Taxonomy Extraction from German Encyclopedic Texts

Michael Granitzer[1], Andreas Augustin[2], Wolfgang Kienreich[2], and Vedran Sabol[2]

[1] Graz University of Technology, Graz, Austria
mgranitzer@tugraz.at,
http://kmi.tugraz.at/
[2] Know-Center Graz, Graz, Austria
aaugustin@know-center.at,
http://www.know-center.at/

Abstract. This paper presents a procedure to construct robust taxonomies from natural language German encyclopedic text. Taxonomic relations are extracted through Hearst patterns, validated via search engines and incorporated into an base ontology. By verifying the relations using an external, web-based source of evidence, the procedure extracted an accurate taxonomy with minimal human intervention. Experimental results revealed that high precision values can be achieved. Results show that a well grounded suggestion of solid patterns for the extraction of German hypernyms relations is possible and that the chosen approach may be efficiently transfered to other western languages.

1 Introduction

Knowledge acquisition still remains a bottle neck for most knowledge based applications as well as in the Semantic Web. However, the large number of unstructured text information in the digital universe [1] suggests itself to be exploited by automatic means in order to extract concept, relations and subsequently ontologies [6]. Focusing on taxonomies, efficient automatic methods for the English language exists. One well known approach are lexico-syntactic patterns [4] aka. Hearst patterns. Through definition of syntactic language patterns, hypernym relations between phrases can be extracted or validated efficiently. Besides machine learning technology, these kind of patterns became widely used in ontology learning from text in the last years.[5]

One drawback of lexico-syntactic patterns is the high degree of false positives. For example, [4] shows that the number of correct to incorrect relations is approximately 3.4 %. The approach taken in [6] to increase this rather low accuracy trains a classifier on multiple sources of evidence. Lexico-syntactic patterns are used to extract hypernym relations from the corpus of interest. The classifier is trained, by applying the same pattern on a set of web pages returned by Google^{TM} queries and some other sources of evidence like WordNet and Wikipedia. The weighted features classify the extracted relations into true and false. Similarly, work in [7] used supervised machine learning to generate lexico-syntactic patterns which results in a further improvement.

However, all these methods were applied to English corpora. Especially in domain dependent knowledge acquisition scenarios (e.g. like in Corporate Intranets), where non-English corpora are present, this may be problematic. Training data is usually sparse and corpora for statistical comparison are usually smaller.

In our work we investigate the definition of lexico-syntactic patterns for German in order to extract and validate taxonomic relationships. We present a method to define lexico syntactic patterns and to validate them on (i) the corpora at hand and (ii) on external sources. The newly extracted relations are combined and mapped to an existing RDF[3] based ontology. Our methodology as well as results may be ported to other languages and, as shown, applied to extract taxonomic relationships in various western languages with reasonable effort and acceptable accuracy.

2 Method

In this section we introduce our methodology to adapt lexico-syntactic patterns to German, which consists of three steps. In a first step we manually derived a set of Hearst patterns from the encyclopedic corpus. Next, we extracted the relationships based on the created patterns using the encyclopedic corpus. As experiments in section 3 show, precision of this extraction process is quite low. Hence, in the last step we validated the extracted taxonomic relationship using an external web source, which increased precision significantly.

2.1 Pattern Creation

To explore German Hearst patterns, we had to find our own approach, as few references exist. In the first attempt pairs of nouns, which are known to reflect a hypernym relation, have been used to extract potential patterns. For example: `Country - Austria`. Sentences with the occurrence of both nouns were retrieved and statistically analysed. The first attempt resulted in too less hits, due to the characteristic of the dataset (see Section 3). The encyclopedia authors intended to compose articles which in general contained as few redundancy as possible - an aspect inherent to the nature of encyclopedia articles.

By altering the query with multiply hyponyms like nations, names and others from a base taxonomy, the recall was sufficient enough to identify two reasonable patterns as seen in Table 1.

Furthermore, to build a larger set of German Hearst patterns, the original patterns from Hearst[4] were translated and validated against the chosen dataset. As a matter of fact not all English patterns could have been translated to meaningful German patterns. The successfully translated patterns were ranked by

[3] Resource Description Framework

	Deutsch (german)
5	NP *wie zum Beispiel* NP... *(for example)*
6	NP *ist ein[e]* NP... *(is a)*

Table 1. Hearst Patterns (extracted by multipy-hyponym search)

the count of hits their lexical part returned. The best four are listed in Table 2. Those patterns and the patterns from Table 1 were used in the second step, the extraction process.

	English	German
1	NP *such as* NP...	NP *wie etwa* NP...
2	NP *including* NP...	NP *einschließlich* NP...
3	NP *or other* NP...	NP *und andere* NP...
4	NP *and other* NP...	NP *oder andere* NP...

Table 2. Hearst Pattern Translation

The total amount of hits found for the best ranked patterns can be seen in Table 3. Other patterns and slight variations of the successful ones did occur far below thousand times. Over all, more than 100 different variations were tested against the dataset. The process of finding the correct pattern was done by hand as the pattern had to be assembled syntactically and semantically correct.

2.2 Extraction of Taxonomic Relations

The process started by collecting all relevant documents which contain at least one lexical part of the Hearst patterns as described in 2.1. For this task a straightforward full text retrieval system has been utilized. All six patterns collected 22,403 occurrences of their lexical parts within 18,000 documents, which covers about 8% of the documents in the dataset.

In the preprocessing step, the documents have been split into sentences and words are tagged with their part-of-speech resp. if they are part of a noun phrase or not. All tasks were performed using a maximum entropy model trained on a German News corpus[8]. On the preprocessed corpus the Hearst patterns were applied and the resulting relations were stored in a triple store along with their location. To enable the later improvement through the Yahoo! Search Web-Service, all extracted relations were also mapped to their extracting Hearst patterns. The hypernym of a relation can exist in multiple documents, so all locations had to be stored with every hyponym. The extracted relations, including their hypernyms and hyponyms are stored using RDF (Resource Description Framework)[2].

To give an example consider the following preprocessed sentence.

```
...(bekannte/ADJA Krankheiten/NN)/NP wie/KOKOM zum/APPRART Beispiel/NN
Malaria/NN ,/COMMA Masern/NN oder/KON Aids/NN ...
```

The tag NP represents a noun-phrase; for the other tags see the STTS - Tagset[4]. The sentence is represented in RDF as

```
<rdf:Description rdf:about="bh://200/term/Diseases">
  <bh:location>bh://200/23423/23</bh:location>
  <bh:location>bh://200/54322/788</bh:location>
  <bh:hyponym>
    <bh:word>bh://200/term/Malaria</bh:word>
    <bh:location>bh://200/23423/27</bh:location>
  </bh:hyponym>
  <bh:hyponym>
<bh:word>bh://200/term/measles</bh:word>
<bh:location>bh://200/23423/28</bh:location>
  </bh:hyponym>
  <bh:hyponym>
<bh:word>bh://200/term/Aids</bh:word>
<bh:location>bh://200/54322/791</bh:location>
  </bh:hyponym>
</rdf:Description>
```

We defined our own RDF Schema to refer hypernyms to the documents they are found in and to provide an extendable knowledge base. Every found hypernym is described by its locations and the related hyponyms. Locations are referenced through a `http://CorporaId/DocumentId/Wordlocation`URI[5] scheme. Hypernyms itself are described using the ***rdf:Description*** element and their locations in the dataset are defined by the ***bh:location*** element. Every found hyponym is added with a ***bh:hyponym*** element, which specifies the term via the ***bh:word*** element and again the location of the term with the ***bh:location*** tag.

Hence, the usage of RDF enables a semantically normalized storage of the relations. Through this schema, other semantic properties can be incorporated later and the relations can easily be linked to different taxonomies or other datasets. As there are many tools to manipulate RDF, the extracted relations can be altered by domain experts as needed. The location of the terms is stored to enable queries on the triple store to directly link to relevant documents.

2.3 Pattern Validation

The extracted taxonomy contains a huge amount of incorrect relations. To improve the ratio of correct to incorrect relations a verification step was added

[4] http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/stts.asc

[5] Uniform Resource Identifier (URI)

which classifies relations into false and true by using an external source of evidence, in our case the Yahoo! Search Web-Service. It provides an interface to send search queries to the Yahoo! Internet search engine. Within the Yahoo! Web-Service the language of the index to be used and the interpretation of the terms as a single phrase query can be set. Those parameters are essential for the used validation process.

After extracting around 39,500 relations from the dataset, every single relation was recreated by combining the hypernym, the lexical part of the patterns and the hyponym. After that, 39,500 phrase queries were created like:

- "Krankheiten *wie etwa* Aids"
  (Diseases *like* aids)
- "Krankheiten *wie etwa* Masern"
  (Diseases *like* measels)
- "Krankheiten *wie etwa* Medikamente“
  (Diseases *like* drugs)

These phrase queries were send to the Yahoo! Search Web-Service and the number of the result list was used as a weighting factor. In the first evaluation all queries with a result more than zero were marked as well-proven relations. The results of the validation are shown in Section 3.2.

In a second evaluation period the threshold to mark a relation as correct was set to a minimum of five hits from the Yahoo! Search Web-Service. This raise resulted in a significant improvement of the precision in which the validated relations correspond to the hand picked correct relations.

False marked relations were removed from the taxonomy. To guarantee a high quality taxonomy, the relations need to be proofed by human intervention. As this task is unpreventable in ontology learning, the Yahoo! validation can be used as filter for high likely correct relations.[3]

3 Evaluation Results

The corpus used to apply our method, consisted of approximately 225,000 documents stored in a relational database. The text origins from a German digital encyclopedia. Hence, the documents are very precise according to their content. An inherent feature of the encyclopedic dataset is the heterogeneous distribution of topics, which potentially leads to an universal taxonomy. The longest document contains 40,000 words, whereas the mean average is around 110 words. Very few filling words and recurrences occur throughout the documents as the dataset was designed for physical print.

The described features were sufficient to claim that machine learning algorithms are not suitable to train a classifier due to the too less similar known relations occuring throughout the corpus. Through an initial dataset survey, we were able to confirm this assumption.

3.1 Accuracy of the Extraction Process

As no predefined ground truth was given, all correct relations had to be annotated by hand. To be able to gain reasonable figures, we first had to declare that our patterns found all their occurrences in the corpus and describe a theoretical recall of 1. We checked each relation against German dictionaries and a German WordNet.

The amount of relations is the sum over all hypernym counts minus wrong hits caused by the part-of-speech tagger and noise in the dataset.

$$Relations = \sum_{i=1}^{rdf:Description} (bh : hyponym)_i \tag{1}$$

As seen in Table 3, the pattern *"wie etwa (such as)"* and *"und/oder andere (and/or other)"* had the best ratio between correct and total relations found, where *"ist ein (is a)"* performed worst. Whereas the overall result of 7.7 percent correct relations turns out to be quite reasonable according to the initial results from the original work. [4]

Pattern	Pattern Hits	Relations	correct Relations	Ratio (%)
ist ein (is a)	3,870	15,342	684	4.5
wie etwa (such as)	3,179	3,248	341	10.5
einschließlich (including)	7,897	6,321	405	6.4
zum Beispiel (for example)	5,827	6,164	367	6.0
und/oder andere (and/or other)	1,630	8,405	939	11.2
Total	**22,403**	**39,480**	**2,736**	**7.7**

Table 3. Hearst pattern extraction results

In the Figures 1 and 2 snippets of the extracted taxonomy are shown. Figure 1 demonstrates a problem which occurred frequently through the extracted relations, where instances are mapped to different levels of category abstraction where none is explicitly wrong. A method to automatically correct this shifts between the levels of a hierarchy can't be applied. This kind of refinement has to be done by a human domain expert.

All relations had to be annotated by hand because many of them were new relations and not coded in machine processable formats yet. This made the whole evaluation process difficult and time consuming, because the relations had to be checked after the Hearst extraction process and after the refinement process with the Yahoo! Web-Service.

Figure 2 presents a small part of the extracted taxonomy. It typified the broad domain variations of taxonomy extracted from an encyclopedic text. Over 60 percent of all extracted relations were merged to one large taxonomy. The rest formed standalone hypernym relations.

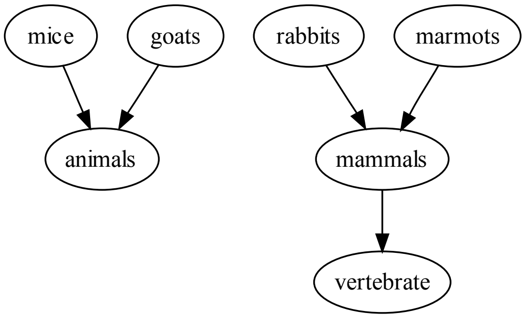

Fig. 1. different levels of taxonomy abstraction (manually translated from German)

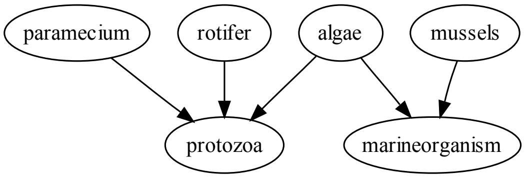

Fig. 2. Illustration of the resulting semantic net (manually translated from German)

To incorporate the extraction into a more automated process of ontology learning, the amount of false hits had to be reduced significantly.

3.2 Accuracy of the Yahoo! Validation

In this section the evaluation of the Yahoo! validation is presented. The focus of the evaluation is set to find those pattern which can be improved best by the Yahoo! validation.

The total amount of correct relations is called *Hearst True Positives*, wheres the *Hearst Positives* are the total amount of relations extracted with the Hearst patterns. The *Hearst Positives* act as ground truth for the evaluation of the Yahoo! results. The *Hearst True Positives* were found by evaluating all relations against German dictionaries and German WordNet like we did with the relations from the Hearst pattern.

As a precondition the recall of the Hearst patterns itself needs to be considered as 1, which reflects the assumption that the pattern did find all occurrences. While we are aware that more correct relations exists, the assumption is required to make reasonable assumption about precision or accuracy on the data set.

As not all positive hits from the Yahoo! refinement could be considered as true, their results had to be validated by hand either, which are called the Yahoo! True Positives.

In the first two columns of the Tables 4 and 5 the Yahoo! Positives and Yahoo! True Positives show the total count of returned queries with a result set larger than zero and the amount of queries which reflect a correct relation. In the column Yahoo! Precision those values are compared and give a first hint of the quality of the Patterns.

Pattern	Yahoo! Positives	Yahoo! True Positives	Yahoo! Hearst Correlation	Yahoo! Precision	Yahoo! Recall
ist ein (is a)	1,917	248	12.5%	12.9%	36.26%
wie etwa (such as)	107	72	3.3%	67.3%	21.11%
einschließlich (including)	328	135	5.2%	41.2%	33.33%
zum Beispiel (for example)	137	108	2.2%	78.8%	29.43%
und/oder andere (and/or other)	1,622	608	19.3%	37.5%	64.75%
Total	**4,111**	**1,171**	**10.4%**	**28.5%**	**42.8 %**

Table 4. Yahoo!-Hearst relation validation with a threshold of 1 hit.

As seen in both Tables the Pattern *"wie etwa (such as)"* and *"zum Beispiel (for example)"* outperformed the others in case of precision. So a first conclusion can be drawn by stating that those two patterns gain a high set of correct relations.

$$Yahoo!\ Hearst\ Correlation = \frac{Yahoo!\ Positives}{Hearst\ Positives} \tag{2}$$

To evaluate how much of the relations extracted by the Hearst patterns were found by the Yahoo! validation step, the Yahoo! Hearst Correlation point out, how good a specific pattern can be verified by the Yahoo! Web-Service. Among this value, the Yahoo! Recall state how much of the correct extracted relations were also found by Yahoo! and classified as correct.

$$Yahoo!\ Recall = \frac{Yahoo!\ True\ Positives}{Hearst\ True\ Positives} \tag{3}$$

To give a qualitative statement it is important to argue on precision and recall. For example, even though the pattern *"ist ein (is a)"* produced the second highest recall, the precision of 12.9 percent was far the worst, which can be seen easily in Figure 3.

Additionally, in Figure 3 the count of total returned hits from Yahoo! and the correct relations are shown. It is notable, that if a threshold with 5 is used, the difference between the total and correct relations vanishes with the pattern *"wie etwa (such as)"* and *"zum Beispiel (for example)"*. Nevertheless, even the other patterns show a reasonable improvement by adapting the threshold.

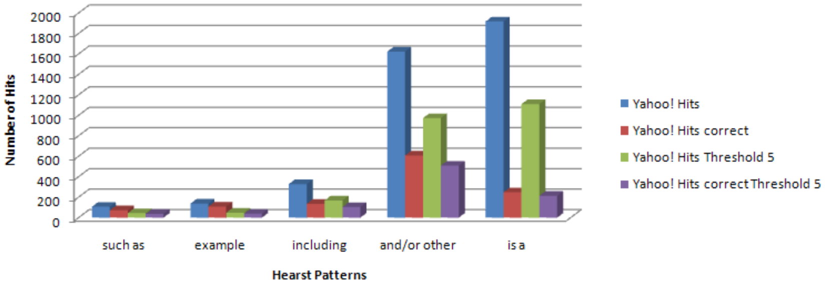

Fig. 3. Yahoo! Validation(total to correct hits with threshold 1 and 5)

Pattern	Yahoo! Positives	Yahoo! True Positives	Yahoo! Hearst Correlation	Yahoo! Precision	Yahoo! Recall
ist ein (is a)	1109	213	7.2%	19.2%	31.14%
wie etwa (such as)	44	37	1.4%	84.1%	10.85%
einschließlich (including)	168	104	2.7%	61.9%	25.68%
zum Beispiel (for example)	48	39	0.8%	81.3%	10.63%
und/oder andere (and/or other)	973	508	11.6%	52.2%	54.10%
Total	**2,342**	**901**	**5.9%**	**38.5%**	**32.93 %**

Table 5. Yahoo!-Hearst relation validation with a threshold of 5 hits

4 Conclusions

Throughout this paper we presented that the combination of a lexico-syntactic extraction of hypernym relations with the use of Hearst Pattern and a validation with an external source of evidence leads to promising results. Both techniques performed well on their field of use.

According the Hearst Pattern it became clear, that the right choice of patterns is an important factor of the overall quality of the extracted relations. As the results have shown, the German patterns *"wie etwa (such as)"* and *"zum Beispiel (for example)"* performed best in both steps of the process in terms of precision.

To sum up, the evaluation showed that, if the right pattern were chosen, the method extracted an accurate set of high quality hypernym relations from the given text. Based upon those relations a taxonomy was built which can provide a navigational entry point to a subset of the corpus as well as an enhancement for queries. Furthermore, based on our results one can argue that Hearst patterns can be successfully translated between western languages especially since validation via search engines shows to be beneficial.

Acknowledgments

The Know-Center GmbH Graz is funded within the Austrian COMET Program - Competence Centers for Excellent Technologies - under the auspices of the Austrian Ministry of Transport, Innovation and Technology, the Austrian Ministry of Economics and Labor and by the State of Styria. COMET is managed by the Austrian Research Promotion Agency FFG.

Furthermore we want to thank the *Bibliographisches Institut & F.A. Brockhaus AG Mannheim* for the provision of their data.

References

1. John F. Gantz. The expanding digital universe. *Technical Report*, March 2007.
2. Frank Manola and Eric Miller. *RDF Primer*. W3C, February 2004.
3. John Davies, Rudi Studer, and Paul Warren. *Semantic Web Technologies: Trends and Research in Ontology-based Systems*. John Wiley and Sons, July 2006.
4. Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics*, pages 539–545, Morristown, NJ, USA, 1992. Association for Computational Linguistics.
5. Wei Liu Wilson Wong and Mohammed Bennamoun. Progress and open problems in ontology engineering. *Technical Report*, July 2006.
6. P. Cimiano, A. Pivk, L. Schmidt-Thieme, and S. Staab. Learning taxonomic relations from heterogeneous sources of evidence. In *Ontology Learning from Text: Methods, Evaluation and Applications*, 2005.
7. Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. Learning syntactic patterns for automatic hypernym discovery. In Lawrence K. Saul, Yair Weiss, and Léon Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 1297–1304, Cambridge, MA, 2005. MIT Press.
8. Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. The TIGER treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*, Sozopol, 2002.