Yang, X. & Liu, Y. (2018). Automatic extraction of theft judgment information in natural language. In *Proceedings of The 18th International Conference on Electronic Business* (pp. 134-142). ICEB, Guilin, China, December 2-6.

Automatic Extraction of Theft Judgment Information in Natural Language

(Full Paper)

Xi Yang, School of Computer Science and Information Engineering
Guangxi Normal University, Guilin, China,543506332@qq.com
Ying Liu*, School of Computer Science and Information Engineering
Guangxi Normal University, Guilin, China,1578694534@qq.com

ABSTRACT

Recently artificial intelligence technology replaces traditional manual methods in many fields. Especially the application of artificial intelligence in the legal field liberates legal people from tedious work. For example, crimes are automatically classified based on the facts of the crime, such as the crime name and sentence prediction. However, the premise of these applications is based on the establishment of case bases. Therefore, this paper studies the issue of automatic extraction of verdict information in natural language. Due to the verbal writing specification, we use regular expressions to construct extraction rule templates for template matching. At the same time, we also use natural language processing technology to extract the relevant semantic information accurately. For further similar case searching. The research focus of this paper is on the theft verdicts, and establish a database of records for the theft of the theft and the prediction of the theft of the theft.

Keywords: Natural language processing; knowledge acquisition; database.

*Corresponding author

INTRODUCTION

With the arrival of the Internet era, the data on the Internet has formed the age of big data, but most of these data are unstructured data. The current situation of data is huge, fast changing, various forms of existence, such as pictures, audio, and video, and many data are of poor quality. The dramatic increase of information quantity highlights the importance of information processing and application in many fields. In the field of justice, legal cases have constituted a fairly large set of cases, but many of them exist in paper documents or text form, from which that is not easy to search or find the information too messy, this makes it unnecessary difficult to automatically judge cases, retrieve similar cases and forecast sentences in the future. These case sets contain the evaluation standards and results of different cases by different legal institutions and lawyers, and contain rich legal knowledge. By extracting and sorting out the information related to the case, and artificial intelligence can be facilitated the retrieval of similar cases of legal institutions, and the recommendation of similar cases can better realise the value of previous cases. How to extract and sort out the massive and jumbled judgment information and build case base is the foundation of artificial intelligence.

To this end, this paper adopts the method of using regular expressions and judgment of part of speech based on natural language processing to extract information about the text features of the formal judgment of theft, and sorts the extracted information into the database to build case base for further use. So, the contribution of this paper is to establish the information database of judgment, so that information retrieval and application become more convenient and fast. Provide more convenient data support to similar case pushing and sentence prediction.

The structure of this paper is: Firstly we describes in detail the structure of the system and the principle of the technology used. The second section we present our prototype system and structure as follows illustrates the operational process and details of the system. Thirdly we explain the experimental results. Fourthly we discuss the content of this paper and finally, it is very important to point out the work that could done in future.

PRINCIPLE

The system structure

First, the original text is obtained from the web page or local file, and then the text is particized. Regular expressions are used to obtain the text and determine the type of words, and finally the corresponding fields are stored in the database.

The system structure is shown in Figure 1.

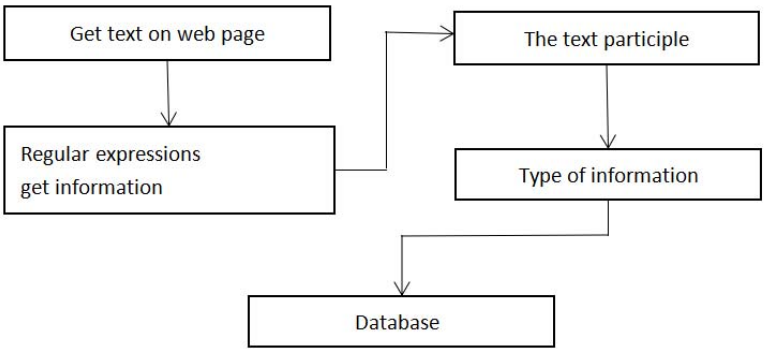


Figure 1: System structure diagram

1. Text segmentation: the first step of text processing is word segmentation. The original corpus is continuous sentences. Chinese word segmentation has its unique difficulty compared to English word segmentation. Most of Chinese word segmentation is based on word frequency statistics (Fei, Kang & Zhu, 2005). But there are many Chinese word packets in python, and this article USES python's third-party library, jieba. The original corpus was segmented by direct use of jieba (Chinese text segmentation: built to be the best Python Chinese word segmentation module.).

2. Judgment of information type: the corpus must be used when processing Chinese text, whether it is word segmentation, part-of-speech tagging or word classification. What is stored in the corpus is the language materials that appear in the actual use of the language. The corpus is the basic resource that carries the language knowledge with the computer as the carrier, while the real corpus needs to be analysed and processed to become a useful resource. Natural language processing is based on corpora. In this paper,WordNet (Fellbaum, 2012) is adopted. WordNet is a tree structure, and each node of it corresponds to a synonym set. The edge represents the upper or lower word relation, i.e., the superior or subordinate relation. The hierarchical fragment of WordNet is shown in figure 2.

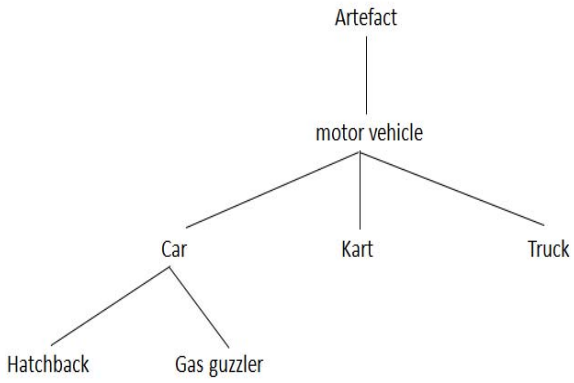


Figure 2: Wordnet hierarchy fragment diagramystem structure diagram

3. Regular expressions: regular expressions (Hu, Qin & Zhang, 2011) are a concept in computer science and are commonly used to retrieve and replace text that meets specified pattern rules. In this paper we use regular expressions to retrieve information that needs to be met, such as dates. The way to construct a regular expression is the same as the way to create a mathematical expression, with multiple meta characters combined with operators to constitute longer expressions. A pattern of text consi of characters in a regular expression. When retrieved with a regular expression template, the template is formed with one or more characters. The template is matched with the string searched, and all matched strings are recorded. In the pattern matching, it is divided into greedy matching or non-greedy matching, which should be handled carefully when searching.

Algorithm

The main algorithm steps of the system are as follows:

```
1：for i in file
2： flag=flag+file[i]
3：       fun1(flag)
4：p1=r” \d*year\d*month\d*day”
5：flag.jieba(cut)
6：flag.cutsy()
7：for i in flag
8：  flag[i].root_hypernyms()
9：  updata(flag[i])
10：end
```

First,It gets the text and stored is as file and store it into flag word by word. When reading the line feed, call the corresponding paragraph processing function and extract the information in the paragraph. (lines 1-3 of the algorithm) . For example, the processing function for the first paragraph, the regular expression gets the time axis, and then the word segmentation function is used to divide the paragraph text. and then filter the punctuation, (line 4-6 of the algorithm). For each word in flag, the function is used to find the upper word of the word, the type of the word is found by the upper word, and the corresponding field is stored in the database (lines 7-10 of the algorithm).

The detail function above only gives the first paragraph, and the following paragraph function is different from the details above.

Criminal judgment on theft

The criminal judgment is a written decision of the people's court on the conviction and sentencing of the defendant based on the facts and evidence ascertained in the end of the trial of the criminal case in accordance with the procedures provided for in the criminal procedure law (Zhang, 2009). After the conclusion of the trial of a criminal case, the people's court will, on the basis of the ascertained facts and the applicable law, make a legally binding judgment on the criminal act committed by the defendant. The judgment clearly state the basic information of the defendant, including the name, gender, age, domicile of origin, address, position, previous experience with criminal punishment, arrest and date of custody of the defendant, and so on. The situation of counsel and prosecutor; determine the facts, reasons and applicable legal basis of the judgment; the judgment and the duration of the appeal and the court of appeal. As soon as the criminal judgment takes legal effect, it has legal binding force.

In this paper we only discuss the verdict of larceny. The main factor in the measurement of theft is the value of stolen goods. Some judgments indicate clearly the total value of stolen goods before the judgment, but some do not. At this time, the part of the description of the crime be analysed by natural language processing technology, and the value of all cash and goods will be added and the total value will be obtained.

There are many cases in the case library, and the amount of larceny is changed with time, because the purchasing power of RMB keeps changing (Yuan, 2011). So time is also a factor in sentencing. The same provinces, nationalities, and so on. And the occupation and education also can influence, and knowing but still break the law to break the law is even more serious. There are also factors that can directly affect the sentence, such as recidivism, surrender, and adulthood. These factors can he used to the determinant the predict similar cases and sentences, but the weights are different, so these are also the information to be extracted.

Information type judgment

When we get the information of the defendant, we get the written information. We do not know whether this information is the educational background, occupation or native place of the defendant, how to judge which type of information is obtained, then we should use corpus or professional documents. The method of professional documents is to list all occupations, provinces, educational backgrounds and so on separately. When getting information, search the corresponding documents one by one. However this method is troublesome and inaccurate and prone to errors that cannot be found. This is the time to use the corpus method.

Wordnet is a tree corpus. The parent node is the big category of the child node. For example, the father node of apple is fruit, and the father node of banana is also fruit. When used, you can find the information word in the corpus and its parent node on the Internet. For example, if the information word is "doctor", the word on the parent node of the doctor can be found to be occupation.

In Python, there is a package called NLTK which can use wordnet directly, and you can simply call the function to get the contents of the parent node. However, wordnet is translated from English, and some words are too specialised to be one-to-one

corresponding. Therefore, we should combine the methods of professional documents and corpus, making up for each other to obtain accurate information classification.

The information type judgment of other literatures is realised by the method of part-of-speech annotation and automatic learning of machine learning, but the combination of natural language processing and professional documents in fixed fields can easily judge the information type.

Regular expression

Regular expressions are especially effective in obtaining information with fixed patterns, such as time, format must be yyyy year mm month and dd day. Regular expressions can be used to obtain all the information in the text that matches the template. Especially the main greedy match and the non-greedy match. Greedy matches match text messages as long as possible, so select the non-greedy pattern.

When obtaining the list of criminal facts, we can use time matching to obtain the list of criminal trends and time clues of the accused. According to the time, the criminal facts of the accused are listed one by one according to the time axis. When the judgment does not specify the total value of stolen properties, the value of stolen ones must be obtained and added to the list of criminal facts to obtain the total value of stolen properties.

Store the data

The storage of data is in MySQL database. The way to store data is to store pages and files. Python3.6 has the pymysql package, which enables the operation of the database. The data is stored in the web page. Django (Django is a high-level Python Web framework that encourages rapid development and clean, pragmatic design. Built by experienced developers, it takes care of much of the hassle of Web development, so you can focus on writing your app without needing to reinvent the wheel. It's free and open source.) is used to obtain the input data through the web page and save it to the local database.

Due to some of the information by calling the function, therefore cannot insert a to all the information, can only be inserted into the court, such as the number of first simple information, and to automatically generate information ID, the ID again passed as a parameter to the function, finally in function with updata instruction update assigned ID data, the function of information in the database. That is, one insert operation, multiple updata operations and one complete data storage.

Some data verdicts are missing, so some database fields may be empty. In the later sentencing prediction, the null field is likely to be divided into zero. If multiple null fields may affect the case reference value, the priority of this data is relatively low.

The data obtained are all strings, but some fields are converted into numeric types, such as the value of stolen properties, sentences, and so on. Converted into Numbers and stored in the database to facilitate the promotion of similar cases and prison term prediction in the future. After conversion into figures, it is convenient to calculate the similarity of average sentence and amount of theft.

Prototype system example

The original text of the judgment is first obtained from the web page, This is the original text of a simple criminal conviction for theft. This example happened in 2017, which is closer to the present. Therefore, this case is used as a simple example to illustrate the research of this paper.The example as Table 1.

The detailed implementation steps are as follows:

1. Get information from the web page; the original text in a file; retrieve the content of each segment in a loop; use different fun() functions to handle flag for different sections.

2. Call fun1(flag) to process the first paragraph. First, use the flag = jieba.cut(flag) text participle to obtain the segmented flag. The flag after the segmentation.

Defendant Du, male, December 8, 1993 born in Jingxing County, Hebei Province, Han, Mass, secondary school culture, jobless. On suspicion of theft, by the Zanhuang County Public Security Bureau decided, on November 18, 2016 by the Zanhuang County Public Security Bureau criminal detention; December 2016 2 was arrested.

Table 1: Judgment

Zanhuang County People's Court
Criminal judgements
(2017) Ji 0129 at the beginning of the sentence No. 6th
Public Prosecution agency Zanhuang County People's Procuratorate, Hebei province.
Defendant Du, male, December 8, 1993 born in Jingxing County, Hebei Province, Han, Mass, secondary school culture, jobless. On suspicion of theft, by the Zanhuang County Public Security Bureau decided, on November 18, 2016 by the Zanhuang County Public Security Bureau criminal detention; December 2016 2 was arrested.
The Zanhuang County People's Procuratorate of Hebei Province filed an indictment with the court on January 4, 2017 on charges of theft on behalf of the defendant, (2017)7, for pleading guilty to prosecution. The court opened the case on January 4, 2017 and applied summary procedure in accordance with the law, with a single trial and a public hearing of the cases. Zanhuang County People's Procuratorate assigned the prosecutor Zhang Limei appeared in court to support the public prosecution, the defendant du appeared in court to participate in the lawsuit, has now heard the end.
After the trial found that November 18, 2016 1:40, the defendant du in Zanhuang County film Internet Café when the theft of the victim Xu a golden oppor9plus mobile phone, after forensics, the price of 2610 yuan.
The above facts are confirmed by the following evidence in court and with proof of quality:
Surveillance video; On-site identification of photos; Zanhuang County Price Bureau Prices Certification Center price verification conclusion; Zanhuang County Public Security Bureau Criminal Police Brigade proof material; defendant du a household registration certificate; Victim Xu's statement; Defendant Du's confession.
The evidence listed above is sufficient to substantiate the fact that the public Prosecution service has accused Du of being guilty of theft.
The Court held that the defendant du for the purpose of illegal possession, the secret theft of other property, and the amount of large, his behavior has constituted the crime of theft, punishable by law, the Public Prosecution Service was charged. In view of the better guilty plea of the accused Du, a lighter penalty may be imposed at his discretion. In accordance with articles No. 264 and 64th of the Criminal Code of the People's Republic of China, the judgment is as follows:
First, the defendant du committed the crime of theft, sentenced to three months ' detention, and a fine of 2000 yuan. (Penalty is payable within 10th after the entry into force of this judgment)
The Court held that the defendant du for the purpose of illegal possession, the secret theft of other property, and the amount of large, his behavior has constituted the crime of theft, punishable by law, the Public Prosecution Service was charged. In view of the better guilty plea of the accused Du, a lighter penalty may be imposed at his discretion. In accordance with articles No. 264 and 64th of the Criminal Code of the People's Republic of China, the judgment is as follows:
First, the defendant du committed the crime of theft, sentenced to three months ' detention, and a fine of 2000 yuan. (Penalty is payable within 10th after the entry into force of this judgment)
(if the sentence is calculated from the date of enforcement of the sentence and is held in advance before the execution of the judgment, the day of detention is discounted from the date of the sentence, that is, from November 18, 2016 to February 17, 2017.))
Second, the defendant du a illegal income of 2610 yuan, in accordance with the law to recover.
If you are not satisfied with this judgment, you may, within 10th from the second day of receipt of the judgment, pass through the court or appeal directly to the Shijiazhuang Municipal Intermediate Court, and a written appeal shall be submitted with one original copy and four copies.
Judge Zhao Ruiying
January 17, 2017
Clerk Wang Zhiwei

3. Repeatedly take the words of flag, read "defendant", and save the next word to the variable name. When you read "male", our system can directly save the gender variable. When our system read "1993", the system can judge it as the date of birth. The next word "Yu" is "Hebei province", which can be found in the professional documents of place names, so it is decided as the address, and the word in front is "birth" and "yu", so it is decided as the native place. When "Han nationality" is read, it can be found in the "nationality" professional document, which exists in the nationality field. When you get the word "unemployed" and look for the word "unemployed" in a professional document, look for synonyms and the word rank in Wordnet. From this position, you know that "unemployed" is a profession, and you store it in the accused's occupation field. When words are not found in professional documents, and the upper word in Wordnet is not required information, so no processing is done.

4. In dealing with the paragraph of criminal facts, the defendant's criminal timeline list is first obtained by using the regular expression "\d* year \d* month \d* day". Each string in the list starts with \d* year \d* month \d* day, and each string represents a criminal fact. When the following paragraphs do not mention the total value, the value of the items in each of the criminal facts in the list is used to obtain the total value of the items. While obtaining the digital amount, and whether the amount of the item is double calculated. Only sum up the value of the non-duplicated stolen item and the cash.

5. The other paragraphs are the same as above. The amount of money obtained, sometimes written in Chinese, is treated as the price of the number type. The digital amount is also converted from character type to numeric type and stored in the database. The database entry is shown in Table 2.

We tested our system with 100 samples convictions for theft, and successfully extracted relevant information. These verdicts include a first offense, a second offense, a repeat offense, and the amount and length of the sentence vary from xx to xx. The database storage is shown in Figure 3.

Table 2: Database case field

The court	People's court of Zanhuang county
The letter of judgment	6 at the beginning of sentence ji 0129
Name of defendant	Du Mou
Nationality	Han
Profession	Unemployed
Native place	Hebei province
Crime or not	0
Criminal facts	After the trial, it was found that on November, 2016, at 1:40 am, when the defendant Du was surfing the Internet at the Ying Le Internet cafe in Zhenhuang county, he stole a gold OPPOR9PLUS mobile phone belonging to the victim Xu, which was verified to be worth 2610 Yuan.
Purpose	For the purpose of illegal possession
Surrender or not	No
Years in prison	0
Penalty	¥2000
Date of judgment	January 17, 2017
Type of judgment	Criminal judgment
The indictment	People's procuratorate of Zhanhuang county, Hebei province
Sex of defendant	man
Birth of the defendant	8 December 1993
Cultural level of the accused	Technical secondary school
Accusation	Larceny
Inspectors	Limei Zhang
Objections	
The total value	2610
Breach of law	Article 264 and 64 of the criminal law of the People's Republic of China
Months in prison	3
Judges	Judge Ruiying Zhao
The clerk	Clerk Zhiwei Wang

ID	fayuan	panjueshuf	panjueshuID	gongsujiguan	name	sex	minzu	birth	zhiye	wenhu	jiguan	crime_	cen	jianchayu	fanzuishishi	isyouyiy	mudi	zongpr	iszish	weifantia	panxin	panx	fajin	shenpanyua	panjuetime	shujiyuan
1	四川省稻城	刑事判决书	(2017) 川3337刑	稻城县人民检察院。	土登曲扎	男	藏族	1997	农民	小学文	四川省	盗窃罪	0	汪前鑫	公诉机关指控被告	无异议	以非法占有	16576	自首	《中华人	1	0	3000	审判员丁真珍	二〇一七年十月二	书记员安清琼
2	福建省惠安	刑事判决书	(2018) 闽0521刑	惠安县人民检察院。	张汉江(曾	男	汉族	1976	无业		福建	盗窃罪	1	陈华南	经审理查明：1、	无异议	以非法占有	6245	没有自	《中华人	0	9	3000	审判员翁添平	二〇一八年一月三	书记员李玲玲
3	北京市密云	刑事判决书	(2018) 京0118刑	北京市密云区人民检	郭某	男	汉族	1965		初中文		盗窃罪	1		公诉机关指控：被	有异议		1940	没有自	《中华人	0	6	1000	审判员单青林	二〇一八年二月十	书记员尹丹梅
4	福建省霞浦	刑事判决书	(2017) 闽0921刑	霞浦县人民检察院。	黄勇	男	汉族	1965	无业		重庆	盗窃罪	1	詹志华	霞浦县人民检察院	无异议	以非法占有	11470	没有自	《中华人	1	0	5000	审判员钟凌云	二〇一七年九月十	书记员汤晶晶
5	四川省叙永	刑事判决书	(2014) 叙永刑初	叙永县人民检察院。	张某某	男	汉族					盗窃罪	0		上述事实，被告人		以非法占有	2000	没有自	《中华人	0	8	4000	代理审判员彭	二〇一四年八月六	书记员曾丽
6	广东省广州	刑事判决书	(2016) 粤0106刑	广州市天河区人民检	罗德顺						湖南	盗窃罪	1	张广斌	广州市天河区人民	无异议	以非法占有	3060	没有自	《中华人	1	2	2000	审判长方洪江	二〇一六年十二月	书记员李丹丹
7	四川省西昌	刑事判决书	(2015) 西昌刑初	西昌市人民检察院。	邓某某	男	汉族	1994	农民	初中文		盗窃罪	0	王元、黎	经审理查明：201	无异议		12300	没有自	《中华人	0	8	10000	审判长何爱红	二0一五年二月十	书记员仁商
8	河南省偃师	刑事判决书	(2015) 偃刑一初	偃师市人民检察院。	任小娃	男	汉族		农民	小学文		盗窃罪	1	高强	偃师市人民检察院	无异议	以非法占有	1100	没有自	《中华人	0	6	2000	代理审判员彭	二〇一五年八月二	代理书记员周彦
9	赞皇县人民	刑事判决书	(2017) 冀0129刑	河北省赞皇县人民检	杜某	男	汉族	1993	无业	中专文	河北	盗窃罪	0	张丽梅	经审理查明，201		以非法占有	2610	没有自	《中华人	0	3	2000	审判员赵瑞强	二〇一七年一月十	书记员王智伟
10	山东省临沂	刑事判决书	(2015) 临兰刑初	临沂市兰山区人民检	王延东				无业			盗窃罪	1	朱丛丛	临沂市兰山区人民	无异议	以非法占有	21783	没有自	《中华人	1	3	15000	审判员董沂强	二〇一五年一月五	书记员徐守丽

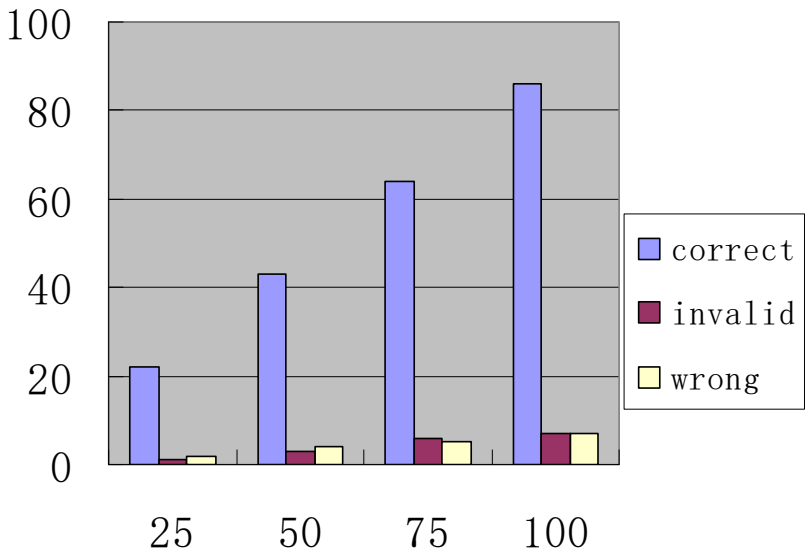
Figure 3: database memory map (in Chinese)

The experimental results

In this experiment, 100 cases of theft were randomly selected from the Chinese judicial network. The contents shown in Table 1 were extracted manually and stored in the database. The case is then copied into the web page one by one, and the application stores the case in database 2 for comparison one by one. Among the 100 samples tested, seven did not run properly because the judgment was not written in strict accordance with the specification. For example, the first four lines of a normal judgment are the court, the type of judgment, ID of judgment and the public prosecution authority. And the case that these cannot run normally is to write public prosecution organ inside the body.

In addition, 86 out of the 93 operational cases were the same as the manually extracted database. Most of the errors were on the theft amount. Many of the final statements do not indicate the total amount of the theft, which can only be found in the facts of the crime, and sum up the amount of the stolen goods according to the time clues obtained from the regular expression. However the description of criminal facts will be mixed with the number of non-theft items, such as the return of part of the money, distribution of stolen goods, and so on. These also need to be subdivided. If the cases that cannot run normally are valid cases, the correct rate is 86%. If the cases that cannot run normally are invalid cases, the correct rate is 92.5%. The distribution of each case is shown in Table 3.

Table 3: Case map



Data would be better obtained if all judgments were written in strict accordance with the standard of writing. Getting the total amount from the facts of the crime, but, requires more elaborate treatment. To distinguish the value of stolen properties, stolen cash, online money transfer, distribution of stolen goods, recovery and other involved amounts of various complex situations, the proper handling is to add data, subtract data, or do not operate data, to get an accurate theft amount.

In addition to the error of the amount, there is also the first time to make the mistake of drawing. Among the one hundred cases, there were two cases where the defendant was recidivist, but the procedure judged as the first offense. Because the description of the criminal record of the defendant was not written in the general position, but in the paragraph of the judgment result, the

criminal record of the defendant was not extracted, so it was judged as the first offense. Therefore, the judgment of "crime or not" should be based on the full text of the judgment.

RElATED WORK

Information extraction technology is applied in various fields. Bao, Huang, and Zhang (2018) studied information extraction in the medical field. In 2018, Luo and Huang (2018) extracted financial events from financial news. Sun and Guan (2004) and Sakamoto and Honiden (2018) studied information extraction methods. Neither has used information extraction techniques in the judicial domain. However, the judicial field has huge data, so it is very troublesome to apply them. The huge original judgment book library and the chaotic sorting are doomed to be useful in the judicial field.

Liu *et al.* (2018) abstracted the information of the accused in a broad way, and did not specifically extract the information of the value of stolen goods in the crime of punishment or theft, as well as the corresponding sentence and fine. The work of Chen (2011) is to extract the information and time and place of the accused, mainly by the characters, time and place of the event.

A study on the extraction method of judicial case information based on GATE (Song, 2016) is to extract the case text information to be used in business and save manpower. This paper is about the crime of theft category case push and prison sentence prediction. Jackson *et al.* (2003) analysed the case and found all relevant historical cases. In this paper, all historical cases are automatically extracted to the database for further inquiry, such as convenient case pushing and sentence prediction.

In the work of Xu and Xu (2018), computer-assisted sentencing has been mentioned and the premise of computer-assisted sentencing is to obtain enough and orderly historical case data are available. Bai has studied sentencing prediction based on the collective experience of judges (Bai, 2016). In fact, the judge's collective experience in sentencing can be added to the computer automatic recommendation of similar cases and sentencing prediction, and then the judge can use the experience to carry out fairer sentencing.

The text extracts information based on the sentencing. For the prediction of the sentence, the sentencing standard of the crime of theft is changing every year. Therefore, not only the value of articles can affect the sentence, time, surrender or not, first offense or recidivism, etc., but also the factors that can affect the sentencing. By getting as much information as possible, similar case push and sentence prediction can be more accurate.

CONCLUSION

In order to forecast the prison sentence of theft crime, this paper studies the problem of extraction automatically. This study is conducive to the efficiency of relevant personnel in sorting out and analyzing larceny cases, and builds a database for category case push and sentence prediction, providing data support.

In future, it is worth applying the information extraction technology to other types of crime cases, extract the key information of different types of crimes by classification, and finally establish a judgment information extraction system applicable to all types of crimes.For example, in cases involving drugs, the amount of drugs sold, transported and produced by the accused shall be extracted; regarding cases involving intentional injury, the degree of injury to the victim should be obtained; and cases involving bribery, the amount of bribery shall be obtained. These are the main factors that determine sentencing.Since it is only information extraction for specific fields and aspects, it does not employ any machine learning method, but natural language processing method. The advantages of natural language processing methods are high accuracy in specific areas, but the disadvantages are poor portability and require a large amount of professional documentation. These can be addressed in future.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China No. 61762016

REFERENCES

[1] Fei, X., Kang, S., & Zhu, X. (2005). Chinese word segmentation research based on statistic the frequency of the word (in Chinese). Computer Engineering and Applications. 41(7), 67-68.
[2] Fellbaum, C. (2012). WordNet. The Encyclopedia of Applied Linguistics. Blackwell Publishing Ltd.
[3] Hu, J., Qin, Y., & Zhang, W. (2011). Regular expression and its application to web information extraction (in Chinese). Journal of Beijing University of Information Technology (natural science edition), 26(6), 86-89.
[4] Zhang, Q. (2009). Analysis of speech act in a court judgment (in Chinese). Tribune of Political Science and Law, 27(3), 144-149.
[5] Yuan, X. (2011). Study on sentencing of larceny (in Chinese). (Doctoral dissertation, China University of Political Science and Law).
[6] Bao, X., Huang, W., & Zhang, K. (2018). A customized method for information extraction from unstructured text data in the electronic medical records (in Chinese). Journal of Peking University (medical edition), 50(2), 256-263.

[7] Luo, M., & Huang, H. (2018). Information extraction method of financial event based on lexical-semantic patterns (in Chinese). Computer Applications, 38(1), 84-90.
[8] Sun, C., & Guan, Y. (2004). A statistical approach for content extraction from web page (in Chinese). Chinese Journal of Information Science, 18(5), 17-22.
[9] Sakamoto, K., & Honiden, S. (2018). Information Extraction Apparatus. Information Extraction Method, and Information Extraction Program. US20180018378.
[10] Liu, W., Wang, J., Li, Y., You, J., & Chen, J. (2018). Design and realization of the key information extraction system of the court judgment (in Chinese). Journal of Hubei University of Technology (1), 63-67.
[11] Chen, H. (2011). Research on text information extraction of criminal cases (in Chinese). (Dissertation, Nanjing Normal University).
[12] Song, C. (2016). Research on the Method of Information Extraction Based on GATE (in Chinese). (Dissertation, Tianjin University).
[13] Jackson, P., Al-Kofahi, K., Tyrrell, A., & Vachher, A. (2003). Information extraction from case law and retrieval of prior cases. Artificial Intelligence, 150(1-2), 239-290.
[14] Xu, W., & Xu, W. (2018). The idea of a computer-aided sentencing system (in Chinese). Computer Knowledge and Technology, 14(6), 239-240, 242.
[15] Bai, J. (2016). Sentencing prediction research based on judges' collective experience (in Chinese). Chinese Journal of Law (6), 140-154.