A METHOD FOR IMPROVING AUTOMATIC WORD CATEGORIZATION

Emin Erkan Korkmaz Göktürk Üçoluk

Department of Computer Engineering
Middle East Technical University

Emails: korkmaz@ceng.metu.edu.tr
ucoluk@ceng.metu.edu.tr

Abstract

This paper presents an approach to automatic word categorization in order to improve both the efficiency of the algorithm and the quality of the formed clusters. The unigram and the biagram statistics of a corpus of about two million words are used with an efficient distance function to measure the similarities of words, and a greedy algorithm to put the words in clusters. The notions of fuzzy clustering like cluster prototypes, degree of membership are used to form up the clusters. The algorithm is of unsupervised type and the number of clusters are determined at run-time.

1 Introduction

Different approaches have been proposed from different disciplines related to language acquisition. However it cannot be claimed that a total theory that can clarify and examine all the processes relevant to natural language acquisition exists. On the other side, researchers working on the subject now underline some aspects of the process.

First of all we know that a child has the capability to map the complex physical signals coming from the outside world onto some representation and by the help of this mapping process, the child can induce the right grammar which lets him/her to understand and produce the utterances. The critical question at this point is "what is the learning mechanism that comes into play?" There are those who believe that the concept of "learning" at this point is very closely related to the concept of "behaving in a typical sence". Marcken[1] states that the goal of the learner is to acquire a grammar under which the evidence is "typical", in a statistical sense. What is more, it is known fact that children can learn language without an explicit, well defined teaching process. By receiving a large number of examples, a child can make up his way through natural language where supervision only takes place while correcting the faulty utterences produced by the child.

These two statements form an approach to language acquisition where learning is visualised as developing a generative, stochastic model of language and putting this model into practice. And this is the key point underlying the studies in the area of statistical natural language processing.

Researchers taking this approach in computer science, have started to develop statistical NLP methods. And it has been shown practically that the usage of such methods can yield better performances for acquiring and representing the structure of language.

Automatic word categorization is an important field in statistical natural language processing where the process is unsupervised and is carried out by working on n-gram statistics to find out the categories of words. The researchers in this area point out that it is possible to determine the structure in a natural language by examining the regularities that is the statistics of that language. And such a structure is preserved in any language[10].

The organisation of this paper is as follows. First of all, related work in the area of word categorisation is presented in section 2. Then in the 3 section, the general background of the categorisation process is described, then the method used for this research is presented in detail. Lastly in sections 4 and 5 the results of the experiments carried out and the discussions for future work are given.

2 Related Work

There exists previous work in which the unigram and the biagram statistics are used for automatic word clustering. That is to say the frequency of single words and the frequencies of occurance of word pairs in a large corpus can give the necessary information to build up the word clusters. Finch[2] uses these biagram statistics for the weight matrix of a neural network. On the other side Brown[3] uses the same biagrams with a greedy algorithm to form up the hirerchical clusters of words.

Genetic algorithms have also been used for the categorisation process. Lanchorst[4] uses genetic algorithms to determine the members of predetermined classes. The drawback of his work is that the number of classes is determined previous to run-time and the genetic algorithm only searches for the membership of those classes.

McMahon and Smith[5] also use the mutual information of a corpus to find the hierarchical clusters. However instead of using a greedy algorithm they use a top-down approach to form the clusters. Firstly by using the mutual information the system divides the initial set containing all the words to be clustered into two parts and then the process continues on these new clusters iteratively.

Statistical NLP methods have been used also together with other methods in NLP. Wilms[6] uses corpus based techniques together with knowledge-based techniques in order to induce a lexical sublanguage grammar. Machine Translation is an other area where knowledge bases and statistics are integrated. Knight[7] tries to scale-up grammar-based, knowledge-based MT techniques with the use of statictical methods.

3 Word Categorization

The words in a natural language can be visualised as consisting of two different sets. The closed class words and the open class ones. New open class words can be added to the language as the language progresses, however the closed class is a fixed one and no new words are added to the set. For instance the prepositions are in the closed class. However nouns are in the open class, since news nouns can be added to the language as the social and economical life progresses. And it is a known fact that some words in a natural language are much more frequent related to the other ones. And these frequent words are commonly from the closed class.

The linguist Zipf[8], who was one of the early researchers on statistical language models, stated that only 2% of the words in a large English corpus form 66% of the total corpus. Therefore by working on a small set consisting of frequent words it is possible to build a framework for the whole natural language.

To build up such a framework n-gram models of language are commonly used. An n-gram model can be formed by collecting the probabilities of word streams $\langle w_i \rangle$ of length n. The probabilities will be used to form the model where we can predict the behaivour of the language up to n words. There exists current research that use biagram statistics for word categorization. That is the probabilities of word pairs in the text are collected.

3.1 Mutual Information

As stated in the related work part these n-gram models can be used with the concept of mutual information to form the clusters. Mutual information is based on the concept of entropy which can be defined informally as the uncertainity of a stochastic experiment. Let X be a stochastic variable defined over the set $X = \{x_1, x_2, ..., x_n\}$ where the probabilities $P_X(x_i)$ are defined for $1 \leq i \leq n$ as $P_X(x_i) = P(X = x_i)$ then the entropy of X, $H(X)$ is defined as:

$$H\{X\} = -\sum_{1 \leq i \leq n} P_X(x_i) \log P_X(x_i) \tag{1}$$

And if Y is another stochastic variable than the mutual information between these two stachostic variables is defined as:

$$I\{X : Y\} = H\{X\} + H\{Y\} - H\{X, Y\} \tag{2}$$

Here $H\{X, Y\}$ is the joint entropy of the stochastic variables X and Y. The joint entropy is defined as:

$$H\{X, Y\} = -\sum_{1 \leq i \leq n} \sum_{1 \leq j \leq m} P_{xy}(x_i, y_j) \log P_{xy}(x_i, y_j) \tag{3}$$

And in this formulation $P_{xy}(x_i, y_j)$ is the joint probability defined as $P_{xy}(x_i, y_j) = P(X = x_i, Y = y_j)$

Given a lexicon space $W = \{w_1, w_2, ..., w_n\}$ consisting of n words to be clustered, we can use the formulation of mutual information for the bigram statistics of a natural language corpus. In this formulation X is defined over the set of the first words appearing in word pairs and Y is defined for the second ones. So the mutual information that is the amount of knowledge that a word in a corpus can give on the proceeding word can be reformulated using the bigram statistics as follows:

$$I\{X : Y\} = \sum_{1 \leq i \leq n} \sum_{1 \leq j \leq n} \frac{N_{ij}}{N_{**}} \log \frac{N_{ij}.N_{**}}{N_{i*}.N_{*j}} \tag{4}$$

In this formulation N_{**} is the total number of word pairs in the corpus and N_{ij} is the number of occurences of word pair (i, j), N_{i*} is the number of occurences of word (i) and N_{*j} is the number of occurences of word (j) respectively. This formulation denotes the amount of linguistic knowledge preserved in biagram of words in a natural language.

3.2 Clustering Approach

When the mutual information is used for clustering, the process is carried out somewhat in a macro-level. Usually search techniques and tools are used with mutual information in order to form some combinations of different sets and then the validity of each configuration is tested. The idea used for the validity testing process is as follows. Since the mutual information denotes the amount of knowledge that a word gives on the proceeding word in a corpus, when a clustering configuration is formed, if similarly behaving words are collected in the same clusters than the loss of mutual information would be minimal. Thus the search is among the possible alternatives for sets or clusters that could yield the minimal loss in mutual information.

However in the presented work a different approach is used. Rather than searching the possible clustering by dividing the whole set into some clusters each time, a constructive bottom up approach, where set prototypes are first built and then combined with other sets or single words, is used. The method is based on the similarities or differences between single words rather than the mutual information of a whole corpus. In combining words into sets a fuzzy set approach is used. This serves to determine the behaviour of the whole set more properly.

Using this constructive approach, it is possible to visualize the word clustering problem as the problem of clustering points in an n-dimensional plane if the lexicon space to be clustered consists of n words. The points that are the words in a corpus for our case are distributed on this n-dimensional plane according to their behaviour related to other words in the lexicon space. Each word is placed on the i^th dimension according to its biagram statistic with the word representing the dimension. So the degree of similarity between two words can be defined informally as having close biagram statistics in the corpus. Words are distributed in the plane according to those biagram statistics. The idea is quite simple: Let w_1 and w_2 be two words from the corpus. Let Z be the stochastic variable ranging over the words to be clustered. Then if $P_X(w_1, Z)$ is close to $P_X(w_2, Z)$ and if $P_X(Z, w_1)$ is close to $P_X(Z, w_2)$ for Z ranging over all the words to be clustered in the corpus, than we can talk about a 'closeness' between the words w_1 and w_2. Here P_X is the probability of occurences of word pairs as stated in section 3.1. $P_X(w_1, Z)$ is the probability where w_1 appears as the first element in a word pair and $P_X(Z, w_1)$ is the reverse probability where w_1 is the second element of the word pair. This is the same for w_2 respectively.

In order to start the clustering process, first of all a distance function is needed between the elements in our plane. Using the idea presented above one can define a simple distance function between words using the bigram statistics. The distance function D between two words w_1 and w_2 is defined as follows:

$$D(w_1, w_2) = D_1(w_1, w_2) + D_2(w_1, w_2) \tag{5}$$

where

$$D_1(w_1, w_2) = \sum_{1 \le i \le n} \mid P_X(w_1, w_i) - P_X(w_2, w_i) \mid \tag{6}$$

and

$$D_2(w_1, w_2) = \sum_{1 \le i \le n} \mid P_X(w_i, w_1) - P_X(w_i, w_2) \mid \tag{7}$$

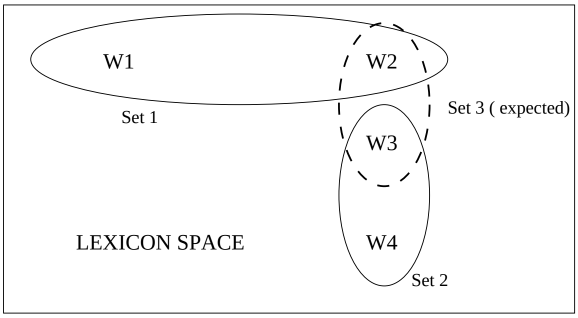

Figure 1: Example for the clustering problem of greedy algorithm in a lexicon space with four different words. Note that d_{w_2,w_3} is the smallest distance in the distribution. However since w_1 is taken into consideration, it forms set1 with its nearest neighbour w_2 and w_3 combines with w_4 and form set2, although w_2 is nearer. And the expected third set is not formed.

Here n is the total number of words to be clustered. Since $P_X(w_i, w_j)$ is defined as $\frac{N_{ij}}{N_{**}}$, the proportion of the number of occurences of word pair w_i and w_j to the total number of word pairs in the corpus, the distance function for w_1 and w_2 reduces down to:

$$D(w_1, w_2) == \sum_{1 \leq i \leq n} \mid N_{w_1 i} - N_{w_2 i} \mid + \mid N_{i w_1} - N_{i w_2} \mid \tag{8}$$

Having such a distance function, it is possible to start the clustering process. The first idea that can be used is to form a greedy algorithm to start forming the hierarchy of word clusters. If the lexicon space to be clustered consists of $\{w_1, w_2, ..., w_n\}$, then the first element from the lexicon space w_1 is taken and a cluster with this word and its nearest neigbour or neighbours is formed. Then the lexicon space is $\{(w_1, w_{s1}, ..., w_{sk}), w_i, ..., w_n\}$ where $(w_1, w_{s1}, ..., w_{sk})$ is the first cluster formed. The process is repeated with the first element in the list that is outside the formed sets, that is w_i for our case and it goes on until no word is outside a set. The formed sets will be the clusters at the bottom of the cluster hiearachy. Then to determine the behaviour of a set, the frequencies of its elements are added and the previous process is carried on the sets this time rather than single words until the cluster hierarchy is formed, that is until a single set is formed that contains all the words in the lexicon space.

In the early stages of this research such a greedy method was used to form the clusters, however although some clusters at the low levels of the tree seemed to be correctly formed, as the number of elements in a cluster increased in the higher levels, the clustering results were unsatisfactory.

Infact there were two main reasons for these unsatisfactory results. The first one was due to the greediness of the process and the second one was due to adding up the frequencies of elements to determine the set behaviour in the second part of the algorithm.

The greedy method results in an nonoptimal clustering in the initial level. This result could be shown using an example. Let us assume that four words w_1,w_2, w_3 and w_4 are in the lexicon space. And let the distances between these words be defined as d_{w_i,w_j}. For instance the distance between w_1 and w_2 is d_{w_1,w_2}. Then consider the distribution in Figure#1. If the greedy method first tries to cluster w_1. Then it will be clustered with w_2, since the smallest d_{w_1,w_i} for the first word is d_{w_1,w_2}. So the second word will be captured in the set and the algorithm will pass w_2 and continue the clustering process with w_3. At this point although w_3 is closest to w_2, since it is captured in a set and w_3 is more closer to w_4 rather than the center of this set a new cluster will be formed with w_3 and w_4. However as it can be obviously seen visually from Figure#1 the first optimal cluster to be formed between these four words is the set which contains w_2 and w_3.

The second problem causing unsatisfactory clustering occurs after the initial sets are formed. According to the algorithm after each cluster is formed, the clusters behave as like other single words and get into clustering with other clusters or single words. However to continue the process, the bigram statistics of the clusters or in other words the common behaviour of the elements in a cluster should be determined so that the distance between the cluster and other elements in the search space could be calculated. One easy way to determine this behaviour is to find the average of the statistics of all the elements in a cluster. However this method has drawbacks. The points in the search space for the natural language application are very close to each other. And if the corpus used for the process is not so large, the proximity problem is more severe. On the other side the linguistic role of a word can change in different contexts in different sentences. Many words can be used as a noun, adjective or as another linguistic category depending on the context. It can be claimed that each word is placed in a cluster initially with its dominant role. However to determine the behaviour of a set the dominant roles of its elements should be used. Somehow the common properties (bigrams) of the elements should be in use and the deviations of each elelment should be eliminated in the process. If such a method is not used the convergence of the sets are disturbed.

3.2.1 Improving the Greedy Method

To improve the clustering process the two drawbacks presented above should be overcome. First a method to overcome the first problem mentioned in figure1 will be presented.

The idea used to find the optimal cluster for each word at the initial step is quite simple. To form up such initial clusters in the algorithm used, being a member of more than one class is allowed for each word in the lexicon space. So after the first pass over the lexicon space, intersecting clusters are formed. For the lexicon space presented in Figure1 with four words, the expected third set is also formed. And as the second step these intersecting sets are combined into a single set. Then the closest two words according to the distance function are searched in each combined set and these two closest words are taken into consideration as the prototype for that set. After finding out the centroids for all sets, the distances between a member and all the centroids are calculated for all the words in the lexicon space. And each word is moved to the set where the distance between this member and the set center is minimal. This procedure is neccessary since the initial sets are formed with combining the intersecting sets. When these intersecting sets are combined the set center of the resulting set might be far away from some elements and there may be other closer set centers formed with other combinations, so the reorganisation of membership is needed.

3.2.2 Fuzzy Membership

As presented in the previous section the clustering process builds up a cluster hierarchy. In the first step words are combined to form the initial clusters, then those clusters become members of the process themselves. To combine clusters into new ones the statistical behaviour of them should be determined, since bigram statistics are used for the process. The statistical behaviour of a cluster is related to the bigrams of words in it. In order to find out the dominant statistical role of each cluster the notion of fuzzy membership is used.

the	5.002056%
and	3.281249%
to	2.836796%
of	2.561952%
a	2.107116%
in	1.591189%
he	1.533916%
was	1.419838%
that	1.306431%
his	1.124362%
it	1.061797%

Table 1: Frequencies of the most frequent ten words

The problem that each word can belong to more than one linguistic category brings up the idea that the sets of word clusters cannot have crisp borderlines and even if a word is in a set due to its dominant linguistic role in the corpus, it can have a degree of membership to the other clusters in the search space. Therefore fuzzy membership can be used for determining the biagram statistics of a cluster.

Researchers working on fuzzy clustering present a framework for defining fuzzy membership of elements. Gath and Geva[9] describe such an unsupervised optimal fuzzy clustering. They present the K-means algorithm based on minimization of an objective function. For the purpose of this research only the membership function of the presented algorithm is used. The memberhip function u_{ij} that is the degree of membership of the i^{th} element to the j^{th} cluster is defined as:

$$u_{ij} = \frac{\left| \frac{1}{d^2(X_i,V_j)} \right|^{\frac{1}{(q-1)}}}{\sum_{k=1}^{K} \left| \frac{1}{d^2(X_i,V_j)} \right|^{\frac{1}{(q-1)}}} \tag{9}$$

Here X_i denotes an element in the search space, V_j is the centroid of the j^{th} cluster. K denotes the number of clusters. And $d^2(X_i, V_j)$ is the distance of $X_i th$ element to the centroid V_j of the j^{th} cluster. The parameter q is the weighting exponent for u_{ij} and controls the fuzziness of the resulting cluster.

After the degree of membership for all the elements of all classes in the search space is calculated, the bigram statistics of the classes are added up. To find those statistics the following method is used: The bigram statistics of each element is multiplied with the degree of the membership of the element in the working set and this is the amount of statistical knowledge passed from the element to that set. So the elements chosen as set centroids will be the ones that affect a set's statistical behaviour mostly. Hence an element away from a centroid will have a lesser statistical contribution.

4 Results

The algorithm is tested on a corpus formed with online novels collected from the web page of the "Book Stacks Unlimited, Inc." The corpus consists of twelve free on-line novels adding up to about 1.700.000 words. The corpus is passed through a filtering process where the special words, useless characters and words are filtered and the frequencies of words are collected. Then the most frequent thousand words are chosen and they are sent to the clustering process described in the previous sections. These most frequent thousand words form the 70.4% of the whole corpus. The percentage goes up to about

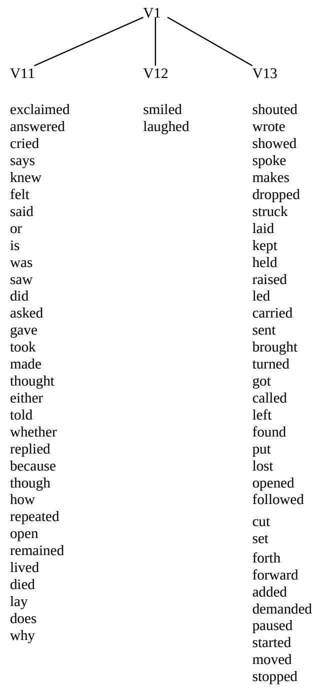

Figure 2: Part of the clustering hierachy

77% if the next most frequent thousand is added to the lexicon space. The first ten most frequent words in the corpora and their frequencies are presented in Table1.

As presented in the previous sections the clustering process builds up a tree of words with words on the leaves and clusters on the inner nodes. The starting node denotes the largest class containing all the lexicon space. Figure 2 is an example from the clustering tree. This part of the tree collects the verbs from the lexicon space. There are three different verb clusters at the lowest level and in the above level these three clusters are combined into one.

Some linguistic categories inferred by the algorithm are listed below:

- prepositions(1): by with in to and of
- prepositions(2): from on at for
- prepositions(3): must might will should could would may
- determiners(1) : your its our these some this my her all any no
- prepositions(4): between among against through under upon over about
- adjectives(1) : large young small good long
- nouns(1) : spirit body son head power age character death sense part case state
- verbs(1) : exclaimed answered cried **says** knew felt said **or is was** saw did asked gave took made thought **either** told **whether** replied **because though how** repeated open remained lived died lay **does why**
- verbs(2) : shouted wrote showed spoke **makes** dropped struck laid kept held raised led carried sent brough rose drove threw drew shook talked **yourself** listened wished meant **ought seem seems** seemed tried wanted began used continued returned appeared **comes knows** liked loved
- adjectives(2) : sad wonderful special fresh serious particular painful terrible pleasant happy easy hard sweet
- nouns(2) : boys girls gentlemen ladies
- adverbs(1) : scarcely hardly neither probably

- verbs(3) : consider remember forget suppose believe say do think know feel understand
- verbs(4) : keeping carrying putting turning **shut** holding getting hearing knowing finding drawing leaving giving taking making having being seeing doing
- nouns(3) : streets village window evening morning night middle rest end road sun garden table room ground door church world name people city year day time house country way place fact river next earth
- nouns(4) : beauty confidence pleasure interest fortune happiness **tears**

The faulty members in the clusters are shown above using bold font. As it can be realised from the above example, the clusters formed are not totally correct. The rate of misplaced words in the clusters above is 9%. However it can be claimed that the clusters represent the linguistic categories with a high success rate (about 91%). Also note the semantic relations in the clusters. For instance group nouns(2) is a good example for such a semantic relation between the words in a cluster.

5 Discussion And Conclusion

It can be claimed that the results obtained in this research are encouraging. Although the corpus used for the clustering is quite small compared to other researches, the clusters formed seem to represent the linguistic categories. The faulty collections seem to depend on the inefficent knowledge passed from the corpus. With a larger training data, an increase in the convergence of frequencies, thus an increase in the quality of clusters is expected. Since the distance function depends on only the difference of the bigram statistics, the running time of the algorithm is quite low compared to algorithms using mutual information. Although the order of the two algorithms are the same there is an increase in the efficiency due to the lack of time consuming mathematical operations like division and multiplication needed to calculate the mutual information of the whole corpus.

For further research the algorithm could be used to infer the phrase structure of a natural language. Finch[10] again uses the mutual information to find out such structures. Using fuzzy membership degrees could be another way to repeat the same process. To find out the phrases, most frequent sentence segments of some length could be collected from a corpus. And in addition to the frequencies and bigrams of words, the statistics for these frequent segments could be gathered and then they could also be passed to the clustering inference mechanism and the resulting clusters would hold such phrases together with the words. For instance noun phrases are to be in clusters with other nouns and verb phrases will be with other verbs.

As a summary it can claimed that automatic word categorization is the initial step for the acqusition of the structure in a natural language and the same method could be used with modifications and improvements to find out more abstract structures in the language and moving this abstraction up to the sentence level succesfully will make it possible for a computer to acquire the whole grammar of any natural language automatically.

References

[1] Carl G. de Marcken Unsupervised Language Acquisition PhD Thesis, 1996.

[2] S.Finch and N.Charter, Automatic methods for finding linguistic categories. In Igor Alexander and John Taylor, editors, Artificial Neural Networks, volume 2. Elsevier Science Publishers, 1992

[3] P.F. Brown, V.J. Della Pietra, P.V. deSouza, J.C. Lai, and R.L. Mercer. Class-based n-gram models of natural language. Computational Linguistics, 18(4):467-477, 1992

[4] M.M. Lankhorst. A Genetic Algorithm for Automatic Word Categorization. In: E. Backer (ed.), Proceedings of Computing Science in the Netherlands CSN'94, SION, 1994, pp. 171-182.

[5] John G. McMahon Francis J. Smith. Improving Statistical Language Model Performance with Automatically Generated Word Hierarchies. Computational Linguistics, 1996.

[6] Wilms, Geert Jan. Automated Induction of a Lexical Sublanguage Grammar Using a Hybrid System of Corpus and Knowledge-Based Techniques. Mississippi State University. PhD Thesis, 1995.

[7] Knight Kevin, Chander Ishwar, Haines Matthew, Hatzivassiloglou Vasieios, Hovy Eduard, Iida Masayo, Luk Steve, Okumura Akitoshi, Whitney Richard, Yamada Kenji. Intagrating Knowledge Bases and Statistics in MT. Proceedings of the 1st AMTA Conference. Columbia, MD. 1994.

[8] Zipf, G. K. The psycho-biology of Language. Boston: Houghton Mifflin. 1935

[9] Gath I, Geva A.B. Unsupervised Optimal Fuzzy Clustering. IEEE Transactions on pattern analysis and machine intelligence, Vol. 11, No. 7, July 1989.

[10] Finch, Paul Steven. Finding Structure in language. PhD Thesis. Centre for Cognitive Science, University of Edinburgh, 1993.