Towards Large-scale High-Performance English Verb Sense Disambiguation by Using Linguistically Motivated Features

Jinying Chen[1], Dmitriy Dligach[2], Martha Palmer[2,3]
[1] *BBN Technologies*
jchen@bbn.com
[2] *Department of Computer Science*
[3] *Department of Linguistics*
University of Colorado at Boulder
{Dmitriy.Dligach, Martha.Palmer}@colorado.edu

Abstract

In this paper we describe the results of training high performance Word Sense Disambiguation (WSD) systems on a new data set based on groupings of WordNet senses. This data set is designed to provide clear sense distinctions with sufficient examples in order to provide high quality training data. The sense distinctions are based on explicit syntactic and semantic criteria. Our WSD features utilize similar syntactic and semantic linguistic information. We demonstrate that this approach, using both Maximum Entropy and SVM models, produces systems whose performance is comparable to that of humans.

1. Introduction

Word sense disambiguation (WSD), determining the meaning a word bears in its given context, has been regarded as essential or necessary in many high-level NLP applications that require a certain degree of semantic interpretation, such as machine translation, information retrieval (IR) and question answering, *etc.* At the same time, high accuracy and broad coverage (disambiguation of a large vocabulary) are two crucial prerequisites for WSD to benefit NLP applications. Previous investigations into the role of WSD in IR have shown that low accuracy in WSD negated any possible performance increase from ambiguity resolution [1,2]. As evidenced by the SENSEVAL exercises[1], supervised WSD systems tend to perform better than unsupervised methods [3-6]. On the other hand, creating the necessary large-scale high-quality sense-tagged corpora is very difficult and time-consuming. In fact, many successful attempts at all-words WSD use unsupervised methods to compensate for the lack of training data[2].

Our current research efforts are aimed at high performance word sense disambiguation and include two major aspects: 1) developing a high-performance WSD system for English verbs by using linguistically motivated features and 2) applying this system to the first large-scale annotation effort aimed specifically at providing suitable training data for high-performance WSD. This effort involves annotating sufficient quantities of instances for English verbs that are linked to a sense inventory based on coarse-grained groupings of fine-grained WordNet senses [7].

We focus on verb sense disambiguation for two reasons. First, the problem caused by high polysemy is more serious for verbs, as indicated by the relatively poorer performance achieved by the best system in the SENSEVAL-2 English lexical sample task for verbs: 56.6% accuracy, in contrast with the 64.2% accuracy for all parts-of-speech [8,9]. Second, accurate verb sense disambiguation is very important, not only for selecting lexical items but also, in many cases, for generating correct and elegant syntactic structures for the target side in machine translation. It is also extremely useful for NLP applications based on deep-level natural language understanding, such as the question answering systems that take full-sentence queries as input or the information extraction systems that find global and/or domain-independent relational information.

Our previous work has shown that linguistically motivated features, such as syntactic and semantic features that capture the information about verb predicate-argument structures and the semantic categories of their noun phrase (NP) arguments, are very useful for verb sense disambiguation [9-12]. Our

[1] http://www.senseval.org/

[2] http://acl.ldc.upenn.edu/acl2004/senseval/index.html

0-7695-2997-6/07 $25.00 © 2007 IEEE
DOI 10.1109/ICSC.2007.69

supervised system based on these features achieved the best performance (64.6% accuracy) for highly polysemous (many multiple meanings) verbs (16.7 senses on average, based on WordNet 1.7 sense distinctions) in an evaluation using English SENSEVAL-2 data (the lexical sample task) [12].

However, 65%, or even 70% WSD accuracy is insufficient for NLP applications. Given that human inter-annotator (ITA) rates for WordNet senses tend to average just above 70%, it is unlikely that, even with vast amounts of training data, systems will be able to improve much on that score. Therefore we are participating in a large scale annotation effort that is based on grouping subtle, fine-grained WordNet senses into coherent semantic sense groups that can be readily distinguished by human annotators. This is part of the OntoNotes project, which also includes Treebanking [13], PropBanking [14], linking to the Omega Ontology [15] and coreference [16]. The goal of this project is to achieve ITA rates of 90%, in order to create training data that can support system performance in the 80+% range.

In this paper, we report our system performance on 217 verbs from the OntoNotes data [16]. Using a set of rich linguistic features we compare the performance of several machine learning algorithms. We find that our two highest performing systems are both very close to the ITA rates for this data, demonstrating that automatic WSD is comparable to human performance.

The rest of the paper is organized as follows. We give a detailed description of our WSD system for English verbs in Section 2 and introduce our current work in verb sense annotation in Section 3. In Section 4, we show the system performance on 217 verbs and discuss the experimental results. We conclude our work in Section 5.

2. A WSD System Using Linguistically Motivated Features

We developed our features for verb WSD using a system based on a smoothed maximum entropy (ME) model with a Gaussian prior [17]. An attractive property of ME models is that there is no assumption of feature independence [18]. Empirical studies have shown that an ME model with a Gaussian prior generally outperforms ME models with other smoothing methods [19].

We added several linguistically motivated features to the set of features associated with the successful WSD system of Dang [9,10] and modified several of Dang's original features. Before discussing our enhancements, we first briefly describe the basic syntactic and semantic features used by our system and Dang's:

Syntactic features:

- 1. Is the sentence passive, semi-passive[3] or active?
- 2. Does the target verb have a subject or object, and what is its head?
- 3. Does the target verb have a sentential complement?
- 4. Does the target verb have a PP adjunct? If so, what is the preposition and what is the head of the NP argument of the preposition?

Semantic features:

- 1. The Named Entity tags of proper nouns (***Person, Organization*** and ***Location***) and certain types of common nouns (***Time, Date, Money*** and ***Percent***)
- 2. The WordNet synsets and hypernyms of head nouns of the NP arguments of verbs and prepositions

In addition to these linguistically motivated features, our system also uses local collocation features (words and their POS's within a 5-word window centered by the target word) and topical features (open-class words in the two sentences preceding and following the sentence containing the target word).

To better explore the advantage of using rich syntactic and semantic features, we focused on three main enhancements: increasing the recall of the extraction of a verb's subject; unifying the treatment of semantic features of pronouns, common nouns and proper nouns; and providing a verb-specific treatment of sentential complements. These are each described in detail in [12] and we repeat the key content below for quick reference.

2.1 Increasing Subject Extraction Recall

To extract a subject, our original system simply checks the left NP siblings of the highest VP that contains the target verb and is within the innermost clause (see Figure 1). This method has high precision but low recall and cannot handle three common cases listed in (1).

[3] Verbs that are past participles and are not preceded by *be* or *have* verbs are semi-passive.

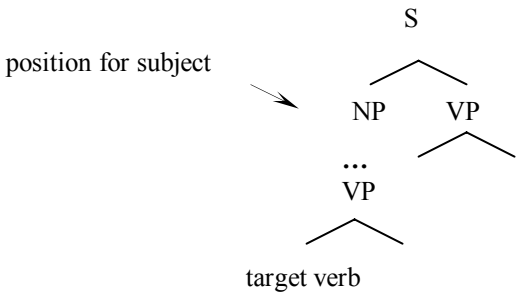

Figure 1. Subject – left NP sibling of highest VP

(1) a. **Relative clauses:** For Republicans$_{sbj}$ [$_{SBAR}$ who began$_{verb}$ this campaign with such high hopes], ...

b. **Nonfinite clauses:** I$_{sbj}$ didn't ever want [$_{S}$ to see$_{verb}$ that woman again].

c.**Verbs within PP's:** Karipo and her women$_{sbj}$ had succeeded [$_{PP}$ in driving$_{verb}$ a hundred invaders from the isle ...]

To increase the recall, we refined the procedure of subject extraction by adding rules based on linguistic knowledge and bracketing labels that can handle relative clauses, nonfinite clauses, and verbs within prepositional phrases (PP's) as illustrated in [12]. For example, if a clause containing the target verb has a bracketing label SBAR and an NP parent, and is headed by a relative pronoun such as *that*, *which* or *who*, then check its left NP siblings for the verb's subject.

Estimation based on SENSEVAL-2 English verb data showed that, with this enhancement, our new system extracts about 35% more subjects than before.

2.2 Unifying Semantic Features

In order to provide a more uniform treatment for the semantic features of the NP arguments of verbs and prepositions, we first merged the semantic features associated with proper nouns and common nouns. We then extended our treatment to include pronouns by adding a pronoun resolution module.

2.2.1 Merging Semantic Features

Our system used an automatic named entity tagger, *IdentiFinder*^{TM} [20], to tag proper nouns with ***Person, Organization*** and ***Location*** and common nouns with ***Date, Time, Percent*** and ***Money***. Additional semantic features are all WordNet synsets and hypernyms[4] of the head nouns of NP arguments, i.e., the system does not disambiguate different WordNet senses of a head noun.

Previously there was no overlap between semantic features generated by the named entity tagger and by WordNet. For example, a personal proper noun only has a ***Person*** tag that has no similarity to the WordNet synsets and hypernyms associated with similar common nouns such as *specialist* and *doctor*, *etc*. This is likely to be a problem for WSD tasks that usually have relatively small amounts of training data, such as SENSEVAL-2. To overcome this problem, our new system associates a common noun (or a noun phrase) with each Named Entity tag (see 2) and adds the WordNet semantic features of these nouns (or noun phrases) to the original semantic feature set.

(2) Person – someone
Organization – organization
Location – location
Time – time unit
Date – time period
Percent – percent
Money – money

2.2.2 Adding Pronoun Resolution

Our original system had no special treatment for pronouns, although a rough count showed that about half of the training instances contain pronominal arguments. Lacking a high performance automatic pronoun resolution module, we adopted a hybrid approach. For personal pronouns, we simply treated them as personal proper nouns. For the rest of the pronouns including *they, them, it, themselves* and *itself*, which occur in about 13% of the training instances, we programmed a rather simple rule-based pronoun resolver. In brief, the resolver searches the parse tree for antecedent candidates similarly to Hobb's algorithm as exemplified in [21] and uses several syntactic and semantic constraints to filter out impossible candidates. The constraints include syntactic constraints for anaphora antecedents [21], number agreement, and whether the candidate is a person. The first candidate that survives the filtering is regarded as the antecedent of the pronoun and its semantic features are added to the original feature set.

2.3 Verb-specific Sentential Complements

Different types of sentential complements can be very useful for distinguishing certain verb senses. For

[4] A unique number defined in WordNet represents each synset or hypernym.

example, (3a-b) show two sentences containing the verb *call* in the SENSEVAL-2 training data. *Call* has WordNet Sense 1 (name) in (3a) and Sense 3 (ascribe) in (3b). In both cases, *call* takes a small clause as its sentential complement, i.e., it has the subcategorization frame X *call* Y Z. The difference is that Z is a Named Entity when *call* is in Sense 1, and Z is usually a common NP or an adjective phrase (ADJP) when *call* is in Sense 3.

(3) a. The slender, handsome fellow was $\underline{\text{called}}_{\text{verb}}$ [$_S$ Dandy Brandon].

b.The White House is purposely not $\underline{\text{calling}}_{\text{verb}}$ [$_S$ the meeting a summit] …

Our original system used a single feature *hasSent* to represent whether the target verb has a sentential complement or not, which cannot capture the rich information that is crucial to distinguishing certain verb senses but is deeply embedded in the sentential complements, as described above. Therefore, we treat sentential complements in a more fine-grained, verb-specific way. We resort to WordNet and PropBank for the information about verb subcategorization frames. Another advantage of this verb-specific treatment is that it can filter out illegal sentential complements generated by the parser.

3. Creating a Sense Inventory that Supports High Quality Annotation

As discussed above, the annotated corpus we are using was developed with the goal of producing high quality supervised training data. A large portion of the Penn Treebank II Wall Street Journal text is being annotated, as well as data from the Brown Corpus and the English/Chinese parallel Treebank. There are two distinct steps involved: sense grouping and annotation. In the grouping process fine-grained sense distinctions listed in WordNet 2.1 are collected into more coarse-grained groupings based on syntactic and semantic criteria, following standard lexicographic practices. For instance, for the verb *call*, Sense1: *I **called** my son David*, and Sense 12: *You can **call** me Sir* are grouped together. Other resources, including PropBank, VerbNet, (based on Levin's verb classes [22]) and online dictionaries are consulted for insights into syntactic and semantic similarities [23-25]. As an aid to annotators, sense groupings are ordered according to saliency and frequency. Detailed comments about distinctions between the groups, including syntactic frames and semantic features as discussed below, are explicitly provided for each group. Several example sentences from WordNet or Google search results are included for further clarification of the sense grouping.

Syntactic criteria: Annotators have found syntactic frames, such as those defining VerbNet classes, to be useful in clarifying boundaries between sense groupings. For example, verbs such as *split* in the "break-45.1" class which participate in the causative/inchoative alternation usually have those usages grouped together: *John split the log/ The log split.*

Semantic criteria: Similar semantic features of verb arguments, such as [+/-attribute], [+/-animate], and [+/-locative] are also frequently used to group senses together. For a more detailed discussion of grouping criteria, see [26]

The annotation process begins with fifty sample sentences being given double blind annotation, and if an ITA rate of 90% or above is achieved, the verb entry is considered complete. The rest of the instances are then given double blind annotation and adjudication. Groupings that receive less than 90% ITA scores are re-grouped and re-annotated. It is sometimes impossible to get ITA scores over 85% for high frequency verbs that also have high polysemy and high entropy. These have to be carefully adjudicated to produce a gold standard. The final versions of the sense groupings are mapped to VerbNet and FrameNet and linked to the Omega Ontology [15].

Verbs are selected based on frequency of appearance in the WSJ corpus. The 740 most frequent verbs were grouped first. They have an average polysemy of 7 senses in WordNet which is reduced to 3.75 by grouping. The 217 verbs used here, which have fairly high frequency, have an average WordNet polysemy of 10.4 which reduces to 5.1. The WordNet senses of these verbs range from 59 to 2 senses per verb, and the groups range from 16 to 2. In addition to reducing polysemy, the clear, explicit criteria for sense distinctions improve annotator productivity up to three-fold [25].

4. Experimental Results

We evaluated our system's performance on the verbs that had at least 50 instances annotated and adjudicated in the OntoNotes project, which amounted to a total of 217 verbs and 35,210 instances. We preprocessed the resulting corpus using Ratnaparkhi's ME sentence boundary detector and POS tagger [27], Bikel's parsing engine [28], and a named entity tagger *IdentiFinder* [20]. After that, we ran our feature extraction module to derive features for each of these instances as described in the previous section.

All of our approaches rely on supervised learning. In the evaluation, we used two machine learning software packages: Mallet [17] and WEKA [29]. The former was used to build a smoothed ME model with a Gaussian prior; the latter was used to build a linear Support Vector Machine (SVM) model (and also the other classifiers mentioned below). The smoothed ME model has been shown to be successful in supervised learning of verb senses [12]. We chose to experiment with SVMs because this machine learning approach has been proven successful in many NLP tasks and we are interested in comparing its performance in WSD with that of the smoothed ME model.[5]

Because the corpus used in the experiments has only recently been created by our team, and thus is new to the WSD community, no other WSD systems have been evaluated against it in the past. Therefore, we compared the performance of our system against the most frequent sense baseline in which all instances were labeled with the most frequent sense of the verb. In the process of annotating our corpus, we collected the ITA (inter-annotator agreement) rates, which reflected the percentage of instances where both annotators agreed in their choice of senses. Because a machine learning system rarely exceeds the performance of a human annotator, ITA can be viewed as a natural way to compare performance of an automatic WSD system to human tagging.

To give the reader an idea of how long the training and the testing phases take for a single verb, we use the verb "require" as an example (128 training instances and 32 test instances). Table 1 shows the time to build and test a smoothed ME model on a machine with 2 1.1GHz CPUs and 2GB RAM and the time to build and test a single linear SVM model on a machine equipped with a 3GHz CPU and 1GB RAM.

ME		SVM	
Train Time	Test Time	Train Time	Test Time
2.04	1.4	3.03	0.23

Table 1 Training and testing time in seconds for the ME model and the linear SVM model for the verb "require"

In our experiments, a separate model was built for each verb; five-fold cross validation was used for testing. Table 2 shows the experimental results. In addition to giving the five-fold cross-validation accuracy for the smoothed ME model and the linear SVM model (columns 6 and 7), the baseline accuracy (column 5) and the ITA (column 8), it also provides the number of grouped senses (column 2), the number of instances (column 4), and sense entropy (column 3). Due to space limitations, Table 2 only shows results for the first 5 and the last 5 verbs (in alphabetic order) used in our experiments. The last row gives the average numbers (weighted by the number of instances) for all the 217 verbs.

As we can see from the table, both classifiers were able to beat the most frequent sense baseline by a wide margin of at least 14 percentage points. The smoothed ME model (average accuracy – 0.8272) is slightly better than the SVM model (average accuracy – 0.8220) (p=0.05). The ME model's performance is slightly above the ITA rate and the SVM's slightly below the ITA. However, both differences are not statistically significant. The fact that there is no statistical significance between the performances of the two classification algorithms and the ITA can be interpreted to mean that the performance of the supervised WSD systems is comparable to human tagging.

By looking at the results in detail, we find that there are 9 verbs that had system accuracies lower than their baselines. 4 of these verbs have very high baselines (above 89%) and the difference between their baselines and system accuracies is lower than 0.01. The others tend to have small sets of training data (50 instances on average).

In a separate experiment, we tested more machine learning models on the 217 verbs by using the same set of features as used in our first experiment. These learning models include: K-nearest Neighbor (KNN) models, Naïve Bayes models, Decision Tree models, AdaBoost with Naïve Bayes as a base classifier, and AdaBoost with decision tree models as base classifiers. AdaBoost works by boosting the output of weak known to improve the performance of weak classifiers, we chose to experiment with AdaBoost in conjunction with naïve Bayes models or decision tree models as base classifiers. The additional motivation for experimenting with AdaBoost is that AdaBoost is known to be less susceptible to overfitting than most learning algorithms. Again, a separate model was built for each verb and five-fold cross validation was used for testing in this experiment.

Table 3 shows the experimental results. To save space, we only report average accuracies for each model. The system accuracies of the smoothed ME model and the linear SVM model, which were obtained in the first experiment, are also included for a comparison. As we can see, the smoothed ME model is the best one, followed by the linear SVM model and the AdaBoost with decision tree models as base classifiers.

[5] In our experiments, we used default parameters provided by the software packages for both models based on our previous experience.

Verb	Polysemy	Sense Entropy	# of Instances	Baseline Accuracy	ME Accuracy	Linear SVM Accuracy	ITA
Accept	5	1.2318	121	0.5207	0.6033	0.5620	0.7661
Account	3	0.3622	119	0.8824	0.9496	0.9412	0.8992
Acquire	4	0.4492	81	0.8765	0.8765	0.8765	0.8235
Act	3	0.7401	86	0.5581	0.7791	0.7907	0.8966
Add	2	0.2107	203	0.9458	0.9557	0.9606	0.7949
…							
Wish	3	0.8799	106	0.5377	0.8962	0.8585	0.9018
Work	9	1.1046	318	0.6478	0.7704	0.7642	0.7933
Worry	3	0.4397	75	0.8400	0.9467	0.9467	0.7426
Write	9	0.4859	268	0.8769	0.8806	0.8955	0.8864
Yield	3	0.4275	72	0.8472	0.9722	0.9583	0.9167
Average	5.1	0.8328		0.6803	0.8272	0.8220	0.8253

Table 2 Performance of our WSD system with ME and a linear SVM model on 217 OntoNotes verbs

Baseline accuracy	KNN	Naïve Bayes	Decision Tree (J48)	AdaBoost w/ Naïve Bayes	AdaBoost w/Decision Tree (J48)	ME	Linear SVM
0.6803	0.6986	0.7396	0.7744	0.7586	0.8038	0.8271	0.8220

Table 3 Performance of different machine learning models on 217 OntoNotes verbs

4.1 Discussion

Table 4 compares the data (our system[6] accuracy, ITA and the most frequent sense baseline) in our current experiments to those in our previous experiments with SENSEVAL-2 verbs [12].

Data set	Baseline Acc.	System Acc.	ITA
SENSEVAL-2 verbs	0.407	0.646	0.713
OntoNotes verbs	0.680	0.827	0.825

Table 4 Baseline Accuracy, System Accuracy and ITA on Two Data Sets with Different Sense Granularity

As we can see from Table 4, the system performance improves by 18 percent (absolute gain) when using coarse-grained senses. This performance gain is lower than that for the baseline accuracy (27 percent) and higher than that for the ITA (9 percent). This result implies that less complicated learning methods (e.g., the most frequent sense heuristic) for WSD could benefit more from adopting more coarse-grained and therefore clearer sense distinctions. However, the fact that our system performance is comparable to that of humans is still very impressive.

It is also worth mentioning that the sense groups are significantly more fine-grained than PropBank and map readily to VerbNet/FrameNet. So, they are still preserving important sense distinctions and are intended to provide an appropriate level for making semantic generalizations [30].

Navigli [31] reported that the accuracy of the best system in SENSEVAL-3 English all-words task improves by 12 percent (absolute gain) by using coarse-grained senses produced through mapping WordNet senses to sense hierarchies of the Oxford Dictionary of English. In our experiments, we focused on English verbs and used different methods for grouping WordNet senses. Despite these differences, both results are consistent and indicate that the accuracies of WSD systems benefit significantly from well-defined coarse-grained sense distinctions. Two coordinated Semeval tasks compare the coarse-grained choices of these approaches [32, 33]

To investigate how much linguistically motivated features contributed to the high accuracy our system

[6] Our system, mentioned here and later, refers to the system that uses the smoothed MaxEnt learning model.

achieved on the OntoNotes verb data used in our experiments, we trained and tested our system by using three different feature sets: all the features (ALL) without semantic features (w/o SEM) and without semantic and syntactic features (w/o SEM+SYN). The semantic and syntactic features were listed in Section 2. Table 5 gives the results. The differences between accuracies are all significant ($p<0.0001$).

ALL	w/o SEM.	w/o SEM+SYN
0.827	0.816	0.789

Table 5 System Accuracy with Different Feature Sets

5. Conclusions and Future Work

This paper presents our efforts at achieving high-performance verb sense disambiguation based on a large-scale annotation effort. Our supervised WSD system uses linguistically motivated features and a smoothed ME model for machine learning. Many of these features have also been used by other successful supervised systems [34, 35]. We enhanced our system's treatment of these features in three specific ways by using linguistic knowledge and rules and automatic pronoun resolution, all of which improved performance [12]. Supervised WSD tasks generally suffer from an insufficiency of sense-tagged training data. Very fine-grained sense distinctions often cause low inter-annotator agreement and slow down the annotation process of large-scale corpora. We address this problem by grouping fine-grained WordNet senses into sense groups with clearly explicated sense distinctions and annotating large numbers of instances with these coarse-grained grouped senses. Following standard lexicographic practice, the groupings are defined by using syntactic and semantic criteria. The automatic system features have been explicitly chosen to capture the same types of information [23]. We evaluated our system performance on 217 verbs annotated by this approach and compared its performance with another well-known machine learning model, the linear SVM model, by using the same feature set. The experimental results show that our system achieved a very high accuracy of 82.71% on these verbs, which is close to ITA and slightly better than the SVM model. In a separate experiment, we tested a set of machine learning models by using the same data set. The smoothed ME model and the linear SVM model were shown to be the two best models.

In future work, we will apply SVMs with non-linear kernels to this task. SVMs with non-linear kernels can map feature vectors to a higher dimensional space and make originally nonlinearly separable data linearly separable in the new feature space. Therefore, they are more powerful than linear SVMs in terms of classification capability and are expected to achieve even better performance. However, we will need sufficient time for parameter tuning to achieve the best performance.

Our preliminary work [36] suggests that active learning works well for learning coarse-grained verb senses. In the future, we will experiment with applying active learning to improve the productivity of our annotation effort. We will also be reexamining the verbs with low ITA rates (16 verbs with ITA < 65%) to see if their sense distinctions can be clarified.

Acknowledgements

We gratefully acknowledge the support of the National Science Foundation Grant NSF-0415923, Word Sense Disambiguation, and the GALE program of the Defense Advanced Research Projects Agency, Contract No. HR0011-06-C-0022, a subcontract from the BBN-AGILE Team.

6. References

[1] Mark Sanderson. 1994. Word sense disambiguation and information retrieval. In *Proceedings of the 17th Int. ACM SIGIR*, Dublin, IE.

[2] Christopher Stokoe, Michael P. Oakes, John Tait. 2003. Word sense disambiguation and information retrieval revisited. In *Proceedings of the 26th annual int. ACM SIGIR conference on research and development in information retrieval*. Toronto, Canada.

[3] Philip Edmonds and Scott Cotton. 2001. SENSEVAL-2: Overview. In *Proceedings of SENSEVAL-2: 2nd Int. Workshop on Evaluating WSD Systems*. ACL-SIGLEX, Toulouse, France.

[4] 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 5-6.

[5] Rada Mihalcea, Timothy Chklovski and Adam Kilgarriff. 2004. The Senseval-3 English lexical sample task. In *Proceedings of Senseval-3: The Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*. Barcelona, Spain. July.

[6] Benjamin Snyder and Martha Palmer. 2004. The English all-words task. In *Proceedings of Senseval-3: The Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*. Barcelona, Spain. July.

[7] Christiane Fellbaum. 1998. *WordNet - an Electronic Lexical Database*. The MIT Press, Cambridge, Massachusetts, London, UK.

[8] David Yarowsky, Silviu Cucerzan, Radu Florian, Charles Schafer and Richard Wicentowski. 2001. The Johns hopkins SENSEVAL2 system description. In *Proceedings of SENSEVAL-2: 2nd Int.Workshop on Evaluating WSD Systems*. Toulouse France.

[9] Hoa T. Dang and Martha Palmer. 2002. Combining contextual features for word sense disambiguation. In *Proceedings of the SIGLEX/SENSEVAL Workshop on WSD: Recent Successes and Future Directions*, in conjunction with ACL-02, Philadelphia.

[10] Hoa T. Dang and Martha Palmer. 2005. The Role of Semantic Roles in Disambiguating Verb Senses. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, Ann Arbor MI, June 26-28.

[11] Jinying Chen. 2006. *Towards High-performance Word Sense Disambiguation by Combining Rich Linguistic Knowledge and Machine Learning Approaches*. PhD Thesis. University of Pennsylvania.

[12] Jinying Chen and Martha Palmer. 2005. Towards Robust High Performance Word Sense Disambiguation of English Verbs Using Rich Linguistic Features. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing*. October 11-13, 2005, Jeju Island, Korea.

[13] Mitchell Marcus, Grace Kim, Mary A. Marcinkiewicz, Robert MacIntyre, Mark Ferguson, Karen Katz and Britta Schasberger. 1994. The Penn Treebank: annotating predicate argument structure. In *Proceedings of the ARPA'94 HLT Workshop*.

[14] Martha Palmer, Dan Gildea and Paul Kingsbury. The Proposition Bank: A Corpus Annotated with Semantic Roles, *Computational Linguistics*, 31:1, 2005.

[15] Andrew Philpot, Eduard Hovy and Patrick Pantel. 2005. The Omega Ontology. In *Proceedings of the ONTOLEX Workshop at the International Conference on Natural Language Processing (IJCNLP05).* Jeju Island, Korea**.**

[16] Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw and Ralph Weischedel. OntoNotes: The 90% Solution. In *Proceedings of HLT-NAACL06*. New York, 2006.

[17] Andrew K. McCallum. 2002. MALLET: A Machine Learning for Language Toolkit. http://www.cs. umass.edu/~mccallum/mallet.

[18] Adam L. Berger, Stephen A. Della Piertra, and Vincent J. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1): 39-71.

[19] Stanley. F. Chen and Ronald Rosenfeld. 1999. *A Gaussian prior for smoothing maximum entropy models.* Technical Report CMU-CS-99-108, CMU.

[20] Daniel M. Bikel, Richard Schwartz and Ralph M. Weischedel. 1999. An algorithm that learns what's in a name. *Machine Learning, 34(1-3). Special Issue on Natural Language Learning*.

[21] Shalom Lappin and Herbert Leass. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4): 535-561.

[22] Beth Levin. 1993. *English Verb Classes And Alternations: A Preliminary Investigation*. Chicago: University of Chicago Press.

[23] Palmer, M., Dang, H.T., and Fellbaum, C., Making Fine-grained and Coarse-grained sense distinctions, both manually and automatically, *Journal of Natural Language Engineering* (to appear, 2007) doi: 10.1017/ S135132490500402X Published Online: 12Jul2006

[24] Kipper, K., A. Korhonen, N. Ryant, and M. Palmer. (2006). Extensive Classifications of English Verbs. *Proceedings of the 12th EURALEX International Congress.* Turin, Italy.

[25] Palmer, M., O. Babko-Malaya, and H.T. Dang. 2004. Different Sense Granularities for Different Applications. *Proceedings of the 2nd Workshop on Scalable Natural Language Understanding Systems (HLT/NAACL 2004).* Boston, MA

[26] Duffield, C. J., Hwang, J. D., Brown, Ss. W., Dligach, D..Vieweg, S. E., Davis, J. L., Palmer, M. S., Criteria for the Manual Grouping of Verb Senses, *Linguistics Annotation Workshop, held in conjunction with ACL-2007*, Prague, The Czech Republic. 2007.

[27] Adwait Ratnaparkhi. 1998. *Maximum entropy models for natural language ambiguity resolution*. Ph.D. thesis, University of Pennsylvania.

[28] Daniel M. Bikel. 2002. Design of a multi-lingual, parallel-processing statistical parsing engine. In *Proceedings of HLT 2002*. San Diego, CA.

[29] Ian H. Witten and Eibe Frank (2005) "*Data Mining: Practical machine learning tools and techniques*", 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

[30] Szu-ting Yi; Edward Loper; Martha Palmer Can Semantic Roles Generalize Across Genres? *In the Proceedings of NAACL 2007, Rochester, NY, April 2007.*

[31] Roberto Navigli, Meaningful clustering of senses helps boost word sense disambiguation performance, Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL, p.105-112, July 17-18, 2006, Sydney, Australia.

[32] Sameer Pradhan; Edward Loper; Dmitriy Dligach; Martha Palmer, SemEval-2007 Task-17: English Lexical Sample, SRL and All Words, *In the Proceedings of SemEval,* held in conjunction with ACL 2007, Prague, Czech Republic, June, 2007

[33] Roberto Navigli; Kenneth C. Litkowski; Orin Hargraves SemEval-2007 Task 07: Coarse-Grained English All-Words Task*, In the Proceedings of SemEval,* held in conjunction with ACL 2007, Prague, Czech Republic, June, 2007

[34] Lee, Yoong Keok, Ng, Hwee Tou, & Chia, Tee Kiah (2004). Supervised Word Sense Disambiguation with Support Vector Machines and Multiple Knowledge Sources. In *Proceedings of SENSEVAL-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*. (pp. 137-140). Barcelona, Spain.

[35] Agirre E., Phil Edmonds 2006 Word Sense Disambiguation: Algorithms and applications. *Text, Speech and Language Technology Series*, Vol. 33. Springer. ISBN: 1-4020-4808-4

[36] Jinying Chen, Andrew Schein, Lyle Ungar and Martha Palmer. 2006. An Empirical Study of the Behavior of Word Sense Disambiguation. In *Proceedings of NAACL-HLT 2006*. NY, 2006.