SENSEVAL-3: Third International Workshop on the Evaluation of Systems
for the Semantic Analysis of Text, Barcelona, Spain, July 2004
Association for Computational Linguistics

SENSEVAL-3 TASK
Word-Sense Disambiguation of WordNet Glosses

Kenneth C. Litkowski
CL Research
9208 Gue Road
Damascus, MD 20872
ken@clres.com

Abstract

The SENSEVAL-3 task to perform word-sense disambiguation of WordNet glosses was designed to encourage development of technology to make use of standard lexical resources. The task was based on the availability of sense-disambiguated hand-tagged glosses created in the eXtended WordNet project. The hand-tagged glosses provided a "gold standard" for judging the performance of automated disambiguation systems. Seven teams participated in the task, with a total of 10 runs. Scoring these runs as an "all-words" task, along with considerable discussions among participants, provided more insights than just the underlying technology. The task identified several issues about the nature of the WordNet sense inventory and the underlying use of wordnet design principles, particularly the significance of WordNet-style relations.

Introduction

In SENSEVAL-2, performance in the lexical sample task dropped considerably (Kilgarriff, 2001). Kilgarriff suggested that using WordNet (Fellbaum, 1998) for SENSEVAL has drawbacks. WordNet was not designed to serve as a lexical resource, but its public availability and reasonable comprehensiveness have been dominant factors in its selection as the lexical resource of choice for Senseval and for many applications. These factors have led to further funding by U.S. government agencies and many improvements are currently underway. Among these improvements is a planned hand-tagging of the WordNet glosses with their WordNet senses. At the same time, sense-tagging of the glosses is being performed in the Extended WordNet (XWN) project under development at the University of Texas at Dallas (Mihalcea and Moldovan, 2001)[1]. The XWN project also parses the WordNet glosses into a part of speech tree and transforms them into a logical predicate form.

More generally, sense disambiguation of definitions in any lexical resource is an important objective in the language engineering community. The first significant disambiguation of dictionary definitions and creation of a hierarchy took place 25 years ago in the groundbreaking work of Amsler (1980). However, while substantial research has been performed on machine-readable dictionaries since that time, technology has not yet been developed to make systematic use of these resources. This SENSEVAL task was designed to encourage the lexical research community to take up the challenge of disambiguating dictionary definitions.

XWN is used as a core knowledge base for applications such as question answering, information retrieval, information extraction, summarization, natural language generation, inferences, and other knowledge intensive applications. The glosses contain a part of the world knowledge since they define the most common concepts of the English language. In the XWN project, many open-class words in WordNet glosses have been hand-tagged and provide a "gold standard" against which disambiguation systems can be judged. The SENSEVAL-3 task is to replicate the hand-tagged results.

The Extended WordNet (XWN) project has disambiguated the content words (nouns, verbs, adjectives, and adverbs) of all glosses, combining human annotation and automated methods using WordNet 1.7.1. A "quality" attribute was given to each lemma. XWN used two automatic systems to disambiguate the content words. When the two systems agreed, the lemma was given a "silver"

[1] http://www.hlt.utdallas.edu/

quality. Otherwise, a lemma was given a “normal” quality (even when there was only one sense in WordNet). In a complex process described in more detail below, certain glosses or lemmas were selected for hand annotation. Lemmas which were hand-tagged were given a “gold” tag.

The WordNet 1.7.1 data were next converted to use WordNet 2.0 glosses. Word senses have been assigned to 630,599 open class words, with 15,179 (less than 2.5 percent) of the open-class words in these glosses assigned manually. Many glosses have more than one word given a “gold” assignment. The resultant test set provided to participants consists of 9,257 glosses, containing 15,179 “gold” tagged content words and a total of 42,491 content words, distributed as follows:

Table 1. Gloss Test Set

POS	Glosses	Golds	Words
Adjective	94	263	370
Adverb	1684	1826	3719
Noun	6706	10985	35539
Verb	773	2105	2863
Total	9257	15179	42491

The disambiguations (and hence the answer key) are available at the XWN web site. Participants were encouraged to investigate the XWN data as well as the methods followed by the XWN team. However, participants were expected to develop their own systems, for comparison with the XWN manual annotations.

1 The Senseval-3 Task

Participants were provided with all glosses from WordNet in which at least one open-class word was given a "Gold" quality assignment. These glosses were provided in an XML file, each with its WordNet synset number, its part of speech, and the gloss itself. Glosses frequently include sample uses. The samples uses were not parsed in the XWN project and were not to be included in the submissions.

The task was configured as essentially identical to the SENSEVAL-2 and SENSEVAL-3 "all-words" tasks, except without any context and with the gloss not constituting a complete sentence. Unlike the all-words task, individual tokens to be disambiguated were not identified, so that participants were required to perform their own tokenization and identification of multiword units. The number of words in a gloss is quite small, but a few glosses do contain the same word more than once. Participants were encouraged to consider a synset's placement within WordNet (its hypernyms, hyponyms, and other relations) to assist in disambiguation. The XWN data contains part of speech tags for each word in the glosses, as well as parses and logical forms, which participants were allowed to use. Most of the glosses in the test set have hand-tagged words as well as words tagged by the automatic XWN systems. The senses assigned to other open-class words have a tag of “silver” or “normal”. In submitting test runs, participants did not know which of the words had been assigned a “gold” quality, but were only scored for the “gold” quality words.[2]

No training data was available for this task since the number of items in the test set was so small. Participants were encouraged to become familiar with the XWN dataset and to make use of it in ways that would not compromise their performance of the task.

2 Submissions

Seven teams participated in the task with one team submitting two runs and one team submitting three runs. A submission contained an identifier (the part of speech of the gloss and its synset number) and a WordNet sense for each content word or phrase identified by the system. The answer key contains part of speech/synset identifier, the XWN quality assignment, the lemma and the word form from the XWN data, and the WordNet sense. The scoring program (a Perl script) stored the answers in three hashes according to quality (“gold”, “silver”, and “normal”) and then also stored the system’s answers in a hash. The program then proceeded through the “gold” answers and determined if a system’s answers included a match for that answer, equaling either the (lemma, sense) or (word form, sense). No system submitted more than one sense for each of its word forms. An exact match received a score of 1.0. If a

[2] The answer key contains all assignments, so it is possible that runs can be analyzed with these other sense assignments with a voting system. However, such an analysis has not yet been performed.

system returned either the lemma or the word form, but had assigned an incorrect sense, the item was counted as attempted.

Precision was computed as the number correct divided by the number attempted. **Recall** was computed as the number correct divided by the total number of "gold" items. The **percent attempted** was computed as the number attempted divided by the total number of "gold" items. Results for all runs are shown in Table 2.

Table 2. System Performance (All Items)

Run	Prec	Rec	Att
01 (UPolitécnica de Valencia)	0.534	0.405	76.0
02 (CL Research)	0.449	0.345	76.8
03 (LanguageComputerCorp)	0.701	0.504	71.9
04a (TALP Research Center)	0.686	0.683	99.5
04b (TALP Research Center)	0.574	0.558	97.2
05 (IRIT-ERSS)	0.388	0.385	99.1
06a (Uni-Roma1-DI)	0.777	0.311	40.0
06b (Uni-Roma1-DI)	0.668	0.667	99.9
06c (Uni-Roma1-DI)	0.716	0.362	50.5
07 (Indian Inst Technology)	0.343	0.301	87.8

Systems 04a and 06b used the part of speech tags available in the XWN files, while the other runs did not.

3 Discussion

During discussions on the SENSEVAL-3 mailing list and in interchanges assessing the scoring of the systems, several issues of some importance arose. Most of these concerned the nature of the XWN annotation process and the "correctness" of the "gold" quality assignments.

Since glosses (or definitions) are only "sentence" fragments, parsing them poses some inherent difficulties. In theory, a proper lexicographically-based definition is one that contains a genus term (hypernym or superordinate) and differentiae. A gloss' hypernym is somewhat easily identified as the head of the first phrase, particularly in noun and verb definitions. Since most WordNet synsets have a hypernym, a heuristic for disambiguating the head of the first phrase would be to use the hypernym as the proper disambiguated sense. And, indeed, the instructions for the task encouraged participants to make use of WordNet relations in their disambiguation.

However, the XWN annotators were not given this heuristic, but rather were presented with the set of WordNet senses without awareness of the WordNet relations. As a result, many glosses had "gold" assignments that seemed incorrect when considering WordNet's own hierarchy. For example, *naught* is defined as "complete failure"; in WordNet, its hypernym *failure* is sense 1 ("an act that fails"), but the XWN annotators tagged it with sense 2 ("an event that does not accomplish its intended purpose").

To investigate the use of WordNet relations heuristics, we considered a set of 313 glosses containing 867 "gold" assignments which team 06 submitted as highly reliant on these relations. As shown in Table 3 (scored on 8944 glosses with 14312 "gold" assignments), precision scores changed most for 03 (0.020), 06b (0.017), and 04a (0.016); these runs had correspondingly much lower scores for the 313 glosses in this set (results not shown). These differences do not appear to be significant. A more complete assessment of the significance of WordNet relations in disambiguation would require a more complete identification of glosses where systems relied on such information.

Table 3. System Performance (Reduced Set)

Run	Prec	Rec	Att
01 (UPolitécnica de Valencia)	0.538	0.407	75.6
02 (CL Research)	0.446	0.342	76.6
03 (LanguageComputerCorp)	0.721	0.516	71.6
04a (TALP Research Center)	0.702	0.698	99.5
04b (TALP Research Center)	0.585	0.568	97.2
05 (IRIT-ERSS)	0.395	0.391	99.1
06a (Uni-Roma1-DI)	0.826	0.323	39.1
06b (Uni-Roma1-DI)	0.685	0.684	99.9
06c (Uni-Roma1-DI)	0.753	0.375	49.7
07 (Indian Inst Technology)	0.346	0.302	87.2

Further discussion with members of the XWN project about the annotation process revealed some factors that should be taken into account when assessing the various systems' performances. Firstly, the annotations of the 9257 glosses with "gold" assignments were annotated using three different methods. The first group of 1032 glosses were fully hand-tagged by two graduate students, with 80 percent agreement and with the project leader choosing a sense when there was disagreement.

For the remaining glosses in WordNet, two automated disambiguation programs were run. When both programs agreed on a sense, they were given a "silver" quality. In those glosses for which all but one or two words had been assigned a "silver" quality, the one or two words were hand-tagged by a graduate student, without any interannotator check or review. There are 4077 noun glosses in this second set.

A third set, the remaining 4738 among the test set, were glosses for which all the words but one had been assigned a "silver" quality. The single word was then hand-tagged by a graduate student, and in some cases by the project leader (particularly when a word had been mistagged by the Brill tagger).

To assess the effect of these three different styles of annotation, we ran the scoring program, restricting the items scored to those in each of the three annotation sets. The scores were changed much more significantly for the various teams for the different sets. For the first set, precision was down approximately 0.07 for three runs, with much lower changes for the other runs. For the second set, precision was up approximately 0.075 for two runs, down approximately 0.08 for two runs, and relatively unchanged for the remaining runs. For the third set, there was relatively little changes in the precision for all runs (with a maximum change of 0.03).

4 Conclusions

The underlying guidance for this SENSEVAL-3 task that, in the absence of significant context, participants make use of WordNet relations for disambiguating glosses has led to some significant insights about the use and importance of wordnets. These insights emerge from the tension between the reliance on WordNet relations and the imprecision of the tagging process.

Many investigators, including several of the participants in this task, are attempting to exploit the kinds of relations between lexical entries that are embodied in WordNet. The use of wordnets in NLP applications has become an important basic construct and increasingly valuable. However, the construction of wordnets is expensive and time-consuming, and without any significant prospects for commercial support. While some dictionary publishers are increasingly incorporating wordnet principles into their lexical resources, this process is slow. At present, the publicly available WordNet remains the wordnet of choice.

The annotation process followed by the XWN project, with the taggings used in this task, has again indicated difficulties with the WordNet sense inventory. The fact remains that WordNet has not had the benefit of sufficient lexicographic resources in the construction of its glosses and in the acquisition of other lexicographic information in its entries. The WordNet project continues its efforts to add information, but with limited resources.

With the diverse set of approaches represented by the participants in this task, it is possible to envision sets of steps that might be employed to improve the details of the WordNet sense inventory. One step would include continued hand-tagging of WordNet glosses without consideration of WordNet relations. Another step would be the use of automated disambiguation routines to act as checks on consistency. Such systems would include those that rely on WordNet relations as well as those that do not, acting as checks on one another.

References

Amsler, Robert A. 1980. *The Structure of the Merriam-Webster Pocket Dictionary*. Ph.D. Thesis., Austin: University of Texas.

Fellbaum, Christiane (ed.). 1998. *WordNet: An Electronic Lexical Database*. The MIT Press: Cambridge, MA.

Kilgarriff, Adam. 2001. English Lexical Sample Task Description. In *Proceedings of SENSEVAL-2: Second International Workshop on Evaluating Word Sense Disambiguation Systems*, Toulouse, France.

Mihalcea, Rada and Dan Moldovan. 2001. eXtended WordNet: Progress Report. In *Proceedings of NAACL Workshop on WordNet and Other Lexical Resources*, Pittsburgh, PA.