UMND2 : SenseClusters Applied to the Sense Induction Task of SENSEVAL-4

Ted Pedersen
Department of Computer Science
University of Minnesota
Duluth, MN 55812
tpederse@d.umn.edu
http://senseclusters.sourceforge.net

Abstract

SenseClusters is a freely–available open–source system that served as the University of Minnesota, Duluth entry in the SENSEVAL-4 sense induction task. For this task SenseClusters was configured to construct representations of the instances to be clustered using the centroid of word co-occurrence vectors that replace the words in an instance. These instances are then clustered using k–means where the number of clusters is discovered automatically using the Adapted Gap Statistic. In these experiments SenseClusters did not use any information outside of the raw untagged text that was to be clustered, and no tuning of the system was performed using external corpora.

1 Introduction

The object of the sense induction task of SENSEVAL-4 (Agirre and Soroa, 2007) was to cluster 27,132 instances of 100 different words (35 nouns and 65 verbs) into senses or classes. The task data consisted of the combination of the test and training data (minus the sense tags) from the English lexical sample task. Each instance is a context of several sentences which contains an occurrence of a given word that serves as the target of sense induction.

SenseClusters is based on the presumption that words that occur in similar contexts will have similar meanings. This intuition has been presented as both the Distributional Hypothesis (Harris, 1968) and the Strong Contextual Hypothesis (Miller and Charles, 1991).

SenseClusters has been in active development at the University of Minnesota, Duluth since 2002. It is an open–source project that is freely–available from sourceforge, and has been been described in detail in numerous publications (e.g., (Purandare and Pedersen, 2004), (Pedersen et al., 2005), (Pedersen and Kulkarni, 2007)).

SenseClusters supports a variety of techniques for selecting lexical features, representing contexts to be clustered, determining the appropriate number of cluster automatically, clustering, labeling of clusters, and evaluating cluster quality. The configuration used in SENSEVAL-4 was just one possible combination of these techniques.

2 Methodology in Sense Induction Task

For this task, SenseClusters represents the instances to be clustered using second order co–occurrence vectors. These are constructed by first identifying word co–occurrences, and then replacing each word in an instance to be clustered with its co-occurrence vector. Then all the vectors that make up an instance are averaged together to represent that instance.

A co–occurrence matrix is constructed by identifying bigrams that occur in the contexts to be clustered two or more times and have a Pointwise Mutual Information (PMI) score greater than five. If the value of PMI is near 1.0, this means that the words in the bigram occur together approximately the number of times expected by chance, and they are not strongly associated. If this value is greater than 1, then the words in the bigram are occurring more often than expected by chance, and they are therefore associated.

The rows of the co–occurrence matrix represent the first word in the selected bigrams, and the columns represent the second word. A window size of 12 is allowed, which means that up to 10 intervening words can be observed between the pair of words in the bigram. This rather large window size was employed since the sample sizes for each word were relatively small, often no more than a few hundred instances.

A stop list was used to eliminate bigrams where either word is a high–frequency low–content word. The particular list used is distributed with the Ngram Statistics Package and is loosely based on the SMART stop list. It consists of 295 words; in addition, all punctuation, single letter words, and numbers (with the exception of years) were eliminated.

Each of the contexts that contain a particular target word is represented by a single vector that is the average (or the centroid) of all the co-occurrence vectors found for the words that make up the context. This results in a context by feature matrix, where the features are the words that occur with the words in the contexts (i.e., second order co–occurrences). The k–means algorithm is used for clustering the contexts, where the number of clusters is automatically discovered using the Adapted Gap Statistic (Pedersen and Kulkarni, 2006). The premise of this method is to create a randomized sample of data with the same characteristics of the observed data (i.e., the contexts to be clustered). This is done by fixing the marginal totals of the context by feature matrix and then generating randomized values that are consistent with those marginal totals. This creates a matrix that is can be viewed as being from the same population as the observed data, except that the data is essentially noise (because it is randomly generated).

The randomized data is clustered for successive values of k from 1 to some upper limit (the number of contexts or the point at which the criterion functions have plateaued). For each value of k the criterion function measures the quality of the clustering solution. The same is done for that observed data, and the difference between the criterion function for the observed data and the randomized data is determined, and the value of k where that difference is largest is selected as the best solution for k, since that is when the clustered data least resembles noise, and is therefore the most organized or best solution. In these experiments the criterion function was intra-cluster similarity.

3 Results and Discussion

There was an unsupervised and a supervised evaluation performed in the sense induction task. Official scores were reported for 6 participating systems, plus the most frequent sense (MFS) baseline, so rankings (when available) are provided from 1 (HIGH) to 7 (LOW). We also conducted an evaluation using the SenseClusters method.

3.1 Unsupervised Evaluation

The unsupervised evaluation was based on the traditional clustering measures of F-score, entropy, and purity. While the participating systems clustered the full 27,132 instances, only the 4,581 instance subset that corresponds to the English lexical sample evaluation data was scored in the evaluation. Table 1 shows the averaged F-scores over all 100 words, all 35 nouns, and all 65 verbs.

In this table the SenseClusters system (UMND2) is compared to the MFS baseline, which is attained by assigning all the instances of a word to a single cluster. We also include several random baselines, where randomX indicates that one of X possible clusters was randomly assigned to each instance of a word. Thus, approximately $100 * X$ distinct clusters are created across the 100 words. The random results are not ranked as they were not a part of the official evaluation. We also present the highest (HIGH, rank 1) and lowest (LOW, rank 7) scores from participating systems, to provide points of comparison.

The randomX baseline is useful in determining the sensitivity of the evaluation technique to the number of clusters discovered. The average number of classes in the gold standard test data is 2.9, so random3 approximates a system that randomly assigns the correct number of clusters. It attains an F-score of 50.0. Note that random2 performs somewhat better (59.7), suggesting that all other things being equal, the F-score is biased towards methods that find a smaller than expected number of clusters.

Table 1: Unsupervised F-Score (test)

	All	Nouns	Verbs	Rank
MFS/HIGH	78.9	80.7	76.8	1
UMND2	66.1	67.1	65.0	4
random2	59.7	60.9	58.4	
LOW	56.1	65.8	45.1	7
random3	50.0	49.9	50.1	
random4	44.9	44.2	45.7	
random10	29.7	28.0	31.7	
random50	17.9	14.9	21.1	

Table 2: Supervised Accuracy (test)

	All	Nouns	Verbs	Rank
HIGH	81.6	86.8	75.7	1
UMND2	80.6	84.5	76.2	2
random2	78.9	81.6	75.8	
MFS	78.7	80.9	76.2	4
LOW	78.5	81.4	75.2	7
random4	78.4	81.1	75.5	
random3	78.3	80.5	75.9	
random10	77.9	79.8	75.8	
random50	75.6	78.5	72.4	

As the number of random clusters increases the F-score declines sharply, showing that it is highly sensitive to the number of clusters discovered, and significantly penalizes systems that find more clusters than indicated in the gold standard data.

We observed for UMND2 that purity (81.7) is quite a bit higher than the F-score (66.1), and that it discovered a smaller number of clusters on average (1.4) than exists in the gold standard data (2.9). This shows that while SenseClusters was able to find relatively pure clusters, it errored in finding too few clusters, and was therefore penalized to some degree by the F-score.

3.2 Supervised Evaluation

A *supervised* evaluation was also carried out on the same clustering of the 27,132 instances as was used in the unsupervised evaluation, following the method defined in (Agirre et al., 2006). Here the train portion (22,281 instances) is used to learn a table of probabilities that is used to map discovered clusters in the test data to gold standard classes. The cluster assigned to each instance in the test portion (4,851 instances) is mapped (assigned) to the most probable class associated with that cluster as defined by this table.

After this transformation is performed, the newly mapped test results are scored using the scorer2 program, which is the official evaluation program of the English lexical sample task and reports the F-measure, which in these experiments is simply accuracy since precision and recall are the same.

In Table 2 we show the results of the supervised evaluation, which includes the highest and lowest score from participating systems, as well as UMND2, MFS, and the same randomX baselines as included in the unsupervised evaluation.

We observed that the difference between the score of the best performing system (HIGH) and the random50 baseline is six points (81.6 - 75.6). In the unsupervised evaluation of this same data this difference is 61 points (78.9 - 17.9) according to the F-score.

The smaller range of values for the supervised measure can be understood by noting that the mapping operation alters the number and distribution of clusters as discovered in the test data. For example, random3 results in an average of 2.9 clusters per word in the test data, but after mapping the average number of clusters is 1.1. The average number of clusters discovered by UMND2 is 1.4, but after mapping this average is reduced to 1.1. For random50, the average number of clusters per word is 24.1, but after mapping is 2.0. This shows that the supervised evaluation has a tendency to converge upon the MFS, which corresponds to assigning 1 cluster per word.

When looking at the randomX results in the supervised evaluation, it appears that this method does not penalize systems for getting the number of clusters incorrect (as the F-score does). This is shown by the very similar results for the randomX baselines, where the only difference in their results is the number of clusters. This lack of a penalty is due to the fact that the mapping operation takes a potentially large number of clusters and maps them to relatively few classes (e.g., random50) and then performs the evaluation.

3.3 SenseClusters Evaluation (F-Measure)

An evaluation was carried out on the full 27,132 instance train+test data set using the SenseClusters evaluation methodology, which was first defined in (Pedersen and Bruce, 1997). This corresponds to an unsupervised version of the F-measure, which in these experiments can be viewed as an accuracy measure since precision and recall are the same (as is the case for the supervised measure).

It aligns discovered clusters with classes such that their agreement is maximized. The clusters and classes must be aligned one to one, so a large penalty can result if the number of discovered clusters differs from the number of gold standard classes.[1]

For UMND2, there were 145 discovered clusters and 368 gold standard classes. Due to the one to one alignment that is required, the 145 discovered clusters were aligned with 145 gold standard classes such that there was agreement for 15,291 of 27,132 instances, leading to an F-measure (accuracy) of 56.36 percent. Note that this is significantly lower than the F-score of UMND2 for the train+test data, which was 63.1. This illustrates that the SenseClusters F-measure and the F-score are not equivalent.

4 Conclusions

One of the strengths of SenseClusters (UMND2) is that it is able to automatically identify the number of clusters without any manual intervention or setting of parameters. In these experiments the Adapted Gap statistic was quite conservative, only discovering on average 1.4 classs per word, where the actual number of classes in the gold standard data was 2.9. However, this is a reasonable result, since for many words there were just a few hundred instances. Also, the gold standard class distinctions were heavily skewed, with the majority sense occurring 80% of the time on average. Under such conditions, there may not be sufficient information available for an unsupervised clustering algorithm to make fine grained distinctions, and so discovering one cluster for a word may be a better course of action that making divisions that are not well supported by the data.

[1] An implementation of this measure is available in the SenseClusters system, or by contacting the author.

5 Acknowledgments

These experiments were conducted with version 0.95 of the SenseClusters system. Many thanks to Amruta Purandare and Anagha Kulkarni for their invaluable work on this and previous versions of SenseClusters.

This research was partially supported by the National Science Foundation Faculty Early Career Development (CAREER) Program (#0092784).

References

E. Agirre and A. Soroa. 2007. Semeval-2007 task 2: Evaluating word sense induction and discrimination systems. In *Proceedings of SemEval-2007: 4th International Workshop on Semantic Evaluations*, June.

E. Agirre, D. Martínez, O. López de Lacalle, and A. Soroa. 2006. Two graph-based algorithms for state-of-the-art wsd. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 585–593, Sydney, Australia, July.

Z. Harris. 1968. *Mathematical Structures of Language*. Wiley, New York.

G.A. Miller and W.G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

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

T. Pedersen and A. Kulkarni. 2006. Automatic cluster stopping with criterion functions and the Gap Statistic. In *Proceedings of the Demo Session of HLT/NAACL*, pages 276–279, New York City, June.

T. Pedersen and A. Kulkarni. 2007. Unsupervised discrimination of person names in web contexts. In *Proceedings of the Eighth International Conference on Intelligent Text Processing and Computational Linguistics*, pages 299–310, Mexico City, February.

T. Pedersen, A. Purandare, and A. Kulkarni. 2005. Name discrimination by clustering similar contexts. In *Proceedings of the Sixth International Conference on Intelligent Text Processing and Computational Linguistics*, pages 220–231, Mexico City, February.

A. Purandare and T. Pedersen. 2004. Word sense discrimination by clustering contexts in vector and similarity spaces. In *Proceedings of the Conference on Computational Natural Language Learning*, pages 41–48, Boston, MA.