SenseClusters: Unsupervised Clustering and Labeling of Similar Contexts

Anagha Kulkarni and **Ted Pedersen**
Department of Computer Science
University of Minnesota
Duluth, MN 55812
{kulka020,tpederse}@d.umn.edu
http://senseclusters.sourceforge.net

Abstract

SenseClusters is a freely available system that identifies similar contexts in text. It relies on lexical features to build first and second order representations of contexts, which are then clustered using unsupervised methods. It was originally developed to discriminate among contexts centered around a given target word, but can now be applied more generally. It also supports methods that create descriptive and discriminating labels for the discovered clusters.

1 Introduction

SenseClusters seeks to group together units of text (referred to as contexts) that are similar to each other using lexical features and unsupervised clustering.

Our initial work (Purandare and Pedersen, 2004) focused on word sense discrimination, which takes as input contexts that each contain a given target word, and produces as output clusters that are presumed to correspond to the different senses of the word. This follows the hypothesis of (Miller and Charles, 1991) that words that occur in similar contexts will have similar meanings.

We have shown that these methods can be extended to proper name discrimination (Pedersen et al., 2005). People, places, or companies often share the same name, and this can cause a considerable amount of confusion when carrying out Web search or other information retrieval applications. Name discrimination seeks to group together the contexts that refer to a unique underlying individual, and allow the user to recognize that the same name is being used to refer to multiple entities.

We have also extended SenseClusters to cluster contexts that are not centered around any target word, which we refer to as *headless clustering*. Automatic email categorization is an example of a headless clustering task, since each message can be considered a context. SenseClusters will group together messages if they are similar in content, without requiring that they share any particular target word between them.

We are also addressing a well known limitation to unsupervised clustering approaches. After clustering contexts, it is often difficult to determine what underlying concepts or entities each cluster represents without manually inspecting their contents. Therefore, we are developing methods that automatically assign *descriptive* and *discriminating* labels to each discovered cluster that provide a characterization of the contents of the clusters that a human can easily understand.

2 Clustering Methodology

We begin with the collection of contexts to be clustered, referred to as the test data. These may all include a given target word, or they may be headless contexts. We can select the lexical features from the test data, or from a separate source of data. In either case, the methodology proceeds in exactly the same way.

SenseClusters is based on lexical features, in particular unigrams, bigrams, co–occurrences, and tar-

Proceedings of the ACL Interactive Poster and Demonstration Sessions,
pages 105–108, Ann Arbor, June 2005.

get co–occurrences. Unigrams are single words that occur more than five times, bigrams are ordered pairs of words that may have intervening words between them, while co-occurrences are simply unordered bigrams. Target co-occurrences are those co–occurrences that include the given target word. We select bigrams and co–occurrences that occur more than five times, and that have a log–likelihood ratio of more than 3.841, which signifies a 95% level of certainty that the two words are not independent. We do not allow unigrams to be stop words, and we eliminate any bigram or co–occurrence feature that includes one or more stop words.

Previous work in word sense discrimination has shown that contexts of an ambiguous word can be effectively represented using first order (Pedersen and Bruce, 1997) or second order (Schütze, 1998) representations. SenseClusters provides extensive support for both, and allows for them to be applied in a wider range of problems.

In the first order case, we create a context (rows) by lexical features (columns) matrix, where the features may be any of the above mentioned types. The cell values in this matrix record the frequencies of each feature occurring in the context represented by a given row. Since most lexical features only occur a small number of times (if at all) in each context, the resulting matrix tends to be very sparse and nearly binary. Each row in this matrix forms a vector that represents a context. We can (optionally) use Singular Value Decomposition (SVD) to reduce the dimensionality of this matrix. SVD has the effect of compressing a sparse matrix by combining redundant columns and eliminating noisy ones. This allows the rows to be represented with a smaller number of hopefully more informative columns.

In the second order context representation we start with creating a word by word co-occurrence matrix where each row represent the first word and the columns represent the second word of either bigram or co–occurrence features previously identified. If the features are bigrams then the word matrix is asymmetric whereas for co-occurrences it is symmetric and the rows and columns do not suggest any ordering. In either case, the cell values indicate how often the two words occur together, or contains their log–likelihood score of associativity. This matrix is large and sparse, since most words do not co–occur with each other. We may optionally apply SVD to this co-occurrence matrix to reduce its dimensionality. Each row of this matrix is a vector that represents the given word at the row via its co–occurrence characteristics. We create a second order representation of a context by replacing each word in that context with its associated vector, and then averaging together all these word vectors. This results in a single vector that represents the overall context.

For contexts with target words we can restrict the number of words around the target word that are averaged for the creation of the context vector. In our name discrimination experiments we limit this scope to five words on either side of the target word which is based on the theory that words nearer to the target word are more related to it than the ones that are farther away.

The goal of the second order context representation is to capture indirect relationships between words. For example, if the word *Dictionary* occurs with *Words* but not with *Meanings*, and *Words* occurs with *Meanings*, then the words *Dictionary* and *Meanings* are second order co-occurrences via the first order co-occurrence of *Words*.

In either the first or second order case, once we have each context represented as a vector we proceed with clustering. We employ the hybrid clustering method known as Repeated Bisections, which offers nearly the quality of agglomerative clustering at the speed of partitional clustering.

3 Labeling Methodology

For each discovered cluster, we create a *descriptive* and a *discriminating* label, each of which is made up of some number of bigram features. These are identified by treating the contexts in each cluster as a separate corpora, and applying our bigram feature selection methods as described previously on each of them.

Descriptive labels are the top N bigrams according to the log–likelihood ratio. Our goal is that these labels will provide clues as to the general nature of the contents of a cluster. The *discriminating* labels are any *descriptive* labels for a cluster that are not *descriptive* labels of another cluster. Thus, the *discriminating* label may capture the content that separates one cluster from another and provide a more

Table 1: Name Discrimination (F-measure)

2-Way Name(M);+	MAJ. (N)	O1 k=2	O2 k=2
AAIRLINES(1075); TCRUISE(1075)	50.0 (2150)	**66.6**	58.8
AAIRLINES(3966); HPACKARD(3690)	51.7 (7656)	**61.7**	59.6
BGATES(1981); TCRUISE(1075)	64.8 (3056)	63.4	53.8
BSPEARS(1380); GBUSH(1380)	50.0 (2760)	56.6	**65.8**
3-Way Name (M);+		k=3	k=3
AAIRLINES(2500); HPACKARD(2500); BMW(2500);	33.3 (7500)	41.4	**45.1**
AAIRLINES(1300); HPACKARD(1300); BSPEARS(1300);	33.3 (3900)	**46.0**	**45.3**
BGATES(1075); TCRUISE(1075); GBUSH(1075)	33.3 (3225)	**53.7**	**53.6**

Table 2: Email Categorization (F-measure)

Newsgroup(M);+	MAJ. (N)	O1 k=2	O2 k=2
comp.graphics(389); misc.forsale(390)	50.1 (779)	**61.1**	**63.9**
comp.graphics(389); talk.pol.mideast(376)	50.8 (756)	**73.6**	54.8
rec.motorcycles(398); sci.crypt(396)	50.13 (794)	**83.1**	60.5
rec.sport.hockey(399); soc.relig.christian(398)	50.1 (797)	**77.6**	58.5
sci.electronics(393); soc.relig.christian(398)	50.3 (791)	**67.8**	52.3

detailed level of information.

4 Experimental Data

We evaluate these methods on proper name discrimination and email (newsgroup) categorization.

For name discrimination we use the 700 million word New York Times portion of the English GigaWord corpus as the source of contexts. While there are many ambiguous names in this data, it is difficult to evaluate the results of our approach given the absence of a disambiguated version of the text. Thus, we automatically create ambiguous names by conflating the occurrences associated with two or three relatively unambiguous names into a single obfuscated name.

For example, we combine *Britney Spears* and *George Bush* into an ambiguous name *Britney Bush*, and then see how well SenseClusters is able to create clusters that reflect the true underlying identity of the conflated name.

Our email experiments are based on the 20-NewsGroup Corpus of USENET articles. This is a collection of approximately 20,000 articles that have been taken from 20 different newsgroups. As such they are already classified, but since our methods are unsupervised we ignore this information until it is time to evaluate our approach. We present results that make two way distinctions between selected pairs of newsgroups.

5 Experimental Results and Discussion

Table 1 presents the experimental results for 2-way and 3-way name discrimination experiments, and Table 2 presents results for a 2-way email categorization experiment. The results are reported in terms of the F-measure, which is the harmonic mean of precision and recall.

The first column in both tables indicates the possible names or newgroups, and the number of contexts associated with each. The next column indicates the percentage of the majority class (MAJ.) and count (N) of the total number of contexts for the names or newsgroups. The majority percentage provides a simple baseline for level of performance, as this is the F–measure that would be achieved if every context were simply placed in a single cluster. We refer to this as the unsupervised majority classifier.

The next two columns show the F–measure associated with the order 1 and order 2 representations of context, with all other options being held constant. These experiments used bigram features, SVD was performed as appropriate for each representation, and the method of Repeated Bisections was used for clustering.

Table 3: Cluster Labels (for Table 1)

True Name	Created Labels
CLUSTER 0: AMERICAN AIRLINES	**Flight 11**, **Flight 587**, **Sept 11**, **Trade Center**, **World Trade**, Los Angeles, New York
CLUSTER 1: TOM CRUISE	**Jerry Maguire**, **Mission Impossible**, **Minority Report**, **Tom Cruise**, **Penelope Cruz**, **Nicole Kidman**, **United Airlines**, **Vanilla Sky**, Los Angeles, New York
CLUSTER 0: GEORGE BUSH	**George Bush** , **George W**, **Persian Gulf**, **President**, **U S**, **W Bush**, **former President**, **lifting feeling**, White House
CLUSTER 1: BILL GATES	**Chairman** , **Microsoft** , **Microsoft Chairman**, **co founder**, **News Service**, **operating system**, **chief executive**, White House
CLUSTER 2: TOM CRUISE	**Jerry Maguire**, **Mission Impossible**, **Minority Report**, **Al Gore**, **New York** , **Nicole Kidman**, **Penelope Cruz**, **Vanilla Sky**, **Ronald Reagan**, White House

Finally, note that the number of clusters to be discovered must be provided by the user. In these experiments we have taken the best case approach and asked for a number of clusters equal to that which actually exists. We are currently working to develop methods that will automatically stop at an optimal number of clusters, to avoid setting this value manually.

In general all of our results significantly improve upon the majority classifier, which suggests that the clustering of contexts is successfully discriminating among ambiguous names and uncategorized email.

Table 3 shows the *descriptive* and *discriminating* labels assigned to the 2–way experimental case of *American Airlines* and *Tom Cruise*, as well as the 3–way case of *George Bush*, *Bill Gates* and *Tom Cruise*. The bold face labels are those that serve as both *descriptive* and *discriminating* labels. The fact that most labels serve both roles suggests that the highest ranked bigrams in each cluster were also unique to that cluster. The normal font indicates labels that are only *descriptive*, and are shared between multiple clusters. There are only a few such cases, for example *White House* happens to be a significant bigram in all three of the clusters in the 3–way case. There were no labels that were exclusively *discriminating* in these experiments, suggesting that the clusters are fairly clearly distinguished.

Please note that some labels include unigrams (e.g., *President* for *George Bush*). These are created from bigrams where the other word is the conflated form, which is not included in the labels since it is by definition ambiguous.

6 Acknowledgements

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

References

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, 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.

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