# Deep learning approach for precedence retrieval systems

Kayalvizhi S (312217405004)
ME CSE, Semester 3

Dr.D.Thenmozhi
Supervisor

**Project Review: 1** (17 August 2018)
Department of Computer Science and Engineering
SSN College of Engineering

## 1 Introduction

In the domain of law,the documents of previous similar cases are often used as references and tagging them with a set of keywords is essential.In making decisions,Judges are obliged to align their decisions to relevant prior cases.Thus,when lawyers prepare for cases, they research extensively on prior cases.In addition,Judges also consult prior cases that had already been decided to ensure that a similar situation is treated similarly in every case.Keywords give a very high-level description of a document.The reader can save himself a lot of time by quickly evaluating the relevance of the document to him by having a look at these keywords.They play a crucial role in reducing the search space and also act as a tool to find similarity between two documents with drastically low cost.In this project,we work to retrieve suitable prior case for the given current case and to extract catchphrases for a given document.

## 2 Motivation

In a Common Law System,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 issue(s) and which can be used as reference in the current case.This is to ensure that a similar situation is treated similarly in every case.If an ongoing case has any related/relevant legal issue(s) that has already been decided,then the court is expected to follow the interpretations made in the prior case.For this purpose,it is critical for legal practitioners to find and study previous court cases,so as to examine how the ongoing issues were interpreted in the older cases.Thus, an automated precedence retrieval approach is desired.

Legal texts (e.g., court case descriptions) are long and have complex structures.This makes their thorough reading time-consuming and strenuous.So,it is essential for legal practitioners to have a concise representation of the core legal issues described in a legal text.One way to list the core legal issues is by keywords or key phrases,which are known as catchphrases in the legal domain.Thus,an automated catchphrase extracion approach is needed.

# 3 Problem statement

Given a set of current case documents and a set of prior case documents,the system will extract the catchphrases present in the documents and retrive the relevant prior case documents for all the current case documents.

# 4 Literature survey

## 4.1 Distributed Representation in Information Retrieval[1]

In this paper,the given training and test documents are represented as vectors using Doc2Vec.Cosine distance between the current and prior document vectors are measured and ranked.To extract the catchphrases,cosine distance between the catchphrase vector and the document vectors are calculated.Based on this cosine distances the catch phrases are ranked.To calculate distance,cosine distance from python scipy pacakge2 have been used.

## 4.2 Detection of Catchphrases and Precedence in Legal Documents[2]

In this paper,the proposed system builds probabilistic model based on this training data,which,in turn,can predict the catchphrases in unseen legal texts.In preprocessing,each of the training statements were tokenized into a list and their Parts-of-Speech (POS) tags were generated.Another sequence of custom NER tagging was made by referring to token list and given catchphrases. B-LEGAL and I-LEGAL tags were employed for Begin and Intermediate of the catchphrases respectively and O for other tokens.In modeling,POS and custom NER(The problem of detecting catchphrases was modeled as customized NER.)tagging performed during pre-processing stage were used to form secondary features.These were used in building CRF model.The generated model file was then used to predict from test data.The CRF++ model was used to predict custom NER tags from the given testing data.The final similarity score is computed as a weighed average of the scores generated by 3 approaches namely regular expressions based,topic modeling based and Doc2Vec similarity based.Each current-prior case pair has a final score based on weighted sum of scores from individual approaches.The results were presented as sorted list of prior case for each current case.

## 4.3 Catchphrase Extraction from Legal Documents Using LSTM Networks[3]

In this paper,the problem is formulated as a classification task and the objective is to learn a classifier using LSTM network. The proposed methodology involves a pipelined approach and is divided into four phases namely pre-processing,candidate phrase generation,creating vector representations for the phrases,training a LSTM network.

## 4.4 Catch Phrase Extraction From Legal Documents Using Deep Neural Network[4]

In this approach,deep neural network provides an elegant way to extract catchphrases. The features used here include grammar,Tf-idf, position in a document etc.In this paper for each file a set of potential meaningful phrases were created and then are classified using deep neural network.Steps involved are preprocessing,create potential meaningful phrases based on common grammar of phrases,feature selection,label the vectors,classification and training the model.

## 4.5 A Text Similarity Approach for Precedence Retrieval from Legal Documents[5]

In this paper,they have proposed 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 were 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 were ranked based on the similarity scores for each current case document.

## 4.6 Text Similarity

### 4.6.1 A Deep Network Model for Paraphrase Detection in Short Text Messages[6]

In this paper,they have proposed a hybrid deep neural architecture composed by a convolutional neural network (CNN) and a long short-term memory (LSTM) model.The proposed paraphrase detection model is composed of two main components,i.e.,pairwise word similarity matching and sentence modelling.The pair-wise similarity matching model is used to extract fine-grained similarity information between pairs of sentences.They have used a CNN to learn the patterns in the semantic correspondence between each pair of words in the two sentences that are intuitively useful for paraphrase identification.

## 5 Existing system

The existing approaches have used cosine similarity approach to retrieve the suitable prior case for the given current case and then rank the prior case documents[1] [5].Probabilistic model have also been used in which the phrases are tokenized, labelled and then classified[2].Deep neural network have been proposed to detect catchphrases in which for each file a set of potential meaningful phrases were created and then are classified using deep neural network[4].A three step approach for extracting catchphrases from legal documents have also been used in which the objective is to learn a classifier using LSTM network[3].Deep learning approach have been used to detect the paraphrase by a hybrid deep neural architecture which composed of CNN and LSTM model.Each sentence in a pair are converted into a semantic representative vector,using a CNN and an LSTM.Then,a semantic pair-level vector is computed by taking the element-wise difference of each vector in the sentence representations.The resulting difference is the discriminating representative vector of the pair of sentences,which is used as feature vector for learning the similarity between the two sentences[6].

## 6 Proposed system

In our work,suitable prior case for the given current case will be retrieved using deep learning approach and the catchphrases from documents will be extracted.

## References

[1] Reshma, U and Kumar, M Anand and Soman, KP,*Distributed Representation in Information Retrieval-AMRITA_CEN_NLP@ IRLeD 2017.*FIRE (Working Notes)69–71,2017.

[2] Kulkarni, Yogesh H and Patil, Rishabh and Shridharan, Srinivasan,*Detection of Catchphrases and Precedence in Legal Documents*.FIRE (Working Notes) 86–89,2017.

[3] Bhargava, Rupal and Nigwekar, Sukrut and Sharma, Yashvardhan,*Catchphrase Extraction from Legal Documents Using LSTM Networks*.FIRE (Working Notes) 72–73,2017.

[4] Das, Sourav and Barua, Ranojoy, *Catch Phrase Extraction from Legal Documents Using Deep Neural Network.*,FIRE (Working Notes)78–79,2017.

[5] Thenmozhi, D and Kannan, Kawshik and Aravindan, Chandrabose, *A Text Similarity Approach for Precedence Retrieval from Legal Documents.*,FIRE (Working Notes)90–91 ,2017.

[6] Agarwal, Basant and Ramampiaro, Heri and Langseth, Helge and Ruocco, Massimiliano,*A Deep Network Model for Paraphrase Detection in Short Text Messages* ,arXiv preprint arXiv:1712.02820,2017.