Multi-document Summarization via LDA and Density Peaks Based Sentence-Level Clustering

Baoyan Wang[1,2], Yuexian Zou[1](✉), Jian Zhang[3], Jun Jiang[4], and Yi Liu[2]

[1] ADSPLAB/Intelligent Lab, School of ECE, Peking University, Beijing, China
zouyx@pkusz.edu.cn
[2] PKU Shenzhen Institute, Shenzhen, China
[3] Dongguan University of Technology, Dongguan, China
[4] Shenzhen Press Group, Shenzhen, China

Abstract. In this paper, we present a novel unsupervised extractive multi-document summarization method by ranking sentences based on the integrated sentence scoring method. The cluster-based methods tend to ignore informativeness of words and Latent Dirichlet Allocation (LDA) based methods are inclined to extract the longish sentences and cannot remove redundancy directly. Those methods select sentences with higher score to generate summaries but not necessarily to the optimal summaries. Our method takes four key issues of sentences into account concurrently by applying LDA to calculate term weighting of words and evaluate the informativeness of sentences and then applying Density Peaks Clustering (DPC) to assess relevance and diversity of sentences simultaneously. Our method achieves the best property on the DUC2004 dataset, which outperforms the state-of-the-art methods, such as DUC2004 Best, R2N2_ILP [3], and WCS [13].

Keywords: Multi-document summarization
The integrated sentence scoring method · Latent Dirichlet Allocation
Density Peaks Clustering

1 Introduction

Given the exponential rate of the volume of information data overload on the WWW, consumers are flooded with a variety of electronic documents i.e. news, blogs, e-books. Therefore, there are urgent requirements for multi-document summarization (MDS) now more than ever, as it focuses on creating a condensed and informative summary for the large set of original documents that facilitates readers quickly grasp the general information of them. Most existing researches are extractive methods, which aim at extracting sentences firsthand from the given documents and assembling sentences together to generate the final summary. In this work, we address the task of generic extractive summarization from multiple documents. An effective summarization method always considers the four key issues properly: [1–3] Relevance, Diversity, Informativeness and Length Constraint.

K. Li et al. (Eds.): ISICA 2017, CCIS 873, pp. 313–323, 2018.
https://doi.org/10.1007/978-981-13-1648-7_27

The methods of extractive-based summarization can be classified into two categories: supervised methods that depend on the provided document-summary pairs, while unsupervised ones based upon the properties derived from document sets. On one hand, the supervised methods generally tend to regard the summarization task as a classification or regression issue [1, 4]. For those supervised methods, a fair amount of annotated data is required, which are costly and time-consuming. On the other hand, the unsupervised methods are continued to be researched. They tend to score and then rank sentences based on semantic, linguistic or statistic grouping extracted from the original documents. Typical existing methods include graph-based methods [5], matrix factorization based methods [6], submodular functions based methods [7], topic model based methods [8, 9], etc. Some of them might just consider one or more key issues, which can be improved farther. Some papers consider reducing the redundancy to hold the diversity of summary [10, 11], i.e. MMR.

It is also appropriate and natural to process MDS with clustering. The cluster based methods [12, 13] tend to divide sentences into groups through clustering method and then rank the sentences on the basis of their saliency scores. [14] ranked sentences with Density Peaks Clustering (DPC) [17] ignoring the informativeness of words and selected sentenced based greedy algorithm, which cannot guarantee the optimal summary. [12] presented their studies about document summarization using the notion of Latent Dirichlet Allocation (LDA) as the representation of documents and mixture models to capture the topics and pick up the sentences, which tend to select longish sentences and cannot remove redundancy directly. Inspired by the applications of cluster-based methods for MDS and LDA for topic model, we propose an integrated sentence scoring method based on LDA combined with DPC to extract sentences with more informativeness, higher relevance, and better diversity under the limitation of length for sentences ordering. Different from their work, our main contributions are:

(1) LDA combined with DPC are firstly adopt for unsupervised multi-document summarization, which improve the performance mutually.
(2) We put forward the integrated sentence scoring method with better scalability to rank sentences, which can be interpreted more intuitively.
(3) We leverage the information of three levels: term level, sentence level and cluster level, which are more comprehensive.

The structure of this paper is as follows. In Sect. 2 we present the integrated sentence scoring method based on LDA and DPC for MDS in detail. Section 3 gives the evaluation of our method on the open benchmark datasets DUC2003 and DUC2004. In Sect. 4, we conclude the paper with directions for the future study.

2 Proposed MDS Method

We propose a new MDS method termed as the integrated sentence scoring method, which use LDA and DPC to take relevance, diversity, informativeness and length constraint into account simultaneously. Sentences are scored in the four aspects, and then the scores are log linearly combined. Finally, the sentences are extracted to generate optimal summary based on dynamic programming algorithm.

2.1 Pre-processing

In order to represent sentences rationally and reprocess them expediently, it is indispensable to carry out t preprocessing step. After the given collection of English documents, $C_{corpus} = \{d_1, d_2,\ldots, d_i,\ldots, d_{cor}\}$, in which d_i denotes the i-th document in C_{corpus}, splitting apart into individual sentences, $S = \{s_1, s_2, \ldots, s_i, \ldots, s_{sen}\}$ where s_i means the i-th sentence in C_{corpus}, all words stemming is performed by Porter's stemming algorithm.

2.2 The Integrated Sentence Scoring Method

(1) Informativeness Score

A good summary sentence should contain enough information. LDA is employed to represent sentences and calculate the informativeness. LDA model is viewed as breaking down the set of documents into themes by representing the document as a mixture of themes with a probability distribution representing the significance of the theme for that document. The themes in turn are represented as a mixture of words with a probability representing the importance of the word for that theme [9]. We utilize Gibbs sampling [8] for the LDA parameters inference including the probability distribution $P(T_k|s_i)$ and $P(W_n|T_k)$. Those variables are always sampled once for each word of the given document. $P(W_n|T_k)$ denotes the distribution matrix of the topics over the words, while $P(T_k|s_i)$ presents the distribution matrix of sentences over the topic. The informativeness of a sentence $SC_I(i)$ is calculated by the sum of the weightings of non-stop words in the sentence.

$$P(T_k|C_{Corpus}) = \sum_{i=1}^{sen} P(T_k|s_i) / \sum_{k=1}^{K} \sum_{j=1}^{sen} P(T_k|s_j) \tag{1}$$

$$w_n = \sum_{k=1}^{K} P(W_n|T_k) P(T_k|C_{Corpus}) \tag{2}$$

$$SC_I(i) = \sum_{j=1}^{num_w} w_{ij},\ w_{ij} = \begin{cases} w_j & W_j \in s_i \\ 0 & else \end{cases} \tag{3}$$

where w_n is the n-th word's weighting, w_{ij} is the weighting of j-th word in i-th sentence, num_w is the non-stop word's number of the corpus and $P(T_k|C_{corpus})$ presents the probability distribution of topics over the whole corpus.

(2) Relevance Score

We employ the relevance score to quantify the degree concerning how much the relevance is between one with the other sentences of the document set. The DPC proposed based on two potential hypotheses. One of the potential hypotheses is that cluster centers should be characterized by a higher density than their neighbors. Proceeding from it we consider that a sentence will be deemed to be more relevant and more representational when it possesses higher density, namely of more similar sentences. In our method, the text representation is the vector space model and the representation vector of each sentence generate from (3). The similarity between sentences is calculated by cosine similarity method. Thus we define the function as follows, which calculate the Relevance Score $SC_R(i)$ in sentences level for every sentence s_i:

$$SC_R(i) = \sum_{j=1}^{K} f(Sim_{ij} - \omega), \quad f(x) = \begin{cases} 1 & x \geq 0 \\ 0 & else \end{cases} \tag{4}$$

where K denotes the number of sentences in the set of documents, and Sim_{ij} is the similarity value between the i-th and j-th sentence. ω represents the predefined threshold of density in DPC.

(3) Diversity Score

We show the diversity score to ensure that the sentences of the optimal summary should not be analogical. The set of documents always includes one central theme and some subthemes. The summary should contain the most evident theme beyond doubt. In order to better comprehend the whole document set, it's also necessary to make the sub-themes displayed in the final summary. Put another way, sentences of the summary ought to be less reduplicated with one another in order to eliminate redundancy. Another underlying assumption of DPC is that cluster centers also are characterized by a relatively large distance from points with higher densities. According to that, the scores of the similar sentences can be ensured to obtain larger gap. Furthermore, the sentences with higher diversity scores could be selected in comparison with all the other sentence of the document set. Therefore the diversity of the summary can be guaranteed globally. In our method, the diversity score $SC_D(i)$ is calculated in clusters level as follows.

$$SC_D(i) = \min_{j:SC_R(j)>SC_R(i)} (1 - Sim_{ij}) \tag{5}$$

We set the diversity value of the sentence maximum of $(1 - Sim_{ij})$ conventionally [15], which possesses the highest relevance score.

(4) Length Constraint

The sentence with the longer length usually possesses the more information. Moreover, the human summarizers always are apt to generate masses of shorter sentence for summary. When only consider the informativeness of sentences, the longish sentences tend to be selected, which is incongruent to the human habit. In the real summary system, it is usually restrained of the amount of words. When we select the longish sentences, the number of selected sentences is fewer. As a consequence, it is in sore need of providing the length constraint score. The length of sentences l_i is distributed in a large scale. In this situation, we bring in the taking logarithm smoothing method to solve this issue. Therefore, the length constraint score SC_L is calculated as follows.

$$SC_L = \log(\max_j l_j / l_i + 1) \tag{6}$$

(5) The Integrated Sentence Scoring Method

In order to extract the sentences with more information, higher relevance, and better diversity under the limitation of length, we proposed an integrated sentence scoring method all sidedly considering the four above objectives. For adapting to the integrated sentence score method, $SC_I(i)$, $SC_R(i)$, $SC_D(i)$ and $SC_L(i)$ should be normalized by divided their own highest values firstly.

$$SC(i) = \alpha \log SC_R(i) + \beta \log SC_D(i) + \gamma \log SC_I(i) + \log SC_L(i) \tag{7}$$

where the parameters α, β, and γ of the integrated scoring method are applied to adjust weightings of the four scores. We conduct a series of experiments on the standard datasets to tune and obtain the optimal the parameters settings.

We should generate a summary by extracting sentences under the restriction of the demanded length L. As every sentence is measure by an integrated score, the score sum of extracted sentences in summary should be as high as possible. Therefore the summary generation is considered as the 0–1 knapsack problem.

$$\begin{gathered} \arg\ \max \sum \left(SC(i) \times x_i\right) \\ \textit{Subject to } \sum_i l_i x_i \le L, x_i = \{0, 1\} \end{gathered} \tag{8}$$

To alleviate the NP-hard problem, we introduce the dynamic programming (DP) algorithm [23] to extract the sentences from the document set until the required length of ultimate summary is reached.

3 Experimental Setup

3.1 Dataset and Evaluation Metrics

The open benchmark datasets DUC2003 and DUC2004, from Document Understanding Conference, are employed in our experiments. DUC2004 consists of 50 news document sets and 10 documents related to each set. Length Limit of summary is 665 bytes. DUC2003 consists of 60 news document sets and about 10 documents for each set. The structures of both datasets are similar. Therefore, we choose DUC2003 as the development dataset for parameters tuning and DUC2004 for evaluation. There are four human generated summaries provided as ground truth for each news set. We observe that the sentences of summaries are not strictly selected in their entirety, but changed considerably.

We apply widely used ROUGE version 1.5.5 toolkit [16] to evaluate the performance of the summary system in our experiments. We select three Rouge evaluation metrics, Rouge-1, 2, SU, in all of embodied evaluation metrics. Rouge-1 is used to measure the co-occurrence of the identical words between the summary of our method and the annotated summary. Rouge-2 is used to measure the co-occurrence of the 2-g while Rouge-SU is applied to measure the co-occurrence of skip-grams. Rouge-2 and Rouge-SU concern more over the readability of the ultimate summary. We show the average value of the F-measure scores of the used metrics in the execute phase. Note that the higher ROUGE scores, the more similar between generated summary and annotated one.

3.2 Parameter Settings

We investigate how parameters α, β and γ, the topic number (T) of LDA and the density threshold ω relate to our method by a series of experiments on DUC2003. The three parameters α, β and γ are set ranging from 0 to 1.5 respectively at the step size of 0.1. The best values of the parameters are selected by comparing all of the results. As shown in Fig. 1, one parameter is tuned while set the others on their best values. The results of tuning parameters are shown in Fig. 1. We find that $\alpha = 0.6$, $\beta = 0.5$ and $\gamma = 0.9$ produce a better performance than $\alpha = 1$, $\beta = 1$ and $\gamma = 1$, which indicates effective degree of the

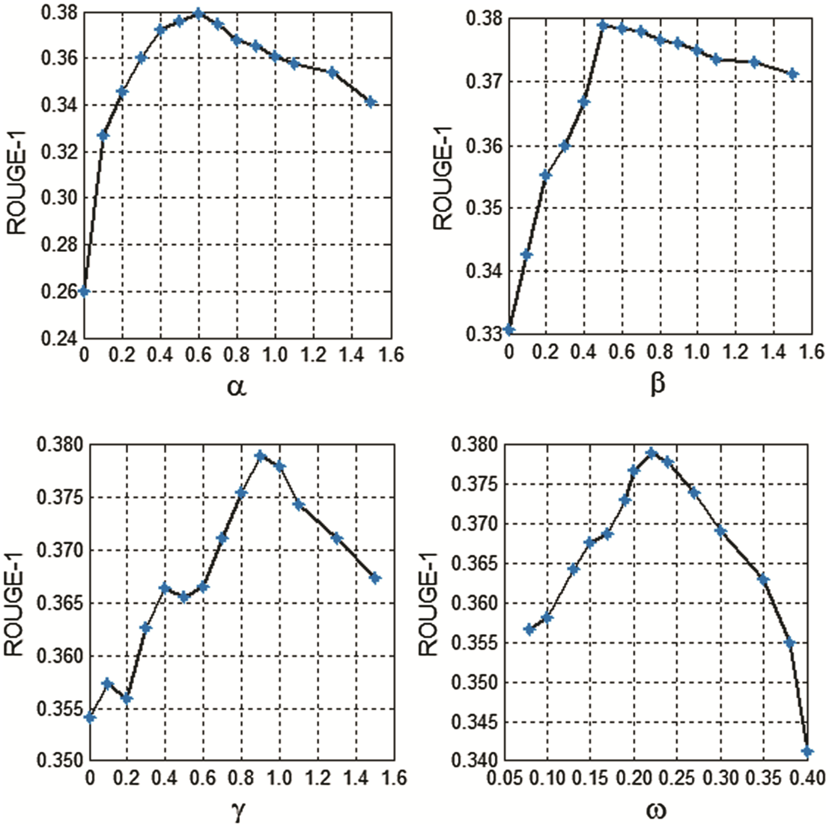


Fig. 1. ROUGE-1 versus parameter α, β, γ and ω on DUC2003.

four scores are different for the integrated sentence scoring method. When α, β, and γ equal zero successively, we observe that the property of our proposed method become worse and dropped at most with α, followed by others. In other words, the relevance score plays a most important role compared with others in our method. The information score enhances the result by term level information while the diversity score eliminates redundancy in cluster level of sentences. Our method works best when density threshold$\omega$is about equal to 0.22. We adopt T as 20, 50 and 100, and choose the best value 50 for the summarization task.

3.3 Experimental Results

We compare different term weighting schemes with ours firstly: (1) BOOL (presence or absence); (2) TF (term frequency); (3) ISF (inverse sentence frequency); (4) TF-ISF (combine TF with ISF). The results of these experiments are listed in Table 1. It can be seen that BOOL term weighting achieves better results compared with that of TF, ISF and TF-ISF. The cause may lie in the frequency of term repetition occur less in sentences. Our method (OURS) gets better results than other rivals. It is probably because LDA weights terms by using its mixture model, which describes the structure of the documents more fully.

Table 1. Validity of different term weighting schemes

Methods	ROUGE-1	ROUGE-2	ROUGE-SU
TF+DPC	0.38756	0.09278	0.13729
ISF+DPC	0.37461	0.08755	0.12863
TF-ISF+DPC	0.38109	0.08934	0.13243
BOOL+DPC	0.39047	0.09559	0.13916
OURS	**0.39893**	**0.09910**	**0.14530**

We choose the following typical or recent published approaches for genic multi-document summarization in comparison with our method. The results of these methods are listed in Table 2. We divided the baseline methods into four categories:

(1) The best human summarizer's performance;
(2) DUC best: The best participating team in DUC 2004;
(3) Cluster based method: RTC (Rank Through Clustering) [12]; WFS-NMF (Weighted Feature Subset Nonnegative Matrix Factorization) [6]; ClusterHITS (Cluster-based HITS) [13]; KM (Kmeans);
(4) Others: BSTM (Topic Model) [9]; MSSF (Submodular Functions) [7]; MCKP (Maximum Coverage Problem) [10]; WCS (Weighted Consensus Scheme) [13]; R2N2_ILP (Recursive Neural Networks) [1].

Table 2. Overall the property comparison of our method and baselines on DUC2004. Remark: "–" indicates that the method does not officially report the results.

Methods	ROUGE-1	ROUGE-2	ROUGE-SU
Best Human	0.41820	0.10500	–
DUC best	0.38224	0.09216	0.13233
KM	0.34872	0.06937	0.12115
RTC	0.37475	0.08973	–
ClusterHITS	0.36463	0.07632	–
WFS-NMF	0.39330	0.11210	0.13540
MCKP	0.38640	0.09240	0.13330
BSTM	0.39065	0.09010	0.13218
MSFF	–	0.09897	0.13951
R2N2_ILP	0.38780	0.09860	–
OURS	**0.39893**	**0.09910**	**0.14530**

For better demonstrating the results, we visually illustrate the comparison between our method with the start-of-art methods in Fig. 2. From Table 3 and Fig. 2, we can have the following observations: the experimental results of all methods are adequate for a 95% confidence interval. Our method achieves a result close to the best human summarizer's performance. What's more, our method distinctly outperforms the DUC04 best participating work. Besides, it also can be seen that our method outperforms other competitors significantly on the ROUGE-1 and ROUGE-SU metric. It can be attributed to the integrated sentence scoring method to combine LDA with DPC, which promotes

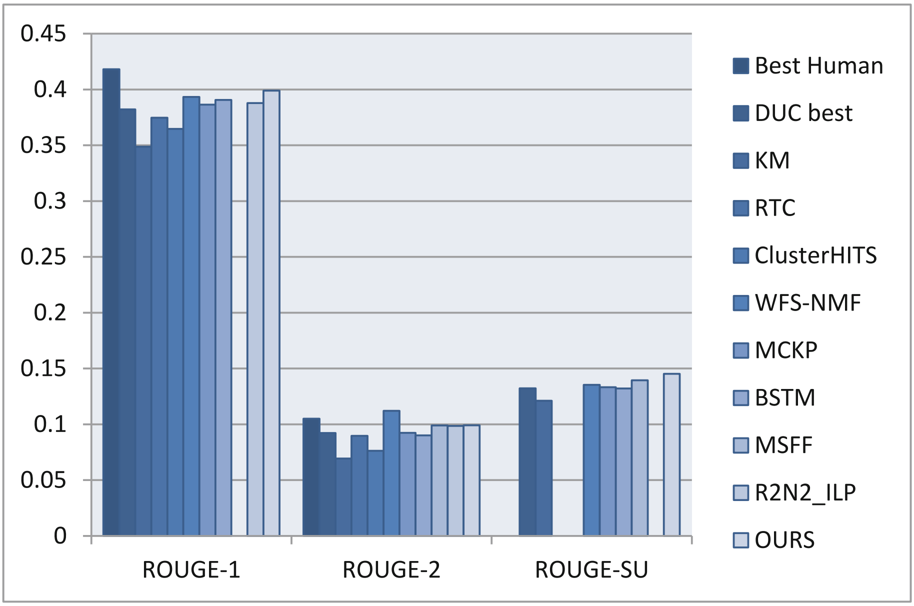
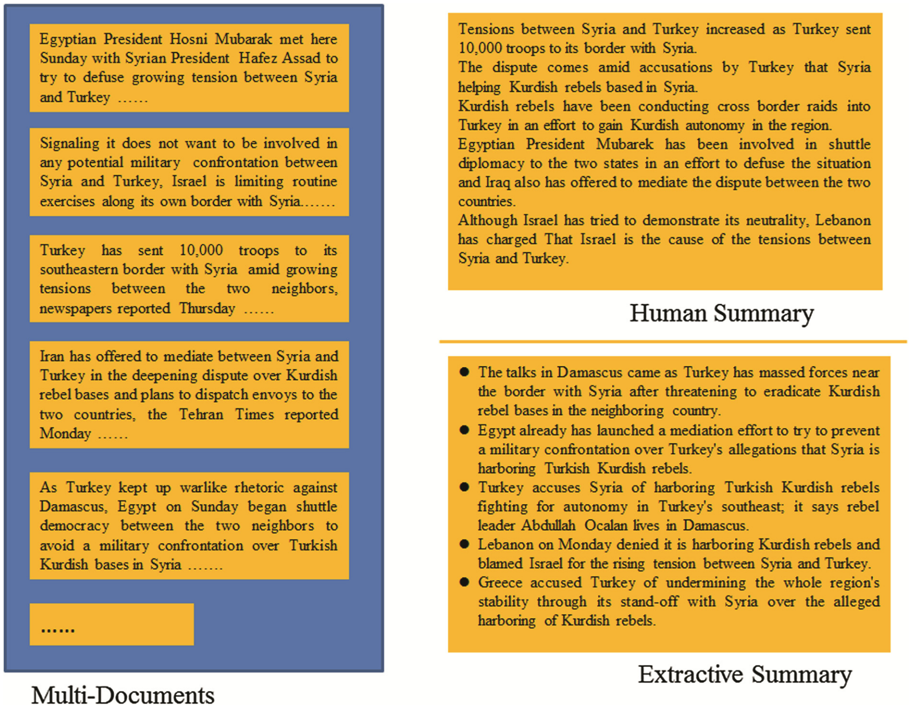


Fig. 2. Summarization results between ours and other state-of-the-art methods

the results mutually and ensure higher quality of the summaries. Compared with other cluster based method, our method removes redundancy when clustering and considers the informativeness of sentences. Our method performs slightly worse than WFS-NMF, MSSF and R2N2_ILP on ROUGE-2 score. It may due to our unigram-centric approach upon which text representation is built. Besides, those methods are complex and even need multiple features and postprocessor. Also, it is extensible for our integrated scoring method to introduce more features like position features.

Figure 3 shows the example of a document set and its result. We pick up one from the document sets randomly, D30028t, to show the human summary and the extractive summary of our method. The D30028t document set talks about the rising tension between Syria and Turkey. Looking at the results by our method in Table 8, each of the sentences represents one cluster respectively and summarizes well specific topics of each cluster. Note that the sentences of the extractive summary for each topic are not just discriminative but they also present the essence of the topic. The contents of the two summaries are very similar. Firstly, all of them talk the current emergency situation between Syria and Turkey. Then, the related countries, such as Egypt, Lebanon, and Greece, expatiate on their standpoints respectively. Besides, the selected summary sentence is completely dissimilar to the summaries of other topics and at the same time it is very relevant to the core event.



Egyptian President Hosni Mubarak met here Sunday with Syrian President Hafez Assad to try to defuse growing tension between Syria and Turkey ……

Signaling it does not want to be involved in any potential military confrontation between Syria and Turkey, Israel is limiting routine exercises along its own border with Syria…….

Turkey has sent 10,000 troops to its southeastern border with Syria amid growing tensions between the two neighbors, newspapers reported Thursday ……

Iran has offered to mediate between Syria and Turkey in the deepening dispute over Kurdish rebel bases and plans to dispatch envoys to the two countries, the Tehran Times reported Monday ……

As Turkey kept up warlike rhetoric against Damascus, Egypt on Sunday began shuttle democracy between the two neighbors to avoid a military confrontation over Turkish Kurdish bases in Syria …….

……

Multi-Documents

Tensions between Syria and Turkey increased as Turkey sent 10,000 troops to its border with Syria.
The dispute comes amid accusations by Turkey that Syria helping Kurdish rebels based in Syria.
Kurdish rebels have been conducting cross border raids into Turkey in an effort to gain Kurdish autonomy in the region.
Egyptian President Mubarek has been involved in shuttle diplomacy to the two states in an effort to defuse the situation and Iraq also has offered to mediate the dispute between the two countries.
Although Israel has tried to demonstrate its neutrality, Lebanon has charged That Israel is the cause of the tensions between Syria and Turkey.

Human Summary

- The talks in Damascus came as Turkey has massed forces near the border with Syria after threatening to eradicate Kurdish rebel bases in the neighboring country.
- Egypt already has launched a mediation effort to try to prevent a military confrontation over Turkey's allegations that Syria is harboring Turkish Kurdish rebels.
- Turkey accuses Syria of harboring Turkish Kurdish rebels fighting for autonomy in Turkey's southeast; it says rebel leader Abdullah Ocalan lives in Damascus.
- Lebanon on Monday denied it is harboring Kurdish rebels and blamed Israel for the rising tension between Syria and Turkey.
- Greece accused Turkey of undermining the whole region's stability through its stand-off with Syria over the alleged harboring of Kurdish rebels.

Extractive Summary

Fig. 3. Example of summary produced by our summarizer and the reference summary

4 Conclusion

In this paper, we proposed an unsupervised method to deal with the issue of multi-document summarization. We take informativeness, relevance, diversity and length constraint into account concurrently by employing LDA and DPC technique and the integrated sentence scoring method to generate the optimal summary. In our method, LDA was applied to acquire the information of sentences in terms level, while DPC was applied to survey the relevance among sentences in sentences level and diversity of sentences in clusters level in the meantime. Considering the length problem, we propose a length constraint score. By combining the four score of sentences, we finally extract sentences based dynamic programming algorithm. A series of experiments on DUC2003 and DUC2004 datasets demonstrate the excellent effectiveness of our method. The performance of our proposed method achieves a significant progress over a set of typical or recent published approaches. In our future work, we will delve into the text representation methods to acquire the semantic of sentence and combine with our integrated sentence scoring method effectively.

Acknowledgements. This work is partially supported by Key Technologies Research & Development Program of Shenzhen (No: JSGG20150512160434776) and Shenzhen Science & Research projects (No: JCYJ20160331104524983).

References

1. Cao, Z., Wei, F., Dong, L., Li, S., Zhou, M.: Ranking with recursive neural networks and its application to multi-document summarization. In: AAAI, pp. 2153–2159 (2015)
2. Li, L., Zhou, K., Xue, G.-R., Zha, H., Yu, Y.: Enhancing diversity, coverage and balance for summarization through structure learning. In: Proceedings of the 18th International Conference on World Wide Web, pp. 71–80 (2009)
3. Ma, T., Wan, X.: Multi-document summarization using minimum distortion. In: 2010 IEEE International Conference on Data Mining. IEEE (2010)
4. Liu, H., Yu, H., Deng, Z.-H.: Multi-document summarization based on two-level sparse representation model. In: AAAI, pp. 196–202 (2015)
5. Mei, Q., Guo, J., Radev, D.: DivRank: the interplay of prestige and diversity in information networks. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1009–1018 (2010)
6. Wang, D., Li, T., Ding, C.: Weighted feature subset non-negative matrix factorization and its applications to document understanding. In: 2010 IEEE International Conference on Data Mining, pp. 541–550 (2010)
7. Li, J., Li, L., Li, T.: Multi-document summarization via submodularity. Appl. Intell. **37**, 420–430 (2012)
8. Arora, R., Ravindran, B.: Latent Dirichlet allocation and singular value decomposition based multi-document summarization. In: 2008 Eighth IEEE International Conference on Data Mining, pp. 713–718 (2008)
9. Wang, D., et al.: Multi-document summarization using sentence-based topic models. In: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers. Association for Computational Linguistics (2009)

10. Takamura, H., Okumura, M.: Text summarization model based on maximum coverage problem and its variant. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, pp. 781–789 (2009)
11. Goldstein, J., Mittal, V., Carbonell, J., Kantrowitz, M.: Multi-document summarization by sentence extraction. In: Proceedings of the 2000 NAACL-ANLP Workshop on Automatic summarization, vol. 4, pp. 40–48 (2000)
12. Cai, X., Li, W.: Ranking through clustering: an integrated approach to multi-document summarization. IEEE Trans. Audio Speech Lang. Process. **21**, 1424–1433 (2013)
13. Wan, X., Yang, J.: Multi-document summarization using cluster-based link analysis. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 299–306. ACM (2008)
14. Zhang, Y., et al.: Clustering sentences with density peaks for multi-document summarization. In: Proceedings of Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL (2015)
15. Wang, B., Zhang, J., Liu, Y., Zou, Y.: Density peaks clustering based integrate framework for multi-document summarization. CAAI Trans. Intell. Technol. **2**(1), 26–30 (2017)
16. Lin, C.-Y: Rouge: a package for automatic evaluation of summaries. In: Text Summarization Branches Out: Proceedings of the ACL-04 Workshop (2004)
17. Rodriguez, A., Laio, A.: Clustering by fast search and find of density peaks. Science **306**, 1910–1913 (2014)