Building Language-Independent Concepts from Wikipedia

Daniel Kinzler, Universität Leipzig

June 30, 2008

Using Interlanguage-Links to Find Articles About the Same Concept in Different Languages.

Abstract

This paper describes a simple method for deriving language-independent concepts by identifying groups of pages about the same topic from Wikipedias in different languages. This allows the information about a concept obtained from different Wikipedias to be merged, and by this provides a way to determine which terms from different languages refer to the same concept. The method presented here was developed as part of WikiWord [Kin08]*, where it was used to create a multilingual thesaurus from multiple monolingual thesauri.*

Introduction

Wikipedia has been identified by recent research as a rich resource for mining knowledge about concepts (e.g. [ZGM07][PS07]). One type of knowledge that can be obtained by analyzing Wikipedia articles is which terms are used to refer to which articles [Mih07] (the *signification* relation). Until now, however, this has only been applied to individual languages, especially for the purpose of *named entity recognition* [Cuc07].

This paper presents a method to combine information about individual languages gathered from different Wikipedias into a single multilingual thesaurus. This method was developed as part of WikiWord [Kin08] and works by identifying and merging equivalent concepts from different sources. Determining which terms are used to refer to which concepts (signification) is another central aspect of WikiWord, but will not be discussed in this paper.

Wikipedia, by virtue of being an encyclopedia, describes concepts in individual articles, each being a web page with a unique URL. Wikipedia contains other types of pages too, most importantly disambiguation pages, redirects, lists and category pages. These are valuable resources for determining which terms are used for which concept or how concepts relate to each other [PS07][ZG07], but they are not discussed here.

In order to gather information about which terms from *different* languages are used for a given concept, it is necessary to determine which language-bound (*local*) concepts from the different Wikipedias are equivalent (or at least very similar) and can be viewed as instances of the same language-independent concept. In other words, the goal is to determine which articles in the different Wikipedias describe the same concept.

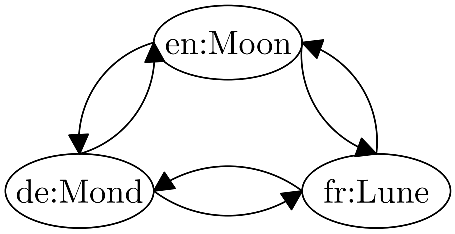
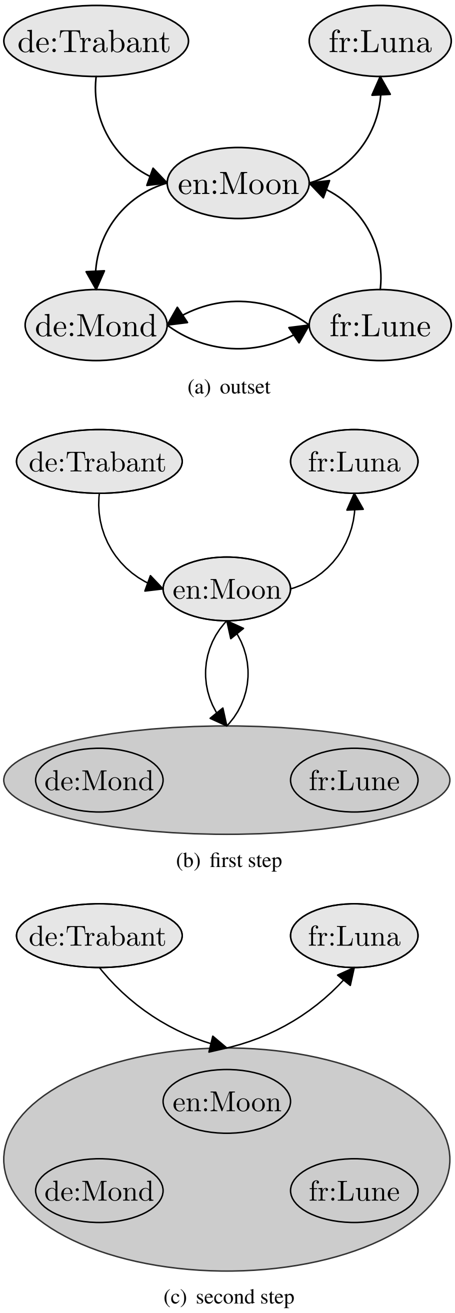


Figure 1: Ideal interlanguage links

The main source of information about how articles (and thus, concepts) in Wikipedias in different languages relate to each other are so called *interlanguage links* (or *language links* for short) [MWI]. Each article may contain interlanguage links referencing articles about the same subject in other Wikipedias (that is, in other languages) [WPI]. So, the article *Moon* in the English language Wikipedia may have interlanguage links to the article *Mond* in the German language Wikipedia, *Lune* in the french language Wikipedia, and so on, as shown in fig. 1. The method presented in this paper will mainly use the information provided by these links.



(a) outset

(b) first step

(c) second step

Figure 2: Merging concepts (contrived example)

Method

The goal of the method described in this section is to identify groups of articles from different Wikipedias that describe the same concept. Such a group contains at most one article from each language, and is thought to *represent* the language-independent concept.

As mentioned above, the approach presented here uses interlanguage links to determine which articles deal with the same subject. Ideally, there is one article in each language describing the subject, so the articles correspond to each other exactly, and each references all the others using interlanguage links, as shown in fig. 1. However, due to differences in organisation and, especially, because of the different size and *granularity* of the Wikipedias, the real structure of interlanguage links often looks different, as shown in fig. 2(a): some links may be missing (like the connection from *de:Mond* to *en:Moon*), some may simply be wrong (like *en:Moon* pointing to *fr:Luna* instead of *fr:Lune*), but most importantly, where in one Wikipedia a topic may be covered in a single article, it may be spread across several in another Wikipedia (like both *de:Mond* and *de:Trabant* referring to *en:Moon*)[1].

The method proposed here for solving this problem is quite simple and will be shown to have a high degree of precision, though it may be overly restrictive for some applications. The idea is based on the notion that articles can be assumed to describe the same concept if they reference *each other* via interlanguage links. Articles connected in this way can be grouped together to form language independent concepts, see fig. 2(b). This can be done as follows:

First, for each article in each Wikipedia, a concept entry is created. Concepts are represented by groups of articles from different Wikis, and we start out with concepts that consist of exactly one article each. The interlanguage links between articles are now considered connections of *similarity* between the concepts that contain the respective articles.

Second, an arbitrary pair of concepts is chosen and merged. A pair can be merged if it satisfies two conditions:

- The concepts must be *mutually* similar, that is, at least one article in the first concept must have an interlanguage link referencing an article in the second concept, and vice versa. By requiring the

[1] Note that this is a contrived example. The reality is usually a bit better than that, but also more complex.

reference to be mutual, erroneous interlanguage links are filtered out and ambiguities are resolved in a way that favors the closest match, as determined by the editors who maintain the interlanguage links. This is analogous to the method of using mutual cross-references to detect *related* concepts, as suggested in [GK06].

- The concept created by merging must not contain two articles from the same Wikipedia, i.e. in the same language. In other words, the sets of languages covered by the two concepts must be disjoint. This reflects the fact that, by convention, in each Wikipedia there should be only one page per subject.

The merging of pairs of concepts is repeated until no pair remains to be merged, as shown in fig. 2(c). Note that through the merging process, new pairs of mutually similar concepts may be created, as seen in fig. 2(b).

The time required by this process is bounded by $O(n \cdot k^2)$, with n being the number of articles and k being the number of Wikipedias being considered. This is because each concept can be merged at most k times, because of the second constraint above: it may only contain at most one article from each Wikipedia (i.e. language), but with each merge, at least one new language is added. The cost of finding pairs to merge is bounded by $O(n \cdot k)$, which is the maximum number of pairs, because each article may only have one interlanguage link to each other Wikipedia, that is, a total of k interlanguage links. This shows that this algorithm scales well enough to be applicable to the millions of articles contained in large Wikipedias.

When analyzing the method presented above, the following problems become apparent:

- The algorithm is not fully deterministic in that the result depends partly on the order in which merges are performed, which the concrete implementation is free to choose. To make the process deterministic, some arbitrary criterion for the order of merges could be imposed, such as the alphanumeric order of page titles. It would however be preferable to find a criterion that would yield the optimal order of merging, or at least one that increases the likelihood of getting a good result.

- Small Wikipedias may have a bad influence on the result, since they have a low granularity, less active users to assure the quality of the interlanguage links, and are also more likely to contain articles to cover multiple subjects. Such a "bad" article may "poison" the corresponding concept(s), leading to unrelated articles to be grouped together, and equivalent articles remaining in separate groups. This can of course be avoided by simply not including very small ("immature") Wikipedias in the process.

- The described method may be overly restrictive. For example, articles that are not yet fully integrated in the network of mutual interlanguage links (e.g. because they are new) often remain isolated. Also, the information contained in interlanguage links to Wikipedias which have not been analyzed is discarded. It would be interesting to investigate alternative methods of determining the similarity concepts in order to merge them. One such method would be to view the set of interlanguage links of an article as a *feature vector* and compare articles (and thus, concepts) based on these. Evaluation shows, however, that there is little hope that this method would yield substantially better results.

This approach is very straightforward and uses the information provided by interlanguage links directly, at "face value". However, to the author's knowledge, it has not before been applied to large amounts of data and evaluated systematically. The following chapter presents the results of some experiments that attempt to provide such an evaluation.

Evaluation

This section evaluates the method described so far in this paper, based on data from experiments conducted with WikiWord [Kin08]. For these experiments, monolingual thesauri were generated from five Wikipedias, namely English, German, French, Dutch, and Norwegian, together covering about 20.0 million terms for 12.5 million concepts, 3.7 million of which have an article describing them[2]. The articles are distributed among the Wikipedias as follows: 2.0 million from the English language Wikipedia, 660 thousand from the German, 500 thousand from the French, 380 thousand from the Dutch, and 160 thousand from the Norwegian language Wikipedia.

[2] The rest result from categories and from "red" links to pages that do not yet exist.

These thesauri were then combined into a single multilingual thesaurus by using the method described above, that is, by identifying and merging equivalent concepts from different Wikipedias. The resulting thesaurus contains 11.5 million concepts, 2.8 million of which have at least one article describing them[3].

To get an impression of the performance of the presented algorithm, the following facts should be considered: The algorithm works on interlanguage links, so it only applies to concepts described by articles in at least *two* languages. By merging equivalent concepts, the 3.7 million concepts that are described by articles have been reduced to 2.8 million, that is, about ¼ (0.9 million) have been "absorbed" by other, equivalent concepts. Only about $1/20$ (146 thousand) of the 2.8 million concepts retain interlanguage links to other concepts in the thesaurus, that is, could have possibly been merged further or differently.

These figures show that the method presented by this paper exhausts the information provided by direct interlanguage 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 interlanguage links. In order to apply further merging, additional information would have to be considered, such as which concepts have interlanguage links in common. But even that would give no immediate benefit: About 160 thousand additional pairs of concepts have interlanguage 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.

As to the quality of the concepts generated this way, a manual evaluation of a sample of 250 concepts revealed no case of an article being assigned to a concept erroneously, and just 6 instances of a concept missing an article. That is, of the 250 concepts were 6 that were connected by an interlanguage link to an article that described the concept in question, but was not merged into the concept. This was generally due to an interlanguage link missing in that article.

In conclusion it can be said that the simple method presented by this paper makes good use of the information available from interlanguage links. Even though alternative methods exist and should be investigated in detail, they are unlikely to yield substantially better results. Further improvements of connecting Wikipedia articles between languages can probably be achieved only by improving and extending the system of interlanguage linking in Wikipedia itself.

[3] The full datasets are available online, see the ▹*WikiWord Resources* section.

References

[Cuc07] S. Cucerzan. Large-Scale Named Entity Disambiguation Based on Wikipedia Data. *EMNLP 2007: Empirical Methods in Natural Language Processing, June 28-30, 2007, Prague, Czech Republic*, 2007.

[GK06] Andrew Gregorowicz and Mark A. Kramer. Mining a large-scale term-concept network from wikipedia. Technical report, Mitre, 2006. Available from: `http://www.mitre.org/work/tech_papers/tech_papers_06/06_1028/06_1028.pdf`.

[Kin08] Daniel Kinzler. Automatischer Aufbau eines multilingualen Thesaurus durch Extraktion semantischer und lexikalischer Relationen aus der Wikipedia. Master's thesis, Universität Leipzig, 2008. Available from: `http://brightbyte.de/papers/2008/DA/WikiWord.pdf`.

[Mih07] Rada Mihalcea. Using wikipedia for automatic word sense disambiguation. In *Proceedings of NAACL HLT 2007*, 2007. Available from: `http://www.cs.unt.edu/~rada/papers/mihalcea.naacl07.pdf`.

[MWI] Help:Interwiki linking: Interlanguage link. Available from: `http://meta.wikimedia.org/wiki/Help:Interwiki_linking#Interlanguage_link`.

[PS07] Simone P. Ponzetto and Michael Strube. Deriving a large scale taxonomy from wikipedia. In *Proceedings of the 22nd National Conference on Artificial Intelligence*, 2007. Available from: `http://www.eml-research.de/english/homes/strube/papers/aaai07.pdf`.

[WPI] Help:Interlanguage links. Available from: `http://en.wikipedia.org/wiki/Help:Interlanguage_links`.

[ZG07] Torsten Zesch and Iryna Gurevych. Analysis of the wikipedia category graph for NLP applications. In *Proceedings of the TextGraphs-2 Workshop (NAACL-HLT)*, 2007. Available from: `http://elara.tk.informatik.tu-darmstadt.de/publications/2007/hlt-textgraphs.pdf`.

[ZGM07] Torsten Zesch, Iryna Gurevych, and Max Mühlhäuser. Comparing wikipedia and german wordnet by evaluating semantic relatedness on multiple datasets. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 2007. Available from: `http://elara.tk.informatik.tu-darmstadt.de/publications/2007/hlt-short.pdf`.

WikiWord Resources

Below is a list of additional resources for the WikiWord project.

Project overview: <`http://brightbyte.de/page/WikiWord`>

Thesis excerpt in english: <`http://brightbyte.de/page/WikiWord/Excerpt`>, *Outline of a method for building a multilingual thesaurus from Wikipedia*.

SQL data dumps: <`http://aspra27.informatik.uni-leipzig.de/~dkinzler/sqldumps/`>, especially the file `full_wikiword-full.sql.bz2` (11GB).

Download area: <`http://brightbyte.de/DA/`>. This contains bundles of the WikiWord source and compliled library files, as well as dumps of small sample data sets.

Source code: <`http://brightbyte.de/repos/DA/WikiWord/`> for online browsing. Bundles are available from the download area.