Providing Multilingual, Multimodal Answers to Lexical Database Queries

Gerard de Melo, Gerhard Weikum

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

Abstract

Language users are increasingly turning to electronic resources to address their lexical information needs, due to their convenience and their ability to simultaneously capture different facets of lexical knowledge in a single interface. In this paper, we discuss techniques to respond to a user's lexical queries by providing multilingual and multimodal information, and facilitating navigating along different types of links. To this end, structured information from sources like WordNet, Wikipedia, Wiktionary, as well as Web services is linked and integrated to provide a multi-faceted yet consistent response to user queries. The meanings of words in many different languages are characterized by mapping them to appropriate WordNet sense identifiers and adding multilingual gloss descriptions as well as example sentences. Relationships are derived from WordNet and Wiktionary to allow users to discover semantically related words, etymologically related words, alternative spellings, as well as misspellings. Last but not least, images, audio recordings, and geographical maps extracted from Wikipedia and Wiktionary allow for a multimodal experience.

1. Introduction

In recent years, the way language users search for information about words and their meanings has evolved significantly. Users are increasingly turning to electronic resources to address their lexical information needs, as traditional print media take more time to consult and are less flexible with respect to the organization of lexical information. Alphabetical ordering, for instance, is not well-suited for conveying conceptual relationships between words. Lexical databases, in contrast, can simultaneously capture multiple forms of organization and multiple facets of lexical knowledge. These include thematic, ontological, derivational, or etymological relationships for words in different languages, as well as multimodal information.

Especially with the advent of the World Wide Web, users are increasingly expecting to be able to lookup words and choose between different types of information, perhaps navigating quickly from one concept to another based on given links of interest. 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*".

In this paper, we discuss techniques to respond to lexical queries by providing multilingual and multimodal information and facilitating navigation along different types of links. For example, our system allows for the retrieval of sense-specific translations in different languages; it delivers images for many concrete objects, and offers audio recordings of pronunciations. We describe in particular how relevant information can be obtained from lexical databases like WordNet (Fellbaum, 1998) and Web sources like Wiktionary[1] and Wikipedia[2].

2. Basic Assumptions

We start out by specifying and clarifying some of the underlying assumptions of our system.

2.1. Lexical Databases

Within the context of this paper, a *lexical database* is a set of entities and relationships, where entities can be words and expressions, word meanings, character strings, or Web URIs, and relationships are labeled links between two entities. We consider word meanings in a somewhat abstract sense, such that words in different languages can (roughly) share the same meaning, just like words within a single language can (roughly) share the same meaning, i.e. be near-synonymous. For more details, please refer to de Melo and Weikum (2010).

2.2. Queries

We assume that end users will generally begin interacting with a lexical database by issuing simple word (or expression) lookup queries, irrespective of their specific information needs. This is usually considered the most convenient way of starting a session. Of course, additional entry points can be provided for browsing.

2.3. User Information Needs

Given the simple form of the user queries, it is not possible to derive the specific information needs of a user from a given query on its own, in the absence of additional contextual information. At times the user might be interested in a simple description of a word's meaning, perhaps complemented with visual information. In other cases, when writing a text, the user might be interested in synonyms, related words, and example sentences. Other users, at times, might be more interested in pronunciation information or translations.

The system will thus have to respond to queries by making a range of different types of information accessible to the user upon demand. While most users appreciate clean,

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

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

simple interfaces, it is also important to pay attention to such diverging information needs. As of 2010, Google's standard Web search interface does not differ very much in appearance from what it was like when the company was founded more than a decade earlier. However, the underlying complexity has increased enormously, as various additional features have been made available. For instance, one can choose to search blogs, maps, and books, or even opt for high-quality, medium-length videos released within the past month.

In a similar vein, our system tries to respond to user queries by providing a simple interface, yet making a diverse range of information available to the user upon request. In the following sections, we will describe the different kinds of information that are offered to the user.

3. Word Meanings

We assume that users querying the lexical database system will in many cases be interested in looking up word meanings, which is one of the standard use cases of conventional dictionary users. Even when this is not their main information need, other types of information are often specific to particular senses of a word. For instance, translations of the word "*bank*" differ depending on whether the financial sense or the sloping land sense of the word is meant. Hence, knowledge about the meanings of words is vital for answering many queries.

3.1. WordNet

Our system is strongly based on WordNet 3.0 (Fellbaum, 1998), a well-known lexical database that describes semantic relationships between English words and their meanings, which are encoded as so-called "synsets". WordNet enumerates the senses of a word, providing a short description text (gloss) and synonyms for each word sense. Additionally, it provides structural information about how senses are related, e.g. via the hyponymy/hypernymy relation that holds when one term is a generalization of another term, e.g. "*publication*" is a hypernym of "*journal*". While inspired by theories in cognitive science and extensively used in computational applications, lexical databases like WordNet have also proven to be very useful for ordinary users of language. For instance, numerous dictionary look-up services on the Web as well as the integrated English-language thesaurus of the OpenOffice.org application suite are based on WordNet.

3.2. Multilingual Words

One of the principal differences of our system in comparison with existing WordNet interfaces is the extensive amount of multilingual knowledge. Rather than being limited to just one or perhaps a couple of languages, our interface relies on our UWN[3] project (de Melo and Weikum, 2009b) to provide words and expressions in over 200 languages. Our work adopts what has been called the expand approach for building wordnets (Vossen, 1998) as the underlying paradigm, where words in different languages are connected to an existing repository of senses, as given by the English WordNet. Previous examples of this approach include the work by Atserias et al. (1997) on the Spanish WordNet. Our approach differs from previous techniques by using a large range of input sources, more sophisticated scores, and by adopting statistical learning techniques to obtain high quality, high recall results.

Table 1: Coverage of UWN

Language	Word-Meaning Links	Unique Words
German	132,523	67,087
French	75,544	33,423
Esperanto	71,247	33,664
Dutch	68,792	30,154
Spanish	68,445	32,143
Turkish	67,641	31,553
Czech	59,268	33,067
Russian	57,929	26,293
Portuguese	55,569	23,499
Italian	52,008	24,974
Hungarian	46,492	28,324
Thai	44,523	30,815
Others	795,782	427,216
Total	1,595,763	822,212

Note that this results in a resource that is much more tightly connected than just a simple union of dictionaries. Whenever two words in different languages share a meaning, they are explicitly connected to the same meaning entity, and hence all information relevant to that meaning entity equally applies to all involved words in different languages. For instance, when a sense of a word in one language is looked up, the user can easily navigate to corresponding words in other languages that share the same meaning, or also to hypernyms in different languages. A small sample of terms from UWN is illustrated in Figure 1, and coverage statistics are given in Table 1.

3.3. Multilingual Glosses

The original WordNet provides English-language glosses for each word sense. These glosses describe a particular meaning in a verbose form, similar to the glosses one would find in a conventional dictionary. For example, one sense of the word "*bank*" is described as "*sloping land (especially the slope beside a body of water)*", while for another one we find "*a financial institution that accepts deposits and channels the money into lending activities*".

Similar glosses are useful to have in languages other than English. Unfortunately, even many of the long-running wordnet projects like GermaNet (Hamp and Feldweg, 1997) still do not provide such glosses for more than a negligibly small subset of senses. We thus turn to alternative sources of sense descriptions.

3.3.1. Wikipedia matching

A large number of multilingual glosses for nouns can be obtained by linking WordNet senses to articles in the open Web-based encyclopedia Wikipedia. In Figure 2, we see

[3] UWN is available at `http://www.mpi-inf.mpg.de/yago-naga/uwn/`

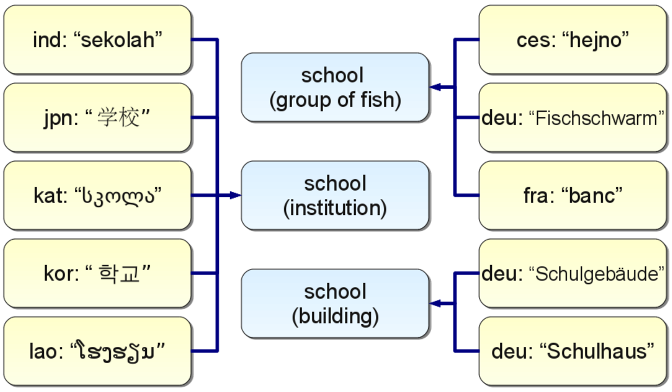


Figure 1: Sample of multilingual terms sharing different meanings.

.

that the first paragraph of an article, or at least the first few sentences are usually suitable gloss descriptions. Such glosses are not only available in English. Articles are often linked to equivalent articles in other languages by means of cross-lingual "interwiki" links. Via such links, we can discover and extract similar non-English gloss descriptions from pages corresponding to the English page. Fortunately, such interwiki links are abundant especially for more commonly used terms like the nouns that are covered by WordNet.

Given a mapping between a Wikipedia article and a WordNet sense, we derive a gloss for the sense by parsing the article's wikitext markup language used by the Mediawiki software as well as filtering out HTML. We then attempt to find the first proper paragraph, skipping preceding infoboxes, pictures, or special notices displayed to users of Wikipedia. If this first paragraph is too long, a sentence boundary is identified in the vicinity of a pre-specified target length position.

To obtain mappings between WordNet and Wikipedia, we average three different heuristic scores.

Label Overlap For every WordNet sense y, we consider the associated words (synonyms) as a set of labels $\mathrm{T}(y)$ from the English WordNet as well as from other languages. Given a Wikipedia article x, we consider a set of term labels $\mathrm{T}(x)$ by taking the respective article title, the titles of corresponding articles in other languages linked via interwiki links, and the titles of redirection pages for any of these articles. Redirection pages often provide synonyms and spelling variations. We rely on a simple similarity measure sim between labels that yields 1 if the languages match and the strings match after making initial characters lower-case as well as removing additional terms in parentheses, and 0 otherwise. For instance, "*Bank (geography)*" matches "*bank*". The label overlap score is then:

$$\sum_{l_y \in \mathrm{T}(y)} \max_{l_x \in \mathrm{T}(x)} w(l_y)\mathrm{sim}(l_x, l_y) \tag{1}$$

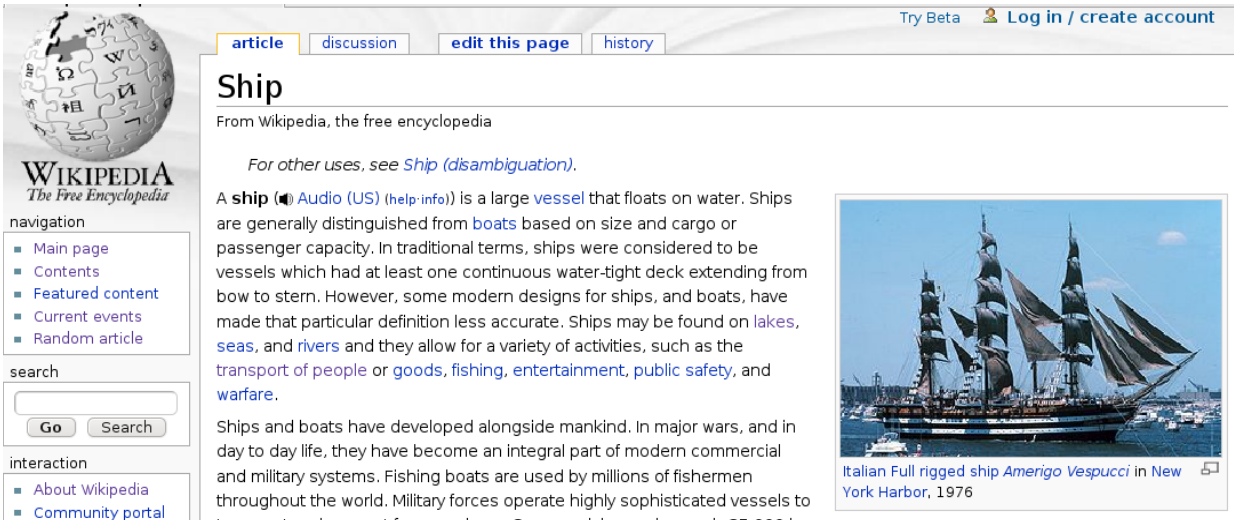
w returns $1/n$ if and only if n different noun meanings of l_y are listed in WordNet.

Gloss Similarity This heuristic considers the cosine $\mathbf{v}_x^T\mathbf{v}_y(||\mathbf{v}_x||\ ||\mathbf{v}_y||)^{-1}$ between vectors $\mathbf{v}_x$, $\mathbf{v}_y$ derived for the gloss extracted from the English Wikipedia and the gloss provided by WordNet, respectively. The vectors are created using TF-IDF scores after Porter stemming.

First Sense Heuristic The first sense heuristic determines the number of Wikipedia labels $l_x \in \mathrm{T}(x)$ of a Wikipedia article x for which the WordNet entity y under consideration is listed as the first sense in WordNet. In WordNet, the first sense listed for a word can generally be assumed to be the most frequent sense in a domain-independent corpus. When looking up labels in WordNet, case differences for the first letter of the label are ignored, but unlike earlier, information in parentheses is not ignored, so only Wikipedia labels in $\mathrm{T}(x)$ that do not contain any additional information in parentheses will be able to match. For example, the label for the Wikipedia article "*House*" can be looked up in WordNet, but labels like "*House (1977 film)*" or "*House (novel)*" will not match anything in WordNet. Hence, in Wikipedia, too, this heuristic only considers what would often be considered the dominant meaning of a label.

Final Output For each Wikipedia article x, we look up the matching WordNet noun synsets y where $\mathrm{T}(x) \cap \mathrm{T}(y) \neq \emptyset$, and compute the weighted average of the aforementioned scores for each y. In the end we choose only those pairs x,y as accepted matches, where there is no other $y' \neq y$ in WordNet with a higher score for x than y, and no other $x' \neq x$ in Wikipedia with a higher score for y than x. We additionally filter out matches with low scores below a predetermined threshold (0.3 used in our experiments).

We experimented with a complete set of Wikipedia dumps from March 2010 in 272 different languages. A total of 38,465 WordNet synsets were connected to Wikipedia, and 433,857 glosses in different languages were obtained. An evaluation of a random sample of 200 mappings between synsets and English Wikipedia pages showed that the precision is 85.8% ± 4.7% (Wilson score interval). Table 2 compares the rather concise WordNet gloss for the main sense of the word "*lake*" with a selection of corresponding glosses extracted from different editions of Wikipedia that were correctly associated with that synset. We find glosses even for smaller language communities like that of Asturian.



Ship

From Wikipedia, the free encyclopedia

For other uses, see Ship (disambiguation).

A **ship** (Audio (US) (help·info)) is a large vessel that floats on water. Ships are generally distinguished from boats based on size and cargo or passenger capacity. In traditional terms, ships were considered to be vessels which had at least one continuous water-tight deck extending from bow to stern. However, some modern designs for ships, and boats, have made that particular definition less accurate. Ships may be found on lakes, seas, and rivers and they allow for a variety of activities, such as the transport of people or goods, fishing, entertainment, public safety, and warfare.

Ships and boats have developed alongside mankind. In major wars, and in day to day life, they have become an integral part of modern commercial and military systems. Fishing boats are used by millions of fishermen throughout the world. Military forces operate highly sophisticated vessels to

Italian Full rigged ship *Amerigo Vespucci* in New York Harbor, 1976

Figure 2: Wikipedia Article example with gloss text in first paragraph and a relevant picture.

3.3.2. Machine translation

Wikipedia does not cover all noun senses in WordNet, and of course users may also desire glosses for verbs, adjectives, and adverbs, which tend not be covered in an encyclopedia. In such cases, we resort to providing lower-quality machine-translated glosses. We informally conducted experiments comparing AltaVista's Babelfish and Google's Translate service on English to German translation, finally settling on the latter service, which is powered by large-scale statistical machine translation techniques, although its translation quality varies significantly and also depends a lot on the language pair being considered.

We proceeded to use the Web API for Google Translate to translating the complete set of 117,659 English glosses given in WordNet into eight target languages. Additional target languages can easily be added. As these translations are often slightly ungrammatical or inaccurate, they are presented to the user only upon request.

3.4. Corpus example sentences

Another addition we considered were multilingual example sentences from corpora. Users appreciate example sentences because they provide an intuitive means of grasping the meaning of a word, and are frequently used to complement conventional word definitions, which some users consider too abstract or even confusing. In some cases, the meaning of a word can directly be determined from its context within the example sentences. In other cases, example sentences can be used in conjunction with conventional definitions to allow users to verify whether they have correctly interpreted a definition.

Example sentences also allow users to see in what circumstances a word would typically be used in practice. For instance, synonymous words such as "*child*", "*kid*", and "*youngster*" can share the same meaning, yet differ significantly with respect to the contexts in which one would consider using them. Example sentences also provide evidence of typical collocations and expressions, e.g. the word "*birth*" often occurs as in "*to give birth*" or "*birth rate*" (but one does not say "*to give nascence*" or "*nascence rate*").

We tackled the challenge of disambiguating example sentences to present them for specific word senses. For instance, for a polysemous word like "*bat*", we would like to have a number of example sentences that refer to the animal sense (e.g. "*There were many bats flying out of the cave.*"), and, separately, a set of example sentences that mention the word in its sports sense (e.g. "*In professional baseball, only wooden bats are permitted.*").

When disambiguating, we found that a higher level of precision can be obtained if we simultaneously look at multiple translations of a text in a parallel corpus. We used word sense disambiguation heuristics and a simple cross-lingual measure of semantic similarity to link example sentences in different languages to specific word senses in WordNet, and additionally investigated techniques to select useful sentences when many are available for a given word sense (de Melo and Weikum, 2009a).

4. Relationships

Relationships between words or between word senses enable Web-like browsing of relevant information, moving from the original user query to related entities that may be closer to the actual user intent and information need.

4.1. WordNet

As mentioned earlier, WordNet provides links from word senses to semantically related word senses, for a number of different relationship types. This allows a user to quickly navigate from a word like "*pessimism*" to its antonym "*optimism*", or from a specific term like "*erythroblast*" to a more generic one like "*cell*".

4.2. Alternative Spellings and Related Words

We relied on the open community-maintained resource Wiktionary to obtain additional lexical information. Wiktionary is a rich source of lexical data, however much of it is only given in a semi-structured or unstructured form. We used custom rule-based extraction code to mine useful information from Wiktionary.

Table 2: WordNet gloss and sample of corresponding multilingual glosses from Wikipedia

English (WordNet)	a body of (usually fresh) water surrounded by land
English (Wikipedia)	A lake (from Latin lacus) is a terrain feature (or physical feature), a body of liquid on the surface of a world that is localized to the bottom of basin (another type of landform or terrain feature; that is, it is not global) and moves slowly if it moves at all. Another definition is, a body of fresh or salt water of considerable size that is surrounded by land.
Asturian	Un llagu ye una masa dagua dulce o salada, embalsada en tierra firme. Los aportes dagua a los llagos vienen-yos de les precipitaciones atmosfériques o de ríos y pequeñes fontes d'agua.
Aymara	Quta, aka pacha uraqinxa umana katuntata chiqa.
Azeri	Göl qurunun səth suları və yeraltı sularının toplandığı çökək hissəsidir. Göllər Dünya Okeanın hissəsi hesab olunmur.
Bashkir	Күл, һыу алмашыныуы аҡрын барған, донъя океаны менән бәйләнеше булмаған тәбиғи һыу ятҡылығы. Әгәр күлдән йылға баш алһа, былдай күл ағын тип атала, әгәр йылға башланмаһа -- тоҡон күл була. Барлыҡҡа килеүе буйынса тектоник, янартау, боҙлоҡ, быуылған, карст, уйһыулыҡ һәм башҡа төрҙәргә бүленә.
Bicolano (Central)	An danaw (Ingles hale sa lacus sa Latin) sarong hawak nin katubigan o iba pang likido sa sarong hawak nin kadagaan. An kadaklan kan mga danaw sa kinâban tubig tâbang asin haros gabos namómogták sa Norteng Hemispero. An mga darakulang danaw paminsan-minsan inaapod na mga "pankadagaan na dagat" asin an mga saradit na dagat paminsan-minsan inaapod na mga danaw.
Belarusian (Taraškievica)	Во́зера — натуральны замкнёны зборнік вады, існаваньне якога забясьпечана існаваньнем паглыбленьня ў глебе, і ў якім зьбіраецца паверхневая вада, якая сілкуецца вадой у большым аб'ёме чым траціцца праз выпарэньне ці выток. Узьнікненьне азёрных катлавін перш за ўсё зьвязана з геалягічным працэсамі. Сілкаваньне перш за ўсё залежыць ад кліматычных умоў.
Bengali	হ্রদ (ইংরেজি ভাষায়: Lake) হল ভূ-বেষ্টিত লবণাক্ত বা মিষ্টি স্থির পানির বড় আকারের জলাশয়। হ্রদ উপসাগর বা ছোট সাগরের মত কোন মহাসমুদ্রের সাথে সংযুক্ত জনয়, তাই এতে জোয়ার ভাটা হয় না। বিভিন্ন ভূ-তাত্ত্বিক কারণে মাটি নিচু হয়ে হ্রদের সৃষ্টি হতে পারে। স্তরীভূত শিলায় ভাঁজের সৃষ্টি হয়ে, অনেক বড় আকারের শিলাস্তর ফল্টের আকারে স্থানচ্যুত হলে, কিংবা ভূমিধ্বসের ফলে পাহাড়ী নদীর গতিপথে প্রতিবন্ধকতার সৃষ্টি হয়ে হ্রদের সৃষ্টি হতে পারে।

The information obtained includes alternative spellings (e.g. "*encyclopædia*" for "*encyclopedia*") and common misspellings (e.g. "*miniscule*" for "*minuscule*"). These are sometimes found in the English glosses of words, and in other cases there are entire sections listing alternative spellings. Capturing them in the lexical database allows answering a greater number of user queries and the user can be directed to the expected information rather than possibly receiving an empty search result list.

A list of relations mined from the English Wiktionary (2009-03-10 XML dump) is given in Table 3.

Table 3: Information mined from the English Wiktionary

Relation	Relationship Instances	Terms
Translation	1,232,167	372,554
Derivation	1,185,851	979,604
Etymology	394,430	326,135
Synonymy	257,372	128,855
Orthographic variant	63,950	43,340
Antonymy	37,026	20,126
Misspelling	1,140	1,099

4.3. Etymological links

In addition to semantic relationships between words, it is useful to have etymological links representing how words originated from other previously existing words. By navigating this network, one can easily see that the English word "*texting*" (the use of a mobile phone to send text messages) derives from the word "*text*" (via "*text messaging*"), which in turn comes from the Latin word "*textus*", which is a derivation of "*texo*", "*texere*" (to weave, intertwine). From there, other cognate forms can be discovered, e.g. the Spanish word "*teja*" (roof tile). Similarly, one can observe the relationship between the English word "*muscular*" and the German word "*Fledermaus*" (bat in its animal sense) in Figure 3. While the two words are semantically unrelated, etymologically, they both evolved out of Indo-European words for "*mouse*".

We obtain etymological links between different words again by mining the English version of Wiktionary. In this case, however, the extraction is even more challenging. The main location for etymological information are the respective Etymology sections within articles, which we recursively parse using a set of regular expressions. Recursion is used because a single etymology can recount a chain of etymological relationships, tracing each word to an earlier form. Since these sections can contain arbitrary written text, the parsing does not always succeed, however there are a significant number of recurring patterns that can be identified in Wiktionary. Often, etymological information about a word is found not on the word's respective Wiktionary page, but on the page of some other word. As an example, the fact that the Latin word "*salarium*" is derived from the word "*sal*" (salt) is found on the page for the English word "*salary*".

Apart from the Etymology sections, there are sometimes separate sections that list derived words and cognate forms. Finally, the glosses given to words in Wiktionary are also parsed, as these often hold links to root forms for deriva-

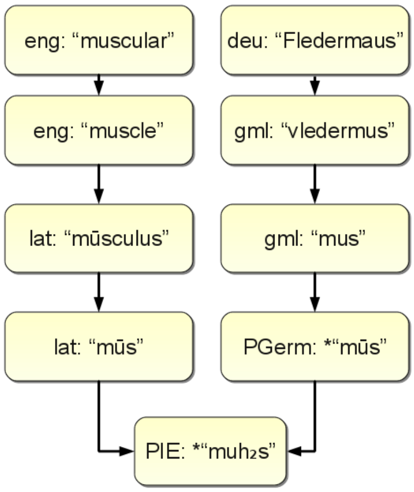


Figure 3: Example of etymological links between words. The Latin word for muscles and the German word for bats (the animals) are both derived from words referring to mice.

tions. For instance, the English word "*booking*" is linked to the verb "*to book*". Extraction result statistics, again, can be found in Table 3. The extracted information has been made available as a lexical resource in its own right, called Etymological Wordnet (de Melo and Weikum, 2010).

4.4. External Links

External links are provided to Wiktionary and Wikipedia whenever a corresponding article has been identified. Additionally, we have extracted links to ontologies and knowledge bases, including SUMO (Niles and Pease, 2001), OpenCyc (Cycorp Inc., 2009), DBpedia (Auer et al., 2007), and YAGO (Suchanek et al., 2007). The latter two are based on the mappings established in Section 3.3.1. While links to ontologies are unlikely to be useful for ordinary users, they can be used to retrieve additional background information that can aid the user.

5. Multimodal Data

5.1. Images and Video

Multimodal data like pictures and videos often are invaluable for highlighting the meaning of a word. For instance, WordNet describes "*jaguar*" as "*a large spotted feline of tropical America similar to the leopard; in some classifications considered a member of the genus Felis*". Such glosses on their own are unlikely to suffice for identification of real world entities, in this case jaguars, without additional information. For this reason, we attempt to augment WordNet with multimodal data by using simple techniques to retrieve representative multimodal data.

5.1.1. Web Services

One approach is to harvest multimodal data using services available on the Web. Currently we use the Flickr[4] API for images and the Google Data API for YouTube[5] videos.

Given a WordNet synset s, we attempt to issue queries to find relevant images. Let $\mathrm{T}(s)$ denote the set of terms of s in WordNet, $\mathrm{T}'(s)$ denote hypernym/class terms for s in WordNet, and $\mathrm{S}(t)$ denote the set of senses for a term t.

As web service queries, we consider all subsets $Q \subseteq \mathrm{T}(s) \cup \mathrm{T}'(s)$ where $Q \cap \mathrm{T}(s) \neq \emptyset$. Each query $q \in Q$ has an inherent score $\mathrm{score}(q) = \sum_{t \in Q} \frac{1}{|\mathrm{S}(t)|}$. The score for a search result r is $\mathrm{score}(q, r) = w(q, r)\mathrm{score}(q)$ with an additional weighting $w(q, r)$ that is service-specific. For Youtube, the weighting function w is based on the user rating of videos (scaled to $[0, 1]$), whereas for Flickr we used the reciprocals of the number of tags and of the description text length to downgrade images with too many tags or overly long description texts.

In order to minimize the search space, we precheck each term $t \in \mathrm{T}(s) \cup \mathrm{T}'(s)$ and each term pair $t, t' \in \mathrm{T}(s) \cup \mathrm{T}'(s)$ $(t \neq t')$ to see which combinations do not occur at all and do not need to be evaluated in combination with other terms. We also optimize the order such that we can decide to avoid issuing a web service query q if its $\mathrm{score}(q)$ dictates that it will not be able to make the top k result list (k was set to 10 in our setup).

We found that this approach works best for concrete nouns. Results for Flickr were superior to those for Youtube, because videos on Youtube tend not to be the type one would use to portray an entity in an encyclopedic manner.

5.1.2. Wikipedia

Another even more reliable source of images is Wikipedia. Images that appear on an article's page tend to be well-suited for explaining the meaning of a word, as they tend to show prototypical and noteworthy examples. An example can be seen in Figure 2.

We use the matching technique from Section 3.3.1 to identify Wikipedia articles matching a WordNet synset, and then extract image links from the article source texts. For roughly 13,000 synsets, we obtain a total of around 57,000 images. Additionally, for around 1000 synsets we obtain around 1,300 relevant video and audio recordings. For instance, the synset for the tambourine instrument is linked to a recording of tambourine music, and the synset for koala bears is linked to a video portraying the animal's behaviour.

Bond et al. (2008) suggest using images also as illustrations of hypernyms of the senses they are assigned to, e.g. an image of a cheetah could also be used as an image for mammal. This leads to a greater coverage.

5.2. Audio Recordings

From Wiktionary, we extract pronunciation audio recordings for many words. Such information is generally found in the Pronunciation sections, embedded using templates.

In total, Wiktionary gives us over 42,000 audio files. In some cases, there are multiple recordings for a single written form due to different pronunciations being possible (e.g. British Received Pronunciation vs. Standard American). In a few cases, there are multiple recordings due to homography. Currently, recordings are tied to words rather than specific word senses. It is not clear whether techniques can be developed to disambiguate such recordings reliably.

[4] http://www.flickr.com/

[5] http://www.youtube.com

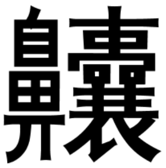


Figure 4: Example Character Image: Unicode character U+9F49, a complex Chinese character (Pinyin: nàng) referring to poor pronunciation due to a blocked nose. Taken from the open source Wen Quan Yi 'Zen Hei' font.

5.3. Geographical Data

In many Wikipedia articles, geographic coordinates are embedded using specific templates. Relying on the Wikipedia links from Section 3.3.1, we are able to assign geographical coordinates to synsets for countries, cities, places, etc. With the March 2010 dumps, we obtained geographical coordinates for 985 synsets.

In our user interface, the user can click on the coordinates to obtain an interactive map and satellite image interface provided by third parties on the Internet. Additionally, the image data from Section 5.1 provides pre-existing schematic maps that quickly allow a user to gain an understanding of an entity's geographical location in the world. The same is possible for language synsets: a user can quickly see where in the world a language is spoken.

5.4. Character Information

For languages that rely on complex scripts, especially ideogrammatic or logographic ones like the Han characters used in Chinese, Japanese, and Korean, it is useful to have additional information about characters.

There are relatively complete open source fonts available for this purpose, e.g. the Freefont Serif font. For all Unicode code points covered, the respective character representations are rendered in a high resolution and stored as image files. Hence, even for complex characters like U+9F49, displayed in Figure 4, which are missing in many fonts, the user can discern the individual strokes that make up the character. The Kanji Stroke Orders Font is particularly useful for this purpose, as it embeds small indications of the standard order in which strokes are written in Japanese.

Additionally, we also capture information from the Unicode and Hanzi Data databases about character composition, e.g. the Chinese "娴" is linked to its radical part "女" and to its pronunciation component "闲". Links between variants are also stored, e.g. between "闲" and "閑". Such composition information is particularly useful for language learners attempting to memorize how to write Chinese characters.

6. Related Work

There are many existing systems for browsing and querying lexical knowledge bases, but most are monolingual. For the MultiWordNet project (Pianta et al., 2002), a simple browsing interface exists that supports multiple languages. As of 2010, less than 10 languages are provided. The WordNet Management System (WNMS) by Robkop et al. (2010) provides a sophisticated graphical way of browsing WordNet in two languages simultaneously. Ayewah et al. (2003) present a user interface for building a Romanian WordNet by showing suggestions to humans. Our system is based on the idea that query interfaces extending WordNet by providing sense-specific translations in many languages and by additionally capturing other types of information like images, pronunciation, and etymological information, lead to a better user experience.

Wu and Weld (2008) link Wikipedia's infobox templates to WordNet, and (Ponzetto and Navigli, 2009) link Wikipedia Category pages to WordNet. However, these studies do not attempt to match regular Wikipedia articles with WordNet synset, which is required in order to obtain large numbers of glosses. Giménez and Màrquez (2006) provided a sophisticated study of statistical machine translation to obtain Spanish glosses. The Google translation services we rely on are also based on statistical machine translation, but are trained on very large corpora.

Zinger et al. (2005) create images for certain WordNet synsets by querying an image search engine for terms. However, the only technique they use to reduce ambiguity of queries is appending the headword of the parent synset whenever the synset itself only has a single term headword. The main strength of their system lies in their subsequent use of image analysis and clustering techniques to find the most prototypical images. Bond et al. (2008) matched file and directory names of clipart images with words and their hypernyms in WordNet. The resulting resource contains very suitable illustrations for certain synset, but the coverage is limited to less than 800 synsets. The most sophisticated association of images to WordNet synsets was provided by Deng et al. (2009). They used Web search engines and human input via Amazon Mechanical Turk to obtain suitable images for around 5,000 synsets. For our interface, the fact that the licensing conditions of these images are unclear was problematic. Approaches that instead rely on Wikipedia or Flickr can be configured to deliver only images with open license models.

Several authors (Schmitz, 2006; Damme et al., 2008) have attempted to start out with tag statistics to induce a lightweight ontology. The resulting taxonomies of tags, however, are still noisy, sparse, and flat compared to WordNet and to many alternative strategies for building ontologies. Furthermore, such works do not address how to assign media to the output taxonomies when tags are ambiguous, which is the main focus of our work.

Buscaldi and Rosso (Buscaldi and Rosso, 2008) used domain-specific heuristics to link WordNet synsets for geographical locations to corresponding Wikipedia articles for those locations.

7. Concluding Remarks

We presented a comprehensive lexical database querying and browsing system that provides multilingual terms and glosses, rich links between different resources, as well as multimodal information like images and audio recordings.

As this is a research prototype, the target group currently consists of people somewhat familiar with lexical databases. However, this could easily be adapted by dropping certain types of information (e.g. links to ontologies) and using simpler but perhaps slightly less accurate labels and descriptions (e.g. language names instead of ISO codes, "broader term" instead of "hypernym"). In future work, we would like to improve the customization capabilities of the interface to cater to different target groups. Also, since our interface includes terms in a lot of rare minority languages, it makes sense to display terms in the user's languages of choice in a more prominent position, possibly blending out all others. Some of this information could be learnt automatically given the users' previous interactions with the system as well their IP addresses, which often reveal their geographical location.

Finally, we would like to explore additional information sources, e.g. one could investigate whether Flickr images can be used to provide geographical locations for places and landmarks, and whether satisfactory audio recordings can be generated using speech synthesis software. Such information could further enhance the user experience.

8. 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. Conference on Recent Advances in NLP 1997*, pages 143–149.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary G. Ives. 2007. DBpedia: A nucleus for a web of open data. In *Proc. ISWC*, volume 4825 of *LNCS*, pages 722–735. Springer.

Nathaniel Ayewah, Rada Mihalcea, and Vivi Nastase. 2003. Building multilingual semantic networks with non-expert contributions over the web. In *Proc. KCAP 2003 Workshop on Distributed and Collaborative Knowledge Capture*.

Francis Bond, Hitoshi Isahara, Kyoko Kanzaki, and Kiyotaka Uchimoto. 2008. Boot-strapping a WordNet using multiple existing WordNets. In *Proc. LREC 2008*, Marrakech, Morocco. European Language Resources Association (ELRA).

Davide Buscaldi and Paolo Rosso. 2008. Geo-WordNet: Automatic georeferencing of WordNet. In *Proc. LREC*, Marrakech, Morocco.

Cycorp Inc., 2009. *OpenCyc*. http://www.opencyc.org/.

Céline Van Damme, Tanguy Coenen, and Eddy Vandijck. 2008. Turning a corporate folksonomy into a lightweight corporate ontology. In Witold Abramowicz and Dieter Fensel, editors, *BIS*, volume 7 of *Lecture Notes in Business Information Processing*, pages 36–47. Springer.

Gerard de Melo and Gerhard Weikum. 2009a. Extracting sense-disambiguated example sentences from parallel corpora. In Gerardo Sierra, María Pozzi, and Juan-Manual Torres-Moreno, editors, *Proc. Workshop on Definition Extraction at RANLP 2009*, pages 40–46.

Gerard de Melo and Gerhard Weikum. 2009b. Towards a universal wordnet by learning from combined evidence. In *Proc. 18th ACM Conference on Information and Knowledge Management (CIKM 2009)*, pages 513–522, New York, NY, USA. ACM.

Gerard de Melo and Gerhard Weikum. 2010. Towards universal multilingual knowledge bases. In *Principles, Construction, and Applications of Multilingual Wordnets. Proceedings of the Fifth Global WordNet Conference (GWC 2010)*, pages 149–156, India. Narosa Publishing.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009)*.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.

Jesús Giménez and Lluís Màrquez. 2006. Low-cost enrichment of spanish wordnet with automatically translated glosses: combining general and specialized models. In *Proc. COLING/ACL 2006*, pages 287–294. ACL.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet — a lexical-semantic net for German. In *Proc. ACL Workshop Automatic IE and Building Lexical Semantic Resources for NLP Applications*.

Ian Niles and Adam Pease. 2001. Toward a Standard Upper Ontology. In Chris Welty and Barry Smith, editors, *Proc. 2nd Intl. Conf. on Formal Ontology in Information Systems (FOIS)*.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. MultiWordNet: Developing an aligned multilingual database. In *Proc. 1st Intl. Global WordNet Conference, Mysore, India*, pages 293–302.

Simone Paolo Ponzetto and Roberto Navigli. 2009. Large-scale taxonomy mapping for restructuring and integrating Wikipedia. In *Proc. IJCAI*, pages 2083–2088, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Kergrit Robkop, Sareewan Thoongsup, Thatsanee Charoenporn, Virach Sornlertlamvanich, and Hitoshi Isahara. 2010. WNMS: Connecting the distributed wordnet in the case of Asian WordNet. In *Principles, Construction, and Applications of Multilingual Wordnets. Proceedings of the Fifth Global WordNet Conference (GWC 2010)*, India. Narosa Publishing.

Patrick Schmitz. 2006. Inducing ontology from Flickr tags. In *Proceedings of the Workshop on Collaborative Tagging at WWW2006*, Edinburgh, Scotland.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. YAGO: A Core of Semantic Knowledge. In *16th International World Wide Web conference (WWW 2007)*. ACM Press.

Piek Vossen, editor. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Springer.

Fei Wu and Daniel S. Weld. 2008. Automatically refining the Wikipedia infobox ontology. In *Proc. WWW*, pages 635–644, New York, NY, USA. ACM.

Svitlana Zinger, Christophe Millet, Benoit Mathieu, Gregory Grefenstette, Patrick Hède, and Pierre-Alain Moëllic. 2005. Extracting an ontology of portrayable objects from wordnet.