ISSN 2278-3091

Volume 9 No.2, March - April 2020

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse208922020.pdf

https://doi.org/10.30534/ijatcse/2020/208922020

Knowledge Extraction from Text Document Using Open Information Extraction Technique

Payal Kadu[1], Ashwini V Zadgaonkar[2]

[1]M.Tech Student, CSE Department, Shri Ramdeobaba College of Engineering and Management, Nagpur, Maharashtra, India, kadupn@rknec.edu

[1]Assistant Professor, CSE Department, Shri Ramdeobaba College of Engineering and Management, Nagpur, Maharashtra, India, zadgaonkarav1@rknec.edu

ABSTRACT

As digital era is growing apace and with this humongous quantity of knowledge is obtainable on the web one such field is legal domain. This paper proposes an approach in which we suggest extraction of knowledge from legal documents with open information extraction technique. In open information extraction technique extraction is done while not requiring dataset that is expounded thereto specific document or that specific domain. Since huge quantity of information is available in legal domain and it is present on the web it has become one of the favored spaces of research. With text mining one can find facts and relationships which is present in the legal textual document and these textual documents has huge amount of information. This extracted knowledge can further be used by ordinary people or practitioner in legal domain. The key purpose of open information extraction is to induce relation tuple while not employing a dataset that is related to particular domain and as a result open information extraction has gained popularity for extracting information. Thus in this paper we extract information from legal documents with open information extraction as legal document can be associated with any domain.

Key words: Legal documents, Open information extraction.

1. INTRODUCTION

1.1 Text Mining

In text mining one could obtain useful facts from text documents. Standard statistics may be derived through the layout of styles which is often referred as statistical learning. Collection of unstructured data from information assets including text documents, webpages, emails, pdf documents and net blogs and plenty of more. Certainly one of the essential steps in textual content mining is facts pre-processing and appearing operation for cleaning, detecting and getting rid of inconsistency in the records. This allows us to acquire base of particular words and you can extract and hold on to useful facts concealed within the data. To acquire this we want the data in structured format as to begin with the data is in unstructured or semi structured format and once to procure the records in structured format you can in addition have a look at the pattern for analysis. After all useful facts are obtained it could be used for making higher decision for an organization also it is able to be used for text summarization or question answer system.

1.2 Information Extraction

Enormous quantity of data is available in different forms and all these data is accessible on the web, for example, news, websites, blogs, and magazines articles however these are in ambiguous format and getting information becomes a tedious task. For this we need some competent and proficient instrument that is termed as information extraction. Essentially we are able to coin information extraction as instrument to gain structured information from unstructured information. Information extraction assimilates natural language processing, name entity recognition, relation extraction.

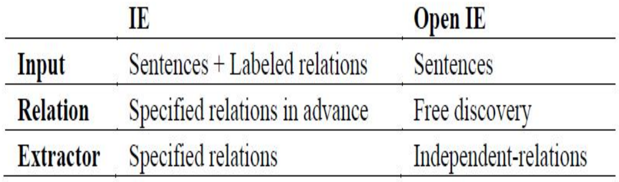
1.3 Open Information Extraction

The task of extracting information from massive amount of data without the requirement of pre-defined data set is termed as open information extraction. Open Information Extraction is one of the crucial and initial step in natural language processing for text mining applications for example relation extraction, text summarizer. Open information extraction is specified for extracting tuples and relations from massive data which is not related to a specific domain. In this era of internet online information is present in enormous amount one such field in which we get large amount of data is legal information that is information related to legal domain which is not bound to a pre specified field. We can use open information extraction to find relation and use it for further analysis since legal documents are large in size relation extraction with open information extraction can help practitioners and ordinary people to get the essence of the document.

Open information extraction as the name suggests it the endeavor of extracting data from tremendous amount of data without the necessity of pre-defined data set. Open Information Extraction is one among the important and initial step in natural language processing for textual content mining applications as an instance relation extraction, text summarizer. Open information extraction is targeted for extracting tuples and relations from huge data which isn't associated with a particular domain. In this era of internet online data is present in widespread amount one such field in which we get massive quantity of data is, legal documents that is records associated with legal domain which isn't always certain to a pre specified field. We can use open information extraction to find relation and use it for further analysis in addition legal documents are big in size, relation extraction with open information extraction can assist practitioners and ordinary human beings to get the essence of the document.

2. LITERATURE REVIEW

Duc-Thuan Vo and Ebrahim Bagheri [7], in their paper discussed the challenges of traditional information extraction that the tradition information extraction is based on some pattern matching technique which is domain specific. Extraction of data which isn't area explicit is cannot be done utilizing traditional methodology of data extraction. Refer figure 1 for differences between information extraction and open information extraction.



	IE	Open IE
Input	Sentences + Labeled relations	Sentences
Relation	Specified relations in advance	Free discovery
Extractor	Specified relations	Independent-relations

Figure 1: IE vs. Open IE [7]

Fei Wu and Daniel S. Weld [3], in their paper presented WOE Extractor, WOE extractor is web based extractor. WOE has 3 important step for information extraction first is pre-processing it uses sentence splitting and NLP , second step is called matcher which does the task of sentence matching third and last step is learner which is associated to pattern classification. It uses Wikipedia to train information extractor.

Luciano Del Corro and Rainer Gemulla [1], in their paper presented ClausIE, a novel, statement based way to deal with open data extraction, which extracts relations and their contentions from characteristic language content. ClausIE in a general sense varies from past methodologies in that it isolates the identification of helpful snippets of data communicated in a sentence from their portrayal as extractions. In more detail,

ClausIE exploits linguistic particulars about the grammar of the English language to primarily recognize statements in an input sentence and eventually recognize the type of each clause according to the grammatical function as defined. In context of this data, ClausIE can create high-accuracy extractions; the portrayal of these extractions can be redone to the basic application. ClausIE depends on reliance parsing and a little arrangement of domain autonomous lexica, works sentence by sentence with no post-handling, and requires no training information.

Galgani F, Compton P, Hoffmann A [15], in their paper proposed an approach for summarization of legal documents by using knowledge base with ripple down rules. These ripple down rules were considered for selection of catchphrases. For the purpose of summarization they have used data set of AustLII that is Australasian Legal Information Institute.

Farzindar A, Lapalme G [11] in their paper they described an approach to summarize the legal documents of federal courts of Canada and they represented this summarized information in a table-style summary in legal domain. They generated summary under four headings namely introduction, context, juridical analysis and conclusion and they have used Federal court of Canada as a data set for their document's summarization.

Kavila SD, Puli V, Raju GP, Bandaru R [13], in their paper of legal document summarization described a hybrid system based on key phrase or key word matching and case based technique. The authors proposed relating roles appeal no of the case, year of proceeding, case, judges, respondent, appellant, sections, facts, judgments, and verdict. They identified various labels using different kind of features like similar words, abbreviated words, and word length.

Kumar R, Raghuveer K [14], in their paper proposed an approach to for short summary generation for legal documents using topics obtained from latent dirichlet allocation formally known as LDA. The author passed a document to LDA as a bunch of words and on basis of probabilistic model different topics were generated. They also described an algorithm for calculation of sentence scores whose scores were based on probability of appearance of a particular word with reference to each topic. They used a data set keralawyer.com which consist 5 sub-domains and about 116 documents.

3. PROPOSED APPROACH

Extraction of information must not depend on relations which are specified beforehand, the relation must be generated automatically this will allow us open extraction of information from text documents. With the use of open information extraction on legal documents ordinary people and practitioners can get essence of the document by using it in text summarization applications.

We have taken raw data from indiankanoon.com. indiankanoon.com is a website which has record of court cases. Input to the system is a text document sample case from indiankanoon.com website.

Pre-processing start with cleaning the data by elimination of ambiguities. We start with elimination of stop words that is words which are commonly used and are not important words. Example for stop words includes words such as are, is, and these qualify for elimination.

3.1 Parts of speech tagging

Once we are done with elimination of stop words the next step is tagging the parts of speech that is POS tagging. Parts of speech tagging is nothing but tagging each word with the part of speech it belong to. This is done with the part of speech tagger which is available in natural language processing toolkit in python. Consider the example for understanding of parts of speech tagging in table 1: "Samuel lives in the red bungalow at the end of the street."

Table 1: Parts of speech tagging

Word	POS	Word	POS
Samuel	Noun (NN)	at	Preposition (IN)
lives	Verb (VBZ)	the	Determiner (DT)
in	Preposition (IN)	end	Noun (NN)
the	Determiner (DT)	of	Preposition (IN)
red	Adjective (JJ)	the	Determiner (DT)
bungalow	Noun (NN)	street	Noun (NN)

3.2 Name entity chunking

In the above example we have seen how parts of speech tagging is accomplished next step is Name Entity Chunking. Chunking in natural language processing refer to chunk of information to find a conclusion this can also be called as shallow parsing the group which shows up as a result is called as a "chunk". Consider the following example statement in figure 2 "The book has many chapters" for understanding how chunking is processed.

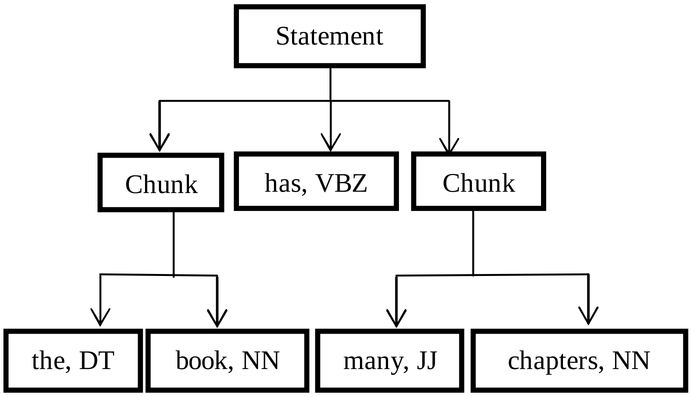


Figure 2: Name entity chunking

3.3 Name entity recognition and relation extraction

Named entity recognition (NER)is most likely the initial move towards information extraction that looks to find and characterize named words in content into pre-characterized classes, for example, the names of people, associations, areas, articulations of times, amounts, money related qualities, rates and others.

For the extraction of relations that is to find the subject, object and predicate value for that we will use some rules like

1. if the word start with a noun (NN) it will be considered as the subject.
2. Secondly if start verb it will be considered as predicate.
3. Third if the tag starts with determinant (DT), noun (NN), and preposition (IN) then we will put that entity in the predicate value category.
4. Fourth if the word is labeled as noun and it belongs to object category as labeled by named entity recognition it be considered as object.

Once we are done with subject, object and predicate identification. We will move on to the final extraction it will be done as extracting verb that will end with preposition and then nearest subject will be identified for more understanding we will take a sample example statement from the text document from indiankanoon.com which is a website that holds the record of court cases in India.

Sample input statement: "The charge sheet reads as under "During investigation, it was disclosed that Krishan Lall was born on 01 st July 1943 at village Guthawali, Khurd District."

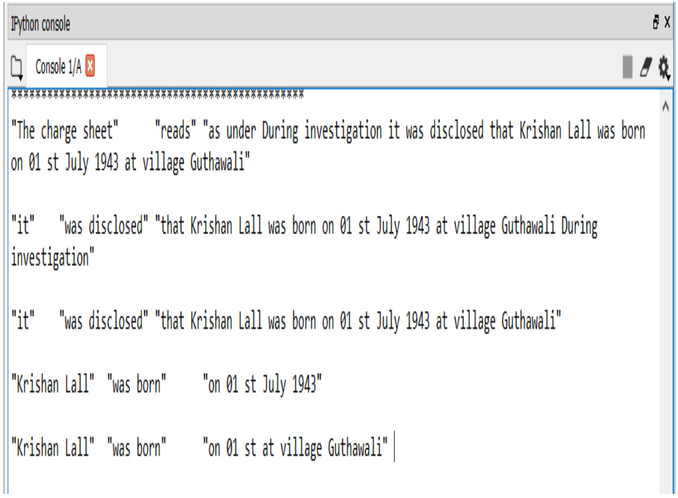
Extracted output 1: "The charge sheet" "reads" "during investigation"

Extracted output 2: "it" "was disclosed" "that Krishan Lall was born on 01 st July 1943 at village Guthawali in Khurd District in Bulandsahar"

Extracted output 3: "Krishan Lall" "was born" "on 01 st July 1943".

4. RESULT AND DISCUSSION

Following figure 3 shows the output screen of the proposed approach:



```
************************************************
"The charge sheet"    "reads" "as under During investigation it was disclosed that Krishan Lall was born
on 01 st July 1943 at village Guthawali"

"it"   "was disclosed" "that Krishan Lall was born on 01 st July 1943 at village Guthawali During
investigation"

"it"   "was disclosed" "that Krishan Lall was born on 01 st July 1943 at village Guthawali"

"Krishan Lall"  "was born"    "on 01 st July 1943"

"Krishan Lall"  "was born"    "on 01 st at village Guthawali"
```

Figure 3: Extracted information with open information extraction on legal document.

5. CONCLUSION AND FUTURE SCOPE

Knowledge extraction form text documents using open information extraction technique is better than traditional information extraction because for open information extraction, the document need not belong to specific domain as open information extraction is unsupervised learning and does not depend on any pattern matching technique. As discussed previously there is a lot of data present in legal domain and legal domain data is not related to a pre specified field so open information extraction is a good way to extract information from legal documents. So in this paper we have extracted information from legal document using open information extraction technique. The proposed system can be used in a question answering system.

REFERENCES

1. Luciano Del Corro, Rainer: **ClausIE based open information extraction**. *International world wide web conference committee* 2013. https://doi.org/10.1145/2488388.2488420
2. Anthony Fader, Stephen Soderland, Oren Etzioni: **Identifying relations for open information extraction**. *Proceedings of 2011 conference on empirical methods in natural language processing*. July 27-31 2011 pp. 1535-1545. https://doi.org/10.18653/v1/W17-5402
3. Fei Wu, Daniel Weld: **Open information extraction using wikipedia.** *Proceedings of the annual meeting of the association for computational linguistics*. 11-16 July 2010 pp. 118-127.
4. Martin Hassel: **Evaluation of automatic text summarization.**
5. Xiang Ren, Yuanhua Lv, Kuansan Wang, Jiawei Han: **Comparative document analysis for large text corpora.** *WSDM 2017 Cambridge United Kingdom*. February 6-10 2017 pp. 325-334.
6. Ah-Hwee Tan: **Text mining the art and the challenges**. Kent ridge digital labs Singapore.
7. Duc-Thuan Vo, Ebrahim Bagheri: **Open information extraction.** *Encyclopaedia with semantic computing* 2016.
8. Christina Niklaus, Matthias Cetto, Andre Freitas, Siegfried Handschuh: **A survey on open information extraction.** *Proceedings of the 27th international conference on computational linguistics*. August 20-26,2018 pp. 3866-3878.
9. Rudolf Schneider, Tom Oberhauser, Tobias Klatt, Felix A. Gers, Alexander Loser: **Analysing errors of open information extraction systems.** July 24, 2017.
10. Ambedkar Kanapala, Sukomal Pal, Rajendra Pamula: **Text summarization from legal documents: a survey**.June 29, 2017. https://doi.org/10.1007/s10462-017-9566-2
11. Farzindar A, Lapalme G: **The use of thematic structure and concept identification for legal text summarization.** *Computational linguistics in the north east, Canada*. August 2004 pp. 67-71.
12. Galgani F, Compton P, Hoffmann A: **Citation based summarization of legal texts.** *Trends in artificial intelligence, Springer* 2012 pp. 40-52.
13. Kavila SD, Puli V, Raju GP, Bandru R: **An automatic legal document summarization search using hybrid system.** *Proceedings of the international conference on frontiers of intelligent computing* 2013 pp. 229-236. https://doi.org/10.1007/978-3-642-35314-7_27
14. Kumar R, Raghuveer K: **Legal document summarization using latent dirichlet allocation.** *Int J computer science telecommunication* 2012 pp. 3:114-117.
15. Galgani F, Compton P, Hoffmann A: **Combining different summarization techniques for legal text.** *Proceedings of the workshop on innovative hybrid approaches to the processing of textual data, association for computational linguistics* 2012 pp. 115-123.
16. Elvan Owen, Dwi H. Widyantoro: **Development of extensible open information extraction.**
17. Arun Krishna Chitturi: **Survey on Abstractive Text Summarization using various approaches.** *International Journal of Advanced Trends in Computer Science and Engineering* Vol. 8 no.6, November-December 2019 pp. 2956-2964. https://doi.org/10.30534/ijatcse/2019/45862019
18. Dian Sa’adillah Maylawati1, Yogan Jaya Kumar, Fauziah Binti Kasmin, Basit Raza: **Sequential Pattern Mining and Deep Learning to Enhance Readability of Indonesian Text Summarization**. *International Journal of Advanced Trends in Computer Science and Engineering* Vol. 8 no.6 , November-December 2019 pp. 3147-3159. https://doi.org/10.30534/ijatcse/2019/78862019
19. Michele Banko, Michael J Cafarella, Stephen Sotherland, Matt Broadhead, Oren Etzioni: O**pen information extraction from the web.** *International Joint Conference on Artificial Intelligence* 2007 pp. 2670-2676.