The SENSEVAL–3 English Lexical Sample Task

Rada Mihalcea
Department of Computer Science
University of North Texas
Dallas, TX, USA
rada@cs.unt.edu

Timothy Chklovski
Information Sciences Institute
University of Southern California
Marina del Rey, CA, USA
timc@isi.edu

Adam Kilgarriff
Information Technology Research Institute
University of Brighton
Brighton, UK
Adam.Kilgarriff@itri.brighton.ac.uk

Abstract

This paper presents the task definition, resources, participating systems, and comparative results for the English lexical sample task, which was organized as part of the SENSEVAL-3 evaluation exercise. The task drew the participation of 27 teams from around the world, with a total of 47 systems.

1 Introduction

We describe in this paper the task definition, resources, participating systems, and comparative results for the English lexical sample task, which was organized as part of the SENSEVAL-3 evaluation exercise. The goal of this task was to create a framework for evaluation of systems that perform targeted Word Sense Disambiguation.

This task is a follow-up to similar tasks organized during the SENSEVAL-1 (Kilgarriff and Palmer, 2000) and SENSEVAL-2 (Preiss and Yarowsky, 2001) evaluations. The main changes in this year's evaluation consist of a new methodology for collecting annotated data (with contributions from Web users, as opposed to trained lexicographers), and a new sense inventory used for verb entries (Wordsmyth).

2 Building a Sense Tagged Corpus with Volunteer Contributions over the Web

The sense annotated corpus required for this task was built using the Open Mind Word Expert system (Chklovski and Mihalcea, 2002) [1]. To overcome the current lack of sense tagged data and the limitations imposed by the creation of such data using trained lexicographers, the OMWE system 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.

The tagging exercise proceeds as follows. For each target word the system extracts a set of sentences from a large textual corpus. These examples are presented to the contributors, who are asked to select the most appropriate sense for the target word in each sentence. The selection is made using checkboxes, which list all possible senses of the current target word, plus two additional choices, "unclear" and "none of the above." Although users are encouraged to select only one meaning per word, the selection of two or more senses is also possible. The results of the classification submitted by other users are not presented to avoid artificial biases.

Similar to the annotation scheme used for the English lexical sample at SENSEVAL-2, we use a "tag until two agree" scheme, with an upper bound on the number of annotations collected for each item set to four.

2.1 Source Corpora

The data set used for the SENSEVAL-3 English lexical sample task consists of examples extracted from the British National Corpus (BNC). Earlier versions of OMWE also included data from the *Penn Treebank* corpus, the *Los Angeles Times* collection as provided during TREC conferences (http://trec.nist.gov), and *Open Mind Common Sense* (http://commonsense.media.mit.edu).

2.2 Sense Inventory

The sense inventory used for nouns and adjectives is WordNet 1.7.1 (Miller, 1995), which is consistent with the annotations done for the same task during SENSEVAL-2. Verbs are instead annotated with senses from Wordsmyth (http://www.wordsmyth.net/). The main reason motivating selection of a different sense inventory is the

[1] Open Mind Word Expert can be accessed at http://teach-computers.org/

Class	Nr of words	Avg senses (fine)	Avg senses (coarse)
Nouns	20	5.8	4.35
Verbs	32	6.31	4.59
Adjectives	5	10.2	9.8
Total	57	6.47	4.96

Table 1: Summary of the sense inventory

weak verb performance of systems participating in the English lexical sample in SENSEVAL-2, which may be due to the high number of senses defined for verbs in the WordNet sense inventory. By choosing a different set of senses, we hope to gain insight into the dependence of difficulty of the sense disambiguation task on sense inventories.

Table 1 presents the number of words under each part of speech, and the average number of senses for each class.

2.3 Multi-Word Expressions

For this evaluation exercise, we decided to isolate the task of semantic tagging from the task of identifying multi-word expressions; we applied a filter that removed all examples pertaining to multi-word expressions prior to the tagging phase. Consequently, the training and test data sets made available for this task do not contain collocations as possible target words, but only single word units. This is a somewhat different definition of the task as compared to previous similar evaluations; the difference may have an impact on the overall performance achieved by systems participating in the task.

2.4 Sense Tagged Data

The inter-tagger agreement obtained so far is closely comparable to 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 taggings (taken in order of submission) entered for each item. About 12% of that 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 between the first two taggings 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. Note that these figures are collected for the entire OMWE data set build so far, which consists of annotated data for more than 350 words.

In addition to raw inter-tagger agreement, the kappa statistic, which removes from the agreement rate the amount of agreement that is expected by chance(Carletta, 1996), was also determined. We measure two figures: *micro-average* κ, where number of senses, agreement by chance, and κ are determined as an average for all words in the set, and *macro-average* κ, where inter-tagger agreement, agreement by chance, and κ are individually determined for each of the words in the set, and then combined in an overall average. With an average of five senses per word, the average value for the agreement by chance is measured at 0.20, resulting in a *micro-*κ statistic of 0.58. For *macro-*κ estimations, we assume that word senses follow the distribution observed in the OMWE annotated data, and under this assumption, the *macro-*κ is 0.35.

3 Participating Systems

27 teams participated in this word sense disambiguation task. Tables 2 and 3 list the names of the participating systems, the corresponding institutions, and the name of the first author – which can be used as reference to a paper in this volume, with more detailed descriptions of the systems and additional analysis of the results.

There were no restrictions placed on the number of submissions each team could make. A total number of 47 submissions were received for this task. Tables 2 and 3 show all the submissions for each team, gives a brief description of their approaches, and lists the precision and recall obtained by each system under fine and coarse grained evaluations.

The precision/recall baseline obtained for this task under the "most frequent sense" heuristic is 55.2% (fine grained) and 64.5% (coarse grained). The performance of most systems (including several unsupervised systems, as listed in Table 3) is significantly higher than the baseline, with the best system performing at 72.9% (79.3%) for fine grained (coarse grained) scoring.

Not surprisingly, several of the top performing systems are based on combinations of multiple classifiers, which shows once again that voting schemes that combine several learning algorithms outperform the accuracy of individual classifiers.

4 Conclusion

The English lexical sample task in SENSEVAL-3 featured English ambiguous words that were to be tagged with their most appropriate WordNet or Wordsmyth sense. The objective of this task was to: (1) Determine feasibility of reliably finding the

System/Team	Description	Fine		Coarse	
		P	R	P	R
htsa3 U.Bucharest (Grozea)	A Naive Bayes system, with correction of the a-priori frequencies, by dividing the output confidence of the senses by $frequency^{\alpha}$ ($\alpha = 0.2$)	72.9	72.9	79.3	79.3
IRST-Kernels ITC-IRST (Strapparava)	Kernel methods for pattern abstraction, paradigmatic and syntagmatic info. and unsupervised term proximity (LSA) on BNC, in an SVM classifier.	72.6	72.6	79.5	79.5
nusels Nat.U. Singapore (Lee)	A combination of knowledge sources (part-of-speech of neighbouring words, words in context, local collocations, syntactic relations), in an SVM classifier.	72.4	72.4	78.8	78.8
htsa4	Similar to htsa3, with different correction function of a-priori frequencies.	72.4	72.4	78.8	78.8
BCU_comb Basque Country U. (Agirre & Martinez)	An ensemble of decision lists, SVM, and vectorial similarity, improved with a variety of smoothing techniques. The features consist of local collocations, syntactic dependencies, bag-of-words, domain features.	72.3	72.3	78.9	78.9
htsa1	Similar to htsa3, but with smaller number of features.	72.2	72.2	78.7	78.7
rlsc-comb U.Bucharest (Popescu)	A regularized least-square classification (RLSC), using local and topical features, with a term weighting scheme.	72.2	72.2	78.4	78.4
htsa2	Similar to htsa4, but with smaller number of features.	72.1	72.1	78.6	78.6
BCU_english	Similar to BCU_comb, but with a vectorial space model learning.	72.0	72.0	79.1	79.1
rlsc-lin	Similar to rlsc-comb, with a linear kernel, and a binary weighting scheme.	71.8	71.8	78.4	78.4
HLTC_HKUST_all HKUST (Carpuat)	A voted classifier combining a new kernel PCA method, a Maximum Entropy model, and a boosting-based model, using syntactic and collocational features	71.4	71.4	78.6	78.6
TALP U.P.Catalunya (Escudero et al.)	A system with per-word feature selection, using a rich feature set. For learning, it uses SVM, and combines two binarization procedures: one vs. all, and constraint learning.	71.3	71.3	78.2	78.2
MC-WSD Brown U. (Ciaramita & Johnson)	A multiclass averaged perceptron classifier with two components: one trained on the data provided, the other trained on this data, and on WordNet glosses. Features consist of local and syntactic features.	71.1	71.1	78.1	78.1
HLTC_HKUST_all2	Similar to HLTC_HKUST_all, also adds a Naive Bayes classifier.	70.9	70.9	78.1	78.1
NRC-Fine NRC (Turney)	Syntactic and semantic features, using POS tags and pointwise mutual information on a terabyte corpus. Five basic classifiers are combined with voting.	69.4	69.4	75.9	75.9
HLTC_HKUST_me	Similar to HLTC_HKUST_all, only with a maximum entropy classifier.	69.3	69.3	76.4	76.4
NRC-Fine2	Similar to NRC-Fine, with a different threshold for dropping features	69.1	69.1	75.6	75.6
GAMBL U. Antwerp (Decadt)	A cascaded memory-based classifier, using two classifiers based on global and local features, with a genetic algorithm for parameter optimization.	67.4	67.4	74.0	74.0
SinequaLex Sinequa Labs (Crestan)	Semantic classification trees, built on short contexts and document semantics, plus a decision system based on information retrieval techniques.	67.2	67.2	74.2	74.2
CLaC1 Concordia U. (Lamjiri)	A Naive Bayes approach using a context window around the target word, which is dynamically adjusted	67.2	67.2	75.1	75.1
SinequaLex2	A cumulative method based on scores of surrounding words.	66.8	66.8	73.6	73.6
UMD_SST4 U. Maryland (Cabezas)	Supervised learning using Support Vector Machines, using local and wide context features, and also grammatical and expanded contexts.	66.0	66.0	73.7	73.7
Prob1 Cambridge U. (Preiss)	A probabilistic modular WSD system, with individual modules based on separate known approaches to WSD (26 different modules)	65.1	65.1	71.6	71.6
SyntaLex-3 U.Toronto (Mohammad)	A supervised system that uses local part of speech features and bigrams, in an ensemble classifier using bagged decision trees.	64.6	64.6	72.0	72.0
UNED UNED (Artiles)	A similarity-based system, relying on the co-occurrence of nouns and adjectives in the test and training examples.	64.1	64.1	72.0	72.0
SyntaLex-4	Similar to SyntaLex-3, but with unified decision trees.	63.3	63.3	71.1	71.1
CLaC2	Syntactic and semantic (WordNet hypernyms) information of neighboring words, fed to a Maximum Entropy learner. See also CLaC1	63.1	63.1	70.3	70.3
SyntaLex-1	Bagged decision trees using local POS features. See also SyntaLex-3.	62.4	62.4	69.1	69.1
SyntaLex-2	Similar to SyntaLex-1, but using broad context part of speech features.	61.8	61.8	68.4	68.4
Prob2	Similar to Prob1, but invokes only 12 modules.	61.9	61.9	69.3	69.3
Duluth-ELSS U.Minnesota (Pedersen)	An ensemble approach, based on three bagged decision trees, using unigrams, bigrams, and co-occurrence features	61.8	61.8	70.1	70.1
UJAEN U.Jaén (García-Vega)	A Neural Network supervised system, using features based on semantic relations from WordNet extracted from the training data	61.3	61.3	69.5	69.5
R2D2 U. Alicante (Vazquez)	A combination of supervised (Maximum Entropy, HMM Models, Vector Quantization, and unsupervised (domains and conceptual density) systems.	63.4	52.1	69.7	57.3
IRST-Ties ITC-IRST (Strapparava)	A generalized pattern abstraction system, based on boosted wrapper induction, using only few syntagmatic features.	70.6	50.5	76.7	54.8
NRC-Coarse	Similar to NRC-Fine; maximizes the coarse score, by training on coarse senses.	48.5	48.5	75.8	75.8
NRC-Coarse2	Similar to NRC-Coarse, with a different threshold for dropping features.	48.4	48.4	75.7	75.7
DLSI-UA-LS-SU U.Alicante (Vazquez)	A maximum entropy method and a bootstrapping algorithm ("re-training") with, iterative feeding of training cycles with new high-confidence examples.	78.2	31.0	82.8	32.9

Table 2: Performance and short description of the supervised systems participating in the SENSEVAL-3 English lexical sample Word Sense Disambiguation task. Precision and recall figures are provided for both fine grained and coarse grained scoring. Corresponding team and reference to system description (in this volume) are indicated for the first system for each team.

System/Team	Description	Fine		Coarse	
		P	R	P	R
wsdiit IIT Bombay (Ramakrishnan et al.)	An unsupervised system using a Lesk-like similarity between context of ambiguous words, and dictionary definitions. Experiments are performed for various window sizes, various similarity measures	66.1	65.7	73.9	74.1
Cymfony (Niu)	A Maximum Entropy model for unsupervised clustering, using neighboring words and syntactic structures as features. A few annotated instances are used to map context clusters to WordNet/Worsmyth senses.	56.3	56.3	66.4	66.4
Prob0 Cambridge U. (Preiss)	A combination of two unsupervised modules, using basic part of speech and frequency information.	54.7	54.7	63.6	63.6
clr04-ls CL Research (Litkowski)	An unsupervised system relying on definition properties (syntactic, semantic, subcategorization patterns, other lexical information), as given in a dictionary. Performance is generally a function of how well senses are distinguished.	45.0	45.0	55.5	55.5
CIAOSENSO U. Genova (Buscaldi)	An unsupervised system that combines the conceptual density idea with the frequency of words to disambiguate; information about domains is also taken into account.	50.1	41.7	59.1	49.3
KUNLP Korea U. (Seo)	An algorithm that disambiguates the senses of a word by selecting a substituent among WordNet relatives (antonyms, hypernyms, etc.). The selection is done based on co-occurrence frequencies, measured on a large corpus.	40.4	40.4	52.8	52.8
Duluth-SenseRelate U.Minnesota (Pedersen)	An algorithm that assigns the sense to a word that is most related to the possible senses of its neighbors, using WordNet glosses to measure relatedness between senses.	40.3	38.5	51.0	48.7
DFA-LS-Unsup UNED (Fernandez)	A combination of three heuristics: similarity between synonyms and the context, according to a mutual information measure; lexico-syntactic patterns extracted from WordNet glosses; the first sense heuristic.	23.4	23.4	27.4	27.4
DLSI-UA-LS-NOSU U.Alicante (Vazquez)	An unsupervised method based on (Magnini & Strapparava 2000) WordNet domains; it exploits information contained in glosses of WordNet domains, and uses "Relevant Domains", obtained from association ratio over domains and words.	19.7	11.7	32.2	19.0

Table 3: Performance and short description for the Unsupervised systems participating in the SENSEVAL-3 English lexical sample task.

appropriate sense for words with various degrees of polysemy, using different sense inventories; and (2) Determine the usefulness of sense annotated data collected over the Web (as opposed to other traditional approaches for building semantically annotated corpora).

The results of 47 systems that participated in this event tentatively suggest that supervised machine learning techniques can significantly improve over the most frequent sense baseline, and also that it is possible to design unsupervised techniques for reliable word sense disambiguation. Additionally, this task has highlighted creation of testing and training data by leveraging the knowledge of Web volunteers. The training and test data sets used in this exercise are available online from http://www.senseval.org and http://teach-computers.org.

Acknowledgments

Many thanks to all those who contributed to the Open Mind Word Expert project, making this task possible. In particular, we are grateful to Gwen Lenker – our most productive contributor. We are also grateful to all the participants in this task, for their hard work and involvement in this evaluation exercise. Without them, all these comparative analyses would not be possible.

We are indebted to the Princeton WordNet team, for making WordNet available free of charge, and to Robert Parks from Wordsmyth, for making available the verb entries used in this evaluation.

We are particularly grateful to the National Science Foundation for their support under research grant IIS-0336793, and to the University of North Texas for a research grant that provided funding for contributor prizes.

References

J. Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

T. Chklovski and R. Mihalcea. 2002. Building a sense tagged corpus with Open Mind Word Expert. In *Proceedings of the ACL 2002 Workshop on "Word Sense Disambiguation: Recent Successes and Future Directions"*, Philadelphia, July.

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

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

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