Unsupervised Event Clustering in Multilingual News Streams

Martijn Spitters, Wessel Kraaij

Department of Multimedia Technology & Statistics
TNO TPD
P. O. Box 155, 2600 AD Delft
The Netherlands
{spitters, kraaij}@tpd.tno.nl

Abstract

The Topic Detection and Tracking (TDT) benchmark evaluation project embraces a variety of technical challenges for information retrieval research. The TDT topic detection task is concerned with the unsupervised grouping of news stories according to the events they discuss. A detection system must both discover new events as the incoming stories are processed and associate incoming stories with the story clusters created so far. The TNO topic detection system is based on a language modeling approach. The system has been evaluated on a multilingual corpus of approximately 80.000 stories from multiple new sources. For the grouping of stories we combined a simple single pass method to establish an initial clustering and a reallocation method to stabilize the clusters within a certain allowed deferral period. The similarity of an incoming story S_n to an existing cluster C is defined as the average of the similarities of S_n to each story $S_i \in C$. These individual similarities are computed by taking the sum of the generative probabilities $P(S_n|S_i)$ and $P(S_i|S_n)$ where S_i and S_n are modeled as unigram language models. Because these story language models are based on extremely sparse statistics, the word probabilities are smoothed using a background model.

1. Introduction

This paper describes the design and development of a system for the unsupervised grouping of news stories according to the events they discuss. The system has been evaluated on an augmented version of the TDT3 corpus which contains approximately 80.000 stories from multiple news sources, including both text and speech. These sources are newswires, radio and television broadcasts, and internet sites. The source languages are English and Mandarin. The TDT3 corpus is annotated for 120 events, each of which spans both English and Mandarin sources.

The TNO topic detection system is based on a language modeling approach. We had good experience with the application of language models for different IR-related tasks, like ad hoc, cross language, web and spoken document retrieval (Hiemstra and Kraaij, 1999; Kraaij et al., 2000; Hiemstra et al., 2001; Kraaij et al., 2002), filtering (Ekkelenkamp et al., 1999), and multi-document summarization (Kraaij et al., 2001). We also successfully applied language models for topic tracking (Spitters and Kraaij, 2001). However, due to the substantially higher computational complexity of topic detection, it was not trivial to convert our tracking approach into a detection algorithm. In the topic tracking task, events are to be followed individually. Each target event is defined by a small set of training stories that discuss it. Our tracking system estimates a single unigram language model based on the union of these on-topic stories and computes for each incoming story the likelihood according to this topic model. The computational complexity of this process is linear to the input. However, the topic detection task is a highly dynamic process. The topic models are constructed on the fly from the incoming stories. Each incoming story is added to a cluster, and thus changes the corresponding topic model. Experiments showed that reclustering the already processed stories (within the allowed deferral window) is important for a good performance. Reclustering is a computationally demanding process, since every change in cluster membership lists is reflected in changes in the cluster models, which form the basis for the similarity computation. Therefore we have chosen for a clustering approach which is independent of the (global) cluster models and instead is based on the similarities between individual stories. The advantage of this approach is that the inter-story similarities can be cached, resulting in a significant speed-up of the clustering process.

The remainder of this paper is organized as follows. To familiarize the reader with the TDT framework, section 2 elaborates on the TDT corpora, the TDT research tasks, and the TDT evaluation method. In section 3 we describe in detail our language model-based approach to topic detection. This section also contains a short study into the influence of two different smoothing methods for language models on the detection performance of our system. In section 4 we try to draw some conclusions.

2. The TDT benchmark test

The topic detection and tracking (TDT) benchmark evaluation project[1] was initiated by DARPA in 1996. After a pilot study in 1997, TDT has continued with annual evaluations conducted by the National Institute of Standards and Technology (NIST). Main purpose of the TDT project is to advance the state-of-the-art in determining the topical structure of multilingual news streams from various sources. See (Wayne, 2000) for a detailed overview of the TDT project.

2.1. TDT corpora

Currently, the Linguistic Data Consortium (LDC) has three corpora available to support TDT research[2] (Cieri et al., 2000). The TDT-Pilot corpus contains newswire and

[1] http://www.nist.gov/speech/tests/tdt

[2] http://www.ldc.upenn.edu/Projects/TDT

transcripts of news broadcasts, all in English, and is annotated for 25 news events. The TDT2 and TDT3 corpora are multilingual (Chinese and English) and contain both audio and text. ASR transcriptions and close captions of the audio data as well as Systran translations of the Chinese data are also provided. TDT2 and TDT3 are completely annotated for 100 and 120 events respectively. Currently, LDC is developing a new TDT corpus (TDT4) which will include Arabic news.

In the TDT evaluation, there are three alternative choices for the form of the audio sources to be processed, namely manual transcriptions, ASR transcriptions, or the sampled audio signal. Three story boundary conditions are supported: reference story boundaries (manually determined correct boundaries), automatic story boundaries (automatically determined errorful boundaries), or no story boundaries (the system must provide its own boundaries). Sites that participate in one of the TDT tasks are required to perform at least one evaluation under shared conditions. See (Doddington and Fiscus, 2001) for the TDT evaluation details.

2.2. TDT research tasks

The TDT benchmark evaluation project embraces a variety of technical challenges for information retrieval research. The goal of *story segmentation* is to segment a stream of data into homogeneous regions, discussing certain events. Given a small number of stories that discuss a certain event, a *tracking* system has the task to detect which stories in the data stream are related to this event and which are not. In *topic detection* there is no knowledge of the events to be detected. A detection system must both discover new events as the incoming stories are processed and associate incoming stories with the event-based story clusters created so far. A task which is very similar to topic detection is *first-story detection*. The goal of this task is to detect, in a chronologically ordered stream of stories, the first story that discusses a certain event. Finally, in *link detection*, the question to be answered is whether or not two stories discuss the same event.

2.3. TDT evaluation method

Topic detection systems are evaluated in terms of their ability to cluster together stories that discuss the same event (or events and activities that are directly connected to the cluster's seminal event). Detection performance is characterized in terms of the probability of miss and false alarm errors (P_{Miss} and P_{FA}). To speak in terms of the more established and well-known precision and recall measures: a low P_{Miss} corresponds to high recall, while a low P_{FA} corresponds to high precision.

These two error probabilities are combined into a single detection cost C_{Det}, by assigning costs to miss and false alarm errors (Doddington and Fiscus, 2001):

$$C_{Det} = C_{Miss} \cdot P_{Miss} \cdot P_{target} + C_{FA} \cdot P_{FA} \cdot P_{\neg target} \tag{1}$$

where C_{Miss} and C_{FA} are the costs of a miss and a false alarm respectively; P_{Miss} and P_{FA} are the conditional probabilities of a miss and a false alarm respectively; P_{target} and $P_{\neg target}$ are the a priori target probabilities ($P_{\neg target} = 1 - P_{target}$).

Then C_{Det} is normalized to:

$$(C_{Det})_{Norm} = \frac{C_{Det}}{min(C_{Miss} \cdot P_{target}, C_{FA} \cdot P_{\neg target})} \tag{2}$$

Detection error probability is estimated by accumulating errors seperately for each topic and by taking the average of the error probabilities over topics, with equal weight assigned to each topic. A set of predefined topics is automatically mapped to the system output topics by choosing for each reference topic the system output topic which produces the lowest evaluation cost.

3. Design of a probabilistic topic detection system

This section describes in detail the design of the TNO topic detection system. 3.1. describes our clustering approach. We combined a simple single pass method to establish an initial clustering and a reallocation method to stabilize the clusters within a certain allowed deferral period. In 3.2. we describe our story-cluster similarity measure. An incoming story is compared to an existing cluster by averaging the similarities of the new story S_n to each story in the cluster S_i. These individual similarities are defined as the sum of the generative probabilities $P(S_n|S_i)$ and $P(S_i|S_n)$ where S_i and S_n are modeled as unigram language models. Because these story language models are based on extremely sparse statistics, the word probabilities are smoothed using a background model. Section 3.3. reports on our experiments concerning the application of two different smoothing methods for language models and some contrastive tests with automatic versus manually determined story boundaries.

3.1. Clustering method

Our clustering procedure combines a simple single pass method and a reallocation method. Because the clusters formed by the single pass method are dependent of the order in which the stories are processed, they are merely used to initiate reallocation clustering. However, because in the TDT evaluation a topic detection system may defer its assignment of stories until a limited amount of subsequent source data (10 source files) is processed, the reallocation is restricted to the stories within that deferral period. More specifically, our clustering process involves the following steps:

1. For each new story within the deferral window, compute its similarity to each cluster the system has created so far. There are two options for a story:
   (a) if the similarity of the story to the closest cluster exceeds a certain threshold, assign the story to that cluster
   (b) else create a new cluster with the concerning story as its seed

2. When the end of the deferral window is reached, loop through the window stories again and compare each story to each existing cluster. There are three options for a story:

   (a) a story may switch to another cluster if the similarity to that cluster exceeds both the similarity to its current cluster and the threshold

   (b) if neither the similarity to its current cluster nor the similarity to any other cluster exceeds the threshold, create a new cluster with the concerning story as its seed

   (c) if the similarity to its current cluster exceeds the threshold as well as the similarities to all other clusters, the story stays in its current cluster

   Step 2 is repeated until all clusters are stable, that is, when 2c is true for each story.

The combination of a cluster initialization step and a reallocation step has previously (successfully) been used for topic detection by a.o. BBN (Walls et al., 1999) and Dragon (Yamron et al., 2000).

The reclustering step is important for a good performance of the detection system. However, the fact that every change in a cluster membership list means that the cluster language model would have to be reestimated, makes it a computationally demanding process. Therefore we have chosen for an approach which does not use the global cluster language models (contrary to our topic tracking approach) but instead is based on the similarities between individual stories. The similarity of an incoming story S_n to an existing cluster C is defined as the average of the similarities of S_n to each story $S_i \in C$. The advantage of this approach is that the inter-story similarities can be cached, resulting in a significant acceleration of the clustering process. These inter-story similarities are computed using a two-way language modeling approach, which is discussed in detail in the following section.

A cluster which has not changed for an uninterrupted period of fifteen days is frozen, which means that it is no longer considered an ‘active event’. The cluster is removed from the list of candidate clusters for new stories. This cluster evolution monitoring has two advantages. First of all it limits the computational complexity, because the number of clusters a story has to be compared with stays within certain bounds. Second, it can be argued that restricting the temporal extent of an event is beneficial for detection performance because it prevents different events with similar vocabulary (like different attacks or political elections) to be grouped together (Yang et al., 1999).

3.2. Language model-based similarity

The basic idea behind the language modeling approach to information retrieval is to estimate a (usually unigram) language model for each document and to rank documents by the probability that the document model generated the query. Absolute probabilities are not important for ranking in the IR situation. For other applications, i.e. topic tracking and also topic detection, scores have to be comparable on an absolute scale. For tracking, we found that modeling similarity as a likelihood ratio and normalizing this likelihood ratio by the (test) story length was adequate (Spitters and Kraaij, 2001). This normalized likelihood ratio is presented in equation (3), where $\mathrm{LLR}_{Norm}(T_1, T_2, ..., T_n|S_k)$ denotes the normalized log likelihood ratio of a story consisting of the terms $T_1, ..T_n$ given the story S_k in comparison with background model $\mathcal{B}$.

$$\mathrm{LLR}_{Norm}(T_1, T_2, ..., T_n|S_k) = \frac{1}{n} \log \sum_{i=1}^{n} \frac{P(T_i|S_k)}{P(T_i|\mathcal{B})} \quad (3)$$

In our clustering approach, the similarity between two stories S_n and S_i is based on a combination of the probability that the language model representing S_n generated story S_i and the reverse: the probability that the language model representing S_i generated story S_n. This approach results in the symmetrical similarity measure, presented in the following equation:

$$Sim(S_n, S_i) = \mathrm{LLR}_{Norm}(S_n|S_i) + \mathrm{LLR}_{Norm}(S_i|S_n) \quad (4)$$

Because the language models are estimated based on very limited amounts of text (single stories), it is very important that the word probabilities are smoothed using some background model. We performed a short study into the influence of two different smoothing methods on the performance of our detection system: Bayesian smoothing using Dirichlet priors and Jelinek-Mercer smoothing. The details of these smoothing methods and the results of our experiments are described in the following section.

3.3. Smoothing

Recent experiments at CMU have shown that different smoothing methods have different characteristics (Zhai and Lafferty, 2001a). For title ad hoc queries, Zhai and Lafferty found Dirichlet smoothing to be more effective than linear interpolation (Jelinek-Mercer smoothing). Both methods start from the idea that the probability estimate for unseen terms: $P_u(T_i|S_k)$ is modeled as a coefficient α_s times the background collection based estimate: $P_u(T_i|S_k) = \alpha_s \cdot P(T_i|\mathcal{B})$. A crucial difference between Dirichlet and Jelinek-Mercer smoothing is that the smoothing coefficient is dependent on the story length for Dirichlet, reflecting the fact that probability estimates are more reliable for longer stories. Formula (5) shows the weighting formula for Dirichlet smoothing, where $c(T_i|S_k)$ is the term frequency of term T_i in story S_k , $\sum_w c(T_i; S_k)$ is the length of story S_k and μ is a constant. The smoothing coefficient α_s is in this case $\frac{\mu}{\sum_w c(T_i;S_k)+\mu}$, whereas the smoothing coefficient is λ in the Jelinek-Mercer based model (formula (6)).

$$P(T_1, T_2, \cdots, T_n|S_k) = \prod_{i=1}^{n} \frac{c(T_i; S_k) + \mu P(T_i|\mathcal{B})}{\sum_w c(T_i; S_k) + \mu} \quad (5)$$

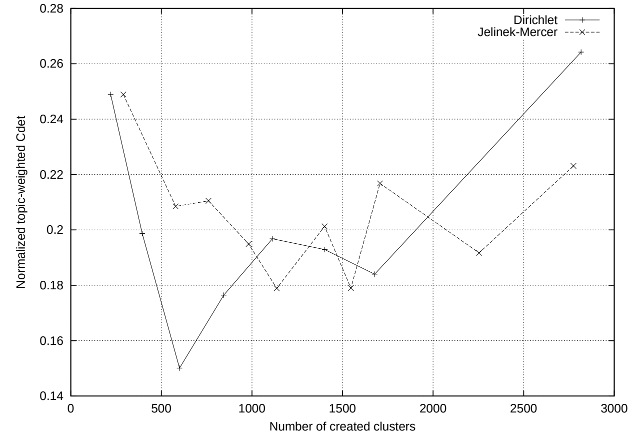

Figure 1: C_{Det} at different decision thresholds for two smoothing methods (Dirichlet and Jelinek-Mercer), performed on the TDT2 stories from April 1998, using automatic boundaries.

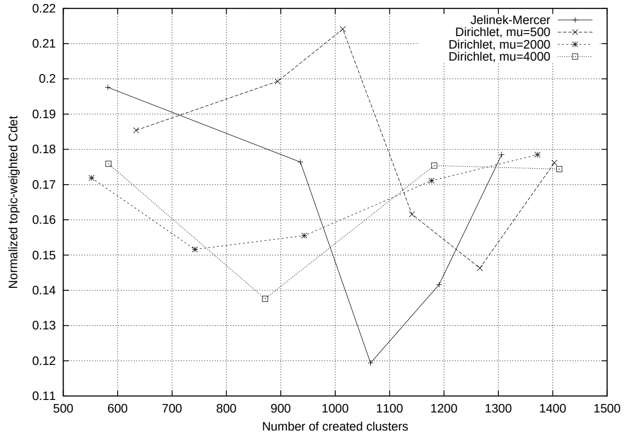

Figure 2: C_{Det} at different decision thresholds for two smoothing methods (Dirichlet and Jelinek-Mercer), performed on the TDT2 stories from April 1998, using reference boundaries.

$$P(T_1, T_2, \cdots, T_n|S_k) = \prod_{i=1}^{n} \lambda P(T_i|\mathcal{B}) + (1-\lambda) P(T_i|S_k) \quad (6)$$

For our official TDT2001 detection run, we applied Dirichlet smoothing with $\mu = 2000$. Our hypothesis was that Dirichlet smoothing would lead to improved performance, since story lengths vary considerably in the TDT corpus, and Dirichlet performed better than Jelinek-Mercer smoothing on a small test corpus (one month of stories from the TDT2 corpus) using the automatic story boundaries and ASR transcriptions of the audio (the primary topic detection evaluation requires these conditions). The results of this experiment are plotted in Figure (1).

We performed some post-hoc experiments on this same test set using reference story boundaries instead of automatic story boundaries and were surprised to find that Jelinek-Mercer performed better than Dirichlet under that condition, even when we varied μ (see equation (5)). Figure (2) shows the results. It is too early to draw conclusions from these experiments, since the test set was small and we did not explore the complete parameter space. However, one explanation could be the observation from Zhai and Lafferty (Zhai and Lafferty, 2001b; Zhai and Lafferty, 2001a) that smoothing has two functions: i) improving the maximum likelihood estimates ii) generate common words in the query. The latter function is especially important for longer queries since they contain more common words.

In the topic detection task we use language models to generate stories instead of queries. Since stories are considerably longer than TREC title queries, it is probably important that the smoothed model generates common words with proper "idf"-like probabilities. The TREC experiments show that the two roles of smoothing have an inverse interaction with the query length. Dirichlet is a good strategy for the first smoothing role (avoiding the assignment of a zero probability to an unseen word) while Jelinek-Mercer is better for the second role (weighting query terms in an idf-like fashion) (Zhai and Lafferty, 2001a). The longer the "queries" are, the more important the second function will become. This phenomenenon might be an explanation for the fact that Dirichlet performs best under the automatic story boundary condition, and Jelinek-Mercer under the reference story boundary condition, since the former has shorter stories than the latter (median: 62 versus 114). Further experiments are needed, including a validation of a combined Dirichlet/Jelinek-Mercer smoothing scheme for the TDT tasks.

4. Conclusions and future work

We think that the choice to use normalized likelihood ratios as the basis of a similarity measure was the key for the good performance of our system. Like in the tracking task, a proper normalized similarity measure is of utmost importance. Simply adding the generative probabilities $P(S_n|S_i)$ and $P(S_i|S_n)$ proved to work well to "symmetrize" the similarity measure. The accuracy of a language model-based clustering approach which is independent of the (global) cluster models and instead is based on the similarities between individual stories surpassed our expectations. However, we intend to check whether a similarity measure based on the global cluster model would enhance the results. The results of some initial post-hoc experiments indicate that the Jelinek-Mercer smoothing method works better than Dirichlet smoothing for manually segmented data, while the Dirichlet method yields better performance than Jelinek-Mercer on automatically segmented data. Further investigation is necessary to draw definite conclusions.

5. References

C. Cieri, D. Graff, M. Liberman, N. Martey, and S. Strassel. 2000. Large multilingual broadcast news corpora for cooperative research in topic detection and tracking: The TDT2 and TDT3 corpus efforts. *Proceedings of the Language Resources and Evaluation Conference (LREC2000).*

G. Doddington and J. Fiscus. 2001. The year 2001 topic detection and tracking (TDT2001) task definition and evaluation plan. Technical Report v. 1.0, National Institute of Standards and Technology.

R. Ekkelenkamp, W. Kraaij, and D. van Leeuwen. 1999. TNO TREC-7 site report: SDR and filtering. *Proceedings of the Seventh Text Retrieval Conference (TREC-7)*, pages 519–526.

D. Hiemstra and W. Kraaij. 1999. Twenty-one at trec-7: Ad hoc and cross language track. *Proceedings of the Seventh Text Retrieval Conference (TREC-7)*, pages 227–238.

D. Hiemstra, W. Kraaij, R. Pohlmann, and T. Westerveld. 2001. Twenty-one at CLEF 2000: Translation resources, merging strategies and relevance feedback. *Proceedings of the CLEF 2000 Cross-Language Text Retrieval System Evaluation Campaign*.

W. Kraaij, R. Pohlmann, and D. Hiemstra. 2000. Twenty-one at TREC-8: using language technology for information retrieval. *Proceedings of the Eighth Text Retrieval Conference (TREC-8)*, pages 282–299.

W. Kraaij, M. Spitters, and M. van der Heijden. 2001. Combining a mixture language model and naive bayes for multi-document summarisation. *Notebook papers of the Document Understanding Conference (DUC 2001)*.

W. Kraaij, T. Westerveld, and D. Hiemstra. 2002. The importance of prior probabilities for entry page search. *Proceedings of the 25th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, To Appear.

M. Spitters and W. Kraaij. 2001. Using language models for tracking events of interest over time. *Proceedings of the Workshop on Language Models for Information Retrieval (LMIR 2001)*, pages 60–65.

F. Walls, H. Jin, S. Sista, and P. van Mulbregt. 1999. Topic detection in broadcast news. *Proceedings of the DARPA Broadcast News Workshop*.

C.L. Wayne. 2000. Multilingual topic detection and tracking: Successful research enabled by corpora and evaluation. *Proceedings of the Language Resources and Evaluation Conference (LREC2000)*, pages 1487–1494.

J.P. Yamron, S. Knecht, and P. van Mulbregt. 2000. Dragon's tracking and detection system for the TDT2000 evaluation. *Notebook papers of the Topic Detection and Tracking Workshop (TDT) 2000*.

Y. Yang, J. Carbonell, R. Brown, T. Pierce, B.T. Archibald, and X. Liu. 1999. Learning approaches for detecting and tracking news events. *IEEE Intelligent Systems: Special Issue on Applications of Intelligent Information Retrieval*, 14(4):32–43.

C. Zhai and J. Lafferty. 2001a. Dual role of smoothing in the language modeling approach. *Proceedings of the Workshop on Language Models for Information Retrieval (LMIR) 2001*, pages 31–36.

C. Zhai and J. Lafferty. 2001b. A study of smoothing methods for language models applied to ad hoc information retrieval. *Proceedings of SIGIR 2001*.