BROWN UNIVERSITY
Department of Computer Science
Master's Project

CS-96-M5

"Clustering Words"

by

Ebru Ersan

Clustering Words

Ebru Ersan

Department of Computer Science
Brown University

Submitted in partial fulfillment of the requirements for the degree of Master of Science in the Department of Computer Science at Brown University

November 1995

Professor Eugene Charniak
Advisor

Clustering Words

Ebru Ersan
Department of Computer Science, Brown University
Sc.M. Project, Final Report

August, 1995

1 Introduction

In a previous study by Carroll and Charniak, a probabilistic context-free grammar (PCFG) was induced for English [2]. Using this grammar, the Wall Street Journal corpus (WSJ) was parsed and statistics on individual words were gathered. These statistics include the number of times a word is the head of a particular PCFG rule. Since it is an important relation used in the clustering process, we will shortly refer to it as "a word heads a PCFG rule". In this study, words that appear in the WSJ corpus are clustered using these statistics. Clustering is the process of classifying words into word classes according to some properties of words. The results of clustering can be used for smoothing language models which usually depend on statistics collected on individual words. WSJ corpus contains about 38 million words. The PCFG has 5057 rules. The statistics used in this study are about 25 Mega bytes.

In the following section, we give a brief overview of clustering and related work. Next, both the development and testing stages of this study are explained in detail. This is followed by the demonstration and evaluation of the results of clustering. Also in this section, some example clusters and test results that measure the correctness of clusters are given.

2 Overview of Clustering and Related Work

Statistics needed to parse and interpret sentences can be gathered on individual words or word classes. We refer to word classes as *clusters* and to the process of forming these clusters as *clustering*. It is a general term in the sense that it comprises all kinds of word classes regardless

of the property(s) on which the similarity is tested or the techniques used for the creation of the clusters.

Totally different clusters can be obtained by looking at different properties of words. One of the important issues is to decide on the correct property according to the task to be accomplished, that is, where the statistical information gathered from clusters will be used. For example, if the property chosen is the spelling of words, clusters will most probably contain words that begin with the same letter of the alphabet. There arises another important issue: the measurement of similarity. How do you decide that two words are similar enough to belong to the same cluster? Again the answer is related to the task at hand and the chosen property. There are many possibilities. Continuing our example, if the property chosen is spelling, two possibilities for the similarity measurement are the distance between two words in an alphabetically ordered list of words or the number of letters in a word. The other important issue is devising the algorithm for clustering. Because the amount of data involved is usually huge, coming up with an algorithm which is efficient in terms of both time and space usage gains more importance.

We will mention two important studies on clustering. Brown et al. [1] looked at the surrounding words of a word to cluster them. They extracted information from the context in which the word appeared. They report two studies, in one of them they look only at the next word and in the other they take into account all the words in a 1001 word window surrounding the word under examination. The similarity measure they used was average mutual information. We will not explain this metric in this report. For efficiency reasons, they first created 1000 clusters using the most common 1000 words in the corpus and then placed the remaining words into one of these clusters. The clusters were tested using a trigram model. The details of the tests are explained in [1]. The results showed that the clusters were of value. Even some of the misspelled words were clustered correctly.

Pereira and Tishby's study [5] is on nouns only. They extracted the verb-direct object information from the corpus and clustered the nouns according to the verbs they were direct objects of. Their distance metric was relative entropy [5]. As for the algorithm, they followed a different path. Instead of predefining a number of clusters, they began with one big cluster and split it into pieces. This behaviour makes it possible to examine the partitioning of clusters into smaller ones. Not only the resulting clusters but the whole tree structure is interesting to examine.

3 Clustering

3.1 Creating Clusters

We used syntactic information to create the clusters. Words are clustered according to the PCFG rules they head. Each word has a vector which contains the counts for every rule applicable to that word. From a graphical point of view, each word is represented with a point in an n dimensional space, n being the number of PCFG rules valid for that word. As hinted above, not every rule is valid for every word, this point will be elaborated when the algorithm is explained.

The distance between two of these points (or equivalently, the similarity between two words) is computed using the cosine of the angle between the two vectors. Cosine is chosen as the metric because it has a very desirable feature as explained in the following example. Consider, for example, the words "fast" and "swift". Although they have the same meaning and therefore head usually the same rules, the word "fast" typically would have much higher number of counts than the word "swift". If the distance metric used were eucledian distance between the points that represented the words, they would be so far away to fall into the same cluster. However, using cosine as the distance metric, two words will be in the same cluster, since their vectors point almost in the same direction.

Having decided on the property to discriminate upon and the metric to compute the similarity, the remaining issue is to devise the algorithm. The important thing is that the algorithm should be efficient enough, in terms of both space and time requirements, to deal with the vast amount of data to be processed. Following the study of Brown et al., we decided to start with a fixed number of clusters and put the remaining words into one of these existing clusters. Because the words are clustered according to the PCFG rules they head, we have a sort of "natural" clustering of the words to begin with, namely the different parts of speech. PCFG rules for different parts of speech are disjoint. Naturally, the rules used for expanding verb phrases are separate from the rules used for expanding noun phrases. Therefore they have discrepant vectors and form separate clusters. There are 27 parts of speech defined in our grammar. Of course, we need more clusters to begin with. A certain number of most common words are found for every part of speech and they are declared as each cluster's leading words. Every other word goes into one of these predefined clusters. Determining the number of clusters a priori brings a huge gain in terms of time. We made different experiments with different number of clusters ranging from 25 to 100 clusters per part of speech. The effect of changing the number of clusters is examined in detail in the following section. It certainly effects the resulting

clusters and the statistics gathered from these clusters, but the changes are not dramatic.

3.2 Testing Clusters

The correctness and the quality of the clusters are tested using a parser developed previously by Charniak [4]. Several experiments were made using this parser with six different probabilistic models. We will not go into the details of these models or the parser. However, a brief description is necessary. In the experiments, parsing is performed on sentences which are not used in training. These sentences are chosen from WSJ text and contain only the 5400 most common words of the corpus. The average sentence length is 17.2 words. Three of the models used in the experiments are of importance for this study. The simplest one uses only the PCFG probabilities of rules. The second model uses the PCFG probabilities and the probabilities of rules given the head of the constituent. These additional probability values are added to the model for smoothing. The third one makes use of the clusters, it uses the PCFG probabilities and the probabilities of rules given the cluster of the head of the constituent.

The parsing success is determined using five different measurements: per-word cross entropy, ratio, accuracy, recall, and error reduction. Ratio is the probability of a sentence divided by the probability of the most likely parse of that sentence. Accuracy gives the percentage of bracketing in the most probable parse which do not violate the correct text as given by the Penn Tree Bank. Recall is the percentage of the correct bracketing which were also in the most probable parse. Error reduction indicates the percentage of reduction in the number of bracket-crossing errors. The results of the experiments are given in the following section.

4 Results and Evaluation

As mentioned in the previous section, we experimented with different number of predefined clusters for every part of speech. Although this effects the test results, the overall structure of the clusters are similar in each case. The clusters will first be examined in general.

Henceforth, we will refer to clusters using their most common words. Because the number of clusters are predefined and forced to be the most common words of the corpus for their part of speech, there are a few number of clusters that could be joined otherwise. For example, for the nouns, there are both "inc." and "corp." clusters. In this case, similar amount of words are classified to be in either of the clusters. However, the situation is different for "million" and

"billion" clusters. The "billion" cluster got all the similar words leaving nothing for the other cluster, the cluster of "million" is empty. Some examples of the clusters are in Figure 1. As seen in the figures, some of the clusters exhibit an obvious similarity among the participating words, whereas some do not. However, the number of "bad" clusters are far less than the number of "good" ones.

Examining the clusters, it is not difficult to see that they do not consist of arbitrary words. We will shortly mention some of the regularities we observed. Plural nouns usually appear in clusters whose most common word is itself a plural noun. Company and brand names are gathered into one cluster the same way as proper nouns are clustered together. Generally, different forms of a verb (e.g., help, helps, helped) are placed into the same cluster. The cluster "York" (York coming from New York) includes words such as jersey, hampshire, orleans, and zealand. Although the clusters are created using syntactic information, words that belong to the same cluster are sometimes also very similar in semantic nature. For example, the cluster "share" includes only the words bushel, acre, barrel, hour, ounce, and bottle.

The correctness and the value of clusters are further supported by the results of experiments conducted using the parser mentioned in the previous section. All PCFG rules have a calculated probability. These probabilities are used while parsing a sentence. For example, a verb phrase can be expanded using the following rules with probabilities as displayed next to them.

```
VB_ -> VB NN_ (0.189156)
VB_ -> VB TO_ (0.057247)
```

The first rule claims that the head verb of the verb phrase is to be followed by a noun phrase with a probability of 0.189156 and the second rules says that the verb will be followed by a phrase beginning with "to" with probability 0.057247. The first rule is more common and more probable. However, if we know that the head of the verb phrase (i.e., the verb) is "go", the second rule is clearly more probable. More accurate results can be obtained by calculating and using the probabilities of rules given the head of the constituent. If, instead, the probabilities of rules given the cluster of the head of the constituent are calculated and used in the parsing process, a little bit of information is lost. However, the results are supposed to be more accurate than the results of the first case, i.e., using only the raw PCFG probabilities. In fact, our experiments gave exactly the expected results. The statistics of experiments with different number of clusters and without the clustering information (using only the head of the constituent information) are

Canadian Iranian Swiss Mexican Nicaraguan Swedish Dutch Chilean protected Argentine 10-year three-year
largest 11th 15th 10th sixth 12th eighth
outstanding payable joint-venture median unadjusted bhopal broke oriented
mr. mrs. ms. e.f. a.h. r.p. dr. c.j. prof. sir lt. st. sen. col. rep.
years nights yards fever sauce sets weeks trails nuclear trailers caps farms shifts rentals languages beers
court war mine station wagon guild club republican herald conference
loss gain decrease amount outline impact shift upturn piece dent degree reversal knock reduction decline
Japan Britain China Pakistan Poland Peru Belgium Colombia India Italy Cuba Indonesia Burma Brazil Iran Ethiopia Egypt Nigeria Norway Austria Israel Chile Honduras France Finland Cambodia Afghanistan Hungary Bolivia Kenya Czechoslovakia Nicaragua Angola Argentina Thailand Zimbabwe Mozambique Greece Haiti Spain Australia Algeria Yugoslavia Syria Turkey Romania Denmark Lebanon Iraq Mexico Venezuela Malaysia Sweden Russia
declined refuse ties planned plan vowed declines continued contributes vows spurt bow bowed
gave granted award loan handed hands seed
asked ordered pressed persuaded

Figure 1. Some example clusters of adjectives, nouns, and verbs. The clusters are given in full length except clusters of nouns which contain large number of words. The examples are taken from the experiment which is listed as "100 clusters" in Figure 2. The first word of every cluster is the most common and the defining word of that cluster. Other words are listed in order of descending similarity to the leading word.

Model	Cross Ent.	Ratio	Accuracy	Recall
PCFG	8.74	10.4	84.4	77.7
Head	8.09	6.78	87.9	80
25 clusters	8.42	7.08	85.6	78.6
50 clusters	8.39	6.65	87.1	79.3
100 clusters	8.35	6.45	86.3	79
varying # clusters	8.38	6.31	86.1	79
changing vector	8.38	6.29	87.2	79.3

Figure 2. Results after parsing 113 sentences. The PCFG rules are assigned different probabilities for each model. The length of sentences vary from 2 to 40 and the average sentence length is 17.2

given in Figure 2.

The first row of the table represents the statistics obtained using the probabilities of rules assigned by the PCFG. The second model uses the probabilities of rules given the head of the constituent. The other five models make use of the word clusters. They mainly differ in the number of predefined clusters. The probability of a rule given the cluster of the head of the constituent is calculated by dividing the total count of that rule in the cluster of the head by the total count of all the rules in that cluster. Using this formula, all the rules are assigned new probabilities for each cluster. The third, fourth, and fifth rows represent the results obtained by clustering with 25, 50, and 100 predefined number of clusters per part of speech, respectively. Of course, these numbers are the upper limits on the number of clusters, not all the parts of speech have that many words. For example, there are only 30 auxiliary verbs. Another experiment is made which takes this fact into account. The number of clusters for every part of speech is determined on the basis of the characteristics of that part of speech. When there are only 30 auxiliaries, if they are forced to make their own clusters, the probabilities calculated using these clusters are no different than calculating the probabilities of rules given the head of the constituent, there is only one word in the cluster. The sixth row displays the result of that

experiment.

During the clustering process, when a word is tried to be classified, a cosine value is calculated between the vector of that word and every cluster for that part of speech. The vector that represents the cluster is the vector of the most common word of that cluster. In all the experiments mentioned above, this vector is not altered as new words are added to that cluster. In other words, the words already decided to be in a cluster did not effect the representation of that cluster in the n dimensional space. The last row in the table represents the results of the experiment in which the vector representing a cluster is altered after each word is added to that cluster. The alteration is simple, the vector of the cluster accumulates the counts of rules for every word in that cluster. The predefined number of clusters is 50 for that experiment.

When the results in Figure 2 are examined, it is seen that the per-word cross entropy values consistently decrease as the number of clusters increases. As expected, per-word cross entropy values for different clustering experiments are better than the PCFG model's value but worse than the value of the model that uses the head of the constituent information. A similar result is observed for the measure "ratio". The decrease in these values are more significant than the previous columns'. Also, the results of the last two experiments are better than the others for that measure. The accuracy of the models exploiting clustering information are very close to each other, and similarly for recall values. In addition to the closeness of the values, they are no better than the value of the model that uses the head of the constituent information.

5 Conclusion

In this study, words of Wall Street Journal corpus are clustered using syntactic information. Clusters are created according to the PCFG rules that the words head. The distance metric used is the cosine of the angle between the vectors of two words. The number of clusters are predefined.

Testing of the resulting clusters are made using a previously developed parser. While parsing sentences, the probabilities of rules given the cluster of the head of the constituent are used as opposed to PCFG probabilities. These results are compared to the results of experiments made using the probabilities of rules given the head of the constituent. Also different experiments are conducted with different number of predefined clusters.

Test results showed that the clusters created in this study are of value. The accuracy

and recall values of models using the clustering information are higher than the PCFG model's values. Although these values are very close to each other for different clustering experiments, the per-word cross entropy and the ratio measurements gave better results for models using higher number of clusters.

References

[1] Brown, P. F., Pietra, V. J. D., DeSouza, P. V., Lai, J. C.,and Mercer, R. L. Class-based n-gram models of natural language. In *Computational Linguistics 18* 4 (1992), 467–479.

[2] Carroll, G. and Charniak, E. *Two experiments on learning probabilistic dependency grammars from corpora. In Workshop Notes, Statistically-Based NLP Techniques.* AAAI, 1-13.

[3] Carroll, G. and Charniak, E. *Learning probabilistic dependency grammars from labeled text. In Workshop Notes, Fall Symposium Series.* AAAI, 1992, 25–32.

[4] Charniak, E. Parsing with context-free grammars and word statistics. Manuscript, 1994.

[5] Pereira, F., and Tishby, N. *Distributional similarity, phase transitions and hierarchical clustering.* In *Working Notes, Fall Symposium Series.* AAAI, 1992, 108–112.