Building a Sense Tagged Corpus with *Open Mind Word Expert*

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

Rada Mihalcea
Department of Computer Science
University of Texas at Dallas
rada@utdallas.edu

Abstract

Open Mind Word Expert is an implemented active learning system for collecting word sense tagging from the general public over the Web. It is available at http://teach-computers.org. We expect the system to yield a large volume of high-quality training data at a much lower cost than the traditional method of hiring lexicographers. We thus propose a Senseval-3 lexical sample activity where the training data is collected via Open Mind Word Expert. If successful, the collection process can be extended to create the definitive corpus of word sense information.

1 Introduction

Most of the efforts in the Word Sense Disambiguation (WSD) field have concentrated on *supervised* learning algorithms. These methods 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. The tagging process is usually done by trained lexicographers, and consequently is quite expensive, limiting the size of such corpora to a handful of tagged texts.

This paper introduces *Open Mind Word Expert*, a Web-based system that aims at creating large sense tagged corpora with the help of Web users. The system has an active learning component, used for selecting the most difficult examples, which are then presented to the human taggers. We expect that the system will yield more training data of comparable quality and at a significantly lower cost than the traditional method of hiring lexicographers.

Open Mind Word Expert is a newly born project that follows the *Open Mind* initiative (Stork, 1999). The basic idea behind *Open Mind* is to use the information and knowledge that may be collected from the existing millions of Web users, to the end of creating more intelligent software. This idea has been used in *Open Mind Common Sense*, which acquires commonsense knowledge from people. A knowledge base of about 400,000 facts has been built by learning facts from 8,000 Web users, over a one year period (Singh, 2002). If *Open Mind Word Expert* experiences a similar learning rate, we expect to shortly obtain a corpus that exceeds the size of all previously tagged data. During the first fifty days of activity, we collected about 26,000 tagged examples without significant efforts for publicizing the site. We expect this rate to gradually increase as the site becomes more widely known and receives more traffic.

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.

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 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.

Lately, the SENSEVAL competitions provide 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 (Kilgarriff, 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, which comprises about 300 word types with 300-1000 tagged instances for each word, selected from a 17 million word corpus.

Sense tagged corpora have thus been central to accurate WSD systems. Estimations made in (Ng, 1997) indicated that a high accuracy domain independent system for WSD would probably need a corpus of about 3.2 million sense tagged words. At a throughput of one word per minute (Edmonds, 2000), this would require about 27 man-years of human annotation effort.

With *Open Mind Word Expert* we aim at creating a very large sense tagged corpus, by making use of the incredible resource of knowledge constituted by the millions of Web users, combined with techniques for active learning.

3 Open Mind Word Expert

Open Mind Word Expert is a Web-based interface where users can tag words with their WordNet senses. Tagging is organized by word. That is, 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.

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 users 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, from a drop-down list that contains all WordNet senses, 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 Open Mind Word Expert may allow contributors to specify more than one sense for any word.

A prototype of the system has been implemented and is available at http://www.teach-computers.org. Figure 1 shows a screen shot from the system interface, illustrating the screen presented to users when tagging the noun "child".

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[1], and *Open Mind Common Sense*[2], a collection of about 400,000 commonsense assertions in English as contributed by volunteers over the Web. A mix of several sources, each covering a different spectrum of usage, is

[1] http://trec.nist.gov

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

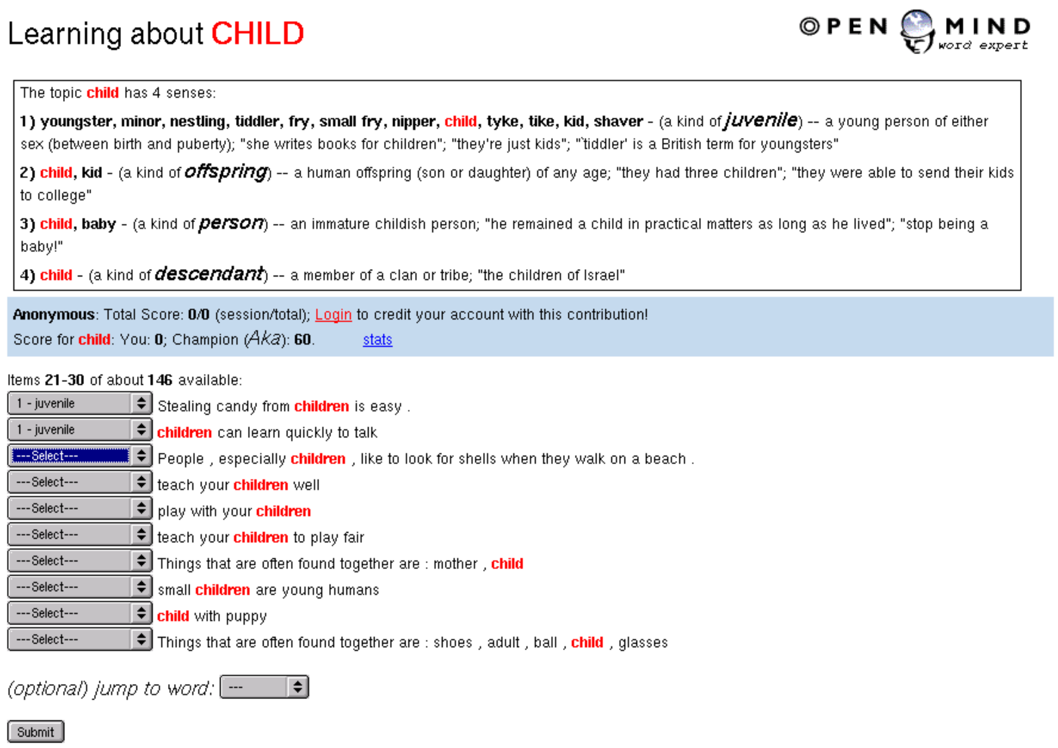

Figure 1: Screen shot from *Open Mind Word Expert*

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. For example, the collection includes such assertions as “keys are used to unlock doors”, and “pressing a typewriter key makes a letter”. We believe these sentences may be a relatively clean source of keywords that can aid in disambiguation. For details on the data and how it has been collected, see (Singh, 2002).

3.2 Active Learning

To minimize the amount of human annotation effort needed to build a tagged corpus for a given ambiguous word, *Open Mind Word Expert* includes 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 *Open Mind Word Expert* system itself, when no other training data is available.

The first classifier is a Semantic Tagger with Active Feature Selection (STAFS). This system (previously known as SMUIs) is one of the top ranked systems in the *English lexical sample* task at SENSEVAL-2. 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. More details on this system can be found in (Mihalcea, 2002b).

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 figure that is close to the 85.5% inter-annotator agreement measured for the SENSEVAL-2 nouns (Kilgarriff, 2002).

System	Precision (fine grained)	(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

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". In *Open Mind Word Expert*, these are the instances that are presented to the users for tagging in the active learning stage.

3.3 Ensuring Quality

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.

We have several steps already implemented and have additional steps we propose to ensure quality.

First, redundant tagging is collected for each item. *Open Mind Word Expert* 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.

Second, inaccurate sessions will be discarded. This can be accomplished in two ways, roughly by checking agreement and precision:

- Using redundancy of tags collected for each item, any given session (a tagging done all in one sitting) will be checked for agreement with tagging of the same items collected outside of this session.

- If necessary, the precision of a given contributor with respect to a preexisting gold standard (such as SemCor or Senseval training data) can be estimated directly by presenting the contributor with examples from the gold standard. This will be implemented if there are indications of need for this in the pilot; it will help screen out contributors who, for example, always select the first sense (and are in high agreement with other contributors who do the same).

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

Additionally, the tagging quality will be estimated by comparing the agreement level among Web contributors with the agreement level that was already measured in previous sense tagging projects. An analysis of the semantic annotation task performed by novice taggers as part of the SemCor project (Fellbaum et al., 1997) revealed an agreement of about 82.5% among novice taggers, and 75.2% among novice taggers and lexicographers.

Moreover, since we plan to use paid, trained taggers to create a separate test corpus for each of the words tagged with *Open Mind Word Expert*, these same paid taggers could also validate a small percentage of the training data for which no gold standard exists.

3.4 Engaging the Contributors

We believe that making the contribution process as engaging and as "game-like" for the contributors as possible is crucial to collecting a large volume of data. With that goal, Open Mind Word Expert tracks, for each contributor, the number of items tagged for each topic. When tagging items, a contributor is shown the number of items (for this word) she has tagged and the record number of items tagged (for this word) by a single user.

If the contributor sets a record, it is recognized with a congratulatory message on the contribution screen, and the user is placed in the Hall of Fame for the site. Also, the user can always access a real-time graph summarizing, by topic, their contribution versus the current record for that topic.

Interestingly, it seems that relatively simple word games can enjoy tremendous user acceptance. For example, WordZap (http://wordzap.com), a game that pits players against each other or against a computer to be the first to make seven words from several presented letters (with some additional rules), has been downloaded by well over a million users, and the reviewers describe the game as "addictive". If sense tagging can enjoy a fraction of such popularity, very large tagged corpora will be generated.

Additionally, NLP instructors can use the site as an *aid in teaching lexical semantics*. An instructor can create an "activity code", and then, for users who have opted in as participants of that activity (by entering the activity code when creating their profiles), access the amount tagged by each participant, and the percentage agreement of the tagging of each contributor who opted in for this activity. Hence, instructors can assign Open Mind Word Expert tagging as part of a homework assignment or a test.

Also, assuming there is a test set of already tagged examples for a given ambiguous word, we may add the capability of showing the increase in disambiguation precision on the test set, as it results from the samples that a user is currently tagging.

4 Proposed Task for SENSEVAL-3

The *Open Mind Word Expert* system will be used to build large sense tagged corpora for some of the most frequent ambiguous words in English. The tagging will be collected over the Web from volunteer contributors. We propose to organize a task in SENSEVAL-3 where systems will disambiguate words using the corpus created with this system.

We will initially select a set of 100 nouns, and collect for each of them $75 + n * 15$ tagged samples (Edmonds, 2000), where n is the number of senses of the noun. It is worth mentioning that, unlike previous SENSEVAL evaluations, where multi-word expressions were considered as possible senses for an constituent ambiguous word, we filter these expressions apriori with an automatic tool for collocation extraction. Therefore, the examples we collect refer only to single ambiguous words, and hence we expect a lower inter-tagger agreement rate and lower WSD tagging precision when only single words are used, since usually multi-word expressions are not ambiguous and they constitute some of the "easy cases" when doing sense tagging.

These initial set of tagged examples will then be used to train the two classifiers described in Section 3.2, and annotate an additional set of $75+n*35$ examples. From these, the users will be presented only with those examples where there is a disagreement between the labels assigned by the two classifiers. The final corpus for each ambiguous word will be created with (1) the original set of $75 + n * 15$ tagged examples, plus (2) the examples selected by the active learning component, sense tagged by users.

Words will be selected based on their frequencies, as computed on SemCor. Once the tagging process of the initial set of 100 words is completed, additional nouns will be incrementally added to the *Open Mind Word Expert* interface. As we go along, words with other parts of speech will be considered as well.

To enable comparison with Senseval-2, the set of words will also include the 29 nouns used in the Senseval-2 lexical sample tasks. This would allow us to assess how much the collected data helps on the Senseval-2 task.

As shown in Section 3.3, redundant tags will be collected for each item, and overall quality will be assessed. Moreover, starting with the initial set of $75 + n * 15$ examples labeled for each word, we will create confusion matrices that will indicate the similarity between word senses, and help us create the sense mappings for the coarse grained evaluations.

One of the next steps we plan to take is to replace the "two tags per item" scheme with the "tag until at least two tags agree" scheme proposed and used during the SENSEVAL-2 tagging (Kilgarriff, 2002). Additionally, the set of meanings that constitute the possible choices for a certain ambiguous example will be enriched with groups of similar meanings, which will be determined either based on some apriori provided sense mappings (if any available) or based on the confusion matrices mentioned above.

For each word with sense tagged data created with *Open Mind Word Expert*, a test corpus will be built by trained human taggers, starting with examples extracted from the corpus mentioned in Section 3.1. This process will be set up independently of the *Open Mind Word Expert* Web interface. The test corpus will be released during SENSEVAL-3.

5 Conclusions and future work

Open Mind Word Expert pursues the potential of creating a large tagged corpus. WSD can also benefit in other ways from the *Open Mind* approach. We are considering using a AutoASC/GenCor type of approach to generate sense tagged data with a bootstrapping algorithm (Mihalcea, 2002a). Web contributors can help this process by creating the initial set of seeds, and exercising control over the quality of the automatically generated seeds.

Acknowledgments
We would like to thank the Open Mind Word Expert contributors who are making all this work possible. We are also grateful to Adam Kilgarriff for valuable suggestions and interesting discussions, to Randall Davis and to the anonymous reviewers for useful comments on an earlier version of this paper, and to all the Open Mind Word Expert users who have emailed us with their feedback and suggestions, helping us improve this activity.

References

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.

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.

C. Fellbaum, J. Grabowski, and S. Landes. 1997. Analysis of a hand-tagging task. In *Proceedings of ANLP-97 Workshop on Tagging Text with Lexical Semantics: Why, What, and How?*, Washington D.C.

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, editor. 2001. *SENSEVAL-2*, Toulouse, France, November.

A. Kilgarriff. 2002. English lexical sample task description. In *Proceedings of Senseval-2, ACL Workshop*.

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. 2002a. Bootstrapping large sense tagged corpora. In *Proceedings of the Third International Conference on Language Resources and Evaluation LREC 2002*, Canary Islands, Spain, May. (to appear).

R. Mihalcea. 2002b. 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. (to appear).

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.

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.

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.

D. Stork. 1999. The Open Mind initiative. *IEEE Expert Systems and Their Applications*, 14(3):19–20.