

THE TOPIC DETECTION AND TRACKING WITH TOPIC SENSITIVE LANGUAGE MODEL

He Ruifang Qin Bing Liu Ting Li Sheng

Information Retrieval Laboratory, School of Computer Science & Technology
Harbin Institute of Technology 321#, China
{rfhe,qinb,tliu,sli}@ir.hit.edu.cn

ABSTRACT

In this paper, we explore the language model with topic sensitive features for the topic detection and tracking, formulate the relationship among the Chinese internet new words, language model with topic sensitive feature and the scheduling logic and the interval temporal reasoning and the key techniques. we use the Chinese internet new words to strengthen the detection and tracking of the topic and try to employ the scheduling logic and interval temporal reasoning to educe the reciprocal influences of events. At last we summarize the potential issues and the future work.

KEYWORDS

TDT, Chinese internet new words recognition, topic-sensitive language modeling, scheduling logic, interval temporal reasoning

1. MOTIVATION

The thriving and prosperous internet brings about the convenience as well as the perplexity for the human being, it is easy to get the information ,but sometimes too much information make us different to decide what we should choose, so the orientation of the internet media can heavily influences peoples' behavior, especially the young. the following information security problem becomes more and more stand out, not only the government, the financial department, but also the enterprise and individual all emphasize on it.

Except for the security of the information system, there is a different problem needed to be handled imminently, that is the content security of the information which is about how to acquire and track the novel content in the formation, at the same time, we need to monitor and give an alarm on the new event with the unhealthy information, this is significant for the security of the state and the healthy body and soul together of the young. Because the content of the special information is mainly expressed as the character, it badly needs the support of the natural language processing techniques, including the topic detection and tracking (short for TDT).

2. RELATED RESEARCH OF THE TDT^[1,2,3]

2.1 Concise History

Topic detection and tracking is a DARPAR-sponsored initiative to investigate the state of the art in finding and following new events in the stream of broadcast news stories, the basic idea for TDT originated in 1996, when the Defense Advanced Research Project Agency (DARPAR) realized that it needed technology to determine the topical structure of news streams without human intervention.

In 1997, a pilot study laid the essential groundwork, producing a small corpus and establishing feasibility. During 1998 and 1999, TDT research blossomed, with new and more challenging tasks, many more participating sites, and considerably larger multilingual corpora (adding Arabic data in 1998 and Chinese data in 1999), which are available from the Linguistic Data Consortium(LDC).

TDT research is continuing under the new DARPA program known as TIDES (Trans-lingual Information Detection, Extraction, and Summarization). And the whole research is driven by the annual TDT evaluations conducted by the National Institute of the Standards and Technology (short for NIST).

The firstly formal evaluation is hold in 1998, from the 1998 to 2004, NIST successfully holds seven TDT evaluations.

2.2 Goal and Value

The purpose of the Topic Detection and Tracking(TDT) Pilot Study is to advance and accurate measure the state of the art in TDT and to assess the technical challenges to be overcome. Its basic idea refers to a variety of automatic techniques for discovering and threading together topically related material in streams of data such as newswire and broadcast news. Such automatic discovery and threading could be quite valuable in many applications where people can have timely and efficient access to large quantities of information. Systems could alert users to new events and to new stories, By examining one or two stories, a user could decide whether to pay attention to the rest of an

evolving thread. Similarly, a user could go to a large archive, find all the stories about a particular event, and learn how it evolved.

2.3 Main Task of TDT and Our Work

The TDT problem consists of five major technical tasks: (1) segmenting a stream of data, especially recognized speech, into distinct stories (segmentation); (2) finding the relevant stories about a given topic (tracking); (3) detecting and threading together new topics (detection); (4) detecting new topics (first story detection); (5) deciding whether stories are on the same topic (linking), the content of the evaluation has little difference every year, up to 2004, there are some changes on the tasks, except for the new event detection, link detection, topic tracking, adding two experiment tasks: supervised adaptive topic tracking; hierarchical topic tracking.

Our work is focused on the multi-document summarization, and we develop our own Chinese system embedded in the news search engine of the IR-Lab, in fact, the TDT is the inheriting and the development of the multi-document technique.

We first formulate the basic idea of the public feelings monitoring and the plan graph of the task in section 3, put forward the three key techniques in section 4, and we conclude with a summary and mention of future work in section 5.

3. THE PRIMARY IDEAS AND THE PLAN GRAPH OF THE TASK

The monitoring of the novel events and the tracking of the topic include the detection of the novel event and the tracking of the reciprocal influence among the relevant topic events and the development of the new events driving from them. As for the different dimensions and the different granularities, We expect to observe the reciprocal influences among the different events, from the point of view of the horizontal line, including synchronization relation and the concurrent relation of the events and so on; we also expect to observe the entity relation graph contained in the different events for special time from the point of view of the vertical, this idea is illustrated as the following figure 1:

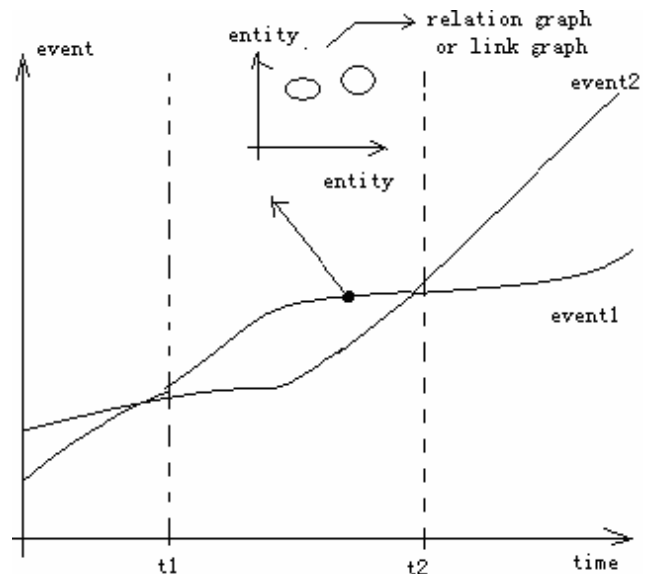


Figure 1 Relation Graph of Events Influences

The novel events include the novel words and the novel topic, because the mutual association and influences is hinted by the core characteristic words of the novel events, based on this assumption, we can study the relationship of the novel events by studying the relationship among the core characteristic words of the novel topic, that is to say, we will adopt the topic sensitive language model and the scheduling logic to resolve this problem.

For the entity relation detection of the special time, we will adopt the information extraction technology, including the corpus-based entity relation detection, which means that we will find the entity relationship in the no-local scope, while the traditional relation extraction don't use the cross-document link information, neither use the entity relation inference contained in the cross-events so comparing with the entity relation extraction, we can the get the knowledge beyond the facts from the corpus-based entity relation extraction, it will become more meaningful.

4. THE KEY TECHNIQUES

4.1 Internet New Methods Recognition ^[5]

The definition of the Chinese word is very ambiguous, it is different to give an accurate definition on the Chinese new words, we think the new characteristic of the Chinese word mainly behaves as the morphology, acceptation, the usage of word, the internet Chinese new words has several difficulties as following: (1) there is no space between the words; (2) there is no common regularity of the new words composite, except for the named entity; (3) the new words recognition of the low frequency is very difficult; (4) the new usage of

the old words is also difficult to detect.

The basic strategy of the Chinese internet new words recognition is to adopt the background and foreground corpus separated by the special time to find the new words, it is based on the following two hypothesis : (1) the occurrence of the new words has repetition; (2) the occurrence of the new words has the strong characteristic of the time.

4.2 The Language Model With Topic Sensitive Features

The goal of the language model is to try to characterize, capture, and exploit regularities in the natural language, and assign a reasonable probability distribution to all words that may follow a given word history, accordingly, it can provide some useful information for the natural language processing tasks with incomplete knowledge, such as automatic speech recognition, machine translation, information retrieval and so on. Even if traditional n-gram model is proved to be very powerful and robust in various tasks, the handicap of the language model based on the Markov assumption is presented obviously, for example, because of the data sparseness, it can't capture the long distance dependency relation among the words string, saying, dependency is limited to very short local context, some non-Markov features is out of consideration, such as lexicon triggers and cache language models. In fact, content words related to a topic can automatically trigger all the other content relevant words, the topic comprises the long-distance dependency relation that n-grams can not capture, what is more, the events with the fine granularity similarly contain the characteristic words dependency with the smaller granularity, hence, we consider the language model with topic sensitive and event sensitive to more accurately characterize the probability distribution of the core words from the different granularity so as to provide a proper words probability for natural language processing with the asymmetric information.

For language model with sensitive topic^[4], the basic model uses the Maximum entropy methods, at the same time, we adopt the internet new words recognition methods to filter the new characteristic strings facing topic sensitive from the corpus of the specific topic, then we add these new strings into the corpus to strengthen the parameter of the model so as to improve the prediction capability, then circulate in turn, we will get the more reasonable language model with topic sensitive features.

For the sensitive events language model with fine granularity, we can strengthen the more important characteristic information for event detection by extend the scope of the sensitive words, such as the person

name, place name, organization name, time and so on.

4.3 Scheduling Logic And The Interval Temporal Logic Reasoning

Advanced artificial intelligence is generally exhibited as knowledge representation and knowledge reasoning, different knowledge inference need the support of the different formal knowledge representation which can construct the different logic system.

Consider the research tasks of the TDT pilot study of the DARPA in the 1996, it is intended to investigate the state of the art in finding and following new events in a stream of broadcast news stories, and tracking the reappearance and evolution of them, Consequently explore the techniques of the information organization based on the events. Just as its name implies, the general task of the TDT is to detect the new events and track the trend and the reciprocal influences of the concerned events and, it is significant for the information security and information protection of the state. The time characteristic of the TDT is very manifest, even if the mutual influences among the events are not the strict formal logic inference, but the intuition inference depending on the common sense, we still can guide this informal logic inference by weakening formal logic inference, that is to say, adopting the interval temporal logic inference^[6].

Scheduling logic has its strict inference system, it drives from the mode logic, this non- classical logic correspond to a ternary atomic group $M(W,R,V)$, W denotes the no-empty set of the possible world, V is the value assignation corresponding to each possible world, say, for each heshi formula, it will denote whether its value is true or false in each possible world, R is the relation assigned to this model, it can be empty. If this relation is explained as the time order, then we get the logic about time, called scheduling logic, while the interval temporal logic is a method of time knowledge representation and reasoning put forward by Allen, though it is not based on the logic, it can refer to the scheduling logic, and it is widely used when published, it is proved to have the great application value.

Allen considered that interval is the basic unit of the time. Time is a kind of scaling standard about occurring sequence of the event, and the happen of the event is always a process, so the interval is not zero, designating by two time point t^- , t^+ , $t^- < t^+$. the relation between intervals has many types, but we can not directly get it, which is extracted by common sense, in addition, relations can comprise the new relation. Allen looked the relation on as a kind of constraint, the combination of the relation can be regarded as the extension of the constraint. Generally speaking, when the relations between intervals is not uniquely

ascertained, we can label all the possible relations on the directed arc, if receiving the new time relation, this uncertainty will decrease step by step. It is a process of building time interval network and reasoning (propagating the time constraint relation) through it.

There are several points to explain: firstly, the time relation summarized from the informal story plots is always uncertain, especially much detail is not depicted. Secondly, the intersection of some intervals is empty, it indicates that the constraints have the conflicts. this discovers a shortcoming of the Allen's method, as an event can be located in the two interval from the common sense, while the Allen assumed that an event can only be located in one interval, so we can consider to improve and extend his theory in the research of the practical question.

5. THE CONCLUSION AND THE FUTURE WORK

TDT is an important area of the research, addressing central application needs, it presents new and interesting technical challenges, the enormous progress of the annual TDT evaluation demonstrated anew the virtue of the formal research task definitions, common data, and common evaluations, at the same time, clearly defined technical tasks and accurately labeled corpus made it possible to move forward and conduct meaningful research.

In the future, we will try to explore the more effective language model for the topic detection and

tracking, including strengthen the capability of the Chinese internet new words and the scheduling logic and interval temporal reasoning.

6. REFERENCES

- [1] James Allan, Jaime Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang, Topic Detection and Tracking Pilot Study: Final Report, *In Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, pages 194-218, San Francisco, CA, 1998, Morgan Kaufmann Publishers, Inc p2-3
- [2] Charles L. Wayne. Multilingual Topic Detection and Tracking: Successful Research Enable y Corpora and Evaluation
<http://www.nist.gov/speech/tests/tdt/index.htm>
2005-5-20 p1-2
- [3] Jonathan Fiscus, Barbara Wheatley. Overview of the TDT 2004 Evaluation and Results, *TDT2004 Workshop*, p2-2, 2004
- [4] Jun Wu A Maximum Entropy. Language Model with Topic Sensitive Features
<http://www.cs.jhu.edu/~junwu/memodell.ps>, 2005-6-22. p13-16
- [5] ZouGang, Liuyang, LiuQun. Internet-oriented Chinese New Words Detection JCIP 2004 p2-6
- [6]. Lu Ruqian .Artificial Intelligence, *Science Publisher* , p424-45, 2002

 文章信息：

He Ruifang, Qin Bing, Liu Ting, Li Sheng. The Topic Detection and Tracking with Topic Sensitive Language Model. International Conference on Multilingual Information Processing, 2005, pp324-327