Duluth-WSI: SenseClusters Applied to the Sense Induction Task of SemEval-2

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

Abstract

The Duluth-WSI systems in SemEval-2 built word co–occurrence matrices from the task test data to create a second order co–occurrence representation of those test instances. The senses of words were induced by clustering these instances, where the number of clusters was automatically predicted. The Duluth-Mix system was a variation of WSI that used the combination of training and test data to create the co-occurrence matrix. The Duluth-R system was a series of random baselines.

1 Introduction

The Duluth systems in the sense induction task of SemEval-2 (Manandhar et al., 2010) were based on SenseClusters (v1.01), a freely available open source software package which relies on the premise that words with similar meanings will occur in similar contexts (Purandare and Pedersen, 2004). The data for the sense induction task included 100 ambiguous words made up of 50 nouns and 50 verbs. There were a total of 8,915 test instances and 879,807 training instances provided. Note that neither the training nor the test data was sense tagged. The training data was made available as a resource for participants, with the understanding that system evaluation would be done on the test instances only. The organizers held back a gold standard annotation of the test data that was only used for evaluation.

Five Duluth-WSI systems participated in this task, six Duluth-Mix systems, and five Duluth Random systems. The WSI and Mix systems almost always represented the test instances using second order co–occurrences, where each word in a test instance is replaced by a vector that shows the words with which it co-occurs. The word vectors that make up a test instance are averaged together to make up a new representation for that instance. All the test instances for a word are clustered, and the number of senses is automatically predicted by either the PK2 measure or Adapted Gap Statistic (Pedersen and Kulkarni, 2006).

In the Duluth systems the co-occurrence matrices are either based on order-dependent bigrams or unordered pairs of words, both of which can be separated by up to some given number of intervening words. Bigrams are used to preserve distinctions between collocations such as *cat house* and *house cat*, whereas co–occurrences do not consider order and would treat these two as being equivalent.

2 Duluth-WSI systems

The Duluth-WSI systems build co-occurrence matrices from the test data by identifying bigrams or co–occurrences that occur with up to eight intermediate words between them in instances of ambiguous nouns, and up to 23 intermediate words for the verbs. Any bigram (bi) or co–occurrence (co) that occurs more than 5 times with up to the allowed number of intervening words and has statistical significance of 0.95 or above according to the left-sided Fisher's exact test was selected (Pedersen et al., 1996). Some of the WSI systems reduce the co–occurrence matrix to 300 dimensions using Singular Value Decomposition (SVD).

The resulting co-occurrence matrix was used to create second order co–occurrence vectors to represent the test instances, which were clustered using the method of repeated bisections (rb), where similarity was measured using the cosine. Table 1 summarizes the distinctions between the various Duluth-WSI systems.

3 Duluth-Mix systems

The Duluth-Mix systems used the combination of the test and training data to identify features to represent the test instances. The goal of this combi-

Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, pages 363–366,
Uppsala, Sweden, 15-16 July 2010.

Table 1: Duluth-WSI Distinctions

name	options
Duluth-WSI	bigrams, no SVD, PK2
Duluth-WSI-Gap	bigrams, no SVD, Gap
Duluth-WSI-SVD	bigrams, SVD, PK2
Duluth-WSI-Co	co-occur, no SVD, PK2
Duluth-WSI-Co-Gap	co-occur, no SVD, Gap

nation was to increase the amount of data that was available for feature identification. Since there was a larger amount of data, some parameter settings as used in Duluth-WSI were reduced.

For example, the Duluth-Mix-PK2 and Duluth-Mix-Gap are identical to the Duluth-WSI and Duluth-WSI-Gap systems, except that they limit both nouns and verbs to 8 intervening words. Duluth-Mix-Narrow-PK2 and Duluth-Mix-Narrow-Gap are identical to Duluth-Mix-PK2 and Duluth-Mix-Gap except that bigrams and co–occurrences must be made up of adjacent words, with no intermediate words allowed.

Duluth-Mix-Uni-PK2 and Duluth-Mix-Uni-Gap are unique among the Duluth systems in that they do not use second order co-occurrences, but instead rely on first order co-occurrences. These are simply individual words (unigrams) that occur more than 5 times in the combined test and training data. These features are used to generate co-occurrence vectors for the test instances which are then clustered (this is very similar to a bag of words model).

4 Duluth-Random systems

Duluth-R12, Duluth-R13, Duluth-R15, and Duluth-R110 provide random baselines. R12 randomly assigns each instance to one of two senses, R13 to one of three, R15 to one of five, and R110 to one of ten senses. Random numbers are generated in the given range with equal probability, so the distribution of assigned senses is balanced.

5 Discussion

The evaluation of unsupervised sense discrimination and induction systems is still not standardized, so an important part of any exercise like SemEval-2 is to scrutinize the evaluation measures used in order to determine to what degree they are providing a useful and reasonable way of evaluating system results.

5.1 Evaluation Measures

Each participating system was scored by three different evaluation methods: the V-measure (Rosenberg and Hirschberg, 2007), the supervised recall measure (Agirre and Soroa, 2007), and the paired F-score (Artiles et al., 2009). The results of the evaluation are in some sense confusing - a system that ranks near the top according to one measure may rank at the bottom or middle of another. There was not any single system that did well according to all of the different measures. The situation is so extreme that in some cases a system would perform near the top in one measure, and then below random baselines in another. These stark differences suggest a real need for continued development of other methods for evaluating unsupervised sense induction.

One minimum expectation of an evaluation measure is that it should expose and identify random baselines by giving them low scores that clearly distinguish them from actual participating systems. The scores of all the evaluation measures used in this task when applied to different random baseline systems are summarized in Table 2. These include a number of post-evaluation random clustering systems, which are referred to as post-R1k, where k is the number of random clusters.

5.1.1 V-measure

The V-measure appears to be quite easily mislead by random baselines. As evidence of that, the Duluth-R (random) systems got increasingly better scores the more random they became, and in fact the post-evaluation random systems reached levels of performance better than any of the participating systems. Table 2 shows that the V-measure continues to improve (rather dramatically) as randomness increases.

The average number of senses in the gold standard data for all 100 words was 3.79. The official random baseline assigned one of four random senses to each instance of a word, and achieved a V-measure of 4.40. Duluth-R15 improved the V-measure to 5.30 by assigning one of five random senses, and Duluth-R110 improved it again to 8.60 by assigning one of ten random senses. The more random the result, the better the score. In fact Duluth-R110 placed sixth in the sense in-

duction task according to the V-measure. In post-evaluation experiments a number of additional random baselines were explored, where instances were assigned senses randomly from 20, 33, and 50 possible values per word. The V-measures for these random systems were 13.9, 18.7, and 23.2 respectively, where the latter two were better than the first place participating system (which scored 16.2). In a post-evaluation experiment, the task organizers found that assigning one sense per instance resulted in a V–measure of 31.7.

5.1.2 Supervised Recall

The supervised recall measure takes the sense induction results (on the 8,915 test instances) as submitted by a participating system and splits that into a training and test portion for supervised learning. The recall attained on the test split by a classifier learned on the training split becomes the measure of the unsupervised system. Two different splits were used, with 80% or 60% of the test instances for training, and the remainder for testing.

This evaluation method was also used in SemEval-1, where (Pedersen, 2007) noted that it seemed to compress the results of all the systems into a narrow band that converged around the Most Frequent Sense result. The same appears to have happened in 2010. The supervised recall of the Most Frequent Sense baseline (MFS) is 58 or .59 (depending on the split), and the majority of participating systems (and even some of the random baselines) fall in a range of scores from .56 to .62 (a band of .06). This blurs distinctions among participating systems with each other and with random baselines.

The number of senses actually assigned by the classifier learned from the training split to the instances in the test split is quite small, regardless of the number of senses discovered by the participating system. There were *at most* 2.06 senses identified per word based on the 80-20 split, and *at most* 2.27 senses per word based on the 60-40 split. For most systems, regardless of their underlying methodology, the number of senses the classifier actually assigns is approximately 1.5 per word. This shows that the supervised learning algorithm that underlies this evaluation method gravitates towards a very small number of senses and therefore tends to converge on the MFS baseline. This could be caused by noise in the induced senses, a small number of examples in the training split for a sense, or it may be that the supervised recall measure is making different distinctions than are found by the unsupervised sense induction method it seeks to evaluate.

Table 2: Evaluation of Random Systems

name	k	V	F	60-40	80-20
MFS	1	0.0	**63.4**	**58.3**	**58.7**
Duluth-R12	2	2.3	47.8	57.7	58.5
Duluth-R13	3	3.6	38.4	57.6	58.0
Random	4	4.4	31.9	56.5	57.3
Duluth-R15	5	5.3	27.6	56.5	56.8
Duluth-R110	10	8.6	16.1	53.6	54.8
post-R120	20	13.9	7.5	46.2	48.6
post-R133	33	18.7	4.0	38.3	42.5
post-R150	50	**23.2**	2.3	30.0	34.2

5.1.3 Paired F-score

The paired F-score was the only evaluation measure that seemed able to identify and expose random baselines. Duluth-R110 was by far the most random of the officially participating systems, and it was by far the lowest ranked system according to the paired F-score, which assigned it a score of 16.1. All the Duluth-R systems ranked relatively low (20th or below). When presented with the 20, 33, and 50 random sense post–evaluation systems, the F-score assigned those scores of 7.46, 4.00, and 2.33, which placed them far below any of the other systems.

However, the paired F-score also showed that the Most Frequent Sense baseline outperformed all of the participating systems. The systems that scored close to the MFS tended to predict very small numbers of senses, and so were in effect acting much like the MFS baseline themselves. The F-score is not bounded by MFS and in fact it is possible (theoretically) to reach a score of 1.00 with a perfect assignment of instances to senses. The lesson learned in this task is that it would have been more effective to simply assume that there was just one sense per word, rather than using the senses induced by participating systems. While this may be a frustrating conclusion, in fact it is a reasonable observation given that in many domains a single sense for a given word can tend to dominate.

5.2 Duluth-WSI and Duluth-Mix Results

The Duluth-WSI systems used the test data to build co-occurrence matrices, while the Duluth-

Mix systems used both the training and test data. Within those frameworks bigrams or co–occurrences were used to represent features, the number of senses was automatically discovered with the PK2 measure or the Adapted Gap Statistic, and SVD was optionally used to reduce the dimensionality of the resulting matrix. Previous studies using SenseClusters have noted that the Adapted Gap Statistic tends to find a relatively small number of clusters, and that SVD typically does not help to improve results of unsupervised sense induction. These findings were again confirmed in this task.

Mixing together all of the training and test data for building the co–occurrence matrices was no more effective than just using the test data. However, the Duluth-Mix systems did not finish before the end of the evaluation period. The Duluth-Mix-Narrow-Gap and PK2 systems were able to finish 8,211 of the 8,915 test instances (92%), the Duluth-Mix-Gap and PK2 systems completed 7,417 instances (83%), and Duluth-Mix-Uni-PK2 and Gap systems completed 2,682 of these instances (30%). While these are partial results they seem sufficient to support this conclusion.

To be usable in practical settings, an unsupervised sense induction system should discover the number of senses accurately and automatically. Duluth-WSI and Duluth-WSI-SVD were very successful in that regard, and predicted 4.15 senses on average per word (with the PK2 measure) while the actual number of senses was 3.79.

The Duluth-WSI systems are direct descendents of UMND2 which participated in SemEval-1 (Pedersen, 2007), where Duluth-WSI-Gap is the closest relative. However, UMND2 used Pointwise Mutual Information (PMI) rather than Fisher's left sided test, and it performed clustering with k-means rather than the method of repeated bisections. Both UMND2 and Duluth-WSI-Gap used the Adapted Gap Statistic, and interestingly enough both discovered approximately 1.4 senses on average per word.

6 Conclusion

The SemEval-2 sense induction task was an opportunity to compare participating systems with each other, and also to analyze evaluation measures. At the very least, an evaluation measure should penalize random results in a fairly significant way. This task showed that the paired F-score is able to identify and expose random baselines, and that it drives them far down the rankings and places them well below participating systems. This seems preferable to the V-measure, which tends to rank random systems above all others, and to supervised recall, which provides little or no separation between random baselines and participating systems.

References

E. Agirre and A. Soroa. 2007. SemEval-2007 Task 02: Evaluating word sense induction and discrimination systems. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 7–12, Prague, Czech Republic, June.

J. Artiles, E. Amigó, and J. Gonzalo. 2009. The role of named entities in Web People Search. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 534–542, Singapore, August.

S. Manandhar, I. Klapaftis, D. Dligach, and S. Pradhan. 2010. SemEval-2010 Task 14: Word sense induction and disambiguation. In *Proceedings of the SemEval 2010 Workshop : the 5th International Workshop on Semantic Evaluations*, Uppsala, Sweden, July.

T. Pedersen and A. Kulkarni. 2006. Automatic cluster stopping with criterion functions and the gap statistic. In *Proceedings of the Demonstration Session of the Human Language Technology Conference and the Sixth Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 276–279, New York City, June.

T. Pedersen, M. Kayaalp, and R. Bruce. 1996. Significant lexical relationships. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, pages 455–460, Portland, OR, August.

T. Pedersen. 2007. UMND2 : SenseClusters applied to the sense induction task of Senseval-4. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 394–397, Prague, Czech Republic, June.

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.

A. Rosenberg and J. 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*, pages 410–420, Prague, Czech Republic, June.