EFFICIENT LITERATURE RESEARCH BASED ON SEMANTIC TAGNETS
Implemented and Evaluated for a German Text-corpus

Uta Christoph, Daniel Götten and Karl-Heinz Krempels
Informatik 4, Intelligent Distributed Systems Group, RWTH Aachen University, Aachen, Germany

Keywords: Semantic networks, Literature research, Text corpus, Text analysis.

Abstract: In this paper we present an approach that is capable to automatically generate semantic tagnets for given sets of german tags (keywords) and an arbitrary text corpus using three different analysis methods. The resulting tagnets are used to estimate similarities between texts that are manually tagged with the keywords from the given tagset. Basically, this approach can be used in digital libraries to provide an efficient and intuitive interface for literature research. Although it is mainly optimized for the german language the proposed methods can easily be enhanced to generate tagnets for a given set of english keywords.

1 INTRODUCTION

Due to the large amount of data available in the world wide web data structuring is an important topic today. One famous approach is the *semantic web* as proposed by Tim Berners-Lee (Berners-Lee et al., 2001). The idea is to structure the available data by its semantic meaning to provide much better access methods and possibilities for automatic analysis. Since no suitable semantic wordnets are available for the german language the idea is to build a system that allows to estimate semantic relations for a given set of german words. In the following an approach will be presented that structures *keywords (tags)* and documents by their semantic information and similarities.

Normally, no semantic information is available for information stored in digital libraries and solely simple search interfaces are provided that allow to search for documents by its title or author names. Apparently, research of similar documents is hard and time-consuming in such cases. One approach to solve this problem is to avail ss that are used to describe the content of single documents and enable the search for documents by their tags.

Although tags allow to categorize documents they do not completely solve the problem described above as illustrated by the following example. Given two similar documents d_1 and d_2 annotated by tags from two disjunct sets S_1 and S_2. It might not be possible to recognize the similarities between d_1 and d_2, if no information about semantic relations between the tags of both sets is given. Especially for a large set of available similar tags to describe each document this problem may occur quite frequently.

The presented approach is based on similarity analysis of documents annotated manually with tags. These document similarities are estimated by automatically extracting semantic relations between the tags of the given set. Both the given tags and the extracted relations form a network called *tagnet*. The derived network of documents is called *similarity net*. Apparently, both networks can be used to make literature research more efficient, by visualizing them in a suitable way. The resulting system provides a quite intuitive interface that enables non-technical users to research literature by surfing through the similarity net.

At first we discuss in Section 2 relevant problems that may occur during automatic text analysis. We show that these problems mainly result from morphological special cases of the german language. As possible solutions the approaches of lemmatization and stemming will be discussed. Afterwards, the idea of the implemented system (the *tagnet builder*) that builds the described networks is introduced in Section 3. Finally, the results are evaluated in Section 4.

Christoph U., GÃűtten D. and Krempels K.
EFFICIENT LITERATURE RESEARCH BASED ON SEMANTIC TAGNETS - Implemented and Evaluated for a German Text-corpus.
DOI: 10.5220/0002805400480054
In *Proceedings of the 6th International Conference on Web Information Systems and Technology (WEBIST 2010)*, page
ISBN: 978-989-674-025-2

2 PROBLEMS

Automatic approaches of text analysis are quite error-prone without adaptation to a given scenario. Especially in the case of automatic text analysis of german[1] texts problems occur due to the morphological special cases in the german language. Concerning statistical approaches in text analysis three problems have to be taken into account:

1. multilingualism
2. morphologicial special cases
3. understanding of texts

Since the considered corpus solely consists of german texts we do not consider multilingualism for the moment. For this reason only the problems (2) and (3) have to be discussed.

2.1 Morphological Special Cases

The main objective of the tagnet builder is the extraction of semantic information from keywords based on statistical frequency analysis on the considered text corpus. These frequencies have to be as accurate as possible, so a reliable mapping of all given keywords to unique stems is required. Furthermore, numbers and stop words can be deleted, since they are not relevant for the analysis.

2.1.1 Stemming

Stemming is an algorithmical approach for the estimation of unambiguous stems for words of a given language. A famous approach is the *porter stemmer* (Porter, 1980). An implementation for german words based on this approach is available in terms of *Snowball*[2].

Unfortunately, stemming algorithms do not generate real stems, but unique pseudo stems by suffix stripping based on predefined rules without consideration of grammatical characteristics. Especially for the case of german words two different words might be mapped to one stem although there exists no semantic similarity. Thus, this simple stemming approach may deteriorate the quality of the semantic tagnet generated by the tagnet builder.

2.1.2 Lemmatization

To provide mappings to real stems *lemmatization* has to be used. This lemmatization is often based on fullform lexicons that have to be defined in prior. Unfortunately, no complete fullform lexicons are available for the german language due to the possibility of compound words. This problem can be solved with the help of a learning lemmatizer as it has been proposed by P. Perera and R. Witté (Perera and Witte, 2005). Basically, this lemmatizer processes words in three phases.

At first it checks if the current word already is part of the existing fullform lexicon to lemmatize it immediately by its corresponding stem. If the current word is not part of the existing lexicon, it will be classified with respect to its POS-Tags[3]. Then suitable rules for suffix stripping will be applied to get candidate stems that are returned and inserted into the lexicon[4].

2.2 Understanding of Texts

Automatic understanding of texts is a complex problem and still a research topic today. Especially if frequency analysis is used, problems may occur that are referable to words with different meanings in subject to the context.

The approach proposed in this paper will assume that the tagnet builder is solely used for texts of one topic. Thus, the described problem normally will not occur in this scenario, but has to be addressed if texts of different topics should be analysed.

3 APPROACH

Given an unstructured list of tags, the described system is able to find semantic relations between these tags and store them in a suitable database. Based on this semantic tagnet text similarities can be calculated for a corpus of tagged texts.

3.1 Tagnet Builder

The tagnet builder uses three different approaches to gain as much information about semantic relations between the given tags as possible. These three approaches are:

1. rule-based extraction of semantic relations
2. lexicon-based extraction of semantic relations
3. statistical estimation of semantic relations based on occurrences in a given set of documents

In the following the approaches will be described in detail.

[1] Or other inflectional complex languages.

[2] http://snowball.tartarus.org

[3] Part-of-speech tagging basically denotes the tagging of words with its grammatical characteristics.

[4] The lexicon may contain intermediately wrong stems, that will be corrected over time.

3.1.1 Rule-based Approach

The problem of compound words has been mentioned in the sections above. Nevertheless, analysis of such compounds can be utilized as a first approach to recognize semantic relations between two words. As an example, *Fachsprache* (terminology) and *Sprache* (language) can be considered. Obviously, *Sprache* is a suffix and *hypernym*[5] of *Fachsprache*.

A suffix matching based on regular expressions is used to recognize such hypernym relations between different tags. Analogously, semantic specification relations, i.e. *Fach* (subject) specifies *Fachtext* (specialized text) can be recognized by such an approach. In this case prefix matching instead of suffix matching has to be used. Nevertheless, this simple matching approach results in so many wrong matches, that it is not applicable.

Example. Given the two tags *Text* and *Kontext* (context). The simple approach will recognize a hypernym relation between these two tags although this is an incorrect matching since no hypernym relation exists in this case.

As a first solution a minimal string length δ of the string before the suffix is demanded. Recalling the example no semantic relation between *Text*, *Kontext* or other similar tags will be detected for $\delta > 3$. Since, $\delta > 4$ proved not to be useful, $\delta = 4$ is assumed in the following. Nevertheless, some correct matchings cannot be recognized due to this restriction, but estimations show that the number of not recognized existing relations is rather small.

As a result the regular expression for the suffix matching for a given alphabet Σ and a suffix s_1 can be denoted as follows:

$$r_s = (((\Sigma+)-)+\Sigma^4\Sigma^*)+(\Sigma+\ \Sigma+))\,s_1\,(\varepsilon|en|e)$$

If $w_1 \in L(r_s)$ for a given tag s_1 then a hypernym relation between s_1 and w_1 exists. Note that this regular expression has already been adapted to several special cases that occured in the given text-corpus. Furthermore, possible inflective suffixes have been removed from s_1 in prior to match as much inflective forms of the word w_1 as possible. Since this is an exclusively rule-based approach, semantic relations cannot be recognized for morphological special cases. If such words should also be recognized a reliable lemmatizer like the learning lemmatizer described above is needed.

In contrast to this suffix matching, the identification of specification relations is much more complex. This mainly results from two facts:

1. inflectional suffixes
2. erroneous splitting of compounds

In german verbs, nouns and adjectives can be used to create compounds or specify other words. Often inflectional suffixes are removed in those cases. As an example *lesen* (read) and *Lesestrategie* (reading strategy) can be considered. Here, the suffix *n* is removed before composing the two words *lesen* and *Strategie* (strategy). Obviously, most of those semantic relations should be recognized by the tagnet builder.

Additionally, specification relations may be recognized wrongly because of equivocal compounds that are mainly determined by the context. One example is *Texterkennung* that can be decomposed to *Text-erkennung* (text recognition) or *Texter-kennung* (writer identification). In this case only the specification relation between *Text* and *Erkennung* is correct, but also the relation between *Texter* and *Kennung* will be recognized. Unfortunately, this problem can solely be solved by manually defined blacklists. Nevertheless, a filter can be defined that filters wrongly detected relations, i.e. between *Text* and *textualisieren* (textualize). This filter checks if for a candidate match the remaining suffix is a known tag or word from a given lexicons. Here, *ualisieren* is not contained in the given lexicon and the relation will be discarded correctly.

Finally, the regular expression can be denoted as follows, where again inflective suffixes are removed from s_1 and the expression is adapted to some special cases:

$$r_p = (\Sigma^4\Sigma^*-)^*s_1\,(ung|en|e|n|\varepsilon)\,(s|-|\varepsilon)\,\Sigma^3\Sigma^*$$

If $w_1 \in L(r_p)$ for a given tag s_1, than a candidate specification relation between s_1 and w_1 exists that has to be filtered using the described filter afterwards.

3.1.2 Lexicon-based Approach

The rule-base approach is solely capable to recognize hypernym and specification relations. Nevertheless, semantic relations of other types should be identified as well by the tagnet builder, i.e. *schreiben* (write) and *lesen* (read). For this reason a lexicon-based approach is presented, where *Wikipedia*[6] is used as the test lexicon. It is assumed that the given lexicon entries have been normalized by the methods of stop word elimination and stemming described above.

The idea of the lexicon-based approach is to extract lexicon entries for each tag of the given set and perform a frequency analysis for all other tags on

[5] or generic term

[6] http://de.wikipedia.org/

these entries. Again, suffixes and prefixes have to be considered during this analysis, but the general approach will be described first.

The main aspect of the lexicon-based approach is to determine the strength of a semantic relation between tags. This significance is estimated based on relative frequencies of tags in the given lexicon entries. The relative frequency γ'_{s_1,s_2} of a tag s_2 in the lexicon entry l_{s_1} for a given tag s_1 is defined as,

$$\gamma'_{s_1,s_2} = \frac{n'_{s_2}}{N} \tag{1}$$

where n'_{s_2} denotes the absolute frequency of the current tag s_2 in the lexicon entry. The relative frequency is then derived by dividing n'_{s_2} by the total number of words N in l_{s_1}. In this simple approach n'_{s_2} is calculated as the frequency of all direct occurrences of s_2 in l_{s_1}.

A further improvement can be achieved by splitting the lexicon entries in headlines and text blocks and giving occurrences in the headlines a higher significance than the ones in the text blocks. This is done by a weight factor $\delta_1 \in [0,1]$. The resulting enhanced relative frequency γ_{s_1,s_2} is denoted as,

$$\gamma_{s_1,s_2} = \delta_1 * \frac{n_{t,s_2}}{N_t} + \frac{n_{h,s_2}}{N_h} \tag{2}$$

where n_{t,s_2} is the absolute frequency of s_2 in the text blocks of l_{s_1}. Analogously, n_{h,s_2} is the absolute frequency of s_2 in the headlines, N_t denotes the number of words in the text blocks, and N_h the number of words in the headlines.

Much more information can be gained by considering occurrences of s_2 in l_{s_1} as prefixe or suffix as well. But occurrences as prefix generally not as relevant as direct or suffix occurrences, since prefixes normally just specify other words in the german language. For this reason the enhanced approach is extended by a weight function which rates the different occurrences in the text blocks by additional parameters δ_2 and δ_3. This distinction is only applied to text blocks in the tagnet builder.

Thus, the enhanced relative frequency within the text blocks n_{t,s_2} is calculated by,

$$n_{t,s_2} = n^{direct}_{t,s_2} + \delta_2 * n^{prefix}_{t,s_2} + \delta_3 * n^{suffix}_{t,s_2} \tag{3}$$

where δ_2 and δ_3 are weight parameters for the absolute frequencies as prefix and suffix, respectively. This allows a much better estimation of semantic relations between tags than the simple approach, since language specific characteristics are taken into account.

3.1.3 Statistical Estimation

In contrast to the approaches described above it is also possible to estimate semantic relations between tags by frequency analysis on texts of a given corpus. Unfortunately, a large text corpus is needed for such statistical approaches in general. Hence, in Section 4.3 it was evaluated how the approach described below works on a corpus of only 200 texts.

Our approach is composed of two estimation steps.

1. estimation of the absolute tag frequencies in the texts
2. estimation of possible semantic relations based on these absolute frequencies

For step (1) basically the approaches described above can be used, so that this step will be skipped here. In the following it is assumed that for the given set of tags S and all texts t_i ($i \in \mathbb{N}$)

$$n_{t_i}(s_j) \forall s_j \in S$$

denotes the absolute tag frequency for tag s_j in the text t_i with $j \in \{1 \dots |S|\}$. Based on these absolute values it is possible to calculate relative tag frequencies for all tags s_j in the text t_i by

$$n'_{t_i}(s_j) = \frac{n_{t_i}(s_j)}{N_{t_i}} \tag{4}$$

where $N_{t_i} = max(n_{t_i}(s_1), \dots, n_{t_i}(s_{|S|}))$. These are used to statistically estimate the significance of possibly existing semantic relations between single tags.

Estimation of Possible Semantic Relations. For all texts t_i and a pair of tags (s_1, s_2) with $s_1, s_2 \in S$ it is checked, if $n_t(s_1) > 0$ and $n_t(s_2) > 0$ holds . The set of all such texts is denoted as T. The strength of the statistical relations between these two tags δ_{s_1,s_2} is calculated recursively for all texts $t_i \in T$ ($i \in \{1 \dots |T|\}$) by

$$\delta^{(k)}_{s_1,s_2} = \frac{\delta^{(k-1)}_{s_1,s_2} * k + n'_{t_k}(s_1) * n'_{t_k}(s_2)}{k+1} \tag{5}$$

where $\delta^{(0)}_{s_1,s_2} = 0$, and $0 \leq k \leq |T|$. The strength of a possible semantic relation between to tags s_1 and s_2 is than given by $\delta_{s_1,s_2} = \delta^{(|T|)}_{s_1,s_2}$.

This results in an histogram of all calculated pairwise statistical relations.

3.1.4 Result Combination

Finally, the tagnet builder combines the three resulting relation sets into one set containing the most relevant relations of all three approaches. Since the quality of this final set mainly depends on the quality of

all three sets, it is once again necessary to introduce weight factors that allow to control the influence of a base set on the final set. In case of relation duplicates occuring in two different base sets the average of both weighted qualities is calculated and stored as the new relation quality in the final set.

3.2 Similarity Analysis

In contrast to other approaches which estimate text similarities (Lee et al., 2005) our approach is based on an uncertainty relation on tagsets S_1 and S_2 of two texts t_1 and t_2 This solves the problem of disjunct tagsets of similar texts and allows to estimate similarity values in such cases. To realize the main objective of our approach the simple estimation is discussed first.

$$\delta'_{t_1,t_2} = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|} \tag{6}$$

Obviously, no similarity can be calculated for disjunct tagsets.

As a solution the tagsets S_1 and S_2 can be extended by information gained from the generated semantic tagnet. These new sets S'_1 and S'_2 are generated recursively for a fixed number of recursion steps $n_{max} \in \mathbb{N}$. In each recursion step the neighbours of each tag in S_i ($i \in \{1,2\}$) are added to the sets S'_i. In our findings it proved that $n_{max} > 4$ is not useful. Thus, $n_{max} < 4$ is assumed in the following.

Although this idea already allows to calculate similarities between two texts t_1 and t_2 with $S'_1 \cap S'_2 = \emptyset$ it not yet satisfying, since no distinctions are made between the tags in the two sets S'_1 and S'_2. For this reason tag significances $p_{i,j}$ were introduced which depend on the current recursion step n and manually defined significances γ_r for all considered relation types r. The similarity of two texts t_1 and t_2 then can be estimated as

$$\delta_{t_1,t_2} = \frac{1}{2M} * \sum_{i=1}^{N} (p_{1,i} + p_{2,i}) \tag{7}$$

where $N = |S'_1 \cap S'_2|$ and $M = |S'_1 \cup S'_2|$. For a given threshold τ and a text t_1 all texts with $\delta t_1, t_i >= \tau$ ($i \in \{1, \cdots, |T|\}, t_i \neq t_1$) can be queried.

4 EVALUATION

The described system has been implemented in Java and evaluated on four different tagsets.

ipTS: This set contains 3087 tags that mainly are part of the topic of text production and writing research which are part of the ipTS[7] research project. Note that this tagset has been preprocessed in terms of removing wrong tags and adjusting flectional suffixes. Thus, the results were somewhat better than the results of the three other tagsets.

II: This set contains 2748 tags from the topic of informatics[8].

IDS: This set contains the 1769 most basic german words[9].

AMSÖ: This set contains more than 10 000 occupational qualifications of different topics and thus is the most comprehensive tagset in this evaluation[10].

The evaluation was made by applying the implemented system to these four tagsets. Subsequently the generated semantic tagnets were checked for wrongly recognized semantic relations. Since manual preprocessing is unwanted, the ipTS tagset was the only preprocessed set in this evaluation.

Fig. 1 depicts a part of a semantic tagnet visualization (Götten, 2009) which was generated by the tagnet builder from the ipTS tagset. The semantic tagnet was embedded into the ipTS[11] project website to simplify the literature research task for domain experts. The arrows represent hyperonym relations between keywords, while the undirected edges express generic semantic relations.

4.1 Rule-based Approach

In the first evaluation step exclusively the rule-based approach was considered. Table 1 contains the evaluation results of the rule-based approach on all four tagsets. Obviously, the rule-based approach works quite well for all given tagsets, since the accuracy is smaller than 1% for most of the sets.

Table 1: Semantic relations extracted by rule-based approach.

Tagset	Entries	Relations	Error Rate
ipTS	3807	1287	0,8 %
II	2748	1656	0,6 %
IDS	1769	502	0,4 %
AMSÖ	10 245	1760	1,5 %

[7] http://www.ipts.rwth-aachen.de/
[8] http://is.uni-sb.de/vibi/
[9] http://www.ids-mannheim.de/oea/
[10] http://www.ams.or.at/bis/
[11] http://www.ipts.rwth-aachen.de/

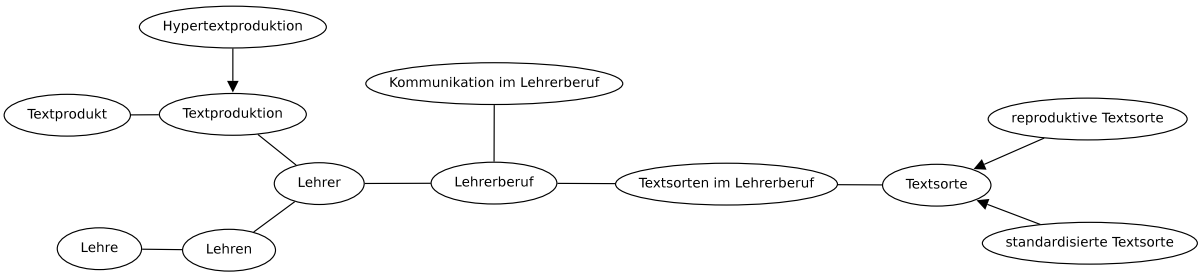

Figure 1: Part of the extracted semantic net.

Only for the AMSÖ tagset 1,5% semantic relations were recognized wrongly, due to the tags of different topics. Note that this accuracy does not consider existing semantic relations between tags that are not recognized. However, on average semantic relations between 33% of the given tags were recognized. This is shown in Table 2.

In the first evaluation step exclusively the rule-based approach is considered. Table 1 contains the evaluation results of the rule-based approach on all four tagsets. Obviously, the rule-based approach works quite well for all given tagsets, since the accuracy is smaller than 1% for most of the sets.

Table 2: Number of linked tags.

Tagset	Linked tags
ipTS	30,4 % (1160)
II	51,6 % (1419)
IDS	29,5 % (522)
AMSÖ	20,7 % (2122)

The results illustrate the benefit of such a pattern-based approach which estimates the possible semantic relations between the tags in different sets in short time. As an example the semantic relations for the ipTS tagset were estimated in less than 20 seconds on the testing machine, while the estimation on the AMSÖ set only took 88 seconds. For a synthetic generated tagset the rule-based approach shows an almost linear growth in runtime for large sets with more than 20 000 entries

4.2 Lexicon-based Approach

In contrast to this fast rule-based approach the lexicon-based approach needs much more calculation time (more than four hours). For all smaller sets the calculation takes less than one hour. Although the calculation time is much longer than in the rule-based approach, the lexicon-based approach allows to recognize semantic relations between tags that cannot be recognized by the plain rule-based approach as described in Section 3.1.1.

Table 3 shows the numbers of extracted relations for the given tagsets. In comparison to the rule-based approach many more relations are extracted especially for the AMSÖ tagset.

Table 3: Semantic relations extracted by lexicon-based approach.

Tagset	Entries	Relations	Error rate
ipTS	3807	2853	2,8 %
II	2748	3337	5,1 %
IDS	1769	1600	4,7 %
AMSÖ	10 245	5494	3,4 %

This mainly results from the type of the contained tags in this set since there is a wide range of words or concepts that describe similar processes or are somehow related to each other. Consequently, the total number of alike tags is higher for this lexicon-based approach compared to the results of the rule-based approach as shown in Table 4.

Table 4: Number of linked tags.

Tagset	Linked tags
ipTS	31,5 % (1198)
II	50,0 % (1376)
IDS	44,7 % (791)
AMSÖ	32,1 % (3286)

4.3 Statistical Approach

The statistical approach has solely been evaluated for the ipTS set due to the lack of a suitable text corpus for the other three tagsets. As expected a very large number of candidate semantic relations were extracted, but mostly with very small semantic strengths. For a proper threshold only 3455 of more than 400 000 candidate relations remained. Unfortunately the number of wrongly recognized semantic relations still were quite high (12%). However it was possible to reduce the error rate to 4,7% by blacklisting problematic tags. It is assumed that the result of

the described statistical approach can be improved by a larger text corpus.

5 CONCLUSIONS

An approach has been described that allows to efficiently generate semantic tagnets for given unstructured lists of tags and an arbitrary text corpus. This semantic tagnet can be used to estimate text similarities for tagged texts in digital libraries to provide a more intuitive way of literature research adapted to the user's cognitive model. In addition to this the generated semantic tagnet could be used to define ontologies or allow users to enhance the network using a suitable interface similar to the idea of *user feedback* as proposed in (Doan and McCann, 2003).

In a future version the tagnet builder may be enhanced by some components that allow to extract synonym relations, too. One idea of a rule-based approach has been proposed in (Ananthanarayanan et al., 2008) for english words. This enhancement would allow to store relations between tags in different languages to create a multilingual semantic net that can be used for a digital library that stores texts in different languages.

ACKNOWLEDGEMENTS

This research was funded in part by the DFG Cluster of Excellence on Ultra-high Speed Information and Communication (UMIC), German Research Foundation grant DFG EXC 89, and by the German Research Foundation grant for the Project *Interdisciplinary Text Production and Writing (ipTS*[12]).

REFERENCES

Ananthanarayanan, R., Chenthamarakshan, V., Deshpande, P. M., and Krishnapuram, R. (2008). Rule based synonyms for entity extraction from noisy text. In *AND '08: Proceedings of the second workshop on Analytics for noisy unstructured text data*, pages 31–38, New York, NY, USA. ACM.

Barnett, B. (2009). Regular expressions, http://www.grymoire.com/unix/regular.html.

Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The semantic web - a new form of web content that is meaningful to computers will unleash a revolution of new possibilities.

[12]http://www.ipts.rwth-aachen.de/

Chaffin, R. (1992). The concept of a semantic relation. In A. Lehrer, E. K., editor, *Frames, Fields and Contrasts*, pages 253–288. Lawrence Erlbaum, Hillsdale, N.J.

Collins, A. and Quillian, M. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, 8(2):240–247.

Doan, A. and McCann, R. (2003). Building data integration systems: A mass collaboration approach. In *IIWeb*, pages 183–188.

Fellbaum, C., editor (1998). *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.

Gaizauskas, R. and Humphreys, K. (1997). Using a semantic network for information extraction. *Nat. Lang. Eng.*, 3(2):147–169.

Götten, D. (2009). Semantische Schlagwortnetze zur effizienten Literaturrecherche. Master's thesis, RWTH Aachen University.

Harris, Z. (1985). In Katz, J. J., editor, *The Philosophy of linguistics*, pages 26–47. Oxford University Press.

Harrison, M. A. (1978). *Introduction to Formal Language Theory*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Lee, M. D., Pincombe, B., and Welsh, M. (2005). An empirical evaluation of models of text document similarity. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, pages 1254–1259, Mahwah, NJ. Erlbaum.

Löbner, S. (2003). *Semantik. Eine Einführung*.

Lovins, J. B. (1968). Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, 11:22–31.

Perera, P. and Witte, R. (2005). A self-learning context-aware lemmatizer for german. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 636–643, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14(3):130–137.

Porter, M. F. (2009). German stemming algorithm.

Quillian, M. R. (1967). Word concepts: A theory and simulation of some basic semantic capabilities. *Behavioral Science*, 12:410–430.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*.

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to german. In *In Proceedings of the ACL SIGDAT-Workshop*, pages 47–50.

Sowa, J., editor (1991). *Principles of Semantic Networks: Explorations in the Representation of Knowledge (Morgan Kaufmann Series in Representation and Reasoning)*. Morgan Kaufmann Pub.

Sowa, J. (2009). Semantic networks, http://www.jfsowa.com/pubs/semnet.htm.