SemEval-2010 Task 14: Evaluation Setting for Word Sense Induction & Disambiguation Systems

Suresh Manandhar
Department of Computer Science
University of York
York, UK, YO10 5DD
suresh@cs.york.ac.uk

Ioannis P. Klapaftis
Department of Computer Science
University of York
York, UK, YO10 5DD
giannis@cs.york.ac.uk

Abstract

This paper presents the evaluation setting for the SemEval-2010 Word Sense Induction (WSI) task. The setting of the SemEval-2007 WSI task consists of two evaluation schemes, i.e. *unsupervised evaluation* and *supervised evaluation*. The first one evaluates WSI methods in a similar fashion to Information Retrieval exercises using F-Score. However, F-Score suffers from the *matching problem* which does not allow: (1) the assessment of the entire membership of clusters, and (2) the evaluation of all clusters in a given solution. In this paper, we present the use of V-measure as a measure of objectively assessing WSI methods in an unsupervised setting, and we also suggest a small modification on the supervised evaluation.

1 Introduction

WSI is the task of identifying the different senses (uses) of a target word in a given text. WSI is a field of significant value, because it aims to overcome the limitations originated by representing word senses as a fixed-list of dictionary definitions. These limitations of hand-crafted lexicons include the use of general sense definitions, the lack of explicit semantic and topical relations between concepts (Agirre et al., 2001), and the inability to reflect the exact content of the context in which a target word appears (Véronis, 2004).

Given the significance of WSI, the objective assessment and comparison of WSI methods is crucial. The first effort to evaluate WSI methods under a common framework (evaluation schemes & dataset) was undertaken in the SemEval-2007 WSI task (SWSI) (Agirre and Soroa, 2007), where two separate evaluation schemes were employed. The first one, *unsupervised evaluation*, treats the WSI results as clusters of target word contexts and Gold Standard (GS) senses as classes. The traditional clustering measure of F-Score (Zhao et al., 2005) is used to assess the performance of WSI systems. The second evaluation scheme, *supervised evaluation*, uses the training part of the dataset in order to map the automatically induced clusters to GS senses. In the next step, the testing corpus is used to measure the performance of systems in a Word Sense Disambiguation (WSD) setting.

A significant limitation of F-Score is that it does not evaluate the make up of clusters beyond the majority class (Rosenberg and Hirschberg, 2007). Moreover, F-Score might also fail to evaluate clusters which are not matched to any GS class due to their small size. These two limitations define the *matching problem* of F-Score (Rosenberg and Hirschberg, 2007) which can lead to: (1) identical scores between different clustering solutions, and (2) inaccurate assessment of the clustering quality.

The supervised evaluation scheme employs a method in order to map the automatically induced clusters to GS senses. As a result, this process might change the distribution of clusters by mapping more than one clusters to the same GS sense. The outcome of this process might be more helpful for systems that produce a large number of clusters.

In this paper, we focus on analysing the SemEval-2007 WSI evaluation schemes showing their deficiencies. Subsequently, we present the use of V-

Proceedings of the NAACL HLT Workshop on Semantic Evaluations: Recent Achievements and Future Directions, pages 117–122,
Boulder, Colorado, June 2009.

measure (Rosenberg and Hirschberg, 2007) as an evaluation measure that can overcome the current limitations of F-Score. Finally, we also suggest a small modification on the supervised evaluation scheme, which will possibly allow for a more reliable estimation of WSD performance. The proposed evaluation setting will be applied in the SemEval-2010 WSI task.

2 SemEval-2007 WSI evaluation setting

The SemEval-2007 WSI task (Agirre and Soroa, 2007) evaluates WSI systems on 35 nouns and 65 verbs. The corpus consists of texts of the Wall Street Journal corpus, and is hand-tagged with OntoNotes senses (Hovy et al., 2006). For each target word tw, the task consists of firstly identifying the senses of tw (e.g. as clusters of target word instances, co-occurring words, etc.), and secondly tagging the instances of the target word using the automatically induced clusters. In the next sections, we describe and review the two evaluation schemes.

2.1 SWSI unsupervised evaluation

Let us assume that given a target word tw, a WSI method has produced 3 clusters which have tagged 2100 instances of tw. Table 1 shows the number of tagged instances for each cluster, as well as the common instances between each cluster and each gold standard sense.

F-Score is used in a similar fashion to Information Retrieval exercises. Given a particular gold standard sense gs_i of size a_i and a cluster c_j of size a_j, suppose a_{ij} instances in the class gs_i belong to c_j. Precision of class gs_i with respect to cluster c_j is defined as the number of their common instances divided by the total cluster size, i.e. $\mathrm{P}(gs_i, c_j) = \frac{a_{ij}}{a_j}$. The recall of class gs_i with respect to cluster c_j is defined as the number of their common instances divided by the total sense size, i.e. $\mathrm{R}(gs_i, c_j) = \frac{a_{ij}}{a_i}$. The F-Score of gs_i with respect to c_j, $F(gs_i, c_j)$, is then defined as the harmonic mean of $P(gs_i, c_j)$ and $R(gs_i, c_j)$.

The F-Score of class gs_i, $F(gs_i)$, is the maximum $F(gs_i, c_j)$ value attained at any cluster. Finally, the F-Score of the entire clustering solution is defined as the weighted average of the F-Scores of each GS sense (Formula 1), where q is the number of GS senses and N is the total number of target word instances. If the clustering is identical to the original classes in the datasets, F-Score will be equal to one. In the example of Table 1, F-Score is equal to 0.714.

	gs_1	gs_2	gs_3
cl_1	500	100	100
cl_2	100	500	100
cl_3	100	100	500

Table 1: Clusters & GS senses matrix.

$$F - Score = \sum_{i=1}^{q} \frac{|gs_i|}{N} F(gs_i) \quad (1)$$

As it can be observed, F-Score assesses the quality of a clustering solution by considering two different angles, i.e. *homogeneity* and *completeness* (Rosenberg and Hirschberg, 2007). Homogeneity refers to the degree that each cluster consists of data points, which primarily belong to a single GS class. On the other hand, completeness refers to the degree that each GS class consists of data points, which have primarily been assigned to a single cluster. A perfect homogeneity would result in a precision equal to 1, while a perfect completeness would result in a recall equal to 1.

Purity and entropy (Zhao et al., 2005) are also used in SWSI as complementary measures. However, both of them evaluate only the homogeneity of a clustering solution disregarding completeness.

2.2 SWSI supervised evaluation

In supervised evaluation, the target word corpus is split into a testing and a training part. The training part is used to map the automatically induced clusters to GS senses. In the next step, the testing corpus is used to evaluate WSI methods in a WSD setting.

Let us consider the example shown in Table 1 and assume that this matrix has been created by using the training part of our corpus. Table 1 shows that cl_1 is more likely to be associated with gs_1, cl_2 is more likely to be associated with gs_2, and cl_3 is more likely to be associated with gs_3. This information from the training part is utilised to map the clusters to GS senses.

Particularly, the matrix shown in Table 1 is normalised to produce a matrix M, in which each entry depicts the conditional probability $P(gs_i|cl_j)$. Given an instance I of tw from the testing corpus, a row cluster vector IC is created, in which

System	**F-Sc.**	Pur.	Ent.	# Cl.	**WSD**
1c1w-MFS	78.9	79.8	45.4	1	78.7
UBC-AS	78.7	80.5	43.8	1.32	78.5
upv_si	66.3	83.8	33.2	5.57	79.1
UMND2	66.1	81.7	40.5	1.36	80.6
I2R	63.9	84.0	32.8	3.08	81.6
UOY	56.1	86.1	27.1	9.28	77.7
1c1inst	9.5	100	0	139	N/A

Table 2: SWSI Unsupervised & supervised evaluation.

each entry k corresponds to the score assigned to cl_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 gs_a is the winning sense for instance I. For example, if we produce the row cluster vector $[cl_1 = 0.8, cl_2 = 0.1, cl_3 = 0.1]$, and multiply it with the normalised matrix of Table 1, then we would get a row sense vector in which gs_1 would be the winning sense with a score equal to 0.6.

2.3 SWSI results & discussion

Table 2 shows the unsupervised and supervised performance of systems participating in SWSI. As far as the baselines is concerned, the *1c1w* baseline groups all instances of a target word into a single cluster, while the *1c1inst* creates a new cluster for each instance of a target word. Note that the *1c1w* baseline is equivalent to the *MFS* in the supervised evaluation. As it can be observed, a system with low entropy (high purity) does not necessarily achieve high F-Score. This is due to the fact that entropy and purity only measure the homogeneity of a clustering solution. For that reason, the *1c1inst* baseline achieves a perfect entropy and purity, although its clustering solution is far from ideal.

On the contrary, F-Score has a significant advantage over purity and entropy, since it measures both homogeneity (precision) and completeness (recall) of a clustering solution. However, F-Score suffers from the *matching problem*, which manifests itself either by not evaluating the entire membership of a cluster, or by not evaluating every cluster (Rosenberg and Hirschberg, 2007). The former situation is present, due to the fact that F-Score does not consider the make-up of the clusters beyond the majority class (Rosenberg and Hirschberg, 2007). For example, in Table 3 the F-Score of the clustering solution is 0.714 and equal to the F-Score of the clustering solution shown in Table 1, although these are two significantly different clustering solutions. In fact, the clustering shown in Table 3 should have a better homogeneity than the clustering shown in Table 1, since intuitively speaking each cluster contains fewer classes. Moreover, the second clustering should also have a better completeness, since each GS class contains fewer clusters.

	gs_1	gs_2	gs_3
cl_1	500	0	200
cl_2	200	500	0
cl_3	0	200	500

Table 3: Clusters & GS senses matrix.

An additional instance of the *matching problem* manifests itself, when F-Score fails to evaluate the quality of smaller clusters. For example, if we add in Table 3 one more cluster (cl_4), which only tags 50 additional instances of gs_1, then we will be able to observe that this cluster will not be matched to any of the GS senses, since cl_1 is matched to gs_1. Although F-Score will decrease since the recall of gs_1 will decrease, the evaluation setting ignores the perfect homogeneity of this small cluster.

In Table 2, we observe that no system managed to outperform the *1c1w* baseline in terms of F-Score. At the same time, some systems participating in SWSI were able to outperform the equivalent of the *1c1w* baseline (MFS) in the supervised evaluation. For example, *UBC-AS* achieved the best F-Score close to the *1c1w* baseline. However, by looking at its supervised recall, we observe that it is below the MFS baseline.

A clustering solution, which achieves high supervised recall, does not necessarily achieve high F-Score. One reason for that stems from the fact that F-Score penalises systems for getting the number of GS classes wrongly, as in *1c1inst* baseline. According to Agirre & Soroa (2007), supervised evaluation seems to be more neutral regarding the number of induced clusters, because clusters are mapped into a weighted vector of senses, and therefore inducing a number of clusters similar to the number of senses is not a requirement for good results.

However, a large number of clusters might also lead to an unreliable mapping of clusters to GS senses. For example, high supervised recall also

means high purity and low entropy as in *I2R*, but not vice versa as in *UOY*. *UOY* produces a large number of clean clusters, in effect suffering from an unreliable mapping of clusters to senses due to the lack of adequate training data.

Moreover, an additional supervised evaluation of WSI methods using a different dataset split resulted in a different ranking, in which all of the systems outperformed the MFS baseline (Agirre and Soroa, 2007). This result indicates that the supervised evaluation might not provide a reliable estimation of WSD performance, particularly in the case where the mapping relies on a single dataset split.

3 SemEval-2010 WSI evaluation setting

3.1 Unsupervised evaluation using V-measure

Let us assume that the dataset of a target word tw comprises of N instances (data points). These data points are divided into two partitions, i.e. a set of automatically generated clusters $C = \{c_j | j = 1 \dots n\}$ and a set of gold standard classes $GS = \{gs_i | gs = 1 \dots m\}$. Moreover, let a_{ij} be the number of data points, which are members of class gs_i and elements of cluster c_j.

V-measure assesses the quality of a clustering solution by explicitly measuring its homogeneity and its completeness (Rosenberg and Hirschberg, 2007). Recall that homogeneity refers to the degree that each cluster consists of data points which primarily belong to a single GS class. V-measure assesses homogeneity by examining the conditional entropy of the class distribution given the proposed clustering, i.e. $H(GS|C)$. $H(GS|C)$ quantifies the remaining entropy (uncertainty) of the class distribution given that the proposed clustering is known. As a result, when $H(GS|C)$ is 0, we have the perfectly homogeneous solution, since each cluster contains only those data points that are members of a single class. However in an imperfect situation, $H(GS|C)$ depends on the size of the dataset and the distribution of class sizes. As a result, instead of taking the raw conditional entropy, V-measure normalises it by the maximum reduction in entropy the clustering information could provide, i.e. $H(GS)$.

Formulas 2 and 3 define $H(GS)$ and $H(GS|C)$. When there is only a single class ($H(GS) = 0$), any clustering would produce a perfectly homogeneous solution. In the worst case, the class distribution within each cluster is equal to the overall class distribution ($H(GS|C) = H(GS)$), i.e. clustering provides no new information. Overall, in accordance with the convention of 1 being desirable and 0 undesirable, the homogeneity (h) of a clustering solution is 1 if there is only a single class, and $1 - \frac{H(GS|C)}{H(GS)}$ in any other case (Rosenberg and Hirschberg, 2007).

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

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

Symmetrically to homogeneity, completeness refers to the degree that each GS class consists of data points, which have primarily been assigned to a single cluster. To evaluate completeness, V-measure examines the distribution of cluster assignments within each class. The conditional entropy of the cluster given the class distribution, $H(C|GS)$, quantifies the remaining entropy (uncertainty) of the cluster given that the class distribution is known.

Consequently, when $H(C|GS)$ is 0, we have the perfectly complete solution, since all the data points of a class belong to the same cluster. Therefore, symmetrically to homogeneity, the completeness c of a clustering solution is 1 if there is only a single cluster ($H(C) = 0$), and $1 - \frac{H(C|GS)}{H(C)}$ in any other case. In the worst case, completeness will be equal to 0, particularly when $H(C|GS)$ is maximal and equal to $H(C)$. This happens when each GS class is included in all clusters with a distribution equal to the distribution of sizes (Rosenberg and Hirschberg, 2007). Formulas 4 and 5 define $H(C)$ and $H(C|GS)$. Finally h and c can be combined and produce V-measure, which is the harmonic mean of homogeneity and completeness.

$$H(C) = -\sum_{j=1}^{|C|} \frac{\sum_{i=1}^{|GS|} a_{ij}}{N} \log \frac{\sum_{i=1}^{|GS|} a_{ij}}{N} \quad (4)$$

$$H(C|GS) = -\sum_{i=1}^{|GS|} \sum_{j=1}^{|C|} \frac{a_{ij}}{N} \log \frac{a_{ij}}{\sum_{k=1}^{|C|} a_{ik}} \quad (5)$$

Returning to our clustering example in Table 1, its V-measure is equal to 0.275. In section 2.3, we also presented an additional clustering (Table 3), which had the same F-Score as the clustering in Table 1, despite the fact that it intuitively had a better completeness and homogeneity. The V-measure

of the second clustering solution is equal to 0.45, and higher than the V-measure of the first clustering. This result shows that V-measure is able to discriminate between these two clusterings by considering the make-up of the clusters beyond the majority class. Furthermore, it is straightforward from the description in this section, that V-measure evaluates each cluster in terms of homogeneity and completeness, unlike F-Score which relies on a post-hoc matching.

3.2 V-measure results & discussion

Table 4 shows the performance of SWSI participating systems according to V-measure. The last four columns of Table 4 show the weighted average homogeneity and completeness for nouns and verbs. Note that the homogeneity and completeness columns are weighted averages over all nouns or verbs, and are not used for the calculation of the weighted average V-measure (second column). The latter is calculated by measuring for each target word's clustering solution the harmonic mean of homogeneity and completeness separately, and then producing the weighted average.

As it can be observed in Table 4, all WSI systems have outperformed the random baseline which means that they have learned useful information. Moreover, Table 4 shows that on average all systems have outperformed the *1c1w* baseline, which groups the instances of a target word to a single cluster. The completeness of the *1c1w* baseline is equal to 1 by definition, since all instances of GS classes are grouped to a single cluster. However, this solution is as inhomogeneous as possible and causes a homogeneity equal to 0 in the case of nouns. In the verb dataset however, some verbs appear with only one sense, in effect causing the *1c1w* homogeneity to be equal to 1 in some cases, and the average V-measure greater than 0.

In Table 4, we also observe that the *1c1inst* baseline achieves a high performance. In nouns only *I2R* is able to outperform this baseline, while in verbs the *1c1inst* baseline achieves the highest result. By the definition of homogeneity (section 3.1), this baseline is perfectly homogeneous, since each cluster contains one instance of a single sense. However, its completeness is not 0, as one might intuitively expect. This is due to the fact that V-measure considers as the worst solution in terms of completeness the one, in which each class is represented by every cluster, and specifically with a distribution equal to the distribution of cluster sizes (Rosenberg and Hirschberg, 2007). This worst solution is not equivalent to the *1c1inst*, hence completeness of *1c1inst* is greater than 0. Additionally, completeness of this baseline benefits from the fact that around 18% of GS senses have only one instance in the test set. Note however, that on average this baseline achieves a lower completeness than most of the systems.

Another observation from Table 4 is that *upv_si* and *UOY* have a better ranking than in Table 2. Note that these systems have generated a higher number of clusters than the GS number of senses. In verbs *UOY* has been extensively penalised by the F-Score. The inspection of their answers shows that both systems generate highly skewed distributions, in which a small number of clusters tag the majority of instances, while a larger number tag only a few. As mentioned in sections 2.1 and 2.3, these small clusters might not be matched to any GS sense, hence they will decrease the unsupervised recall of a GS class, and consequently the F-Score. However, their high homogeneity is not considered in the calculation of F-Score. On the contrary, V-measure is able to evaluate the quality of these small clusters, and provide a more objective assessment.

Finally, in our evaluation we observe that *I2R* has on average the highest performance among the SWSI methods. This is due to its high V-measure in nouns, but not in verbs. Particularly in nouns, *I2R* achieves a consistent performance in terms of homogeneity and completeness without being biased towards one of them, as is the case for the rest of the systems. For example, *UOY* and *upv_si* achieve on average the highest homogeneity (42.5 & 32.8 resp.) and the worst completeness (11.5 & 13.2 resp.). The opposite picture is present for *UBC-AS* and *UMND2*. Despite that, *UBC-AS* and *UMND2* perform better than *I2R* in verbs, due to the small number of generated clusters (high completeness), and a reasonable homogeneity mainly due to the existence of verbs with one GS sense.

3.3 Modified supervised WSI evaluation

In section 2.3, we mentioned that supervised evaluation might favor methods which produce many

System	V-measure			Homogeneity		Completeness	
	Total	**Nouns**	**Verbs**	Nouns	Verbs	Nouns	Verbs
1c1inst	21.6	19.2	24.3	100.0	100.0	11.3	15.8
I2R	16.5	22.3	10.1	31.6	27.3	20.0	10.0
UOY	15.6	17.2	13.9	38.9	46.6	12.0	11.1
upv_si	15.3	18.2	11.9	37.1	28.0	14.5	11.8
UMND2	12.1	12.0	12.2	18.1	15.3	55.8	63.6
UBC-AS	7.8	3.7	12.4	4.0	13.7	90.6	93.0
Rand	7.2	4.9	9.7	12.0	30.0	14.1	14.3
1c1w	6.3	0.0	13.4	0.0	13.4	100.0	100.0

Table 4: V-Measure, homogeneity and completeness of SemEval-2007 WSI systems. The range of V-measure, homogeneity & completeness is 0-100.

clusters, since the mapping step can artificially increase completeness. Furthermore, we have shown that generating a large number of clusters might lead to an unreliable mapping of clusters to GS senses due to the lack of adequate training data.

Despite that, the supervised evaluation can be considered as an application-oriented evaluation, since it allows the transformation of unsupervised WSI systems to semi-supervised WSD ones. Given the great difficulty of unsupervised WSD systems to outperform the MFS baseline as well as the SWSI results, which show that some systems outperform the MFS by a significant amount in nouns, we believe that this evaluation scheme should be used to compare against supervised WSD methods.

In section 2.3, we also mentioned that the supervised evaluation on two different test/train splits provided a different ranking of methods, and more importantly a different ranking with regard to the MFS. To deal with that problem, we believe that it would be reasonable to perform k-fold cross validation in order to collect statistically significant information.

4 Conclusion

We presented and discussed the limitations of the SemEval-2007 evaluation setting for WSI methods. Based on our discussion, we described the use of V-measure as the measure of assessing WSI performance on an unsupervised setting, and presented the results of SWSI WSI methods. We have also suggested a small modification on the supervised evaluation scheme, which will allow for a more reliable estimation of WSD performance. The new evaluation setting will be applied in the SemEval-2010 WSI task.

Acknowledgements

This work is supported by the European Commission via the EU FP7 INDECT project, Grant No. 218086, Research area: SEC-2007-1.2-01 Intelligent Urban Environment Observation System. The authors would like to thank the anonymous reviewers for their useful comments.

References

Eneko Agirre and Aitor Soroa. 2007. Semeval-2007 task 02: Evaluating word sense induction and discrimination systems. In *Proceedings of the 4rth International Workshop on Semantic Evaluations*, pages 7–12, Prague, Czech Republic, June. ACL.

Eneko Agirre, Olatz Ansa, David Martinez, and Eduard Hovy. 2001. Enriching wordnet concepts with topic signatures. In *Proceedings of the NAACL workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*. ACL.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: The 90% solution. In *Proceedings of the Human Language Technology / North American Association for Computational Linguistics conference*, New York, USA.

Andrew Rosenberg and Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 410–420.

Jean Véronis. 2004. Hyperlex: lexical cartography for information retrieval. *Computer Speech & Language*, 18(3):223–252.

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