Disambiguating Noun and Verb Senses Using Automatically Acquired Selectional Preferences*

Diana McCarthy and **John Carroll**
Cognitive & Computing Sciences
University of Sussex
Brighton BN1 9QH, UK
{dianam,johnca}@cogs.susx.ac.uk

Judita Preiss
Computer Laboratory
University of Cambridge, JJ Thomson Avenue
Cambridge CB3 0FD, UK
Judita.Preiss@cl.cam.ac.uk

Abstract

Our system for the SENSEVAL-2 all words task uses automatically acquired selectional preferences to sense tag subject and object head nouns, along with the associated verbal predicates. The selectional preferences comprise probability distributions over WordNet nouns, and these distributions are conditioned on WordNet verb classes. The conditional distributions are used directly to disambiguate the head nouns. We use prior distributions and Bayes rule to compute the highest probability verb class, given a noun class. We also use anaphora resolution and the 'one sense per discourse' heuristic to cover nouns and verbs not occurring in these relationships in the target text. The selectional preferences are acquired without recourse to sense tagged data so our system is unsupervised.

1 Introduction

In the first SENSEVAL, we used automatically acquired selectional preferences to disambiguate head nouns occurring in specific grammatical relationships (Carroll and McCarthy, 2000). The selectional preference models provided co-occurrence behaviour between WordNet synsets[1] in the noun hyponym hierarchy and verbal predicates. Preference scores, based on mutual information, were attached to the classes in the models. These scores were conditioned on the verbal context and the grammatical relationship in which the nouns for training had occurred. The system performed comparably to the other system using selectional preferences alone.

The work here is an extension of this earlier work, this time applied to the English all words task. We use probability distributions rather than mutual information to quantify the preferences. The preference models are modifications of the Tree Cut Models (TCMs) originally proposed by Li and Abe (1995; 1998). A TCM is a set of classes cutting across the WordNet noun hypernym hierarchy which covers all the nouns of WordNet disjointly, i.e. the classes in the set are not hyponyms of one another. The set of classes is associated with a probability distribution. In our work, we acquire TCMs conditioned on a verb class, rather than a verb form. We then use Bayes rule to obtain probability estimates for verb classes conditioned on co-occurring noun classes.

Using selectional preferences alone for disambiguation enables us to investigate the situations when they are useful, as well as cases when they are not. However, this means we loose out in cases where preferences do not provide the necessary information and other complementary information would help. Another disadvantage of using selectional preferences alone for disambiguation is that the preferences only apply to the grammatical slots for which they have been acquired. In addition, selectional preferences only help disambiguation for slots where there is a strong enough tie between predicate and argument. In this work, we use subject and object relationships, since these appear to work better than other relationships (Resnik, 1997; McCarthy, 2001), and we use argument heads, rather than the entire argument phrase.

Our basic system is restricted to using only selectional information, and no other source of disambiguating information. However, we ex-

* This work was supported by UK EPSRC projects GR/L53175 'PSET: Practical Simplification of English Text' and GR/N36462/93 'Robust Accurate Statistical Parsing (RASP)'.

[1] We will hereafter refer to WordNet synsets as classes.

perimented with two methods of extending the coverage to include other grammatical contexts. The first of these methods is the 'one sense per discourse' heuristic (Gale et al., 1992). With this method a sense tag for a given word is applied to other occurrences of the same word within the discourse. The second method uses anaphora resolution to link pronouns to their antecedents. Using the anaphoric links we are able to use the preferences for a verb co-occurring with a pronoun with the antecedent of that pronoun.

2 System Description

There is a training phrase and a run-time disambiguation phase for our system. In the training phase a preprocessor and parser are used to obtain training data for selectional preference acquisition. At run-time the preprocessor and parser are used for identifying predicates and argument heads for application of the acquired selectional preferences for disambiguation. Anaphora resolution is used at run-time to make links between antecedents of nouns, where the antecedents or the predicates may be in subject or object relationships.

2.1 Preprocessor and Parser

The preprocessor consists of three modules applied in sequence: a tokeniser, a part-of-speech (PoS) tagger, and a lemmatiser. The tokeniser comprises a small set of manually-developed finite-state rules for identifying word and sentence boundaries. The tagger (Elworthy, 1994) uses a bigram HMM augmented with a statistical unknown word guesser. When applied to the training data for selectional preference acquisition it produces the single highest-ranked tag for each word; at run-time it returns multiple tags whose associated forward-backward probabilities are incorporated into parse probabilities. The lemmatiser (Minnen et al., 2001) reduces inflected verbs and nouns to their base forms.

The parser uses a 'shallow' unification-based grammar of English PoS tags, performs disambiguation using a context-sensitive probabilistic model (Carroll and Briscoe, 1996), and recovers from extra-grammaticality by returning partial parses. The output of the parser is a set of *grammatical relations* specifying the syntactic dependency between each head and its dependent(s), read off from the phrase structure tree that is returned from the disambiguation phase. For selectional preference acquisition we applied the analysis system to the 90 million words of the written portion of the British National Corpus (BNC); both in the acquisition phase and at run-time we extracted from the analyser output only subject-verb and verb-direct object dependencies[2]. Thus we did not use the SENSEVAL-2 Penn Treebank-style bracketings supplied for the test data.

2.2 Selectional Preferences

A TCM provides a probability distribution over the noun hyponym hierarchy of WordNet. We acquire TCMs conditioned on WordNet verb classes to represent the selectional preferences of the verbs in that verb class. The noun frequency data used for acquiring a TCM is that occurring with verbs from the target verb class. The verb members for training are taken from the class directly and all hyponym classes. However not all verbs in a verb class are used for training. We use verbs which have a frequency at or above 20 in the BNC, and belong to no more than 10 WordNet classes.

The noun data is used to populate the hypernym hierarchy with frequencies, where the frequency count for any noun is divided by the number of noun classes it is a member of. A hyperonym class includes the frequency credit attributed to all its hyponyms.

A portion of two TCMs is shown in figure 1. The TCMs are similar as they both contain direct objects occurring with the verb *seize*; the TCM for the class which includes *clutch* has a higher probability for the **entity** noun class compared to the class which also includes *assume* and *usurp*. This example includes only classes at WordNet roots, although it is quite possible for the TCM to use more specific noun classes. The method for determining the generalisation level uses the minimum description length principle and is a modification of that proposed by Li and Abe (1995; 1998). In our modification, all internal nodes of WordNet have their synonyms placed at newly created leaves. Doing this ensures that all nouns are

[2] In a previous evaluation of grammatical relation accuracy, the analyser returned subject-verb and verb-direct object dependencies with 84–88% recall and precision (Carroll et al., 1999).

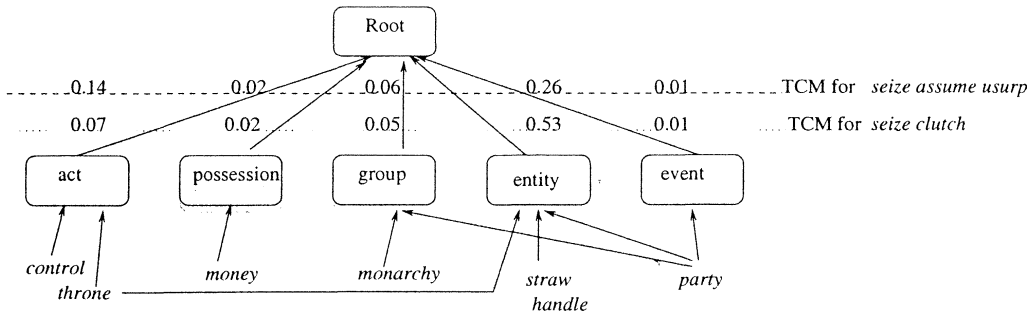

Figure 1: TCMs for the direct object slot of two verb classes which include the verb *seize*.

covered by the probability distribution specified by the TCM.

2.3 Disambiguation

The probability distributions enable us to get estimates for $p(noun\ class|verb\ class)$ for disambiguation. To disambiguate a noun occurring with a given verb, the noun class ($n1$) out of all those to which the noun belongs that gives the largest estimate for $p(n1|v1)$ is taken, where the verb class ($v1$) is the one for the co-occurring verb which maximises this estimate. The selectional preferences provide an estimate for $p(n1|v1)$. The probability estimate of the hyperonym noun class ($n2$) occurring above $n1$ on the TCM for $v1$ is multiplied by the ratio of the prior probability estimate for the hyponym divided by that for the hyperonym on the TCM, i.e. by $\frac{p(n1)}{p(n2)}$. These prior estimates are taken from populating the noun hypernym hierarchy with the prior frequency data.

To disambiguate a verb occurring with a given noun, the verbclass ($v2$) which gives the largest estimate for $p(v2|n3)$ is taken. The noun class ($n3$) for the co-occurring noun is taken as the one that maximises this estimate. Bayes rule is used to obtain this estimate:

$$p(v2|n3) = p(n3|v2)\ \frac{p(v2)}{p(n3)}$$

The TCMs for the candidate verb classes are used for the estimate of $p(n3|v2)$. The estimate for $p(n3)$ is taken from a frequency distribution stored over the entire noun hyponym hierarchy for the prior noun data for the target grammatical slot. The estimate $p(v2)$ is taken from a frequency distribution over the entire verb hyponym hierarchy for the given grammatical slot.

2.4 Increasing Coverage – OSPD and anaphora resolution

When applying the one sense per discourse (OSPD) heuristic, we simply used a tag for a noun, or verb to apply to all the other nouns (or verbs) in the discourse, provided that there was not more than one possible tagging provided by the selectional preferences for that discourse.

In order to increase coverage of the selectional preferences we used anaphoric links to allow preferences of verbs occurring with pronouns to apply to antecedents.

The anaphora resolution algorithm implemented is due to Kennedy and Boguraev (1996). The algorithm resolves third person pronouns, reciprocals and reflexives, and its cited accuracy is 75% when evaluated on various texts taken from the World Wide Web.

The algorithm places each discourse referent into a coreference class, where discourse referents in the same class are believed to refer to the same object. The classes have a salience value associated with them, and an antecedent for a pronoun is chosen from the class with the highest salience value. The salience value of a class is computed by assigning weights to the grammatical features of its discourse referents, and these grammatical features are obtained from the Briscoe and Carroll (1996) parser.

3 Evaluation

We entered three systems for the SENSEVAL-2 English all words task:

sussex-sel Selectional preferences were used alone. Preferences at the subject slot were applied first, if these were not applicable then the direct object slot was tried.

System (sussex-)	Precision (%)	Recall (%)	Attempted (%)
sel	59.8	14.0	23
sel-ospd	56.6	16.9	30
sel-ospd-ana	54.5	16.9	31

Table 1: English all words fine-grained results

Slot	Nouns (%)	Verbs (%)
subject	34	36
direct object	28	45
random baseline	24	25

Table 2: Analysis of sussex-sel precision for polysemous nouns and verbs

sussex-sel-ospd The selectional preferences were applied first, followed by the one sense per discourse heuristic. In the English all words task a discourse was demarcated by a unique text identifier.

sussex-sel-ospd-ana The selectional preferences were used, then the anaphoric links were applied to extend coverage, and finally the one sense per discourse was applied.

The results are shown in table 1. We only attempted disambiguation for head nouns and verbs in subject and direct object relationships, those tagged using anaphoric links to antecedents in these relationships and those tagged using the one sense per discourse heuristic. We do not include the coarse-grained results which are just slightly better than the fine-grained results, and this seems to be typical of other systems. We did not take advantage of the coarse grained classification as this was not available at the time of acquiring the selectional preferences.

From analysis of the fine-grained results of the selectional preference results for system sussex-sel, we see that nouns performed better than verbs because there were more monosemous nouns than verbs. However, if we remove the monosemous cases, and rely on the preferences, the verbs were disambiguated more accurately than the nouns, having only a 1% higher random baseline. Also, the direct object slot outperformed the subject slot. In future it would be better to use the preferences from this slot first.

4 Conclusions

Given that this method is unsupervised, we feel our results are promising. The one sense per discourse heuristic works well and increases coverage. However, we feel that anaphora resolution information has not reached its full potential. There is plenty of scope for combining evidence from several anaphoric links, especially once we have covered more grammatical relationships. We hope that precision can also be improved by combining or comparing several pieces of evidence for a single test item. We are currently acquiring preferences for adjective-noun relationships.

References

John Carroll and Ted Briscoe. 1996. Apportioning development effort in a probabilistic LR parsing system through evaluation. In *ACL/SIGDAT Conference on Empirical Methods in Natural Language Processing*, pages 92–100, University of Pennsylvania, PA.

John Carroll and Diana McCarthy. 2000. Word sense disambiguation using automatically acquired verbal preferences. *Computers and the Humanities. Senseval Special Issue*, 34(1–2):109–114.

John Carroll, Guido Minnen, and Ted Briscoe. 1999. Corpus annotation for parser evaluation. In *EACL-99 Workshop on Linguistically Interpreted Corpora*, pages 35–41, Bergen, Norway.

David Elworthy. 1994. Does Baum-Welch re-estimation help taggers? In *4th ACL Conference on Applied Natural Language Processing*, pages 53–58, Stuttgart, Germany.

William Gale, Kenneth Church, and David Yarowsky. 1992. A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26:415–439.

Chris Kennedy and Bran Boguraev. 1996. Anaphora for everyone: Pronominal anaphora resolution without a parser. In *16th International Conference of Computational Linguistics, COLING-96*, pages 113–118.

Hang Li and Naoki Abe. 1995. Generalizing case frames using a thesaurus and the MDL principle. In *International Conference on Recent Advances in Natural Language Processing*, pages 239–248, Bulgaria.

Hang Li and Naoki Abe. 1998. Generalizing case frames using a thesaurus and the MDL principle. *Computational Linguistics*, 24(2):217–244.

Diana McCarthy. 2001. *Lexical acquisition at the syntax-semantics interface: diathesis alternations, subcategorization frames and selectional preferences.* Ph.D. thesis, University of Sussex.

Guido Minnen, John Carroll, and Darren Pearce. 2001. Applied morphological processing of English. *Natural Language Engineering*, 7(3):207–223.

Philip Resnik. 1997. Selectional preference and sense disambiguation. In SIGLEX *Workshop on Tagging Text with Lexical Semantics: Why What and How?*, pages 52–57, Washington, DC.