Research on Summary Sentences Extraction Oriented to Live Sports Text

Liya Zhu[1], Wenchao Wang[1], Yujing Chen[1], Xueqiang Lv[1(✉)], and Jianshe Zhou[2]

[1] Beijing Key Laboratory of Internet Culture and Digital Dissemination Research, Beijing Information Science and Technology University, Beijing, China
lxq@bistu.edu.cn

[2] Beijing Advanced Innovation Center for Imaging Technology, Beijing, China
zhoujianshe@cnu.edu.cn

Abstract. In order to enable automatic generation of sports news, in this paper, we propose an extraction method to extract summary sentences from live sports text. After analyzing the characteristics of live sports text, we regard extraction of summary sentence as the sequence tagging problem, and decide to use Conditional Random Fields (CRFs) as the extraction model. Firstly, we expend the correlated words of keywords using word2vec. Then, we select positive correlated words, negative correlated words, time and the window of score changes as features to train the model and extract summary sentences. This method get good results on the evaluation indicators of ROUGE-1, GOUGE-2 and ROUGE-SU4. And it shows that this method has a meaningful influence on automatic summarization and automatic generation of sports news.

Keywords: Sports news · Live sports text · Conditional Random Fields · Word2vec

1 Introduction

With the rapid development of information technology, the internet information, as a brand new information communication platform, is now spreading its influence on every aspect of daily life. Under this circumstance, the information acquisition is becoming more and more convenient. Through the network media, sports news has become one of the main ways to know the sports games. However, compared with the live broadcasts of sports events, the sport news reports have a drawback of hysteresis. So how to improve the efficiency of the news writing, and handle the processing of information collection, news writing and news arrangement in a unified frame work, so to realize the two-step automatic news from "data extraction" to "document generation" will become a popular research direction in the future. At present, the "data extraction" of sports events includes game entity extraction, data mining and the events of dynamic information extraction. Among them, the events of dynamic information extraction is one of the hotspots in the current research, through the extraction of dynamic information, we can easily get the important events, such as the wonderful ball-passing, the

C.-Y. Lin et al. (Eds.): NLPCC-ICCPOL 2016, LNAI 10102, pp. 798–807, 2016.
DOI: 10.1007/978-3-319-50496-4_72

goal-scoring, the interception and the foul goals in a football game. Text Summarization achieves the extraction of information through the text mining, and it is an important means of information filtering and effective way to solve the information overload in the field of Natural Language Processing.

Automatic Summarization was proposed by Luhn [1] in 1985, he put forward an automatic summarization method based on keyword frequency statistics. He weighted each keyword by the word frequency, graded and ranked for the sentences according to the word weights, and extracted the sentence as the summary sentence if it reach the threshold. Prasad Pingali and Varma [2] proposed a feature that separated from the query and another feature that depends on the query, and rated each of the two characteristics, then calculated the score of each sentence by the linear combination of the two characteristics and extracted the sentence as the summary sentence if its score reach the threshold. Lin et al. [3] proposed a method which combine a graph model with time-stamped and the MMR technique to extract the summary sentences. He et al. [4] proposed a strategy of summary sentence selection based on the multi-feature fusion, and they finally realized the extraction through the fusion of two features, the first feature is the correlation characteristics of sentences and query, the second is the correlation characteristics of global sentences. And their method achieved good results. Liu et al. [5] proposed a method based on the HMM model, in his method, the assumption of theme independent in LDA model has been eliminated, and a multi-feature fusion has been used to improve the quality of summary. Cheng et al. [6] built a weight function on multi-features, used a mathematical regression model to train the corpus, then removed the redundant sentences and realized the summaries generation.

In this paper a new method on the dynamic information extraction from the live sport text is been presented, it transforms the dynamic information extraction into summary sentences extraction. First, build a positive keywords set and a negative keywords set by hand, and extend this keywords in semantic level to get more related words. Then the positive correlated words, the negative correlated words, the time and the window of score changes are treated as features in the Conditional Random Fields model for model training. Finally use the model to extract the summary sentences from the live sport text.

2 Expanding Correlated Words Method Based on Word2vec

In this thesis, an efficient method of expanding the correlated words based on word2vec is presented. This method uses a vector model trained by word2vec to represent the words in corpus, and transforms the problem of text-processing into the vector operations in space vector. It computes the text similarity by using vector space model and the cosine distance to realize the expansion of related words, and to strengthen the indication role of keywords in the extraction of summary sentences.

2.1 Model Training on Word2vec

Word2vec is an open source released by Google in 2013, it can translate the words into vectors by using a deep learning algorithm. There are two kinds of training models in

Wrod2vec, the Continuous Bag-Of-Words Model (CBOW) [7] and Skip-gram model [8]. Both of them use a shallow neural networks training algorithm. The basic principle of CBOW is to predict the probability of the word according to the context, however the Skip-gram is to forecast the probability of context according to the word. This paper establishes a prediction model based on Skip-gram, and the model is optimized by the Hierarchical Softmax method. Assume that the training data is $w_1, w_2, w_3 \ldots w_t$, the objective function of Skip-gram is as follows:

$$J(\theta) = \frac{1}{T} \sum_{t=1}^{T} \sum_{-c \le k \le c} log\, p(w_{t+k}|w_t) \tag{1}$$

In the formula (1), $J(\theta)$ represents the objective function, T is the total number of data, c is an important parameter which determines the neighborhood size, and the bigger the value of c is, the longer it takes for data training, therefore the more accurate the results will be.

In respect of optimization, the paper adopts the Hierarchical Softmax algorithm which realizes the representation of characteristic words using the Huffman Binary Tree. It treats the words in output layer as leaf nodes, then weights the words according to the frequency and codes them. In the Huffman binary tree, high frequency words are assigned the shorter paths, low frequency words are assigned the longer paths, and each word has a unique path that can be accessed. So the function of $p(u|w)$ is defined as the formula (2):

$$\mathrm{p(u|w)} = \prod_{\mathrm{j=2}}^{\mathrm{L(u)}} \mathrm{p}(\mathrm{d}_{\mathrm{j}}^{\mathrm{u}} | v(\mathrm{w}), \theta_{\mathrm{j-1}}^{\mathrm{u}}) \tag{2}$$

In the function, L(u) is the path length of root node to u node, θ_j^u is the vector of the *jth* non-leaf node in the path of root node to u node, d_j^u is the code of jth node in the path of root node to u node, v(w) is the vector of w. Finally, we use the algorithm of gradient descent to solve the objective function, and the word vector is generated.

2.2 Correlated Words Extension

Generally speaking, in the field of live sports text, the keywords can express the action theme of the sentence. For example, we can speculate a series of events through the words of "传中", "攻门", "出底线" in the sentence of "桑切斯右路的传中, 吉鲁俯身头球攻门打在冯特身上出底线". It can be seen that some keywords are decisive roles in judging whether a sentence is important or not. On the other hand, if some words co-occurrence frequently in the sentence, there must exist some relevance between them. Therefore, we propose to build a positive keywords set and a negative keywords set manually, and extend these keywords set according to the semantic relevance, finally use the keywords and the extended words to improve the extraction effect of summary sentences.

In the big data environment, the distance between two points in the vector space is exactly the correlation of these two words. So when the vector model training on word2vec is completed, we use cosine distance to measure the relevant weight of

keywords in relation to other words, and the greater the cosine distance is, the more relevant the words are, so we can select Top N the most relevant words to realize words extension. In addition, the calculation formula of cosine function is shown as formula (3), and $distance(w_1, w_2)$ is the cosine distance between the word w_1 and word w_2, the v_{w_1}, v_{w_2} is the vector of w_1 and w_2 respectively.

$$distance(w_1, w_2) = v_{w_1} \cdot v_{w_2} \tag{3}$$

The Tables 1 and 2 shows the related words of "进球" and "拦截" respectively, which obtained by means of the method in this paper.

Table 1. The related words of "进球"

Related words	Cosine distance
射门	0.6894
直射	0.6828
攻门	0.6623
追回	0.6498
领先	0.6408
打门	0.6396
打破	0.6294
僵局	0.6251

Table 2. The related words of "拦截"

Related words	Cosine distance
截断	0.6529
断球	0.6397
挡出	0.6363
扑住	0.6256
扑出	0.6208
解围	0.6107
没收	0.6084
破坏	0.6053

3 Summary Sentence Extraction Based on CRFs

In this paper a new method of summary sentences extraction on live sport text is been presented, it transforms the summary sentences extraction into an equivalent sequence tagging problem, and builds up an automatic extraction model through the Conditional Random Fields. The output of automatic extraction model is a sequence of "1" and "0", if a sentence is judged as the summary sentence, its label is "1", otherwise "0". While it is affected by multiple factors to determine whether a sentence is a summary sentence

or not, according to the characteristics of the live texts for football matches, we select four kinds of features for model training: the positive correlated words, the negative correlated words, the time and the window of score changes.

3.1 Conditional Random Fields

The Conditional Random Fields (CRFs) is a probability statistic model, which was first proposed by Lafferty [9] in 2001. It combines the advantages of the Maximum Entropy Model (MEM) and the Hidden Markov Model (HMM), which overcomes the limitation for the strong independence assumption in HMM, it has a strong ability of feature fusion and can accommodate rich contextual information. On the other hand, the CRFs adopts the global normalization method, and overcomes the making bias problem in MEM. CRFs is one of the best machine learning models which can effectively solve the problem of serialized data partitioning and data annotation, and has been widely applied in the field of Natural Language Processing, such as the task of Named Entity Recognition (NER), Chunk Parsing and Part-of-Speech Tagging, etc.

3.2 Extraction Model

In our method, we transform the problem into an equivalent sequence tagging problem, and build up the automatic extraction model through the Conditional Random Fields. The input of the model is a set of documents that composed of sentences, the output is a sequence of "0" and "1", the tag is "1" if the sentence can be summary sentence, otherwise "0". Assume that the input is $X = \{x_1, x_2, x_3, \ldots, x_n\}$, the output sequence is $Y = \{y_1, y_2, \ldots, y_n\}$, value of y_i is 1 or 0. From the basic principle of random field theory, the probability of y under the given conditions of x is shown as formula (4).

$$\mathrm{P(y|x;w)} = \frac{1}{\mathrm{Z(x|w)}} \exp\left(\sum\nolimits_{\mathrm{j}} \mathrm{w_j F_j(x,y)}\right) \tag{4}$$

$Z(x|w)$ is the normalized constant to ensure the sum of probabilities is 1, the calculation formula is shown as formula(5). $F_j(x, y)$ is the jth feature of X, and its

$$\mathrm{Z(x|w)} = \sum\nolimits_{\mathrm{y}} \exp \sum\nolimits_{\mathrm{j}} \mathrm{w_j F_j(x,y)} \tag{5}$$

$$\mathrm{F_j(x,y)} = \sum\nolimits_{\mathrm{i}} \mathrm{f_j(y_{i-1}, y_i, x, i)} \tag{6}$$

Our goal is to find the weight vector w and make the formula (7) be workable. Finally, we use the gradient ascent method to estimate the CRF parameters and get the weight vector w.

$$\mathrm{y^* = argmax\ P(y|x,w)} \tag{7}$$

3.3 Feature Selection

- The positive correlated words
  The ultimate goal is to extract the sentences that reflect the key events in the live text of football match. Through the observation of the live text, we found that the words such as "进球", "犯规" can be used to identify the key actions, these words bring important guiding role for the extraction of summary sentences, we call them the positive correlated words. In our method, we collect the positive keywords from live text, then use the method of word2vec mentioned above to extend the keywords, thus to get the positive correlated words. Finally we statistics on the number of positive correlated words in every sentence, and treat the number as a training feature, join it into the training model.
- The negative correlated words
  Contrary to the positive correlated words, there also exist some words such as "收看", "嘉士伯" in the live text, these words will lead to the information redundancy, and reduce the accuracy of extraction, we call these words the negative correlated words. We get the negative correlated words through the same method as the positive correlated words, then statistics on the number of negative correlated words in every sentence, and treat the number as a training feature, join it into the training model.
- The time
  Through the comprehensive statistical analysis of the scoring time in soccer competition and live texts, we found that there exist important information and important comments in some periods of time, these periods are the minutes after game starting, the midfield time and a few minutes before match ending. So we select the time as a feature for the model training, and the functions of characteristic time are defined as follows (8).

$$F(s) = \alpha f_1(s) + \beta f_2(s) + \gamma f_3(s) \tag{8}$$

$$f_1(s) = \begin{cases} 1, & 0 < x \le T \\ 0, & else \end{cases} \tag{9}$$

$$f_2(s) = \begin{cases} 1, & s \text{ in the break time} \\ 0, & else \end{cases} \tag{10}$$

$$f_3(s) = \begin{cases} 1, & endTime - T_3 \le x \le endTime \\ 0, & else \end{cases} \tag{11}$$

In the above formulas, s is a sentence, F(s) refers to the characteristic time of s, which consists of $f_1(s)$, $f_2(s)$, $f_3(s)$. $f_1(s)$ is the function to judge whether s is in the period of T_1 minutes after game starting, $f_2(s)$ is the function to judge whether s is in the midfield time, and $f_3(s)$ is the function to judge whether s is in the period of T_3 minutes before the match ending. endTime is the time of match ending in live text, and the weight of three periods is α, β, γ respectively. In our experiment, we set $\alpha = 0.18$, $\beta = 0.32$, $\gamma = 0.5$.

- The window of score changes
  The information is especially important before and after the goal, and the goal-scoring means the score changes between the two teams. So we propose to set context window according to the score changes in the live text, and judge whether the sentence is contained in a context window, if in, mark the sentence "1", otherwise mark "0". Finally, we treat the marks as a training feature and join it into the training model.

4 Experiment and Results

4.1 Data Set

The training data this experiment used are 900 live texts that crawled from web, the test data are the 30 sample files provided by NLPCC-ICCPOL 2016. Accordingly, we select key sentences from the standard news in the 30 sample files, and take them as the reference summary.

4.2 Evaluating Indicator

We use the ROUGE-1.5.5 toolkit [10] for evaluation, the toolkit uses multiple evaluation indexes to evaluate the results. It measures summary quality by counting overlapping units such as the n-gram, word sequences and word pairs between the summary results and the reference summary. Here we use the ROUGE metrics–Recall and F-scores in ROUGE-1、ROUGE-2 and ROUGE-SU4 to evaluate the result of this experiment comprehensively.

4.3 Result and Analysis

In the experiment, we manually constructed a positive keywords set and a negative keywords set, used the word2vec to build a word vector on training corpus, and used vector result and cosine distance to achieve the lexical semantic computation, we will select the top 8 words that ranking by the cosine value from big to small for each keyword, so we can get the related words set. Finally, we filtered both two related words set by removing the words that the semantic error are obvious, then got 179 positive correlated words and 43 negative correlated words. Some positive and negative correlated words are shown in Table 3. The next step, we put the positive correlated words, the negative correlated words, the time and the window of score changes as features in CRFs model for training, the trained model is used to the extraction of summary sentence on the test data set.

To verify the effect of related words extension at different number on experimental result, we conducted some comparative experiments while other experimental parameters being unchanged. In the comparative experiments 0, 5 and 15 correlated words are extended, the Baseline is the number of 8 that we extended, the comparison results are shown in Table 4:

Table 3. Positive and negative correlated words set

Word set classes	The words
Positive correlated words set	进球 球门 扑出 旋向 角球 边线 射门 没收 犯规 踢倒 解围 拦截 断球 推射 直射 打门 底线 攻门 挡出 绊倒 打飞 换人 换下 受伤 倒地 流血 包扎 冲撞 黄牌 警告 拉倒
Negative correlated words set	网友 进场 更正 回复 镜头 如何 办法 休息 直播 大家 感谢 收看 结束 再见 图文 关注

Table 4. Results in different number of extensions

Number of extensions	ROUGE-1		ROUGE-2		ROUGE-SU4	
	Recall	F-value	Recall	F-value	Recall	F-value
Baseline	0.587	0.674	0.252	0.269	0.248	0.275
0	0.328	0.362	0.117	0.163	0.212	0.197
5	0.504	0.513	0.146	0.175	0.233	0.247
15	0.556	0.603	0.245	0.254	0.263	0.243

The figures in Table 4 indicate that compared to Baseline, the extracting effect is poor when the keywords are not extended or the number of extensions is small, the reason is that many important words in sentences have not been found, so the accuracy and Recall rate are low. On the other hand, when extending too much of the correlated words, there will be a lot redundant information being extracted, therefore the Recall rate and F-value is low.

To verify the effect of CRF machine learning method on summary sentences extraction from the live sports text, we conducted some comparative experiments while other experimental parameters being unchanged. In the comparative experiments, the Hidden Markov Model (HMM) and the Maximum Entropy Model (MEM) are used to train the model on corpus, the extraction results are shown in Table 5:

Table 5. Results on different models

Method	ROUGE-1		ROUGE-2		ROUGE-SU4	
	Recall	F-value	Recall	F-value	Recall	F-value
CRFs	0.556	0.603	0.245	0.261	0.248	0.266
HMM	0.392	0.477	0.184	0.231	0.197	0.206
MEM	0.385	0.361	0.191	0.226	0.188	0.223

As seen in Table 5, the effect of HMM is not so good, that is because, in the HMM model, each sentence in the corpus is considered as an independent individual, it can not effectively use the complex features, however, there is a certain correlation between sentences in the corpus. And the effect of MEM is not so obvious also, although it can solve the complex problems which combine multi-characteristics well, but it can only use the feature of binarization which only records characteristics appear or not, there is no way to record the strength of the characteristics, so there exists biases in the annotation results.

Through the above comparing experiments, it can be concluded that the proposed method, which based on the correlated words extension and the CRFs machine learning method achieved a good result on the summary sentences extraction in the field of live sport text.

5 Conclusion

From the perspectives of the semantics, the vector representation of words and correlated words extension based on word2vec can effectively solve the synonym and correlated words problem. So the experimental result has a good performance by applying the word2vec to extend the keywords in the live sports text. On the other hand, the CRFs can transform the extraction problem into an equivalent sequence tagging and binary classification problem. In our method, we select positive correlated words, negative correlated words, time and the window of score changes as features to train a CRFs model, and use the model to extract the summary sentences. Experiment shows that it not only improves training efficiency, but also has high precision. And the proposed method has a meaningful influence on automatic summarization and automatic generation of sports news.

Acknowledgements. This work is supported by the National Natural Science Foundation of China under Grants Nos. 61271304, 61671070, Beijing Advanced Innovation Center for Imaging Technology BAICIT-2016003, National Social Science Foundation of China under Grants Nos. 14@ZH036, 15ZDB017, National Language Committee of China under Grants No. ZDA125-26.

References

1. Luhn, H.P.: The automatic creation of literature abstracts. IBM J. Res. Dev. **2**(2), 159–165 (1958)
2. Prasad Pingali, R.K., Varma, V.: IIIT Hyderabad at DUC 2007. In: Proceedings of DUC 2007 (2007)
3. Lin, Z., Chua, T.S., Kan, M.Y., et al.: NUS at DUC 2007: using evolutionary models of text. In: Proceedings of Document Understanding Conference (DUC) (2007)
4. He, T., Shao, W., Xiao, H.S., et al.: The implementation of a query-directed multi-document summarization system. In: 6th International Conference on Advanced Language Processing and Web Information Technology, ALPIT 2007, pp. 105–110. IEEE (2007)
5. Liu, J., Xu, J., Zhang, Y.: Summarization based on hidden topic Markov model with multi-features. Acta Scientiarum Naturalium Universitatis Pekinensis **1**, 027 (2014)
6. Cheng, Y., Silamu, W., Hasimua, M.: Automatic text summarization based on comprehensive characteristics of sentence. Comput. Sci. **42**(4), 226–229 (2015)
7. Mikolov, T., Chen, K., Corrado, G., et al.: Efficient estimation of word representations in vector space (2013). arXiv preprint arXiv:1301.3781
8. Mikolov, T., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems (2013)

9. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of 18th International Conference on Machine Learning, ICML, vol. 1, pp. 282–289 (2001)
10. Lin, C.Y.: Rouge: a package for automatic evaluation of summaries. In: Workshop Text Summarization Branches Out: Proceedings of ACL-2004, vol. 8 (2004)