SemEval-2013 Task 13: Word Sense Induction for Graded and Non-Graded Senses

David Jurgens
Dipartimento di Informatica
Sapienza Università di Roma
jurgens@di.uniroma1.it

Ioannis Klapaftis
Search Technology Center Europe
Microsoft
ioannisk@microsoft.com

Abstract

Most work on word sense disambiguation has assumed that word usages are best labeled with a single sense. However, contextual ambiguity or fine-grained senses can potentially enable multiple sense interpretations of a usage. We present a new SemEval task for evaluating Word Sense Induction and Disambiguation systems in a setting where instances may be labeled with multiple senses, weighted by their applicability. Four teams submitted nine systems, which were evaluated in two settings.

1 Introduction

Word Sense Disambiguation (WSD) attempts to identify which of a word's meanings applies in a given context. A long-standing task, WSD is fundamental to many NLP applications (Navigli, 2009). Typically, each usage of a word is treated as expressing only a single sense. However, contextual ambiguity as well as the relatedness of certain meanings can potentially elicit multiple sense interpretations. Recent work has shown that annotators find multiple applicable senses in a given target word context when using fine-grained sense inventories such as WordNet (Véronis, 1998; Murray and Green, 2004; Erk et al., 2009; Passonneau et al., 2012b; Jurgens, 2013; Navigli et al., 2013). Such contexts would be better annotated with multiple sense labels, weighting each sense according to its applicability (Erk et al., 2009; Jurgens, 2013), in effect allowing ambiguity or multiple interpretations to be explicitly modeled. Accordingly, the first goal of this task is to evaluate WSD systems in a setting where instances may be labeled with one or more senses, weighted by their applicability.

WSD methods are ultimately defined and potentially restricted by their choice in sense inventory; for example, a sense inventory may have insufficient sense-annotated data to build WSD systems for specific types of text (e.g., social media), or the inventory may lack domain-specific senses. Word Sense Induction (WSI) has been proposed as a method for overcoming such limitations by learning the senses automatically from text. In essence, a WSI algorithm acts as a lexicographer by grouping word usages according to their shared meaning. The second goal of this task is to assess the performance of WSI algorithms when they are able to model multiple meanings of a usage with graded senses.

Task 12 focuses on disambiguating senses for 50 target lemmas: 20 nouns, 20 verbs, and 10 adjectives (Sec. 2). Since the Task evaluates only unsupervised systems, no training data was provided; however, to enable more comparison, Unsupervised WSD systems were also allowed to participate. Participating systems were evaluated in two settings (Sec. 3), depending on whether they used induced senses or WordNet 3.1 senses for their annotations. The results (Sec. 5) demonstrate a substantial improvement over the competitive most frequent sense baseline.

2 Task Description

This task required participating systems to annotate instances of nouns, verb, and adjectives using WordNet 3.1 (Fellbaum, 1998), which was selected due to its fine-grained senses. Participants could label each instance with one or more senses, weighting

*Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 290–299, Atlanta, Georgia, June 14-15, 2013.

We all are relieved to lay aside our fight-or-flight reflexes and to commemorate our births from out of the **dark** centers of the women, to feel the complexity of our love and frustration with each other, to stretch our cognition to encompass the thoughts of every entity we know.
`dark%3:00:01::` – devoid of or deficient in light or brightness; shadowed or black `dark%3:00:00::` – secret
I **ask** because my practice has always been to allow about five minutes grace, then remove it.
`ask%2:32:02::` – direct or put; seek an answer to `ask%2:32:04::` – address a question to and expect an answer from

Table 1: Example instances with multiple senses due to intended double meanings (top) or contextual ambiguity (bottom). Senses are specified using their WordNet 3.1 sense keys.

each by their applicability. Table 1 highlights two example contexts where multiple senses apply. The first example shows a case of an intentional double meaning that evokes both the physical aspect of *dark.a* as being devoid of light and the causal result of being secret. In contrast, the second example shows a case of multiple interpretations from ambiguity; a different preceding context could generate the alternate interpretations "I ask [*you*] because" (sense `ask%2:32:04::`) or "I ask [*the question*] because" (sense `ask%2:32:02::`).

2.1 Data

Three datasets were provided with the task. The trial dataset provided weighted word sense annotations using the data gathered by Erk et al. (2009). The trial dataset consisted of 50 contexts for eight words, where each context was labeled with WordNet 3.0 sense ratings from three untrained lexicographers.

Due to the unsupervised nature of the task, participants were not provided with sense-labeled training data. However, WSI systems were provided with the ukWaC corpus (Baroni et al., 2009) to use in inducing senses. Previous SemEval WSI tasks had provided participants with corpora specific to the task's target terms; in contrast, this task opted to use a large corpus to enable WSI methods that require corpus-wide statistics, e.g., statistical associations.

Test data was drawn from the Open American National Corpus (Ide and Suderman, 2004, OANC) across a variety of genres and from both the spoken and written portions of the corpus, summarized in Table 2. All contexts were manually inspected to ensure that the lemma being disambiguated was of the correct part of speech and had an interpretation that matched at least one WordNet 3.1 sense. This filtering also removed instances that were in a collocation, or had an idiomatic meaning. Ultimately, 4664 contexts were used as test data, with a minimum of 22 and a maximum of 100 contexts per word.

2.2 Sense Annotation

Recent work proposes to gather sense annotations using crowdsourcing in order to reduce the time and cost of acquiring sense-annotated corpora (Biemann and Nygaard, 2010; Passonneau et al., 2012b; Rumshisky et al., 2012; Jurgens, 2013). Therefore, we initially annotated the Task's data using the method of Jurgens (2013), where workers on Amazon Mechanical Turk (AMT) rated all senses of a word on a Likert scale from one to five, indicating the sense does not apply at all or completely applies, respectively. Twenty annotators were assigned per instance, with their ratings combined by selecting the most frequent rating. However, we found that while the annotators achieved moderate inter-annotator agreement (IAA), the resulting annotations were not of high enough quality to use in the Task's evaluations. Specifically, for some senses and contexts, AMT annotators required more information about sense distinctions than was feasible to integrate into the AMT setting, which led to consistent but incorrect sense assignments.

Therefore, the test data was annotated by the two authors, with the first author annotating all instances and the second author annotating a 10% sample of each lemma's instances in order to calculate IAA. IAA was calculated using Krippendorff's α (Krippendorff, 1980; Artstein and Poesio, 2008), which is an agreement measurement that adjusts for chance,

	Spoken		Written						
Genre	Face-to-face	Telephone	Fiction	Journal	Letters	Non-fiction	Technical	Travel Guides	*All*
Instances	52	699	127	2403	103	477	611	192	4664
Tokens	1742	30,700	3438	69,479	2238	11,780	17,337	4490	141,204
Mean senses/inst.	1.17	1.08	1.15	1.13	1.31	1.10	1.11	1.11	1.12

Table 2: Test data used in Task 12, divided according to source type

ranging in $(-1, 1]$ for interval data, where 1 indicates perfect agreement and -1 indicates systematic disagreement; two random annotations have an expected α of zero. We treat each sense and instance combination as a separate item to rate. The total IAA for the dataset was 0.504, and on individual words, ranged from 0.903 for *number.n* to 0.00 for *win.v*. While this IAA is less than the 0.8 recommended by Krippendorff (2004), it is consistent with the IAA distribution for the sense annotations of MASC on other parts of the OANC corpus: Passonneau et al. (2012a) reports an α of 0.88 to -0.02 with the MASI statistic (Passonneau et al., 2006).

Table 2 summarizes the annotation statistics for the Task's data. The annotation process resulted in far fewer senses per instance in the trial data, which we attribute to using trained annotators. An analysis across the corpora genres showed that the multiple-sense annotation rates were similar. Due to the variety of contextual sources, all lemmas were observed with at least two distinct senses.

3 Evaluation

We adopt a two-part evaluation setting used in previous SemEval WSI and WSD tasks (Agirre and Soroa, 2007; Manandhar et al., 2010). The first evaluation uses a traditional WSD task that directly compares WordNet sense labels. For WSI systems, their induced sense labels are converted to WordNet 3.1 labels via a mapping procedure. The second evaluation performs a direct comparison of the two sense inventories using clustering comparisons.

3.1 WSD Task

In the first evaluation, we adopt a WSD task with three objectives: (1) detecting which senses are applicable, (2) ranking senses by their applicability, and (3) measuring agreement in applicability ratings with human annotators. Each objectives uses a specific measurement: (1) the Jaccard Index, (2) positionally-weighted Kendall's τ similarity, and (3) a weighted variant of Normalized Discounted Cumulative Gain, respectively. Each measure is bounded in $[0, 1]$, where 1 indicates complete agreement with the gold standard. We generalize the traditional definition of WSD Recall such that it measures the average score for each measure across *all* instances, including those not labeled by the system. Systems are ultimately scored using the F1 measure between each objective's measure and Recall.

3.1.1 Transforming Induced Sense Labels

In the WSD setting, induced sense labels may be transformed into a reference inventory (e.g., WordNet 3.1) using a sense mapping procedure. We follow the 80/20 setup of Manandhar et al. (2010), where the corpus is randomly divided into five partitions, four of which are used to learn the sense mapping; the sense labels for the held-out partition are then converted and compared with the gold standard. This process is repeated so that each partition is tested once. For learning the sense mapping function, we use the distribution mapping technique of Jurgens (2012), which takes into account the sense applicability weights in both labelings.

3.1.2 Jaccard Index

Given two sets of sense labels for an instance, X and Y, the Jaccard Index is used to measure the agreement: $\frac{|X \cap Y|}{|X \cup Y|}$. The Jaccard Index is maximized when X and Y use identical labels, and is minimized when the sets of sense labels are disjoint.

3.1.3 Positionally-Weighted Kendall's τ

Rank correlations have been proposed for evaluating a system's ability to order senses by applicability; in previous work, both Erk and McCarthy (2009) and Jurgens (2012) propose rank correlation coefficients that assume all positions in the ranking are equally important. However, in the case of graded

sense evaluation, often only a few senses are applicable, with the applicability ratings of the remaining senses being relatively inconsequential. Therefore, we consider an alternate rank scoring based on Kumar and Vassilvitskii (2010), which weights the penalty of reordering the lower positions less than the penalty of reordering the first ranks.

Kendall's τ distance, K, is a measure of the number of item position swaps required to make two sequences identical. Kumar and Vassilvitskii (2010) extend this distance definition using a variable penalty function δ for the cost of swapping two positions, which we denote K_δ. By using an appropriate δ, K_δ can be biased towards the correctness of higher ranks by assigning a smaller δ to lower ranks. Because K_δ is a distance measure, its value range will be different depending on the number of ranks used. Therefore, to convert the measure to a similarity we normalize the distance to $[0, 1]$ by dividing by the maximum K_δ distance and then subtracting the distance from one. Given two rankings x and y where x is the reference by which y is to be measured, we may compute the normalized similarity using

$$K_\delta^{\text{sim}} = 1 - \frac{K_\delta(x, y)}{K_\delta^{\max}(x)}. \quad (1)$$

Equation 1 has its maximal value of one when ranking y is identical to ranking x, and its minimal value of zero when y is in the reverse order as x. We refer to this value as the positionally-weighted Kendall's τ similarity, K_δ^{sim}. As defined, K_δ^{sim} does not account for ties. Therefore, we arbitrarily break ties in a deterministic fashion for both rankings. Second, we define δ to assign higher cost to the first ranks: the cost to move an item into position i, δ_i, is defined as $\frac{n-(i+1)}{n}$, where n is the number of senses.

3.1.4 Weighted NDCG

To compare the applicability ratings for sense annotations, we recast the annotation process in an Information Retrieval setting: Given an example context acting as a query over a word's senses, the task is to retrieve all applicable senses, ranking and scoring them by their applicability. Moffat and Zobel (2008) propose using Discounted Cumulative Gain (DCG) as a method to compare a ranking against a baseline. Given (1) a gold standard weighting of the k senses applicable to a context, where w_i denotes the applicability for sense i in the gold standard, and (2) a ranking of the k senses by some method, the DCG may be calculated as $\sum_{i=1}^{k} \frac{2^{w_i+1}-1}{log_2(i+1)}$. DCG is commonly normalized to $[0, 1]$ so that the value is comparable when computed on rankings with different k and weight values. To normalize, the maximum value is calculated by first computing the DCG on the ranking when the k items are sorted by their weights, referred as the Ideal DCG (IDCG), and then normalizing as $NDCG = \frac{DCG}{IDCG}$.

The DCG only considers the weights assigned in the gold standard, which potentially masks importance differences in the weights assigned to the senses. Therefore, we propose weighting the DCG by the relative difference in the two weights. Given an alternate weighting of the k items, denoted as $\hat{w}_i$,

$$WDCG = \sum_{i=1}^{k} \frac{\frac{min(w_i,\hat{w}_i)}{max(w_i,\hat{w}_i)}\left(2^{w_i+1}-1\right)}{log_2(i)}. \quad (2)$$

The key impact in Equation 2 comes from weighting an item's contribution to the score by its relative deviation in absolute weight. A set of weights that achieves an equivalent ranking may have a low WDCG if the weights are significantly higher or lower than the reference. Equation 2 may be normalized in the same way as the DCG. We refer to this final normalized measure as the Weighted Normalized Discounted Cumulative Gain (WNDCG).

3.2 Sense Cluster Comparisons

Sense induction can be viewed as an unsupervised clustering task where usages of a word are grouped into clusters, each representing uses of the same meaning. In previous SemEval tasks on sense induction, instances were labeled with a single sense, which yields a *partition* over the instances into disjoint sets. The proposed partition can then be compared with a gold-standard partition using many existing clustering comparison methods, such as the V-Measure (Rosenberg and Hirschberg, 2007) or paired FScore (Artiles et al., 2009). Such cluster comparison methods measure the degree of similarity between the sense boundaries created by lexicographers and those created by WSI methods.

In the present task, instances are potentially labeled both with multiple senses and with weights

reflecting the applicability. This type of sense labeling produces a fuzzy clustering: An instance may belong to one or more sense clusters with its cluster membership relative to its weight for that sense. Formally, we refer to (1) a solution where the sets of instances overlap as a *cover* and (2) a solution where the sets overlap and instances may have partial memberships in a set as *fuzzy cover*.

We propose two new fuzzy measures for comparing fuzzy sense assignments: Fuzzy B-Cubed and Fuzzy Normalized Mutual Information. The two measures provide complementary information. B-Cubed summarizes the performance per instance and therefore provides an estimate of how well a system would perform on a new corpus with a similar sense distribution. In contrast, Fuzzy NMI is measured based on the clusters rather than the instances, thereby providing a performance analysis that is independent of the corpus sense distribution.

3.2.1 Fuzzy B-Cubed

Bagga and Baldwin (1998) proposed a clustering evaluation known as B-Cubed, which compares two partitions on a per-item basis. Amigó et al. (2009) later extended the definition of B-Cubed to compare overlapping clusters (i.e., covers). We generalize B-Cubed further to handle the case of fuzzy covers. B-Cubed is based on precision and recall, which estimate the fit between two clusterings, X and Y at the item level. For an item i, precision reflects how many items sharing a cluster with i in X appear in its cluster in Y; conversely, recall measures how many items sharing a cluster in Y with i also appear in its cluster in X. The final B-Cubed value is the harmonic mean of the two scores.

To generalize B-Cubed to fuzzy covers, we adopt the formalization of Amigó et al. (2009), who define item-based precision and recall functions, P and R, in terms of a correctness function, $C \rightarrow \{0, 1\}$. For notational brevity, let avg be a function that returns the mean value of a series, and $\mu_x(i)$ denote the set of clusters in clustering X of which item i is a member. B-Cubed precision and recall may therefore calculated over all n items:

$$\text{B-Cubed Precision} = \underset{i}{\text{avg}}[\underset{j \neq i \in \cup\mu_y(i)}{\text{avg}} P(i,j)] \quad (3)$$

$$\text{B-Cubed Recall} = \underset{i}{\text{avg}}[\underset{j \neq i \in \cup\mu_x(i)}{\text{avg}} R(i,j)]. \quad (4)$$

When comparing partitions, P and R are defined as 1 if two items cluster labels are identical. To generalize B-Cubed for fuzzy covers, we redefine P and R to account for differences in the partial cluster membership of items. Let $\ell_X(i)$ denote the set of clusters of which i is a member, and $w_k(i)$ denote the membership weight of item i in cluster k in X. We therefore define C with respect to X of two items as

$$\text{C}(i,j,X) = \sum_{k \in \ell_X(i) \cup \ell_X(j)} 1 - |w_k(i) - w_k(j)|. \quad (5)$$

Equation 5 is maximized when i and j have identical membership weights in the clusters of which they are members. Importantly, Equation 5 generalizes to the correctness operations both when comparing partitions and covers, as defined by Amigó et al. (2009). Item-based Precision and Recall are then defined using Equation 5 as $P(i,j,X) = \frac{\text{Min}(\text{C}(i,j,X),\text{C}(i,j,Y))}{\text{C}(i,j,X)}$ and $R(i,j,X) = \frac{\text{Min}(\text{C}(i,j,X),\text{C}(i,j,Y))}{\text{C}(i,j,Y)}$, respectively. These fuzzy generalizations are used in Equations 3 and 4.

3.2.2 Fuzzy Normalized Mutual Information

Mutual information measures the dependence between two random variables. In the context of clustering evaluation, mutual information treats the sense labels as random variables and measures the level of agreement in which instances are labeled with the same senses (Danon et al., 2005). Formally, mutual information is defined as $I(X;Y) = H(X) - (H(X|Y)$ where $H(X)$ denotes the entropy of the random variable X that represents a partition, i.e., the sets of instances assigned to each sense. Typically, mutual information is normalized to $[0, 1]$ in order to facilitate comparisons between multiple clustering solutions on the same scale (Luo et al., 2009), with $Max(H(X), H(Y))$ being the recommended normalizing factor (Vinh et al., 2010).

In its original formulation Mutual information is defined only to compare non-overlapping cluster partitions. Therefore, we propose a new definition of mutual information between fuzzy covers using extension of Lancichinetti et al. (2009) for calculating the normalized mutual information between covers. In the case of partitions, a clustering is represented as a discrete random variable whose states denote the probability of being assigned to each cluster. In

the fuzzy cover setting, each item may be assigned to multiple clusters and no longer has a binary assignment to a cluster, but takes on a value in $[0, 1]$. Therefore, each cluster X_i can be represented separately as a continuous random variable, with the entire fuzzy cover denoted as the variable $\mathbf{X}_{1...k}$, where the ith entry of $\mathbf{X}$ is the continuous random variable for cluster i. However, by modeling clusters using continuous domain, differential entropy must be used for the continuous variables; importantly, differential entropy does not obey the same properties as discrete entropy and may be negative.

To avoid calculating entropy in the continuous domain, we therefore propose an alternative method of computing mutual information based on discretizing the continuous values of X_i in the fuzzy setting. For the continuous random variable X_i, we discretize the value by dividing up probability mass into discrete bins. That is, the support of X_i is partitioned into disjoint ranges, each of which represents a discrete outcome of X_i. As a result, X_i becomes a *categorical distribution* over a set of weights ranges $\{w_1, \ldots, w_n\}$ that denote the strength of membership in the fuzzy set. With respect to sense annotation, this discretization process is analogous to having an annotator rate the applicability of a sense for an instance using a Likert scale instead of using a rational number within a fixed bound.

Discretizing the continuous cluster membership ratings into bins allows us to avoid the problematic interpretation of entropy in the continuous domain while still expanding the definition of mutual information from a binary cluster membership to one of degrees. Using the definition of X_i and Y_j as a categorical variables over discrete ratings, we may then estimate the entropy and joint entropy as follows.

$$H(X_i) = \sum_{i=1}^{n} p(w_i) \log_2 p(w_i) \quad (6)$$

where $p(w_i)$ is the probability of an instance being labeled with rating w_i Similarly, we may define the joint entropy of two fuzzy clusters as

$$H(X_k, Y_l) = \sum_{i=1}^{n} \sum_{j=1}^{m} p(w_i, w_j) \log_2 p(w_i, w_j) \quad (7)$$

where $p(w_i, w_j)$ is the probability of an instance being labeled with rating w_i in cluster X_k and w_j in cluster Y_l, and m denotes the number of bins for Y_l. The conditional entropy between two clusters may then be calculated as

$$H(X_k|Y_l) = H(X_k, Y_l) - H(Y_l).$$

Together, Equations 6 and 7 may be used to define $I(X, Y)$ as in the original definition. We then normalize using the method of McDaid et al. (2011). Based on the limited range of fuzzy memberships in $[0, 1]$, we selected uniformly distributed bins in $[0, 1]$ at 0.1 intervals when discretizing the membership weights for sense labelings.

3.3 Baselines

Task 12 included multiple baselines based on modeling different types of WSI and WSD systems. Due to space constraints, we include only the four most descriptive here: (1) **Semcor MFS** which labels each instance with the most frequent sense of that lemma in SemCor, (2) **Semcor Ranked Senses** baseline, which labels each instance with all of the target lemma's senses, ranked according to their frequency in SemCor, using weights $\frac{n-i+1}{n}$, where n is the number of senses and i is the rank, (3) **1c1inst** which labels each instance with its own induced sense and (4) **All-instances, One sense** which labels all instances with the same induced sense. The first two baselines directly use WordNet 3.1 senses, while the last two use induced senses.

4 Participating Systems

Four teams submitted nine systems, seven of which used induced sense inventories. **AI-KU** submitted three WSI systems based on a lexical substitution method; a language model is built from the target word's contexts in the test data and the ukWaC corpus and then Fastsubs (Yuret, 2012) is used to identify lexical substitutes for the target. Together, the contexts of the target and substitutes are used to build a distributional model using the S-CODE algorithm (Maron et al., 2010). The resulting contextual distributions are then clustered using K-means to identify word senses. The University of Melbourne (**Unimelb**) team submitted two WSI systems based on the approach of Lau et al. (2012). Their systems use a Hierarchical Dirichlet Process (Teh et al., 2006) to automatically infer the number of senses from contextual and positional features. Un-

		WSD F1			Cluster Comparison			
Team	System	Jac. Ind.	K_{δ}^{sim}	WNDCG	Fuzzy NMI	Fuzzy B-Cubed	#Cl	#S
AI-KU	Base	0.197	0.620	**0.387**	**0.065**	0.390	7.76	6.61
AI-KU	add1000	0.197	0.606	0.215	0.035	0.320	7.76	6.61
AI-KU	remove5-add1000	**0.244**	**0.642**	0.332	0.039	0.451	3.12	5.33
Unimelb	5p	0.218	0.614	0.365	0.056	0.459	2.37	5.97
Unimelb	50k	0.213	0.620	0.371	0.060	**0.483**	2.48	6.08
UoS	#WN Senses	0.192	0.596	0.315	0.047	0.201	8.08	6.77
UoS	top-3	0.232	0.625	0.374	0.045	0.448	3.00	5.44
La Sapienza	system-1	0.149	0.507	0.311	-	-	-	8.69
La Sapienza	system-2	0.149	0.510	0.383	-	-	-	8.67
All-instances, One sense		0.192	0.609	0.288	0.0	0.623	1.00	6.62
1c1inst		0.0	0.0	0.0	0.071	0.0	1.00	0.0
Semcor MFS		0.455	0.465	0.339	-	-	-	1.00
Semcor Ranked Senses		0.149	0.559	0.489	-	-	-	8.66

Table 3: Performance on the five evaluation measures for all system and selected baselines. Top system performances are marked in bold.

like other teams, the Unimelb systems were trained on a Wikipedia corpus instead of the ukWaC corpus. The University of Sussex (**UoS**) team submitted two WSI systems that use dependency-parsed features from the corpus, which are then clustered into senses using the MaxMax algorithm (Hope and Keller, 2013); the resulting fine-grained clusters are then combined based on their degree of separability. The **La Sapienza** team submitted two Unsupervised WSD systems based applying Personalized Page Rank (Agirre and Soroa, 2009) over a WordNet-based network to compare the similarity of each sense with the similarity of the context, ranking each sense according to its similarity.

Team	System	F1	NMI	B-Cubed
AI-KU	Base	**0.641**	**0.045**	0.351
AI-KU	add1000	0.601	0.023	0.288
AI-KU	remove5-add1000	0.628	0.026	0.421
Unimelb	5p	0.596	0.035	0.421
Unimelb	50k	0.605	0.039	**0.441**
UoS	#WN Senses	0.574	0.031	0.180
UoS	top-3	0.600	0.028	0.414
La Sapienza	System-1	0.204	-	-
La Sapienza	System-2	0.217	-	-
All-instances, One sense		0.569	0.0	0.570
1c1inst		0.0	0.018	0.0
Semcor MFS		0.477	0.0	0.570

Table 4: System performance in the single-sense setting. Top system performances are marked in bold.

5 Results and Discussion

Table 3 shows the main results for all instances. Additionally, we report the number of induced clusters used to label each sense as #Cl and the number of resulting WordNet 3.1 senses for each sense with #S. As in previous WSD tasks, the MFS baseline was quite competitive, outperforming all systems on detecting which senses were applicable, measured using the Jaccard Index. However, most systems were able to outperform the MFS baseline on ranking senses and quantifying their applicability.

Previous cluster comparison evaluations often faced issues with the measures being biased either towards the 1c1inst baseline or labeling all instances with the same sense. However, Table 3 shows that systems are capable of performing well in both the Fuzzy NMI and Fuzzy B-Cubed measures, thereby avoiding the extreme performance of either baseline.

An analysis of the systems' results showed that many systems labeled instances with a high number of senses, which could have been influenced by the trial data having significantly more instances labeled with multiple senses than the test data. Therefore, we performed a second analysis that partitioned the test set into two sets: those labeled with a single sense and those with multiple senses. For single-sense set, we modified the test setting to have systems also label instances with a single sense: (1) the sense mapping function for WSI systems (Sec. 3.1.1) was modified so that after the mapping,

		WSD F1			Cluster Comparison	
Team	System	Jac. Ind.	K_{δ}^{sim}	WNDCG	Fuzzy NMI	Fuzzy B-Cubed
AI-KU	Base	0.394	0.617	0.317	0.029	0.078
AI-KU	add1000	0.394	0.620	0.214	0.014	0.061
AI-KU	remove5-add1000	0.434	0.585	0.290	0.004	0.116
Unimelb	5p	**0.436**	0.585	0.286	0.019	0.130
Unimelb	5000k	0.414	0.602	0.298	0.021	**0.134**
UoS	#WN Senses	0.367	0.627	0.313	**0.036**	0.037
UoS	top-3	0.421	0.574	0.302	0.006	0.113
La Sapienza	system-1	0.263	0.660	0.447	-	-
La Sapienza	system-2	0.412	**0.694**	**0.536**	-	-
All-instances, One sense		0.387	0.635	0.254	0.0	0.130
1c1inst		0.0	0.0	0.0	0.300	0.0
Semcor MFS		0.283	0.373	0.197		
Semcor Ranked Senses		0.263	0.593	0.395		

Table 5: System performance on all instances labeled with multiple senses. Top system performances are marked in bold.

only the highest-weighted WordNet 3.1 sense was used, and (2) the La Sapienza system output was modified to retain only the highest weighted sense. In this single-sense setting, systems were evaluated using the standard WSD Precision and Recall measures; we report the F1 measure of Precision and Recall. The remaining subset of instances annotated with multiple senses were evaluated separately.

Table 4 shows the systems' performance on single-sense instances, revealing substantially increased performance and improvement over the MFS baseline for WSI systems. Notably, the performance of the best sense-remapped WSI systems surpasses the performance of many supervised WSD systems in previous WSD evaluations (Kilgarriff, 2002; Mihalcea et al., 2004; Pradhan et al., 2007; Agirre et al., 2010). This performance suggests that WSI systems using graded labels provide a way to leverage huge amounts of unannotated corpus data for finding sense-related features in order to train semi-supervised WSD systems.

Table 5 shows the performance on the subset of instances that were annotated with multiple senses. We note that in this setting, the mapping procedure transforms the All-Instances One Sense baseline into the average applicability rating for each sense in the test corpus. Notably, the La Sapienza systems sees a significant performance increase in this setting; their systems label each instance with all of the lemma's senses, which significantly degrades performance in the most common case where only a single sense applies. However, when multiple senses are known to be present, their method for quantifying sense applicability appears closest to the gold standard judgments. Furthermore, the majority of WSI systems are able to surpass all four baselines on identifying which senses are present and quantifying their applicability.

6 Conclusion

We have introduced a new evaluation setting for WSI and WSD systems where systems are measured by their ability to detect and weight multiple applicable senses for a single context. Four teams submitted nine systems, annotating a total of 4664 contexts for 50 words from the OANC. Many systems were able to surpass the competitive MFS baseline. Furthermore, when WSI systems were trained to produce only a single sense label, the performance of resulting semi-supervised WSD systems surpassed that of many supervised systems in previous WSD evaluations. Future work may assess the impact of graded sense annotations in a task-based setting. All materials have been released on the task website.[1]

Acknowledgments

We thank Rebecca Passonneau for her feedback and suggestions for target lemmas used in this task.

[1] http://www.cs.york.ac.uk/semeval-2013/task13/

References

Eneko Agirre and Aitor Soroa. 2007. Semeval-2007 task 2: Evaluating word sense induction and discrimination systems. In *Proceedings of the Fourth International Workshop on Semantic Evaluations*, pages 7–12. ACL.

Eneko Agirre and Aitor Soroa. 2009. Personalizing PageRank for Word Sense Disambiguation. In *Proceedings of EACL*, pages 33–41. ACL.

Eneko Agirre, Oier López De Lacalle, Christine Fellbaum, Andrea Marchetti, Antonio Toral, and Piek Vossen. 2010. SemEval-2010 task 17: All-words word sense disambiguation on specific domains. In *Proceedings of SemEval-2010*. ACL.

Enrique Amigó, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. 2009. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, 12(4):461–486.

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

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *Proceedings of the Linguistic Coreference Workshop at LREC*, pages 563–566.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Chris Biemann and Valerie Nygaard. 2010. Crowdsourcing wordnet. In *The 5th International Conference of the Global WordNet Association (GWC-2010)*.

Leon Danon, Albert Díaz-Guilera, Jordi Duch, and Alex Arenas. 2005. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(09):P09008.

Katrin Erk and Diana McCarthy. 2009. Graded word sense assignment. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 440–449. ACL.

Katrin Erk, Diana McCarthy, and Nicholas Gaylord. 2009. Investigations on word senses and word usages. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 10–18. ACL.

Christine Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press.

David Hope and Bill Keller. 2013. MaxMax: A Graph-Based Soft Clustering Algorithm Applied to Word Sense Induction. In *Proceedings of CICLing*, pages 368–381.

Nancy Ide and Keith Suderman. 2004. The american national corpus first release. In *Proceedings of the Fourth Language Resources and Evaluation Conference*, pages 1681–1684.

David Jurgens. 2012. An Evaluation of Graded Sense Disambiguation using Word Sense Induction. In *Proceedings of *SEM, the First Joint Conference on Lexical and Computational Semantics*. ACL.

David Jurgens. 2013. Embracing Ambiguity: A Comparison of Annotation Methodologies for Crowdsourcing Word Sense Labels. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*. ACL.

Adam Kilgarriff. 2002. English lexical sample task description. In *Proceedings of ACL-SIGLEX SENSEVAL-2 Workshop*.

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to Its Methodology*. Sage, Beverly Hills, CA.

Klaus Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology*. Sage, Thousand Oaks, CA, second edition.

Ravi Kumar and Sergei Vassilvitskii. 2010. Generalized distances between rankings. In *Proceedings of the 19th International Conference on World Wide Web (WWW)*, pages 571–580. ACM.

Andrea Lancichinetti, Santo Fortunato, and János Kertész. 2009. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, 11(3):033015.

Jey Han Lau, Paul Cook, Diana McCarthy, David Newman, and Timothy Baldwin. 2012. Word sense induction for novel sense detection. In *Proceedings of the 13th Conference of the European Chapter of the Association for computational Linguistics (EACL 2012)*.

Ping Luo, Hui Xiong, Guoxing Zhan, Junjie Wu, and Zhongzhi Shi. 2009. Information-theoretic distance measures for clustering validation: Generalization and normalization. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1249–1262.

Suresh Manandhar, Ioannis P. Klapaftis, Dmitriy Dligach, and Sameer S. Pradhan. 2010. SemEval-2010 task 14: Word sense induction & disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 63–68. ACL.

Yariv Maron, Michael Lamar, and Elie Bienenstock. 2010. Sphere embedding: An application to part-of-speech induction. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*.

Aaron F. McDaid, Derek Greene, and Neil Hurley. 2011. Normalized mutual information to evaluate overlapping community finding algorithms. arXiv:1110.2515.

Rada Mihalcea, Timothy Chklovski, and Adam Kilgarriff. 2004. The Senseval-3 English lexical sample task. In *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 25–28. ACL.

Alistair Moffat and Justin Zobel. 2008. Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems (TOIS)*, 27(1):2.

G. Craig Murray and Rebecca Green. 2004. Lexical knowledge and human disagreement on a wsd task. *Computer Speech & Language*, 18(3):209–222.

Roberto Navigli, David Jurgens, and Daniele Vanilla. 2013. Semeval-2013 task 12: Multilingual word sense disambiguation. In *Proceedings of the 7th International Workshop on Semantic Evaluation*.

Roberto Navigli. 2009. Word Sense Disambiguation: A Survey. *ACM Computing Surveys*, 41(2):1–69.

Rebecca Passonneau, Nizar Habash, and Owen Rambow. 2006. Inter-annotator agreement on a multilingual semantic annotation task. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, pages 1951–1956.

Rebecca J Passonneau, Collin Baker, Christiane Fellbaum, and Nancy Ide. 2012a. The MASC word sense sentence corpus. In *Proceedings of LREC*.

Rebecca J. Passonneau, Vikas Bhardwaj, Ansaf Salleb-Aouissi, and Nancy Ide. 2012b. Multiplicity and word sense: evaluating and learning from multiply labeled word sense annotations. *Language Resources and Evaluation*, pages 1–34.

Sameer S. Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. SemEval-2007 task 17: English lexical sample, SRL, and all-words. In *Proceedings of the 4th International Workshop on Semantic Evaluations*. ACL.

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)*. ACL.

Anna Rumshisky, Nick Botchan, Sophie Kushkuley, and James Pustejovsky. 2012. Word sense inventories by non-experts. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*.

Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2006. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

Jean Véronis. 1998. A study of polysemy judgments and inter-annotator agreement. In *Program and advanced papers of the Senseval workshop*.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. 2010. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *The Journal of Machine Learning Research*, 11:2837–2854.

Deniz Yuret. 2012. FASTSUBS: An Efcient Admissible Algorithm for Finding the Most Likely Lexical Substitutes Using a Statistical Language Model. *Computing Research Repository (CoRR)*.