FIRE2019@AILA: Legal Information Retrieval Using Improved BM25

Zicheng Zhao², Hui Ning², Liang Liu¹, Chengzhe Huang¹, Leilei Kong^{1,*}, Yong Han¹, and Zhongyuan Han¹

> ¹ Heilongjiang Institute of Technology. Harbin, China ² Harbin Engineering University, Harbin, China

> > kongleilei1979@gmail.com

Abstract. This paper details the approaches of implementing the tasks of identifying relevant precedents and identifying relevant statues in the evaluation of Artificial Intelligence for Legal Assistance proposed by Forum of Information Retrieval Evaluation in 2019(AILA@Fire2019). We formalize the two tasks as the issue of information retrieval, and present the improved BM25 models to retrieve the prior cases and identify the relevant statues. For the task of identifying relevant precedents, the proposed improved BM25 model integrates the relevance scores of the original current case and the filtered current case. For the task of identifying relevant statues, the proposed improved BM25 models exploit the search results as the reference documents of the current case and integrate the ranking information of search results into the BM25 model. Comparisons to the other submissions for the same tasks, our improved BM25 model achieves the top performers for the task of identifying relevant precedents on all evaluation measures. For the task of identifying relevant statues, the improved BM25 model wins the second place on 1/rank of first relevant document and the third place on BPREF.

Keywords: Legal Information Retrieval, Prior Case Identifying, Statues Identifying, BM25.

1 Introduction

In a Common Law System¹, great importance is given to prior cases. A prior case (also called a precedent) is an older court case related to the current case, which discusses similar issues and which can be used as reference in the current case [1]. A prior case is treated as important as any law written in the law book called statutes. This is to ensure that a similar situation is treated similarly in every case. If any relevant legal

^{*} Corresponding author

Copyright 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0). FIRE 2019, 12-15 December 2019, Kolkata, India. ¹ https://en.wikipedia.org/wiki/Common_law/

issues have been decided in the ongoing case, the court should follow the interpretation in the previous case. For this purpose, it is critical for legal practitioners to find and study previous court cases, to examine how the ongoing issues were interpreted in the older cases [2].

With the recent developments in information technology, the number of digitally available legal documents has rapidly increased. It is, hence, imperative for legal practitioners to have an automatic precedent retrieval system. The task of identifying relevant prior case can be modeled as a task of information retrieval, where the current case document (or a description of the current situation) will be used as the query, and the system should return relevant prior cases as results [2]. Additionally, identifying the most relevant statutes is to identify the most relevant statutes for each query.

Usually, legal document retrieval considered as a rank task. Early approaches for handling term dependencies in IR considered extensions of the bag of word representation of texts, by including bi-grams to the vocabulary. Such an approach was taken by Fagan [3] for vector space models, while the language model counterpart was proposed in the late of 90s [4, 5, 6] where the authors proposed to use a mixture of the bigram and unigram language models. Multiple weighted fields base on BM25 were proposed by Robertson [7].

2 Method of Identifying Relevant Precedents

For the task of Identifying Relevant Precedents, we apply the approach based on the information retrieval to obtain the prior cases of the current case.

Given a current case, denoted as q, and a collection of prior cases denoted as D, the goal of Identifying Relevant Precedents is to retrieve the relevant document d in collection D when given the current case q.

For the submission HLJIT2019-AILA_task1_1, we choose the BM25 model as the retrieval model, defined in Eq. (1)

$$rel(q,d) = BM25(q,d) = \sum_{w_i \in q} \log(\frac{tf(w_i,d) \cdot idf(w_i) \cdot (k_1+1)}{tf(w_i,d) + k_1 \cdot (1-b+b \cdot \frac{LEN(d)}{avdl})})$$
(1)

where w_i is the word in q, avdl is the average length of the document, k_1 and b are the parameters of BM25.

In the evaluation, the q and D is preprocessed firstly as follows: Porter stemmer [8] is used for stemming and the stop words, the punctuation characters, and the numbers are filtered by using the Lucene toolkit². Especially, for the current case q, we rank the words in q according to their IDF (Inverted Document Frequency) scores (the collection of prior cases D is chosen to compute the IDF). Then the top m% words are chosen to represent the query q. We set the parameter m = 50, $k_1 = 1.2$, and b = 0.75.

For the submission HLJIT2019-AILA_task1_2, we modify the relevance computation to get an improved BM25 as follows:

² http://lucene.apache.org/

$$rel(q,d) = BM25(q',d) + BM25(q'',d)$$
 (2)

where q' is the same as HLJIT2019-AILA_task1_1, and q'' is the original current case without IDF filtering. All the other settings are followed HLJIT2019-AILA_task1_1.

In addition, we also experimented with Word2vec method as the submission HLJIT2019-AILA_task1_3. After choosing the top m% words with high TF-IDF to represent the query q and document d, we represent q and d as vectors using Word2vec method, shown in Eq.(3):

$$V(x) = \frac{1}{n} \sum_{i=1}^{n} t_i$$
(3)

where t_i denotes word vector of *i*-th term, n is the top m% words. In experiments, we set the parameter = 50.

Then, we use the Euclidean distance to calculate the similarity between q and d.

3 Method of Identifying Relevant Statues

For the task of Identifying Relevant Statues, we apply the approach based on the information retrieval to obtain the relevant statues.

Given a current case q, and a collection of statues, denoted as S, the goal of Identifying Relevant Statues is to retrieve the relevant statues s in collection S when given the current case q.

For the submission HLJIT2019-AILA_task2_1, we also choose the BM25 model as the retrieval model. We use the description part in statues to construct the document collection S. And all the parameters setting and pre-processing are followed HLJIT2019-AILA_task1_1.

For the submission HLJIT2019-AILA_task2_2, we use the top-n relevant prior cases associated with the current case p as the reference document p_i , and integrate the relevant scores obtained by p_i into the original relevant score, shown as Eq.(4):

$$rel(q,s_j) = BM25(q,s_j) + \frac{BM25(p_i,s_j)}{rank(p,p_i)}$$
(4)

where s_j denotes the *j*-th statue in S, p_i is the *i*-th relevant prior case associated with p, and $rank(q, p_i)$ is the rank of p_i in the search results of p. In the evaluation, we select the top-10 search results as the reference documents.

Furthermore, considering the ranking information of relevant statues, we modify the Eq.(4) as follows

$$rel(q,s_j) = BM 25(q,s_j) + \frac{BM 25(p_i,s_j)}{rank(p,p_i) \cdot rank(q,s_j)}$$
(5)

where $rank(q, s_i)$ is the rank of s_i in the search results of p. In the evaluation, the number of reference documents is 10, and the number of re-ranking relevant statues is 197.

4 Results

Table 1 shows the experimental results of the task of precedent retrieval.

Team name	Run ID	P@10	MAP	BPREF	1 / rank
HLJIT2019-AILA	HLJIT2019-AILA_task1_2	0.0700	0.1492	0.1286	0.288
JiamingGao-HGC	HGC_1	0.0575	0.1382	0.1207	0.28
HLJIT2019-AILA	HLJIT2019-AILA_task1_1	0.0600	0.1335	0.1134	0.282
JiamingGao-HGC	HGC_2	0.0500	0.1263	0.1092	0.256
BabanGain-IITP	IITP_BM25_case	0.0275	0.0984	0.0869	0.175
TRDDCPune	TFIDF	0.0500	0.0956	0.0670	0.203
JiamingGao-HGC	HGC_3	0.0316	0.0946	0.0804	0.180
TRDDCPune	ENSEMBLE	0.0400	0.0817	0.0591	0.162
TRDDCPune	BM25	0.0375	0.0773	0.0547	0.151
YunqiuShao-thuir_legal	thuir_legal_3	0.0425	0.0689	0.0434	0.121
BabanGain-IITP	IITP_Doc2Vec_case	0.0175	0.0677	0.0552	0.138
YunqiuShao-thuir_legal	thuir_legal_1	0.0375	0.0599	0.0316	0.149
SaraRenjit-CUSAT_NLP	Task1_CUSAT_NLP_1	0.0300	0.0481	0.0412	0.166
SoumilMandal-JU_SRM	JU_SRM_1	0.0250	0.0478	0.0284	0.131
KavyaSGanesh	R1	0.0100	0.0416	0.0131	0.069
KayalvizhiS-SSN_NLP	SSN_NLP_1	0.0300	0.0405	0.0277	0.091
YunqiuShao-thuir_legal	thuir_legal_2	0.0225	0.0405	0.0221	0.095
SaraRenjit-CUSAT_NLP	Task1_CUSAT_NLP_2	0.0200	0.0264	0.0227	0.102
SoumilMandal-JU_SRM	JU_SRM_2	0.0175	0.0228	0.0163	0.065
HLJIT2019-AILA	HLJIT2019-AILA_task1_3	0.0150	0.0220	0.0066	0.065
SoumilMandal-JU_SRM	JU_SRM_3	0.020	0.0181	0.006	0.044
KayalvizhiS-SSN_NLP	SSN_NLP_2	0	0.0026	0	0.003
KayalvizhiS-SSN_NLP	SSN_NLP_3	0	0.0025	0	0.003

Table 1. Results of the AILA Task 1 - Precedent Retrieval.

The experimental results show that the improved BM25 method achieves the highest result in all evaluation measures. The BM25 method achieves the second-highest in 1/rank of first relevant document and P@10, MAP, BPREF has third place in all run. The Euclidean distance method has not achieved good results. The main reason is that we use General Corpus for word vector pre-training.

Table 2 shows the results of the task of statute retrieval.

The experimental results show that the HLJIT2019-AILA_task2_3 method achieves the second highest result in 1 / rank of first relevant document and the third place on BPREF. The HLJIT2019-AILA_task2_2 wins the third place on 1/rank of first relevant

document and the fourth place on BPREF. According to the results, the relevant prior case information is helpful to guide the judgment of current case.

Team name	Run ID	P@10	MAP	BPRE F	1/rank
YunqiuShao-thuir_legal	thuir_legal_2	0.0975	0.1566	0.0961	0.281
HLJIT2019-AILA	HLJIT2019-AILA_task2_3	0.0675	0.0819	0.0703	0.279
HLJIT2019-AILA	HLJIT2019-AILA_task2_2	0.0675	0.0773	0.0671	0.263
YunqiuShao-thuir_legal	thuir_legal_3	0.0900	0.1318	0.0742	0.247
YunqiuShao-thuir_legal	thuir_legal_1	0.0650	0.1115	0.0653	0.23
UBLTM	UBLTM1	0.0725	0.1022	0.0571	0.214
UBLTM	UBLTM2	0.0725	0.1023	0.0571	0.211
UBLTM	UBLTM3	0.0725	0.1023	0.0571	0.211
SaraRenjit-CUSAT_NLP	Task2_CUSAT_NLP_1	0.0550	0.0866	0.0412	0.202
SoumilMandal-JU_SRM	JU_SRM_5	0.0600	0.0918	0.0402	0.201
HLJIT2019-AILA	HLJIT2019-AILA_task2_1	0.0675	0.0606	0.0516	0.2
SaraRenjit-CUSAT_NLP	Task2_CUSAT_NLP_2	0.0550	0.0967	0.0377	0.199
KayalvizhiS-SSN_NLP	SSN_NLP_2	0.0475	0.0778	0.0494	0.191
SoumilMandal-JU_SRM	JU_SRM_6	0.0600	0.0831	0.0285	0.162
SoumilMandal-JU_SRM	JU_SRM_4	0.0600	0.0767	0.0309	0.146
KavyaSGanesh	R1	0.0350	0.0682	0.054	0.136
BabanGain-IITP	IITP_BM25_statutes	0.0200	0.0360	0.0397	0.129
KayalvizhiS-SSN_NLP	SSN_NLP_1	0.0250	0.0518	0.0285	0.128

 Table 2. Results of the AILA Task 2 - Statute Retrieval.

5 Conclusion

We describe an approach to Precedent Retrieval and Statute Retrieval that makes use of the improved BM25. Comparisons to the other submissions for the same tasks, our improved BM25 model achieves the top performers for the task of identifying relevant precedents on all evaluation measures. For the task of identifying relevant statues, the improved BM25 model wins the second place on 1/rank of first relevant document and the third place on BPREF.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (No.61806075, No.61772177), the National Social Science Fund of China (No.18BYY125), the Natural Science Foundation of Heilongjiang province of China

(F2018029) and the Social Science Fund of Heilongjiang Province of China (No.18TQB103).

References

- P. Bhattacharya, K. Ghosh, S. Ghosh, A. Pal, P. Mehta, A. Bhattacharya., P. Majumder, Overview of the Fire 2019 AILA tracks: Artificial Intelligence for Legal Assistance. In Proc. of FIRE 2019 - Forum for Information Retrieval Evaluation, Kolkata, India, December 12-15, 2019.
- 2. Mandal A, Ghosh K, Bhattacharya A, et al. Overview of the FIRE 2017 IRLeD Track: Information Retrieval from Legal Documents[C]//FIRE (Working Notes). 2017: 63-68.
- J. L. Fagan. Automatic Phrase Indexing for Document Retrieval: An Examination of Syntactic and Non-Syntactic Methods. In SIGIR'87, 1987.
- 4. J. Gao, J.-Y. Nie, G. Wu, and G. Cao. Dependence language model for information retrieval. In SIGIR'04, 2004.
- 5. F. Song and W. B. Croft. A general language model for information retrieval. In CIKM'99, 1999.
- M. Srikanth and R. K. Srihari. Biterm language models for document retrieval. In SIGIR'02, 2002.
- Robertson, Stephen, Zaragoza, et al. Simple BM25 extension to multiple weighted fields[C]// Thirteenth Acm International Conference on Information & Knowledge Management. ACM, 2004.
- Willett P. The Porter stemming algorithm: then and now[J]. Program Electronic Library & Information Systems, 2006, 40(3):p ágs. 219-223.