Semeval-2007 Task 2:
Evaluating Word Sense Induction and Discrimination Systems

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

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

Abstract

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 (Martinez 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 the task, and the official results. Finally, Section 4 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 verify each occurrence of the word, and different runs of the algorithm need to be evaluated in turn.

\(^1\)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.
ally 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 $n_{ir}$ 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 $P(s_r, h_i) = \frac{n_{ir}}{n_r}$ is the precision value and $R(s_r, h_i) = \frac{n_{ir}}{n_i}$ 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, $M = \{m_{ij}\}$, $1 \leq i \leq m, 1 \leq j \leq n$, 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.

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 (S07LS for short). The texts come from the Wall Street Journal and Brown corpora, and were hand-annotated with OntoNotes senses (REF***). 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.

|          | 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. Train corresponds to the training part of S07LS, and so does test.

### Participant systems

In total there were 6 participant systems: them:

- **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
Table 3: Unsupervised evaluation on the test corpus (Fscore). Note that UBC-AS* is the system submitted by the organizers of the task. UofL** is not a sense induction system.

| System           | Rank | All Nouns | All Verbs |
|------------------|------|-----------|-----------|
|                  | Fscore | Purity | Entropy | Fscore | Fscore |
| 1clusterPerWord  | 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   | 38.0   | 37.66  |
| 1clusterPerInst  | 9     | 9.5      | 100      | 0      | 6.6    | 12.7   |

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. UofL** is not a sense induction system.

| system  | All clusters | nouns | verbs |
|---------|--------------|-------|-------|
| I2R     | 3.08         | 3.11  | 3.06  |
| UBC-AS* | 1.32         | 1.63  | 1.15  |
| UMND2   | 1.36         | 1.71  | 1.17  |
| UofL**  | 4.68         | 5.17  | 4.42  |
| upv_si  | 5.57         | 7.2   | 4.69  |
| UOY     | 9.28         | 11.28 | 8.2   |

Gold standard

- **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.

- **UofL**: This system is not a sense induction system, but rather an unsupervised knowledge-based system, i.e., they don’t induce the senses but take them right away from WordNet2.1 and eXtended WordNet. Each target word is then expanded to all its senses and a disambiguation graph is created where nodes are words and associated senses and (weighted) edges are semantic relations among them (which are also extracted from the lexicon). Nodes whose out-edge sum is maximum are selected as the right senses. If there is a tie, the sense which comes first in WordNet is selected. Note that this system should not have taken part in the competition, and although we report their results, we don’t mention it in the comments.

- **upv_si**: A self-term expansion method based on co-occurrence, where the terms of the corpus are expanded by its best co-occurrence 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.

- **OUY**: 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.

### Results

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

Table 3 shows the unsupervised evaluation of the systems when the evaluation is performed exclusively in the test corpus. We also include three baselines: the “one cluster per word” baseline, which groups all instances of a word into a single cluster, the “1 cluster per instance” baseline where each instance is a distinct cluster, and a random baseline, where the induced word senses and their associated weights have been randomly produced.

As shown in Table 3, no system outperforms the IclusterPerWord baseline which indicates that this baseline is quite strong, perhaps due to the relatively small number of classes in the gold standard. However, all systems outperform by far the random and IclusterPerInst baselines, meaning that the system are able to induce correct senses.

Overall, the systems that induced a lower number of clusters yielded best results. This can be explained by the fact that the FScore measure penalizes a system inducing a different number of clusters than those present in the gold standard. The UOY system was the one which induced the most number of clusters, and, despite of having a good purity and entropy figures, it got the worst FScore. upv, si and UMND2 seem to have managed to automatically approximate the number of classes (see Table 2), and they yield clustering solutions with high purity, relatively low entropy, yet yielding a high FScore. Note that UBC-AS made their system

| System   | Rank | All | Nouns | Verbs |
|----------|------|-----|-------|-------|
| UBC-AS*  | 1    | 76.9| 78.3  | 75.0  |
| upv_s1   | 2    | 63.7| 65.8  | 60.7  |
| UMND2    | 3    | 63.1| 62.5  | 64.0  |
| I2R      | 4    | 58.0| 61.6  | 52.8  |
| UofL**   | 5    | 57.3| 57.5  | 56.9  |
| UOY      | 6    | 49.0| 56.5  | 38.3  |

Table 4: Unsupervised evaluation on the whole corpus (Fscore). Note that UBC-AS is the system submitted by the organizers of the task. UofL** is not a sense induction system.

The results of the supervised evaluation can be seen in Table 5. The evaluation is also performed over the test corpus. There are two baselines: the most frequent sense (MFS), which tags every test instance with the sense that occurred most often in the train split, and the random baseline.

### Discussion

### 4 Conclusions

### Acknowledgment

We want to thank the organizers of SemEval-2007 task 17 for kindly letting us use their corpus. We are also grateful to all the participants in this task, for their hard work and involvement in this evaluation exercise.

### References

Eneko Agirre, David Martı́nez, Oier López de Lacalle, and Aitor Soroa. 2006. Evaluating and optimizing
| 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    | 79.5                   | 82.6 | 76.0  |
| MFS     | 4    | 78.7                   | 80.9 | 76.2  |
| Random  | 5    | 78.7                   | 80.9 | 76.1  |
| UBC-AS* | 6    | 78.6                   | 80.7 | 76.2  |
| UOY     | 7    | 78.5                   | 81.8 | 74.9  |
| UofL**  | 8    | 78.5                   | 81.4 | 75.2  |

Table 5: Supervised evaluation as recall. As the coverage is 100%, precision equals recall. Note that UBC-AS* is the system submitted by the organizers of the task. UofL** is not a sense induction system.

the parameters of an unsupervised graph-based wsd algorithm. In *Proceedings of TextGraphs: the Second Workshop on Graph Based Methods for Natural Language Processing*, pages 89–96, New York City, June. Association for Computational Linguistics.

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.

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.