Extending Synsets with Medical Terms

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

Abstract

An important problematic issue with general semantic lexicons like WordNet or GermaNet is that they do not cover many terms and concepts specific to certain domains. Therefore, these resources need to be tuned to a specific domain at hand. This involves selecting those senses that are most appropriate for the domain, as well as extending the sense inventory with novel terms and novel senses that are specific to the domain. In this paper we focus on extending GermaNet synsets with domain specific terms, taking into account the domain relevance of senses (i.e. synsets).

1 Introduction

Natural language applications, such as information extraction and machine translation, require a certain level of semantic analysis. An important part of this process is *semantic tagging*: the annotation of each content word with a semantic category. Semantic categories are assigned on the basis of a semantic lexicon like WordNet for English (Miller et al., 1995) or similar resources like GermaNet for German (Hamp and Feldweg, 1997).

A problematic issue, however, is that general semantic lexicons like WordNet or GermaNet do not cover many terms and concepts specific to certain domains. Therefore, these resources need to be tuned to a specific domain at hand. This involves selecting those senses that are most appropriate for the domain, as well as extending the sense inventory with novel terms and novel senses that are specific to the domain.

Some work in this area has been reported, with an emphasis on domain specific sense selection (Basili et al., 1997; Cucchiarelli and Velardi, 1998; Turcato et al., 2000). In (Buitelaar and Sacaleanu, 2001) a bottom up approach to sense selection was reported, which determines the domain specific relevance of (WordNet, GermaNet) synsets on the basis of the relevance of their constituent synonyms that co-occur within representative domain corpora.

In this paper we focus on extending GermaNet synsets with domain specific terms, taking into account the domain relevance of concepts (i.e. senses) as computed by the method described in (Buitelaar and Sacaleanu, 2001). We approach the extension task from two angles: through morphological analysis (decomposition) and through learning semantic similarity from co-occurrence patterns on domain specific corpora.

The system includes a linguistic preprocessing step in which all words are annotated with part-of-speech and morphological information. We used the TnT tagger (Brants, 2000) for part-of-speech tagging and the MMORPH package (Petitpierre and Russell, 1995) for morphological analysis.

The medical domain corpus used for the experiments reported here has been 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[1].

2 Extension by Decomposition

As German is a highly compositional language, morphological decomposition is the most intuitive way of acquiring novel terms from German domain specific corpora. Every compound is a specification of its head (i.e.

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

stem). Therefore, compounds can be easily added to GermaNet as hyponyms of this head word. For instance, some compounds with head *Therapie* (*therapy*) in the medical corpus are:

Antibiotikatherapie	*(anti-biotics--)*
Gentherapie	*(gene --)*
Lasertherapie	*(laser --)*
Sauerstofftherapie	*(oxygen --)*
Toxoplasmosetherapie	*(toxoplasmosis --)*

In order to limit this process to domain relevant terms and synsets, each term is assigned a term relevance relative to its occurrence in other domain corpora as described in (Buitelaar and Sacaleanu, 2001). The relevance measure 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. This formula gives full weight to words that occur in just one domain and a weight of zero to those occurring in all domains.

The term relevance of each term is used to compute a relevance measure also for each synset (i.e. sense) in which these terms occur as a synonym. According to this relevance measure, synsets are ranked and the top most synsets selected as domain relevant. The extension process described in this paper is restricted to these top most domain relevant synsets and top most novel terms.

2.1 Heads with One Sense

Adding compounds to GermaNet is straightforward, if the head word in question has only one sense. For instance, *Tumor (tumor)* has only the following sense:

#1 [Geschwulst, Geschwür, Tumor]
(blastoma, ulcer, tumor)

Hence, the following compounds with head *Tumor* can simply be added to GermaNet through the hyponymy relation:

Blasentumor	*(blatter --)*
Magentumor	*(stomach --)*
Schädelbasistumor	*(cranial base --)*
Talgdrüsentumor	*(sebaceous glands --)*
Wirbelsäulentumor	*(spinal --)*

2.2 Heads with More Senses

The acquisition process can be easily automated for those head words that have only one sense. More frequently, however, the head words have at least two senses. This introduces an ambiguity in adding compounds as hyponyms to one of the senses. We distinguish two cases: 1. only one sense is relevant to the domain; 2. two or more senses are equally relevant to the domain.

2.2.1 One Domain Specific Sense

The first case applies if only one sense of a given head word was determined to be domain relevant by the automatic method described above. It is then to be expected that all compounds of the head word also refer to this sense [2]. Take for example *Gewebe (tissue),* with the following two senses:

#1 [Gewebe, Körpergewebe]
(tissue, body tissue)

#2 [Gewebe, Stoff, Textilstoff]
(tissue, cloth, textile)

Since only the first sense applies to the medical domain, all compounds that were automatically extracted from the medical corpus can be acquired as hyponyms of sense #1:

Entzündungsgewebe	*(infection --)*
Gehirngewebe	*(brain --)*
Karzinomgewebe	*(carcinoma --)*
Pankreasgewebe	*(pancreas --)*
Schilddrüsengewebe	*(thyroid gland --)*

2.2.2 More than one Domain Specific Sense

Compounds of a head word may be added as hyponyms of either sense if two or more senses were determined to be equally relevant within the domain. Consider for instance the noun *Infektion*, which has the following two senses:

#1 [Entzündung,Infektion, Infektionskrankheit]
(infection,inflammation, infectious disease)

#2 [Ansteckung, Infektion, Übertragung]
(infection, transmission)

[2] Unfortunately, even if the head word has a clearly dominant sense within the domain some instances of other senses may occur as well -- i.e. *Polstergewebe (cushion/pad tissue)* in our medical corpus.

Some of the compounds extracted from the medical corpus with this noun as head are given below. The list contains hyponyms of both senses, with underlined terms corresponding to the second sense.

Blutstrominfektion	*(blood flow --)*
Erstinfektion	*(initial --)*
Hautinfektion	*(skin --)*
Krankenhausinfektion	*(hospital --)*
Luftweginfekion	*(airborne --)*

To add the right compounds as hyponyms to the right sense, some additional processing is needed. Clustering techniques could be used to automatically separate the compounds in several groups, each of which corresponding to a sense of the head word. In the current system, however, clustering has not yet been implemented. For now, a supervised process is assumed in which a domain expert decides which compounds are added as hyponyms of which sense.

3 Extension by Similarity

Much of the work on the acquisition of semantic classes has been based on statistics over co-occurrence of words within a fix window of text, where a window can be a number of words, a sentence, a paragraph, or even an entire document (e.g. Church and Hanks, 1990; Brown et al., 1992). The results of these approaches have shown that a simple frequency analysis of words co-occurring with other words may indicate classes of similar meanings.

Here we present an approach to semantic classification that uses patterns of lexico-syntactic context to discover semantic similarities between classes in GermaNet (i.e. synsets) and novel terms that are not currently in GermaNet.

The hypothesis on which this work is based is that words used in similar *syntactic* contexts and with a large overlap in *lexical* information will be semantically similar. In other words, we intend to classify words by means of their lexical contexts under consideration of syntactic constraints.

3.1 System Overview

The system assumes a set of novel terms and a set of domain specific synsets to which these terms will be assigned (classified). Both, novel terms and domain specific synsets, are selected using the methods discussed in (Buitelaar and Sacaleanu, 2001).

For each of the novel terms and for each of the synonym terms of the synsets, lexico-syntactic patterns are extracted from the corpus and a co-occurrence measure is computed on each of their instances.

Finally, an instance-based learning algorithm is used to generate a classifier for each of the patterns, which is used to automatically assign a novel term to one of the synsets.

3.2 Lexico-Syntactic Patterns

For each novel term and for each synonym term in a domain relevant synset a set of lexico-syntactic patterns within a window of *n* words[3] is extracted. Here, we only consider nouns, adjectives and verbs as relevant, with all other word classes being marked as irrelevant (“null”). For instance, the following pattern represents the context of a term (*T*) with two irrelevant words and an adjective on its left side, and an irrelevant word, followed by an adjective and a noun on the right:

[null, null, ADJ, T, null, ADJ, NN]

Then, for each pattern all corpus instances are extracted. In this way, each novel term and each synset is represented by a number of lexico-syntactic context instances that are used in classification. For each context instance, we compute a mutual information score for all co-occurrence pairs. A *co-occurrence pair*, written as (*x*, *y*), represents the co-occurrence of a *term*, *x*, with a *context word*, *y*, within a context instance. Let *N* be the total number of words that occur in all instances of a given pattern. Using the Maximum Likelihood Estimator (MLE), the probability of a pair, *P*(*x*, *y*), is estimated by its *relative frequency*:

$$P(x, y) \approx \frac{f(x, y)}{N}$$

where *f*(*x*, *y*) is the count of (*x*, *y*) pair over all instances of the pattern. Similarly, the

[3] In all experiments reported here n = 7.

probability for an occurrence of a word, *P(w)*, is estimated by:

$$P(w) \approx \frac{f(w)}{N}$$

The mutual information of a co-occurrence pair *MI*(*x*, *y*) is then estimated by:

$$MI(x,y) = \log_2 \frac{P(x,y)}{P(x)P(x)} \approx \log_2 \left(\frac{f(x,y)}{f(x)f(y)} N \right)$$

where *x* is a *term* and *y* a *context word*.

The frequency of synsets is defined by the sum of the frequencies of its component synonym terms. The co-occurrence frequency of a synset with a context word is then defined by the co-occurrence of the context word with the synonym terms. Thus, for a given synset *C*:

C: $[t_1, t_2, t_3, \ldots, t_n]$

and a context word *w*, the mutual information will be defined as follows:

$$MI(C,w) \approx \log_2 \left(\frac{\sum_{i=1}^{n} f(t_i,w)}{\sum_{i=1}^{n} f(t_i) f(w)} N \right)$$

where t_i are synonym terms of synset *C*.

To arrive at a mutual information score between synsets and their contexts as a whole, we decided to simply take the sum of the mutual information for all context words.

3.3 Instance-Base Learning

Deciding on similarity between terms and term classes (i.e. synsets) according to a shared context is a task well suited for machine learning. More specifically, we decided to use an instance-based -- *k-nearest neighbor* -- classifying algorithm that uses all of the context instances to assign the most similar class (synset) to a novel term (Witten, 2000).

3.3.1 Data Models

For each of the syntactic patterns the learning system creates a data model that consists of all context instances for each synset. For instance, for a lexico-syntactic pattern of the form:

[NN, null, ADJ, (T|C), null, NN, null]

context instances are represented for synset *C* in the following format[4]:

C, MI, $Noun_i$*,* $Adjective_j$*,* $Noun_k$

Similarly, a data model of context instances is created for each novel term *T* in the following format:

T, MI, $Noun_l$*,* $Adjective_m$*,* $Noun_n$

3.3.2 Classification

In classification, only assignments that were made uniformly by different *k*-values are considered. Results take the form of a list with one assignment for each pattern. In order to obtain the most likely one among these, we introduce a simple bagging strategy, which selects the most frequent assignment.

3.4 Experiments

In order to test the classification system, we ran an experiment on a corpus of medical abstracts (Buitelaar, 2000). Using the methods discussed in (Buitelaar and Sacaleanu, 2001), we automatically extracted a set of domain relevant synsets and domain relevant novel terms.

3.4.1 Evaluation Set

For evaluation purposes, we asked a medical domain expert to manually classify the top 150 novel terms, given a selection of the most domain relevant synsets.

From the top 25 synsets, as proposed by the system, the medical domain expert discarded 4. Further, only 56 novel terms could be manually classified given these 21 synsets. The evaluation set therefore consists of 56 novel terms classified in the following synsets (In order to increase coverage of each synset on the medical corpus, we included besides synonyms also direct hyponyms. Synonyms are in bold.):

[4] As n*ull* attributes play no further role in the classification process, they are discarded in the representation of the context instances,.

*C_1: [**Geschwulst, Geschwür, Tumor**, Abszeß, ...]*
(blastoma, ulcer, tumor, abscess)

*C_2: [**Krankheit**, Abhängigkeit, Anfall, Attacke, ...]*
(disease, addiction, seizure, attack)

*C_3: [**Gewebe, Körpergewebe**, Bindegewebe, ...]*
(tissue, body tissue, connective tissue)

*C_4: [**Entzündung, Infekt, Infektion**, ...]*
(infection, inflammation)

*C_5: [**Krankheitsbild, Syndrom**, ...]*
(clinical syndrom)

*C_6: [**Symptom**]*
(symptom)

*C_7: [**Gelenk**, Ellbogen, Fingergelenk, ...]*
(joint, elbow, finger joint)

*C_8: [**Reduktion, Reduzierung**, Abbau, ...]*
(reduction, decrease, atrophy)

*C_9: [**Anordnung, Aufstellung, Formation**, ...]*
(order, disposition, formation)

*C_{10}: [**Medizin**, Chirurgie, Frauenheilkunde, ...]*
(medicine, surgery, gynecology)

*C_{11}: [**Quote, Rate**, Beschleunigungsquote, ...]*
(proportion, rate)

*C_{12}: [**Parameter**]*
(parameter)

*C_{13}: [**Blutung, Blutverlust**]*
(bleeding, loss of blood)

*C_{14}: [**Facharzt**, Augenarzt, Chirurg, ...]*
(specialist, ophthalmologist, surgeon)

*C_{15}: [**Leiden**, Allergie, Anämie, Artrose, ...]*
(ailment, allergy, anemia, arthosis)

*C_{16}: [**Zelle**, Körperzelle, Pflanzenzelle]*
(cell, body cell, plant cell)

*C_{17}: [**Eingriff, Operation**, Abtreibung, ...]*
(operation, abortion)

*C_{18}: [**Abhandlung, Studie**]*
(survey, case study)

*C_{19}: [**Prophylaxe**, Empfängnisverhütung, ...]*
(prophylaxis, contraception)

*C_{20}: [**Drüse**, Bauchspeicheldrüse, ...]*
(gland, pancreas)

*C_{21}: [**Krankheitssymptom, Symptom**]*
(disease symptom, symptom)

3.4.2 Classification and Results

For each of the patterns, a classifier assigns a synset to each of the 150 novel terms. We used five different values of *k* (3, 6, 9, 12, 15) to validate results. Only assignments that were invariant for all values of *k* are kept. Through our bagging strategy we then select from among all the assignments the most frequent one. Some examples of correctly classified novel terms are given below. A full account of results is presented in Table 1.

C_1:	*Karzinom*	*(carcinoma)*
	Metastase	*(metastasis)*
	Neoplasie	*(neoplasia)*
C_{11}:	*Prävalenz*	*(prevalence)*
	Spezifität	*(specificity)*
C_{17}:	*Resektion*	*(resection)*
	Transplantation	*(transplantation)*

To test our approach in using lexico-syntactic patterns, we also ran an experiment that takes into account the lexical context but with more flexible syntactic constraints. For this purpose we extracted contexts in windows of 3 words on each side of the novel term, as in the original approach, but instead of taking into account the position of these words we now only consider their order of occurrence. Consider the same example as before:

[NN, null, ADJ, (T|C), null, NN, null]

Ignoring syntactic constraints, we now only consider the occurrence of *NN* and *ADJ* on the left and *NN* on the right. Therefore, this pattern is then equivalent to all of the following patterns and many more:

[NN, ADJ, (T|C), NN]
[null, NN, ADJ, (T|C), NN]
[NN, null, ADJ, (T|C), null, NN]

As shown by the results below, our original approach outperforms this alternative approach, which indicates that next to lexical context also a representation of syntactic constraints on this context is an important source of information.

Finally, we also evaluated our strategy to simply sum the mutual information scores for all (relevant: ADJ, NN, VERB) context words. Instead, we ran an experiment with our original approach, keeping mutual information scores separate for each context word. Given the same example again, a context instance then has the following format:

C, MI_1, MI_2, MI_3, *Noun, Adjective, Noun*

Results for each of the three approaches are as follows:

	Manual	System Correct
Approach1	56	23 (41.07%)
Approach2	56	12 (21.43%)
Approach3	56	18 (32.14%)

Table 1: Overall Results for Each Approach

3.4.3 Discussion

Our original approach gives the best results (about 41% precision), which we may compare with a completely random classification of only 5%. A comparison with other systems is almost impossible. First of all, to our knowledge, no other work exists on the automatic classification of terms to WordNet/GermaNet synsets. Work on term clustering is related to our work but not directly comparable. Also, comparing classification results between domains is not straightforward.

Our best result is about 41%, which is relatively high given the completely unsupervised nature of our approach. Classification is performed without any prior sense disambiguation.

The main problem we encountered in this work was the fact that a lot of the synsets in GermaNet were deemed problematic from the medical point of view. For instance, the synset *[Abhandlung, Studie]* connects *Studie (case study)* with *Abhandlung (survey)*.

Additional problems arose in connection with PoS tagging and morphological analysis, specifically concerning compounds, both of which need to be further adapted to the medical domain.

Finally, from the machine learning point of view, we decided that the attributes we use (a context word and its mutual information score) are dependent on each other, which needs to be reflected in the data model. The instance-based algorithm we currently use does not provide such an option.

4 Acknowledgements

We would like to thank Jörg Bay and Oktavian Weiser for providing us with medical expertise.

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

References

Basili R., Della Rocca M. and Pazienza M.-T. *Contextual Word Sense Tuning and Disambiguation*. Applied Artificial Intelligence, vol. 11, 1997.

Brown P., Pietra, V., deSouza P. V., Lai J., and Mercer R. L. *Class-based n-gram models of natural language*. Computational Linguistic, 18:467--479, 1992.

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

Buitelaar P. and Sacaleanu B. *Ranking and Selecting Synsets by Domain Relevance*. In: Proceedings NAACL WordNet Workshop, 2001.

Church, K. and Hanks, P. *Word Association Norms, Mutual Information, and Lexicography*. Computational Linguistics, vol. 16:1, 22-29, 1990.

Cucchiarelli A. and Velardi P. *Finding a Domain-Appropriate Sense Inventory for Semantically Tagging a Corpus*. In: Journal of Natural Language Engineering, 1998.

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

Hamp, B. and Feldweg, H. *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, 1997.

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

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

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

Turcato D., Popowich F., Toole J., Fass D., Nicholson D. and Tisher G. *Adapting a synonym database to specific domains*. In: Proceedings of the ACL workshop on recent advances in NLP and IR. Hong Kong, 2000.

Witten, Ian H., Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. The Morgan Kaufmann Series in Data Management Systems, Jim Gray, 2000.