SemEval-2010 Task 14: Word Sense Induction & Disambiguation

Suresh Manandhar
Department of Computer Science
University of York, UK

Ioannis P. Klapaftis
Department of Computer Science
University of York, UK

Dmitriy Dligach
Department of Computer Science
University of Colorado, USA

Sameer S. Pradhan
BBN Technologies
Cambridge, USA

Abstract

This paper presents the description and evaluation framework of SemEval-2010 Word Sense Induction & Disambiguation task, as well as the evaluation results of 26 participating systems. In this task, participants were required to induce the senses of 100 target words using a training set, and then disambiguate unseen instances of the same words using the induced senses. Systems' answers were evaluated in: (1) an unsupervised manner by using two clustering evaluation measures, and (2) a supervised manner in a WSD task.

1 Introduction

Word senses are more beneficial than simple word forms for a variety of tasks including Information Retrieval, Machine Translation and others (Pantel and Lin, 2002). However, word senses are usually represented as a fixed-list of definitions of a manually constructed lexical database. Several deficiencies are caused by this representation, e.g. lexical databases miss main domain-specific senses (Pantel and Lin, 2002), they often contain general definitions and suffer from the lack of explicit semantic or contextual links between concepts (Agirre et al., 2001). More importantly, the definitions of hand-crafted lexical databases often do not reflect the exact meaning of a target word in a given context (Véronis, 2004).

Unsupervised Word Sense Induction (WSI) aims to overcome these limitations of hand-constructed lexicons by learning the senses of a target word directly from text without relying on any hand-crafted resources. The primary aim of SemEval-2010 WSI task is to allow comparison of unsupervised word sense induction and disambiguation systems.

The target word dataset consists of 100 words, 50 nouns and 50 verbs. For each target word, participants were provided with a training set in order to learn the senses of that word. In the next step, participating systems were asked to disambiguate unseen instances of the same words using their learned senses. The answers of the systems were then sent to organisers for evaluation.

2 Task description

Figure 1 provides an overview of the task. As can be observed, the task consisted of three separate phases. In the first phase, *training phase*, participating systems were provided with a training dataset that consisted of a set of target word (noun/verb) instances (sentences/paragraphs). Participants were then asked to use this training dataset to induce the senses of the target word. No other resources were allowed with the exception of NLP components for morphology and syntax. In the second phase, *testing phase*, participating systems were provided with a testing dataset that consisted of a set of target word (noun/verb) instances (sentences/paragraphs). Participants were then asked to tag (disambiguate) each testing instance with the senses induced during the *training phase*. In the third and final phase, the tagged test instances were received by the organisers in order to evaluate the answers of the systems in a supervised and an unsupervised framework. Table 1 shows the total number of target word instances in the training and testing set, as well as the average number of senses in the gold standard.

The main difference of the SemEval-2010 as compared to the SemEval-2007 sense induction task is that the training and testing data are treated separately, i.e the testing data are only used for sense tagging, while the training data are only used

Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, pages 63–68,
Uppsala, Sweden, 15-16 July 2010.

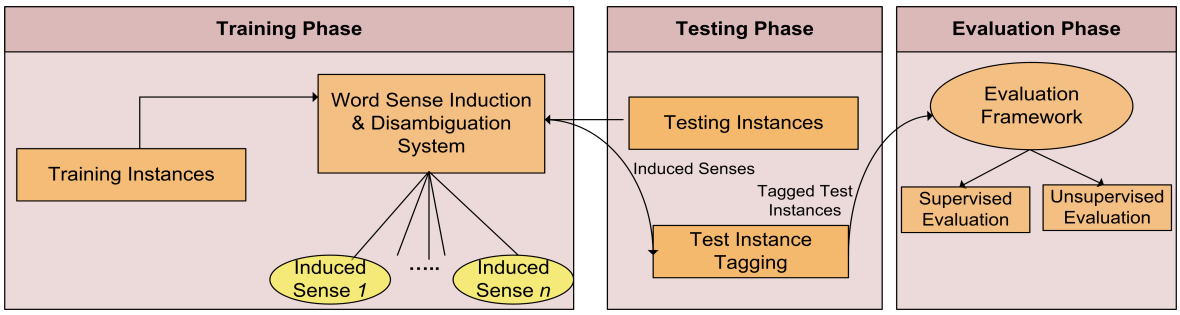

Figure 1: Training, testing and evaluation phases of SemEval-2010 Task 14

	Training set	Testing set	Senses (#)
All	879807	8915	3.79
Nouns	716945	5285	4.46
Verbs	162862	3630	3.12

Table 1: Training & testing set details

Word Sense	Query
Sense 1	failure AND (loss OR nonconformity OR test OR surrender OR "force play" OR ...)
Sense 2	failure AND (ruination OR flop OR bust OR stall OR ruin OR walloping OR ...)

Table 2: Training set creation: example queries for target word *failure*

for sense induction. Treating the testing data as new unseen instances ensures a realistic evaluation that allows to evaluate the clustering models of each participating system.

The evaluation framework of SemEval-2010 WSI task considered two types of evaluation. In the first one, *unsupervised evaluation*, systems' answers were evaluated according to: (1) *V-Measure* (Rosenberg and Hirschberg, 2007), and (2) *paired F-Score* (Artiles et al., 2009). Neither of these measures were used in the SemEval-2007 WSI task. Manandhar & Klapaftis (2009) provide more details on the choice of this evaluation setting and its differences with the previous evaluation. The second type of evaluation, *supervised evaluation*, follows the supervised evaluation of the SemEval-2007 WSI task (Agirre and Soroa, 2007). In this evaluation, induced senses are mapped to gold standard senses using a mapping corpus, and systems are then evaluated in a standard WSD task.

2.1 Training dataset

The target word dataset consisted of 100 words, i.e. 50 nouns and 50 verbs. The training dataset for each target noun or verb was created by following a web-based semi-automatic method, similar to the method for the construction of *Topic Signatures* (Agirre et al., 2001). Specifically, for each WordNet (Fellbaum, 1998) sense of a target word, we created a query of the following form:

<Target Word> ***AND*** *<Relative Set>*

The *<Target Word>* consisted of the target word stem. The *<Relative Set>* consisted of a disjunctive set of word lemmas that were related to the target word sense for which the query was created. The relations considered were WordNet's hypernyms, hyponyms, synonyms, meronyms and holonyms. Each query was manually checked by one of the organisers to remove ambiguous words. The following example shows the query created for the first[1] and second[2] WordNet sense of the target noun *failure*.

The created queries were issued to Yahoo! search API[3] and for each query a maximum of 1000 pages were downloaded. For each page we extracted fragments of text that occurred in   html tags and contained the target word stem. In the final stage, each extracted fragment of text was POS-tagged using the Genia tagger (Tsuruoka and Tsujii, 2005) and was only retained, if the POS of the target word in the extracted text matched the POS of the target word in our dataset.

2.2 Testing dataset

The testing dataset consisted of instances of the same target words from the training dataset. This dataset is part of OntoNotes (Hovy et al., 2006). We used the sense-tagged dataset in which sentences containing target word instances are tagged with OntoNotes (Hovy et al., 2006) senses. The texts come from various news sources including CNN, ABC and others.

[1] An act that fails
[2] An event that does not accomplish its intended purpose
[3] http://developer.yahoo.com/search/ [Access:10/04/2010]

	G_1	G_2	G_3
C_1	10	10	15
C_2	20	50	0
C_3	1	10	60
C_4	5	0	0

Table 3: Clusters & GS senses matrix.

3 Evaluation framework

For the purposes of this section we provide an example (Table 3) in which a target word has 181 instances and 3 GS senses. A system has generated a clustering solution with 4 clusters covering all instances. Table 3 shows the number of common instances between clusters and GS senses.

3.1 Unsupervised evaluation

This section presents the measures of unsupervised evaluation, i.e *V-Measure* (Rosenberg and Hirschberg, 2007) and (2) *paired F-Score* (Artiles et al., 2009).

3.1.1 V-Measure evaluation

Let w be a target word with N instances (data points) in the testing dataset. Let $K = \{C_j | j = 1 \dots n\}$ be a set of automatically generated clusters grouping these instances, and $S = \{G_i | i = 1 \dots m\}$ the set of gold standard classes containing the desirable groupings of w instances.

V-Measure (Rosenberg and Hirschberg, 2007) assesses the quality of a clustering solution by explicitly measuring its *homogeneity* and its *completeness*. Homogeneity refers to the degree that each cluster consists of data points primarily belonging to a single GS class, while completeness refers to the degree that each GS class consists of data points primarily assigned to a single cluster (Rosenberg and Hirschberg, 2007). Let h be homogeneity and c completeness. V-Measure is the harmonic mean of h and c, i.e. $VM = \frac{2 \cdot h \cdot c}{h+c}$.

Homogeneity. The homogeneity, h, of a clustering solution is defined in Formula 1, where $H(S|K)$ is the conditional entropy of the class distribution given the proposed clustering and $H(S)$ is the class entropy.

$$h = \begin{cases} 1, & \text{if } H(S) = 0 \\ 1 - \frac{H(S|K)}{H(S)}, & \text{otherwise} \end{cases} \quad (1)$$

$$H(S) = -\sum_{i=1}^{|S|} \frac{\sum_{j=1}^{|K|} a_{ij}}{N} \log \frac{\sum_{j=1}^{|K|} a_{ij}}{N} \quad (2)$$

$$H(S|K) = -\sum_{j=1}^{|K|} \sum_{i=1}^{|S|} \frac{a_{ij}}{N} \log \frac{a_{ij}}{\sum_{k=1}^{|S|} a_{kj}} \quad (3)$$

When $H(S|K)$ is 0, the solution is perfectly homogeneous, because each cluster only contains data points that belong to a single class. However in an imperfect situation, $H(S|K)$ depends on the size of the dataset and the distribution of class sizes. Hence, instead of taking the raw conditional entropy, V-Measure normalises it by the maximum reduction in entropy the clustering information could provide, i.e. $H(S)$. When there is only a single class ($H(S) = 0$), any clustering would produce a perfectly homogeneous solution.

Completeness. Symmetrically to homogeneity, the completeness, c, of a clustering solution is defined in Formula 4, where $H(K|S)$ is the conditional entropy of the cluster distribution given the class distribution and $H(K)$ is the clustering entropy. When $H(K|S)$ is 0, the solution is perfectly complete, because all data points of a class belong to the same cluster.

For the clustering example in Table 3, homogeneity is equal to 0.404, completeness is equal to 0.37 and V-Measure is equal to 0.386.

$$c = \begin{cases} 1, & \text{if } H(K) = 0 \\ 1 - \frac{H(K|S)}{H(K)}, & \text{otherwise} \end{cases} \quad (4)$$

$$H(K) = -\sum_{j=1}^{|K|} \frac{\sum_{i=1}^{|S|} a_{ij}}{N} \log \frac{\sum_{i=1}^{|S|} a_{ij}}{N} \quad (5)$$

$$H(K|S) = -\sum_{i=1}^{|S|} \sum_{j=1}^{|K|} \frac{a_{ij}}{N} \log \frac{a_{ij}}{\sum_{k=1}^{|K|} a_{ik}} \quad (6)$$

3.1.2 Paired F-Score evaluation

In this evaluation, the clustering problem is transformed into a classification problem. For each cluster C_i we generate $\binom{|C_i|}{2}$ instance pairs, where $|C_i|$ is the total number of instances that belong to cluster C_i. Similarly, for each GS class G_i we generate $\binom{|G_i|}{2}$ instance pairs, where $|G_i|$ is the total number of instances that belong to GS class G_i.

Let $F(K)$ be the set of instance pairs that exist in the automatically induced clusters and $F(S)$ be the set of instance pairs that exist in the gold standard. Precision can be defined as the number of common instance pairs between the two sets to the total number of pairs in the clustering solution (Equation 7), while recall can be defined as the number of common instance pairs between the two sets to the total number of pairs in the gold

standard (Equation 8). Finally, precision and recall are combined to produce the harmonic mean ($FS = \frac{2 \cdot P \cdot R}{P+R}$).

$$P = \frac{|F(K) \cap F(S)|}{|F(K)|} \quad (7)$$

$$R = \frac{|F(K) \cap F(S)|}{|F(S)|} \quad (8)$$

For example in Table 3, we can generate $\binom{35}{2}$ instance pairs for C_1 , $\binom{70}{2}$ for C_2, $\binom{71}{2}$ for C_3 and $\binom{5}{2}$ for C_4, resulting in a total of 5505 instance pairs. In the same vein, we can generate $\binom{36}{2}$ instance pairs for G_1, $\binom{70}{2}$ for G_2 and $\binom{75}{2}$ for G_3. In total, the GS classes contain 5820 instance pairs. There are 3435 common instance pairs, hence precision is equal to 62.39%, recall is equal to 59.09% and paired F-Score is equal to 60.69%.

3.2 Supervised evaluation

In this evaluation, the testing dataset is split into a mapping and an evaluation corpus. The first one is used to map the automatically induced clusters to GS senses, while the second is used to evaluate methods in a WSD setting. This evaluation follows the supervised evaluation of SemEval-2007 WSI task (Agirre and Soroa, 2007), with the difference that the reported results are an average of 5 random splits. This repeated random sampling was performed to avoid the problems of the SemEval-2007 WSI challenge, in which different splits were providing different system rankings.

Let us consider the example in Table 3 and assume that this matrix has been created by using the mapping corpus. Table 3 shows that C_1 is more likely to be associated with G_3, C_2 is more likely to be associated with G_2, C_3 is more likely to be associated with G_3 and C_4 is more likely to be associated with G_1. This information can be utilised to map the clusters to GS senses.

Particularly, the matrix shown in Table 3 is normalised to produce a matrix M, in which each entry depicts the estimated conditional probability $P(G_i|C_j)$. Given an instance I of tw from the evaluation corpus, a row cluster vector IC is created, in which each entry k corresponds to the score assigned to C_k to be the winning cluster for instance I. The product of IC and M provides a row sense vector, IG, in which the highest scoring entry a denotes that G_a is the winning sense. For example, if we produce the row cluster vector $[C_1 = 0.8, C_2 = 0.1, C_3 = 0.1, C_4 = 0.0]$, and multiply it with the normalised matrix of Table 3, then we would get a row sense vector in which G_3 would be the winning sense with a score equal to 0.43.

System	VM (%) (All)	VM (%) (Nouns)	VM (%) (Verbs)	#Cl
Hermit	16.2	16.7	15.6	10.78
UoY	15.7	20.6	8.5	11.54
KSU KDD	15.7	18	12.4	17.5
Duluth-WSI	9	11.4	5.7	4.15
Duluth-WSI-SVD	9	11.4	5.7	4.15
Duluth-R-110	8.6	8.6	8.5	9.71
Duluth-WSI-Co	7.9	9.2	6	2.49
KCDC-PCGD	7.8	7.3	8.4	2.9
KCDC-PC	7.5	7.7	7.3	2.92
KCDC-PC-2	7.1	7.7	6.1	2.93
Duluth-Mix-Narrow-Gap	6.9	8	5.1	2.42
KCDC-GD-2	6.9	6.1	8	2.82
KCDC-GD	6.9	5.9	8.5	2.78
Duluth-Mix-Narrow-PK2	6.8	7.8	5.5	2.68
Duluth-MIX-PK2	5.6	5.8	5.2	2.66
Duluth-R-15	5.3	5.4	5.1	4.97
Duluth-WSI-Co-Gap	4.8	5.6	3.6	1.6
Random	4.4	4.2	4.6	4
Duluth-R-13	3.6	3.5	3.7	3
Duluth-WSI-Gap	3.1	4.2	1.5	1.4
Duluth-Mix-Gap	3	2.9	3	1.61
Duluth-Mix-Uni-PK2	2.4	0.8	4.7	2.04
Duluth-R-12	2.3	2.2	2.5	2
KCDC-PT	1.9	1	3.1	1.5
Duluth-Mix-Uni-Gap	1.4	0.2	3	1.39
KCDC-GDC	7	6.2	7.8	2.83
MFS	0	0	0	1
Duluth-WSI-SVD-Gap	0	0	0.1	1.02

Table 4: V-Measure unsupervised evaluation

4 Evaluation results

In this section, we present the results of the 26 systems along with two baselines. The first baseline, Most Frequent Sense (*MFS*), groups all testing instances of a target word into one cluster. The second baseline, *Random*, randomly assigns an instance to one out of four clusters. The number of clusters of *Random* was chosen to be roughly equal to the average number of senses in the GS. This baseline is executed five times and the results are averaged.

4.1 Unsupervised evaluation

Table 4 shows the V-Measure (VM) performance of the 26 systems participating in the task. The last column shows the number of induced clusters of each system in the test set.The *MFS* baseline has a V-Measure equal to 0, since by definition its completeness is 1 and homogeneity is 0. All systems outperform this baseline, apart from one, whose V-Measure is equal to 0. Regarding the *Random* baseline, we observe that 17 perform better, which indicates that they have learned useful information better than chance.

Table 4 also shows that V-Measure tends to favour systems producing a higher number of clus-

System	FS (%) (All)	FS (%) (Nouns)	FS (%) (Verbs)	#Cl
MFS	63.5	57.0	72.7	1
Duluth-WSI-SVD-Gap	63.3	57.0	72.4	1.02
KCDC-PT	61.8	56.4	69.7	1.5
KCDC-GD	59.2	51.6	70.0	2.78
Duluth-Mix-Gap	59.1	54.5	65.8	1.61
Duluth-Mix-Uni-Gap	58.7	57.0	61.2	1.39
KCDC-GD-2	58.2	50.4	69.3	2.82
KCDC-GDC	57.3	48.5	70.0	2.83
Duluth-Mix-Uni-PK2	56.6	57.1	55.9	2.04
KCDC-PC	55.5	50.4	62.9	2.92
KCDC-PC-2	54.7	49.7	61.7	2.93
Duluth-WSI-Gap	53.7	53.4	53.9	1.4
KCDC-PCGD	53.3	44.8	65.6	2.9
Duluth-WSI-Co-Gap	52.6	53.3	51.5	1.6
Duluth-MIX-PK2	50.4	51.7	48.3	2.66
UoY	49.8	38.2	66.6	11.54
Duluth-Mix-Narrow-Gap	49.7	47.4	51.3	2.42
Duluth-WSI-Co	49.5	50.2	48.2	2.49
Duluth-Mix-Narrow-PK2	47.8	37.1	48.2	2.68
Duluth-R-12	47.8	44.3	52.6	2
Duluth-WSI-SVD	41.1	37.1	46.7	4.15
Duluth-WSI	41.1	37.1	46.7	4.15
Duluth-R-13	38.4	36.2	41.5	3
KSU KDD	36.9	24.6	54.7	17.5
Random	31.9	30.4	34.1	4
Duluth-R-15	27.6	26.7	28.9	4.97
Hermit	26.7	24.4	30.1	10.78
Duluth-R-110	16.1	15.8	16.4	9.71

Table 5: Paired F-Score unsupervised evaluation

System	SR (%) (All)	SR (%) (Nouns)	SR (%) (Verbs)	#S
UoY	62.4	59.4	66.8	1.51
Duluth-WSI	60.5	54.7	68.9	1.66
Duluth-WSI-SVD	60.5	54.7	68.9	1.66
Duluth-WSI-Co-Gap	60.3	54.1	68.6	1.19
Duluth-WSI-Co	60.8	54.7	67.6	1.51
Duluth-WSI-Gap	59.8	54.4	67.8	1.11
KCDC-PC-2	59.8	54.1	68.0	1.21
KCDC-PC	59.7	54.6	67.3	1.39
KCDC-PCGD	59.5	53.3	68.6	1.47
KCDC-GDC	59.1	53.4	67.4	1.34
KCDC-GD	59.0	53.0	67.9	1.33
KCDC-PT	58.9	53.1	67.4	1.08
KCDC-GD-2	58.7	52.8	67.4	1.33
Duluth-WSI-SVD-Gap	58.7	53.2	66.7	1.01
MFS	58.7	53.2	66.6	1
Duluth-R-12	58.5	53.1	66.4	1.25
Hermit	58.3	53.6	65.3	2.06
Duluth-R-13	58.0	52.3	66.4	1.46
Random	57.3	51.5	65.7	1.53
Duluth-R-15	56.8	50.9	65.3	1.61
Duluth-Mix-Narrow-Gap	56.6	48.1	69.1	1.43
Duluth-Mix-Narrow-PK2	56.1	47.5	68.7	1.41
Duluth-R-110	54.8	48.3	64.2	1.94
KSU KDD	52.2	46.6	60.3	1.69
Duluth-MIX-PK2	51.6	41.1	67.0	1.23
Duluth-Mix-Gap	50.6	40.0	66.0	1.01
Duluth-Mix-Uni-PK2	19.3	1.8	44.8	0.62
Duluth-Mix-Uni-Gap	18.7	1.6	43.8	0.56

Table 6: Supervised recall (SR) (test set split:80% mapping, 20% evaluation)

ters than the number of GS senses, although V-Measure does not increase monotonically with the number of clusters increasing. For that reason, we introduced the second unsupervised evaluation measure (paired F-Score) that penalises systems when they produce: (1) a higher number of clusters (low recall) or (2) a lower number of clusters (low precision), than the GS number of senses.

Table 5 shows the performance of systems using the second unsupervised evaluation measure. In this evaluation, we observe that most of the systems perform better than *Random*. Despite that, none of the systems outperform the *MFS* baseline. It seems that systems generating a smaller number of clusters than the GS number of senses are biased towards the *MFS*, hence they are not able to perform better. On the other hand, systems generating a higher number of clusters are penalised by this measure. Systems generating a number of clusters roughly the same as the GS tend to conflate the GS senses lot more than the *MFS*.

4.2 Supervised evaluation results

Table 6 shows the results of this evaluation for a 80-20 test set split, i.e. 80% for mapping and 20% for evaluation. The last columns shows the average number of GS senses identified by each system in the five splits of the evaluation datasets. Overall, 14 systems outperform the *MFS*, while 17 of them perform better than *Random*. The ranking of systems in nouns and verbs is different. For instance, the highest ranked system in nouns is *UoY*, while in verbs *Duluth-Mix-Narrow-Gap*. It seems that depending on the part-of-speech of the target word, different algorithms, features and parameters' tuning have different impact.

The supervised evaluation changes the distribution of clusters by mapping each cluster to a weighted vector of senses. Hence, it can potentially favour systems generating a high number of homogeneous clusters. For that reason, we applied a second testing set split, where 60% of the testing corpus was used for mapping and 40% for evaluation. Reducing the size of the mapping corpus allows us to observe, whether the above statement is correct, since systems with a high number of clusters would suffer from unreliable mapping.

Table 7 shows the results of the second supervised evaluation. The ranking of participants did not change significantly, i.e. we observe only different rankings among systems belonging to the same participant. Despite that, Table 7 also shows that the reduction of the mapping corpus has a different impact on systems generating a larger number of clusters than the GS number of senses.

For instance, *UoY* that generates 11.54 clusters outperformed the *MFS* by 3.77% in the 80-20 split and by 3.71% in the 60-40 split. The reduction of the mapping corpus had a minimal impact on its performance. In contrast, *KSU KDD* that generates 17.5 clusters was below the *MFS* by 6.49%

System	SR (%) (All)	SR (%) (Nouns)	SR (%) (Verbs)	#S
UoY	62.0	58.6	66.8	1.66
Duluth-WSI-Co	60.1	54.6	68.1	1.56
Duluth-WSI-Co-Gap	59.5	53.5	68.3	1.2
Duluth-WSI-SVD	59.5	53.5	68.3	1.73
Duluth-WSI	59.5	53.5	68.3	1.73
Duluth-WSI-Gap	59.3	53.2	68.2	1.11
KCDC-PCGD	59.1	52.6	68.6	1.54
KCDC-PC-2	58.9	53.4	67.0	1.25
KCDC-PC	58.9	53.6	66.6	1.44
KCDC-GDC	58.3	52.1	67.3	1.41
KCDC-GD	58.3	51.9	67.6	1.42
MFS	58.3	52.5	66.7	1
KCDC-PT	58.3	52.2	67.1	1.11
Duluth-WSI-SVD-Gap	58.2	52.5	66.7	1.01
KCDC-GD-2	57.9	51.7	67.0	1.44
Duluth-R-12	57.7	51.7	66.4	1.27
Duluth-R-13	57.6	51.1	67.0	1.48
Hermit	57.3	52.5	64.2	2.27
Duluth-R-15	56.5	50.0	66.1	1.76
Random	56.5	50.2	65.7	1.65
Duluth-Mix-Narrow-Gap	56.2	47.7	68.6	1.51
Duluth-Mix-Narrow-PK2	55.7	46.9	68.5	1.51
Duluth-R-110	53.6	46.7	63.6	2.18
Duluth-MIX-PK2	50.5	39.7	66.1	1.31
KSU KDD	50.4	44.3	59.4	1.92
Duluth-Mix-Gap	49.8	38.9	65.6	1.04
Duluth-Mix-Uni-PK2	19.1	1.8	44.4	0.63
Duluth-Mix-Uni-Gap	18.9	1.5	44.2	0.56

Table 7: Supervised recall (SR) (test set split:60% mapping, 40% evaluation)

in the 80-20 split and by 7.83% in the 60-40 split. The reduction of the mapping corpus had a larger impact in this case. This result indicates that the performance in this evaluation also depends on the distribution of instances within the clusters. Systems generating a skewed distribution, in which a small number of homogeneous clusters tag the majority of instances and a larger number of clusters tag only a few instances, are likely to have a better performance than systems that produce a more uniform distribution.

5 Conclusion

We presented the description, evaluation framework and assessment of systems participating in the SemEval-2010 sense induction task. The evaluation has shown that the current state-of-the-art lacks unbiased measures that objectively evaluate clustering.

The results of systems have shown that their performance in the unsupervised and supervised evaluation settings depends on cluster granularity along with the distribution of instances within the clusters. Our future work will focus on the assessment of sense induction on a task-oriented basis as well as on clustering evaluation.

Acknowledgements

We gratefully acknowledge the support of the EU FP7 INDECT project, Grant No. 218086, the National Science Foundation Grant NSF-0715078, Consistent Criteria for Word Sense Disambiguation, and the GALE program of the Defense Advanced Research Projects Agency, Contract No. HR0011-06-C-0022, a subcontract from the BBN-AGILE Team.

References

Eneko Agirre and Aitor Soroa. 2007. SemEval-2007 Task 02: Evaluating Word Sense Induction and Discrimination Systems. In *Proceedings of SemEval-2007*, pages 7–12, Prague, Czech Republic. ACL.

Eneko Agirre, Olatz Ansa, David Martinez, and Eduard Hovy. 2001. Enriching Wordnet Concepts With Topic Signatures. *ArXiv Computer Science e-prints*.

Javier Artiles, Enrique Amigó, and Julio Gonzalo. 2009. The role of named entities in web people search. In *Proceedings of EMNLP*, pages 534–542. ACL.

Christiane Fellbaum. 1998. *Wordnet: An Electronic Lexical Database*. MIT Press, Cambridge, Massachusetts, USA.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: the 90% solution. In *Proceedings of NAACL, Companion Volume: Short Papers on XX*, pages 57–60. ACL.

Suresh Manandhar and Ioannis P. Klapaftis. 2009. Semeval-2010 Task 14: Evaluation Setting for Word Sense Induction & Disambiguation Systems. In *DEW '09: Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pages 117–122, Boulder, Colorado, USA. ACL.

Patrick Pantel and Dekang Lin. 2002. Discovering Word Senses from Text. In *KDD '02: Proceedings of the 8th ACM SIGKDD Conference*, pages 613–619, New York, NY, USA. ACM.

Andrew Rosenberg and Julia Hirschberg. 2007. V-measure: A Conditional Entropy-based External Cluster Evaluation Measure. In *Proceedings of the 2007 EMNLP-CoNLL Joint Conference*, pages 410–420, Prague, Czech Republic.

Yoshimasa Tsuruoka and Juníchi Tsujii. 2005. Bidirectional Inference With the Easiest-first Strategy for Tagging Sequence Data. In *Proceedings of the HLT-EMNLP Joint Conference*, pages 467–474, Morristown, NJ, USA.

Jean Véronis. 2004. Hyperlex: Lexical Cartography for Information Retrieval. *Computer Speech & Language*, 18(3):223–252.