Information Extraction Framework to Build Legislation Network

Neda Sakhaee · Mark C Wilson

Abstract This paper concerns an Information Extraction process for building a dynamic Legislation Network from legal documents. Unlike supervised learning approaches which require additional calculations, the idea here is to apply Information Extraction methodologies by identifying distinct expressions in legal text and extract quality network information. The study highlights the importance of data accuracy in network analysis and improves approximate string matching techniques for producing reliable network data-sets with more than 98 percent precision and recall. The values, applications, and the complexity of the created dynamic Legislation Network are also discussed and challenged.

1 Introduction

Legislation Networks were first introduced in 2015 [23] and further discussed in [34] [33]. These networks are essential to explore the relationship between legislation and societies' evolution [34]. There are many obvious benefits from studying Legislation Networks [34] [33] [39] [14], but building these

N Sakhaee
Computer Science Department, University of Auckland
E-mail: nsak206@aucklanduni.ac.nz

M C Wilson
Computer Science Department, University of Auckland
E-mail: mc.wilson@auckland.ac.nz

networks is not always a straightforward task, as only a few legislation systems provide machine-readable documents or structured databases [13] [3]. The majority of legislation systems supply legal documents in human-readable format. For example the New Zealand Parliamentary Council Office provides machine-readable XML files [30] only for the current active Acts, which constitute around ten percent of the entire set of Acts. All other historic documents are scanned and supplied by a third party institute in Portable Document Format (PDF) [20].

To extract information from legal documents, the first step is the conversion of images into text if the text is not available. This concept is well studied as **optical character recognition (OCR)** [36], and there are several techniques and tools developed to convert typewritten or handwritten images to text. OCR is the first step of the proposed framework, and for the case study we selected ABBYY FineReader [27].

Information Extraction (IE) involves locating and extracting specific information from text [2]. Information Extraction assumes that in each individual text file there are one or more entities that are similar to those in other text documents but differing in the details [15].

IE approaches in the legal domain are considerably different from other knowledge areas because of the two main characteristics of legal texts. Legal documents exhibit a wide range of internal structure, and they often have a significant amount of manual editorial value added. One of the earliest information retrieval approaches for legal materials based on searching inside the document was proposed in 1978 [16]. Later works mainly used natural supervised learning techniques to retrieve the required data from legal texts, but with a substantial error [35] [22]. In the proposed framework of this study several IE tasks are used, and more are described later in this section.

Named entity recognition (NER) is one of the main sub-tasks of IE. The goal of this task is to find each occurrence of a named entity in the text [21]. Entities usually include people, locations, quantities, and organizations but also more specific entities such as the names of genes and proteins [10], the names of college courses [26], and drugs [24]. In the New Zealand legislation corpus, entities could be the name of legislative documents such as Acts, Regulations, Bills, Orders, or Case-Laws [34]. In the case study which is discussed section 3, the main required entities inside the text documents are is the names of the New Zealand Acts.

The main traditional NER algorithm that identifies and classifies the named entities is statistical sequence modeling [21]. But there are other modern approaches based on combinations of lists, rules, and supervised machine learning [9]. To extract the require information for Legislation Network, there are clear rules to identify the named entities, and the classification of the entities is not needed. Therefore the second NER approach is more appropriate and discussed further for the proposed framework.

The next IE task which is used in our study is to detect the relationships that exist among the recognized entities. This task is called **relation extraction (RE)** [21]. The earliest algorithm for relation extraction is the use of lexico-syntactic patterns [18]. This algorithm is still valid and widely used, but there are other algorithms introduced later such as supervised learning [21] and bootstrapping [7][6]. Considering that legislation texts are well structured, it is assumed that there is a large collection of previously annotated material that can define the rules for classifiers.

Approximate string matching techniques find items in a database when there may be a spelling mistake or other error in the keyword [17]. This is becoming a more relevant issue for fast-growing knowledge areas such as information retrieval [28]. Various techniques are studied to address the identity uncertainty of the objects, and briefly are reviewed in this study. These techniques could be distance based, token based, or a hybrid model of the distance and token based models.
Damerau-Levenshtein metrics are the main approximate string matching techniques to address the distance functions between the two strings [12] [25]. The most famous function in this category is *edit-distance*, and it is defined as the minimum number of changes required to convert one string into the other [25]. Several alternative distance functions to the edit-distance have been proposed such as q-gram and maximal matches [37].
The next set of techniques are token based or probabilistic object identification methods adapted for string matching tasks [11][31] [17]. *Jaccard similarity* and *cosine similarity* are common token based measures widely used in the information retrieval community[11]. Hybrid techniques combine distance-based and token-based string matching measures such as Jaro-Winkler [38]. All of the string matching algorithms have been developed by filtering and bit-parallelism approaches.
The fastest algorithms use a combination of filters to discard most of the text by focusing on the potential matches. Hybrid models significantly improve precision and recall reducing the error in a range between 0.1 to 0.2 [28].

Network inferences require high accuracy of data [5]. For this study various string matching techniques are examined for the Legislation Network and comparing the results, a hybrid model of *Jaccard similarity* and *edit-distance* is used as described in the next section.

The main contribution of this study is the proposed Information Extraction framework which engages several processes and enables the researcher to have access to the network information from historic documents. This framework makes it possible to study the Legislation Network as a dynamic graph. In this paper the case study covers all Acts in New Zealand legislation corpus including historic, expired, repealed and consolidated Acts as at end of September 2018. This comes to a set of 23870 PDF files of which about 87% are in scanned image format.
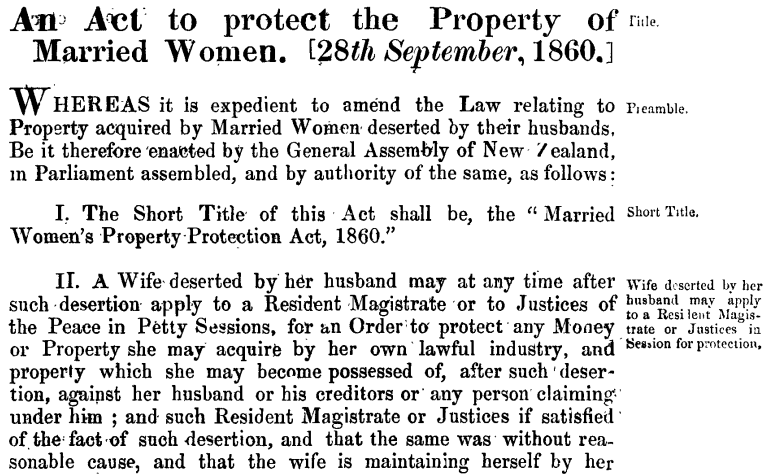
An Act to protect the Property of Title.
Married Women. [*28th September*, 1860.]

Whereas it is expedient to amend the Law relating to Preamble.
Property acquired by Married Women deserted by their husbands.
Be it therefore enacted by the General Assembly of New Zealand,
in Parliament assembled, and by authority of the same, as follows:

I. The Short Title of this Act shall be, the "Married Short Title.
Women's Property Protection Act, 1860."

II. A Wife deserted by her husband may at any time after Wife deserted by her
such desertion apply to a Resident Magistrate or to Justices of husband may apply
the Peace in Petty Sessions, for an Order to protect any Money to a Resident Magistrate or Justices in
or Property she may acquire by her own lawful industry, and Session for protection.
property which she may become possessed of, after such deser-
tion, against her husband or his creditors or any person claiming
under him ; and such Resident Magistrate or Justices if satisfied
of the fact of such desertion, and that the same was without rea-
sonable cause, and that the wife is maintaining herself by her

Fig. 1: Married Women Property Protection Act 1860

Figure 1 shows a sample image of an average quality scanned PDF document. The proposed framework suggests a high-performance procedure to derive network information from such poor quality documents. In the following sections, examples and the experimental results are used to illustrate the framework, its performance, and its potential applications.

In this section a summary of the required Information Extraction processes and methodologies is discussed. In the next section the proposed framework is presented and various examples are explained. Then the case study analysis and the application of the proposed framework are examined. Next, a number of experiments are designed and studied to evaluate the accuracy of the extracted information and to study the robustness of Legislation Network. The study finishes with a quick review on the novelty and the importance of discovering the time-varying behaviour of the Legislation Network.

2 The proposed Information Extraction framework

In this section, the Information Extraction framework to build the Legislation Network is discussed. Figure 2 depicts the overview of the proposed framework. The process starts with the conversion of non machine readable files to text by using *OCR* available tools. This step is relatively straightforward, but could be time-consuming considering the number of documents in the study. As mentioned earlier, in the case study the tool named ABBYY FineReader [27] is used. The average accuracy of this step is just above 80 percent and implies the need for a typos analysis step that is discussed in section 2.3.

2.1 Text Canonicalization

The next step in the proposed framework is *text canonicalization* [40]. There are several required tasks to convert all of the text files into a unique format, so the rules can be defined more easily while running the Information Extraction tasks. The text canonicalization step could be implemented via different approaches depending to the text style and language. In this paper, some of the common tasks are suggested, and two potentially required tasks are described.

In the case study the designed system transfers all letters to *Lowercase*. This transition applies a level of consistency across the text documents and the Information Extraction rules. In the experiments, the system also replaces *Special Characters* with generic tags in the text. The only character which is not replaced is the parenthesis, as it is often used in the title of legislation. The other suggested generic text canonicalization task is to replace multiple spaces with one space.
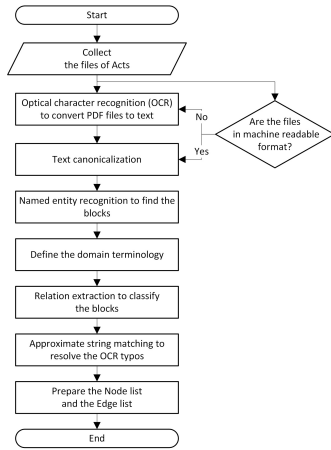
Apart from the general text canonicalization steps, there are other potential corrections that shape the text to make it a better input for the Information Extraction process. The first is to remove text margins that OCR mistakenly merges to the main body of text. Figure 1 includes examples of these margins that might impact the Information Extraction rules and result in error. As an example in the case study, the phrase *short title* is often used in the margin, and OCR merges it to the nearest part of the text. This might impact the named entity recognition task, so the system removes this phrase and many of its possible misspelled forms from the text. The next recommended text canonicalization step is to resolve the misspelling issues for the keywords used in the Information Extraction process. As an example in this study, Acts are the main entities, and the rule to recognize them uses the keyword *Act*. So the system corrects some of the possible misspelling forms of the word *-act-*. To provide a better explanation of this step, Table 1 provides an example

Fig. 2: Legal Text IE Framework

Table 1: OCR and Text canonicalization result comparison

OCR	I~ The Short Title' of this' Act shall be, the ((Married Short TItle,'Vomen's 'Property'Protectio£Aot, 1860."
Text canonicalization	I the short title of this act shall be the ((married vomens propertyprotectio act 1860

referring to third paragraph of the Figure 1 image. As can be seen the text canonicalization converts the text to a simpler unique structure prepares it for the next steps of *named entity recognition* and *relation extraction*. Typo resolution is not expected at this step, being covered under the last step via Approximate String Matching.

2.2 Named Entity Recognition and Relation Extraction

As explained, the text canonicalization step normalizes the text files to a unique format and prepares them for in-depth information extraction steps. To extract the network node information, a combined Named Entity Recognition approach is suggested which engages rules and supervised learning. To identify the rules, a sample set of the documents should be reviewed. The sample size is not necessarily large, but a stratified sampling approach is suggested to eliminate the impact of time-period style and author's writing style.

Table 2 shows examples of the entities in the case study based on the Acts recognition rules. Acts are the main part of the New Zealand legislation system and as explained before the case study only considers the Acts. In the case

Table 2: Entities, types and examples

Type	Tag	Sample	Canonicalized text
Year	YR	1860	the short title of this act shall be the ((married vomens propertyprotectio act 1860
Act	ACT	married vomens propertyprotectio act 1860	the short title of this act shall be the ((married vomens propertyprotectio act 1860

study stratified sampling method is used and the strata are five different time periods in a range of more than 200 years. A total number of 55 text files are reviewed, and several clear rules engaging a set of keywords and lists are built to identify the named entities. Table 3 provides examples of the rules in each stratum and y represents the year in which the Act is commenced.

Table 3: Examples of the Named Entity Recognition rules

Stratum	Keyword example	Rule example	Sample document
$y < 1850$	*ordinance*	an [keyword] to *any phrase* of [act name] [date]	Police Magistrates Act 1841
$1850 < y < 1900$	*short title, shall be*	the [keyword] of this act [keyword] the [act name] [year]	Customs Tariff Act 1873
$1900 < y < 1950$	*amend, consolidate*	an act to [keyword] *any phrase* of the [act name] [date]	Mining Act 1926
$1950 < y < 2000$	*meaning, section*	same [keyword] as in [keyword] [any number] of the [act name] [year]	Copyright Act 1962
$2000 < y$	act	this [keyword] is the [act name] [year]	Social Security Act 2018

Alongside recognizing the entities, to extract the network edge information, a rule based Relation Extraction approach is suggested considering that legislation texts are contextually structured. To identify the rules, this study suggests to use the same sample set which is used for the Named Entity Recognition. From the case study it is observed that the style of writing legislation has changed considerably over time, so the sampling approach is very important to minimize the impact of various text styles. By reviewing the sample files,

there is a large collection of previously annotated material that can define the rules for relation classifiers.

For the case study, as explained total number of 55 text files are reviewed, and several classifier rules engaging a set of keywords are built to identify the relations between the named entities. Table 4 summarizes the entity relation list for the case study and provides examples. This suggested process can be

Table 4: Relations example

Relation	Type	Canonicalized text	Sample document
Title	TIT	the short title of this act shall be the ((married vomen propertyprotectio act 1860	Married Women Property Protection Act 1860
Citation	CIT	within the meaning of section 5 of the companies act 1993	Trade Marks Act 2002
Amendment	AMD	section 25.1b amended, by section 5.2 of the trade marks amendment act 2005	Trade Marks Act 2002
Partial Repeal	PRP	section 5(1) repealed, by section 4(8) of the trade marks amendment act 2011	Trade Marks Act 2002
Repeal	FRP	acts repealed. 1860, No. 9.the married vyomens pfoperty protection act, 1860.	Married Women Property Protection Act 1880

generalized for any other case study in Legislation Network building process considering that legislation texts are coherently structured. So there is always a large collection of previously annotated material that can define the rules for entity recognition and relation classifiers.

2.3 Approximate String Matching

Named entity recognition identifies the Acts and relation extraction recognizes the relationship between them. So these two steps result in an initial version of the node list and the edge list of the intended Legislation Network. However testing this network shows that the extracted data is shoddy with an average error rate of 12 [1] percent, so another step is required to resolve typo issues and imperfect entities. This poor-quality data implies the need for the approximate string matching step. To run this step two main components are required, the technique and the correct pattern. Table 5 provides an example which shows the first match as the output of the proposed approximate string matching technique.

Table 5: Approximate string matching example

Technique	Extracted entity	First match
Hybrid Model	married vomens propertyprotectio act 1860	married women property protection act 1860

As mentioned earlier after the implementation of different approximate string matching techniques, a hybrid model of *Jaccard* and *Edit-Distance* is designed and proposed. Algorithm 1 shows the proposed hybrid model, and Figure 3 compares the results of the hybrid model with Edit-Distance and Jaccard

---

[1] To estimate this error rate, a cluster sampling method is used to randomly choose ten sets of 30 entities. By manual check of the samples, the rate of incorrectly matched entities is observed.
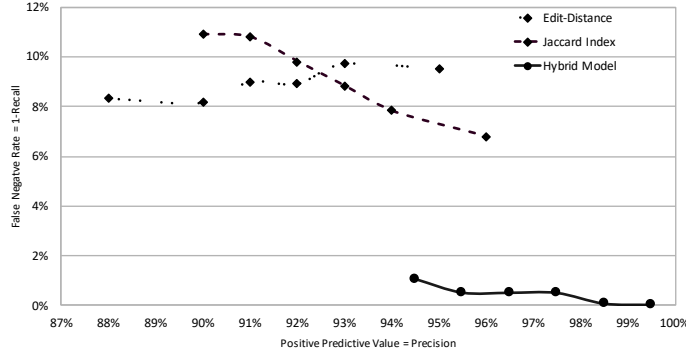
Fig. 3: Precision and Recall comparison of the approximate string matching techniques

techniques in terms of precision and recall of the approximate string matching step. To run this comparison, a stratified sampling technique is used with different time periods being the groups[2].

Algorithm 1 Approximate String Matching of Legislation

1: **procedure** Legislation Name Matching
2:     $string1 \leftarrow$ *Extracted legislation name*
3:     $masterlist \leftarrow$ Open *Legislation Title Master List*
4:     $j \leftarrow 1$
5:     $tline \leftarrow$ The first line of *masterlist*.
6:     $GetOut \leftarrow 0$
7:     **while** $< GetOut \neq 0 >$ **and** $< tline \neq 0 >$ **do**
8:
9:         $string2 \leftarrow tline$.
10:         $m(j) \leftarrow Jaccard(string1,\ string2)$
11:         $n(j) \leftarrow EditDistance(string1,\ string2)$
12:         **if** $< m(j) = 0.5 >$ **or** $< n(j) = 0 >$ **then**
13:             *GetOut*
14:         $j \leftarrow j + 1$.
15:         $tline \leftarrow$ The next line of *masterlist*
16:         **close**;
17:     $[x1, I1] \leftarrow max(m)$
18:     $[y1, I2] \leftarrow min(n)$
19:     **if** $y1$ *is smaller than or equal to 5* **then**
20:         $match \leftarrow I2$
21:     **else if** $x1$ *is bigger than 0* **then**
22:         $match \leftarrow I1$

The graph shows the error rates in each time sample of documents based on the chosen approximate string matching model. For example for the documents commenced prior to 1850, the first marker point at each graph line shows the false negative error and the precision of the chosen approximate string

---

[2] Time periods: before 1800, 1800-1850, 1850-1900, 1900-1950, 1950-2000, 2000-2018

matching method. As can be seen samples from this oldest groups of acts show a higher error rate regardless of the approximate string matching method, and Edit-Distance performs slightly better for the old documents comparing to the Jaccard index. In summary the proposed hybrid model performs significantly better than the other two methods for all documents regardless of their age with less than two percent of false-negative error and average precision of more than 98 percent.

In the case study the pattern which is used for the approximate string matching step is the list of all NZ Acts provided by NZLII [20]. In case of not having access to such a master list, the typo resolution could be more time consuming. Approximate string matching considerably improves the quality of the extracted information, result in reliable edge list and node list. Later in this study the evaluation of the final extracted data set and the robustness of the network is discussed. The robustness study proves the value of a high performing approximate string matching technique which improves the data quality significantly.

3 Application

The proposed Information Extraction framework resolves the historic data limitation in previous studies [34][33] and results in a large and reliable dynamic network data set which is called LegiNet and is available at [32]. This dynamic [8] and complex network has a very intersecting range of characteristics and behaviours. To maintain the subject consistency of this paper, more in-depth analysis of network behaviours are delayed to the future studies. In this section, generic network science characteristics of the case studied network are discussed, and an overall view of the structural and node importance evolution is presented.

Table 6 compares the produced network based on the Information Extraction process with the earlier versions of the network that was built with parsing of limited available XML resources. As illustrated the network size and structure is significantly changed comparing to its earlier versions. Figure 4 and Table 7

Table 6: NZ Legislation Network, this study versus previous studies

Network	Nodes	Edges	Average degree	Average CC[3]	Average path-length	Network type
This study	16385	137751	8.407	0.216	4.873	dynamic
Previous studies	3856	33884	8.878	0.39	3.569	one snapshot

capture the overall evolution of the Legislation Network in New Zealand from 1267 to the second quarter of 2018. To visualize the data, a network force-directed approach is used. In the layouts in Figure 4 each node is placed

---

[3] The average clustering coefficient (CC) is calculated based on the assumption that the network is directed using the approach that discussed in [34]
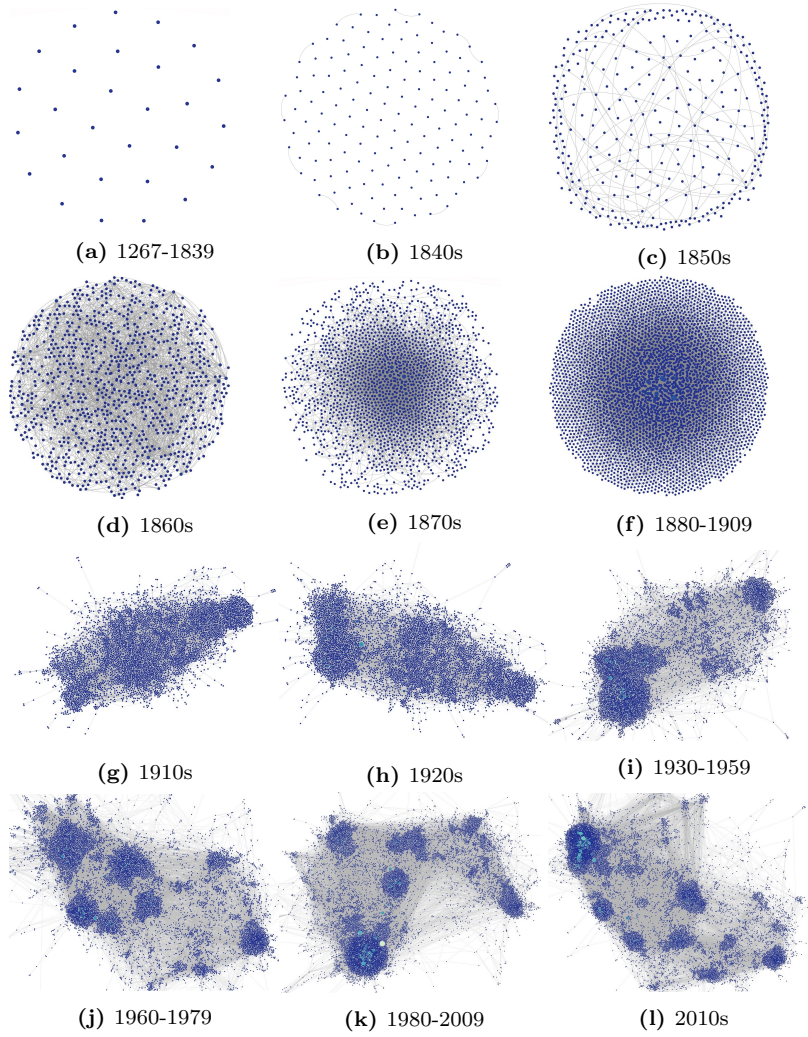
(a) 1267-1839   **(b)** 1840s   **(c)** 1850s

(d) 1860s   **(e)** 1870s   **(f)** 1880-1909

(g) 1910s   **(h)** 1920s   **(i)** 1930-1959

(j) 1960-1979   **(k)** 1980-2009   **(l)** 2010s

Fig. 4: Overview of the network structure evolution

depending on their connection to the other nodes. As can be seen references between the Acts first appear in the 1840s, but the data-set visually looks like a graph since the 1850s and it gets denser from the 1870s. As can be seen in Table 7 the graphs show some small-world properties from 1860s with $\sigma > 1$ and small-world property of the graphs is significant from 1970s comparing to 50 random graphs. As illustrated overlay the network gets denser and the average degree is growing. More significant clusters are observed during the

---

[4] The small-world sigma σ is calculated by comparing clustering coefficient and average path length of each network to 50 equivalent random network with same average degree as suggested by [19]

Table 7: Overview of the network measures evolution

Time	1267-1839	1840s	1850s	1860s	1870s	1880-1909	1910s	1920s	1930-1959	1960-1979	1980-2009	2010s
Number of nodes	28	148	315	939	1945	4712	5473	6292	8622	11940	15524	16199
Number of edges	0	12	64	1252	4756	14851	18767	24538	44859	70683	121019	130969
Average degree	0	0.081	0.203	1.333	2.445	3.152	3.429	3.900	5.203	5.920	7.796	8.085
Average path length	0	1	1.046	2.605	3.592	8.061	7.514	8.301	6.164	5.554	5.051	4.927
Directed CC	0	0	0.001	0.007	0.12	0.13	0.133	0.143	0.161	0.193	0.213	0.212
Small-world[4] σ	NA	0	0.447	1.165	15.587	75.173	82.519	80.122	24.239	16.131	195.837	208.084

most recent decades which can be seen in Figure 4. These clusters could be the outcome of housekeeping activities such as edge and node removal, or it could be the result of mature referencing approach in legal drafting process. Both of the above hypotheses should be examined in future studies.

Based on the network structure information provided in Table 7 and Figure 4, six different time periods are chosen for the centrality evolution analysis. Figure 5 captures the time evolution of the top 10 nodes and the most frequent words [5] in the top 20 nodes based on Katz prestige centrality measure.

As mentioned earlier, prior to the 1860s the graphs don't show significant small-world properties. The visual presentation in Figure 4a to Figure 4c also reflects that the network can be considered as a random graph during this period. So the Katz centrality degree distribution is nearly a uniform distribution in these time periods and is excluded from Figure 5.

Figure 5a shows the most important nodes with the impression that **Land** was the most important law subject back at that time. In the next selected time period the network shows small-world properties, and as can be seen in Figure 5b the centrality measure shows a higher kurtosis with the word **Council** being the most frequent topic in legal domain.

Similarly the other graphs reflect the change in the network structure and highlights the relationship between the laws and the socio-economic requirements of the country. In the current decade, with the new sets of legislation being introduced and referenced to the older documents, the centrality measure is increased comparing to the previous decade, and the hot legal topics show a change which could be a good reflection of the society's needs.

4 Evaluation and Robustness

In this section the performance of the proposed framework is discussed. As explained in the previous section, the main goal of the study is to extract the information to build Legislation Network. The framework includes Named Entity Recognition, Relation Extraction, and Approximate String Matching jointly to extract the network's node and edge information. In this section the proposed framework is evaluated and the related errors are calculated. The familiar metrics of recall and precision measures are used to evaluate the

---

[5] To find the frequent words, Textalyzer Python module is used. The frequent prepositions, conjunctions and articles are excluded from the analysis.
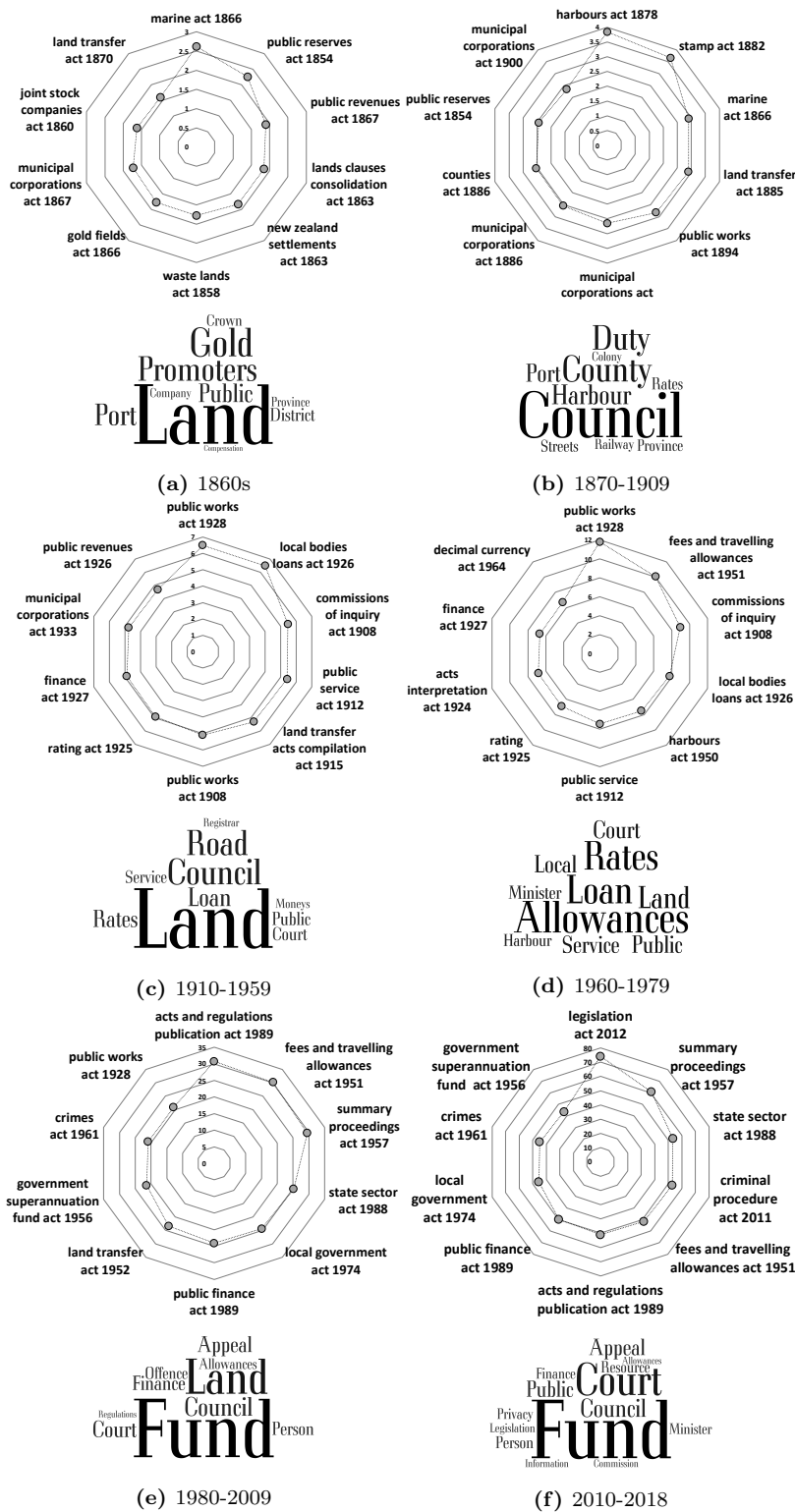
(a) 1860s

(b) 1870-1909

(c) 1910-1959

(d) 1960-1979

(e) 1980-2009

(f) 2010-2018

Fig. 5: Time evolution of the top ten legislation and the top subjects in the top twenty legislation

system. High **precision** means that the framework returns substantially more relevant results than irrelevant ones, while high **recall** means that the process returns most of the relevant results. At the end of this section the impact of the identified errors on the network structure is explained, and the robustness is assessed.

4.1 Error Estimation, Precision, and Recall

In the proposed Information Extraction process, Named Entity Recognition is combined with Approximate String Matching to recognize, validate and optimize the entities (the nodes and the edges). Figure 6 illustrates the occurrence of the false-positive and false-negative errors in this process and helps in studying the robustness of the network.
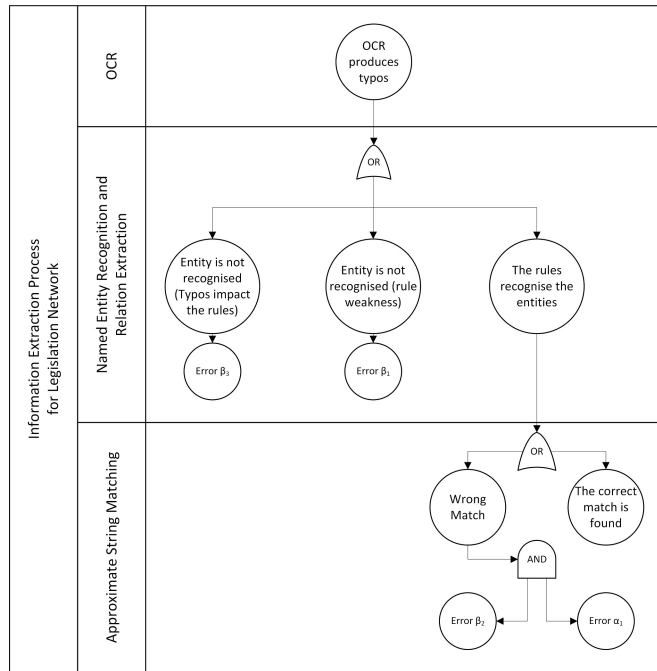

Fig. 6: Error Diagram

If the Entity Recognition process finds the entities, then there is a possibility that the Approximate String Matching process fails to find the correct match. A type I error α_1 occurs when the approximate string matching process fails to find the correct match. From one side this issue contributes to the false-positive error because it adds invalid entities to the output. These invalid entities impact the accuracy of the node list and the edge list of Legislation

Network. To estimate the $\bar{\alpha}_1$ in the case study, a cluster sampling method is used to randomly choose ten sets of 30 entities. By manual check of the samples, the rate of incorrectly matched entities is observed. A Kolmogorov-Smirnov test suggests that the estimated error $\bar{\alpha}_1$ has a normal distribution with the parameters in Table 8.

β_1 also occurs when approximate string matching system picks a wrong match for the entities. This issue can contribute to the false-negative error because those entities that are wrongly matched to other entities are missing from the data set. The estimation methodology and the estimated value of $\bar{\beta}_1$ are equal to that of $\bar{\alpha}_1$ as indicated in Table 8.

β_2 is measuring the Information Extraction rules' performance. If it fails to recognize entities, then those entities are missed, and it results in another type of false-negative error. The estimation process for $\bar{\beta}_2$ is different from the previous two errors, and it is harder to address. For the case study a sample set of 30 text files are randomly chosen using cluster sampling method. Then all of the extracted entities for each document is compared to the actual entities in a human involving process. The list of missing entities is categorized into two parts: caused by a typo, or caused by insufficient rules to recognize the entity. The rate of missing entities caused by weak or missing rules is calculated for each document and denoted by $\bar{\beta}_2$. The Kolmogorov-Smirnov results through all of the 30 documents show that $\bar{\beta}_2$ has a normal distribution with the parameters in Table 8.

Table 8: Errors, sensitivity and specificity

Measure	$\bar{\alpha}_1$	$\bar{\beta}_1$	$\bar{\beta}_2$	$\bar{\beta}_3$	$\bar{\alpha}$	$\bar{\beta}$
μ	0.0160	0.0160	0.0012	0.0007	0.0160	0.0179
σ	0.0012	0.0012	0.0001	0.0001	0.0012	0.0012

β_3 addresses the error when typos cause problem for recognizing the entities. The estimation process is very similar to that of $\bar{\beta}_2$. A sample of 30 text files are collected. Then the rate of missing entities caused by OCR typos is calculated for each document and addressed as $\bar{\beta}_3$. The Kolmogorov-Smirnov results through the selected documents show that $\bar{\beta}_3$ has a normal distribution with the parameters in Table 8. In the sample it is observed that the typos that cause the entity recognition failure are only numeric typos. For example OCR might produce an error and convert 1987 to l987 by misspelling number 1 to letter l. Then the Information Extraction rules are impacted to recognize l987 as a year, so the entity is missed.

As Figure 6 shows α_1 is the only false positive error which contributes to the **overall false positive error** of the system. Table 8 captures $\bar{\alpha}$, assuming that $\bar{\alpha}$ estimates the overall type one error. To estimate the **overall false negative error** of the system, β_1, β_2, and β_3 are considered as mutually exclusive events. From Figure 6 it is also clear that the intersections of each two of these errors are empty, so they are independent. Table 8 shows $\bar{\beta}$, the estimated value for the overall false negative, or the type II error.

To calculate the Precision and Recall, Equation 1 and Equation 2 are used.

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} = \frac{1 - \bar{\alpha} - \bar{\beta}}{1 - \bar{\beta}} \tag{1}$$

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative} = \frac{1 - \bar{\alpha} - \bar{\beta}}{1 - \bar{\alpha}} \tag{2}$$

Referring to the above equations and Table 8, the Precision of 98.37% and Recall of 98.18% are calculated. These outcomes suggest the high performance of the Proposed Information Extraction framework that results in a high data reliability of the output Legislation Network.

As explained in section 2.3, the proposed hybrid Approximate String Matching technique substantially reduces the errors. It is important to mention that at the earlier stages of the study by using the classic string matching techniques, the error rates were considerably higher, and the accuracy of the network was questionable. A time consuming examination process engaging manual checks was applied to propose the hybrid model which resulted in impressive performance and high precision and recall. The improvement obviously involved a lot of efforts and time, but resulted in accuracy and confidence in Legislation Network studies.

4.2 Robustness

With a coherent understanding of the errors, it is very important to study the robustness of the network to the error. The robustness study proves the importance of the data accuracy which supports the value of the proposed hybrid model for the approximate string matching. In this section to study the network robustness, diameter and three major centrality measures are used.

To understand the diameter robustness of the network, attack and failure analysis is required. As discussed earlier Legislation Network in general show scale-free characteristics. So it is expected to observe a reasonable error tolerance of the network as the result of random failures, but vulnerability as the result of attacks[1]. To study the network robustness to node failures the method is to randomly remove a fraction of nodes f and recalculate the diameter of the network d. To study the network robustness to attack by removing

a fraction f of the largest nodes [6] and observe the change to the the diameter d. The results of both failure and attack to the nodes are captured in Figure 7. The observed tolerance to failures and the vulnerability to attacks shows that the connectivity is provided by a few highly connected nodes, and majority of nodes have only few edges. As can be seen the vulnerability to the
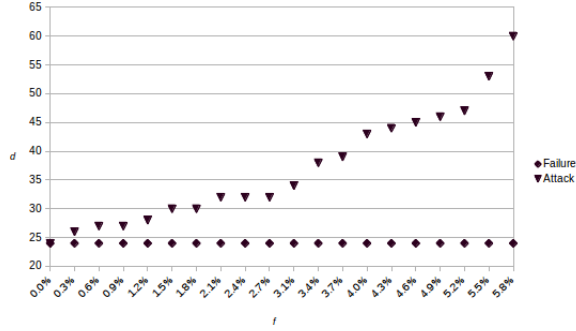

Fig. 7: Changes of the network diameter d as the function of fraction f

attacks starts immediately after removing a small fraction $f = 0.3\%$ of the highly connected nodes. This scenario of attack is highly unlikely in Legislation Network considering the high Precision and Recall of the proposed data extraction process.

As discussed in the previous studies [34] [33], the most relevant centrality measure for Legislation Network is the Katz second prestige measure. In recent studies, the reliability of different centrality measurements against network manipulation has been addressed [4] [29], but Katz prestige centrality is not much discussed. In this paper the Katz centrality, betweenness centrality, and degree centrality robustness of Legislation Network against edge deletion error is studied. To address the robustness, four major measures of accuracy that proposed in [4] and [29] are used. These measures are Top 1, Top 3, Top 10 percent, and the Pearson correlation to compare the centrality measures between the true network and the manipulated network.

The error level is considered as a specific percentage value from the set of 1%, 5%, 10%, 20%, that is relative to the number of manipulated edges from the original true network. Figure 8 shows the results of the different Centrality measure as the function of the fraction of manipulated edges f. For each fraction level, the test is repeated for 100 times, and the graphs show the average of the all sampled sets. Table 9 shows the Pearson correlation between the nodes centrality in manipulated network and the original network when 10% of the edges are randomly deleted.

---

[6] Based on their connectivity (total degree)

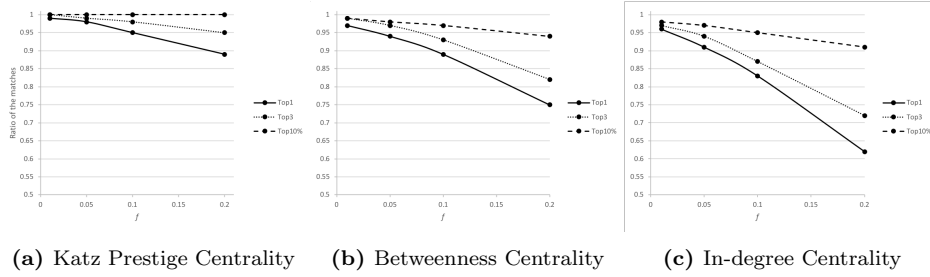**(a)** Katz Prestige Centrality **(b)** Betweenness Centrality **(c)** In-degree Centrality

Fig. 8: Robustness of the top nodes as the function of fraction of manipulated edges f

Table 9: Node centrality Pearson correlation between the manipulated network and the original network

Measure	Katz prestige centrality	Betweenness centrality	Degree centrality
Significance (p-value)	$2.2e^-16$	0.001	0.003
Correlation	0.939	0.948	0.67

The pattern and level of robustness of the three selected centrality measures considered in this paper are are not as similar as suggested in [4]. In-degree centrality shows more fragility comparing to betweenness and Katz measures. This difference could be related to the network topology as suggested by [29]. The results also confirm the findings of [29] [4] that accuracy declines monotonically with increasing error.

As can be seen in the graph, the Katz centrality is fairly robust to the edge deletion when less than 20 percent of the network structure is touched. The graphs indicate a moderate fragility when the network structure is hugely manipulated. For example the removal of 20 percent of the edges somehow impacts the in-degree centrality. However in more than 90 percent of these extreme samples, the top 1 node in the manipulated network is a member of top ten percent of nodes in the original graph. The results imply that centrality measures on Legislation Network are quite robust under small amounts of error (such as 5 percent or under) and to some extent fragile under bigger data errors. So the reliability of the network information is very important for in-depth network studies. As explained earlier the the precision and the recall of the proposed Information Extraction process is above 98 percent, so it is reasonable to compute the centrality measures when studying the Legislation Network.

5 Conclusion

This study focused on the **time** as a very important attribute in understanding and analyzing legislation. Legislation Network has been discussed in recent years, but the importance of having access to the historic legislation was never discussed much. This paper underlined the value of studying legislation as

dynamic networks, and proposed a new Information Extraction process to achieve a highly accurate Legislation Network. The performance of the data extraction framework is examined, is compared to the previous studies and proved to be considerably high. This work contributed to the literature of network Information Extraction from old documents, and insisted on the value and applications of the dynamic Legislation Network. The proposed process can be used not only in the legal domain but also in various research areas involving documented knowledge, facts, and cases.

Analyzing a dynamic Legislation Network is a novel approach to understand the underlying process behind the generation of the laws, and to study the behaviour, culture and growth of societies. This subject is very interesting, but mathematically complicated. So it will be discussed in a separate study.

References

1. Réka Albert, Hawoong Jeong, and Albert-László Barabási. Error and attack tolerance of complex networks. *nature*, 406(6794):378, 2000.
2. Peggy M Andersen, Philip J Hayes, Alison K Huettner, Linda M Schmandt, Irene B Nirenburg, and Steven P Weinstein. Automatic extraction of facts from press releases to generate news stories. In *Proceedings of the third conference on Applied natural language processing*, pages 170–177. Association for Computational Linguistics, 1992.
3. Alexander Boer, R Hoekstra, E De Maat, Fabio Vitali, Monica Palmirani, and B Ratai. Metalex (open xml interchange format for legal and legislative resources). *Management Center, Akon*, 2010.
4. Stephen P Borgatti, Kathleen M Carley, and David Krackhardt. On the robustness of centrality measures under conditions of imperfect data. *Social networks*, 28(2):124–136, 2006.
5. Carter T Butts. Network inference, error, and informant (in) accuracy: a bayesian approach. *social networks*, 25(2):103–140, 2003.
6. Sander Canisius and Caroline Sporleder. Bootstrapping information extraction from field books. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2007.
7. Andrew Carlson and Charles Schafer. Bootstrapping information extraction from semi-structured web pages. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 195–210. Springer, 2008.
8. Arnaud Casteigts, Paola Flocchini, Walter Quattrociocchi, and Nicola Santoro. Time-varying graphs and dynamic networks. *International Journal of Parallel, Emergent and Distributed Systems*, 27(5):387–408, 2012.
9. Laura Chiticariu, Yunyao Li, and Frederick R Reiss. Rule-based information extraction is dead! long live rule-based information extraction systems! In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 827–832, 2013.
10. Kevin Bretonnel Cohen and Dina Demner-Fushman. *Biomedical natural language processing*, volume 11. John Benjamins Publishing Company, 2014.
11. William Cohen, Pradeep Ravikumar, and Stephen Fienberg. A comparison of string metrics for matching names and records. In *Kdd workshop on data cleaning and object consolidation*, volume 3, pages 73–78, 2003.
12. Fred J Damerau. A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3):171–176, 1964.
13. EUR-Lex. Access to european union law. eur-lex.europa.eu/homepage.html, Accessed: 2017-09-10.

14. James H Fowler, Timothy R Johnson, James F Spriggs, Sangick Jeon, and Paul J Wahlbeck. Network analysis and the law: Measuring the legal importance of precedents at the us supreme court. *Political Analysis*, 15(3):324–346, 2007.
15. Dayne Freitag. Machine learning for information extraction in informal domains. *Machine learning*, 39(2-3):169–202, 2000.
16. Carole Diane Hafner. An information retrieval system based on a computer model of legal knowledge. *Ann Arbor, MI*, 1978.
17. Patrick AV Hall and Geoff R Dowling. Approximate string matching. *ACM computing surveys (CSUR)*, 12(4):381–402, 1980.
18. Marti A Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics-Volume 2*, pages 539–545. Association for Computational Linguistics, 1992.
19. Mark D Humphries and Kevin Gurney. Network small-world-ness: a quantitative method for determining canonical network equivalence. *PloS one*, 3(4):e0002051, 2008.
20. New Zealand Legal Information Institute. Free access to legal information in new zealand. www.nzlii.org, Accessed: 2018-10-31.
21. Dan Jurafsky and James H Martin. *Speech and language processing*, volume 3. Pearson London, 2014.
22. Uri Kartoun. Text nailing: an efficient human-in-the-loop text-processing method. *interactions*, 24(6):44–49, 2017.
23. Marios Koniaris, Ioannis Anagnostopoulos, and Yannis Vassiliou. Network analysis in the legal domain: A complex model for european union legal sources. *Journal of Complex Networks*, 6(2):243–268, 2017.
24. Martin Krallinger, Florian Leitner, Obdulia Rabal, Miguel Vazquez, Julen Oyarzabal, and Alfonso Valencia. Overview of the chemical compound and drug name recognition (chemdner) task. In *BioCreative challenge evaluation workshop*, volume 2, page 2, 2013.
25. Vladimir I Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710, 1966.
26. Andrew McCallum. Information extraction: Distilling structured data from unstructured text. *Queue*, 3(9):4, 2005.
27. Edward Mendelson. Abbyy finereader professional 9.0. *PC Magazine*, 2008.
28. Gonzalo Navarro. A guided tour to approximate string matching. *ACM computing surveys (CSUR)*, 33(1):31–88, 2001.
29. Qikai Niu, An Zeng, Ying Fan, and Zengru Di. Robustness of centrality measures against network manipulation. *Physica A: Statistical Mechanics and its Applications*, 438:124–131, 2015.
30. New Zealand Parliamentary Counsel Office. The authoritative source of new zealand legislation. www.legislation.govt.nz, Accessed: 2018-10-31.
31. Hanna Pasula, Bhaskara Marthi, Brian Milch, Stuart J Russell, and Ilya Shpitser. Identity uncertainty and citation matching. In *Advances in neural information processing systems*, pages 1425–1432, 2003.
32. Neda Sakhaee. Leginet new zealand, first outcome of the new information extraction framework proposed to build legislation network. 10.7910/dvn/ib3qsf, Published: 2018-09-21.
33. Neda Sakhaee, Mark Wilson, Shaun Hendy, and Golbon Zakeri. Network analysis of new zealand legislation. *NZ Law Journal*, 10(2017), 2017.
34. Neda Sakhaee, Mark C Wilson, and Golbon Zakeri. New zealand legislation network. In *Legal Knowledge and Information Systems: JURIX 2016: The Twenty-Ninth Annual Conference*, volume 294, page 199. IOS Press, 2016.
35. Cheng Tin Tin, Leonard Cua Jeffrey, Davies Tan Mark, Gerard Yao Kenneth, and EditaRoxas Rachel. Information extraction from legal documents. In *2009 Eighth International Symposium on Natural Language Processing*, 2009.
36. Oivind Due Trier, Anil K Jain, Torfinn Taxt, et al. Feature extraction methods for character recognition-a survey. *Pattern recognition*, 29(4):641–662, 1996.
37. Esko Ukkonen. Approximate string-matching with q-grams and maximal matches. *Theoretical computer science*, 92(1):191–211, 1992.
38. William E Winkler. The state of record linkage and current research problems. In *Statistical Research Division, US Census Bureau*. Citeseer, 1999.

39. Paul Zhang and Lavanya Koppaka. Semantics-based legal citation network. In *Proceedings of the 11th international conference on Artificial intelligence and law*, pages 123–130. ACM, 2007.
40. Yitao Zhang and Jon Patrick. Paraphrase identification by text canonicalization. In *Proceedings of the Australasian Language Technology Workshop 2005*, pages 160–166, 2005.