Measuring the Impact of Sense Similarity on Word Sense Induction

David Jurgens[1,2]
[1]HRL Laboratories, LLC
Malibu, California, USA
jurgens@cs.ucla.edu

Keith Stevens[2]
[2]University of California, Los Angeles
Los Angeles, California, USA
kstevens@cs.ucla.edu

Abstract

Word Sense Induction (WSI) is an unsupervised learning approach to discovering the different senses of a word from its contextual uses. A core challenge to WSI approaches is distinguishing between related and possibly similar senses of a word. Current WSI evaluation techniques have yet to analyze the specific impact of similarity on accuracy. Therefore, we present a new WSI evaluation that quantifies the relationship between the relatedness of a word's senses and the ability of a WSI algorithm to distinguish between them. Furthermore, we perform an analysis on sense confusions in SemEval-2 WSI task according to sense similarity. Both analyses for a representative selection of clustering-based WSI approaches reveals that performance is most sensitive to the clustering algorithm and not the lexical features used.

1 Introduction

Many words in a language have several distinct meanings. For example, "earth" may refer to the planet Earth, dirt, or solid ground, depending on the context. The goal of Word Sense Induction (WSI) is to automatically discover the different senses by examining how a word is used. This unsupervised discovery process produces a sense inventory where the number of senses is corpus-driven and where senses may reflect additional usages not present in a predefined sense inventory, such as those for medicine or law (Dorow and Widdows, 2003). Furthermore, these discovered senses can be used to automatically expand lexical resources such as WordNet or FrameNet (Klapaftis and Manandhar, 2010).

Discovering the multiple senses is frequently confounded by the relationships between a word's senses. While homonyms such as "bass" or "bank" have unrelated senses, many polysemous words have interrelated senses, with lexicographers often in disagreement for the number of fine-grained senses (Palmer et al., 2007). For example, the most frequent four senses for "law" according to WordNet, shown in Table 1, are similar in several aspects and could be ascribed interchangeably in some contexts. The difficulty of automatically distinguishing two senses is proportional to their similarity because of the increasing likelihood of the two senses sharing similar contexts.

While the issue distinguishing between related senses is a recognized issue for Word Sense Disambiguation (Chugur et al., 2002; McCarthy, 2006), which uses supervised training to learn sense distinctions, measuring the impact of sense relatedness on the harder problem of WSI remains unaddressed. The recent SemEval WSI tasks (Agirre and Soroa, 2007; Manandhar and Klapaftis, 2009) have provided a standard framework for evaluating WSI systems, with a controlled training corpus designed to limit sense ambiguity in the example contexts. However, given the potential relatedness of a word's senses, we view it necessary to consider how WSI methods perform relative to the degree of contextual ambiguity. Our goal is therefore to quantify the similarity at which a WSI approach is unable to distinguish between two senses, which reflects the sense granularity at which the approach operates.

We propose two new evaluations. The first, described in Section 4, uses a similarity-based pseudoword discrimination task to measure the discrimination capability for related senses along a graded scale of similarity. As a second evaluation, in

Proceedings of EMNLP 2011, Conference on Empirical Methods in Natural Language Processing, pages 113–123,
Edinburgh, Scotland, UK, July 27–31, 2011.

1	the collection of rules imposed by authority
2	legal document setting forth rules governing a particular kind of activity
3	a rule or body of rules of conduct inherent in human nature and essential to or binding upon human society
4	a generalization that describes recurring facts or events in nature

Table 1: Definitions for the top four senses of "law" according to WordNet

Section 5 we perform an error analysis using the SemEval-2010 WSI task, examining sense confusion relative to the sense similarities. For both evaluations, we examine twenty different WSI clustering-based models through combining five feature types and four clustering algorithms. These models were selected to be representative of a wide class of existing algorithms as a way of influence future algorithmic directions based on the current model's performance.

2 Clustering Contexts to Discover Senses

Frequently, WSI is treated as an unsupervised clustering problem: The contexts in which a word appears are clustered in order to discover its senses (Navigli, 2009). We selected four diverse clustering algorithms for evaluation based on three criteria: (1) the ability to automatically determine the final number of clusters given an upper bound or a set of parameters, (2) an efficient run time, and (3) high quality results in either WSI or other fields related to text analysis. The first criteria is essential for WSI; the final number of senses must be derived without supervision in order to reflect the true number of senses present in the corpus.

K-Means K-Means builds clusters based on the similarity between two data points. Clusters grow by assigning data points to the cluster with the most similar centroid. After every data point is assigned, each cluster's centroid is recalculated to be the average of all the data points assigned to the cluster. This process repeats until the centroids converge to a fixed point. We choose initial seeds at random and use the H2 criterion function (Zhao and Karypis, 2001). Although K-Means is efficient and widely used, it requires the number of clusters to be specified a priori. Therefore, we follow the WSI model of Pedersen and Kulkarni (2006) and use the Gap Statistic (Tibshirani et al., 2000) to automatically determine the number of clusters.

The Gap Statistic runs K-Means repeatedly with different values of K, ranging from 1 to some sensible maximum. The Gap Statistic first induces a data model from the feature distributions of the initial dataset and then for each K, creates a set of artificial datasets by sampling from the derived model. K is increased until the "gap", i.e. the distance between the objective function of the original dataset and the average objective function of the artificial datasets, is larger then the gap for the previous K value. We calculate the gap using 10 artificial data sets sampled from the model.

Spectral Clustering Spectral Clustering interprets a dataset's elements as vertices in graph with edges based on their similarity (Ng et al., 2001). Clusters are found by identifying the graph partition that produces the minimum conductance between every partition. This can be thought of as trying to find small islands that are connected by as few bridges as possible. We refer the reader to (von Luxburg, 2007) for further technical details. To our knowledge, only He et al. (2010) have applied spectral clustering to WSI, which was performed on a Chinese dataset. However, the algorithm used by He et al. requires the number of clusters to be specified.

We instead use a hybrid spectral clustering algorithm, first applied to information retrieval (Cheng et al., 2006), that automatically selects the number of clusters. This algorithm recursively partitions a dataset in half by finding the cut that produces the minimum conductance, which builds a tree of partitions. This split is done until either every data point is in its own partition or a maximum number of partitions is found. Partitions are then dynamically merged, starting at leaf partitions, based on a clustering criteria. We use the relaxed correlation criteria (Cheng et al., 2006), which tries to maximize both inter cluster similarity and intra cluster dissimilarity. The final cluttering generated is then the best tree-respecting partition of the original data set.

Clustering By Committee Pantel and Lin (2002) found that K-Means clustering folded all features found in a cluster into the centroid, many of which are not useful for identifying the desired word sense.

To overcome this, they proposed a novel clustering algorithm for WSI, Clustering by Committee (CBC), which includes only the most distinguishing features for a cluster into the centroid.

For each context, an initial set of "committees" is formed by clustering the most similar contexts to each context, with the resulting committees ranked to prefer larger, highly similar clusters. The final set of committees (sense clusters) are selected by recursively identifying the highest ranking committees that are dissimilar to each other and then repeating the process for any contexts not similar to existing committees. In essence, CBC aims to find the clusters that are similar to the largest set of contexts, while keeping clusters dissimilar from each other. CBC's recursion ensures that contexts dissimilar to the large committees are still grouped into their own smaller committees, which enables the discovery of infrequent senses with distinct contexts. We use a hard sense assignment for each context, i.e., a context is labeled with only one sense according to the most similar cluster.

Streaming K-Means As WSI moves into inducing senses from Web-scale amounts of data, existing clustering algorithms that keep all contexts in memory become impractical. Jurgens and Stevens (2010a) proposed an on-line hybrid clustering solution using on-line K-Means and Hierarchical Agglomerative Clustering, which automatically decided the number of clusters without retaining all the contexts. To the best of our knowledge, theirs is the only work using an on-line approach. We extend this work by applying a more theoretically sound online K-Means algorithm, called Streaming K-Means (Braverman et al., 2011), to WSI. We use Streaming K-Means to conduct a direct algorithmic comparison with K-Means in the hopes that online approaches can be made just as effective as off-line approaches.

Streaming K-Means processes each data point only once, thus reducing the memory overhead dramatically. Instead of recording each data point, it immediately assigns each data point to a cluster and maintains $K \cdot C$ clusters. C varies as the algorithm runs, initially being set to 0. When assigning a data point, it is only assigned to an existing cluster when their similar is above some threshold, otherwise the data point becomes the centroid of a new cluster. Once C reaches a threshold, based on an estimate of the number of data points, or the overall K-Means clustering cost reaches some limit, the centroids are treated as new data points and re-clustered, with the goal of merging some centroids. We follow (Jurgens and Stevens, 2010a) and cluster the final centroids with Hierarchical Agglomerative Clustering, with the average link criteria as suggested by (Pedersen and Bruce, 1997).

3 Modeling Context

For each clustering algorithm, we consider five context models that represent the types of lexical features used by the majority of WSI approaches.

Co-Occurrence Contexts formed from word co-occurrence are the most common in WSI algorithms. For each occurrence of a word, those words within a certain range are counted as features. Prior work has used a variety of context sizes, e.g. words in the same sentence (Bordag, 2006), in nearby lexical positions (Gauch and Futrelle, 1993), or within a paragraph-sized context window (Pedersen, 2010).

We consider two co-occurrence context models: a 5-word and a 25-word window. We note that in co-occurrence-based word space algorithms, smaller context sizes have shown to better capture paradagmatic similarity, while larger sizes capture semantic associativity (Peirsman et al., 2008; Utsumi, 2010).

Dependency-Relations Dependency parsing creates a syntax tree where words are directly linked according to their relation. These links refine co-occurrence based contexts by utilizing syntactic indications of how words are related. Dependency parsed features have proven highly effective for word representations in many NLP applications, e.g., (Padó and Lapata, 2007; Baroni et al., 2010). We follow Pantel and Lin (2002) and Dorow and Widdows (2003) using the sentence as contexts and all words with a dependency path of length 3 or less, with the last word and its relation as a feature. We note that recently Kern et al. (2010) achieved good WSI performance with only a small, manually-tuned subset of all relations as context.

Word Ordering Word ordering can provide a mild form of syntactic information (Jones et al., 2006; Sahlgren et al., 2008). While other syntac-

tic features may provide significantly more information, word ordering is efficient to compute and provides an alternative source of syntactic information for knowledge-lean systems or for languages where NLP tools are not readily available.

Because we treat word ordering as a syntactic feature, we limit the context to words occurring in the same sentence. A feature is the combination of a co-occurring word and its relative position, i.e. the same word in different positions is treated as two separate features.

Parts of Speech Part of speech tagging can provide a preliminary coarse-grained sense disambiguation of a word's contextual features, where a word may have as many senses as it does parts of speech. For example, consider an occurrence of "house" in the context of "address" as a noun and verb: "I went to his house address," and "I heard the legislator address the house." Labeling "address" with its part of speech provides for more semantic information on its meaning, which further constrains the sense of "house." Prior work (Pedersen and Bruce, 1997) has suggested that this information can improve performance, but to our knowledge, the impact of POS features has not been evaluated in isolation.

Each context is formed from the containing sentence; a feature is a combination of each word and its part of speech, e.g., "board-NOUN" is distinct from "board-VERB."

4 WSI Performance on Related Senses

The proposed methodology measures the ability of a WSI approach to distinguish between related senses. However, generating a large corpus with manually labeled sense assignments and sense similarity judgements is prohibitively expensive. Therefore, we employ a pseudo-word discrimination task where a base word and a second word, its *confounder*, are replaced throughout the corpus with a pseudo-word. The objective is then to determine which of the words was originally present given the context of an occurrence of the pseudo-word. Due to not requiring manual annotation, this type of task was initially proposed as a substitute for word sense disambiguation (Schütze, 1992; Gale et al., 1992) and for selectional preferences (Clark and Weir, 2002).

Following the suggestions of Chambers and Jurafsky (2010) on designing pseudo-words, pseudo-words were created from words with the same part of speech and equal frequency in the training corpus. We selected nouns occurring more than 5,000 times in a 2009 Wikipedia snapshot and then drew 5,000 contexts for each. The snapshot was tagged with the Stanford Part of Speech Tagger (Toutanova et al., 2003) and parsed with the Malt Parser (Nivre et al., 2006).

festival		**laws**	
offices	0.13660	interests	0.18289
play	0.13751	politics	0.20440
convention	0.20296	governments	0.29125
tournament	0.29007	regulations	0.40761
concerts	0.48348	legislation	0.56112

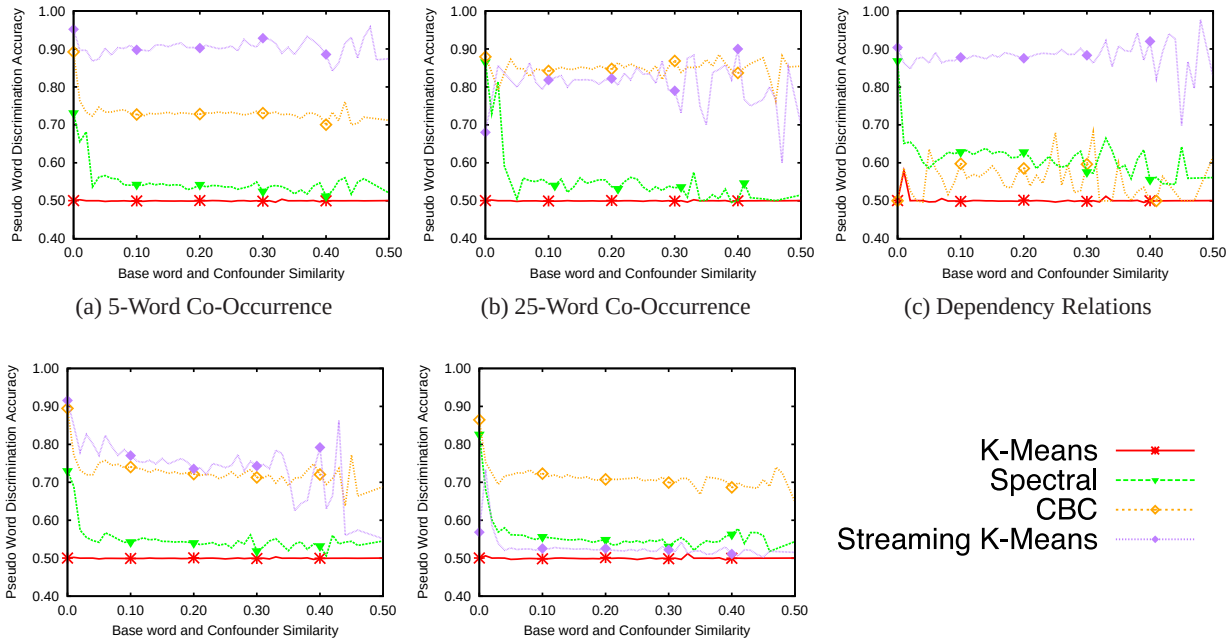
Table 2: Example confounders for "festival" and "laws" and their similarities

To evaluate the impact of sense similarity, pseudo-words were created from word pairs with a broad range of lexical similarities. We selected lexical similarity as an approximation of sense similarity in order to model the hypothesis that similar senses may appear in similar contexts. Similarity scores were calculated using cosine similarity on contextual distributions built from a sliding ± 2 word window over the Wikipedia corpus. Table 2 highlights several example confounders and their similarities with the base term. In total, we generated 5000 term-confounder pairs from 98 base terms, with a mean of 51 confounders per term.

All clustering parameters were chosen using the default values provided in the original papers. K-means and Streaming K-Means were both set with a maximum of 15 clusters, with the final number of clusters being determined by the data itself.

4.1 Evaluation

The pseudo-word's senses are induced from a training segment using each feature and clustering combination. Given that both words making up the pseudo-word may be polysemous, more than two senses may be induced. Each sense cluster is labeled according to which of the original words was present in the majority of its contexts. For testing, each instance of the pseudo-word in a previously unseen context is assigned the label of the cluster



(a) 5-Word Co-Occurrence (b) 25-Word Co-Occurrence (c) Dependency Relations

(d) Word Order (e) Parts of Speech

Figure 1: Pseudo-word discrimination performance

to which it is most similar. We perform five-fold cross-validation, using 4,000 contexts for training and 1,000 contexts for testing. Discrimination accuracy is reported as the average of all five runs. Since an equal number of contexts are used for each term, the base line accuracy of a most frequent sense model is 50% for each pseudo-word.

4.2 Results and Discussion

Figure 1 shows the discrimination accuracy relative to the similarity of a base pair and confounder, for each feature and clustering algorithm combination. Similarity values were binned at the 0.01 level with a mean of 39.0 scores per bin (median=11). Because most word pairs are not related, the distribution of similarity values is biased towards lower values. Therefore, we omit similarity ranges above 0.5, as too few confounders occurred in that range to draw reliable conclusions. The standard error (not shown) is < 1 for all measurements.

The general trends suggests that the clustering algorithm impacts the sense discriminatory ability far more than the lexical feature choice. Furthermore, sense similarity affects most clustering algorithms, with most systems seeing a noticeable performance drop when pseudo-word similarity is increased just beyond 0. Performance at high similarity becomes more variable for all algorithms and features.

For each clustering algorithm, we see dramatically different trends. Streaming K-Means performs well with co-occurrence based features and it does poorly when either contexts have too many features, as in the 25 Window Co-Occurrence feature space, or the feature space overall is too sparse, as in the Parts of Speech and Ordering feature spaces.

K-Means with the gap statistic converges to the most frequent sense baseline for nearly every confounder pair. We note that this behavior significantly differs from that seen in (Pedersen and Kulkarni, 2006), which clustered second-order co-occurrence vectors rather than the first-order features that we use. Our analysis showed that the H2 criterion was responsible for this behavior. A subsequent analysis revealed that K-Means still converged to MFS for the E1, E2, I1, and I2 criterion functions (Zhao and Karypis, 2001) as well as when the number of artificial datasets was increased up to 100. However, additional tests using the same features on the SemEval-1 WSI task did not converge to MFS. Further investigation is needed to identify the cause of convergence and what types of data are appropriate

the Gap Statistic.

Clustering by Committee performs well on most models, but significantly worse on dependency relation features. A subsequent analysis showed that CBC generates significantly more clusters than all other models. For the POS, 5 word window, and 25 window Co-Occurrence feature spaces, CBC generated between 205 and 247 clusters on average, per word. With the order feature space, CBC generated 1087 clusters per word. However, when paired with dependency relation features, the number of clusters drops to only 78 per word.

Spectral Clustering is most affected by sense similarity, performing competitively for unrelated senses but dropping significantly when words become even slightly similar. This performance drop is seen across all features. Performance is therefore low, with the exception of dependency relations.

Overall, these results suggest that sense relatedness is a important factor in WSI performance and its impact should be considered in future WSI evaluations. A potential next step is to vary the proportion of contexts from the confounder. The current method intentionally uses a uniform distribution to avoid potential bias; however, word sense distributions are rarely equal, and a varied distribution would more closely model real world distributions. Similarly, the current method tested only two senses, whereas an n-way disambiguation between multiple confounders should also provide further insight into a WSI approach's discriminatory abilities.

5 Sense Confusion in SemEval-2 Task 14

As a second experiment, we analyze incorrect sense assignments on SemEval-2 Task 14 (Manandhar et al., 2010) to measure whether sense-relatedness biases which sense was incorrectly selected. For WSI systems, a similarity bias would indicate that similar senses are more likely to be incorrectly identified as a single sense.

We summarize Task 14 as follows. Systems are provided with an unlabeled training corpus consisting of 879,807 multi-sentence contexts for 100 polysemous words, comprised of 50 nouns and 50 verbs. Systems induce sense representations for target words from the training corpus and then use those representations to label the senses of the target words in unseen contexts from a test corpus.

The induced senses are then evaluated against the gold standard labels OntoNotes (Hovy et al., 2006) senses labels for the test corpus. For our evaluation, we use both the two contrasting unsupervised measures, the paired FScore (Artiles et al., 2009) and the V-Measure (Rosenberg and Hirschberg, 2007), and a supervised measure. For each metric, we use the evaluation framework provided by the organizers of SemEval-2 Task 14.[1]

The V-Measure rates the homogeneity and completeness of a clustering solution. Solutions that have word clusters formed from one gold-standard sense are homogeneous; completeness measures the degree to which a gold-standard sense's instances are assigned to a single cluster. The paired FScore measures two types of overlap of a solution and the gold standard in cluster assignments for all in pairwise combination of instances. This score tends to penalize solutions with many small clusters and highly heterogeneous clusters (Manandhar and Klapaftis, 2009).

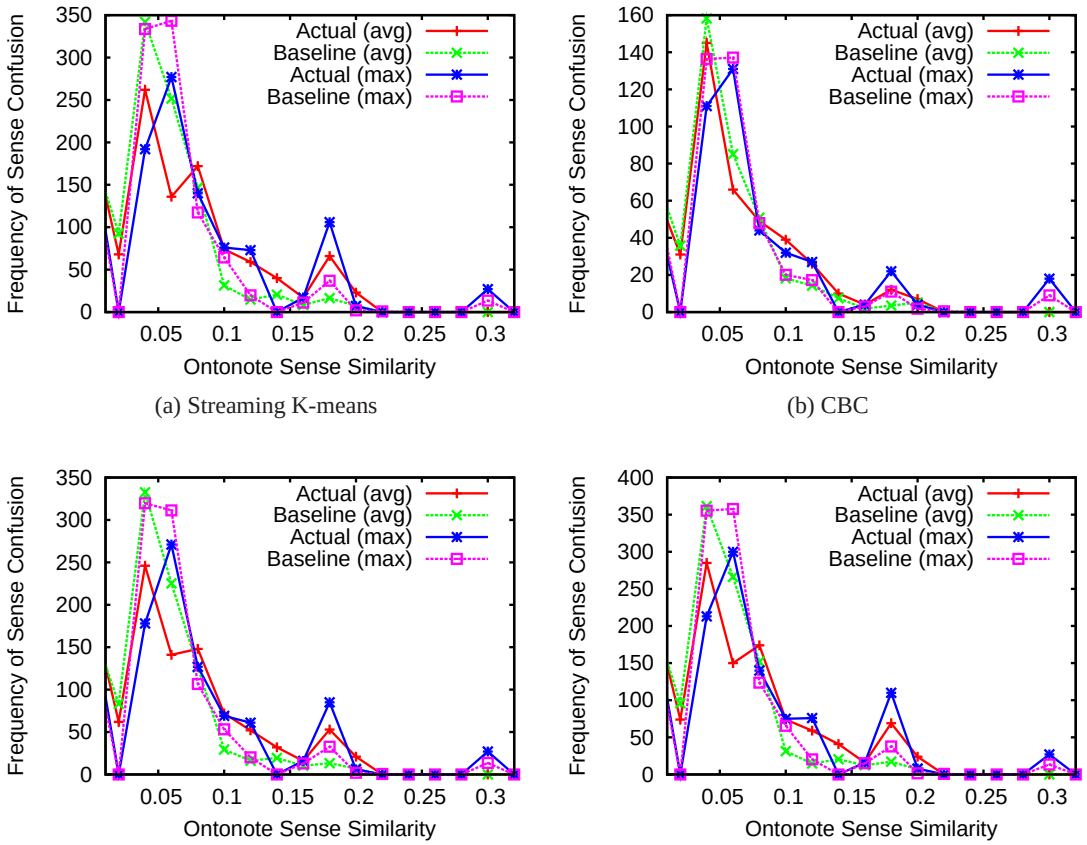
The supervised evaluation measures the recall when building a Word Sense Disambiguation classifier from the induced senses. The WSI system labels the entire corpus, which is then divided into training and test portions. The sense labels in the training portion are used to construct a mapping from induced senses to the gold standard OntoNotes labels. This mapping is then evaluated for the induced labels in the test. We report the scores for the 80% training and 20% testing scenario.

5.1 Evaluation

We expect that if sense similarity is a factor in sense confusion, the probability of confusion will increase with sense similarity. Therefore, we measure the probability of labeling an instance with the incorrect OntoNotes sense relative to the sense similarity with the gold standard sense.

In order to calculate the incorrect assignments, the induced senses must be mapped to OntoNotes senses. Each induced sense, s_i, is mapped to the OntoNotes sense that occurs most frequently among the instances in the test corpus that are assigned induced sense s_i. We note that this labeling process is only an approximate solution to assigning gold standard labels to induced senses. A more robust

[1] http://www.cs.york.ac.uk/semeval2010_WSI/



(a) Streaming K-means (b) CBC

(c) Spectral Clustering (d) K-means

Figure 2: The error frequency distributions for confusing the correct sense with another sense of the given similarity when using a 5-word co-occurrence window as context. Dashed lines indicate the null models.

labeling could take into account the distribution of gold standard senses labels in the corpus from which the senses are induced; however, such labels are not available in the Task 14 training corpus.

For each incorrect sense assignment, we measure the similarity of the confused sense to the correct sense. To our knowledge, no work has been done on calculating sense similarity within the OntoNotes sense hierarchy.[2] Therefore, we approximate OntoNotes sense similarity by using sense similarity in the WordNet ontology, on which has many similarity measures have been defined. Following Budanitsky and Hirst (2006), we estimate the WordNet sense similarity using the method proposed by Jiang and Conrath (1997).

Each OntoNotes sense s^i is mapped to a set of WordNet 3.0 senses $S^i = \{wn_1, \ldots, wn_n\}$ using the sense mapping provided by the CoNLL shared task.[3] The sense similarity for two OntoNotes senses is computed using one of two methods:

$$sim = \frac{1}{|S^1||S^2|} \sum_{wn^i \in S^1, wn^j \in S^2} JCN(wn^i, wn^j), \quad (1)$$

or

$$sim = \underset{wn^i \in S^1, wn^j \in S^2}{\text{argmax}} JCN(wn^i, wn^j), \quad (2)$$

where JCN indicates the Jiang-Conrath similarity of two WordNet senses, calculated using WordNet::Similarity (Pedersen et al., 2004). Eq. 1 computes similarity as the average similarity of all pairwise WordNet sense combinations, while Eq. 2 uses the highest similarity. The resulting OntoNote sense similarities range from 0 to 1, with 1 being maximally similar. We excluded 10 words from the test

[2] We suspect that this is in part because a word's OntoNotes senses have been designed to minimize sense confusion.

[3] http://conll.bbn.com/index.php/data.html

Context Feature	Clustering	V-Measure	F-Score	Recall	# Clusters	Purity	GoF p-Value
5-Word Co-Occurrence	Streaming	6.7	55.5	54.8	4.74	0.103	$p < 2.07e-37$
	Spectral	10.8	39.2	54.3	8.41	0.194	$p < 1.11e-25$
	CBC	23.9	8.2	39.5	39.7	0.665	$p < 0.916$
	K-Means	2.5	**61.8**	**55.6**	1.68	0.020	$p < 1.20e-37$
25-Word Co-Occurrence	Streaming	2.6	61.7	55.5	1.7	0.020	$p < 1.20e-37$
	Spectral	5.0	48.6	55.9	3.3	0.083	$p < 4.36e-32$
	CBC	21.3	11.6	45.0	32.2	0.561	$p < 0.011$
	K-Means	2.5	**61.8**	**55.6**	1.68	0.020	$p < 1.20e-37$
Dependency Relations	Streaming	3.0	61.5	**55.6**	1.9	0.022	$p < 7.33e-38$
	Spectral	8.5	46.8	55.3	5.9	0.134	$p < 5.45e-14$
	CBC	12.9	31.3	52.4	11.4	0.259	$p < 4.07e-12$
	K-Means	2.5	**61.8**	**55.6**	1.6	0.020	$p < 1.20e-37$
Word Order	Streaming	10.8	43.1	54.2	10.8	0.300	$p < 4.46e-24$
	Spectral	12.2	32.4	53.7	10.0	0.26	$p < 3.27e-20$
	CBC	**27.2**	11.8	30.3	54.9	**0.857**	$p < 0.999$
	K-Means	2.5	**61.8**	**55.6**	1.6	0.020	$p < 1.20e-37$
Parts of Speech	Streaming	6.6	53.0	54.5	4.7	0.117	$p < 1.06e-39$
	Spectral	10.9	39.4	53.7	8.3	0.201	$p < 2.38e-13$
	CBC	23.8	08.0	40.1	39.7	0.678	$p < 1.04e-2$
	K-Means	2.5	**61.8**	**55.6**	1.6	0.020	$p < 1.20e-37$
SemEval-2 Most Frequent Sense		0.0	63.4	58.6	1.0	0.0	$p < 4.244e-23$
Best SemEval-2 FScore		0.0	63.3	58.6	1.0	0.0	$p < 2.893e-23$
Best SemEval-2 VMeasure		16.2	26.7	58.3	10.7	0.367	$p < 1.956e-14$
Best SemEval-2 Supervised Recall		15.7	49.7	62.4	11.5	0.187	$p < 8.910e-19$

Table 3: Unsupervised and Supervised scores on the SemEval-2010 WSI Task for each feature and clustering models, with reference scores for the top performing systems for each evaluation shown below.

set that did not have mappings from OntoNotes to WordNet 3.0 senses, and additional 23 words that only had two senses, which prevented testing for a similarity bias. The remaining 67 words yielded 4,097 test instances for evaluation.

Each instance of the test corpus was tested for sense confusion, recording the similarity of the incorrectly assigned sense and the gold standard sense. The resulting incorrect assignments are transformed into an error distribution according by accumulating error counts into similarity bins where each bin has a range of 0.02. We analyze the WSI systems defined in section 4 as well as the results of three systems that participated in Task 14 and scored the highest on the paired FScore, V-measure, or Supervised Recall evaluations.

To quantify the impact, we compare each system's error distribution against a null model over the set of incorrect test instances missed by that system. In the null model, the incorrect sense for each instance is selected with uniform probability from the available senses. This behavior produces a distribution with no similarity bias. The cumulative error distribution for the null model is not uniform due to multiple sense pairings having the same similarity.[4] To quantify the difference between a system's error distribution and corresponding null model, we calculate the G-test as a measure of Goodness of Fit (GoF). The resulting p-values reflect the probability of observing the system's error distribution if there was no bias from sense-similarity.

5.2 Results and Discussion

We compare the error analysis against the evaluation measures of Task 14. Table 3 displays the eval-

[4]Verb senses often have a JCN similarity of 0 due to having no shared parent within the WordNet verb sense hierarchy, which results in high frequency distribution around 0.

uation measures. We also report the average number of clusters per word, the cluster purity, and the p-value when using Eq. 2 to measure sense similarity. Figure 2 visualizes the error distributions for the four clustering algorithms on 5-word co-occurrence features. The distributions in Figure 2 are representative of those of the other context models, which we omit due to space. Each plot reflects the frequency at which a sense with the specified similarity was confused for the correct sense.

The low p-values in Table 3 indicate a significant deviation from the null model. Examining the shape of the error distribution in Figure 2 reveals a noticeable skew towards higher similarity when an incorrect sense assignment is made. This distribution skew is also consistent for both similarity measures.

Comparing the Task 14 results in Table 3 to the sense confusion trends in in Figure 2 highlights an interesting pattern among the various models: as the number of induced sense clusters increases, the error distribution better approximates the null model. Specifically, the GoF for all models was well correlated with cluster purity (ρ=0.66), and the number of clusters (ρ=0.76). CBC generated the highest number of clusters and has a sense confusion distribution that closely matches the null model, indicating that it is less affected by sense similarity. In comparison, all of the Streaming K-Means models, which have the fewest clusters, differ noticeably from the null model. Spectral Clustering, which also generates fewer clusters than CBC, has an observed confusion rate that differs from the baseline. K-Means again reduces to the MFS baseline.

When comparing along the feature sets, we see that on average Word Order features generate the highest V-Measure scores, highest purity, and highest p-values for Streaming K-Means and CBC. This result correlates well with the average number of features seen per context: Word Order contexts used 0.03% of the feature space while contexts in other feature spaces used between 0.07% and 0.12% of the feature space, suggesting that the SemEval measures are determined in part by feature space density. Similarly, 25-word co-occurrence features had the highest percentage of features used per context, 0.12%, and generated the lowest V-Measure, purity score, and p-value for 3 clustering models.

These scores support another known trend in the SemEval-2 evaluation: the performance on the V-Measure is proportional to the number of induced sense clusters, while the paired FScore is inversely proportional. But what is surprising is that models which perform well against the V-Measure also exhibit a smaller sense similarity bias, suggesting that CBC and similar clustering methods are suitable for situations where competing senses of a word have a high degree of overlap.

As a final comparison, we also computed the sense bias for the top 3 SemEval systems under each measure. The best of these models are listed in Table 3. We did not find any consistent trends between the V-Measure, purity, and p-value among these models. The top F-Scoring models all used either a first or second order co-occurrence feature space similar to ours (Kern et al., 2010; Pedersen, 2010), whereas the top supervised score was achieved by a graph-based system (Klapaftis and Manandhar, 2008).

6 Future Work and Conclusion

We presented a two evaluation for WSI approaches and examined the performance of a wide range of algorithms. The results raise a potential issue for clustering-based WSI approaches: sense discrimination degrades notably as the sense relatedness increases. We highlight three potential avenues for future research. First, this methodology should be applied to additional WSI models, such as graph-based (Klapaftis and Manandhar, 2008; Navigli and Crisafulli, 2010) and probabilistic models (Dinu and Lapata, 2010; Elshamy et al., 2010). Second, we plan to extend the analysis to different sense distributions, varying number of senses, and for human annotated sense similarity data. Third, this evaluation makes the simplifying assumption of one sense per instance; however, Erk et al. (2009) note that the relations between senses may cause a single word instance to evoke multiple senses within the same context. Therefore, a future experiment should consider how WSI systems might address learning senses given the presence of multiple, similar senses for a single instance.

All models, associated data sets, testing framework, and scores have been released as a part of the open-source S-Space Package (Jurgens and Stevens, 2010b).[5]

[5] http://code.google.com/p/airhead-research/

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.

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

Marco Baroni, Brian Murphy, Eduard Barbu, and Massimo Poesio. 2010. Strudel: A corpus-based semantic model based on properties and types. *Cognitive Science*, 34(2):222–254.

Stefan Bordag. 2006. Word sense induction: Triplet-based clustering and automatic evaluation. In *Proceedings of the 11th EACL*, pages 137–144.

Vladimir Braverman, Adam Meyerson, Rafail Ostrovsky, Alan Roytman, Michael Shindler, and Brian Tagiku. 2011. Streaming k-means on Well-Clusterable Data. In *Proceedings of SODA 2011*.

Alexander Budanitsky and Graeme Hirst. 2006. Evaluating wordnet-based measures of semantic distance. *Computational Linguistics*, 32(1):13–47, March.

Nathanael Chambers and Dan Jurafsky. 2010. Improving the Use of Pseudo-Words for Evaluating Selectional Preferences. In *ACL 2010*.

David Cheng, Ravi Kannan, Santosh Vempala, and Grant Wang. 2006. A divide-and-merge methodology for clustering. *ACM Transactions on Database Systems (TODS)*, 31(4):1499–1525.

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.

Stephen Clark and David Weir. 2002. Class-based probability estimation using a semantic hierarchy. *Computational Linguistics*, 28(2):187–206.

Georgiana Dinu and Mirella Lapata. 2010. Measuring distributional similarity in context. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1162–1172. Association for Computational Linguistics.

Beate Dorow and Dominic Widdows. 2003. Discovering corpus-specific word senses. In *Proceedings of the 10th EACL*, pages 79–82.

Wesam Elshamy, Doina Caragea, and William H. Hsu. 2010. KSU KDD: Word sense induction by clustering in topic space. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 367–370. 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.

William A. Gale, Kenneth W. Church, and David Yarowsky. 1992. Work on statistical methods for word sense disambiguation. In *AAAI Fall Symposium on Probabilistic Approaches to Natural Language*, pages 54–60.

Susan Gauch and Robert P. Futrelle. 1993. Experiments in automatic word class and word sense identification for information retrieval. In *Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 425–434.

Zhengyan He, Yang Song, and Houfeng Wang. 2010. Applying Spectral Clustering for Chinese Word Sense Induction. In *Proceedings of the CIPS-SIGHAN Joint Conference on Chinese Language Processing*.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: the 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL*, pages 57–60. Association for Computational Linguistics.

Jay J. Jiang and David W. Conrath. 1997. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In *Proceedings of International Conference on Research in Computational Linguistics*,.

Michael N. Jones, Walter Kintsch, and Doughlas J. K. Mewhort. 2006. High-dimensional semantic space accounts of priming. *Journal of Memory and Language*, 55:534–552.

David Jurgens and Keith Stevens. 2010a. HERMIT: Using word ordering applied to the Sense Induction task of SemEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluations*. Association for Computational Linguistics.

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

Roman Kern, Markus Muhr, and Michael Granitzer. 2010. KCDC: Word sense induction by using grammatical dependencies and sentence phrase structure. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 351–354. Association for Computational Linguistics.

Ioannis P. Klapaftis and Suresh Manandhar. 2008. Word sense induction using graphs of collocations. In *Proceeding of the 2008 conference on ECAI 2008: 18th European Conference on Artificial Intelligence*, pages 298–302.

Ioannis P. Klapaftis and Suresh Manandhar. 2010. Taxonomy learning using word sense induction. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 82–90, Morristown, NJ, USA. Association for Computational Linguistics.

Suresh Manandhar and Ioannis P. Klapaftis. 2009. SemEval-2010 Task 14: Evaluation Setting for Word Sense Induction & Disambiguation Systems. In *NAACL-HLT 2009 Workshop on Semantic Evaluations: Recent Achievements and Future Directions*.

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.

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*, 41(2):1–69.

Andrew Ng, Michael Jordan, and Yair Weiss. 2001. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems 14: Proceeding of the 2001 Conference*, pages 849–856.

Joakim Nivre, Johan Hall, and Jens Nilsson. 2006. MaltParser: A data-driven parser-generator for dependency parsing. In *Proceedings of LREC*, pages 2216–2219.

Sebastian Padó and Mirella Lapata. 2007. Dependency-Based Construction of Semantic Space Models. *Computational Linguistics*, 33(2):161–199.

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.

Patrick Pantel and Dekang Lin. 2002. Discovering word senses from text. In *Proceedings of ACM Conference on Knowledge Discovery and Data Mining (KDD-02)*, pages 613–619.

Ted Pedersen and Rebecca Bruce. 1997. Distinguishing word senses in untagged text. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, pages 197–207, August.

Ted Pedersen and Anagha Kulkarni. 2006. Automatic cluster stopping with criterion functions and the gap statistic. In *In Proceedings of the Demo Session of HLT/NAACL*, pages 276–279.

Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. 2004. Wordnet:: Similarity: measuring the relatedness of concepts. In *Demonstration Papers at HLT-NAACL 2004 on XX*, pages 38–41. Association for Computational Linguistics.

Ted Pedersen. 2010. Duluth-WSI: SenseClusters Applied to the Sense Induction Task of SemEval-2. In *Proceedings of the SemEval 2010 Workshop : the 5th International Workshop on Semantic Evaluations*, pages 363–366, July.

Yves Peirsman, Kris Heylen, and Dirk Geeraerts. 2008. Size matters. Tight and loose context definitions in English word space models. In *ESSLLI Workshop on Distributional Lexical Semantics*, pages 34–41.

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.

Magnus Sahlgren, Anders Holst, and Pentti Kanerva. 2008. Permutations as a means to encode order in word space. In *Proceedings of the 30th Annual Meeting of the Cognitive Science Society (CogSci'08)*.

Hinrich Schütze, 1992. *Context Space*, pages 113–120. AAAI Press, Menlo Park, CA.

Robert Tibshirani, Guenther Walther, and Trevor Hastie. 2000. Estimating the number of clusters in a dataset via the gap statistic. *Journal Royal Statistics Society B*, 63:411–423.

Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. pages 252–259.

Akira Utsumi. 2010. Exploring the Relationship between Semantic Spaces and Semantic Relations. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'2010)*, pages 257–262.

Ulrike von Luxburg. 2007. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416.

Ying Zhao and George Karypis. 2001. Criterion functions for document clustering: Experiments and analysis. Technical Report UMN CS 01-040, University of Minnesota.