Semantic Indexing with Typed Terms Using Rapid Annotation

CHRIS BIEMANN

Illustrated with an example of a manual semantic resource for German, the use of typed index terms for semantic indexing is proposed. Types group index terms in the same semantic category and can be used by any search or cluster mechanism. To obtain semantic resources, a method to rapidly annotate large corpora is described in detail.

1 Introduction

In the literature, the term "semantic indexing" is used in a variety of definitions. They all have in common that in some way index terms are assigned to a document that are not necessarily contained in the document itself, but related to its important terms.

Most publications discuss the application of "semantic indexing" in order to improve Information Retrieval (IR) - in this case, the retrieval performance is maximized regardless of what exact index terms are used.

According to (Mihalcea & Moldovan 2000), there are three main approaches to incorporate semantic information in the IR process: conceptual indexing, semantic indexing and query expansion.

While query expansion leaves the representation of the document unchanged and tries to add search terms to the user's query, conceptual and semantic indexing both add information to the index terms of the document. The difference between the latter two is rather artificial: conceptual indexing uses a domain ontology or taxonomy for assigning concepts to documents (although other authors determine what they call "concept" in a more statistical way, e.g. Kang 2003), semantic indexing uses lexical-semantic resources like WordNet (Vorhees 1998) for assigning synset labels to previously disambiguated words (see Sanderson 2000 for a survey). The WordNet (Miller 1990) hierarchy and synset structure is used to enrich the document with additional information.

But also other applications like document clustering, categorization (see e.g. Rosso et al. 2004) and summarization can benefit from semantic indexing as a means to reduce data sparseness.

Alleviating the bottleneck that resources like WordNet are not available in large scale for most natural languages or specific domains, a method to rapidly create semantic resources is introduced in section 3. Before, the need and motivation for typed index terms is pointed out in section 2.

2 Typed Index Terms

Approaches to enrich documents with semantic information, for example the Semantic Web movement, do not simply calculate some vector representation of the document (cf. Salton et al. 1975), but specify the entities and relations between them by assigning types. A well known example of a retrieval system that uses typed information is the CiteSeer portal (Giles et al. 1998), where information like URL, title, authors, citations and citation contexts are automatically obtained from computer science research papers. These fields can be used to inter-connect the documents and to specify the retrieval.

While these pieces of information are stored separately and serve as a characterization of the document, the Semantic Web community favours the annotation to be done in-place as semantic annotation. Having a document collection fully annotated by assigning correct ontological types to entities in the running text, an index with typed terms can be realized by simply collecting the entities and sorting them according to their types. But as (Erdmann et al. 2000) point out: Despite convenient tools like OntoBroker (Decker et al. 1999) exist, authors are hard to convince to actually carry out the annotation of their document. Even if they do, a large portion is partially or totally incorrect, mostly because the types are not well defined, giving rise to a variety of interpretations and possibilities.

In this essay, the way of separately storing index terms and their types as meta-information is followed. When having types for index terms at hand, it is not only possible to search document collections in a structured way, but also calculate similarities for clustering and classification on a specific subset of terms. This facilitates answering questions like "*Which documents contain the persons Bill Gates and Steve Jobs in connection with at least two computer companies?*" - assuming our index knows about persons and computer companies. Documents get comparable not only in the keywords itself, but in index term categories of different types as well. For example, It might be interesting to cluster documents mentioning similar persons, regardless of what situations are described. Given a first article on politics where company leader A receives a weapon purchase order from politician B and a second article where A invites B on a holiday trip, the similarity between the articles regarding index terms of only type person exhibits interesting relations that will be obfuscated when taking untyped index terms on politics and holiday places into account.

The granularity needed for solving specific problems is defined by the questions asked and can be split into a generic part and a domain-specific part. The generic part that is almost independent of document type, genre and domain is the identification and categorization of proper nouns like persons, organizations and locations (Named Entity Extraction), domain specific types include e.g. names of genes in the medical domain or spare parts in the automotive domain, while domain specific relations of those domains can be e.g. treatment-for-illness or part-for-model.

Due to the efforts undertaken in competition tasks like the MUC shared task (Grishman & Sundheim 1996), Named Entity Recognition is more or less solved for English and some more common languages. For these languages, there are also generic lexical word nets like WordNet to enrich index terms with e.g. more generic terms or hypernyms. How to obtain typed terms and relations for the domain specific part and in a language-independent way will be subject of the next section.

3 Rapidly Annotating Corpora

In the following, some data sources that provide input for a set of annotation tools that aid (probably unskilled) users in the annotation process are described. The process of annotation is understood as assigning types to single terms (semantic primitives) or pairs of terms (semantic relations) on a generic level. Unlike the annotation tools mentioned in the introduction, types are not assigned to single instances in documents, but once and for all per term. Although not dealing with word ambiguities and situative contexts, the method assigns the 'main meaning' of terms and has considerable benefits in annotation speed that makes it possible to annotate a whole domain in a very short time. Rather than annotating single documents, a whole set (corpus) of documents is annotated at the same time. What specific primitives and relations are to be annotated is specified beforehand and handled flexible by the tools. In this work, a configuration for the semantic annotation of a large German corpus of "Projekt Deutscher Wortschatz"[1] is presented throughout all examples.

The goal of corpus annotation is to assign all primitives and relations needed by the application in question. When annotating primitives, a list of all words in the corpus could be presented to the user, asking her to assign all matching primitives to each word. When sorted by word frequency in de-

[1] see *http://www.wortschatz.uni-leipzig.de*

scending order, a text coverage of 70%-80% is reached by annotating merely the 20,000 most frequent terms. But a presentation mechanism for relations is more complicated: Presenting all possible pairs of the most frequent N words will leave N·N pairs to be looked through by the annotator, of which most of the word pairs are not related at all.

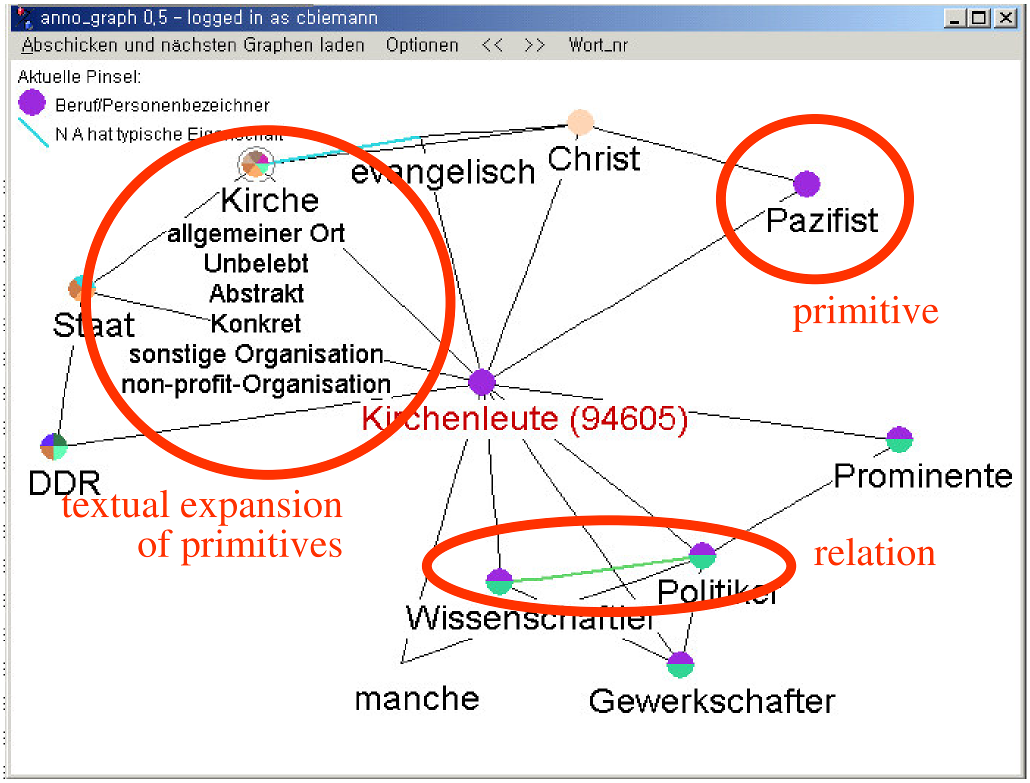


Figure 1: Graphical annotation tool displaying co-occurrences of "Kirchenleute" (church people). Coloured nodes reflect different primitives, coloured edges, e.g. between "Wissenschaftler" (scientist) and "Politiker" (politician), are typed relations.

To extract word pairs from corpora, co-occurrence analysis as described in (Biemann et al. 2004a) can provide word pairs that tend to occur more often together than to be expected from their frequencies. By a significance measure, the related words can be ordered by strength. Co-occurrences reflect human associations (e.g. butter-bread, dog-cat, ...) and bear many more relations of any kind then randomly pairing words for relation annotation. A major advantage of statistical methods like co-occurrence analysis is its language-independence: For words are treated as strings with no assumptions made on the specific language, this kind of analysis can be carried out for all natural languages[2] in the same way.

Exploiting this fact, a graphical annotation tool was constructed that presents highly significant co-occurrences of a reference word and allows the annotation of relations and primitives, see the screenshot in figure 1.

In the graphical annotation tool, a line between two terms means that the connected terms co-occur significantly often together. The example of "Kirche"(church) shows that multiple primitives can be assigned to one term. Here, "Kirche" is annotated with "general location", "animate -", "abstract", "concrete", "organization" and "non-profit organization" as a part of the semantic primitives defined. Due to the inherent ambiguity of church as house and organization, the primitives are contradictory in this case.

While co-occurrences reflect associations between words, they rather reflect syntagmatic than paradigmatic relations (following Saussure 1916): In co-occurrences, more relations between typical heads and modifiers are found than classical semantic relations like synonymy or hyponymy. On the other hand, exactly the latter relations are the most interesting to cover phenomena like subsumption and to construct a term hierarchy.

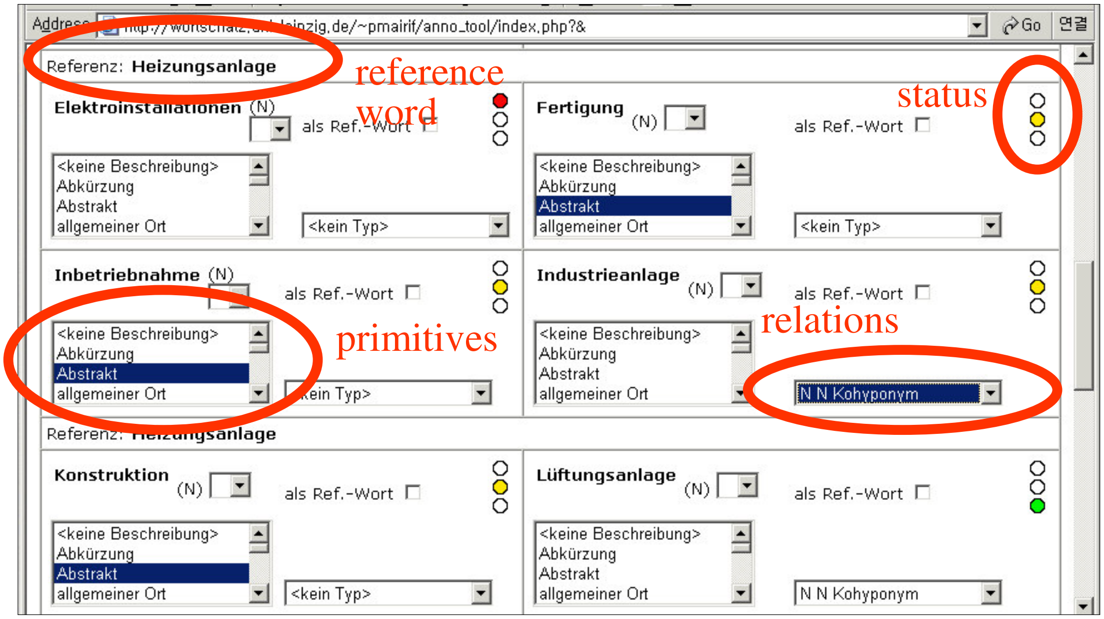
As (Biemann et al. 2004b) shows, an iteration step in the calculation of co-occurrences yields a higher rate of paradigmatic pairs in what is called co-occurrences of higher orders. Intuitively, a term pair co-occurs in a higher order, if the terms tend to occur in similar contexts. The most frequent relation found in this data source is co-hyponymy. Table 1 shows on some examples how the number of relations per type differ in the different data sources.

Ratio of relations Co-occurrences : Co-occ. of higher order	**Relations** (selection)
10 : 1	Typical-place-for Typical-activity-for-place
5 : 1	typical-Object-of-Verb typical-property
1 : 1	part-of, antonym-of
1 : 5	is-a, synonym-of cohyponym-of
1 : 10	adjective-verb-derivate

Table 1: Different data sources prefer different relations - Co-occurrences of higher orders propose more paradigmatic than syntagmatic relations.

[2] although languages without whitespac between words like Chinese and Japanese need a preprocessing step to identify word boundaries

To present co-occurrences of higher orders as well as other data sources of arbitrary source to the user, a web-based annotation tool was developed that allows assigning the same primitive and relation types as the graphical tool, as depicted in figure 2.



Referenz: **Heizungsanlage**

Elektroinstallationen (N) als Ref.-Wort ☐ ● ○ ○
<keine Beschreibung>
Abkürzung
Abstrakt
allgemeiner Ort
<kein Typ>

Fertigung (N) als Ref.-Wort ☐ ○ ● ○
<keine Beschreibung>
Abkürzung
Abstrakt
allgemeiner Ort
<kein Typ>

Inbetriebnahme (N) als Ref.-Wort ☐ ○ ● ○
<keine Beschreibung>
Abkürzung
Abstrakt
allgemeiner Ort
<kein Typ>

Industrieanlage (N) als Ref.-Wort ☐ ○ ● ○
<keine Beschreibung>
Abkürzung
Abstrakt
allgemeiner Ort
N N Kohyponym

Referenz: **Heizungsanlage**

Konstruktion (N) als Ref.-Wort ☐ ○ ● ○
<keine Beschreibung>
Abkürzung
Abstrakt
allgemeiner Ort
<kein Typ>

Lüftungsanlage (N) als Ref.-Wort ☐ ○ ○ ●
<keine Beschreibung>
Abkürzung
Abstrakt
allgemeiner Ort
N N Kohyponym

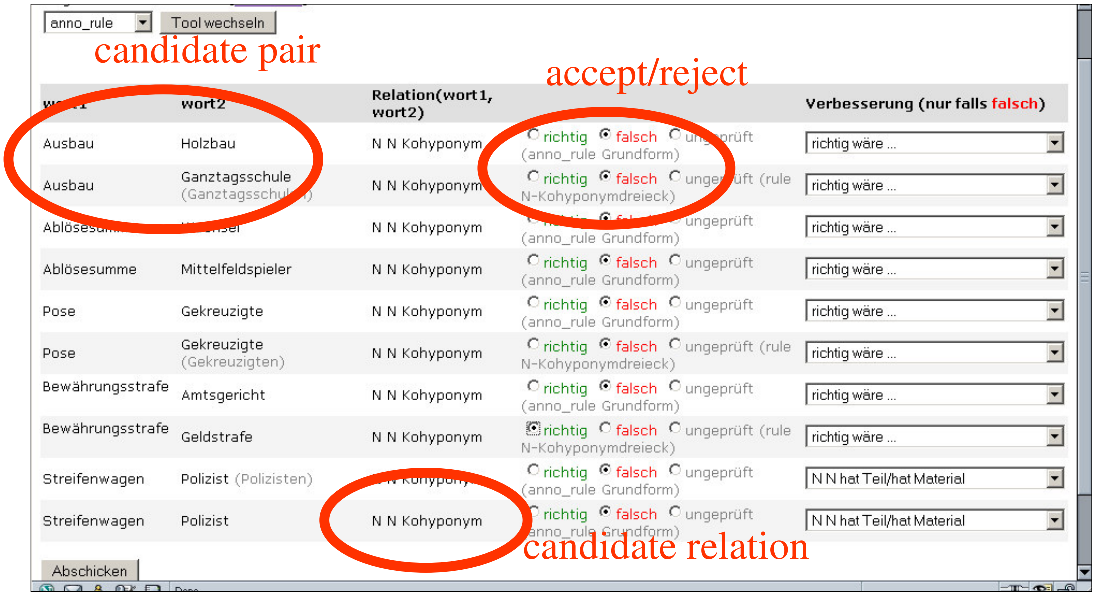
Figure 2: Web-based annotation tool displaying co-occurrences of higher orders of "Heizungsanlage" (heating system). Relations are defined from reference word to the terms displayed in the boxes below.

These two modes of presenting candidate word pairs to the annotators has the advantage over presenting simple lists of words, that words belonging to one semantic field are presented together, providing context and aiding understanding of terms that are not recognized when presented isolated.

But there are more possibilities to ease and speed up the annotation process. In the following, a method will be described that results in high-quality candidate primitives and relations and uses the already existing annotations and some rules in order to find these.

As an illustrating example for the terms *cat*, *animal* and *dog*, let us assume that the following primitives and relations have been annotated already: LIVING(*cat*), LIVING(*animal*), IS-A(*dog*, *animal*) and COHYPONYM(*dog*, *cat*).

The IS-A relation usually inherits properties from a hypernym to its hyponyms. Defining a corresponding rule yields LIVING(*dog*) as a candidate. Another rule stating that co-hyponyms have the same hypernyms comes up with IS-A(*cat*, *animal*) as a candidate relation. These candidates can be easily accepted or rejected by the annotator in a tool shown in figure 3.



wort1	wort2	Relation(wort1, wort2)		Verbesserung (nur falls falsch)
Ausbau	Holzbau	N N Kohyponym	○ richtig ● falsch ○ ungeprüft (anno_rule Grundform)	richtig wäre ...
Ausbau	Ganztagsschule (Ganztagsschulen)	N N Kohyponym	○ richtig ● falsch ○ ungeprüft (rule N-Kohyponymdreieck)	richtig wäre ...
Ablösesumme	[illegible]	N N Kohyponym	○ richtig ● falsch ○ ungeprüft (anno_rule Grundform)	richtig wäre ...
Ablösesumme	Mittelfeldspieler	N N Kohyponym	○ richtig ● falsch ○ ungeprüft (anno_rule Grundform)	richtig wäre ...
Pose	Gekreuzigte	N N Kohyponym	○ richtig ● falsch ○ ungeprüft (anno_rule Grundform)	richtig wäre ...
Pose	Gekreuzigte (Gekreuzigten)	N N Kohyponym	○ richtig ● falsch ○ ungeprüft (rule N-Kohyponymdreieck)	richtig wäre ...
Bewährungsstrafe	Amtsgericht	N N Kohyponym	○ richtig ● falsch ○ ungeprüft (anno_rule Grundform)	richtig wäre ...
Bewährungsstrafe	Geldstrafe	N N Kohyponym	● richtig ○ falsch ○ ungeprüft (rule N-Kohyponymdreieck)	richtig wäre ...
Streifenwagen	Polizist (Polizisten)	N N Kohyponym	○ richtig ● falsch ○ ungeprüft (anno_rule Grundform)	N N hat Teil/hat Material
Streifenwagen	Polizist	N N Kohyponym	○ richtig ● falsch ○ ungeprüft (anno_rule Grundform)	N N hat Teil/hat Material

Figure 3: Tool for accepting or rejecting candidates obtained by rules

With this set of tools presenting terms and term pairs extracted by the appropriate statistical data sources, an average speed of about 5 units (primitive or relation) per minute is reached after a small number of training hours.

In total, we achieved about 150,000 primitives and 150,000 relations for over 80'000 distinct terms in about 1000 hours of annotation, covering to a large extent world-knowledge semantics of German.

Conclusion

Assigning and using typed index terms and relations between them was motivated by defining different views on document sets for similarity measures and by enhanced retrieval possibilities. For acquiring these types, a framework for rapidly annotating corpora was proposed. Unlike other annotation approaches, information is assigned to terms and not to their actual manifestations within documents, speeding up the annotation process significantly. A set of graphical tools was introduced that made it possible to obtain a semantic network of general German language and is easily applicable to other languages.

References

Biemann, C., Bordag, S., Heyer, G., Quasthoff, U., Wolff, Chr.: *Language-independent Methods for Compiling Monolingual Lexical Data*. Proceedings of CicLING 2004, Seoul, Korea and Springer LNCS 2945, pp. 215-228, Springer Verlag Berlin Heidelberg, 2004(a)

Biemann, C.; Bordag, S.; Quasthoff, U. (2004*): Automatic Acquisition of Paradigmatic Relations using Iterated Co-occurrences*, Proceedings of LREC2004, Lisboa, Portugal, 2004(b)

Decker, S.; Erdmann, M.; Fensel, D. & Suder, R. *Ontobroker: Ontology Based Access to Distributed and Semi-Structured Information.* In Meersman, R. et al. (Eds), Database Semantics: Semantic Issues in Multimedia Systems, pp. 351-369, Kluwer Academic Publisher, 1999

Erdmann, M.; Maedche, A.; Schnurr, H.-P. & Staab, S: *From Manual to Semi-automatic Semantic Annotation*: *About Ontology-based Text Annotation Tools.* COLING-2000 Workshop on Semantic Annotation and Intelligent Content, Luxembourg, 2000

Giles, C.L; Bollacker, K.D. & Lawrence, S. *CiteSeer: An Automatic Citation Indexing System.* Third ACM conference in Digital Libraries, ACM Press, New York, pp. 89-98, 1998

Grishman, R. & Sundheim, B. *Message Understanding Conference-6: A Brief History*. Proceedings of COLING-96, Copenhagen, Denmark 1996
Kang, B.-Y. *A Novel Approach to Semantic Indexing Based on Concept*. Proceedings of the 41st Annual Meeting of the ACL, Sapporo, Japan 2003

Mihalcea, R. & Moldovan, D. *Semantic Indexing Using WordNet Senses*. proceedings of the ACL Workshop on IR. & NLP, Hong Kong, 2000.

Miller, G. A.. *Wordnet - an on-line lexical database*. International Journal of Lexicography 3(4):235-312. 1990

Rosso, P; Ferretti, E.; Jimènez, D. and Vidal, V. *Text Categorization and Information Retrieval Using WordNet Senses*. Proceedings of the Second Global Wordnet Conference, Brno, Czech Republic 2004

Salton, G., Wong A. & Yang, C.S. *A Vector Space Model for Automatic Indexing*. Communications of the ACM, 18(11), pp. 613 - 620, 1975.

Sanderson, M. *Retrieving with good sense.* Information Retrieval, 2(1):49-69
Vorhees, E,M. *Using WordNet for text retrieval. In WordNet*, An Electronic Lexical Database, pp. 285-303, The MIT Press 1998