Second Asia Pacific Conference on Contemporary Research (APCCR, Malaysia, 2016)
ISBN:978 0 994365 606
www.apiar.org.au

SIMPLIFYING LAW STATEMENTS USING NATURAL LANGUAGE PROCESSING

Nayana Dharmasiri [a], Bashitha Gunathilake [b], Umavi Pathirana [c], Sajani Senevirathne [d]
Anupiya Nugaliyadde [e], Samantha Thellijagoda [f]
[abcdef] Sri Lanka Institute of Information Technology, Malabe, Sri Lanka
Corresponding email: nayanad.own@gmail.com

Abstract

Understanding the law statements for general public is evidently complex. The research derives a computational solution on reducing the complexity of the law statements. Given a law statement, the research will use both wordnet and "LawNet" to create a simpler meaning. The research will focus on information extraction, information retrieval, question analysis and answer generation techniques to derive better meaning of law statements. The law statement will be treated as a question and the "LawNet" and wordnet will be used in as information extraction points. The law statement will be analyzed as a question; more information will be retrieved through the wordnet and "LawNet". This process mostly acts similar to a search engine's process. The results provide on average 80% accuracy for a 1500 dataset.

Keywords: Natural Language Processing, Natural Language Understanding, Ranking, Stemming

1. Introduction

Law is one of the most important aspects for a citizen because it serves as a norm of conduct for all citizens and residents. Thus, there is much information about law only a few can understand since the complexity of the usage of language. Therefore, most of the people miscommunicate what is accepted in the society.

At present, after utilizing the existing search engines, we can discover information with respect to law of Sri Lanka. Despite the fact that we have the essential information, we are not aware of the precise meaning of the existing information because of the complexity of it (Srilankalaw.lk, n.d.). For any citizen to understand the law of Sri Lanka, the information should have a simpler format. This paper focuses on describing the processes followed for building flexibility, ease of use and productive search engine. The search engine will provide the user with the most accurate, reliable and up to date information of law in a simple understandable format. Furthermore, it provides the user an ability to perceive whether an act or a conflict is a legal one and whether a legitimate solution is available. By taking the above-mentioned specific benefits into consideration, this document will present the most suitable and optimized methodology to implement functional areas of the Search Engine. When we consider the Sri Lankan law, there are documents everywhere throughout the web. But, the huge amounts of information are in the most complex format, which has a major effect on legal awareness and legal literacy. Lawyers are aware of the law yet, as they can't discover data with respect to previous cases and law documents instantaneously, this leads them to lose the current cases. Through this product "Justice", problems faced by normal citizen, law students and lawyers are mostly focused and provides the flexible, user friendly, cost-effective and productive search engine to educate every citizen of the country regarding law of the country and make it beneficial for lawyers and law students in learning law related commodities.

ISBN:978 0 994365 606
www.apiar.org.au

2. Methodology

All information used for this particular search engine is stored in a customized file system. This system consists of Law Net, Law cases, Law Reports, Constitution of Sri Lankan Law, Penal code, Code of Criminal procedure and Acts.

"LawNet" is made up of a glossary and definitions of words found in the law domain. This is in .xml format. The Law cases found in the engine are dispute between two parties resolved by a court or some similar legal procedure. It can be either a civil or criminal case. For each case, a summary and the final verdict are included.

The law reports included in the file system are a series of documents that contains judicial opinions from a group of selected law cases. The constitution of Sri Lankan law is a document of fundamental principles according to which Sri Lanka is governed. The 126th Amendment and all restrictions included in the Constitution are used in this Search engine. The Penal code documents a significant amount of a particular jurisdiction's criminal law. The Code of Criminal

Procedure is the main legislation on the process for administration of substantive criminal law. The Acts included are documents that record facts of something either said or done.

The search engine also includes WordNet, which is a lexical database for English language. It is used in this search engine to obtain definitions and synonyms of key words (Srilankalaw.lk, n.d.).

For this Research, the methodology used is Natural Language Processing (NLP), which involves the processing of any natural language text. Natural Language Understanding (NLU), which is a sub field of NLP in artificial intelligence that deals with machine reading comprehension.

Natural Language Tool Kit (NLTK) is the tool used for NLP and it provides a platform for building python programs to work with Human Language data.

In this Search engine, the user is allowed to search for any keyword. Crawling, Indexing and retrieval will be the three basic processes carried out in the search engine. Once the keywords are recognized, the search engine crawls through the file system to discover the required content and then it will be indexed. During indexing, the information obtained from the file system is analyzed and ranked according to the ranking model. And finally it retrieves the relevant information from the file system.

Figure 1: Information Extraction

In the proposed search engine, all information generated will be relevant for the law domain. But, all such information will be in its simplest format so that anyone even with a minimal knowledge of law can understand the content. All such information generated is extracted from documents and resources saved in the inbuilt file System.

ISBN:978 0 994365 606
www.apiar.org.au

The final implementation of the proposed System will consist of four major functions mentioned below;

Information Extraction Information Retrieval Question Analysis

Answer Generation

Once the user enters a question in the system it is analyzed under three phases.

Question processing

Question type detection Query formulation.

Question processing phase involves the extraction of the relevant pieces of information from the question entered. Next the Answer type is determined by the Question type Detection Phase. Finally, a list of keywords will be determined from the question entered during the Query Formulation phase. These keywords are then used to retrieve information from the in-built file system.

Question processing can be further broken down into Sentence segmentation, Tokenization, Stemming and Removal of stop words. During the process of segmentation, questions entered will be broken down into meaningful units such as words, sentences or topics. This is a part of natural language processing. Tokenization separates the question entered into tokens mainly by depending on previously obtained results (Techipedia, n.d.).

It follows the steps mentioned below.

The question is divided into tokens or words at all whitespaces, punctuation marks or line break present.

The inclusion of white spaces or punctuation marks depends on the requirements of the system

A token may consist of alpha characters, alphanumeric characters or numeric characters only. Tokens also act as separators. For example, white spaces are not required when placing identifiers with arithmetic operators.

Stemming is the removal of morphological affixes from words in order to obtain the main word stem (nltk, n.d.). The algorithm used for stemming in this particular search engine is Snowball. In Information Retrieval algorithmic stemmers are of great use. But, a few algorithmic descriptions of stemmers are liable to misinterpretation (snowball.tartarus, n.d.).

Removal of stop words is the next step in question processing. Stop words are the words or adjectives, which are very frequently used. Words such as ‘the’, ‘is’, ‘or’ etc. interferes with the Search engine optimization effort and they use character spaces for the particular task (blogs.iit.edu, 2013).

The question-processing phase comes to an end once the above four steps completed. At the end of this phase, useful information is extracted from the question entered by the user. This information is then used in the question Detection phase.

In the question detection phase the question is analyzed. The type of question entered by the user is determined during this phase. For example, if the question entered is “What is Law?” the keyword extracted in the earlier phase will be “What”.

Query Formulation phase comes after the question detection phase. Query Formulation is the task of creating a list of keywords from the question entered and forms a query that will be sending to the information Retrieval system. The query formulation always depends on the question entered. When this method is applied to the system, it simply create a keyword list from every keyword in the question letting the search engine removes the stop words according to the above method mentioned. In this case we always leave the Question word

ISBN:978 0 994365 606
www.apiar.org.au

such as what, where, when. As a substitute keywords can be arranged only using the terms found in the noun phase of the question, and also by applying the stop word detection it ignores function words low-content, high frequency verbs.

The Question Processing phase will be ending after the Query Formulation is done.

Answer Generation is the next main function of this research. It includes two main functions namely;

Answer type Detection Answer Processing Once question processing is completed, answer type detection is the next immediate function. This function is carried out according to a knowledge-based technique.

This is because the file system used by the search engines consists of a vast amount of information of which about 95% is in a well-structured form while the rest of the 5% is in semi structured form

All possible questions entered by the user fall under one of the three main categories mentioned below. The search engine is able to produce the most suitable answer according to the category of the question.

Namely they are;

Rule Based Supervised
Semi- Supervised

Methods required to answer all three categories of questions were used in this particular research. Rule based method is used for the relations that occur frequently. It is also known as hand-written rules since the relation can be extracted right away from the question entered. For example, to extract a definition from the file system, this method includes written patterns that search for the question word "what", a main verb and then extract the name entity argument of the verb.

In some scenarios, the file system contains supervised data, which consists of a question paired with the most correct and best logical form. The task of this method is to extract pairs of training tuples and try to reproduce the functionality that can then be mapped from new questions to get their logical form.

The most supervised algorithms are aligned to the parse tree in a logical form to learn to answer the simple questions about relations. Normally, this system bootstrap by having a small set of constraints to create this mapping and also an initial lexicon.

Semi-supervised method is used to deal with variations. Since it is difficult to train data sets with labeled questions with representing the meaning of them, supervised data sets can't cover the vast amount of question forms even the simplest factoid questions. Because of this reason, most of the methods for mapping factoid questions to the established relations or other structures are used to find the use of textual redundancy.

Once the answer type detection is completed, Information Retrieval is the next functionality that should be done. Information Retrieval is the action of getting information resources significant to an information need from a collection of files or from another data storage. In this research the file system is used to retrieve the information such as Law Net, Law cases, Law Reports, Constitution of Sri Lankan Law, Penal code, Code of Criminal procedure and Acts.

Once the user inputs the user query to the system "Justice" it will go through the above methods question processing and answer type detection. The query that was created in above methods is used for the Information Retrieval function. In Information Retrieval, a query does not particularly recognize a solitary object in the collection. Rather, in most cases, several objects may match the query with different levels of significance.

ISBN:978 0 994365 606
www.apiar.org.au

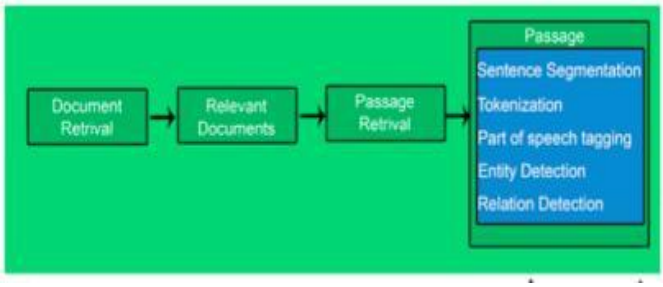

Figure 2: Information Retrieval

The file system should be indexed to initiate the retrieval process. As the user specify their need through the search query which will be then parsed and transformed by text operations applied to the text.

Despite the fact that the reports are arranged according to the importance of parts positioned, the top-positioned document may not be the relevant response to the entered question. This is because the documents are not the appropriate unit to rank with respect to the objectives of a question-answering framework. A highly relevant and large document that does not prominently answer a question may not be the perfect possibility for further processing (Jurafsky & Martin, 2015).

In "Justice", all the files are indexed. Hence, the next step is to extract a set of potential answer passages from the retrieved set of law documents. The meaning of a passage is fundamentally system dependent. However, the regular units include segments, passages, and sentences. This research have used paragraph segmentation algorithm to extract the relevant passages. Mainly in retrieving the passages, this particular search engine uses the Law cases, Law Reports and Penal Code as the files.

Retrieval of passages is the next immediate function in "Justice". In this stage, first the system shift through passages in the returned documents that don't contain potential answers. Afterward, rank the rest of the documents, which are prone to contain a response to the query. The initial phase in this procedure is to run a named entity or answer type classification on the retrieved passages.

Answer type detection phase decide the answer type from the entered question and let the search engine know the possible answer type, which will help to find in the answer. The System; therefore, filters out the documents that don't contain any entities of the right type. And finally, the remaining passages are then ranked by relying on a small set of components that can be extracted from the candidate answer passages, which were extracted from the file system.

The number of named entities of the right type in the passage.

The number of question keywords in the passage.

The longest exact sequence of question keywords that occurs in the passage.

The rank of the document from which the passage was extracted.

The N-gram overlap between the passage and the question.

Count the N-grams in the question and the N-grams in the answer passages.

Once ranking is completed, answer processing is the next immediate function, which comes under Answer Generation.

ISBN:978 0 994365 606
www.apiar.org.au

Figure 3: Answer Generation

Since the final output of the system "Justice" consists of several segments such as 'Definition', 'Law cases', 'Cases Referred to' and 'Punishments' the system should extract information from text files as well as from .xml files. And additionally, the system gives the opportunity to get registered to "Justice". When the users get registered, the user can either register as a lawyer, law student or as a law firm. If the user is registering as a lawyer, the user should enter the Membership.

ID, which was provided by the Bar Association of Sri Lanka, for the verification.

The system has used simplification algorithms to simplify the extracted content of the law cases, to give the maximum understanding to the users. This algorithm uses the synonyms of the words, which are used in the law cases. Also, the system ignores all the stop words while simplifying.

At the end of the main functions, the system gives the user the most reliable answers, which are easily understandable for all the citizens in Sri Lanka.

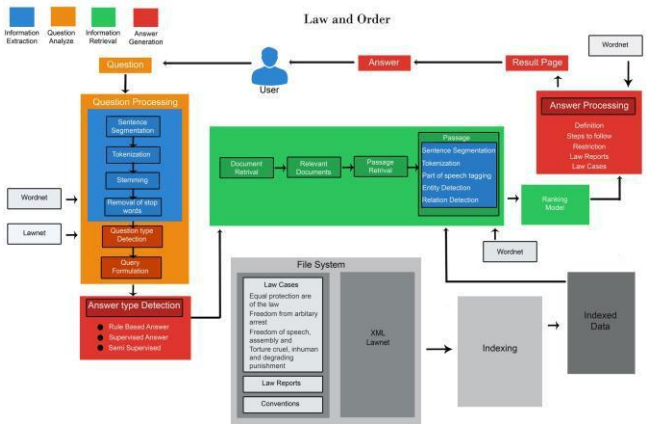

Figure 4: System Overview

3. Result and Discussion

This search engine acts as an effective and accurate answer generator. As the final result, the engine produces simplified versions of 'Held' in law cases, Definition, Punishments and Restrictions.

According to the keywords searched for, the top five of the most relevant simplified versions of "Held" in law cases are indexed. The accuracy of relevance for these cases is around 60% in this system.

ISBN:978 0 994365 606
www.apiar.org.au

If any of the keywords searched for are found in the law domain, then the definitions are also shown. Also, all punishments relevant to the keywords are extracted from the penal code and are shown.

Restrictions are also extracted from the 126th amendment of the constitution according to the searched keywords.

Lawyers, law firms, law students and clients can be registered as users in the system. The first year after the launch of the search engine, is free of charge to its users. But, a membership fee will be charged annually afterwards.

If a law firm registers, it can also advertise its vacancies on the system.

The LawNet is available online for further research (srilankalaw, n.d.).

4. Conclusion

This research provides possible methodologies, which are best suitable to implement information extraction technique, information retrieval technique, question analysis technique and answer generation technique. The Search engine, itself is an exceptional engine from existing products being as a comprehensive product which includes simplification, instantaneous retrieval, complex question analysis and exact sentence extraction all in one.

When it comes to Law of Sri Lanka, information exists in all available search engines but most of the required information are still in books to refer manually. Hence, for a common citizen, it has become a difficult task to find information instantaneously. This has cause a drastic fall in the awareness of law of Sri Lanka among the common citizen. Therefore, utilizing all the information available in a way that it will enhance the awareness of Law would have been an extraordinary matter of significance nowadays. At present, utilizing existing search engines we can discover information regarding law of Sri Lanka. Despite the fact that we have the essential information, we are not aware of the precise meaning of the existing information because of the complexity of it. For any citizen to understand the law of Sri Lanka, the information should have a simple format, which is understandable by every citizen.

Considering the huge amount of information of Sri Lankan Law available in the web, all of them are in the most complex format which has a major effect on legal awareness and legal literacy. While lawyers are aware of the law yet as they can't discover data with respect to previous cases and law documents instantaneously, it leads them to lose the current case. In order to increase the awareness of law of Sri Lanka, all the available information must be in a simplest form of English language.

“Justice” caters the need of solving the problems faced by the common citizen and also the lawyers and law students.

“Justice” comes with all the value addition feature such as simplification of complex information, instantaneous retrieval, complex question analysis and exact sentence extraction. Law documents such as law cases, law reports, acts and law codes are the major sources for “Justice” which are stored in the file system. Without a proper maintenance of the file system, the system won’t be able to provide the exact requested information. Hence, proper maintenance of information and availability will be the main requirement for “Justice”. Continuous change of law documents might be a challenge for this kind of a technological solution. The world is changing day by day. Therefore, more research should be performed in order to identify new technological ways to face the challenges of the world.

ISBN:978 0 994365 606
www.apiar.org.au

5. Future Work

This research was carried out only for fundamental human rights, one of the many areas in the law domain. As such, in the future, this system could be developed to cover all such areas in law domain. Additionally, the System could be developed in Sinhala and Tamil languages, to make it user-friendlier for the citizens in Sri Lanka.

As of now, the System only produces simplified law information of about 60% accuracy. But, the system can be improved to provide information of higher accuracy.

The LawNet can also be improved by adding more law terms providing a vast area in law to be covered.

References

i. Angeli, G., Nayak N. & Manning, C.D. (2016) 'Combining Natural Logic and Shallow Reasoning for Question Answering', *Association for Computational Linguistics*.

ii. Available at: https://docs.python.org/2/library/xml.etree.elementtree.html. Accessed 22nd August 2016.

iii. blogs.iit.edu (2013) 'SEO: The Evil Stop Words.' [Online] Available at: http://blogs.iit.edu/iit_web/2013/04/29/seo-the-evil-stop-words/ Accessed 24th August 2016.

iv. Chowdhury, G.G. (2005) 'Natural language processing', *Annual Review of Information Science and Technology*, 37(1), pp. 51–89.

v. Fishkin, R. (2014) 'Why SEO is important - the beginners guide to SEO.' *Moz*. [Online] Available at: https://moz.com/beginners-guide-to-seo/why-search-engine-marketing-is-necessary. Accessed: 29th August 2016.

vi. Greenwood, M.A. (2005) *Open-Domain Question Answering* (1st Ed.) UK: University of Sheffield.

vii. Jurafsky, D. & Martin, J.H. (2008) *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (2nd Ed.) United States: Pearson Prentice Hall.

viii. Jurafsky, D. & Martin, J.H. (2015) *"Question Answering" in Speech and Language Processing*

ix. Knowles, M. (2008) 'Short history of early search engines – the history of SEO.' In *thehistoryofseo*. [Online] Available: http://www.thehistoryofseo.com/The-Industry/Short_History_of_Early_Search_Engines.aspx. Accessed: 28th August 2016.

x. Levene, M. (2005) *An introduction to search engines and web navigation* (2nd Ed). United States: Addison-Wesley Educational Publishers.

xi. Manning C.D., Raghavan, P. & Schütze, H. (2008) *Introduction to Information Retrieval*. Cambridge University Press.

xii. Manning, C.D. & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

xiii. Neves, M. (2016) *NLP09_Questionanswering* (1st Ed.) Potsdam, Germany Web. [Online] Accessed 31st August 2016.

xiv. nlp.stanford.edu (n.d.) *The Stanford Natural Language Processing Group*. [Online] Available at: http://nlp.stanford.edu/. Accessed: 22 August 2016.

xv. nltk.org (n.d.) 'Stemmers.' [Online] Available at: http://www.nltk.org/howto/stem.html. Accessed: 28th August 2016.

xvi. Python.org (n.d.) 'The ElementTree XML API.' [Online]

xvii. Roberts, K. & Hick, A. (2000) *Scaling Answer Type Detection to Large Hierarchies*. USA: Language Computer Corporation.

xviii. Rubin, Y. & Engineer, S. (2010) 'Various client-server communication mechanisms in an Ajax-based web application.' [Online] Available at: http://www.ibm.com/developerworks/library/wa-aj-ajaxcomm/ Accessed 31st August 2016.

xix. Srilankalaw.lk (n.d.) 'Introduction'. [Online] Available at: http://www.srilankalaw.lk Accessed 25th August 2016.

xx. Tambimuttu, A.V. (2009) 'Sri Lanka: Legal research and legal system – GlobaLex.' [Online] Available at: http://www.nyulawglobal.org/globalex/Sri_Lanka.html. Accessed 29th August 2016.

xxi. Wikipedia (n.d.) 'Natural language processing.' *Wikimedia Foundation* [Online] Available at: https://en.wikipedia.org/wiki/Natural_language_processing. Accessed 27th August 2016.

xxii. Wordstream (2005) 'The history of search engines – an Infographic.' [Online] Available at: http://www.wordstream.com/articles/internet-search-engines-history. Accessed 24th August 2016.

xxiii. www.github.com, 'Law Glossary', [Online]. Available: https://github.com/sajani93/Law-Glossary. Accessed: 20 August 2016.

xxiv. www.osac.gov (2015) 'Sri Lanka 2015 crime and safety report'. [Online] Available at: https://www.osac.gov/pages/ContentReportDetails.aspx?cid=17573, Accessed 26th August 2016.

xxv. www.snowball.tartus.org (n.d.) 'Snowball: A language for stemming algorithms.' [Online]. Available at: http://snowball.tartarus.org/texts/introduction.html. Accessed 22nd August 2016.

xxvi. www.techipedia.com (n.d.) 'Tokenization.' [Online]. Available at: https://www.techopedia.com/definition/13698/tokenization Accessed 25th August 2016.