Name Discrimination and E-mail Clustering Using Unsupervised Clustering of Similar Contexts

Anagha Kulkarni and Ted Pedersen

Department of Computer Science, University of Minnesota, Duluth, MN 55812, USA

ABSTRACT

In this paper, we apply an unsupervised word-sense discrimination technique based on clustering similar contexts (Purandare & Pedersen, 2004) to the problems of name discrimination and e-mail clustering. Names of people, places, and organizations are not always unique. This can create a problem when we refer to or seek out information about such entities. When this occurs in written text, we show that we can cluster ambiguous names into unique groups by identifying which contexts are similar to each other. It has been previously shown by Pedersen et al. (2005) that this approach can be successfully used for discrimination of names with two-way ambiguity. Here we show that it can be extended to multi-way distinctions as well. On the similar lines of contextual similarity, we also observe that e-mail messages can be treated as contexts, and that in clustering them together we are able to group them based on their underlying topic.

KEYWORDS

word sense discrimination, proper name discrimination, e-mail clustering, unsupervised clustering, contextual similarity

Reprint requests to: Dept of Computer Science, Univ of Minnesota, Duluth, MN 55812, USA; tpederse@d.umn.edu; http://senseclusters.sourceforge.net

1. INTRODUCTION

Humans and systems alike have long encountered the problem of name ambiguity caused by multiple people or places or organizations sharing the same name. With the perpetual growth of World Wide Web, this problem is becoming more and more pervasive and acute. For example, a Google search for the name *George Miller* returns web-pages related to the famous *Psychologist* from *Princeton University* but also returns web pages about an *Australian movie director*. Such unresolved ambiguity can lead to the degradation of information retrieval systems, for instance.

We extend and adapt the methods proposed by (Purandare & Pedersen, 2004) for unsupervised word sense discrimination to this problem of name discrimination. The authors base their approach on the methods proposed by (Schüzte, 1998) and (Pedersen & Bruce, 1997). The philosophy underlying these methods is a hypothesis proposed by Miller and Charles (1991), in which they state that *two words are semantically similar to the extent that their contextual representations are similar*. For instance, all the occurrences of *George Miller* that occur along with *Princeton University* or *WordNet* can be expected to refer to the *Psychologist* whereas the ones co-occurring with *Australia* or *Mad Max* are highly likely to refer to the *Movie Director*.

We also present the preliminary exploration of the e-mail clustering domain by adapting the unsupervised word-sense discrimination methods. The main objective is to be able to cluster a given set of e-mails based upon the overall topic of the e-mail. This approach can be very useful for managing e-mails when we do not want to group or separate e-mails based only upon the presence or absence of a particular string of words but want to cluster them based on the similarity of the underlying topic of thee-mail.

For this work, we adapted and extended an Open Source suite of Perl programs developed by Pedersen and Purandare namely the SenseClusters package.

Related Work

Bagga and Baldwin (1998) proposed a method using the Vector Space Model to disambiguate references to a person, place, or event across

documents. The proposed approach uses their previously developed system CAMP (from the University of Pennsylvania) to find *within document* co-references. 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. The authors 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, familial relationships, and associations that the entity frequently shows. 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.

Gooi and Allan (2004) present a comparison of Bagga and Baldwin's approach to two variations of their own. The authors use the John Smith Corpus and create their own corpus which is called the Person-X corpus. Gooi and Allan re-implement Bagga and Baldwin's context vector approach, and compare it with another context vector approach that groups vectors together using agglomerative clustering. The authors also group instances together based on the Kullback-Liebler Divergence. Their conclusion is that the agglomerative clustering works particularly well.

Some research has been conducted on automatically organizing e-mail based on topic or category. However, many of these techniques use supervised learning, which requires an existing pool of labeled examples to serve as training data, and the learned model is limited to assigning incoming e-mail to an existing category. For example, Bekkerman et al. (2004) propose a supervised approach for categorizing e-mails into predefined folders. The authors apply Maximum Entropy, Naive Bayes, Support Vector Machine (SVM), and Wide-Margin Winnow classifiers to the Enron and SRI[1] e-mail corpora. In their results, although the SVMs achieved the best accuracy, most

[1] http://www.ai.sri.com/project/CALO

of the times, their Wide-Margin Winnow classifier compared fairly well given its simplicity and speed.

Kushmerick and Lau (2005) automate e-mail management based on the structured activities that occur via e-mail. The authors use finite-state automata to formalize this problem, where the states of the automata are the status of the process and the transitions are the e-mail messages. Kushmerick and Lau divide the problem into four tasks of Activity Identification, Transition Identification, Automaton Induction, and Message Classification. They report the results in terms of the accuracy (86% to 97%) with which the methods were able to predict—the next state, the end of activity, and the overlap between the predicted and correct transition message.

METHODOLOGY

Our methodology consists of the following three main phases: feature identification, context representation, and clustering. Next, we describe each step of our methodology in detail.

Feature Identification

SenseClusters supports four different types of lexical features, and each one of them captures slightly different information from the others. The supported features are unigrams, bigrams, co-occurrences, and target co-occurrences. Unigrams are individual words that occur in the corpus more often than some cut-off frequency. Bigrams are ordered pairs of words, while co-occurrences are unordered word-pairs. The word-pairs can optionally have intervening words between them in the actual corpus, which are ignored while forming the bigram/co-occurrence. For bigrams, by retaining the order of occurrence of the words, we expect to capture phrases or collocations whereas with co-occurrences, we identify the word-pairs that tend to occur together. Finally, the target co-occurrences are unordered pairs of words where one of the words is the ambiguous word (target-word). This feature is based on the reasoning that the words near to the ambiguous word are more related to it than are the words that are farther away. These features can be

specified to be selected from the test data or from a separate set of raw data, which is not clustered but is used only for feature identification and is thus referred to as the *feature selection corpus*. In either case, there is no manually annotated information about the underlying entity of ambiguous names or the correct clustering of e-mail messages available. For the filtering of features, we use either rank-cutoffs or statistical tests of association like log-likelihood ratio, mutual information (MI), point-wise MI, etc.

Context Representation

Once we have the set of identified lexical features, we proceed to represent each context in terms of these features either directly or indirectly. The contexts can be represented using either first-order or second-order context representation. The first-order representation is based upon the technique that Pedersen & Bruce (1997) adopt. In this representation, we create a matrix with each context representing a row, each identified feature representing a column and each cell representing the frequency of occurrence of the feature (column) in the context (row). We typically use Singular Value Decomposition (SVD) to reduce the dimensionality of this matrix. Each row of this reduced matrix can be now looked at as a reduced vector representing the context at the row. Thus, the matrix translates into context vectors at each row of the matrix which are later clustered.

Alternatively, one can use the second order context representation which is adapted from Schüzte (1998). We start by representing the identified bigram or co-occurrence features in a word by word matrix format where the first word of the feature is represented across the row, the second word across the column, and the cell values are either their co-occurrence frequencies or the statistical scores of test of associativity. Note that this matrix does not incorporate any information from test data in it. SVD is employed to this matrix for dimensionality reduction and smoothing of values. Each row of this reduced matrix can be interpreted as the word vector for the word it represents. This word vector carries the information about the co-occurrence pattern of the word, that is, it gives the information about which other words co-occur with the word at the row. These word-vectors are used to create the context vector representation. A context vector is created by averaging the

word-vectors for the words that occur in the context. The motivation behind this context representation is to capture direct as well as indirect relations among words. For example, if the word *ergonomics* occurs with *work-place* and also with *science,* then *work-place* and *science* will be indirectly related by the virtue of *ergonomics*.

For name discrimination problems, we can restrict the words around the target-word that would contribute toward the formation of the context vector. Once the contexts are represented in vector format, clustering follows.

Clustering. Here the context vectors created in the previous phase are clustered into different clusters based on similarity/dissimilarity among them. Various different types of clustering methods can be used. However, for our domain we have found that partitional methods that produce hierarchical clustering solutions using repeated bisections usually give the best results. SenseClusters integrates CLUTO[2], a suite of clustering algorithms to provide clustering functionality.

Cluster Labeling. We try to address the commonly faced problem of identifying the underlying entity that a cluster represents without having to manually examine the cluster contents. We achieve this by assigning a label to each discovered cluster. The labels are classified into two types specifically Descriptive and Discriminating. The Descriptive labels are the top N bigrams from the contents of the cluster and Discriminating labels are the top N bigrams from the contents of the cluster that are unique to the cluster. The Descriptive labels capture the main concept or entity of the cluster while the Discriminating clusters highlight the distinctive characteristics of the cluster.

Test Data

For the name discrimination problem, only a very few well-known sets of data, like the John Smith corpus compiled by Bagga and Baldwin and the name data by Mann and Yarowsky, can be used. Thus to overcome this deficit of test data, we create a test dataset by conflating two or more unambiguous names. For example, we replace all the occurrences of *Tony Blair* and *Vladimir Putin* in a given corpus with *BlairPutin* and thus create artificial ambiguity.

[2] http://www-users.cs.umn.edu/~karypis/cluto

Relatively more data is available for e-mail clustering experiments. Resources like 20 NewsGroup Data[3] and Enron Data[4] are widely used for research purposes.

Experimental Data

For the experiments, we have used the data for which the true entities in case of the name discrimination problem and the groups in case of the e-mail clustering problem are already known so that we can automatically evaluate our methods and do not have to rely on manual evaluation. This information is strictly ignored until the evaluation stage.

Name Discrimination Data

For the name discrimination experiments, we used Agence France Press English Service (AFE) portion of the English GigaWords Corpus, as distributed by the Linguistic Data Consortium. In particular, we use the corpus with 234,162,179 words that had appeared as AFE newswire data from January 2002 to June 2002. We categorized the experiments into two types namely 2-way and 3-way experiments. In the 2-way experiments, we conflate 2 unambiguous names whereas for 3-way we conflate three names.

E-mail Clustering Data

We used the 20 NewsGroup corpus of USENET articles for the e-mail clustering experiments. The 20 NewsGroup corpus consists of approximately 20,000 articles already classified into 20 categories such as *computer graphics*, *recreational motorcycles*, *science electronics,* and so on. We ignore this categorization information and use it only in the evaluation stage. Similar to name discrimination we have categorized these experiments into 2-way and 3-way categories.

[3] http://people.csail.mit.edu/jrennie/20Newsgroups
[4] http://www.cs.cmu.edu/~enron

Experimental Setup

We use features of type bigram in all the experiments. Bigrams capture more information than the unigrams and are less restrictive than the target co-occurrences that mandate the word-pairs to contain the target-word. We use the log-likelihood ratio with cutoff of 3.841 for ranking the bigrams according to the associativity between the two words of the bigram. The cutoff of 3.841 signifies 95% certainty that the two words in the bigrams are not occurring together just by chance. The bigrams that occur less than five times in the corpus are ignored. We also employ OR stop-listing which means that if either word in a bigram is a stop word, then the bigram is filtered out. After enforcing all these elimination rules, what we are left with is a rich set of features. We experiment with both the context representations—the first-order and the second-order context representation, for all the experiments. We performed Singular Value Decomposition on all the experiments. We restricted the scope for the second-order name discrimination experiments to five words on either sides of the target-word. By doing so, we restrict the number of word-vectors that will be averaged to generate the context vector to ten word-vectors. Each experiment is performed once with number of clusters set to actual number of entities or news groups present in the data and once with number of clusters set to six. The theory behind setting the cluster number to artificially high value is to test the method in the situation where the user does not know how many entities are present in the test data. The expectation in such cases where the test data has n real entities but the number of clusters specified is c (where c>n) is that finally even if c number of clusters are generated only n of them would be populated and the c-n clusters would be empty and thus can be ignored.

EXPERIMENTAL RESULTS

The results for the experiment are specified in terms of F measure, which is a harmonic mean of Precision (P) and Recall (R) values. The Precision value is the percentage of contexts clustered correctly out of those that were attempted, while Recall is the percentage of contexts clustered correctly out

TABLE 1

Experimental Results for name discrimination in terms of F measure.

Target Word	M	MAJ. (N)	K	Order1	Order2
Tony Blair & Vladimir Putin	1436 1788	55.45 (3224)	2 6	94.88 61.44	96.22 76.17
Mexico & Uganda	1256 1256	50.00 (2512)	2 6	60.11 51.37	59.16 51.89
Microsoft & Compaq	380 380	50.00 (760)	2 6	68.42 54.37	70.26 57.57
Serena Williams & Tiger Woods	308 291	51.41 (599)	2 6	53.09 51.23	68.95 63.39
Sonia Gandhi & Leonid Kuchma	112 110	50.45 (222)	2 6	89.15 60.12	91.03 54.37
Tony Blair, Vladimir Putin & Saddam Hussein	1436 1788 1048	41.85 (4272)	3 6	72.66 62.23	75.68 67.31
Mexico, Uganda & India	1256 1256 1256	33.34 (3768)	3 6	44.75 37.66	46.44 45.25
Microsoft, Compaq & Serena Williams	380 380 380	33.34 (1140)	3 6	51.95 56.62	52.60 52.08

M = number of instances per original category; MAJ = majority classifier F-measure; N = total instances for conflated categories (sum of M); K = number of clusters; Order 1 = F-measure for first order method; Order 2 = F-measure for second order method

of the total number of contexts. The F measure values can be interpreted as the agreement between the clustering proposed by the methods and that proposed by the answer key of the dataset.

The baseline used in these experiments is the Majority Classifier, which is calculated by clustering all the contexts in a single cluster and then calculating the Precision, Recall and F measure for this single cluster. In other words this baseline specifies the result that one can expect without clustering the contexts at all which is the lowest threshold for effectiveness of any method.

Table 1 summarizes the experimental results for name discrimination. The first column of the table specifies the original names that were conflated. The next column (M) gives the count of the contexts per word in the test data,

that is, the test data for *Tony Blair* and *Vladimir Putin* contains 1436 contexts about *Tony Blair* and 1788 contexts about *Vladimir Putin*. Thus, this column shows the distribution of the names in the test data. The third column specifies two values - the first one (MAJ.) is the majority classifier for the experiment which is the baseline and the next (N) is the sum of earlier column which is the total number of contexts in the test data. The fourth column shows the number of clusters we sought to discover (K) for the experiment. The next to last column (Order 1) and the last column (Order 2) indicates the results for experiment with first order and second order context representation, respectively. Table 2 summarizes the results for the e-mail clustering experiments and the column heading interpretation is same as that for Table 1.

TABLE 2

Experimental Results for email clustering in terms of F measure.

NewsGroup	M	MAJ. (N)	K	Order1	Order2
Comp.Graphics &	584	50.04	2	50.90	62.19
Misc.ForSale	585	(1169)	6	34.49	41.37
Comp.Graphics &	584	50.87	2	68.82	57.06
Talk.Politics.MidEast	564	(1148)	6	41.23	47.81
Rec.Motorcycles &	598	50.12	2	61.53	62.70
Sci.Crypt	595	(1193)	6	44.27	42.54
Rec.Sport.Hockey &	600	50.04	2	55.55	63.14
Soc.Religion.Christian	599	(1199)	6	40.71	41.29
Sci.Electronics &	591	50.33	2	50.25	54.87
Soc.Religion.Christian	599	(1190)	6	38.18	46.85
Comp.Graphics,	584	33.70	3	36.69	39.84
Rec.Autos &	594	(1777)	6	37.07	34.76
Soc.Religion.Christian	599				
Misc.ForSale, Sci.Med	585	33.63	3	40.15	40.20
& Rec.Sport.Baseball	594	(1775)	6	33.43	40.08
	597				
Talk.Politics.Guns,	546	35.80	3	40.06	38.35
Talk.Politics.MidEast	564	(1575)	6	35.26	30.58
& Talk.Politics.Misc	465				

The labels assigned to the clusters of Tony Blair, Vladimir Putin and Saddam Hussein for the 3-way name discrimination experiment, are shown in Table 3. The bigrams in bold face indicate those word-pairs that were selected as Descriptive as well as Discriminating labels for that cluster and bigrams in normal font face are the Descriptive labels. As we can see majority of the labels are Descriptive as well as Discriminating. This is expected and in fact indirectly indicates the effectiveness of the clustering algorithm because if the clustering algorithm is able to separate the contexts correctly then the contents of the clusters that are unique to it should be the ones which also are commonly occurring in it. In short overlapped Descriptive and Discriminating labels indicate that the identified clusters are clearly distinct.

DISCUSSION

As we can see from the Table 1 almost all the results are above the baseline, especially for the experiments where the number of clusters specified is equal to the actual senses in the test data and is not an artificially high value.

Another trend that can be clearly seen from Table 1 results is that not just people names but particularly names of personalities related to politics are disambiguated more effectively. By that we are referring to the 2-way experiments about *Tony Blair*, *Vladimir Putin* and *Sonia Gandhi*, *Leonid. Kuchma*, which show a remarkable increase in the F measure over the baseline. But at the same time, the *Serena Williams* and *Tiger Woods* experiment, although about names of personalities which are often discussed in news paper, does not perform as well as we would expect. This trend is driven by the nature of the data used for feature identification. The data used for feature identification is a newswire data that usually contains political articles in much more details than article on any other topic. Thus, contexts containing politics related words generate richer context vectors by virtue of richer feature set. Had we used a corpus compiled from some sports magazines, we would expect the sport related features to be much stronger than any other features.

In almost all the experiments for which the results are significantly above the majority classifier, the second-order vector representation outperforms the

first-order representation, which can be attributed to its ability of capturing the direct as well as the indirect relationships.

Although the results for some of the e-mail clustering experiments are above the majority classifier, they are not as good as the name-discrimination results, for several reasons.

- First of all, the words contributing toward the creation of context vectors cannot be restricted to the words nearer to the target-word because we do not have a target-word to be discriminated and as a result, a high amount of noise gets induced in the context vectors.
- The next important factor is the effect of the writing styles for e-mails versus news paper article. Newspaper articles have a more organized and defined writing style that reflects in the vocabulary being used consistently, the tense being used and the active or passive phrasing of sentences.

Consistency in all these factors in the case of newspaper articles helps build rich features. E-mails, on the other hand, tend to be more informal and loose with the possible occurrences of slang and regional and very domain-specific vocabulary that is not widely known and used. This results in a very large set of weak features. Currently we do not filter out the e-mail headers and as a result, e-mail-stoplist words like *Subject*, *Reply* etc. are not removed from the set of features. These words occur in all the e-mails and thus have heavy and highly skewed vectors. We plan to create a separate e-mail specific stoplist that would filter out all such words.

Table 3 shows that even the simple scheme of labeling the clusters with significant bigrams helps in identifying the underlying entity for a cluster.The top 3 Descriptive and Discriminating labels assigned to the Cluster-0 of the 2-way experiment, shown in Table 3, *British Prime, BlairPutin Minister* and *Downing Street* clearly suggests that the true underlying entity for this cluster has to be *Tony Blair* and not *Vladimir Putin*. For the 3-way experiment, we can see that although Cluster-0 can be confidently assigned to *Tony Blair*, the next two clusters cannot be easily identified based on the labels assigned, suggesting that the clusters for *Vladimir Putin* and *Saddam Hussein* are not very pure

TABLE 3

Cluster labels for the 2-way experiment - *Tony Blair* and *Vladimir Putin,* for the 3-way experiment - *Tony Blair, Vladimir Putin and Saddam Hussein*

True Name	Created Labels
Cluster 0: Tony Blair	**British Prime, Minister, Downing Street, Middle East, President George, words moved,** George W, Prime Minister, United States, W Bush
Cluster 1: Vladimir Putin	**Cold War, President, Russian President, Saint-Petersburg, TV 6, news agency,** George W, Prime Minister, United States, W Bush
Cluster 0: Tony Blair	**British Prime, Minister, Downing Street, Middle East, words moved,** President George, George W, Prime Minister, United States, W Bush
Cluster 1: Vladimir Putin	**Iraqi President, President, Russian President, TV 6, news agency,** George W, Prime Minister, United States, W Bush
Cluster 2: Saddam Hussein	**Iraqi leader, Russian counterpart, US President, counterpart, leader,** George W, President George, Prime-Minister, United States, W Bush

CONCLUSION

We have shown in this paper that the Word Sense Discrimination techniques proposed by Purandare and Pedersen (2004) can be effectively applied to the problems of name discrimination and e-mail clustering. The results obtained with second-order context representation out-perform the results obtained using first-order context representation. The simple yet elegant cluster labeling technique helps identify the underlying entity that a cluster represents via the Descriptive and Discriminating labels.

ACKNOWLEDGMENT

This research was supported by a National Science Foundation Faculty Early CAREER Development Award (#0092784). All experiments were done with SenseClusters-v0.69.

REFERENCES

Kushmerick N. and Lau T. 2005. Automated email activity management: an unsupervised learning approach, *International Conference on Intelligent User Interfaces*, 67-74.

Bekkerman R., McCallum A. and Huang G. 2004. Automatic Categorization of Email into Folders: Benchmark Experiments on Enron and SRI Corpora, *Center for Intelligent Information Retrieval, Technical Report IR-418.*

Gooi C. and Allan J. 2004. Cross-document co-reference on a large scale corpus, *The Proceedings of HLT-NAACL*, 9-16.

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

Pedersen T., Purandare A. and Kulkarni A. 2005. Name discrimination by Clustering Similar Contexts, *The Proceedings of the Sixth International Conference on Intelligent Text Processing and Computational Linguistics*, 226-237.

Mann G. and Yarowsky D. 2003. Unsupervised personal name disambiguation, *The Proceedings of the Conference on Computational Natural Language Learning*, 33-40.

Bagga A. and Baldwin B. 1998. Entity-based cross-document co-referencing using the vector space model, *The Proceedings of the Seventeenth International Conference on Computational Linguistics*, 79-85.

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

Pedersen T. and Bruce R. 1997. Distinguishing word senses in untagged text. *The Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, 197-207.

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