

Major Topic Detection and Its Application to Opinion Summarization

Lun-Wei Ku, Li-Ying Lee, Tung-Ho Wu, and Hsin-Hsi Chen

Department of Computer Science and Information Engineering

National Taiwan University

Taipei, Taiwan

{lwku, lylee, dhwu}@nlg.csie.ntu.edu.tw; hhchen@csie.ntu.edu.tw

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval-*Selection process*.

General Terms

Algorithms, Design, Experimentation.

Keywords

Opinion Summarization, Sentence Retrieval, Topic Detection.

1. Introduction

Watching specific information sources and summarizing the newly discovered opinions is important for governments to improve their services and companies to improve their products [1, 3]. Because no queries are posed beforehand, detecting opinions is similar to the task of topic detection on sentence level. Besides telling which opinions are positive or negative, identifying which events correlated with such opinions are also important. This paper proposes a major topic detection mechanism to capture main concepts embedded implicitly in a relevant document set. Opinion summarization further retrieves all the relevant sentences related to the major topic from the document set, determines the opinion polarity of each relevant sentence, and finally summarizes positive and negative sentences, respectively.

2. Representative Terms of an Implicit Topic

Choosing representative words that can exactly present the main concepts of a relevant document set is the spirit of our major topic detection. A term is considered to be representative if it appears frequently across documents or appears frequently in each document [2]. Such terms form the major topic of the relevant document set. How to choose the major topic is described as follows. We assign weights to each word both at document level and paragraph level. In the following formulas, W denotes weights; S is document level while P is paragraph level. TF is term frequency, and N is word count. In the subscripts, symbol i is the document index, symbol j is the paragraph index, and symbol t is the word index. Formulas (1) and (2) compute TF*IDF scores of term t in document i and paragraph j , respectively. Formulas (3) and (4) denote how frequently term t appears across documents and paragraphs. Formulas (5) and (6) denote how frequently term t appears in each document and in each paragraph.

$$W_{S,t} = TF_{S,t} \times \log \frac{N}{N_{S_i}} \quad (1)$$

$$W_{P,t} = TF_{P,t} \times \log \frac{N}{N_{P_j}} \quad (2)$$

$$Disp_{S_t} = \sqrt{\frac{\sum_{i=1}^m (W_{S,t} - mean)^2}{m}} \times TH \quad (3)$$

$$Dev_{S,t} = \frac{W_{S,t} - mean}{Disp_{S_t}} \quad (4)$$

$$Disp_{P_t} = \sqrt{\frac{\sum_{j=1}^n (W_{P,t} - mean)^2}{n}} \quad (5)$$

$$Dev_{P,t} = \frac{W_{P,t} - mean}{Disp_{P_t}} \quad (6)$$

A term is thought as representative if it satisfies either Formulas (7) or (8). Terms satisfying Formula (7) tend to appear in few paragraphs of many documents, while terms satisfying Formula (8) appear in many paragraphs of few documents. The score of a term, defined as the absolute value of $Dev_{P,t}$ minus $Dev_{S,t}$, measures how significant it is to represent the main concepts of a relevant document set.

$$Disp_{S_t} \leq Disp_{P_t} \quad \exists S_i, \forall P_j \in S_i, Dev_{S,t} \leq Dev_{P,t} \quad (7)$$

$$Disp_{S_t} > Disp_{P_t} \quad \exists S_i, \forall P_j \in S_i, Dev_{S,t} > Dev_{P,t} \quad (8)$$

Comparing with Fukumoto and Suzuki [2], we modify the scoring function in paragraph level. All documents in the same corpus are concatenated into a bigger one, i.e., the original boundaries between documents are neglected, so that words which repeat frequently among paragraphs will be chosen. We also use threshold TH to control the number of representative terms in a relevant corpus. The larger the threshold TH is, the more the number of terms will be included. The value of this parameter is trained in the experiments.

3. Relevant Sentence Retrieval

Relevant sentence retrieval aims to retrieve sentences satisfying a specific topic from a document collection. In our experiment, those sentences that contain a sufficient number of representative terms are considered as relevant sentences to the major topic. WordNet synonyms are adopted to expand the coverage of representative terms. All relevant sentences are collected in a topical set and are called *topical sentences* hereafter.

The test beds of novelty track in TREC (Text REtrieval Conference) 2003 and 2004 [4] are used as the experiment corpus to verify the performance of our relevant sentence retrieval. There are 50 document sets in 2003 TREC novelty corpus, and each set has 25 documents. All documents in the same set are relevant. That meets the assumption of our major topic detection. Take set 2 (“clone Dolly sheep”) as an example. It discussed the feasibility of the gene cloning and the perspectives of the authority. The extracted terms and the related scores to represent the major topic of this set are shown in the left part of Table 1. For comparison, the original topic narrative is listed in the right part.

Table 1. Major Terms Extracted vs. Original Topic Narrative

adult	232.64	To be relevant information there must be specific reference to “Dolly” or “the first cloned sheep” or “large animal.” References to Dolly's children are relevant if Dolly's name is included. Mention of the company that cloned Dolly is not relevant if nothing more is said about Dolly. References to the consequences of her being a clone are relevant. Mention of Polly and Molly are not relevant.
aging	174.52	
dolly	168.08	
DNA	153.73	
cell	153.19	
nucleus	131.25	
genetic	106.91	
human	99.24	
medical	82.58	
lamb	63.89	
clone	27.92	

Because we employ the terms of the major topic found in Section 2 rather than the topic given in TREC novelty corpus, we can examine the performance of the relevant sentence identification to verify the accuracy of the major topic detection. The more accurate the relevant sentence retrieval is, the more precise the major topic detection is. Sentences with more qualified terms are considered as relevant. Table 2 shows the experiment results. The average F-measure is 0.617 for 2003 TREC novelty corpus, which is better than most runs submitted by the participants [4]. It shows that our major topic identification can capture the concepts embedded in documents.

Since a “majority” approach is adopted, the system performance depends on the quantity of relevant documents. Experiments on 2004 TREC novelty corpus demonstrate the point. This corpus contains relevant and non-relevant documents in the same set. If all documents are used, F-measure is only around 0.38 (TREC 2004 A). However, if non-relevant documents are filtered in advance, F-measure 0.46 is achieved (TREC 2004 B).

4. Opinion Summarization

The opinion-oriented sentences are extracted from topical set and their tendencies are determined. The procedure is as follows.

For every topical sentence

For every sentiment word in this sentence

If a negation operator appears nearby, reverse the sentiment tendency. Every sentiment word contributes its sentiment score to this sentence.

Decide the opinion tendency of a sentence by the functional composition of sentiment words.

Sentiment words are indispensable to decide opinion orientation. If the score is positive/negative, the sentence is positive/negative-oriented. Besides, we also consider opinion keywords, e.g., “say”, “present”, “show”, “suggest”, etc. If a sentence contains such opinion keywords which follow a named entity with zero opinion score, it is regarded as a neutral opinion.

Each sentence has two scores denoting topical degree and opinion-oriented degree. We distinguish positive and negative

documents. A document is positive if it has more positive-topical sentences than negative-topical ones; and vice versa. Among positive and negative documents, two types of opinion summarizations are proposed, that is, brief and detailed opinion summary. For brief summary, we pick up the document with the largest number of positive or negative sentences and use its headline to represent the overall summary of positive-topical or negative-topical sentences. For detailed summary, we list positive-topical and negative-topical sentences with higher opinion-oriented degree. Examples are shown in Tables 3 and 4.

Table 2. TREC 2003 & 2004 Relevant Sentences Retrieval

	Precision	Recall	F-Measure
TREC 2003	0.56	0.85	0.617
TREC 2004 A	0.28	0.87	0.38
TREC 2004 B	0.34	0.86	0.46

Table 3. Brief Opinion Summary

Positive	Chinese Scientists Suggest Proper Legislation for Clone Technology
Negative	UK Government Stops Funding for Sheep Cloning Team

Table 4. Detailed Opinion Summary

Positive	Ahmad Rejai Al-Jundi, Assistant Secretary General of the Islamic Organization, declared earlier that the seminar would be aimed at shedding light on medical and legal aspects of the internationally controversial issue and seeking to take a stand on it.
Negative	Dolly the cloned sheep is only 3, but her genes are already showing signs of wear and she may be susceptible to premature aging and disease -- all because she was copied from a 6-year-old animal, Scottish researchers say.

5. Conclusion

A four layered opinion summarization system is proposed in this paper, including major topic detection, relevant sentence retrieval, opinion-oriented sentence identification, and summarization. Our system achieved 61.7% F-measure using 2003 TREC novelty corpus, and 46% in 2004 TREC novelty corpus if non-relevant documents are filtered in advance.

This system provides an opinion-based summarization. Major topics and the summarization of the corresponding opinions from a large quantity of documents are very useful for the government, institutes, companies, and the concerned public.

6. REFERENCES

- [1] Dave, K., Lawrence, S., and Pennock, D.M. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. *WWW 2003*, pp 519-528.
- [2] Fukumoto, F. and Suzuki, Y. Event Tracking based on domain dependency. *SIGIR 2000*, pp 57-64.
- [3] Morinaga, S., Yamanishi, K., Tateishi, K. and Fukushima, T. Mining product reputations on the web. *ACM SIGKDD 2002*, pp 341-349.
- [4] Soboroff, I. and Harman, D. Overview of the TREC 2003 novelty track. *The Twelfth Text REtrieval Conference*, National Institute of Standards and Technology, pp 38-53.