International Journal of Security and Its Applications
Vol. 10, No. 7 (2016) pp.307-318
http://dx.doi.org/10.14257/ijsia.2016.10.7.27

Research on Text Proofreading Method for Judgment Document

XU Yabin[1,2] and Ji Xuan [2]

[1]Beijing Key Laboratory of Internet Culture and Digital Dissemination Research, Beijing Information Science &Technology University, Beijing 100101, China
[2]Institute of Computer, Beijing Information Science &Technology University, Beijing 100101, China
xyb@bistu.edu.cn, 404494394@qq.com

Abstract

Automatic proofreading of judgment document can effectively overcome human factors and ensure the quality of proofreading. The outstanding work in this paper includes following two aspects: The first aspect is to proofread the mistakes of legal file name and legal provision reference by establishing the proofreading knowledge base of judgment document. The second aspect is to maximizing identify legal terminology and common name entities, then proofread the collocation mistakes between words, phrases and legalese by using Markov Model. The experiment results show that this method can basically meet the needs of the judgment documents proofreading.

Keywords: *judgment documents; text proofreading; proofreading knowledge base; legal terminology recognition; Markov model*

1. Introduction

The supreme people's court about the regulation of the people's court announced the judgment documents on the Internet was formally implemented from January 1, 2014. It clearly state that effective judgment documents of the people's courts at all levels shall be published on the Internet, in addition to (1) involving state secrets, personal privacy and minor crime; (2) concluding by mediation; (3) The parties have expressly requested not to be published on the Internet and have a legitimate reason, as well as not involving the interest of public; (4) Others not be published on the Internet.

The judgment document is a document that not only a proof that the parties enjoy in rights and obligations of the burden, but also an important basis for the superior people's court to supervise the civil trial activities of the lower people's court. Promulgating the judgment document on the Internet can effectively protect the public’s right to know and to supervise, and further promote the openness, fairness and justice in judicature. Unfortunately, some of the judgment documents not only exist in various types of vocabulary, syntactic, logical and other expression errors, even there are some reference error in legal documents and legal provisions. The existence of these errors seriously damaged the image of the people's court, undermined the authority and credibility of the people's court, and even affect the effectiveness of the implementation of judgment document. The traditional way of manual correction will inevitably exist errors due to the judge’s limited energy, constrained time, negligence, thinking inertia and other factors. Thus, the proofreading quality is difficult to guarantee. So, proofreading the errors in judgment document by using the text proofreading technology not only can improve the proofreading speed, but also can effectively overcome the human factors, improve the quality of proofreading. Therefore, the study is of great significance.

At present, a lot of researches have been carried out on the text proofreading technology at home and abroad. Research on English text proofreading by foreign scholars start earlier. The dictionary searching method adopted in literature [1], is mainly

ISSN: 1738-9976 IJSIA

used to check whether the word in the text is in the dictionary. The word that does not exist in the dictionary is regarded as an error word. In the English text, the dictionary searching method is of high accuracy, and is a relatively popular error detection method at present. The disadvantage of the dictionary searching method is that we need to build a large dictionary, which results in the decrease of the efficiency of the error detecting system. N-gram model adopted in literature [2], is used in the query by the established N-gram table, those N-gram substring do not exist or have a very low frequency may be identified as an error. For example, "dfc" is a string identified as wrong tri-gram substring (due to its low frequency). In this method, with increase of N, although the more extensive context information can be inspected, the data sparse problem of the matrix can be caused.

The domestic research mainly revolves around the Chinese text. Unlike English, Chinese text between words haven't a natural space division. Although it can be resolved by the Chinese word segmentation technology, however, the difference between Chinese grammar and English grammar is big, some of the method in English text proofreading is difficult to apply to the Chinese text proofreading. The idea of chunk parsing adopted in literature [3], is used to correct the error through constructing a common collocations knowledge base, and combined with a dictionary of Chinese characters and Pinyin. But the overall computing cost is relatively large. In literature [4], through analyzing the common text error type in CSSCI (Chinese Social Science Citation Index), designed and implemented an automatic acquiring subsystem of proofread rules. The higher test result has been made, but with limitations which the method is difficult to apply to other types of texts. In literature [5], through training Bi-gram model, the Chinese text has been revised from two aspects of punctuation errors and typos. What's more, there design and implement a Chinese spelling check system based on the conversion in Chinese characters-Pinyin-Chinese characters. But the noise of the training corpus will be constantly accumulated in the training process by this method. Finally, the proofread accuracy is affected. In literature [6], through constructing collocation knowledge base of words and their meanings, implemented a checking algorithm of errors in text. Although the experiment achieved good results, the construction of knowledge base is relatively complex, and the proofreading effectiveness of complex sentence is poor. The rules and linguistic knowledge, adopted in literature [7], is used to construct knowledge base of political news and text error detection algorithm in different types of errors. Although it has certain practical value, but with strong target.

In conclusion, there is no relevant research on text proofreading of judgment documents. Some existing text correction methods are not designed for the judgment documents. Because of the lack of legal background knowledge and with little target, applying above methods to judgment documents will cause poor proofreading effect.

2. Error Types Analysis of Judgment Document and Our Working Idea

Based on analyzing a large number of the contents and features of judgment document, we find that the high frequency of error types in the judgment document is following 6 classes:

Legal document reference error. Such errors mainly related to reference error about the name and order of Legal document in judgment document. The main problem caused by the reference error of legal document name is legal document reference error. The main problem caused by legal document reference sequence error is that such reference is not in accordance to the force's scale of law and the level of law. This is not consistent with the provisions of the legal document reference [8].

Legal clause reference error. Such errors appear occasionally, including: ①The legal provisions cited in judgment document are not matched with the legal documents and the legal provisions mentioned in the preceding article. ②Legal provisions reference are improper and contrary to the jurisprudence. According to Jurisprudence, entire reference only cite whole legal provision, not clauses and items; Provision-clause style, provision-item style or both style reference only cite part of the legal provisions [14].

Legal terminology confusion error. Such errors are caused by improper use of legal terminology in judgment document [9-10].Legal terminology has a strict legal basis, it can't be arbitrarily fabricated or tampered.

Words, phrases or legal terminology collocation error. Such errors are caused by using incorrect match in words, phrases or legal terminology. For example, "declare" is a word matched with "declaratory judgement", "declaratory bankruptcy", "declaration of death" and so on, but sometimes it is mistaken for "announce judgment", "announce bankruptcy and other collocation errors.

More than words, missing words or typos error. Such errors is of the highest frequency in the text. Some of the above errors may be caused by them. In order to distinguish between them from specific errors in judgment document, we list them here separately.

Punctuation error. Such errors can be detected by computer, mainly including: punctuation repetition, loss, improper collocation, *etc*. For example, the punctuation pairs did not appear in pairs in a sentence; Some words follow the ellipsis.

By deep analysis of the error types in judgment document, we found that types of error (1), (2) and (3) are peculiar, and those of error types have not yet been involved in present study on text proofreading. To this end, we will take it as the emphasis of this paper. Although the common Chinese character and word collocation errors in the error type (4) has been studied and developed by literatures and scholars, and the available research results have been made[11], there are plenty of legal terminologies, legal document names, legal documents and legal provisions, which are composed of a number of compound nouns, and named entity which is composed of place names, organization names and personal names in a judgment document, it maybe not constitute a complete legal terminology or named entity after the processing of word segmentation. In addition, judgment document also involve some proprietary legal terminologies that require strictly both to meet the grammatical relation and to fit in with legal knowledge. Therefore, how to effectively organize the legal terminology or named entity, and accurately identify their collocation is a difficult problem placed in front of us and needed to give effective solutions. As for error types (5) and (6), because there are many scholars that have carried out research[3-5, 12] on these, and the related method have been more mature. So, this paper is only to apply the relevant method other than give further study and discussion.

Through some study on existing methods, we find that the knowledge base which is used by general text proofreading system is a common knowledge base[4,13]. If we directly use it to proofread judgment document, proofreading effect is poor due to lack of professional knowledge of legal field. Therefore, in order to improve the effect of proofreading, we must construct the domain knowledge base for judgment.

On this basis, we determine our research work idea as follows: First, we should construct legal document knowledge base for judgment document, and proofread reference error of the legal document name and legal provisions. And then, to improve the efficiency and accuracy of proofreading, we design an algorithm for maximizing recognition legal terminologies and named entities in a judgment document, and thus build a Markov model based this to proofread legal terminology, name of the entity and collocation errors of words.

3. Error Proofreading for Legal Documents and Legal Provisions

3.1. Legal Document Reference Error Proofreading

The legal documents used in this paper come from the current regulatory legal documents database in China published in February 26, 2015[15]. According to this database, we can construct the knowledge base of legal document (LDB) and use it in this paper. The database includes four aspects of legal document, laws, administrative regulations, local regulations and judicial interpretations, which fields include legal document name, legal items, legal provisions and effect level and so on.

According to above analysis, reference error for a legal document should be proofread from two aspects: legal document name and reference sequence. For the reference error of legal document name, we can use the short text comparison algorithm. If the legal document name queried in the knowledge base LDB can be exactly matched, the reference is right, otherwise, the reference is wrong. For the case of reference error, we need to find the legal document name in LDB which has the longest common substring with the document name to be proofread, and then give corrective advice. The algorithm is as follows:

Algorithm 1: The correction algorithm for reference sequence error of legal documents

Step1: Get the legal name *Ln* in sentence *S*, and inquire *Ln* in the legal document database (Ldb). If there is a complete match in the knowledge base, then end;

Step2: Carry out word segmentation for *Ln*, and get the keyword sequence *Kw{1,2,... N}*, then remove the stop words, respectively retrieve in the Ldb according to key words, get legal document name sequence *List<Ln>*;

Step3: Use Nakatsu algorithm[16] to calculate the longest common substring in *Ln* and *List<n>*, then common substring sequence *List<LCS>* is obtained;

Step4: Arrangement *List<LCS>* in descending order, end.

It can be seen by this algorithm, the time complexity depends on the number of records in Ldb, which is *O(N)*; The time complexity of corrective recommendations is determined by the *m* of the number of the legal file name sequence *List <Ln>* and the complexity of Nakatsu algorithm, which is *O(m*N)*.

Proofreading for legal document reference sequence errors need to be carried out according to the following rules[14]:

(1) Principle that a higher level law is superior to a lower level law: the laws are prior to administrative regulations, local regulations and judicial interpretations, then administrative regulations are prior to local regulations;

(2) Principle of judicial final settlement: judicial interpretations are behind administrative regulations and local regulations;

(3) Principle of the procedure produce the entity result: procedural laws are prior to substantive laws;

(4) Principle that the particularity of contradiction takes precedence over the universality of contradiction: special laws are prior to common laws;

(5) Principle that the applicable circumstance of main laws affect applicable circumstance of subordinate laws: main laws are prior to subordinate laws;

(6) Principle of judicial sovereignty: domestic law is prior to international treaty and foreign law.

Specific to the actual algorithm, as long as we compare the level of legal effectiveness in accordance with the order appeared of legal document. If it fits in with the above rules, it is determined as correct, otherwise determined as wrong.

3.2. Legal Provisions Reference Error Proofreading

For the above two kinds of reference errors in legal provisions, we design a corresponding proofreading algorithm. The main idea of this algorithm is as follows: legal provisions extracted from the judgment document (pattern string *P*) should be aligned to the left with the legal provisions (text string *T*) retrieved from the legal document database (LDB) according to the names and terminologies of legal document, and be matched by the order from right to left. If the match is successful, then continue to move to the left; If not, then the pattern string (*P*) should be slipped to right. Loop until the mode string is moved to the end of the text string.

Suppose *P* is legal provisions extracted from judgment document, *T* is the legal provisions retrieved from the legal document database (LDB) according to the names and terminologies of legal document, *i* is used to save the cumulative slip of the mode string which the initial value is the length of the pattern string, *n*. Specific processes such as algorithm 2:

Algorithm 2: Legal provisions proofreading algorithm

Step 1: Text *T*, which length is *n*, will be left aligned with text *P*, which length is *m*;

Step 2: When $i<n$, please skip to step 3, otherwise, exit;

Step 3: Comparing from right to left, if *P* not yet reach the end, and the match is successful, please skip to step 4; if *P* has come to the end, please skip to step 5, if match unsuccessful, please skip to step 6;

Step 4: Please continue to compare *P* with *T* from right to left, skip to step 3;

Step 5: The match is successful, exit;

Step 6: If j $(0<=j<m)$, the position of *T*, does occur a mismatch character *x*, and if *x* does not occur in *P*, please skip to step 7; if *x* does occur in *P*, please skip to Step 8;

Step 7: *P* directly skip the area and make its header be left aligned with the next position of *x* in *T*;

Step 8: If *x* does occur in the position of *k* ($k<j$ and $k=mas\{0,k\}$) in *P*, then *P* should be moved in order to *k* aligned with *j*, skip to step 2; if *x* only occurs in *P'* which has matched in *P*, and the whole *P'* started with the position of *t* can be found in position *t'* (*t'*<*t*) , and the leading char both in *t* and in *t'* are different, please skip to step 9; If *P'* does not exist in *P*, then find largest prefix *s* of *P* in order to make it equal to *P*'', which is a suffix of *P'*, please skip to step 10; If *s* does not exist, please skip to step 11;

Step 9: *P* should be moved in order to *t'* aligned with a position where *t* was located, skip to step 2;

Step 10: Move *P* to the right, so as to let *s* aligned with place where *P'* was located, skip to step 2;

Step 11: Make *P'*s header be left aligned with the next position of the tail of matched *T'*s suffix.

By this algorithm, we may conclude this method can check wrong with legal provisions reference, when *P* and *T* can't match, the time complexity is the worst, it is $O(n\times m)$. While the best time complexity is $O(n/(m+1))$.

4. Identification of Legal Terminologies and Proofreading of Collocation Error

This study is not directly to construct a Morkov proofreading model based on the results of word segmentation, but to maximize identification of legal terminology and common named entities in sentence, and then use these elements to build a Morkov proofreading model. In this way, on the one hand, it can avoid the problem of excessive computation, while on the other hand, it can avoid the error identification results due to the low probability of co-occurrence between vocabularies.

In the part of common named entity recognition, LTP (language technology platform) provided by HIT-SCIR (Harbin Institute of Technology-Society Computing and Information Retrieval research center) is applied to recognize in this paper, while legal terminology will use the conditional random field (CRF) model to recognize. Since the former basically follows the existing method[21] and therefore it is not specifically described. This article only focuses on the latter.

4.1. Identification of Legal Terminologies based on CRF Model

Conditional random field (CRF) model is an information extraction model based on statistic, which can be used for the identification of legal terminologies, and has incomparable advantages over other methods. The CRF model we have constructed to identify the legal terminologies is shown in Figure 1:

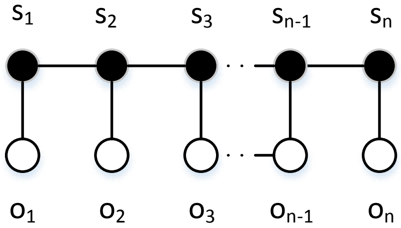


Figure 1. Graph Structure of Linear Chain CRF

The CRF model for the identification of legal terminologies can be viewed as a simple chain graph. In this model, $O=\{o_1,o_2,...,o_n\}$ means a text sequence to be recognized, such as the ordered words, phrases or terminologies that are contained in input statement. $S=\{s_1,s_2,...,s_n\}$ means a predicted state sequence which each state is associated with a observed values; When the observation sequence and the parameter $\Lambda=\{\lambda_1,\lambda_2,...,\lambda_k\}$ of this linear chain CRF are determined, the conditional probability of output sequence is shown in the formula (1):

$$P(S\mid O)=\frac{1}{Z_o}\exp(\sum_{n=1}^{N}\sum_{k=1}^{K}\lambda_k f_k(s_{n-1},s_n,o,n)) \tag{1}$$

In this formula (1), $f_k(S_{n-1},s_n,o,n)$ is sub characteristic function, on behalf of any one of all features. λ_k is a parameter after training corpus, and is a weight of K and the corresponding characteristic function; n is a length of observation data sequence; Z_o is a normalized function of all possible state sequences, its calculation formula is as follows:

$$Z_o=\sum_{s}(\sum_{n}\sum_{k}\lambda_k f_k(s_{n-1},s_n,o,n)) \tag{2}$$

By word segmentation processing, we can get this sequence f_1, f_2,…, f_n, which is the input sequence of CRF.

There are many features that can be used to identify the terminologies, such as part of speech, word boundaries, characters, Pinyin, the list of all kinds of name entities, grammatical dependency and sentence orientation, and so on. According to the composition characteristics of legal terminologies[17], we choose the following features:

Feature 1: Current Word (CurWord). In legal terminologies, the phrase terminology is about 91% of all legal terminologies. Therefore, this paper will focus on phrase terminology. But, because the phrase terminology is composed of a plurality of words, so we use the current word as an observation value feature by word-based approach, and compound phrases which are combined with the word can also be seen as a legal terminology.

Feature 2: Part of speech (POS). The feature of POS refers to the part of speech of current word. In legal terminologies, only by the composition of words. Part of speech can be regarded as an important feature of legal terminologies.

Feature 3: Word Boundaries. In this study, the "BMES" method is used to identify the location of a single word in the terminology, and the boundary of the terminology is identified. Among them, B can be indicated the first word of a multiple words, E is the last word, M is the middle word, and S is the single word.

Feature 4: The number of word in a terminology (WordsCount). The more number of word in legal terminology, the meaning of the expression is clearer, this feature is conductive to recognize the "longest" legal terminology.

Feature 5: Simple terminology (InDict). Due to a regular word formation method in legal terminology, a word can be extended to a number of legal terminologies which combine the "simple terminology" with the front and back words.

Feature 6: Affix of legal terminology (Affixed). In legal terminologies, there is often a kind of affix. Which forms legal terminologies with the front and back words, so it is necessary to recognize this feature as one of the features of the legal terminology.

Feature 7: Verb-object combination (VO). In terminologies of the structure of legal terminologies, the structure of the phrase type is mainly about the positive structure, which accounts for 90% of all the phrase terminologies. So, by analyzing the syntactic dependencies of the phrase terminologies, the verb-object combination can be used as one of the features.

In order to verify the above characteristics, this paper uses the cross experiment to verify, which can be seen on the experiment and analysis section.

4.2. Collocation Error Proofreading based on Markov Chain Model

Definition 1 The text element is divided into two categories, one is the legal terminologies and named entities after identification of legal terminologies and named entity recognition, the other is the rest of individual phases and words.

Based on statistical text proofreading method, a Markov chain model is constructed which use statistical properties of text, according to a time before and after a moment occurred state transition probability value, then to predict the input text content according to the maximum probability sequence. In this paper, we use the order 1 Markov chain model which can be shown in the formula (3).

$$\mathrm{p}\left(w_1, w_2, \ldots, w_n\right) = \prod_{i=1}^{n} p(w_i \mid w_{i-1}) \tag{3}$$

In this formula, $\mathrm{p}\left(w_1, w_2, \ldots, w_n\right)$ means the co-occurrence probability of text string $w_1, w_2, \ldots, w_n$, $p(w_i \mid w_{i-1})$ means the probability that the text element w_i appears before $w_{i\text{-}1}$. Take text string $\mathrm{wList} = \{w_1, w_2, \ldots, w_n\}$ for example, we can conclude that the occurrence probability of the current text element w_i is only related to the occurrence probability of the previous text element $w_{i\text{-}1}$ according to formula (3).

The collocation error proofreading algorithm based on Markov chain model is as follows:

Algorithm 3: Collocation error proofreading algorithm based on Markov chain model

Step1: Establish a Markov chain model by using the corpus of partial judgment documents, with all kinds of named entities, legal terminologies and independent words as unit;

Step2: Enter a sequence of pre proofreading text element, $\text{Sen} = \{s_1, s_2, \ldots, s_n\}$.

Step3: For a certain moment of the text element Si, please query the probability value about transferring the previous text element S_{i-1} to S_i. If the probability value is 0, then query a candidate elements of the second elements and the transition probability of the element with the previous element in this candidate elements. So, we can choose the maximum number of transfer probability as a candidate text element, while retaining the candidate elements which their transfer probability is of the previous ranking.

Step4: Compare the best sequence searched with original input string, then mark inconsistent content as wrong, and choose candidate elements as the result of error correction, exit.

From the above algorithm, under the transfer probability matrix constructed by the Markov chain model, the probability of the input text string is predicted. Then select one of the best path as output result by using Viterbi algorithm, and proofread according to the contents of the candidate text element.

5. Experiment

5.1. Experimental Data Sets

In this experiment, the data of the judgment document comes from the Chinese judgment document network[18] and the OpenLaw[19]. Due to the strict supervision of the legal institutions on the judgment document issued by the network, fewer errors can be found in the judgment document. In order to better simulate the real data, we select 200 cases of all kinds of civil cases from January 2014 to June 2014 in our country as training set and test set respectively. Among them, training set contains 4936 sentences, and test set contains 5211 sentences, in which randomly distributed 500 sentences with various errors. Relatively, the legal document knowledge base contains 223 parts, which are the civil laws and regulations of our country.

For the validation of method for identification of legal terminologies, we select altogether 300 types judgment by the first trial about civil case with a total of 628154 Chinese characters. What’s more, we divide the data into 5 groups to carry out the experiment and then take the average in order to reduce the recognition result due to the imbalance of the data.

5.2. Experimental Result Analysis

In order to verify the validity of identification method of legal terminologies proposed in this paper, we add the selected features to the feature set. The combination feature template is shown in Table 1.

Table 1. The Combination Feature Template for Legal Terminology Identification

Combination feature template	Features used
(1)	CurWord、POS、Boundary
(2)	CurWord、POS、Boundary、WordsCount
(3)	CurWord、POS、Boundary、WordsCount、InDict
(4)	CurWord、POS、Boundary、WordsCount、InDict、Affixed
(5)	CurWord、POS、Boundary、WordsCount、InDict、Affixed、VO

The observed experimental results, through using CRF++0.58 toolkit[20] to test above combination features, are shown in Figure 2.

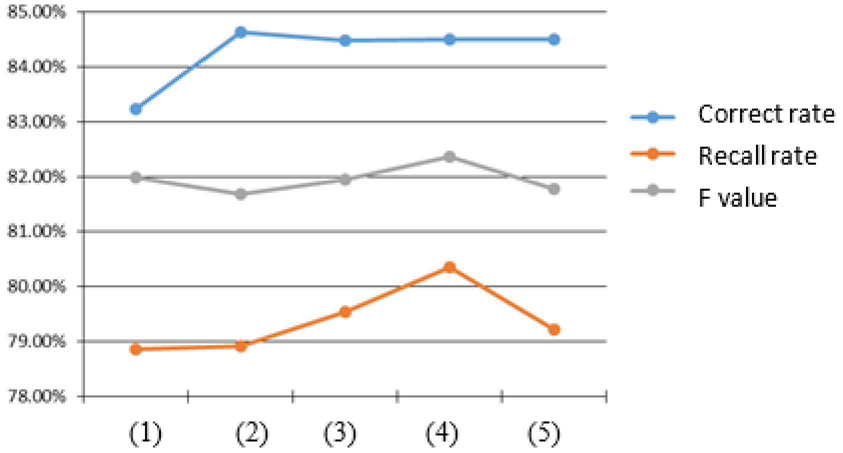


Figure 2. The Legal Terminologies Identification Results with Different Features

It can be seen from Figure 2, the use of the combination of features (4), which consist current word, part of speech, word boundaries and the number of words, will obtain the highest accuracy, about 84.63%. The correct rate was reduced after adding into dictionary, but the recall rate was improved. We suspect that there are some simple terminologies that are part of the terminology in some phrases, but not in others. After adding feature 6, the recall rate has increased, reaching 80.35%, while the correct rate is not significant, about 85.50%. In combination (5), correct rate has no significant change, the recall rate declined. This is because the feature of syntactic dependency is helpful to the recognition of the two word form, while it can be suppressed to the other. In contrast, feature combination (4) is most consistent with the purpose of maximizing the terminology.

The proofreading of legal document name and legal provisions of the legal documents use legal documents name and legal provisions proofreading algorithm, and reference sequence can also be collated. The results of the experiment are shown in Table 2.

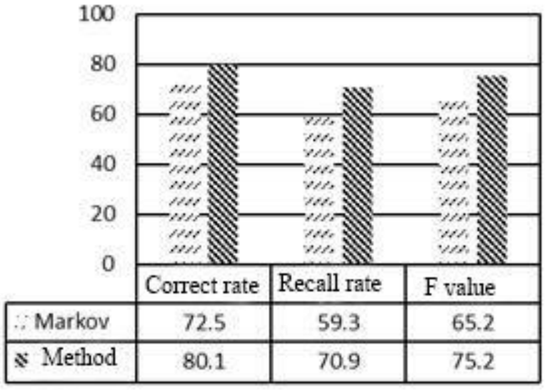
Table 2. The Proofreading Result of Legal Document Name and Legal Provisions

	Identifiable	Correct	Accuracy rate
Legal document name	42	39	93%
Legal document reference sequence	35	35	100%
Legal provision	33	31	94%
Legal provision reference sequence	24	22	100%

We can see from Table 2, there is a high accuracy of Legal documents reference sequence and Legal provision reference sequence. But, there are some mistakes in proofreading Legal document name and Legal provision, which is caused by some of laws that have been discontinued or have just begun to come into force, and the knowledge base of legal document has not been updated in time.

Finally, we separately use word segmentation results to construct the Markov chain model and apply this method mentioned in paper to construct Markov chain model, and then to proofread collocation error. Comparison experiment results shows in Figure 3.

From Figure 3, we can conclude that Markov chain model, although, can be used to correct collocation error after training judgment document, the overall accuracy rate, recall rate and F value are low, while the method mentioned in paper can obtain a good result in terminologies of accuracy rate, recall rate and F value.



	Correct rate	Recall rate	F value
⁘ Markov	72.5	59.3	65.2
▧ Method	80.1	70.9	75.2

Figure 3. Comparison Experiment Results in Markov Model and Method of this Paper

From Table 3, we can conclude the method mentioned in paper can reach a higher accuracy rate, but recall rate is lower. Reasons for this phenomena are as follows:

Table 3. Comprehensive Results of the Proofreading Method of this Paper

	Count	Error	Actual error	Correct	Recall rate	Accuracy rate	Error rate
Right	4711	48	0	0	—	—	0.01%
Wrong	500	341	500	341	68.2%	—	—
Total	5211	428	500	341	68.2%	79.7%	20.3%

(1) Due to the lack of error corpus, some collocation patterns are not in the model, which some collocation relations with a small probability of occurrence do not consider into this constructed model. So this model can’t cover all kinds of complicated language phenomenon.

(2) Although judgment document has certain regularity, there are always plenty of scattering mistakes that are hard to count. And lots of sentences have no problem, they

are just common sense error. It is very difficult to proofread this kind of error using computer.

6. Conclusion

If the general text proofreading technology is used to collate judgment document, the effect of proofreading is not satisfactory because of no strong pertinence. In order to solve this problem in judgment document proofreading, this paper fully considers the features of judgment document, and puts forward a specific proofreading method. In order to check legal documents and legal provisions reference error, we construct the knowledge base of legal document and reference rule base, and design a corresponding proofreading algorithm. What's more, it can effectively reduce the amount of checking calculation based on the Markov model and thus improve the efficiency and accuracy by maximizing the identification of legal terminologies and common named entities in judgment document. The experimental results show that the recall rate is 68.2% and the accuracy rate is 79.7%, which is based on the method proposed in this paper. It is proved that this method can be used to proofreading all kinds of error in judgment document, and it has a certain practicality.

Acknowledgements

The National Natural Science Foundation of China under Grant No. 61370139;

The Beijing Key Laboratory of Internet Culture and Digital Dissemination Research under Grant No. ICDD201506;

The Project of Construction of Innovative Teams and Teacher Career Development for Universities and Colleges Under Beijing Municipality No. IDHT20130519;

The Project of Beijing "3-Commnication & 2-Platform" standard demonstration project No. BJW201601.

References

[1] Kukich K., "Techniques for automatically correcting words in text [J]", ACM Computing Surveys (CSUR), vol. 24, no. 4, **(1992)**, pp. 377-439.

[2] Huang G., Huang Y. and Zhang Y., "A misspelling intelligent analysis approach for correcting misspelled words in English text [J]", Journal of Convergence Information Technology, vol. 5, no. 5, **(2010)**.

[3] L. Kong, "The Study on Automatic Checking Technology of Lexical Errors in Chinese [D]", Nation University of Defense Technology, **(2012)**.

[4] W. Siyu and S. Bo, "Construction of Automatic Proofreading System based on CSSCI [J]", Library Work in Colleges and Universities, vol. 34, no. 6, **(2014)**, pp. 50-54.

[5] L. Bailing, "Design and implementation of Chinese opinion text collation system based on statistics [D]", Heilongjiang University, **(2014)**.

[6] G. Jun, X. Wei and Z. Yang-sen, "Construction and application of semantic collocation knowledge base based on multiple knowledge bases [J]", Computer Engineering and Design, vol. 34, no. 6, **(2013)**, pp. 2136-2140.

[7] Z. Yangsen, T. Anjie and Z. Zewei, "Chinese Text Proofreading for Political News Field [J]", Journal of Chinese Information Processing, vol. 28, no. 6, **(2014)**, pp. 79-84.

[8] The Website of The Supreme People's Court [EB/OL]. [2015-04-25]. http://www.chinacourt.org/law/detail/2009/10/id/ 137892.shtml

[9] Civil Procedure Law of the People's Republic of China [EB/OL]. [2015-02-25].http://www.gov.cn/flfg/2012-09/01/ content_2214662.htm

[10] Civil Procedure Law of the People's Republic of China [EB/OL]. [2015-02-25].http://www.gov.cn/flfg/2012-03/17/ content_2094354.htm

[11] D. Hao, "Text automatic checking system based on Natural Language Processing [D]", University of Electronic Science and Technology of China, **(2006)**.

[12] W. Lin and Z. Yang-sen, "Reasoning Model of Multi-level Chinese Text Error-detecting Based on Knowledge Bases [J]", Computer Engineering, vol. 38, no. 20, **(2012)**.

[13] L. Rong, "A Chinese Spelling Check System for the OCR Output [J]", Journal of Chinese Information Processing, vol. 23, no. 5, **(2009)**, pp. 92-97.

[14] W. Zhaoxiang, "Understanding and Application of the Provisions of the Laws, Regulations and Other

Normative Legal Documents Concerning the Judgment Document [J]", People's Justice, vol. 52, no. 23, **(2009)**, pp. 29-33.

[15] China current normative legal document database [EB/OL]. [2015-02-26].http://www.chinalawindex.cn/

[16] X. Anqi, "An Algorithm of Computing String Similarity Based on Improved Levenshtein Distance [D]", Northeast Normal University, **(2013)**.

[17] C. Lu, "Analysis on the characteristics of Chinese legal terminologies and related problems [D]", China University of Political Science and Law, **(2010)**.

[18] Chinese judgment document network [EB/OL]. [2015-02-26].http:// http://www.court.gov.cn/

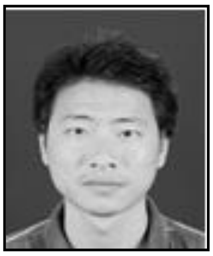
[19] OpenLaw [EB/OL]. [2015-03-14].http://openlaw.cn/

[20] CRF++0.54 Tools [EB/OL]. [2015-06-25]. http://sourceforge.net/projects/crfpp/.

[21] Z. Guanglei and X. Yabin, "A Mining Method on Hot-Topic-Oriented for Microblog View", Chinese Journal of Communications, vol. 35(z2), **(2014)**.

Authors

XU Yabin was born in 1962. He received the M.S. degree in Beijing from Beijing Jiaotong University in 1989. He is a professor at Beijing Information Science &Technology University. His research interests include Big data, Social network, Cloud computing, Future network, *etc*.



Ji Xuan was born in 1987. He received the M.S. degree in Beijing from Beijing Information Science &Technology university in 2016. His research interests include big data and text proofreading, *etc*.