Reducing Lexical Semantic Complexity with Systematic Polysemous Classes and Underspecification

Paul Buitelaar
DFKI Language Technology Lab
Stuhlsatzenhausweg 3,
66123 Saarbrücken, Germany
paulb@dfki.de

Abstract

This paper presents an algorithm for finding systematic polysemous classes in WordNet and similar semantic databases, based on a definition in (Apresjan 1973). The introduction of systematic polysemous classes can reduce the amount of lexical semantic processing, because the number of disambiguation decisions can be restricted more clearly to those cases that involve real ambiguity (homonymy). In many applications, for instance in document categorization, information retrieval, and information extraction, it may be sufficient to know if a given word belongs to a certain class (underspecified sense) rather than to know which of its (related) senses exactly to pick. The approach for finding systematic polysemous classes is based on that of (Buitelaar 1998a, Buitelaar 1998b), while addressing some previous shortcomings.

Introduction

This paper presents an algorithm for finding systematic polysemous classes in WordNet (Miller et al 1990) and GermaNet (Hamp and Feldweg 1997) -- a semantic database for German similar to WordNet. The introduction of such classes can reduce the amount of lexical semantic processing, because the number of disambiguation decisions can be restricted more clearly to those cases that involve real ambiguity (homonymy). Different than with homonyms, systematically polysemous words need not always be disambiguated, because such words have several related senses that are shared in a systematic way by a group of similar words. In many applications then, for instance in document categorization and other areas of information retrieval, it may be sufficient to know if a given word belongs to this group rather than to know which of its (related) senses exactly to pick. In other words, it will suffice to assign a more coarse grained sense that leaves several related senses underspecified, but which can be further specified on demand[1].

The approach for finding systematic polysemous classes is based on that of (Buitelaar 1998a, Buitelaar 1998b), but takes into account some shortcomings as pointed out in (Krymolowski and Roth 1998) (Peters, Peters and Vossen 1998) (Tomuro 1998). Whereas the original approach identified a small set of top-level synsets for grouping together lexical items,

[1] As pointed out in (Wilks 99), earlier work in AI on 'Polaroid Words' (Hirst 87) and 'Word Experts' (Small 81) advocated a similar, incremental approach to sense representation and interpretation. In line with this, the CoreLex approach discussed here provides a large scale inventory of systematically polysemous lexical items with underspecified representations that can be incrementally refined.

the new approach compares lexical items according to all of their synsets on all hierarchy levels. In addition, the new approach is both more flexible and precise by using a clustering algorithm for comparing meaning distributions between lexical items. Whereas the original approach took into account only identical distributions (with additional human intervention to further group together sufficiently similar classes), the clustering approach allows for completely automatic comparisons, relative to certain thresholds, that identify partial overlaps in meaning distributions.

1 Acquisition and Application of Systematic Polysemous Classes

In lexical semantics, a distinction can be made between senses that are of a *contrastive* and those that are of a *complementary* nature (Weinreich 1964). Contrastive senses are unrelated to each other as with the two meanings of ”bank”. However, such clear-cut (contrastive) meaning distinctions are rather the exception than the rule. Often, words have a number of (complementary) senses that are somehow related to each other in systematic ways (Pustejovsky 1995). For instance, a word like ”mouth” has several senses that are all somehow related (after Cruse 1986):

John opened his mouth.
This parasite attaches itself to their mouths.
The mouth of the cave resembles a bottle.
The mouth of the river starts here.

2 CoreLex

Related senses are, however, only systematic (or regular) if more than one example in a language can be found as formulated in (Apresjan 1973):

Polysemy of the word A with the meanings a_i and a_j is called regular if, in the given language, there exists at least one other word B with the meanings b_i b_j, which are semantically distinguished from each other in exactly the same way as a_i and a_j and if a_i and b_i, a_j and b_j are nonsynonymous.

With this definition, we can construct classes of systematically polysemous words as shown in the CoreLex approach (Buitelaar 1998a) (Buitelaar 1998b). This method takes WordNet sense assignments and compares their distribution by reducing them to a set of basic types. For instance, WordNet assigns to the noun ”book” the following senses:

1. publication
2. product, production
3. fact
4. dramatic_composition, dramatic_work
5. record
6. section, subdivision
7. journal

At the top of the WordNet hierarchy these seven senses can be reduced to two basic types: the content that is being communicated and the medium of communication. We can arrive at systematically polysemous classes by investigating which other words share these same senses and are thus polysemous in the same way. For instance, the seven different senses that WordNet assigns to ”book” can be reduced to two basic types: artifact and communication. We do this for each noun and then group them into classes according to their combination of basic types. Finally, by human introspection several classes were grouped together, because their members seemed sufficiently similar.

Among the resulting classes are a number that are to be expected given the literature on systematic polysemy. For instance, the classes animal / food and plant / natural_product have been discussed widely. Other classes are less

expected, but seem quite intuitive. The class artifact / attribute / substance for instance includes a number of nouns (”chalk, charcoal, daub, fiber, fibre, tincture”) that refer to an object that is at the same time an artifact made of some substance and that is also an attribute.

3 CoreLex-II

3.1 A More Flexible Approach

The CoreLex database has been used and/or evaluated in a number of projects, leading to some criticisms of the approach in (Krymolowski and Roth 1998) (Peters, Peters and Vossen 1998) (Tomuro 1998) and in personal communication. Primarily it was argued that the choice of basic types is arbitrary and on too high a level. Systematic class discovery in the original approach is dependent on this set of basic types, which means that classes on lower levels are not captured at all. Further criticism arose on the arbitrariness (and inefficiency) of human intervention in grouping together resulting classes into more comprehensive ones based on the similarity of their members.

In response to this, a new approach was formulated and implemented that addresses both these points. Comparison of sense distributions is now performed over synsets on all levels, not just over a small set on the top levels. In addition, similarity on sense distribution between words need no longer be complete (100%), as with the former approach. Instead, a threshold on similarity can be set that constraints a clustering algorithm for automatically grouping together words into systematic polysemous classes. (No human intervention to further group together resulting classes is required.) This approach took inspiration from the pioneering work by (Dolan 1994), but it is also fundamentally different, because instead of grouping similar senses together, the CoreLex approach groups together words according to all of their senses. Thereby following Apresjan`s definition of systematic polysemy discussed above.

3.2 The Algorithm

The algorithm works as follows (for example for nouns):

1. **foreach** noun
2. get all $level_1$ synsets (senses)
3. **if** number of $level_1$ synsets > 1
   then put noun in *list*
4. **foreach** $level_1$ synset
5. get all higher level synsets (hypernyms)

6. **foreach** $noun_1$ in *list*
7. **foreach** $noun_2$ in *list*
8. compute similarity $noun_1$ and $noun_2$
9. **if** similarity > *threshold*
   then put $noun_1$ and $noun_2$ in *matrix*

10. **foreach** $noun_1$ in *matrix*
11. **if** $noun_1$ not assigned to a cluster
    then construct a new cluster C_i and assign $noun_1$ to it
12. **foreach** $noun_2$ similar to noun1
13. **if** $noun_2$ not assigned to a cluster
    then assign $noun_2$ to new cluster C_i

For every noun in the WordNet or GermaNet index, get all of its senses (which are in fact $level_1$ synsets). If a noun has more than one sense put it in a separate list that will be used for further processing. Nouns with only one sense are not used in further processing because we are only interested in systematic distributions of more than one sense over several nouns. In order to compare nouns not only on the sense level but rather over the whole of the WordNet hierarchy, also all higher level synsets (hypernyms) for each sense are stored.

Then, for each noun we compare its ”sense distribution” (the complete set of synsets derived in the previous steps) with each other noun. Similarity is computed using the Jaccard score, which compares objects

according to the attributes they share and their unique attributes. If the similarity is over a certain threshold, the noun pair is stored in a matrix which is consequently used in a final clustering step.

Finally, the clustering itself is a simple, single link algorithm that groups together objects uniquely in discrete clusters.

3.3 Quantitative and Qualitative Analysis

Depending on the threshold on similarity, the algorithm generates a number of clusters of ambiguous words that share similar sense distributions, and which can be seen as systematic polysemous classes. In the following table an overview is given of results with different thresholds. The number of nouns in WordNet that were processed is 46.995, of which 10.772 have more than one sense.

Threshold	**Number of Clusters** *(Systematic Polysemous Classe)s*	**Ambiguos Nouns in Clusters** *(Systematic Polysemous Nouns)*
0,70	1.793	4.391
0,75	1.341	3.336
0,80	1.002	2.550
0,90	649	1.449

A qualitative analysis of the clusters shows that best results are obtained with a threshold of 0,75. Some of the resulting clusters with this threshold are:

- ball / game
  baseball, basketball, football, handball, volleyball

- fish / food
  albacore, blowfish, bluefin, bluefish, bonito, bream, butterfish, crappie, croaker, dolphinfish, flatfish, flounder, grouper, halibut, lingcod, mackerel, mahimahi, mullet, muskellunge, pickerel, pompano, porgy, puffer, rockfish, sailfish, scup, striper, swordfish, tuna, tunny, weakfish
- plant / nut
  almond, butternut, candlenut, cashew, chinquapin, chokecherry, cobnut, filbert, hazelnut, pistachio
- plant / berry
  bilberry, blueberry, checkerberry, cowberry, cranberry, currant, feijoa, gooseberry, huckleberry, juneberry, lingonberry, serviceberry, spiceberry, teaberry, whortleberry
- vessel / measure
  bottle, bucket, cask, flask, jug, keg, pail, tub
- cord / fabric
  chenille, lace, laniard, lanyard, ripcord, whipcord, worsted
- taste_property / sensation
  acridity, aroma, odor, odour, pungency
- communication / noise
  clamor, hiss, howl, roar, roaring, screaming, screech, screeching, shriek, sigh, splutter, sputter, whisper

4 Application

Systematic polysemous classes that are obtained in this way can be used as filters on sense disambiguation in a variety of applications in which a coarse grained sense assignment will suffice in many cases, but where an option of further specification exists. For instance, in information retrieval

it will not always be necessary to distinguish between the two interpretations of ”baseball, basketball, football, ...”[2].Users looking for information on a baseball-game may be interested also in baseball-balls. On the other hand, a user may be interested specifically in buying a new baseball-ball and does not wish to be flooded with irrelevant information on baseball-games. In this case, the underspecified ball / game sense needs to be further specified in the ball sense only. Similarly, it will not always be necessary to distinguish exactly between the vessel interpretation of ”bottle, bucket, cask, ...” and the measure interpretation, or between the communication interpretation of a ”clamor, hiss, roar, ...” and the noise interpretation.

Currently, a query expansion module based on the approach described here is under development as part of the prototype systems of two EU funded projects: MIETTA[3] (a cross-lingual search engine in the tourism domain – Buitelaar et al 1998) and OLIVE[4] (a cross-lingual video retrieval system).

Also in shallow processing applications like semantic pre-processing for document categorization it will be sufficient to use an underspecified sense instead of needless disambiguation between senses that are roughly equal in their relevance to a certain document category. Similarly, in shallow syntactic processing tasks, like statistical disambiguation of PP-attachment, the use of underspecified senses may be preferable as shown in experiments by (Krymolowski and Roth 1998).

[2] Compare (Schütze 1997) for a similar, but purely statistical approach to underspecification in lexical semantic processing and its use in machine learning and information retrieval.

[3] http://www.mietta.net/mietta

[4] http://twentyone.tpd.tno.nl/olive/

In order to train systems to accurately perform syntactic analysis on the basis of semantic classes, semantically annotated corpora are needed. This is another area of application of the research described here. CoreLex clusters can be considered by annotators as alternatives to WordNet or GermaNet synsets if they are not able to choose between the senses given and instead prefer an underspecified sense. This approach is currently tested, in cooperation with the GermaNet group of the University of Tübingen, in a preliminary project on semantic annotation of German newspaper text.

Conclusion

We presented a new algorithm for generating systematic polysemous classes from existing resources like WordNet and similar semantic databases. Results were discussed for classes of English nouns as generated from WordNet. With a threshold of 75% similarity between nouns, 1341 classes could be found covering 3336 nouns. Not discussed were similar experiments for verbs and adjectives, both in English and German. The resulting classes can be used as filters on incremental sense disambiguation in various applications in which coarse grained (underspecified) senses are preferred, but from which more fine grained senses can be derived on demand.

References

J. Apresjan (1973) *Regular Polysemy*. Linguistics, 142.

Paul Buitelaar (1998a) *CoreLex: Systematic Polysemy and Underspecification.* PhD Thesis, Brandeis University.

Paul Buitelaar (1998b) *CoreLex: An Ontology of Systematic Polysemous Classes.* In: Formal Ontology in Information Systems. IOS Press, Amsterdam.

Paul Buitelaar, Klaus Netter and Feiyu Xu (1998) *Integrating Different Strategies In Cross-Language Information Retrieval in the MIETTA Project.* In: Proceedings of TWLT14, Enschede, the Netherlands , December.

D. A. Cruse (1986) *Lexical Semantics.* Cambridge University Press.

Bill Dolan (1994) *Word Sense Ambiguation: Clustering Related Senses.* In: Proceedings of COLING-94. Kyoto, Japan.

Birgit Hamp and Helmut Feldweg (1997) *GermaNet-a Lexical Semantic Net for German.* In: Proceedings of the ACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications. Madrid,.

G. Hirst (1987) *Semantic Interpretation and the Resolution of Ambiguity.* Cambridge University Press.

Yuval Krymolowski and Dan Roth (1998) *Incorporating Knowledge in Natural Language Learning: A Case Study.* In: Proceedings ACL-98 Workshop on the Use of WordNet in NLP.

G. A. Miller and R. Beckwith and Ch. Fellbaum and D. Gross and K. Miller (1990) *Introduction to WordNet: An On-line Lexical Database.* International Journal of Lexicography, 3,4.

Wim Peters, Ivonne Peters and Piek Vossen (1998) *Automatic Sense Clustering in EuroWordNet.* In: Proceedings of LREC. Granada.

James Pustejovsky (1995) *The Generative Lexicon.* MIT Press.

Hinrich Schütze (1997) *Ambiguity Resolution in Language Learning.* Volume 71 of CSLI Publications. Chicago University Press.

S. Small (1981) *Viewing Word Expert Parsing as Linguistic Theory.* In: Proceedings of IJCAI.

Noriko Tomuro (1998) *Semi-Automatic Induction of Systematic Polysemy from WordNet.* In: Proceedings ACL-98 Workshop on the Use of WordNet in NLP.

Uriel Weinreich (1964) *Webster's Third: A Critique of its Semantics.* International Journal of American Linguistics, 405-409, 30.

Yorick Wilks (1999) *Is Word Sense Disambiguation just one more NLP task?* Cs.CL/9902030.