Topic Tracking using Subject Templates and Clustering Positive Training Instances

Yoshimi Suzuki **Fumiyo Fukumoto** **Yoshihiro Sekiguchi**
Department of Computer Science and Media Engineering,
Yamanashi University
4-3-11, Takeda, Kofu, 400-8511, Japan
{ysuzuki@alps1.esi, fukumoto@skye.esb, sekiguti@alps1.esi}.yamanashi.ac.jp

Abstract

Topic tracking, which starts from a few sample stories and finds all subsequent stories that discuss the same topic, is a new challenge for the text categorization task and is useful for timeline-based IR systems. Much previous research on topic tracking use machine learning techniques. However, the small size of the training data, especially positive training stories, presents difficulties in training the parameters of the topic tracking system to produce optimal results. In this paper, we present a method for topic tracking using subject templates and k-means clustering algorithm to select a suitable training set. The method was tested on the TDT1 corpus, and the result shows the effectiveness of the method.

1 Introduction

Topic tracking of news stories, which starts from a few sample stories and finds all subsequent stories that discuss the same topic, is a new challenge and is useful for multi-document summarization as well as timeline-based IR systems such as archives of news and e-mails filtering. Topic Tracking is studied by many researchers including the TDT (Topic Detection and Tracking) project (Allan et al., 1998a). Most of them use machine learning techniques. The main task of these techniques is to tune the parameters, and they obtained high accuracy (Allan et al., 1998b), (Schultz and Liberman, 1999), (Yang et al., 2000). Suzuki showed that Support Vector Machine is a very effective machine learning technique for topic tracking (Suzuki et al., 2001). However, Yang claimed that optimal parameter settings for early and later stories are often very different. This may cause the fact that the discussion of a subject changes over time.

In the TDT1 corpus, many stories discuss a same topic. For instance, the topic "Kobe Japan quake" consists of 89 stories and "OK-City bombing" consists of 295 stories. A set of stories concerning a topic can be classified into some groups. This is particularly well illustrated by the topic: "Kobe Japan quake" in the TDT1 data. The first story concerning the "Kobe Japan quake" says that a severe earthquake shook the city of Kobe. It continues until the 5th story. The 6th through 17th stories report damage, location and nature of quake. The 18th story, on the other hand, states that the Osaka area suffered much less damage than Kobe. In a similar way, the 19th story discusses a new subject: President Clinton offers aid to Japan and stumps for support. The subjects of these stories are different from each other, while all of these stories are related to the topic: Kobe Japan quake topic.

In the topic tracking task where the number of initial positive training stories is 4, all of the initial positive training stories discuss that a severe earthquake shook the city of story and these stories do not discuss other subjects of "Kobe Japan quake". Also, the small size of the training data, especially positive training story presents special difficulties in tuning the parameters of the tracking system to produce optimal results. In addition, in incremental topic tracking, if once the system misjudges, error data are included in the positive training data ever since, and it gives negative effect for latter topic tracking.

In order to deal with these difficulties, we propose using subject templates and classifying positive training stories for selecting training data of SVMs.

We also use a broader class than topic: subject templates and topic class. For instance,

the topic class 'Earthquake' has subjects "occurrence of an earthquake", "occurrence of tsunami", "influence of earthquake", rescue efforts", "life of victims", "explanation of mechanism of earthquake", "explanation of damage from the earthquake", and etc. Although there are few studies using a broader class than topic, like a subject templates, using broader classes can ease data sparseness.

2 A Topic Class and A Subject Template

Suppose that stories of "Kobe Japan quake" is similar to the news stories concerning earthquakes occurred at other places, e.g., China or Mexico. The stories in a stream of news about these earthquake occurred at several places consist of some subjects, e.g., "early reports of damage from the earthquake", "location and nature of quake", "rescue efforts", "consequences of the quake", etc. We call these subjects *subject templates*.

In this paper, we use 4 kinds of clusters, e.g., a *topic class*, a *topic*, a *subject* and a *story*. Figure 1 illustrates the relation among these clusters. A *topic class* is a set of stories concerning a topic which occurs at general place and general time, e.g., 'Earthquake' and 'Criminal Trial'. A *topic* is an event or an activity, along with all directly related events and activities. We call a set of stories about an event or an activity in a topic *subject*. A *story* is a newswire article or a broadcast news story.

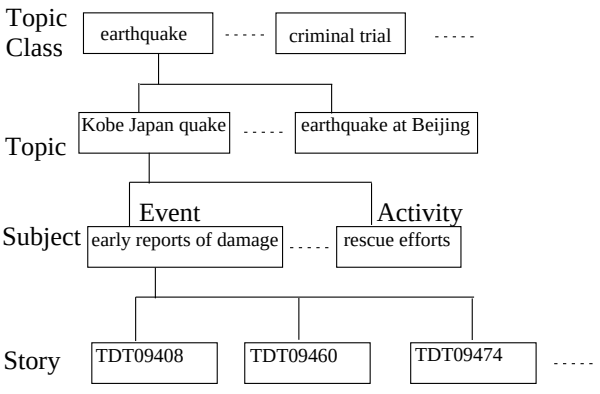

Figure 1: Relation among topic class, topic, subject and story

In order to extract stories which are sufficiently related to the current subject, we use *topic classes* and *subject templates*. Figure 2 illustrates the relationship between a topic class and subject templates. a topic class consists of some subject templates.

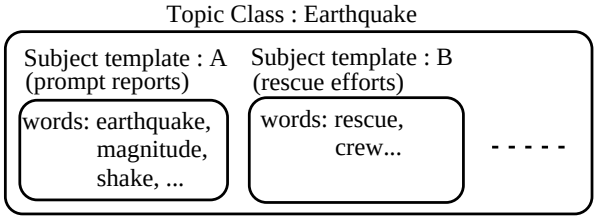

Figure 2: A topic class and subject templates

3 Extracting Subject Templates

In the topic tracking task, it is often the case that words which feature the subject of a test story are not included in the positive training stories. Because the subjects of the stories which discuss the same topic often shift to other subjects. In this case, the semantic distance between the test story and each of the positive training stories is not very close. In order to deal with the problem, we use *subject templates*.

Since classification of stories into subject templates needs costly human intervention, we classify stories of each topic in the corpora for subject templates into several clusters using k-means algorithm, and these clusters are substituted for subject templates.

The procedure for selecting subject templates of 'Earthquake' is as follow: We first select news concerning the earthquake from the corpus, next we extract common nouns, verbs and adjectives, finally we classify the stories into some subject templates using k-means algorithm. To select a suitable subject template for each story, we employ the k-Nearest Neighbors (kNN) method.

4 Selecting Training Data

In our topic tracking task, we use the combined positive training stories. The data is first created from the initial training data. During the tracking phase, if a test story is judged to be positive, we assume that the test story is relevant to the target topic, and the training data is regenerated by adding the test story to the initial training data.

In a stream of news, news stories closer together in the stream are more likely to discuss related subject than stories further apart. However, some news stories concerning the same subject are not close in chronological order. Besides, some selected stories may not discuss related topic with each other. In order to deal with the problem, we use a clustering technique.

We use k-means algorithm for clustering positive training stories. Yang et al. addressed the issue of difference between early and later stories related to the target event in the TDT tracking task. They adapted several machine learning techniques, including k-Nearest Neighbors (kNN) algorithm and Rocchio approach(Yang et al., 2000). Their method combines the output of a diverse set of classifiers and tuning parameters for the combined system on a retrospective corpus. The idea comes from the well-known practice in information retrieval and speech recognition of combining the output of a large number of systems to yield a better result than the individual system's output. They reported that the new variants of kNN reduced up to 71% in weighted error rates on the TDT3-dryrun corpus. In our method, the system classifies positive training stories into some subjects for positive training instances of SVMs.

There are three cases in our clustering method.

Figure 3 illustrates the case that a topic of positive stories include a subject and the test story is a member of it. In this case, it is easy to perform topic tracking, because all stories in positive stories are close to each other.

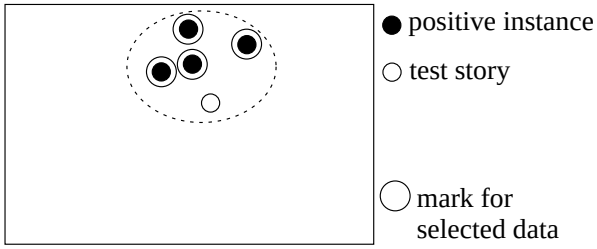

Figure 3: Clustering (when the test story is close to all positive instances)

Figure 4, on the other hand, illustrates the case that a topic of positive stories has a subject and the subject of a test story is not the same as the subject. In this case, the distance between the test story and a subject cluster is larger than that of between an story in a subject and a story with different subject. We use k nearest neighbors of test story.

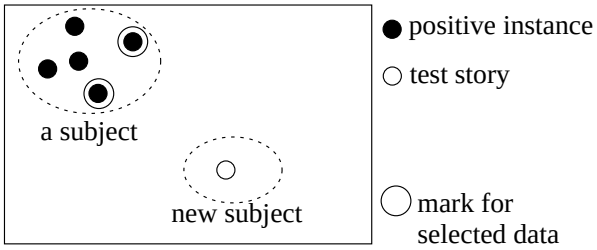

Figure 4: Clustering (when the test story is far from all positive instances)

Figure 5 shows the case where a topic of positive stories has some subject clusters and the subject of a test story is a member of a subject cluster. In this case, we will use a subject cluster which has a test story as a member by kNN algorithm.

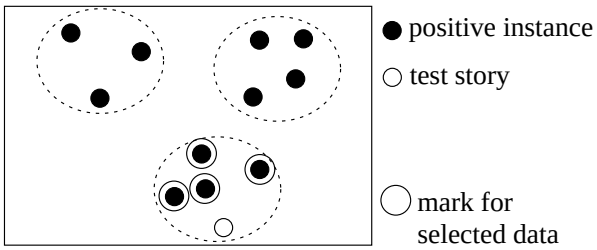

Figure 5: Clustering (when the test story is close to all positive instance in a cluster)

We select a set of negative training stories as well as the second case of clustering positive training stories.

5 Topic Tracking

We apply Support Vector Machines to the results of clustering. As positive training instances, we used all stories in a cluster which are closest to the target story, and as negative training instances, we used k nearest neighbors of the test data in the negative stories. Figure 6 illustrates the flow of our tracking method.

5.1 Applying SVMs to Topic Tracking

Support Vector Machines (SVMs) is a relatively new learning approach introduced by

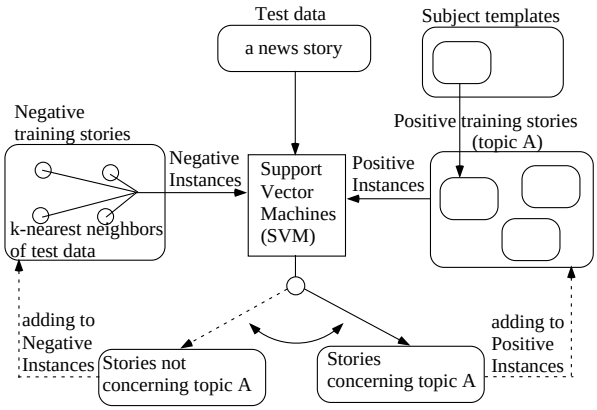

Figure 6: Topic Tracking

(V.Vapnik, 1995) for solving two-class pattern recognition problems. It is based on the Structural Risk Minimization principle for which error-bound analysis has been theoretically motivated. The method is defined over a vector space where the problem is to find a decision surface that 'best' separates a set of positive examples from a set of negative examples by introducing the maximum margin between two sets. The margin is defined as the distance from the hyperplane to the nearest of the positive and negative examples.

We applied SVMs to topic tracking using positive and negative instances which are selected by the clustering method.

6 Experiments

We first performed topic tracking experiments using 5 kinds of methods (Experiment A). Then we conducted experiments using 5 sets of subject templates and the best method of the former experiments (Experiment B).

The TDT1 corpus consists of 7,898 CNN TV news and 7,965 Reuters newswire articles. They are classified into 25 topics. We used 17 out of 25 topics, because 8 topics have less than 16 stories.

In order to make subject templates, we used the TDT2 (LDC, 1998) and TDT3 (LDC, 1999) corpora. We selected 9,168 stories which are evaluated as 'YES'(the story discusses the topic). We use 139 topics which has at least one story which are evaluated as 'YES'. For the training and test data, nouns, verbs and adjectives words are extracted by a part-of-speech tagger (H.Schmid, 1995) from the TDT1 corpus. Common nouns, verbs and adjectives are extracted from the TDT2 and TDT3 corpora to generate some subject templates.

Table 1: Methods in experiment A

Method	PTS	NTS
A	centroids	100 stories at random
B	4 NNs	100 stories at random
C	members of the closest cluster	100 stories at random
D	members of the closest cluster	100 NNs
E	all stories	all stories

Table 2: Results (selecting training data)

method	Recall	Precision	F1
A	52%	68%	0.59
B	52%	61%	0.56
C	53%	67%	0.59
D	64%	70%	0.66
E	49%	62%	0.55

We used SVM^{light} (Joachims, 1998) for topic tracking. We evaluated 5 methods to produce training data which is used for SVMs learning. Table 1 illustrates methods of experiment A. PTS and NTS indicates positive training stories and negative training stories, respectively. NNs denotes nearest neighbors.

We also performed another experiment (Experiment B) using 5 sets of subject templates with method D, which produced the best result among all methods in the experiment A.

6.1 Experimental Results

Table 2 illustrates the results using 5 kinds of methods for selecting training data for SVMs. The number of initial positive training stories is 4.

where

$$F1 = \frac{2 \times precision \times recall}{recall + precision}.$$

In Table 2, the result of method D is better than any other results. Table 3 shows the results with and without subject templates. The number of initial positive training stories is 4. We selected method D, which is the best among the results of the former experiment. "# of ST"

Table 3: Results (Subject templates)

# of ST	Recall	Precision	F1
0	64%	70%	0.66
1	69%	72%	0.70
2	69%	72%	0.71
3	67%	71%	0.69
4	66%	71%	0.68

denotes the number of subject templates in each topic of the TDT2 and TDT3 corpora. 0 in the column of "# of ST" denotes the method without subject templates. If the number of stories shown in Table 3 is larger than that of stories in a topic, we use all stories with the same topic. For example, when we performed an experiment using 4 subject templates but only 3 stories are in the topic, we assigned 3 stories to the 3 different subject templates.

In Table 3, the result using 2 subject templates per topic is better than the other results.

7 Concluding Remarks

The choice of *good* training data is an important issue for a binary classifier such as SVMs to produce a better result on topic tracking. In this paper, we presented a method for classifying training data and extracting subject templates from other corpora to select good training data. The results shows that clustering training data is effective for topic tracking by experiment A. Throughout the experiments, we found that the closest cluster of test story as the positive training data and k nearest neighbors of each test story as the negative training data are the best among the others. The results suggest that selecting training data is important for topic tracking, although the number of training data becomes small.

The results also suggest that using subject templates from other news corpora in order to enlarge the number of the positive training instances is effective for topic tracking by experiment B. We classified stories of other news corpora, i.e., TDT2 and TDT3 corpora into some subjects using k-means algorithm. The result using 2 subject templates per topic is better than any other results. This indicates that some topics has a small number of stories in the TDT2 and TDT3 corpora.

Future work includes (i) selecting the optimal number of the subject templates per topic, (ii) selecting the optimal number of the negative training instances, (iii) applying our method to different number of initial positive training stories.

References

J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Yang Y. 1998a. Topic detection and tracking pilot study final report. In *the DARPA Broadcast News Transcription and Understanding Workshop*, page http://www.itl.nist.gov/iaui/894.01/proc /darpa98/index.htm.

James Allan, Ron Papka, and Victor Lavrenko. 1998b. On-line new event detection and tracking. In *Proc. of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 37–45.

H.Schmid. 1995. Improvements in Part-of-Speech Tagging with an Application to German. In *Proc. of the EACL SIGDAT Workshop.*

Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In *Proc. of the Conference on Machine Learning*, pages 96–103.

LDC. 1998. Topic detection and tracking 2. In *http://morph.ldc.upenn.edu/Projects/TDT2.*

LDC. 1999. Topic detection and tracking 3. In *http://morph.ldc.upenn.edu/Projects/TDT3.*

J. Michael Schultz and Mark Liberman. 1999. Topic detection and tracking using idf-weighted cosine coefficient. In *Proc. of the DARPA Broadcast News Workshop.*

Yoshimi Suzuki, Fumiyo Fukumoto, and Yoshihiro Sekiguchi. 2001. Event tracking using wordnet meronyms. In *Proceedings of NAACL 2001 Workshop, WordNet and Other Lexical Resources: Applications, Extensions and Customizations.*

V.Vapnik. 1995. *The Nature of Statistical Learning Theory.* Springer.

Yiming Yang, Tom Ault, Thomas Pierce, and Charles W. Lattimer. 2000. Improving text categorization methods for event tracking. In *Proc. of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 65–72.