How many different "John Smiths", and who are they?

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

Abstract

In this work we propose three unsupervised measures to automatically identify the number of distinct entities a given ambiguous name refers to in a corpus. We experiment with 22 artificially created name conflations and observe that the measure ($PK2$) formulated as the ratio of two successive clustering criterion function values outperforms the other two measures. We also describe a method to assign a unique label to each discovered cluster so as to identify the underlying entity that it refers to.

Introduction

As the World Wide Web (WWW) grows and the information available online increases, the problem of proper name ambiguity becomes more acute. For example, a Google search for the name *John Smith* comes up with links to 6 different entities (people, organizations or places) sharing the same name - few refer to a captain from 17th century, few to a musician, few to a shop and so forth. Such ambiguity evidently degrades the performance of information retrieval systems, search engines and can potentially lead to confusion.

In our previous work (Pedersen, Purandare, & Kulkarni 2005) we propose an unsupervised method for the name discrimination problem. This is the problem of separating references to an ambiguous name into different clusters without using any training data or any prior knowledge and where each resulting cluster relates to one unique entity. In this method we follow (Schütze 1998) & (Pedersen & Bruce 1997) and appeal to the principle of contextual similarity to perform unsupervised clustering of the various references to an ambiguous name. For example, with the *John Smith* example if we observed words like *soldier*, *colonists* in the surrounding context of the name then one can conclude that such references are highly likely to be related to the captain rather than the musician.

Currently the above described method of name discrimination expects the number of clusters i.e. the number of unique entities for the given ambiguous name to be specified by the user and thus assumes that the user has some prior knowledge about the data and the ambiguous name at hand. However, this might not always be true and thus here we propose 3 measures that each try to predict the number for clusters (k) that a given data naturally separates into for the given ambiguous name. These measures rely exclusively on the information present or inferable from the given data.

Cluster Stopping Measures

We formulate our cluster stopping measures using clustering criterion functions (crfun) which are functions that are used by the clustering algorithms to obtain an optimal clustering solution. The crfun are categorized into three groups namely, internal, external and hybrid. The internal crfun produce a clustering solution that maximizes the similarity among the members of any given cluster. The external crfun minimizes the similarity between any two clusters and the hybrid crfun combine internal and external crfun and thus aim at achieving clustering solution which lead to clusters that are internally tightly bound and externally clearly distinct. We recommend using the hybrid crfun with our measures, although internal crfun can be used with equal ease. For further discussion we will use a hybrid crfun $H2$ which is formulated as: $H2(m) = \frac{I2(m)}{E1(m)}$, where $I2(m)$ is an internal and $E1(m)$ is an external crfun.

One could start with, repeatedly clustering the given data containing N references to the ambiguous name into m clusters, where $m = 1, .., N$ and recording the $H2(m)$ values, however this could be computationally very intense for large N. The plot of $H2$ values over $m = 1, .., N$ looks like a knee i.e. the $H2$ values initially increase as m increases and then the increase slows down and finally plateaus out. We take advantage of this fact to identify the upper bound on k ($deltaK$) beyond which the $H2$ values change negligibly and perform clustering of data only from $m = 1, .., deltaK$.

Our first measure $PK1$ is based on (Mojena 1977). Here we calculate the mean and the standard deviation of $H2(1, .., deltaK)$ values and then compute the $PK1(m)$ scores:

$$PK1(m) = \frac{H2(m) - mean(H2[1...deltaK])}{std(H2[1...deltaK])} \quad (1)$$

A threshold on $PK1(m)$ value has to be set to predict k. The above formulation is very similar to the z-score computation and thus we have hopes that there might be some way of automating the threshold selection. But as yet we have not

identified any strategy to automate the threshold selection and thus for the current experiments we have used -0.7 as the threshold which was empirically established for this study. Although, of the 3 measures, $PK1$ is the only measure that requires the user to set a threshold.

In our second measure $PK2$ we take ratio of two consecutive $H2$ scores to compare the improvement gained by going from $m-1$ clusters to m clusters.

$$PK2(m) = \frac{H2(m)}{H2(m-1)} \quad (2)$$

The m values at which this ratio is approximately 1 are all candidate k values because the ratio value of 1 indicates that the improvement was negligible which implies that the points are on the plateau. To choose one of the candidate k values we calculate the mean and the standard deviation of the $PK2$ scores and pick the m value which has a score that is outside (but is closest) the interval defined by one standard deviation. We adopt this selection procedure to ensure that the selected k value is a point neither on the rising edge of the knee nor on the plateau region but the is right on the knee. $PK2$ is similar in spirit with (Hartigan 1975).

We formulate $PK3$ using 3 consecutive $H2$ scores.

$$PK3(m) = \frac{2 \times H2(m)}{H2(m-1) + H2(m+1)} \quad (3)$$

$PK3$ value of 1 or more indicates that the 3 points are linear, either on the rising edge of the knee or on the plateau but we are interested in the knee point. To select this point we use the similar strategy as in $PK2$ - we calculate the mean and the standard deviation of $PK3$ values and select the m value for which the $PK3$ score is greater than 1 and is closest from outside the interval of one standard deviation.

Experimental Results

For generating experimental data we conflated together unambiguous names so as to artificially create name ambiguity. For example, we replaced all the occurrences of the names *Tony Blair* and *Bill Clinton* with *Blair_Clinton* to create a pseudo name ambiguity. We created 22 such name conflations from the English GigaWord Corpus, 10 name conflations were created by conflating 2 names, 6 with 3 names and 6 with 4 names. Each name conflation was experimented with two different clustering configurations leading to 44 experiments. One of the configuration consists of first order context representation and unigram features with cutoff frequency of 10. While the second configuration uses second order context representation and bigram features with cutoff frequency of 10 and log-likelihood score cutoff of 3.841. Further details about the configurations can be found at (Pedersen, Purandare, & Kulkarni 2005). The value predicted by $PK1$ agreed with the expected value in 15 experiments. $PK3$'s predictions were correct in 24 experiments and $PK2$'s predictions were correct in 31 experiments. All the incorrect predictions made by $PK2$ were off only by +/-1 cluster while those made by $PK1$ and $PK3$ were much more spread-out. All the 3 measures did relatively well with 2 name cases as compared to 3 and 4 name cases.

Cluster Labeling

The solution to the name discrimination problem should not stop at predicting the number of clusters and clustering the data into those many clusters but should also help in identifying the unique underlying entity that each of the clusters represent. For this purpose we assign a label to each of the cluster. Such a label, in the best situation, would specify the underlying entity, for example, the captain *John Smith* could be labeled as *Captain John Smith (1580-1631): Jamestown Leader*. Our name discrimination method currently supports a basic cluster-labeling technique where the labels are generated by selecting the most frequent and most unique word-pairs from the contents of the clusters. We wish to augment the current cluster-labeling strategy with the information from WWW and other structured knowledge resources like WordNet.

Future Work

One of the main directions of our future work will be experimenting with the Gap Statistic (Tibshirani, Walther, & Hastie 2001). This uses *within cluster dispersion (error)* which is inversely related to criterion functions. The main idea of this approach is to standardize the graph of *error* by comparing with the expected graph under appropriate null reference distribution. The adopted null model is the case of single cluster (k=1) which is rejected in favor of k (k>1) if sufficient evidence is present. Another interesting direction is to absorb and make use of user's knowledge of data, if available, into the measures. We also plan on more experiments with name-conflated data and with real data, namely the *John Smith* data compiled by Bagga and Baldwin. For our new set of experiments we plan on experimenting with names that might have higher degree of ambiguity.

Acknowledgments

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

References

Hartigan, J. 1975. *Clustering Algorithms*. NY: Wiley.

Mojena, R. 1977. Hierarchical grouping methods and stopping rules: An evaluation. *The Computer Journal* 20(4):359–363.

Pedersen, T., and Bruce, R. 1997. Distinguishing word senses in untagged text. In *Proceedings of the 2nd Conference on EMNLP*, 197–207.

Pedersen, T.; Purandare, A.; and Kulkarni, A. 2005. Name discrimination by clustering similar contexts. In *Proceedings of the Sixth International Conference on Intelligent Text Processing and Computational Linguistics*, 220–231.

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

Tibshirani, R.; Walther, G.; and Hastie, T. 2001. Estimating the number of clusters in a dataset via the Gap statistic. *Journal of the Royal Statistics Society (Series B)* 63:411–423.