Corpus Based Unsupervised Labeling of Documents

Delip Rao, Deepak P, Deepak Khemani

Department of Computer Science and Engineering
Indian Institute of Technology Madras
delip@cse.iitm.ernet.in , deepakp7@gmail.com , khemani@iitm.ac.in

Abstract

Text categorization involves mapping of documents to a fixed set of labels. A similar but equally important problem is that of assigning labels to large corpora. With a deluge of documents from sources like the World Wide Web, manual labeling by domain experts is prohibitively expensive. The problem of reducing effort in labeling of documents has warranted a lot of investigation in the past. Most of this work involved some kind of supervised or semi-supervised learning. This motivates the need to find automatic methods for annotating documents with labels. In this work we explore a novel method of assigning labels to documents without using any training data. The proposed method uses clustering to build semantically related sets that are used as candidate labels to documents. This technique could be used for labeling large corpora in an unattended fashion.

Introduction

The World Wide Web and the Internet has resulted in an explosion of data sources like web pages, weblogs, newsfeeds and email to name a few. These documents span over a wide range of topics and are very dynamic in nature. Text categorization can be used as a tool to organize and manage this data. The dynamic nature of these data sources make it difficult to define a closed set of labels that could be assigned to documents. One approach to tackle this would be to assign labels using a few important keywords from the document. For example, the articles on weblogs could be labeled so that they are served to a wider audience. The current method optionally makes use of the "tags" feature where labels are manually assigned by the blog writer. Similarly web search results can be categorized for efficient browsing. Traditional, hand labeled techniques for text categorization makes it impossible to handle such copious data. Besides most manual techniques are laborious and error prone. Several methods have been suggested in the past to alleviate the labeling problem. Many of these methods rely on the availability of some kind of training data, building a classifier and using the classifier to further label the unseen data. Training data might be sparse or difficult and expensive to obtain. Given the wide scope of the documents on the web and their dynamic nature it is not possible to rely on a model that has been trained on a single corpus. As (Nigam K. et al., 1998) have noted, the diversification of applications of automatic text categorization makes it difficult create training data for each application area. Attempts have been made to reduce the amount of training data like using a combination of labeled and unlabeled data. We review some of these methods in the following section.

In this work we propose an unsupervised attempt to labeling documents. We use the traditional bag-of-words representation of text and represent each word as a vector which reflects the distribution of words in the different documents. So this method can be readily incorporated in any of the existing IR systems that use the same representation. These word vectors are then clustered using the k-means (McQueen, 1967) clustering algorithm. We draw a set of representative words from each cluster as a label and derive a set of candidate labels. A label from the set of candidate labels is assigned to each document that maximizes the norm of the dot-product of the document vector and the label vector. The method presented here is significantly different from the previous works as it does not require any manual intervention or labels.

Related Work

An entire gamut of machine learning techniques like supervised, semi-supervised and unsupervised learning has been applied at various levels to the task of text categorization. We review some of these techniques and contrast how our work differs from them.

The supervised selection techniques rely on the presence of training data. The training data is usually in the form of a few labeled documents. A classifier is trained from these labeled documents is used for further classifying of unseen documents. Work done by (Lewis & Gale 1994; McCallum & Nigam 1998a) using Naïve Bayes classifiers, (Joachims, 1998) using support vector machines, (Lang, 1995) using classifiers based on the MDL principle, (Koller & Sahami, 1997) using probabilistic models and by InfoSeek using neural networks. Although these methods perform well they require training data which might be difficult to obtain.

The problems with manual labeling resulted in development of semi-supervised techniques by Denis 1998, Blum & Mitchell 1998; Goldman & Zhou, 2000; Nigam et al., 2000; Bockhorst & Craven, 2002; Ghani, 2002; Liu et al, 2002 and Yu, Han & Chang, 2002. These methods are characterized by the use of both labeled and unlabeled data.

The above methods, viz, supervised and semi-supervised learning make use of labeled training documents, although in differing quantities. Our approach differs from these as we do not make use of labeled training data.

(Bing Liu et al, 2004) and (Youngjoong & Jungyan, 2000) use unsupervised methods to text categorization. (Bing Liu et al, 2004) make use of labeled words instead of labeled documents. They expect the user to provide a few "representative words" for each class and use this information along with the clustered results to build a document classifier. Our method differs from this as we do not take any additional input from the users apart from the unlabeled corpus. (Youngjoong & Jungyan, 2000) on the other hand create "training sentence sets" using keyword lists of each category and use them for training and classifying text documents. This scheme, as a part of its preprocessing step, derives features by part-of-speech tagging of the text. We do not make use of such features.

Proposed Approach

Broadly, our approach can be divided into four sequential phases, as below. In the following subsections, we go on to describe each of the different phases in greater detail.

- Clustering of Words to Arrive at Semantically related Word Clusters
- Generating lists of representative words for each Semantically Related Word Cluster
- Tagging documents with clusters
- Building labels for each document from the clusters it is tagged with

Clustering of Words

Given a corpus, the text clustering task usually starts off with building the term document matrix which has as much rows as the number of documents and as much columns as the number of words. Each entry in the matrix indicates the number of times the corresponding word has occurred in the corresponding document. Each row corresponds to the TF (term frequency) vector of the particular document. Further techniques to process the matrix involves normalization of each document vector to add up to a constant, whereby we get the normalized TF (nTF) vector. An additional step of Inverse Document Frequency (IDF) weighting may be incorporated before normalization, whereby we get the normalized TF-IDF vector. Given the TF, nTF or nTFIDF vectors of the documents, clustering is a straightforward task. Having outlined the document clustering task, an analogous method of word clustering is not very difficult to perceive. In the term document matrix, each column corresponds to a term and the transpose can be used as a Document Frequency (DF) vector, whereas a normalized version of the DF vector could analogously be termed the nDF vector. We cluster the set of nDF vectors as defined above using the k-means algorithm. k-means takes the number of clusters (to be generated in the output) as a parameter. Given that we do not have any knowledge of the number of clusters that exist, we tried out different values of k. The values chosen were 10, 20, 50 and 100. The actual k-means clustering was done using the WEKA Toolkit[1] for Data Mining developed by the University of Waikato, New Zealand. The k-means algorithm assigns a word to the cluster, whose centroid (k-means builds convex clusters each of which can be represented by a single centroid in the vector space) it is closest to. It includes an iterative process which refines the centroid at the end of each iteration. Thus, at the end of this phase, we have a set of word clusters, each having its own centroid.

Building Representative Words for Each Cluster

Each cluster built in the previous phase, would be a collection of words, and the results (presented in a later section) confirm our hypothesis that semantically related words would be placed in the same cluster. Given that our aim is to build labels, we would like to concisely represent each cluster using a "manageable" number of words for each cluster. We put forward a hypothesis, which has an intuitive basis and aids us in our aim to build representative words for each cluster.

Hypothesis: The points closest to the cluster center are representative of the cluster.

As this hypothesis is intuitively justifiable to an extent, we choose not to further explain it here. We take the m closest points to the cluster center (for each of the clusters generated) and use them to represent the cluster. We call these words as "Representative Words." If the words thus selected are semantically related to each other then the representative words can be said to be semantically coherent. As we move away from the cluster center, the semantic coherence of the words is expected to decrease. For the purpose of our experiments, we take five words (m=5) closest to the cluster center. All representative words of a cluster are assigned equal weighting in all the subsequent phases irrespective of their distances from the cluster center. Thus, this phase of the approach assigns

[1] http://www.cs.waikato.ac.nz/ml/weka/

representative words for each of the clusters generated in the previous phase.

Tagging Documents with Clusters

Having arrived at a concise representation for each semantically related cluster of words, this phase starts off with assigning a score for each document-cluster pair. The following formula computes the score for the <d,C> pair, where d is a specific document and C is the set of representative words for a specific cluster.

$$\text{Score}(d,C) = \sum_{c \in C} (\text{frequency of } c \text{ in } d)$$

As can be seen, it just computes the sum of frequencies of each word in C, in the document d. This is done for every document in the corpus (used in the first phase) and every word cluster (generated in the first phase). Thus, for each document, we have an array of scores, with one entry per cluster. We choose to map a document to the cluster(s), with whom, it has the highest score, provided the highest score is greater than zero. Thus, if there is a tie and a document has multiple highest scores, all the clusters with the highest scores are taken to tag the document. Further, it may be noted that a document may not be tagged with any clusters, if it has no occurrences of any of the representative words in any cluster. We expect that such a case would be very rare. At the end of this phase, we have each document tagged with clusters. This phase is represented in pseudo-code as shown in Figure 1.

```
Tag_Doc_With_Clusters(Corpus COR, Clusters CL)
{
    for each document d in Corpus COR,
    {
        for each Cluster C in CL,
            $\text{Score}(d,C) = \sum_{c \in C} (\text{frequency of } c \text{ in } d)$
        Clusters to tag d with is defined as
         { C ∈ CL | (Score(d,C) > 0 &&
            Score(d,C) >= Score(d,C1))  ∀ C1 ∈ CL }
    }
}
```

Figure 1. Algorithm to tag documents with clusters

Building Labels for documents

This phase assigns labels (a word or multiple words) to each document in the corpus. The label would always be a subset of the union of the representative words of the clusters with which the document is tagged. We acknowledge that choosing such a limited vocabulary of labels might be too restrictive in some cases. For instance, the document which talks about the Suez Canal may be labeled with Egypt and not Suez because Suez may not among the most representative words for any cluster (Ref. Table 2, Document named TIME068). A cluster may get a high score with a document if one of its representative words occurs very frequently in the document, even if none of the other representative words occur in the document at all. This phase shields against such hostile cases. For each cluster that a document is tagged with, the average number of occurrences (in the document) of words in the representative word-set is computed and all words (among the representative words) which have at least as many occurrences as half of the average so obtained, are added to the label of the document. Consider a hostile case where, among the set of representative words {p, q, r, s, t} for a cluster which occurs among the tags of a document d, word 'p' occurs 100 times in d (inducing a score of 100) and all others do not have any occurrences. Only 'p' would be added to the label as none of the other words have more than 10 occurrences, 20 being the average number of occurrences for that cluster-document pair. Although the algorithm is prone to less hostile cases, our results reaffirm that half the average number of occurrences is a good enough threshold. The algorithm is represented in pseudo-code as shown in Figure2.

```
Build_Label_for_Documents(Corpus COR, Clusters CL)
{
    for each document d in Corpus COR,
    {
        Label(d) = Ø
        for each Cluster C among the tags of d,
        {
            Average_Score(d,C) = Score(d,C)/|C|
            Label(d) = Label(d) ∪
              {c ∈ C | (frequency of c in d) >
                    (0.5*Average_Score(d,C))}
        }
        Output <d,Label(d)>
    }
}
```

Figure 2. Algorithm to Build Labels for Documents

Experiments and Results

Experiment Setup

We chose to test our approach on the Time corpus (Bergler 1990), a popular dataset in the Information Retrieval and Data Mining communities which consists of 423 articles published by the Time magazine during the cold-war period (1960s). The entire dictionary of words in the corpus, after stop-word removal, is of size 20000. We chose not to use the labeled corpuses popular in literature as those corpuses mostly had very abstract labels whereas out approach generated very specific labels. For instance, the article which talks about Indian Prime Minister Jawaharlal Nehru's talks with Pakistan counterparts on the Kashmir issue would most probably, be labeled just "Kashmir" in a labeled corpus, whereas our approach generates "India" , "Pakistan" , "Kashmir" , "Indian" and "Nehru" as labels. Further, manually assigned labels tend to have words not in the document. Just to cite an example, an article on a company buying stakes in another company would most probably be labeled "acquisition" whereas our approach can, at best, come up with "buy", "stakes" among the labels.

Word Clusters and Semantic Coherence

The central idea of our approach is that clustering of nDF vectors would cluster semantically coherent words together and that words closer to the centroid of a cluster are most representative of the cluster (and hence can be used for labeling the documents). In this section, we present results to assert that our assumption does indeed work. We manually verified the semantic coherence using independent knowledge sources such as Google and WikiPedia. We hereby present the table of representative words (Table 1) on clusters gathered using k = 10 (in k-means) and the semantic relationship mined from independent knowledge sources as cited above. Given that our corpus is the set of news articles in the cold war period, we expect to get clusters with words bearing semantic relationship in the context of the Cold War period. Our results show that our assumption is indeed very true.

Labeling of Documents

We proceed to illustrate the labeling of documents that we arrived at using our approach. We present the <document name, extract from document contents, labels, score> triplets for a random sample of the results from our experiments. Due to space constraints, we present the results of our experiments in Table 2 where k was set to 100 in k-means.

Conclusion and Future Work

Firstly, the experiments confirm that paritional clustering of normalized DF vectors does reveal the semantic relationship and groups the semantically coherent words together. Secondly, the experiments testify our idea that words around the cluster center can be used as representative words. Thirdly, the collection of representative words of various clusters, seem to be abstract enough to label documents. This is particularly interesting since, in the course of our experiments, we use a maximum of 500 words (5 each from 100 clusters) for

Table 1. Representative words derived from the clusters

Cluster #	Representative Words (m=5)	Descriptions from Independent Knowledge Sources
0	Damascus, Arabs, Syrian, Egyptian, Jordan	**Syria**, **Egypt** and **Jordan** are **Arab** nations. **Damascus** is the capital of **Syria**
1	time, minister, years, labor, week	**Labor** is a political party which had **ministers** in power during the cold war **years**
2	European, charles, nuclear, market, french	**French** are the peoples of the **European** nation of France
3	lemass, Ireland, irish, Dublin	**Ireland**, whose peoples are called **Irish** has its capital at **Dublin**. Sean **Lemass** was an Irish political leader
4	Saigon, Vietnam, cong, Buddhist, nhu	**Saigon** is district one of ho-chi-min city, the capital of **Vietnam**. Madame **Nhu**, the first lady, was a member of the Viet **Cong**, which had anti-**Buddhist** policies
5	small, including, finally, high, united	
6	Brunei, malay, Malayan, borneo, Singapore	**Malay** is the language spoken by the **Malayan** people and is the official language of Malaysia, **Brunei** and **Singapore**. The Malaysian city of Sabah was called British **Borneo** when it was a British colony
7	constantly, ability, mistakes, endless, aide	
8	peking, red, mao, soviet, communist	**Peking** was the former name of the Beijing, the capital of china where the book called the little **red** book of quotations by **Mao** Zedong was published in 1962. he was trying to drive a wedge between Moscow of **soviet** Russia and Peking of China. Both China and Soviet Russia were **communist** nations.
9	famine, densely, Malthusian, ecological, bachelors	

labeling documents out of the entire dictionary which comprises 20000 words. Thus, we have been successful in reducing the dictionary by 1/40th without any significant loss of words that could be used as a label (as our experiments show). Lastly, our experiments validate the utility of term frequencies as a meaningful and simple statistic in assigning clusters to documents and thus assigning labels to documents.

We have used partitional clustering techniques in the course of our experiments. We would like to use soft clustering techniques where a word can be assigned to more than one cluster to extend this work. As a motivating example (from our experiments) towards the same, "south" is related to Vietnam (as a lot of cold war events are centered around south Vietnam) and to "Africa" (south Africa as a country features in the corpus, although very rarely). We find that "south" has been clustered with "Vietnam" in the same cluster. We would like to devise soft clustering techniques which make use of co-occurence frequencies so that even the slightest semantic relationships (such as that between "south" and "Africa" which occur together very rarely) could be made explicit. We would also like to work with clustering of n-grams rather than single words. Although, it would obviously be more computationally expensive compared to the word clustering approaches, phrases such as "south Africa" and "cold war" would arguably have much more descriptive power compared to sets of words.

Table 2. Results of labeling documents

Document Name: TIME071 ***Extract from the document:*** … EUROPE A NEW & OBSCURE DESTINATION IN AN ALLIANCE IN WHICH PARTNERS HAD BECOME INCREASINGLY MINDFUL OF ONE ANOTHER'S SENSITIVITIES, IN WHICH VICTORIES WERE TACTFULLY NOT CROWED OVER, AND TOGETHERNESS IN ITSELF WAS REGARDED AS A GOOD THING, CHARLES DE GAULLE LAST WEEK REMINDED THE WORLD OF WHAT ONE … ***Labels:*** **gaulle,france,europe,de** ***Score:*** 168
Document Name: TIME370 ***Extract from the document:*** … IN 1845, BEFORE THE POTATO FAMINE DECIMATED ITS POPULATION, IRELAND WAS WESTERN EUROPE'S MOST DENSELY SETTLED COUNTRY; SINCE THEN, ITS 9,000,000 INHABITANTS HAVE DWINDLED TO 2,824,000 . IRELAND IS THE ONLY NATION IN EUROPE WHOSE POPULATION HAS SHRUNK IN THAT TIME . WHILE IRISHMEN LEFT THE COUNTRY IN WAVES, THEY ENTERED IT … ***Labels:*** **ireland,irish,lemass** ***Score:***135
Document Name: TIME024 ***Extract from the document:*** … KASHMIR TALKING AT LAST THE BRITISH RAJ, WHICH ONCE CONTROLLED INDIA'S NORTHWEST FRONTIER PROVINCE OF KASHMIR, EXACTED A TOKEN ANNUAL TRIBUTE OF TWO KASHMIRI SHAWLS AND THREE HANDKERCHIEFS FROM THE MAHARAJAH . NEVER SINCE HAS THE PRICE OF PEACE BEEN AS SMALL . IN THE YEARS AFTER INDEPENDENCE IN 1947 SPLIT THE INDIAN SUBCONTINENT … ***Labels:*** **indian,Pakistan,india,kashmir,nehru** ***Score:***64
Document Name: TIME464 ***Extract from the document: …*** SOUTH VIET NAM REPORT ON THE WAR OVERSHADOWED BY THE POLITICAL AND DIPLOMATIC TURMOIL IN SAIGON, THE ALL BUT FORGOTTEN WAR AGAINST THE VIET CONG CONTINUES ON ITS UGLY, BLOODY AND WEARISOME COURSE . THE DRIVE AGAINST THE COMMUNISTS HAS NOT DIMINISHED IN RECENT WEEKS ; IN FACT, IT HAS INTENSIFIED . FEARS THAT THE … ***Labels:*** **Vietnam,south** ***Score:*** 50
Document Name: TIME381 ***Extract from the document: …***COMMUNISTS WAIT TILL NEXT YEAR SCARCELY HAD THE SINO-SOVIET TALKS GOTTEN UNDERWAY THAN THE MEETING HEADED FOR COLLAPSE . IT DID NOT MUCH MATTER WHEN RED CHINA'S SEVEN-MAN DELEGATION WOULD PACK THEIR BAGS AND ACTUALLY LEAVE MOSCOW ; BACK HOME PEKING'S PEOPLE'S DAILY SEEMED READY TO CALL IT QUITS . " WE WANT UNITY, NOT A SPLIT, " SAID THE.. ***Labels:*** **peking,red,soviet** ***Score:*** 28
Document Name: TIME302 ***Extract from the document:*** …KENYA THE RETURN OF BURNING SPEAR IN DAZZLING SUNLIGHT LAST WEEK, 30,000 SINGING, DANCING AFRICANS GATHERED BEFORE NAIROBI'S MINISTRY OF WORKS . A GREAT ROAR WENT UP AS TWO SOLEMN MEN EMERGED . ONE WAS KENYA'S BRITISH GOVERNOR MALCOLM MACDONALD . THE OTHER, WEARING HIS CUSTOMARY LEATHER JACKET AND BEADED BEANIE, WAS BURLY JOMO … ***Labels:*** **kenyatta,kenya** ***Score:*** 25
Document Name: TIME400

Extract from the document: …GREAT BRITAIN THE SAGA OF POLISH PETER LIKE THE OVERTURNING OF A DEEPLY EMBEDDED ROCK, THE PROFUMO SCANDAL CAUSED A FRANTIC SCURRYING OF A GREAT MANY ODD HUMAN INSECTS . ONE OF THE CRAWLIEST FIGURES TO EMERGE WAS THAT OF PETER RACHMAN, WHO MAY, OR MAY NOT, BE DEAD . LAST WEEK PRESS AND PARLIAMENT WERE ABUZZ WITH HIS SORDID STORY . RACHMAN.. ***Labels:*** **rachman** ***Score:*** 21
Document Name: TIME391 ***Extract from the document: …*** SOUTH AFRICA FAMILY TROUBLES FAMILY DAY IN SOUTH AFRICA IS AN EXPANDED VERSION OF MOTHER'S OR FATHER'S DAY A TIME FOR ALL KINFOLK TO GET TOGETHER . SOUTH AFRICA'S WHITES AND BLACKS LAST WEEK CELEBRATED THE HOLIDAY IN IRONICALLY CONTRASTING WAYS . WHILE WHITES PICNICKED OR FROLICKED ON BEACHES, THOUSANDS OF BLACKS MOURNED THE ABSENCE OF … ***Labels:*** **south** ***Score:*** 13
Document Name: TIME149 ***Extract from the document: …***EAST AFRICA THE ASIANS IN THEIR MIDST FOR MANY EUROPEAN SETTLERS, AFRICA FOR THE AFRICANS " SIMPLY MEANS PACKING UP AND GOING HOME, PAINFUL THOUGH IT MAY BE . THE FUTURE IS FAR DARKER FOR THE ASIANS IN EAST AFRICA, WHO HAVE LONG FORMED A PRECARIOUS MIDDLE CLASS . DESPISED BY COLOR-CONSCIOUS WHITES, … ***Labels:*** **African** ***Score:*** 10
Document Name: TIME068 ***Extract from the document:*** …EGYPT SURPRISE AT SUEZ WHEN EGYPT'S PRESIDENT GAMAL ABDEL NASSER GRABBED THE SUEZ CANAL 6F YEARS AGO, HIS BITTER ENEMIES IN EUROPE PREDICTED THAT THE BIG DITCH WOULD SOON BE FILLED WITH SILT AND THAT UNTRAINED EGYPTIAN PILOTS WOULD NEVER BE ABLE TO STEER SHIPPING THROUGH SAFELY . THE CRITICS TURNED OUT TO BE WRONG ON BOTH COUNTS . EGYPT HAS … ***Labels:*** **egypt** ***Score:*** 4

References

McCallum, A., and Nigam. K., 1999, Text classification by bootstrapping with keywords, EM and shrinkage, In *Proceedings of the ACL Workshop on Unsupervised Learning in Natural Language Processing.*

McQueen, J.B. 1967, Some Methods of Classification and Analysis of Multivariate Observations, In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281-297

Lewis, D., and Gale, W., 1994. A sequential algorithm for training text classifiers. In *Proceedings of the Annual ACMSIGIR Conference on Research and Development on Information Retrieval (SIGIR)*

McCallum, A., and Nigam, K., 1998, A comparison of event models for naïve Bayes text classification. In *Proceedings of the AAAI-98 Workshop on Learning for Text Categorization*

Joachims, T.,1998, Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the European Conference on Machine Learning.*

Lang, 1995, NewsWeeder: Learning to Filter Netnews, In *Proceedings of the International Conference on Machine Learning*

Koller & Sahami, 1997, Hierarchically Classifying Documents Using Very Few Words, In *Proceedings of the International Conference on Machine Learning*

Nigam, K., McCallum, A., Thrun, S., and Mitchell, T., 2000, Text classification from labeled and unlabeled documents using EM. *Journal of Machine Learning*, 39(2-3), pp. 103-134.

Blum, A., and Mitchell, T., 1998, Combining labeled and unlabeled data with co-training, In *Proceedings of the Conference on Computational Learning Theory(COLT)*

Bockhorst, J., and Craven, M., 2002, Exploiting relations among concepts to acquire weakly labeled training data. In *Proceedings of the International Conference on Machine Learning*

Ghani, R., 2002, Combining labeled and unlabeled data for multiclass text categorization, In *Proceedings of the International Conference on Machine Learning*

Goldman, S. and Zhou, Y., 2000, Enhancing supervised learning with unlabeled data, In *Proceedings of the International Conference on Machine Learning*

Denis, F., 1998, PAC learning from positive statistical queries. In *Proceedings of Algothmic Learning Theory*

Liu, B., Lee, W. S., Yu, P., and Li, X., 2002, Partially supervised classification of text documents, In *Proceedings of the Intl Conference on Machine Learning*

Yu, H., Han, J., Chang, K., 2002, PEBL: Positive example based learning for Web page classification using SVM, In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining*

Bing Liu et al, 2004, Text Classification by Labeling Words, In *Proceedings of AAAI National Conference*

Youngjoong & Jungyan, 2000, Automatic Text Categorization by Unsupervised Learning, In *Proceedings of COLING*

Sabine Bergler. 1990, Collocation patterns for verbs of reported speech--a corpus analysis of time Magazine corpus, Technical: report, Brandeis University Computer Science,. 1990.