# An Evaluation of Graded Sense Disambiguation using Word Sense Induction

**David Jurgens**[1,2]
[1]HRL Laboratories, LLC
Malibu, California, USA
[2]Department of Computer Science
University of California, Los Angeles
jurgens@cs.ucla.edu

## Abstract

Word Sense Disambiguation aims to label the sense of a word that best applies in a given context. Graded word sense disambiguation relaxes the single label assumption, allowing for multiple sense labels with varying degrees of applicability. Training multi-label classifiers for such a task requires substantial amounts of annotated data, which is currently not available. We consider an alternate method of annotating graded senses using Word Sense Induction, which automatically learns the senses and their features from corpus properties. Our work proposes three objective to evaluate performance on the graded sense annotation task, and two new methods for mapping between sense inventories using parallel graded sense annotations. We demonstrate that sense induction offers significant promise for accurate graded sense annotation.

## 1 Introduction

Word Sense Disambiguation (WSD) aims to identify the sense of a word in a given context, using a predefined sense inventory containing the word's different meanings (Navigli, 2009). Traditionally, WSD approaches have assumed that each occurrence of a word is best labeled with a single sense. However, human annotators often disagree about which sense is present (Passonneau et al., 2010), especially in cases where some of the possible senses are closely related (Chugur et al., 2002; McCarthy, 2006; Palmer et al., 2007).

Recently, Erk et al. (2009) have shown that in cases of sense ambiguity, a *graded* notion of sense labeling may be most appropriate and help reduce the ambiguity. Specifically, within a given context, multiple senses of a word may be salient to the reader, with different levels of applicability. For example, in the sentence

- The athlete **won** the gold metal due to her hard work and dedication.

multiple senses could be considered applicable for "won" according to the WordNet 3.0 sense inventory (Fellbaum, 1998):

1. win (be the winner in a contest or competition; be victorious)
2. acquire, win, gain (win something through one's efforts)
3. gain, advance, win, pull ahead, make headway, get ahead, gain ground (obtain advantages, such as points, etc.)
4. succeed, win, come through, bring home the bacon, deliver the goods (attain success or reach a desired goal)

In this context, many annotators would agree that the athlete has both won an object (the gold metal itself) and won a competition (signified by the gold medal). Although contexts can be constructed to elicit only one of these senses, in the example above, a graded annotation best matches human perception.

Graded word sense (GWS) annotation offers significant advantages for sense annotation with a fine-grained sense inventory. However, creating a sufficiently large annotated corpus for training supervised GWS disambiguation models presents a significant challenge, i.e., the laborious task of gathering annotations for all combinations of a word's senses, along with variation in those senses applicabilities. To our knowledge, Erk et al. (2009) have provided the only data set with GWS annotations for 11 terms.

*First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 189–198,
Montréal, Canada, June 7-8, 2012.

Therefore, we consider the use of Word Sense Induction (WSI) for GWS annotation. WSI removes the need for substantial training data by automatically deriving a word's senses and associated sense features through examining its contextual uses. Furthermore, the data-driven sense discovery defines senses as they are present in the corpus, which may identify usages not present in traditional sense inventories (Lau et al., 2012). Last, many WSI models represent senses loosely as abstractions over usages, which potentially may transfer well to expressing GWS annotations as a blend of their sense usages.

In this paper, we consider the performance of WSI models on a GWS task. The contributions of this paper are as follows. First, in Sec. 2, we motivate three GWS annotation objectives and propose corresponding measures that provide fine-grained analysis of the capabilities of different WSI models. Second, in Sec. 4, we propose two new sense mapping procedures for converting an induced sense inventory to a reference sense inventory when GWS annotations are present, and demonstrate significant performance improvement using these procedures on GWS annotation. Last, in Sec. 5, we demonstrate a complete evaluation framework using three graph-based WSI models as examples, generating several insights for how to better evaluate GWS disambiguation systems.

## 2 Evaluating GWS Annotations

Graded word sense annotation conveys multiple levels of information, both in which senses are present and their relative levels of applicability; and so, no single evaluation measure alone is appropriate for assessing GWS annotation capability. Therefore, we propose three objectives for the evaluating the sense labeling: (1) Detection of which senses are present, (2) Ranking senses according to applicability, and (3) Perception of the graded presence of each sense. We separate the three objectives as a way to evaluate how well different techniques perform on each aspect individually, which may encourage future work in ensemble WSD methods that use combinations of the techniques. Figure 1 illustrates each evaluation on example annotations. We note that Erk and McCarthy (2009) have also proposed an alternate set of evaluation measures for GWS annotations. Where applicable, we describe and compare their measures to ours for the three objectives.

In the following definitions, let $S_G^i$ refer to the set of senses $\{s_1, \ldots, s_n\}$ present in context $i$ according to the gold standard, and similarly, let $S_L^i$ refer to the set of senses for context $i$ as labeled by a WSD system using the same sense inventory. Let $per_i(s_j)$ refer to the perceived numeric applicability rating of sense $s_j$ in context $i$.

**Detection** measures the ability to accurately identify which senses are applicable in a given context, independent of their applicability. While the most basic of the evaluations, systems that are highly accurate at multi-sense detection could be used for recognizing ambiguous contexts where multiple senses are applicable or for evaluating the granularity of sense ontologies by testing for correlations between senses in a multi-sense labeling. Detection is measured using the Jaccard Index between $S_G^i$ and $S_L^i$ for a given context $i$: $\frac{S_G^i \cap S_L^i}{S_G^i \cup S_L^i}$

**Ranking** measures the ability to order the senses present in context $i$ according to their applicability but independent of their quantitative applicability scores. Even though multiple senses are present, a context may have a clear primary senses. By providing a ranking in agreement with human judgements, systems create a primary sense label for each context. When the induced senses are mapped to a sense inventory, selecting the primary sense is analogous to non-graded WSD where a context is labeled with its most applicable sense.

To compare sense rankings, we use Goodman and Kruskal's $\gamma$, which is related to Kendall's $\tau$ rank correlation. When the data has many tied ranks, $\gamma$ is preferable to both Kendall's $\tau$ as well as Spearman's $\rho$ rank correlation (Siegel and Castellan Jr., 1988), the latter of which is used by Erk and McCarthy (2009) for evaluating sense rankings. The use of $\gamma$ was motivated by our observation that in the GWS dataset (described later in Section 5.1), roughly 65% of the instances contained at least one tied ranking between senses.

To compute $\gamma$, we examine all pair-wise combinations of senses $(s_i, s_j)$ of the target word. Let $r_G(s_i)$ and $r_L(s_i)$ denote the ranks of sense $s_i$ in the gold standard and provided annotations. In the event that a ranking does not include senses, all of the inapplicable senses are assigned a tied rank

| Instance | Gold Standard Annotation |
|---|---|
| The athlete **won** the gold metal due to her hard work and dedication. | win.v.1: 0.6, win.v.2: 0.4<br>(not applicable: win.v.3, win.v.4) |

| Test Annotation | Detection | Ranking | Perception |
|---|---|---|---|
| win.v.1: 0.7, win.v.2: 0.3 | 1.0 | 1.0 | 0.983 |
| win.v.1: 1.0 | 0.5 | 1.0 | 0.832 |
| win.v.2: 1.0 | 0.5 | 0.333 | 0.554 |
| win.v.3: 0.5, win.v.1: 0.3, win.v.4: 0.2 | 0.25 | -0.2 | 0.405 |

Figure 1: Example annotations of the same context compared with the gold standard according to Detection, Ranking, and Perception.

lower than the least applicable sense; i.e., for $m$ applicable senses, all inapplicable senses have rank $m$+1. A pair of senses, $(s_i, s_j)$ is said to be concordant if $r_G(s_i) < r_G(s_j)$ and $r_L(s_i) < r_L(s_j)$ or $r_G(s_i) > r_G(s_j)$ and $r_L(s_i) > r_L(s_j)$, and discordant otherwise. $\gamma$ is defined as $\frac{c-d}{c+d}$ where $c$ is the number of concordant pairs and $d$ is the number of discordant.

**Perception** measures the ability to equal human judgements on the levels of applicability for each sense in a context. Unlike ranking, this evaluation quantifies the difference in sense applicability. As a potential application, these differences can be used to quantify the contextual ambiguity. For example, the relative applicability differences can be used to distinguish between ambiguous contexts where multiple highly-applicable senses exist and unambiguous contexts where a single main sense exists but other senses are still minimally applicable.

To quantify Perception, we compare sense labelings using the cosine similarity. Each labeling is represented as a vector with a separate component for each sense, whose value is the applicability of that sense. The Perception for two annotations of context $j$ is then calculated as

$$\frac{\sum_i per_j(s_i^G) \times per_j(s_i^L)}{\sqrt{\sum_i per_j(s_i^G)^2} \times \sqrt{\sum_i per_j(s_i^L)^2}}.$$

Note that because all sense perceptibilities are non-negative, the cosine similarity is bounded to $[0, 1]$.

Erk and McCarthy (2009) propose an alternate measure for comparing the applicability values using the Jensen-Shannon divergence. The sense annotations are normalized to probability distributions, denoted $G$ and $L$, and the divergence is computed as:

$$JSD(G||L) = \frac{1}{2}D_{KL}(G||M) + \frac{1}{2}D_{KL}(L||M)$$

where $M$ is the average of the distributions $G$ and $L$ and $D_{KL}$ denotes the Kullback-Leibler divergence. While both approaches are similar in intent, we find that the cosine similarity better matches the expected difference in Perception for cases where two annotations use different numbers of senses. For example, the fourth test annotation in Fig. 1 has a $JSS$[1] of 0.593, despite its significant differences in ordering and the omission of a sense. Indeed, in cases where the set of senses in a test annotation is completely disjoint from the set of gold standard senses, the $JSS$ will be positive due to comparing the two distributions against their average; In contrast, the cosine similarity in such cases will be zero, which we argue better matches the expectation that such an annotation does not meet the Perception objective.

## 3 WSI Models

For evaluation we adapt three recent graph-based WSI methods for the task of graded-sense annotation: Navigli and Crisafulli (2010), referred to as *Squares*, Jurgens (2011), referred to as *Link*, and UoY (Korkontzelos and Manandhar, 2010). At an abstract level, these methods operate in two stages. First, a graph is built, using either words or word pairs as vertices, and edges are added denoting some form of association between the vertices. Second, senses are derived by clustering or partitioning the graph. We selected these methods based on their superior performance on recent benchmarks and also

[1] The $JSD$ is a distance measure in $[0, 1]$, which we convert to a similarity $JSS = 1 - JSD$ for easier comparison.

for their significant differences in approach. Following, we briefly summarize each method to highlight its key parameters and then describe its adaptation to GWS annotation.

**Squares** Navigli and Crisafulli (2010) propose a method that builds a separate graph for each term for sense induction. First, a large corpus is used to identify associated terms using the Dice coefficient: For two terms $w_1$, $w_2$, $Dice(w_1, w_2) = \frac{2c(w_1,w_2)}{c(w_1)+c(w_2)}$ where $c(w)$ is the frequency of occurrence. Next, for a given term $w$ the initial graph, $G$, is constructed by adding edges to every term $w_2$ where $Dice(w, w_2) \geq \delta$, and then the step is repeated for the neighbors of each term $w_2$ that was added.

Once the initial graph is constructed, edges are pruned to separate the graph into components. Navigli and Crisafulli (2010) found improved performance on their target application using a pruning method based on the number of squares (closed paths of length 4) in which an edge participates. Let $s$ denote the number of squares that an edge $e$ participates in and $p$ denote the number of squares that would be possible from the set of neighbors of $e$. Edges with $\frac{s}{p} < \sigma$ are removed. The remaining connected components in $G$ denote the senses of $w$.

Sense disambiguation on a context of $w$ is performed by computing the intersection of the context's terms with the terms in each of the connected components. As originally specified, the component with the largest overlap is labeled as the sense of $w$. We adapt this to graded senses by returning all intersecting components with applicability proportional to their overlap. Furthermore, for efficiency, we use only noun, verb, and adjective lemmas in the graphs.

**Link** Jurgens (2011) use an all-words method where a single graph is built in order to derive the senses of all words in it. Here, the graph's clusters do not correspond to a specific word's senses but rather to contextual features that can be used to disambiguate any of the words in the cluster.

In its original specification, the graph is built with edges between co-occurring words and edge weights corresponding to co-occurrence frequency. Edges below a specified threshold $\tau$ are removed, and then link community detection (Ahn et al., 2010) is applied to discover sense-disambiguating word communities, which are overlapping cluster of vertices in the graph, rather than hard partitions. Once the set of communities is produced, communities with three or fewer vertices are removed, under the assumption that these communities contain too few features to reliably disambiguate.

Senses are disambiguated by finding the community with the largest overlap score, computed as the weighted Jaccard Index. For a context with the set of features $F_i$ and a community with features $F_j$, the overlap is measured as $|F_j| \cdot \frac{|F_i \cap F_j|}{|F_i \cup F_j|}$.

We adapt this algorithm in three ways. First, rather than use co-occurrence frequency to weight edges between terms, we weight edges accord to their statistical association with the G-test (Dunning, 1993). The G-test weighting helps remove edges whose large edge weights are due to high corpus frequency but provide no disambiguating information, and the weighting also allows the $\tau$ parameter to be more consistently set across corpora of different sizes. Second, while Jurgens (2011) used only nouns as vertices in the graph, we include both verbs and adjectives due to needing to identify senses for both. Third, for graded senses, we disambiguate a context by reporting all overlapping communities, weighted by their overlap score.

**UoY** Korkontzelos and Manandhar (2010) propose a WSI model that builds a graph for each term for disambiguation. The graph is built in four stages, with four main tuning parameters, summarized next. First, using a reference corpus, all contexts of the target word $w$ are selected to build a list of co-occurring noun lemmas, retaining all those with frequency above $P_1$. Second, the Log-Likelihood ratio (Dunning, 1993) is computed between all selected nouns and $w$, retaining only those with an association above $P_2$. Third, all remaining nouns are used to create all $\binom{n}{2}$ noun pairs. Next, each term and pair is mapped to the set of contexts in the reference corpus in which it is present. A pair $(w_i, w_j)$ is retained only if its set of contexts is dissimilar to the sets of contexts of both its member terms, using the Dice coefficient to measure the similarity of the sets. Pairs with a Dice coefficient above $P_4$ with either of its constituent terms are removed. Last, edges are added between nouns and noun pairs according to their conditional probabilities of occurring with each other. Edges with a conditional probability less than

$P_3$ are not included.

Once the graph has been constructed, the Chinese Whispers graph partitioning algorithm (Biemann, 2006) is used to identify word senses. Each graph partition is assigned a separate sense of $w$. Next, each partition is mapped to the set of contexts in the reference corpus in which at least one of its vertices occurs. Partitions whose context sets are a strict subset of another are merged with the subsuming partition.

Word sense disambiguation occurs by counting the number of overlapping vertices for each partition and selecting the partition with the highest overlap as the sense of $w$. We extend this to graded annotation by selecting all partitions with at least one vertex present and set the applicability equal to the degree of overlap.

## 4 Evaluation Across Sense Inventories

Directly comparing GWS annotations from the induced and gold standard sense inventories requires first creating a mapping from the induced senses to the gold standard inventory. Agirre et al. (2006) propose a sense-mapping procedure, which was used in the previous two SemEval WSI Tasks (Agirre and Soroa, 2007; Manandhar et al., 2010). We consider this procedure and two extensions of it to support learning a mapping from graded sense annotations.

The procedure of Agirre et al. (2006) uses three corpora: (1) a base corpus from which the senses are derived, (2) a mapping corpus annotated with both gold standard senses, denoted $gs$, and induced senses, denoted $is$, and (3) a test corpus annotated with $is$ senses that will be converted to $gs$ senses.

Once the senses are induced from the base corpus, the mapping corpus is annotated with $is$ senses and a matrix $M$ is built where cell $i, j$ initially contains the counts of each time $gs_j$ and $is_i$ were used to label the same instance. The rows of this matrix are then normalized such that each cell now represents $p(gs_j|is_i)$. The final mapping selects the most probable $gs$ sense for each $is$ sense.

To label the test corpus, each instance that is labeled with $is_i$ is relabeled with the $gs$ sense with the highest conditional probability given $is_i$. When a context $c$ is annotated by a set of labels $L = \{is_i, \ldots, is_j\}$, the final sense labeling contains the set of all $gs$ to which the $is$ senses were mapped, weighted by their mapping frequencies: $per_c(gs_j) = \frac{1}{|L|} \sum_{is_i \in L} \delta(is_i, gs_j)$ where $\delta$ returns 1 if $is_i$ is mapped to $gs_j$ and 0 otherwise.

The original algorithm of Agirre et al. (2006) does not consider the role of applicability in evaluating whether an $is$ sense should be mapped to a $gs$ sense; $is$ senses with different levels of applicability in the same context are treated equivalently in updating $M$. Therefore, as a first extension, referred to as $Graded$, we revise the update rule for constructing $M$ where for the set of contexts $C$ labeled by both $is_i$ and $gs_j$, $M_{i,j} = \sum_{c \in C} per_c(is_i) \times per_c(gs_j)$. As in (Agirre et al., 2006), $M$ is normalized and each $is$ sense is mapped to its most probable $gs$ sense.

To label the test corpus using the $Graded$ method, the applicability of the $is$ sense is also included. For a context $c$ is annotated with senses $L = \{is_i, \ldots, is_j\}$, the final sense labeling contains the set of all $gs$ senses to which the $is$ senses were mapped, weighted by their mapping frequencies: $per_c(gs_j) = \sum_{is_i \in L} [\delta(is_i, gs_j) \times per_c(is_i)]$. The applicabilities are then normalized to sum to 1.

The prior two methods restrict an $is$ sense to mapping to only a single $gs$ sense. However, an $is$ sense may potentially correspond to multiple $gs$ senses, each with different levels of applicability. Therefore, we consider a second extension, referred to as $Distribution$, that uses the same matrix construction as the Graded procedure, but rather than mapping each $is$ to a single sense, maps it to a distribution over all $gs$ senses for which it was co-annotated, which is the normalized row vector in $M$ for an $is$ sense. Labeling in the test corpus is then done by summing the distributions of the $is$ senses annotated in the context and normalizing to create a probability distribution over the union of their $gs$ senses.

## 5 Experiments

We adapt the supervised WSD setting used in prior SemEval WSI Tasks (Agirre and Soroa, 2007; Manandhar et al., 2010) to evaluation the models according to the three proposed objectives. In the supervised setting, WSI systems provide GWS annotation of their induced senses for the test corpus, which is already labeled with the gold-standard GWS annotations. Then, a portion of the test corpus with gold standard annotations is used to build a mapping from induced senses to the reference sense inven-

| Term | PoS | # senses | Avg. # Senses per Instance |
|---|---|---|---|
| add | verb | 6 | 4.18 |
| ask | verb | 7 | 5.98 |
| win | verb | 4 | 3.98 |
| argument | noun | 7 | 5.18 |
| interest | noun | 7 | 5.12 |
| paper | noun | 7 | 5.54 |
| different | adj. | 5 | 4.98 |
| important | adj. | 5 | 4.82 |

Table 1: The terms from the GWS dataset (Erk et al., 2009) used in this evaluation

tory using one of the three algorithms described in Section 4. The remaining, held-out test corpus instances have their induced senses converted to the gold standard sense inventory and the sense labelings are evaluated for the three objectives from Section 2. In our experiments we divide the reference corpus into five evenly-sized segments and then use four segments (80% of the test corpus) for constructing the mapping and then evaluate the converted GWS annotations of the remaining segment.

### 5.1 Graded Annotation Data

The gold standard GWS annotations are derived from a subset of the GWS data provided by Erk et al. (2009). Here, three annotators rated the applicability of all WordNet 3.0 senses of a word in a single sentence context. Ratings were done using a 5-point ordinal ranking according to the judgements from 1 – this sense is not applicable to 5 – this usage exactly reflects this sense. Annotators used a wide-range of responses, leading to many applicable senses per instance. We selected the subset of the GWS dataset where each term has 50 annotated contexts, which were distributed evenly between SemCor (Miller et al., 1993) and the SENSEVAL-3 lexical substitution corpus (Mihalcea et al., 2004). Table 1 summarizes the target terms in this context.

To prepare the data for evaluation, we constructed the gold standard GWS annotations using the mean applicability ratings of all three annotators for each context. Senses that received a mean rating of 1 (not applicable) were not listed in gold standard labeling for that instance. All remaining responses were normalized to sum to 1.

### 5.2 Model Configuration

For consistency, all three WSI models were trained using the same reference corpus. We used a 2009 snapshot of Wikipedia,[2] which was PoS tagged and lemmatized using the TreeTagger (Schmid, 1994). All of target terms occurred over 12,000 times. The G-test between terms was computed using a three-sentence sliding window within each article in the corpus. The Dice coefficient was calculated using a single sentence as context.

For all three models, we performed a limited grid search to find the best performing system parameters, within reasonable computational limits. We summarize the parameters and models, selecting the configuration with the highest average Perception score. For all models, the applicability ratings for each instance are normalized to sum to 1.

| **Model** | **Parameter Range** | **Selected** |
|---|---|---|
| Squares | $\delta$={0.008, 0.009, . . . , 0.092} | 0.037 |
| | $\sigma$={0.25, 0.30, . . . , 0.50, 0.55} | 0.55 |
| Link | $\tau=\{400, 500, \ldots, 900, 1000\}$ | 500 |
| UoY | $P_1=\{10, 20\}$ | 20 |
| | $P_2=\{10, 20, 30\}$ | 20 |
| | $P_3=\{0.2, 0.3, 0.4\}$ | 0.3 |
| | $P_4=\{0.4, 0.6, 0.8\}$ | 0.4 |

### 5.3 Baselines

Prior WSI evaluations have used the Most Frequent Sense (MFS) labeling a strong baseline in the supervised WSD task. For the GWS setting, we consider five other baselines that select one, some, or all of the sense of the target word, with different ordering strategies. In the six baselines, each instance is labeled as follows:

**MFS:** the most frequent sense of the word
**RS:** a single, randomly-selected sense
**ASF:** all senses, ranked in order of frequency starting with the most frequent
**ASR:** all senses, randomly ranked
**ASE:** all senses, ranked equally
**RSM:** a random number of senses, ranked arbitrarily

To establish applicability values from a ranking of $n$ senses, we set applicability to the $i^{th}$ ranked sense of $\frac{(n-i)+1}{\sum_{k=1}^{n} k}$, where rank 1 is the highest ranked sense.

[2] http://wacky.sslmit.unibo.it/

|  | Agirre et al. (2006) Mapping |  |  | Graded Mapping |  |  | Distribution Mapping |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|
| **Model** | **D** | **R** | **P** | **D** | **R** | **P** | **D** | **R** | **P** | **Recall** |
| Squares | 0.192 | -0.024 | 0.382 | 0.198 | 0.555 | 0.504 | **0.879** | **0.562** | **0.925** | 0.560 |
| Link | 0.282 | 0.081 | 0.454 | 0.335 | 0.436 | 0.528 | 0.854 | 0.503 | 0.907 | 0.800 |
| UoY | 0.238 | 0.116 | 0.445 | 0.244 | 0.486 | 0.528 | 0.848 | 0.528 | 0.907 | **0.940** |

Table 2: Average performance of the three WSI models according to **D**etection, **R**anking, and **P**ercetion

| **Baseline** | **Detection** | **Ranking** | **Perception** |
|---|---|---|---|
| MFS | 0.204 | **0.334** | 0.469 |
| RS | 0.167 | -0.036 | 0.363 |
| ASF | **0.846** | 0.218 | 0.830 |
| ASR | **0.846** | 0.006 | 0.776 |
| ASE | **0.846** | 0.000 | **0.862** |
| RSM | 0.546 | 0.005 | 0.632 |

Table 3: Average performance of the six baselines

### 5.4 Results and Discussion

Each WSI model was trained and then used to label the sense of each target term in the GWS corpus. The three sense-mapping procedures were then applied to the induced sense labels on the held-out instances to perform a comparison in the graded sense annotations. Table 2 reports the performance for the three evaluation measures for each model and mapping configuration on all instances where the sense mapping is defined. The sense mapping is undefined when (1) a WSI model cannot match an instance's features to any of its senses therefore leaves the instance unannotated or (2) when an instance is labeled with an $is$ sense not seen in the training data. Therefore, we report the additional statistic, Recall, that indicates the percentage of instances that were both labeled by the WSI model and mapped to $gs$ senses. Table 3 summarizes the baselines' performance.

The results show three main trends. First, introducing applicability into the sense mapping process noticeably improves performance. For almost all models and scores, using the Graded Mapping improves performance a small amount. However, the largest increase comes from using the Distribution mapping where induced senses are represented as distributions over the gold standard senses.

Second, performance was well ahead of the baselines across the three evaluations, when considering the models' best performances. The Squares and Link models were able to outperform the baselines that list all senses on the Detection objective, which the UoY model only improves slightly from this baseline. For the Ranking objective, all models substantially outperform the best baseline, MFS; and similarly, for the Perception objective, all models outperform the best performing baseline, ASE. Overall, these performance suggest that induce senses can be successfully used to produce quality GWS annotations.

Third, the WSI models themselves show significant differences in their recall and multi-labeling frequencies. The Squares model is only able to label approximately 56% of the GWS instances due to sparseness in its sense representation. Indeed, only 12 of its 237 annotated instances received more than one sense label, revealing that the model's performance is mostly based on correctly identifying the primary sense in a context and not on identifying the less applicable senses. The UoY model shows a similar trend, with most instances being assigned a median of 2 senses. However, its sense representation is sufficiently dense to have the highest recall of any of the models. In contrast to the other two models, the Link model varies significantly in the number of induced senses assigned: "argument," "ask," "different," and "win" were assigned over 60 senses on average to each of their instances, with "different" having an average of 238, while the remaining terms were assigned under two senses on average.

Furthermore, the results also revealed two unexpected findings. First, the ASE baseline performed unexpectedly high in Perception, despite its assignment of uniform applicability to all senses. We hypothesize this is due to the majority of instances in the GWS dataset being labeled with most of a word's senses, as indicated by Table 1, which results in their perceptibilities becoming normalized to small values. Because the ASE solution has applicability ratings for all senses, normalization brings the ratings close to those of the gold standard solution, and furthermore, the difference in score between applicable and inapplicable senses become too small to significantly affect the resulting cosine similarity. As an alternate model, we reevaluated the baselines against the gold standard using the Jensen-Shannon divergence as proposed by Erk and McCarthy (2009). Again, ASE is still the highest performing baseline on Perception. The high performance for both evaluation measures suggests that an alternate measure may be better suited for quantifying the difference in solutions' GWS applicabilities.

Second, performance was higher on the Perception task than on Ranking, the former of which was anticipated being more difficult. We attribute the lower Ranking performance to two factors. First, the GWS data contains main tied rank senses; however, ties in sense ranks after the mapping process are relatively rare, which reduces $\gamma$. Second, instances in the GWS often have senses within close applicability ranges. When scoring an induced annotation that swaps the applicability, the Perception is less affected by the small change in applicability magnitude, whereas Ranking is more affected due to the change in ordering.

## 6 Conclusion and Future Work

GWS annotations offer great potential for reliably annotating using fine-grained sense inventories, where word instance may elicit several concurrent meanings. Given the expense of creating annotated training corpora with sufficient examples of the graded senses, WSI offers significant promise for learning senses automatically while needing only a small amount GWS annotated data to learn the sense mapping for a WSD task.

In this paper, we have carried out an initial study on the performance of WSI systems on a GWS annotation task. Our primary contribution is an end-to-end framework for mapping and evaluating induced GWS data. We first proposed three objectives for graded sense annotation along with corresponding evaluation measures that reliably convey the effectiveness given the nature of GWS annotations. Second, we proposed two new mapping procedures that use graded sense applicability for converting induced senses into a reference sense inventory. Using three graph-based WSI models, we demonstrated that incorporating graded sense applicability into the sense mapping significantly improves GWS performance over the commonly used method of Agirre et al. (2006). Furthermore, our study demonstrated the potential of WSI systems, showing that all the models were able to outperform all six of the proposed baseline on the Ranking and Perception objectives.

Our findings raise several avenues for future work. First, our study only considered three graph-based WSI models; future work is needed to assess the capabilities other WSI approaches, such as vector-based or Bayesian. We are also interested in comparing the performance of the Link model with other recently developed all-words WSI approaches such as Van de Cruys and Apidianaki (2011).

Second, the proposed evaluation relies on a supervised mapping to the gold standard sense inventory, which has potential to lose information and incorrectly map new senses not in the gold standard. While unsupervised clustering evaluations such as the V-measure (Rosenberg and Hirschberg, 2007) and paired Fscore (Artiles et al., 2009) are capable of evaluating without such a mapping, future work is needed to test extrinsic soft clustering evaluations such as BCubed (Amigó et al., 2009) or develop analogous techniques that take into account graded class membership used in GWS annotations.

Last, we note that our setup normalized the GWS ratings into probability distribution, which is standard in the SemEval evaluation setup. However, this normalization incorrectly transforms GWS annotations where no predominant sense was rated at the highest value, e.g., an annotation of only two senses rated as 3 on a scale of 1 to 5. While these perceptibilities may be left unnormalized, it is not clear how to compare the induced GWS annotations with such mid-interval values, or when the rating scale of the WSI system is potentially unbounded. Future work is needed both in GWS evaluation and in quantifying applicability along a range in GWS-based WSI systems to address this issue.

All models and data will be released as a part of the S-Space Package (Jurgens and Stevens, 2010).[3]

[3] https://github.com/fozziethebeat/S-Space

## References

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

Eneko Agirre, David Martínez, Oier ó de Lacalle, and Aitor Soroa. 2006. Evaluating and optimizing the parameters of an unsupervised graph-based WSD algorithm. In *Proceedings of TextGraphs: the First Workshop on Graph Based Methods for Natural Language Processing*, pages 89–96. Association for Computational Linguistics.

Yong-Yeol Ahn, James P. Bagrow, and Sune Lehmann. 2010. Link communities reveal multiscale complexity in networks. *Nature*, (466):761–764, August.

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.

Chris Biemann. 2006. Chinese whispers: an efficient graph clustering algorithm and its application to natural language processing problems. In *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing*, pages 73–80. Association for Computational Linguistics.

Irina Chugur, Julio Gonzalo, and Felisa Verdejo. 2002. Polysemy and sense proximity in the senseval-2 test suite. In *Proceedings of the ACL-02 workshop on Word sense disambiguation: recent successes and future directions - Volume 8*, WSD '02, pages 32–39, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19(1):61–74.

Katrin Erk and Diana McCarthy. 2009. Graded word sense assignment. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1*, pages 440–449. Association for Computational Linguistics.

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: Volume 1*, pages 10–18. Association for Computational Linguistics.

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

David Jurgens and Keith Stevens. 2010. The S-Space Package: An Open Source Package for Word Space Models. In *Proceedings of the ACL 2010 System Demonstrations*. Association for Computational Linguistics.

David Jurgens. 2011. Word sense induction by community detection. In *Proceedings of Sixth ACL Workshop on Graph-based Methods for Natural Language Processing (TextGraphs-6)*. Association for Computational Linguistics.

Ioannis Korkontzelos and Suresh Manandhar. 2010. Uoy: Graphs of unambiguous vertices for word sense induction and disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 355–358. Association for Computational Linguistics.

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)*.

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. Association for Computational Linguistics.

Diana McCarthy. 2006. Relating WordNet senses for word sense disambiguation. *Making Sense of Sense: Bringing Psycholinguistics and Computational Linguistics Together*, page 17.

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. Barcelona, Spain, Association for Computational Linguistics.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A semantic concordance. In *Proceedings of the workshop on Human Language Technology*, pages 303–308. Association for Computational Linguistics.

Roberto Navigli and Giuseppe Crisafulli. 2010. Inducing word senses to improve web search result clustering. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 116–126. Association for Computational Linguistics.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, 41(2):10.

Martha Palmer, Hoa Trang Dang, and Christiane Fellbaum. 2007. Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Natural Language Engineering*, 13(02):137–163.

Rebecca J. Passonneau, Ansaf Salleb-Aoussi, Vikas Bhardwaj, and Nancy Ide. 2010. Word sense annotation of polysemous words by multiple annotators. In *Proceedings of Seventh International Conference on Language Resources and Evaluation (LREC-7)*.

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, June.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*.

Sidney Siegel and N. John Castellan Jr. 1988. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, second edition.

Tim Van de Cruys and Marianna Apidianaki. 2011. Latent Semantic Word Sense Induction and Disambiguation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL/HLT)*, pages 1476–1485.