The Web as Collective Mind: Building Large Annotated Corpora with Web Users' Help

Rada Mihalcea
Department of Computer Science
University of North Texas
rada@cs.unt.edu

Timothy Chklovski
Artificial Intelligence Laboratory
Massachusetts Institute of Technology
timc@ai.mit.edu

Abstract

This paper presents an approach for building large annotated corpora by tapping into the world's vast pool of knowledge. It does this by relying on the vast number of Web users who contribute their knowledge to data annotation. We focus on building semantically annotated corpora, by collecting word sense tagging from the general public over the Web. The paper addresses the various issues faced when collecting annotated data from Web users, and evaluates the quality of these data through several Word Sense Disambiguation experiments.

1 Introduction

Good performance in Natural Language Processing (NLP) applications often requires large amounts of annotated data. While there were significant recent developments in the performance of NLP methods and algorithms, there has been relatively little progress on addressing the problem of obtaining annotated data required by some of the highest-performing algorithms. Consequently, many of today's NLP applications experience severe training data bottlenecks.

One notoriously difficult problem in understanding text is Word Sense Disambiguation (WSD). Ambiguity is very common, especially among the most common words. Humans, however, are so competent at figuring out word senses from context that they usually do not even notice the ambiguities. While a large number of efficient WSD algorithms have been designed and implemented to date within the recent SENSEVAL evaluation frameworks (Kilgarriff and Palmer, 2000), (Preiss and Yarowsky, 2001), and elsewhere, the availability of sense tagged data is still a significant problem.

Most of the efforts in WSD have focused on *supervised* learning algorithms, which usually achieve the best performance at the cost of low recall. The main weakness of these methods is the lack of widely available semantically tagged corpora and the strong dependence of disambiguation accuracy on the size of the training corpus. For instance, one study reports that high precision WSD requires an average of 500 examples per ambiguous word, depending on the word entropy (Ng, 1997). At a throughput of one tagged example per minute (Edmonds, 2000), and with about 20,000 ambiguous words in the common English vocabulary, a simple calculation leads to about 160,000 hours of tagging, which is nothing less than 80 man-years of human annotation work.

Similar data bottleneck problems are faced by many other NLP applications. High quality part of speech tagging for English requires about 3 million words annotated with their part of speech. The state-of-the-art in syntactic parsing in English is close to 88-89%(Collins, 1996), performance attainable by training parser models on a corpus of about 600,000 words manually parsed within the Penn Treebank project, an annotation effort that required approximately 2 man-years of work (Marcus et al., 1993). Information extraction, automatic summarization,

anaphora resolution, and other tasks also strongly require large annotated corpora. Since the tagging process is usually done by trained lexicographers, it is very expensive, and limits the size of such corpora to a handful of tagged texts.

In this paper, we present a Web-based system that aims to create large sense tagged corpora with the help of Web users. The annotation workload is distributed among millions of potential human annotators, which is likely to significantly reduce the cost and the duration of the annotation process. We investigate the amount and quality of the data produced during first year of deployment of the activity, and present results obtained during WSD experiments that rely on this sense tagged data.

The paper is organized as follows. We first review the sense tagged corpora currently available, and show the limitations imposed by the traditional method of annotating data by hiring lexicographers. Next, we describe the Web-based system that we employ to collect lexically annotated corpora from Web users, and evaluate the quantity and quality of the data collected during the first year of activity. Finally, we present the results obtained during WSD experiments relying on these data, which represent an additional proof towards the validity of the annotation process.

2 Sense Tagged Corpora

The availability of large amounts of semantically tagged data is crucial for creating successful WSD systems. Yet, as of today, only few sense tagged corpora are publicly available (See http://www.senseval.org for currently available sense tagged corpora).

One of the first large scale hand tagging efforts is reported in (Miller et al., 1993), where a subset of the Brown corpus was tagged with WordNet (Miller, 1995) senses. The corpus includes a total of 234,136 tagged word occurrences, out of which 186,575 are polysemous. There are 88,058 noun occurrences of which 70,214 are polysemous.

The next significant hand tagging task was reported in (Bruce and Wiebe, 1994), where 2,476 usages of *interest* were manually assigned with sense tags from the Longman Dictionary of Contemporary English (LDOCE). This corpus was used in various experiments, with classification accuracies ranging from 75% to 90%, depending on the algorithm and features employed.

The high accuracy of the LEXAS system (Ng and Lee, 1996) is due in part to the use of large corpora. For this system, 192,800 word occurrences have been manually tagged with senses from WordNet. The set of tagged words consists of the 191 most frequently occurring nouns and verbs. The authors mention that approximately one man-year of effort was spent in tagging the data set.

Recently, the SENSEVAL competitions have been providing a good environment for the development of supervised WSD systems, making freely available large amounts of sense tagged data for about 100 words. During SENSEVAL-1 (Kilgarriff and Palmer, 2000), data for 35 words was made available adding up to about 20,000 examples tagged with respect to the Hector dictionary. The size of the tagged corpus increased with SENSEVAL-2 (Preiss and Yarowsky, 2001), when 13,000 additional examples were released for 73 polysemous words. This time, the semantic annotations were performed with respect to WordNet.

Additionally, (Kilgarriff, 1998) mentions the Hector corpus, with about 300 word types with 300-1000 tagged instances for each word, selected from a 17 million word corpus.

With the approach described in this paper, we aim at creating a very large sense tagged corpus by making use of the incredible resource of knowledge constituted by millions of Web users. We use techniques for active learning to utilize this resource efficiently.

3 Building Sense Tagged Corpora with the Help of Web' Users

To overcome the current lack of sense tagged data and the limitations imposed by the creation of such data using trained lexicographers, we designed a system that enables the collection of semantically annotated corpora over the Web.

Sense tagged examples are collected using a Web-based application that allows contributors to annotate words with their meanings. Tagging is organized by word: for each ambiguous word for which we want to build a sense tagged corpus, users are presented with a set of natural language (English)

sentences that include an instance of the ambiguous word.

The overall process proceeds as follows. Initially, example sentences are extracted from a large textual corpus. If other training data is not available, a number of these sentences are presented to the users for tagging in *Stage 1*. Next, this tagged collection is used as training data, and active learning is used to identify in the remaining corpus the examples that are "hard to tag". These are the examples that are presented to the contributors for tagging in *Stage 2*. For all tagging, users are asked to select the sense they find to be the most appropriate in the given sentence. The selection is made from a drop-down list containing all WordNet senses of the current word, plus two additional choices, "unclear" and "none of the above." The results of any automatic classification or the classification submitted by other users are not presented so as to not bias the contributor's decisions. Based on early feedback from both researchers and contributors, a future version of the system may allow contributors to specify more than one sense for a given instance. As will be elaborated below, the current approach of collecting redundant tagging already addresses this to some degree.

3.1 Data

The starting corpus we use is formed by a mix of three different sources of data, namely the *Penn Treebank* corpus (Marcus et al., 1993), the *Los Angeles Times* collection, as provided during TREC conferences, and *Open Mind Common Sense*[1], a collection of about 500,000 commonsense assertions in English as contributed by volunteers over the Web (Singh, 2002) [2]. A mix of several sources, each covering a different spectrum of usage, is used to increase the coverage of word senses and writing styles. While the first two sources are well known to the NLP community, the *Open Mind Common Sense* constitutes a fairly new textual corpus. It consists mostly of simple single sentences. These sentences tend to be explanations and assertions similar to glosses of a dictionary, but phrased in a more common language and with many sentences per sense.

[1] http://commonsense.media.mit.edu

[2] See also (Singh et al., 2002) for additional details regarding the quality of free-form entered information, evaluation, bias, and the level of difficulty of the collected knowledge.

We are currently in the process of integrating the *British National Corpus*, and we are planning to also integrate the *American National Corpus* as soon as it will become available.

3.2 Sense Inventory

The sense inventory used in the current system implementation is WordNet (Miller, 1995). Users are presented with the current sense definitions from WordNet, and asked to decide on the most appropriate meaning in the given context. Future versions of the system may adopt a new sense inventory, or use the coarse sense classes from WordNet, since the current fine granularity of WordNet was occasionally a source of confusion and eventually discouraged some of them to return to the tagging task.

3.3 Active Learning

To minimize the amount of human annotation effort needed to build a tagged corpus for a given ambiguous word, we have an active learning component that has the role of selecting for annotation only those examples that are the most informative.

According to (Dagan et al., 1995), there are two main types of active learning. The first one uses memberships queries, in which the learner constructs examples and asks a user to label them. In natural language processing tasks, this approach is not always applicable, since it is hard and not always possible to construct meaningful unlabeled examples for training. Instead, a second type of active learning can be applied to these tasks, which is *selective sampling*. In this case, several classifiers examine the unlabeled data and identify only those examples that are the most informative, that is the examples where a certain level of disagreement is measured among the classifiers.

We use a simplified form of active learning with selective sampling, where the instances to be tagged are selected as those instances where there is a disagreement between the labels assigned by two different classifiers. The two classifiers are trained on a relatively small corpus of tagged data, which is formed either with (1) Senseval training examples, in the case of Senseval words, or (2) examples obtained with the system itself, when no other training data is available.

The first classifier is a Semantic Tagger with Active Feature Selection (STAFS) (Mihalcea, 2002). The system consists of an instance based learning algorithm improved with a scheme for automatic feature selection. It relies on the fact that different sets of features have different effects depending on the ambiguous word considered. Rather than creating a general learning model for all polysemous words, STAFS builds a separate feature space for each individual word. The features are selected from a pool of eighteen different features that have been previously acknowledged as good indicators of word sense, including: part of speech of the ambiguous word itself, surrounding words and their parts of speech, keywords in context, noun before and after, verb before and after, and others. An iterative forward search algorithm identifies at each step the feature that leads to the highest cross-validation precision computed on the training data.

The second classifier is a COnstraint-BAsed Language Tagger (COBALT). The system treats every training example as a set of soft constraints on the sense of the word of interest. WordNet glosses, hyponyms, hyponym glosses and other WordNet data is also used to create soft constraints. Currently, only "keywords in context" type of constraint is implemented, with weights accounting for the distance from the target word. The tagging is performed by finding the sense that minimizes the violation of constraints in the instance being tagged. COBALT generates confidences in its tagging of a given instance based on how much the constraints were satisfied and violated for that instance.

Both taggers use WordNet 1.7 dictionary glosses and relations. The performance of the two systems and their level of agreement were evaluated on the Senseval noun data set. The two systems agreed in their classification decision in 54.96% of the cases. This low agreement level is a good indication that the two approaches are fairly orthogonal, and therefore we may hope for high disambiguation precision on the agreement set. Indeed, the tagging accuracy measured on the set where both COBALT and STAFS assign the same label is 82.5%, a fairly high figure.

Table 1 lists the precision for the agreement and disagreement sets of the two taggers. The low precision on the instances in the disagreement set justifies referring to these as "hard to tag". These are the instances that are presented to the users for tagging in the active learning stage.

System	Precision (fine grained)	Precision (coarse grained)
STAFS	69.5%	76.6%
COBALT	59.2%	66.8%
STAFS∩COBALT	**82.5%**	**86.3%**
STAFS - STAFS∩COBALT	52.4%	63.3%
COBALT - STAFS∩COBALT	30.09%	42.07%

Table 1: Disambiguation precision for the two individual classifiers and their agreement and disagreement sets

4 Quality and Quantity of Semantically Annotated Corpora Collected over the Web

Collecting from the general public holds the promise of providing much data at low cost. It also makes attending to two aspects of data collection more important: (1) ensuring contribution quality, and (2) making the contribution process engaging to the contributors.

To ensure contribution quality, redundant tagging is collected for each item. The system currently uses the following rules in presenting items to volunteer contributors:

- Two tags per item. Once an item has two tags associated with it, it is not presented for further tagging.
- One tag per item per contributor. We allow contributors to submit tagging either anonymously or having logged in. Anonymous contributors are not shown any items already tagged by contributors (anonymous or not) from the same IP address. Logged in contributors are not shown items they have already tagged.

In all, automatic assessment of the quality of tagging seems possible, and, based on the experience of similar volunteer contribution projects (Singh, 2002), the rate of maliciously misleading or incorrect contributions has been surprisingly low.

During the first year of activity, we collected almost 100,000 individual sense taggings from contributors. Of that number, approximately 16,500 tags came from using the system in the classrooms

as a teaching aid (the web site provides special features for this). Future rate of collection depends on the site being listed in various directories and on the contributor repeat visit rate. We are also experimenting with prizes to encourage participation.

There are two main figures that we measured to evaluate the quality of the annotation task. One is *inter tagger agreement*, which represents the agreement between the tags assigned by two different annotators. The other is *replicability*, which measures the degree to which an annotation experiment can be replicated. According to (Kilgarriff, 1999), the capability of recreating a set of annotated data is an even more telling indicator of annotation quality than the inter-annotator agreement.

4.1 Inter-Tagger Agreement

In terms of inter-annotator agreement, the results obtained so far can be directly compared with the agreement figures previously reported in the literature. (Kilgarriff, 2002) mentions that for the SENSEVAL-2 nouns and adjectives there was a 66.5% agreement between the first two tags collected for each item. About 12% of their tagging consisted of multi-word expressions and proper nouns, which are usually not ambiguous, and which are not considered during our data collection process. So far we measured a 62.8% inter-tagger agreement for single word tagging, plus close-to-100% precision in tagging multi-word expressions and proper nouns (as mentioned earlier, this represents about 12% of the annotated data). This results in an overall agreement of about 67.3% which is reasonable and closely comparable with previous figures.

4.2 Replicability

To measure the replicability of the tagging process performed by Web users, we had to replicate a tagging experiment where the annotation was performed with "trusted humans." To this end, we used the data set for the noun "interest," made available by (Bruce and Wiebe, 1994). In this data set, consisting of 2,369 examples, the annotation was done with respect to LDOCE, and therefore we had first to map the sense entries from this dictionary to WordNet. The mapping did not pose any particular problems, and consists of one-to-one mappings for the six LDOCE entries, plus one WordNet entry not defined in LDOCE, for which we discarded all corresponding examples from the Open Mind annotation.

Number of training examples	Precision		Error rate reduction
	baseline	STAFS	
any	63.32%	66.23%	9%
> 100	75.88%	80.32%	19%
> 200	63.48%	72.18%	24%
> 300	45.51%	69.15%	43%

Table 2: Precision and error rate reduction for various sizes of the training corpus.

Next, we identified and eliminated all the examples in the corpus that contained collocations (e.g. "interest rate"); these examples account for more than 35% of the data. Finally, the remaining 1,438 examples were displayed on the Web-based interface for tagging.

Out of the 1,438 examples, 1,066 had two tags that agreed, therefore a 74% inter-annotator agreement for single words tagging[3]. Out of these 1,066 items, 967 have a tag that coincides with the tag assigned in the experiments reported in (Bruce and Wiebe, 1994), which leads to an 90.8% replicability for single words tagging (note that the 35% monosemous multi-word expressions are not taken into account by this figure). This is close to the 95% replicability scores mentioned in (Kilgarriff, 1999) for annotation experiments performed by lexicographers.

5 Word Sense Disambiguation Experiments

For additional evaluations of the quality of the data collected from Web users, we used these data sets in disambiguation experiments, performed using the STAFS WSD system. Two sets of experiments were performed. One set of experiments involved the corpus constructed by Web users, and evaluations were performed using ten-fold cross validations runs. This is the *intra-corpus* experiment, were both training and test sets are from the same source. The second experimental set involved *inter-corpora* evaluations, where the training corpus provided during SENSEVAL-2 evaluation exercise was

[3]Adding the 35% monosemous multi-word expressions tagged with 100% precision, leads to an overall 83% inter-tagger agreement for this particular word

augmented with examples contributed by Web users, and subsequently tested on the SENSEVAL-2 test data.

5.1 Intra-corpus WSD

In this experiment, we employ the STAFS WSD system, with a fixed set of features, and evaluated the WSD performance in 10-fold cross validation runs. We also compute a simple baseline, consisting of a simple heuristic that assigns the most frequent sense by default (also computed during 10-fold cross validation runs). Table 3 lists: all words for which we collect sense tagged data with at least 100 annotated examples available; the number of items with full inter-annotator agreement; the most frequent sense baseline; and the precision achieved with STAFS.

For the total of 280 words for which data were collected from Web users, the average number of examples per word is 87. The most frequent sense heuristic yields correct results in 63.32% overall. When disambiguation is performed using STAFS, restricting the system to a simple set of features consisting of the word itself, the word's part of speech, and a surrounding context of two (words and their corresponding parts of speech), the overall precision is 66.23%, which represents an error reduction of about 9% with respect to the most frequent sense heuristic.

Moreover, the average for the 72 words which have at least 100 training examples (the words listed in Table 3) is 75.88% for the most frequent sense heuristic, and 80.32% when using STAFS, resulting in an error reduction of 19%. When at least 200 examples are available per word, the most frequent sense heuristic is correct 63.48% of the time, and STAFS is correct 72.18% of the time, which represents a 24% reduction in disambiguation error. Table 2 lists the precisions obtained with the most frequent sense heuristic and STAFS, as a function of corpus size. The error reduction rates grow steadily with the number of training examples.

For the words for which more data was collected from Web users, the improvement over the most frequent sense baseline was larger. This agrees with prior work by other researchers (Ng, 1997), (Banko and Brill, 2001), who noted that additional annotated data is likely to bring significant improvements in disambiguation quality.

Word	SENSEVAL-2		SENSEVAL-2 + OMWE	
	Fine	Coarse	Fine	Coarse
art	60.2%	65.3%	61.2%	68.4%
authority	70.7%	85.9%	76.1%	90.2%
bar	45.7%	58.3%	46.4%	57.0%
bum	75.6%	77.8%	66.7%	78.9%
chair	81.2%	81.2%	82.6%	82.6%
channel	52.1%	54.8%	49.2%	56.2%
child	60.9%	62.5%	56.2%	57.8%
church	62.5%	62.5%	67.2%	67.2%
circuit	49.4%	49.4%	48.2%	50.6%
day	65.5%	65.5%	66.2%	67.6%
detention	68.8%	68.8%	71.9%	71.9%
dyke	82.1%	82.1%	82.1%	82.1%
facility	58.6%	93.1%	48.3%	94.8%
fatigue	79.1%	83.7%	76.7%	81.4%
feeling	64.7%	64.7%	64.7%	64.7%
grip	54.7%	74.5%	62.7%	70.6%
hearth	71.9%	87.5%	71.9%	87.5%
holiday	77.4%	83.9%	77.4%	87.1%
lady	77.4%	86.8%	77.4%	88.7%
material	50.7%	60.9%	53.6%	66.7%
mouth	56.7%	85.0%	66.7%	90.0%
nation	70.3%	73.0%	70.3%	73.0%
nature	65.2%	76.1%	69.6%	84.8%
post	58.2%	62.0%	57.0%	60.8%
restraint	48.9%	60.0%	57.8%	68.9%
sense	54.7%	54.7%	58.5%	60.4%
spade	57.6%	57.6%	51.5%	51.5%
stress	56.4%	82.1%	56.4%	82.1%
yew	78.6%	96.4%	78.6%	96.4%
Average	63.99%	72.27%	64.58%	73.78%

Table 4: Evaluation using SENSEVAL-2 and Web-users examples

5.2 Inter-corpora experiments

In these experiments, we enlarge the set of training examples provided within the Senseval evaluation exercise with the examples collected from Web users, and evaluate the impact on performance of these new training examples. Table 4 shows the results obtained on the test data when only SENSEVAL-2 training data were used, and the results obtained with both SENSEVAL-2 and Web-users training examples. The same system as in the previous experiments is used in this experiment. Only examples pertaining to single words were used (that is, we eliminate the SENSEVAL-2 examples pertaining to collocations).

There is a small error rate reduction of 2% for the fine grained scoring, but a more significant error reduction of 5.7% for coarse grained scoring. Notice that the examples used in our Web-based system are drawn from a corpus completely different than the corpus used for SENSEVAL-2 examples, and therefore the sense distributions are often different, and do not always match the test data sense distributions (as is the case when train and test data are drawn from the same source).

Word	Set size	Baseline	STAFS	Word	Set size	Baseline	STAFS	Word	Set size	Baseline	STAFS
activity	103	90.00%	90.00%	arm	142	52.50%	80.62%	art	107	30.00%	63.53%
attitude	107	100.00%	100.00%	bank	160	91.88%	91.88%	bar	107	61.76%	70.59%
bed	142	98.12%	98.12%	blood	136	91.05%	91.05%	brother	101	95.45%	95.45%
building	114	87.33%	88.67%	captain	101	47.27%	48.18%	car	144	99.44%	99.44%
cell	126	89.44%	88.33%	chance	115	56.25%	81.88%	channel	103	84.62%	86.15%
chapter	137	68.50%	71.50%	child	105	55.33%	84.67%	circuit	197	31.92%	45.77%
coffee	115	95.00%	95.00%	day	192	34.76%	44.76%	degree	140	71.43%	82.14%
device	106	98.12%	98.12%	doctor	133	100.00%	100.00%	dog	130	100.00%	100.00%
door	112	54.62%	45.38%	eye	117	96.11%	96.11%	facility	205	81.60%	74.40%
father	160	96.88%	96.88%	function	105	24.67%	32.00%	god	110	71.82%	81.82%
grip	239	45.94%	61.88%	gun	143	94.71%	94.71%	hair	147	96.67%	96.67%
horse	138	100.00%	100.00%	image	120	36.67%	71.67%	individual	103	96.15%	96.15%
interest	1066	39.91%	71.08%	kid	106	83.75%	84.38%	law	106	38.12%	66.88%
letter	137	85.00%	81.00%	list	102	100.00%	100.00%	material	196	77.60%	76.40%
mother	119	99.00%	99.00%	mouth	151	74.38%	77.50%	name	136	98.42%	98.42%
object	183	96.19%	96.19%	office	209	62.76%	61.03%	officer	103	56.15%	55.38%
people	120	99.17%	99.17%	plant	126	98.89%	98.89%	pressure	106	72.50%	70.62%
product	216	80.74%	81.48%	report	101	66.36%	60.91%	rest	360	51.11%	67.22%
restraint	204	22.92%	46.25%	room	124	100.00%	100.00%	sea	205	90.80%	90.80%
season	102	92.50%	92.50%	song	116	92.35%	92.35%	structure	112	75.38%	72.31%
sun	101	63.64%	66.36%	term	125	71.18%	90.59%	treatment	108	67.78%	66.67%
tree	105	100.00%	100.00%	trial	109	87.37%	86.84%	type	135	92.78%	92.78%
unit	108	54.44%	46.67%	volume	103	63.85%	54.62%	water	103	53.85%	72.31%

Table 3: Words with more than 100 sense tagged examples: (1) set size, (2) precision attainable with the most frequent sense heuristic, (3) precision attainable with the STAFS WSD system

6 Conclusions and future work

Collecting data from Web users has the potential of creating a large sense tagged corpus. In this paper we investigated the amount and quality of data collected during the first year of deployment of the activity. The experiments performed showed that inter-tagger agreement, replicability, and disambiguation results obtained on these data are comparable with what can be achieved with data collected using the traditional method of hiring lexicographers, at a much lower cost.

Acknowledgments

We want to thank the Open Mind Word Expert contributors who are making all this work possible. We are also grateful to Ted Pedersen and the NLP group at University of Minnesota at Duluth for interesting discussions and important contributions to this data collection process, to Adam Kilgarriff for valuable suggestions, and to all the Open Mind Word Expert users who have emailed us with their feedback and suggestions, helping us improve this activity.

References

M. Banko and E. Brill. 2001. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Lingusitics (ACL-2001)*, Toulouse, France, July.

R. Bruce and J. Wiebe. 1994. Word sense disambiguation using decomposable models. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL-94)*, pages 139–146, LasCruces, NM, June.

M. Collins. 1996. A new statistical parser based on bigram lexical dependencies. In *Proceedings of the 34th Annual Meeting of the ACL*, Santa Cruz.

I. Dagan, , and S.P. Engelson. 1995. Committee-based sampling for training probabilistic classifiers. In *International Conference on Machine Learning*, pages 150–157.

P. Edmonds. 2000. Designing a task for Senseval-2, May. Available online at http://www.itri.bton.ac.uk/events/senseval.

A. Kilgarriff and M. Palmer, editors. 2000. *Computer and the Humanities. Special issue: SENSEVAL. Evaluating Word Sense Disambiguation programs*, volume 34, April.

A. Kilgarriff. 1998. Gold standard datasets for evaluating word sense disambiguation programs. *Computer Speech and Language*, 12(4):453–472.

A. Kilgarriff. 1999. 95% replicability for manual word sense tagging. In *Proceedings of European Association for Computational Linguistics*, pages 277–278, Bergen, Norway, June.

A. Kilgarriff. 2002. English lexical sample task description. In *Proceedings of Senseval-2 Workshop, Association of Computational Linguistics*, pages 17–20, Toulouse, France.

M.P. Marcus, B. Santorini, and M.A. Marcinkiewicz. 1993. Building a large annotated corpus of english: the Penn Treebank. *Computational Linguistics*, 19(2):313–330.

R. Mihalcea. 2002. Instance based learning with automatic feature selection applied to Word Sense Disambiguation. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING-ACL 2002)*, Taipei, Taiwan, August.

G. Miller, C. Leacock, T. Randee, and R. Bunker. 1993. A semantic concordance. In *Proceedings of the 3rd DARPA Workshop on Human Language Technology*, pages 303–308, Plainsboro, New Jersey.

G. Miller. 1995. Wordnet: A lexical database. *Communication of the ACM*, 38(11):39–41.

H.T. Ng and H.B. Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense: An examplar-based approach. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL-96)*, Santa Cruz.

H.T. Ng. 1997. Getting serious about word sense disambiguation. In *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?*, pages 1–7, Washington.

J. Preiss and D. Yarowsky, editors. 2001. *Proceedings of SENSEVAL-2, Association for Computational Linguistics Workshop*, Toulouse, France.

P. Singh, T. Lin, E. Mueller, G. Lim, T. Perkins, and W. Li Zhu. 2002. Open mind common sense: Knowledge acquisition from the general public. In *Proceedings of the First International Conference on Ontologies, Databases, and Applications of Semantics for Large Scale Information Systems*. Lecture Notes in Computer Science, Heidelberg: Springer-Verlag.

P. Singh. 2002. The public acquisition of commonsense knowledge. In *Proceedings of AAAI Spring Symposium: Acquiring (and Using) Linguistic (and World) Knowledge for Information Access.*, Palo Alto, CA. AAAI.