Analyzing and Accessing Wikipedia as a Lexical Semantic Resource

Torsten Zesch, Iryna Gurevych, Max Mühlhäuser

Department of Telecooperation
Ubiquitous Knowledge Processing Group
Darmstadt University of Technology, Hochschulstraße 10
D-64289 Darmstadt, Germany
{zesch,gurevych,max} (at) tk.informatik.tu-darmstadt.de

Abstract. We analyze Wikipedia as a lexical semantic resource and compare it with conventional resources, such as dictionaries, thesauri, semantic wordnets, etc. Different parts of Wikipedia reflect different aspects of these resources. We show that Wikipedia contains a vast amount of knowledge about, e.g., named entities, domain specific terms, and rare word senses. If Wikipedia is to be used as a lexical semantic resource in large-scale NLP tasks, efficient programmatic access to the knowledge therein is required. We review existing access mechanisms and show that they are limited with respect to performance and the provided access functions. Therefore, we introduce a general purpose, high performance Java-based Wikipedia API that overcomes these limitations. It is available for research purposes at http://www.ukp.tu-darmstadt.de/software/WikipediaAPI.

1 Introduction

Wikipedia[1] is a free, multilingual online encyclopedia constructed in a collaborative effort of voluntary contributors. It grows exponentially and has probably become the largest collection of freely available knowledge. Wikipedia shares many of its properties with other well known lexical semantic resources (such as dictionaries, thesauri, semantic wordnets, and conventional encyclopedias), but combines them in a unique way. This makes Wikipedia a promising lexical semantic resource that has already been used for such different natural language processing (NLP) tasks as question answering (Ahn et al., 2004), text classification (Gabrilovich and Markovitch, 2006), or named entity disambiguation (Bunescu and Pasca, 2006).

As Wikipedia is relatively new,[2] a detailed analysis of its properties is necessary to enable its use as a lexical semantic resource for NLP. In section 2, we compare Wikipedia and conventional lexical semantic resources and show that different parts of Wikipedia reflect different aspects of these resources. Section 3

[1] http://www.wikipedia.org
[2] The German Wikipedia was founded in March 2001.

reviews existing access mechanisms for Wikipedia and shows that they are limited with respect to performance and the provided access methods. Therefore, section 3.2 introduces a general purpose, high performance Java-based API that overcomes these limitations.

2 Wikipedia as a Lexical Semantic Resource

2.1 Comparison of Resources

In order to compare Wikipedia with other lexical semantic resources, we first characterize these resources and the types of knowledge represented therein.

The simplest form of a lexical semantic resource is a **dictionary** (e.g., the Longman Dictionary of Contemporary English)[3]. It lists all lexical entities in a domain, connects them with their semantic meaning via a defining gloss, and enumerates all senses in case of polysemous entities. Like a dictionary, a **thesaurus** (e.g., Roget's Thesaurus)[4] lists lexical entities, but additionally categorizes them into topical groups by means of lexical relations like synonymy, or hypernymy and hyponymy. A **semantic wordnet** (e.g., GermaNet (Kunze, 2004)) displays features of the aforementioned simpler resources. Like a dictionary, it offers an account of lexical units, their senses and sometimes even short glosses. Additionally, lexical units and senses are organized in a thesaural structure. Furthermore, Ruiz-Casado et al. (2005) have proposed to add encyclopedic features to wordnets by augmenting WordNet (Fellbaum, 1998) entries with Wikipedia articles.

Encyclopedias (e.g., Encyclopædia Britannica)[5] offer a detailed description of each lexical entry. Encyclopedias are often not freely available and lack the coverage of Wikipedia, but they are subject to strict editorial control by the publisher resulting in high quality articles. However, Giles (2005) showed that the quality of Wikipedia articles is comparable to those in Encyclopædia Britannica. Furthermore, Wikipedia articles are up-to-date, reflecting also changes caused by very recent events.

Wikipedia is primarily an encyclopedia with the additional benefit of heavy linking between articles and without the size constraints of paper articles. Recent work has explored the use of explicitly labeled links between articles (Völkel et al., 2006). This would turn Wikipedia into a huge semantic net, but this feature has not been added to the Wikipedia software yet.

Due to an editorial decision,[6] Wikipedia contains only terms of encyclopedic interest. Hence, it is not a dictionary.[7] Wikipedia covers mainly nouns and

[3] `http://www.ldoceonline.com`

[4] `http://thesaurus.reference.com`

[5] `http://www.britannica.com`

[6] `http://en.wikipedia.org/wiki/WP:WWIN`

[7] Wiktionary `http://www.wiktionary.org/` is an online dictionary, constructed in the same collaborative, bottom-up process like Wikipedia. It can also be used in NLP, e.g., for sentiment classification (Chesley et al., 2006).

only few adjectives and verbs. In most cases, they redirect to their corresponding nouns, e.g., the verb "sehen" (Eng. *to see*) redirects to the phrase "Visuelle Wahrnehmung" (Eng. *visual perception*) in the German Wikipedia.

Dictionaries, thesauri, and wordnets focus on general vocabulary, while Wikipedia covers a larger number of named entities and domain specific terms, such as: *Gentest* (Eng. *DNA test*), *Makake* (*Macaque*), *Kortex* (*Cortex*), *Kompaktvan* (*Compact van*), *Nanopartikel* (*Nanoparticle*), or *Welthungerhilfe* (*German Agro Action*).

Wikipedia also covers specific senses of common terms. The term "Wald" (Eng. *forest*) has only one sense ("an area with trees") in GermaNet. In contrast, four senses are listed in Wikipedia, including, e.g., a special sense denoting "data structure in computer science". Additionally, Wikipedia lists more than ten geographical entities with the name "Wald" and four persons with exactly that surname.

Another excellent source of lexical semantic information in Wikipedia are article redirects, as they express synonymy, spelling variations and abbreviations. For example, the article about the current pope *Benedikt XVI* has more than 10 redirects including spelling variations like *Papst Benedikt XVI.*, or *Papst Benedikt 16.* Furthermore, his secular name *Joseph Ratzinger* and various combinations like *Kardinal Joseph Ratzinger* or *Joseph Kardinal Ratzinger*, as well as common misspellings like *Josef Ratzinger* are included. This example indicates the potential of Wikipedia redirects to improve named entity recognition and co-reference resolution.

2.2 Types of Lexical Semantic Information in Wikipedia

Table 1 gives an overview of the types of lexical semantic information found in Wikipedia, which will be described in the following in more detail.

The **first paragraph** of a Wikipedia article usually contains a short definition of the term the article is about. The **full article** text contains related terms and describes the meaning of the article term in detail. It may even contain translations of the article term encoded in the links to Wikipedia in other languages, turning Wikipedia into a valuable multilingual resource.

Article links are another source of lexical semantic information in Wikipedia. Article links point from one article to another article. Therefore, a link establishes a relation between the two terms the articles are about. Links between Wikipedia articles are untyped. Thus, they express semantic relatedness, but the type of the relation and the degree of strength is unknown. All links between Wikipedia articles form a graph that can be used, e.g., to compute the similarity of two terms from their positions in the graph (Page et al., 1998, Jeh and Widom, 2002). On a Wikipedia HTML page, each link is visualized as a highlighted text that can be clicked. The highlighted text (also called **link label**) does not necessarily have to be the same as the title of the article that it points to. For example, many links referring to the article with the title *Deutschland* are actually labeled *Bundesrepublik Deutschland*. As a result, a link label may provide information about synonyms, spelling variations or related terms.

Sources	Lexical semantic information
Articles	
- First paragraph	Definition
- Full text	Description of meaning; related terms; translations
- Redirects	Synonymy; spelling variations, misspellings; abbreviations
- Title	Named entities; domain specific terms or senses
Article links	
- Context	Related terms; co-occurrences
- Label	Synonyms; spelling variations; related terms
- Target	Link graph; related terms
Categories	
- Contained articles	Semantically related terms (siblings)
- Hierarchy	Hyponymic and meronymic relations between terms
Disambiguation pages	
- Article links	Sense inventory

Table 1. Sources of lexical semantic information in Wikipedia.

Related and co-occurring terms can be extracted from a context window around a link label, e.g., the link label *Benedikt XVI.* is often preceded by *Papst* (Eng. *pope*).

The **category system** in Wikipedia (Voss, 2006) can be viewed from two perspectives. From an article-centric perspective, each article can have an arbitrary number of categories, where a category is a semantic tag. Thus, the category system is a kind of collaborative tagging. From a category-centric perspective, each category can contain an arbitrary number of articles that semantically belong to this category. A category can have subcategories expressing meronymic or hyponymic relations. For example, a category *Fahrzeug* (Eng. *vehicle*) has subcategories like *Luftfahrzeug* (Eng. *aircraft*) or *Wasserfahrzeug* (Eng. *watercraft*). Thus, the category system forms a thesaurus. Consequently, the Wikipedia category system was called "collaborative thesaurus tagging" by Voss (ibid.). Thesaurus tagging differs from Web 2.0 (O'Reilly, 2005) collaborative tagging used by Flickr[8] or del.icio.us[9]: Tags in Wikipedia have to be chosen from the category thesaurus which is agreed upon by the community of Wikipedia users, while each user defines his own tags in collaborative tagging.

Wikipedia represents polysemous terms by using **disambiguation pages**. A disambiguation page lists all articles that exist for a certain term. As each article must have a unique title, articles about polysemous terms are usually differentiated by adding a disambiguation tag in parentheses, e.g., "Wald", "Wald (Graphentheorie)" (Eng. "Forest", "Forest (Graph theory)"). As a result, a disambiguation page forms a sense inventory for a given term. The article without disambiguation tag is usually about the most common sense of the term, i.e. it could be used as a most-frequent-sense baseline in word sense disambiguation.

[8] `http://www.flickr.com`
[9] `http://del.icio.us`

However, disambiguation pages may also contain additional links to pages that do not point to a sense of the term. Therefore, extracting a sense inventory for a given term is not straightforward. It requires to differentiate disambiguation links from other links.

2.3 Wikipedia Mining

Many types of semantic knowledge in Wikipedia are not directly available in machine readable form, but have to be extracted from Wikipedia's content or structure. We call this process **Wikipedia mining** and differentiate between **content mining** and **structure mining**. *Content mining* refers to searching the article content for relevant knowledge. This includes, e.g., using the first paragraph as a definition, using redirects for finding spelling variations, or analyzing link labels for finding synonyms. *Structure mining* refers to extracting knowledge from structural features of Wikipedia, such as the link graph or the inner structure of an article. This includes, e.g., determining the meaning of a page by means of ingoing and outgoing links on that page, or measuring the semantic similarity of two terms by computing the distance between the corresponding Wikipedia articles or categories.

The properties of Wikipedia can be summarized as follows: It contains a wide variety of lexical entities, but mainly nouns. It contains domain specific terms and senses, but lacks coverage on common concepts that are not of encyclopedic interest. Wikipedia articles express the meaning of a term by means of a gloss and describing text. The meaning is also implicitly expressed via the position of an article in the article graph or in the category graph. Links in Wikipedia are untyped, except for the links between categories that encode either a taxonomic or a meronymic relation.

3 Accessing Wikipedia

In the previous section, we have shown that Wikipedia contains many different types of lexical semantic information. If Wikipedia is to be used in large-scale NLP tasks, efficient programmatic access to these knowledge sources is required. We review existing access mechanisms (Riddle, 2006, Strube and Ponzetto, 2006, Summers, 2006, Wikimedia Foundation, 2006), and show that they suffer from insufficient performance and provide access only to some types of the available lexical semantic information.

3.1 Overview of Existing Access Mechanisms

The simplest way to retrieve a Wikipedia page is to enter a search term on the Wikipedia site (Wikimedia Foundation, 2006). However, this approach is not suited for automatic access to pages by an application. The Perl module WWW::Wikipedia (Summers, 2006) offers simple means for retrieving Wikipedia

pages by programmatically querying the Wikipedia web site. This poses enormous load on the Wikipedia servers, when used in large-scale applications. Therefore, it is discouraged by the Wikimedia foundation (Wikimedia Foundation, 2006).

A solution to the problem of high server load is to run an own Wikipedia server using publicly available database dumps. The system developed by Strube and Ponzetto (2006) follows that approach, relying on a modified version of the WWW::Wikipedia module to retrieve articles. This gives better, but still insufficient performance due to the overhead of using web protocols for delivering the retrieved pages. This may represent and efficiency bottleneck for large-scale NLP tasks.

The Perl module Parse::MediaWikiDump (Riddle, 2006) parses the Wikipedia XML dump to retrieve articles. As Wikipedia dumps are very large (over 2 GB for the snapshot of the German Wikipedia from May 2006), the performance of parsing is not sufficient for large-scale NLP tasks. Additionally, the time that is required to retrieve an article is not predictable, but depends on the position of an article in the XML dump.

Alternatively, the database dumps can be directly accessed using database to guarantee nearly constant retrieval time for each article. This approach is superior to web retrieval, as it is more efficient. In technical terms: At the time of writing, retrieving a Wikipedia page from the web usually involves a transfer of the request from an application to the web server via HTTP. The web server then executes a PHP script that accesses a database. The database returns the text with Wiki markup to the PHP script. It converts the Wiki markup to HTML. Finally, the web server delivers the data back to the application via HTTP. In contrast, directly accessing the database involves querying the database and delivering the results to the application. Using a fixed database dump has the additional benefit of making the obtained results reproducible. This is an important feature, if Wikipedia is employed for research purposes, as it is very likely that the online Wikipedia would have changed between two runs of a certain experimental setting. Another benefit of accessing the database is that it contains explicitly stored information about a page's links or categories, while they are only implicitly encoded in the HTML structure of an article retrieved via HTTP.

In the following, we introduce a system architecture that directly accesses the Wikipedia database to provide fast access to multiple types of lexical semantic information therein.

3.2 Wikipedia API system architecture

First, we transform the Wikipedia database scheme into a different representation that can be more efficiently accessed and yields optimal access to all types of lexical semantic information that were identified in section 2. Then, we access the transformed database using object-relational mapping explained below. The **advantages** of our system architecture as opposed to the approaches outlined above are: i) decoupling the API implementation from changes in the MediaWiki

software underlying Wikipedia, ii) making research results reproducible, iii) explicitly storing the information, scattered in the original database structure, like redirects, and iv) computational efficiency for large-scale NLP tasks. Figure 1 gives an overview of the system architecture.

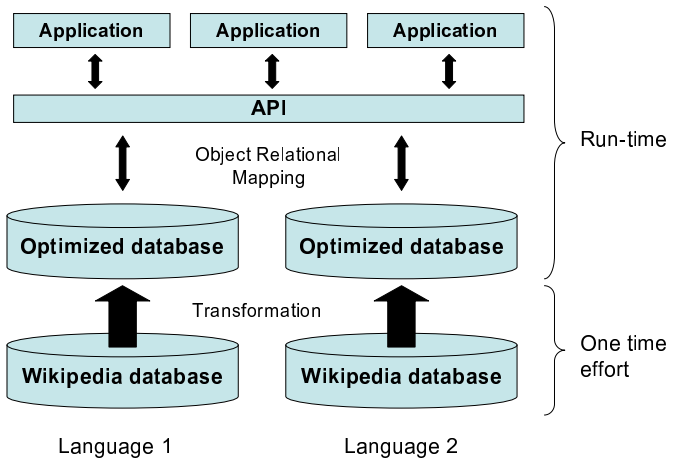

Fig. 1. System architecture of the Wikipedia API.

Database transformation Even if accessing articles using the Wikipedia database is more efficient than access via HTTP or parsing XML dumps, some types of lexical semantic information for NLP purposes can only be extracted with high computational and implementation effort. This is because the original Wikipedia database structure is optimized for page retrieval and page editing, which are the most common activities when using Wikipedia. For example, article redirects are not explicitly stored in the database. Each redirect is only implicitly represented as an article, which solely contains a link to another page. A page with the title *Benedikt 16.* may contain the text "[[Redirect: Benedikt XVI]]" meaning that it does not contain any content, but redirects to the page *Benedikt XVI.* This is a good representation for efficiently retrieving Wikipedia articles given a search term, but getting all redirects for a given page is very costly. We would have to parse all redirect pages and extract the articles they point to. Given the high number of redirects, this cannot be done at runtime.

In the transformed database scheme, redirects and other sources of lexical semantic information are explicitly stored with the corresponding article making them easily accessible. The transformation step has the additional benefit of decoupling the API implementation from possible changes of the Wikipedia database scheme. It has changed fundamentally with Version 1.4 of the Media-

wiki software underlying the Wikipedia website, and further changes are to be expected.

Wikipedia is available in different languages and each version has its peculiarities. For example, the top-most category is called *Categories* in English Wikipedia, while it is *!Hauptkategorie* in German Wikipedia. The transformation step maps all language specific features into a generalized representation and externalizes necessary language dependent settings. We plan to make the transformation software available, too. Then, necessary changes for other languages can be performed by the research community, turning Wikipedia in a multilingual lexical semantic resource.

Transformation of the database structure is a one-time effort. Afterwards, the database can be accessed using *object-relational mapping* as explained in the next paragraph.

Object-relational mapping (ORM) bridges the impedance mismatch between relational databases and object-oriented programming languages. The impedance mismatch occurs because relational databases store data as "rows and columns", while object-oriented programming relies on complex objects. We cannot read these objects directly from the database. At this point, ORM can be applied. A mapping file tells the relational database how a complex object should be mapped to a relational database scheme. Thus, objects can be read from the database in a transparent manner. ORM even transparently updates the database, when the object is changed in the Java program. This guarantees a high stability and maintainability of the Wikipedia API. Additionally, ORM abstracts further from the actual database structure and, thus, fully decouples the API design from a particular database or a particular underlying database scheme.

Wikipedia Application Programming Interface (API) We have developed a Java-based API relying on the previously introduced system architecture that uses an optimized database scheme. The API enables fast and efficient access to different types of lexical semantic information encoded in Wikipedia, as identified in section 2. In particular, the API provides access to Wikipedia articles, links, categories and redirects. Thereby, the category system is converted into a graph representation. On that representation, a whole range of standard graph algorithms can be applied, e.g., finding the shortest path between two given categories. A detailed description of the provided functionality and an introductory tutorial can be found at `http://www.ukp.tu-darmstadt.de/software/WikipediaAPI`.

4 Conclusion

In this article, we performed an analysis of Wikipedia as *an emerging lexical semantic resource*. We compared Wikipedia to a number of conventional lexical

semantic resources and showed that, due to the collaborative way of construction and the use of Web 2.0 principles, Wikipedia displays a set of unique features. It contains not only lexical semantic information, but also vasts amounts of domain and world knowledge. This makes Wikipedia a promising knowledge resource, which bears the potential to eliminate knowledge acquisition bottlenecks and coverage problems pertinent to existing lexical semantic resources.

As Wikipedia was not constructed specially for NLP purposes, a research effort is required to turn Wikipedia into a lexical semantic resource and further into an easily accessible knowledge base. In order to be employed in large-scale NLP tasks, different types of lexical semantic information therein have to be *identified and extracted* in a stable and computationally efficient manner. We reviewed a set of existing access mechanisms to Wikipedia and introduced *a highly efficient and extensible Java-based API*. It overcomes the limitations of the available tools and displays the following advantages: (i) decoupling the API implementation from changes in the MediaWiki software underlying Wikipedia, (ii) making research results reproducible, (iii) explicitly storing the information, scattered in the original database structure, and (iv) computational efficiency for large-scale NLP tasks.

So far, the Wikipedia API is able to extract lexical semantic information, explicitly represented in Wikipedia. Our *ongoing work* investigates the use of knowledge in Wikipedia to compute semantic relatedness of words. Measures of semantic relatedness based on Wikipedia will soon extend the API. Further extensions of the API are underway, which exploit the lexical semantic knowledge represented in the inner structure of Wikipedia articles. We call the process of explicating the knowledge encoded in Wikipedia and making it accessible to computational programs *Wikipedia mining*. We differentiate between content mining and structure mining, and expect to extend the API in both of these directions.

We believe that Wikipedia is *an invaluable multilingual NLP resource*, having the potential to substantially improve NLP applications by utilizing broad coverage lexical semantic and world knowledge. This exploration has just begun and has an exciting future ahead of it. In order to foster the research on using Wikipedia in NLP, we made the Wikipedia API *freely available* to the research community. The API and the underlying Wikipedia database are available at `http://www.ukp.tu-darmstadt.de/software/WikipediaAPI`. We hope this will help the research community to achieve rapid advances in lexical semantic processing and NLP in general.

Acknowledgments

This work was carried out as part of the project "Semantic Information Retrieval from Texts in the Example Domain *Electronic Career Guidance*" (SIR) funded by the German Research Foundation under the grant GU 798/1-2.

Bibliography

Ahn, D., Jijkoun, V., Mishne, G., Müller, K., de Rijke, M., and Schlobach, S. (2004). Using Wikipedia at the TREC QA Track. In *Proceedings of TREC 2004*.

Bunescu, R. and Pasca, M. (2006). Using Encyclopedic Knowledge for Named Entity Disambiguation. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 9–16, Trento, Italy.

Chesley, P., Vincent, B., Xu, L., and Srihari, R. (2006). Using Verbs and Adjectives to Automatically Classify Blog Sentiment. Technical Report SS-06-03, AAAI Spring Symposium.

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

Gabrilovich, E. and Markovitch, S. (2006). Overcoming the Brittleness Bottleneck using Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge. In *AAAI*, pages 1301–1306, Boston, MA.

Giles, J. (2005). Internet encyclopaedias go head to head. *Nature*, 438(7070):900–901.

Jeh, G. and Widom, J. (2002). SimRank: A Measure of Structural-Context Similarity. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Edmonton, Alberta, Canada.

Kunze, C. (2004). *Computerlinguistik und Sprachtechnologie*, chapter Lexikalisch-semantische Wortnetze, pages 423–431. Spektrum Akademischer Verlag.

O'Reilly, T. (2005). What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. URL `http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web%-20.html`.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1998). The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford Digital Library Technologies Project.

Riddle, T. (2006). Parse::MediaWikiDump. URL `http://search.cpan.org/~triddle/Parse-MediaWikiDump-0.40/`.

Ruiz-Casado, M., Alfonseca, E., and Castells, P. (2005). Automatic Assignment of Wikipedia Encyclopedic Entries to WordNet Synsets. In *AWIC*, pages 380–386.

Strube, M. and Ponzetto, S. P. (2006). WikiRelate! Computing Semantic Relatedness Using Wikipedia. In *AAAI*, pages 1419–1424, Boston, Massachusetts.

Summers, E. (2006). WWW:Wikipedia. URL `http://search.cpan.org/~esummers/WWW-Wikipedia-1.9/`.

Völkel, M., Krötzsch, M., Vrandecic, D., Haller, H., and Studer, R. (2006). Semantic Wikipedia. In *Proceedings of the 15th International Conference on World Wide Web*, Edinburgh, Scotland.

Voss, J. (2006). Collaborative thesaurus tagging the Wikipedia way. *ArXiv Computer Science e-prints*, URL `http://arxiv.org/abs/cs.IR/0604036`.

Wikimedia Foundation (2006). Wikipedia. URL `http://en.wikipedia.org/wiki/Wikipedia:Searching`.