Unsupervised Discrimination and Labeling of Ambiguous Names

Anagha K. Kulkarni
Department of Computer Science
University Of Minnesota
Duluth, MN 55812
kulka020@d.umn.edu
http://senseclusters.sourceforge.net

Abstract

This paper describes adaptations of unsupervised word sense discrimination techniques to the problem of name discrimination. These methods cluster the contexts containing an ambiguous name, such that each cluster refers to a unique underlying person or place. We also present new techniques to assign meaningful labels to the discovered clusters.

1 Introduction

A name assigned to an entity is often thought to be a unique identifier. However this is not always true. We frequently come across multiple people sharing the same name, or cities and towns that have identical names. For example, the top ten results for a Google search of *John Gilbert* return six different individuals: A famous actor from the silent film era, a British painter, a professor of Computer Science, etc. Name ambiguity is relatively common, and makes searching for people, places, or organizations potentially very confusing.

However, in many cases a human can distinguish between the underlying entities associated with an ambiguous name with the help of surrounding context. For example, a human can easily recognize that a document that mentions *Silent Era*, *Silver Screen*, and *The Big Parade* refers to John Gilbert the actor, and not the professor. Thus the neighborhood of the ambiguous name reveals distinguishing features about the underlying entity.

Our approach is based on unsupervised learning from raw text, adapting methods originally proposed by (Purandare and Pedersen, 2004). We do not utilize any manually created examples, knowledge bases, dictionaries, or ontologies in formulating our solution. Our goal is to discriminate among multiple contexts that mention a particular name strictly on the basis of the surrounding contents, and assign meaningful labels to the resulting clusters that identify the underlying entity.

This paper is organized as follows. First, we review related work in name discrimination and cluster labeling. Next we describe our methodology step-by-step and then review our experimental data and results. We conclude with a discussion of our results and outline our plans for future work.

2 Related Work

A number of previous approaches to name discrimination have employed ideas related to context vectors. (Bagga and Baldwin, 1998) proposed a method using the vector space model to disambiguate references to a person, place, or event across multiple documents. Their approach starts by using the CAMP system to find related references within a single document. For example, it might determine that *he* and *the President* refers to *Bill Clinton*. CAMP creates co-reference chains for each entity in a single document, which are then extracted and represented in the vector space model. This model is used to find the similarity among referents, and thereby identify the same referent that occurs in multiple documents.

(Mann and Yarowsky, 2003) take an approach to

name discrimination that incorporates information from the World Wide Web. They propose to use various contextual characteristics that are typically found near and within an ambiguous proper-noun for the purpose of disambiguation. They utilize categorical features (e.g., age, date of birth), familial relationships (e.g., wife, son, daughter) and associations that the entity frequently shows (e.g. country, company, organization). Such biographical information about the entities to be disambiguated is mined from the Web using a bootstrapping method. The Web pages containing the ambiguous name are assigned a vector depending upon the extracted features and then these vectors are grouped using agglomerative clustering.

(Pantel and Ravichandran, 2004) have proposed an algorithm for labeling semantic classes, which can be viewed as a form of cluster. For example, a semantic class may be formed by the words: *grapes, mango, pineapple, orange* and *peach*. Ideally this cluster would be labeled as the semantic class of *fruit*. Each word of the semantic class is represented by a feature vector. Each feature consists of syntactic patterns (like verb-object) in which the word occurs. The similarity between a few features from each cluster is found using point-wise mutual information (PMI) and their average is used to group and rank the clusters to form a grammatical template or signature for the class. Then syntactic relationships such as *Noun like Noun* or *Noun such as Noun* are searched for in the templates to give the cluster an appropriate name label. The output is in the form of a ranked list of concept names for each semantic class.

3 Feature Identification

We start by identifying features from a corpus of text which we refer to as the feature selection data. This data can be the test data, i.e., the contexts to be clustered (each of which contain an occurrence of the ambiguous name) or it may be a separate corpus. The identified features are used to translate each context in the test data to a vector form.

We are exploring the use of bigrams as our feature type. These are lexical features that consist of an ordered pair of words which may occur next to each other, or have one intervening word. We are interested in bigrams since they tend to be less ambiguous and more specific than individual unigrams. In order to reduce the amount of noise in the feature set, we discard all bigrams that occur only once, or that have a log-likelihood ratio of less than 3.841. The latter criteria indicates that the words in the bigram are not independent (i.e., are associated) with 95% certainty. In addition, bigrams in which either word is a stop word are filtered out.

4 Context Representation

We employ both first and second order representations of the contexts to be clustered. The first order representation is a vector that indicates which of the features identified during the feature selection process occur in this context.

The second order context representation is adapted from (Schütze, 1998). First a co-occurrence matrix is constructed from the features identified in the earlier stage, where the rows represent the first word in the bigram, and the columns represent the second word. Each cell contains the value of the log-likelihood ratio for its respective row and column word-pair.

This matrix is both large and sparse, so we use Singular Value Decomposition (SVD) to reduce the dimensionality and smooth the sparsity. SVD has the effect of compressing similar columns together, and then reorganizing the matrix so that the most significant of these columns come first in the matrix. This allows the matrix to be represented more compactly by a smaller number of these compressed columns.

The matrix is reduced by a factor equal to the minimum of 10% of the original columns, or 300. If the original number of columns is less than 3,000 then the matrix is reduced to 10% of the number of columns. If the matrix has greater than 3,000 columns, then it is reduced to 300.

Each row in the resulting matrix is a vector for the word the row represents. For the second order representation, each context in the test data is represented by a vector which is created by averaging the word vectors for all the words in the context.

The philosophy behind the second order representation is that it captures indirect relationships between bigrams which cannot be done using the

first order representation. For example if the word *ergonomics* occurs along with *science*, and *workplace* occurs with *science*, but not with *ergonomics*, then *workplace* and *ergonomics* are second order co-occurrences by virtue of their respective co-occurrences with *science*.

Once the context is represented by either a first order or a second order vector, then clustering can follow. A hybrid method known as Repeated Bisections is employed, which tries to balance the quality of agglomerative clustering with the speed of partitional methods. In our current approach the number of clusters to be discovered must be specified. Making it possible to automatically identify the number of clusters is one of our high priorities for future work.

5 Labeling

Once the clusters are created, we assign each cluster a *descriptive* and *discriminating* label. A label is a list of bigrams that act as a simple summary of the contents of the cluster.

Our current approach for *descriptive* labels is to select the top N bigrams from contexts grouped in a cluster. We use similar techniques as we use for feature identification, except now we apply them on the clustered contexts. In particular, we select the top 5 or 10 bigrams as ranked by the log-likelihood ratio. We discard bigrams if either of the words is a stop-word, or if the bigram occurs only one time. For *discriminating* labels we pick the top 5 or 10 bigrams which are unique to the cluster and thus capture the contents that separates one cluster from another.

6 Experimental Data

Our experimental data consists of two or more unambiguous names whose occurrences in a corpus have been conflated in order to create ambiguity. These conflated forms are sometimes known as pseudo words. For example, we take all occurrences of Tony Blair and Bill Clinton and conflate them into a single name that we then attempt to discriminate.

Further, we believe that the use of artificial pseudo words is suitable for the problem of name discrimination, perhaps more so than is the case in word sense disambiguation in general. For words there is always a debate as to what constitutes a word sense, and how finely drawn a sense distinction should be made. However, when given an ambiguous name there are distinct underlying entities associated with that name, so evaluation relative to such true categories is realistic.

Our source of data is the New York Times (January 2000 to June 2002) corpus that is included as a part of the English GigaWord corpus.

In creating the contexts that include our conflated names, we retain 25 words of text to the left and also to the right of the ambiguous conflated name. We also preserve the original names in a separate tag for the evaluation stage.

We have created three levels of ambiguity: 2-way, 3-way, and 4-way. In each of the three categories we have 3-4 examples that represent a variety of different degrees of ambiguity. We have created several examples of intra-category disambiguation, including Bill Clinton and Tony Blair (political leaders), and Mexico and India (countries). We also have inter-category disambiguation such as Bayer, Bank of America, and John Grisham (two companies and an author).

The 3-way examples have been chosen by adding one more dimension to the 2-way examples. For example, Ehud Barak is added to Bill Clinton and Tony Blair, and the 4-way examples are selected on similar lines.

7 Experimental Results

Table 1 summarizes the results of our experiments in terms of the F-Measure, which is the harmonic mean of precision and recall. Precision is the percentage of contexts clustered correctly out of those that were attempted. Recall is the percentage of contexts clustered correctly out of the total number of contexts given.

The variable M in Table 1 shows the number of contexts of that target name in the input data. Note that we divide the total input data into equal-sized test and feature selection files, so the number of feature selection and test contexts is half of what is shown, with approximately the same distribution of names. (N) specifies the total number of contexts in the input data. MAJ. represents the percentage of the majority name in the data as a whole, and can be viewed as a baseline measure of performance that

Table 1: Experimental Results (F-measure)

Target Word(M);+	MAJ. (N)	K	Order 1		Order 2			
			FSD	TST	FSD	FSD/S	TST	TST/S
BAYER(1271);	60.0	2	67.2	68.6	**71.0**	51.3	**69.2**	53.2
BOAMERICA(846)	(2117)	6	37.4	33.9	47.2	**53.3**	42.8	49.6
BCLINTON(1900);	50.0	2	82.2	**87.6**	81.1	81.2	81.2	70.3
TBLAIR(1900)	(3800)	6	58.5	61.6	61.8	**71.4**	61.5	**72.3**
MEXICO(1500);	50.0	2	42.3	52.4	52.7	**54.5**	52.6	**54.5**
INDIA(1500)	(3000)	6	28.4	36.6	37.5	49.0	37.9	**52.4**
THANKS(817);	55.6	2	61.2	**65.3**	61.4	56.7	61.4	56.7
RCROWE(652)	(1469)	6	36.3	41.2	38.5	**52.0**	39.9	47.8
BAYER(1271);BOAMERICA(846);	43.2	3	69.7	**73.7**	57.1	54.7	55.1	54.7
JGRISHAM(828);	(2945)	6	31.5	38.4	32.7	**53.1**	32.8	**52.8**
BCLINTON(1900);TBLAIR(1900);	33.3	3	51.4	**56.4**	47.7	44.8	47.7	44.9
EBARAK(1900);	(5700)	6	**58.0**	54.1	43.8	48.1	43.7	48.1
MEXICO(1500);INDIA(1500);	33.3	3	**40.4**	**41.7**	38.1	36.5	38.2	37.4
CALIFORNIA(1500)	(4500)	6	31.5	**38.4**	32.7	**36.2**	32.8	**36.2**
THANKS(817);RCROWE(652);	35.4	4	42.7	**61.5**	42.9	38.5	42.7	37.6
BAYER(1271);BOAMERICA(846)	(3586)	6	47.0	**53.0**	43.9	34.0	43.5	34.6
BCLINTON(1900);TBLAIR(1900);	25.0	4	48.4	**52.3**	44.2	**50.1**	44.7	**51.4**
EBARAK(1900);VPUTIN(1900)	(7600)	6	**51.8**	47.8	43.4	**49.3**	44.4	**50.6**
MEXICO(1500);INDIA(1500);	25.0	4	**34.4**	**35.7**	29.2	27.4	29.2	27.1
CALIFORNIA(1500);PERU(1500)	(6000)	6	**31.3**	**32.0**	27.3	27.2	27.2	27.2

Table 2: Sense Assignment Matrix (2-way)

	TBlair	BClinton	
C0	784	50	834
C1	139	845	984
	923	895	1818

Table 3: Sense Assignment Matrix (3-way)

	BClinton	TBlair	EBarak	
C0	617	57	30	704
C1	65	613	558	1236
C2	215	262	356	833
	897	932	944	2773

would be achieved if all the contexts to be clustered were placed in a single cluster.

K is the number of clusters that the method will attempt to classify the contexts into. FSD are the experiments where a separate set of data is used as the feature selection data. TST are the experiments where the features are extracted from the test data. For FSD and TST experiments, the complete context was used to create the context vector to be clustered, whereas for FSD/S and TST/S in the order 2 experiments, only the five words on either side of the target name are averaged to form the context-vector.

For each name conflated sample we evaluate our methods by setting K to the exact number of clusters, and then for 6 clusters. The motivation for the higher value is to see how well the method performs when the exact number of clusters is unknown. Our belief is that with an artificially- high number specified, some of the resulting clusters will be nearly empty, and the overall results will still be reasonable. In addition, we have found that the precision of the clusters associated with the known names remains high, while the overall recall is reduced due to the clusters that can not be associated with a name.

To evaluate the performance of the clustering,

Table 4: Labels for Name Discrimination Clusters (found in Table 1)

Original Name	Type	Created Labels
CLUSTER 0: TONY BLAIR	Desc.	Britain, British Prime, Camp David, Middle East, Minister, New York, Prime, Prime Minister, U S, Yasser Arafat
	Disc.	Britain, British Prime, Middle East, Minister, Prime, Prime Minister
CLUSTER 1: BILL CLINTON	Desc.	Al Gore, Ariel Sharon, Camp David, George W, New York, U S, W Bush, White House, prime minister
	Disc.	Al Gore, Ariel Sharon, George W, W Bush
CLUSTER 2: EHUD BARAK	Desc.	Bill Clinton, Camp David, New York, President, U S, White House, Yasser Arafat, York Times, minister, prime minister
	Disc.	Bill Clinton, President, York Times, minister

a contingency matrix (e.g., Table 2 or 3) is constructed. The columns are re-arranged to maximize the sum of the cells along the main diagonal. This re-arranged matrix decides the sense that gets assigned to the cluster.

8 Discussion

The order 2 experiments show that limiting the scope in the test contexts (and thereby creating an averaged vector from a subset of the context) is more effective than using the entire context. This corresponds to the findings of (Pedersen et.al., 2005). The words closest to the target name are most likely to contain identifying information, whereas those that are further away may be more likely to introduce noise.

As the amount and the number of contexts to be clustered (and to be used for feature identification) increases, the order 1 context representation performs better. This is because in the larger samples of data it is more likely to find an exact match for a feature and thereby achieve overall better results. We believe that this is why the order 1 results are generally better for the 3-way and 4-way distinctions, as opposed to the 2-way distinctions. This observation is consistent with earlier findings by Purandare and Pedersen for general English text.

An example of a 2-way clustering is shown in Table 2, where Cluster 0 is assigned to Tony Blair, and Cluster 1 is for Bill Clinton. In this case the precision is 89.60 ((1629/1818)*100), whereas the recall is 85.69 ((1629/1818+83)*100). This suggests that there were 83 contexts that the clustering algorithm was unable to assign, and so they were not clustered and removed from the results.

Table 3 shows the contingency matrix for a 3-way ambiguity. The distribution of contexts in cluster 0 show that the single predominant sense in the cluster is Bill Clinton, but for cluster 1 though the number of contexts indicate clear demarcation between BClinton and TBlair, this distinction gets less clear between TBlair and EBarak. This suggests that perhaps the level of details in the New York Times regarding Bill Clinton and his activities may have been greater than that for the two non-US leaders, although we will continue to analyze results of this nature.

We can see from the labeling results shown in Table 4 that clustering performance affects the quality of cluster labels. Thus the quality of labels for cluster assigned to BClinton and TBlair are more suggestive of the underlying entity than are the labels for EBarak clusters.

9 Future Work

We wish to supplement our cluster labeling technique by using World Wide Web (WWW) based methods (like Google-Sets) for finding words related to the target name and other significant words in the context. This would open up a venue for large and multi-dimensional data. We are cautious though that we would have to deal with the problems of noisy data that WWW brings along with the good data. Another means of improving the clustering labeling will be using WordNet::Similarity to find the relatedness amongst the words from the cluster using the knowledge of WordNet as is also proposed by (McCarthy et.al., 2004).

Currently the number of clusters that the contexts should be grouped into has to be specified by the user. We wish to automate this process such that the clustering algorithm will automatically determine the optimal number of clusters. We are exploring a number of options, including the use of GAP statistic (Tibshirani et.al., 2000).

For the order 2 representation of the contexts there is considerable noise induced in the resulting context vector because of the averaging of all the word-vectors. Currently we reduce the noise in the averaged vector by limiting the word vectors to those associated with words that are located near the target name. We also plan to develop methods that select the words to be included in the averaged vector more carefully, with an emphasis on locating the most content rich words in the context.

Thus far we have tested our methods for one-to-many discrimination. This resolves cases where the same name is used by multiple different people. However, we will also test our techniques for the many-to-one kind ambiguity that occurs when the same person is referred by multiple names, e.g., President Bush, George Bush, Mr. Bush, and President George W. Bush.

Finally, we will also evaluate our method on real data. In particular, we will use the John Smith Corpus as compiled by Bagga and Baldwin, and the name data generated by Mann and Yarowsky for their experiments.

10 Conclusions

We have shown that word sense discrimination techniques can be extended to address the problem of name discrimination. The experiments with second order context representation work better with limited or localized scope. As the dimensionality of the ambiguity increases first order context representation out-performs second order representation. The labeling of clusters using the simple technique of significant bigram selection also shows encouraging results which highly depends on the performance of the clustering of contexts.

11 Acknowledgments

I would like to thank my advisor Dr. Ted Pedersen for his continual guidance and support.

I would also like to thank Dr. James Riehl, Dean of the College of Science and Engineering, and Dr. Carolyn Crouch, Director of Graduate Studies in Computer Science, for awarding funds to partially cover the expenses to attend the Student Research Workshop at ACL 2005.

I am also thankful to Dr. Regina Barzilay and the ACL Student Research Workshop organizers for awarding the travel grant.

This research has been supported by a National Science Foundation Faculty Early CAREER Development Award (#0092784) during the 2004-2005 academic year.

References

Amruta Purandare and Ted Pedersen. 2004. Word sense discrimination by clustering contexts in vector and similarity spaces. *The Proceedings of the Conference on Computational Natural Language Learning*, pages 41-48. Boston, MA.

Gideon Mann and David Yarowsky. 2003. Unsupervised personal name disambiguation. *The Proceedings of the Conference on Computational Natural Language Learning*, pages 33-40. Edmonton, Canada.

Amit Bagga and Breck Baldwin. 1998. Entity–based cross–document co–referencing using the vector space model. *The Proceedings of the 17th international conference on Computational linguistics*, pages 79-85. Montreal, Quebec, Canada.

Patrick Pantel and Deepak Ravichandran. 2004. Automatically Labeling Semantic Classes. *The Proceedings of HLT-NAACL*, pages 321-328. Boston, MA.

Diana McCarthy, Rob Koeling, Julie Weeds and John Carroll. 2004. Finding Predominant Word Senses in Untagged Text. *The Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 279-286. Barcelona, Spain.

Robert Tibshirani, Guenther Walther and Trevor Hastie. 2000. Estimating the number of clusters in a dataset via the Gap statistic. *Journal of the Royal Statistics Society (Series B)*, 2000.

Ted Pedersen, Amruta Purandare and Anagha Kulkarni 2005. Name Discrimination by Clustering Similar Contexts. *The Proceedings of the Sixth International Conference on Intelligent Text Processing and Computational Linguistics*, pages 226-237. Mexico City, Mexico.

Schütze H. 1998. Automatic Word Sense Discrimination *Computational Linguistics*, 24(1):97-124.