Proceedings of the SIGLEX/SENSEVAL Workshop on Word Sense Disambiguation: Recent Successes and Future Directions, Philadelphia, July 2002, pp. 88-94. Association for Computational Linguistics.

Combining Contextual Features for Word Sense Disambiguation

Hoa Trang Dang and Martha Palmer
Department of Computer and Information Sciences
University of Pennsylvania
Philadelphia, PA, USA, 19104
{htd,mpalmer}linc.cis.upenn.edu

Abstract

In this paper we present a maximum entropy Word Sense Disambiguation system we developed which performs competitively on SENSEVAL-2 test data for English verbs. We demonstrate that using richer linguistic contextual features significantly improves tagging accuracy, and compare the system's performance with human annotator performance in light of both fine-grained and coarse-grained sense distinctions made by the sense inventory.

1 Introduction

Highly ambiguous words pose continuing problems for Natural Language Processing (NLP) applications. They can lead to irrelevant document retrieval in IR systems, and inaccurate translations in Machine Translation systems (Palmer et al., 2000). While homonyms like *bank* are fairly tractable, polysemous words like *run*, with related but subtly distinct meanings, present the greatest hurdle for Word Sense Disambiguation (WSD). SENSEVAL-1 and SENSEVAL-2 have attempted to provide a framework for evaluating automatic systems by creating corpora tagged with fixed sense inventories, which also enables the training of supervised WSD systems.

In this paper we describe a maximum entropy WSD system that combines information from many different sources, using as much linguistic knowledge as can be gathered automatically by current NLP tools. Maximum entropy models have been applied to a wide range of classification tasks in NLP (Ratnaparkhi, 1998). Our maximum entropy system performed competitively with the best performing systems on the English verb lexical sample task in SENSEVAL-1 and SENSEVAL-2. We compared the system performance with human annotator performance in light of both fine-grained and coarse-grained sense distinctions made by WordNet in SENSEVAL-2, and found that many of the system's errors on fine-grained senses stemmed from the same sources that caused disagreements between human annotators. These differences were partially resolved by backing off to more coarse-grained sense-groups, which are sometimes necessary when even human annotators cannot make the fine-grained sense distinctions specified in the dictionary.

2 Related Work

While it is possible to build an automatic sense tagger using only the dictionary definitions, the most accurate systems tend to take advantage of supervised learning. The system with the highest overall performance in SENSEVAL-1 used Yarowsky's hierarchical decision lists (Yarowsky, 2000); while there is a large set of potential features, only a small number is actually used to determine the sense of any given instance of a word. Chodorow, Leacock and Miller (Chodorow et al., 2000) also achieved high accuracy using naive bayesian models for WSD, combining sets of linguistically impoverished features that were classified as either topical or local. Topical features consisted of a bag of open-class words in a wide window covering the entire context provided; local features were words and parts of speech within a small window or at particular offsets

from the target word. The system was configured to use only local, only topical, or both local and topical features for each word, depending on which configuration produced the best result on a held-out portion of the training data.

Previous experiments (Ng and Lee, 1996) have explored the relative contribution of different knowledge sources to WSD and have concluded that collocational information is more important than syntactic information. Additionally, Pedersen (Pedersen, 2001; Pedersen, 2000) has pursued the approach of using simple word bigrams and other linguistically impoverished feature sets for sense tagging, to establish upper bounds on the accuracy of feature sets that do not impose substantial pre-processing requirements. In contrast, we wish to demonstrate that such pre-processing significantly improves accuracy for sense-tagging English verbs, because we believe that they allow us to extract a set of features that more closely parallels the information humans use for sense disambiguation.

3 System Description

We developed an automatic WSD system that uses a maximum entropy framework to combine linguistic contextual features from corpus instances of each verb to be tagged. Under the maximum entropy framework (Berger et al., 1996), evidence from different features can be combined with no assumptions of feature independence. The automatic tagger estimates the conditional probability that a word has sense x given that it occurs in context y, where y is a conjunction of features. The estimated probability is derived from feature weights which are determined automatically from training data so as to produce a probability distribution that has maximum entropy, under the constraint that it is consistent with observed evidence.

In order to extract the linguistic features necessary for the model, all sentences were first automatically part-of-speech-tagged using a maximum entropy tagger (Ratnaparkhi, 1998) and parsed using the Collins parser (Collins, 1997). In addition, an automatic named entity tagger (Bikel et al., 1997) was run on the sentences to map proper nouns to a small set of semantic classes. Following work by Chodorow, Leacock and Miller, we divided the possible model features into topical and local contextual features. Topical features looked for the presence of keywords occurring *anywhere* in the sentence and any surrounding sentences provided as context (usually one or two sentences). The set of 200-300 keywords is specific to each lemma to be disambiguated, and is determined automatically from training data so as to minimize the entropy of the probability of the senses conditioned on the keyword.

The local features for a verb w in a particular sentence tend to look only within the smallest clause containing w. They include *collocational* features requiring no linguistic preprocessing beyond part-of-speech tagging (1), *syntactic* features that capture relations between the verb and its complements (2-4), and *semantic* features that incorporate information about noun classes for objects (5-6):

1. the word w, the part of speech of w, and words at positions -2, -1, +1, +2, relative to w
2. whether or not the sentence is passive
3. whether there is a subject, direct object, indirect object, or clausal complement (a complement whose node label is S in the parse tree)
4. the words (if any) in the positions of subject, direct object, indirect object, particle, prepositional complement (and its object)
5. a Named Entity tag (PERSON, ORGANIZATION, LOCATION) for proper nouns appearing in (4)
6. WordNet synsets and hypernyms for the nouns appearing in (4)[1]

This set of local features relies on access to syntactic structure as well as semantic class information, and represents our move towards using richer syntactic and semantic knowledge sources to model human performance.

[1]Nouns were not disambiguated in any way, and all possible synsets and hypernyms for the noun were included. No separate disambiguation of noun complements was done because, given enough data, the maximum entropy model should assign high weights to the correct semantic classes of the correct noun sense if they represent defining selectional restrictions.

4 Evaluation

In this section we describe the system performance on the verbs from SENSEVAL-1 and SENSEVAL-2. The system was built after SENSEVAL-1 but before SENSEVAL-2.[2]

SENSEVAL-1 SENSEVAL-1 used a DARPA-style evaluation format where the participants were provided with hand-annotated training data and test data. The lexical inventory used was the Hector lexicon, developed jointly by DEC and Oxford University Press (Kilgarriff and Rosenzweig, 2000). By allowing for discussion and revision of confusing lexical entries during tagging, before the final test data was tagged, inter-annotator agreement of over 90% was eventually achieved. However, the Hector lexicon was very small and under proprietary constraints, making it an unsuitable candidate for applications requiring a large-scale, publicly-available dictionary.

SENSEVAL-2 The subsequent SENSEVAL-2 exercise used a pre-release version of WordNet1.7 which is much larger than Hector and is more widely used in NLP applications. The average training set size for verbs was only about half of that provided in SENSEVAL-1, while the average polysemy of each verb was higher[3]. Smaller training sets and the use of a large-scale, publicly available dictionary arguably make SENSEVAL-2 a more indicative evaluation of WSD systems in the current NLP environment than SENSEVAL-1. The role of sense groups was also explored as a way to address the popular criticism that WordNet senses are too vague and fine-grained. During the data preparation for SENSEVAL-2, previous WordNet groupings of the verbs were carefully re-examined, and specific semantic criteria were manually associated with each group. This occasionally resulted in minor revisions of the original groupings (Fellbaum et al., 2001). This manual method of creating a more coarse-grained sense inventory from WordNet contrasts with automatic methods that rely on existing semantic links in WordNet (Mihalcea and Moldovan, 2001), which can produce divergent dictionaries.

Our system performs competitively with the best performing systems in SENSEVAL-1 and SENSEVAL-2. Measuring accuracy as the recall score (which is equal to precision in our case because the system assigns a tag to every instance), we compare the system's coarse-grained scores using the revised groupings versus random groupings, and demonstrate the coherence and utility of the groupings in reconciling apparent tagging disagreements.

4.1 SENSEVAL-1 Results

The maximum entropy WSD system's performance on the verbs from the evaluation data for SENSEVAL-1 (Kilgarriff and Rosenzweig, 2000) rivaled that of the best-performing systems. Table 1 shows the performance of variants of the system using different subsets of possible features. In addition to experimenting with different combinations of local/topical features, we attempted to undo passivization transformations to recover underlying subjects and objects. This was expected to increase the accuracy with which verb arguments could be identified, helping in cases where selectional restrictions on arguments played an important role in differentiating between senses.

The best overall variant of the system for verbs did not use WordNet class features, but included topical keywords and passivization transformation, giving an average verb accuracy of 72.3%. This falls between Chodorow, Leacock, and Miller's accuracy of 71.0%, and Yarowsky's 73.4% (74.3% post-workshop). If only the best combination of feature sets for each verb is used, then the maximum entropy models achieve 73.7% accuracy. Even though our system used only the training data provided and none of the information from the dictionary itself, it was still competitive with the top performing systems which also made use of the dictionary to identify multi-word constructions. As we show later, using this additional piece of information improves performance substantially.

In addition to the SENSEVAL-1 verbs, we ran the system on the SENSEVAL-1 data for *shake*, which contains both nouns and verbs. The system simply excluded verb complement features whenever the part-of-speech tagger indicated that the word

[2]The system did not compete officially in SENSEVAL-2 because it was developed by people who were involved in coordinating the English verbs lexical sample task.

[3]The average number of senses per verb in the training data was 11.6 using the Hector dictionary in SENSEVAL-1, and 15.6 using WordNet1.7 in SENSEVAL-2.

task	lex	lex+topic	lex+trans+topic	wn	wn+topic	wn+trans+topic
amaze	0.957	0.928	0.942	0.957	0.899	0.913
bet-v	0.709	0.667	0.667	0.718	0.650	0.650
bother	0.866	0.852	0.847	0.837	0.828	0.823
bury	0.468	0.502	0.517	0.572	0.537	0.532
calculate	0.867	0.902	0.904	0.862	0.881	0.872
consume	0.481	0.492	0.508	0.454	0.503	0.454
derive	0.682	0.682	0.691	0.659	0.664	0.696
float-v	0.437	0.441	0.445	0.406	0.445	0.432
invade	0.560	0.522	0.531	0.580	0.551	0.536
promise-v	0.906	0.902	0.902	0.888	0.893	0.893
sack-v	0.972	0.972	0.972	0.966	0.966	0.966
scrap-v	0.812	0.866	0.871	0.796	0.876	0.882
seize	0.653	0.741	0.745	0.660	0.691	0.703
verbs	0.705	0.718	0.723	0.703	0.711	0.709
shake-p	0.744	0.725	0.742	0.767	0.770	0.758

Table 1: Accuracy of different variants of maximum entropy models on SENSEVAL-1 verbs. Only local information was used, unless indicated by "+topic," in which case the topical keyword features were included in the model; "wn" indicates that WordNet class features were used, while "lex" indicates only lexical and named entity tag features were used for the noun complements; "+trans" indicates that an attempt was made to undo passivization transformations.

to be sense-tagged was not a verb. Even on this mix of nouns and verbs, the system performed well compared with the best system for *shake* from SENSEVAL-1, which had an accuracy of 76.5% on the same task.

4.2 SENSEVAL-2 Results

We also tested the WSD system on the verbs from the English lexical sample task for SENSEVAL-2. In contrast to SENSEVAL-1, senses involving multi-word constructions could be directly identified from the sense tags themselves (through the WordNet sense keys that were used as sense tags), and the head word and satellites of multi-word constructions were explicitly marked in the training and test data. This additional annotation made it much easier for our system to incorporate information about the satellites, without having to look at the dictionary (whose format may vary from one task to another). The best-performing systems on the English verb lexical sample task (including our own) filtered out possible senses based on the marked satellites, and this improved performance.

Table 2 shows the performance of the system using different subsets of features. While we found little improvement from transforming passivized sentences into a more canonical form to recover underlying arguments, there is a clear improvement in performance as richer linguistic information is incorporated in the model. Adding topical keywords also helped.

Incorporating topical keywords as well as collocational, syntactic, and semantic local features, our system achieved 59.6% and 69.0% accuracy using fine-grained and coarse-grained scoring, respectively. This is in comparison to the next best-performing system, which had fine- and coarse-grained scores of 57.6% and 67.2% (Palmer et al., 2001). Here we see the benefit from including a filter that only considered phrasal senses whenever there were satellites of multi-word constructions marked in the test data; had we not included this filter, our fine- and coarse-grained scores would have been only 56.9% and 66.1%.

Table 3 shows a breakdown of the number of senses and groups for each verb, the fine-grained accuracy of the top three official SENSEVAL-2 systems, fine- and coarse-grained accuracy of our maxi-

Feature Type (local only)	Accuracy	Feature Type (local and topical)	Accuracy
collocation	47.6	collocation	49.8
+ syntax	54.9	+ syntax	57.1
+ syntax + transform	55.1	+ syntax + transform	57.3
+ syntax + semantics	58.3	+ syntax + semantics	59.6
+ syntax + semantics + transform	58.9	+ syntax + semantics + transform	59.5

Table 2: Accuracy of maximum entropy system using different subsets of features for SENSEVAL-2 verbs.

Verb	Senses	Groups	SMULS	JHU	KUNLP	MX	MX-c	ITA	ITA-c
begin	8	8	87.5	71.4	81.4	83.2	83.2	81.2	81.4
call	23	16	40.9	43.9	48.5	47.0	63.6	69.3	89.2
carry	27	17	39.4	51.5	45.5	37.9	48.5	60.7	75.3
collaborate	2	2	90.0	90.0	90.0	90.0	90.0	75.0	75.0
develop	15	5	36.2	42.0	42.0	49.3	68.1	67.8	85.2
draw	32	20	31.7	41.5	34.1	36.6	51.2	76.7	82.5
dress	14	8	57.6	59.3	71.2	61.0	89.8	86.5	100.0
drift	9	6	59.4	53.1	53.1	43.8	43.8	50.0	50.0
drive	15	10	52.4	42.9	54.8	59.5	78.6	58.8	71.7
face	7	4	81.7	80.6	82.8	81.7	90.3	78.6	97.4
ferret	1	1	100.0	100.0	100.0	100.0	100.0	100.0	100.0
find	17	10	29.4	26.5	27.9	27.9	39.7	44.3	56.9
keep	27	22	44.8	55.2	44.8	56.7	58.2	79.1	80.1
leave	14	10	47.0	51.5	50.0	62.1	66.7	67.2	80.5
live	10	8	67.2	59.7	59.7	68.7	70.1	79.7	87.2
match	8	4	40.5	52.4	52.4	47.6	69.0	56.5	82.6
play	25	18	50.0	45.5	37.9	50.0	51.5	*	*
pull	33	28	48.3	55.0	45.0	53.3	68.3	68.1	72.2
replace	4	2	44.4	57.8	55.6	62.2	93.3	65.9	100.0
see	21	13	37.7	42.0	39.1	47.8	55.1	70.9	75.5
serve	12	7	49.0	54.9	68.6	68.6	72.5	90.8	93.2
strike	26	21	38.9	48.1	40.7	33.3	44.4	76.2	90.5
train	9	4	41.3	54.0	58.7	57.1	69.8	28.8	55.0
treat	6	5	63.6	56.8	56.8	56.8	63.6	96.9	97.5
turn	43	31	35.8	44.8	37.3	44.8	56.7	74.2	89.4
use	7	4	72.4	72.4	65.8	65.8	78.9	74.3	89.4
wander	4	2	74.0	78.0	82.0	82.0	90.0	65.0	90.0
wash	13	10	66.7	58.3	83.3	75.0	75.0	87.5	90.6
work	21	14	43.3	45.0	45.0	41.7	56.7	*	*
TOTAL	15.6	10.7	56.3	56.6	57.6	59.6	69.0	71.3	82.0

Table 3: Number of senses and sense groups in training data for each SENSEVAL-2 verb; fine-grained accuracy of top three competitors (JHU, SMULS, KUNLP) in SENSEVAL-2 English verbs lexical sample task; fine-grained (MX) and coarse-grained accuracy (MX-c) of maximum entropy system; inter-tagger agreement for fine-grained senses (ITA) and sense groups (ITA-c). *No inter-tagger agreement figures were available for “play” and “work”.

mum entropy system, and human inter-tagger agreement on fine-grained and coarse-grained senses. Overall, coarse-grained evaluation using the groups improved the system's score by about 10%. This is consistent with the improvement we found in inter-tagger agreement for groups over fine-grained senses (82% instead of 71%). As a base-line, to ensure that the improvement did not come simply from the lower number of tag choices for each verb, we created random groups. Each verb had the same number of groups, but with the senses distributed randomly. We found that these random groups provided almost no benefit to the inter-annotator agreement figures (74% instead of 71%), confirming the greater coherence of the manual groupings.

4.3 Analysis of errors

We found that the grouped senses for *call* substantially improved performance over evaluating with respect to fine-grained senses; the system achieved 63.6% accuracy with coarse-grained scoring using the groups, as compared to 47.0% accuracy with fine-grained scoring. When evaluated against the fine-grained senses, the system got 35 instances wrong, but 11 of the "incorrect" instances were tagged with senses that were actually in the same group as the correct sense. This group of senses differs from others in the ability to take a small clause as a complement, which is modeled as a feature in our system. Here we see that the system benefits from using syntactic features that are linguistically richer than the features that have been used in the past.

29% of errors made by the tagger on *develop* were due to confusing Sense 1 and Sense 2, which are in the same group. The two senses describe transitive verbs that create new entities, characterized as either "products, or mental or artistic creations: CREATE (Sense 1)" or "a new theory of evolution: CREATE BY MENTAL ACT (Sense 2)." Instances of Sense 1 that were tagged as Sense 2 by the system included: *Researchers said they have developed a genetic engineering technique for creating hybrid plants for a number of key crops; William Gates and Paul Allen developed an early language-housekeeper system for PCs*. Conversely, the following instances of Sense 2 were tagged as Sense 1 by the tagger: *A Purdue University team hopes to develop ways to magnetically induce cardiac muscle contractions; Kobe Steel Ltd. adopted Soviet casting technology used it until it developed its own system.* Based on the direct object of *develop*, the automatic tagger was hard-pressed to differentiate between developing a *technique/system* (Sense 1) and developing a *way/system* (Sense 2).

Analysis of inter-annotator disagreement between two human annotators doing double-blind tagging revealed similar confusion between these two senses of *develop*; 25% of the human annotator disagreements on *develop* involved determining which of these two senses should be applied to phrases like *develop a better way to introduce crystallography techniques.* These instances that were difficult for the automatic WSD system, were also difficult for human annotators to differentiate consistently.

These different senses are clearly related, but the relation is not reflected in their hypernyms, which emphasize the differences in what is being highlighted by each sense, rather than the similarities. Methods of evaluation that automatically back off from synset to hypernyms (Lin, 1997) would fail to credit the system for "mistagging" an instance with a closely related sense. Manually created sense groups, on the other hand, can capture broader, more underspecified senses which are not explicitly listed and which do not participate in any of the WordNet semantic relations.

5 Conclusion

We have demonstrated that our approach to disambiguating verb senses using maximum entropy models to combine as many linguistic knowledge sources as possible, yields state-of-the-art performance for English. This may be a language-dependent feature, as other experiments indicate that additional linguistic pre-processing does not necessarily improve tagging accuracy for languages like Chinese (Dang et al., 2002).

In examining the instances that proved troublesome to both the human taggers and the automatic system, we found errors that were tied to subtle sense distinctions which were reconciled by backing off to the more coarse-grained sense groups. Achieving higher inter-annotator agreement is necessary in order to provide consistent training data

for supervised WSD systems. Lexicographers have long recognized that many natural occurrences of polysemous words are embedded in underspecified contexts and could correspond to more than one specific sense. Annotators need the option of selecting, as an alternative to an explicit sense, either a group of specific senses or a single, broader sense, where specific meaning nuances are subsumed. Sense grouping, already present in a limited way in WordNet's verb component, can be guided and enhanced by the analysis of inter-annotator disagreements and the development of explicit sense distinction criteria that such an analysis provides.

6 Acknowledgments

This work has been supported by National Science Foundation Grants, NSF-9800658 and NSF-9910603, and DARPA grant N66001-00-1-8915 at the University of Pennsylvania. The authors would also like to thank the anonymous reviewers for their valuable comments.

References

Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1).

Daniel M. Bikel, Scott Miller, Richard Schwartz, and Ralph Weischedel. 1997. Nymble: A high-performance learning name-finder. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, Washington, DC.

Martin Chodorow, Claudia Leacock, and George A. Miller. 2000. A topical/local classifier for word sense identification. *Computers and the Humanities*, 34(1-2), April. Special Issue on SENSEVAL.

Michael Collins. 1997. Three generative, lexicalised models for statistical parsing. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, Madrid, Spain, July.

Hoa Trang Dang, Ching yi Chia, Martha Palmer, and Fu-Dong Chiou. 2002. Simple features for chinese word sense disambiguation. In *Proceedings of Coling-02*, Taipei, Taiwain.

Christiane Fellbaum, Martha Palmer, Hoa Trang Dang, Lauren Delfs, and Susanne Wolf. 2001. Manual and automatic semantic annotation with WordNet. In *Proceedings of the Workshop on WordNet and Other Lexical Resources*, Pittsburgh, PA.

A. Kilgarriff and J. Rosenzweig. 2000. Framework and results for English SENSEVAL. *Computers and the Humanities*, 34(1-2), April. Special Issue on SENSEVAL.

Dekang Lin. 1997. Using syntactic dependency as local context to resolve word sense ambiguity. In *Proceedings of the 35th Annual Meeting of the ACL*, Madrid, Spain.

Rada Mihalcea and Dan I. Moldovan. 2001. Automatic generation of a coarse grained WordNet. In *Proceedings of the Workshop on WordNet and Other Lexical Resources*, Pittsburgh, PA.

Hwee Tou Ng and Hian Beng Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, Santa Cruz, CA, June.

M. Palmer, Chunghye Han, Fei Xia, Dania Egedi, and Joseph Rosenzweig. 2000. Constraining lexical selection across languages using tags. In Anne Abeille and Owen Rambow, editors, *Tree Adjoining Grammars: formal, computational and linguistic aspects*. CSLI, Palo Alto, CA.

Martha Palmer, Christiane Fellbaum, Scott Cotton, Lauren Delfs, and Hoa Trang Dang. 2001. English tasks: All-words and verb lexical sample. In *Proceedings of SENSEVAL-2: Second International Workshop on Evaluating Word Sense Disambiguation Systems*, Toulouse, France, July.

Ted Pedersen. 2000. A simple approach to building ensembles of naive bayesian classifiers for word sense disambiguation. In *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics*, Seattle, WA.

Ted Pedersen. 2001. A decision tree of bigrams is an accurate predictor of word sense. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics*, Pittsburgh, PA.

Adwait Ratnaparkhi. 1998. *Maximum Entropy Models for Natural Language Ambiguity Resolution*. Ph.D. thesis, University of Pennsylvania.

David Yarowsky. 2000. Hierarchical decision lists for word sense disambiguation. *Computers and the Humanities*, 34(1-2), April. Special Issue on SENSEVAL.