2017 2nd International Conference on Computer Engineering, Information Science and Internet Technology (CII 2017)
ISBN: 978-1-60595-504-9

# Research on Automatic Writing of Football News based on Deep Learning

KAI ZHANG, JIANSHE ZHOU, WENCHAO WANG, XUEQIANG LV,
WENYAN ZHANG, WEILI WANG, NAREN TUYA and JINSHENG SHI

## ABSTRACT

This paper enclosed the exploration and analysis on the Sports News Generation from Live Webcast Scripts, which is a sub-project in the competition conducted by Shared Tasks in NLPCC- ICCPOL 2016. The task mainly focuses on evaluating document summarization techniques for producing Chinese sports news articles from live webcast scripts. Due to the analytical characteristics of input data set and output news articles, the team found that it is crucial to precisely classify the type of each sentence of the live webcast. Thus, the Character-Level Convolution Networks for Sentence Classification was developed and applied to distinguish the category of sentences. Comparing with the traditional machine learning, our model offers higher accuracy on results but lacks of further processing after extracting key sentences. The team won the second prize in the evaluation.

## KEYWORDS

Automatic Writing, SVM, CNN, Language Intelligence, Semantic Understanding.

## INTRODUCTION

Automatic text summarization is one of the standard tasks in the field of NLP. Starting from Luhn's seminal work (1958), research in automatic summarization has made great progress with many different variants of the problem and several classes of methods explored in the community. The Text Analytics Conference (TAC) and first the Document Understanding Conferences (DUC) have provided the main thrust of research in this area by creating standardized data and evaluation methods.

A large number of recent summarization methods are extractive, which means they pick the most relevant sentences from the original document and aggregate them in some order to produce the output summary.

The first task is to assign sentences importance scores. Most early document summarization methods used the word distribution statistics to first find the most relevant words, and then pick sentences that contain these words (Luhn, 1958;

Zhang Kai, Jianshe Zhou, Wenyan Zhang, Weili Wang, Naren Tuya, Jinsheng Shi, Captial Normal University, Beijing, China
Zhang Kai, Beijing Chinese Language Test Center, Beijing, China
Wenchao Wang, Xueqiang Lv, Beijing Information Science and Technology University, Beijing, China

Baxendale, 1958). In later work, more sophisticated methods were explored that range from using external knowledge (Barzilay & Elhadad, 1997), supervised machine learning based methods (Kupiec et al., 1995), methods based on discourse properties of input text (Marcu, 1995; Mani et al., 1998) and network based methods (Erkan & Radev, 2004).

We used different method to manage this task and encounter various issues, the solution will be described in the next chapter. For improve correct rate of sentence classification which is the most challenging task, we also trained a svm sentence classifier and a c-cnn sentence classifier .at the last of this paper we will compare this two model and try to analysis why c-cnn model works better .

## KEY MEANS

This project is one of Limited style automatic writing, which is also considered as a special case of single document summarization. The syntactic structures of news are kind of fixed: the first and last paragraph usually introduces the time, address of the game, and historical records of corresponding teams. As for the content of body paragraphs, all events are truly presented based on the timeline. The further step is to embed the key information and key sentences from live webcast scripts to generate the sports news with specific templates.

### Three Techniques in the First Paragraph Generation

1. Robust regular expression was improved for recognizing structural information in live webcast scripts, and this information is used to produce in the first and last paragraph. However, extracting information base on the regular expression. Thus, the next approach will be used when the regular expression failed.

2. Through the Web crawler technology, the team gets the plenty historical game live scripts, and the Chinese sports news published by Sina and Sohu. The training set contains the first and last paragraph of sports news, and the input text before the game starts. The main idea is to use CRF to obtain time, address, and other major information in live webcast scripts.

As mentioned before, robust regular expression has the highest priority to extract key information, and conditional random field algorithm has the second priority. Even though the combination of the first two techniques provides pretty high accuracy, the team still explores another backup technique to warrant the news could be smoothly generated.

3. There are two characteristics: the content of live webcast scripts is uncertain, which could lead the bad results of extraction with insufficient key information. Furthermore, the syntactic structure of first and last paragraph in football news is fairly fixed. Based on the two characteristics mentioned before, the first and last paragraph of sports news are finally constructed by automatically matching diverse model with various input.

Methods described above just like Fig.1.

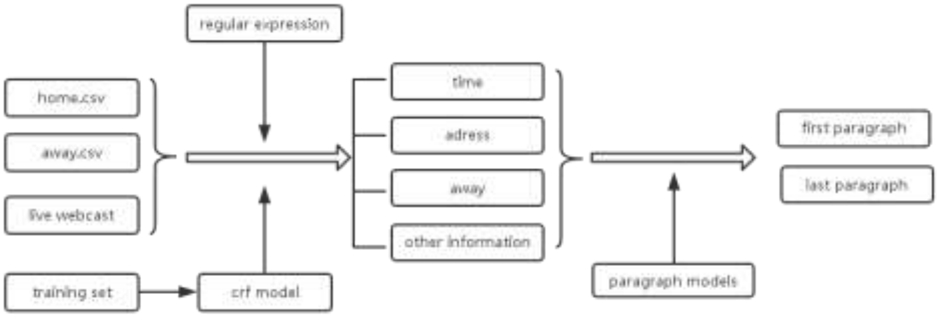

Figure 1. Methods described.

## Techniques in Body Paragraph Generation

Through reading abundant sports news published by web portals, the key points are analyzed as following:

The information of goal in

The information of wonderful shot

The information of flagrant foul

The information of substitution

The information of extra time

The information of penalty

A high quantity news should contain some information of pass and interception when describing the first three key information.

Steps for obtaining the above-mentioned key points.

1. Construct a custom dictionary includes players' name, Chinese name, and abbreviation.

2. Use the custom dictionary to segment the live webcast scripts in order to improve the accuracy of segmentation.

3. After segmentation, the further step is to analyze the word frequency. The 1500 intermediate frequency words are filtered after remove high and low-frequency words.

4. Based on step 3, we develop forward direction dictionary and reverse direction dictionary to develop a recognition system. All verbs are saved in the forward direction dictionary, and negative words are kept in the reverse direction dictionary.

5. Combine the intermediate frequency vocabulary with manually marked keywords, an SVM model is trained for classifying significant sentence.

6. Translate Hanzi to the five-stroke method as input to achieve a Character-level Convolutional Networks for classifying key sentences in all live webcast scripts.

7. Synthesized the flexibility of step 4, the high compatibility of step 5 and 6, the category of sentences can be determined.

The flow chart of this system is shown as Fig.2.

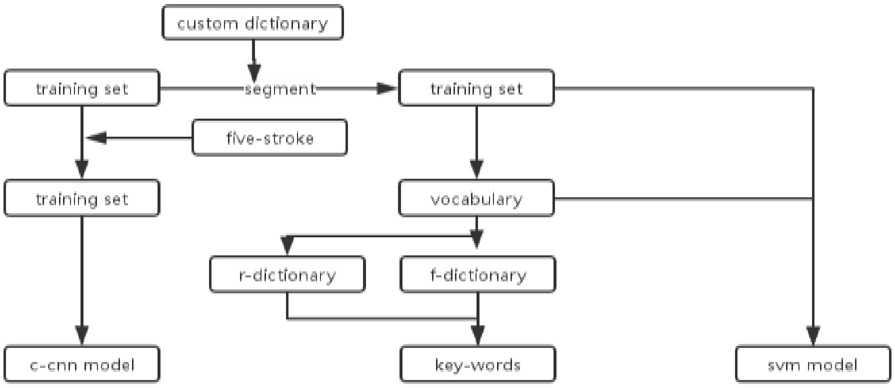

Figure 2. The flow chart of this system.

After determining the above-mentioned program, the system is generated by coding. Our system can fast output two different types of football news in a .txt format file with the information of away team in away.csv, the information of home team in home.csv, and live webcast scripts in live.csv. The team did the following procedure in order to warrant the accuracy of results: a sentences recognition system is developed based on forward reverse direction keyword, an SVM short text categorization evaluation is trained, and the Character-level convolutional Network is improved. The experiment results show that after improving characteristic-level CNN, the short text categorization evaluation holds the highest accuracy.

## SENTENCE CLASSIFIER BASED ON KEYWORDS

Word segmentation is a fundamental problem of the Chinese Natural Language Processing, which is also the major difficulty for automatic football news writing. Differ from English, whose words separated by spaces in the text, but there is no obvious segmentation between Chinese words. It is acknowledged that the low recognition accuracy of Chinese transliterations of member names, places, and teams in English could lead to wrong outputs. Therefore, a web crawler is implemented to collect all transliterations of teams, member names, and the commonly used acronyms, which formed a custom dictionary separated with newline symbol. Such custom dictionary substantially improved the accuracy of segmenting words for live webcast scripts using the classical word segmentation algorithm.

The first problem is to develop a text classifier based on key words Next, the following will describe how to process the corpus and extract the key words using such text classifier.

Firstly, all corpus is processed via the method of word segmentation mentioned before. After analyzing the frequency of all words by python, it is easy to see that some punctuation marks and auxiliary verbs occurs in high frequency. However, these words are removed from the high frequency vocabulary, since they are meaningless for text classifier based on key words. Furthermore, most of low frequency words are member names, places, teams, and other inactive words. And most of middle frequency words are terminology words in football, which well represent the character of whole sentence and determine the content of this sentence. Thus, these middle frequency words are collected in the vocabulary list and manually divided into six classes, including Substitution, Goal, Foul, Shoot, Dribble, Punish. But it cannot well recognize the type of each sentence only depend on these forward direction key words. Consequently, a reverse direction vocabulary is generated for each class, which improved perfectly as tests going.

The calculation is made according to the distribution of words occurrences and the forward/ reverse direction key word lists. So far the text classifier is completed. The following Table 1 shows part of our vocabulary.

## MACHINE LEARNING MODELS

### SVM Text Classifier

Support vector machine (hereinafter referred to as SVM) is a kind of second class classification model.

The basic principle of SVM is shown in the following Fig.3.

The svm classifier was generated by Python and open source libsvm. Libsvm is a simple, easy-to-use, and efficient software for SVM classification and regression. It solves C-SVM classification, nu-SVM classification, one-class-SVM, epsilon-SVM regression, and nu-SVM regression. It also provides an automatic model selection tool for C-SVM classification. Follow the subsequent steps:

Firstly, determine the data structure of input. The format of training and testing data file is:

<label> <index1>:<value1> <index2>:<value2> ...

TABLE I. PART OF VOCABULARY.

| | forward direction | reverse direction |
|---|---|---|
| Substitution | Replaced, substitutions, put on ... no, ready ... | Replaced, substitutions, put on ... no, ready ... |
| Goal | Scored, broken, netted ... goals, goals ... | Scored, broken, netted ... goals, goals ... |
| Foul | Kick turn, overturned, shovel turn ... Sorry, wrong ... | Kick turn, overturned, shovel turn ... Sorry, wrong ... |
| Shoot | Hit the door, push, throw ... ready to hit the door ... | Hit the door, push, throw ... ready to hit the door ... |
| Dribble | Give it, give it, give it a shot ... give it another ... | Give it, give it, give it a shot ... give it another ... |
| Punish | Yellow card, red card, whistled ... should ... | Yellow card, red card, whistled ... should ... |

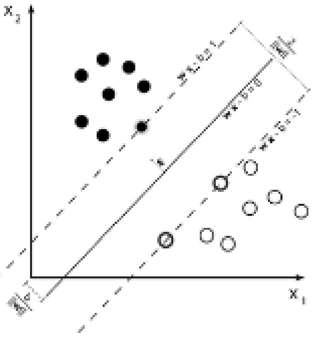

Figure 3. The basic principle of SVM.

Each line contains an instance and is ended by a '\n' character. For classification, <label> is an integer indicating the class label Set “label” as the corresponding number of class and “value” as the index number of this word in the vocabulary list. By this way, a completed input file of training set is generated as a.txt. While training, y, x = svm_read_problem('./a.txt');m = svm_train(y[9000:],x[9000:], '-c 8 '). After training svm_save_model('./football.model', m), the classifier model is made. It is crucial to adjust the parameters while training the model using libsvm. The main idea is to use cross validation to determine the value of c and g, which could lead to the highest accuracy. Thus, the smallest value of c and g are traded as the best, because although high parameter of penalty could get high accuracy of validation data, it will lead to learning status.

There are few word vectors in each sentence after classifying, since the main object is short Chinese phrase. Thus, there is less characters to learn for svm model, which influence the accuracy of svm classifier.

## CNN Text Classifier

It is shown by recent research that cnn can solve not only computer vision, but also nlp problem. Our CNN model is implemented as the description in Kim's Convolutional neural networks for sentence classification. The input model was regulated in order to process Chinese corpus.

The following Fig.4 shows the basic structure of Convolutional Neural Networks.

The corresponding equation is shown as:

$h_{W,b}(x) = f(W^T x) = f(\sum_{i=1}^{3} W_i x_i + b)$

Among them, this unit is also referred as the logistic regression model. The neural network model is produced when multiple cells combined with hierarchical structure. The Fig.5 below shows a neural network with one hidden layer.

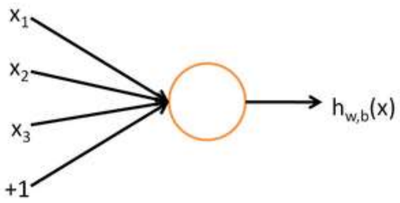

Figure 4. The basic structure of Convolutional Neural Networks.

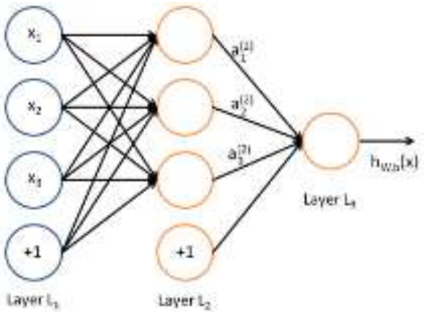

Figure 5. A neural network with one hidden layer.

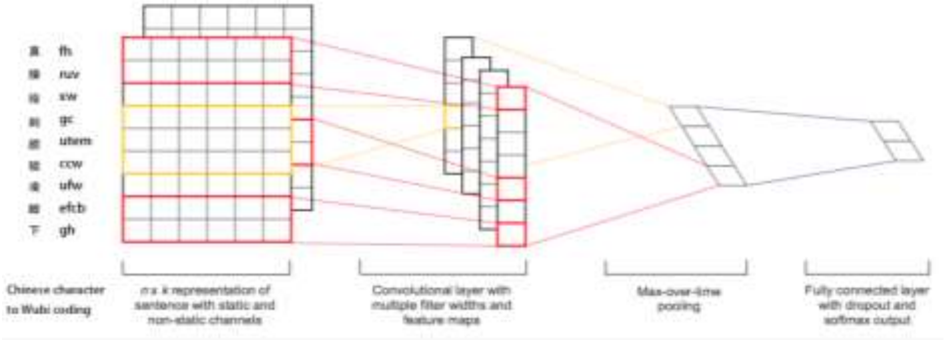
fh
ruv
sw
gc
utem
ccw
ufw
efcb
gh
Chinese character to Wubi coding
n x k representation of sentence with static and non-static channels
Convolutional layer with multiple filter widths and feature maps
Max-over-time pooling
Fully connected layer with dropout and softmax output

Figure 6. The structure of model.

The corresponding equation is shown as:

$$a_1^{(2)} = f(W_{11}^{(1)}x_1 + W_{12}^{(1)}x_2 + W_{13}^{(1)}x_3 + b_1^{(1)})$$

$$a_2^{(2)} = f(W_{21}^{(1)}x_1 + W_{22}^{(1)}x_2 + W_{23}^{(1)}x_3 + b_2^{(1)})$$

$$a_3^{(2)} = f(W_{31}^{(1)}x_1 + W_{32}^{(1)}x_2 + W_{33}^{(1)}x_3 + b_3^{(1)})$$

$$h_{W,b}(x) = a_1^{(3)} = f(W_{11}^{(2)}a_1^{(2)} + W_{12}^{(2)}a_2^{(2)} + W_{13}^{(2)}a_3^{(2)} + b_1^{(2)})$$

The training method of neural network is similar to logistic, but due to its multiple layers, the chain derivative method is used for derivation of hidden layer nodes, which is also called back propagation. This paper does not enclose the training algorithm.

The following Fig.6 shows the structure of our model.

As the figure shows, the input level is corpus, whose characters are encoded to matrix in five strokes (from top to bottom). Assume there are nn words in a sentence, and the dimension of vector is kk, then the matric is n * kn * k.

First convolutional layer

Several Feature Maps are obtained by input level convolution, and the size of convolution window is h * kh * k, which hh represents the number of vertical words, kk represents the dimension of words. Several Feature Maps with one columns are obtained through such a large convolution window.

Pooling layer

Furthermore, the method called Max-over-time pooling is used for pooling layer, which simply extract the max value (represent the most significant signal) from the previous layer of Feature Map. This pooling method can solve the problem of convertible length of sentence. Eventually, the outputs of pooling layer are the max values of each Feature Map, which is a one dimensional vector.

All connection + Softmax layer

The output of pooling layer, which is a one dimensional vector, connect to Softmax layer through all connection. Softmax layer is set according to the needs of the tasks.

Dropout technology is used for the penultimate part of all connection, which means weight the parameter of all layer connection with L2 regularization restriction. The advantage is to prevent the hidden layer unit adaptive, thus reducing the degree of fitting.

## Experimental Result

The experiment data chose the sport news with manually annotation, in order to test the performance of classifying Chinese texts using the models mentioned before.

As known that the quantity of samples in different classes is essential to the result of classification, thus the quantity of sentences of six classes are set to 10000 after randomly discard some data. We use 8900 samples as training set, 1000 samples as val set, 100 samples as test set. Training set looks like Fig.7.

As for svm model, Python jieba, word segmentation tool, is used to load the custom dictionary, segment the words, get the word vector of each word, and input these vector data and manually label data to libsvm for training. As for cnn model, all words in corpus are translated to pypinyin, which is character-level input of cnn. And then translate words into five strokes as character-level input for cnn to produce a set of result, which is shown as Table 2 and the training process of cnn model looks like Fig.8.

For cnn model we achieve Top-1 accuracy 94.0% but the Top-1 accuracy on svm model is 91.08%.

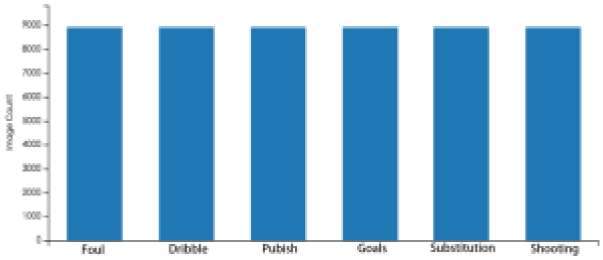

Figure 7. Training result.

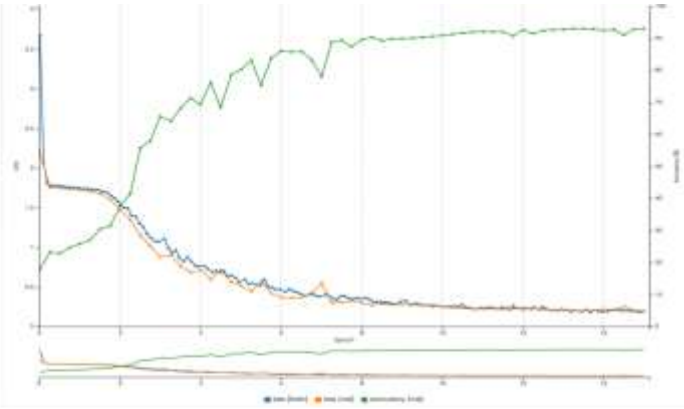

Figure 8. The training process of cnn model.

TABLE 2. CONFUSION MATRIX.

| | Dribble | Foul | Goals | Punish | Shooting | Substitution | Per-class accuracy |
|---|---|---|---|---|---|---|---|
| Dribble | 92 | 1 | 1 | 5 | 1 | 0 | 92% |
| Foul | 3 | 96 | 0 | 0 | 1 | 0 | 96% |
| Goals | 0 | 0 | 96 | 1 | 0 | 3 | 96% |
| Punish | 8 | 1 | 0 | 91 | 0 | 0 | 91% |
| Shooting | 3 | 0 | 0 | 0 | 97 | 0 | 97% |
| Substitution | 0 | 0 | 8 | 0 | 0 | 92 | 92% |

The experiment results show that the performance of traditional text classifier svm is perfect with high accuracy text classification, but is poor with error text classification. The accuracy is high when the input of cnn model is five strokes

According to the character of cnn model, cnn is good at process the raw data, which extract the implicit characters in raw signal by simulating the exchange of information between neurons. However, Chinese characters are hieroglyphs, whose components combine in its own set of deep logic. After modifying the input from pinyin to five-strokes, the character-level cnn can learn more about this kind of character, thus, the accuracy of classification be high.

## Conclusion and Outlook

Automatically generated Football News is one of refined field in document summary and natural language generation. The biggest challenge we faced is that how wherein the maximum extent possible to extract all the key information. When humans writing news, the first step is to grasp the overall context of the whole game, then locking key information base on their experience and intuition, after that they are able to organize an overview of all key information in appropriate form, appear a wonderful game of football news. The system can’t identify important information point game Intuitively as a professional editor at the present stage. Our approach at this stage is to build a model base on machine learning algorithm to learn the contact and difference between critical and non-critical information by reading the manual tagging corpus, and then try to identify the key sentence from all input. However, this method extracted key sentences have some problems. However we found some problems in extracted key sentences when we review them. The quality of key sentences selected by the model is not perfect. The information described by those sentence sometimes mediocre shot in reader angle, on the other hand wonderful and difficult shot was ignored by the model.

In the identification of key words, the human is not simply based on a sentence, but the comprehensive context, and even the whole football game. Therefore, we believe that if we can make the system understand the ins and outs of the whole game, we will be able to establish appropriate model then make the system lock key information from the global. We can then go along this idea, develop our key word recognition model, manual tagging more corpus. It can identify category of all phrases in the broadcast. Finally establish a recognition model kick-off from beginning to end of the game. After understand the progress of the whole game, our system has the ability to grasp the overall situation like the human editor.

## ACKNOWLEDGEMENTS

Correspondent author: Zhang Kai, Wenchao Wang; This paper is supported by the following projects: Beijing Advanced Innovation Center for Imaging Technology (BAICIT-2016019), National Social Science Foundation (No.14&ZH036, No. 17BYY211), National Language Committee Foundation (No.ZDA125-26).

## REFERENCES

1. Kim, Y. (2014). Convolutional neural networks for sentence classification. Eprint Arxiv.
2. Zeng, D., Liu, K., Lai, S., Zhou, G., & Zhao, J. (2014). Relation classification via convolutional deep neural network.

3. Kalchbrenner, N., Grefenstette, E., & Blunsom, P. (2014). A convolutional neural network for modelling sentences. Eprint Arxiv, 1.
4. Zhang, X., Zhao, J., & Lecun, Y. (2015). Character-level Convolutional Networks for Text Classification. Neural Information Processing Systems.
5. Joachims, T. (2002). Learning to classify text using support vector machines. Springer Berlin, 29(4), 655-661.
6. Dash, M., & Liu, H. (1997). Feature selection for classification. Intelligent Data Analysis, 1(3), 131-156.
7. Santos, C. N. D., & Gattit, M. (2014). Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts. International Conference on Computational Linguistics.
8. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, 25(2), 2012.
9. Kishore, A., Singh, S., & Jindal, S. (2010). Designing deep learning neural networks using caffe.