

Thesis Proposal

Unsupervised Discrimination and Labeling of Ambiguous Names in Large Corpora

Author: Anagha K Kulkarni

Advisor: Dr. Ted Pedersen

1) Provide a brief but complete statement of the research problem to be solved. In a separate paragraph, specify the goals of this research.

The research problem to be solved is how to discriminate between various references in a large sample of text to people sharing the same name and then to label these disambiguated names by their distinguishing characteristics. The problem is to be solved by unsupervised methods, which means not using pre-compiled knowledge bases or training data.

The Name Discrimination problem has two aspects to it. The first aspect can be referred to as *one-to-many* association. The object of this task is to classify references about different people sharing the same name into different clusters. Thus we want to classify this sentence, “Do you believe in Gandhi’s philosophy of non-violence?” into a cluster associated with ‘Mahatma Gandhi’ and not ‘Indira Gandhi’ or ‘Rajiv Gandhi’.

The problem of ambiguous names arises commonly in web searches or document categorization. For example, a Google search for ‘George Miller’ results in 341,000 hits (as of 11/17/2004) of which the top ten links refer to five different George Millers.

The second aspect of the Name Discrimination problem can be referred to as *many-to-one* association. This task is to identify that multiple names are being used for the same person. For example, President Clinton, Bill Clinton, Mr. Clinton, President Bill Clinton, all likely refer to the same person.

The next important task will be to label the clusters created by the previous

step. These labels will be a short description that identifies each cluster by its distinguishing characteristics. For example, a cluster for George H. W. Bush would be ideally labeled as ‘the 41st president of USA’, or that for Mahatma Gandhi as ‘Indian Humanitarian’.

The contents of the clusters would be used to analyze and extract unique characteristics about the identity using Ngram Statistics Package (A package developed at University of Minnesota, Duluth) and CLUTO (A package developed at University of Minnesota, Twin Cities).

The specific goals of this research are...

- To create baseline performance measures for one-to-many Name Discrimination results obtained by extending the clustering techniques developed by Purandare and Pedersen [5].
- Label the clusters by applying statistical methods and text summarization methods to the contents of the clusters. External knowledge sources like the WWW or WordNet will also be tapped for extracting distinguishing characteristics needed for the label.
- Extending and evaluating the one-to-many Name Discrimination approach for the many-to-one aspect.
- Adapting the approaches and implementations to issues that arise when dealing with very large corpora of more than 250,000 different instances of the one-to-many or many-to-one Name Discrimination problem.
- Automatically deciding the number of clusters depending upon the input corpora. A statistical method proposed by Tibshirani [?], could be one approach to handling this problem.

2) Provide a descriptive, one or two paragraph review, which sets the framework for the research and cites recent references which establish the importance of the work or problem to be solved. (Three to five references from recent publications are sufficient.)

Mann and Yarowsky [3] have proposed an approach for disambiguating personal names with multiple real referents using unsupervised clustering technique over a rich feature space of biographic facts, like birth-date, birth-place, occupation, relatives, collegiate information etc. A seed fact pair (e.g., Mozart, 1776), is queried on the web and the sentences returned as search results are used to generate the patterns which are then used to extract the biographical information from the data. Once these features are extracted clustering follows. The clustering method used is the bottom-up centroid agglomerative clustering. In this method each document is assigned a vector of extracted features. At each stage of clustering two most similar vectors are merged to produce a new cluster. This step is repeated until all the documents or references to be disambiguated get clustered.

Landauer and Dumais [2] developed a mathematical/statistical technique called Latent Semantic Analysis (LSA), for extracting the meaning or sense of a word in a given context and not just as an independent word in a large corpus of text. The motivation was that the similarity between meanings of words can be judged more precisely by looking at the context around the word. LSA does not use any kind of pre-compiled knowledge base or dictionary. It represents the text in form of word by context co-occurrence matrix. Each row represents unique word and each column represents text passage or any form of context. The cell values are the frequency counts of the word occurring in the context. These cell values maybe be subjected to transformations (like taking log of the values, computing entropy, etc.), which brings out the word's importance in the particular context and the degree to which the word type carries information in the domain of the text. Singular Value Decomposition (SVD) is applied next to the matrix to reduce the dimensionality of the matrix. An optimal dimensionality reduction should cause noise reduction but at the same time no loss of vital information and thus to bring to surface the underlying relations. Finally the cosine measure is used to find the similarity between words. Thus LSA provides a framework, which has many implementation variants. It has proved to be good at synonym tests, solving multiple choice questions, essay grading, information retrieval and various other applications. We will see next an implementation of LSA used by a clustering technique along with the traditional form of clustering.

Purandare and Pedersen [5] developed an unsupervised clustering technique

which clusters instances from the corpus using both vector and similarity spaces. This is achieved by representing context of each instance as a vector in a high dimensional feature space and then clustering these context vectors directly in vector space and also by finding pairwise similarities. Thus given a target word used in number of different contexts, word sense discrimination groups these instances of target word together by determining which contexts are most similar to each other. Thus the problem reduces to finding classes of similar contexts such that each class represents a single word sense. Authors have incorporated in their technique, two ways of representing the context vector (namely first-order (direct) and second-order (indirect)), three categories of clustering algorithms (hierarchical, partitional and hybrid). For the Name Discrimination problem the target word would be the proper noun to be discriminated and we can expect each class or cluster returned by the clustering technique, to refer to a unique person.

Pantel and Ravichandran [4] have proposed an algorithm for labeling semantic classes. For example, a class may be formed by the words grapes, mango, pineapple, orange, peach, and would be ideally labeled as a semantic class of fruits. In the first phase of algorithm, each word of the semantic class is represented as feature vector. A feature is a syntactic patterns like verb-object, in which the word occurs. To find the relatedness or independence between two words point-wise mutual information is used. The anomalous behavior of PMI for infrequent features is corrected by an additional factor they introduce. Second phase proceeds to preparing a square matrix of few representative words from the class and finding similarity between each pair. Grouping the words and ranking these grouped clusters using the number of member words and the average pairwise similarity between words is the next step. These representatives help to form a grammatical template or signature for the class. 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.

Bagga and Baldwin [1] have 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” coreference. CAMP

creates coreference chains for each entity in the document. The CAMP processed document is given to the SentenceExtractor which extracts sentences related to the entity of current interest from the documents. Thus SentenceExtractor creates summaries about the entity for each document. Then Vector Space Model uses these summaries to find similarity between each reference. Vector Space Model finds weight of the terms occurring in both the summaries being compared currently. Then takes product of these weights for each term and sums this product for all such common terms.

3) List the resources (in terms of software and hardware) required to complete the work. If not available within the department, indicate how/where such resources will be obtained. If only one instance of a required resource is available, indicate how you will account for failure of that resource.

Hardware Requirements: Existing facilities provided in the department are sufficient for the research.

Software Requirements: Existing software facilities provided in the department are sufficient for the research. We propose to use the Perl programming language which is freely available. We also plan to use the Perl::LWP module to interact with the Web which is freely available at CPAN.

Data Requirements: We will need large amount of conflated data for the evaluation purposes. We are using a Perl program written by us to conflate proper nouns from the English Giga Word Corpus. We will also be using the “John Smith” corpus which was created by Bagga and Baldwin. We also plan to use the data created by Mann and Yarowsky. All the mentioned data is available with the department for the research purposes.

4) In a numbered list below, specify each step required to solve the problem along with a specific target date for the completion of that step. (The last two steps should be the write-up of the thesis and the completion of oral exam and colloquium, respectively, along with their scheduled dates.)

1. Fall 2004
 - (a) November 29:
 - i. Prepare and Submit Thesis Proposal.
 - ii. Finish writing the program to create name conflated data from English Giga Word Corpus.
 - (b) December 22:
 - i. Baseline Name Discrimination results. Use various data sources mentioned in the Data Requirement section to evaluate the results obtained by the proposed approach for Name Discrimination.
 - ii. Modify existing performance evaluation methods to adapt to large data (more than 10 number of clusters), by using Munkres algorithm for assignment problem.
2. Spring 2005
 - (a) January 15:

Design and implement method for labeling the clusters. Statistical methods and text summarization will be used on the context data of each reference to extract distinguishing characteristics of the identities.
 - (b) March 31:

Finish the adaptation for many-to-one aspect from one-to-many part of the Name Discrimination Problem.
 - (c) May 31:

Complete the incorporation of extending labeling of clusters using external knowledge sources like WWW and WordNet.
3. Fall 2005
 - (a) October 15:

Finish adapting of labeling methods for many-to-one from one-to-many labeling approaches.
 - (b) November 30:

Complete design and implementation of the approach to identify

the number of Clusters that the instances should be classified into depending upon the corpora.

4. Spring 2006

- (a) January 15:
Evaluate and extend first year approaches for very large data. Test the clustering approach for very large corpora and implement the necessary changes.
- (b) February 28:
Finish evaluation and adaptation of labeling approaches for very large data.
- (c) April 15:
Turn in the Thesis to the review committee.
- (d) May early:
Complete the Thesis Defense.

5) Specify clearly how the results of the research will be evaluated. What objective measures will be used to establish that the goals of the research have been met?

During the course of research we will evaluate our approach at every important stage. As described in the Data Requirement section various manually created corpora are available for evaluation purposes. We will also use the name conflated data created from English Giga Word Corpus using the the NameConflate program. So these corpora will be the gold standard for our experiments and evaluation.

The majority classifier will be the baseline for the performance evaluation. This is a method that simply groups all of the instances of a name into a single cluster, and will have accuracy equal to the percentage of instances that belong to the most common underlying identify.

6) State clearly the contribution to research that this work will make.

As the World Wide Web grows so does the problem of Name Discrimination. This research work will alleviate web users' sufferings by automating the process of reading (partially) through the contents of each returned hit page and categorizing them into labeled classes.

When dealing with proper nouns there is no pre-existing inventory which identifies and classifies underlying possible identities. Our research will not only allow users to identify the underlying entity but would also create an inventory of classified and labeled entities using unsupervised methods.

STUDENT

THESIS ADVISOR

References

- [1] A. Bagga and B. Baldwin. Entity-based cross-document co-referencing using the vector space model. In *Proceedings of the 17th international conference on Computational linguistics*, pages 79–85. Association for Computational Linguistics, 1998.
- [2] T. Landauer and S. Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104:211–240, 1997.
- [3] G. Mann and D. Yarowsky. Unsupervised personal name disambiguation. In W. Daelemans and M. Osborne, editors, *Proceedings of CoNLL-2003*, pages 33–40. Edmonton, Canada, 2003.
- [4] P. Pantel and D. Ravichandran. Automatically labeling semantic classes. In S. Dumais, D. Marcu, and S. Roukos, editors, *HLT-NAACL 2004*:

Main Proceedings, pages 321–328, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for Computational Linguistics.

- [5] 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.
- [6] R. Tibshirani, G. Walther, and T. Hastive. Estimating the number of clusters in a dataset via the Gap statistic. *Journal of the Royal Statistics Society (Series B)*, 2000.