Ph.D. Research Proposal

Information Extraction for Legal Documents

Haojie Huang

School of Computer Science

University of New South Wales

May 10, 2017

Abstract

Information Extraction technique, which is a process of converting unstructured text into structured data, has been used in a variety range of applications such as traffic monitoring systems, biomedical systems and more general open knowledge bases for open language learning. However, the use of information extraction in legal area is still quite limited and suffers difficulties. For example, the existing system AustLII[1], which is developed and maintained by University of Technology Sydney and University of New South Wales, uses traditional information retrieval approach to organise all the legal documents and the system only supports keywords search. Another example, the LexML Web project in Brazil also applies similar approach. Although more sophisticated techniques such as case based reasoning, semantic network and ontology based approaches are introduced, not many of them are actually used in an integrated application. Hence, the aim of this study would be to come up with new design that applies information extraction techniques in order to construct a more accurate legal text management system. The new system is designed to manage laws as well as legal cases for both legal study and use. Comparing with AustLII, the new system will also focus on Australia laws but provide more accurate search results because of the use of information extraction methods.

[1]Australasian Legal Information Institute

Contents

Chapter 1

Introduction

Currently, information extraction has become an attractive research field for researchers. More and more intelligent systems such as a question-answer system are built depends on the development of information extraction techniques. Examples of these techniques are name entity recognition (NER), co-reference resolution, ontology extraction etc. With such methods, relations between entities can be identified and saved in a structured format, which allows more complex queries rather than plain text search, sort of like letting the machine understand the content of the text.

Because of these properties, research on information extraction based systems such as traffic monitoring and medical document extraction has been done in recent years. However, not much work is done in legal documents. Traditionally, to support fast search on laws, an inverted index is created [7], which is like a search engine that allows users to search a particular law by typing in keywords. Such systems can be used by experienced users, for example, laywers, but not for users who are not familiar with the local law.

Using the AustLII System [7] as a starting point, I would like to develop a law knowledge base system that supports more complex and intelligent queries. This proposal starts by describing the techniques in both information retrieval and information extraction, then followed by comparing different methods that researchers used for indexing legal documents. These materials are discussed in Chapter 2. Then in Chapter 3, I will talk about the procedure for development my own system, and the total research plan and workload is described in Chapter 4.

Chapter 2

Literature Review

2.1 Background

Information explosion is what people used to describe the rapidly increasing data in the information age today. It is no doubt that more data are available than ever before. Because of the size and growth speed, it is not possible for people to manually read through everything manually. Hence, techniques are applied to help people organise, search, filter and manage electronic documents.

In the early ages, Information Retrieval (IR) is one of the most important ways being researched in both research field and used industry. It has been researched for a long time, and now more works turn to focus on a related but different field called Information Extraction (IE). This section provides an overview of both information retrieval and information extraction. More details such as real applications that use these techniques are also introduced in the following section.

2.1.1 Information Retrieval

Salton [10] defines information retrieval as the field of study dealing with the representation, storage, organisation of and access to documents in the year 1968. In contrast to a relational database, information retrieval is dealing with unstructured text documents. The most widely known examples of IR systems are search engines, where Google is the most successful one as most people know.

In addition, IRs are also used by different technical groups in constructing more specific index systems in their own research field. For example, the Global Legal Research project WorldLII[6] manages a database of legal documents over the world and it aims to index these documents for experts in laws.

However, IR systems suffer issues in a large amount of text data. Most people may have the same experience that in order to search one particular result, the query highly depends on the keywords chosen. Moreover, users are unlikely to read the search results if it is too long. In specific research areas, accuracy is another issue that information retrieval may suffer.

2.1.2 Information Extraction

To overcome the problems in information retrieval, researchers started to focus on a new type of text process technique named Information Extraction (IE). Rather than building indices for all documents, information extraction attempts to fetch useful information from raw text documents and store it into a structured format[13]. It involves more tasks which require natural language processing techniques to achieve the goal. For example, the most basic task is named entity recognition (NER), which tries to identify all names, locations, times and other important noun phrases in a text document.

Obviously, a system which applies IE techniques works more accurately than a system using IR. However, it requires more effort to gain higher accuracy. More specifically, an IE system usually focuses on a narrower area of study. For example, Zhou et al.[17] introduced a framework for biomedical event extraction in 2015. Alternatively, another example is that Liu et al.[11] used NER and a weighted undirected bipartite graph to extract key entities from the online daily news. More generally both their work is called event extraction as they both focus on one particular type of event. As [9] states, event extraction is useful in risk analysis applications, monitoring systems and tools that help to make decisions.

2.1.3 Open Information Extraction

In order to make wider use of information extraction, researchers have introduced Open Information Extraction (OIE) [2], which is a paradigm that extracts a large set of relational tuples in a much more general domain of articles without human input. Michele et al. [2] also introduce their implementation TextRunner which they claimed to be the state-of-the-art at the time they published their paper. In their work, TextRunner is compared to KnowItAll [5], which is an earlier domain-independent web extraction system. One important feature of KnowItAll is that it uses regular expressions based on part-of-speech tags to identify noun phrases instead

of NER. This does extend the scope of content to be identified; however, it requires much longer time up to several weeks to complete a single task.

To overcome these issues, [2] designed three main modules in TextRunner, which are the Self-Supervised Learner, the Single-Pass Extractor and the Redundancy-Based Assessor. In more detail, the Self-Supervised Learner trains a small set of input data and outputs a classifier. The Single-Pass Extractor makes a single pass over all corpus and makes a part-of-speech tagging on all the words. Then all the noun phrases can be found and relations can be formed by eliminating those non-essential phrases. Finally, the Redundancy-Based Accessor eliminates some of the redundant relations.

2.2 Information Extraction in My Own Research

Open Information Extractors like TextRunner [2] does extend the ability of information extraction from specific areas to a wider scope in extracting relations from the open web comparing to tradition IE. However, the downside can be quite obviously in this stage, whereas the scope extended the accuracy decreased. Indeed, for some of the more specific research areas, it is not necessary to use OIE since they require more professional structured data, or sometimes called a knowledge base, to make decisions. Traditionally, in order to get higher accuracy, it requires experts to pre-define a set of rules for the extraction process, and the whole process is time-consuming. By reviewing the techniques used in KnowItAll and TextRunner, it is possible to apply such methods on a specific area of study. In my research, I will start from this idea and try to build a knowledge base using information extraction techniques for **legal documents**.

2.3 Information Extraction in Legal Area

The reason that I chose the legal area to be a start is because it is highly related to information retrieval and information extraction. First of all, there are a large amount of documents available in this field, and in most cases, these documents are strictly written in the formal form, which makes it easier for Natural Language Processing (NLP) to process than those extremely informal social media data. People uses legal documents for many reasons. For example, some of them want to study laws; some of them want to organise and build argumentation; some amateur users want to know how seriously the case is when they involved in a traffic accident. Secondly, the changing law systems make it even more difficult for users to study. Hence, a

knowledge base which may automatically manage new laws and cases will be helpful for all those who want to make use of legal documents.

2.3.1 Review of Existing Systems

Currently, there are several systems that achieved the representation of legal documents. One example is the LexMl[1] project in Brazil released in 2009. LexMl stores the documents in XML format. Its objective is to establish data in the context of identifying and structuring legal documents. Alternatively, FindLaw.com[2] is another legal information site which allows consumers, small-business owners, students and legal professionals to find solutions to legal questions and legal counsel. It was found in 1995 and now it is one of the largest website of legal documents in the world.

Moreover, the Cato system [1] is a case-based reasoning system that stores a collection of case and the court opinions. It can produce a judgement for new cases by comparing with the old ones. The Salomon [14] system divides law documents into frames and relations, and it stores these frames in a graph data structure. More details of these two systems are given in the following section.

2.3.2 Review of Exsiting Approaches

There are several existing approaches that may use for representation of legal knowledge. They are Case-based reasoning (CBR), Semantic Network and Ontology-Based Approach[4].

2.3.2.1 Case-based reasoning (CBR)

Case-based reasoning (CBR) is the process of solving new problems based on the solutions of similar known problems. In the legal study, this method is widely used manually by lawyers where they usually promote a particular outcome based on legal precedents. Similar to rule induction algorithms in machine learning, CRB starts with a set of training examples, which represent the previous cases. These cases all consist of the problem statement as well as a given solution, sometimes with extra data such as annotations about how the solution is found. Then it tried to generalise these examples by classification approaches. The generalised cases are the experience that can be used if a new case comes in.

[1] http://projeto.lexml.gov.br
[2] http://findlaw.com

As it points out in [3], CBR systems are not simply retrieving text cases from plain text documents, they summarise abstract fact patterns contained in a case. One typical example is the CATO system [1], which is an intelligent tutoring environment for teaching new law students making arguments with cases. CATO collects about 150 cases and these patterns are represented in more detailed patterns. More precisely, legal cases are represented by facts and decisions. It organises these facts and decisions into a hierarchical structure called Factor, which contains 26 base-level factors for trade secret law, 16 abstract factors and 50 links [1]. Refering back to CBR, when new cases come in, CATO uses these factors for computing similarity between the new case and the factors, so that a relevant judgement can be produced.

2.3.2.2 Semantic Network

Using Semantic Networks is another approach for representation of legal knowledge. A semantic network is a graph that represents semantic relations between concepts. The graph can either be directed or undirected, where the vertices represent concepts and the edges represent relations between those concepts. Semantic Networks are widely used in artificial intelligence fields, especially in natural language processing. A typical example of a semantic network is WORDNET [12], which is a lexical database for English language.

In the legal area, [4] states that a semantic network can be used to represent the causal aspect between facts. For example, SALOMON[14] is a criminal case summary system that applies semantic network.

SALOMON stands for "Summary and Analysis of Legal texts for Managing On-line Needs". It formalises the text grammar as a semantic network of **Segment Frames**, where each segment frame is a short noun phrase text. All these segment frames are connected together hierarchically, sequentially or conditionally. More specifically, their relations can be categorised to "has a", "precedes" and "if ... then" respectively. Figure 2.1 shows an example of such semantic network in SALOMON. By representing legal knowledge in this way, similar cases can be found if they both have similar semantic structure. And even further, the structure helps summarise criminal cases and hence could help quickly identify and locate relevant cases.

To achieve the goal of constructing a semantic network, [4] implements a parser to process the original legal text data based on text grammar. The parser starts by recognising the head segment frame, which refers to the complete text or an overview of the text. Then the parser identifies the subsegments in a depth-first approach. The result forms those "has a" relation shown in Figure 2.1. After that, a

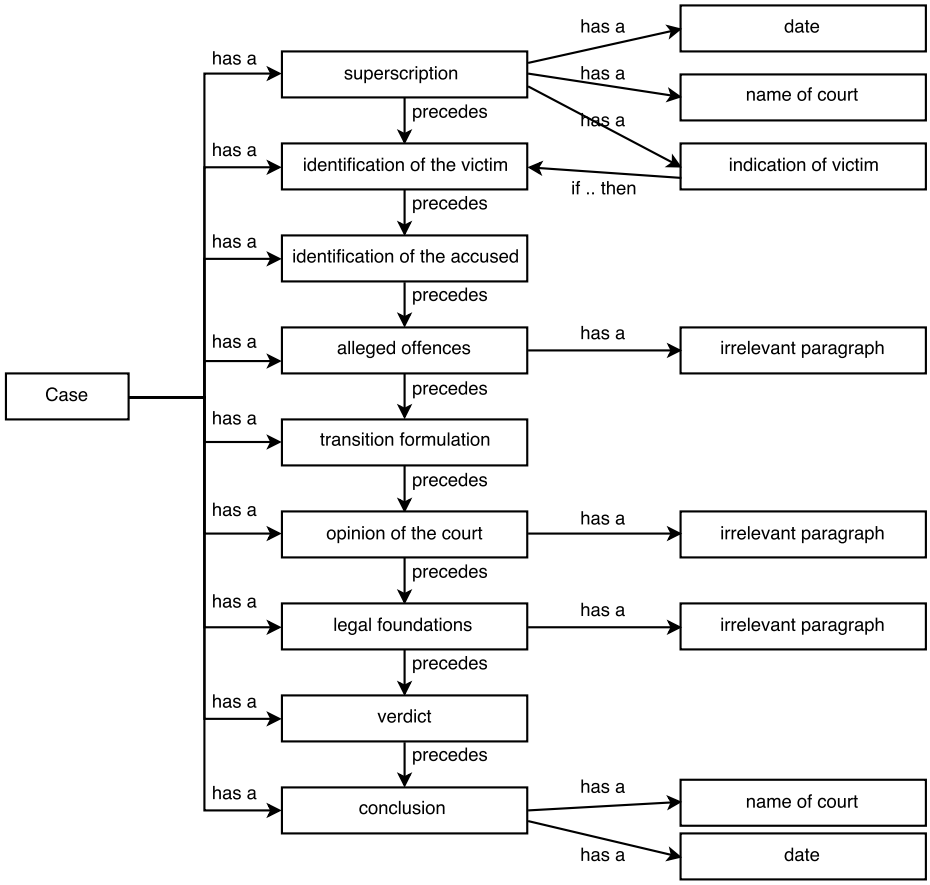

Figure 2.1: Representation of the text structure of a criminal case

recognition procedure identifies the "precedes" and "if ... then" relations by a specific text grammar.

2.3.2.3 Ontology Approach

Ontology is a term borrowed from philosophical study. In information science, it refers to formal naming and definition of types, properties or any entities that really or fundamentally exist in a particular domain. Similar to its original meaning in philosophy, it represents entities, ideas and events in a system. For example, in geopolitical domain an ontology can be concepts like countries, states and cities.

An ontology-based information extraction (OBIE), as stated in [16], is a system that process unstructured or semi-structured natural language data over a mechanism guided by ontologies to extract certain types of information. Normally, OBIE systems have different depending on how people want to use them. Howerver, [16] gives a common archotecture of a OBIE system, which is shown in Figure 2.2. In practise

some of the components are possibly not presented in some real OBIE applications, but such architecture is sufficient to do most of the ontology extraction and query tasks.

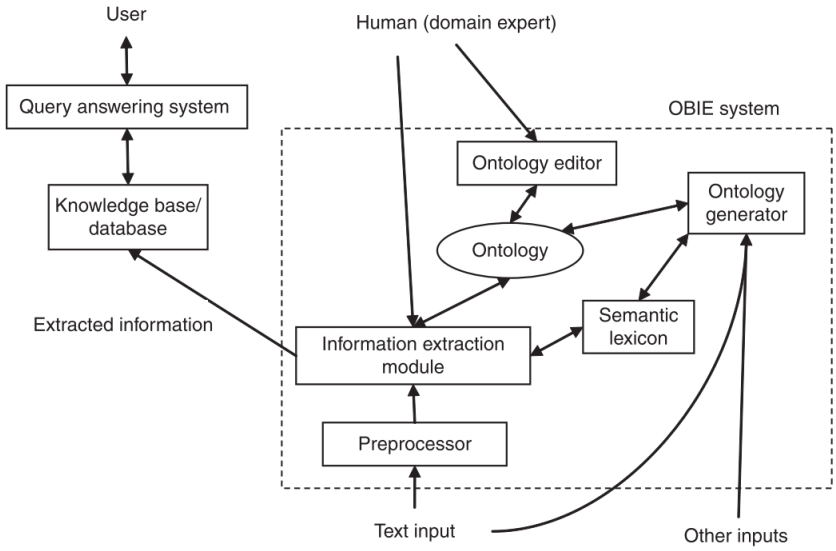

Figure 2.2: General Architecture of an OBIE System

In an OBIE system, text input is firstly passed through the preprocessor for removing meaningless symbol characters such as HTML tags and converted into a pure text format. Then extraction procedure takes place in the information extraction module. In different tasks, specific extraction modules will be applied for a particular task for extracting the knowledge required. In order to support this task, human experts may be involved in the extraction task, or they may pre-define a set of ontologies for the specific task, such as legal knowledge. A semantic lexicon module is also applied for helping information extraction. After all, the information extraction module stores its output to a knowledge base to allow users doing queries.

Although ontology approach is suitable for knowledge to represent and query in the legal area, it is hard to find new examples. The well known ones are those in the 1990s, for instance, the Frame Based Ontology (FBO) [15] and Ontolingua [8].

2.3.3 More Information Extraction in Legal Study

As stated in the beginning of this section, information extraction techniques help to grab useful information from plain text documents and re-organise it into a structured format. It is also seen from the previous work that a lot of systems has come up with their own approach to do the work. Locally in Australia, **The Australasian Legal**

Information Institue [7] (AustLII) is a joint facility of Law Faculties of several top universities in Sydney. It is best known as a website, where it maintains a large database of legislations, case laws and indices for legal study. All the materials in AustLII are still organised using traditional information retrieval techniques, just like a search engine.

Comparing to systems such as Cato, AustLII has obviously a much greater amount of text documents. It is hard to do case summarise like Cato did in traditional approaches. As a result, modern techniques that have been applied to OIE may be considered as a way to achieve information extraction on the entire database in AustLII. In the next section, methods which are going to be used to do information extraction, as well as queries, will be discussed.

Chapter 3

Research Methodology

In this section, the proposed methodology to achieve the research aims is described. The sequence of theoretical and practical research activities are also included in the following sub-sections.

As discussed in the last part of the previous section, the main goal of this research is to construct a similar law management system like AustLII by using information extraction technologies instead of information retrieval which is what currently used by the system. We aim to take advantage of the more accurate and more structured characteristics of IE systems and apply it to the AustLII law documents. Different from the old IE attempts in the 1990s, OIE techniques will be considered in this research. More specifically, the whole methodology procedure can be seen in Figure 3.1. More specific description of how this system is constructed is described in the following sections.

3.1 Analysis of existing information extraction methods

The most effective way to construct a new application is to firstly build one use the existing techniques and follow what the previous researchers did. In this case, my first step will be to go through some of the open information extraction methods. For example, I would review the core modules in the source code of the Ollie [1] system, since its code is open sourced.

From the literature review section, it is found that the Ollie System extracts relations in any plain texts and produces its outputs in forms of ‘¡obj¿ ¡relation¿ ¡sub¿.

[1] The upgraded version of the KnowItAll Project

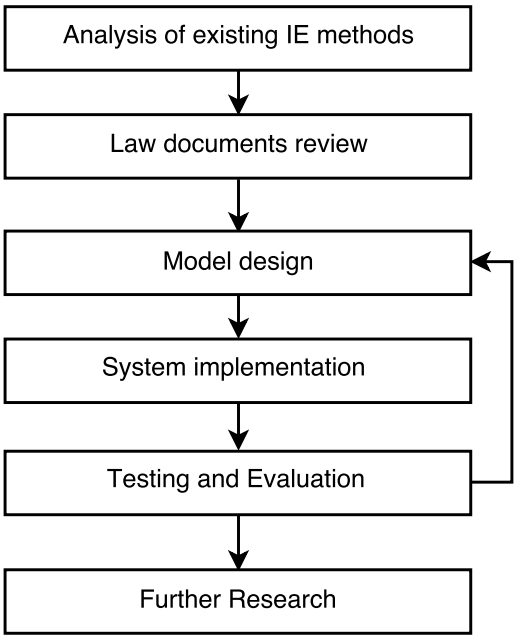

Figure 3.1: Research Process

To achieve this goal, it uses a large set of linguistic relations written in regular expression form and applies that to the text. The regular expressions help it accurately captures the relations in plenty of plain text documents. So it is also essential to know how these regular expressions is used in the system structure, and try to eliminate the redundant components as well as the redundant rules in the original system.

3.2 Overview of Law documents in the AustLII system

In this stage, some of the law documents in the AustLII System need to be reviewed as I am not a professional in legal documents.

There are 2 main reasons that I need to do a brief review on some of the legal documents. Firstly, it is helpful for getting better understanding of the law structures. In order to make the final application more professional, I required to learn a bit about laws in Australia, since the AustLII System is a huge index of Australian laws. Example of questions that I need to know are the difference between case laws and legislations, and what is added to a particular law in which year.

Secondly, this stage helps me design a data structure for storing laws and cases. It is found that AustLII stored the large amount of laws in inverted index, which

is traditional information retrieval approach. To get more accurate result in queries, more details such as time, location and even meaning of the content need to be recorded as well. This is also the final stored pattern for users' queries.

3.3 Framework design of the new IE system

This stage is immediately followed by reviewing law documents. The primary goal of the framework is to decompose a complex system into modules so that different time can be allocated to different modules in the implementation stage by the importance. In this case, I have supposed a simple framework for the law IE system as a starting point, wherein this basic version there are 5 modules in the backend: a preprocessor, an information extractor, a grammar database, a knowledge database for storing index and a query processor. A more detailed diagram is shown in Figure 3.2.

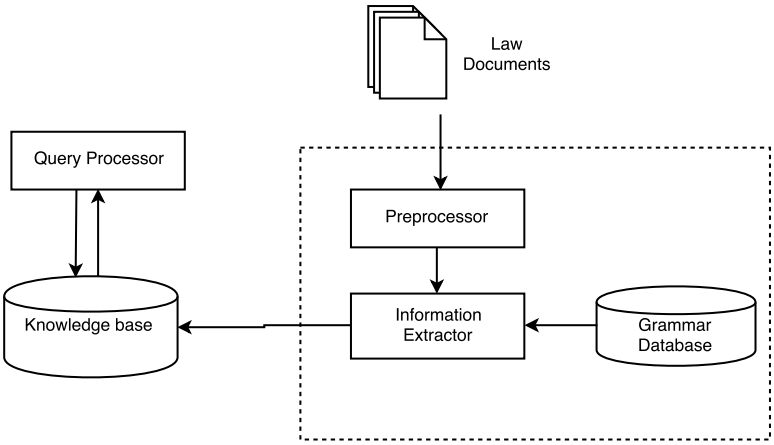

Figure 3.2: Framework Design

It is no doubt that the extractor and the grammar database are the two key components in the system. As a result, more time will be allocated to design a more detailed inner structure for them. For instance, the grammar database may require a lot more research on keywords that are most commonly used in law documents.

3.4 Framework implementation

This stage involves the implementation of all the components which are listed in the design stage. In my current plan, Python will be the main programming language to

be used. The code of each functional component will be done in object-oriented programming, where for the 2 databases, any kinds of the traditional relational database system can be considered.

More specially, the implementation procedure will be started from the preprocessor, and then the grammar database, followed by the information extractor. As stated before, most of the time will be spend and focus on the grammar information extractor, where if possible, redesign steps will be taken. On the other hand, the query system will be another major part of the system that requires some efforts on both design and implementation too. After that, it comes to the final step where a user interface may be taken into account. The final step is optional depending on how much time it will take.

3.5 Testing and Evaluation

As shown in Figure 3.1, this is a stage with an extra arrow pointing back to the design stage, because the redesign of the algorithms and even the data structure may be required if the test result is not accurate enough. The testing and evaluation procedure consists of the integration test of the system itself, as well as a comparison with the AustLII System.

In terms of integration testing, test cases will range from simple queries with 2 words up to complex sentences that describe a case. Similar to previous information extraction systems, precision and recall will be taken to be tested as the primary features. Then more will be on the relevance of search results.

Moreover, in comparison with the AustLII System, the same set of queries would be used to test both of the systems. In this part, the accuracy will be mainly focused. As mentioned before, it is possible that more than one design of the new law IE system will be taken, the precision and recall of different design will also be compared.

3.6 Further Research

As mentioned before, the law IE System is the very first step of my whole research subject. It is no doubt that new problem would appear in the process of implementing the whole system, some of these problems may be recorded down and provide a new direction for the rest of my research, so that the idea of information extraction can be applied to a wider scope and more applications in the real world.

Chapter 4

Research Plan

This section defines a schedule by semester for the 3 years of research. Under each semester, a list of percentage of the time allocated is marked and more key features are fully explained following the table.

	Lit. Review	**Model Dev.**	**Exp**	**Software Dev.**	**Thesis/Paper Writing**	**Course Work**
2016 S2	20%	5%	5%	5%	-	65%
2017 S1	20%	20%	-	-	10%	40%
2017 S2	20%	20%	20%	20%	20%	-
2018 S1	20%	20%	10%	40%	10%	-
2018 S2	10%	-	40%	-	40%	-
2019 S1	10%	-	40%	-	40%	-
2019 S2	10%	-	40%	-	40%	-

Table 4.1: Research schedule and tasks weighted for each semester

4.1 Key Features

- **Lit. Review** Literature Review of existing publications and recources that are related to my research topic.
- **Model Dev.** Development of framework and model of the expected system.
- **Exp** Experiment of the model that developed in the model development step.
- **Software Dev.** Development of the whole software system. More specifically, it is an integration step for grouping the models together to form a complete system.

- **Thesis/Paper Writing** Writing of any publications and the final thesis of interim results.
- **Course Work** Coursework courses that are required to complete the task.

Chapter 5

Conclusion

Overall, this proposal presents the background of information extraction research as well as a proposed law information extraction system that I would complete in my research.

The works that have been done in information retrieval and information extraction provide the various amount of methods that I can use in my further research. Moreover, the AustLII System creates a good starting point for guiding my research direction since it has some aspects that can be improved. Its great amount of text resource is also helpful in designing and implementing my law information extraction system. In order to achieve the final goal, this proposal provides a brief guideline for the next couple of years' research tasks, but feedback from the public and other researchers can be taken into account so that it leads to modification of the whole research process.

Bibliography

[1] Kevin D Ashley and Vincent Aleven. Reasoning symbolically about partially matched cases. In *IJCAI (1)*, pages 335–341, 1997.

[2] Michele Banko, Michael J Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. Open information extraction from the web. In *IJCAI*, volume 7, pages 2670–2676, 2007.

[3] Stefanie Brüninghaus and Kevin D Ashley. The role of information extraction for textual cbr. In *International Conference on Case-Based Reasoning*, pages 74–89. Springer, 2001.

[4] Denis Andrei de Araujo, Carolina Müller, Rove Chishman, and Sandro José Rigo. Information extraction for legal knowledge representation–a review of approaches and trends. *Revista Brasileira de Computação Aplicada*, 6(2):2–19, 2014.

[5] Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S Weld, and Alexander Yates. Unsupervised named-entity extraction from the web: An experimental study. *Artificial intelligence*, 165(1):91–134, 2005.

[6] G Greenleaf. Global legal research: Worldlii and the future. *Internet Newsletter for Lawyers*, 2005.

[7] Graham Greenleaf, Andrew Mowbray, and Philip Chung. Austlii: Thinking locally, acting globally. *Australian Law Librarian*, pages 101–116, 2011.

[8] Thomas R Gruber. *Ontolingua: A mechanism to support portable ontologies*, volume 27. Citeseer, 1992.

[9] Frederik Hogenboom, Flavius Frasincar, Uzay Kaymak, and Franciska De Jong. An overview of event extraction from text. In *Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2011) at*

Tenth International Semantic Web Conference (ISWC 2011), volume 779, pages 48–57. Citeseer, 2011.

[10] Michael E Lesk and Gerard Salton. Relevance assessments and retrieval system evaluation. *Information storage and retrieval*, 4(4):343–359, 1968.

[11] Mingrong Liu, Yicen Liu, Liang Xiang, Xing Chen, and Qing Yang. Extracting key entities and significant events from online daily news. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 201–209. Springer, 2008.

[12] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[13] Sharon Gower Small and Larry Medsker. Review of information extraction technologies and applications. *Neural computing and applications*, 25(3-4):533–548, 2014.

[14] Caroline Uyttendaele, Marie-Francine Moens, and Jos Dumortier. Salomon: automatic abstracting of legal cases for effective access to court decisions. *Artificial Intelligenceand Law*, 6(1):59–79, 1998.

[15] Robert van Kralingen. A conceptual frame-based ontology for the law. In *Proceedings of the First International Workshop on Legal Ontologies*, pages 6–17. Citeseer, 1997.

[16] Daya C Wimalasuriya and Dejing Dou. Ontology-based information extraction: An introduction and a survey of current approaches. *Journal of Information Science*, 2010.

[17] Deyu Zhou and Dayou Zhong. A semi-supervised learning framework for biomedical event extraction based on hidden topics. *Artificial intelligence in medicine*, 64(1):51–58, 2015.