EINDHOVEN UNIVERSITY OF TECHNOLOGY

Department of Mathematics and Computer Science

temporary_title

NLP and Hybrid Approaches for Legal Information Retrieval

Bachelor End Project Report

Giacomo Grazia

Supervisors:
Prof. Dr. Johan Wolswinkel, LL.M.
Harry Nan, PhD ......

January xy, 2025

Abstract

» TBD AT THE END «

Write an abstract for your work. Replace each of the points below with one sentence (two if you must) and you have your abstract. Write it when you finished your entire report.[1]

Introduction. In one sentence, what's the topic? Phrase it in a way that your reader will understand. If you're writing a PhD thesis, your readers are the examiners – assume they are familiar with the general field of research, so you need to tell them specifically what topic your thesis addresses. Same advice works for scientific papers – the readers are the peer reviewers, and eventually others in your field interested in your research, so again they know the background work, but want to know specifically what topic your paper covers.

State the problem you tackle. What's the key research question? Again, in one sentence. (Note: For a more general essay, I'd adjust this slightly to state the central question that you want to address) Remember, your first sentence introduced the overall topic, so now you can build on that, and focus on one key question within that topic. If you can't summarize your thesis/paper/essay in one key question, then you don't yet understand what you're trying to write about. Keep working at this step until you have a single, concise (and understandable) question.

Summarize (in one sentence) why nobody else has adequately answered the research question yet. For a PhD thesis, you'll have an entire chapter, covering what's been done previously in the literature. Here you have to boil that down to one sentence. But remember, the trick is not to try and cover all the various ways in which people have tried and failed; the trick is to explain that there's this one particular approach that nobody else tried yet (hint: it's the thing that your research does). But here you're phrasing it in such a way that it's clear it's a gap in the literature. So use a phrase such as "previous work has failed to address...". (if you're writing a more general essay, you still need to summarize the source material you're drawing on, so you can pull the same trick – explain in a few words what the general message in the source material is, but expressed in terms of what's missing)

Explain, in one sentence, how you tackled the research question. What's your big new idea? (Again for a more general essay, you might want to adapt this

[1] https://www.easterbrook.ca/steve/2010/01/how-to-write-a-scientific-abstract-in-six-easy-steps/

slightly: what's the new perspective you have adopted? or: What's your overall view on the question you introduced in step 2?)

In one sentence, how did you go about doing the research that follows from your big idea. Did you run experiments? Build a piece of software? Carry out case studies? This is likely to be the longest sentence, especially if it's a PhD thesis – after all you're probably covering several years worth of research. But don't overdo it – we're still looking for a sentence that you could read aloud without having to stop for breath. Remember, the word 'abstract' means a summary of the main ideas with most of the detail left out. So feel free to omit detail! (For those of you who got this far and are still insisting on writing an essay rather than signing up for a PhD, this sentence is really an elaboration of sentence 4 – explore the consequences of your new perspective).

As a single sentence, what's the key impact of your research? Here we're not looking for the outcome of an experiment. We're looking for a summary of the implications. What's it all mean? Why should other people care? What can they do with your research. (Essay folks: all the same questions apply: what conclusions did you draw, and why would anyone care about them?)

Contents

Chapter 1

Introduction

Purpose and scope of Chapter 1. The introduction chapter is a summary of your work and your scientific argument that shall be understandable to anyone in your scientific field, e.g., anyone in Data Science. A reader must be able to comprehend the problem, method, relevant execution details, results, and their interpretation by reading the introduction and the introduction alone. Section 1.1 introduces the general topic of your research. Section 1.2 discusses the state of the art and identifies a research. Section 1.3 then states the research problem to investigate. Section 3.4 explains the research method that was followed, possibly with execution details. Section 1.5 then presents the results and their interpretation. Only if a reader thinks they are not convinced or they need more details to reproduce your study, they shall have to read further. The individual chapters and sections provide the details for each of the steps in your scientific argument.

You usually write the introduction chapter *after* you wrote all other chapters, but you should keep on making notes for each of the sections as you write the later chapters..

Purpose and scope of the introduction paragraph to a chapter. The paragraph you are reading above is a typical introductory paragraph to a chapter. It is a high-level summary of the chapters' topic (SM1 and SM2). It gives the reader some guidance by breaking down the chapter topic into subtopics that are clearly named (SM3) in the right order with forward references to the corresponding sections (SM4). It may close with announcing the result you obtain (SM6) but this is usually not done in the opening paragraph of the introduction.

1.1 Context and Relevance of Legal Information Disclosure - DRAFT

Advancements in Government Transparency Policies. Given recent advancements in transparency policies, particularly with the Open Government Act, proactive disclosure of government documents has been prioritized to enhance public oversight. Such policies enable citizens to assess governmental actions, fostering accountability and reducing opacity in administrative processes. However, implementing these disclosures remains challenging due to the sheer volume of data generated and the historical lack of standardized formatting in legal documents, which complicates large-scale analyses.

Challenges in Legal Data Standardization and Analysis. Furthermore, legal data often lacks uniformity, as much of it has not been structured in a way that facilitates easy access or comparative study. Structured metadata, which refers to organized data elements that describe key information (such as decision dates, parties involved, and legal references), is often absent in publicly available administrative decisions, like permits. This lack of organization makes comprehensive analysis labor-intensive and, at times, impractical. Conducting comparative studies on these decisions, although complex, can significantly improve the consistency of legal application and reduce risks of arbitrary rulings. However, achieving this through traditional legal research alone is increasingly impractical due to time constraints and resource limitations.

Leveraging NLP for Efficient Legal Data Processing. The recent developments in natural language processing (NLP)—a field of artificial intelligence focused on enabling computers to interpret and process human language—offer powerful tools for automating the extraction and structuring of information from legal texts. Specifically, large language models (LLMs), such as those developed by OpenAI and other organizations, have demonstrated impressive capabilities in understanding and analyzing text with high accuracy. These advanced information extraction techniques allow the processing of vast quantities of legal data with minimal human intervention, enabling accurate and detailed analysis at scale. This evolution in NLP represents a viable solution for handling large datasets in the legal domain, aligning with the growing demand for transparency and accessibility in governmental decision-making.

1.2 State of the Art DRAFT

In recent years, there has been a growing emphasis on automated information extraction from legal documents to meet transparency requirements, streamline legal analysis, and facilitate comparative studies. However, significant challenges persist in achieving accurate, scalable solutions in this domain.

Manual annotation has traditionally played a central role in labeling legal factors within documents, especially when considering judicial opinions and complex regulatory texts. Although human annotation provides high accuracy, it is inherently resource-intensive and difficult to scale, as highlighted by recent studies exploring legal text processing [1]. This demand for manual input underscores the need for more efficient annotation systems, particularly given the ever-increasing volume of legal data.

Current methods for legal information extraction predominantly rely on rule-based and traditional NLP techniques. Rule-based approaches operate on predefined linguistic patterns to identify specific legal terms and concepts, but these systems often struggle with the nuanced language and complex structures typical of legal documents [2]. Such limitations have led researchers to investigate the potential of more machine learning-oriented approached, and, more recently, large language models (LLMs) as tools to assist in legal annotation and extraction tasks. Initial findings suggest that LLMs like GPT-4 have made strides in preliminary labeling, potentially serving as a foundation for automated annotation. However, there are notable concerns regarding the accuracy and reliability of these models, particularly around potential biases and their inability to capture intricate legal nuances without human oversight [1].

A key challenge in automating legal information extraction lies in the diversity and inconsistency of legal data. This issue is especially pressing in the context of transparency legislation, such as the Open Government Act, which mandates proactive disclosure of government decisions. While this legislation supports public access to government records, a lack of standardized formatting across legal documents complicates large-scale analysis and comparative studies [3]. Without structured metadata—elements like decision dates, involved parties, and legal references—publicly available legal documents remain difficult to navigate and analyze.

The gap between transparency mandates and the current capabilities of legal information extraction (IE) underscores the need for hybrid annotation models. Combining the scalability of large language models (LLMs) with the expertise of human annotators could significantly improve both efficiency and accuracy in processing legal documents. However, relying solely on human annotators is often impractical due to the costs and time-intensive nature of manual annotation. Therefore, it is crucial to integrate rule-based methods to streamline the annotation process, with human intervention focused primarily on oversight.

While LLMs can provide valuable preliminary results, they are sometimes unreliable due to risks like hallucination. By first applying rule-based methods, relevant information can be extracted to create a prompt that gives the LLM more context, which helps limit the risk of generating unrelated or inaccurate outputs. This integration enables human experts to validate and refine the extracted information with fewer risks of irrelevant content. However, verification on a large scale remains a challenge when the volume of data to review becomes substantial.

Introducing a rule-based component alongside LLMs allows for broader and more accurate applications, enhancing overall reliability. This hybrid approach could offer a promising solution to improve the accessibility, consistency, and transparency of legal information systems, aligning with modern governance and research needs. Over time, automated IE could enable more seamless comparative studies across legal decisions, fostering consistent legal applications and reducing the risk of arbitrary rulings.

I WANT TO ADD more from my notes on the papers to give a clearer overview of the state of the art.

1.3 Research Question (SM2) DRAFT

This research addresses the gap in effectively extracting and structuring metadata from Dutch administrative decisions, particularly energy permits, to support transparency mandates like the Open Government Act. As transparency policies increasingly require government bodies to disclose documents in accessible formats, there is a growing need for structured data that enables public scrutiny and comparative analysis. However, implementing such disclosures faces significant challenges. The sheer volume of data and the lack of standardized formatting in legal documents complicate large-scale analysis and make comprehensive manual processing impractical.

This study aims to leverage natural language processing (NLP) techniques to systematically categorize and extract key metadata—such as issuance dates, involved parties, and regulatory references—from administrative decisions issued by the Autoriteit Consument & Markt (ACM). Recognizing the constraints of incomplete labeling in legal data, weak labeling techniques will be explored to achieve systematic categorization with minimal manual intervention. This approach seeks to improve data accessibility, consistency, and transparency, thus enabling a deeper understanding of administrative decision-making patterns. This led to the following research question:

How can NLP techniques be applied to effectively extract and structure key metadata from Dutch administrative decisions (energy permits), while overcoming the challenge of incomplete labels using weak labeling techniques, and comparing the performance of different approaches, including rule-based methods and machine learning?

In this context, *effectiveness* refers to the ability of NLP techniques to extract and structure meta-data from administrative decisions in a way that achieves satisfactory outcomes across multiple dimensions: accuracy, efficiency, scalability, and adaptability. An approach is considered effective if it yields high-quality results, is feasible to implement, and can be applied to various types of decisions with minimal supervision.

To answer this question comprehensively, the research question can be further broken down into the following sub-questions:

- **Weak labeling**: *How can weak labeling techniques be applied to systematically categorize Dutch administrative decisions, addressing the issue of incomplete data labeling?* THIS ANSWER WILL present how the data has been labeled in different ways depending on the metadata (e.g. dates with gpt, legal ground with LX + gpt, NER for ontvanger etc.)
- **Comparison of techniques**: *Which NLP techniques—whether rule-based, machine learning, or a hybrid approach—are most effective in extracting key metadata from administrative decisions in the absence of fully labeled data?* THIS IS VERY CONNECTED TO PREVIOUS SUBQUESTION. However, focus is on the outcomes. WOULD like to have feedback on this.

This study hypothesizes that a hybrid approach integrating rule-based methods with large language models (LLMs) could provide a scalable solution without compromising accuracy. For instance, rule-based methods can assist in generating contextual information for LLMs, reducing the likelihood of hallucinations and ensuring more relevant, accurate extractions. This research has the potential to support automated information extraction at scale, making it feasible to conduct comparative studies across legal decisions, fostering a more consistent application of the law, and reducing the risk of arbitrary rulings. Ultimately, insights from this work will contribute to the broader goal of achieving accessible and transparent government information, advancing both policy analysis and academic research in legal studies.

1.4 Method or Approach (SM3, SM4) DRAFT

To address the research questions outlined in Section 1.3, a structured workflow was developed, encompassing data preparation, information extraction, and final analysis. This approach applies natural language processing (NLP) techniques to extract and structure key metadata from Dutch administrative decisions, specifically focusing on energy permits issued by the Autoriteit Consument & Markt (ACM), sourced directly from the website `www.acm.nl`.

Data preparation involved cleaning and formatting the initial dataset, which was provided in CSV format. Since the raw data contained OCR errors [ADD DETAILS ON COLLECTION? or at least more on what OCR is and its limitations/issues i will have to fix], a combination of rule-based techniques (e.g., regular expressions for newline or non-ASCII character removal) and the OpenAI API was used for correction. Minor errors, such as misspellings or misformatted company names (e.g., ”totalenergies” corrected to “TotalEnergies”) and split words (e.g., “art icle” to “article”), were addressed using a custom prompt (HERE MENTION PROMPT OR add it to later and more detailed section) designed to minimize unnecessary alterations. For entries with significant errors where automated recovery was infeasible, such as unreadable text, manual flagging and deletion were applied, affecting around 0.06% of the total dataset.

The information extraction phase utilized a combination of NLP models and regular expressions to identify and extract key data points, such as issuance dates, legal grounds, and parties involved (CONSIDERING ADDING ORIGINAL DUTCH NAMES FOR THEM). These structured elements are essential for any subsequent analysis, enabling the comparison of decisions across multiple categories.

THIS PART GIVES A GENERAL IDEA OF WHAT WOULD BE THE TEXT IN CASE I CAN GO FURTHER with an analysis based on the extracted metadata.

Finally, the processed data was analyzed to evaluate cohesiveness among decisions and distinctness among classes of decisions.

The entire process emphasizes reproducibility, ensuring that each step can be systematically followed to handle similar datasets effectively. This approach aligns with the objectives of transparency and accessibility, providing structured insights from unstructured data and supporting comparative legal analysis.

1.5 Findings (SM5, SM6) TO BE DONE AT THE END

Purpose and scope. You close the introduction by clearly stating the evaluation setup you designed to evaluate the success of your study regarding the research objective, which comes in two steps. It is most likely a summary of your evaluation in Chapter 6.

Results (SM5)

You state the evaluation method that is in line with your research question from Sect. 1.3 and summarize the measurements you obtained but you do not interpret them, i.e., you only report the numbers but you do not include judging statements.

Interpretation (SM6)

You summarize your interpretation of the results and draw conclusions. State whether and to which degree the research question from Sect. 1.3 has been answered successfully or not.

Finally state briefly how much closer society and science have come in answering the general objective you outlined in Sect. 1.1.

Chapter 2

Background (SM1)

Purpose and scope. The background chapter has multiple roles.

- Preliminaries. It has to provide all (and exactly the) information that is necessary to understand the methodological and technical parts of your work in the specific area of study. Assume as starting point another student in your degree who did not study the specific subject you are studying but has the task to understand your work. Which concepts, terms, definitions, etc. does the student have to know? Which formulas, symbols, etc. are standard in this topic? Only introduce definitions if you actually need them in any of the subsequent chapters.

- Related Work. It has to provide a comprehensive discussion of all prior work in the area on this subject. Your discussion has to summarize these prior works and has to explain in which way the research question you are solving (Sect. 1.3) has not been solved yet because prior work had more limiting assumptions, addressed a different angle, their results are not complete etc. Depending on the subject you are studying, the related work part can be larger and warrant an entire chapter on its own, or be fully concluded within Sect. 1.2.

  You can close the related work discussion by clarifying the positioning and formulation of your research question (SM2) in relation to all the prior work, making more explicit whether you address an existing research question under different premises or whether you work on a modified or completely new research question.

Chapter 3

Problem Exposition (optional -> PROBABLY NOT NEEDED)

Purpose and scope. Introduce the problem context in more detail if Sect. 1.1 does not provide all necessary information about the problem to follow the rest of the report. This can include further details on the data you studied, context assumptions and requirements, etc.

If you have to expose the problem in more detail here, then this chapter should also provide a more detailed explanation of research question and the method you are applying, i.e., you can now provide more concrete subproblems compared to Sect. 1.3 more details for the method Sect. 1.4 because you now have explained the problem much better. A typical structure can be.

3.1 Context/Business Understanding (SM1)

provide details

3.2 Data Understanding (SM1)

provide details

3.3 Detailed Research Questions (SM2)

provide details based on Sect. 3.1 and 3.2

3.4 Detailed Method (SM3)

provide details based on Sect. 3.1 and 3.2

Chapter 4

First Real Chapter addressing first Research Problem

Purpose and scope. After you stated research context (SM1), research problem (SM2), and research method (SM3) in Chapter 1 and possibly Chapter 3, the remainder of your entire report addresses execution (SM4), results (SM5), and interpretation (SM6). You usually do this by addressing various sub-problems again through scientific arguments following the 6 steps SM1-SM6.

Have a short chapter introduction that recalls and explains the first research problem of your thesis. The problem has to show up in the introduction in Sect. 1.3 or in Sect. 3.3 already. This provides the background (SM1) for this chapter while the first research problem of the thesis becomes the research question/hypothesis (SM2) for this chapter.

Next, explain in the chapter intro how you solve the research problem in this chapter by breaking it down in further sub-problems. By this, you outline the method (SM3) through which you are going to solve the problem of this chapter. This is necessary to give the reader guidance of what's to come in this chapter and how it fits into the thesis as a whole. Explain that you will address the first sub-problem in Sect. 4.1 and the second sub-problem in Sect. 4.2, etc. The sections then provide the details for execution and results.

4.1 First Sub-Problem

The first paragraph describes the first sub-problem and develops the requirements a solution has to satisfy (SM2 for this section). The requirements have to be based on the knowledge and reasoning developing in the preceding chapters

Figure 4.1: A scientific figure that has to be explained in the text

and sections. Try to use an example to illustrate the problem and the desired properties of the solution. Check that every term/concept you use here has already been defined already in a previous section. If you cannot describe your problem without defining new terms, you may have to add another section before this one that develops the terms and concepts you need to explain the problem.

The second paragraph describes the method/approach how you address the problem (SM3 for this section). Describe the method in a level of detail that allows another student to reproduce your steps. Make use of appendices (see Sect. A) if certain details take too much space.

The third, fourth, and following paragraph provides details on applying the method or developing a new approach, i.e., execution (SM4) and may explain results (SM5), i.e. details on the steps needed to reproduce the results.

Results (SM5) can come in many forms, e.g., conceptual diagrams, algorithms, tables, charts, a list of articles from a literature research etc. You must reference them ("Figure 4.1 shows...") and describe the results in text. If you use diagrams, tables, or charts, you cannot expect the reader to know what to you expect them to see in a diagram, table or chart. Describe to them how to read these, explain the meaning of particular elements, point out special observations. But you may only describe the results you must not interpret them. Make use of appendices if certain details take too much space.

After describing the results, you may interpret them (SM6). Here you can infer what a particular observation means (for you), how it can be applied, or what others can do with it. You must not write interpretations before completely describing your results. This is a common mistake done by most beginner writers. You want to quickly get to the point, which is the final finding or interpretation. But you forget that your reader does not understand yet what you are interpreting - they do not know yet what you do know. An interpretation can only be followed after all results have been described. The interpretation must be based on the written description only. Then you can be sure that your readers can follow your interpretation and reach the same conclusions as you have.

Ideally, your interpretation leads to the next sub-problem in Sect. 4.2.

4.2 Second Sub-Problem

You now build on the solution to the first sub-problem of Sect. 4.1 (SM1) and recall second sub-problem (SM2, you detailed in the introduction of this chapter) and follow the same pattern as before (SM3-SM6).

Note that not all sections may not include all parts SM1-SM6 in all detail. Some sections do not require to repeatedly state the background (SM1) or the research problem (SM2) if they were already clearly defined in a previous section. Sometimes, a section is only dedicated to describing the method (SM3) and execution (SM4) and does not contain any results or interpretations. Sometimes results (SM5) and interpretations (SM6) only come in the evaluation chapter.

What is important for you when you are writing a scientific argument is not to slavishly have SM1-SM6 in each section explicitly, but that you are always fully aware of the following:

- Which step of a scientific argument am I currently writing (SM1, SM2, ..., SM6)?
- Does the step that I am writing come in the right order, i.e., if you are writing about execution (SM4, e.g., details of building a model), is there a preceding paragraph or section that describes the method (SM3) and is that one preceded by a clear statement of the (sub-)problem addressed (SM2)?
- Are you really *not* writing interpretation SM6 before SM5, SM4, or SM3?
- Is it clear to the reader which part of the scientific argument you are currently making?

Chapter 5

Second Real Chapter

Have a short chapter introduction that recalls what you already achieved in Chapter 4 and explain the second research problem of your thesis. The problem has to show up in the introduction in Sect. 1.3 or in Sect. 3.3 already. etc.

Chapter 6

Evaluation

Purpose and scope. The evaluation chapter should be the most formal and rigorously structured chapter of your thesis as the validity of your evaluation argument depends on it.

6.1 Objective (SM2)

Clearly state what you want to evaluate and what you want to measure.

6.2 Setup (SM3)

State which data, participants, tools, etc. you chose and why. Clearly state how you measure outcomes and how you compare them to baselines, reference groups, etc.

6.3 Execution (SM4)

Provide all details on the execution that are necessary to allow another person to reproduce your results at a later point.

6.4 Results (SM5)

You only report the measurements. You must present and reference them ("Figure 6.1 shows...") and describe the results in text. If you use diagrams, tables, or charts, you cannot expect the reader to know what to you expect them to see in a diagram, table or chart. Describe to them how to read these, explain the meaning of particular elements, point out special observations.

Figure 6.1: Another scientific figure that has to be explained in the text

But you may only describe the results you must not interpret them. Make use of appendices if certain details take too much space.

6.5 Discussion (SM6)

An interpretation can only be followed after all results have been described. The interpretation must be based on the written description in Sect. 6.4 only. Then you can be sure that your readers can follow your interpretation and reach the same conclusions as you have.

Chapter 7

Conclusion

Your conclusions are not just a factual summary of your work, but they position, interpret, and defend your findings against the state of the art that you discussed in Sect. 1.2. You specifically outline which concrete findings or methodological contributions advance our knowledge towards the general objective you introduced in Sect. 1.1. Objectively discuss which parts you solved and in which parts you failed.

You should explicitly discuss limitations and shortcomings of your work and detail what kind of future studies are needed to overcome these limitations. Be specific in the sense that your arguments for future work should be based on concrete findings and insights you obtained in your report.

Bibliography

[1] Morgan Gray, Jaromir Savelka, Wesley Oliver, and Kevin Ashley. *Can GPT Alleviate the Burden of Annotation?* 12 2023.

[2] Carlo Sansone and Giancarlo Sperlí. Legal information retrieval systems: State-of-the-art and open issues. *Information Systems*, 106:101967, 2022.

[3] Johan Wolswinkel. Actieve openbaarmaking van beschikkingen: Op weg naar transparante besluitvorming 'nieuwe stijl'? *Nederlands Juristenblad*, 2024(24):1851–1857, July 2024.

List of Figures

List of Tables

Appendix A

My first appendix