Automatic Sublanguage Identification for a New Text

Satoshi SEKINE
Computer Science Department
New York University
715 Broadway, Room.709
New York, NY 10003, USA

Abstract

A number of theoretical studies have been devoted to the notion of sublanguage, which mainly concerns linguistic phenomena restricted by the domain or context. Furthermore, there are some successful NLP systems which have explicitly or implicitly addressed the sublanguage restrictions (e.g. TAUM-METEO, ATR). This suggests the following two objectives for future NLP research: 1) automatic linguistic knowledge acquisition for sublanguage, and 2) automatic definition of sublanguage and identification of it for a new text. The two issues become realistic owing to the appearance of large corpora. Despite of the recent bloom of the research on the first objective, there are few on the second objective. If this objective is achieved, NLP systems will be able to optimize to the sublanguage before processing the text, and this will be a significant help in automatic processing. A preliminary experiment aiming at the second objective is addressed in this paper. It is conducted on about 3 MB of Wall Street Journal corpus. We made up article clusters (sublanguages) based on word appearance, and the closest article cluster among the set of clusters is chosen for each test article. The comparison between the new articles and the clusters shows the success of the sublanguage identification and also the promising ability of the method. Also the result of an experiment using the first two sentences in the articles indicates the feasibility of applying this method to speech recognition or other systems which can't access the whole article prior to the processing.

1 Introduction

A number of theoretical studies have been devoted to the notion of sublanguage, which mainly concerns linguistic phenomena, including syntax, semantics or pragmatics, restricted by the domain, context or

discourse of the text or the utterance. In particular, sublanguage studies indicated that several kinds of restrictions or deviations are characteristic for each sublanguage [1] [2] [3] [4].

Furthermore, there are some successful natural language processing systems which have explicitly or implicitly utilized sublanguage restrictions. For example, TAUM-METEO [5] is a machine translation system in which the translation is aimed only at sentences in the weather forecast domain, and it works remarkably well . Also, recently ATR [6] built a translation system for the conference registration task, and it works well, too.

This suggests the following two objectives in order to make a breakthrough on current NLP research.

1. Automatic linguistic knowledge acquisition for sublanguages.

2. Automatic definition of sublanguages and identification of the sublanguage of a new text.

These two goals become realistic owing to the appearance of large corpora. Using large corpora, preliminary experiments to meet the first objective have been conducted [7] [8]. Although these are still small experiments, in terms of the accuracy and the coverage for practical applications, their objectives address the first goal above, and could make a breakthrough for future N.L.P. systems by reducing the costly and errorsome linguistic knowledge encoding task by human linguists.

The second objective has not received so much attention. In the previous sublanguage N.L.P. systems, the domain the system is dealing with is predefined. For example, the definitions of "weather forecast domain", "medical report domain" or "computer manual text" are artificially or intuitively defined by a human. This is actually one method to define the sublanguage of the text, and it seems to work well. However, it is not easy and not always possible. The processing of newspaper articles is one such example. Since the range of articles is normally wide, a human can't prepare or intervene at each article to decide which sublanguage the text belongs to and it is impossible to utilize the sublanguage knowledge. In this paper, we will propose a objective way of defining sublanguage and automatically identifying the sublanguage for a new text. Then, comparison of the identified sublanguage and the test text will be reported, which reveal the method is promising.

2 Sublanguage

There has been a significant amount of research on the notion of sublanguage. In the literatures, sublanguage was defined by the name of the domain, (e.g. 'computer manual domain', 'weather report') or type of the document, (e.g. 'fiction', 'exposition'). Although comparative studies indicated that there exist several distinctive features in each sublanguage, there is no guarantee that these

features are uniform over a sublanguage defined in this way. The computer manual domain may contains several sublanguages in it. Therefore we need an objective and linguistic measurement of sublanguage to define it.

In our experiment, automatic article clustering based on word appearance is used to generate a set of sublanguages. The method we used is based on a document clustering metric [9] [10]. Although word appearance is one of the main features of sublanguage, other kinds of phenomena, including syntactic and semantic features, have been reported as sublanguage features. It would be better to utilize these phenomena as well in defining sublanguage, but several problems, in particular ambiguity problems, make it difficult for us to do so initially. Furthermore, word appearance is an important factor in N.L.P. systems. For example, word ambiguity in speech recognition and optical character recognition are good examples. These systems often produce several candidates for a portion of utterance or character sequence, and we sometime find that some of the candidates are totally irrelevant to the topic of the speech or the document (see the following example. Extracted from UNIX manual: command *banner*). This kind of ambiguity can be eliminated based on the sublanguage method which is being proposed in this paper.

```
Display  a  string in large letters.
Display aster ring in large lettuce.
```

3 Experiments

As we mentioned before, the experiments can be divided into the following three phases.

1. Sublanguage Definition: cluster articles based on word appearance
2. Sublanguage Identification: find the closest cluster to a new text
3. Evaluation: compare the cluster and the text

In addition, we will report another experiment which uses only the first two sentences of each article to determine its sublanguage. This experiment addresses the possibility of dynamic language adaptation in text processing.

4 Article Clustering

The corpus for the experiment consists of 1106 articles, extracted from a week of the Wall Street Journal (3 MB including header information). The statistics for the corpus are as follows:

- Number of articles: 1106
- Number of words (as tokens): 421281
- Number of words (as types): 25622

Clusters are produced by a similarity measure calculated between every two articles in the corpus. The formula for the similarity measure is based on the combination of inverse document frequencies of words and normalization by the number of the words in a text [10]. The formal definition similarity measure between article A_i and A_j is the following:

$$Similarity(A_i, A_j) = \frac{1}{|A_i||A_j|} \sum_{\{w|w \in A_i, A_j\}} \frac{1}{|A^w|} \quad (1)$$

Here, $|A|$ is the number of distinct words in article A, $|A^w|$ is the number of articles which contain word w. In the following explanation, values of the similarity measure are multiplied by 1,000,000, to aid readability.

The Cut off number for $|A^w|$ is set to 50, i.e. the sum is over words which occur in 50 or less articles throughout the corpus. This condition is introduced in order to avoid frequent words which may have no role in this calculation. Also, the minimum overlap is set to 3, i.e. the similarity measure between two articles becomes 0, if they have only 2 or less words in common. This condition is useful to avoid over-generation of accidental clusters.

Then a matrix of the similarity measure is obtained, whose values range from 0.0 to 1181.7 (recall that this value is 1,000,000 times the original similarity value). Clusters are generated based on this data. A simple algorithm, hierarchical clustering with single linkage method, is adopted for cluster generation. We set a threshold, i.e. any two articles for which the similarity measure is greater than the threshold belong to a same cluster. The value of the threshold is set to 50.0 experimentally. According to the algorithm above, we generated 129 clusters which have more than one article. The average number of articles in a cluster is 4.30 and the maximum number of articles in a cluster is 31.

5 Cluster Identification

We chose 50 test articles from new Wall Street Journal articles which are not used in the clustering experiment described above. In this section, we will describe the method to identify the closest

cluster for each test article. It is basically the same method as the calculation of the similarity measure explained in the previous section. For each test article, similarity measures to all clusters are calculated. Here, the similarity between an article and a cluster is set to equal to the maximum similarity between the article and an article in the cluster. This calculation is not so expensive, because we have a limited set of words which have to be taken into account in the calculation. The number of the words is normally much less than the number of tokens in the test articles and the number of clusters to be examined for each word is less than the cut off number (50 in the experiment). So this calculation is almost linear in the number of articles if there is enough space to store the word index.

We also set a threshold to decide if the test article and the cluster are similar enough. The threshold is set at the same level as the threshold which is used to produce the clusters (50.0). The result of the cluster identification experiment is shown in Table 1.

	Number of articles
Found the closest cluster	21
Found the closest article	9
Can't find anything	20

Table 1: Results of cluster identification

In Table 1, "Found the closest cluster" means that the closest cluster which contains 2 or more articles is found with greater similarity measure than the threshold value (50.0). "Found the closest article" means that it found the closest cluster satisfying the threshold, but the cluster contains only one article. "Can't find cluster" means that the closest cluster can't be found because of the threshold.

6 Evaluation

Evaluation of the experiment will be described in this section. The 21 articles which find the closest cluster containing more than 2 articles are examined at this evaluation. The evaluation measure is based on how many tokens in the test article also exist in the closest cluster, i.e. tokens which overlap between the test article and the closest cluster. A straightforward measurement is its coverage, which counts how many tokens are overlapping out of all the tokens in the article. This figure could help to decide how useful this method is, but it might still difficult to make an objective decision. So the number of overlapping tokens is compared with an expected value computed on the condition that tokens are randomly distributed in the 3 MB corpus (keeping word frequency distribution are the same). This expected value is calculated by the following formula.

$$E = \sum_{w} \frac{n_t * n_w}{N} (1 - (1 - \frac{n_c}{N})^{n_w}) \quad (2)$$

Here, n_t is the number of tokens in the test article, n_c is the number of tokens in the cluster, N is the number of tokens in the entire corpus, and n_w is the frequency of word w. The first factor inside the sum shows the expected frequency of a word w in the test article. The second term shows the probability that the word occurs at least once in the cluster in which n_c tokens exist. So the sum of the product for all the words in the corpus computes the expected number of tokens occuring in the test article which also occur in the closest cluster.

It is well known that high frequency words like "the" or "of", occur constantly regardless of the topic. On the other hand, low frequency words, which are often regarded as reflecting the topic, are expected to concentrated in similar articles. So, we can anticipate in this experiment that many low frequency words in the test article can also be found in the closest clusters. (Note that, because words with $|A^w| < 50$ are used in the similarity calculation, those words must be specially treated in the evaluation.) To observe such details, we classified the result by word frequency obtained on the 3MB corpus. The expected occurrence can be classified by limiting words in the sum to those whose frequencies are in the given range. Table 2 shows an example of the result. In the sample result, the number of tokens in the article is 164 and the number of tokens in the cluster is 621.

The first column shows the number of tokens and its expected number in the bracket below each number. For example, the expected number of tokens whose frequency ranges from 100 to 299 is 26.7, but is actually 24 in the article. These two figures indicate the balance of the word frequency in the article, and the averages of the ratio over all the 21 test articles are from 0.90 to 1.14, so we can say the articles are well balanced.

The second column shows information about overlapping tokens. For example, 13 tokens out of 24 tokens in frequency range from 100 to 299 are overlapping, and 106 out of 164 in the article are overlapping to the closest cluster.

The third column shows the coverage of overlapping tokens in the test article. We can see the overall coverage was 64.6% in this sample.

The figure in the fourth column indicates that tokens in the article are overlapping 135% of the expected value. Also 999% and 678% of expected overlap tokens are found in the article in frequency ranges from 1 to 49 and from 50 to 99, respectively. As mentioned before, the result for words whose frequency ranges between 1 to 49 can not be evaluated directly by the figure. As the method to find the closest cluster is based on similarity to each article, so the closest article in the cluster to the test article is the key in the cluster search. The number of tokens overlapping between the test article and the closest article in frequency 1 to 50, i.e. number of tokens used in the similarity calculation, is 6

Word frequency	Number of words	Overlap words	Coverage (%)	Ratio (%)
0	10 (5.6)			
1-49	32 (42.5)	11 (1.1)	34.4	998.8
50-99	22 (14.6)	10 (1.5)	45.5	678.2
100-299	24 (26.7)	13 (6.2)	54.2	209.0
300-999	22 (19.3)	18 (10.8)	81.8	167.2
1000-	54 (60.9)	54 (58.9)	100.0	91.6
Total	164 (164.0)	106 (78.5)	64.6	135.0

Table 2: An example of the result

for this sample (not shown in the table). This means that out of 11 overlapping tokens between the test article and the closest cluster, 6 tokens are found in the closest article and 5 others are found in the rest of the articles in the cluster. So these figures show the benefit of clustering in increasing the overlapping tokens.

The average coverage and ratio between the number of the overlapping tokens and the expected value throughout the 21 sample articles are given in Table 3.

The second column of the table shows the average coverage over the 21 articles in the experiment, and the third column shows the average ratio of number of overlap tokens to expected overlap tokens. From the figures in Table 3, we can tell the success of the method. For example, tokens whose frequency range from 50 to 99 are found with 473% of the expected value in the closest cluster. As these words are not used in the similarity calculation, it proves the existence of sublanguage at least with respect to word distribution. The same thing can apply to the data in range from 100 to 299.

For the words whose frequency ranges from 1 to 49, the total number of overlapping tokens between the test articles and their closest clusters is 256, and the number of overlapping tokens between the test articles and their closest articles, i.e. tokens used in the similarity calculation is 198. So, 58 new

Word Frequency	Coverage (%)	Ratio (%)
1 - 49	36.29	1077.56
50 - 99	46.25	473.45
100 - 299	54.11	230.16
300 - 999	78.52	140.47
1000 -	95.36	108.33
Total	67.18	142.04

Table 3: Average coverage and ratio

tokens are introduced by the clustering effect. These may be useful in language processing based on this sublanguage method.

Examining the values of the ratio in Table 3, it is intuitively understandable that the lower frequency words tend to have a large ratio and the higher frequency words tend to be close to 1.0. For instance, almost all of the words whose frequency is more than 1000 are closed class word, like "the", "of" or "it" (there are only 46 words which have frequency over 1000). The result that the ratio is close to 1.0 indicates that the distribution of these words is balanced, as usually assumed. On the other hand, the ratio in lower frequencies shows that lower frequency words tend to occur together in similar articles. This is the very fact that we had hypothesized before the experiment.

7 Experiment with first two sentences

For some N.L.P. applications, it may be impossible to see all the sentences in an article before processing. For example, spontaneous speech recognition systems have to process each utterance at a time. We will call these applications 'dynamic applications' in comparison to 'static applications' in which the system can pre-scan the material before the actual processing. The algorithm described above can't apply to dynamic applications directly. Therefore we conducted an experiment using the first two sentences in the test articles, instead of all the sentences in the article, for finding the closest cluster. If it can find a good cluster (i.e. sublanguage) for the rest of the article by using only the first two sentences, dynamic applications can adopt this method for the processing of sentences after the second sentence.

For this experiment, since the number of sentences to be examined is reduced, the minimum requirement of overlapping words is set to 2 instead of 3, and the threshold of similarity in this

experiment is deleted. Table 4 shows the evaluation of this experiment. The results are derived from the 21 test articles which are also used in the previous experiment.

Word Frequency	Coverage (%)	Ratio (%)
1 - 49	31.68	892.14
50 - 99	44.18	453.12
100 - 299	50.20	218.00
300 - 999	74.55	120.84
1000 -	96.48	104.07
Total	65.46	130.81

Table 4: Average coverage and ratio

Surprisingly, the result is almost the same as the previous one. Actually, more than half of them select the same closest cluster as that selected in the previous experiment by using all the sentences in an article. This fact is also intuitively understandable, because the first sentences in an article normally indicate its topic. So the first two sentences could help to find its sublanguage. This result is strongly encouraging that this method can be applicable to speech recognition systems or others in which it is impossible to pre-scan all the sentences before processing.

8 Discussion

The main discussion will be on the notion of sublanguage. In comparison to the traditional definition of sublanguage, we propose an objective and empirical definition of sublanguage. As we discussed, statistical methods using a large corpora are surely useful to define a sublanguage. However, we are not against the traditional way of sublanguage definition, since it well appeal to our intuition and actually it is useful for a certain purpose. Rather, our contention is that the definition may be depending on the purpose of processing.

The Wall Street Journal, which we used in the experiment, is normally regarded as a homogeneous material and a sublanguage itself. However, the results of the experiment shows that we can find several clusters in it and an NLP system will benefit if the system regards WSJ as a bunch of sublanguages.

Since the algorithm has worked successfully on this homogeneous corpus, it may be interesting to apply it to more general materials. For example, a daily newspaper includes wider range of topics, so it is easy to imagine that this algorithm works better in such a corpus.

The clustering method is problematic. In the experiment, parameters are set to make clusters for which the average number of articles in a cluster becomes 4.3. The bigger the clusters are, the bigger the coverage might be, which means the more tokens overlap between a test article and the closest cluster. This leads sublanguage knowledge more useful, but at the same time amount of the knowledge become larger and more ambiguity may be contained. This trade-off has to be settled by the feature of sublanguage and also the feature of the system. This is not easy to solve, but has to be considered at its implementation.

9 Conclusion

In conclusion, it appears that the sublanguage definition and identification works by using large scale corpus, and the evaluation results show that the automatically found sublanguage knowledge is useful in processing of a new text. The second experiment prove that the method is applicable not only for static applications which can pre-scan the material before its actual processing, but also for dynamic applications which process sequentially the sentences from the top to the bottom, like a speech recognition system.

We can find several directions of future study. One is, of course, to refine the algorithm and to make larger experiment. Also an experiment with different kind of corpus could give us fruitful prospects. With regard to applications, as has been repeatedly mentioned, speech recognition is one of the most interesting. Although the utility of this approach may heavily depend on the characteristics of the speech recognition system to find out how useful this method is, we believe that a certain type of ambiguity can be resolved and furthermore it may become possible to enlarge the vocabulary size of the word dictionary. Finally, it might be interesting to explore the same technique towards not only word occurrence but also other kinds of linguistic knowledge, including syntax, semantics or others.

10 Acknowledgments

The work reported here was supported by the Advanced Research Projects Agency under contract DABT63-93-C-0058 from the Department of the U.S. Army. We would like to thank our colleagues at NYU, in particular Prof.Grishman, whose comments have been very useful.

References

[1] R.Kittredge, J.Lehrberger ed.: "Sublanguage: Study of language in restricted semantic domain" (1982)

[2] R.Grishman, R.Kittredge ed.: "Analyzing language in restricted domains" (1986)

[3] D.Biber: "Using Register-Diversified Corpora for General Language Studies" *Computational Linguistics Vol.19, Number 2* (1993)

[4] W.Gale, K.Church, D Yarowsky: "One Sense Per Discourse" *4th DARPA Speech and Natural Language Workshop* (1992)

[5] P.Isabelle: "Machine Translation at the TAUM group" *The ISSCO Tutorial on Machine Translation* (1984)

[6] T.Morimoto et al.: "ATR's speech translation system: ASURA" 42.2, *Eurospeech* (1993)

[7] R.Grishman: "Discovery Procedures for Sublanguage Selectional Patterns: Initial Experiments" *Comp. Linguistics Vol.12 No.3* (1986)

[8] S.Sekine, S.Ananiadou, J.J.Carroll, J.Tsujii: "Linguistic Knowledge Generator" *COLING-92* (1992)

[9] P.Willett: "Resent trends in hierarchic document clustering: a critical review" *Information Processing and Management Vol.24, No.5 pp.577-597* (1988)

[10] K.Sparck-Jones: "Index Term Weighting" *Information Storage and Retrieval, Vol.9, p619-633* (1973)