Improving Word Sense Discrimination with Gloss Augmented Feature Vectors

Amruta Purandare[1] and Ted Pedersen[2]

[1] University of Pittsburgh, Pittsburgh, PA 15260 USA
[2] University of Minnesota, Duluth, MN 55812 USA
http://senseclusters.sourceforge.net

Abstract. This paper presents a method of unsupervised word sense discrimination that augments co–occurrence feature vectors derived from raw untagged corpora with information from the glosses found in a machine readable dictionary. Each content word that occurs in the context of a target word to be discriminated is represented by a co-occurrence feature vector. Each of these vectors is augmented with the content words that occur in the glosses of the different possible meanings of the word it represents. Then these vectors are averaged to create a vector that represents that context of the target word. Discrimination is carried out by clustering all of the vectors associated with the contexts in which the target word occurs. We show via an evaluation with the SENSEVAL-2, *line*, *hard* and *serve* corpora that feature vectors augmented with gloss information from WordNet significantly improve discrimination performance when limited data is available.

1 Introduction

Word sense discrimination is the task of grouping multiple occurrences of a given target word into clusters, where each cluster represents a distinct meaning or sense of that word (e.g., [PB97], [Sch98], [PP04]). Approaches to this problem rely on the notion that words that are used in similar contexts will have the same or a closely related meaning [MC91]. Note that this is not the same as word sense disambiguation, in that there are no sense tags attached to the clusters. Rather instances of a word that occur in similar contexts are grouped together.

We take a context vector approach to sense discrimination. Each context in which a target word occurs in a set of test data is represented by a vector, which is in turn the average of a set of feature vectors that represent each word that occurs in that context. Each feature vector represents the co–occurrence behavior of a word and is derived from a separate corpus of training data. Note that both the training and test data are simply raw text, and the process is completely unsupervised. This paper describes recent enhancements to this approach, where we augment feature vectors with words derived from dictionary glosses.

The motivation for our approach is to improve discrimination accuracy when only limited amounts of corpora are available. For example, in certain specialized domains or minority languages dictionaries may be available, but the amount

of online corpora might be rather small. Our approach offers a way to exploit unsupervised techniques using modestly sized corpora.

This paper begins with a discussion of context vector sense discrimination in general, and then presents our method of enhancing this process with dictionary glosses. We describe an evaluation with the SENSEVAL-2, *line*, *hard*, and *serve* corpora, and then conclude with a discussion of future work.

2 Context Vector Sense Discrimination

We have developed a method of context vector sense discrimination (e.g., [Pur03], [PP04]) that originally followed from [Sch98], and is related to Latent Semantic Indexing [DDF$^+$90] and Latent Semantic Analysis [LFL98]. The object of the algorithm is to take a set of instances of a particular target word, and cluster them such that instances with similar or related meanings of that target word are grouped together.

Discrimination starts by building a co–occurrence matrix of words from a training corpus. This may be a collection of 2-3 sentence instances, each of which contain a particular target word, or it may be a more general resource such as the Wall Street Journal or the British National Corpus. In the latter case the target word may occur in that corpus, although that is not strictly necessary.

The rows and columns of the co–occurrence matrix represent words, and the cells in the matrix indicate if those words co–occur in the training data. The words on the rows and columns are selected from the training corpus based on a combination of frequency cutoffs and measures of association. Each cell in this matrix contains a binary value indicating if the pair of words co–occur (or not). We adopt the convention that two words that occur within five positions of each other (i.e., with up to three words between them) are co–occurrences. Each row of the resulting matrix serves as a feature vector showing the co–occurrence behavior of the associated word. As the co–occurrence matrices are usually very large and sparse, the dimensionality reduction techniques such as Singular Value Decomposition (SVD) are often employed. SVD converts a word level feature space into a concept level semantic space and thus addresses both the problems of polysemy (using same word with different meanings) and synonymy (using multiple words to describe the same concept) (e.g., [DDF$^+$90], [BDO95], [BDS95]).

Once the word co–occurrence matrix is ready, the attention then shifts to a separate set of test data, which consists of the instances of a given target word that are to be discriminated. For each instance in the test data, the content words are represented by their associated feature vectors taken from the co–occurrence matrix. All of the feature vectors associated with the words in the context of the test instance are averaged to create a single vector that represents that context. Thus, a set of test data that consists of multiple instances is converted into a set of context vectors, each of which represents an instance of the target word.

Discrimination actually takes place by clustering the context vectors using a partitional or agglomerative algorithm (e.g., [JMF99], [JD88], [ZK02]). The

resulting clusters are made up of instances that are used in similar contexts, and the presumption is that each cluster will represent a different meaning of the target word.

3 Gloss Augmented Feature Vectors

The context vector word sense discrimination algorithm requires a large amount of training corpora to learn the co–occurrence behavior of words. This method therefore does not perform very well if the available corpus is too small to study this behavior. The feature vectors generated using a small corpus are represented in a very small dimensional vector space (of few hundred dimensions) that does not completely describe the word co–occurrence patterns as seen in natural language text. This results into very sparse context vectors that, in turn, do not accurately capture the meaning of these contexts. In order to address this problem while dealing with a small corpus, we augment the feature vector of each word with the content words that occur in the glosses of the various senses of that word. The intuition here is that, with each word there is an associated set of words that strike either because these words heavily co–occur with that word in the text or because they are seen in its dictionary definition.

In our experiments we utilize WordNet 2.0 as the source of our glosses, but any machine readable dictionary would suffice. We scan the co–occurrence matrix, and for each word represented by a row we look up all the senses in WordNet in which that word can be used. Then we take the content words from the glosses of those senses and include them in the feature vector for that word. If a word in the co–occurrence matrix does not appear in WordNet then there is no augmentation.

For example, suppose that the feature vector (i.e., the corresponding row in the co–occurrence matrix) of *history* has non-zero entries associated with the words *arts, world, museum, books,* and *education*. Suppose that the various dictionary glosses of *history* include the words *century, record, past, events, time, arts, discipline* and *world*. The feature vector associated with *history* would be augmented with those words from the gloss that are not already present, resulting in : *arts, world, museum, books, education, century, record, past, events, time* and *discipline*.

Since many words have multiple senses, this can lead to a significant augmentation to a feature vector. This can also have the effect of greatly expanding the size of the co–occurrence matrix (if there are words in the dictionary glosses that were not observed in the training data), so after gloss augmentation we employ Singular Value Decomposition [BDO$^+$93] to reduce the feature space, and thereby smooth out the many zero valued entries. We use a reduction factor of 50, which means that the dimensions (the columns) in the co–occurrence matrix are reduced down to 2% of their original number.

Then, the context of each test instance is represented by averaging the feature vectors of all the words that appear in the context. In these experiments the context vectors are clustered using the Unweighted Pair Group Method with

Arithmetic mean (UPGMA), which is an agglomerative algorithm [Kar02]. The clusters of instances that are discovered are contextually more similar and hence are more likely to be semantically related.

To evaluate our results, we use manually assigned sense tags (if available) in the test data. We assign each cluster to the sense with which it shares the most instances. Each cluster can be assigned at most one sense. We measure precision and recall, where precision is the number of instances correctly clustered among the total number of instances clustered. Recall is the number of instances correctly clustered divided by the total number of instances available. The harmonic mean of these values, the F-measure, is used to summarize these results.

4 Experimental Methodology

Seventy–five discrimination experiments were carried out, using 72 words from the SENSEVAL-2 corpus, and the *line*, *hard* and *serve* corpora.[1] The words and their corresponding parts of speech are shown in Table 1. Note that each word is used in only one part of speech: noun, verb, or adjective. In all of this data, each instance is two or three sentences long, and contains a single usage of a given target word. Note that each target word is treated separately, and that at no time do we mix data from multiple target words.

The SENSEVAL-2 data has an existing division into test and training portions that we utilize. The training data contains 8,609 instances, while the test data includes 4,327 instances. However, on a word by word basis this data is relatively small. Each word has approximately 50 to 200 training and test instances. This is particularly challenging for unsupervised techniques since many of the target words have a large number of fine grained senses, generally from 8 to 12. The limited amount of data combined with so many possible senses leads to some senses that have small numbers of associated instances. As a result, we filter out any sense that constitutes less than 5% of the available senses in the training data and also in the test data. This reduced the number of possible senses to approximately 4 to 7 in the most typical cases, which is still a challenging number.

By contrast, the *line*, *hard*, and *serve* data is much larger, each with approximately 4,200 instances. The number of possible senses is smaller, where *line* has 6, *hard* has 3 and *serve* has 4. Each sense is well represented in the data, so no filtering is performed. This data does not have a standard test and training split, so we randomly divided each corpus into an 60-40 training-test split. This results in approximately 2,520 training instances and 1,680 test instances per word.

In our experiments, we always identify 10 clusters regardless of the number of senses associated the target word. This reflects the fact that we do not know a–priori how many senses a word will have, since we ignore the sense tags

[1] The SENSEVAL-2 data contains 73 words, but one of them (*ferret*) consists of three instances and was omitted from this study.

present in the data we are using until evaluation. In addition, it allows us to test the hypothesis that a good clustering approach will automatically discover approximately the same number of clusters as there are senses, and that the extra clusters (10–#senses) will contain very few instances.

word	F-nogl	F-gl	word	F-nogl	F-gl	word	F-nogl	F-gl
art.n	40.00	**50.95**	authority.n	49.70	40.00	bar.n	54.39	50.44
begin.v	49.69	**59.88**	blind.a	32.43	**45.00**	bum.n	60.32	36.36
call.v	35.44	**37.11**	carry.v	44.74	40.97	chair.n	48.00	**71.03**
channel.n	45.16	32.81	child.n	56.86	50.91	church.n	41.76	**54.37**
circuit.n	42.46	34.24	collaborate.v	40.00	**59.09**	colourless.a	56.00	**58.62**
cool.a	31.32	**35.56**	day.n	44.15	**65.31**	detention.n	62.22	42.55
develop.v	34.55	**39.64**	draw.v	41.86	**52.38**	dress.v	37.89	37.50
drift.v	39.29	**46.43**	drive.v	45.61	**54.54**	dyke.n	48.78	**60.00**
face.v	41.79	**77.01**	facility.n	43.90	**46.00**	faithful.a	42.42	42.42
fatigue.n	49.18	**64.79**	feeling.n	33.90	**46.58**	find.v	30.23	**41.86**
fine.a	41.51	**48.21**	fit.a	40.91	40.91	free.a	45.61	**47.79**
graceful.a	38.89	38.89	green.a	56.21	55.07	grip.n	41.46	**53.33**
hearth.n	57.70	44.90	holiday.n	37.74	**44.89**	keep.v	35.82	**67.50**
lady.n	37.34	**54.54**	leave.v	50.98	39.60	live.v	36.36	31.77
local.a	44.07	41.94	match.v	41.27	**52.94**	material.n	38.71	**41.60**
mouth.n	33.71	**39.21**	nation.n	59.26	**76.67**	natural.a	33.07	**34.78**
nature.n	36.84	33.73	oblique.a	40.00	**54.55**	play.v	48.72	37.33
post.n	47.70	39.39	pull.v	45.28	44.44	replace.v	38.24	**52.38**
restraint.n	40.54	35.90	see.v	33.34	**34.70**	sense.n	32.19	**39.08**
serve.v	50.64	45.98	simple.a	33.96	**47.06**	solemn.a	25.00	**47.06**
spade.n	44.90	**48.14**	stress.n	42.86	36.07	strike.v	37.50	**40.62**
train.v	41.13	41.13	treat.v	47.76	47.37	turn.v	40.00	34.62
use.v	31.20	**62.12**	vital.a	5.56	5.56	wander.v	30.13	**56.41**
wash.v	66.67	60.00	work.v	39.21	**49.18**	yew.n	56.41	**68.19**
line.n	43.13	43.04	hard.a	67.25	67.09	serve.v	38.54	36.60

Table 1. F–measures with *(F-gl)* and without *(F-nogl)* gloss augmentation

5 Experimental Results

Table 1 shows the F–measure of word sense discrimination attained for each word, with *(F-gl)* and without *(F-nogl)* gloss augmentation. Entries in bold type show the experiments where gloss augmented feature vectors resulted in significantly better performance than using feature vectors derived strictly from training data.

Out of the 72 SENSEVAL-2 words, a total of 43 showed improved F-measures using gloss augmented feature vectors. There were seven words that showed no significant change. In addition, all of these 43 words also showed improved recall

when using gloss augmented feature vectors, which shows that the number of instances correctly clustered was increased due to the use of the gloss augmentation.

A further analysis showed that not all of these 43 words showed a similar increase in their precision, which further indicates that the gloss augmentation not only increased the number of instances correctly clustered but also increased the total number of instances attempted by the algorithm. This is because the rise in the total number of instances correctly clustered was accompanied by a rise in the total number of instances attempted, resulting in relatively steady precision.

Our hypothesis is that the sparsity in the feature vectors without gloss augmentation left large number of instances unclustered due to very low levels of similarity with any of the other instances. We believe that gloss augmentation increases the likelihood of discriminating instances that have a very distinct set of features that may not be shared by other instances. Thus, the gloss augmentation allowed for a certain amount of standardization in the feature vectors, which raised the number of instances that were successfully clustered.

However, the results for *line*, *hard* and *serve* do not show any clear improvement when using gloss augmented feature vectors. We believe that this is due to the fact that most of the words that occur in the dictionary glosses of these words have already occurred in these larger samples of training data, so the gloss information is essentially redundant. Thus, we believe that gloss augmented feature vectors are particularly useful for situations where unsupervised discrimination must be performed using smaller samples of training data.

6 Future Directions

Our next round of experiments will compare the discrimination results attained by gloss augmented feature vectors with feature vectors derived from very large corpora such as the British National Corpus. In the latter case the feature vectors for each word will represent general co–occurrence behavior, and will not be specific to particular target words as they were in this paper. This experiment will allow us to test the hypothesis that relatively small amounts of focused corpora, when augmented with dictionary gloss content, are more effective for discrimination than more generic word vectors from very large corpora.

In certain circumstances it is possible to create very large corpora from the Web for certain words in major languages. However, the quality of such data may be unpredictable, and there could be a significant amount of noise. In order to test this hypothesis, we will use the Web as a source of training data for selecting features and constructing feature vectors. We will then compare the performance of Web derived feature vectors with those augmented with dictionary glosses.

We believe that smaller amounts of higher quality corpora may be more useful in certain circumstances, so we will continue to explore methods of augmenting such data with data from other resources so as to improve word sense discrim-

ination. In addition to dictionary glosses, we also plan to incorporate classes of words that are discovered via other unsupervised techniques (e.g., [PL02]).

7 Conclusion

There may be situations when an unsupervised learning approach to word sense discrimination does not have sufficient training data from which to learn a truly discriminating set of features. In such situations, if an external resource such as a machine readable dictionary is available, we have shown that augmenting the feature vectors of words in a co–occurrence matrix with words that occur in dictionary glosses results in improved discrimination performance.

8 Acknowledgments

This research is supported by a National Science Foundation Faculty Early CAREER Development Award (#0092784).

All of the experiments in this paper were carried out with version 0.47 of the SenseClusters package, freely available from the URL shown on the title page.

This work was completed while the first author was at the University of Minnesota, Duluth.

References

[BDO+93] M. Berry, T. Do, G. O'Brien, V. Krishna, and S. Varadhan. SVDPACK (version 1.0) user's guide. Technical Report CS-93-194, University of Tennessee at Knoxville, Computer Science Department, April 1993.

[BDO95] M.W. Berry, S. Dumais, and G. O'Brien. Using linear algebra for intelligent information retrieval. *SIAM Review*, 37(4):573–595, 1995.

[BDS95] M.W. Berry, S. Dumais, and A. Shippy. A case study of latent semantic indexing. Technical Report CS-95-271, University of Tennessee at Knoxville, Computer Science Department, January 1995.

[DDF+90] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391–407, 1990.

[JD88] A. Jain and R. Dubes. *Algorithms for Clustering Data*. Prentice–Hall, Inc., Upper Saddle River, NJ, 1988.

[JMF99] A. Jain, M. Murthy, and P. Flynn. Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323, September 1999.

[Kar02] G. Karypis. CLUTO - a clustering toolkit. Technical Report 02-017, University of Minnesota, Department of Computer Science, August 2002.

[LFL98] T.K. Landauer, P.W. Foltz, and D. Laham. An introduction to latent semantic analysis. *Discourse Processes*, 25:259–284, 1998.

[MC91] G.A. Miller and W.G. Charles. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28, 1991.

[PB97] T. Pedersen and R. Bruce. 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 1997.

[PL02] P. Pantel and D. Lin. Discovering word senses from text. In *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining-2002*, 2002.

[PP04] A. Purandare and T. Pedersen. 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, 2004.

[Pur03] A. Purandare. Discriminating among word senses using McQuitty's similarity analysis. In *Proceedings of the HLT-NAACL 2003 Student Research Workshop*, pages 19–24, Edmonton, Alberta, Canada, May 27 - June 1 2003.

[Sch98] H. Schütze. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123, 1998.

[ZK02] Y. Zhao and G. Karypis. Evaluation of hierarchical clustering algorithms for document datasets. In *Proceedings of the 11th International Conference on Information and Knowledge Management*, pages 515–524, McLean, VA, 2002.