Legal Knowledge and Information Systems
E. Francesconi et al. (Eds.)

doi:10.3233/FAIA220468

Functional Classification of Statements of Chinese Judgment Documents of Civil Cases

Chao-Lin Liu[†] Hong-Ren Lin[‡] Wei-Zhi Liu[⁋] Chieh Yang
National Chengchi University, Taiwan
{chaolin[†],109753156[‡],109753157[⁋]}@nccu.edu.tw

Abstract. Enabling the inference systems for assisting legal decisions to identify the functions of sentences and paragraphs in documents of legal judgments can enhance the justifiability of their algorithmic recommendations. The information about the functions of larger linguistic constituents complements the information at the word level like NER, and provides more clues about the arguments for the legal decisions. We explore this venue for the civil cases, which is a relatively uncommon choice in legal informatics and more challenging than working on the criminal cases. Current experimental results are promising.

Keywords: civil cases, functional classification, semantic classification, machine learning, justification

1 Introduction

There is a long history of the applications of artificial intelligence (AI) to the legal domain, and the first AI and law conference was held in 1987. In contrast, the field has attracted a large-scale attention of the Chinese community only until recently (e.g., [7] [13]). In this short manuscript, we are unable to review the field, but to focus on our research topic.

Offering both the recommendations and supportive justifications for the legal queries is important for people to accept the algorithmic recommendations. Take the task of the legal judgment prediction (LJP) for example. The LJP task aims at predicting the outcomes of lawsuits (e.g., [4][8]). The majority of the current work is about criminal cases, and the goals are to predict the penalties against the defendants. So far, work on civil cases like [6] is relatively uncommon. When addressing the applications of machine learning techniques to the prediction tasks, Mumford et al. [9] commented that "*without an explanation of why the case was so classified, the adjudicator has no reason to follow*."

Evidences also show that enabling computers to read and understand some details in judgment documents appears to be necessary for improving the quality of legal judgment predictions [3]. It may be relatively easy to determine the types of charges and the citing statutes for criminal cases, but the quality of the predictions for the penalties remains to be expected [4].

Hence, techniques for strengthening computers' competence to understand and explain details in judgment documents are needed for both the acceptance and the quality of the prediction systems [1].

Detailed information can refer to such word level information as named entities [3]. In our work, we focus on the functions of the sentences in legal documents. The concept is similar and related to the research that hopes to find the legal arguments in judgement documents ([2][11][12]).

Information about the functions of sentences in judgment documents complements the information of the word or phrase levels, and may be useful for explaining the algorithmic recommendations of AI systems.

2 Problem Definition

The judgment documents that were published by the Judicial Yuan in Taiwan do not follow a very strict format, although the documents typically contain some common sections. Some sections provide the meta data of a case, such as the court name, the time of the judgement, and the summary of the judgments of the case. The central part of a judgment document is the section that records the facts and the information about the reasoning for reaching the judgments, e.g., the criminal activities for criminal cases, the conflicts in interests for the involving parties for civil cases, the opinions of the court, and various considerations for reaching judgments (subsuming).

We are working on the alimony problems, which is a specific type of the civil cases. Our first goal is to classify the sentences of the central part of a judgment document into **four categories** of functions, namely, the pleadings of the applicants, the responses of the opposite parties, the opinions of the court, and the subsuming part. We denote these four categories as **C1**, **C2**, **C3**, and **C4**, respectively. These are the most central factors for judging civil cases.

We will extend our classifiers to categorize the sentences into **five categories**, and add the category of "conflicts", denoted by **C5**, to indicate the disagreeing items between the applicants and the opposite parties.

Since sentences of C5 are about the conflicts of the two sides of the disputes, it was easy to confuse sentences of C1 and C2 with C5. In addition, the courts often mention the conflicts in reaching decisions, so it may not be easy to tell specific sentences of C4 and C5 apart.

3 Data Source and Preprocessing

The main source of data for this research is the open data websites of the Judicial Yuan of Taiwan. The websites provide judgment documents of three levels of courts, i.e., the district, high, and supreme courts, and some special types of courts, e.g., for business and intellectual issues, for juvenile and family issues, and for administrative issues. Due to certain privacy and protection considerations, not all of the judgements of the courts are released, and few of the previously published documents may be retracted due to a wide variety of legal reasons. Therefore, the total number of published documents is stable only within a reasonable range.

We may access the data via the batch site or the interactive site.[1] For computational efficiency, we downloaded the documents from the batch site and will refer to it as TWJY, henceforth. TWJY updates the corpus monthly. At the time of this writing, we can find about 17.6 million documents for all cases that were judged since January 1996 in TWJY.

Although the total number of published documents in TWJY is huge, these public documents belong to a myriad of types of lawsuits. The number of cases that is related to a specific category may not be large, and we have to search the TWJY data to identify the documents that are at least seemingly relevant to our research needs. We extracted 6679 cases of which the "cause of judgment" (裁判案由) were for "the issues of alimony"

[1] https://opendata.judicial.gov.tw/ and https://law.judicial.gov.tw/FJUD/default_AD.aspx

(給付扶養費). These cases were judged between January 2000 and December 2021, and Figure 1 depicts the trends of the annual numbers of the extracted cases.

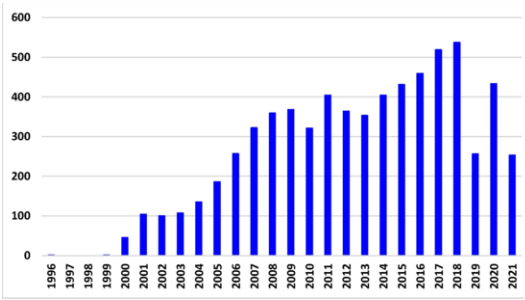

Figure 1. Annual number of cases.

We further filter the files for more detailed reasons. We counted the number of main paragraphs in the judgment documents of these 6679 cases, and Table 1 offers the counts of the number of main paragraphs in the documents. We inspected the paragraphs, and found that the cases with only one, two, or three main paragraphs are related to relatively simple and stereotypical problems, and are not relevant to our current studies. The cases with only three main paragraphs are commonly related to specific types of legal issues, such as that the applicants did not pay for the legal fees or that the litigants requested to transfer the cases to other courts. There were 11 cases which had more than 10 main paragraphs. We did not use these rare cases in our experiments.

Par.	Counts	Par.	Counts
1	2021	6	452
2	500	7	226
3	1541	8	106
4	1073	9	40
5	677	10	32

Table 1. Dist. of the number of paragraphs.

We use the cases with three or four main paragraphs in most of our experiments, and will use cases with six paragraphs also. When we examined the contents of the judgment documents more closely, we could find cases that are not appropriate for our study due to the main issues of the cases, so we could use only 820 cases. This small number of cases can be surprising and even frustrating, particularly when one compares this number with the size of TWJY. However, this is a fact that one can verify with professors of law [5][14].

We offer two possible reasons for this relative few number of cases. The first is that, in Taiwan and in Asian cultures, resolving family problems in courts can be a shameful problem for families, so, unless the problems are really not resolvable in private, the number of lawsuits for issues of alimony is suppressed for cultural reasons. The second is that, even after the cases have been submitted to the courts, the legal procedure encourages family members to resolve their problems privately via the mediation process that is assisted by the courts. This face-saving alternative can contribute to the reduction of the number of cases that have to be judged by the courts.

The definition of "sentence" in Chinese is much vaguer than that in English, although there is a symbol for sentence period in Chinese punctuation. In short, we segmented texts by the comma ("，"), the period ("。"), the semicolon ("；"), and the quotation marks. From these cases, we had 114204 sentences.

4 Labeling the Data

We can label the categories of our data in two different ways. The most common way is to ask domain professionals to categorize the individual statements. We refer to this dataset as **DE**, where "D" and "E" denote "**D**omain" and "**E**xpert", respectively. We used 80% and 20% of DE for training and testing, respectively. We refer to these subsets as **DEA** and **DEE**, respectively, where the ending "A" and "E" are for "tr**A**in" and "t**E**st", respectively.

The main annotator owns a college degree in law, and we have only limited data for inter-rater agreement at this stage.

An intriguing alternative is to take advantage of the regularities of the writing styles in the judgment documents. It is possible to find common patterns of collocations that are indicative of the high-level functions of the paragraphs. Hence, we relied on specific keywords, phrases, and collocations to algorithmically label the paragraphs to bootstrap our classification tasks, and treated all sentences in a labeled paragraph to belong to the category of the paragraph.

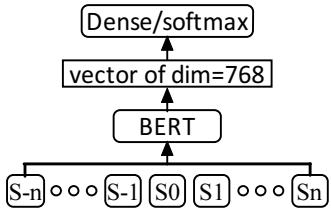

Figure 2. An intuitive classifier with BERT.

Paragraphs for categories C1 and C2 usually begin with "**聲請人**" and "**相對人**", respectively. It is very common for the courts to use "按" in the beginning statements in paragraphs of category C3. The collocations that follow the regular expression "**(本|法)院 .* (判斷|心證|經查|據)**" strongly suggests paragraphs of category C4. We refer to this algorithmically labeled dataset as **KP**, where "K" and "P" denote "**K**eywords" and "**P**atterns", respectively. Again, we used 80% and 20% of KP for training and testing, respectively. We will refer to these subsets of data as **KPA** and **KPE**, respectively, for analogous naming principle.

5 The Classifiers and Settings

In this section, we report results of the applications of machine learning methods to legal informatics. We applied established tools, including those offered by scikit-learn[2] and TensorFlow[3]. More specifically, we employed the tools for TFIDF vectorizer, logistic regression, and support vector machines of scikit-learn. We used the pretrained BERT for Chinese, including its tuning, that was demonstrated in the interface in TensorFlow Hub. Unless stated explicitly, we used the default parameters for the tools in the experiments reported in this short manuscript.

We have explored many different network structures for this study, including some intuitive ones and more academic options (e.g., [10]). We chose to report the results of the intuitive network that is shown in Figure 2. When training this model, we used the class label of a target sentence, S0, for the context of S0, where the context contains ***n*** sentences on both sides of S0. We made the BERT itself trainable, set the batch size to 64, and set the patience to stop training to five. In this report, we set *n* to be 1, 2, 3, 4, and 5.

In the experiments, 80% of the training data were used for validation. Hence, we had more than 70000, 18000, and 23000 instances for training, validation, and testing, respectively, but the exact numbers of the instances may vary slightly due to the settings of the environments.

6 Sentence Classification with DE

In this section, we report results of using **DEA** and **DEE** for training and testing, respective.

[2] https://scikit-learn.org/stable/

[3] https://tfhub.dev/tensorflow/bert_zh_L-12_H-768_A-12/4

6.1 *Classification of Four Categories*

Table 2 shows the confusion matrix of a DEA-DEE experiment for a four-category classification, and n was 4. The middle four rows of Table 3 list the F_1 values for these four categories and different values of n, while the last row lists the values of the macro F_1 for different values of n.

In Table 2, we can see that it was relatively difficult to distinguish sentences of C1 and C2 because their similarity in nature. Analogously, it was not easy to differentiate sentence of C3 and C4. One might have expected that, the quality of classification might improve as we expend the widths of the contexts, and this is supported partially by the statistics in Table 3. However, as we enlarged the contexts, the classifiers might be confused by some misleading statements in far-away contexts.

	categorized			
actual	C1	C2	C3	C4
C1	3402	198	7	623
C2	460	1667	56	838
C3	6	3	4330	298
C4	322	252	369	10962

Table 2. DEA-DEE, four categories and $n = 4$.

n	1	2	3	4	5
C1	0.717	0.785	0.760	0.808	0.747
C2	0.548	0.638	0.612	0.649	0.569
C3	0.917	0.910	0.897	0.921	0.897
C4	0.858	0.879	0.872	0.890	0.881
MF	0.760	0.803	0.785	0.817	0.773

Table 3. DEA-DEE, F_1 and macro F_1

	categorized				
actual	C1	C2	C3	C4	C5
C1	3673	351	7	189	55
C2	675	1403	38	322	102
C3	6	1	4482	103	46
C4	704	594	453	6635	1022
C5	196	23	20	491	1686

Table 4. DEA-DEE, **five** categories and $n = 4$.

6.2 *Classification of Five Categories*

Table 4 shows the confusion matrix of a DEA-DEE experiment for five-category classification, and n was 4. The statistics supported our explanation and expectation which we stated at the end of Section 2. The F_1 values of C1, C2, and C4 dropped to 0.771, 0.571, and 0.774, respectively; all are smaller than their counterparts in Table 3. The F_1 value of C3 was 0.930 and relatively stable. The F_1 value of C5 was only 0.633.

7 Training the Classifiers with KPA

In this section, we report results of using KPA to train the classifiers, and test the classifiers with DEE and KPE. Recall that the dataset KP was labeled by heuristic rules, and we did not have rules for C5, so the experiments reported in this section can involve only four categories.

7.1 *Testing with DEE*

Table 5 shows the confusion matrix of a KPA-DEE experiment for a four-category classification, and n was 4. The middle four rows of Table 6 list the F_1 values for these four categories and different values of n, while the last row lists the values of the macro F_1 for different values of n.

The statistics in Table 5 indicate similar trends as those indicated by Table 2. It was relatively difficult to distinguish sentences of C1 and C2, and it was not easy to differentiate sentence of C3 and C4. Statistics in Table 6 also support that expanding the widths of contexts benefited the F_1 to an extent.

It is interesting to compare the statistics in Tables 3 and 6. Training the classifiers with the DEA, which is annotated by domain professionals, offered better classification results than training the classifiers with heuristically labeled data across the board. The average gain in macro F_1 is above 0.05.

	categorized			
actual	C1	C2	C3	C4
C1	3221	228	47	734
C2	574	1633	69	745
C3	4	57	4150	426
C4	276	433	1070	10126

Table 5. KPA-DEE, four categories and $n = 4$.

n	1	2	3	4	5
C1	0.705	0.703	0.688	0.776	0.700
C2	0.538	0.578	0.383	0.608	0.624
C3	0.834	0.810	0.797	0.832	0.802
C4	0.827	0.816	0.775	0.846	0.752
MF	0.726	0.727	0.661	0.765	0.719

Table 6. KPA-DEE, F_1 and macro F_1.

n	1	2	3	4	5
DEA	0.675	0.710	0.697	0.720	0.679
KPA	0.675	0.697	0.621	0.721	0.662

Table 7. Testing with KPE, macro F_1.

7.2 *Testing with KPE*

Table 7 shows only the values of the macro F_1 when we tested two classifiers that were trained by DEA and by KPA and test both classifiers with KPE. Since we already had human annotated data, this comparison is only of theoretical interest. Even if the KPA and the KPE datasets were labeled with the same principles, the statistics in Table 7 indicate that using DEA to train can still help us achieve slightly better qualities than when we train the classifiers with KPA.

8 Concluding Remarks

Justifications are required for algorithmic recommendations for legal decisions. We compared effects of different ways to annotate the raw data, and evaluated a few classification models for categorizing the legal functions of the sentences and paragraphs in judgment documents of the civil cases, and the current results are promising. We have conducted more experiments, and can report their results during the Conference.

Acknowledgements

This research was supported in part by the project 110-2221-E-004-008-MY3 of the National Science and Technology Council of Taiwan. We are very thankful to the reviewers' comments.

References

[1] Atkinson K, Bench-Capon T, and Bollegala D. Explanation in AI and law: Past, present and future. *Artificial Intelligence*. 2020; 289:103387.

[2] Aumiller D, Almasian S, Lackner S, and Gertz M. Structural text segmentation of legal documents. *Proc. of the 18th Int'l Conf. on Artificial Intelligence and Law*. 2021; p. 2–11.

[3] Chen Y, Sun Y, Yang Z, and Lin H. Join entity and relation extraction for legal documents with legal feature enhancement. *Proc. of the 28th Int'l Conf. on Computational Linguistics*. 2020; p. 1561–1571.

[4] Feng Y, Li C, and Ng V. Legal judgment prediction via event extraction with constraints. *Proc. of the 60th Annual Meeting of the Association for Computational Linguistics (vol. 1)*. 2022. p. 648–664.

[5] Huang S-C. An empirical study of the legal judgments of alimony for elderly parents in Taiwan. presented in the Conf. on Technologies for Law and Access to Justice. 25 April 2022.

[6] Huang S-C, Shao H-L, and Leflar RB. Applying decision tree analysis to family court decisions: factors determining child custody in Taiwan. *Proc. of the 18th Int'l Conf. on Artificial Intelligence and Law*. 2021. p. 258–259.
[7] Liu C-L, Chang C-T, and Ho J-H. Classification and clustering for case-based criminal summary judgments, *Proc. of the Ninth Int'l Conf. on Artificial Intelligence and Law*. 2003; p. 252–261.
[8] Long S, Tu C, Liu Z, and Sun M. Automatic judgment prediction via legal reading comprehension. *Proc. of the 2019 China National Conf. on Chinese Computational Linguistics*. 2019; p. 558–572.
[9] Mumford J, Atkinson K, and Bench-Capon T. Machine learning and legal argument. *Proc. of the 21st Workshop on Computational Models of Natural Argument*. 2021; p. 47–56.
[10] Rao G, Huang W, Feng Z, and Cong Q. LSTM with sentence representations for document-level sentiment classification. *Neurocomputing*. 2018; 308:49–57.
[11] Stab C and Gurevych I. Parsing argumentation structures in persuasive essays. *Computational Linguistics*. 2017; 43(3):619–659.
[12] Wyner A, Mochales-Palau R, Moens M-F, and Milward D. Approaches to text mining arguments from legal cases. In *Semantic Processing of Legal Texts*. 2010; p. 60–79.
[13] Xiao C, Zhong H, Guo Z, Tu C, Liu Z, Sun M, Feng Y, Han X, Hu Z, Wang H, and Xu J. CAIL2018: A large-scale legal dataset for judgment prediction. 2018; arXiv:1807.02478.
[14] Xu H, Savelka J, and Ashley KD. Accounting for sentence position and legal domain sentence embedding in learning to classify case sentences. *Proc. of the 34th Int'l Conf. on Legal Knowledge and Information Systems*. 2021; p. 33–42.