Comparative News Summarization Using Linear Programming

Xiaojiang Huang **Xiaojun Wan*** **Jianguo Xiao**
Institute of Computer Science and Technology, Peking University, Beijing 100871, China
Key Laboratory of Computational Linguistic (Peking University), MOE, China
{huangxiaojiang, wanxiaojun, xiaojianguo}@icst.pku.edu.cn

Abstract

Comparative News Summarization aims to highlight the commonalities and differences between two comparable news topics. In this study, we propose a novel approach to generating comparative news summaries. We formulate the task as an optimization problem of selecting proper sentences to maximize the comparativeness within the summary and the representativeness to both news topics. We consider semantic-related cross-topic concept pairs as comparative evidences, and consider topic-related concepts as representative evidences. The optimization problem is addressed by using a linear programming model. The experimental results demonstrate the effectiveness of our proposed model.

1 Introduction

Comparative News Summarization aims to highlight the commonalities and differences between two comparable news topics. It can help users to analyze trends, draw lessons from the past, and gain insights about similar situations. For example, by comparing the information about mining accidents in Chile and China, we can discover what leads to the different endings and how to avoid those tragedies.

Comparative text mining has drawn much attention in recent years. The proposed works differ in the domain of corpus, the source of comparison and the representing form of results. So far, most researches focus on comparing review opinions of products (Liu et al., 2005; Jindal and Liu, 2006a; Jindal and Liu, 2006b; Lerman and McDonald, 2009; Kim and Zhai, 2009). A reason is that the aspects in reviews are easy to be extracted and the comparisons have simple patterns, e.g. positive vs. negative. A few other works have also tried to compare facts and views in news article (Zhai et al., 2004) and Blogs (Wang et al., 2009). The comparative information can be extracted from explicit comparative sentences (Jindal and Liu, 2006a; Jindal and Liu, 2006b; Huang et al., 2008), or mined implicitly by matching up features of objects in the same aspects (Zhai et al., 2004; Liu et al., 2005; Kim and Zhai, 2009; Sun et al., 2006). The comparisons can be represented by charts (Liu et al., 2005), word clusters (Zhai et al., 2004), key phrases(Sun et al., 2006), and summaries which consist of pairs of sentences or text sections (Kim and Zhai, 2009; Lerman and McDonald, 2009; Wang et al., 2009). Among these forms, the comparative summary conveys rich information with good readability, so it keeps attracting interest in the research community. In general, document summarization can be performed by extraction or abstraction (Mani, 2001). Due to the difficulty of natural sentence generation, most automatic summarization systems are extraction-based. They select salient sentences to maximize the objective functions of generated summaries (Carbonell and Goldstein, 1998; McDonald, 2007; Lerman and McDonald, 2009; Kim and Zhai, 2009; Gillick et al., 2009). The major difference between the traditional summarization task and the comparative summarization task is that traditional summarization task places equal emphasis on all kinds of information in

*Corresponding author

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers, pages 648–653,
Portland, Oregon, June 19-24, 2011.

the source, while comparative summarization task only focuses on the comparisons between objects.

News is one of the most important channels for acquiring information. However, it is more difficult to extract comparisons in news articles than in reviews. The aspects are much diverse in news. They can be the time of the events, the person involved, the attitudes of participants, etc. These aspects can be expressed explicitly or implicitly in many ways. For example, "*storm*" and "*rain*" both talk about "*weather*", and thus they can form a potential comparison. All these issues raise great challenges to comparative summarization in the news domain.

In this study, we propose a novel approach for comparative news summarization. We consider comparativeness and representativeness as well as redundancy in an objective function, and solve the optimization problem by using linear programming to extract proper comparable sentences. More specifically, we consider a pair of sentences comparative if they share comparative concepts; we also consider a sentence representative if it contains important concepts about the topic. Thus a good comparative summary contains important comparative pairs, as well as important concepts about individual topics. Experimental results demonstrate the effectiveness of our model, which outperforms the baseline systems in quality of comparison identification and summarization.

2 Problem Definition

2.1 Comparison

A comparison identifies the commonalities or differences among objects. It basically consists of four components: the **comparee** (i.e. what is compared), the **standard** (i.e. to what the compare is compared), the **aspect** (i.e. the scale on which the comparee and standard are measured), and the **result** (i.e. the predicate that describes the positions of the comparee and standard). For example, "*Chile is richer than Haiti.*" is a typical comparison, where the comparee is "*Chile*"; the standard is "*Haiti*"; the comparative aspect is *wealth*, which is implied by "*richer*"; and the result is that *Chile is superior to Haiti*.

A comparison can be expressed explicitly in a comparative sentence, or be described implicitly in a section of text which describes the individual characteristics of each object point-by-point. For example, the following text

> *Haiti is an extremely poor country.*
> *Chile is a rich country.*

also suggests that *Chile is richer than Haiti*.

2.2 Comparative News Summarization

The task of comparative news summarization is to briefly sum up the commonalities and differences between two comparable news topics by using human readable sentences. The summarization system is given two collections of news articles, each of which is related to a topic. The system should find latent comparative aspects, and generate descriptions of those aspects in a pairwise way, i.e. including descriptions of two topics simultaneously in each aspect. For example, when comparing the earthquake in Haiti with the one in Chile, the summary should contain the intensity of each temblor, the damages in each disaster area, the reactions of each government, etc.

Formally, let t_1 and t_2 be two comparable news topics, and D_1 and D_2 be two collections of articles about each topic respectively. The task of comparative summarization is to generate a short abstract which conveys the important comparisons $\{< \ t_1, t_2, r_{1i}, r_{2i} \ >\}$, where r_{1i} and r_{2i} are descriptions about topic t_1 and t_2 in the same latent aspect a_i respectively. The summary can be considered as a combination of two components, each of which is related to a news topic. It can also be subdivided into several sections, each of which focuses on a major aspect. The comparisons should have good quality, i.e., be clear and representative to both topics. The coverage of comparisons should be as wide as possible, which means the aspects should not be redundant because of the length limit.

3 Proposed Approach

It is natural to select the explicit comparative sentences as comparative summary, because they express comparison explicitly in good qualities. However, they do not appear frequently in regular news articles so that the coverage is limited. Instead,

it is more feasible to extract individual descriptions of each topic over the same aspects and then generate comparisons.

To discover latent comparative aspects, we consider a sentence as a bag of concepts, each of which has an atom meaning. If two sentences have same concepts in common, they are likely to discuss the same aspect and thus they may be comparable with each other. For example,

> *Lionel Messi named FIFA Word Player of the Year 2010.*
> *Cristiano Ronalo Crowned FIFA Word Player of the Year 2009.*

The two sentences compare on the "*FIFA Word Player of the Year*", 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*". Thus the pairs of semantic related concepts can be considered as evidences of comparisons.

A comparative summary should contain as many comparative evidences as possible. Besides, it should convey important information in the original documents. Since we model the text with a collection of concept units, the summary should contain as many important concepts as possible. An important concept is likely to be mentioned frequently in the documents, and thus we use the frequency as a measure of a concept's importance.

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, named entities and bigrams to simply represent concepts, and leave the more complex concept extraction for future work.

Based on the above ideas, we can formulate the summarization task as an optimization problem. Formally, let $C_i = \{c_{ij}\}$ be the set of concepts in the document set $D_i, (i = 1, 2)$. Each concept c_{ij} has a weight $w_{ij} \in \mathbb{R}$. $oc_{ij} \in \{0, 1\}$ is a binary variable indicating whether the concept c_{ij} is presented in the summary. A cross-topic concept pair $< c_{1j}, c_{2k} >$ has a weight $u_{jk} \in \mathbb{R}$ that indicates whether it implies a important comparison. op_{jk} is a binary variable indicating whether the pair is presented in the summary. Then the objective function score of a comparative summary can be estimated as follows:

$$\lambda \sum_{j=1}^{|C_1|} \sum_{k=1}^{|C_2|} u_{jk} \cdot op_{jk} + (1-\lambda) \sum_{i=1}^{2} \sum_{j=1}^{|C_i|} w_{ij} \cdot oc_{ij} \quad (1)$$

The first component of the function estimates the comparativeness within the summary and the second component estimates the representativeness to both topics. $\lambda \in [0, 1]$ is a factor that balances these two factors. In this study, we set $\lambda = 0.55$.

The weights of concepts are calculated as follows:

$$w_{ij} = tf_{ij} \cdot idf_{ij} \quad (2)$$

where tf_{ij} is the term frequency of the concept c_{ij} in the document set D_i, and idf_{ij} is the inverse document frequency calculated over a background corpus.

The weights of concept pairs are calculated as follows:

$$u_{jk} = \begin{cases} (w_{1j} + w_{2k})/2, & \text{if } rel(c_{1j}, c_{2k}) > \tau \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

where $rel(c_{1j}, c_{2k})$ is the semantic relevance between two concepts, and it is calculated using the algorithms basing on WordNet (Pedersen et al., 2004). If the relevance is higher than the threshold τ (0.2 in this study), then the concept pair is considered as an evidence of comparison.

Note that a concept pair will not be presented in the summary unless both the concepts are presented, i.e.

$$op_{jk} \leq oc_{1j} \quad (4)$$

$$op_{jk} \leq oc_{2k} \quad (5)$$

In order to avoid bias towards the concepts which have more related concepts, we only count the most important relation of each concept, i.e.

$$\sum_{k} op_{jk} \leq 1, \forall j \quad (6)$$

$$\sum_{j} op_{jk} \leq 1, \forall k \quad (7)$$

The algorithm selects proper sentences to maximize the objective function. Formally, let $S_i =$

$\{s_{ik}\}$ be the set of sentences in D_i, ocs_{ijk} be a binary variable indicating whether concept c_{ij} occurs in sentence s_{ik}, and os_{ik} be a binary variable indicating whether s_{ik} is presented in the summary. If s_{ik} is selected in the summary, then all the concepts in it are presented in the summary, i.e.

$$oc_{ij} \geq ocs_{ijk} \cdot os_{ik}, \forall 1 \leq j \leq |C_i| \quad (8)$$

Meanwhile, a concept will not be present in the summary unless it is contained in some selected sentences, i.e.

$$oc_{ij} \leq \sum_{k=1}^{|S_i|} ocs_{ijk} \cdot os_{ik} \quad (9)$$

Finally, the summary should satisfy a length constraint:

$$\sum_{i=1}^{2} \sum_{k=1}^{|S_i|} l_{ik} \cdot os_{ik} \leq L \quad (10)$$

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

The optimization of the defined objective function under above constraints is an integer linear programming (ILP) problem. Though the ILP problems are generally NP-hard, considerable works have been done and several software solutions have been released to solve them efficiently.[1]

4 Experiment

4.1 Dataset

Because of the novelty of the comparative news summarization task, there is no existing data set for evaluating. We thus create our own. We first choose five pairs of comparable topics, then retrieve ten related news articles for each topic using the Google News[2] search engine. Finally we write the comparative summary for each topic pair manually. The topics are showed in table 1.

ID	Topic 1	Topic 2
1	Haiti Earth quake	Chile Earthquake
2	Chile Mining Accident	New Zealand Mining Accident
3	Iraq Withdrawal	Afghanistan Withdrawal
4	Apple iPad 2	BlackBerry Playbook
5	2006 FIFA World Cup	2010 FIFA World Cup

Table 1: Comparable topic pairs in the dataset.

4.2 Evaluation Metrics

We evaluate the models with following measures:

Comparison Precision / Recall / F-measure: let a_a and a_m be the numbers of all aspects involved in the automatically generated summary and manually written summary respectively; c_a be the number of human agreed comparative aspects in the automatically generated summary. The comparison precision (CP), comparison recall (CR) and comparison F-measure (CF) are defined as follows:

$$CP = \frac{c_a}{a_a}; \quad CR = \frac{c_a}{a_m}; \quad CF = \frac{2 \cdot CP \cdot CR}{CP + CR}$$

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 (Lin and Hovy, 2003). In the experiment, we report the f-measure values of ROUGE-1, ROUGE-2 and ROUGE-SU4, which count overlapping unigrams, bigrams and skip-4-grams respectively. To evaluate whether the summary is related to both topics, we also split each comparative summary into two topic-related parts, evaluate them respectively, and report the mean of the two ROUGE values (denoted as MROUGE).

4.3 Baseline Systems

Non-Comparative Model (NCM): The non-comparative model treats the task as a traditional summarization problem and selects the important sentences from each document collection. The model is adapted from our approach by setting $\lambda = 0$ in the objection function 1.

Co-Ranking Model (CRM): The co-ranking model makes use of the relations within each topic and relations across the topics to reinforce scores of the comparison related sentences. The model is adapted from (Wan et al., 2007). The

[1] We use IBM ILOG CPLEX optimizer to solve the problem.

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

SS, WW and SW relationships are replaced by relationships between two sentences within each topic and relationships between two sentences from different topics.

4.4 Experiment Results

We apply all the systems to generate comparative summaries with a length limit of 200 words. The evaluation results are shown in table 2. Compared with baseline models, our linear programming based comparative model (denoted as LPCM) achieves best scores over all metrics. It is expected to find 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, it does not guarantee to choose pairwise sentences to form comparisons. The LPCM model focus on both comparativeness and representativeness at the same time, and thus it achieves good performance on both comparison extraction and summarization. Figure 1 shows an example of comparative summary generated by using the CLPM model. The summary describes several comparisons between two FIFA World Cups in 2006 and 2010. Most of the comparisons are clear and representative.

5 Conclusion

In this study, we propose a novel approach to summing up the commonalities and differences between two news topics. We formulate the task as an optimization problem of selecting sentences to maximize the score of comparative and representative evidences. The experiment results show that our model is effective in comparison extraction and summarization.

In future work, we will utilize more semantic information such as localized latent topics to help capture comparative aspects, and use machine learning technologies to tune weights of concepts.

Acknowledgments

This work was supported by NSFC (60873155), Beijing Nova Program (2008B03) and NCET (NCET-08-0006).

Model	CP	CR	CF	ROUGE-1	ROUGE-2	ROUGE-su4	MROUGE-1	MROUGE-2	MROUGE-su4
NCM	0.238	0.262	0.247	0.398	0.146	0.174	0.350	0.122	0.148
CRM	0.313	0.285	0.289	0.426	0.194	0.226	0.355	0.146	0.175
LPCM	**0.359**	**0.419**	**0.386**	**0.427**	**0.205**	**0.234**	**0.380**	**0.171**	**0.192**

Table 2: Evaluation results of systems

World Cup 2006	**World Cup 2010**
The 2006 Fifa World Cup drew to a close on Sunday with Italy claiming their fourth crown after beating France in a penalty shoot-out.	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.
Zidane won the Golden Ball over Italians Fabio Cannavaro and Andrea Pirlo.	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.	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
Germany striker Miroslav Klose was the Golden Shoe winner for the tournament's leading scorer.	Among the winners were goalkeeper and captain Iker Casillas who won the Golden Glove Award.
England's fans brought more colour than their team.	Only four of the 212 matches played drew more that 40,000 fans.

Figure 1: A sample comparative summary generated by using the LPCM model

References

Jaime Carbonell and Jade Goldstein. 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*, pages 335–336. ACM.

Dan Gillick, Korbinian Riedhammer, Benoit Favre, and Dilek Hakkani-Tur. 2009. A global optimization framework for meeting summarization. In *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP '09, pages 4769–4772, Washington, DC, USA. IEEE Computer Society.

Xiaojiang. Huang, Xiaojun. Wan, Jianwu. Yang, and Jianguo. Xiao. 2008. Learning to Identify Comparative Sentences in Chinese Text. *PRICAI 2008: Trends in Artificial Intelligence*, pages 187–198.

Nitin Jindal and Bing Liu. 2006a. Identifying comparative sentences in text documents. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 244–251. ACM.

Nitin Jindal and Bing Liu. 2006b. Mining comparative sentences and relations. In *proceedings of the 21st national conference on Artificial intelligence - Volume 2*, pages 1331–1336. AAAI Press.

Hyun Duk Kim and ChengXiang Zhai. 2009. Generating comparative summaries of contradictory opinions in text. In *Proceeding of the 18th ACM conference on Information and knowledge management*, pages 385–394. ACM.

Kevin Lerman and Ryan McDonald. 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, Companion Volume: Short Papers*, pages 113–116. Association for Computational Linguistics.

Chin-Yew Lin and Eduard Hovy. 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 - Volume 1*, NAACL '03, pages 71–78, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bing Liu, Minqing Hu, and Junsheng Cheng. 2005. Opinion observer: analyzing and comparing opinions on the Web. In *Proceedings of the 14th international conference on World Wide Web*, pages 342–351. ACM.

Inderjeet Mani. 2001. *Automatic summarization*. Natural Language Processing. John Benjamins Publishing Company.

Ryan McDonald. 2007. A study of global inference algorithms in multi-document summarization. In *Proceedings of the 29th European conference on IR research*, ECIR'07, pages 557–564, Berlin, Heidelberg. Springer-Verlag.

Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. 2004. WordNet:: Similarity: measuring the relatedness of concepts. In *Demonstration Papers at HLT-NAACL 2004 on XX*, pages 38–41. Association for Computational Linguistics.

Jian-Tao Sun, Xuanhui Wang, Dou Shen, Hua-Jun Zeng, and Zheng Chen. 2006. CWS: a comparative web search system. In *Proceedings of the 15th international conference on World Wide Web*, pages 467–476. ACM.

Xiaojun Wan, Jianwu Yang, and Jianguo Xiao. 2007. Towards an iterative reinforcement approach for simultaneous document summarization and keyword extraction. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 552–559, Prague, Czech Republic, June. Association for Computational Linguistics.

Dingding Wang, Shenghuo Zhu, Tao Li, and Yihong Gong. 2009. Comparative document summarization via discriminative sentence selection. In *Proceeding of the 18th ACM conference on Information and knowledge management*, pages 1963–1966. ACM.

ChengXiang Zhai, Atulya Velivelli, and Bei Yu. 2004. A cross-collection mixture model for comparative text mining. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 743–748. ACM.