Semeval-2007 Task 02:
Evaluating Word Sense Induction and Discrimination Systems

Eneko Agirre
IXA NLP Group
Univ. of the Basque Country
Donostia, Basque Country
e.agirre@ehu.es

Aitor Soroa
IXA NLP Group
Univ. of the Basque Country
Donostia, Basque Country
a.soroa@ehu.es

Abstract

The goal of this task is to allow for comparison across sense-induction and discrimination systems, and also to compare these systems to other supervised and knowledge-based systems. In total there were 6 participating systems. We reused the SemEval-2007 English lexical sample subtask of task 17, and set up both clustering-style unsupervised evaluation (using OntoNotes senses as gold-standard) and a supervised evaluation (using the part of the dataset for mapping). We provide a comparison to the results of the systems participating in the lexical sample subtask of task 17.

1 Introduction

Word Sense Disambiguation (WSD) is a key enabling-technology. Supervised WSD techniques are the best performing in public evaluations, but need large amounts of hand-tagging data. Existing hand-annotated corpora like SemCor (Miller et al., 1993), which is annotated with WordNet senses (Fellbaum, 1998) allow for a small improvement over the simple most frequent sense heuristic, as attested in the all-words track of the last Senseval competition (Snyder and Palmer, 2004). In theory, larger amounts of training data (SemCor has approx. 500M words) would improve the performance of supervised WSD, but no current project exists to provide such an expensive resource. Another problem of the supervised approach is that the inventory and distribution of senses changes dramatically from one domain to the other, requiring additional hand-tagging of corpora (Martínez and Agirre, 2000; Koeling et al., 2005).

Supervised WSD is based on the “fixed-list of senses” paradigm, where the senses for a target word are a closed list coming from a dictionary or lexicon. Lexicographers and semanticists have long warned about the problems of such an approach, where senses are listed separately as discrete entities, and have argued in favor of more complex representations, where, for instance, senses are dense regions in a continuum (Cruse, 2000).

Unsupervised Word Sense Induction and Discrimination (WSID, also known as corpus-based unsupervised systems) has followed this line of thinking, and tries to induce word senses directly from the corpus. Typical WSID systems involve clustering techniques, which group together similar examples. Given a set of induced clusters (which represent word *uses* or senses[1]), each new occurrence of the target word will be compared to the clusters and the most similar cluster will be selected as its sense.

One of the problems of unsupervised systems is that of managing to do a fair evaluation. Most of current unsupervised systems are evaluated in-house, with a brief comparison to a re-implementation of a former system, leading to a proliferation of unsupervised systems with little ground to compare among them. The goal of this task is to allow for comparison across sense-induction and discrimination systems, and also to compare these systems to other supervised and knowledge-based systems.

The paper is organized as follows. Section 2 presents the evaluation framework used in this task. Section 3 presents the systems that participated in

[1] WSID approaches prefer the term ’word uses’ to ’word senses’. In this paper we use them interchangeably to refer to both the induced clusters, and to the word senses from some reference lexicon.

Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pages 7–12,
Prague, June 2007.

the task, and the official results. Finally, Section 5 draws the conclusions.

2 Evaluating WSID systems

All WSID algorithms need some addition in order to be evaluated. One alternative is to manually decide the correctness of the clusters assigned to each occurrence of the words. This approach has two main disadvantages. First, it is expensive to manually verify each occurrence of the word, and different runs of the algorithm need to be evaluated in turn. Second, it is not an easy task to manually decide if an occurrence of a word effectively corresponds with the use of the word the assigned cluster refers to, especially considering that the person is given a short list of words linked to the cluster. We also think that instead of judging whether the cluster returned by the algorithm is correct, the person should have independently tagged the occurrence with his own senses, which should have been then compared to the cluster returned by the system. This is paramount to compare a corpus which has been hand-tagged with some reference senses (also known as the gold-standard) with the clustering result. The gold standard tags are taken to be the definition of the classes, and standard measures from the clustering literature can be used to evaluate the clusters against the classes.

A second alternative would be to devise a method to map the clusters returned by the systems to the senses in a lexicon. Pantel and Lin (2002) automatically map the senses to WordNet, and then measure the quality of the mapping. More recently, the mapping has been used to test the system on publicly available benchmarks (Purandare and Pedersen, 2004; Niu et al., 2005).

A third alternative is to evaluate the systems according to some performance in an application, e.g. information retrieval (Schütze, 1998). This is a very attractive idea, but requires expensive system development and it is sometimes difficult to separate the reasons for the good (or bad) performance.

In this task we decided to adopt the first two alternatives, since they allow for comparison over publicly available systems of any kind. With this goal on mind we gave all the participants an unlabeled corpus, and asked them to induce the senses and create a clustering solution on it. We evaluate the results according to the following types of evaluation:

1. Evaluate the induced senses as clusters of examples. The induced clusters are compared to the sets of examples tagged with the given gold standard word senses (classes), and evaluated using the FScore measure for clusters. We will call this evaluation *unsupervised*.
2. Map the induced senses to gold standard senses, and use the mapping to tag the test corpus with gold standard tags. The mapping is automatically produced by the organizers, and the resulting results evaluated according to the usual precision and recall measures for supervised word sense disambiguation systems. We call this evaluation *supervised*.

We will see each of them in turn.

2.1 Unsupervised evaluation

In this setting the results of the systems are treated as clusters of examples and gold standard senses are classes. In order to compare the clusters with the classes, hand annotated corpora is needed. The test set is first tagged with the induced senses. A perfect clustering solution will be the one where each cluster has exactly the same examples as one of the classes, and vice versa.

Following standard cluster evaluation practice (Zhao and Karypis, 2005), we consider the FScore measure for measuring the performance of the systems. The FScore is used in a similar fashion to Information Retrieval exercises, with precision and recall defined as the percentage of correctly "retrieved" examples for a cluster (divided by total cluster size), and recall as the percentage of correctly "retrieved" examples for a cluster (divided by total class size).

Given a particular class s_r of size n_r and a cluster h_i of size n_i, suppose n_r^i examples in the class s_r belong to h_i. The F value of this class and cluster is defined to be:

$$f(s_r, h_i) = \frac{2P(s_r, h_i)R(s_r, h_i)}{P(s_r, h_i) + R(s_r, h_i)}$$

where $P(s_r, h_i) = \frac{n_r^i}{n_r}$ is the precision value and $R(s_r, h_i) = \frac{n_r^i}{n_i}$ is the recall value defined for class s_r and cluster h_i. The FScore of class s_r is the maximum F value attained at any cluster, that is,

$$F(s_r) = \max_{h_i} f(s_r, h_i)$$

and the FScore of the entire clustering solution is:

$$\text{FScore} = \sum_{r=1}^{c} \frac{n_r}{n} F(s_r)$$

where q is the number of classes and n is the size of the clustering solution. If the clustering is the identical to the original classes in the datasets, FScore will be equal to one which means that the higher the FScore, the better the clustering is.

For the sake of completeness we also include the standard entropy and purity measures in the unsupervised evaluation. The entropy measure considers how the various classes of objects are distributed within each cluster. In general, the smaller the entropy value, the better the clustering algorithm performs. The purity measure considers the extent to which each cluster contained objects from primarily one class. The larger the values of purity, the better the clustering algorithm performs. For a formal definition refer to (Zhao and Karypis, 2005).

2.2 Supervised evaluation

We have followed the supervised evaluation framework for evaluating WSID systems as described in (Agirre et al., 2006). First, we split the corpus into a train/test part. Using the hand-annotated sense information in the train part, we compute a mapping matrix M that relates clusters and senses in the following way. Suppose there are m clusters and n senses for the target word. Then, $M = \{m_{ij}\} \quad 1 \leq i \leq m, 1 \leq j \leq n$, and each $m_{ij} = P(s_j | h_i)$, that is, m_{ij} is the probability of a word having sense j given that it has been assigned cluster i. This probability can be computed counting the times an occurrence with sense s_j has been assigned cluster h_i in the train corpus.

The mapping matrix is used to transform any cluster score vector $\bar{h} = (h_1, \ldots, h_m)$ returned by the WSID algorithm into a sense score vector $\bar{s} = (s_1, \ldots, s_n)$. It suffices to multiply the score vector by M, i.e., $\bar{s} = \bar{h}M$.

We use the M mapping matrix in order to convert the cluster score vector of each test corpus instance into a sense score vector, and assign the sense with maximum score to that instance. Finally, the resulting test corpus is evaluated according to the usual precision and recall measures for supervised word sense disambiguation systems.

	All	Nouns	Verbs
train	22281	14746	9773
test	4851	2903	2427
all	27132	17649	12200

Table 1: Number of occurrences for the 100 target words in the corpus following the train/test split.

3 Results

In this section we will introduce the gold standard and corpus used, the description of the systems and the results obtained. Finally we provide some material for discussion.

Gold Standard

The data used for the actual evaluation was borrowed from the SemEval-2007 “English lexical sample subtask” of task 17. The texts come from the Wall Street Journal corpus, and were hand-annotated with OntoNotes senses (Hovy et al., 2006). Note that OntoNotes senses are coarser than WordNet senses, and thus the number of senses to be induced is smaller in this case.

Participants were provided with information about 100 target words (65 verbs and 35 nouns), each target word having a set of contexts where the word appears. After removing the sense tags from the train corpus, the train and test parts were joined into the official corpus and given to the participants. Participants had to tag with the induced senses all the examples in this corpus. Table 1 summarizes the size of the corpus.

Participant systems

In total there were 6 participant systems. One of them (UoFL) was not a sense induction system, but rather a knowledge-based WSD system. We include their data in the results section below for coherence with the official results submitted to participants, but we will not mention it here.

I2R: This team used a cluster validation method to estimate the number of senses of a target word in untagged data, and then grouped the instances of this target word into the estimated number of clusters using the sequential Information Bottleneck algorithm.

UBC-AS: A two stage graph-based clustering where a co-occurrence graph is used to compute similarities against contexts. The context similarity matrix is pruned and the resulting associated graph is clustered by means of a random-walk type algorithm. The parameters of the system are tuned against the Senseval-3 lexical sample dataset, and some manual tuning is performed in order to reduce the overall number of induced senses. Note that this system was submitted by the organizers. The organizers took great care in order to participate under the same conditions as the rest of participants.

UMND2: A system which clusters the second order co-occurrence vectors associated with each word in a context. Clustering is done using k-means and the number of clusters was automatically discovered using the Adapted Gap Statistic. No parameter tuning is performed.

upv_si: A self-term expansion method based on co-ocurrence, where the terms of the corpus are expanded by its best co-ocurrence terms in the same corpus. The clustering is done using one implementation of the KStar method where the stop criterion has been modified. The trial data was used for determining the corpus structure. No further tuning is performed.

UOY: A graph based system which creates a co-occurrence hypergraph model. The hypergraph is filtered and weighted according to some association rules. The clustering is performed by selecting the nodes of higher degree until a stop criterion is reached. WSD is performed by assigning to each induced cluster a score equal to the sum of weights of hyperedges found in the local context of the target word. The system was tested and tuned on 10 nouns of Senseval-3 lexical-sample.

Official Results

Participants were required to induce the senses of the target words and cluster all target word contexts accordingly[2]. Table 2 summarizes the average number of induced senses as well as the real senses in the gold standard.

system	All	nouns	verbs
I2R	3.08	3.11	3.06
UBC-AS*	1.32	1.63	1.15
UMND2	1.36	1.71	1.17
upv_si	5.57	7.2	4.69
UOY	9.28	11.28	8.2
Gold standard			
test	2.87	2.86	2.86
train	3.6	3.91	3.43
all	3.68	3.94	3.54

Table 2: Average number of clusters as returned by the participants, and number of classes in the gold standard. Note that *UBC-AS** is the system submitted by the organizers of the task.

System	R.	All			Nouns	Verbs
		FSc.	Pur.	Entr.	FSc.	FSc.
1c1word	1	**78.9**	79.8	45.4	80.7	76.8
UBC-AS*	2	**78.7**	80.5	43.8	80.8	76.3
upv_si	3	**66.3**	83.8	33.2	69.9	62.2
UMND2	4	**66.1**	81.7	40.5	67.1	65.0
I2R	5	**63.9**	84.0	32.8	68.0	59.3
UofL**	6	**61.5**	82.2	37.8	62.3	60.5
UOY	7	**56.1**	86.1	27.1	65.8	45.1
Random	8	**37.9**	86.1	27.7	38.1	37.7
1c1inst	9	**9.5**	100	0	6.6	12.7

Table 3: Unsupervised evaluation on the test corpus (FScore), including 3 baselines. Purity and entropy are also provided. *UBC-AS** was submitted by the organizers. *UofL*** is not a sense induction system.

System	Rank	Supervised evaluation		
		All	Nouns	Verbs
I2R	1	**81.6**	86.8	75.7
UMND2	2	**80.6**	84.5	76.2
upv_si	3	**79.1**	82.5	75.3
MFS	4	**78.7**	80.9	76.2
UBC-AS*	5	**78.5**	80.7	76.0
UOY	6	**77.7**	81.6	73.3
UofL**	7	**77.1**	80.5	73.3

Table 4: Supervised evaluation as recall. *UBC-AS** was submitted by the organizers. *UofL*** is not a sense induction system.

Table 3 shows the unsupervised evaluation of the systems on the test corpus. We also include three baselines: the "one cluster per word" baseline (*1c1word*), which groups all instances of a word into a single cluster, the "one cluster per instance" baseline (*1c1inst*), where each instance is a distinct cluster, and a random baseline, where the induced word senses and their associated weights have been randomly produced. The random baseline figures in this paper are averages over 10 runs.

As shown in Table 3, no system outperforms the *1c1word* baseline, which indicates that this baseline

[2] They were allowed to label each context with a weighted score vector, assigning a weight to each induced sense. In the unsupervised evaluation only the sense with maximum weight was considered, but for the supervised one the whole score vector was used. However, none of the participating systems labeled any instance with more than one sense.

is quite strong, perhaps due the relatively small number of classes in the gold standard. However, all systems outperform by far the *random* and *1c1inst* baselines, meaning that the systems are able to induce correct senses. Note that the purity and entropy measures are not very indicative in this setting. For completeness, we also computed the FScore using the complete corpus (both train and test). The results are similar and the ranking is the same. We omit them for brevity.

The results of the supervised evaluation can be seen in Table 4. The evaluation is also performed over the test corpus. Apart from participants, we also show the most frequent sense (MFS), which tags every test instance with the sense that occurred most often in the training part. Note that the supervised evaluation combines the information in the clustering solution implicitly with the MFS information via the mapping in the training part. Previous Senseval evaluation exercises have shown that the MFS baseline is very hard to beat by unsupervised systems. In fact, only three of the participant systems are above the MFS baseline, which shows that the clustering information carries over the mapping successfully for these systems. Note that the *1c1word* baseline is equivalent to MFS in this setting. We will review the random baseline in the discussion section below.

Further Results

Table 5 shows the results of the best systems from the lexical sample subtask of task 17. The best sense induction system is only 6.9 percentage points below the best supervised, and 3.5 percentage points below the best (and only) semi-supervised system. If the sense induction system had participated, it would be deemed as semi-supervised, as it uses, albeit in a shallow way, the training data for mapping the clusters into senses. In this sense, our supervised evaluation does not seek to optimize the available training data.

After the official evaluation, we realized that contrary to previous lexical sample evaluation exercises task 17 organizers did not follow a random train/test split. We decided to produce a random train/test split following the same 82/18 proportion as the official split, and re-evaluated the systems. The results are presented in Table 6, where we can see that all participants are above the MFS baseline, showing that all of them learned useful clustering information. Note that UOY was specially affected by the original split. The distribution of senses in this split did not vary (cf. Table 2).

System	Supervised evaluation
best supervised	**88.7**
best semi-supervised	**85.1**
best induction (semi-sup.)	**81.6**
MFS	**78.7**
best unsupervised	**53.8**

Table 5: Comparing the best induction system in this task with those of task 17.

System	Supervised evaluation
I2R	**82.2**
UOY	**81.3**
UMND2	**80.1**
upv_si	**79.9**
UBC-AS	**79.0**
MFS	**78.4**

Table 6: Supervised evaluation as recall using a random train/test split.

Finally, we also studied the supervised evaluation of several random clustering algorithms, which can attain performances close to MFS, thanks to the mapping information. This is due to the fact that the random clusters would be mapped to the most frequent senses. Table 7 shows the results of random solutions using varying numbers of clusters (e.g. random2 is a random choice between two clusters). Random2 is only 0.1 below MFS, but as the number of clusters increases some clusters don't get mapped, and the recall of the random baselines decrease.

4 Discussion

The evaluation of clustering solutions is not straightforward. All measures have some bias towards certain clustering strategy, and this is one of the reasons of adding the supervised evaluation as a complementary information to the more standard unsupervised evaluation.

In our case, we noticed that the FScore penalized the systems with a high number of clusters, and favored those that induce less senses. Given the fact that FScore tries to balance precision (higher for large numbers of clusters) and recall (higher for small numbers of clusters), this was not expected. We were also surprised to see that no system could

System	Supervised evaluation
random2	**78.6**
random10	**77.6**
ramdom100	**64.2**
random1000	**31.8**

Table 7: Supervised evaluation of several random baselines.

beat the "one cluster one word" baseline. An explanation might lay in that the gold-standard was based on the coarse-grained OntoNotes senses. We also noticed that some words had hundreds of instances and only a single sense. We suspect that the participating systems would have beaten all baselines if a fine-grained sense inventory like WordNet had been used, as was customary in previous WSD evaluation exercises.

Supervised evaluation seems to be more neutral regarding the number of clusters, as the ranking of systems according to this measure include diverse cluster averages. Each of the induced clusters is mapped into a weighted vector of senses, and thus inducing a number of clusters similar to the number of senses is not a requirement for good results. With this measure some of the systems[3] are able to beat all baselines.

5 Conclusions

We have presented the design and results of the SemEval-2007 task 02 on evaluating word sense induction and discrimination systems. 6 systems participated, but one of them was not a sense induction system. We reused the data from the SemEval-2007 English lexical sample subtask of task 17, and set up both clustering-style unsupervised evaluation (using OntoNotes senses as gold-standard) and a supervised evaluation (using the training part of the dataset for mapping). We also provide a comparison to the results of the systems participating in the lexical sample subtask of task 17.

Evaluating clustering solutions is not straightforward. The unsupervised evaluation seems to be sensitive to the number of senses in the gold standard, and the coarse grained sense inventory used in the gold standard had a great impact in the results. The supervised evaluation introduces a mapping step which interacts with the clustering solution. In fact, the ranking of the participating systems varies according to the evaluation method used. We think the two evaluation results should be taken to be complementary regarding the information learned by the clustering systems, and that the evaluation of word sense induction and discrimination systems needs further developments, perhaps linked to a certain application or purpose.

Acknowledgments

We want too thank the organizers of SemEval-2007 task 17 for kindly letting us use their corpus. We are also grateful to Ted Pedersen for his comments on the evaluation results. This work has been partially funded by the Spanish education ministry (project KNOW) and by the regional government of Gipuzkoa (project DAHAD).

References

E. Agirre, D. Martínez, O. López de Lacalle, and A. Soroa. 2006. Evaluating and optimizing the parameters of an unsupervised graph-based wsd algorithm. In *Proceedings of the NAACL TextGraphs workshop*, pages 89–96, New York City, June.

D. A. Cruse, 2000. *Polysemy: Theoretical and Computational Approaches*, chapter Aspects of the Microstructure of Word Meanings, pages 31–51. OUP.

C. Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

E. Hovy, M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel. 2006. Ontonotes: The 90% solution. In *Proceedings of HLT/NAACL*.

R. Koeling, D. McCarthy, and J.D. Carroll. 2005. Domain-specific sense distributions and predominant sense acquisition.

D. Martínez and E. Agirre. 2000. One sense per collocation and genre/topic variations.

G.A. Miller, C. Leacock, R. Tengi, and R.Bunker. 1993. A semantic concordance. In *Proc. of the ARPA HLT workshop*.

C. Niu, W. Li, R. K. Srihari, and H. Li. 2005. Word independent context pair classification model for word sense disambiguation. In *Proc. of CoNLL-2005*.

P. Pantel and D. Lin. 2002. Discovering word senses from text. In *Proc. of KDD02*.

A. Purandare and T. Pedersen. 2004. Word sense discrimination by clustering contexts in vector and similarity spaces. In *Proc. of CoNLL-2004*, pages 41–48.

H. Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

B. Snyder and M. Palmer. 2004. The english all-words task. In *Proc. of SENSEVAL*.

Y Zhao and G Karypis. 2005. Hierarchical clustering algorithms for document datasets. *Data Mining and Knowledge Discovery*, 10(2):141–168.

[3] All systems in the case of a random train/test split