Acquiring a Taxonomy from the German Wikipedia

Laura Kassner[‡], Vivi Nastase[*], Michael Strube[*]

[‡] Seminar für Sprachwissenschaft
University of Tübingen, Tübingen, Germany
Laura.Kassner@gmx.de

[*]EML Research gGmbH, Heidelberg, Germany
{nastase,strube}@eml-research.de

Abstract

This paper presents the process of acquiring a large, domain independent, taxonomy from the German Wikipedia. We build upon a previously implemented platform that extracts a semantic network and taxonomy from the English version of the Wikipedia. We describe two accomplishments of our work: the semantic network for the German language in which *isa* links are identified and annotated, and an expansion of the platform for easy adaptation for a new language. We identify the platform's strengths and shortcomings, which stem from the scarcity of free processing resources for languages other than English. We show that the taxonomy induction process is highly reliable – evaluated against the German version of WordNet, GermaNet, the resource obtained shows an accuracy of 83.34%.

1. Introduction

The multitude of projects in which WordNet is being used show how important a semantic network is for a wide variety of applications – translation, question answering, summarization, and many others. Finding and understanding connections between words is a crucial aspect of semantic analysis. The success of WordNet has given rise to the development of similar resources for other languages. Building a resource is an effort-intensive process, which may reflect itself in the end result in two ways: the resource obtained is not free (mostly because of the financing necessary to obtain it in the first place), and is limited in coverage (because development stops when financing stops, and because of choices relative to the linguistic inventory within the resource).

The World Wide Web is not only a repository of a huge amount of potential data, but also a collaborative medium, through which people can contribute towards the development of large-scale, up-to-date, resources, such as Wikipedia. People contribute to the growth of this encyclopedia on a volunteer basis, sharing their knowledge freely. Guidelines assist them in making their contributions consistent and in connecting them with existing material. Out of these efforts has emerged a large collection of encyclopedic knowledge, which is just the tip of the iceberg. Underneath, there is a large semantic network, in which categories and pages are connected to each other based on individual decisions made by the contributors.

Strube and Ponzetto (2006) and Ponzetto and Strube (2007) have shown that the "hidden" part of Wikipedia can be used as a semantic network in a similar way to WordNet. The network built from the category links for the English Wikipedia performs competitively with WordNet on semantic relatedness and similarity judgements, and better on high end applications such as coreference resolution. The big advantage such an approach has over using WordNet is the fact that Wikipedia is constantly growing, and is extremely up-to-date. For high end applications such as question answering and summarization – which are likely to involve recent events, situations and personalities – this last feature of the resource is crucial.

Zirn et al. (2008) further refined the taxonomy extracted from the English Wikipedia (Ponzetto and Strube, 2007) by introducing a distinction between instance and class-type categories[1].

Having shown that a resource built from Wikipedia is useful, the next logical step is to exploit the multi-linguality of the Wikipedia project. In this paper we present the steps for the development and the evaluation of a German Wikipedia taxonomy. To obtain this resource we use the system developed to induce the English Wikipedia taxonomy, and adapt it for the new language. We show what the adaptation entailed, and evaluate the end result by comparing with GermaNet – the German WordNet (Lemnitzer and Kunze, 2002).

2. Wikipedia Taxonomy

Wikipedia is a multi-lingual online encyclopedia, grown through volunteer contributions over the Internet. Contributors are given guidelines for categorizing articles and naming new categories[2]. This has led to the emergence of large category networks underlying the Wikipedia articles in various languages. Categories are connected in these networks through unnamed links, that may represent different types

---

[1] http://www.eml-research.de/nlp/download/ wikitaxonomy.php
[2] http://en.wikipedia.org/wiki/Wikipedia: Categorization
http://en.wikipedia.org/wiki/Wikipedia: Naming_conventions_(categories)

of relations: THUMB *isa* FINGER *part of* HAND[3]. It is often useful to be able to filter specific types of semantic relations, such as *isa*, from the ones that are available. Ponzetto and Strube (2007) show how to induce *isa* links in the category network of the English Wikipedia, and shape it into a taxonomy. Their method goes through the following steps:

1. **Filter:** Filter out meta-categories that pertain to management issues in Wikipedia using key words (e.g. `template`, `user`, `portal`).

2. **ByMatcher:** Filter out links between categories C_1 and C_2 whose names match the patterns: C_1 = X `by` Y, C_2 = Y X (e.g. C_1 = ALBUMS BY ARTIST, C_2 = MILES DAVIS ALBUMS). These links are labeled *is refined by*.

3a. **HeadMatcher:** Assign *isa* label to the link between two categories that have the same syntactic head (e.g. BRITISH COMPUTER SCIENTISTS and COMPUTER SCIENTISTS).

3b. **ModifierMatcher:** Assign *notisa* label to the link between categories C_1 and C_2 if C_1's syntactic head's lemma appears in non-head position in C_2, or the other way around (e.g. CRIME COMICS and CRIME).

4a. **PluralMatcher:** Assign *isa* label to the link between a category C and its super-category SC, if C has a homonymous page categorized under category C_{SC} that has the same head and SC, and this head is a plural noun. Figure 1 shows such an example, for C = MICROSOFT, SC = COMPUTER AND VIDEO GAME COMPANIES and C_{SC} = COMPANIES LISTED ON NASDAQ.
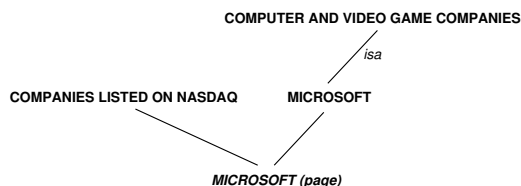

Figure 1: *isa* links induced using structural information

4b. **CooccurrenceMatcher:** Assign *isa* label to the link between a category C and its super-category SC if a page is redundantly categorized under both of them.

5. **Pattern:** Use lexico-syntactic patterns indicative of *isa* (e.g. `X is a Y`; `Y, such as X`) and *notisa* (e.g. `X is part of Y`; `X in Y`) relations to find evidence for and against an *isa* relation from a corpus. The label is assigned based on majority voting.

6. **HierarchicalPropagation:** Draw links between categories in an *isa* chain, and label them as *isa*.

---

[3]We adopt the following notation conventions: `Sans Serifs` for words and patterns, *Italics* for relations, SMALL CAPS for categories and pages.

3. Language specific issues

The processing steps applied by Ponzetto and Strube (2007) to induce a taxonomy from the category network can be split into three main types: pattern-based, structure-based, language (morpho-syntax) based.

Steps 1, 2 and 5 are pattern-based. They rely on lexicographic patterns, easily adaptable when processing a new language. Steps 4b and 6 are structure-based, and thus rely exclusively on link configuration in the category network. The steps that pose most problems for adapting to a new language are the ones requiring morpho-syntactic processing: 3a, 3b, 4a. Let us look in a bit more detail at the steps that are more challenging to adapt.

A first requirement is a parser, to determine the syntactic head of the category phrase. Compared to English, German has the additional problem of ubiquitous complex noun compounds. While such situations do appear in English as well (`birthday`), in German the problem is on a much larger scale. The processing steps described in Section 2 require access to the head of a phrase and its modifier. Both the head and the modifier may be embedded in a compound. Splitting a compound is not trivial, because of several issues: (i) a compound may have several possible splits into simple words, but only one makes sense in a given context – e.g. `Spielerfolge` can be split into `Spieler` (player) – `Folge` (sequence) or `Spiel` (game/play) – `Erfolge` (successes); (ii) compounds are not always "pure", in the sense that there is more to the compound than just the words that compose it. The extra elements that come in are called *Fugenelemente* (`e`, `en`, `es`, `er`, `ens`, `n`). In some cases these elements can be interpreted as case inflections – `Amtsblatt` (official gazette) = `Blatt` eines `Amts`; `Kinderwagen` (stroller) = `Wagen` für ein `Kind` (actually used in Plural form `Kinder` because it sounds better). In other case the Fugenelemente serve to facilitate pronunciation – `Informationsdienst` (information service) = `Dienst` zur `Information`; `Mausefalle` (mouse trap) = `Falle` für eine `Maus`.

Another problem we encounter in parsing German phrases are the noun cases. Noun cases are indicated by articles. Without context, and occasionally even with, or without a good lexicon with gender information, it may be hard to tell which case an article indicates. For example, the article `der` is the singular nominative masculine article, but can also be the singular dative or genitive feminine article, or the plural genitive article. This kind of ambiguity is reflected in the results of the parser we used, where genitival feminine nouns are not identified correctly, and usually interpreted as nominal masculine nouns. Because of this, the noun in the genitive case is erroneously identified as the head of the phrase (as it is interpreted to be in the nominative case). For example, `Volkskammer` (People's Chamber) is given as the head of the phrase `Abgeordneter der Volkskammer` (delegate of the People's Chamber), whereas the correct head would be `Abgeordneter` (delegate).

Step 5, the pattern-based step that searches for patterns in-

dicating *isa* and *notisa* relations between a pair of categories, is also more challenging in German than it is in English. German is a morphologically more complex language. Because of case inflections and compound nouns we need more than just a simple translation of the pattern from English to German. For example, the words `leaf` and `tree`, and the pattern `the leaves of the tree` correspond to the German versions `Blatt`, `Baum`, `die Blätter des Baumes`. In translating the pattern, then, we must pay attention to determiners (their case and gender), and the changes according to case of the nouns. To be able to use better the information in the corpus, we need to allow `Baum` to appear as part of a larger compound, as in the fragment: `die Blätter des Kastanienbaumes`.

4. Adapting the platform

We adapt the platform developed by Ponzetto and Strube (2007) to induce a taxonomy from category networks for the multi-language Wikipedia. After identifying the steps that require modification, as described in Section 3, we proceed to adapt the platform to make it flexible relative to the language of the network it is run on. The system is implemented in Java, and each processing step corresponds to a Java class, and they all share a common parent. In the new version of the platform, the classes that have language specific requirements are now parents, with language specific descendants. We review the processing steps and show the language specific changes and the German resources used:

1. **Filter** – remove meta-categories using language-specific keywords (e.g. `Artikel` (article), `Benutzer` (user), `Begriffsklärung` (disambiguation), `Kategorie` (category)).

2. **ByMatcher** – the matching pattern `/.* by .*/` has the following German equivalents: `/*. nach .*/` and `/.* als Thema/`.

3a. **HeadMatcher** – identifies linked categories that have the same phrasal head, and assigns the link the label *isa*. In this step the category names are parsed and tagged with named entity information. The original system used the Stanford Parser (Klein and Manning, 2003), and the Stanford Named Entity Tagger (Finkel et al., 2005). The Stanford Parser can be used for German, if we provide it with a German grammar model. We have build models by training the parser on the TüBa-D/Z (Hinrichs et al., 2004) and Negra (Skut et al., 1997) tagged corpora. We have found that in our case Tüba-D/Z works slightly better. This step and the next could benefit from a morphological analyzer, in the case in which the head noun is a compound. The analyzer would split the compound word into its components. For the lack of such a tool, we have implemented a method that checks whether two words share an end-substring which is a noun. If they do, we assume the head matching has succeeded.

3b. **ModifierMatcher** – identifies *notisa* links, between two categories X and Y, where X is an ancestor of Y, and Y's name is XX_1. We add a modifier matching method for the case that X and Y are compounds. We verify whether a substring at the beginning of Y is a noun and it appears at the end of X, or the other way around. If we find such a substring we assume the modifier matching has succeeded.

4a. **PluralMatcher** – identifies *isa* links using the convention that in the English version of the Wikipedia, categories are written in plural, whereas pages are usually in the singular. Such a convention does not exist in the guidelines for building the German Wikipedia, so this step is bypassed in processing the German network.

4b. **CooccurrenceMatcher** – identifies *isa* links based on structural information, in particular, overlap in links from the two categories on the edge under analysis. This step is language independent, and is used as in the original system.

5. **Patterns** – we have translated the English *isa* and *notisa* patterns to German, and used them to find evidence for and against *isa* labels in a corpus formed from the German Wikipedia articles. The patterns were adapted to allow for case variations, reflected in determiners and noun endings.

6. **Link Propagators** – propagate *isa* links based on transitivity and structural similarity for two nodes. These are not language specific and were not changed.

5. Evaluation

It is crucial to evaluate a new resource, to show how reliable it is. The two main options are evaluating the resource in isolation, or within the context of an application. While evaluating through an application is more realistic, it makes sense to assess the resource in isolation beforehand, to make sure the effort of including it in an application is justified. In the case of resources as large as the one we built, manual evaluation is not a realistic option. Ponzetto and Strube (2007) evaluate the English semantic network using ResearchCyc, the version of Cyc licensed for research (Lenat and Guha, 1990). The largest semantic network available for German is GermaNet – the German version of WordNet.

We evaluate the result of the taxonomy acquisition process using GermaNet. We started with a network consisting of 686,751 nodes and 2,014,357 links, covering both categories and pages in the German Wikipedia version of 25.09.2007. Filtering unwanted categories (with names containing `wikipedia, wikimedia, liste, mediawiki, vorlage, artikel, benutzer, user ...`) slightly reduces the network to 670,213 nodes and 1,901,448 links.

The by-category filter identifies 17,716 links that are labeled *is refined by*. The HeadMatcher labels as *isa* 98,679

links. The overlap with GermaNet is 1139 links, on which we obtained a precision of 76.03%. This evaluation is strict, in the sense that we constrain that the complete compound or multi-word category appears in GermaNet, not just the head of the compound. The ModifierMatcher labels 40,089 links as *notisa*. The overlap between pairs in the Wikipedia semantic network and GermaNet is 980 in this case (we test that both elements of the pair appear in GermaNet, not necessarily linked by some relation). The evaluation of the *notisa* relations against GermaNet reveals a precision of 91.83%. The head and modifier matching methods combined give an overall accuracy of 83.34%.

The link labels induced through the by-category, head and modifier matching apply mostly to categories in the network. After these three processing steps we are left with 932,106 unique pairs. These pairs consist of category/page syntactic heads, to improve recall results. We use search patterns to find evidence for *isa* and *notisa* relations in the Wikipedia corpus. Despite the fact that the German Wikipedia contains approximately 600,000 articles, the average article length is approximately 450 words. Because of this, very few pairs linked by patterns from our list of *isa* and *notisa* patterns were actually found. We plan to explore other solutions for identifying the type of links between nodes in this large semantic network. One option is to use a larger corpus or the Web, another is to induce link types through inter-lingual links in Wikipedia. We currently have a large semantic network for the English Wikipedia from which we have induced a taxonomy. We will map the English and German networks using Wikipedia's cross-language article links, not only to enhance the German network, but also to obtain a larger, multi-lingual taxonomy.

Evaluation of the full network is hard because of the lack of a comparable, already annotated, resource. The evaluation on the portion that overlaps with GermaNet shows very high precision, which justifies the incorporation of the network in applications, and performing an evaluation in context. This will show the usefulness of the built taxonomy for high end NLP applications, and is our aim for future work.

6. Conclusion

We have presented the induction of and evaluated a taxonomy for the German language. It was built using the category network underlying the German version of Wikipedia and a processing platform originally developed for English. We have identified the requirements for obtaining a taxonomy in another language than English, and have adjusted the system to allow for fast adaptation for new language processing. The resource obtained is evaluated against the German WordNet, with high accuracy (83.34%) after the first processing steps. The process described can be used for new releases of Wikipedia, thus keeping the resource obtained as up-to-date as the online encyclopedia.

7. References

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics,* Ann Arbor, Mich., 25–30 June 2005, pages 363–370.

Erhard Hinrichs, Sandra Kübler, Karin Naumann, Heike Telljohann, and Julia Trushkina. 2004. Recent developments in linguistic annotations of the TüBa-D/Z treebank. In *Proceedings of the 3rd Workshop on Treebanks and Linguistic Theories,* Tübingen, Germany, 10–11 December 2004.

Dan Klein and Christopher D. Manning. 2003. Fast exact inference with a factored model for natural language parsing. In Suzanna Becker, Sebastian Thrun, and Klaus Obermayer, editors, *Advances in Neural Information Processing Systems 15 (NIPS 2002),* pages 3–10. MIT Press, Cambridge, Mass.

Lothar Lemnitzer and Claudia Kunze. 2002. GermaNet – representation, visualization, application. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation,* Las Palmas, Canary Islands, Spain, 29–31 May 2002, pages 1485–1491.

Douglas B. Lenat and R. V. Guha. 1990. *Building Large Knowledge-Based Systems: Representation and Inference in the CYC Project.* Addison-Wesley, Reading, Mass.

Simone Paolo Ponzetto and Michael Strube. 2007. Deriving a large scale taxonomy from Wikipedia. In *Proceedings of the 22nd Conference on the Advancement of Artificial Intelligence,* Vancouver, B.C., Canada, 22–26 July 2007, pages 1440–1445.

Wojciech Skut, Brigitte Krenn, Thorsten Brants, and Hans Uszkoreit. 1997. An annotation scheme for free word order languages. In *Proceedings of the Fifth Conference on Applied Natural Language Processing ANLP-97,* pages 88–95, Washington, DC.

Michael Strube and Simone Paolo Ponzetto. 2006. WikiRelate! Computing semantic relatedness using Wikipedia. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI-06),* pages 1419–1424, Boston, Mass., July.

Cäcilia Zirn, Vivi Nastase, and Michael Strube. 2008. Distinguishing between instances and classes in the Wikipedia taxonomy. In *Proceedings of the 5th European Semantic Web Conference,* Tenerife, Spain, 1–5 June 2008. To appear.