A Text Similarity Approach for Precedence Retrieval from Legal Documents

D. Thenmozhi
SSN College of Engineering
Chennai, Tamilnadu
theni_d@ssn.edu.in

Kawshik Kannan
SSN College of Engineering
Chennai, Tamilnadu
kawshik98@gmail.com

Chandrabose Aravindan
SSN College of Engineering
Chennai, Tamilnadu
aravindanc@ssn.edu.in

ABSTRACT

Precedence retrieval of legal documents is an information retrieval task to retrieve prior case documents that are related to a given case document. This helps in automatic linking of related documents to ensure that identical situations are treated similarly in every case. Several methodologies, such as information extraction based on natural language processing, rule-based method, and machine learning techniques, are used to retrieve the prior cases with respect to the current case. In this paper, we propose a text similarity approach for precedence retrieval to retrieve older cases that are similar to a given case from a set of legal documents. Lexical features are extracted from all the legal documents and the similarity between each current case document and all the prior case documents are determined using cosine similarity scores. The list of prior case documents are ranked based on the similarity scores for each current case document. We have evaluated our approach using the data set given by IRLeD@FIRE2017 shared task.

KEYWORDS

Precedence Retrieval, Information Retrieval, Document Similarity, Legal Documents

1 INTRODUCTION

Precedence retrieval is the process of retrieving relevant prior documents with respect to a current document. This is very important in common law system where a prior case which discusses similar issues can be used as a reference in the current case. This is to ensure that identical situations are treated similarly in every case. Recently, the number of digitally available legal documents has increased rapidly due to the developments in information technology. An automatic precedence retrieval system from legal documents helps legal practitioners to easily refer to the earlier cases that are related to the current case. Such a precedence retrieval system has several applications such as case based reasoning [2][8], legal citations and legal information retrieval [9]. Several approaches, such as information extraction based on natural language processing [4], rule-based aprroach [3], and machine learning techniqes [1], are used to retrieve the prior cases with respect to the current case. We propose to use a text/document similarity approach for precedence retrieval to retrieve relevant older cases for the current case from legal documents. In this work, we have focused on the shared task of IRLeD@FIRE2017[1] [6] which aims to retrieve prior case documents for a given current case document. IRLeD@FIRE2017 is a shared Task on Information Retrieval from Legal Documents (IRLeD) collocated with the Forum for Information Retrieval Evaluation (FIRE), 2017. The track has two tasks. Given a set of training cases with annotated catchphrases and a set of test cases, the first task is to extract the catchphrases present in the test cases. The second task is to retrieve all the relevant prior cases for a given current case. Our focus is on the second task of IRLeD@FIRE2017.

[1]https://sites.google.com/view/fire2017irled

2 PROPOSED APPROACH

We have implemented a document similarity approach for this IRLeD precedence retrieval task. We have used three variations of our approach namely i. Method-1 with concepts and TF-IDF (Term Frequency - Inverse Document Frequency) scores, ii. Method-2 with concepts, relations and TF-IDF scores, and iii. Method-3 with concepts, relations and Word2Vec. We have implemented our methodology in Python for this IRLeD task. The data set used to evaluate the Task 2 (Precedence retrieval task) of IRLeD shared task consists of 200 current case documents and 2000 prior case documents. The steps used in our approach are given below.

- Preprocess the given text
- Extract linguistics features from both current case documents and prior case documents
- Construct feature vectors for the documents using TF-IDF score or Word2Vec
- Find cosine similarity score between each current case with all the prior cases
- Rank prior cases based on the similarity score for each current case

The steps used in all the three methods are explained in detail in the sequel.

2.1 Method-1 with concepts and TF-IDF scores

The prior case documents and the current case documents are preprocessed by removing the punctuations like “, ”, - , ‘, ’, _, and the string ‘[?CITATION?]’ which is part of the text. The text is annotated with parts of speech (POS) information such as noun, verb, determiner, adverb, and adjective. In this method, only nouns are considered to obtain the concepts. All forms of nouns (NN*) namely NN, NNS and NNP are extracted from both current case text and prior case text and are lemmatized. The feature set is constructed by eliminating all duplicate terms from the lemmatized terms. The feature vector for each document is constructed using TF-IDF scores with respect to the features from the feature set. The cosine similarity scores between each current case document and

all the prior case documents are determined. The prior case documents are ranked based on the similarity score and are retrieved for each current case document.

We have used *NLTK* tool kit[2] to preprocess and annotate the given data with POS information. The extracted concepts from POS information are lemmatized using *Wordnet Lemmatizer*. The TF-IDF scores are obtained for the features by using *scikit-learn*[3] library (*TfidfVectorizer* from *sklearn.feature_extraction.text*). The similarity between each current case and the prior cases are obtained using scikit *cosine_similarity* from *sklearn.metrics.pairwise*. The prior cases for each current case are ranked based on the similarity scores (the prior case with highest similarity score is retrieved first).

2.2 Method-2 with concepts, relations and TF-IDF scores

In Method-2, we have considered both concepts and relations as features. All forms of nouns (NN*) namely NN, NNS and NNP to obtain the concepts and all forms of verbs (VB*) namely VB, VBZ, VBN, and VBD to obtain the relations are extracted from both current cases and prior cases POS information. The other steps like lemmatization, construction of feature vectors using TF-IDF, finding cosine similarity and ranking are similar to Method-1.

2.3 Method-3 with concepts, relations and Word2Vec

In Method-3, the key terms are extracted by using concepts and relations for each case from current set and prior set. The terms with respect to a particular case are lemmatized and vectorized into an array of dimensions 300 using Word2Vec [7]. The average of all the term vectors of the document is determined and that average represents the vector for the document. Likewise, the vector representations of all the prior case documents and the current case documents are obtained. Similar to the other two methods, the cosine similarity scores between each current case document and all the prior case documents are determined. The prior case documents are ranked based on the similarity scores and are retrieved for each current case document.

In this method, the key terms are obtained by extracting the terms that are tagged with NN, NNS, NNP, VB, VBN, VBZ and VBD. Each key term is lemmatized using Wordnet Lemmatizer and vectorized using Word2Vec *KeyedVectors.load_word2vec_format* from *gensim.models.keyedvectors* with 300 dimensions. We have used *GoogleNews-vectors-negative300.bin.gz*[4] for this vectorization.

3 RESULTS AND DISCUSSIONS

We have evaluated our document similarity approach for precedence retrieval of legal documents based on the metrics namely Mean Average Precision (MAP), Mean Reciprocal Rank (MRR), Precision@10 and Recall@10. The results of our approach are given in Table 1.

Method-2 which considers both concepts and relations from the text as features performs better than the other methods in terms of mean average precision and mean reciprocal rank with the values 0.2677 and 0.5457 respectively. Method-1 that considers only concepts as features gives better results for precision@10 and recall@10 with the values 0.1795 and 0.681 respectively. However, our third method does not perform well for this precedence retrieval of legal documents. The average of vectors used to represent the documents may not be a suitable solution. The performance may be improved if we use Doc2Vec [5], an extension of Word2Vec for vector representation.

Table 1: IRLeD Task 2 Performance

Method	MAP	MRR	Precision@10	Recall@10
Method 1	0.2633	0.5176	0.1795	0.681
Method 2	0.2677	0.5457	0.178	0.669
Method 3	0.101	0.277	0.0755	0.435

[2]http://www.nltk.org/

[3]http://scikit-learn.org/

[4]https://github.com/mmihaltz/word2vec-GoogleNews-vectors

ACKNOWLEDGMENTS

We would like to thank the management of SSN Institutions for funding the High Performance Computing (HPC) lab where this work is being carried out.

REFERENCES

[1] Khalid Al-Kofahi, Alex Tyrrell, Arun Vachher, and Peter Jackson. 2001. A machine learning approach to prior case retrieval. In *Proceedings of the 8th international conference on Artificial intelligence and law*. ACM, 88–93.

[2] Ramon Lopez De Mantaras, David McSherry, Derek Bridge, David Leake, Barry Smyth, Susan Craw, Boi Faltings, Mary Lou Maher, MICHAEL T COX, Kenneth Forbus, et al. 2005. Retrieval, reuse, revision and retention in case-based reasoning. *The Knowledge Engineering Review* 20, 3 (2005), 215–240.

[3] Filippo Galgani, Paul Compton, and Achim Hoffmann. 2015. Lexa: Building knowledge bases for automatic legal citation classification. *Expert Systems with Applications* 42, 17 (2015), 6391–6407.

[4] Peter Jackson, Khalid Al-Kofahi, Alex Tyrrell, and Arun Vachher. 2003. Information extraction from case law and retrieval of prior cases. *Artificial Intelligence* 150, 1-2 (2003), 239–290.

[5] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*. 1188–1196.

[6] Arpan Mandal, Kripabandhu Ghosh, Arnab Bhattacharya, Arindam Pal, and Saptarshi Ghosh. 2017. Overview of the FIRE 2017 track: Information Retrieval from Legal Documents (IRLeD). In *Working notes of FIRE 2017 - Forum for Information Retrieval Evaluation (CEUR Workshop Proceedings)*. CEUR-WS.org.

[7] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[8] Chieh-Yuan Tsai and Chuang-Cheng Chiu. 2009. Developing a Significant Nearest Neighbor Search Method for Effective Case Retrieval in a CBR System. In *Computer Science and Information Technology-Spring Conference, 2009. IACSITSC'09. International Association of*. IEEE, 262–266.

[9] Marc Van Opijnen and Cristiana Santos. 2017. On the concept of relevance in legal information retrieval. *Artificial Intelligence and Law* 25, 1 (2017), 65–87.