WikiWord
Multilingual Image Search and More

Daniel Kinzler

Wikimedia Deutschland e.V., Eisenacher Str. 2, 10777 Berlin, Germany

SUMMARY

This paper describes a method for building a multilingual thesaurus from the data of Wikipedia projects in different languages as well as the use of such a thesaurus to facilitate language independent retrieval of media content such as images. Of special interest shall be the relations of terms (labels, words) to language-independent concepts as well as relations between such concepts, such as abstraction and similarity. This paper describes the methods used to build the thesaurus and gives an overview of the actual data collected using that method. It also describes a web application prototype for retrieving media content from Wikimedia Commons using one of the supported languages to describe concepts.

KEY WORDS: Wikipedia; Thesaurus; Information Retrieval; Multilingual Search
CR-Classification: H.3.1, H.3.3, I.2.7

1. INTRODUCTION

A multilingual thesaurus is a powerful tool to map words and phrases from documents and user queries to concepts from an indexing vocabulary. This paper focuses on using such a thesaurus to find images on Wikimedia Commons, using search terms in a language of the user's choice as well as topic based navigation.

Wikipedia is well suited as a corpus for mining such a multilingual thesaurus (Nakayama et al. 2007A, Gregorowicz and Kramer 2006): it describes on concept per page, and the pages (and thus concepts) are highly interlinked. Also, Wikipedia is available in multiple languages, and mappings are provided between the concepts described in individual languages. This makes it easy to derive the conceptual structure of a thesaurus, as well as the lexical mapping of terms to concepts.

This paper is structured into three parts:

1.1. Building the Thesaurus

The first section describes how a thesaurus can be derived from Wikipedia. The focus is on what kinds of pages exist on on Wikipedia, how they relate to each other, and how different types of links between Wikipedia pages are interpreted and processed. Later in the section we

discuss how a multilingual thesaurus may be derived by merging the thesauri built for individual languages.

1.2. Using the Thesaurus

In this section, a prototype for a web application is described, which may be used to find media on Wikimedia Commons, and to browse the repository by topic. The key feature of this application is that it is designed to allow the user to search and navigate using their language of preference, even though Wikimedia Commons uses the English language nearly exclusively.

1.3. Evaluating the Thesaurus

Finally, the methods used to evaluate the thesaurus will be explained. Further, the result of the evaluation is discussed and opportunities for further research identified. The focus will be on the quality of the aspects of the thesaurus that will be used for the task at hand, namely, searching and navigating media on Wikimedia Commons.

2 BUILDING THE THESAURUS

This section describes how to extract a multilingual thesaurus from Wikipedia. For this purpose, a software system called *WikiWord* has been designed (Kinzler 2008). WikiWord extracts the relations needed for a multilingual concept-based thesaurus in an efficient manner. The basic idea is to limit the extraction to mining the different types of hyperlinks that may be used in Wikipedia articles:

- Simple wiki links provide information about the semantic context of concepts — they can be used to derive semantic relatedness, among other things (compare Gregorowicz and Kramer 2006).
- The link text provides information about which terms or labels (the link text) refer to which concept (the link target) (see Mihalcea 2007, Eiron and Mccurley 2003, as well as Kraft and Zien 2004).
- Categorization links provide a relation of abstraction (compare Ponzetto and Strube 2007).
- Language links connect descriptions of the same concept (or similar concepts) in different languages — this allows semantic similarity of concepts to be detected, between languages as well as within a single language (i.e. within a wiki). They can thus be used to build a multilingual thesaurus by combining concepts from different languages into language independent (or *pan-lingual*) concepts (compare section 2.3). This appears to be a novel approach introduced by Kinzler (2008).

Besides wiki links, additional properties of wiki pages may be used, for example, the templates used on on the page. These can be used especially well for classifying pages and concepts. Also, redirects, disambiguation pages as well as some special "magic words" may be used to assign additional terms to a concept.

Focusing on wiki links and other markup elements avoids a number of difficulties encountered by classical methods of automatic thesaurus generation (Nakayama et al. 2007a): most importantly, no natural language processing needs to be performed, and even problematic tasks on the lexical level, like stemming, are avoided. Thus, the method is completely indepentent of the content language. Also, human knowledge encoded in the application of markup as well as rules and conventions is utilized as directly as possible, instead of relying on statistics-based heuristics.

2.1 Analysis of Wiki Text

The first step is parsing and analyzing the wiki text (markup) from the XML-dump as provided on download.wikimedia.org. In this step, relevant information is extracted from the wiki text and recorded in the *resource model* which represents the wiki page and provides direct access to the properties of the page. Most of the textual content is discarded here, the page is regarded as an unsorted collection of *features* such as links, categories, templates, etc.
The resource model offers access to a number of properties of the page, most importantly the following: the title and namespace, all links (including categorization and inter-language links), and all templates (including parameters).

This information is extracted mainly using simple *pattern matching* applied to wiki text. Based upon this information, more specific properties are determined (mostly also using pattern matching):

- the *type of the page*. This property determines how the page is further processed. Possible page types (resource types) are: *article*, *redirect*, *disambiguation*, *list*, and *category*.
- a set of terms that can be determined as labels for the concept directly from the page itself, especially from the title.
- the first sentence of the page's text, for use as the *definition* (*gloss*) of the concept described by an article.
- if the page is a disambiguation page, a list of disambiguation-links, that is, links to pages describing possible meanings of the term that is being disambiguated.
- if the page is a redirect, the target of the redirect.

Some of the patterns used to detect specific pieces of information in the wiki text are specific to a wiki project. Note that such patterns are generally not determined *ad hoc*, but are governed by the wiki markup or model explicit conventions and guidelines of individual wiki projects.

2.2 Building the Thesaurus Structure

The thesaurus structure is maintained in a relational database. It represents the relations between terms and concepts (the meaning-of or *signification* relation) as well as several types of relations between concepts (such as *abstraction, similarity* or *relatedness*). These relations are derived from the properties in the resource model – mainly from the different types of links, but in case of the signification, also from page titles and other sources.

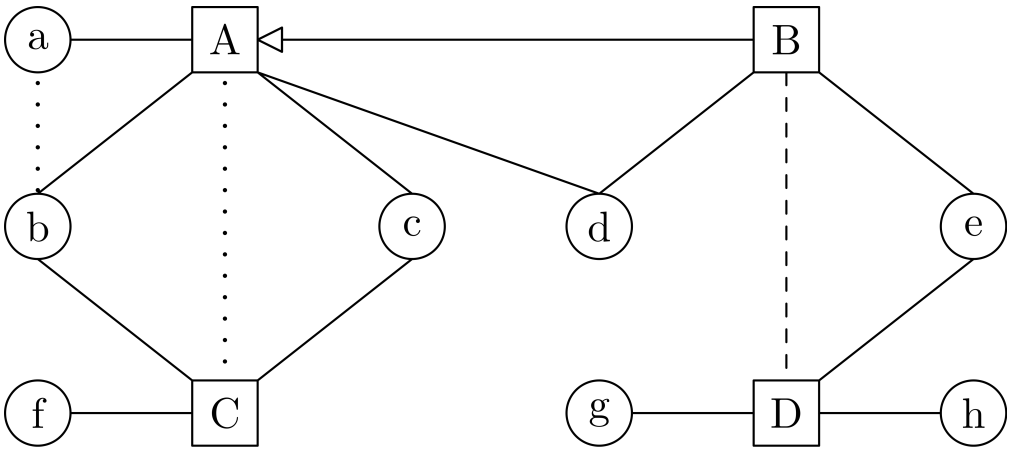

Fig. 1. This diagram shows an example structure of terms and concepts: Circles are terms and rectangles are concepts. Solid lines show the meaning-of (*signification*) relation between terms and concepts. The arrow between A and B shows the broader-than relation (*abstraction*), the dashed line between B and D shows relatedness. The dotted lines show implicit relations: a and b are synonyms because they both signify A; A and B are homonyms because both are meaning of (i.e. *signified by*) b (and c). Several more homonymous and synonymous pairs in the diagram remain unmarked for reasons of clarity.

In this step, a mono-lingual thesaurus is derived from each Wikipedia. This is done by interpreting the information from the resource model as follows:

- Pages that are not redirects, disambiguations, lists, categories or otherwise special, are considered to be articles. That is, it's assumed that they describe exactly one concept. Consequently, a concept record is created for each such page (similar to Gregorowicz and Kramer 2006).
- The relation of subsumption is derived directly from the categorization of pages. No difference is made with respect to categorization between concept pages and articles, that is, categories are simply handled as concepts (compare Suchanek et al. 2007).
- The relatedness of concepts is determined based on cross-references (wiki links) between articles: if two articles refer to *each other* using wiki links, it is assumed that they describe related concepts (again following Gregorowicz and Kramer 2006).
- The similarity of concepts is determined using language-links: if two articles refer to the same article in another language using language-links, these two articles are considered similar. The reason is that, according to Wikipedia convention, language-links should always point to similar (or, ideally, equivalent) articles, and the relation of similarity is considered to be transitive and symmetrical.
- The labels (terms) for a concept are determined from a variety of sources, among others:
    - from the title of the article, as well as the use of the magic word DISPLAYTITLE.
    - from the title of redirect pages that refer to the article.
    - from the title of disambiguation pages that refer to the article (compare Ponzetto and Strube 2007).

- from the link text of wiki links that refer to the article. This information has rarely been utilized by prior research (with the notable exception of Mihalcea 2007), even though it can be very valuable for information retrieval (compare Eiron and Mccurley 2003 as well as Kraft and Zien 2004).
- from sort-keys used when categorizing articles, as well as the use of the magic word DEFAULTSORT.

The result of this interpretation is a local (per-wiki) dataset which constitutes a monolingual thesaurus.

2.3 Building a Multilingual Thesaurus

The multilingual thesaurus is created by merging several mono-lingual thesauri: groups of similar concepts from different languages are combined into one language-independent concept each. This method was first described by Kinzler (2008).

The skeleton of the multilingual thesaurus is represented by the *global data model*, which contains mainly information about which concepts from the individual language-specific data sets have been combined into a language-independent (pan-lingual) concept and how the relations between the local concepts are mapped to relations of language-independent concepts. Together with the local data models that it references, the global dataset constitutes a *collection* of data models that represents a full multilingual thesaurus.

Each concept in the global data set is a set of local concepts, with at most one concept from each language. Language-specific properties (especially terms and definition glosses) are not again stored in the global dataset — instead, they are taken from the respective local data set when needed.

The main task when creating the global data set is thus to find groups of concepts from the different languages (i.e. local data set), such that concepts that are as similar as possible (or, ideally, equivalent) to each other are grouped together. The (possibly severe) differences in granularity and coverage of the Wikipedia in the different languages has to be taken into account when doing this. The algorithm used can be summarized as follows:

1. import all concepts from each language in the collection.
2. determine which concepts refer directly (using inter-language links) to which other concept in the

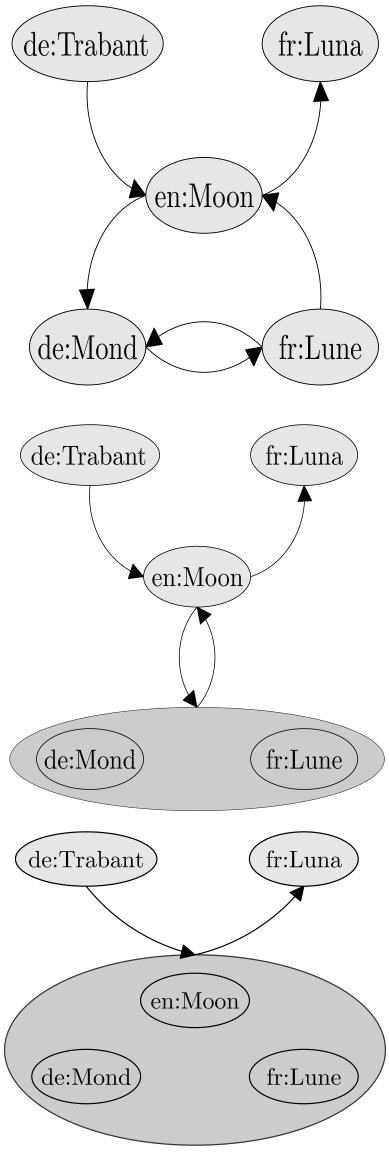

Fig. 2. Merging concepts from different languages.

multilingual thesaurus. This way, pairs of similar concepts from different languages are marked in the thesaurus.

3. determine which pairs of concepts refer to *each other* in this way. Because language links generally refer to equivalent or more general concepts, it can be assumed that, if two concepts reference each other via language links, they are equivalent or at least very similar. This approach is analogous to the method used to determine related concepts within one language, following Gregorowicz and Kramer (2006).
4. merge pairs of equivalent concepts, combining all the concepts' properties. While doing this, it is recorded which language-independent concept covers which languages. Each language must be present in each resulting pan-lingual concept no more than once.
5. merge such pairs until no more suitable pairs of equivalent concepts are available. If two concepts are connected by language-links in both ways but the sets of the languages they cover overlap, those concepts are *conflicting* and cannot be combined. They can, however, be marked as being *similar* to each other.

It should be noted that the order in which pairs are merged influences the outcome – the method is not stable against the order of processing. Heuristically, pairs from larger wikis should be merged first, because larger wikis tend to have higher granularity. This way, more precise matches are preferred.

To enable reuse of the thesaurus data by third parties, it can be mapped to the RDF data model using the SKOS vocabulary. RDF can then be serialized to XML or other formats suitable for storage and transmission. Representing a thesaurus as RDF/SKOS is the state of the art, as suggested by Gregorowicz and Kramer (2006), Assem et al. (2006), and Miles (2005).

3 USING THE THESAURUS TO SEARCH WIKIMEDIA COMMONS

The multilingual thesaurus generated by the process described above now contains all information needed to search and navigate in a language independent manner: Terms and phrases in different languages can be mapped to pan-lingual concepts. These pan-lingual concepts can then be mapped to language-specific (resp. wiki-specific) concepts, which in turn are connected to the actual wiki pages that describe these concepts.

The method described however applies to Wikipedias. In order to perform a search on Wikimedia Commons, we have to integrate it into the thesaurus schema. This is done by a simple trick: we pretend that *"commons"* is a language. This way, we end up with concepts "described" on Commons, just the way we have concepts described in English or French. And we then know which articles (galleries) and categories on Commons correspond to these concepts. Once we have the gallery page or category on Commons, we can easily find the images they contain.

3.1 Finding Meanings and Images

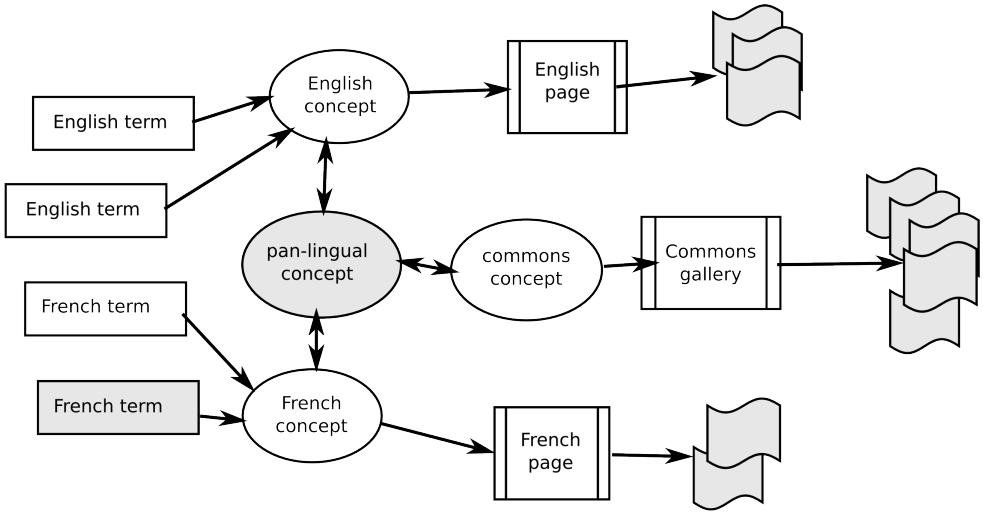

Fig. 3. A lookup path: from the French term the concept from the French Wikipedia is found, which is linked to the corresponding pan-lingual concept. From there, concepts from all wikis can be found, including commons. For each such concept, the corresponding wiki page is determined, and the pictures contained on that page collected. This diagram only shows the process for a single meaning of the initial term – there may be multiple.

A search (or rather, a *look-up*) is performed as follows: the user selects a language and enters a term or phrase. All possible meanings (associated concepts) of that term in that language are looked up and the corresponding "global" (pan-lingual) concepts are determined. Each pan-lingual concept found this way will be listed as one entry in the list of search results together with some images associated with the concept.

The images associated with a pan-lingual concept are the images associated with any of the per-wiki concepts that make up the pan-lingual "global" concept. The images associated with a concept are simply the images used on the wiki page associated with that concept (directly, not via a template) – or, in case the page is a category page, the images contained in that category.

To see all images associated with a given topic, as opposed to only the top rated pictures shown in the overview of possible meanings of the search term, the user can select a detailed view of any topic offered. In the detail view, a full gallery of images is shown. This view also offers navigational links to other topics (concepts), namely broader and narrower concepts as well as related and similar concepts, as defined by the thesaurus model.

For each concept, labels and definitions are available in several languages. WikiWord will attempt to show them in the user's selected language – if this is not available for a concept, a list of fall-back languages could be used to find some information that is still meaningful to the user – however, this has not yet been implemented.

Using the method outlined above, users are enabled to search Wikimedia Commons in their native language and navigate the topics without having to know English.

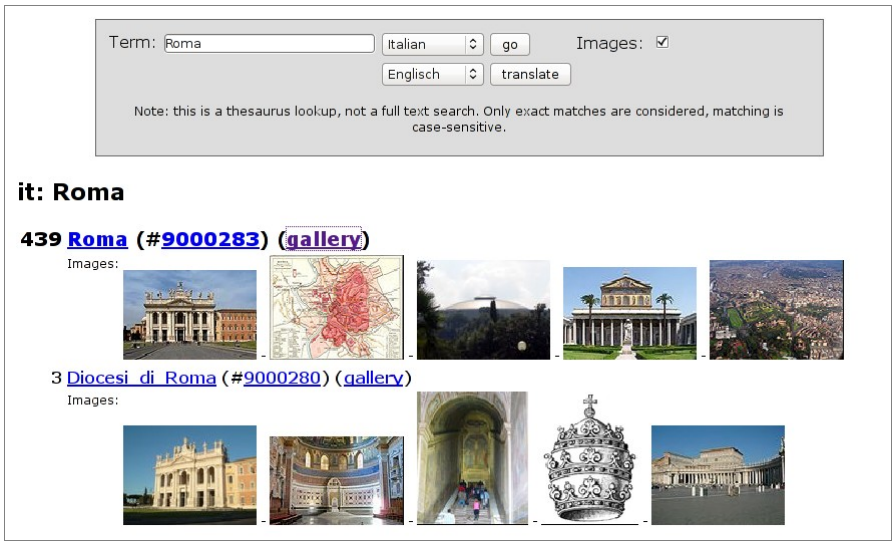

Fig. 4. Screenshot of a WikiWord search for "Roma" in Italian, turning up two results. Five sample images are shown for each result. Note that this search was run on a small sample data set, on a full thesaurus there would be much more results and frequency numbers would be much greater.

3.2. Ranking Results

There are two kinds of rankings used on WikiWord's result page: ranking the meanings of the terms the user entered, and ranking the images associated with each meaning.

For ranking the meanings, the weight of the association of the given term with the respective concept entry is used. That weight is derived directly from the frequency with which this term is used to refer to that concept – that is, roughly, how often this term is used as the link text of a link to that concept's wiki page.

For ranking the images for each concept, it is counted on how many of the pages associated with this concept (in different languages resp. wikis) the image is used. Usage on an associated gallery page on Commons has less weight, and inclusion in an associated category on Commons is rated a lot less important, because they include a much broader range of images. By contrast, the images actually used on Wikipedia pages are usually the most relevant to the given topic.

Another criterion that could be used but is currently not analyzed by WikiWord are quality markers on the images themselves. Commons has a variety of projects for rating images for different aspects of quality, including *"valued images"*, *"quality images"* and *"featured images"*.

4 EVALUATING THE THESAURUS

As of the time of writing this paper, there is still no complete multilingual thesaurus that includes Wikimedia Commons. Tests have been performed on smaller extracts from Wikipedia. Thus, an evaluation of the particular use case presented by this paper, namely,

searching Wikimedia Commons using the WikiWord thesaurus, was not yet possible. Below are some of the findings from the thesis for which WikiWord was originally created (Kinzler 2008). The values relate to the data mined from dumps from late 2007.

4.1. Statistical Overview

The thesaurus created for the original WikiWord thesis in 2007, as described by Kinzler (2008), covered the languages English (en), German (de), French (fr), Dutch (nl), traditional Norwegian (no), simple English (simple) and Nether-German (nds). The latter two are very small Wikipedias that were used mainly to determine how WikiWord behaves for small data sets. Two multi-lingual thesauri were created: one including all languages, and one only including the five large ones (en, de, fr, nl, and no).

Table 1. Size of the mono-lingual thesauri.

language	Pages	Articles	Concepts	Terms/Concept
en	4 936 730	2 007 044	6 319 639	2.05
de	1 316 554	656 703	2 127 028	2.20
fr	1 446 357	508 741	2 108 017	1.86
nl	640 222	382 550	1 341 691	1.63
no	267 827	156 725	626 982	1.64
simple	46 050	25 217	196 372	1.50
nds	14 839	10 895	80 473	1.28

It can be observed that in all wikis, about half of all pages in the main namespace are articles (the others being mainly redirects and disambiguation pages). Also, there are about three times as many concepts referenced than articles exist – concepts without an article result from “red” links, that is, they are concepts that are referenced but not yet described. Considering the subjective experience that nearly all links one sees on Wikipedia are blue (meaning their target page exists), this is quite interesting – basically, we are seeing the long tale of pages that get only referenced once or twice, while most links we would come across in existing Wikipedia articles would be to pages that are linked lots of times. This matches the expectation that the hyper-link network of Wikipedia pages be a *scale free small world network* (Barabasi and Albert 1999, Steyvers and Tenenbaum 2005), a finding confirmed by various studies (Capocci et al. 2006, Bellomi and Bonato 2005).

The number of terms associated with each concept is about 2, with numbers roughly declining with the size of the wiki resp. thesaurus. German and French are interesting cases: German has a surprisingly large number of terms per concept and French much less, even though it has nearly the same number of pages. This is especially surprising since French has a particularly large percentage of redirect and disambiguation pages.

4.2. Evaluation of the merging algorithm

When combining the thesauri of the five big languages into a multilingual thesaurus and merging concepts as described in section 2.3, about 11.6 million pan-lingual concepts remain of the total of 12.5 million concepts (if the two small languages are included, these figures are only insignificantly larger). This means that about 0.9 million concepts have been "absorbed" by other, equivalent concepts.

This number may seem small on a first glance, until some important facts are considered: The algorithm works on inter-language links, so it only applies to concepts described by articles in at least two languages. But only 3.7 million concepts are even described by an article even in one language, the rest are referenced without having a definition. These 3.7 million have been reduced to 2.8 million by merging, that is, about 1/4 have been "absorbed" by other, equivalent concepts.

Only about 1/20 (146'000) of the 2.8 million concepts retain inter-language links to other concepts in the thesaurus, that is, could have possibly been merged further or differently. This shows that the method presented by this paper exhausts the information provided by direct inter-language links. The fact that a lot of concepts remain isolated appears to be caused by many concepts really being described only in one Wikipedia, or at least missing sufficient inter-language links. This seems likely considering that the English Wikipedia is more than twice as large as the next runner-up, the German Wikipedia, and thus more than half of the pages on the English Wikipedia will probably have no inter-language link at all, or at least not an unambiguous one.

In order to apply further merging, additional information would have to be considered, such as which concepts have inter-language links in common. But even that would give no immediate benefit: About 160'000 additional pairs of concepts have inter-language links in common, none of which however could be merged directly, because they all conflict with regards to the set of languages they already cover. So while it might be possible to get better matches between languages, it's unlikely that it is possible to find more.

4. Accuracy of signification

The most important relation for the application at hand, namely finding possibly meanings of a term in a given language, is the signification relation, associating terms with concepts. The author evaluated the accuracy of this relation in some detail in the original thesis on WikiWord (Kinzler 2008) in section 12.5.

For the sake of this evaluation, a derived term referring to a generic concept (e. g. "physicist" referring to physics) is considered correct, and a broader term referring to a narrower concept (e. g. "America" for USA) is also acceptable. While these kinds of inaccuracies may be prohibitive for building a traditional dictionary, they are tolerable or even useful in the context of indexing and retrieval (compare Ide and Veronis 1998 as well as Wilks et al. 1996).

Other kinds of problems were considered unacceptable, including of course plain wrong associations, but also the case when a specific term referred to a more general concept (e. g. "Manhattan" referring to New York). The latter kind of problem is quite common when a

specific concept does not have its own article, but is covered in a more general article. This happens particularly often for parts of small towns.

A manual evaluation of a sample of a few hundred concepts showed that the percentage of bad associations of terms to concepts is up to 20%. However, it was also evident that concepts that were rarely used suffered more inaccuracies. Taking into account the frequency with which each association is used, the percentage of bad terms dwindles to about 5%.

To improve the accuracy, the law of great numbers can be utilized: we simply remove all associations with a low frequency. Requiring a minimum of two occurrences improves the accuracy from 80% to about 85%, however removing about half of all term associations. This seems a great loss for a small gain, but again, when we consider how often an association occurs (and thus how likely it will be used), things look better: we then see the accuracy improve from 95% to 97% while retaining a coverage of 95%, that is, only losing 5% of the expected lookups.

In addition to the mere frequency of a term being used to refer to a given concept, we can also consider how this reference takes place – that is, which of the rules described in section 2.2 is applied to construct the signification relation. We can then apply the frequency-based cut-off only to term associations that stem from a "weak" rule, namely, the rule that uses the link-text. All other rules, such as the page title, category sort keys, etc, are considered "strong" and override the cut-off. Using this method, we can still improve accuracy from 80% to about 83% while retaining 70% coverage. Scaling these figures by association frequency, we get an improvement from 95% to 96% with a coverage of 97%. Since this method seems to strike a good balance between accuracy and coverage, it will be used for the Wikimedia Commons search.

4.3. Coverage and Granularity

An important aspect to consider is the nature of concepts covered by Wikipedia. Since Wikipedia is an encyclopedia, it mainly contains nouns; WikiWord thus covers few verbs or adjectives. For an application in information retrieval this is however not problematic – to the contrary, it might even be considered an advantage: classification and indexing of documents is traditionally based on nouns, because they represent the topics of a document particularly well (Syed et al. 2008, Hepp et al. 2006, Eiron and Mccurley 2003).

In contrast to traditional thesauri and dictionaries, Wikipedia contains a wealth of entries about people, places, works of art and organizations (these make up more than half of all pages on Wikipedia). Because of this, Wikipedia is suited particularly well for indexing media files, since these tend to cover just those kinds of subjects. Another interesting property is the granularity of the terms covered by Wikipedia: even though Wikipedia covers a wide variety of topics, the thesaurus we derive from it is rather rough from a lexical point of view: the terms "physics", "physicist" and "physical" are treated as synonyms all referring to the concept of *physics*. Even opposites like "inconsistent" and "consistent" may be mapped to the same entry (in this case, *consistency*).

This type of inaccuracy however is not at all harmful in the context of information retrieval (compare Ide and Veronis 1998 as well as Wilks et al. 1996): when looking up the term "inconsistent", a document dealing with the topic of consistency will likely be a useful result.

The taxonomic granularity (e.g. if there is a separate article about nuclear physics, or if that is covered in the general article about physics) largely depends on the size of the individual Wikipedias and varies greatly between languages. This fact was one of the determining factors when developing the algorithm described in 2.3.

5 CONCLUSION

In this paper, a method for extracting a multilingual thesaurus from Wikipedia has been presented and the application of that thesaurus to facilitate language independent search for Wikimedia Commons described. To that end, the first section described the extraction of the relevant information from wiki text, the internal model of the thesaurus data, and the algorithm for merging language-specific thesauri into a multi-lingual thesaurus. Next, a method for finding sets of images for a given search term was described, based on the data supplied by the thesaurus. Finally, the thesaurus extracted from Wikipedia was analyzed and discussed.

The final analysis showed that the thesaurus derived from Wikipedia is well suited for the purpose of multilingual information retrieval in general. The specific use case of a multilingual image search for Wikimedia Commons could not yet be evaluated, since the project is not yet complete.

4. Outlook

WikiWord is an ongoing project, the software is permanently under development. The goal for the near future is to generate a full multilingual thesaurus from fresh dumps in eight languages (English, German, French, Dutch, Italian, Spanish, Portuguese, and Polish) plus Commons. Based on this data, a working search and navigation interface for Wikimedia Commons will be developed and made available to the Wikimedia community for day to day use.

For the future, it is intended to make the thesaurus itself available for use by third parties, especially for research purposes. This will be done in the form of RDF/SKOS dumps which will be updated every few months.

ACKNOWLEDGEMENTS

WikiWord was developed as part of the author's diploma thesis (Kinzler 2008) at the Department for Automatic Language Processing of the University of Leipzig, Germany. This paper is mainly based on the research conducted for this diploma thesis and some parts of it were taken from the thesis' text.

Development of WikiWord has been continued by the author for Wikimedia Deutschland e.V. as well as for some freelance engagements. Development is ongoing.

REFERENCES

Assem et al. 2006. A Method to Convert Thesauri to SKOS. The Semantic Web: Research and Applications: 95–109.

Barabasi and Albert 1999. Emergence of scaling in random networks. Science vol. 286, Nr. 5439: 509–512.

Bellomi und Bonato 2005. Network Analysis for Wikipedia. Proceedings of Wikimania 2005.

Capocci et al. 2006. Preferential attachment in the growth of social networks: the case of Wikipedia. arXiv:physics/0602026v2.

Eiron and Mccurley 2003. Analysis of anchor text for web search.

Gregorowicz and Kramer 2006. Mining a Large-Scale Term-Concept Network from Wikipedia. Mitre.

Hepp et al. 2006. Harvesting Wiki Consensus – Using Wikipedia Entries as Ontology Elements. First Workshop on Semantic Wikis.

Ide and Veronis 1998. Word Sense Disambiguation: The State of the Art. Computational Linguistics vol. 24: 1–40.

Kinzler 2008. Automatischer Aufbau eines multilingualen Thesaurus durch Extraktion semantischer und lexikalischer Relationen aus der Wikipedia. Diploma Thesis, Department for Automatic Language Processing of the University of Leipzig, Germany.

Kraft and Zien 2004. Mining anchor text for query refinement. WWW '04: Proceedings of the 13th international conference on World Wide Web, ACM Press: 666–674.

Mihalcea 2007. Using Wikipedia for Automatic Word Sense Disambiguation. Proceedings of NAACL HLT.

Miles 2005. Quick Guide to Publishing a Thesaurus on the Semantic Web. W3C Working Draft WD-swbp-thesaurus-pubguide-20050517.

Nakayama et al. 2007a. A Thesaurus Construction Method from Large Scale Web Dictionaries. Aina 00 (2007): 932–939.

Nakayama et al. 2007b. Wikipedia Mining for an Association Web Thesaurus Construction. WISE vol. 4831, Springer: 322–334.

Ponzetto and Strube 2007. Deriving a large scale taxonomy from Wikipedia.

Steyvers and Tenenbaum 2005. The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth. Cognitive Science 29, No. 1: 41–78.

Suchanek 2007. Yago: a core of semantic knowledge. Proceedings of the 16th international conference on World Wide Web (2007): 697–706.

Syed et al. 2008. Wikipedia as an Ontology for Describing Documents. Proceedings of the Second International Conference on Weblogs and Social Media.

Wilks et al. 1996. Electric Words: Dictionaries, Computers, and Meanings. MIT Press: 59.

COPYRIGHT