Football News Generation from Chinese Live Webcast Script

Tang Renjun[1,2], Zhang Ke[1,2], Na Shenruoyang[1], Yang Minghao[1,3], Zhou Hui[1,2], Zhu Qingjie[1,2], Zhan Yongsong[2],Tao Jianhua[1,3,4]

1. National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China
2. Guangxi Key Laboratory of Trusted Software, Guilin University Of Electronic Technology, Guilin, China
3. CAS Center for Excellence in Brain Science and Intelligence Technology, Institute of Automation, Chinese Academy of Sciences, Beijing, China
4. School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing, China

Abstract. Challenges exist in the field of sports news generation automatically from webcast that (1) finding hot events and sentences accurately; (2) organizing the selected sentences with highly readability. This paper proposes a framework to generate sports news automatically. First, to obtain accurate hot events and sentences, we design a neural network to predict the probabilities that each statement in live webcast script appears in the writing news, where the inputs of the neural network are weighed word vectors obtained from football keywords dictionary, and the outputs the similarity of statements in training live webcast script and sentences in training news. In this way, the "good" sentences selected from webcast contribute to the semi-finished sport news. To make the generated news to be possibly similar to human writing, we adopt idioms often appeared in football game to describe or summarize the games' development or turns between the selected sentences, and come into being the final sport news. The proposed framework are validated on the training and test data set proved by "Sports News Generation from Live Webcast scripts" task of NLPCC 2016, the experiments show that the proposed method present good performance.

Keywords: Football news generation, neural network, live webcast script

1 Introduction

Football news always are written by football news journalist after the game. So it is necessary for a journalist to record the key events happened in the match. However, it is inefficient to spend time on writing the report after the game. How to write and publish football news immediately became a challenge for journalist and media website. Fortunately, many Chinese media provides football game live webcast script on the websites. It is possible to generate a football news by using the live webcast script and automatic summary technique.

To generate a Chinese football news automatically, two problems should be solved: (1) extract appropriate information from the live webcast script and (2) improve the readability of the generated news. To solve the first problem, we applied keyword weight to evaluate the important of the sentences in the script, because a particular football game can be described through some important sentences. Furthermore, a football related keywords dictionary and a neural network are established to improve the performance of the information extraction. As for the second problem, we introduced a correlation sentence detection method to improve the description of the football news. Besides that, several event recognition methods and preinstall templates are designed to enhance the readability of the whole football news.

According to the experiment part and conclusion part in this paper, our system obtains quite well results in ROUGE evaluation and artificial scoring. Our system can be applied to generate Chinese football news from live webcast scripts. Due to the improvement we made, the generated news is able to descript the football game correctly and vividly. With the help of our system, journalists will be liberated from the repeated and heavy works.

2 Related work

Recently years, news generation and automatic writing have made good progress in fields such as economy, finance, opening speech and so on[1,2]. For example, Dixon implemented a financial news generation system which generates financial news automatically. Automated Insights, an American company, writes prose narrative of 150 ~ 300 words automatically according to the data of Zacks Investment Research[3]. A group from Capital Normal University implemented automatic writing of opening speech by their Chinese intelligent authoring system. The mentioned systems reflected the latest research progress of news generation and automatic writing. However, sports news generation is still a challenge for researchers in the field of Natural Language Processing.

Heretofore, automatic summarization technology has been developed for a long time, and is widely applied in generating news[4]. Two main methods are employed in this area, which are extraction method and generation method.

Generation method is mainly based on natural language understanding. It focuses on analyzing the grammar and meaning of the text. It uses the information fusion method to generate the summary. Comparing to generation method, extraction method is much simpler and easier for practical application. Extraction method tries to divide the original text into small units, and each unit is given a weight. Extraction method follows some rules to select the most important units and assemble them into summarization text. There are some famous automatic summarization systems in the world, such as *NeATS*[*5*,6]，*NewsBlaster*[7] and *NewsInEssence*[8]. In China, Tencent has developed an automatic summarization system called dreamwriter, which published financial news in September 2015.[①] As for the field of sports, an automated

[①] http://js.qq.com/a/20150911/016002.htm storytelling system called *Heliograf* is used by Washington Post for reporting the 2016 Rio Olympics. So far, it can only generated several short sentences about the key information.[②]

Our task for generating sports news report is similar to single document summary. Our method combines the advantages of these two methods, and we also make improvements on them. We adopt the concepts of extraction method to build the main structure of the news, which is easy to be realized. Furthermore, the generation method is applied to improving the readability of the news.

3 Sports news generation

The proposed method consists of four steps: keyword processing, sentence selection, paragraphs writing and report refinement. On the first two steps, we focus on extracting information from webcast script accurately. As for the other two steps, our target is to improve the readability of the generated news.

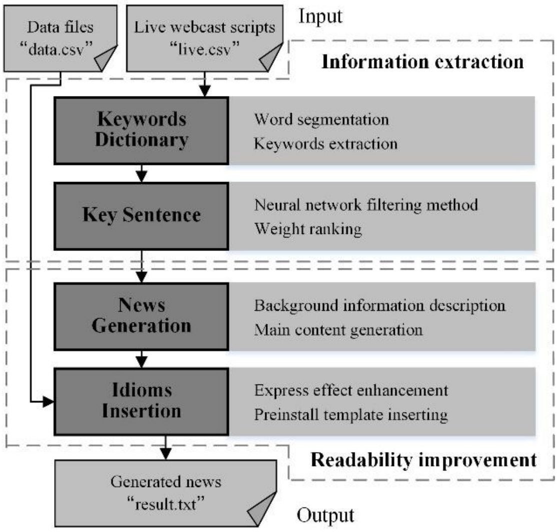


Fig. 1 Framework of the proposed method

3.1 Keywords dictionary

This part mainly introduces the solutions for establishing keywords dictionary. In order to obtain the keywords, the words must be segmented in Chinese sentences first[9]. To ensure the football player names and team names are segmented correctly.

[②] https://www.washingtonpost.com/pr/wp/2016/08/05/the-washington-post-experiments-with-automated-storytelling-to-help-power-2016-rio-olympics-coverage/

We collect plenty of data of proper names and the corresponding part-of-speech (POS) tags to enrich the word segmentation dictionary.

To evaluate the importance of keywords, Term Frequence-Inverse Document Frequency (TF-IDF) method was used to calculate words weight. TF-IDF is a statistical method which tends to filter common words and reserve important keywords. The weights are obtained from the following formula:

$$W_{i,j} = TF_{i,j} \times IDF_i = \frac{n_i}{\sum_k n_{k,j}} \times \log \frac{|d|}{\left|1 + \sum\left\{j\middle|T_i \in D_j\right\}\right|}$$

In this formula, $W_{i,j}$ is the weight value of word T_i in document D_j. n_i is the sum of T_i , $\sum_k n_{k,j}$ is the total number of words in D_j. d means the total number of documents in corpus, $\sum\left\{j\middle|T_i \in D_j\right\}$ means the sum of the document which contain T_i .

We extract the top 30 keywords from each reference news, then reserve verbs and football-specific words, which remove unrelated word and meaningless single-character word.

$$\mathrm{L} = \left\{word_i | pos_i \in \{verb, noun\}\right\}$$

L is the extracted keywords dictionary. In this dictionary, the POS tag of word is verb or noun. However, the words such as red card, injury and own goal appear in low frequency. But these words are also important for football news. Therefore, these words are added to the keywords dictionary particularly. The weights of these words are set at the maximum value. Part of keywords dictionary is shown in the table below.

Table.1 Part of keywords dictionary

Word	Part of speech	Frequency	Weighting
起脚	v	18	0.5
绝杀	v	12	1.0
门框	n	43	0.5

3.2 Key sentences

On sentence level, our work concentrates on calculating the importance of live webcast sentence. Two methods are employed to extract key sentences. One is basing on keyword weights, the other is neural network filtering method.

Sentence weight calculation

The importance of sentence depends on how many keywords it contains. Basically, sentence weight is the algebra sum of the weights of each word. The weight of the sentence is calculated as following:

$$W_{sentence} = \sum_k w_k \, , w_k \in sentence$$

w_k stands for the weight of k th word in sentence. Finally, key sentences are selected by weight ranking. Sentences of high weight usually descript an important event in football news.

Good sentences selected by neural network

However, it is unreliable to extract sentences or information only by weight ranking method. Therefore we try to train a neural network model to decide whether the sentence is the description of events which often appear in football news. Each sentence is a sequence of words, and the word embedding method is applied to represent the words in vector. The vectors of words are used to represent the sentence vector.

$$V_{sentence} = (V_{W_1}, V_{W_2}, \dots, V_{W_n}) \text{ , } W_k \in sentence \text{ , } k \leq 20$$

W_k is the k th keyword of sentence. V_{W_k} is the vector of W_k. $V_{sentence}$ is the vector representation of the sentence. The length of the vector of each word is 50. The vector of sentence is consisted of 20 vectors of the keyword. Stop words such as player names and team names are abandoned.

The neural network model is designed as Restricted Boltzmann Machines[10]. The input of network is the vector of sentence, the output of network is an estimated similarity which indicates the correlation degree between input sentence and reference news[11]. The structure of network is shown below:

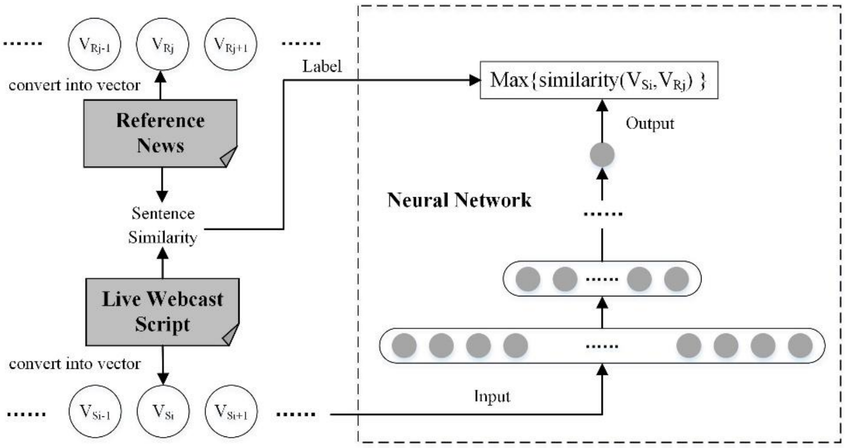


Fig. 2. Structure of the neural network

The network is trained by sentence in live webcast script and its corresponding similarity estimation. The similarity is evaluated by the cosine similarity between the sentence in the script and the sentence from reference news. In order to obtain enough reference sentences of football news, we collected 600 additional football news that contain nearly 24000 sentences. A threshold is set to extract important sentences. Some rules are built to improve the result of the network extraction. For example, if the sentence contains keywords of shooting or foul, then the weight of the sentence is increased. As a result, weights of all sentences have been adjusted and sentences could be ranked by weights to extract the most important one.

3.3 News generation

Reference football news has a uniform format. The first part is the description of basic background information such as game time, place and teams. The second part is the description of the process of the game, and the third part is the starting line-up and player list.

In the first paragraph, sentence weights are ranked to extract the top sentences. In the main paragraph, the events may not be described completely only with the key sentences. Thus, a correlation sentence detection method is proposed to choose the correlating sentences of the key sentences[12,13]. The correlating sentences are added as the supplement of the event. For example, a key sentence may possess very high weighting value, but it is an incomplete description of event. It is possible to calculate the correlation by using sentence vector. Sentence correlation can be represented by the following:

$$R_{n,k} = (1 - \log_d(|k - n| + 1))$$

In the formula, $R_{n,k}$ is the correlation value between sentence n and k. $|k - n|$ means the distant between sentence k and n. The correlation of two neighboring sentences depends on the distant between them. The $R_{n,k}$ above the certain threshold would be chose as the correlating sentences. Finally, the selected sentences are sorted by time to formulate a whole paragraph. What's more, the time information is added before each event to enrich the event completeness. In the final paragraph, the starting line-up and player list can be generated by extracting information from data file of the participating teams.

3.4 Readability improvement

After the football news is generated, several phrases have to be inserted to improve the readability of news. According to the different type of events in the game, several kinds of templates are designed to generate colloquial phrases. The phrases are made up of preinstall text template whose subject is vacant. Those templates are classified as two different types, one type is motion action which contains shooting, foul, cutting and so on, the other type is score rewriting such as first goal, final-hitting, overtaking, draw, etc.

The motion action event is recognized by detect the motion keywords in the sentences. Thereby, the behavior body of the motion is extracted and added to the suitable templates. Score rewriting event is recognized by analyzing the data statistics file and score recording history. The templates are chose to describe the score events according to the different score changing conditions.

4 Experiments

This part, we will demonstrate some experiment results from three aspects.

4.1 Keywords collection

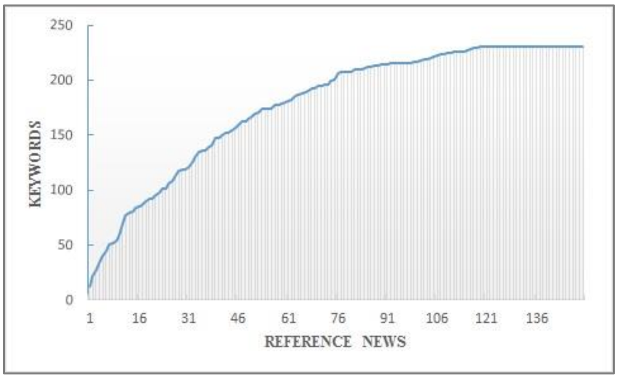


Fig. 3 The number of keyword changes with the total number of the reference news.

As the number of the football news is increasing, the extracted number of keywords would converge to a certain value. In our experiment, the number of keywords would achieve 230 or so, when we extract more than 120 football news. All these keywords is stored in keywords dictionary. The relation between the scale of the keywords dictionary and the scale of the sports news is shown below. According to the result of Fig.3, the number of keywords stored in our keywords dictionary is enough for information extraction.

4.2 Weights alignment between hot events and key sentences

To test the accuracy of our method, we introduce time-weight graph to directly show the comparison between event and sentence's weight.

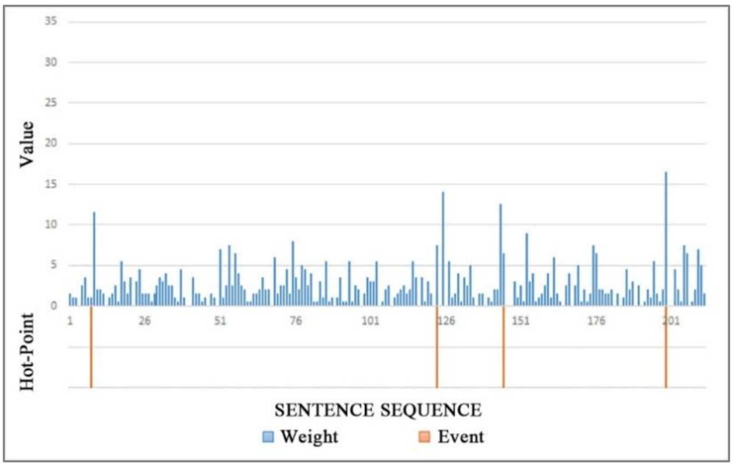
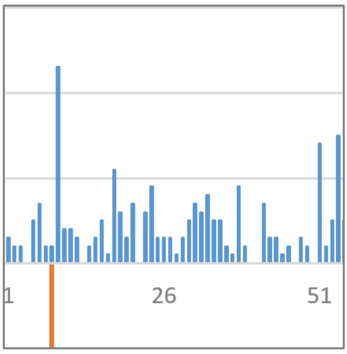


Fig. 4 (a)



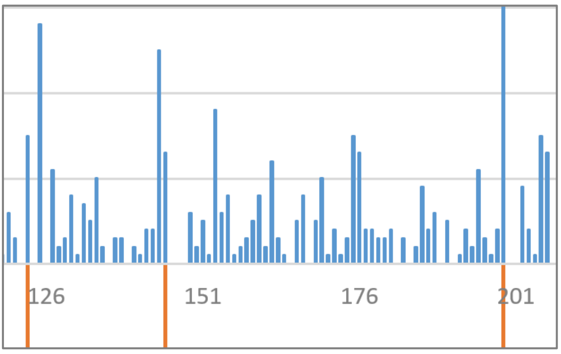


Fig. 4 (b)

Fig. 4 (c)

Fig. 4 (a): Sentences weight calculated by keyword weight summation and scoring event. Fig. 4 (b) and Fig. 4 (c) are the enlarge view of Fig. 4 (a).

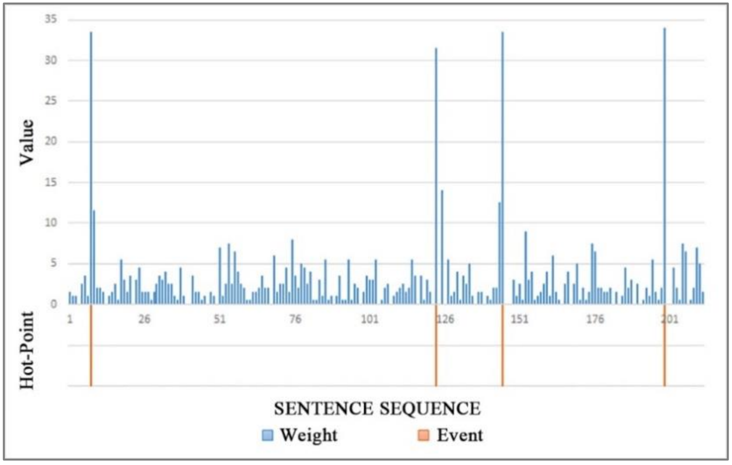
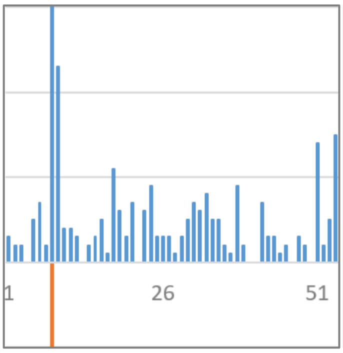


Fig. 5 (a)



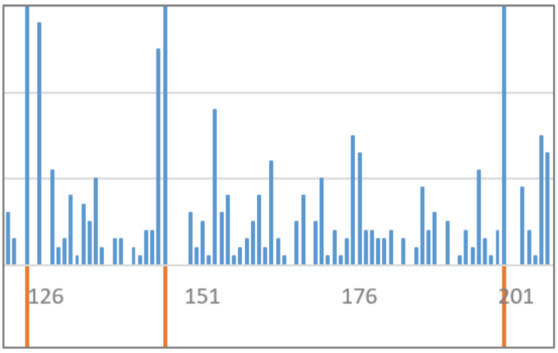


Fig. 5 (b)

Fig. 5 (c)

Fig. 5 (a) is the amendment of Fig. 4 (a), Fig. 5 (b) and Fig. 5 (c) are the enlarge view of Fig. 5 (a).

Fig. 4 (a) shows the relation between scoring event and weights of the sentences which are not adjusted by neural network and rules. The scoring event usually appears on the peak of the weight or near it. If sentence extraction is only based on weight, some important events will be missed. Fig. 5 (a) demonstrates the adjusted result by neural network and rules.

According to Fig. 4 (a) and Fig. 5 (a), our method is reliable for extracting important information. The other peaks of weight columns represent for special event such as foul, passing and attack.

We tested on 110 live webcast scripts, as a result, our method successfully extracts the hot point in the football game from live webcast scripts. The extracted high value sentences usually describe shooting, stealing, saving and so on, which are similar to normal news.

4.3 Scores of automatic evaluation

ROUGE toolkit is widely used for automatic evaluation of summaries. In this paper, ROUGE2.0 is chosen to evaluate our generated news[14,15]. 30 football news are set as reference document and all these news are provided by NLPCC 2016 shared task. Therefore, this news is professional and typical. We use ROUGE toolkit to test the performance of initial generated news, final generated news and a manually writing version. Initial version is the result without any weight adjustment and readability improvement, Final version is the result for weight adjustment and readability improvement, manual version is one of the reference documents which are also provided by NLPCC 2016 shared task. In the properties file of the ROUGE toolkit, we set configuration parameter *ngram* as *1* and *2*, *pos_tagger_name* as *chinese-nodistsim.tagger*. Other configuration parameters are set as default. The performance is shown in the following tables:

Table 2 Average scores of the ROUGE-1

ROUGE-Type	**System Name**	**Avg Recall**	**Avg Precision**	**Avg F-Score**
ROUGE-1	*Initial version*	0.135278	0.531720	0.214376
ROUGE-1	*Final version*	0.494006	0.506041	0.496974
ROUGE-1	*Manual version*	0.501425	0.501036	0.497021

Table 3 Average scores of the ROUGE-2

ROUGE-Type	**System Name**	**Avg Recall**	**Avg Precision**	**Avg F-Score**
ROUGE-2	*Initial version*	0.069832	0.300113	0.112575
ROUGE-2	*Final version*	0.247545	0.385886	0.296926
ROUGE-2	*Manual version*	0.217605	0.211780	0.212743

According to Table 2 and Table 3, our final version of generated news has a better performance than initial version. The improvement of final version is obviously. In the Table 1, the score of the generated news of our final version is close to the manual version. In the Table 2, our final version of generated news acquires a higher score than manual version.

Besides, Table 4 is automatic and manual evaluation results provided by organizers. Table 4 list the automatic evaluation results of 7 teams who participated in the share task5 of NLPCC2016 and Table 5 is the manual evaluation results of top 3 teams. IACAS_Human_HCI is our team.

Table 4 Automatic evaluation results

Team		**ROUGE-1**		**ROUGE-2**		**ROUGE-SU4**	
		Recall	F-measure	Recall	F-measure	Recall	F-measure
IACAS_Human_HCI	1	**0.57782**	0.59846	0.24998	0.26293	**0.25464**	0.26652
	2	0.55643	**0.60331**	0.24448	0.26092	0.24777	0.26581
ICDD_SportsNews	1	0.56515	0.59261	0.25235	0.26444	0.25404	0.26613
	2	0.56768	0.59179	0.25059	0.26119	0.25438	0.26497
RDNH	1	0.55235	0.5865	**0.25527**	**0.27081**	0.25333	**0.26863**
BIT_Coder	1	0.49728	0.55851	0.22524	0.25333	0.22484	0.25263
CQUT_AC996	2	0.5222	0.55728	0.22182	0.23688	0.22689	0.2422
CCNU2016NLP	1	0.46105	0.52478	0.19486	0.22128	0.19322	0.21947
	2	0.4948	0.52425	0.20894	0.22123	0.21102	0.22325
BIT_Hunter	2	0.36532	0.47758	0.16072	0.2106	0.16504	0.21589

Table 5 Manual evaluation results

Team	**Run**	**Aspect**	**Average**
IACAS_Human_HCI	1	Read.	3.84444
		Cont.	3.54444
		Overall	3.63333
	2	Read.	**3.88889**
		Cont.	**3.64444**
		Overall	**3.73333**
ICDD_SportsNews	1	Read.	3.34444
		Cont.	3.32222
		Overall	3.24444
	2	Read.	3.46667
		Cont.	3.32222
		Overall	3.28889
BIT_Coder	1	Read.	2.55556
		Cont.	2.74444
		Overall	2.45556

Each team is permitted to submit two version, our results win three best one in Automatic evaluation results. In manual evaluation, results are evaluated by three factors: readability (Read.), content coverage (Cont.) and overall score. Because of our efforts in the readability improvement, our results win the highest and the second-highest average scores in manual evaluation. In summary, our method has a good performance both in automatic evaluation and manual evaluation.

5 Conclusion

In this paper, a method of neural network and key weighting is proposed for football news generation. It utilizes live webcast scripts to generate a sport news automatically, all the data related to the game are available on the Internet. The football news is generated by the method in this paper has obtained a good evaluation in the competition of NLPCC2016 task5 —Sports News Generation from Live Webcast Scripts, which is held by Technical Committee of Chinese Information, China Computer Federation. We obtained good performance both in automatic evaluation and manual evaluation.

Acknowledge

This work is supported by the National Key Research & Development Plan of China (No. 2016YFB1001404), the Strategic Priority Research Program of the CAS (Grant XDB02080006), the National High-Tech Research and Development Program of China(863 Program) (No.2015AA016305), the National Natural Science Foundation of China (NSFC) (No.61425017, No.61332017, No.61375027, No.61203258, No.61273288), the Strategic Priority Research Program of the CAS (Grant XDB02080006), and the Guangxi Science and Technology Development Project(No: 1598018-6), the Guangxi Key Laboratory of Trusted Software of Guilin University of Electronic Technology(KX201514).

References

[1] Schiller V H. System, report, and method for generating natural language news-based stories: US, US8494944[P]. 2013.

[2] Dixon T. FINANCIAL NEWS GENERATION SYSTEM:, WO/2012/119247[P]. 2012.

[3] Tornoe R. Learn to Stop Worrying and Love Robot Journalists[J]. Editor & Publisher, 2014.

[4] Wan X, Yang J, Xiao J. Manifold-Ranking Based Topic-Focused Multi-Document Summarization[C]//IJCAI. 2007, 7: 2903-2908.

[5] Hovy E, Lin C Y. Automated text summarization and the SUMMARIST system[C]//Proceedings of a workshop on held at Baltimore, Maryland: October 13-15, 1998. Association for Computational Linguistics, 1998: 197-214.

[6] Lin C Y, Hovy E. From single to multi-document summarization: A prototype system and its evaluation[C]//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2002: 457-464.

[7] Evans D K, Klavans J L, McKeown K R. Columbia newsblaster: Multilingual news summarization on the web[C]//Demonstration Papers at HLT-NAACL 2004. Association for Computational Linguistics, 2004: 1-4.

[8] Radev D, Otterbacher J, Winkel A, et al. NewsInEssence: summarizing online news topics[J]. Communications of the ACM, 2005, 48(10): 95-98.

[9] Min K, Ma C, Zhao T, et al. BosonNLP: An Ensemble Approach for Word Segmentation and POS Tagging[C]//National CCF Conference on Natural Language Processing and Chinese Computing. Springer International Publishing, 2015: 520-526.

[10] Chuang W T, Yang J. Extracting sentence segments for text summarization: a machine learning approach[C]//Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2000: 152-159.

[11] Zhang Q, Huang X, Wu L. A new method for calculating similarity between sentences and application on automatic text summarization[C]//Proceedings of the first National Conference on Information Retrieval and Content Security. 2004.

[12] Conroy J M, O'leary D P. Text summarization via hidden markov models[C]//Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2001: 406-407.

[13] Zhang P, Li C. Automatic text summarization based on sentences clustering and extraction[C]//Computer Science and Information Technology, 2009. ICCSIT 2009. 2nd IEEE International Conference on. IEEE, 2009: 167-170.

[14] Lin, Chin-Yew. 2004a. ROUGE: a Package for Automatic Evaluation of Summaries. In Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004), Barcelona, Spain, July 25 - 26, 2004.

[15]Lin C Y, Hovy E. Automatic evaluation of summaries using n-gram co-occurrence statistics[C]//Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. Association for Computational Linguistics, 2003: 71-78.