims at the second question. Section 5 presents a lexical network that encodes etymological and derivational relationships between words, answering the third question. Finally, Section 6 provides concluding remarks.

2 Data Organization Framework

We begin with a few basic assumptions that define the general framework.

2.1 Preliminaries

Definition 2.1. A *statement* is an item from $\mathcal{U} \times \mathcal{R} \times \mathcal{U} \times [0, 1] \times \Sigma$, where the universe \mathcal{U} is a set of entities, \mathcal{R} is a set of relations, and Σ is a labelling alphabet. A statement (x, r, y, c, a) expresses that two entities x, y stand in relation r to each other with degree of confidence c and additional attributes given by $a \in \Sigma$.

For example, one can specify that the English word “snow” stands in an `etymology` relation to the reconstructed Proto-Germanic form `*“snai-waz”` with confidence $c < 1$, and an attribute a denoting the source of this claim. The universe \mathcal{U} may include both real world entities as well as abstractions and conceptualizations. We use entity identifier strings to refer to them. In Semantic Web knowledge bases, the entities can be arbitrary

real-world entities or Web resources. Relations like `dc:creator` express that one entity is the creator of another entity. In WordNet, the entities one deals with are mainly words and word senses, i.e. meanings of words. Relations include word-to-sense relations that connect words to their meanings and vice versa. Additionally, there are sense-to-sense relations like the hyponymy/hypernymy relation, which connects a word or word sense to a more general word or word sense, e.g. “school” is a hyponym of “*educational institution*”. The statement attributes can for instance be used to capture data provenance, or to specify that a relation between two words applies with respect to specific senses of those words.

Definition 2.2. A *knowledge base* is a set of statements that are asserted to be true (to the extent given by their respective degrees of confidence).

A knowledge base of this form can also be seen as a graph or network, and statements can be viewed as edges or links in the network. Note that statements not in the knowledge base are not assumed to be false, i.e. there is no formal commitment to the Closed World Assumption. Hence one can freely extend a knowledge base with whatever information is available or required for a specific task, without implicitly asserting that other statements are false. For example, in Section 3, we extend WordNet with extensive information about a specific domain, and in Section 4 we add new terms to WordNet without being able to guarantee that all senses of those terms are covered. A knowledge base may also be created collaboratively by multiple stakeholders with different foci.

Up to this point, the definitions are generally compatible with the W3C RDF standard (Hayes, 2004). The following principle goes beyond the common practices on the Semantic Web.

Principle 2.3. $x = y$ should hold for any two entities $x, y \in U$ considered semantically identical.

This means that, within a single knowledge base, ideally only a single, shared set of entities should be used, without semantic duplicates. For example, when linking word senses to specific categories such as law, sports, etc., some knowledge bases rely on a separate vocabulary of domain labels, e.g. Bentivogli et al. (2004). We instead advocate following WordNet in using identifiers already present in the knowledge base instead of a separate vocabulary. In WordNet, the sense for “*plaintiff*” is connected to the primary sense of

“*law*”. This has the advantage of extensive information about the domains being readily available, e.g. the hypernym hierarchy can be used to relate domains to each other.

2.2 Representation Choices for Entities

In what follows, we elaborate on how specific real-world entities can be represented.

2.2.1 Terms

When considering entities for words, expressions, or more generally ‘terms’, different levels of abstraction can be considered. For the term entities, we choose to consider two homonyms, e.g. the animal noun “*bear*” and the verb “*bear*”, as the same term, because, typically, one wishes to look up terms in the lexical knowledge base without already knowing what senses exist. This distinction is instead made at the level of sense entities instead of for term entities. In contrast, we do consider the Spanish term “*con*”, which means “*with*”, distinct from the French term “*con*”, which means “*idiot*”. This level of abstraction allows us to model relationships between words in different languages using statements like `(eng:"digital",etymology,lat:"digitus",1,∅)` to express that the English word “*digital*” stems from the Latin word “*digitus*” (finger or toe). If one instead used pure string literals without language information, it would be necessary to specify the two respective languages as additional attributes of the statement.

We consider different word forms distinct terms. There are a few minor subtleties of term identity regarding string encoding. For multilingual applications, the ISO 10646 / Unicode standards offer an appropriate set of characters for encoding strings. Since Unicode allows encoding a character such as “*à*” in either a composed or in a decomposed form, NFC normalization (Davis and Dürst, 2008) is applied to avoid duplicate entities.

2.2.2 Senses

Lexical knowledge bases are generally based on the assumption that the meanings of a word can be enumerated as a list of word senses. In the EuroWordNet approach (Vossen, 1998), also adopted for BalkaNet and other related projects (Tufiş et al., 2004; Atserias et al., 2004), each individual wordnet has its own inventory of senses, and a separate interlingual index (ILI) is intended to serve as a language-neutral repository of senses. When-

ever possible, senses in the individual wordnets are linked to the ILI by means of synonymy, near-synonymy, hyponymy, or other relations.

Such a representation can be transformed into one that is in accordance with Principle 2.3, where sense identifiers are directly shared whenever these can be thought of as existing in multiple languages. Such sharing is in fact one major difference between WordNet and traditional dictionaries: In WordNet, synonymous terms are tied to a single shared sense identifier, while in conventional dictionaries the respective senses have distinct, unconnected entries. What WordNet does for synonymous terms within a language can be generalized to terms across languages by allowing a sense entity to apply to words in more than one language. The general idea is that the set of terms associated with a sense should be either near-synonymous or translational equivalents (with respect to specific contexts).

Note that this principle does not imply that language-specific subtleties be neglected, since distinct entities may co-exist whenever semantic differences persist. For example, if in one language the word for “tree” has a meaning that includes shrubs, then that meaning should not be conflated with the meaning of the English word “tree”, which generally does not include shrubs. In a similar vein, if in one language birds and insects are considered animals and in another they are not, then there are actually distinct concepts that need to be demarcated. This is similar to how the vernacular English concept of “nuts” should be distinguished from the corresponding botanical concept, which excludes peanuts and almonds.

2.2.3 Languages and Language Collections

Sense entities for individual human languages are of particular interest in a multilingual knowledge base. The English word “language” can be viewed from either a countable or an uncountable perspective. One might think of Spanish, Hindi, and so on, as individual *instances* of languages. Alternatively, language can be conceived as a phenomenon, and words like “Spanish” as referring to certain varieties of that phenomenon. In this latter conception, “Spanish”, “Hindi”, etc. can be regarded as hyponyms of “language”, as in WordNet. This allows us to easily model a hierarchy that keeps making finer distinctions as one follows hyponymy links. For instance, from language families like the Semitic or Sinitic languages one

may move down to macrolanguages like Arabic or Chinese, and then to more specific forms like Moroccan Arabic or Mandarin Chinese, dialect groups like Ji-Lu Mandarin, or even dialects of particular cities. Similar distinctions can be made with respect to temporal classifications, or writing systems and orthographies. Subjective or controversial distinctions between language families and macrolanguages, or between languages and dialects or sociolects can be avoided.

3 Extension of WordNet with Language-Related Information

Our first case study deals with modelling relationships between human languages. More specifically, it involves enriching WordNet with domain-specific information about languages and their relationships, as elaborated earlier in Section 2.2.3. WordNet already contains certain languages and language families as hyponyms of “language”. We extend WordNet’s language hierarchy to cover a significantly larger range of languages, with additional background information. This allows multilingual applications to use language identifiers specified within the knowledge base in accordance with Principle 2.3, while simultaneously also facilitating interoperability with international standards. An application can look up information about a language, e.g. where it is spoken.

3.1 Knowledge Extraction

We draw on the following sources to extract relevant information:

- the ISO 639-3 specification¹, which defines codes for around 7,000 languages and lists relationships between macrolanguages and individual languages
- the ISO 639-5 specification², which describes a limited number of language families (e.g. Tai languages) and other collections (e.g. sign languages)
- the ISO 15924 specification³, which lists a number of writing systems, e.g. Cyrillic, Devanagari, and Hangul
- the Ethnologue language codes database (Lewis, 2009), which provides additional language names, geographical regions, etc.

¹<http://www.sil.org/iso639-3/>

²<http://www.loc.gov/standards/iso639-5/>

³<http://unicode.org/iso15924/>

- the Linguist List⁴, which contributes information on extinct languages as well as constructed languages
- the Unicode Common Locale Data Repository⁵ (CLDR), which connects languages to their geographical regions and writing systems, and delivers names in many languages
- the English Wikipedia⁶, from which we can extract multilingual names, glosses, and language family information for several hundred languages

In order to abide to Principle 2.3, we attempt to merge duplicates.

1. Those resources that rely on codes defined by ISO 639 Part 1, 2, or 3 are consolidated simply by means of those codes, possibly relying on the ISO 639-3 mapping tables.
2. Wikipedia’s languages are merged with languages from ISO 639-3 by extracting the codes from the respective Wikipedia articles.
3. Wikipedia’s language families are merged with corresponding families from ISO 639-5 where possible, by extracting links from Wikipedia’s “*List of ISO 639-5 codes*” article, which also provides equivalences between ISO 639-5 and ISO 639-3.
4. Finally, we attempt to map each sense entity x derived from the resources to existing WordNet senses y , using scores computed as

$$m(x, y) = \sum_{t \in \Gamma(x)} \frac{\mathbf{1}_{\Gamma(t) \cap \Delta(x)}(y)}{|\Gamma(t) \cap \Delta(x)|}$$

Here, $\Delta(x)$ returns the set of all WordNet senses in the same WordNet branch where x will be placed. These branches are defined as hyponyms of the “*language*” or “*script*” senses, or as meronyms of the hemisphere senses for geographical areas (parts of one of the hemispheres). The function Γ yields the set of terms for a sense x , or the set of senses of a term t (i.e. the out-neighbourhood in the graph of all term-sense links). For a given set S , $\mathbf{1}_S$ is the corresponding set membership indicator function.

Those languages and writing systems (scripts) that could not be mapped to WordNet are connected to

⁴<http://linguistlist.org/>

⁵<http://cldr.unicode.org/>

⁶<http://en.wikipedia.org/>

the WordNet hypernym hierarchy as new senses in accordance with Section 2.2.3. The language senses are made hyponyms of the respective language family sense if such information is available, or simply added as direct hyponyms of “*language*” or similar words (e.g. “*artificial language*”) if no explicit language family information is available.

Similarly, the writing systems defined by ISO 15924, e.g. Cyrillic and Devanagari, are made new instances of the sense for “*script*”, and geographical regions are made new instances of “*geographical area*”. These, too, are merged with existing entries already in WordNet when possible.

3.2 Results

Even for scores with a low threshold $m(x, y) > 0$, an accuracy rate of $94.3\% \pm 4.1\%$ is obtained for 100 random WordNet language mappings. Ambiguous and low-score mappings were corrected manually by an annotator to ensure the quality of the resulting extension. The process also adds over 7,000 new languages to the roughly 600 existing ones in WordNet, as well as smaller numbers of language families and scripts. Languages often have their name provided not only in English but in many different languages, sometimes over 100.

When new terms are added, these may not satisfy the lexicographic inclusion criteria that other entries are subjected to, e.g. certain language names may not be sufficiently lexicalized within a language to warrant an inclusion in WordNet. This problem is addressed by flagging the newly added term-sense statements appropriately.

The languages are integrated into WordNet’s hypernym hierarchy, using macrolanguages and language families as intermediate hypernyms when possible. In addition to the hypernymy links, the language senses are also equipped with other statements that provide further background information, for instance geographical regions, identification codes, writing systems (links to writing system entities), etc. Table 1 shows an example for the African Bemba language (ChiBemba). Geographical regions are provided by the CLDR and Ethnologue based on ISO 3166 / UN M.49, and the respective entities are merged with the corresponding WordNet senses using the mapping procedure described above.

In the future, we would like to address automatically mapping ISO 639-6 identifiers to WordNet

| Relation | Values* |
|-----------------|---|
| has_gloss | "The Bemba language, Chibemba, also known as Cibemba, Ichibemba, Icibemba and Chiwemba, is a Bantu language that is spoken primarily in Zambia by the Bemba people and about 18 related ethnic groups. [...]" (eng) |
| lexicalization | eng:"Bemba", ukr:"бемба", cmn:"姆巴文", many more |
| hypernym | Central_Bantu_languages |
| iso_639_2B_code | "bem" |
| iso_639_3_code | "bem" |
| region | Zambia, etc. |
| described_by | http://en.wikipedia.org/wiki/Bemba_language |
| described_by | http://www.ethnologue.com/show_language.asp?code=bem |
| script | Latin_script |
| ... | ... |

* The entity identifiers are presented here in a slightly more human-readable form than actually stored in the KB.

Table 1: Example Language Entity: Bemba language (ChiBemba)

to cover dialects and additional variations, once the respective data has been made publically available.

4 Multilingual WordNet Translation

The next case study addresses how semantic relationships between terms in different languages can be captured. As explained earlier in Section 2.2.2, the principles that WordNet is based on can be extended to the multilingual case by treating semantic relationships between terms in different languages in the same way as semantic relationships between terms of a single language: Terms with the same meaning are linked to the same sense node, and terms with related meanings are connected indirectly via connected sense nodes. In order to accomplish this at a large scale, we automatically link terms in different languages to the word senses 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 (de Melo and Weikum, 2009).

4.1 Knowledge Extraction

Following Principle 2.3 and Section 2.2.2, we share sense identifiers between languages where appropriate. In the past, several authors have assumed a similar stance and proposed using translation dictionaries to attach non-English terms to sense identifiers from the English WordNet, e.g. Atserias et al. (1997), Isahara et al. (2008). Such techniques fall within what has been called the

‘expand’ paradigm for building wordnets (Vossen, 1998). Unfortunately, a straightforward translation runs into major difficulties because of synonyms and homonyms. For example, a word such as “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. This challenge can be approached by harnessing machine learning techniques.

An initial input knowledge base graph G_0 is constructed by extracting information from existing wordnets, translation dictionaries including Wiktionary⁷, and the FreeDict project dictionaries⁸, multilingual thesauri and ontologies like the GEMET thesaurus⁹, and parallel corpora like the OpenSubtitles corpus (Tiedemann, 2004). Additional heuristics are applied to increase the density of the graph and merge similar statements.

A sequence of knowledge graphs G_i are iteratively derived by evaluating paths from a new term x to an existing WordNet sense z via some English translation y covered by WordNet. For instance, the German word “Fledermaus” has the English word “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

⁷<http://www.wiktionary.org>

⁸<http://www.freedict.org>

⁹<http://www.eionet.europa.eu/gemet/>

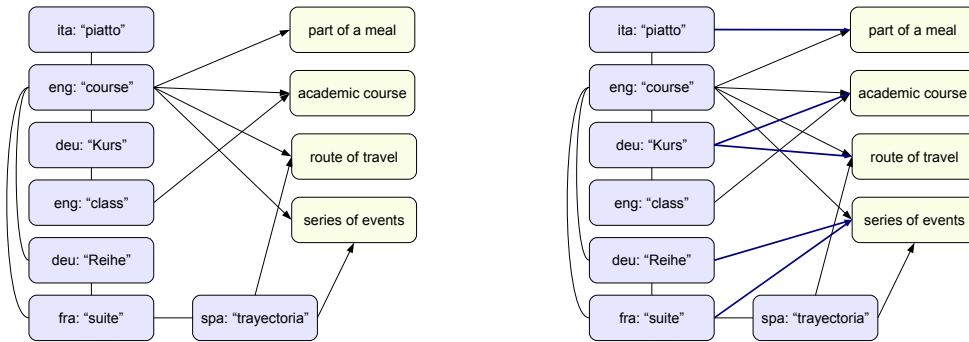


Figure 1: Connections in the input graph G_0 (left) and the desired output graph G_i (right). Lines with arrows represent links from terms to senses, while lines without an arrow represent translation links.

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 significant features $x_i(x, z)$ are computed as:

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

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

The formulas 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, due to mutual reinforcement and propagation effects. For instance, in the first iteration, one might determine that the German word “*Fledermaus*” is linked to the animal sense of “*bat*” with high probability, and then in the next iteration this can aid in inferring that the Turkish translation “*yarasa*” has the same meaning. For further details of this approach, please refer to de Melo and Weikum (2009).

| | Term-Sense Links | Distinct Terms |
|------------|------------------|----------------|
| Nouns | 1,048,003 | 589,536 |
| Verbs | 221,916 | 88,189 |
| Adjectives | 289,328 | 147,257 |
| Adverbs | 36,095 | 26,254 |
| Overall | 1,595,763 | 822,212 |

Table 2: Coverage of multilingual wordnet graph

4.2 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 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, as shown in Table 2. The three most well-represented languages are currently German, French, and Esperanto, which is largely due to the choice of input dictionaries. These figures can easily grow even further as the input is extended by tapping on additional sources.

The structure of the extended wordnet is reasonably rich, including hyponymy/hypernymy and several other generic relations for which it is fair to assume that they apply to the new terms as well. The next step would involve manual revision and extension, since our approach does not necessarily generate complete sense listings and the set of senses associated with a word may not always result in sense distinctions that would seem perfectly adequate to a lexicographer compiling a monolingual dictionary. Additional experiments however

have shown that the wordnet is already beneficial in several application tasks even in this raw form. Examples studied include cross-lingual text classification and semantic relatedness estimation, where high-quality manually created resources are outperformed (de Melo and Weikum, 2009).

5 An Etymological Word Network

As a final case study, we investigate capturing relationships between multilingual word forms, i.e. etymological and derivational information. Traditionally, lexical knowledge bases have focussed on synchronic relationships. We produce an etymological word network that additionally captures diachronic information by representing how words originated from other previously existing words. By navigating this network, one can easily see that the English “*doubtless*” is derived from “*doubt*”, which in turn comes from Old French “*douter*”, which evolved from the Latin word “*dubitare*”. Starting from these latter entities, cognate forms are also discoverable.

5.1 Knowledge Extraction

The knowledge base is mined from the English version of Wiktionary using custom pattern matching techniques. We process the XML dump of Wiktionary, and segment articles by language, since a single article can cover unrelated words in different languages. The “Etymology” sections in the articles may contain arbitrary text describing the roots of a word. Fortunately, certain patterns are very frequent, as one can observe in Figure 2. We thus recursively parse the section using a set of regular expressions that cover many of the etymological relationships described in Wiktionary. Regular expressions extract the language (if mentioned), the original term, and the rest, i.e. the next element in an etymological chain. In addition, the English glosses of words are also parsed, as these often hold links to root forms for derivations, or links to standard forms when there are orthographic variations or other alternative forms. For instance, the English word “*booking*” is attached to the verb “*to book*”. Many articles also have separate sections listing derived forms and alternative spellings, which we harvest as well.

Etymological print dictionaries often do not cite their sources due to space constraints. In our case, the Wiktionary page that provided the etymological link can be referenced. Frequently, this is not

the article page for the word itself, but rather some other page that references that word while tracing a longer etymological history. For example, the etymological link from Anglo-Norman “*estorie*” back to the Latin “*historia*” is found on the page for the English word “*story*”.

Another issue arising in etymology is that some words are known only as reconstructed forms. We represent this at the statement level, adding attributes that specify that the links as well as the unattested forms are hypothetical.

5.2 Results

We obtain a lexical network with over 1,000,000 terms, 200,000 etymological links between terms, and 1,700,000 derivational links between terms. Note however that the distinction between derivational and etymological relations is not always completely clear. For example, many words developed due to quite regular processes of affixation or compound formation, e.g. “*sexism*”, “*microwave*”, and “*website*”. In this regard, our knowledge base follows the conventions adopted in Wiktionary.

Existing standards like TEI P5 (Burnard and Bauman, 2009) define a semi-structured representation of etymological data, rather than a genuinely structural one that exposes relationships between words using a network-like graph model. Graph representations expose the connections between words much more explicitly. Due to affixes such as “*non-*”, “*-ize*”, etc., it turns out that much of the graph actually constitutes a single connected component that can be navigated by following links. In addition, graph representations are machine-readable and more language-neutral, which makes them reusable in different contexts. Information that they cannot directly capture faithfully can still be retained in textual form, e.g. using additional statement attributes. Fortunately, most forms of etymological information, including e.g. when a word’s use was first attested, historic examples of a word’s use, or even the presence of multiple conflicting etymological hypotheses could easily be couched in a machine-readable graph representation without resorting to textual comments.

Etymological relationships are essentially links between words in different language, which can naturally be modelled as relations between terms as defined in Section 2.2.1. Of course, statement


| | |
|---|------------------------|
| English | [edit] |
| Most common English words: hard « ask « question « #410: doubt » around » black » lady | |
| Pronunciation | [edit] |
| <ul style="list-style-type: none"> enPR: dout, IPA: /daʊt/, SAMPA: /daʊt/ <p>Rhymes: -aʊt</p> <ul style="list-style-type: none">  Audio (US)^{help}.file | |
| Etymology | [edit] |
| From Middle English <i>douten</i> from Anglo-Norman <i>douter</i> from Old French <i>douter</i> , from Latin <i>dubitare</i> . Replaced Middle English <i>tweonien</i> "to doubt" (from Old English <i>twēonian</i> , cf Old English <i>twēo</i> "doubt, duplicity"). | |

Figure 2: Excerpt from Wiktionary article on “doubt”, which explains the etymological roots going back to the Latin “dubitare”

attributes could be added to specify that an etymological relationship only applies to specific senses of a term. Indeed, one could also specify relationships of regular polysemy between senses, which would enable a clearer distinction between genuine homonyms and polysemy in the narrow sense than is currently possible in WordNet. Such issues are possible directions for future work.

6 Conclusion

We have analysed principles for representing multilingual knowledge and proposed a general framework as well as techniques to organize existing knowledge within this framework. The first case study involved enriching WordNet with additional information about the vast number of languages in the world. and their relationships. The second demonstrated the use of machine learning to bootstrap a preliminary version of a generic multilingual wordnet describing relationships between terms in different languages. Our final study examined how derivational information between terms in different languages can be extracted from Wiktionary to produce a lexical network of etymological relationships. Together, they demonstrate not only how knowledge bases can universally capture multiple languages simultaneously, but also the additional level of interlinking that this enables.

References

- Jordi Atserias, Salvador Climent, Xavier Farreres, German Rigau, and Horacio Rodríguez. 1997. Combining multiple methods for the automatic construction of multilingual WordNets. In *Proc. International Conference on Recent Advances in NLP (RANLP)*, pages 143–149.
- Jordi Atserias, Luís Villarejo, German Rigau, Eneko Agirre, John Carroll, Bernardo Magnini, and Piek Vossen. 2004. The MEANING Multilingual Central Repository. In *Proc. 2nd Global WordNet Conference (GWC)*, pages 80–210.
- Luisa Bentivogli, Pamela Forner, Bernardo Magnini, and Emanuele Pianta. 2004. Revising the Wordnet Domains hierarchy: semantics, coverage and balancing. In *Proc. COLING 2004 Workshop on Multilingual Linguistic Resources*, pages 94–101, Geneva, Switzerland.
- Lou Burnard and Syd Bauman, 2009. *TEI P5: Guidelines for Electronic Text Encoding and Interchange, Version 1.4.1*. TEI Consortium, July.
- Mark Davis and Martin Dürst. 2008. Unicode normalization forms, rev. 29. Technical report, Unicode.
- Gerard de Melo and Gerhard Weikum. 2009. Towards a universal wordnet by learning from combined evidence. In *Proc. 18th ACM Conference on Information and Knowledge Management (CIKM 2009)*. ACM.
- Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.
- Patrick Hayes. 2004. RDF semantics. W3C recommendation, World Wide Web Consortium, February.
- Hitoshi Isahara, Francis Bond, Kiyotaka Uchimoto, Masao Utiyama, and Kyoko Kanzaki. 2008. Development of the Japanese WordNet. In *Proc. LREC 2008*, Marrakech, Morocco.
- M. Paul Lewis. 2009. Ethnologue: Languages of the world, sixteenth edition (online version).
- Jörg Tiedemann. 2004. The OPUS corpus - parallel & free. In *Proc. LREC 2004*.
- Dan Tufiş, Dan Cristea, and Sofia Stamou. 2004. BalkaNet: Aims, Methods, Results and Perspectives. A General Overview. *Romanian J. Information Science and Technology*, 7(1–2):9–34, 4.
- Piek Vossen, editor. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Springer.