Ranking and Selecting Synsets by Domain Relevance

Paul Buitelaar, Bogdan Sacaleanu
DFKI GmbH
Stuhlsatzenhausweg 3
D-66123 Saarbruecken, Germany
{paulb,bogdan}@dfki.de

Abstract

The paper presents a novel method for domain specific sense assignment. The method determines the domain specific relevance of GermaNet synsets on the basis of the relevance of their constituent terms that co-occur within representative domain corpora. The approach is task independent and completely automatic. Experiments show results on three selected domains: *business*, *soccer* and *medical*.

1 Introduction

This paper presents a novel method for domain specific sense assignment using WordNet (Miller et al. 1995) and similar resources like GermaNet -- a lexical semantic resource for German (Hamp and Feldweg, 1997) with a structure similar to that of WordNet. Within the context of a restricted domain, many ambiguous terms (i.e. nouns) may have a strong preference for one of their senses. A system for automatically determining this most likely sense is therefore a useful tool in semantic processing like semantic tagging for information extraction or machine translation.

The method we describe determines the domain specific relevance of GermaNet synsets on the basis of the relevance of their constituent terms that co-occur within representative domain corpora. Although some research has been done in the direction of domain specific tuning of lexical semantic resources (see e.g. deliverable D3.1.1 of the ECRAN project[1] and Turcato et al. 2000), using term relevance for this is to our knowledge a novel approach. Also, we believe our approach to be more general, because it is task independent and completely automatic.

The particular research described here is part of a larger effort to develop semi-automatic methods for domain specific semantic lexicon construction that builds on the reuse of existing resources. Adapting these resources to a specific domain includes selecting those terms and meanings that are relevant for the domain and adding new terms and meanings that are missing from the existing resource. The research described here deals primarily with the former aspect. Adding new terms and meanings through term classification and clustering is discussed in more detail in (Sacaleanu, 2001).

The experiments described here show results on three selected domains: *business* (newspaper reports on economic and financial policy), *soccer* (newspaper reports and live tickers), *medical* (abstracts of articles on medical research).

2 System Description

The system starts with extracting terms (i.e. nouns) from the given domain corpora. It then proceeds by computing the domain relevance of each term and using this information to compute the cumulative relevance of each of the synsets in which the terms occur relative to a particular domain. This allows for a ranking of synsets according to domain relevance.

2.1 Preprocessing

The system first annotates all words with part-of-speech and morphological information. This allows us to extract noun stems that are used as terms in further processing (to look up the lemma in GermaNet). We used the TnT package (Brants, 2000) for part-of-speech tagging, trained on a general corpus in order to make it unbiased to any of the domain specific corpora used in the experiments.

For morphological analysis we used the MMORPH package (Petitpierre and Russel, 1995). An important aspect of this, specifically for a highly compositional language like German, is compound analysis. For example, in the medical domain there are many compounds involving the noun `Patient` (patient):

`Poliklinik-Patienten`	(polyclinic -)
`Notfallpatient`	(emergency -)
`Placebopatienten`	(placebo -)

All of these need to be taken into account in order to reflect the frequency of `Patient` in the medical

[1] http://www.dcs.shef.ac.uk/research/ilash/Ecran/

domain. Therefore, in counting term frequency for a term, we summed up the number of times it appeared in the domain corpus as a term by itself as well as its morphological derivations.

2.2 Term Relevance

The next step is to compute an index that defines the relevance of each term to a specific domain. We adopt this approach from information retrieval, in which a relevance measure is computed between each term and each document of a retrieval corpus. In this way, documents can be assigned to a query term that is known to be relevant to them. Our approach builds on this idea by computing the relevance of each term to each domain-specific corpus.

The relevance measure we use is a slightly adapted version of standard *tf.idf*, as used in vector-space models for information retrieval (Salton and Buckley, 1988):

$$rlv(t \mid d) = \log(tf_{t,d})\log(\frac{N}{df_t})$$

where *t* represents the term, *d* the domain, *N* is the total number of domains (here *N=3*). This formula gives full weight to words that occur in just one domain and a weight of zero to those occurring in all domains.

2.3 Concept Relevance

Given term relevance, the next step is to determine the relevance of each concept (a GermaNet synset[2]) in which the relevant terms occur. The most intuitive way for this is to sum up the relevance of each term in the synset as reflected by the following definition for concept relevance:

$$rlv(c \mid d) = \sum_{t \in c} rlv(t \mid d)$$

2.3.1 Lexical Coverage

Now, suppose we want to estimate concept relevance for the medical domain. Consider for example the term `Zelle` which occurs in the following two synsets:

[*Zelle, Gefängniszelle*]	("prison cell")
[*Zelle*]	("living cell")

Although `Zelle` will have a high relevance in the medical domain, the occurrence of `Gefängniszelle` in this domain is very unlikely and therefore the relevance value of both concepts will be equal. Although the latter concept is more relevant to the medical domain, we would not be able to automatically determine this by merely adding up the relevance of the terms in each of the synsets. Therefore we reconsidered the concept relevance definition to take into account so-called *lexical coverage*, that is the number of terms in the synset that actually occur in the domain:

$$rlv(c \mid d) = \sum_{t \in c} \frac{T}{|c|} rlv(t \mid d)$$

where *T* represents the lexical coverage, and |*c*| is the length of concept *c*. This relevance measure reflects the intuition that if many terms in the synset occur in the domain, then the more likely it is that the synset is relevant for that domain.

2.3.2 Hyponyms

However, the lexical coverage measure has two drawbacks. First, it has a preference for synsets with only one element, as lexical coverage (*T*) is always maximal, relative to the length of *c*. In our example this means that the medical sense of `Zelle` will always be favored, even in non-medical domains, unless the "prison cell" sense is clearly represented in the domain through the co-occurrence of `Gefängniszelle`. Secondly, the measure assigns equal values to synsets of equal length if only one of their terms occurs in the domain. For example consider the two senses of `Geschlecht`:

[*Geschlecht, Haus*]	("family line")
[*Geschlecht, Sexus*]	("gender")

When neither `Haus` nor `Sexus` occur in the domain, both concepts will get equally weighted. One way to overcome these problems is to add more lexical information. GermaNet encodes two relations that are suitable for this purpose: hypernymy and hyponymy. Hypernymy is frequently associated with *concept generalization* and hyponymy with *concept specialization*. As for our purposes concept generalization would endanger the domain-specificity of our methods, we decided to draw upon hyponymy by attaching to each synset all of its direct hyponyms. The senses of `Zelle` can therefore be supplemented with further lexical information:

[*Zelle,Gefängniszelle,**Todeszelle***]
[*Zelle,**Körperzelle**,**Pflanzenzelle***]

Unfortunately not all concepts in GermaNet have direct hyponyms, as with one of the senses of `Geschlecht`:

[*Geschlecht,Haus,**Adel**,
Adelgeschlecht,**Fürstenhaus**,
Herrscherhaus, **Königshaus***]
[*Geschlecht, Sexus*]

[2] We will interchangeably use s*ynset* and *concept* to refer to a list of synonyms as defined by GermaNet.

Adding hyponyms slightly changes the concept relevance formula. In addition to summing up the relevance of each term in the synset, now also each direct hyponym adds to its relevance:

$$rlv(c+ \mid d) = \sum_{t \in c+} \frac{T}{|c|} rlv(t \mid d)$$

where c+ is the extended concept. Note that in the definition *T* (number of terms in the concept that occur in the domain) and |*c*| (number of terms in the concept) have not changed. That is, hyponyms do not affect lexical coverage, but only add to the summed weight of the concept.

Nevertheless, we would like to express a notion of lexical coverage also for the hyponyms that extend a concept. For this purpose we introduced a penalty on missing hyponyms. That is, if a hyponym does not occur in the domain it is assigned a constant negative weight. This value is dynamically determined by taking care that a concept relevance should never become smaller than zero. Therefore, for a concept X: [$x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_m$], where x_i are the constituent terms and y_i are its direct hyponyms, it holds:

$$\sum_{i=1}^{n} rlv(x_i \mid d) - \sum_{j=1}^{m} rlv(y_j \mid d) > 0 \cdot$$

In the worst case (*n*=1, *m*=30[3]) with all hyponyms missing, we have:

$$x_1 > 30 * k \Leftrightarrow k < x_1 / 30$$

That is, the value of the constant negative weight *k* we are looking for, should be smaller than the 30th of any term domain relevance.

3 Experiments

In order to assess the correctness of our methods we set up an experiment that would show two things: 1. How well our method selects domain specific concepts, and 2. How well the most specific sense (synset) is selected for domain specific terms.

3.1 Domain Specific Concepts

For the first experiment we selected for each domain the top 100 concepts from a ranked list as computed by the system and presented them to two judges for inspection. The judges were asked to rate each concept on a scale from 1 to 3 indicating how relevant it is for the domain. To avoid any subjectivity in this task, the judges were presented with guidelines to establish uniform criteria. The instructions that were given to the judges are as follows. On a scale from 1 to 3, rate each concept's relevance to the given domain.

- **3: Relevant** The concept belongs to the domain.
  Example: [Spiel (game), Match, ...]
  ➔ Relevant in Soccer

- **2: Associated** The concept belongs to the domain and to one or more of the other domains.
  Example: [Gegner (rival), Opponent, ...]
  ➔ Relevant in Soccer and in Financial

- **1: Not Relevant** The concept does not belong to the domain.
  Example: [Szene (scene), Sketch]
  ➔ Not Relevant in Soccer

As the relevance measures used are relative to the domains at hand, the judges were requested to constrain their decisions also to the three given domains. That is, in judging concept relevance, they had to consider the domains as a closed world: a concept can be assigned only to one or more of the given domains. Therefore, concepts like:

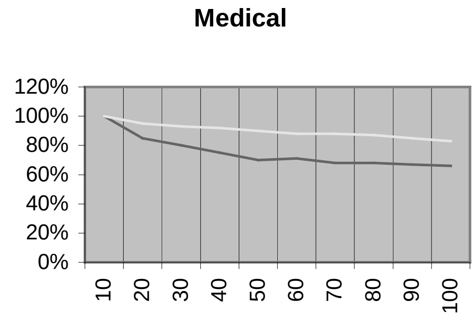
[*Endspiel, Finale*]	(finals)
[*Viertelfinale*]	(quarter finals)

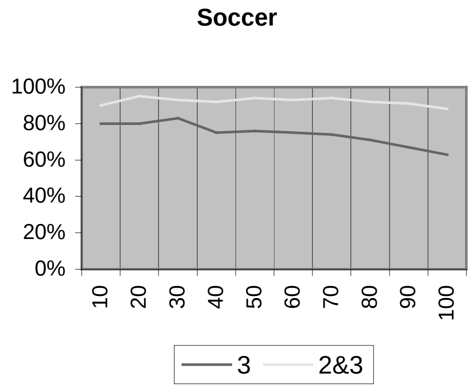
were assigned as strongly relevant for the soccer domain, although they could also occur in other sports related domains, like baseball or volleyball.

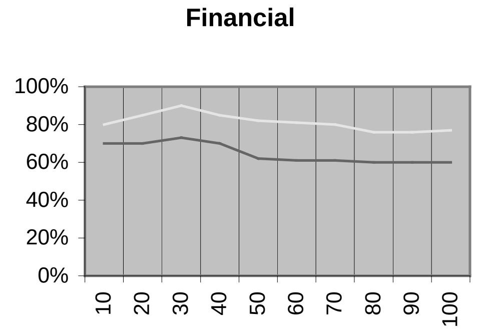
The results (i.e. the rate of concepts in the top-concept lists that were judged relevant and/or associated) were computed for category 3 (relevant concepts), as well as categories 2&3 (relevant + associated concepts). The plotting distance is 10 words, stepping down the ranked list of concepts. The purpose of our experiment was to evaluate consistency in suggesting relevant concepts for all three domains. The results are shown in the graphs below.

The outcome for the concepts that were judged to be relevant (category 3) was about 65% for the medical and soccer domains, and about 60% for the financial domain. If we look only at the first 20 relevant concepts (category 3), best results were in the medical domain at 85% (even 100% for the first 12 concepts), for soccer at 80% and for financial at 70%. Considering all of the concepts that are either relevant or associated to the domain (categories 2&3), we achieved an overall score of 80% to 90% for all three domains.

[3] Because very general concepts with a large number of hyponyms are well covered by any domain (many of the terms occur in the domain) and therefore consistently assigned a high relevance, we removed any concepts with more than 30 direct hyponyms.







3.2 Domain Specific Sense

A second experiment was set up to evaluate the effectiveness of the concept relevance measure in assigning the most likely domain specific sense. Out of the first 100 concepts previously evaluated to be relevant (category 3), we extracted all constituent terms that actually occur in the domain. From among

	Ranked Terms -- with English translation(s)	Ranked Concepts
yes	Eingriff (operation, intervention)	1. [Eingriff:c, Operation:c, Abtreibung, Biopsie, ...] 2. [Eingreifen:c, Eingriff:c, Intervention:c]
all	Infektion (infection)	1. [Entzündung:c, Infektion:c, Infektionskrankheit:c, ...] 2. [Ansteckung:c, Infektion:c, Übertragung:c]
all	Studie (study, report)	1. [Experiment:c, Studie:c, Test:c, Versuch:c,...] 2. [Abhandlung:c, Studie:c]
all	Prophylaxe (prophylaxis)	1. [Prophylaxe:c, Empfängnisverhütung, Impfung, Verhütung] 2. [Prophylaxe:c, Vorbeugung:c, Vorsorge:c, ...]
yes	Gewebe (tissue)	1. [Gewebe:c, Körpergewebe:c, Bindegewebe, Tumor, ...] 2. [Gewebe:c, Kleiderstoff:c, Stoff:c, Textilstoff:c, ...]
all	Medizin (medicin)	1. [Medizin:c, Chirurgie, Frauenheilkunde, Gynäkologie, ...] 2. [Arznei:c, Arzneimittel:c, Heilmittel:c, Medikament:c, ...]
yes	Gefäß (vascular, container)	1. [Gefäß:c, Blutgefäß, Haargefäß, Herzkranzgefäß, Lymphgefäß] 2. [Gefäß:c, Container, Form, Pokal, Schale, Schüssel, Tonne, ...]
yes	Zelle (cell)	1. [Zelle:c, Körperzelle, Pflanzenzelle] 2. [Gefängniszelle:c, Zelle:c, Todeszelle]
all	Einschränkung (constraint, restriction)	1. [Beschränkung:c, Einschränkung:c, Vorbehalt:c] 2. [Beschränkung:c, Degression:c, Drosselung:c, Einschränkung:c]
all	Aufnahme (intake, reception)	1. [Aufnahme:c, Aufzeichnung:c, Mitschnitt:c, Protokoll, ...] 2. [Aufnahme:c, Beherbergung:c, Unterbringung:c, Notaufnahme, ...]
yes	Sektion (section)	1. [Autopsie:c, Leichenöffnung:c, Obduktion:c, Sektion:c] 2. [Amtsbereich:c, Dezernat:c, Geschäftsbereich:c, Sektion:c, ...]
all	Ausdehnung (spread, dimensions)	1. [Ausdehnung:c, Rauminhalt:c, Volumen:c] 2. [Ausdehnung:c, Ausweitung:c, Dehnung:c, Erweiterung:c, ...]
yes	Geburt (birth, rebirth)	1. [Geburt:c, Fehlgeburt, Frühgeburt] 2. [Geburt:c, Wiedergeburt]
yes	Abweichung (abnormality, divergence)	1. [Abweichung:c, Differenz:c, Abnormität, Anomalie, ...] 2. [Abweichung:c, Differenz:c, Meinungsverschiedenheit]
yes	Probe (test, rehearsal)	1. [Probe:c, Blutprobe, Gesteinsprobe, Urinprobe, Wasserprobe] 2. [Bühnenprobe:c, Probe:c, Chorprobe, Generalprobe]
all	Verhütung (prevention)	1. [Abwendung:c, Vereitelung:c, Verhinderung:c, Verhütung:c] 2. [Empfängnisverhütung:c, Verhütung:c]
all	Empfindung (feeling, sensation)	1. [Empfindung:c, Hören, Riechen, Schmecken, Sehen, Spüren, ...] 2. [Emotion:c, Empfindung:c, Gefühl:c, Gemütsbewegung:c, ...]
all	Beschränkung (constraint, restriction)	1. [Beschränkung:c, Einschränkung:c, Vorbehalt:c] 2. [Beschränkung:c, Degression:c, Drosselung:c, Einschränkung:c]
yes	Wirkstoff (active component)	1. [Wirkstoff:c, Hormon] 2. [Wirkstoff:c, Koffein, Teein]
yes	Verordnung (prescription, regulation)	1. [Medikation:c, Verordnung:c] 2. [Verfügung:c, Verordnung:c, Stadtverordnung]
yes	Krebs (cancer, crustacean)	1. [Krebs:c, Krebserkrankung:c, Krebsgeschwür:c, Blutkrebs, ...] 2. [Krebs:c, Krebstier:c, höherer_Krebs, niederer_Krebs, ...]
all	Ausnahme (exception)	1. [Ausnahme:c, Besonderheit:c, Irregularität:c, Sonderfall:c, Rarität, ...] 2. [Ausnahme:c, Ausnahmeerscheinung:c, Ausnahmefall:c, ...]
yes	Operation (operation, procedure)	1. [Eingriff:c, Operation:c, Abtreibung, Amputation, Autopsie, Biopsie, ...] 2. [Operation:c, Prozedur:c, Bearbeitung, Behandlung, Verarbeitung]
all	Besonderheit (peculiarity, anomaly)	1. [Ausnahme:c, Besonderheit:c, Irregularität:c, Sonderfall:c, Rarität, ...] 2. [Außergewöhnlichkeit:c, Besondere:c, Besonderheit:c]

these we then selected the ambiguous ones and ranked for each of them the concepts (all concepts, not only the top 100 -- i.e. their senses) in which they occur by their domain relevance. This produced a list of 24 domain specific, ambiguous terms for the medical domain, 17 for the financial domain and 8 for the soccer domain. They were evaluated by annotating them with one of:

- *yes* (most likely domain specific sense was correctly predicted)
- *no* (most likely domain specific sense was not correctly predicted)
- *all* (all senses apply equally to the domain)

The table shows that out of 24 terms in the medical domain 12 have at least one sense that is specific for this domain. All of these were determined correctly by our automatic method. For the financial domain, 6 out of 17 terms had at least one domain specific sense, of which 5 were determined correctly. For the soccer domain, these figures were: 6 out of 8 terms of which 5 were determined correctly. These results indicate a consistently accurate prediction of domain specific senses by the automatic method described.

4 Discussion

Obviously, the results discussed above depend on many underlying factors. These include: 1. The adequacy of the domain corpora in representing each of the domains; 2. The accuracy of pre-processing (tokenizing, part-of-speech tagging, morphology); 3. The coverage of terms and concepts (senses) in the lexical semantic resource. Here we discuss each of these issues in more detail.

4.1 Domain Corpora

All domain corpora used in the experiments are manually constructed and may therefore be assumed to represent the domain in a reliable way.

The medical domain corpus is collected in the context of the MUCHMORE project on cross-lingual retrieval of medical information (Buitelaar, 2000). The corpus consists of abstracts of scientific articles in various areas of medical research as obtained from the Springer LINK website[4]. The soccer domain corpus is collected for the MUMIS project on retrieval of multi-media soccer documents (Declerck and Wittenburg 2001). The corpus consists of live tickers and newspaper reports of the Euro Cup 2000 and World Cup 1998 games. The corpus for the financial domain was collected for the PARADIME project (Declerck et al., 1998) and consists of articles from the business journal *Wirtschaftswoche*.

All three corpora consist of about 200,000 tokens. This number was determined by the soccer corpus because this one was the smallest. A comparison with larger corpora showed that the distribution of (relevant) terms remained rather stable between smaller and larger samples of the same corpora. Therefore we decided to cut off the other two corpora (medical, financial) to match the soccer corpus in size.

Manually constructed domain specific corpora are not always readily available for most domains. Therefore we performed some small experiments in the automatic construction of such corpora by identifying relevant sections in a general newspaper corpus (*Frankfurter Rundschau*) through a combination of domain specific keywords. Future work could improve on this by incrementally using the output of our system for identifying relevant terms that could then be used to extract more relevant sections.

4.2 POS Tagging, Morphology

As mentioned before, the part-of-speech tagger as used in the experiments was not specifically trained for each domain. This has the advantage that the tagger does not have any bias to any of the domains. At the same time however we may expect an improvement in tagging accuracy if the tagger would be trained on each of the domains, which will therefore be included in future work.

The main problem in morphology for German is compound analysis. The following example illustrates that also this level of pre-processing needs to be adapted to the domain in future work. The medical term `Oedem` (oedema, edema[5]) may be part of many compounds like `Hirnoedem` (brain oedema), or `Lungenoedem` (lung oedema). Therefore, a correct analysis of the term `Schleimhautoedem` would be

```
Schleimhaut - Oedem
```

Unfortunately, however, our morphological analysis tool proposes

```
Schleimhaut - Oe - dem
```

where `dem` is the abbreviation for "Deutsche Mark".

4.3 GermaNet Coverage

A final important influence on the results of our experiments is the coverage of the lexical semantic resource used, both of terms and of concepts. The coverage of terms by GermaNet for each domain can

[4] http://link.springer.de/

[5] "swelling from excessive accumulation of serous fluid in tissue"

be easily counted. The following table gives an overview of the number of different terms that were extracted for each domain and the number of these that are in GermaNet. These numbers include both simple terms (stems) and complex terms, because GermaNet includes both simple and complex terms as individual lemmas. For instance, in case of the complex term `Karzinompatient` both its stem, the simple term `Patient` and the complex term itself are listed as different terms.

	Medical	Financial	Soccer
Terms	18,204	15,161	10,332
Terms in GermaNet	1,429 (7.8%)	2,895 (19.1%)	2,056 (19.9%)

As the table shows, both the financial and soccer domains are covered at about 19%-20% by GermaNet. The medical domain, however, has only a coverage of about 8% which (not surprisingly) indicates that GermaNet is less tuned towards highly specific terminology.

It is hard to indicate which concepts, that is, which senses are missing from a resource. The number of concepts that need to be covered cannot readily be obtained from a (domain specific) corpus. Only relative to a certain task there could be an indication of which concepts are needed. This is for instance the case with a natural language interface to a database, where all possible objects in the database need to correspond to concepts in the lexical semantic resource that is used in the analysis of the natural language queries.

For our purposes we attempt to model sense assignment relative to a domain, but independent of a specific task. Therefore we can only indicate senses that are missing as they arise in evaluation of our results. For instance `Lappe` (lobe) occurs as the stem of many terms in the medical domain:

> `Okzipitallappe`
> (occipital lobe)
>
> `Prostatamittellappe`
> (prostate middle lobe)

In GermaNet, however, there is only one sense of `Lappe` represented, which is the non-medical one of: "Lapp, Lapplander".

5 Conclusions

We presented a novel method for domain specific sense assignment using GermaNet in combination with relevance measures as used in information retrieval. The method determines the domain specific relevance of GermaNet synsets on the basis of the relevance of their constituent terms that co-occur within representative domain corpora. Results show that this allows for an accurate prediction of domain specific senses.

6 Acknowledgements

Thanks to Detlef Prescher, Daniel Sonntag, Markus Becker and Oliver Scherf for evaluating the results. Thanks also to Detlef Prescher for useful comments on earlier versions of the paper.

This research has in part been supported by EC/NSF grant *IST-1999-11438* for the MUCHMORE project.

References

Brants, T. 2000. *TnT - A Statistical Part-of-Speech Tagger*. In: Proceedings of 6th ANLP Conference, Seattle, WA.

Buitelaar, P. 2000. *MUCHMORE: Multilingual Concept Hierarchies for Medical Information Organization and Retrieval*. In: Proceedings of ASIS 2000, Chicago, USA.

Declerck, Th. and Wittenburg, P. 2001. *MUMIS -- A Multimedia Indexing and Searching Environment*. In: Proceedings of the First International Workshop on MultiMedia Annotation, Tokyo, Japan.

Declerck, Th., Klein, J. and Neumann G. 1998. *Evaluation of the NLP Components of an Information Extraction System for German*. In: Proceedings of LREC, Granada, Spain.

Hamp, B. and Feldweg, H. 1997. *GermaNet: a Lexical-Semantic Net for German.* In: Proceedings of the ACL/EACL97 workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications, Madrid.

Miller, G.A. 1995. *WordNet: A Lexical Database for English*. Communications of the ACM 11.

Petitpierre, D. and Russell, G. 1995. *MMORPH - The Multext Morphology Program.* Multext deliverable report for the task 2.3.1, ISSCO, University of Geneva.

Sacaleanu B. 2001. *Domain specific Tuning and Extension of Lexical Semantic Resources*. Master Thesis, University of the Saarland, Germany.

Salton, G. and Buckley, C. 1988. *Term-Weighting Approaches In Automatic Text Retrieval*. In: Information Processing & Management. 24, 5, pp.515-523.

Turcato, D., Popowich F., Toole J., Fass D., Nicholson D., and Tisher G. 2000. *Adapting a synonym database to specific domains*. In: Proceedings of the ACL'2000 Workshop on Recent Advances in Natural Language Processing and Information Retrieval, Hong Kong.