Active Learning for Story Segmentation of Spoken Documents

Shun-Chuan Chen
Speech Laboratory, College of EECS, National Taiwan University, Taipei, Taiwan, ROC
hypno@speech.csie.ntu.edu.tw

Abstract

In story segmentation, it is often difficult to gather the segmented data to learn a new model for the purpose of supervised learning. Therefore how to gather the useful data and to reduce the human effort on segmenting stories is an important issue. In this paper, we apply active learning to selecting the most informative examples to train a supervised model more efficiently. Active Learning aims to minimize the number of segmented examples by automatically selecting the examples that are most informative for the story segmentor. By this method, we can decrease the labor effort in boring story segmentation and get the same or better performance in automatic story segmentation In this paper, we also consider another problem that whether training data in story segmentation can be reusable or not and that how the different special structures of the language influence the performance of active learning for story segmentation. The experimental results show that active learning can reduce the number of segmented examples to reach a given level of performance and that reusable training data in story segmentation still contain information that improves the performance. It is a satisfactory result.

1. Introduction

Story segmentation is a research application that is used to detect changes between topically cohesive sections. It is one of the five main research applications for Topic Detection and Tracking Technologies (TDT) directed by NIST [1]. Currently there exist several methods to segment stories, and most of them use supervised learning methods, which require a lot of efforts to segment examples to prepare for training the segmentor. Because segmentation by humans is particularly time-consuming, we apply the new machine learning method called "active learning" to accomplish the task rather than rely on an expert or random sampling. An active learning system is designed to construct its own examples, to request certain types of examples, or to determine which set of unsupervised examples is most usefully segmented. The active learning we mention in this paper focuses mainly on the last approach: the selective sampling [2].

In our proposed approach, a hidden Markov model is used to simulate the topic change model. It is assumed that as the state changes, the topic changes. Next, the Viterbi algorithm is applied to find out the best path of a new sequence. The special structure of the language is also considered [3].

The rest of the paper is divided into five parts. In section 2, we describe the background of active learning. In section 3, we present our story segmentation system using a hidden Markov model. We describe the overall system and active learning for story segmentation in section 4. In section 5, the experimental results are given. We present the conclusion and suggestion for future work in section 6.

2. Active Learning

As mentioned in the introduction, the method we proposed focuses mainly on one important feature of active learning, the selective sampling method. In this approach, the initial learning phase begins with a small number of segmented examples and a large pool of unsegmented data. We then use the selective sampling method, an adaptive learner, to select the most informative data among the unsegmented examples. There are two main approaches to accomplish this work: certainty-based active learning method [4] and committee-based active learning method [5]. Both of these two methods are designed to choose the most informative data for learning.

In the certainty-based active learning approach, a simple model is trained with a small number of segmented examples. By this simple model, we can attach different certainties to each example. The k examples with the lowest certainties are then segmented

0-7803-8603-5/04/$20.00 ©2004 IEEE

and added to the training examples to retrain a new learner.

In committee-based approach, we can use several different methods to do the segmentation, such as the hidden Markov model [6], the probabilistic latent semantic model [7], the similarity-based method [8], etc. Each method is called a committee member. After the segmentation by these methods, there are different segmental results for each member. We finally pick up examples with the most disagreements among the members.

Figure 1 shows the algorithm of two selective samplings for story segmentation. In an ideal situation, the number of examples picked up in each selection, the batch size k, will be set to one to make the most intelligent decision in future selection. However, it is inefficient for doing the retraining with such a high frequency. Therefore, a reasonable batch size instead of one is adopted.

[Algorithm of Two Selective Samplings]

1. **Apply the learner to n bootstrap examples, creating one learner with N states (N categories).**
2. **Until there are no more examples or the annotator is unwilling to segment more examples, do:**
   - **2.1 Use the recent learner(s) to segment each unsegmented example.**
   - **2.2 Find the k examples with the lowest certainty/disagreement amongst all instances.**
   - **2.3 Annotate the boundaries of these examples.**
   - **2.4 Train the new learner on the bootstrap examples and all instances segmented to this point.**

Figure 1. The algorithm of two selective sampling methods for story segmentation.

In this paper, we only take into account the certainty-based selective sampling and mention two different approaches to certainty-based active learning: the reusable certainty-based active learning approach and the unreusable certainty-based active learning approach. It is easy to distinguish the two methods from each other. The reusable approach means that we can reuse the data used to train the previous model and then add them into the new learner. The reason why we can still gain new information from previous training data is because that the segmentation here is not just a classifier. Some steps of training a story segmentation learner such as clustering are influenced not only by straight data but also by what is transformed during the steps of the training process. On the other hand, the unreusable approach is, obviously, the traditional approach to the certainty-based method that cannot use the previous data any more.

3. Story Segmentation System

Story segmentation system [3] [6] is a useful tool to segment documents into stories suitable for different topic clusters. For example, we segment the broadcast news into units. Every unit seems to be a meaningful news story. To reach this goal, we identify a topic cluster for each sentence. When two adjacent sentences belong to two different topic clusters, a topic segmentation boundary is identified. Therefore, the problem of how to segment each story can be transformed into the problem of how to identify the topic change boundary.

Figure 2 is the diagram of the hidden Markov model. The probability of each state means the probability that the new sentence belongs to each topic cluster. The transition probability between two states is the probability that the new sentence changes its topic cluster from the original one to another. If the topic clusters between two sentences are the same, the probability equals to 1.0.

The state probabilities of the HMM are trained by the unigram language model for spoken documents, which are clustered to N topic categories. On the other hand, the transition probabilities here are assumed as a constant number.

By this method, we can find the boundaries between two stories with two different topic clusters.

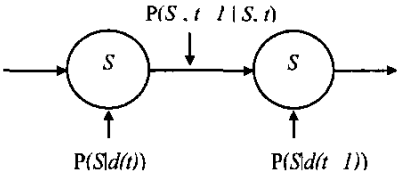

Figure 2. Hidden Markov model of story segmentation system.

4. Active Learning for Story Segmentation

The following figure 3 is the overall diagram of the active learning approach to story segmentation.

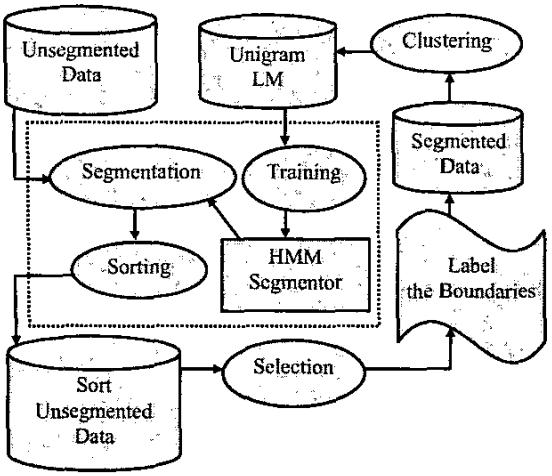

Figure 3. Overall diagram of the active learning approach to story segmentation.

To apply certainty-based selective sampling to the story segmentation system, the most important thing is to determine the certainty of each spoken document. In this paper, the basic type of each example is a long spontaneous spoken document. Each of them has a lot of stories with different topic clusters. During the processing of an example, our approach here is to compute each certainty for each decision that determines whether there is a boundary or not and to combine these certainties to obtain the overall certainty of the example. Therefore, during the processing of segmentation, we compute a confidence score for each decision and assume that there is a reverse correlation with the confidence score and the informativeness of each example.

The equation of the confidence score for each decision is as follows:

$$conf(d(t)) = (2 \times p(s(t) \mid d(t)) - \sum_{s' \in S} p(s' \mid d(t))), \quad (1)$$

where $d(t)$ is the t^{th} sentence in a spoken document and $s(t)$ is the state of this sentence. It means that

$$s(t) = \arg\max_{s \in S} p(s \mid d(t)).$$

Then, we can use this equation to compute the informativeness of each example as follows:

$$Info(d) = KL(P \| \quad) = KL(P(d(\cdot)) \| conf(d(\cdot)))$$

$$= \sum_{t=1}^{|d|} \left\{ p(d(t)) \times \log\left(\frac{p(d(t))}{conf(d(t))} \right) + (1 - p(d(t))) \times \log\left(\frac{1 - p(d(t))}{1 - conf(d(t))} \right) \right\}, \quad (2)$$

where $p(d(t)) = 1.0$, if $d(t)$ is a story boundary; otherwise, $p(d(t)) = 0.0$.

The informativeness of each example is computed during the process of segmentation. After that, we choose the examples with top k informative scores and add to the original corpus to retrain a new learner. After the iterative running process, we can get a satisfactory performance with fewer examples.

In this paper, we examine two different methods for certainty-based sample selection: reusable certainty-based approach and unreusable certainty-based approach. Both of them use the same certainty-based selection methods; however, the word "reusable" means that we can reuse the training examples, which have been used for training before. On the contrary, we cannot use any training examples that have ever been used for the unreusable approach. In the following section, we observe some experimental results for these two different approaches, and we also experiment these approaches using different feature units for the special structure of Chinese spoken documents [3].

5. Experimental Results

The spontaneous spoken documents we used here are TDT 2001 evaluation data [9]. TDT-2 was used as the training corpus including 3,320 minutes of audio signals, or 2,936 news stories, while TDT-3 as the testing corpus, including 7,620 minutes of audio signals, or 4,578 news stories. All of them are in Mandarin Chinese from the radio station "Voice of America". The segmentation cost, Cseg, defined by the TDT evaluation [9] was also used here, which includes the cost of false alarm and missing. The speech recognition was performed with the Dragon recognizer, with word, character, and syllable error rates begin 36.97%, 19.78%, and 15.06%, respectively, for the testing corpus and the training corpus.

For our story segmentation system, we use N different topics which we get from the K-means clustering method. Then we assume that the probability of each state is based on the unigram language model of each different topic cluster. The number of topics N we use here is 100, which performs well in general conditions.

For active learning results, we measured performance at 200 example intervals. We applied both reusable and unreusable approaches in each experiment and observed the different results caused by different kinds of feature units [3]. The first experimental result is shown in Figure 4.

Based on the basic type, the character, we observed that the best result of the reusable approach was better

than that of the unreusable one or random sampling, which is set to be the baseline. Of course, unreusable approach also performed better than random sampling. The best results of both methods of certainty-based active learning are better than the final result that we used total examples in the corpus. It can be proved that active learning can choose the examples with high informativeness from the training corpus. Nevertheless, the performance curve of the reusable method was a little unstable. It might be resulted from that the equation which matches the unreusable approach is not very suitable for the reusable one.

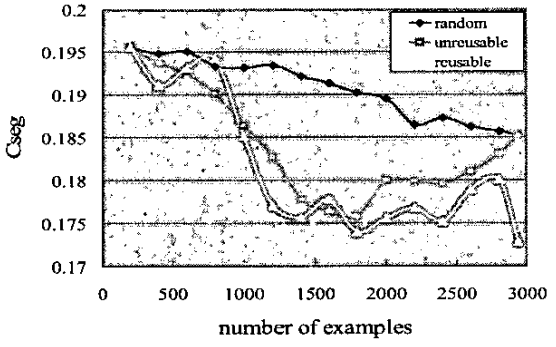

Figure 4. The experimental result for certainty-based selective sampling using character feature unit.

Figure 5 shows the result of using a different type of special feature unit, C(2), which means the overlapped two-character unit meaningful in Chinese language [3]. The result is just like the result for character. Therefore, active learning is still useful when different kinds of special structures are used.

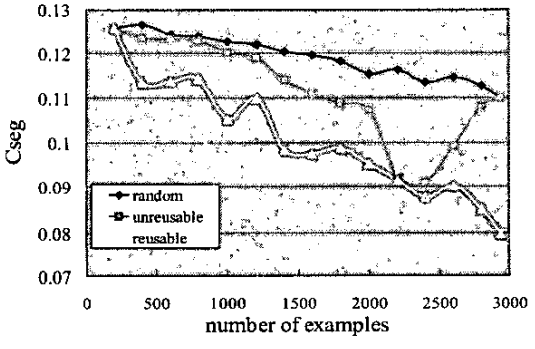

Figure 5. The experimental result for certainty-based selective sampling using overlapped two-character feature unit.

6. Conclusions and Future Work

In this paper, active learning is proved to be useful for story segmentation problems. We can see that reusable certainty-based selective sampling can reach a very satisfactory performance. However, it does not seem to be as stable as the unreusable one. It is because that the formula of informativeness suitable for the unreusable method may not match the reusable one. Reusable and unreusable certainty-based selective samplings can help improve the performance of segmentation. In the future, we will try to find the different certainty equations for reusable and unreusable approaches. Then, we would like to combine other features in spoken documents and implement committee-based selective sampling with other segmentation approaches.

7. References

[1] Topic Detection and Tracking (TDT) directed by NIST http://www.nist.gov/speech/tests/tdt/.

[2] Cohn, D., Atlas, L., and Ladner, R., "Improving Generalization with Active Learning", Machine Learning", 15(2), 201-221.

[3] Lin-shan Lee, Yuan Ho, Jia-fu Chen, and Shun-Chuan Chen, "Why is the Special Structure of the Language Important for Chinese Spoken Language Processing? - Examples on Spoken Document Retrieval, segmentation and Summarization", Eurospeech 2003, Geneva, Switzerland.

[4] Lewis, D. D. and Catlett, J., "Heterogeneous uncertainty sampling for supervised learning", ICML 1994, pp. 148-156, San Francisco, CA, U.S.A.

[5] Liere, R., and Tadeaplli, P., "Active Learning with committees for text categorization", AAAI 1997, pp. 591-596, Providence, RI.

[6] P. van Mulbregt, I. Carp, L. Gillick, S. Lowe, and J. Yamron, "Text Segmentation and Topic Tracking on Broadcast News via a Hidden Markov Model Approach", 1998.

[7] David M. Blei,Pedro J. Moreno, "Topic Segmentation with an Aspect Hidden Markov Model", ACM SIGIR'01, Sep.9-12, 2001, New Orleans, USA, pp343-348.

[8] Seiichi Takao, Jun Ogata and Yasuo Arikl, "Topic Segmentation of News Speech Using Word Similarity", Proc. of the Int'l Workshop on Multimedia Information Retrieval (ACM MIR2000), pp.442-444, (Nov. 2000).

[9] "The Topic Detection and Tracking 2001 (TDT-2001) Evaluation Plan", http://www.nist.gov/speech/tests/tdt/tdt2001/evalplan.htm.