Knowl Inf Syst (2014) 38:691–716
DOI 10.1007/s10115-012-0604-8

REGULAR PAPER

Comparative news summarization using concept-based optimization

Xiaojiang Huang · Xiaojun Wan · Jianguo Xiao

Received: 19 March 2012 / Revised: 5 November 2012 / Accepted: 27 December 2012 /
Published online: 18 January 2013

Abstract Comparative news summarization aims to highlight the commonalities and differences between two comparable news topics by using human-readable sentences. The summary ought to focus on the salient comparative aspects of both topics, and at the same time, it should describe the representative properties of each topic appropriately. In this study, we propose a novel approach for generating comparative news summaries. We consider cross-topic pairs of semantic-related concepts as evidences of comparativeness and consider topic-related concepts as evidences of representativeness. The score of a summary is estimated by summing up the weights of evidences in the summary. We formalize the summarization task as an optimization problem of selecting proper sentences to maximize this score and address the problem by using a mixed integer programming model. The experimental results demonstrate the effectiveness of our proposed model.

Keywords Comparative news summarization · Comparative text mining · Multi-document summarization · Mixed integer programming

1 Introduction

Along with the development of Internet and web media, we are able to know all events around the world. It becomes a critical problem for readers to get useful information efficiently from

X. Huang · X. Wan (✉) · J. Xiao
Institute of Computer Science and Technology, Peking University, Beijing 100871, China
e-mail: huangxiaojiang@pku.edu.cn

X. Wan
e-mail: wanxiaojun@pku.edu.cn

J. Xiao
e-mail: xiaojianguo@pku.edu.cn

X. Huang · X. Wan · J. Xiao
Key Laboratory of Computational Linguistics, Peking University, MOE, Beijing, China

these mass data, and a lot of work has been done to help achieve this goal. News topic detection is one of the most widely used technologies by many news websites and search engines. It groups articles into event clusters and presents them with short abstracts. Related news recommendation is also popular recently. It provides similar or related topics based on the currently browsed article. These techniques raise a great opportunity to mine useful knowledge from news documents. We can discover trends among related topics, for example, the trend of market prices in the past periods. We can learn lessons by comparing similar events, for example, what causes the revolutions in Tunisia, Egypt and Libya, and what leads to the different endings in the mining accidents in Chile and China. We can judge the pros and cons of competitors according to related words, for example, which of Putin and Medvedev is better to be the next president of Russia. However, it requires thorough knowledge of all involved topics in order to make good comparisons, and thus, it is usually very time-consuming and labor-intensive for manual analysis. If a summary can be generated automatically to highlight the comparative information among news topics, it can obviously help people analyze those topics in a much easier and more efficient way.

Literally, a comparison identifies the similarities or differences among two or more objects. It basically consists of the three components: the compared objects, the scale (i.e., the aspect on which the objects are measured and the comparison is made), and the result (i.e., the predicate that describes the positions of objects on the comparative scale). Comparisons can be represented by using comparative sentences. They use particular syntactic forms and/or words and describe a simple relation between the compared objects, for example, equative, greater or less, non-gradable differential, etc. Comparisons can also be formed by describing each object in a text section (e.g., a sentence, or a paragraph) respectively. This kind of expression is not as explicit as the comparative sentence, but it can describe the features of compared objects in more details. For example,

(i) Chile is richer than Haiti.
(ii) Haiti is an extremely poor country.
(iii) Chile is a rich country.

Sentence (i) is a typical comparative sentence, where the objects are *Chile* and *Haiti*; the comparative scale is *wealth*, which is implied by "*richer*"; and the result is that *Chile is superior to Haiti*. Sentence (ii) does not describe any comparison when it is regarded individually, and neither does sentence (iii). However, when we take them both into account, we are implied with a comparison on the wealth of the two countries by the "*poor – rich*" pair, and we are conveyed with the similar information as sentences (i).

The comparison should meet some semantic conditions. First, the objects must be comparable, that is, they all have some common aspects and usually belong to the same concept category. For example, the sentence "*Then the righteous will shine like the sun in the kingdom of their father.*" is an allegory but not a comparison, because "*righteous*" and "*sun*" are different concepts and usually not comparable. Second, the comparative scale must be shared by all the objects. For example, "*A pigeon can fly faster than a man*" is semantically improper because men cannot fly. Third, the comparative result must describe the relation among all objects clearly. Take "*Haiti is a poor country.*" and "*India's wealth drop 32 %.*" for example. Although these two sentences both talk about the wealth of countries, they do not compose a good comparison because we cannot extrapolate the relations between Haiti and India, that is, whether India is richer than Haiti, or as poor as Haiti, or even poorer than Haiti.

A news topic consists of stories about events and activities that are directly connected to a central event. The comparative news summarization task aims to extract salient comparisons among comparable news topics and convey such information using human readable sentences.

Recently, this task has drawn much attention, and a few algorithms have been proposed. However, most previous studies have focused on comparing review opinions of products, because the aspects in reviews are easy to extract and the comparisons in reviews have simple patterns, for example, positive vs. negative. In contrast, the aspects are much more diverse in news documents. They can be the time of the events, the persons involved, or the attitudes of participants, etc. These aspects can be expressed explicitly or implicitly in various ways. These issues raise great challenges to comparative summarization in the news domain.

In this study, we propose a novel approach for comparative news summarization. A good comparative summary should contain sentences that convey both the comparative information among topics and the representative information about each individual topic. A set of sentences is considered comparative if the sentences share comparative concepts, and a sentence is considered representative if it contains important concepts about the topic. We take into account these two criteria in an objective function and solve the optimization problem using linear programming. Experimental results demonstrate the effectiveness of our model, which outperforms the baseline systems in quality of comparison identification and summarization.

The rest of this article is organized as follows: first, we briefly review the related works in Sect. 2. Then, we put forward the definition of comparative news summarization task in Sect. 3. Section 4 describes our proposed approach in detail. The performance evaluation of our system follows on, and effects of key parameters are discussed in Sect. 5. Finally, Sect. 6 concludes our work and talks about the possible future works.

2 Related work

2.1 Comparative text analysis

Comparisons have been researched for a long time in the linguistic and literature fields. Many works have studied the connotation, extension, forms and usages of comparisons [20,24,32,45]. The comparative analysis has been applied in many domains, and several related academic subjects have been founded, such as comparative linguistic [2] comparative literature ([55]), comparative history [6] and comparative politics [26].

The comparative analysis is also widely used in web applications. Many electronic commerce systems, for example Amazon[1] and Newegg,[2] provide commodities comparisons on the prices and the functionalities basing on the underlying structural data. More recently, mining comparative information from unstructured data has drawn much attention. Several researchers propose to find comparable objects by using linguistic patterns [4,25] or distributional similarities [14,30]. Some studies try to identify explicit linguistic comparative sentences and extract components of comparisons from them [16,17]. Other studies make contrasts by extracting features of individual objects and then matching them up [22,42,59]. While most of studies concentrate on comparing the common aspects of objects, there are also some researches focusing on detecting the unique points of topics [52] or the novelty of documents [49], which can be considered as a special kind of comparison "*with* vs. *without*".

Researchers have developed various forms of presentations for discovered comparisons. Liu et al. [29] simplify the opinion into a real value indicating the polarity and strength of sentiment and present the comparisons using histograms. Zhai et al. [59] propose a topic

[1] http://www.amazon.com.

[2] http://www.newegg.com.

model to discover comparative themes and present them using word distributions. More recently, the comparative summary becomes popular because of its rich informativeness and high readability [22,23,42,56].

Witte and Bergler [56] introduce a contrastive summarization task to indicate the common themes across all documents and document-specific contrastive themes. They use a fuzzy set theory-based clustering algorithm to generate topic clusters. The topics that span a high percentage (e.g., $>90\,\%$) of all documents are extracted as the common topics, while the topics covered only in a subset (e.g., $<5\,\%$) of documents are extracted as distinguishing topics. The main drawback of their approach is that the distinguishing topics are not aligned, that is, they may talk about different aspects of different documents, rather than contrast the divergences of documents on the same aspects. Lerman and McDonald [23] propose a model for generating pairs of summaries that highlights differences between two products. The basic idea of the model is to reward sentences that are similar to their own product's reviews (i.e., have a low KL-divergence with respect to the own product's reviews) and different to the other product's reviews (i.e., have a high KL-divergence with respect to the other product's reviews). This model also prefers different aspects between products rather than divergences on the same aspects. Kim and Zhai [22] propose a method to extract comparable sentences from two sets of positive and negative opinions and generate a comparative summary containing a set of contrastive sentence pairs. The task is formalized as an optimization problem of maximizing the content similarity and the contrastive similarity, and the optimization is solved by using a greedy algorithm. Paul et al. [42] propose a random walk formulation called *Comparative LexRank* to score sentences and pairs of sentences from opposite viewpoints. So far, the study of comparative summarization mostly focuses on product reviews, where the aspects and sentiments are relatively easy to extract. Wan et al. [50] propose a system to summarize the differences in the news reports of the same topics in different languages by using a constrained co-ranking method. In contrast, our task focuses on summarizing commonalities and differences of two comparable topics.

2.2 Multiple document summarization

The automatic summarization aims to generate a short description that conveys important information in the original texts in natural language [33]. Several specific subtasks have been proposed. The traditional summarization task places equal emphasis on different information and provides balanced coverage. The guided summarization (also called topic-focused summarization) makes summaries that focus on some user's current context, which are usually provided in keywords and/or short narratives [56]. In comparison, the comparative summarization focuses on a particular kind of information, instead of any specific context. The updating summarization emphasizes on detecting novelty of new articles over the earlier articles [8], which can be regarded as a special kind of comparative summary.

A variety of summarization methods have been proposed recently. Generally speaking, the summarization task can be performed by extraction, which identifies important sections of the text and then produce them verbatim, or by abstraction, which involves generating novel sentences with important material. Extraction-based methods usually assign a saliency score to each sentence and then rank sentences in the document. The scores are usually computed based on statistical and linguistic features, including term frequency, sentence position, cue words, topic signature, etc. [13,35,44]. Machine learning methods have also been employed to extract sentences, including unsupervised methods [40] and supervised methods [47,58]. More recently, graph-based methods have been proposed for document summarization

[12,36,46,51,54]. These methods build a graph based on the similarity between sentences and calculate the importance of a sentence regarding global information of the graph.

So far, most of summarization models consider a sentence as an information unit. Sentences are selected under the "maximal marginal relevance" (MMR) criterion [7] to maximize the involved information while minimizing the redundancy. Mcdonald [34] adapts the MMR framework and gives an Integer Linear Programming (ILP) formulation with explicit relevance and redundancy terms. Gillick et al. [11] propose a linear programming model on sub-sentence units without explicit redundancy term. Instead, the redundancy is limited by the fact that each kind of unit is counted only once, combined with a length constraint so that the solution prefers diverse information more. In this study, we expand the concept-based ILP model in [11] for comparative news summarization.

2.3 News article analysis

News article analysis is a hot area in both the academic circle and the industrial community for several reasons. First, there are strong demands of news analysis techniques for alleviating the burdens of getting information from massive news resources. Second, the news documents are good corpora for data mining research because of their large amount and easy accessibility. Third, the high linguistic quality of news articles and the abundant linguistic phenomenon in news articles makes the news domain suitable for natural language processing.

So far, many text mining and natural language processing tasks have been applied to the news domain, including classification [38], sentiment analysis [3], summarization [5], named entity recognition [41], relation extraction [48], etc. Some specific tasks on news analysis have also been proposed. The Topic Detection and Tracking (TDT) aims to find topically related material (i.e., an event) in streams of news data [1,53]. The news trend analysis aims to study the developing behavior of the society interests, that is, determining if they change or remain considerably stable from one period to another [37]. The News Personalization generalizes personalized recommendation for each reader based on their preferences [10,19].

News analysis techniques have been successfully applied on real applications. Google News[3] is a well-known online news service, providing news categorization, topic aggregation and summarization, related news recommendation, news personalization, news retrieval, etc. There are also many similar systems, such as Bing News[4] and Yahoo! News.[5] There is a significant difference between the news summarization provided by these systems and our proposal. The summarization in these systems is based on the articles of a single topic, while we propose to generate a comparative summary of two comparable topics. To the best of our knowledge, yet there is no public accessible news service that provides comparative analysis of two news topics.

3 Problem definition

3.1 News topic comparison

A news topic is "a seminal event or activity, along with all directly related events and activities" [39]. It contains a collection of stories which discuss events and activities that

[3] http://news.google.com.

[4] http://www.bing.com/news.

[5] http://news.yahoo.com.

are directly connected to the seminal event. For example, a story about search for survivors of an earthquake will be considered to belong to the earthquake topic. However, two stories about different earthquakes are usually not considered to belong to the same news topic.

Similar events happen from time to time, and it is interesting to compare those events to discover latent knowledge. Different from comparisons of product reviews which focus on a few features of products, the comparisons of news topics involve many various aspects. For example, a news topic may involve the causes and consequences of the event, the attitudes and actions of involved persons, as well as many details of the event. Any of these aspects can be a comparative scale, if it occurs in the comparable topics simultaneously. For example, when comparing the earthquake in Haiti with the one in Chile, we can compare on the *intensity of the temblors*, the *damages in the disaster areas*, the *rescue efforts of local governments*, the *international assistances*, etc.

Note that the number of news topics in a comparison is not limited to be two. We can compare three or even more topics, for example, the "*Color Revolutions*" in *Georgia*, *Ukraine*, *Kyrgyzstan* and *Tunisia*. A strict comparison of several topics ought to contain information of all topics. Thus the more topics involved, the fewer comparisons can be extracted. A looser definition allows absence of some topics, and thus, it actually degenerates to a combination of several comparisons between two topics. In this study, we focus on the comparison of two news topics and leave the study of comparison of more topics as future work.

In addition to comparing different events, we can compare the different periods of a continuous event. For example, by comparing the Libya's turmoil before and after March, we can find completely different situations caused by NATO's Airstrikes. It is also possible to compare articles about the same topic written by different news agencies and analyze their different views and attitudes. The comparison on periods and versions of a topic emphasizes more on the differences and contradictions, while the comparison of different topics places equal balances on the commonalities and differences. In this study, we focus on the comparison of different topics.

3.2 Comparative news summarization

A comparative news summary of two comparable news topics highlights the commonalities and differences between them. It consists of two blocks of texts. Each block is concentrated on a single topic, while both blocks refer to the comparable aspects of two topics. For example, Table 1 illustrates a comparative summary about two mine accidents. The left column describes the Chilean copper mine accident, while the right columns describes the New Zealand coal mine accident. Both blocks mention the *names of the mines* (*San José copper–gold mine* vs. *Pike River Coal mine*), the *numbers of victims* (33 *miners* vs. *29 workers*), the *efforts of rescues* (*drill a new hole … to extract the miners* vs. *wait for tests … before entering the coal mine*), and the *endings of accidents* (*brought safely to the surface* vs. *died after … blast*).

Formally, let $topic_1, topic_2$ be two comparable news topics, where each topic is described by a document collection $D_i (i = 1, 2)$. The task of comparative summarization is to extract two blocks of texts $B_1 = S_{11} \cup \cdots \cup S_{1n},\ B_2 = S_{21} \cup \cdots \cup S_{2n}$, where $S_{ij} \subset D_i (i = 1, 2)$ is a set of representative sentences about $topic_i$ on the aspect $aspect_j$. In other words, S_{1j} and S_{2j} describe a comparison of the two topics on $aspect_j$. Meanwhile, the length (i.e., the number of words) of the summary should not exceed a limit L.

Generally speaking, a good comparative summary should meet several criteria. First, the summary ought to focus on the comparison between news topics. Second, the information in the summary should be salient and representative to the news topics. Third, the summary

Table 1 Comparative summary about "*Chilean mine accident* versus *New Zealand mine accident*"

On August 5, 33 miners were trapped more than 700 m (2,300 ft) underground, in the San José copper–gold mine, located about 40 kilometers north of Copiapó, Chile. Rescue of the miners will take 3–4 months, given the instability of the mine and the time needed to drill a new hole, 2.5 feet in diameter, to extract the miners. After 69 days trapped deep underground, all 33 men were brought safely to the surface on 13 October 2010 over a period of almost 24 h	Up to 29 workers were trapped by the explosion took place at the Pike River Coal mine, 160 miles west of Christchurch. Police said the miners, aged 17–62, are believed to be about 1.2 miles (2 km) down the main tunnel. Rescue workers have been forced to wait for tests to show the air is safe before entering the coal mine. All 29 men missing in a New Zealand coal mine have died after a powerful second blast tore through the pit

should convey as much information as possible within a length limit. Finally, the summary must have good linguistic quality, that is, it should be fluent and can be easily understood by human.

4 Proposed approach

A naive idea of comparative summarization is to extract the comparative sentences from original articles. Unfortunately, it is not practical because the comparative sentences do not appear frequently in news articles, and those which are relevant to desired competitors are even fewer. Instead, we extract proper non-comparative sentences which talk about the similar aspects in the two topics from the news documents and organize them appropriately to form comparisons.

The processing procedure of our comparative summarization system is illustrated in Fig. 1. The main steps include pre-processing, sentence selection and sentence ordering. The pre-processing step cleans the texts and segments them into information units. The sentence selection step extracts proper sentences from the original documents by considering both information representativeness and comparativeness. The sentence ordering step reorganizes the sentences in the summary to improve the readability of the summary. The details of each step are described in the following subsections.

4.1 Pre-processing

The pre-processing step aims to clean the texts and extract information units from the texts. In theory, the semantic information of a text is related to the meaning of words, the syntactic structure of the sentence, the relation among the sentences, and even the contexts around the text. To simplify the model, we only take into account the meaning of words in the text. Similar to the bag-of-words model, we represent the text with a bag-of-concepts, where each concept is an information unit. Obviously, the more accurate the extracted concepts are, the better we can represent the meaning of a text. However, it is not easy to extract semantic concepts accurately. In this study, we use words/stems, named entities and bigrams to simply represent concepts, and leave the more complex concept extraction for future work.

In the preprocessing step, we first break each document into sentences and then tokenize each sentence into words and recognize the named entities (We use the Stanford CoreNLP toolkit[6] for NER in English, and an in-house CRF-based NER tool for NER in Chinese. Other

[6] http://nlp.stanford.edu/software/corenlp.shtml.

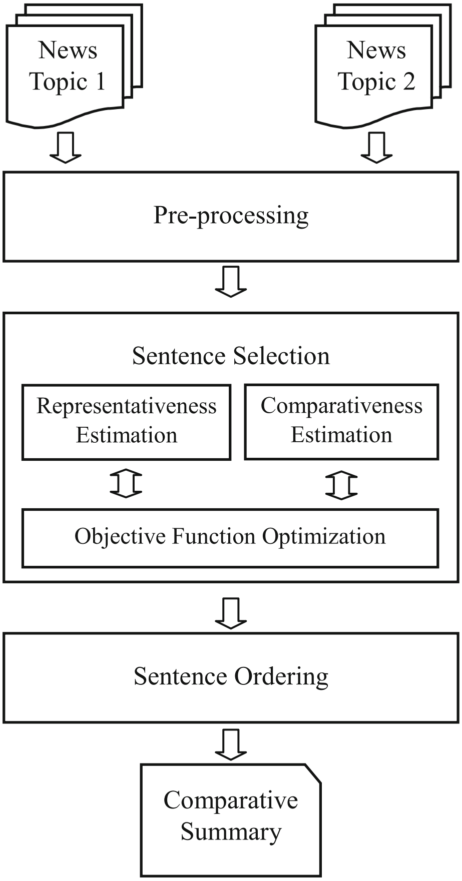


Fig. 1 The processing procedure of proposed system

NER toolkits, for example GATE[7] and LinPipe,[8] will also work). The common stop words are discarded. For English documents, the stems are used instead of the original words. We filter out the sentences that contain no more than three non-stop words. The unigram and bigram of tokens are extracted, and the frequencies of their occurrences are counted.

4.2 Sentence selection

The step of sentence selection aims to extract proper sentences from original documents to compose a comparative summary. The comparative summary ought to be representative to the information in each topic, and emphasize the comparisons between the two topics. We formalize the sentence selection as an optimization problem of selecting sentences to maximize the representativeness and the comparativeness of the summary. The estimations of a summary's representativeness and comparativeness, as well as the optimization model, are described in the following texts, respectively.

[7] http://gate.ac.uk/.

[8] http://alias-i.com/lingpipe/.

4.2.1 Representativeness estimation

As the essential requirement for text summarization, the comparative summary should represent the salient aspects of each individual news topic, that is, convey the important information in the documents. The representativeness of the summary can be estimated by the amount and saliency of information contained in the summary. Note that a concept is a unit of information. Therefore, the amount and saliency of information can be estimated by the aggregation of weights of the concepts.

Intuitively, as the keynotes of the topic, the important information will be mentioned many times across the articles. Thus the more frequently a concept occurs in the documents, the more likely it represents a salient piece of information. Based on this assumption, we can estimate a concept's saliency with its frequency of occurrences. Formally, the weight w_{ij} of a concept c_{ij}, is calculated as follows:

$$w_{ij} = freq\left(c_{ij}, D_i\right) \cdot idf\left(c_{ij}\right) \tag{1}$$

where *freq*(c_{ij}, D_i) is the frequency of occurrences of c_{ij} in D_i; *idf* (c_{ij}) is the inverse document frequency of c_{ij}that penalizes the topic-independent common words, as defined in Eq. 2:

$$idf\left(c_{ij}\right) = \log \frac{|D_B|}{1 + \sum_{d \in D_B} I\left(c_{ij}, d\right)} \tag{2}$$

where D_B is a large background corpus, $|D_B|$ is the amount of documents in D_B, and $I(c_{ij}, d) \in \{0, 1\}$ is an indicator function of whether c_{ij} occurs in document d:

$$I\left(c_{ij}, d\right) = \begin{cases} 1, & \text{if } freq\left(c_{ij}, d\right) > 0 \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

The representativeness of a comparative summary *sum* is defined as the aggregation of the concepts' weights as follows:

$$Rep\left(sum\right) = \sum_{i=1}^{2} \sum_{j=1}^{|C_i|} w_{ij}\, f\left(c_{ij}, sum\right) \tag{4}$$

where C_i is the collection of concepts in D_i; $f(c_{ij}, sum)$ is a function to calculate c_{ij}'s contribution of information in the summary.

A simple form of f is the frequency function*freq*(c_{ij}, sum). The basic idea is that the more frequently a salient concept occurs in the summary, the more sentences focus on the salient aspects in the topics, and thus the more representative the summary is. The disadvantage of frequency function is that it does not penalize the redundancy of sentences of in the summary, and thus, the coverage of information will be low.

Another form of f is the indicator function $I(c_{ij}, sum)$, as used in [11]. The advantage of indicator function is that it only counts each concept once, so that the optimized summary tends to contain sentences without overlapping concepts. However, because the summary should focus on some certain topics, a few overlaps of topic words are necessary. For example, many sentences in the summary about an earthquake will contain the word "*earthquake*", because it is the topic' centroid concept that is inescapable.

To balance the information saliency and redundancy, the f should increase as the frequency grows, but the growth rate of f should decrease gradually, as illustrated in Fig. 2.

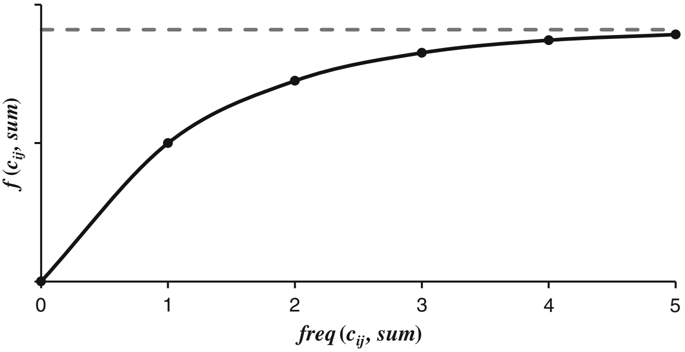


Fig. 2 Graph of function $f(x)$

In this study, we defined f as follows

$$f\left(c_{ij}, sum\right) = \begin{cases} 0 & freq\left(c_{ij}, sum\right) = 0 \\ \sum_{k=0}^{freq(c_{ij},sum)-1} \alpha^k, & 0 < freq\left(c_{ij}, sum\right) \leq N \\ \sum_{k=0}^{N} \alpha^k, & freq\left(c_{ij}, sum\right) > N \end{cases} \tag{5}$$

where $\alpha \in [0, 1]$ (we define $0^0 = 1$ here). The lower α is, the less a redundant concept instance contributes, and thus the summary tends to contain less redundancy. When $\alpha = 0$, then f degenerates to the indicator function. When $\alpha = 1$, then f is actually the frequency function which does not avoid any redundancy. N controls the upper boundary of a concept's contribution. In this study, we set N= 3.

4.2.2 Comparativeness estimation

According to the definition of comparison, a set of texts can form a comparison only if they discuss the common aspects of objects. For example,

> *Lionel Messi named FIFA World Player of the Year 2010. Cristiano Ronaldo Crowned FIFA World Player of the Year 2009.*

The above two sentences compare on the aspect "*FIFA World Player*", which is contained in both sentences. Furthermore, semantic-related concepts can also represent comparisons. For example, "*snow*" and "*sunny*" can indicate a comparison on *weather*; "*alive*" and "*death*" can imply a comparison on *rescue result*. If two sentences share a pair of semantic-related concepts, we consider them as potential comparisons. In other words, semantic-related concept pairs are evidences of comparisons.

Formally, let $C_i = \{c_{ij}\}$ (i =1, 2) be the set of concepts in the document set D_i. A sentence $s_{ij} \in D_i$ is represented with a subset of C_i. Then a pair of sentences $\langle s_{1a}, s_{2b}\rangle$ is a potential comparison iff

$$\exists c_{1k_1} \exists c_{2k_2} \left(c_{1k_1} \in s_{1a} \wedge c_{2k_2} \in s_{2b} \wedge rel\left(c_{1k_1}, c_{2k_2}\right) \geq \tau\right)$$

where $rel\left(c_{1k_1}, c_{2k_2}\right)$ is the semantic relevance between c_{1k_1} and c_{2k_2}, and τ is a minimal threshold. The c_{1k_1} and c_{2k_2} make up an evidence of the comparativeness between s_{1a} and s_{2b}. In the latter of this paper, we denote a comparative evidence, that is, a semantically related concept pair, as $ce_k = \langle c_{1k_1}, c_{2k_2}\rangle$, where $c_{1k_1} \in D_1$ and $c_{2k_2} \in D_2$.

To extract these comparative evidences, we extract the concepts from the documents and then calculate the semantic relevance between each couple of concepts. The semantic relevance between two English words is calculated using the algorithms based on WordNet [43]. The relevance between two Chinese words is calculated based on Hownet [31]. The relevance between two out-of-vocabulary bigrams $wd_{11}wd_{12}, wd_{21}wd_{22}$ is calculated as $[rel\,(wd_{11}, wd_{21}) + rel\,(wd_{12}, wd_{22})]\,/2$, where $wd_{ij}\,(i,\ j = 1, 2)$ is a word. The collection of obtained comparative evidences (i.e., semantic-related concept pairs) is denoted as $CE = \{ce_k\}$.

The weight u_k of a comparative evidence $ce_k = \langle c_{1k_1}, c_{2k_2}\rangle$ is calculated as:

$$u_k = \frac{w_{1k_1} + w_{2k_2}}{2} \tag{6}$$

where w_{1k_1} and w_{2k_2} are the weights of c_{1k_1} and c_{2k_2}, calculated using Eq. 1, respectively.

The comparativeness of the summary *sum* is calculated as:

$$Cmp\,(sum) = \sum_{k=1}^{|CE|} u_k g\,(ce_k, sum) \tag{7}$$

where $g(ce_k, sum)$ is a function to calculate ce_k's contribution of information. Similar to the $f(c_{ij}, sum)$ discussed in Sect. 4.2.1, the $g(ce_k, sum)$ is defined as follows:

$$g\,(ce_k, sum) = \begin{cases} 0, & freq\,(ce_k, sum) = 0 \\ \sum_{i=0}^{freq(ce_k,sum)-1} \alpha^i, & 0 < freq\,(ce_k, sum) \le N \\ \sum_{i=0}^{N} \alpha^i, & freq\,(ce_k, sum) > N \end{cases} \tag{8}$$

where $freq(ce_k, sum)$ is the frequency of ce_k's occurrences in the summary, that is,

$$freq\,(ce_k, sum) = Min\left(freq(c_{1k_1}, sum), freq(c_{2k_2}, sum)\right) \tag{9}$$

4.2.3 Concept-based comparative summary model

To highlight the comparison among news topics, a comparative summary should contain as many salient comparative evidences as possible. Besides, it should represent the topics well, that is, convey the important information of each individual news topic. Thus, the score of a comparative summary can be calculated as the sum of the comparativeness score and the representativeness score:

$$\begin{aligned} Score\,(sum) &= (1-\lambda)\cdot Rep\,(sum) + \lambda\cdot Cmp\,(sum) \\ &= (1-\lambda)\cdot \sum_{i=1}^{2}\sum_{j=1}^{|C_i|} w_{ij} f\left(c_{ij}, sum\right) + \lambda\cdot\sum_{k=1}^{|CE|} u_k g\,(ce_k, sum) \end{aligned} \tag{10}$$

where $\lambda \in [0, 1]$ is a coefficient that balances the comparativeness and representativeness.

The summary consists of several sentences extracted from the original documents. The aim of sentence selection is to maximize the score of the summary by selecting proper sentences, that is,

$$ComparativeSummary\,(D_1, D_2) = \mathop{\arg\max}_{sum*\subset D_1\cup D_2} Score\,(sum*) \tag{11}$$

The objective function can be optimized using the linear programming algorithm. Let f_{ij} be a numeric variable that indicates the function value of $f(c_{ij}, sum)$, and g_k be a numeric variable that indicates the function value of $g(ce_k, sum)$. Then the optimization subjection can be defined as:

$$\text{Max } (1-\lambda)\sum_{i=1}^{2}\sum_{j=1}^{|C_i|} w_{ij} f_{ij} + \lambda \sum_{k=1}^{|CE|} u_k g_k \tag{S.1}$$

Let n_{ij} be an integer variable denoting the value of *freq*(c_{ij}, sum). The piecewise function Eq. 5 is equivalent to the following constraints (under $N = 3$):

$$f_{ij} \leq n_{ij} \tag{C.1}$$

$$f_{ij} \leq \alpha \cdot n_{ij} + 1 - \alpha \tag{C.2}$$

$$f_{ij} \leq \alpha^2 \cdot n_{ij} + 1 + \alpha - 2\alpha^2 \tag{C.3}$$

$$f_{ij} \leq \alpha^2 + \alpha + 1 \tag{C.4}$$

Similarly, let m_k be an integer variable denoting the value of *freq*(ce_k, sum). According to Eq. 8, each g_k should satisfy the following constraints (under $N = 3$):

$$g_k \leq m_k \tag{C.5}$$

$$g_k \leq \alpha \cdot m_k + 1 - \alpha \tag{C.6}$$

$$g_k \leq \alpha^2 \cdot m_k + 1 + \alpha - 2\alpha^2 \tag{C.7}$$

$$g_k \leq \alpha^2 + \alpha + 1 \tag{C.8}$$

According to Eq. 9, m_k should satisfy the following constraints:

$$m_k \leq n_{1k_1} \tag{C.9}$$

$$m_k \leq n_{2k_2} \tag{C.10}$$

Furthermore, let o_{iq} be a binary variable indicating whether the sentence s_{iq} is presented in the summary, and $I(c_{ij}, s_{iq})$ be a binary constant indicating whether the concept c_{ij} occurs in the sentence s_{iq}, then the frequencies of concepts should meet the following constraints:

$$n_{ij} = \sum_{s_{iq} \in D_i} I\left(c_{ij}, s_{iq}\right) \cdot o_{iq} \tag{C.11}$$

Finally, the summary should satisfy a length constraint:

$$\sum_{i=1}^{2} \sum_{s_{iq} \in D_i} l_{iq} \cdot o_{iq} \leq L \tag{C.12}$$

where l_{iq} is the length of sentence s_{iq}, and L is the maximal summary length.

Note that the variables f_{ij}, g_k, n_{ij},m_k and o_{iq} are all linear in S.1 and C.1–C.12, and their coefficients are all constants. Thus, the optimization problem is a mixed integer programming (MIP) problem. Though the MIP problems are generally NP-hard, considerable works have been done [9,18,21,57], and several software solutions have been released to solve them efficiently. In this study, we use the IBM ILOG CPLEX Optimizer[9] to solve this problem. Other MIP optimizer, for example Gurobi Optimizer[10] and GPLK,[11] will also work.

4.3 Sentence ordering

For better intelligibility, it is necessary to organize the summary according to the comparative aspects. In this study, we use an aggregative clustering algorithm [15] to group the sentence into several bunches based on the similarity of the sentences. A sentence in the summary is represented as a weighted vector of the concepts that is contained in the sentence and the comparative concept pairs where one of the concepts is contained in the sentence. The similarity of two sentences is calculated using the cosine value of the two vectors. After the clusters are obtained, we order them according to the numbers of sentences they contain. The final summary contains two blocks. Each block consists of sentences which are selected from the same document set. In each block of the summary, we organize the sentences according to the order of the clusters which they belong to, that is,

$$\forall s_i \in cluster_a \forall s_j \in cluster_b : cluster_a \prec cluster_b \rightarrow s_i \prec s_j$$

where $x \prec y$ means x is arranged before y.

5 Experiment

5.1 Dataset

Because of the novelty of the comparative news summarization task, there is no existing dataset yet for evaluation, and thus, we create our own. We first choose twenty pairs of comparable topics (ten in English and ten in Chinese, as shown in Tables 2 and 3) and then retrieve ten related news articles for each topic using the Google News search engine.[12] Finally, we write the comparative summary for each topic pair manually. Note that every reference summary also contains two blocks, each of which concentrates on a single topic in the pair.

[9] http://www-01.ibm.com/software/integration/optimization/cplex-optimizer/.

[10] http://www.gurobi.com/.

[11] http://www.gnu.org/software/glpk/.

[12] http://news.google.com.

Table 2 Comparable topic pairs in the English dataset

ID	Topic 1	Topic 2
1	Haiti earthquake	Chile earthquake
2	Chilean mining accident	New Zealand mining accident
3	Nuclear accident of Fukushinia	Chernobyl nuclear disaster
4	Iraq withdrawal	Afghanistan withdrawal
5	Iran nuclear issue	Korea nuclear issue
6	Libya Turmoil	Syria Turmoil
7	Nobel price 2009	Nobel prize 2010
8	Oscar/academy award 2009	Oscar/academy award 2010
9	Apple iPad 2	BlackBerry playbook
10	2006 FIFA world cup	2010 FIFA world cup

Table 3 Comparable topic pairs in the Chinese dataset

ID	Topic 1	Topic 2
1	汶川地震 (Wenchuan Earthquake)	玉树地震 (Yushu Earthquake)
2	王家岭矿难 (Wangjialing Mining Accident)	智利矿难 (Chilean Mining Accident)
3	福岛核事故 (Nuclear accident of Fukushinia)	切尔诺贝利核事故 (Chernobyl Nuclear Disaster)
4	伊拉克撤军 (Iraq Withdrawal)	阿富汗撤军 (Afghanistan Withdrawal)
5	伊朗核问题 (Iran Nuclear Issue)	朝鲜核问题 (Korea Nuclear Issue)
6	突尼斯骚乱 (Tunisia Turmoil)	埃及骚乱 (Egypt Turmoil)
7	2009年诺贝尔奖 (Nobel Prize 2009)	2010年诺贝尔奖 (Nobel Prize 2010)
8	2009年奥斯卡奖 (Oscar Award 2009)	2010年奥斯卡奖 (Oscar Award 2010)
9	苹果 iPad 2 (Apple iPad 2)	黑莓 Playbook (BlackBerry Playbook)
10	2006年世界杯 (World Cup 2006)	2010年世界杯 (World Cup 2010)

5.2 Evaluation metrics

Comparative Aspect Recall (CAR): It is defined as the number of human agreed comparative aspects in the summary. This metric evaluates the performance of comparative extraction.
Overall Responsiveness (OR): The assessors will give an overall responsiveness score to each summary, based on both content and readability/fluency. It is judged on the 5-point scale indicating very poor, poor, barely acceptable, good, very good, respectively.
ROUGE: The ROUGE is a widely used metric in summarization evaluation. It measures summary quality by counting overlapping units between the candidate summary and the reference summary [27,28]. The *n-gram* based ROUGE value is calculated as follows:

$$ROUGE\text{-}N = \frac{\sum_{S\in\{RefSum\}} \sum_{n-gram\in S} Count_{match}\,(n-gram)}{\sum_{S\in\{RefSum\}} \sum_{n-gram\in S} Count\,(n-gram)} \tag{12}$$

where n stands for the length of the n-gram, $Count_{match}(n\text{-}gram)$ is the maximum number of n-grams co-occurring in a candidate summary and a set of reference summaries, and *Count* (*n-gram*) is the number of n-grams in the reference summaries. The *ROUGE-S* is similar to *ROUGE-N* ($N = 2$), but based on the skip-bigrams (i.e., pairs of words in their sentence order, allowing for a certain gap) instead of the regular bigrams (i.e., pairs of continuous words). The ROUGE-SU evaluates the co-occurrence of both skip-bigrams and unigrams.

In our experiment, we use the *ROUGE*-2 (i.e., *ROUGE-N* with $N = 2$) values and the *ROUGE-SU*4 (i.e., *ROUGE-SU* allowing gaps within 4 terms) values to evaluate the systems' performance. In addition, we evaluate each block in the summary respectively and report the mean of two ROUGE values (denoted as *M-ROUGE*-2 and *M-ROUGE-SU*4) to evaluate whether the comparative summary is related to each topics.

For English dataset, we use the ROUGE toolkit[13] to calculate the values. For Chinese dataset, we first tokenize both the reference summary and the automatically generated summary and then use a modified version of ROUGE toolkit[14] to calculate the values.

5.3 Baseline systems

Baseline 1: Non-Comparative Model (NCM): The non-comparative model treats the comparative summarization task as a traditional summarization problem. It merges all the documents into a single document and selects the most salient and representative sentences as the summary. In this study, we use the model proposed in [11] because of its simplicity and good performance.

Baseline 2: Co-Ranking Model (CRM): This model is adapted from [51]. It makes use of the relations within each topic and the relations across the topics to reinforce scores of the comparison-related sentences. More specifically, a sentence's score consists of two parts, that is, the contribution of related in-topic sentences (representativeness), and the contribution of related cross-topic sentences (comparativeness).

Formally, let $s_{1i} \in D_1$ and $s_{2j} \in D_2$ denote sentences in D_1 and D_2, respectively. us_i denotes the score of s_{1i}, and vs_j denote the score of s_{2j}. $sim(s,\ s')$ denote the similarity of two sentence s and s'. Then the scores of sentences are computed as follows:

$$us_i = \theta \cdot \sum_{s_{1k} \in D_1} \frac{sim\left(s_{1i}, s_{1k}\right) \cdot us_k}{\sum_{s_{1p} \in D_1} sim\left(s_{1p}, s_{1k}\right)} + (1-\theta) \cdot \sum_{s_{2j} \in D_2} \frac{sim\left(s_{1i}, s_{2j}\right) \cdot vs_j}{\sum_{s_{1p} \in D_1} sim\left(s_{1p}, s_{2j}\right)} \quad (13)$$

$$vs_i = \theta \cdot \sum_{s_{2k} \in D_2} \frac{sim\left(s_{2j}, s_{2k}\right) \cdot vs_k}{\sum_{s_{2q} \in D2} sim\left(s_{2q}, s_{2k}\right)} + (1-\theta) \cdot \sum_{s_{1i} \in D_1} \frac{sim\left(s_{1i}, s_{2j}\right) \cdot us_j}{\sum_{s_{2q} \in D_2} sim\left(s_{1i}, s_{2q}\right)} \quad (14)$$

where $\theta \in [0, 1]$ is a parameter to balance the influence of representativeness and comparativeness. In this study, we set $\theta = 0.5$. The values of us_i and vs_j can be computed iteratively. The algorithm first assigns random initial values to us_i and vs_j and then recursively compute the new estimations of $us_i^{(n+1)}$, $vs_j^{(n+1)}$ using Eqs. 13 and 14 until the values are convergent.

After that, we estimate the score of a cross-topic sentence pair $sp_k = \langle s_{1k_1}, s_{2k_2} \rangle$ as follows:

$$score(sp_k) = \eta \cdot sim(s_{1k_1}, s_{2k_2}) + (1-\eta)(us_{k_1} + vs_{k_2}) \quad (15)$$

where $\eta \in [0, 1]$ is a factor that balance the comparativeness and the saliency of sentences ($\eta = 0.5$ in this study). The most salient sentence pairs are selected iteratively, and the scores of remained sentence pairs are updated using the MMR algorithm [7].

[13] http://www.berouge.com/.

[14] The modification only alters the word filter to allow Chinese words.

Table 4 The evaluation results on the english dataset

Model	CAR	OR	ROUGE-2	ROUGE-SU4	M-ROUGE-2	M-ROUGE-SU4
Baseline 1 /NCM	1.9	2.6	0.219	0.256	0.152	0.182
Baseline 2 /CRM	2.3	2.9	0.215	0.259	0.167	0.207
CCM	**3.5**	**3.4**	**0.274**	**0.305**	**0.221**	**0.249**

The best result in each metric is marked in bold

Table 5 The evaluation results on the Chinese dataset

Model	CAR	OR	ROUGE-2	ROUGE-su4	M-ROUGE-2	M-ROUGE-su4
Baseline 1/NCM	1.6	2.4	0.164	0.175	0.132	0.141
Baseline 2/CRM	2.0	2.7	0.181	0.203	0.170	0.184
CCM	**3.1**	**3.3**	**0.216**	**0.225**	**0.190**	**0.200**

The best result in each metric is marked in bold

5.4 Experimental results

We apply all the systems to generate comparative summaries with a length limit of 400 words. The evaluation results are shown in Tables 4 and 5. The NCM and CRM models are described in Sect. 5.3, and the concept-based comparative model (CCM model) is our proposed model. In this experiment, the λ and α in CCM are set as follows: $\lambda = 0.3$, $\alpha = 0.5$ for English dataset (Table 4), and $\lambda = 0.5$, $\alpha = 0.5$ for Chinese dataset (Table 5).

Compared with the baseline systems, our proposed model achieves the best scores over all metrics. It is not surprising that the NCM model does not perform well in this task, because it does not focus on the comparisons. The CRM model utilizes the similarity between two topics to enhance the score of comparison-related sentences. However, the sentences in a comparison usually share only a few words in common, and thus the similarities among them are low. In such cases, the co-ranking algorithm can barely benefit from the cross-topic relations. The CCM model calculates the score of summary explicitly using scores of comparative concepts and representative concepts. It balances these two factors and is not affected by the low similarities among sentences. Besides, it uses a mixed integer programming model to find a globally optimized solution. Thus it achieves good performance on both comparison extraction and summarization.

5.5 Parameter effect

In our model, there are two important parameters, λ and α. λ balances the importance of comparativeness and the importance of representativeness, while α balances the information saliency and the redundancy. Intuitively, both λ and α should be medium, neither too big nor to small. To verify this assumption, we run the system with different settings of λ and α and evaluate them using the ROUGE values.

First, we set $\alpha = 0.5$ and range λ from 0 to 1 in step of 0.1. The results in English and Chinese corpus are shown in Figs. 3 and 4, respectively. In both corpuses, the performance first increases as λ grows and then reaches top at a medium value of λ. After that, it decreases instead as λ continuously grows.

It is interesting that the performance of considering representativeness only ($\lambda = 0$) is superior to the performance of considering comparativeness only ($\lambda = 1$) in the English

Fig. 3 The system performances on the English dataset with different λ settings

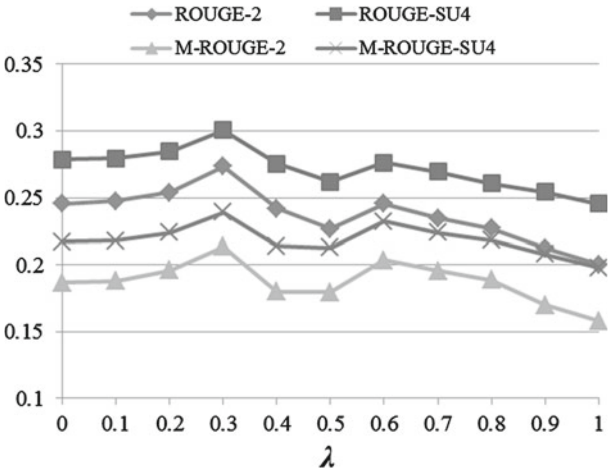
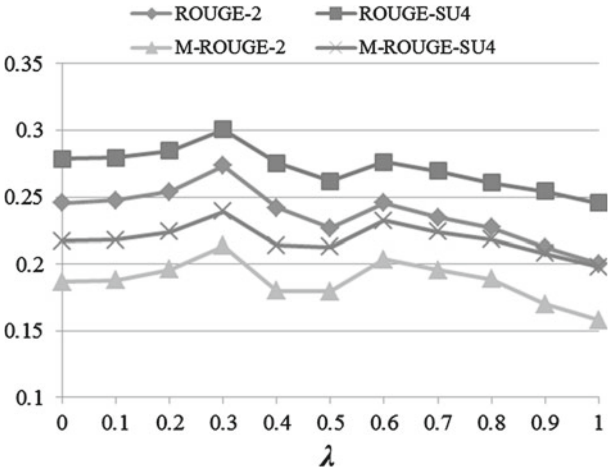


Fig. 4 The system performances on the Chinese dataset with different λ settings



dataset. The possible reason is that the English news articles contain more wide information, for example, interviews and quotations. These less important information leads to some inessential comparisons and thus harm the performance. On the other hand, the focused points of related news topics are much the same, and thus the representative summaries of each topic are likely to be comparative. The Chinese news articles mostly consist of objective descriptions of the news event, and thus the comparisons are more likely to focus on the salient aspects. Generally speaking, λ should be small (i.e., emphasize the representativeness) on the news articles of divergent themes, and be medium on the news articles of compact themes. In practice, the optimized parameters can be learned by using an evaluation dataset.

To investigate the effect of α, we set λ to the best setting according to the previous experiment and range α from 0 to 1. The results are illustrated in Figs. 5 and 6. Similar to the effects of λ, the system also performs best at a medium value of α. Notice that the fluctuation of performances in English corpus is little when $\alpha \in [0.2, 0.5]$, and thus $\alpha = 0.5$ is an acceptable setting for both corpuses.

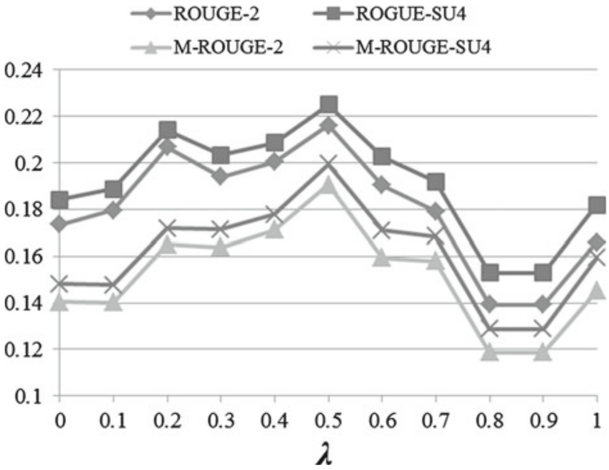


Fig. 5 The system performances on the English dataset with different α settings

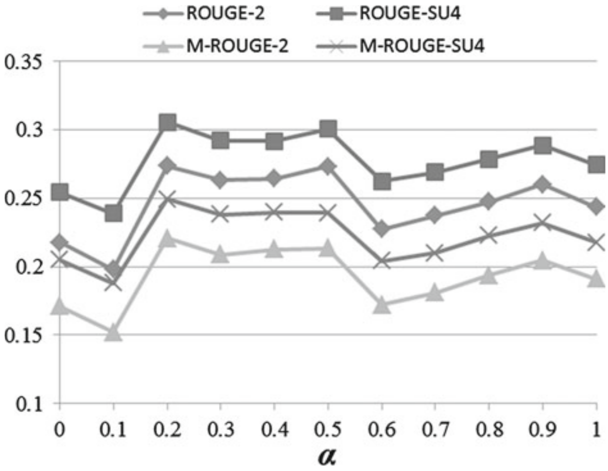


Fig. 6 The system performances on the Chinese dataset with different α settings

5.6 Case study

Tables 6 and 7 show the reference summary and the system-generated summary for *World Cup 2006* versus *World Cup 2010*, respectively. The generated summary introduces the champions, the Golden Ball winners, the Young Players of two World Cup matches, the effects to the economy, etc. However, it also makes a comparison on Golden Shoe winner and Golden Glove winner. The primary cause is the word "Golden", which is not an appropriate unit of concept. To overcome these defects, more precise concept extraction and relation extraction should be applied.

Tables 8 and 9 show the reference summary and the system-generated summary for *Wenchuan Earthquake* vs. *Yushu Earthquake*. The generated comparative summary presents the times, locations, rescue efforts & effects, and responses of foreign countries in the two earthquakes in China. Overall, it has good quantity in comparison. However, the sentences (1.8) and (1.9) are not quite appropriate to be compared with the sentence (2.7), because the former two sentences describe the responses of foreign governments, while the latter one describes the response of a foreign media. The representativeness of some aspects is a

Table 6 The reference comparative summary about two World Cup matches

World cup 2006	World cup 2010
Italy claimed a fourth world title in a penalty shoot-out victory over France after the two sides finished a goal apiece following extra-time in Berlin's Olympic Stadium on Sunday	Spain have won the 2010 FIFA World Cup South Africa final, defeating Netherlands 1-0 with a wonderful goal from Andres Iniesta deep into extra-time.
France captain Zinedine Zidane won the Golden Ball award for the tournament's best player	Uruguay star striker Diego Forlan won the Golden Ball Award as he was named the best player of the tournament at the FIFA World Cup 2010 in South Africa
Lukas Podolski was named the inaugural Gillette Best Young Player by FIFA's TSG after scoring three goals and contributing boundless energy to Germany's enthralling FIFA World Cup campaign	German youngster Thomas Mueller got double delight after his side finished third in the tournament as he was named Young Player of the World Cup by the FIFA Technical Study Group (TSG) and he also won the Golden Boot Award for the tournament's top-scorer.The net economic benefit from hosting the World Cup for South Africa, in terms of current and future tourism impact, is unclear
Germany striker Miroslav Klose was the Golden Shoe winner for the tournament's leading scorer	South Africa will have five brand new state of the art football stadiums that seat an average of 50,400 spectators and five newly renovated stadiums that seat an average of 53,300
Germany's minister of economics and technology, Michael Glos, says he is confident the World Cup will boost the economy.	In Berlin, about 3,50,000 people watched Germany at the FIFA fan fest on Wednesday night, while 56,836 people attended the fan fest in Durban
An average of 52,500 fans packed into the 12 stadiums for the 64 matches	A global TV audience of more than 700 million watched Sunday's World Cup final, according to the tournament's organizers
In Berlin, for example, police estimated that up to one million fans converged on the official Fan Fest public viewing venue in front of the Brandenburger Tor on Saturday to watch the host nation beat Sweden for a quarterfinal berth	
Television audiences for the 2006 FIFA World cup^{TM} in Germany are being collated as the tournament progresses and it already looks as if they are heading for the record books	

bit low. For example, the sentence (1.2) mentions the property damage, but does not report the exact amount of the damage. The features of representativeness need to be further studied.

6 Conclusion & future work

The comparative analysis of related news topics is useful in many applications. In this study, we propose a novel approach to summing up the commonalities and differences among comparable news topics. We formalize the task as a problem of selecting sentences to maximize both the comparativeness and the representativeness, propose an estimation function based on concept-level evidences, and solve it using linear programming. The experiment results show that our model outperforms the baseline systems in comparison extraction and summarization.

Table 7 The system-generated comparative summary about two world cup matches

World cup 2006	World cup 2010
(1.1) Italy claimed a fourth world title in a penalty shoot-out victory over France after the two sides finished a goal apiece following extra-time in Berlin's Olympic Stadium on Sunday	(2.1) Spain have won the 2010 FIFA World Cup South Africa final, defeating Netherlands 1-0 with a wonderful goal from Andres Iniesta deep into extra-time
(1.2) France captain Zinedine Zidane, sent off for head-butting Marco Materazzi late in Sunday's World Cup final loss to Italy, won the Golden Ball award for the tournament's best player	(2.2) Uruguay star striker Diego Forlan won the Golden Ball Award as he was named the best player of the tournament at the FIFA World Cup 2010 in South Africa
(1.3) Lukas Podolski was named the inaugural Gillette Best Young Player by FIFA's TSG after scoring three goals and contributing boundless to Germany's enthralling FIFA World Cup campaign	(2.3) German youngster Thomas Mueller got double delight after his side finished third in the tournament as he was named Young Player of the World Cup
(1.4) Germany striker Miroslav Klose was the Golden Shoe winner for the tournament's leading scorer	(2.4) Mueller scored five goals in the tournament just like Forlan, Sneijder and Villa
(1.5) Germany's minister of economics and technology, Michael Glos, says he is confident the World Cup will boost the economy	(2.5) Among the winners were goalkeeper and captain Iker Casillas who won the Golden Glove Award
(1.6) In 1998, the first four German games attracted a cumulative audience of 85.5 million, on average half a million viewers fewer per match than in 2006	(2.6) The net economic benefit from hosting the World Cup for South Africa, in terms of current and future tourism impact, is unclear
(1.7) England's fans brought more color than their team.	(2.7) A global TV audience of more than 700 million watched Sunday's World Cup final, according to the tournament's organizers
	(2.8) Only four of the 212 matches played drew more than 40,000 fans

Due to the complexity of news topics and semantic requirements of comparisons, the comparative summaries are still far from satisfactory. In future work, we are going to utilize more semantic information in this task to enhance the quality of summarization. First, the comparative evidences need to be further anatomized. Intuitively, adverbs will not lead to comparisons unless they modify similar actions. We will introduce syntactic structures into comparative evidence extraction in future work. We are also going to use more resources, such as Wikipedia, to calculate the semantic relatedness among concepts. Second, the representativeness can also take into account relatedness among concepts. When a concept is presented in the summary, its similar concepts are not necessary to be selected. The weights of concepts can be tuned using machine learning techniques. Finally, we are going to formalize the sentence ordering step in more compact way, in the consideration of both comparativeness and coherence.

In the future, we are also going to extend the task of comparative summarization. First, comparisons of more than two topics can be studied. Second, it is better to separate the commonality and difference of news topics and extract the key phrases indicating the compared aspects. Third, a comparison of the snapshots along the timeline of a continual event can be studied. This particular kind of comparison mainly focuses on the difference and evolutions among snapshots. Finally, it is an interesting problem to compare cross-lingual news topics, where contradictions are the most important information.

Table 8 The reference comparative summary about two earthquakes

Wenchuan Earthquake	Yushu Earthquake
北京时间5月12日14时28分，位于北纬31度、东经103.4度的四川省汶川县发生里氏7.8级地震，造成重大人员伤亡和重大财产损失。(At 12:14:28 on May 12, a magnitude 7.8 earthquake happened in Wenchuan country (103.4E, 31W), Sichuan province, caused heavy casulities and significant property damage) 截至13日7时，四川汶川县地震已造成四川、甘肃、陕西、重庆、云南、山西、贵州、湖北8省市共11921人遇难，倒塌房屋50余万间。(Up to 7:00 on the 13th, the earthquake in Wenchuan country, Sichuan province has killed 11921 people and collapsed more than half million houses in Sichuan, Gansu, Shanxi, Chongqing, Yunnan, Shanxi, Guizhou and Hubei) 汶川地震专家委员会副主任史培军表示，在汶川地震造成直接经济损失8451亿元。(Shi Peijun, the deputy director of Wenchuan Earthquake Expert Committee, said that the Wenchuan earthquake had caused direct economic losses of 845.1 billion yuan.) 四川省汶川县发生强烈地震后，成都军区紧急出动6100余名官兵赶赴灾区参加抗震救灾。(After the strong earthquake in Whenchuan country, Sichuan Province, Chengdu Military Region urgently dispatched more than 6100 troops to the disaster area to participate in earthquake relief.) 截至16日0时38分，部队采取机械和强行突破方式向原来没有到达的58个乡镇派出3208个士兵强迫到达，出动直升机97架，现在各大医院，部队医院共80名医疗机构对灾区支援。(Up to 0:38 on the 16th, 3208 soldiers have arrived at the 58 towns unreached before by machinery and breakthrough way. 97 helicopters have been sent. A total of 80 medical institutions, including major hospitals and military hospitals have supported the disaster region.) 至15日上午，参与救援行动的解放军和武警官兵、公安民警、干部群众和医务工作者等救援人员已从灾区抢救出伤员6万多人，伤员全部得到及时救治和妥善安置。(Up to the morning of the 15th, the rescue workers of the PLA soldiers, armed police officers and medical worker have salvaged more than 60,000 wounded. All injured have receive timely medical treatment and proper rehabilitation.)	4月14日7时49分,青海省玉树藏族自治州玉树县发生7.1级地震,给当地人民群众生命财产造成严重损失。(At 7:49 on April 14, a 7.1 magnitude earthquake happened in Yushu country, Yushu Tibetan autonomous prefecture, Qinghai province, and caused great damages to lives and properties of local people.) 截至北京时间15日上午9时，发生在中国青海玉树的地震已经造成617人遇难，313人失踪，9110人受伤，其中，970人伤势严重。(Up to 9:00AM on the 15th, 617 people were killed by the earthquake in Yushu, Qinghai Province, China; 313 people were missing; 9110 people were injured, of which 970 people were seriously injured.) 地震造成大量房屋破坏，当地的教育、卫生、电力、通讯、公路、水利等基础设施也受到严重破坏。(The earthquake has caused damages to a large number of houses. The local infrastructures of educations, health, electricity, communications, roads, and water conservancy have been seriously damaged.) 据初步统计，灾害造成直接经济损失近3亿元人民币。(According to preliminary statistics, the disaster has caused direct economic losses of nearly 300 million yuan.) 截至17日，已调集至玉树震灾现场的各类救援人员达到15000余人，累计搜救营救被困群众17000人。(Up to the 17th, more than 15,000 rescue works have arrived at the scene of the Yushu earthquake, and have rescued 17 000 trapped people.) 中国军队和武警部队还派出了13支医疗救援队和2个方舱医院，抢救被埋群众，救治伤员；空军和陆航部队共飞行89架次执行救灾任务运送帐篷、食品等救灾物资5196吨。(The Chinese army and armed police have sent 13 medical team and 2 cabins hospital to rescue the buried masses and treat the wounded. The Air Force and Army Aviation have flown 89 sorties of disaster relief missions to deliver tent, food and other relief supplies for 5196 tons.) 救灾部队行动迅速，目前已救出被压埋群众1564人，并对2万多名伤员进行了救治。(The relief forces act quickly, have rescued 1564 buried masses, and treated more than 20,000 wounded.)

Table 9 The system-generated comparative summary about two earthquakes

Wenchuan Earthquake	Yushu Earthquake
(1.1) 2008年5月12日，四川汶川发生8. 0级大地震。(On May 12, 2008, a 8 magnitude earthquake happened in Wenchuan, Sichuan province) (1.2) 地震造成的人员财产损失正在进一步统计中。(The statistics of damage to lives and properties caused by the earthquake is still being gathered) (1.3) 中国国家减灾委于12日22点15分紧急启动一级救灾应急响应。(China National Committee for Disaster Reduction activated the first level emergency response system at 22:15, May 12.) (1.4) 目前，国务院总理温家宝已火速抵达四川地震灾区，直接靠前指挥抗震救灾工作。（Premier Wen Jiabao has arrived the disaster area, and command the relief work at the frontline.） (1.5) 四川省汶川县发生强烈地震后，成都军区紧急出动6100余名官兵赶赴灾区参加抗震救灾。(After the great earthquake at Wenchuan country, Sichuan province, about 6100 soldiers has set out for the disaster area to join relief work.) (1.6) 汶川地震灾区救援人员至 15 日抢救伤员6万多人，15日，救援人员向58个重灾乡镇挺进。(The rescue workers have rescue more than 60 thousands wonders by May 15, and set out for 58 towns hard hit on 15.) (1.7) 中国红十字会总会从中国红十字会成都备灾救灾中心迅速调拨价值78万余元的救灾物资。(The Red Cross Society of China promptly allocate relief supplies of 78 million yuan from the disaster preparedness center in Chendu.） (1.8) 他表示斯政府将尽最大努力帮助中国政府和人民应对此次地震灾害。(He said the Sri Lanka government will do its utmost to help Chinese government and people respond to the earthquake disaster.) (1.9) 老挝副总理兼政府常设委员会主席宋萨瓦20日前往中国驻老挝大使馆吊唁四川汶川大地震遇难者。(Somsavat, Lao Deputy Prime Minister and Chairman of the Standing Committee on Government, lamented the victims of the Wenchuan earthquake at the Chinese Embassy in Laos on May 20.)	(2.1) 4月14日7时49分,青海省玉树藏族自治州玉树县发生 7.1 级地震,给当地人民群众生命财产造成严重损失。(At 7:49, April 14, a 7.1 magnitude earthquake happened in Yushu country, Yushu Tibetan autonomous prefecture, Qinghai province, and cause a great damage to lives and properties of local people.) (2.2) 救灾部队行动迅速，目前已救出被压埋群众1564人，并对2万多名伤员进行了救治。(The rescue troops acted in an instant. They have rescued 1564 person buried under debris and treated more than 20 thousands wounded by now.) (2.3) 4月15日民政部和中国地震局联合举行新闻发布会，通报了青海玉树地震灾害和抗震救灾的进展情况。(On April 15, Ministry of Civil Affairs and China Earthquake Administration held a joint press conference to announce the earthquake in Yushu, Qinghai province and the process of relief works.) (2.4) 中国地震局震灾应急救援司副司长苗崇刚指出，救灾工作的一个困难就是灾区环境。(Miao Chonggang, the deputy director of earthquake emergency rescue department of China Earthquake Administration, pointed out the environment in the disaster area was one of the difficulties.) (2.5) 相关部门应注意天气变化，做好灾区群众和救援人员的保暖防寒工作。(The relevant departments should pay attention to the weather changes and warm-keeping of victims and relief workers in the disaster area.) (2.6) 据悉，发改委紧急下达了中央预算内补助资金5000万元，专项用于青海玉树地震抗震救灾和损毁设施恢复重建；同时协调国家电网公司、中石油等部门，向灾区调运物资。(It is reported that the NDRC issued an emergency grant funds of 50 million yuan from the central budget for the relief and reconstruction in Yushu, Qinghai, and coordinate the transportation of materials with SGCC, CNPC, etc.) (2.7) 泰国媒体今天迅速报道了中国青海玉树发生 7.1 级地震以及中国政府积极组织救援工作的消息。（The medias in Thailand reported the news of magnitude 7.1 earthquake in Yushu, Qinghai and the rescue work actively organized by the Chinese government today.）

The evaluation methods of comparative summarization can also be further studied. In this study, we use the automatically calculated ROUGE values, manually annotated Comparative Aspect Recall and Overall Responsiveness to evaluate the systems' performance. Additional independent human evaluation should be involved to strengthen the reliability of evaluation. We also plan to investigate the correlation between automatically calculated metrics and manual ratings to verify the reliability of those metrics.

Acknowledgments This work is supported by NSFC (61170166), Beijing Nova Program (2008B03), and National High Technology Research and Development Program of China (2012AA011101). We thank IBM for making the CPLEX optimizer freely access for academic use.

References

1. Allan J (ed) (2002) Topic detection and tracking: event-based information organization. Kluwer Academic Publishers, Dordrecht
2. Anttila R (1989) Historical and comparative linguistics (current issues in linguistic theory). John Benjamins Pub Co., Amsterdam
3. Balahur A, Steinberger R, Kabadjov M et al. (2010) Sentiment analysis in the news. In: Proceedings of the seventh international conference on language resources and evaluation
4. Bao S, Li R, Yu Y et al. (2008) Competitor mining with the web. IEEE Trans Knowl Data Eng 20(10): 1297–1310
5. Barzilay R, McKeown KR (2005) Sentence fusion for multidocument news summarization. Comput Linguist 31(3):297–328
6. Black CE (1966) Dynamics of modernization. A study in comparative history. Harper & Row, New York
7. Carbonell J and Goldstein J (1998) The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval. ACM, New York, NY, USA, pp 335–336
8. Dang HT and Owczarzak K (2008) Overview of the TAC 2008 update summarization task. In: Proceedings of the first text analysis conference (TAC 2008). National Institute of Standards and Technology, Gaithersburg, MD, USA
9. Dantzig GB (1949) Programming of interdependent activities: II mathematical model. Econometrica 17(3):200–211
10. Das AS, Datar M, Garg A et al. (2007) Google news personalization: scalable online collaborative filtering. In: Proceedings of the 16th international conference on World Wide Web. AMC, New York, NY, USA, pp 271–280
11. Gillick D, Favre B, Hakkani-Tur D (2008) The ICSI Summarization System at TAC 2008. In: Proceedings of the first text analysis conference (TAC 2008). National Institute of Standards and Technology, Gaithersburg, MD, USA
12. Gunes E, Radev DR (2004) LexPageRank: prestige in multi-document text summarization. In: Proceedings of the 2004 conference on empirical methods in natural language processing. Association for Computational Linguistics, Stroudsburg, PA, USA, pp 365–371
13. Hovy E, and Lin C-Y (1997) Automated text summarization in SUMMARIST. In: Proceedings of ACL'1997/EACL'1997 workshop on intelligent scalable text summarization
14. Jain A and Pantel P (2009) How do they compare? Automatic identification of comparable entities on the web. In: Proceedings of the 18th ACM conference on information and knowledge management. ACM, New York, NY, USA, pp 1661–1664
15. Jain AK, Dubes RC (1988) Algorithms for clustering data. Prentice-Hall, Inc, Upper Saddle River
16. Jindal N and Liu B (2006a) Identifying comparative sentences in text documents. In: Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval. ACM, New York, USA, pp 244–251
17. Jindal N and Liu B (2006b) Mining comparative sentences and relations. In: Proceedings of the 21st national conference on, artificial intelligence (AAAI-06). pp 1331–1336
18. Karmarkar N (1984) A new polynomial-time algorithm for linear programming. Combinatorica 4(4): 373–395
19. Kawai Y, Kumamoto T, Tanaka K (2007) Fair news reader: recommending news articles with different sentiments based on user preference. Knowl Intell Inf Eng Syst 4692:612–622

20. Kennedy C (2005) Comparatives, semantics of. In: Allen K (ed) Enclycopedia of language and linguistics, 2nd edn. Elsevier, Oxford
21. Khachian LG (1979) A polynomial algorithm in linear programming. Doklady Akademiia Nauk SSSR 244:1093–1096
22. Kim HD, Zhai C (2009) Generating comparative summaries of contradictory opinions in text. In: Proceedings of the 18th ACM conference on information and knowledge management. ACM, New York, NY, USA, pp 385–394
23. Lerman K, McDonald R (2009) Contrastive summarization: an experiment with consumer reviews. In: Proceedings of human language technologies: the 2009 annual conference of the North American chapter of the association for computational linguistics. Association for Computational Linguistics, Stroudsburg, PA, USA, pp 113–116
24. Lerner J-Y and Pinkal Ms (2004). Comparatives and nested quantifications. In: Semantics: critical concepts in linguistics vol V, operators and sentence types, pp 70–87
25. Li R, Bao S, Wang J et al (2006) CoMiner: an effective algorithm for mining competitors from the web. In: Proceedings of the sixth IEEE international conference on data mining (ICDM'06). IEEE Computer Society, Washington, DC, USA, pp 948–952
26. Lijphart A (1971) Comparative politics and the comparative method. Am Polit Sci Rev 65(3):682–693
27. Lin C-Y and Hovy E (2003) Automatic evaluation of summaries using N-gram co-occurrence statistics. In: Proceedings of the 2003 conference of the North American chapter of the association for computational linguistics on human language technology. Association for Computational Linguistics, Stroudsburg, PA, USA, pp 71–78
28. Lin C-Y and Och FJ (2004) Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In: Proceedings of the 42nd annual meeting on association for computational linguistics. Association for Computational Linguistics, Stroudsburg, PA, USA, pp 605
29. Liu B, Hu M, Cheng J (2005) Opinion observer: analyzing and comparing opinions on the web. In: Proceedings of the14th international conference on world wide web. ACM, New York, NY, USA, pp 342–351
30. Liu J, Wagner E, Birnbaum L (2007) Compare & contrast: using the web to discover comparable cases for news stories. In: Proceedings of the 16th international conference on World Wide Web. ACM, New York, NY, USA, pp 541–550
31. Liu Q, Li S (2002) Word similarity computing based on How-net. Comput Linguist Chinese Lang Process 7(2):59–76
32. Ma J (1898) Ma Shi Wen Tong. Commercial Press, Shanghai
33. Mani I (2001) Automatic summarization, vol 3. John Benjamins Pub Co., Amsterdam
34. Mcdonald R (2007) A study of global inference algorithms in multi-document summarization. In: Proceedings of the 29th European conference on IR research. Springer, Berlin, Heidelberg, pp 557–564
35. Mei J-P, Chen L (2012) SumCR: a new subtopic-based extractive approach for text summarization. Knowl Inf Syst 31(3):527–545
36. Mihalcea R (2004) Graph-based ranking algorithms for sentence extraction, applied to text summarization. In: Proceedings of the ACL 2004 on interactive poster and demonstration sessions. Association for Computational Linguistics, Stoudsburg, PA, USA, pp 20
37. Montes-y-Gómez M, Gelbukh A, López-López A (2001) Mining the news: trends, associations, and deviations. Computación y Sistemas 5(1):14–24
38. Nigam K, McCallum AK, Thrun S et al (2000) Text classification from labeled and unlabeled documents using EM. Mach Learn 39(2–3):103–134
39. NIST (2002) The 2002 topic detection and tracking (TDT2002) task definition and evaluation plan
40. Nomoto T, Matsumoto Y (2001) A new approach to unsupervised text summarization. In: Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval. AMC, New York, NY, USA, pp 26–34
41. Pablo-Sánchez Cd, Segura-Bedmar I, Martínez P et al (2012) Lightly supervised acquisition of named entities and linguistic patterns for multilingual text mining. Knowl Inf Syst, pp 1–23
42. Paul MJ, Zhai C, Girju R (2010) Summarizing contrastive viewpoints in opinionated text. In: Proceedings of the 2010 conference on empirical methods in natural language processing. Association for Computational Linguistics, Stroudsburg, PA, USA, pp 66–76
43. Pedersen T, Patwardhan S, Michelizzi J (2004) WordNet:Similarity: measuring the relatedness of concepts. In: Demonstration papers at HLT-NAACL 2004. Association for Computational Linguistics, Stroudsburg, PA, USA, pp 38–41
44. Rau LF, Jacobs PS, Zernik U (1989) Information extraction and text summarization using linguistic knowledge acquisition. Inf Process Manag 25(4):419–428

45. Roy JR (2000) Compare and contrast, Grades 5-6: using comparisons and contrasts to build comprehension. Instructional fair, Inc., Grand Rapids, MI
46. Salton G, Singhal A, Mitra M et al (1997) Automatic text structuring and summarization. Inf Process Manag 33(2):193–207
47. Shen D, Sun J-T, Li H et al (2007) Document summarization using conditional random fields. In: Proceedings of the 20th international joint conference on artifical intelligence. Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, pp 2862–2867
48. Sun A, Grishman R, Sekine S (2011) Semi-supervised relation extraction with large-scale word clustering. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies, vol 1 . Association for Computational Linguistics, Stroudsburg, PA, USA, pp 521–529
49. Tsai F, Zhang Y (2011) D2S: document-to-sentence framework for novelty detection. Knowl Inf Syst 29(2):419–433
50. Wan X, Jia H, Huang S et al (2011) Summarizing the differences in multilingual news. In: Proceedings of the 34th annual ACM SIGIR conference (SIGIR 2011). ACM, New York, NY, USA, pp 735–744
51. Wan X, Yang J, Xiao J (2007) Manifold-ranking based topic-focused multi-document summarization. In: Proceedings of the 20th international joint conference on artifical, intelligence, pp 2903–2908
52. Wang D, Zhu S, Li T et al (2009) Comparative document summarization via discriminative sentence selection. In: Proceedings of the 18th ACM conference on information and knowledge management. ACM, New York, NY, USA, pp 1963–1966
53. Wayne CL (1998) Topic detection & tracking (TDT) overview & perspective. In: Proceedings of the broadcast news transcription and understanding workshop
54. Wei F, Li W, Lu Q et al (2010) A document-sensitive graph model for multi-document summarization. Knowl Inf Syst 22(2):245–259
55. Weisstein U (1974) Comparative literature and literary theory: survey and introduction. Indiana University Press, Indiana
56. Witte R, Bergler S (2007) Next-generation summarization: contrastive, focused, and update summaries. In: International conference on recent advances in natural language processing (RANLP 2007)
57. Wood MK, Dantzig GB (1949) Programming of interdependent activities: I general discussion. Econometrica 17(3):193–199
58. Yeh J-Y, Ke H-R, Yang W-P et al (2005) Text summarization using a trainable summarizer and latent semantic analysis. Inf Process Manag 41(1):75–95
59. Zhai C, Velivelli A, Yu B (2004) A cross-collection mixture model for comparative text mining, In: Proceedings of the 10th ACM SIGKDD international conference on knowledge discovery and data mining . ACM, New York, NY, USA, pp 743–748

Author Biographies

Xiaojiang Huang received a B.E. degree from Beijing Information Technology Institute, Beijing, China, in 2005. He is currently a Ph.D. student at Institute of Computer Science and Technology, Peking University, Beijing, China. His research interests include natural language processing and web mining.

Xiaojun Wan is an associate professor at Institute of Computer Science and Technology of Peking University. He received his B.S., M.S. and Ph.D. degrees from Peking University in 2000, 2003 and 2006, respectively. His main research interests are natural language processing and text mining, including topics like document summarization, sentiment analysis and knowledge extraction.

Jianguo Xiao is an professor at Institute of Computer Science and Technology of Peking University. His research interests include image and video processing, web information processing and text mining.