From Case Law to Ratio Decidendi

student: Josef Valvoda
supervisor: Dr Oliver Ray

15th September 2017

Executive Summary

This aim of this project is to apply recent advances in information mining from legal documents to automatically identify legally binding principles, known as ratio decidendi or just ratio, from transcripts of court judgements, also known as case law or just cases. To realise this aim we present a novel system for automatically extracting ratio from cases using a combination of natural language processing (NLP) and machine learning (ML). Our method differs from previous work by looking for ratio in the text of the cited paragraphs (in the given case) as opposed to the text of the citing paragraphs (in a subsequent case). Our main objective is to investigate our hypothesis that ratio can be reliably obtained by identifying the statements of legal principles in paragraphs that are cited by subsequent cases.

The motivation to identify ratio in particular stems from the legal doctrine of *stare decisis*. Under this doctrine, a legal case forms a precedent. The precedent requires that ratio is applied on subsequent cases with similar facts to decide their outcome. In current legal practice, identifying ratio usually involves manually reading through a number of related cases. In addition to analysing the text of a case of interest it is often necessary to consult several cited and citing cases along with many other cases related by topic. Since a single case can spread over 30 pages, this is an extremely time consuming process for human to perform.

We deliver on our objective, testing the hypothesis the ratio is in paragraphs that are cited by subsequent cases, by doing the following. First, we conduct our own independent, small-scale manual annotation study which reveals that our hypothesis, shifting the focus from citing to cited paragraphs, substantially increases reliability of finding the ratio, which we identify using the legally sound Wambaug's test. Second, we automate the identification of cited paragraphs by building upon the previous work of Shulayeva et al. [1] in the automatic detection of legal principles employing ML. Third, we automate cited paragraph identification by adapting the research in cross reference identification by Adedjouma et al. [2] using NLP. Finally, by combining the two classifiers above, we present a fully automated system that successfully identifies the ratio with an accuracy of 72%.

The project is 75% type I (software development) and 25% type II (investigatory). The software development component consists of implementing two classifiers capable of automatically identifying principles and cited paragraphs in case law, further combining them to identify ratio. The investigatory component consists of an independent small-scale annotation study investigating our hypothesis, developing new features and training corpus for principle identification and adapting cross reference identification for cited paragraph identification.

Our main contributions are as follows:

- We conducted an independent small-scale annotation study revealing that our hypothesis substantially increases reliability of finding the ratio.
- We have improved upon Shulayeva et al.'s research on principle identification in case law by 11%, setting a benchmark for this task at 96% accuracy, using ML.
- We have set a benchmark for a new task of cited paragraph identification at 94% accuracy by adapting Adedjouma et al.'s research, using NLP.
- We contribute towards proving our hypothesis by combining the above classifiers to identify ratio in case law, setting a benchmark for this task with 72% accuracy.

Acknowledgements

I would like to thank my supervisor, Dr Oliver Ray, for his critique that guided me through this project, Anastasia Vrublevska for bringing a note of levity into my life at times that seemed difficult, family and friends for their constant moral support and Shulayeva et al. for providing me with their case law corpus.

Contents

1 Introduction

This Section introduces the problem we are trying to solve and our approach to solve it. We outline our aims and objectives, list the main deliverables and the added value of our work. Finally, we elaborate on the scope of the project and the structure of the argument presented in this document.

1.1 Aim

In common law *ratio decidendi*, or just *ratio* [3], are the key principles used to decide the outcome of a legal case and the doctrine of *stare decisis*, or simply *precedent* [4], requires that ratio is applied on subsequent cases with similar facts to decide their outcome. Thus the identification of ratio is crucial to the work of lawyers and judges in common law countries.

In current legal practice, identifying ratio usually involves manually reading through a number of related cases [5]. In addition to analysing the text of a case of interest it is often necessary to consult several cited and citing cases along with many other cases related by topic. Although legal search engines, such as Westlaw[1], are typically used to find the related cases relatively efficiently, there is currently a lack of technological support for the task of actually identifying the ratio contained within the transcripts. Since a single case can consist of more than 150 paragraphs spread over 30 pages of text, this is an extremely time consuming and error-prone process for humans to perform. Automating this process would be of great value to legal practitioners.

The aim of this project is to automatically identify the ratio by applying recent advances in ML and NLP. To do this it is necessary to address two issues. The first issue is to distinguish statements of legal principles from discussion of specific facts to which those principles are applied in a particular case. The second issue is to determine which of those principles are pivotal to the outcome of the case (and therefore constitute the legally binding ratio), from those principles which are merely incidental and which are formally known as *obiter* [3].

1.2 Objectives

We hypothesise that the ratio of a given case can be reliably obtained by automatically identifying the statements of legal principles in paragraphs that are cited by subsequent cases. We also hypothesise, that NLP and ML techniques, namely in research of Adedjouma et al. and Shulayeva et al. could be applied to automate this objective [2, 1].

The following list breaks down our core objectives:

- Compare the correlation between principles in cited and citing paragraphs to ratio. To do this we have manually identified ratio using the legally sound Wambaugh's test, see Section 2.2. Then we manually annotated principles in cited paragraphs and citing paragraphs, see Section 4.4. Finally we compare the correlation between principles in cited and citing paragraphs with ratio.

- Improve over Shulayeva et. al.'s ML model performance in identifying principles in case law. We have done this by testing new sentential features, achieving only a small improvement, and repurposing their data corpus for the task of principle identification, achieving a large improvement. See Section 4.5.

[1] Westlaw UK, Online legal research from Sweet & Maxwell, http://westlaw.co.uk

- Adapt Adedjouma et al.'s NLP work for cited paragraph identification. We do this by creating our own corpus of citing cases, the Cited corpus, and apply a schema, automated by regular expression, to find cited paragraphs of a case of interest. See Section 4.6.

- Automate ratio identification by combining the above classifiers. Using our ML principle classifier and NLP cited paragraph classifier, we automatically identify ratio.

1.3 Deliverables

The main deliverables are as follows:

- An empirical proof of correlation between principles in cited paragraphs and ratio in a form of a manual annotation study. We demonstrate principles in cited paragraphs identify ratio with 76% accuracy, a theoretical ceiling of our method.

- A ML model identifying principles in case law with high accuracy. We present a ML classifier identifying principles with 96% accuracy, and 11% improvement over previous work by Shulayeva et al. [1].

- A program capable of identifying cited paragraphs from subsequent case law using regular expression and legal text schema with high accuracy. We present a NLP classifier identifying cited paragraphs with 94% accuracy, setting a benchmark for this novel task.

- A program capable of automatically identifying ratio from case law. Our program is capable of automatically identifying ratio with 72% accuracy, annotating cases in HTML format.

1.4 Added value

This paper presents a novel system for automatically extracting ratio from cases using a combination of NLP and ML to solve the two issues noted in Section 1.1 above. To achieve this we essentially build upon and integrate the recent work of Shulayeva et al. on the detection of principles and facts in case law using ML [1] and the earlier work of Adedjouma et al. on cross reference resolution in academic papers using NLP [2]. We improve Shulayeva et al.'s ML model performance on principle identification by 11% by finding five new informative sentential features and repurposing their training corpus to focus on principles. This sets a new benchmark of 96% for this task. We adapt Adedjouma et al.'s research for cited paragraph identification, achieving comparable precision to their original task (cross reference identification and resolution) of 98.7% and setting a new benchmark for this task at 94% accuracy.

It should be noted from the outset that although many aspects of our work may appear to be closely based on Shulayeva et al. the true contribution of out work actually stems from several fundamental differences. Whereas Shulayeva et al. specifically emphasise that they are *not* attempting to identify ratio, we specifically emphasise that we *are*. Whereas Shulayeva et al. are primarily concerned with the distinction between *facts* and *principles*, we are primarily concerned with the distinction between (principles that are) *orbiter* and (and principles that are) *ratio*. Whereas the method of Shulayeva et al. is mainly concerned with text in *citing* paragraphs, our method is mainly concerned with text in *cited* paragraphs.

To demonstrate the effectiveness of our approach, we conducted our own independent small-scale annotation study which supports our hypothesis that the ratio of a given case can be reliably

obtained by identifying the statements of legal principles in paragraphs that are cited by subsequent cases. In fact, our investigation shows that the principles in cited paragraphs correspond to ratio (as manually determined by a human expert using Wambaugh's Inversion test [3]) with a precision of 47% whereas the principles extracted by Shulayeva et al. from the citing paragraphs correspond to ratio with a precision of only 27%.

Finally, the novel research above is combined to identify ratio in case law with 72% accuracy. To the best of the authors knowledge this is also the only successful approach in identifying ratio according to a legally sound definition, creating a benchmark in the area and opening up new roads for approaching this difficult problem.

1.5 Research scope

The focus of the research in this project falls between two areas of computer science; natural language processing and machine learning. Under NLP, the cross reference analysis research is used to identify paragraph citations and link them to their targets, to identify cited paragraphs of a case of interest. We primarily draw from the research of Adedjouma et al. for this task [2]. With regards to ML, Shulayeva et. al's research on facts and principles classification is adapted to identify principles in case law [1].

1.6 Thesis structure

The core argument of this document is contained in Sections 2 to 4. In Section 2 we present legal research background later used to reason there is a need for identifying ratio in case law and that this need is not addressed by current legal research technology. We also justify the legal validity and objectivity of the definition of ratio we are employing and introduce the key distinction, between citing and cited paragraph in case law. Finally, we explain our choice of Westlaw[2] as the provider of cases for the Cited corpus developed by this project.

Section 3 presents legal information mining background required to later identify there is no legally sound scientific research addressing the gap in legal research tools. We introduce the research on principle identification and different sentential features used by ML algorithms used for identifying rhetorical roles of sentences in case law. We also introduce the research in cross reference resolution. These two research areas later form the basis of our new classifier.

In Section 4 we argue that there is a need for identifying ratio in case law and this need is not addressed by current legal research technology or scientific research. We define the task, in light of a legally sound definition, as identifying principles and distinguishing the pivotal ones from the rest to identify the ratio. We hypothesise that this task can be automated by identifying principles, but distinguishing them by focusing only on those in cited paragraphs. We engage with Shulayeva et al.'s research to demonstrate the novelty of our approach and prove validity of our hypothesis by conducting an empirical study. Drawing from the scientific research on principle identification and cross reference resolution, we envision new ways of improving the former and refocus the latter on the task of identifying cited paragraphs in case law.

In Section 5 we implement our methodology. In Section 6 we present and evaluate our results, report on the limitations of our work and propose future work. In Section 7 we conclude the paper.

[2]Westlaw legal search engine - http://www.westlaw.co.uk

1.7 Guide to corpuses

This project deals with several legal corpuses. To prevent any confusion, we list each corpus below, with a little description, as a quick point of reference.

- Golden Standard corpus - this is the corpus Shulayeva et al. [1] use in their research, and we recreate in Section 5.1.1, from the data provided by them. It consists of 50 common law taken from the British and Irish Legal Institute (BAILII) website in RTF format [1].
- New corpus - this is the Golden Standard corpus relabelled by us to address the issues Shulayeva et al. [1] report in their error analysis. We use it to improve over Shulayeva et al.'s performance in identifying principles in case law.
- Cited corpus - this is a corpus of 41 Westlaw cases collected and annotated by us, used to measure the accuracy of our cited paragraph identifier on the task of finding cited paragraphs in Stack v Dowden.

2 Legal research background

In this Section we briefly introduce the common law legal system with its inner workings, focusing in particular on the definition of ratio decidendi and distinction of citing and cited paragraphs. We also provide an overview of the current tools available for legal research. The aim is to identify two important hurdles of legal research and argue the latter has not yet been solved or even addressed by the contemporary legal research tools. Finally, we will elaborate on the reasons for using Westlaw for sourcing the case law for this project.

2.1 Common law

Different States do use different kinds of legal systems. From a short inspection of the index of United Nations Member states and Corresponding Legal Systems[3], we can glean that the five major legal systems used throughout the world are: Civil law, Common law, Customary law, Muslim law and Jewish law. For this project, we will be focusing on the Common-law jurisdiction. Among the countries using Common law are Canada, United States of America, Australia and of course United Kingdom. For this project, we will focus on the law of England, Scotland and Northern Wales in particular, henceforth referred to as English law or simply the law.

In Common law jurisdictions, the law is created by the judges, courts and tribunals in decisions they make when deciding an individual case. The decision made serves two purposes. Firstly, to decide the matter at hand. Secondly, to set the precedent for future cases. This is known as the doctrine of *stare decisis* [6] or *precedent*. It is this precedential effect of the decision that distinguishes the common law from some other types of jurisdictions, by placing the precedent on equal footing with law contained within Statutes and Regulations [6]. The bodies deciding the cases are therefore a part of law creation as much as legislative and executive branches of the government. Transcripts of decisions, also known as *judgements*, are recorded and kept for future reference as *case law*.

Legal professionals cite the case law to support their arguments. According to the doctrine of *stare decisis*, cases turning on similar facts should be decided according to the same principles. However, within one judgement a judge does not simply deliver only these general principles. Instead the judgement contains both the binding law principles, known as *ratio decidendi*, or *ratio*, but also the throwaway commentary, with only persuasive power, known as *obiter dicta*, or *obiter*. Branting [7], investigated identification of the ratio, and Plug [8] identification of the obiter.

A legal professional therefore needs to read through the related case law to be able to determine what the ratio is and to subsequently build a legal argument from it. This entails a creation of a mental model of the current law, lucidly described by Zhang et al. as *"[going] through a repetitive mental process of forward and backward searching in the imaginary space of legal issues embodied mainly by previous cases"* [5]. This has been dubbed, *"gathering citations"* [9], *"chaining"* [10] and *"footnote chasing and citation searching"* [11], and is the reason why this project is focused on identifying ratio, to aid what Zhang et al. [12] called the *"exhaustive shepardizing"*. Only at the end of this process the legal professional knows what the law is.

[3]Alphabetical Index of the 192 United Nations Member States and Corresponding Legal Systems - http://www.juriglobe.ca/eng/syst-onu/index-alpha.php

A non obvious specificity of law citations, is that they are semantically multi-dimensional [5]. This means that one case can cite other case on several occasions and each time it can be for a different reason. As Zhang et al. puts it: *"Two citations pointing to the same case may not necessarily be semantically related because they may each be based on a different legal issue"*. This is why resolving case citations superficially on the level of case is not enough and one of the reasons we focus on what is specifically inside the case rather than what the case is about overall.

Given the sheer amount of text that needs to be subjected to this analysis, it's not surprising that legal professionals do spend up to 31% of their time researching what the law is instead of applying it [13]. The two tasks a legal professional faces during that time can be simplified as: One, to find the cases that contain the law, by looking for related cases on facts. Two, to find the ratio within these cases. As we will see below, the current legal tools are equipped for solving the first task, while offer very little help in solving the second.

2.2 Defining ratio via Wambaugh's inversion test

Before evaluating current legal research technology, it is important to establish clearly what ratio actually is. There are many different interpretations [4, 14]. Branting's research elegantly summarises the differing points of view on the matter [3]. He identifies two general areas of focus. One is on the material facts, the other on the deciding principles. These further translate into five different viewpoints [3]:

1. *The ratio decidendi of a precedent consists of propositions of law explicit or implicit in the opinion that are necessary to the decision.*
2. *A unique proposition of law necessary to a decision can seldom be determined. Instead a gradation of propositions ranging in abstraction from the specic facts of the case to abstract rules can satisfy this condition.*
3. *The ratio decidendi of a precedent must be grounded in the specific facts of the case.*
4. *The ratio decidendi of a precedent includes not only the precedents material facts and decisions but also the theory under which the material facts lead to the decision.*
5. *Subsequent decisions can limit, extend, overturn earlier precedents.*

In this paper we have choose definition number one from above, since, unlike the other four, it can be objectively tested for using Wambaugh's *inversion* test [3]. Under this test, a principle in the case is inverted in its meaning; and if such inversion would affect the outcome of the case, then the principle is deemed a ratio. On the other hand, if the inversion would not affect the outcome, the principle is deemed an obiter. Currently this test is one of the popular ways of determining the ratio by legal practitioners, and lends our research a solid grounding.

2.3 Legal research tools

Figure 1: *Westlaw search index: http://legalresearch.westlaw.co.uk*

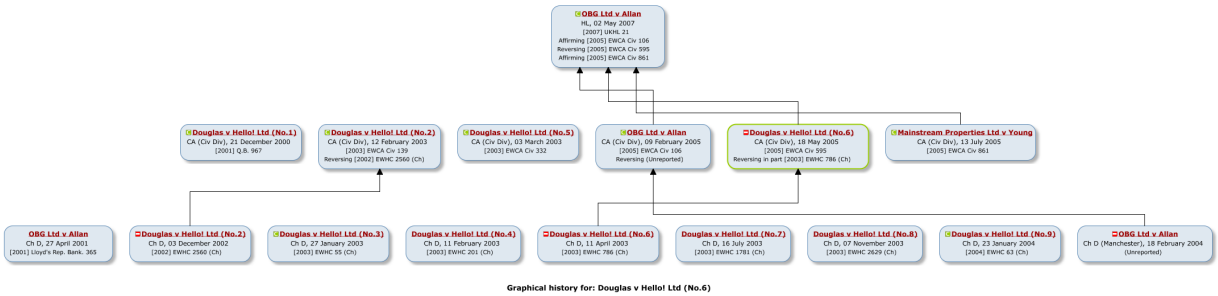

Figure 2: *Westlaw search index's graphical history tool.*

Shepardization is the classification of all the cases that cite the case which is being *shepardized*, based on their legal treatment of it. This information guides a lawyer towards cases which are approved by the judiciary as a source of a valid law. Professional services such as Westlaw[4], see Figure 1, and LexisNexis[5], see Figure 3, further classify cases into topics, so that it is easy to find the leading case of a specific area of law. These services also provide a summary of each case, insights into legal areas and briefs on current awareness. On top of it, Westlaw allows the user to browse a graphical history of a case, see Figure 2. Legal professionals are also trained to know which cases are important in each area of law they advise on, which makes, together with the wealth of tools above, solving the first task of where the law is on the level of case index comparatively as easy as searching Google[6] for a website.

[4]Westlaw legal search engine - http://www.westlaw.co.uk
[5]LexisNexis legal search engine - https://www.lexisnexis.com
[6]Google search engine - https://www.google.com

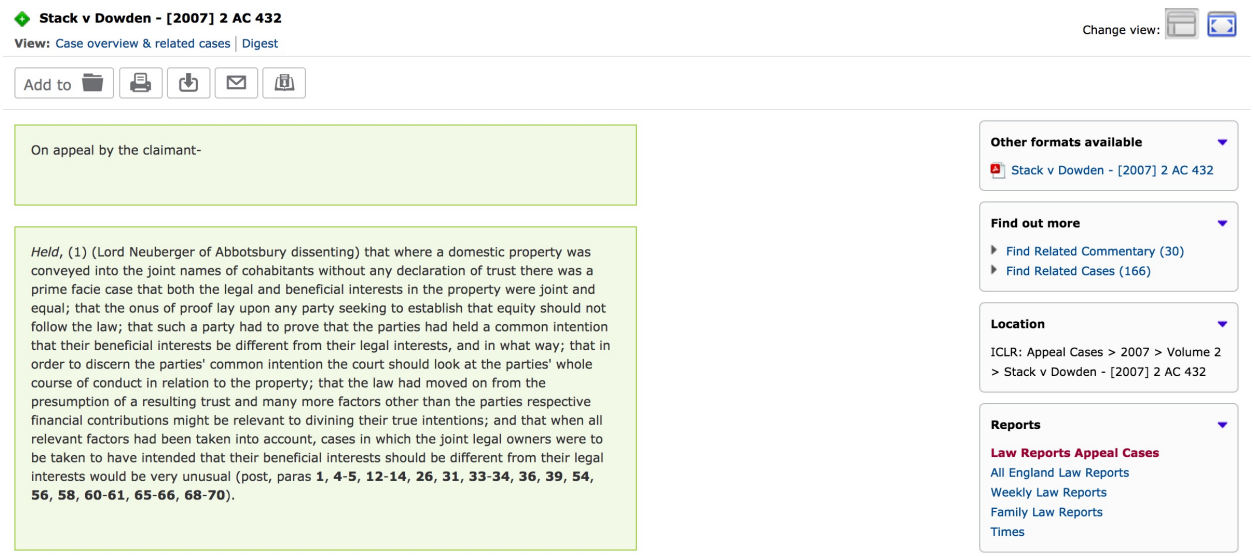

Figure 3: *LexisNexis search index, highlighting of the summary: https://www.lexisnexis.com/en-us/home.page*

While this treatment is often sufficient when searching for a website, and helpful for legal and scientific searches, unlike a website a case is a long and highly specific piece of text, which does not at a glance make it easy for you to find what you came looking for. To solve the second task, of identifying ration in the judgement itself, there is nothing more effective at the moment than reading the judgement, and manually searching through using the Wambaug's inversion test.

One improvement in the area is LexisNexis's highlighting of the summary, see Fig. 3, which, although visually separates the summarised outcome of the case from the rest of the case, is of course is only a very small step towards the goal of this project.

Because an importance of a principle in the overall argument might be highlighted only by a subsequent case, the time the task of ratio identification takes can easily multiply. A prudent lawyer will investigate what the subsequent judgements thought were the important principles in the case of interest, by looking which paragraphs were later cited and for what reason, and only then she can conclude if a principle is indeed a ratio or an obiter. This involves reading tens of cases, each easily stretching more than 30 pages, just to understand what the law is in one.

While Westlaw and LexisNexis provide a list of cases citing the case of interest, they do not show which paragraphs have been subsequently cited and apart from reading the subsequent judgements, there is no other way of obtaining this information. The second problem can therefore be broken down into two subtasks. One, to quickly identify the principles. Two, to discriminate between ratio and obiter. Neither of these is addressed by the current legal indexes.

Apart from LexisNexis and Westlaw, there are two other tools which are seldom used by lawyers, but we will mention them here for the sake of completeness. They are the British and Irish Legal Information Institute (BAILII)[7], see Figure 4, and JustCite[8], see Figure 5. BAILII only offers the

[7]BAILII legal search engine - http://www.bailii.org
[8]JustCite legal search engine - https://www.justcite.com

plain text of the case itself without further guidance and does provide free but smaller index compared to the services above. JustCite only provides the Precedent Map, a visual representation of how are cases related and does not actually provide any case law. Neither provides the summaries or insights, and therefore do not represent the wealth of information their more popular counterparts provide for solving the first task of identifying important cases. Needless to say, neither does anything to help the navigation within a case or in any way resolve the second task of ratio identification.

This exhausts the tools for legal research currently available. While the paid professional tools have carried Shepard's vision, and provide enough information to guide legal professional to select the important cases. She must figure out herself the non-trivial task of what the law actually is.

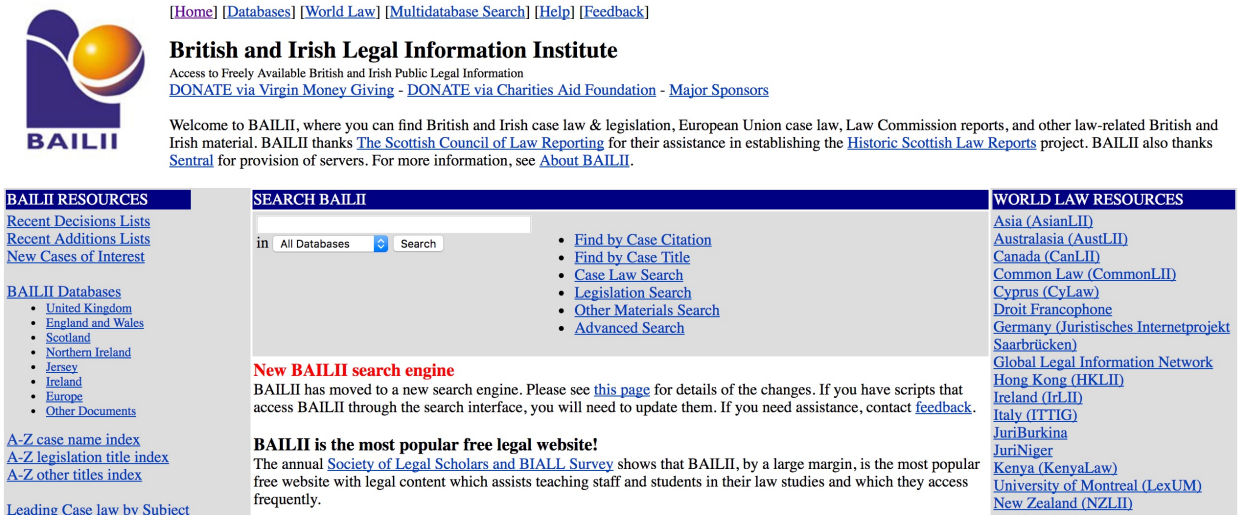

Figure 4: *BAILII, http://www.bailii.org*

2.4 Access to the case law

In our project we use two legal corpuses. The first one was graciously provided by Shulayeva et al., the Golden Standard corpus, and serves the purpose of direct comparison to their work. Then there is a second corpus, the Cited corpus, which was collected by us. This subsection applies to the latter corpus.

A hurdle for this project was to choose which of the resources should be used to form our legal corpus. Thanks to the library resources of the University of Bristol, both the freely available and the paid for options were possible candidates. Since the intended product of this research is a tool helping legal professionals, it would be preferable if the resource would cover as much legal knowledge as possible, in which respect paid services provided by Westlaw and LexisNexis far outweigh the free database of BAILII. Since JustCite does not provide the actual case content it would not be a good resource and so it was not further considered.

Westlaw, LexisNexis and BAILII were further evaluated based on their copyright policies. An ideal resource would allow direct access for the program to download searched case and the cases it cites. Unfortunately, all three resources do not allow automatic collection of data in their database

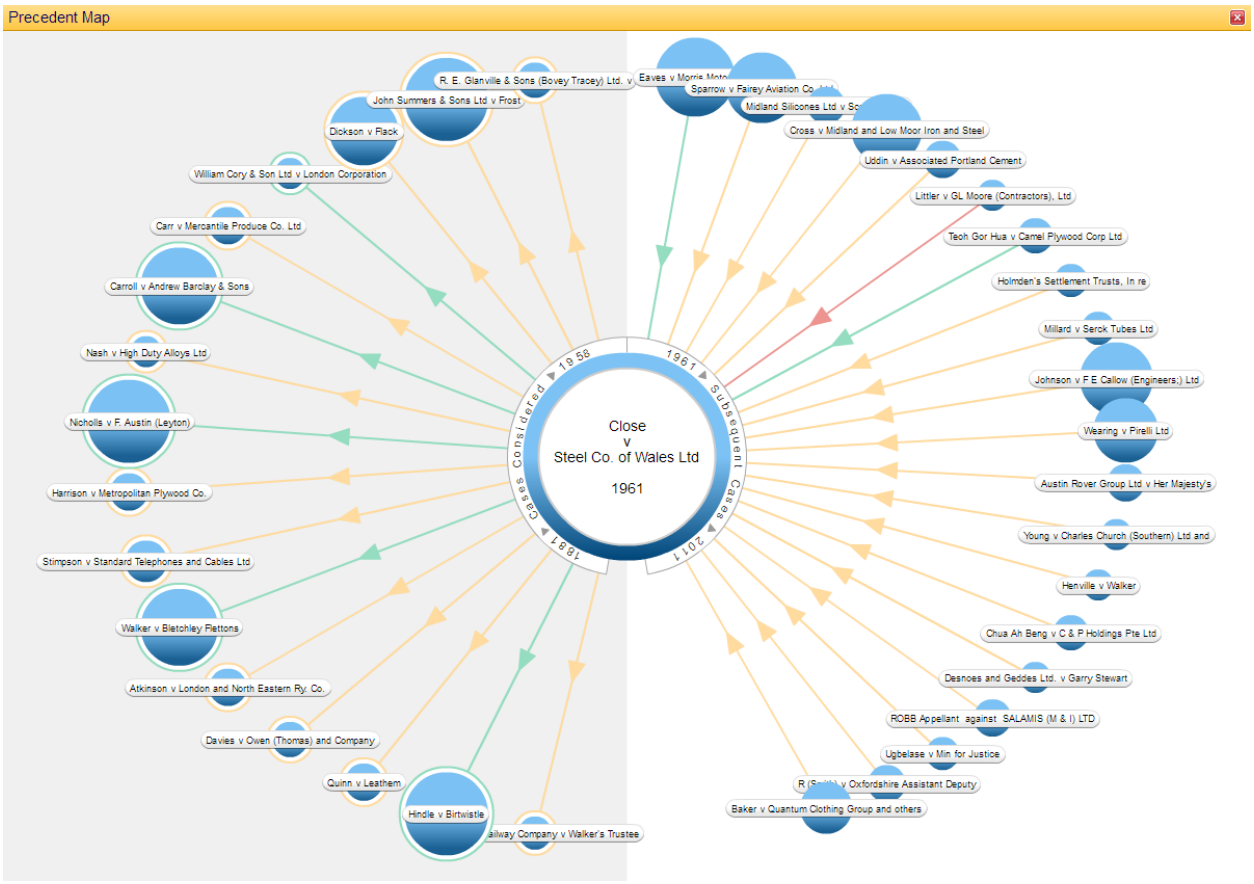

Figure 5: *JustCite's Precedent Map. https://www.justcite.com*

and directly prohibit this.

The next best solution would be to have a manually created corpus of case law which the program could work on. Only Westlaw allows the users to *"put a collection of material on your intranet"*, *"storing information in a know how database"* and download the information if it is done in a way that cannot be used to substitute their database [9]. Westlaw also allows to share this information with people not subscribed to its services, if it cannot be used to substitute their database. Being the only legal search engine directly not prohibiting collection of data, Westlaw was selected as the source for the Cited corpus used in this project.

Apart from being the only provider not explicitly forbidding any manipulation with their data, Westlaw has the advantage of being well known resource that most legal professionals are acquainted to. It also provides a list of citing cases of a case of interest, this will prove important in Section 4.6, where we hypothesise the identification of cited paragraphs could help identifying the ratio.

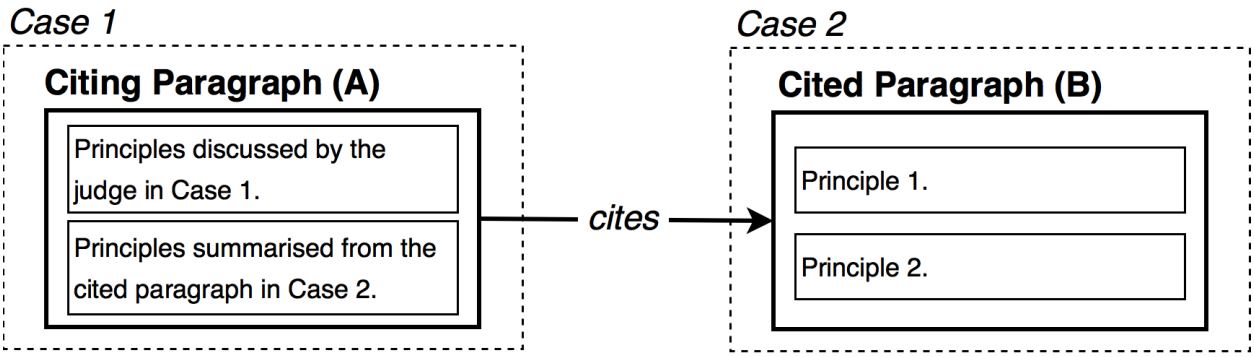

Figure 6: *The distinction between cited and citing paragraphs.*

2.5 Citing vs. cited: cases, paragraphs and principles

When referring to paragraphs in legal cases connected by a citation we will be careful to use the terminology in Fig. 6, which illustrates the distinction between citing and cited paragraphs. Although this is conceptually trivial it helps to avoid confusion when comparing our approach to that of Shulayeva et al.. Quite simply, suppose that a paragraph (A) in a case (1) contains a citation to some paragraph (B) in another case (2). Then the former is called the citing paragraph (A) while the latter is called the cited paragraph (B). This is a natural extension of the nomenclature in Westlaw which refers to former case (1) as the citing case and the latter case (2) as the cited case. Obviously both paragraphs (A and B) can simultaneously be cited and cite many other paragraphs in many other cases.

As highlighted by Shulayeva et al., the citing paragraph (A) might contain rephrasings of principles mentioned in the cited paragraph (B) (i.e. Principles 1 and 2) but it can also contain discussion of other principles being discussed by the judge specific to case 1 itself. It is important to note that Shulayeva et al. refer to the rephrasing of principles 1 and 2 found in the *citing* paragraph as the *cited* principles (which can sometimes lead to confusion unless one is careful).

2.6 Summary

In conclusion, the contemporary legal research tools are focused on an overall relationship between cases and summarization of what each case is in general about. They do provide very little help in the extraction of ratio decidendi from the case, making the task of law identification inefficient. Therefore, the aim of this project is to fill this gap in legal research tools. To fully address this issue, our program needs to identify all principles and distinguish the ones with ratio from the ones with obiter. Finally, from the four resources reviewed, only one can be used for corpus collection carried out in this project, and that is Westlaw.

[9]Westlaw's Terms of Use - http://legalsolutions.thomsonreuters.com/law-products/about/legal-notices/terms-of-use

3 Legal information mining background

There are two areas of research we are interested in, principle identification and cited paragraph identification, both of which could be broadly described as legal information mining. For identifying principles, we explore the work of Shulayeva et al. [1] and related work in case law summarisation to identify the ML model to use for this task as well as the new features which could improve it. This background information will later inform our new approach to ratio identification and possible ways of improving principle identification. We explore the work of Adedjouma et al. [2] on using NLP on identifying and resolving cross references and contrasting it with the work of Tran et al. who uses ML on the same task, for what will become our basis for identifying cited paragraphs.

3.1 Principle identification

The identification of principles is crucial for this project and no research comes closer to this task than Shulayeva et al.'s [1]. This is why we will employ their ML model, described below, for identifying principles. Shulayeva et al. have set out on the task of identifying and distinguishing cited facts and principles in citing paragraphs, i.e. looking at citing paragraph (A) to find principles and facts in cited paragraph (B) [1]. They formulate this task as a classification problem.

Their research demonstrates an inter-annotator agreement[10] between two human annotators on annotating cited facts and principles of $\kappa = 0.65$ and intra-annotator agreement of $\kappa = 0.95$. The inter-annotator study compares two human annotators, one with legal background the other without, on identifying facts, principles and neither sentences in case law paragraphs containing at least one citation. The intra-annotator study does the same, however with the same annotator annotating the same cases several months apart. The detailed results are in the Table 1 below.

Table 1: Shulayeva et al.'s inter and intra annotator agreement study.

	Annotator1 (original annotation)	Annotator2 (inter-annotator study)	Annotator1 (intra-annotator study)
Principles	266 (32%)	211(26%)	258(31%)
Facts	56 (7%)	20 (2%)	54 (7%)
Neithers	499 (61%)	590 (72%)	509 (62%)

Shulayeva et al. have automated their task with 85% accuracy using supervised machine learning framework based on linguistic features. The features they use are: part of speech tags, unigrams, dependency pairs, length of sentence, position in the text and an indicator the sentence contains a full case citation. Their method is described below [1]:

1. *Feature counts were normalised by tf and idf.*

[10] κ is the predominant agreement measure that corrects raw agreement P(A) for agreement by chance P(E) [1, 15] :

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

2. *Attribute selection (InfoGainAttributeEval in combination with Ranker (threshold = 0) search method) was performed over the entire dataset.*

3. *The Naive Bayes Multinomial classifier was used for the classification task.*

4. *Results are reported for tenfold cross-validation. The 2659 sentences in the dataset were randomly partitioned into 10 subsamples. In each fold one of the subsamples was used for testing after training on the remaining 9 subsamples. Results are reported over the 10 testing subsamples, which constitute the entire dataset.*

Shulayeva et al. have applied their framework on their Gold Standard corpus comprising of 2659 sentences selected from 50 common law reports that had been taken from the British and Irish Legal Institute (BAILII) website in RTF format. The corpus contained human annotated sentences labeled 60% as neutral, 30% as principles and 10% as facts. Their complete results are in Table 2 below.

Table 2: Per category and aggregated statistics for the Shulayeva et al.'s principle and fact classifier trained on their Gold Standard corpus.

Classified as →	Principle	Fact	Neither
Principle	646	5	160
Fact	4	198	41
Neither	135	38	1432
Type	Precision	Recall	F-Measure
Principle	0.823	0.797	0.810
Facts	0.822	0.815	0.818
Neither	0.877	0.892	0.884
Accuracy	0.85	κ	0.72

3.2 Sentential features

However, approaching data mining task as a classification problem, as Shulayeva et al. does above, isn't new. Teufel et al. have used it to classify citation functions and argumentative zones in scientific reports [16]. Inspired by Teufel's argumentative zones, Hachey et al. have used it to summarize House of Lords case law [17], Farzindar et al. to find arguments in Canadian case law [18], and Kuhn to extract content zones in German court decisions [19]. All of them use different machine learning models, however the features they train their models with are interesting to us, since they could be used to improve Shulayeva et al.'s model as well. We will describe the features previously used in such research below.

The first approach complementing unigrams is the use of bigrams. This was employed by Pang et al. to capture some *"context"* in their research on sentiment analysis. Yet, from Figure 7, if we compare line 2 and 4, we can see the accuracy of sentiment analysis using bigrams declines. Pang concluded, that in their setting bigrams were insufficient in capturing the grammatical context. While Pang et al. were interested in classifying sentiment in movie reviews, their observation is

universal, since two subsequent words upon inspection do not provide enlightening grammatical context. Shulayeva et al. instead use a more sophisticated dependency pairs, which could be obtained from the Stanford CoreNLP[11], and would be grammatically linked [1]. This approach helps the problem of capturing of grammatical context, which Pang et al. identified.

	Features	# of features	frequency or presence?	NB	ME	SVM
(1)	unigrams	16165	freq.	**78.7**	N/A	72.8
(2)	unigrams	"	pres.	81.0	80.4	**82.9**
(3)	unigrams+bigrams	32330	pres.	80.6	80.8	**82.7**
(4)	bigrams	16165	pres.	77.3	**77.4**	77.1
(5)	unigrams+POS	16695	pres.	81.5	80.4	**81.9**
(6)	adjectives	2633	pres.	77.0	**77.7**	75.1
(7)	top 2633 unigrams	2633	pres.	80.3	81.0	**81.4**
(8)	unigrams+position	22430	pres.	81.0	80.1	**81.6**

Figure 7: *Pang et al. accuracy of applying ML tools for sentiment analysis.*

A second, similar approach to remedy the problem above is the use of Parts of Speech tags (POS). POS could help recognise the difference between a sentence such as: *"I love this film"* and *"this film is about love"*. Pang did not find major improvements using these: *"as depicted in line (5) [of Figure 7], the accuracy improves [only] slightly for Naive Bayes."* But as Shulayeva et al. discovered, these can be used for distinguishing facts and principles as *"sentences that introduce facts are most often presented in the Past Indefinite tense"* while *"both epistemic and deontic modal qualifiers that use modal verbs are common in sentences containing legal principles"*.

Apart from unigrams, as we can from Fig. 8, taken from Shulayeva et al.'s paper, POS and dependencies have been particularly strong features in their research. Of course, because these features are further vectorised and pruned so that the Multinomial Naive Bayes algorithm can be applied on them, as per their methodology, they do get translated into thousand of individual features. From Fig 9 taken from Shulayeva et al.'s paper, the particularly informative strings per feature were identified. We will later compare these with regard to our improved model, see Section 6.1.

	Majority class	Part-of-speech	Unigrams	Dependencies	All
Accuracy	0.60	0.63	0.77	0.81	0.85
κ	0.00	0.18	0.58	0.63	0.72

Figure 8: *Shulayeva et al.'s reported accuracy and κ for each type of feature.*

A third approach is using location within the text. Pang [38] again did not find any great improvements in their research, however Teufel et al. [20], in their efforts to identify a rhetorical

[11]Stanford CoreNLP - Natural language software v 3.8.0 - https://CoreNLP.github.io/CoreNLP/

Part-of-speech tags: VBZ, NN, JJ, VB, MD, DT, IN, CC, NNP, VBN, COMM, RB, WRB, SEMM, FS, NNS, TO, QUOT, WDT, VBG, WP, POS, VBP

Unigrams: is, a, the, or, be, Mr, was, must, of, had, may, has, see, 0, it, 100, where, I, were, other, are, if, will, to, concerned, general, person, 300, an, Mrs, and, judgment, party, that, planning, principle, letter, company, one, If, circumstances, which, per, money, whether, Hawk, always, submissions, not, Charles, Arista, court, jurisdiction, forum, can, pictures, Akzo, Miss, ordinary, fund, man, S1, contained, 000, wholesalers, this

Dependency pairs: case-that, is-there, concern-was, concern-case, case-was, parties-the, condition-a, court-the, is-if, letter-the

Figure 9: *Shulayeva et al.'s reported top 100 features by information gain.*

status of a sentence, identified that this feature is relevant in scientific papers. Teufel et al.'s research formed a basis for Hachey et al. [17] who tried the same in the domain of case law, also identifying this as an important feature. Shulayeva et al. uses location within the text too, measuring the position from the beginning of the case on a scale from 0 to 1, however they do not report individual performance of this feature like they do for unigrams, dependencies and POS tags. This is probably because this feature can help in conjunction with other features, but on its own is not very informative.

Like Shulayeva et al., Teufel et al. and Hachey et al. [1, 20, 17] both use the length of a sentence to complement their models. Shulayeva et al. uses number of words to capture the difference in sentence length between sentences containing facts and principles. Hachey et al. even reports the performance of this feature when used by Naive Bayes algorithm in Fig. 10. We highlight particularly the results for Naive Bayes classifier in the red rectangle, which show that individual (Ind) performance of length of sentence and location in text is quite low, but improves the overall performance none the less.

The last feature Shulayeva et al. [1] uses is presence of a citation. This feature does not have a direct counterpart in Hachey et al.'s or Teufle and Moens research. It maps the distribution of citations of other cases on sentences with facts and principles, helping to distinguish them. The individual performance of this feature in not reported by Shulayeva et al..

	C4.5		NB		Winnow		SVM		ME	
	Ind	Cum	Ind	Cum	Ind	Cum	Ind	Cum	Ind	Cum
Cue Phrase	47.8	47.8	39.6	39.6	31.1	31.1	52.1	52.1	48.1	48.1
Location	**65.4**	54.9	34.9	47.5	34.2	40.2	35.9	55.0	42.5	51.9
Entities	35.5	54.4	32.6	48.8	26.0	40.2	33.1	56.5	35.8	53.7
Sent. Lngth	27.2	55.1	20.0	49.1	27.0	40.4	12.0	56.8	21.5	54.0
Quotations	28.4	59.5	29.7	**51.8**	23.3	41.1	27.8	60.2	25.7	57.3
Them. Wds	30.4	59.7	21.2	51.7	25.7	**41.4**	12.0	**60.6**	27.7	**57.5**

Figure 10: *Hachey et al.'s micro-averaged F-score results for rhetorical classification. Ind stands for individual, cum for cumulative result.*

There are also three features that Shulayeva et al. [1] do not use, but which are employed in research on case law summarisation. The first one is lemmatised text. While strictly speaking not typically used as an individual feature, Hachey et al. in their research lemmatises the text to group together words with the same core or lemma. This approach helps to reduce the noise in their dataset improving the performance of their model. Another features used by Hachey is quotation. Hachey captures the percentage of the sentence in quotation marks, and reports this as the single most informative feature employed.

Saravanan et al. [37], whom we will introduce in more detail in Section 3.3, also use named entity recognition in their research on summarising case law. Named entity could be ranging from person to a company or a date, and helps teaching the algorithm to distinguish sentences based on the occurrence of such entities.

It should be noted on the outset, that the features we are exploring can not increase the performance of Shulayeva et al.'s model substantially. This is because they will be used together with many other features, to compensate for the bag of word approach deficiencies we report on in Section 4.5.

3.3 Ratio identification

There has been previous research on summarising case law, by Saravanan et al., who as part of their work claim to have identified the ratio with a high accuracy [37]. Their work serves as an example of how easy it is to confuse what ratio is, rather than as a comparison to our own work and we further criticise it in Section 4.2.

Saravanan et al. [37] report that a legal judgement can be separated into segments of rhetorical roles. They present a system for annotating these rhetorical notes and composing automatically a case headnote (i.e. a summary of the case). They achieve this employing ML technique (Conditional Random Fields), on document segmentation and probabilistic models for extraction of identification of key sentences. Their rhetorical status number five, *"Ratio of the decision"* is of a particular interest to us. As can be seen from the confusion matrix in Fig. 11, they identify 206 out of 228 ratio's correctly, giving them an accuracy of 90% for identifying this particular rhetorical status.

3.4 Cross reference resolution

Citation analysis was pioneered by Frank Shepard in 1873 [21], full hundred years before Eugene Garfield [22, 23, 24], inspired by Shepard, helped to develop citation analysis in the domain of scientific reports. Automated identification of in text citations in law date at least to 1969 [25]. It is of no surprise, that every legal research service mentioned in Section 2, including Westlaw, provides their users with an already annotated text, linking within their respective database, as well as a list of citing and cited cases for every major judgement. However, while Borkowsky had already in 1969 extracted full case citations with better than 99% accuracy, there has been no effort to the authors knowledge to go a step further and automatically link the specific citations, pointing to pages, paragraphs and sentences [25]. Indeed, none of the providers mentioned above do give their users an easy access to these. Yet, as will become apparent in Section 4, identifying cited paragraphs is crucial to our project. Fortunately, there has been a research into a related area of cross references [26, 27, 28, 29, 30, 31, 32, 33, 34, 2]. The task of cross reference resolution has been attempted both by applying the rule based approach as well as the ML approach. We contrast

Actual \ System		Identifying the case	Establishing the facts of the case	Arguing the case	History of the case	Arguments	Ratio of the decision	Final decision		Total
Identifying the case		**53**	2	0	4	1	1	0		61
Establishing the facts of the case		2	**358**	1	24	7	4	8		404
Arguing the case		2	1	**128**	7	19	2	0		159
History of the case		51	54	20	**2048**	218	143	14		2548
Arguments		2	11	58	40	**697**	13	2		823
Ratio decidendi		1	2	0	10	3	**206**	6		228
Final decision		0	0	0	2	0	2	**101**		105
Total		111	428	207	2135	945	371	131		**4328**

Figure 11: *Saravanan et al. present their results in identifying rhetorical categories using CRF in the confusion matrix above.*

these two approaches below.

3.4.1 Rule based approach

Adedjouma et al.'s aim was to *"enable easier navigation and handling of cross references"*, by an *"automated detection and resolution of legal cross references"* [2]. In their research, cross reference was understood mainly as a reference to another part of the same text. The task is accomplished in two steps, first identifying the references and second breaking them down into individual components.

Art. 2. [...] Individuals are considered non-resident taxpayers if they do not reside in Luxembourg but have a local income as per the definition of Art. 156.

Figure 12: *CRI in Dutch legislation.*

First lets consider step one. On the example in Figure 12, we can see a typical reference Adedjouma et al. is working with. Their research automatically detects the *"Art. 156"* and links it with the target provision contained in the same document. To do this Adedjouma et al., have identified natural language patterns in Luxembourg's legislation and wrote them down as a grammar in Figure 13, which was used to write regular expressions to automate the task.

Figure 13, distinguishes between two types of reference patterns, simple and complex. This draws on the research done by De Maat et al., on Dutch tax and customs legislation [29]. Both De Maat et al. and Adedjouma et al. distinguish several types of citations. For example, Adedjouma et al. distinguishes between the simple and complex, implicit and explicit, and internal and external. De Maat [29] on the other hand distinguishes four types of simple citations and two, multivalued and multilayered, complex citations. The following paragraphs further describe Adedjouma et al.'s findings.

Simple citations can be implicit or explicit. An explicit cross reference comprises of a concept marker, for example *article* and a number, or an ordinal expression, for example *first* or *1st* followed by the concept marker. The numbers might be in round brackets, written in roman number a text or a letter. An implicit cross reference will not give a direct number for the concept marker, and might be vague, for example *this article* or *following paragraphs.*

As per Adedjouma et al., *"complex [cross reference']s enhance simple [cross reference']s with three additional features: enumerations, ranges, and navigation through levels"*[2], complex cross references further subdivide into multivalued and multilayered.

A multivalued cross reference *"cites many provisions in the same expression by specifying only once a concept marker followed by a numerical expression."* There are three types of numerical expression, as per Adedjouma et al. [2]:

1. *an AND/OR enumeration, e.g., "numbers 1, 2 and 3" and "articles 22 or 102"*
2. *a range, e.g., "numbers 1 to 3"*
3. *a combination of enumerations and ranges, e.g., "articles 119 to 121 and 124"*

A multilayered cross reference *"describes a navigation path through the hierarchy of a legal text".* There are three types of multilayered cross references, as per Adedjouma et al. [2]:

1. *The navigation may be from an upper to a lower level, e.g., "article 91, 1st alinea, No 2".*
2. *The navigation can be from a lower to an upper level, e.g., "second sentence of article 10 of the law of 23 may 1964".*
3. *The navigation can be mixed-mode. That is, a CRE may start at a convenient hierarchical level, navigate upward or downward in the hierarchy, and then go in the reverse direction.*

The research done by Adedjouma et al. has achieved a precision of 99.7%, recall of 97.9% and F-measure of 98.8% in identifying the references, and it took 34 seconds in a task involving a total of 1640 pages from Luxembourg's legal corpora.

For the second step, the identified references must further be broken down into individual article numbers. This is usually as simple as breaking down a list of numbers into individual components. However there are two potentially problematic areas Adedjouma et al. identifies.

First, there are two instances of simple implicit references that need to be resolved:

1. *Cross references that are semantically equivalent to current, previous, or next followed by a concept marker. For example "next article", "next" needs to be resolved as the number of next article.*
2. *Cross references that are semantically equivalent to same or this followed by a concept marker. For example "this article", "this" needs to be resolved as the current article number.*

Line	Simple cross reference patterns		
1	⟨*simple-ref-expr*⟩	::=	⟨*explicit-expr*⟩ | ⟨*implicit-expr*⟩
2	⟨*explicit-expr*⟩	::=	⟨*internal-expr*⟩ | ⟨*external-expr*⟩
3	⟨*internal-expr*⟩	::=	⟨*marker-term*⟩⟨*num-expr*⟩ | ⟨*ordinal-expr*⟩⟨*marker-term*⟩ | ⟨*generic-term*⟩⟨*num-expr*⟩
4	⟨*marker-term*⟩	::=	"article" | "articles" | "art." | "paragraph" | ...
5	⟨*num-expr*⟩	::=	⟨NUMBER⟩ | ⟨LETTER⟩ | ⟨ALPHANUM⟩
6	⟨*ordinal-expr*⟩	::=	⟨TEXT-ORDINAL⟩ | ⟨NUM-ORDINAL⟩
7	⟨*generic-term*⟩	::=	"sub" | "under"
8	⟨*external-expr*⟩	::=	⟨*external-expr*$_1$⟩ | ⟨*external-expr*$_2$⟩
9	⟨*external-expr*$_1$⟩	::=	⟨*name-term*⟩ | ⟨*category-term*⟩⟨*link-term*⟩ ⟨DATE⟩| ⟨*adj-term*⟩⟨*category-term*⟩⟨*link-term*⟩ ⟨DATE⟩ | ⟨*name-term*⟩⟨*link-term*⟩ ⟨DATE⟩ | ⟨*delegating-expr*⟩
10	⟨*external-expr*$_2$⟩	::=	⟨*internal-expr*⟩⟨*auxiliary-term*⟩⟨*external-expr*$_1$⟩
11	⟨*delegating-expr*⟩	::=	⟨*delegation-term*⟩ | ⟨*adj-term*⟩⟨*delegation-term*⟩
12	⟨*category-term*⟩	::=	"law" | "decree" | "directive" |...
13	⟨*name-term*⟩	::=	"social insurance code"|"complementary pension law"| ...
14	⟨*adj-term*⟩	::=	"modified" | "grand-ducal" | "ministerial"
15	⟨*auxiliary-term*⟩	::=	"as it was introduced by the" | ...
16	⟨*delegation-term*⟩	::=	"regulation" | "memorial" | ...
17	⟨*implicit-expr*⟩	::=	⟨*implicit-term*⟩⟨*marker-term*⟩ | ⟨*implicit-term*⟩⟨*category-term*⟩| ⟨*marker-term*⟩⟨*implicit-term*⟩ | ⟨*category-term*⟩⟨*implicit-term*⟩| ⟨*internal-expr*⟩⟨*implicit-term*⟩| ⟨*implicit-term*⟩⟨*unspecific-term*⟩| ⟨*implicit-term*⟩⟨*num-expr*⟩⟨*marker-term*⟩ | ⟨*unspecific-term*⟩⟨*implicit-term*⟩
18	⟨*implicit-term*⟩	::=	"above"| "below" | "preceding" | "following" | "that follows"| "next" | "previous" | "this" | "in question" | "same" | ...
19	⟨*unspecific-term*⟩	::=	"provision"
20	⟨*link-term*⟩	::=	"of" | "of the" | "of a"
	Complex cross reference patterns		
21	⟨*complex-ref-expr*⟩	::=	⟨*multivalued-expr*⟩ | ⟨*multilayered-expr*⟩
22	⟨*multivalued-expr*⟩	::=	⟨*multivalued-expr*$_1$⟩ | ⟨*multivalued-expr*$_2$⟩
23	⟨*multivalued-expr*$_1$⟩	::=	⟨*internal-expr*⟩⟨*sep-term*⟩⟨*num-expr*⟩ | ⟨*external-expr*⟩⟨*sep-term*⟩⟨*num-expr*⟩⟨*sep-term*⟩ ⟨DATE⟩
24	⟨*multivalued-expr*$_2$⟩	::=	⟨*multivalued-expr*$_1$⟩⟨*sep-term*⟩⟨*num-expr*⟩ | ⟨*multivalued-expr*$_1$⟩⟨*sep-term*⟩⟨*implicit-term*⟩
25	⟨*multilayered-expr*⟩	::=	⟨*multilayered-expr*$_1$⟩ | ⟨*multilayered-expr*$_2$⟩
26	⟨*multilayered-expr*$_1$⟩	::=	⟨*internal-expr*⟩ ⟨*sep-term*⟩ ⟨*internal-expr*⟩
27	⟨*multilayered-expr*$_2$⟩	::=	⟨*multilayered-expr*$_1$⟩⟨*sep-term*⟩⟨*internal-expr*⟩ | ⟨*multilayered-expr*$_1$⟩⟨*link-term*⟩⟨*internal-expr*⟩ | ⟨*multilayered-expr*$_1$⟩⟨*link-term*⟩⟨*multivalued-expr*⟩
28	⟨*sep-term*⟩	::=	";" | "–" | "and" | "or" | "to" | ...

Figure 13: *Luxembourg's legislation grammar, by Adedjouma et al.*

Second, for a resolution of the complex multivalued cross references such as *Articles 10-15 or Articles 1 to 3*, the reference needs to expanded accordingly to cover the entire range of articles cited. The same applies for multilayered references such as *Article 10 a, b, c.* These need to be further resolved accordingly as *Article 10a, 10b and 10c.*

The research done by Adedjouma et al. has achieved a precision of 99.9%, a recall of 97.5%, and F-measure of 98.7% for this second step. De Maat reports accuracy of 99% for simple references and 95% for complex ones, for both tasks. While De Maat's work is very similar to Adedjouma et al.'s, due to Adedjouma et al.'s more careful analysis, which includes the multivalued cross references, we will further concentrate only on their work.

3.4.2 Statistical approach

Similar studies to Adedjouma et al. have been conducted on Italian [35][30][26], Spanish [34], US [26] and Japanese [32] legal corpora. Tran et al. [32] is particularly interesting for contrast with the

research above as it uses machine learning, specifically sequence labelling using conditional random fields [36], to identify citations. Tran et al. criticizes the above research for *"limiting resolvers to identifying only the referred documents but not to which parts of texts in these documents"*.

The first reason Tran et al. decided to go in this direction was to not limit themselves to the extraction of only *"normative references"*. Instead their aim was to *"extract the smallest fragments of texts that are actually referred to by references"*. What Tran et al. mean by normative references can be best understood from the Figure 14. *"Normative references would appear in the forms of 'para 1' and previous article, para 4"*, targeting a full paragraph, whereas the smallest possible fragment these references target would be highlighted by the green brackets in Figure 14 instead. This could be a full paragraph, but could also be only a part of a sentence as in the A12P1-1 example below.

The second reason Tran et al. decided to use statistical methods, is to be able to analyse the legal text without any domain knowledge. To port their method from one language to another for example should theoretically take only *"a minimal amount of human intervention"*. Their *"framework using a machine learning approach"*, let's the system to be *" automatically trainable from a corpus"* [32].

Tran et al. reports 80% accuracy at the core task of identification, over 19% less than what Adedjouma et al. report. This is ultimately why we later adapt Adedjouma's research instead of Tran et al.'s. See Section 4.6.

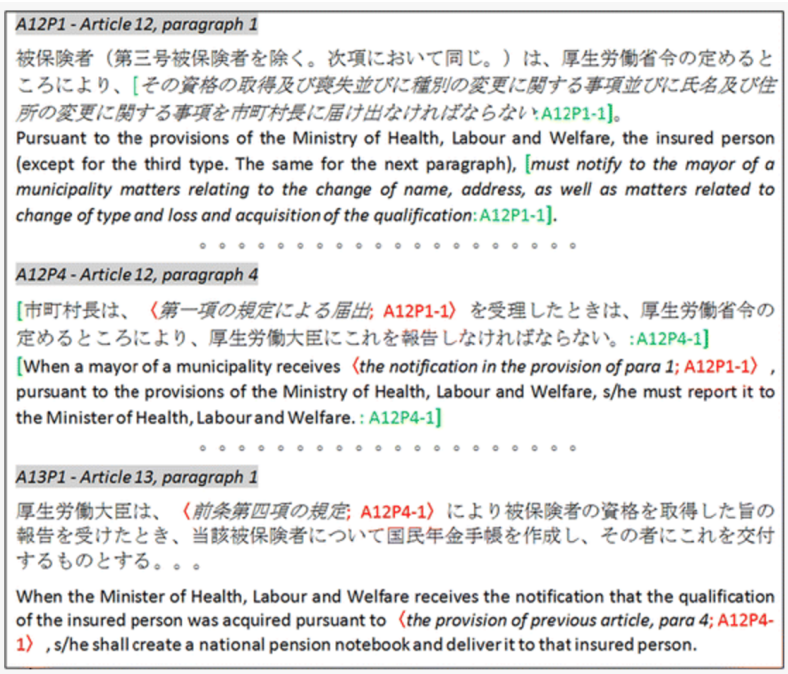

Figure 14: *Tran et al. example from Japanese legal text. References are in the red brackets, referents are in the green brackets.*

3.5 Summary

In conclusion, this Section identifies the relevant research for both of our core tasks. We identify Shulayeva et al. as important research in finding paragraphs in case law. We further identify how different research interested in classifying rhetorical function of a sentence in case law or scientific reports have identified different features to help their ML task. Crucially, some of these features have not been used by Shulayeva et al.. We further compared the two approaches to dealing with cross references, to demonstrate that NLP approach has superior results to ML approach for this task.

4 Towards a new approach

In light of the background information above, this section identifies a gap in current legal research tools and argues this gap has not been addressed by computer science research. We further present a hypothetical solution to this problem and test this solution is viable by a manual annotation study. We elaborate on how to improve the previous research on identifying principles by Shulayeva et al.. Finally, we explain why we will be adapting the NLP approach of Adedjouma et al. [2] instead of ML approach of Tran et al. [32] to identify cited paragraphs and highlight the similarities between the the task of cross reference resolution and identification of cited paragraphs.

4.1 Limitation of legal research tools

As we have established in the discussion of legal research in Section 2.3, the two tasks a legal professional faces when conducting research can be described as: One, to find the cases that contain the law by looking for related cases on facts. Two, to find the ratio within these cases.

The current legal research tools are well adapted to help with task number one. They do so by allowing lawyers to find cases related on facts of the case. Notably Westlaw and Lexisnexis provide a summary of what has happened in the case, so that without reading the judgement a lawyer can identify if such case contains similar circumstances to the problem he or she is trying to resolve.

Yet, the current tools are ill-adapted for resolving the second task. While it is easy to glean the outcome of the case from the summary, there is no other way of identifying the reasoning in a case than reading it. Further, sometimes it is necessary to read numerous other cases, to understand which principles in the judgement of interest are indeed the important ones, and which are simply an obiter. All this is very time consuming.

4.2 Limitation of current legal information mining research

Saravanan et al. are the only recent work, as far as we are aware, engaging at least partly with solving the problem of ratio identification [37]. As part of their work in summarising cases, they report to have incidentally discovered a way of identifying ratio with high accuracy. However as per the paragraphs below, while their findings seem valid in terms of identifying a way of summarising cases, the rhetorical function they call a ratio of the case, is simply misnomer.

Saravan et. al. reports identifying the ratio in Indian case law with 90% accuracy, focusing on rhetorical roles, employing Conditional Random Fields algorithm [37]. The problem with their approach is that it conflates the doctrine or stare decisis or precedent with that of ratio decidendi. We can notice this in their definition of ratio: *"A judge generally follows the reasoning used by earlier judges in similar cases. This reasoning is known as the reason for the decision (Ratio decidendi)."* [37]. These are however two distinct concepts. One is of precedent, under which the reasoning in previous judgements is applied on cases with similar facts. The other is ratio, the principles deciding the outcome of the case are the binding law. As we will see in the discussion of our hypothesis, there is a connection between the two. However, not every deciding principle is taken from previous case and not always is a case cited for the ratio. Certainly, they are not the same concept.

The oversimplification in Saravan et al.'s work is further underlined by their lack of discussion of the opposite of ratio, the obiter. Not a single time they elaborate on how are they distinguishing

the deciding principles from any other principle they find. Only examining their examples can therefore give us a clue about what they in fact identify. We cite these examples below [37]:

1. *We are of the view that the order under challenge does not require any interference by this court*
2. *Looking at the question in the above perspective, we find no infirmity in the order passed by the Chief Commissioner in transferring the case to the assistant Commissioner of Income Tax, Calicut*
3. *We are clearly of the view that all these statutory remedies could not be allowed to be bypassed and, therefore it was not appropriate for the learned judge to have entered into the merits of the controversy at the state when the Company had been issued a notice under Section 17 of the Act*

These sentences are concluding statements. This is especially noticeable on the example number two, which specifically refers to the *"above perspective"*, that presumably contains the reasons for reaching this conclusion. However, these decisions on individual points of law do not contain the reasoning that we come to expect from a ratio under the Wambaught's test. While there are indeed different interpretations of ratio, these statements lacking both facts and principles do not seem to fall under any legally sound definition we came across, since all these definitions require focus on either facts, principles or both.

Furthermore, the only way these examples could be connected with the definition of ratio Saravanan et al. themselves propose, is if they are somehow the results of following the precedent in previous cases. Yet there is no evidence that their method is taking previous cases into account. On contrary, it seems that the features they employ (i.e. named entities, upper case word, position within text etc.) can not capture the relationship these sentences have with previous case law.

Saravanan et al.'s research is therefore valuable when related to other research in case law summarisation, such as Hachey et al.'s efforts to identify general segments of the case [17], yet can not be considered a research on ratio identification, such as the theoretical work carried out by Branting [3]. This makes sense since both Saravanan et al. and Hachey et al. focus primarily on generating case law summaries, or headnotes. The criticism of Saravanan et al.'s work above, highlights how difficult it is to stay true to at least one of the legally sound definitions and identifies the gap in research of automatic ratio identification.

4.3 A new approach to the problem

A new approach is therefore necessary. As per Wambaugh's test, to identify ratio it is imperative to address two issues, see Section 2.2. The first issue is to distinguish statements of legal principles from discussion of specific facts to which those principles are applied in a particular case. The second issue is to determine which of those principles are pivotal to the outcome of the case (and therefore constitute the legally binding ratio), from those principles which are merely incidental and which are formally known as *obiter* [3].

According to the doctrine of stare decisis, cases are decided based on the precedent set out in the case law preceding them. Judges cite previous cases to apply the ratio from preceding cases on the facts before them [6]. Therefore, when judges cite specific paragraphs, it is reasonable to expect they are citing the paragraphs mostly for the ratio contained in them. To find the ratio we

would thus like to identify paragraphs in a case that have been cited by other cases and identify the principles in them. Therefore, much like a lawyer, we first must identify principles. However, instead of conducting the complex task of evaluating whether a principle is ratio or not by employing the Wambaught's Inversion test, we find only those principles contained in cited paragraphs.

Upon the first inspection we could attempt to find ratio of a particular case using the methodology of Shulayeva et al., by identifying principles in the paragraphs of a subsequent case that cite the case of interest (i.e. citing paragraph A in Fig. 6). But as they report, this is not effective because it tends to confuse the ratio of the earlier case with that of the subsequent case: *"The main cause of error for the automatic annotation of principles was that the gold standard only annotated principles from cited cases, but often these were linguistically indistinguishable (in our machine learning approach) from discussions of principles by the current judge; i.e principles expressed by the current judge should have been annotated as neither, but were frequently annotated as principles."* [1]. Perhaps for this reason, Shulayeva et al. do explicitly say that *"the term ratio will be avoided"* [1].

To find the ratio of a given case we agree it is important to identify principle (as opposed to facts) and in particular those principles cited by subsequent cases (as opposed to those which are not). But, in contrast to Shulayeva et al. we believe it is better to extract the ratio from the paragraphs of the given case to which the subsequent citation refers (i.e. cited paragraph B in Fig. 6) than from the text of the subsequent case in which the citation occurs (i.e. citing paragraph A in Fig. 6).

4.4 Manual annotation study

To test if cited paragraphs correlate with ratio more than citing paragraphs, we have conducted an empirical study on the Cited corpus. We have created the Cited corpus ourselves by collecting and analysing cases from Westlaw. Westlaw, a major legal search engine, contains a list of key citing cases for every major case in its database, see Section 2.3. As the case of interest, Stack v Dowden [2007] UKHL 17 (*Stack*) was selected. For Stack this list contains 51 cases, out of these, Westlaw has 41 in their database. The rest are unreported. These 41 cases contain 798 specific citations form our Cited corpus.

We selected Stack as the case of interest, for two reasons. First, because it is both cited and cites frequently as a major Supreme Court judgement and therefore gives us plenty of data to work with. Second, because it contains a famous dissenting judgement of Lord Neuberger, containing principles reaching conclusion disagreeing with the rest of the judgement, a good example of an obiter. It is therefore as harsh a case as can be for the purposes of this task.

Our study compares the occurrence of ratio in cited paragraphs containing a principle, with the occurrence of ratio in citing paragraphs containing a principle. To do this, we manually identified the citing and cited paragraphs, paragraphs containing ratio and paragraphs with principles.

To manually determine whether a paragraph contains ratio or obiter, we have applied the Wambaugh's Inversion test. Under the inversion test, the meaning of the principle is inverted and if this changes the outcome of the case, the principle is ratio, otherwise it is an obiter. Paragraphs without a principle were labeled as an obiter. Employing the Inversion test we have found out that out of 158 paragraphs in Stack, only 34 contain ratio.

Manually analysing all the 41 citing cases in Case Corpus we found all the paragraphs in Stack which have been cited. Out of 798 citations in the cases, we identified 72 distinct cited paragraphs in Stack. Out of these 62 contain principles. Going through Stack paragraph by paragraph we identified 46 citing paragraphs, out of these 34 contain principles. We retained Shulayeva et al.'s definition of a principle, under whom it is defined as *"any statement which is used, along with facts, to reach a conclusion"* [1]. Note that this is different from a definition of a ratio, as the conclusion does not have to be the decision of the case for a statement to be a principle, however it does for a principle to be a ratio under the Wambaugh's test.

The confusion matrix in Table 5.2 shows the results for cited paragraphs. The data suggest correlation between cited paragraphs and the ratio. Applying this method achieves 76% accuracy in identifying paragraphs containing ratio and $\kappa = 0.45$. We have further tested the opposite assumption looking at citing paragraphs. As per Table 4 there is a drop in accuracy to 68% and most importantly the κ coefficient falls down to mere 0.06. These results prove our hypothesis that cited paragraphs are a better indicator of ratio than citing paragraphs. Cited paragraphs are almost twice as precise and have more than three times the recall in identifying ratio than citing paragraphs.

Table 3: Distribution of ratio and obiter between cited and not cited paragraphs containing principle.

Contained in →	Cited	Not-Cited	
Ratio	31	3	
Obiter	41	83	
Type	Precision	Recall	F-Measure
Cited	0.468	0.853	0.604
Non-Cited	0.948	0.734	0.827
Accuracy	0.76	κ	0.45

Table 4: Distribution of ratio and obiter between citing and not citing paragraphs containing principle.

Classified as →	Citing	Not-Citing	
Ratio	9	25	
Obiter	25	99	
Type	Precision	Recall	F-Measure
Citing	0.265	0.265	0.265
Non-Citing	0.798	0.798	0.798
Accuracy	0.68	κ	0.06

4.5 Improving principle identification ML model

While our aim is different to Shulayeva et al.'s, we still need to identify principles and Shulayeva et al.'s ML classification model capable of identifying facts and principles is therefore important source to us. Since Shulayeva et al. use only six sentential features, see Section 3.1, we elaborate below on which features, employed by the research on case law summarisation, we later implement to improve Shulayeva et al.'s model performance. In particular Hachey's and Saravanan et al.'s research on case law summarisation, see Section 3.2, is useful in this respect, due to the similarity between some of the rhetorical roles they extract and principles. As Hachey recognises, *"Although there is a significant distance in style between scientific articles and legal texts, we have found it useful to build upon the work of Teufel and Moens and to pursue the methodology of investigating the usefulness of a range of features in determining the argumentative role of a sentence"*. In the same way the research on case law summarisation can inform our features. To understand how new features can be used to improve upon Shulayeva et al.'s research we will first look at the limitations of the Naive Bayes approach they employ.

Naive Bayes, relies on a simple representation of a text, commonly referred to as bag of words model. In Shulayeva et al.'s research this corresponds to the unigram feature. Unigrams are simply individual words of the text. Pang et al. [38] identified the pitfall of the bag of word approach when dealing with more complex types of classification than topic based. They have focused in their research on the sentiment analysis, but their discussion is enlightening for our purposes as well. For example, in a sentence such as: *"This movie should be terrible, it has all the signs of an awful comedy, but to my surprise I have enjoyed it very much."* as per Pang et al.: *"a human would easily detect the true sentiment of the review, but bag-of-features classifiers would presumably find these instances difficult, since there are many words indicative of the opposite sentiment to that of the entire review."* This downside can be attributed to the false assumption of the bag of word approach *"that word order and grammatical relations are not significant"*. Today, classifiers select features specifically to provide algorithms with some level of context awareness, and not just evaluate specific word frequency. This is how Pang et al. addressed this problem, by introducing multiple features other than unigrams. And it is also how Shulayeva et al. remedies this shortcoming. Interestingly this is a direct violation of the Naive Bayes assumption of conditional independence of the features [38], [39], but as per Pang et al. *"this does not imply that Naive Bayes will necessarily do poorly"* [40].

The first novel feature we believe could improve Shulayeva et al.'s model is lemmatising the text. Lemmatisation removes the variants of words with the same lemma, by grouping together the inflected forms of such words [17]. Shulayeva et al. have noted the importance of tense and voice in classification of facts and principles, both of which are removed by lemmatisation. This is perhaps why their research did not use lemmatised unigrams as a feature. We demonstrate this assumption is true in Section 6.1. However, the expectation is that combining the lemmatised sentences with unigrams, part of speech tags and dependency pairs, all of which do contain the tense and voice, can still improve the performance by better capturing indicative lemma. Hachey [17] has used this feature to identify rhetorical function of a sentence in case law. Since the rhetorical function category of framing, described as Hachey as *"part of the law lord's argumentation"* [17], is essentially identifying principles, we believe this feature can be applied for our purposes as well.

Hachey [17] also uses a feature capturing percentage of the sentence in quotation marks, reporting this as a single most informative feature for Naive Bayes, on their task. We will use quotation

feature, which captures those sentences which contain text in quotation marks, returning a binary result. The idea behind this feature is that it captures sentences referencing a ratio of another case, and therefore could be indicative of a principle as much as it was for rhetorical status of a sentence in Hachey's research.

Much like quotation, sentences citing a specific paragraph should be more likely to contain a principle, since they are probably citing the ratio. This feature does not have a direct counterpart in Hachey et al.'s, Teufel et al.'s or Moen et al.'s research [17, 20, 39], but should be similar to the quotation feature they employ described above. Unlike the full text citations, used by Shulayeva et al., both the quotation and paragraph citation are specifically targeting instances where the reference is picking out subpart of the case, and not it's entirety. We believe that due to the doctrine of stare decisis, see Section 2.1, this should be a more useful feature than the full case citation, which points to the entire case instead of it's sub-section.

Shulayeva et al. already uses length of a sentence as a feature. From analysing case law, it seems that rather than the length of a sentence, which in general is long, it is the sub-clauses of a sentence that indicate argumentative nature. The sub-clauses were therefore selected to test whether they might help to indicate principles better than length of a sentence.

Finally, we decided to add named entity recognition (NER). We have experimented with naming people, places, dates, organisations and numbers. The hope was to better capture sentences which talk about facts. NER was employed before by Hachey et al. as well as Saravanan et al. for rhetorical roles identification [17, 37]. Since NER also captures words with capital first letter, this feature largely covers Saravanan et al.'s upper case word feature.

However, there is another way we implement to improve the performance. Shulayeva et al. identify two problems with their work. One, as already mentioned, their model has trouble identifying principles given by the current judge, from discussions of principles from cited cases. However, since we are interested in extracting the ratio from the paragraphs of the given case to which the subsequent citation refers (i.e. cited paragraph B in Fig. 6) rather than from the text of the subsequent case in which the citation occurs (i.e. citing paragraph A in Fig. 6), we do want to use their model to identify all the principles. To do that we can repurpose their training corpus to focus on principles by relabelling the training data in their Golden Standard corpus so that all the principles, not just the cited ones, are labeled as such. We have done this with our New corpus.

The second problem Shulayeva et al. identifies is that the *"confusion between fact and principle though rare overall may be typical in sentences whose aim is not to introduce facts and where factual information is used as a part of reasoning"* [1]. But the inevitable misclassification, that arose for Shulayeva et al., because a sentence that is both a fact and principle can only have a single label, doesn't arise for us, since such sentences can be indisputably relabeled as principles in our New corpus.

All of these features and changes combined should improve Shulayeva et al.'s model performance on our task, we implement these in Sections 5.1.3 and 5.1.4 and report our results in Section 6.1.

4.6 Adapting cross reference resolution

The research on identifying and resolving cross references conducted by Tran et al. and Adedjouma et al. is very similar to the problem of identifying cited paragraphs we are facing [32, 2]. Westlaw provides a list of cases citing a case of interest. Our task is therefore to build a program which is capable of identifying those citations in the cases from Westlaw which are referring to specific paragraphs of our case of interest and extract from them the paragraph numbers.

As we have established in Section 3.4, Adedjouma et al.'s research, utilising NLP, yields superior results to that of Tran et al. using ML approach on the same problem [32, 2]. There are however the two advantages of the ML approach over the NLP approach Tran et al. highlights in their paper we need to address.

The first advantage of Tran et al.'s approach is the ability to identify the smallest portion of text the citation refers to. They call this smallest portion a non-normative reference. While Tran et al. focus on the non-normative references is advantageous in the cross reference resolution of legislative texts, where the intention of the drafters of such legislation is to refer to a specific sentence but instead they end up citing a full paragraph, when judges cite other cases, they purposefully choose the level at which they cite. If a judge cites a paragraph, the entire paragraph is up for an interpretation. The citation style is therefore normative by its design, for the purpose to relieve the judiciary from the fear of potentially changing the law by citing only a specific section, or even worse paraphrasing it. This makes the first motivation behind the statistical approach to cross references not applicable on the domain of case law.

The second reason Tran et al. decided to use statistical methods, is to be able to analyse the legal text without any domain knowledge. While lack of domain knowledge is not a problem the author of this project is facing, the implication is that statistical methods could lead to a domain independent reference identifier. Even an identifier which could 'just' adjust to the inevitable changes of the reference style of the judiciary, would be an improvement over the rigid rule based identifier described by Adedjouma et al.. This prospect of Tran et al. work is appealing to this project.

However, the statistical approach seems to yield significantly worse results than the approach using the regular expression, with 80% accuracy at the core task of identification, they demonstrate over 19% less accuracy than what Adedjouma et al. report. Despite its potential to be more versatile, for the purposes of this project the accuracy takes precedent over versatility and so does the rule based approach. A detailed explanation of how we adapted Adedjouma et al.'s work for our purposes, analysing the differences and similarities between case law and statutes, is in Section 5.2.

4.7 Summary

This Section presented and justified the methodology we employ to automatically identify ratio. Our method involves two tasks, one identifying principles and two identifying cited paragraphs. We show an empirical study to support our hypothesis and evaluation on how to apply previous research to resolve the two tasks we are facing. The following Section will elaborate on the actual implementation of this methodology.

5 Implementing automatic ratio identifier

To automate the methodology described in the Section 4 above we resolved two problems. First, how to automatically identify principles and second, how to automatically identify cited paragraphs. Each is explored in a subsection below. Finally, at the end of this Section, we combine the two solutions to automatically identify the ratio. The actual results are presented and evaluated separately in Section 6.

5.1 Implementing principle identifier

To identify principles we apply and improve upon Shulayeva et al.'s framework [1]. Before improving on their model we first recreate it and reconstruct their data set, and as we report in the next Section, this already improves over Shulayeva et al.'s results. Then we implement the five new sentential features identified in Section 4.5 above. The new features result only in a small performance improvement. Finally, we relabel the Gold Standard corpus refocusing it on principles to create the New corpus. This leads to a major performance improvement.

5.1.1 Replicating Shulayeva et al.'s Golden Standard corpus

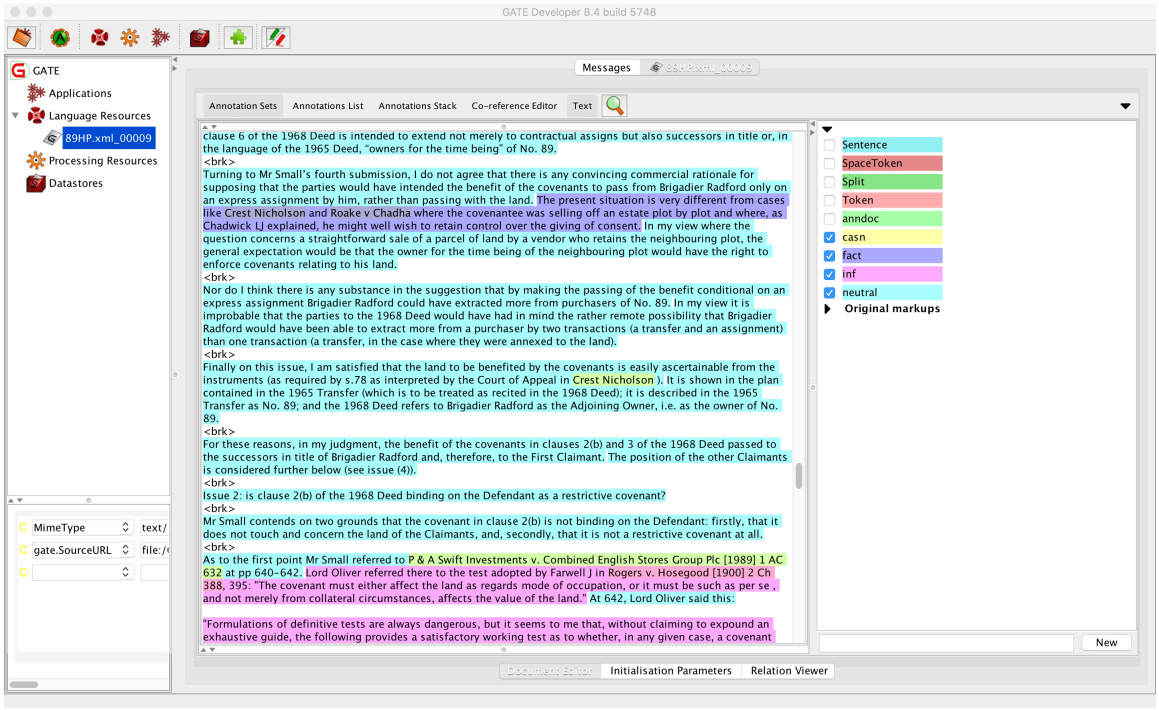

Figure 15: *Golden Standard corpus in GATE.*

Shulayeva et al. have kindly provided us with their Golden Standard corpus. This corpus of 50 English cases have been collected by Shulayeva et al. from the British and Irish Legal Institute (BAILII) website in RTF format [1]. Their paper informs us that *"most reports used for this study only provided the leading opinion and were narrated by the court in the form of monologue speech"* [1]. The topics in these cases range in issues concerning contract, trust and property law, and vary in length [1].

However, the files we obtained were not the final corpus. Instead they were the cases themselves saved in the Unix executable format, which is a format MacOS computers assign to a file with an unknown type. The first task was therefore to open these files and recreate the Golden Standard corpus. Shulayeva et al. reports that their dataset was annotated in GATE[12], a widely used text analysis tool [1]. After changing the file extension to .xml, it turned out that the files were indeed in a proprietary GATE XML formatting. Otherwise the dataset was as described by Shulayeva et al., with annotated facts, principles and neither areas of each case. Fig. 15 shows an example of a case opened in GATE. Curiously, Shulayeva et al. labeled principles as *inf.*

```
% 1. Title: Iris Plants Database
%
% 2. Sources:
%      (a) Creator: R.A. Fisher
%      (b) Donor: Michael Marshall (MARSHALL%PLU@io.arc.nasa.gov)
%      (c) Date: July, 1988
%
@RELATION iris

@ATTRIBUTE sepallength  NUMERIC
@ATTRIBUTE sepalwidth   NUMERIC
@ATTRIBUTE petallength  NUMERIC
@ATTRIBUTE petalwidth   NUMERIC
@ATTRIBUTE class        {Iris-setosa,Iris-versicolor,Iris-virginica}
```

The **Data** of the ARFF file looks like the following:

```
@DATA
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
5.4,3.9,1.7,0.4,Iris-setosa
4.6,3.4,1.4,0.3,Iris-setosa
5.0,3.4,1.5,0.2,Iris-setosa
4.4,2.9,1.4,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
```

Figure 16: *Weka ARFF example, the top half is the head of the file, the bottom is where the actual data goes.*

Before recreating their ML framework in Weka[13], a java ML toolkit used by Shulayeva et al. [1], we needed to reformat the data to an ARFF file, which is the proprietary format required by Weka. Fig. 16 shows the example of an ARFF file from Weka documentation. As per Fig. 16, an ARFF file is made of two parts, first part is the head, which describes individual attributes by a type (i.e. numeric, string etc.), where the class attribute contains the possible labels for the data (i.e. for Shulayeva et al.'s Golden Standard corpus it would be *"{Fact, Principle, Neither}"*). The second part is the data itself, which is formatted as comma separated values, where one row represents an instance of data, in our case a sentence, and one column represents one feature (referred to as attribute in ARFF head). Some columns need to be contained in quotation marks, for example if the attribute is of a string type, but others must not be in quotation marks, for example if the attribute is numeric. Otherwise the file is not accepted by Weka as well formed ARFF. The

[12]GATE Developer 8.4.1 - https://gate.ac.uk
[13]Weka 3.8 - http://www.cs.waikato.ac.nz/ml/weka/index.html

paragraphs below describe how we first extracted all the features, or columns, and then how we combined them to get an ARFF file.

To convert the GATE files into an ARFF file, we needed to first export each case individually as an inline XML file (as opposed to the proprietary GATE XML, which is poorly recognised by the tools we use to navigate the XML later on). This was done manually by opening each file in GATE and exporting it from there as an inline XML (GATE gives users two options for saving files, one is GATE XML, the other is Inline XML). We then needed to extract the sentences labelled as principles, facts and neither from these XML files. In the dataset, every sentence which was not a fact or a principle was labelled as neutral. This was a problem, because Shulayeva et al. only count the sentences in annotation areas, i.e. paragraphs with at least one citation, as part of their corpus [1], making the majority of neutral labelled sentences not part of their corpus.

To extract only the neutral sentences in citing paragraphs out of the dataset, in order to stay true to the Golden Standard corpus, we wrote a script in Python to extract only paragraphs with a citation. Our *corpusparser.py* program uses split("<brk>") function, where *"<brk>"* is the indicator of a paragraph in our .xml files, to identify individual paragraphs. If a paragraph contains *"casn"* tag, indicating a citation, it is further processed. Our program then uses the Beautiful Soap[14] library to find the sentences labeled as fact, inf and neutral in the XML structure of the selected paragraphs. All the extracted sentences are written in a text file, henceforth referred to as unigrams, with each sentence inserted on an individual line. The unigrams had been further manually checked and corrected for extra line breaks, that sometimes occurred in the source data. We have extracted into another file, in the same order, the classes (i.e. fact, principle, neither) corresponding to each sentence, again line by line. For details see *corpusparser.py*.

We then focused on extracting Shulayeva et al.'s other features [1], each feature was saved, maintaining the order of the files above, as a separate text file under the name of the feature (i.e. POS feature was saved as pos.txt). Shulayeva et al. briefly report on extracting these features: *"We used NLTK (Bird 2006) to extract part of speech tags and Stanford CoreNLP (Manning et al. 2014) to extract grammatical relations or dependencies. The other features were derived by means of a python script."* [1]. We therefore use the same tools for the same task below, keeping our method consistent with Shulayeva et al.'s..

To extract the part of speech (POS) tags we apply NLTK[15] toolkit in Python, processing and writing each unigram sentence as POS tags in a separate text file. The list of possible POS tags and explanation of what they mean is in Fig. 18 below. For details see the *pos.py* program.

Dependency pairs: case-that, is-there, concern-was, concern-case, case-was, parties-the, condition-a, court-the, is-if, letter-the

Figure 17: *Shulayeva et al.'s examples of dependencies [1].*

The dependency feature Shulayeva et al. employs has been extracted using the CoreNLP[16] Java toolkit. It was important to use the command *-ssplit.eolonly*, when running CoreNLP, otherwise

[14]Beautiful Soup 4.6.0 - https://www.crummy.com/software/BeautifulSoup/
[15]Natural Language Toolkit 3.2.4 - http://www.nltk.org
[16]Stanford CoreNLP 3.8.0 - https://CoreNLP.github.io/CoreNLP/

the CoreNLP would further parse the sentences on each line, often wrongly, instead of only taking each line as an individual sentence. This toolkit returns the annotated file in an XML format. We therefore had to write another Python script, this time using ElementTree library, which comes as part of Python 3, to extract the dependencies. Our program runs the CoreNLP by invoking it using *os.system* function, and setting the necessary flags for extraction of dependencies, using unigrams as an input. The XML file generated by CoreNLP contains the dependency pairs where each word has both a *governor* and a *dependent*. Our program extracts both and combines them into a single word. This is an important step, not described by Shulayeva et al.. By gluing the two words together we create for each word a feature capturing the relationship of the word in the sentence. We know this is the correct way of handling dependencies, because it corresponds to the examples of dependencies reported by Shulayeva et al. [1]. We include these examples here in Fig. 17. For more details see *dependencies.py*.

POS tag list:

CC coordinating conjunction
CD cardinal digit
DT determiner
EX existential there (like: "there is" ... think of it like "there exists")
FW foreign word
IN preposition/subordinating conjunction
JJ adjective 'big'
JJR adjective, comparative 'bigger'
JJS adjective, superlative 'biggest'
LS list marker 1)
MD modal could, will
NN noun, singular 'desk'
NNS noun plural 'desks'
NNP proper noun, singular 'Harrison'
NNPS proper noun, plural 'Americans'
PDT predeterminer 'all the kids'
POS possessive ending parent's
PRP personal pronoun I, he, she
PRP$ possessive pronoun my, his, hers
RB adverb very, silently,
RBR adverb, comparative better
RBS adverb, superlative best
RP particle give up
TO to go 'to' the store.
UH interjection errrrrrrrm
VB verb, base form take
VBD verb, past tense took
VBG verb, gerund/present participle taking
VBN verb, past participle taken
VBP verb, sing. present, non-3d take
VBZ verb, 3rd person sing. present takes
WDT wh-determiner which
WP wh-pronoun who, what
WP$ possessive wh-pronoun whose
WRB wh-abverb where, when

Figure 18: *Guide to POS tags.*

There are three types of Stanford dependencies [41]. The *standard dependencies*, *enhanced dependencies* and *enhanced dependencies++*. The standard dependency is a surface-structure dependency tree and as such *"tend[s] to follow the linguistic structure of sentences too closely and*

frequently fail[s] to provide direct relations between content words" [41]. This is why Stanford introduced enhanced dependencies, which are supposed to make *"implicit relations between content words more explicit by adding relations and augmenting relation names"* [41]. These have been further improved by enhanced dependencies++, which improve recognition of partitives and light noun constructions, multi-word prepositions, conjoined prepositions and prepositional phrases and relative pronouns [41]. Since Shulayeva et. al. do not specify which type of dependency they use, we choose the enhanced dependencies++, as they are the most sophisticated in capturing the relationships between words [41]. As will become apparent in Section 6.1, this decision most likely contributes to the improvement of our recreated model over the originally reported results by Shulayeva et al..

The length of a sentence feature was extracted using another Python script. Each sentence was split on an empty character, using *split()* method. The length of the resulting list gave us the number of words. See *length.py* for details.

The position feature was a little harder. Using a Python script we identified for each sentence a position of a paragraph in the case and given it a value corresponding to the index of its paragraph divided by total number of paragraphs in the case. While the XML files did not integrate paragraphs as part of their tree structure, they could be separated into paragraphs by using the aforementioned *split("<brk>")* function. Much like with extraction of unigrams we extracted the number of paragraphs only for sentences in annotation areas by checking if a paragraph contains a citation. The position value was rounded to a single decimal place, as per Shulayeva et al. [1], leaving us with position indicated on scale of 0 to 1. See *position.py* for details.

For full case citation feature, we went through the original dataset, much like with unigrams and positions, finding the paragraphs with citation, already annotated in the data we received from Shulayeva et al. as *casn*, and then for every sentence in these paragraphs checking if it contains *casn* tag, using Beautiful Soup library. If it did contain a citation, a 1 was written for this sentence in the *citation.txt* file, if it did not a 0 was written instead. We were again careful to maintain the order in the *citation.txt* file to match that of the unigrams. For details see *citation.py*.

Finally, we have combined the files together using *combine.py* and *adjust.py* programs. This is where the maintained order of the files containing the features is important. Each feature must correspond to the sentence it was extracted from, otherwise the corpus would be invalid. Since we do maintain the order throughout, we can simply read the files into a table, where a row corresponds to a single sentence of a case, and a column is a single feature extracted from that sentence.

The *combine.py* program uses the Pandas library[17] to read the text file in a data frame corresponding to the ARFF file @DATA notation. As stressed above, each feature is read as a single column. Since all text files are in the same order, with each sentence on a separate line, the data frame contains in each row the features corresponding to a single sentence of the dataset. This data frame was then saved as a text file, containing the comma separated value formatting required by the ARFF format.

[17]Pandas 0.20.3 - http://pandas.pydata.org

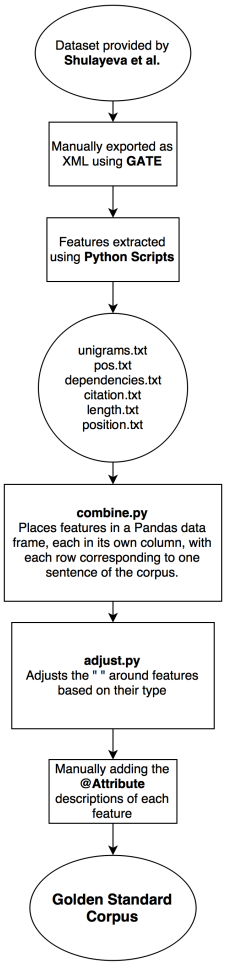

Figure 19: *The process by which we formatted the raw data supplied by Shulayeva et al. into a dataset replicating their Golden Standard corpus.*

The issue that arose was that by default Pandas *to_csv()* function, used to save data frame as a text file, did add quotation marks only around some of the strings, creating an invalid ARFF file. ARFF file requires that you specify the type of each column in the head of the case. As per Fig. 16, strings must be quoted, numerical values must not. However, Pandas does not allow detailed specification of which columns are to be quoted, and which are not, when exporting the data frame. We have therefore exported the data frame with all columns quoted, later removing the quotes from numeric values and types (i.e. fact, principle and neither) using our *adjust.py* program. This program finds, using regular expression, all our numerical features, such as lenght, position or citation, and removes the quotation marks from around these columns. For detail see *adjust.py*

We manually added the necessary description of attributes, forming the head of the file, and saved the file as an ARFF document. The resulting file is our replicated Golden Standard corpus. The whole process is illustrated in Fig. 19.

5.1.2 Replicating Shulayeva et al.'s ML model

With the dataset recreated, we have implemented the Shulayeva et al.'s model in Weka toolkit[18] in Java. As per Shulayeva et al., we *"relied on automatic feature selection to prune the feature set, ran a single machine learning algorithm with default settings and report results using a cross-validation methodology, as detailed below:"* [1]

1. *Feature counts were normalised by tf and idf.*
2. *Attribute selection (InfoGainAttributeEval in combination with Ranker (threshold = 0) search method) was performed over the entire dataset.*
3. *The Naive Bayes Multinomial classifier was used for the classification task. This has been widely used in text classification tasks (Teufel et al. 2006; Mitchell 1997), and its performance is often comparable to more sophisticated learning methods (Schneider 2005).*
4. *Results are reported for tenfold cross-validation. The 2659 sentences in the dataset were randomly partitioned into 10 subsamples. In each fold one of the subsamples was used for testing after training on the remaining 9 subsamples. Results are reported over the 10 testing subsamples, which constitute the entire dataset.*

There are however few steps missing in Shulayeva's description we needed to discover in order to replicate their results [1]. Firstly the *TF* and *IDF* normalisation is part of a broader step of vectorising the data, which transforms our unigrams, POS tags and dependencies, into a vector representing word occurrence. To both vectorise and normalise data, Weka already contains the *StringToWordVector* function. However to also prune the feature set a *MultiFilter* function must be applied, which takes as an argument both the *StringToWordVector*, but also the *AttributeSelection* method. The *AttributeSelection* method then takes in as an argument the *InfoGainAttributeEval* and *Ranker* functions. Additionally, the number of words to keep for the *StringToWordVector* needs to be set at least to 3000 in order to get 85% accuracy (to be safe, we set ours to 5000). This is as opposed to the default value, suggested by Shulayeava et al.'s description of methodology, which is only 1000 and performs significantly worse, see Table 15 in the Appendix. For details see *Classifier.java*.

We have considered using different ML toolkits. In particular we have started implementing the model using Scikit-learn[19], because it is modern, build in Python and widely used. However, we found that the attribute selection process used by Weka, particularly the InfoGainAttributeEval method in the implementation, does not have a direct counterpart in Scikit-learn. While it would be possible to adapt similar method in Scikit for the same effect, it seemed unnecessary, since Weka does have a well documented and supported java API and the methodology is described by Shulayeva et al. in terms of individual Weka functions [1]. We therefore decided to use Weka in the end as well.

5.1.3 Implementing new features

The implementation of the new features, which we have identified for this task in Section 4.5, is below. Like Shulayeva et al. we use Python scripts, NLTK and CoreNLP to extract our features [1]. Much like above, we have saved each in a separate text file, maintaining the order of the rest of the

[18] Weka 3.8 - http://www.cs.waikato.ac.nz/ml/weka/index.html
[19] Scikit-learn v0.19.0 - http://scikit-learn.org/stable/

corpus, and later combined them in an ARFF file as extra columns via our *combine.py* program. The results and evaluation are in Section 6.1.

1. **Lemmatisation:** We have lemmatised the unigrams using the CoreNLP parser. Since the parser returns XML file, we have used a Python script to extract the lemma from the XML, much like we have done for the dependencies. The lemma of each word was placed in the same order of the original sentence. See *lemma.py* for details.

2. **Quotation:** Quotation captures those sentences which contain text in quotation marks. We have written a Python script, using regular expression, to find instances of quotation marks. We looked for both " and " in a each sentence, if the result was positive, we wrote 1 if not we wrote 0. For details see *quote.py*.

3. **Paragraph Citations:** We wrote a regular expression identifying paragraph citations. Since we have primarily done this for identifying cited paragraphs, the detail explanation of the process via which we arrived at the regular expressions is in Section 5.2 below. If a paragraph citation was present in a sentence, we wrote 1, otherwise a 0. For details see *paras.py*.

4. **Sub-Clauses:** Sub-clauses were identified by counting the number of commas in a sentence. Much like with counting the length of a sentence, we use split() to divide the sentence on commas and count the length of the list produced. These were recorded as an integer. See *sub.py* for details.

5. **Named Entity Recognition:** We used CoreNLP to identify named people, places, dates, organisations and numbers. Again, the results from XML were extracted with a Python script and as with all features written in a text file, sentence by sentence, in the same order as unigrams. We record each NER in binary, 1 for it being present, 0 for not.

   While implementing this feature, we ran into a problem with the length of the files in the corpus being too large, causing CoreNLP to run out of memory. CoreNLP API[20] recommends using commands to specify only the features a user wants annotated to mitigate this issue, however we already specify only the bare minimum of features necessary while using CoreNLP throughout our project and still met with this problem. The API also suggests splitting the files into smaller sections in order to avoid running out of memory, which is the solution that made it possible for us resolve our large corpus. We have applied our *ner.py* program section by section on roughly 500 lines of our unigrams at a time to resolve this issue, combining the results together after the whole corpus was processed.

5.1.4 Constructing new corpus

Since Shulayeva et al.'s Gold Standard corpus is manually annotated for extraction of cited facts and principles, we re-annotate it to retrain their ML model on principle identification only. This was also done to address the two issues Shulayeva et al. report in their error analysis [1], see Section 4.5 for description of the problem and Section 6.1.3 for our evaluation on how we dealt with it. We relabelled 96 sentences originally labelled as neutral or fact, 4% of the Golden Standard corpus, as principles and created the New corpus. We again implemented Shulayeva et al.'s framework in Weka according to their instructions, as described above, and trained it on the New corpus. We report our results in Section 6.1.

[20]Stanford CoreNLP API - https://CoreNLP.github.io/CoreNLP/

Our New corpus annotates certain sentences differently from Shulayeva et al.. For example, Shulayeva et al. have decided to annotate sentences with both facts and principles as facts. We on the other hand annotate them as principles. An example of a sentence Shulayeva et al. would annotate as a fact that we re-annotated is below, the bold part of the sentence is a principle, while the rest is a fact.

"Thus for example if the manager explains either when making the request for payment to a third party or on it being questioned by the customer that the third party is a supplier and that the object is to obtain necessary materials for the work more quickly or that the third party is an associated company carrying on the same business ***such an explanation might well bring the request for the payment to the third party within the usual authority of a person in his position and therefore within his apparent authority****"* [1]

In contrast to Shulayeva et al., we also believe it is better to extract the ratio from the paragraphs of the given case to which the subsequent citation refers (i.e. cited paragraph B in Fig. 1) than from the text of the subsequent case in which the citation occurs (i.e. citing paragraph A in Fig. 1), see Section 4 for our reasoning behind this decision. Shulayeva et al.'s approach, which tries to find principles of cited cases in the citing paragraphs, results in arguably mislabeling principles which are not part of the reasoning in the cited case, as neither. For example the sentence below would be labeled neither, because the principle contained in this sentence originate in the citing case (as opposed to the cited case), yet it is clearly a principle for our purposes. We have therefore relabelled such sentences as well.

"He is not obliged to comply with any request that may be made to him by the borrower let alone by a surety if he judges it to be in his own interests not to do so." [1]

In identifying principles we have followed Shulayeva et al.'s definition, but ignored the distinction between cited and novel principles mentioned above. According to Shulayeva et al., a legal principle is an argument used with facts to reach a conclusion. This must not be mistaken with our definition of ratio, which, despite being similar, is focused on a broader story the case is trying to tell. A principle, unlike ratio, might argue to any conclusion, not necessary leading to the outcome of the case. As Shulayeva et al. note, a principle is indicated by deontic modality *"expressions of must for obligation, must not for prohibition, or may for permission, which contrast with epistemic modalities for necessity and possibility."* [1]. They also provide an example of a principle, which we have used for a comparison with the principles we have identified, to ensure as much consistency as possible in our relabelling of data. The paragraph below contains this example.

"As a matter of principle no order should be made in civil or family proceedings without notice to the other side unless there is a very good reason for departing from the general rule that notice must be given (Gorbunova v Berezovsky (aka Platon Elenin) & Ors, 2013)." [1]

Shulayeva et al.'s ML model was applied on the New corpus, implemented above. Since we have only changed the class attribute in the ARFF file, all the features extracted remained unchanged and there was no need to re-extract them.

5.2 Implementing cited paragraph identifier

To identify cited paragraphs we have analysed 798 citations in Cited corpus, created by us for our empirical study in Section 4.4. These citations come from 41 cases manually downloaded from Westlaw. They represent the full list of important cases citing Stack v Dowden. We know this because such list is provided by Westlaw with every major case, such as Stack v Dowden. The 41 cases in the corpus are concerned with different issues, some citing to distinguish their problem, some to criticise it, and some to apply the law. They also come from a variety of courts and judges. Together, this gives us a representative sample of the variety of approaches judges and transcript writers could use to cite another case. This corpus is also comparable in size with Shulayeva et al.'s Golden Standard corpus, which uses 50 cases. The results are reported in Section 6.2.

para(.) paras(.) paragraph paragraphs	**followed by:**	*n* *n(1)* to *n(x)* *n(1)* and *n(2)* *n(1)* - *n(x)* *n(1), n(2)*

Figure 20: *Possible combinations for standard cited paragraph notation.*

at	**followed by:**	*[n]* *[n(1)] – [n(x)]* *[n(1)] to [n(x)]*

Figure 21: *Possible combinations for squared cited paragraph notation.*

Identifying cited paragraphs can be broken down in three steps. First the paragraph citations must be identified in citing cases. Second, the number of paragraphs must be extracted from these citations. Third, a schema must be devised to establish which of these cited paragraphs should be attributed to our case of interest. We implement each step below.

5.2.1 Identifying paragraph citations

Just like Adedjouma et al. we start by identifying all the patterns possible for recognising the citation itself. Just like they identify for cross references, we find paragraph citations in case law can be divided into simple and complex citations. The terminology used below purposefully follows that of Adedjouma et al. [2]. The discussion serves as a close comparison between paragraph citations and cross references as well as an explanation of our implementation. As such it is best read together with Section 3.4.

First we consider the simple citations. While simple cross references can be either explicit or implicit, all citations in law are always explicit. Unlike the cross references however, citations can be further subdivided as either *standard* or *squared*. The standard simple citations provide a paragraph marker, which is *paragraph*, *para* or *para.*, followed by a number, for example *"para 4"*. Unlike a cross reference however, the number is always arabic, never roman, text or a letter. There is also a square simple citation style used in some of the judgements, were paragraph marker is *at* and the number of paragraph cited is presented immediately afterwards in square brackets, for

example *"at [2]"*.

Second we consider the complex citations. There are only multivalued complex citations, unlike the cross references which can also be multilayered. Just like with cross references, the complex citations enhance simple citations with enumerations and ranges, but unlike cross reference they lack navigation through levels, since paragraphs are always on the same level. This is why paragraph citations can only be multivalued but can't be multilayered. The numerical expressions in multivalued citations are nearly the same as in cross references. There are only three options:

1. An enumeration, for example *"paras. 1, 2, 3"* or *"paras. 1 and 3"*.
2. A range, for example *"paragraphs 1 to 3"* or *"paragraphs 1 - 3"*.
3. A combination of both of the above, for example *"paragraphs 1, 2 and 3"* or *"paragraphs 1, 2, 3 to 20 and 30"*.

The only difference between citations and cross references in this respect is the lack of *or* enumeration in paragraph citations. There are no instances we came across where a judge would cite *paragraph 1 or 2*.

Further, while all the options above apply to the complex standard notation, the complex square notation used in some cases is limited only to option 2, since when manually collecting the data we did not find any examples corresponding to options number 1, and 3. However, these could be easily added to expand our model if necessary. For an illustration of both simple and complex citations, see Fig. 20 and 21.

The analysis of case law citations above lead to the development of the two regular expressions below. The first one captures all the simple and complex standard paragraph citations, the second captures all the simple square paragraph citations as well as complex option 2 above, adapted for square notation, for example *"at [1] to [4]* or *"at [1] - [4]"*. We illustrate these in Fig. 22 and 23.

- `para(graph)?s?\.?(((\s)|(,)|(-)|(?)|(and)|(to)|(\d+)|(\.))+)`
- `at\s(((\[\d{1,3}\])|(\s\-\s)|(\s?\s)|(\sto\s))+)`

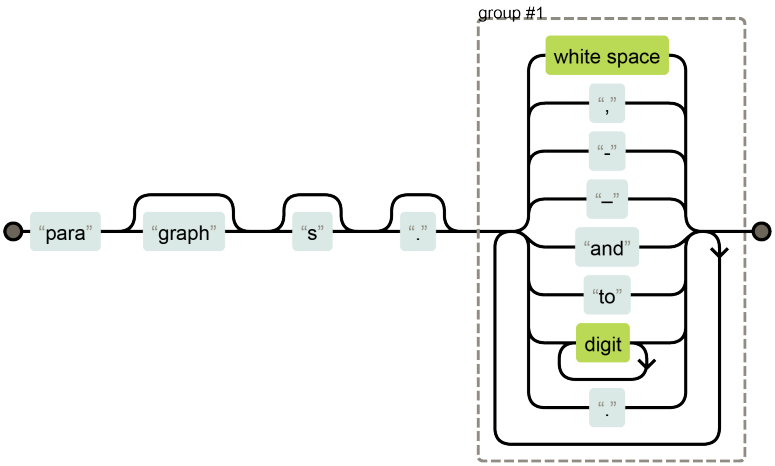

Figure 22: *Regular expression capturing the standard citations.*

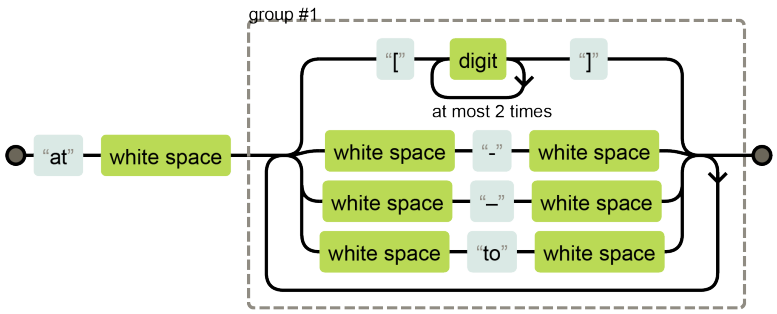

Figure 23: *Regular expression capturing the squared citations.*

5.2.2 Resolving paragraph citations

As with Adedjouma et al.'s research the second step is to break these citations into individual numbers. To do this, we identify instances where a range is suggested by *"-"* or *"and"* and expand such ranges, filling the missing numbers. For example, *"paras. 1 to 3"* is expanded into a list of numbers *"[1, 2, 3]"*. The enumerated paragraphs get split by column into a list of individual numbers, for example *"paragraphs 4, 5, 6"* become *"[4, 5, 6]"*.

5.2.3 Attributing paragraph citations

This is as far as Adedjouma et al.'s research could serve as a guide. The remaining task was how to connect the paragraph numbers with the cases they refer to. For example in a sentence *"In paragraph 42, Lady Hale rejected this approach."* to identify that paragraph 42 is from Stack v Dowden, where one of the judges is Baroness Hale. For our purposes this task could be simplified by focusing only on attributing those paragraphs that belong to the case of interest, as opposed to connecting all paragraphs with all cases. This makes the task significantly easier because we know the name of the case we are looking for and the names of the judges that made judgement in that case. For majority of citing sentences, the name of the case, or its abbreviation, for example *Stack* for *Stack v Dowden*, is enough to attribute the paragraphs. A usual sentence will have the name of the case as well as the cited paragraphs. Yet, as we see from the example above, sometimes instead of the name of a case, a name of a judge, or its abbreviation is used. Knowing the names of the judges in the case of interest, which are easily extracted from below the title of the case, makes it seem easy to attribute even these sentences.

However, there is a problem. Judges do judge many cases. And therefore a judges name is not exclusive for a single judgement. Going back to the example above, how can we tell if the paragraph containing the citation and Lady Hale's name, is in the case of interest as opposed to any other case? The simple answer is we can't for certain, however in our approach we work with the knowledge the case we analyse is citing the case of interest at some point. We know this since the cases we are analysing are selected from the list of cases citing Stack v Dowden, Westlaw provides, see Section 4.6. Therefore assuming that the case cited with a judge of the case of interest is indeed the case of interest, is a reasonable assumption to make. And as we report in Section 6.2, while a weakness of our approach, it still allows us to identify cited paragraphs with very high accuracy. Moreover, this is the best we can do without engaging with full analysis of the paragraph, or indeed the full case, which would be necessary to fully resolve this problem.

There are two ideas we implement to mitigate this shortcoming of our approach. One, we identify those instances where there is a full citation in the sentence citing paragraph which is not a citation of the case of interest. For example *"In paragraph 42, Lady Hale rejected this approach in Jones v Kernott."* we can identify the *Jones v Kernott* citation and, since it is a full case citation, conclude the paragraph is cited from Jones v Kernott, instead of assuming Lady Hale's judgement is from the case of Stack v Dowden.

Two, while there are no implicit paragraph citations as Adedjouma et al. describes them (i.e. *"this paragraph"* or *"following paragraph"*), there are indicators of internal citations, which are similar to implicit cross references identified by Adedjouma et al. [2]. These internal citations can be explicit or implicit. Explicit ones will have simple or complex citation followed by *"above"*. Implicit ones will in the same sentence as a citation say *"this court"* or *"present case"*. We can use these to identify when the judge is referring to another part of the judgement. For example in the sentence: *"In paragraph 42 above, Lady Hale rejected this approach."*, we can infer from the internal paragraph citation that the sentence is not referring to our case of interest, but instead to Lady Hale's judgement in the same case from which the sentence comes from.

Employing these findings, we constructed and implemented the Schema in Fig. 24 and applied it on the Cited corpus, for details see *citations.py*.

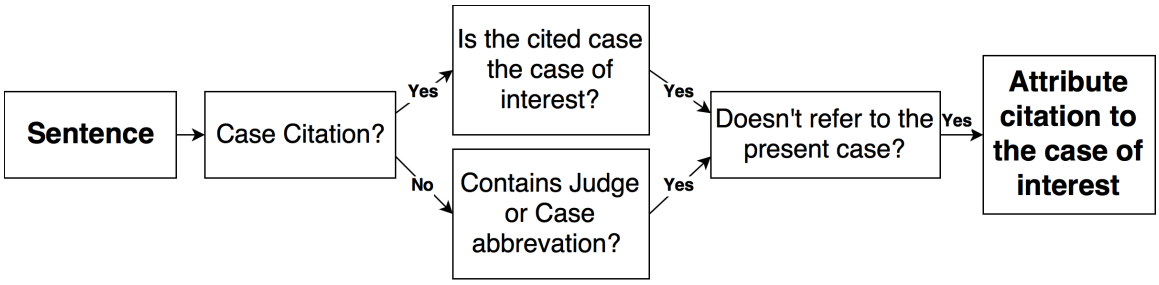

Figure 24: *Schema for attributing the citations to the case of interest.*

5.3 Implementing automatic ratio identifier

Finally, combining the principle and cited paragraph classifier described above, we create our program automating our hypothesis that ratio in a case can be identified by finding principles in cited paragraphs. As will become apparent in the next Section, the principle classifier trained on the New corpus improves the results significantly on its own, with just the core features of unigrams, dependencies and POS tags. The effect of the other features, both Shulayeva et al.'s and our own, have in fact slightly negative impact on the performance, decreasing the accuracy by around 0.1% depending on the combination of features used, see Table 16 in the Appendix for detailed results. Since we would need to extract all the extra features, some of which like NER potentially increase the time the program takes to run significantly, we decided for our final program to train and classify on the core set of features (i.e. dependencies, POS and unigrams), which makes it run fast and increases the accuracy. Instead we used the time to focus on building a pipeline which allows the user to simply insert cases as found on Westlaw and get annotated HTML in return. We describe our pipeline below, and illustrate it in Fig. 25.

Our final program takes a case in HTML along with all the cases that cite it, both of which

are readily available from Westlaw, as an input. It records the number of sentences per paragraph. It processes the case of interest to find principles, classifying each sentence as either a principle or none. It processes all the citing cases to find cited paragraphs in the case of interest, creating a list of cited paragraphs. Using this information the sentences of a case of interest are written back in a new HTML, retaining the paragraph structure by inserting a break where appropriate, using the number of sentences per paragraph recorded earlier. Paragraphs which are cited and contain at least one sentence classified as a principle are highlighted blue, while the principles in them are highlighted yellow. For details see *casetoratio.py*.

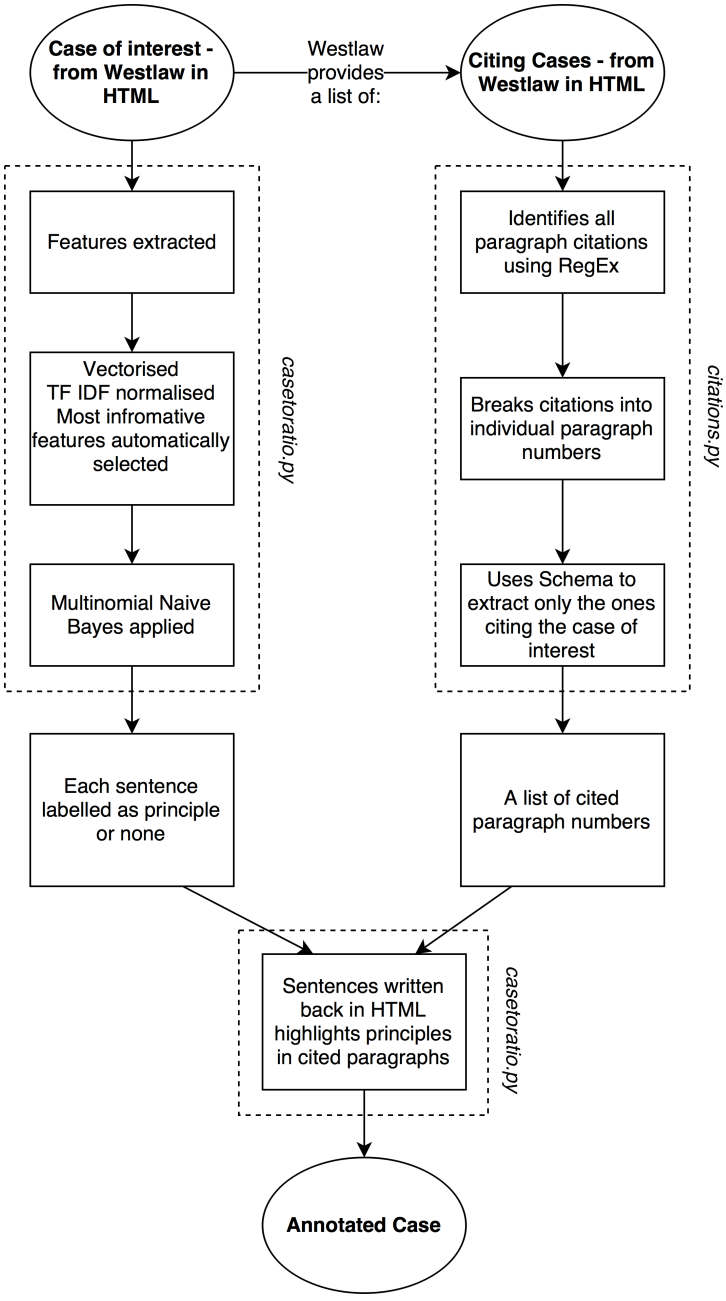

Figure 25: *The principle classifier and cited paragraph identifier described above are combined for attributing the citations to the case.*

5.4 Summary

In conclusion we have successfully implemented the automatic ratio identifier, dealing with several difficulties along the way. In particular, recreating the Golden Corpus from the data provided by Shulayeva et al. [1] and extending the cross reference research by Adedjouma et al. [2] on cited paragraph identification required careful consideration. We have combined our work to create a user friendly program, that when run from a command line breaks down a case, saved as HTML, into features, analyses all the citing cases and outputs an annotated HTML, visualising our findings.

6 Results and evaluation

In this Section, we presents the results we have achieved by implementing out methodology above and we evaluate our hypothesis articulated in Section 4, that finding principles in cited paragraphs can identify ratio decidendi. We report an improvement over previous research in principle identification and nearly equally high precision for the new task of cited paragraph identification as was achieved in the cross reference resolution research we base out method on. Finally, we evaluate our ratio identifier, discuss the limitations of our work as well as potential ways of building on it and reflect, in light of the results of the project, on our aim.

6.1 Principle identifier

We first evaluate the performance of our replicated model on our replicated corpus, reporting a slight improvement over Shulayeva et al. [1]. Then we report the performance of the new features, demonstrating another slight improvement in accuracy. Finally, we report a significant improvement in accuracy of 11%, achieved by training our ML model on the New corpus, demonstrating we have found a way of solving the issues Shulayeva et al. identified with their approach. For our evaluation we use the same statistics used by Shulayeva et al., including the κ coefficient[21], which measures the predominant agreement, to enable a clear comparison between our results and the ones reported by Shulayeva et al. [1].

6.1.1 Replicated Shulayeva et al. model

Table 5: Per category and aggregated statistics for the original Shulayeva et al.'s principle and fact classifier trained on the Gold Standard corpus.

Classified as →	Principle	Fact	Neither
Principle	646	5	160
Fact	4	198	41
Neither	135	38	1432
Type	**Precision**	**Recall**	**F-Measure**
Principle	0.823	0.797	0.810
Facts	0.822	0.815	0.818
Neither	0.877	0.892	0.884
Accuracy	0.85	κ	0.72

We have successfully implemented Shulayeva et al.'s ML model on the replicated corpus, and report slightly improved performance just on our implementation alone. The possible reasons for the difference of 1% in accuracy between the original Shulayeva et al.'s results in Table 5, and the

[21] κ is the predominant agreement measure that corrects raw agreement P(A) for agreement by chance P(E) [1, 15] :

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

results we report when recreating their model in Table 7, are investigated in the paragraphs below.

First, the difference could be in the version of software we are using. Unfortunately, since Shulayeva et al. don't report on their version of Weka, CoreNLP or NLTK in their paper, this can not be easily verified.

Second, from Table 6, where we present the detailed side by side comparison of the performance of the individual features, we can see our unigrams and dependencies outperform, while our POS tags underperform, Shulayeva et al.'s. Since our dependency feature particularly performs well with +2% increase in accuracy, while POS tags and unigrams balance each other out, dependencies are the most likely culprit in the overall increase of performance. The most likely reason for this is our use of the *dependencies++*, as opposed to just dependencies used by Shulayeva et al., see implementation above at 5.1.1.

Another reason might be the slight differences in our methodology. As per our Section on implementation above, we set the words to keep option, under the StringToWordVector function in Weka, at 5000. Shulayeva et al. on the other hand claim they use the default setting on every option, where not specified otherwise, which in this case would be only 1000. However, when we use the default setting of 1000 words, we only get overall accuracy of the model of 79%, see Appendix 15. This suggests the difference in our performance is perhaps rather due to the deviations between software versions and our use of the improved dependency feature. Shulayeva et al. might have simply omitted specifying this option in the description of their method, after all they barely spend half a page in their paper elaborating on how they implemented their software, and as we will see below, don't report individually on the performance of half of the features they claim to have implemented.

In conclusion we demonstrate that we can more than match Shulayeva et al.'s performance on their own task by replicating their methodology and present the results of our replicated ML classifier trained on our own replicated Golden Standard corpus, in the Table 7 below.

Table 6: Per features statistics comparing the performance of our replicated Shulayeva et al.'s principle and fact classifier with the originally reported results.

Our Features:	POS	Unigrams	Dependencies
Accuracy	0.62	0.78	0.83
κ	0.38	0.60	0.67
Shulayeva et al.'s Features:	**POS**	**Unigrams**	**Dependencies**
Accuracy	0.63	0.77	0.81
κ	0.18	0.58	0.63

Table 7: Per category and aggregated statistics for the replicated Shulayeva et al.'s principle and fact classifier trained on Gold Standard corpus.

Classified as →	Principle	Fact	Neither
Principle	675	4	167
Fact	5	194	47
Neither	122	37	1473
Type	Precision	Recall	F-Measure
Principle	0.826	0.804	0.815
Facts	0.809	0.809	0.809
Neither	0.878	0.890	0.884
Accuracy	0.86	κ	0.74

6.1.2 New features

We could have used the Shulayeva et al.'s classifier, in combination with the cited paragraph identifier, to test our hypothesis without doing any further work. However, we have set ourself the objective of improving on the performance of Shulayeva et al's model on principle identification, since we saw two areas where improvement seemed feasible. The first area we thought we could improve on was the selection of sentential features. We selected our new sentential features based on the sentential features that proved useful in related research on case law summarisation, see Section 4.5. We report on their performance below.

What is noticeable in Shulayeva et al.'s paper, is the lack of any specific data on how did three out of the six features they implement perform. While, as per Table 6, they present the individual accuracy and κ of the POS tags, unigrams and dependencies, henceforth referred to as core features, the other features they report as implemented (i.e. full citation, position in text and length of a sentence), go unreported. We will refer to these as extra features from now on. On the one hand, this is reasonable, since these extra features are meant to improve the performance of the model in conjunction with the core features, and therefore their individual performance is both not very informative and most likely very low. On the other hand, there is no way of knowing how did any of the extra features perform from Shualyeva et al.'s paper alone.

Since our new features are meant to improve the core features much like the extra features in Shulayeva et al.'s paper did, we decided to report on their performance when each new feature is used on top of the core features. Otherwise the benefit or detriment of their use would not be possible to determine. Since Shulayeva et al. did not, we also report, using the same method, the performance of the extra features from their paper [1]. As a base line to compare the improvement or detriment of each feature we use the accuracy and κ of our replicated classifier, trained on the replicated corpus reduced to only the three core features. This baseline is 85.61% and $\kappa = 0.73$ for accuracy and κ respectively.

This way we can evaluate how our features compare to their equivalents in Shulayeva et al.'s research. Even better we can also report on how does the combination of all new and extra features

improve the performance. The complete results are in Table 8.

The best performing new feature is quoted text indicator (Quotation), improving over the baseline by 0.40% achieving 86% accuracy. It is also the single most informative feature automatically selected by Weka. Since under Shulayeva et al.'s methodology each unique word is used as an individual feature, the performance improvement of even the most informative feature, such as Quotation, is limited. While 0.40% improvement in accuracy might seem only a small increase, especially compared to the improvements we report achieving by creating the New corpus below, it is only fair to compare it to the performance of Shulayeva et al.'s own extra features. As can be seen from Table 8, the 0.40% improvement is the performance of their maximum performing extra feature, which Weka evaluates as only the 5th most informative feature. For ranking of top 40 features in the dataset, based on their information gain, see Fig. 28 in the Appendix.

Table 8: The performance of new features when combined with Shulayeva et al.'s base features (POS, Unigrams and Dependencies).

Feature	Accuracy	Improvement	κ
Baseline	85.61%	+-0%	0.73
New Features			
Lemmatisation	85.72%	+0.11%	
Quotation	86.01%	+0.40%	
Paragraph Citations	85.61%	+-0%	
Sub-Clauses	85.50%	-0.11%	
NER - All	85.68%	+0.07%	
—– person	85.50%	-0.11%	
—– location	85.61%	+-0%	
—– date	85.61%	+-0%	
—– organisation	85.61%	+-0%	
—– number	85.43%	-0.18%	
Shulayeva et al.'s unreported features			
Full citation	85.54%	-0.07%	
Length	86.01%	+0.40%	
Position	85.61%	+-0%	
All Combined	**86.71%**	**+1.10%**	**0.75**

In the context of our project, the performance of Quotation feature suggests that perhaps on top of looking for cited paragraphs in cases, in order to discriminate between ratio and obiter, the future research might benefit from looking at quoted sentences as well, see Section 6.5. After all, just like paragraph citations, quoted sentences or passages are simply another way of pointing to a specific parts of previous cases, and are arguably easier to identify from citing cases than cited paragraphs.

Second highest performing feature was lemmatisation. Usually unigrams are lemmatised as part of preprocessing to improve the classifier performance, however such classification tasks are

not relying so much on tense and voice as our research, see Section 4.5. Our results show that lemma can, ever so slightly, improve the overall performance if combined as a separate feature with unigrams. As far as we know, this is a novel, or at least atypical approach to using lemma. We have tested using lemmatised unigrams in place of unigrams as well, finding the performance of lemmatised unigrams is lower by 1% and $\kappa = 0.03$, compared to using unigrams. This result proves Shulayeva et al.'s assumptions about the importance of tense and voice, since when tense and voice are removed via lemmatisation, the unigram feature performs less well. See Table 14, in the Appendix.

The third best performing new feature are the named entities. Our hope that NER feature will help distinguish sentences primarily introducing facts, and therefore improve the model performance significantly, was misplaced. In hindsight the underwhelming results of this feature does make sense in light of the confusion between facts and principles Shulayeva et al. report [1]. As perviously stated in Section 4.5, there is a considerable overlap between facts and principles, because principles are applied on facts in the same sentence. Therefore when looking at probability of the feature given the class, NER is suggestive of the neither class. In other words, sentences with named entities are more likely to be classified as neither, rather than as we hoped a principle or a fact.

Not all named entities we extracted are retained by the ML model. As we describe in implementation, all the features are pruned in Weka using the InforGainAtributeEval function in conjunction with the Ranker search method. As per Shulayeva et al., the Ranker threshold option is set to 0, discarding any feature that does not provide any information gain. After the feature pruning, of our ARFF file, only the features which overcome the threshold are used in the ML model. Some of our new features, as well as one of the Shulayeva et. al.'s extra features, were filtered our this way and therefore do not contribute to the overall performance at all. From our features, these were some of the named entities (location, data, organisation) and the paragraph citations. From Shulayeva et al.'s extra features, it was the position within the text. We can therefore conclude with confidence that these features are not useful for distinguishing facts from principles in case law.

Interestingly, the named entities that were not filtered out, slightly decreased the overall performance, when evaluated individually. The same applies to our sub-clause feature and Shulayeva et al.'s full citation feature. Despite the slight negative individual performance, these features can still improve accuracy when combined with other features, much like lemmatisation does when combined with unigrams. For example, as can be see in Table 8, when combining the person and number named entity, both slightly negative individually, we get an increase in accuracy.

We have tried all the combinations of the extra features and the new features, and found that the best performance is achieved when all of the features are combined (the aforementioned features that are pruned out by the ML model naturally do not count). This way we get an increase of over 1% in accuracy and a jump in κ from 0.73 to 0.75, see Table 8. If we were only to combine the two highest performing features (i e. the new quotation feature with Shulayeva et al.'s length feature) the accuracy increases only by 0.70% and κ by 0.01, see Table 17 in the Appendix.

Table 9: Per category and aggregated statistics for Shulayeva et al.'s principle and fact classifier trained on Gold Standard corpus with new features.

Classified as →	Principle	Fact	Neither
Principle	693	4	149
Fact	7	191	48
Neither	117	37	1478
Type	Precision	Recall	F-Measure
Principle	0.848	0.819	0.833
Facts	0.823	0.776	0.799
Neither	0.882	0.906	0.894
Accuracy	0.87	κ	0.75

Therefore our investigated features, except the ones that are filtered out by Weka, do increase the overall accuracy and κ of Shulayeva et al.'s model. Some, like Lemmatisation or Quotation, archive this individually, others, like NER and sub-clauses only in conjunction with the rest. The new sentential features performed best when combined, improving accuracy by over 1% and κ coefficient by 0.03.

6.1.3 New corpus

The second area of improvement came directly from Shulayeva et al.'s error analysis [1], see Section 4.5. Shulayeva et al.'s error analysis revealed the issue of fact and principle overlap and the issue of inability to distinguish cited from novel principles. We demonstrate that it is possible to fix these issues by relabelling the corpus, re-focusing it only on principles, but all principles (as opposed to on both facts and principles, but only on the restated facts and principles originating from cited cases, which is what Shulayeva et al. did, see Section 2.5 for more detail on this fine distinction). The accuracy increased to 96% and κ to 0.90 an 11% increase in accuracy over Shulayeva et al.'s research. The full breakdown of the results is in Table 10.

Table 10: Per category and aggregated statistics for Shulayeva et al.'s classifier trained on New corpus focused on extraction of principles only.

Classified as →	Principle	Neither	
Principle	837	70	
Neither	48	1769	
Type	Precision	Recall	F-Measure
Principle	0.946	0.923	0.934
Neither	0.962	0.974	0.968
Accuracy	0.96	κ	0.90

Because our New corpus focuses only on the principles, a different task from Shulayeva et al.'s, the new features we explored above on the Golden Standard corpus as well as the extra features used by Shulayeva et al. (i.e. all features apart from unigrams dependencies and POS tags), contribute minimally to the overall performance. In fact, when these features are combined together in the same way they were combined to improve the performance with the Golden Standard corpus, but are used on the New corpus instead, the performance ever so slightly decreases from 95.668% to 95.595%, in comparison to when they are not used at all. See Table 16 in the Appendix. Some of the new features, like NER, also require longer time to extract, see Section 5.1.3. Therefore, to increase the speed and performance, our final principle classifier, the one we used in combination with the cited paragraph identifier to annotate ratio in case law, uses only the core features of unigrams, part of speech tags and dependencies.

Table 11: Per features statistics comparing the performance of our principle classifier trained on the New corpus with the originally reported results in Shulayeva et al. [1].

Our Features:	POS	Unigrams	Dependencies
Accuracy	0.69	0.89	0.94
κ	0.15	0.75	0.87
Shulayeva et al.'s Features:	**POS**	**Unigrams**	**Dependencies**
Accuracy	0.63	0.77	0.81
κ	0.18	0.58	0.63

When inspecting the performance of using only the core features individually, trained on the New corpus side by side with Shulayeva et al.'s results, the improvement across features is clear. See Table 11. Yet notably, POS tags, while increasing in accuracy, decrease in κ coefficient slightly. Otherwise all features of the New corpus outperform their equivalents in Shulayeva et al.'s research.

To fully explore whether our solution, identifying principles separately, addresses the error analysis Shulayeva et al. provide in their paper, we have also implemented a version of their model only trained on corpus with labelled facts (as opposed to both facts and principles in Golden Standard corpus and only principles in our New corpus). To our satisfaction, this model also shows an increase in accuracy of over 10% over the results of looking for facts and principles together, see Table 18 in the Appendix for full results. Therefore finding facts and principles separately helps to address Shulayeva et al.'s observation, that *"sentences often only contain a short clause containing information about facts, so that in a small dataset, statistical weights associated with the rest of the sentence may outweigh those associated with the clause"* [1]. We can therefore conclude that in order for the statistical weights in the factual clause not to be outweighed by the statistical weights of the principle clause and vice versa, the model must be trained to search for fact or principle individually. This is reasonable conclusion to make as human would equally struggle to label a sentence containing both fact and principle as only a fact or principle.

Overall we are pleased to report setting a new benchmark for the task of principle identification at 96% accuracy. We have showed some improvement with new features, but more importantly identified that focusing on principles alone can significantly improve the performance of Shulayeva

et al.'s classifier on the task of principle identification.

6.2 Cited paragraph identifier

Table 12: Per category and aggregated statistics for cited paragraph classifier.

Classified as →	Cited	Non-Cited	
Cited	64	8	
Non-Cited	1	85	
Type	Precision	Recall	F-Measure
Principle	0.985	0.889	0.935
Neither	0.914	0.988	0.950
Accuracy	0.94	κ	0.88

As per Section 4, to automate our hypothesis, we also need to identify cited paragraphs. Since Shulayeva et al. discovered sentential features alone are not enough to make such identification possible, see Section 4.3, we have investigated how related research in cross reference resolution could be adapted on the case law domain, to give us the ability to find all the paragraphs cited in a case of interest, by looking through all the cases that cite it, see Section 4.5. We report the results of the schema we developed for this task in the paragraphs below.

First we report on how does our schema compare to Adedjouma et al.'s [2] research, since it is the closest research to our task. To do this we evaluate on the precision, a metric Adedjouma et al. use to report their results, with which the citations of paragraphs in Stack v Dowden were identified from citing cases. To do this we needed to establish the number of citations of Stack from the total number of citations in all the cases we are applying our cited paragraph identifier on. We have manually identified all the cited paragraphs to establish that from total of 798 citations in the Cited corpus, there are 183 citations of Stack v Dowden.

Our *citations.py* program automatically identifies 176 citations of Stack. When we have compared these to the paragraph numbers identified manually, we found 175 out of these automatically identified citations, were correctly identified. The single false positive is of a sentence where the same Judge, who gave the judgement in the case of interest, is reported citing a paragraph but of a different case that is not mentioned in the same sentence. This is an unavoidable problem of our approach, since we analyse only a single sentence at at a time, and a schema analysing the context of every sentence would be required to resolve this issue. Such schema would require a high level of natural language understanding and is well beyond the scope of this project.

There are also 8 citations that the program fails to recognise as citing Stack. These false negatives, have not been extracted because the citing entity is contained in a different sentence or paragraph from the citing expression. Despite these shortcomings of our schema we achieve 98.7% precision, comparable to the 99.9% precision reported by Adedjouma et al. [2].

While it is conceivable to develop a schema taking the deviations causing misclassifications above into account, we believe that there will always be new ways of breaking out of such schema. Without analysing the content of the sentence, paragraph and the whole case, this seems unavoidable. Therefore we believe our results, including this shortcoming of our schema, are representative of the near-maximum performance of the rule based approach adapted from Adedjouma et al. [2] on the task of cited paragraph identification.

However, since we are focused on identifying paragraphs citing Stack, it's better to evaluate on how accurate is the classifier at identifying cited paragraphs, rather than how well it performs compared to Adedjouma et al.'s research. Out of 72 paragraphs we have manually identified as cited from 158 paragraphs in Stack, the classifier identifies 64 true positives and 85 true negatives giving it a decent accuracy of 94%. With only 9 cited paragraphs misclassified, our program performs well. The full results are in Table 12.

6.3 Automatic ratio identifier

Table 13: Per category and aggregated statistics for Ratio Decidendi classifier.

Classified as →	Ratio	Obiter	
Ratio	22	12	
Obiter	33	91	
Type	Precision	Recall	F-Measure
Principle	0.400	0.647	0.494
Neither	0.884	0.734	0.802
Accuracy	0.72	κ	0.31

The automatic ratio identifier is the result of combining the principle and cited paragraph classifiers evaluated above. The Fig. 26 and Fig. 27, demonstrate a sample from Stack v Dowden case subjected to our program. Fig. 26 shows a paragraph from the original HTML, Fig. 27 of the annotated one. There are two things happening, the paragraph is highlighted in blue because it was cited and couple of sentences are highlighted yellow to identify the principles in the paragraph. Without engaging with the legal side of the case too much, it is safe to say this is an example where the software successfully identifies ratio, since the case is turning on the decision whether indirect contributions should be considered in dividing property between an unmarried couple.

12 The result might have been different if greater weight could have been given to the inclusion in the transfer of the standard-form receipt clause. But English property law does not permit this, for the reasons explained in *Mortgage Corpn v Shaire [2001] Ch 743* , 753. I think that indirect contributions, such as making improvements which added significant value to the property, or a complete pooling of resources in both time and money so that it did not matter who paid for what during their relationship, ought to be taken into account as well as financial contributions made directly towards the purchase of the property. I would endorse Chadwick LJ's view in *Oxley v Hiscock [2005] Fam 211* , para 69 that regard should be had to the whole course of dealing between them in relation to the property. But the evidence in this case shows that there never was a stage when both parties intended that their beneficial interests in the property should be shared equally. Taking a broad view of the matter, therefore, I agree that the order that the Court of Appeal made provides the fairest result that can be achieved in the circumstances.

Figure 26: *A sample paragraph from an input HTML case (Stack v Dowden).*

12 The result might have been different if greater weight could have been given to the inclusion in the transfer of the standard-form receipt clause. But English property law does not permit this, for the reasons explained in Mortgage Corpn v Shaire [2001] Ch 743 , 753. I think that indirect contributions, such as making improvements which added significant value to the property, or a complete pooling of resources in both time and money so that it did not matter who paid for what during their relationship, ought to be taken into account as well as financial contributions made directly towards the purchase of the property. I would endorse Chadwick LJ's view in Oxley v Hiscock [2005] Fam 211 , para 69 that regard should be had to the whole course of dealing between them in relation to the property. But the evidence in this case shows that there never was a stage when both parties intended that their beneficial interests in the property should be shared equally. Taking a broad view of the matter, therefore, I agree that the order that the Court of Appeal made provides the fairest result that can be achieved in the circumstances.

Figure 27: *A sample paragraph from an output of the same HTML case (Stack v Dowden).*

We evaluate the performance of our ratio classifier on the Stack case, for which we have manually annotated paragraphs containing ratio, and which we use in our manual annotation study in Section 4.4. Our ratio classifier identifies principles in cited paragraphs. We record the numbers of the paragraphs automatically identified and compare them with our manually identified paragraphs containing ratio decidendi.

There are 158 paragraphs in Stack and only 34 contain ratio. As per Table 13, our program identifies ratio with 72% accuracy. This is an improvement over the direct application of Shulayeva et al.'s model of 4%, see Section 4.4. More importantly, as can be seen from the jump in κ coefficient from Shulayeva et al.'s 0.06 to our 0.31, our method has a much higher agreement with human annotator and is not simply identifying ratio by pure chance like principles in citing paragraphs are, see Section 4.4 for our empirical study on this matter.

Combining the principle and cited paragraph classifier therefore not only pin-points the position of the ratio in the paragraph, but also filters the cited paragraphs that do not contain principles, removing the instances where only facts are cited, further improving the performance.

Analysing the mislabeled paragraphs, we came up with two common reasons why judges cite paragraphs without the ratio, one of these is partially addressed by our current approach.

1. The judge might refer to obiter of the cited case. This is a situation where a judge explores an area of law which is not settled, or decides on dissenting from previously recognised law. Interestingly, a judge might also be using parts of the logic of the dissenting judge, in our case Lord Neuberger, to explain the reasoning of the majority of judges in the case. See Section 6.6 below for our reflection on this.

2. The judge might refer to the facts of the cited case. This happens when the judge is linking two cases on their facts to establish the applicability of the precedent. This situation is resolved in our approach if cited paragraph does not contain any principles (in which case it is no longer considered a ratio).

Our automated approach therefore nearly matches it's theoretical ceiling performance of 76% in the task of ratio identification, we established in Section 4.4, proving our hypothesis that focus on principles in cited paragraphs is a possible way of tackling the difficult task of automatic ratio identification. This in turn is a successful step towards our aim of automatically identifying ratio decidendi. As far as we are aware, this project is the first successful approach to this problem, and our novel understanding of relationships between cases and their implication of finding ratio opens up new doors, to further investigate automating this issue based on our method.

6.4 Limitations

There are three fundamental limitations to our work. First there is an implicit limitation to our methodology. For our approach to work, we need a case which has been already cited. The quality of our approach will greatly depend on quantity of citations. That is to say it is important to get a variety of cited paragraphs, rather than one or two frequently cited. These citations might come even from a single case, as long as it cites widely. While not every case gets cited, the interest the case generates in subsequent judgements is proportional to the interest a lawyer or a judge will have in its ratio. This is given by the doctrine of precedent. If a case sets a new precedent, it will be cited a lot, if it does not, then it is safe to assume the case does not establishes anything new and it isn't of an interest to lawyers and judges. Therefore, while still a limitation of our method, this is not as constraining limitation as it might initially appear.

The second limitation of our work is the breadth of our study. While we have spend a lot of time to create two new data corpuses, each spanning over 40 cases, the final evaluation is carried on a single case. This is because we had to identify the ratio in that case, manually analysing over 150 paragraphs and we had to manually analyse near 800 citations in over 40 cases to find the cited paragraphs of this one case. Both tasks were time demanding and to complete the project in time we could not continue building our corpus any further. We have chosen our case wisely to be as representative as possible, with 5 judges giving their judgement in different styles and with the dissenting judgement of Lord Neuberger, to mitigate this shortcoming. But our corpus is limiting none the less.

Finally, there is a limitation in our annotation of ratio. We have spend a long time evaluating which definition to use for ratio identification, to be as objective as possible. A lot of effort was placed on consistently evaluating which paragraphs contain the ratio under the Wambaug's test. However, as any lawyer would concede, there are some grey areas. Some principles are borderline between contributing to the overall story and only adding an extra point of interest. This is inherent in the ambiguity of human language and can be described as the discrepancy between what the judge meant and the variety of ways in which her sentence can be interpreted. In a common law system, there is no way around this problem. Our automated ratio classifier in many ways is an exploration of what the ratio actually is and could feedback into legal theory as much as it draws from it. We elaborate on this in Section 6.6.

6.5 Future work

There are three areas the future work could focus on. First, on adressing the shortcommings of our work, described above. Second, on improving our method. Third, by expanding our model further towards the aim of identifying the valid law in English case law. We elaborate on each of these approaches in the following paragraphs.

In future we would like to address the shortcomings identified above. To address the first limitation, we would like to conduct an empirical study to investigate what are the cases that lawyers are interested in and whether indeed focusing on major judgements is a limitation or not. To address the second shortcoming we would like to build a large dataset of varied case law to fully test our hypothesis across the entire judicial spectrum. To address the third limitation, we would like to carry out a comparative study between human annotators to see what is the deviation between

ratio identified by different annotators.

To continue our work, we would like to focus on improving the principle and cited paragraph identification. For example, unsupervised ML combined with a larger data set, might identify principles with even higher accuracy. Or perhaps a different schema would better capture cited paragraphs. However, since we are extracting principles, and cited paragraphs, at a high accuracy already, any improvement in this area will not have a major impact on our accuracy of ratio identification. Yet, we believe that there is a way of improving our method that could significantly heighten the theoretical ceiling of 76% accuracy. In our project, we have focused on all cited paragraphs without discrimination. However, cited paragraphs are not created equal. They are cited with different frequency, they are cited by cases from courts of different importance (Supreme court might cite differently than the Court of Appeal), the citing cases themselves might be reported immediately after the case comes out, as well as several years or even decades later and the citing case might approve as well as disapprove of a paragraph. All of the above could be used as features discriminating between cited paragraphs. Further, it is not only paragraphs that are cited, a sentence can by directly quoted and paragraphs might be cited individually or in a range. Discriminating between the precision with which text is referred to could again help distinguish between ratio and obiter.

Finally, not all ratio is a valid law. While not a limitation of our work, it should be noted that ratio, while binding at the time it is developed, might be later rejected or modified. For a legal practitioner this means she needs to be constantly aware of the changes in case law, how different cases treat each other and to what extend is the past reasoning rejected or replaced, to find ratio which is currently the valid law. Usually the research interested in this is focused on Shepardization of whole cases, i.e. identifying whether the whole case is valid or invalid. But reality is far more nuanced. Applying similar techniques used on Shepardization to trace the validity of precedent, would therefore be of a great help to anyone conducting legal research.

6.6 Reflection

The definition of ratio we employ has been chosen for the objectivity it brings to our human annotation and it's root in legal academic discussion. However, in many ways the Wambaugh's test is not perfect and our project reveals some of it's shortcomings. In our study it was particularly noticeable on the obiter of Lord Neuberger in Stack, what are the limits of the test. While Lord Neuberger's obiter, an opinion disagreeing with the rest of judges on the outcome of the case, can not satisfy the Wambaug's test, since his reasoning is not deciding or contributing to the decision, there have been multiple citations of some of his paragraphs by later judgements.

Many of these citations pointed out the problems with the final evaluation Lord Neuberger makes, yet lot of the cited paragraphs are discussing a distinction between imputed, inferred and expressed intention. This is a distinction which is crucial for the case, and subsequent judgements have praised Lord Neuberger's reasoning in these paragraphs. In many ways these (i.e. paragraphs 124-129 in Stack), are the reasoning for the outcome of subsequent cases. Therefore these principles, although an obiter, can be as important if not more, than the ratio for lawyers and judges. While this is fine in the context of the aim of our project, in the broader aim of identifying law in English case law, it is important to recognise the limitation of the Wambaugh's test.

However, looking over the alternative definitions of ratio, none seem to be able to remedy this

issue. The problem is perhaps not as much fault of the definition of ratio we employ, as the limitation of our aim of focusing on distinguishing ratio from obiter. In identifying what is helpful for lawyers and judges, our research identified the task of finding ratio in case law. In the light of our research, the ratio, while the closest definition of what the legal practitioner is looking for, lacks the nuance of what the lawyer really looks for in a case. Obiter is after all persuasive component of common law. To relive lawyers of arduous task of reading so many cases to get nuanced view of the law, perhaps expanding the focus on both ratio and obiter, but highlighting only the ratio and obiter that later gains traction in subsequent judgements through the virtue of being cited, would be of more help to lawyers. The example of a highly informative obiter above, identified in our research, shows that in practice such approach might have already emerged in courts.

Our project therefore comes a full circle, with the results observed informing the academic discussion of what is law in case law, the same discussion that served as the point of departure for our project.

6.7 Summary

In this section we have presented the results of our work, demonstrating success in both the area of principle identification, cited paragraph identification as well as for our main task of ratio identification. Learning from the limitation of our project, we further explained different paths of improving upon our work as well as reflected on our original aim.

7 Conclusion

The aim of this project was to apply recent advances in information mining to identify the ratio decidendi in case law. We have contextualised this aim in the core doctrine of stare decisis or precedent, due to which lawyers search for law in past legal judgement transcripts. This discussion led to identifying ratio decidendi, the valid law, as sought after by legal professionals conducting their research. Our project has via a careful analysis of the aforementioned legal research process identified the two parts such research comprises of. The first part is finding relevant cases, and we describe a wealth of support that is offered in this respect by the major legal search engines. The second part is looking inside those cases and finding the actual law or ratio in them. We report this part of the research is not supported by any commercially available tool. We have therefore turned to scientific research to find approaches claiming to identify ratio are misrepresenting what the ratio actually is. Mindful of a legally sound definition, under which only the deciding principles of a case constitute the ratio, we embarked towards a new approach of automatically identifying the ratio by finding principles in cited paragraphs of a case. While not the first to explore the relationship between cited principles and ratio, learning from the limitations of Shulayeva et al. [1], we instead of looking at citing paragraphs search for cited ones. We prove our hypothesis by conducting a small scale manual annotation study, demonstrating this shift of focus significantly increases the precision of our approach. With this manual study, we establish a theoretical performance ceiling of our approach at 76% accuracy.

To automate our method we had to overcome several technical challenges. We investigated new sentential features to improve principle identification, reporting small improvements in accuracy. However, we have significantly improved over the past research on principle identification in case law, by re-annotating Shulayeva et al.'s Gold Standard corpus, setting an new benchmark for this task at 96%. Then we successfully demonstrated an approach to identifying cited paragraphs in case law using regular expression and a schema, reporting the similarities and differences between resolving citation, a new tasks only attempted by us, and cross reference resolution a well established research. Our results show that we have nearly matched the performance of similar studies on cross reference resolution, and more importantly that we are therefore able to find cited paragraphs in a case with a decent accuracy of 94%. When we have combined the principle with cited paragraph classifier, we achieved 72% accuracy in automatically identifying ratio decidendi. Automating our method to near the maximum of its theoretical performance.

This project has therefore successfully moved towards the goal of automating ratio identification, opening up a new way of approaching this problem. In this thesis we provide a benchmark that future research can use as a comparative measure to investigate how our method performs applied on a wider legal corpus. We do hope, the solid starting ground provided by this project, will in the future translate directly into improving legal practice.

8 Bibliography

References

[1] O. Shulayeva, A. Siddharthan, and A. Wyner, "Recognizing cited facts and principles in legal judgements," *Artificial Intelligence and Law*, vol. 25, no. 1, pp. 107–126, 2017.

[2] M. Adedjouma, M. Sabetzadeh, and L. C. Briand, "Automated detection and resolution of legal cross references: Approach and a study of luxembourg's legislation," in *2014 IEEE 22nd International Requirements Engineering Conference (RE)*, pp. 63–72, Aug 2014.

[3] K. Branting, "Four challenges for a computational model of legal precedent," *THINK (Journal of the Institute for Language Technology and Artificial Intelligence)*, vol. 3, pp. 62–69, 1994.

[4] K. Greenawalt, "Interpretation and judgment," *Yale Journal of Law*, vol. 9, no. 2, 2013.

[5] P. Zhang and L. Koppaka, "Semantics-based legal citation network," in *Proceedings of the 11th International Conference on Artificial Intelligence and Law*, ICAIL '07, (New York, NY, USA), pp. 123–130, ACM, 2007.

[6] C. Elliott and F. Quinn, *English legal system*. Pearson Education, 1 ed., 2012.

[7] L. Branting, "A reduction-graph model of precedent in legal analysis," *Artificial Intelligence*, vol. 150, no. 1, pp. 59 – 95, 2003.

[8] J. Plug, "Indicators of obiter dicta." unpublished, 1997.

[9] S. K. Stoan, "Research and library skills: An analysis and interpretation," *College and Research Libraries*, vol. 45, pp. 99–109, March 1984.

[10] D. Ellis, "A behavioral approach to information retrieval system design," *J. Doc.*, vol. 45, pp. 171–212, Oct. 1989.

[11] M. J. Bates, "Where should the person stop and the information search interface start?," *Inf. Process. Manage.*, vol. 26, pp. 575–591, Oct. 1990.

[12] S. M. Marx, "Citation networks in the law," *Jurimetrics Journal*, vol. 10, no. 4, pp. 121–137, 1970.

[13] J. Davis and C. Levitt, "Internet legal research on a budget," 2014.

[14] M. Raz, "Inside precedents: The ratio decidendi and the obiter dicta," *Common Law Review*, p. 21, 2002.

[15] J. Carletta, "Assessing agreement on classification tasks: the kappa statistic," *Computational linguistics*, vol. 22, no. 2, pp. 249–254, 1996.

[16] S. Teufel, A. Siddharthan, and C. Batchelor, "Towards discipline-independent argumentative zoning: Evidence from chemistry and computational linguistics," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3*, EMNLP '09, (Stroudsburg, PA, USA), pp. 1493–1502, Association for Computational Linguistics, 2009.

[17] B. Hachey and C. Grover, "Extractive summarisation of legal texts," *Artif. Intell. Law*, vol. 14, pp. 305–345, Dec. 2006.

[18] A. Farzindar and G. Lapalme, "Letsum, an automatic legal text summarizing system," *Legal knowledge and information systems, JURIX*, pp. 11–18, 2004.

[19] F. Kuhn, "A description language for content zones of german court decisions," in *Proceedings of the LREC 2010 Workshop on the Semantic Processing of Legal Texts*, pp. 1–7, 2010.

[20] S. Teufel, A. Siddharthan, and D. Tidhar, "Automatic classification of citation function," in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, (Stroudsburg, PA, USA), pp. 103–110, Association for Computational Linguistics, 2006.

[21] P. Ogden, "Mastering the lawless science of our law: A story of legal citation indexes," *Law Library Journal*, vol. 85, pp. 88–91, 1993.

[22] E. Garfield, "Citation indexes for science: A new dimension in documentation through association of ideas," *Science*, vol. 122, no. 3159, pp. 108–111, 1955.

[23] E. Garfield, *Citation Indexing: Its Theory and Application in Science, Technology, and Humanities*. Information sciences series, Isi Press, 1979.

[24] E. Garfield, "How isi selects journals for coverage: Quantitative and qualitative considerations," *Current Contents*, vol. 13, p. 185, May 1990.

[25] C. Borkowski, "Structure, effectiveness, and uses of the citation identifier, an operational computer program for automatic identification of case citations in legal literature," in *Proceedings of the 1969 Conference on Computational Linguistics*, COLING '69, (Stroudsburg, PA, USA), pp. 1–22, Association for Computational Linguistics, 1969.

[26] N. Kiyavitskaya, N. Zeni, T. D. Breaux, A. I. Antón, J. R. Cordy, L. Mich, and J. Mylopoulos, "Automating the extraction of rights and obligations for regulatory compliance," in *International Conference on Conceptual Modeling*, pp. 154–168, Springer, 2008.

[27] T. Breaux and A. Antón, "Analyzing regulatory rules for privacy and security requirements," *IEEE transactions on software engineering*, vol. 34, no. 1, pp. 5–20, 2008.

[28] T. D. Breaux, *Legal requirements acquisition for the specification of legally compliant information systems*. North Carolina State University, 2009.

[29] E. de Maat, R. Winkels, and T. van Engers, "Automated detection of reference structures in law," in *Proceedings of the 2006 Conference on Legal Knowledge and Information Systems: JURIX 2006: The Nineteenth Annual Conference*, (Amsterdam, The Netherlands, The Netherlands), pp. 41–50, IOS Press, 2006.

[30] M. Palmirani, R. Brighi, and M. Massini, "Automated extraction of normative references in legal texts," in *Proceedings of the 9th International Conference on Artificial Intelligence and Law*, ICAIL '03, (New York, NY, USA), pp. 105–106, ACM, 2003.

[31] M. Hamdaqa and A. Hamou-Lhadj, "An approach based on citation analysis to support effective handling of regulatory compliance," *Future Gener. Comput. Syst.*, vol. 27, pp. 395–410, Apr. 2011.

[32] O. T. Tran, M. Le Nguyen, and A. Shimazu, "Reference resolution in legal texts," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law*, ICAIL '13, (New York, NY, USA), pp. 101–110, ACM, 2013.

[33] C. Nentwich, L. Capra, W. Emmerich, and A. Finkelsteiin, "Xlinkit: A consistency checking and smart link generation service," *ACM Trans. Internet Technol.*, vol. 2, pp. 151–185, May 2002.

[34] M. Martínez-González, P. de la Fuente, and D.-J. Vicente, *Reference Extraction and Resolution for Legal Texts*, pp. 218–221. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005.

[35] A. Bolioli, L. Dini, P. Mercatali, and F. Romano, "For the automated mark-up of italian legislative texts in xml," in *Legal Knowledge and Information Systems (Jurix 2002*, pp. 21–30, IOS Press, 2002.

[36] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, (San Francisco, CA, USA), pp. 282–289, Morgan Kaufmann Publishers Inc., 2001.

[37] M. Saravanan and B. Ravindran, "Identification of rhetorical roles for segmentation and summarization of a legal judgment," *Artif. Intell. Law*, vol. 18, pp. 45–76, Mar. 2010.

[38] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: Sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, (Stroudsburg, PA, USA), pp. 79–86, Association for Computational Linguistics, 2002.

[39] R. M. Palau and M.-F. Moens, "Argumentation mining: The detection, classification and structure of arguments in text," in *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, ICAIL '09, (New York, NY, USA), pp. 98–107, ACM, 2009.

[40] P. Domingos and M. Pazzani, "On the optimality of the simple bayesian classifier under zero-one loss," *Machine Learning*, vol. 29, no. 2, pp. 103–130, 1997.

[41] S. Schuster and C. D. Manning, "Enhanced english universal dependencies: An improved representation for natural language understanding tasks.," in *LREC*, 2016.

9 Appendix

Table 14: Lemmatised unigrams perform less well than normal unigrams.

Features:	Lemma	Unigrams
Accuracy	0.77	0.78
κ	0.57	0.60

Table 15: Shulayeva et al.'s principle and fact classifier with "words to keep" option set to only 1000 (as opposed to 5000) reduces the overall performance dramatically. Compare with Table 5.

Classified as →	Principle	Fact	Neither
Principle	650	4	192
Fact	9	193	44
Neither	271	42	1319
Type	Precision	Recall	F-Measure
Principle	0.699	0.768	0.732
Facts	0.808	0.785	0.796
Neither	0.848	0.808	0.828
Accuracy	0.79	κ	0.62

Table 16: New features applied on the New corpus reduce the models performance slightly. Compare with Table 10.

Classified as →	Principle	Neither	
Principle	819	88	
Neither	32	1785	
Type	Precision	Recall	F-Measure
Principle	0.962	0.903	0.932
Neither	0.953	0.982	0.967
Accuracy	0.96	κ	0.90

Table 17: The performance of the best performing new feature when combined with the best performing extra feature is lower than for all features together. See Table 8

Feature	Accuracy	Improvement	κ
Baseline	85.61%	+-0%	0.73
Quotation	86.01%	+0.40%	
Length	86.01%	+0.40%	
Combined	**86.31%**	**+0.70%**	**0.74**

Table 18: Shulaeyva et al.'s model trained to recognise facts only on Golden Standard corpus.

Classified as →	Fact	Neither	
Fact	102	144	
Neither	3	2475	
Type	Precision	Recall	F-Measure
Fact	0.971	0.415	0.581
Neither	0.945	0.999	0.971
Accuracy	0.95	κ	0.56

Informative Ranking	Feature	Principle	Fact	Neither
1	Quote	0.007954203	0.004197786	0.002493346
2	VBD	0.001946499	0.004731086	0.004623112
3	VBZ	0.003868197	0.001263954	0.004065678
4	MD	0.00475889	0.0016756	0.003971663
5	Length	0.4912159	0.433473838	0.554131423
6	is	0.004655933	0.001415914	0.004122042
7	VB	0.003126833	0.00224015	0.003391296
8	or	0.004356749	0.001744681	0.002323945
9	a	0.003779859	0.003678898	0.003812923
10	Mr	4.50E-04	0.002108239	0.00441439
11	NNP	0.001988084	0.003175284	0.004510942
12	was	0.001819372	0.006420261	0.004552658
13	JJ	0.002011274	0.001831842	0.002485191
14	must	0.002803815	5.25E-05	9.55E-04
15	be	0.001961051	0.001815149	0.002463111
16	Person	0.002047038	0.003778007	0.006414805
17	CC	0.003640187	0.00376627	0.003945657
18	had	0.001022002	0.005746327	0.002511821
19	may	0.003051047	6.58E-04	0.001272174
20	has	0.003300244	6.51E-04	0.002203711
21	if	0.00341852	0.001056481	0.002158185
22	of	0.002626619	0.002703731	0.00324049
23	it	0.004144436	0.003224913	0.003824984
24	see	0.002743922	7.74E-04	0.001345721
25	where	0.002798011	0.001263217	0.001339699
26	I	0.001630486	0.001030582	0.004718649
27	were	9.71E-04	0.004757292	0.002220197
28	WRB	0.003435484	0.001705053	0.002399269
29	RB	0.002803871	0.002211108	0.003557427
30	other	0.002756997	0.001823935	0.001416821
31	party	0.002952656	8.78E-04	0.001954773
32	NN	4.05E-04	4.61E-04	5.83E-04
33	are	0.002885341	4.97E-04	0.002309948
34	the	0.001183478	0.001275	0.001646868
35	concern	7.37E-04	0.003554399	8.80E-04
36	TO	0.002924655	0.002889314	0.003666473
37	to	0.002924655	0.002889314	0.003666473
38	concerned	3.37E-04	0.00309086	7.22E-04
39	Sub-Clause	0.044932474	0.04161055	0.059734941
40	will	0.00257846	6.34E-04	0.00193463
41	VBG	0.003804737	0.003849193	0.004181532

Figure 28: *Top 41 ranked features in the dataset. The highlighted rows are new or extra features.*