SemEval-2007 Task 17: English Lexical Sample, SRL and All Words

Sameer S. Pradhan
BBN Technologies,
Cambridge, MA 02138

Edward Loper
University of Pennsylvania,
Philadelphia, PA 19104

Dmitriy Dligach and Martha Palmer
University of Colorado,
Boulder, CO 80303

Abstract

This paper describes our experience in preparing the data and evaluating the results for three subtasks of SemEval-2007 Task-17 – Lexical Sample, Semantic Role Labeling (SRL) and All-Words respectively. We tabulate and analyze the results of participating systems.

1 Introduction

Correctly disambiguating words (WSD), and correctly identifying the semantic relationships between those words (SRL), is an important step for building successful natural language processing applications, such as text summarization, question answering, and machine translation. SemEval-2007 Task-17 (*English Lexical Sample, SRL and All-Words*) focuses on both of these challenges, WSD and SRL, using annotated English text taken from the Wall Street Journal and the Brown Corpus. It includes three subtasks: i) the traditional All-Words task comprising fine-grained word sense disambiguation using a 3,500 word section of the Wall Street Journal, annotated with WordNet 2.1 sense tags, ii) a Lexical Sample task for coarse-grained word sense disambiguation on a selected set of lexemes, and iii) Semantic Role Labeling, using two different types of arguments, on the same subset of lexemes.

2 Word Sense Disambiguation

2.1 English fine-grained All-Words

In this task we measure the ability of systems to identify the correct fine-grained WordNet 2.1 word sense for all the verbs and head words of their arguments.

2.1.1 Data Preparation

We began by selecting three articles `wsj_0105.mrg` (on homelessness), `wsj_0186.mrg` (about a book on corruption), and `wsj_0239.mrg` (about hot-air ballooning) from a section of the WSJ corpus that has been Treebanked and PropBanked. All instances of verbs were identified using the Treebank part-of-speech tags, and also the headwords of their noun arguments (using the PropBank and standard headword rules). The locations of the sentences containing them as well as the locations of the verbs and the nouns within these sentences were recorded for subsequent sense-annotation. A total of 465 lemmas were selected from about 3500 words of text.

We use a tool called `STAMP` written by Benjamin Snyder for sense-annotation of these instances. `STAMP` accepts a list of pointers to the instances that need to be annotated. These pointers consist of the name of the file where the instance is located, the sentence number of the instance, and finally, the word number of the ambiguous word within that sentence. These pointers were obtained as described in the previous paragraph. `STAMP` also requires a sense inventory, which must be stored in XML format. This sense inventory was obtained by querying WordNet 2.1 and storing the output as a

Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pages 87–92,
Prague, June 2007. ©2007 Association for Computational Linguistics

set of XML files (one for each word to be annotated) prior to tagging. STAMP works by displaying to the user the sentence to be annotated with the target word highlighted along with the previous and the following sentences and the senses from the sense inventory. The user can select one of the senses and move on to the next instance.

Two linguistics students annotated the words with WordNet 2.1 senses. Our annotators examined each instance upon which they disagreed and resolved their disagreements. Finally, we converted the resulting data to the Senseval format. For this dataset, we got an inter-annotator agreement (ITA) of 72% on verbs and 86% for nouns.

2.1.2 Results

A total of 14 systems were evaluated on the All Words task. These results are shown in Table 1. We used the standard Senseval scorer – scorer2[1] to score the systems. All the F-scores[2] in this table as well as other tables in this paper are accompanied by a 95% confidence interval calculated using the bootstrap resampling procedure.

2.2 OntoNotes English Lexical Sample WSD

It is quite well accepted at this point that it is difficult to achieve high inter-annotator agreement on the fine-grained WordNet style senses, and without a corpus with high annotator agreement, automatic learning methods cannot perform at a level that would be acceptable for a downstream application. OntoNotes (Hovy et al., 2006) is a project that has annotated several layers of semantic information – including word senses, at a high inter-annotator agreement of over 90%. Therefore we decided to use this data for the lexical sample task.

2.2.1 Data

All the data for this task comes from the 1M word WSJ Treebank. For the convenience of the participants who wanted to use syntactic parse information as features using an off-the-shelf syntactic parser, we decided to compose the training data of Sections 02-21. For the test sets, we use data from Sections 01, 22, 23 and 24. Fortunately, the distribution of words was amenable to an acceptable number of instances for each lemma in the test set. We selected a total of 100 lemmas (65 verbs and 35 nouns) considering the degree of polysemy and total instances that were annotated. The average ITA for these is over 90%.

	Train	Test	Total
Verb	8988	2292	11280
Noun	13293	2559	15852
Total	22281	4851	

Table 2: The number of instances for Verbs and Nouns in the Train and Test sets for the Lexical Sample WSD task.

The training and test set composition is described in Table 2. The distribution across all the verbs and nouns is displayed in Table 4

2.2.2 Results

A total of 13 systems were evaluated on the Lexical Sample task. Table 3 shows the Precision/Recall for all these systems. The same scoring software was used to score this task as well.

2.2.3 Discussion

For the all words task, the baseline performance using the most frequent WordNet sense for the lemmas is 51.4. The top-performing system was a supervised system that used a Maximum Entropy classifier, and got a Precision/Recall of 59.1% – about 8 points higher than the baseline. Since the coarse and fine-grained disambiguation tasks have been part of the two previous Senseval competitions, and we happen to have access to that data, we can take this opportunity to look at the disambiguation performance trend. Although different test sets were used for every evaluation, we can get a rough indication of the trend. For the fine-grained All Words sense tagging task, which has always used WordNet, the system performance has ranged from our 59% to 65.2 (Senseval3, (Decadt et al., 2004)) to 69% (Seneval2, (Chklovski and Mihalcea, 2002)). Because of time constraints on the data preparation, this year's task has proportionally more verbs and fewer nouns than previous All-Words English tasks, which may account for the lower scores.

As expected, the Lexical Sample task using coarse

[1] http://www.cse.unt.edu/~rada/senseval/senseval3/scoring/

[2] scorer2 reports Precision and Recall scores for each system. For a system that attempts all the words, both Precision and Recall are the same. Since a few systems had missing answers, they got different Precision and Recall scores. Therefore, for ranking purposes, we consolidated them into an F-score.

Rank	Participant	System ID	Classifier	F
1	Stephen Tratz <stephen.tratz@pnl.gov>	PNNL	MaxEnt	59.1±4.5
2	Hwee Tou Ng <nght@comp.nus.edu.sg>	NUS-PT	SVM	58.7±4.5
3	Rada Mihalcea <rada@cs.unt.edu>	UNT-Yahoo	Memory-based	58.3±4.5
4	Cai Junfu <caijunfu@gmail.com>	NUS-ML	naive Bayes	57.6±4.5
5	Oier Lopez de Lacalle <jibloleo@si.ehu.es>	UBC-ALM	kNN	54.4±4.5
6	David Martinez <davidm@csse.unimelb.edu.au>	UBC-UMB-2	kNN	54.0±4.5
7	Jonathan Chang <jcone@princeton.edu>	PU-BCD	Exponential Model	53.9±4.5
8	Radu ION <radu@racai.ro>	RACAI	Unsupervised	52.7±4.5
9	*Most Frequent WordNet Sense*	Baseline	N/A	51.4±4.5
10	Davide Buscaldi <dbuscaldi@dsic.upv.es>	UPV-WSD	Unsupervised	46.9±4.5
11	Sudip Kumar Naskar <sudip.naskar@gmail.com>	JU-SKNSB	Unsupervised	40.2±4.5
12	David Martinez <davidm@csse.unimelb.edu.au>	UBC-UMB-1	Unsupervised	39.9±4.5
14	Rafael Berlanga <berlanga@uji.es>	tkb-uo	Unsupervised	32.5±4.5
15	Jordan Boyd-Graber <jbg@princeton.edu>	PUTOP	Unsupervised	13.2±4.5

Table 1: System Performance for the All-Words task.

Rank	Participant	System	Classifier	F
1	Cai Junfu <caijunfu@gmail.com>	NUS-ML	SVM	88.7±1.2
2	Oier Lopez de Lacalle <jibloleo@si.ehu.es>	UBC-ALM	SVD+kNN	86.9±1.2
3	Zheng-Yu Niu <niu_zy@hotmail.com>	I2R	Supervised	86.4±1.2
4	Lucia Specia <lspecia@gmail.com>	USP-IBM-2	SVM	85.7±1.2
5	Lucia Specia <lspecia@gmail.com>	USP-IBM-1	ILP	85.1±1.2
5	Deniz Yuret <dyuret@ku.edu.tr>	KU	Semi-supervised	85.1±1.2
6	Saarikoski <harri.saarikoski@helsinki.fi>	OE	naive Bayes, SVM	83.8±1.2
7	University of Technology Brno	VUTBR	naive Bayes	80.3±1.2
8	Ana Zelaia <ana.zelaia@ehu.es>	UBC-ZAS	SVD+kNN	79.9±1.2
9	Carlo Strapparava <strappa@itc.it>	ITC-irst	SVM	79.6±1.2
10	*Most frequent sense in training*	Baseline	N/A	78.0±1.2
11	Toby Hawker <toby@it.usyd.edu.au>	USYD	SVM	74.3±1.2
12	Siddharth Patwardhan <sidd@cs.utah.edu>	UMND1	Unsupervised	53.8±1.2
13	Saif Mohammad <smm@cs.toronto.edu>	Tor	Unsupervised	52.1±1.2
-	Toby Hawker <toby@it.usyd.edu.au>	USYD*	SVM	89.1±1.2
-	Carlo Strapparava <strappa@itc.it>	ITC*	SVM	89.1±1.2

Table 3: System Performance for the OntoNotes Lexical Sample task. Systems marked with an * were post-competition bug-fix submissions.

grained senses provides consistently higher performance than previous more fine-grained Lexical Sample Tasks. The high scores here were foreshadowed in an evaluation involving a subset of the data last summer (Chen et al., 2006). Note that the best system performance is now closely approaching the ITA for this data of over 90%. Table 4 shows the performance of the top 8 systems on all the individual verbs and nouns in the test set. Owing to space constraints we have removed some lemmas that have perfect or almost perfect accuracies. At the right are mentioned the average, minimum and maximum performances of the teams per lemma, and at the bottom are the average scores per lemma (without considering the lemma frequencies) and broken down by verbs and nouns. A gap of about 10 points between the verb and noun performance seems to indicate that in general the verbs were more difficult than the nouns. However, this might just be owing to this particular test sample having more verbs with higher perplexities, and maybe even ones that are indeed difficult to disambiguate – in spite of high human agreement. The hope is that better knowledge sources can overcome the gap still existing between the system performance and human agreement. Overall, however, this data indicates that the approach suggested by (Palmer, 2000) and that is being adopted in the ongoing OntoNotes project (Hovy et al., 2006) does result in higher system performance. Whether or not the more coarse-grained senses are effective in improving natural language processing applications remains to be seen.

Lemma	S	s	T	t	1	2	3	4	5	6	7	8	Average	Min	Max
turn.v	13	8	340	62	58	**61**	40	55	52	53	27	44	49	27	61
go.v	12	6	244	61	64	**69**	38	66	43	46	31	39	49	31	69
come.v	10	9	186	43	49	46	56	**60**	37	23	23	49	43	23	60
set.v	9	5	174	42	**62**	50	52	57	50	57	36	50	52	36	62
hold.v	8	7	129	24	58	46	50	54	54	38	50	**67**	52	38	67
raise.v	7	6	147	34	**50**	44	29	26	44	26	24	12	32	12	50
work.v	7	5	230	43	**74**	65	65	65	72	67	46	65	65	46	74
keep.v	7	6	260	80	56	54	52	**64**	56	52	48	51	54	48	64
start.v	6	4	214	38	53	50	47	**55**	45	42	37	45	47	37	55
lead.v	6	6	165	39	69	69	**85**	69	51	69	36	46	62	36	85
see.v	6	5	158	54	56	54	46	54	**57**	52	48	48	52	46	57
ask.v	6	3	348	58	**84**	72	72	78	76	52	67	66	71	52	84
fi nd.v	5	3	174	28	**93**	**93**	86	89	82	82	75	86	86	75	93
fi x.v	5	3	32	2	**50**	**50**	**50**	**50**	**50**	0	0	**50**	38	0	50
buy.v	5	3	164	46	**83**	80	80	**83**	78	76	70	76	78	70	83
begin.v	4	2	114	48	**83**	65	75	69	79	56	50	56	67	50	83
kill.v	4	1	111	16	**88**	**88**	**88**	**88**	**88**	**88**	**88**	81	87	81	88
join.v	4	4	68	18	44	50	50	39	56	**57**	39	44	47	39	57
end.v	4	3	135	21	**90**	86	86	**90**	62	87	86	67	82	62	90
do.v	4	2	207	61	92	90	90	**93**	**93**	90	85	84	90	84	93
examine.v	3	2	26	3	**100**	**100**	67	**100**	**100**	67	**100**	33	83	33	100
report.v	3	2	128	35	89	**91**	**91**	**91**	**91**	**91**	**91**	86	90	86	91
regard.v	3	3	40	14	**93**	**93**	86	86	64	86	57	**93**	82	57	93
recall.v	3	1	49	15	**100**	**100**	87	87	93	87	87	87	91	87	100
prove.v	3	2	49	22	**90**	88	82	80	**90**	86	70	74	82	70	90
claim.v	3	2	54	15	67	73	80	80	80	80	80	**87**	78	67	87
build.v	3	3	119	46	**74**	67	**74**	61	54	**74**	61	72	67	54	74
feel.v	3	3	347	51	71	69	69	74	**76**	69	61	71	70	61	76
care.v	3	3	69	7	43	43	43	43	**100**	29	57	57	52	29	100
contribute.v	2	2	35	18	67	**72**	**72**	67	50	61	50	67	63	50	72
maintain.v	2	2	61	10	80	80	70	**100**	80	90	90	80	84	70	100
complain.v	2	1	32	14	**93**	86	86	86	86	86	86	79	86	79	93
propose.v	2	2	34	14	**100**	86	**100**	86	**100**	93	79	79	90	79	100
promise.v	2	2	50	8	**88**	**88**	75	**88**	75	75	62	**88**	80	62	88
produce.v	2	2	115	44	**82**	**82**	77	73	75	75	77	80	78	73	82
prepare.v	2	2	54	18	**94**	83	89	89	83	86	83	83	86	83	94
explain.v	2	2	85	18	**94**	89	**94**	89	**94**	89	89	**94**	92	89	94
believe.v	2	2	202	55	**87**	78	78	86	84	78	74	80	81	74	87
occur.v	2	2	47	22	86	73	91	**96**	86	**96**	86	82	87	73	96
grant.v	2	2	19	5	**100**	80	80	80	40	80	60	80	75	40	100
enjoy.v	2	2	56	14	50	57	57	50	**64**	57	50	57	55	50	64
need.v	2	2	195	56	**89**	82	86	**89**	86	78	70	70	81	70	89
disclose.v	1	1	55	14	**93**	**93**	**93**	**93**	**93**	**93**	**93**	**93**	93	93	93
point.n	9	6	469	150	91	91	89	91	**92**	87	84	79	88	79	92
position.n	7	6	268	45	**78**	**78**	**78**	53	56	65	58	64	66	53	78
defense.n	7	7	120	21	**57**	48	52	43	48	29	48	48	46	29	57
carrier.n	7	3	111	21	**71**	**71**	**71**	**71**	67	**71**	**71**	62	70	62	71
order.n	7	4	346	57	93	**95**	93	91	93	92	90	91	92	90	95
exchange.n	5	3	363	61	**92**	90	**92**	85	90	88	82	79	87	79	92
system.n	5	3	450	70	**79**	73	66	67	59	63	63	61	66	59	79
source.n	5	5	152	35	**86**	80	80	63	83	68	60	29	69	29	86
space.n	5	2	67	14	93	**100**	93	93	93	86	86	71	89	71	100
base.n	5	4	92	20	75	**80**	75	50	65	40	50	75	64	40	80
authority.n	4	3	90	21	**86**	**86**	81	62	71	33	71	81	71	33	86
people.n	4	4	754	115	**96**	96	95	96	95	90	91	91	94	90	96
chance.n	4	3	91	15	60	67	60	60	67	**73**	20	**73**	60	20	73
part.n	4	3	481	71	90	90	92	**97**	90	74	66	66	83	66	97
hour.n	4	2	187	48	83	85	**92**	83	77	90	58	**92**	83	58	92
development.n	3	3	180	29	**100**	79	86	79	76	62	79	62	78	62	100
president.n	3	3	879	177	**98**	97	**98**	97	93	96	97	85	95	85	98
network.n	3	3	152	55	91	87	**98**	89	84	88	87	82	88	82	98
future.n	3	3	350	146	97	96	94	97	83	**98**	89	85	92	83	98
effect.n	3	2	178	30	**97**	93	80	93	80	90	77	83	87	77	97
state.n	3	3	617	72	85	**86**	**86**	83	82	79	83	82	83	79	86
power.n	3	3	251	47	**92**	87	87	81	77	77	77	74	81	74	92
bill.n	3	3	404	102	98	**99**	98	96	90	96	96	22	87	22	99
area.n	3	3	326	37	**89**	73	65	68	84	70	68	65	73	65	89
job.n	3	3	188	39	85	80	77	**90**	80	82	69	82	80	69	90
management.n	2	2	284	45	89	78	87	73	**98**	76	67	64	79	64	98
condition.n	2	2	132	34	**91**	82	82	56	76	78	74	76	77	56	91
policy.n	2	2	331	39	95	**97**	**97**	87	95	**97**	90	64	90	64	97
rate.n	2	2	1009	145	90	88	**92**	81	92	89	88	91	89	81	92
drug.n	2	2	205	46	94	94	**96**	78	94	94	87	78	89	78	96
			Average	Overall	86	83	83	82	82	79	76	77			
				Verbs	78	75	73	76	73	70	65	70			
				Nouns	89	87	86	81	83	80	77	76			

Table 4: All Supervised system performance per predicate. (Column legend – S=number of senses in training; s=number senses appearing more than 3 times; T=instances in training; t=instances in test.; The numbers indicate system ranks.)

3 Semantic Role Labeling

Subtask 2 evaluates Semantic Role Labeling (SRL) systems, where the goal is to locate the constituents which are arguments of a given verb, and to assign them appropriate semantic roles that describe how they relate to the verb. SRL systems are an important building block for many larger semantic systems. For example, in order to determine that question (1a) is answered by sentence (1b), but not by sentence (1c), we must determine the relationships between the relevant verbs (*eat* and *feed*) and their arguments.

(1) a. What do lobsters like to eat?
b. Recent studies have shown that lobsters primarily feed on live fish, dig for clams, sea urchins, and feed on algae and eel-grass.
c. In the early 20th century, Mainers would only eat lobsters because the fish they caught was too valuable to eat themselves.

Traditionally, SRL systems have been trained on either the PropBank corpus (Palmer et al., 2005) – for two years, the CoNLL workshop (Carreras and Màrquez, 2004; Carreras and Màrquez, 2005) has made this their shared task, or the FrameNet corpus – Senseval-3 used this for their shared task (Litkowski, 2004). However, there is still little consensus in the linguistics and NLP communities about what set of role labels are most appropriate. The PropBank corpus avoids this issue by using theory-agnostic labels (ARG0, ARG1, ..., ARG5), and by defining those labels to have only verb-specific meanings. Under this scheme, PropBank can avoid making any claims about how any one verb's arguments relate to other verbs' arguments, or about general distinctions between verb arguments and adjuncts.

However, there are several limitations to this approach. The first is that it can be difficult to make inferences and generalizations based on role labels that are only meaningful with respect to a single verb. Since each role label is verb-specific, we can not confidently determine when two different verbs' arguments have the same role; and since no encoded meaning is associated with each tag, we can not make generalizations across verb classes. In contrast, the use of a shared set of role labels, such as VerbNet roles, would facilitate both inferencing and generalization. VerbNet has more traditional labels such as Agent, Patient, Theme, Beneficiary, etc. (Kipper et al., 2006).

System	Type	Precision	Recall	F
UBC-UPC	Open	84.51	82.24	83.36±0.5
UBC-UPC	Closed	85.04	82.07	83.52±0.5
RTV	Closed	81.82	70.37	75.66±0.6
Without "say"				
UBC-UPC	Open	78.57	74.70	76.60±0.8
UBC-UPC	Closed	78.67	73.94	76.23±0.8
RTV	Closed	74.15	57.85	65.00±0.9

Table 5: System performance on PropBank arguments.

Therefore, we chose to annotate the corpus using two different role label sets: the PropBank role set and the VerbNet role set. VerbNet roles were generated using the SemLink mapping (Loper et al., 2007), which provides a mapping between PropBank and VerbNet role labels. In a small number of cases, no VerbNet role was available (e.g., because VerbNet did not contain the appropriate sense of the verb). In those cases, the PropBank role label was used instead.

We proposed two levels of participation in this task: i) Closed – the systems could use only the annotated data provided and nothing else. ii) Open – where systems could use PropBank data from Sections 02-21, as well as any other resource for training their labelers.

3.1 Data

We selected 50 verbs from the 65 in the lexical sample task for the SRL task. The partitioning into train and test set was done in the same fashion as for the lexical sample task. Since PropBank does not tag any noun predicates, none of the 35 nouns from the lexical sample task were part of this data.

3.2 Results

For each system, we calculated the precision, recall, and F-measure for both role label sets. Scores were calculated using the `srl-eval.pl` script from the CoNLL-2005 scoring package (Carreras and Màrquez, 2005). Only two teams chose to perform the SRL subtask. The performance of these two teams is shown in Table 5 and Table 6.

System	Type	Precision	Recall	F
UBC-UPC	Open	85.31	82.08	83.66±0.5
UBC-UPC	Closed	85.31	82.08	83.66±0.5
RTV	Closed	81.58	70.16	75.44±0.6
Without "say"				
UBC-UPC	Open	79.23	73.88	76.46±0.8
UBC-UPC	Closed	79.23	73.88	76.46±0.8
RTV	Closed	73.63	57.44	64.53±0.9

Table 6: System performance on VerbNet roles.

3.3 Discussion

Given that only two systems participated in the task, it is difficult to form any strong conclusions. It should be noted that since there was no additional VerbNet role data to be used by the Open system, the performance of that on PropBank arguments as well as VerbNet roles is exactly identical. It can be seen that there is almost no difference between the performance of the Open and Closed systems for tagging PropBank arguments. The reason for this is the fact that all the instances of the lemma under consideration was selected from the Propbank corpus, and probably the number of training instances for each lemma as well as the fact that the predicate is such an important feature combine to make the difference negligible. We also realized that more than half of the test instances were contributed by the predicate "say" – the performance over whose arguments is in the high 90s. To remove the effect of "say" we also computed the performances after excluding examples of "say" from the test set. These numbers are shown in the bottom half of the two tables. These results are not directly comparable to the CoNLL-2005 shared task since: i) this test set comprises Sections 01, 22, 23 and 24 as opposed to just Section 23, and ii) this test set comprises data for only 50 predicates as opposed to all the verb predicates in the CoNLL-2005 shared task.

4 Conclusions

The results in the previous discussion seem to confirm the hypothesis that there is a predictable correlation between human annotator agreement and system performance. Given high enough ITA rates we can can hope to build sense disambiguation systems that perform at a level that might be of use to a consuming natural language processing application. It is also encouraging that the more informative VerbNet roles which have better/direct applicability in downstream systems, can also be predicted with almost the same degree of accuracy as the PropBank arguments from which they are mapped.

5 Acknowledgments

We gratefully acknowledge the support of the Defense Advanced Research Projects Agency (DARPA/IPTO) under the GALE program, DARPA/CMO Contract No. HR0011-06-C-0022; National Science Foundation Grant NSF-0415923, Word Sense Disambiguation; the DTO-AQUAINT NBCHC040036 grant under the University of Illinois subcontract to University of Pennsylvania 2003-07911-01; and NSF-ITR-0325646: Domain-Independent Semantic Interpretation.

References

Xavier Carreras and Lluís Màrquez. 2004. Introduction to the CoNLL-2004 shared task: Semantic role labeling. In *Proceedings of CoNLL-2004*.

Xavier Carreras and Lluís Màrquez. 2005. Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling. In *Proceedings of CoNLL-2005*.

Jinying Chen, Andrew Schein, Lyle Ungar, and Martha Palmer. 2006. An empirical study of the behavior of active learning for word sense disambiguation. In *Proceedings of HLT/NAACL*.

Timothy Chklovski and Rada Mihalcea. 2002. Building a sense tagged corpus with open mind word expert. In *Proceedings of ACL-02 Workshop on WSD*.

Bart Decadt, Véronique Hoste, Walter Daelemans, and Antal Van den Bosch. 2004. GAMBL, genetic algorithm optimization of memory-based wsd. In *Senseval-3*.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: The 90% solution. In *Proceedings of HLT/NAACL*, June.

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2006. Extending VerbNet with novel verb classes. In *LREC-06*.

Ken Litkowski. 2004. Senseval-3 task: Automatic labeling of semantic roles. In *Proceedings of Senseval-3*.

Edward Loper, Szu ting Yi, and Martha Palmer. 2007. Combining lexical resources: Mapping between propbank and verbnet. In *Proceedings of the IWCS-7*.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: A corpus annotated with semantic roles. *Computational Linguistics*, 31(1):71–106.

Martha Palmer. 2000. Consistent criteria for sense distinctions. *Computers and the Humanities*, 34(1-1):217–222.