Automatic Extension of Feature-based Semantic Lexicons via Contextual Attributes

Chris Biemann[1] and Rainer Osswald[2]

[1] Institut für Informatik, Abteilung Automatische Sprachverarbeitung,
Universität Leipzig, 04109 Leipzig, Germany
[2] Fachbereich Informatik, Lehrgebiet Intelligente Informations- und
Kommunikationssysteme, FernUniversität in Hagen, 58084 Hagen, Germany

Abstract. We describe how a feature-based semantic lexicon can be automatically extended using large, unstructured text corpora. Experiments are carried out using the lexicon HaGenLex and the Wortschatz corpus. The semantic classes of nouns are determined via the adjectives that modify them. It turns out to be reasonable to combine several classifiers for single attributes into one for complex semantic classes. The method is evaluated thoroughly and possible improvements are discussed.

1 Introduction

Natural language processing systems for text retrieval and question answering that go beyond mere statistical pattern matching require the semantic analysis of large collections of text. In particular, such systems rely on a reasonably large computational lexicon that provides not only morphosyntactic but also semantic information about lexical units. While building a high quality semantic lexicon might presumably not be possible without manually created lexical entries, there is no doubt that, especially in the case of nouns, automatic classification methods have to be exploited for reasons of quantity and coverage. This paper describes how an automatic semantic classification using co-occurrence statistics on very large text corpora can successfully extend a manually created semantic lexicon.

2 Resources

2.1 The computational lexicon HaGenLex

The lexicon used for our experiments is the semantically based computational lexicon HaGenLex (Hartrumpf et al. 2003). HaGenLex is a domain independent lexicon for German that currently comprises about 25,000 lexical entries, roughly half of which are nouns. All HaGenLex entries are semantically annotated, where the semantic description is based on the MultiNet paradigm, a knowledge representation formalism developed for the representation of natural language semantics (Helbig 2001).

MultiNet provides classificatory as well as relational means of representation. The experiments reported here are restricted to the classification of nouns with respect to their *ontological sort* and *semantic features.* MultiNet defines a hierarchy of 45 ontological sorts like *d* (*discrete object*) and *abs* (*situational object*), of which 17 apply to nouns (cf. Figure 4). In addition, nouns are classified with respect to 16 binary semantic features like HUMAN and MOVABLE (cf. Figure 3). These feature and sorts are not independent of each other; e.g., HUMAN+ implies ANIMATE+, ARTIFICIAL−, and *discrete object.* In order to exclude inconsistent choices, all possible combinations are explicitly combined into (complex) semantic classes, on which a natural specialization hierarchy is defined. In total, there are 50 semantic classes, of which the most frequent 22 in our training data are listed in Figure 5.

2.2 The German corpus 'Projekt Deutscher Wortschatz'

Our text resource is the German main corpus of the 'Projekt Deutscher Wortschatz' (35 million sentences, 500 million tokens).[1] By calculating statistically significant neighboring co-occurrences (Biemann et al. 2004) and part-of-speech filtering, *pairs of adjectives and nouns* are determined that typically co-occur next to each other. If two words A and B are in subsequent position in a corpus, then A is called the left neighbor of B and B the right neighbor of A. To determine pairs of statistically significant neighbors, a significance measure is applied that indicates the amount of "surprise" of seeing frequent co-occurrences of A and B under the assumption of independence – the larger the significance value, the less is the probability that they co-occurred just by chance. If this measure exceeds a certain threshold, we call A a *(left) neighboring co-occurrent* of B and define the *(left) neighboring profile* of B as the set of all (left) neighboring co-occurrents.

Our method for classifying nouns is based on the *Distributional Hypothesis* (Harris 1968), which implies that semantic similarity is a function over global contexts (cf. Miller and Charles 1991). Concretely, we try to classify nouns by considering their modifying adjectives. The set of modifying adjectives for a given noun is here approximated by the statistical *adjective profile* of the noun, which is defined as the set of adjectives in the left neighboring profile of the noun. (Correspondingly, the *noun profile* of an adjective is the set of nouns in its right neighboring profile.) These profiles contain lemmatized words and consist of the union of the full form profiles. From our corpus we extracted over 160,000 nouns that co-occur with one or more of 23,400 adjectives (where half of the nouns have only one adjective in their profile). It has turned out that taking into account the actual significance values has no impact on the classification results; what is important is merely that adjective-noun pairs show up multiple times and typically in the corpus.

[1] See `www.wortschatz.uni-leipzig.de`.

3 Method

3.1 Constructing a classifier for single attributes

For every relevant semantic attribute of nouns, a classifier is constructed in the following way: For every adjective that modifies at least one noun from the training set, a profile is calculated stating the proportion how often this adjective favors which class (class probabilities). The classifier is not limited in the number of classes. Unclassified nouns are then classified on the basis of their adjective profiles; this change between profile calculation and classification of new nouns is iterated in an EM-bootstrapping style (cf. Dempster et al. 1977) until no more nouns can be classified.

```
Initialize adjective and noun profiles;
Initialize the training set;
As long as new nouns get classified:
    Calculate adjective class probabilities;
    For each unclassified noun n:
        Multiply class probabilities class-wise;
        Assign class with highest probability to noun n;
```

Fig. 1. Bootstrapping algorithm for assigning semantic attributes to nouns

Figure 1 gives an overview of the algorithm. In the outer loop, class probabilities are assigned to each adjective that indicate how often this adjective can be found in adjective profiles of nouns of the respective class, i.e., how strong this adjective votes for which class. The probability is calculated from the frequency distribution per class, divided by the total number of nouns per class and normalized in sum to one. Division by the total number of nouns per class is motivated by distributing the same probability mass for all classes and has turned out to be crucial when dealing with skewed class distributions. Because the number of classified nouns increases in every iteration step, the class probabilities per adjective have to be re-calculated in each iteration.

Within the inner loop, the algorithm tries to assign classes to nouns that have not been classified in the previous steps: the class probabilities of the adjectives occurring in the respective adjective profile are multiplied class-wise. Only adjectives occurring in at least one adjective profile of an already classified noun are taken into consideration. The class with the highest value is then assigned to the noun. To increase classificatory precision, one can introduce a threshold α for the minimal number of adjectives in the adjective profile of a noun. The experiments described in Section 4 make use of such a threshold.

3.2 Combining attribute classifications

The overall goal is to classify nouns with respect to the (complex) semantic classes introduced in Section 2.1. In principle, such a classifier could be constructed along the lines of Section 3.1. However, first experiments in that direction have led to a rather unsatisfying precision (tradeoff between 60% precision at 45% recall and 76% precision at only 2.8% recall). The method described here, in contrast, uses separate classifiers for each semantic feature and each ontological sort and combines their results as follows:

(1) Determine all complex semantic classes that are compatible with all results of the individual classifiers.
(2) From the results of (1) select those classes that are minimal with respect to the specialization relation on the set of complex semantic classes.
(3) If the set determined in (2) contains exactly one element, then take this as the result class, otherwise refuse a classification.

The classifier is weak in the sense that it does not always assign a class (which is already the case for the individual classifiers).

The results presented in Section 4.2 are based on the combination method just described. In order to improve the recall, the following two modifications suggest themselves for future experiments: If the set determined in step (2) contains more than one element, select the most specialized semantic class that is more general than all elements in the set. If no class can be found by step (1) then ignore the results of the most unreliable single classifiers step-by-step until a compatible class is found and proceed with (2).

4 Evaluation

For evaluation, we used 10-fold cross validation on a set of 6045 HaGenLex nouns in all experiments. In the preselection of the training set, care was taken to exclude polysemous nouns. The precision (number of correct classifications divided by number of total classifications) was calculated on the basis of the unification of the test sets, although in all experiments a much larger number of nouns could be classified. The threshold α for the minimum number of adjectives in the adjective profile of a noun was varied from 2 to 20, which led to different numbers of total classifications, as shown in Figure 2.

For all further experiments, we (arbitrarily) fixed the minimum number of classifying adjectives to five, which lead to a classification of over 31,000 nouns in all experiments. Since for only 5133 nouns from the HaGenLex training set more than four co-occurring adjectives could have been extracted from the corpus, the *a priori* upper bound on the recall (number of correctly classified divided by number of total items) is 84.9%.

Section 4.1 discusses the results for the individual classifiers for semantic features and ontological sorts, Section 4.2 presents the results for the combined classifier for complex semantic classes.

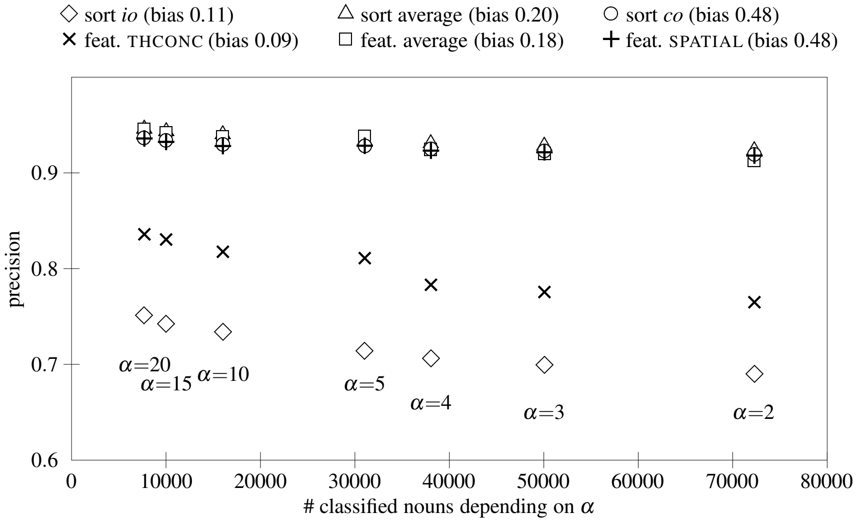

Fig. 2. Minimal adjective number α vs. corpus coverage and classifier precision

4.1 Assignment of semantic features and ontological sorts

As mentioned in Section 3, a separate binary classifier was constructed for all 16 features. Figure 3 shows the distribution in the training data for the semantic features and the fraction of the smaller class (*bias*). It can be seen that the classifiers are able to assign the right features to the test nouns if their bias is not smaller than 0.05. In the other cases we observe a high total precision per feature (METHOD, INSTIT, MENTAL, INFO, ANIMAL and GEOGR) which was more or less obtained by always assigning the more frequent attribute. The less frequent +-attribute is recognized poorly in these cases. The overall precision is 93.8% (87.6% for +-attributes), overall recall is 75.8% (76.9% for +-attributes).

As for the ontological sorts, we constructed for each of the 17 sorts a binary training set that contains words where the sort is present (attribute +) or absent (attribute −). Nouns not specified with respect to the respective sort were excluded from the training set. Figure 4 shows a similar picture as Figure 3: sorts having a bias over 0.1 can be differentiated well or even very well, less frequent sorts lead to problems. Notice that for the sorts *ab* and *o*, the attribute − was taken into consideration in the diagram in Figure 4, because this was the less frequent attribute. Overall precision is 93.3% (90.35% for attribute +) at an overall recall of 79.2% (76.3% for attribute +).

It is worthwhile to recall from Section 2.1 that neither the semantic features nor the ontological sorts are independent of each other. (The ontological sorts are even arranged in a tree hierarchy.) Ideally, the individual classifiers respect these dependencies, which is prerequisite for combining their results to (complex) semantic classes.

feature	#	+	−	bias
METHOD	6004	12	5992	0.0020
INSTIT	6032	39	5993	0.0065
MENTAL	9008	162	8846	0.0180
INFO	6015	119	5896	0.0198
ANIMAL	5995	143	5852	0.0239
GEOGR	6015	188	5827	0.0313
THCONC	6028	518	5510	0.0859
INSTRU	5932	969	4963	0.1634
HUMAN	5995	1313	4682	0.2190
LEGPER	6009	1352	4657	0.2250
ANIMATE	6010	1505	4505	0.2504
POTAG	6015	1664	4351	0.2766
ARTIF	5864	2204	3660	0.3759
AXIAL	5892	2260	3632	0.3836
MOVABLE	5827	2345	3482	0.4024
SPATIAL	6033	2910	3123	0.4823

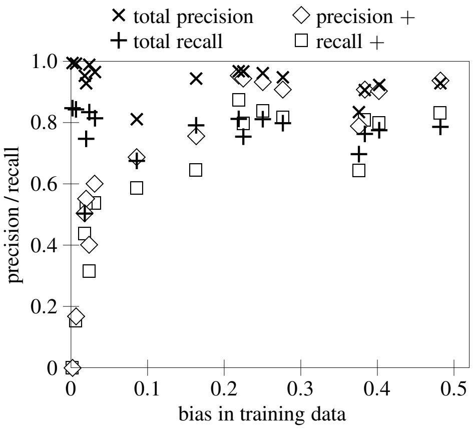

Fig. 3. Left: distribution of features in the training set; right: total precision and recall and precision and recall of +-attributes versus bias in training set

sort	#	+	−	bias
re	6033	7	6026	0.0012
mo	6033	8	6025	0.0013
oa	6033	39	5994	0.0065
o−	6033	5994	39	0.0065
me	6045	41	6004	0.0068
qn	6045	41	6004	0.0068
ta	6033	107	5926	0.0177
s	6010	224	5786	0.0373
as	6031	363	5668	0.0602
na	6033	411	5622	0.0681
at	6033	450	5583	0.0746
io	6033	664	5369	0.1101
ad	6031	1481	4550	0.2456
abs	6033	1846	4187	0.3060
d	6010	2663	3347	0.4431
co	6033	2910	3123	0.4823
ab−	6033	3082	2951	0.4891

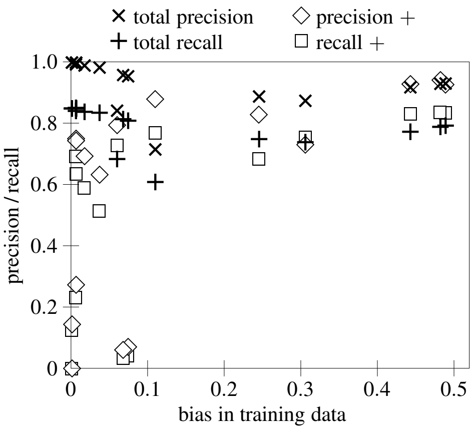

Fig. 4. Left: Distribution of +/− attributes in training sets; right: precision and recall in total per sort and for attributes + versus bias in training data.

4.2 Assignment of complex semantic classes

With respect to the task of extending the given semantic lexicon, the most important point of our approach is the quality of the assignment of complex semantic classes as described in Section 3.2. Figure 5 lists the cross-validation results for all complex semantic classes with at least 40 ($\approx 0.68\%$) occurrences in the training set. For the remaining classes, which comprise about 5.9% of the training set, Figure 5 presents a collective evaluation (class "rest").

An obvious thing to notice is the fact that certain semantic classes are assigned with very good precision whereas others show a rather bad performance. A first conclusion could be that certain semantic properties of nouns

class	#	prec	rec
nonment-dyn-abs-situation	1421	92.25	26.81
human-object	1313	95.05	78.98
prot-theor-concept	516	59.05	12.02
nonoper-attribute	411	0.00	0.00
ax-mov-art-discrete	362	51.94	37.02
nonment-stat-abs-situation	226	48.39	6.64
animal-object	143	100.00	16.08
nonmov-art-discrete	133	57.41	23.31
ment-stat-abs-situation	126	70.00	5.56
nonax-mov-art-discrete	108	40.82	18.52
tem-abstractum	107	97.06	30.84
mov-nonanimate-con-potag	98	73.21	41.84
art-con-geogr	96	55.26	21.88
abs-info	94	35.71	10.64
art-substance	88	65.52	21.59
nat-discrete	88	100.00	25.00
nat-substance	86	64.29	10.47
prot-discrete	73	100.00	53.42
nat-con-geogr	63	80.00	19.05
prot-substance	50	94.44	34.00
mov-art-discrete	45	100.00	31.11
meas-unit	41	100.00	2.44
rest	357	52.17	10.08

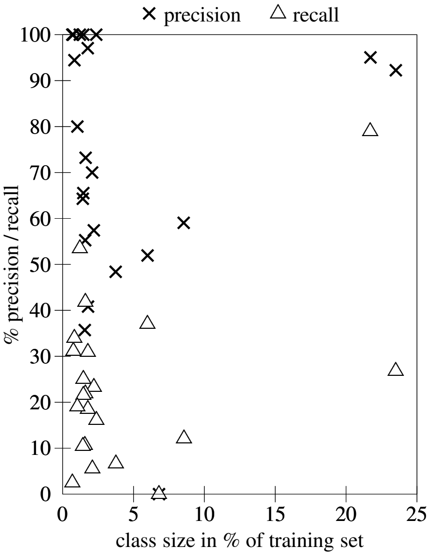

Fig. 5. Precision and recall for complex semantic classes

are reflected by modifying adjectives while others are not. Notice that the assignment of complex semantic classes does not show the same close correspondence between class size and precision that has been observed in the previous section on the classification by single attributes.

The overall precision of the assignment of semantic classes is about 82.3% at a recall of 32.8%. The fairly low recall is due to the fact that the method of Section 3.2 refuses a classification in case the results of the single attribute classifiers are not fully consistent with each other. Despite of this low recall, our approach gives us classification results for about 8500 unknown nouns. If we relax the minimal number α of co-occurring adjectives from five to two, the number of newly classified nouns rises even to almost 13,000, with a reduction of precision of only 0.2%.

5 Conclusion and future work

We have presented a method to automatically extend noun entries of semantic lexica via modifying adjectives. Given a moderate number of training items, the approach is able to classify a high number of previously unclassified nouns at more than 80% overall precision. An evaluation for the different semantic noun classes shows that certain semantic classes can be characterized by modifying adjectives while others can not. It would be interesting to see whether there is a similar distinction for other contextual constellations as, for instance, role filler positions in verb frames, but this requires much more preprocessing.

To improve the recall of our method, the combination of the single attribute classifiers as described in Section 3.2 could be relaxed by taking the quality of the classifiers into account. Another way to circumvent the sparse data problem is to abstract from single adjectives by means of semantic adjective classes like 'physical property'; cf. (Biemann and Osswald 2005, Sect. 6.1). However, this would require a large scale classification of adjectives by appropriate semantic classes.

A further important issue for the extension of the method is the treatment of polysemy: If a word has multiple readings that differ in at least one attribute, the method as proposed here classifies the word according to the most frequent reading in the corpus in the best case. In the worst case, the word will not get classified at all, because the adjectives seem to contradict each other in some attributes. A possibility to split an adjective profile into several profiles, which reflect the different readings, is shown in (Bordag 2003) for untyped co-occurrences and can be paraphrased for the task described here as follows: Presuming one reading per sentence, weak co-occurrence between the context words of the different readings, and strong co-occurrence within the context words of the same reading, the adjective profiles can be split in disjoint subsets that collect modifiers of different noun readings, respectively.

References

BIEMANN, C., BORDAG, S., HEYER, G., QUASTHOFF, U. and WOLFF, C. (2004): Language-independent Methods for Compiling Monolingual Lexical Data. In: *Proceedings of CicLING 2004*. LNCS 2945, Springer, Berlin, 215–228.

BIEMANN, C. and OSSWALD, R. (2005): Automatische Erweiterung eines semantikbasierten Lexikons durch Bootstrapping auf großen Korpora. In: B. Fisseni, H.-C. Schmitz, B. Schröder and P. Wagner (Eds.): *Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen – Beiträge zur GLDV-Tagung 2005 in Bonn*. Peter Lang, Frankfurt am Main, 15–27.

BORDAG, S. (2003): Sentence Co-Occurrences as Small-World-Graphs: A Solution to Automatic Lexical Disambiguation. In: *Proceedings of CicLING 2003*. LNCS 2588, Springer, Berlin, 329–333.

DEMPSTER, A.P., LAIRD, N.M. and RUBIN, D.B. (1977): Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B, 39(1):1–38.*

HARRIS, Z. (1968): *Mathematical Structures of Language*. John Wiley & Sons, New York.

HARTRUMPF, S., HELBIG, H. and OSSWALD, R. (2003): The Semantically Based Computer Lexicon HaGenLex – Structure and Technological Environment. *Traitement automatique des langues, 44(2), 81–105.*

HELBIG, H. (2001): *Die semantische Struktur natürlicher Sprache: Wissensrepräsentation mit MultiNet*. Springer, Berlin

MILLER, G.A. and CHARLES, W.G. (1991): Contextual correlates of semantic similarity. *Language and Cognitive Processes, 6(1):1–28.*