Automated Construction of Legal Terminology Thesaurus

Zoltán Szoplák[1], Peter Gurský[1], Šimon Horvát[1], Dávid Varga[1] and Stanislav Krajči[1]

[1]Institute of Computer Science, Faculty of Science, Pavol Jozef Šafárik University in Košice, Jesenná 5, 040 01 Košice, Slovakia

Abstract

A dictionary of legal terms along with their definitions is not merely a powerful resource for legal experts and researchers or laymen, but also for many automated tasks, namely for generating embeddings of a term via a language model, through which legal documents and their snippets can be filtered or clustered or for LLM prompting when a specific definition of a term is required. Therefore, in this article, we focus on the automated extraction of legal terms defined in the laws of the Slovak Republic. We present our efforts collecting these terms from several publicly available databases as well as the headlines of legal paragraphs using several automated filtering methods and metrics to evaluate the validity and usefulness of these potential legal terms based on their wording, definition, and usage in legal texts. In addition to the legal terms themselves, we are also concerned with extracting their definitions, which we attempt to do using a mix of rule-based extraction systems as well as methods reliant on language models.

Keywords

legal terms, term definition, automated extraction

1. Introduction

The long-term goal of our project is to enable the filtering and retrieval of documents and relevant chunks of documents for legal research and (Retrieval Augmented Generation) RAG systems. One of the main approaches to legal text classification is to assign legal terms that describe the legal qualification of the matter under decision. These terms can then be used when searching relevant legal texts. In our previous research, we presented several methods for extracting key terms or keyphrases from court decisions [1]. Although the keyphrases obtained in this way summarized the text of the court decision well, the extracted phrases mainly described the circumstances of the legal case under discussion, and only some of them specified the legal qualification.

To improve our methods, we must create a thesaurus of legal terms that would contain a set of keyphrases suitable for describing judicial decisions. Using legal terms such as this as candidates, we can use more robust keyphrase extraction methods, such as those described in [2]

In addition to the term itself, we also need its definition. For one, they serve as a good text representation of the term that can be used to create a semantic embedding, which can then be used to look up related documents or chunks of documents via a vector database.

There are several ways that we can use to obtain such a dictionary. For one, there are a multitude of online sources such as SLOV-LEX, the Slovak Law Thesaurus, the list of terms from Najprávo.sk, a law information system for experts and the public alike as well as thesaurus of legal terms we obtained from the Analytical department of the Supreme Court of the Slovak Republic. However, these only contained a little over three thousand terms total. And while that's not an insignificant amount, there are many viable legal phrases not found among them. Therefore we have opted to use the headlines of legal paragraphs as a pool of potential candidate keyphrases and the headlines themselves as their definitions. In this article, we mainly focus on the automated creation of a dictionary of legal terms using such a source. We present two approaches that reflect two common ways to define legal terms in the laws of the Slovak Republic.

The first way of defining legal terms in the laws of the Slovak Republic is that in some laws, there are paragraphs with a title such as "definition of terms", "basic terms", "interpretation of terms" etc. In this

ITAT'25: Information Technologies -– Applications and Theory, September 26–-30, 2025, Telgárt, Slovakia

*Corresponding author.

†These authors contributed equally.

case, the individual terms are presented as a definitions list. Each item in such a list usually contains a legal term along with its definition. The biggest challenge of the automated processing of such legal definitions is precisely revealing where the term is found in the text and where its definition is.

The second way of defining legal terms is that a separate paragraph is dedicated to the term, which also contains its legal use in addition to the definition of the term. In such a case, the term is given as the name of the paragraph, while its definition is usually at the beginning of the text of this paragraph. The automated processing of such terms faces mainly the problem of eliminating those paragraph names that are not legal terms.

2. Related work

Several studies have used various methodologies and technologies to address the automatic extraction of legal terms and their definitions from laws.

A statistical method called *C-value/NC-value* [3] is a domain-independent method for the automated extraction of multi-word terms. Specifically, *C-value* considers the frequency and co-occurrence of terms and is sensitive to nested terms, which are composed of more general terms. *NC-value* incorporates context information by ranking words by importance in the context window of extracted terms.

In work [4], authors extracted concepts, their relations, and definitions from sources of law in the tax domain using simple NLP and semantic web technology. To extract concepts, authors used Part-of-speech tagging on every article and then used regular expressions to extract nouns and noun phrases and marked them as concepts. For definition extraction, they suggest that a definition consists of three model fragments: concept, definition, and scope declaration (conditions and scopes for the definition to apply, e.g., current law of the article). To extract definitions, the authors used the MetaLex Annotator tool. However, they achieved a recall of only 42%.

Authors of [5] presented extraction of terms and their definitions from Australian contracts to visualize their definition networks. First, they used regular expressions to search for the terms 'means' and 'includes' and classified these sentences as definitions. In the next step, they segmented definitions from each other by identifying a full-stop punctuation mark. Then, they extracted the defined term and the defining text for each definition. This tool prototype was published online; users could edit the automatically generated results and achieve 100% Authors previous research [6] with a fully automated solution using machine learning methods achieved an accuracy of around 80%, making the semi-automatic solution more usable.

Another study, [7], aimed to produce a Japanese legal terminology comprising legal terms and their explanations, along with accessible citations. While the authors successfully identified over 14,000 terms with high precision, they encountered a challenge where 23.1% of the correct explanations included inaccessible citations due to their context-dependent format. To address this issue, the paper proposed a method that involved revising explanatory sentences by considering XML-tag annotation for context-independent formatting of all citations. Experimental results confirmed the effectiveness of this approach, highlighting its potential for improving the accessibility and comprehensibility of legal terminologies.

The same group of researchers studied the development of diachronic changes in Japanese legal terminology [8]. By using regular expressions, they were able to search for articles containing definitions. They could extract legal terms and explanatory sentences with another set of regular expressions. They also used regular expressions to extract the IS-A relationship between terms from explanatory sentences, precisely one regular expression to extract hypernym and another for hyponym.

In works [9, 10, 11], definition extraction was taken as a classification task, where the input was short snippets or individual sentences from legal texts. Using traditional machine learning and natural language processing, they classified sentences such as prohibitions, delegations, obligations, citations, and, as of our interest, definitions.

3. Methods

In this section, we present our methods for automatic extraction of legal terms and their semantics from the laws of the Slovak Republic. First, we focus on methods applicable to special paragraphs dedicated to terms definitions.

3.1. From the Law to the structured dictionary

In the laws of the Slovak Republic, it is common to find multiple definitions of legal terms grouped in a paragraph labeled with a header containing the word *pojem* (meaning term) in various inflections. We have identified 40 different headers that contain terms' definitions. However, the content of these paragraphs varies in a template of definitions list.

In the following paragraphs, we will describe rule-based methods based on linguistic analysis. If these disjoint sets of rules are applied to the definition, they can identify the boundaries between terms and their definitions.

3.1.1. Separators

One approach to identifying the separation between terms and definitions is using an ordered list of commonly used separators. These separators can include phrases such as *je na účely tohto zákona* (meaning "for the purposes of this law"), *podľa tohto zákona* (meaning "by this law"), *je aj* (meaning "is also") or *je* (meaning "is"). It is important to note that the separators in the list should be searched in the text in the given order. This is because some text samples contain more than one separator from the list. For example, if the text includes both "je" and "podľa tohto zákona", the suitable separator of a term and its definition is "podľa tohto zákona". The ordering of separators is made by a domain expert. Totally we have identified 10 different separators.

3.1.2. Regular expressions

Another method involves regular expression to identify the position of a term. Specifically, we are using pattern

`\(ďalej len [^)]+\)`[1]

which means "in the following used as". It can be used to indicate the whole term. In this example, the definition of the term or a synonym of the term is usually present before the parenthesis matched by regular expression.

3.1.3. POS as a term identifier

Since Slovak is an inflected language, many times, only the change of the declension of words can indicate where the term ends and its definition begins. Therefore, part-of-speech (POS) tags can be utilized as separators. Examining the POS tags assigned to each word in the text makes it possible to identify specific patterns that indicate the separation between terms and definitions. For instance, POS tag $S7$ (noun, instrumental) or $A7$ (adjective, instrumental) at the first place (corresponding with the first word) followed by a series of ...7 (words with the same declension - instrumental) ended by $S/A1$ (noun or adjective, nominative) could signal the beginning of a definition (with the word corresponding with the $S/A1$ tag). Currently, we recognize four different POS tags patterns in total.

3.2. Legal term extraction from law paragraph headlines

The headlines of law paragraphs can often be considered law terms of their own that have their meaning defined in the text of the paragraph. However, not all headlines are suitable to be used as legal terms to

[1]The regular expression matches a specific pattern in a text where it starts with the phrase *(ďalej len* and is followed by one or more characters that are not the closing parenthesis and ends with the closing parenthesis character.

annotate documents with. Sometimes the headline is not semantically bonded to the law and is more general, e.g. "Definitions of terms" or "Final provisions". In some cases, the headlines have too specific meaning and cannot be considered as legal terms, e.g. "How the tax is determined" or "Action plan to ensure a multimodal approach in the provision of on-demand audiovisual media services"

We have implemented several metrics designed to test the suitability of terms for the description and categorization of court decisions. These metrics can be used individually or combined with one another to determine the suitability of the given headline to be used as a law term.

3.2.1. Determining validity through term name analysis

The first and simplest of these metrics is *the number of words that make up the potential term.* Headlines of paragraphs that are longer have a higher probability of being sentence-like or too specific, which makes them less suitable to use as term and too specific to use for categorizing court decisions.

3.2.2. Determining validity through term frequency

The second type of metric relies of the occurrences of a given potential term from a headline within collections of legal texts. The most simplistic are those that rely on the number of times a given term appears in a given collection of legal texts.

The simplest metric is obtained by calculating the number of times the given phrase appears in the collection of laws. The more times a given term appears in the collection of terms, the higher the probability that term is relevant for being used as a keyphrase to describe court decisions.

One other way of determining metrics is to use the dictionary-based graph generation method described in [12]. Let $G = term_1, term_2, ..., term_n$ be the set of all potential terms. This method creates a directed weighted graph where G can be viewed as the set of all vertices. Let E be defined as the set of all relations between terms of G. Let $D_i = d_{i1}, d_{i2}, ...d_{im}$ be the set of all law paragraphs that have the term $term_i$ as a headline. If there exists a law paragraph d_{ij} that contains $term_l$ then a directed edge $(term_i, term_l)$ is added to the set of all relations E. After constructing such a graph the indegree for a given $term_i$ tells us in how many law paragraphs belonging to other terms from G is that $term_i$ mentioned in. The outdegree of a specific $term_i$ on the other hand tells us how many other terms from G are cited in the various law paragraphs $term_i$ serves as a headline for. Both of these values can be used as separate metrics.

3.2.3. Determining validity through term frequency combined with name analysis

An improved version of this idea relies on the calculation of the occurrences of N-grams of words in the Slovak collection of laws. We have calculated the frequencies of all unigrams, bigrams and trigrams located within the text. Calculating n-grams of higher order was computationally infeasible. We then use this list to calculate what we refer to as an inseparability metric. The idea is to look at a word as an n-gram of terms.

Let $term_i = w_{i1}, w_{i2}, ..., w_{ik_i}$ be i-th term from our database of headline, where w_{ij} denotes the j-th word making up the i-th term.

1. If $k_i = 1$ or the term consists of a single word, then the inseparability metric is incalculable
2. If $k_i = 2$ or the term consists of two words, the inseparability metric is calculated as

$$\text{inseparibility}(term_i) = \frac{2 \times \text{freq}(w_{i1}, w_{i2})}{\text{freq}(w_{i1}) + \text{freq}(w_{i2})} \tag{1}$$

3. If $k_i = 3$ or the term consists of three words, the inseparability metric is calculated as

$$\text{inseparibility}(term_i) = \frac{2 \times \text{freq}(w_{i1}, w_{i2}, w_{i3})}{\text{freq}(w_{i1}, w_{i2}) + \text{freq}(w_{i2}, w_{i3})} \tag{2}$$

4. If $k_i > 3$ or the term consists of more than three words, the metric is calculated as the term is split into trigrams of words. Let $T_i = t_{i1}, t_{i2}, ..., t_{ik_{i-2}}$ be the set of word trigrams created from $term_i$, where $t_{ij} = (w_{ij}, w_{ij+1}, w_{ij+2})$ The inseparability metric is calculated for every trigram $t_{i1}, t_{i2}, ..., t_{ik_{i-2}}$ the following way:

$$\text{inseparablity}(t_{ij}) = \frac{2 \times \text{freq}(w_{ij}, w_{ij+1}, w_{ij+2})}{\text{freq}(w_{ij}, w_{ij+1}) + \text{freq}(w_{ij+1}, w_{ij+2})} \tag{3}$$

After calculating the inseparability metric for every trigram of the term, we obtain the final inseparability metric for the given term by calculating the mean of the inseparability values of its trigrams.

$$\text{inseparablity}(term_i) = \frac{\sum_{j=1}^{k_{i-2}} \text{inseparablity}(tr_{ij})}{k_{i-2}} \tag{4}$$

3.2.4. Determining validity through statistical definition analysis

We can also take into account metrics that are reliant on exploring the text of the law paragraphs that the potential term appears in. Since there are multiple law paragraphs corresponding to the same headline and therefore potential terms, these metrics are considered fulfilled if it fulfills this condition in any one of them. One such type metric lies in taking into account whether the headline, our potential term, appears in the law paragraph. Naturally, these metrics are calculated over lemmatized text. For the lemmatization, we make use of the Slovak word form dictionary called Tvaroslovník [13]. There are three metrics we can derive from this idea:

1. whether a paragraph contains its given headline
2. whether the first sentence of a paragraph contian its headline
3. whether the paragraph begins with its given headline.

3.2.5. Determining validity through definition suitability analysis via language models

We can also make use of language models. we use the SlovakBERT model described in [14] fine-tuned on all Slovak legal texts we possess including all judicial decisions and collection of laws. We then created a classification model with the task of classifying whether the paragraph corresponding to the headline contains a ”definition”. Since the paragraph is quite long, we split it into window-sized chunks and if any meets the threshold of a definition, we set the value to one. Since we don’t have enough manually labelled definitions from the collection of law paragraphs, especially negative examples, we have opted to use datasets that we consider representative enough for the task. As positive examples of texts containing a definition, we have used the legal definitions contained in the Slovak Law Thesaurus, SLOV-LEX [15]. As negative examples, texts not containing definitions, we used those law paragraphs that did not have a headline. The model trained on this dataset was then used as to predict whether the given law paragraph contained a definition.

The advantages of such a method lie in the fact that we can use the classifier to extract the parts of the paragraph it labels as a definition to obtain a definition of the paragraph, although the problem of finding the exact cutoff point remains as our spitting into chunks based on sentence separators is not always accurate.

We can employ a similar approach using foundation models/LLMs. We have experimented with the 7B version of Llama 3.1. We have prompted the model to return a YES or NO answer whether the description of the term is its definition. We have set the model temperature to 0 in order to force the most probable reply and limit the verbosity of the model by setting the max token length to two. In case it returned any other answer than yes or no, we discarded it and ran the same prompt again.

We also made a few-shot variant of the same prompt, giving it two examples with the definitions extracted from other sources and two negative examples from the law paragraphs with no headlines that we manually checked did not contain definitions.

This method can also be used to extract definitions, though it suffers from a similar issue as the SlovakBERT approach, that being the inaccurate exact chunking. A snippet of a legal paragraph extracted only depending on whether it fulfilled the condition of containing a definition, might only contain an incomplete fraction of it or on the flip side, contain excess text.

So, for this particular task, we can modify the prompt to instead extract the definition out of law paragraphs chunks that our previous approach deemed to contain one.

4. Evaluation

4.1. Rule-based methods

These rule-based methods for extracting terms and their definitions have been evaluated on a manually annotated dataset created from actual paragraphs of laws in the Slovak Republic. The achieved success rate is 78%. The success of individual methods is presented in the table 1. This validation demonstrates the effectiveness of these techniques in identifying and separating legal terms and their corresponding definitions.

Although the dataset consists of only 375 samples, the rule-based methods presented here can contribute to creating a larger dataset suitable for training more advanced models, such as neural networks.

The simple rule-based method based on extracting repeating patterns and language analysis does not cover the complexity of our problem. The 22% error is caused by the significant variability of the definition structure and indicates the need to use more sophisticated methods.

By leveraging these techniques to generate a comprehensive dataset, it becomes possible to enhance the training process and develop more complex automated annotation systems for legal texts, including court decisions.

Table 1
Percentage success distribution of individual methods

Submethods	Covering of success
Regular Expression	6%
Separators	67%
POS Tags	27%

4.2. Legal term extraction from law paragraph headlines

We have created a dataset of potential legal terms from the collection of law paragraphs headlines. This is a dataset comprised of 34628 entries total. We combined and cross referenced these entries with other sources of legal terms obtained from other external dictionaries. The first and most relevant of these is the collections of terms from the Slovak Law Thesaurus, SLOV-LEX [15]. We have combined this with the thesaurus of legal terms we obtained from the Analytical department of the Supreme Court of the Slovak Republic. Finally, we have also extracted the list of legal terms from Najprávo.sk, a law information system for experts and the public alike.

We have tested the performance of our metrics on a dataset of terms that are both in the collection of potential terms from law paragraph headlines, as well as being present in at least one of the sources above. This dataset consists of 1290 entries total.

We have then opted to test out the various filters described above. We can calculate the estimated recall value of our filters by seeing how much of the 1290 manually extracted entries - which, unlike the headlines, we can use as ground truth- remain after being passed through our filters. We will refer to this metric as Recall. For the sake of interpretability, we will display the recall values as percentages. As for precision, we couldn't calculate the exact metric since we have no labels we can use for the potential legal terms from law paragraph headlines, so we randomly chose 10 entries that the filter

discarded and manually evaluated how many of those were discarded correctly. We will refer to this metric as Specficity Estimate (SE). Finally, we will also display the number of discarded potential terms by using one of our metrics.

The first filter is the one based on the number of words that comprise a given term. The idea of this metric is to filter out terms that are too long as they might be overly specific or even sentence like in nature, as described in subsection 3.1.1. We have plotted out our results into table 4.2

Table 2
Table of results for the number of words metric.

N of words:	5+	7+	9+	11+	12+	13+
Recall:	69.31	95.66	98.29	99.53	99.77	100.0
SE:	1/10	3/10	6/10	8/10	8/10	10/10
Discarded:	22042	9936	5923	3261	2383	1992

We have found that 1199 out of those 1290 entries are comprised of five words or less and there are no terms comprised of more than 12 words within the 1290 entries. And the entries that such a filter discards are usually terms or sentences in the Czech language from the times before the separation of Czechoslovakia or terms written with spaces between their letters. Despite the simplicity of the metric it seems surprisingly effective, even though it doesn't discard that many terms.

We have also opted to test the metric that relies on the number of citations of a given potential legal term within the Slovak collection of laws. The idea behind this metric is that the more times a given potential term is cited, the more probable it is a valid legal term. We have plotted the results into table 3.

Table 3
Table of results for filtering potential terms based on the number of their citations in the Slovak collection of laws.

Citation count	500+	100+	20+	5+
Recall:	11.86	24.96	44.65	51.08
SE:	0/10	0/10	1/10	1/10
Discarded:	34186	33128	30530	28926

As we can see, while the general idea seems to hold up, it seems that even if the potential term is not one often cited in the collection of laws, it is still often times a valid legal term. Filtering using such a metric has little merit. Just because a given term is not cited frequently doesn't mean that it isn't a valid legal term.

Furthermore, we have tested out the our graph-based metrics for filtering out potential terms. One of them involves filtering based on the indegree and outdegree of the term graph the construction of which we describe in subsection 3.2.2. We have plotted these results into table 4

Table 4
Table of results for filtering potential terms based on node incidence in definition graph

Metric	IN: 200+	IN: 50+	IN: 5+	OUT: 200+	OUT: 50+	OUT: 5+
Recall:	12.64	17.91	23.95	2.09	17.91	61.16
SE:	0/10	1/10	0/10	0/10	0/10	2/10
Discarded:	34348	33886	33401	34598	32944	23669

From the table we can see the graph based metrics are also not that suited to filter through our dataset, as they discard far too many entities. However interestingly enough it was the outdegree metric, which calculates how many other terms are located within the law paragraphs corresponding to a given term that achieved the best results. It makes sense, considering the less other terms used to define it, the more it can be considered a core or base term, while a term being used in the definition of others, as

generic as it may be, is probably a valid one to annotate documents with, even if it doesn't provide too much information.

Next, we have evaluated the results using the inseparability metric described in subsection 3.2.3. and plotted them into table 5.

Table 5
Table of results for filtering potential terms based on their inseparability metric.

Inseparability rate	0.5+	0.1+	0.01+	0.001+
Recall:	4.88	14.10	28.83	32.64
SE:	0/10	0/10	1/10	1/10
Discarded:	33775	31903	30312	29752

As we can see, this metric is also not ideal for filtering out invalid terms. Part of it is due to the fact that this metric cannot be used for legal terms made up of single words, so this metric can only be used on specific entries of the dataset. Moreover, even if a term is created by using common words, it does not necessarily indicate the term is invalid, as those legal terms might describe very specific situations.

The following filters are those that examine the law paragraphs, for the citation of the term as well as its as a definition. There are three metrics overall, described in subsection 3.2.4. as:

- One of the paragraphs with the term as its headline having its first word be the aforementioned term
- The first sentence of one of the paragraphs with the term as its headline containing the aforementioned term
- One of the paragraphs with the term as its headline containing the aforementioned term

Since these metrics are binary in nature, there's no need to explore their hyperparameters, therefore we decided to plot the results into table 6

Table 6
Table of results for all statistical methods for definition validity

Metric	Starts with term	Term in first sentence	Contains term
Recall:	12.56	60.31	62.32
SE:	0/10	2/10	1/10
Discarded:	32898	29328	27354

As we can see, the term being the first word of the paragraph is fairly uncommon, as such it isn't really a good idea to discard a term that doesn't meet this criterion. The term being contained in the first sentence of the paragraph seems to be a slightly better metric, however it still seems to discard far too many valid terms. There's not much difference if we further loosen our condition by looking at whether the term appears in the definition at all. As a general rule, when a term is contained in its decision, it seems to be at the beginning.

Finally, we've evaluated the metrics using transformer-based models and LLMs which try to determine the suitability of a headline by determining whether its description contains a description.

There are three methods we've tested, described in subsection 3.2.5. as:

- Classification whether any part of the law paragraphs contains a definition via SlovakBERT
- Classification whether any part of the law paragraph contains a definition via zero-shot Llama 3.1 prompting
- Classification whether any part of the law paragraph contains a definition via few-shot Llama 3.1 prompting

Table 7
Table of results for all language model based methods for definition validity

Method	SlovakBERT	Llama 3.1 zero-shot	Llama 3.1 few-shot
Recall:	98.68	95.73	99.24
SE:	9/10	9/10	10/10
Discarded:	2718	2853	2649

We have plotted our result into table 7

As we can see from the results of this table, determining whether the headline is a suitable legal term via the validity of its corresponding paragraph to serve as its definition using language models seems to be the most efficient approach to resolve this problem. As we can gather from the results, the best results were achieved using the Llama 3.1 large language model with multi-shot prompting. It had the highest Recall and PE values of the two, with the number of discarded terms being the smallest in its category, though still considerably more than the naive method of only discarding terms consisting of 13 or more words. The second best results weren't achieved by the LLM however, but rather by the SlovakBERT method, achieving higher recall than the zero-shot Llama 3.1 and a matching SE, and even though the latter is less of an objective metric, the fact that the SlovakBERT method discards less phrases overall means that its SE might be at least equal if not higher. We believe this could be due to a number of factors, such as more compatibility with the language and legal texts or having been trained as a classifier for this specific purpose. However, perhaps due to the fact that it's possible not all law paragraphs without a headline truly lack a definition, it might be the case that the training data is itself faulty or the language model is overall far inferior compared to the LLM. Still, it managed to achieve higher results as when we simply ask the model to decide whether a given text is a definition of that headline without any examples provided.

5. Conclusion and Future work

The overall conclusion we can draw from these results is that the majority of headlines seem to be suitable as legal terms and key phrases to annotate documents. The best results were conversely seemingly achieved by the simplest and most complex of methods (in terms of architecture at least). The method to discard all terms with a high word count and to check whether their paragraphs contain a definition seem to be the most surefire way to distinguish headlines suitable to be legal terms. The latter methods have the advantage that they have mechanisms through which they're able to extract the definitions of these terms as well.

As the final algorithm to determine the validity of terms, we chose to first discard all the terms that have a length of 13 or more terms and then ran what we presume to be our most successful method for headline suitability based on whether its corresponding paragraph contains a headline, the few-shot Llama 3.1 prompting on the truncated dataset discarding the longest of terms. This combined approach resulted in a Recall of **99.76**, with an PE of **10/10** and a total of **2520** discarded terms. We use the dataset obtained by this method and combined with the terms from other, externally sourced legal databases as our final one, extracting its definitions using the Llama 3.1 few-shot prompting as well.

The rule-based methods discussed in this paper can not only help identify term and definition boundaries but also facilitate the creation of a more extensive dataset. This expanded dataset can be utilized for training advanced machine learning models, thereby enhancing their effectiveness. The methods for extracting definitions from the law paragraphs themselves haven't been thoroughly evaluated but early results appear promising.

As our future work, we wish to evaluate the validity of the definitions extracted by methods using language models, even testing a wider variety such as BGE-M3, modernBERT or the slovak 7B Mistral model. We further wish to obtain more labelled data, using our dictionary as a recommender system and have legal experts annotate all terms and their extracted definition. Once complete we can speculate on

ways to improve our methods, such as by utilizing a neural network that would take the values of all calculated metrics as inputs and create a complex classifier that takes multiple metrics into account with variable weights. We also plan to combine our legal dictionary with vector databases and test it for retrieval tasks and clustering tasks.

`\bibliography{bibfile}`

where "`bibfile`" is the name, without the "`.bib`" suffix, of the BibTEX file.

Acknowledgments

This work was supported by the Slovak Research and Development Agency under contract No. APVV-21-0336 *Analysis of Court Decisions by Methods of Artificial Intelligence*, and by the Ministry of Education, Science, Research and Sport of the Slovak Republic under contract No. VV-MVP-24-0038 *Analysis of Liability for Internet Torts with Machine Learning Methods*.

Declaration on Generative AI

Either:
The author(s) have not employed any Generative AI tools.

Or (by using the activity taxonomy in ceur-ws.org/genai-tax.html):
During the preparation of this work, the author(s) used X-GPT-4 and Gramby in order to: Grammar and spelling check. Further, the author(s) used X-AI-IMG for figures 3 and 4 in order to: Generate images. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] D. Varga, Š. Horvát, Z. Szoplák, L. Antoni, S. Krajči, P. Gurský, L. B. Rózenfeldová, Keyphrase extraction from slovak court decisions, CEUR Workshop Proceedings Vol-3226 (2022) 142–150.

[2] Z. Szoplák, P. Gurský, D. Varga, Optimizing Keyphrase Extraction for Court Decisions Using Legal References, 2024. doi:10.3233/FAIA241265.

[3] K. T. Frantzi, S. Ananiadou, The c-value/nc-value domain-independent method for multi-word term extraction, Journal of Natural Language Processing 6 (1999) 145–179.

[4] R. Winkels, R. Hoekstra, Automatic extraction of legal concepts and definitions, in: Legal Knowledge and Information Systems: JURIX 2012: the Twenty-Fifth Annual Conference, volume 250, IOS Press, 2012, pp. 156–165.

[5] M. Curtotti, E. McCreath, S. Sridharan, Software tools for the visualization of definition networks in legal contracts, in: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law, 2013, pp. 192–196.

[6] M. Curtotti, E. Mccreath, Corpus based classification of text in australian contracts, in: Proceedings of the Australasian Language Technology Association Workshop, 2010.

[7] M. Nakamura, Y. Ogawa, K. Toyama, Extraction of legal definitions and their explanations with accessible citations, in: AI Approaches to the Complexity of Legal Systems: AICOL 2013 International Workshops, AICOL-IV@ IVR, Belo Horizonte, Brazil, July 21-27, 2013 and AICOL-V@ SINTELNET-JURIX, Bologna, Italy, December 11, 2013, Revised Selected Papers, Springer, 2014, pp. 157–171.

[8] M. Nakamura, T. Ogawa, K. Toyama, Development of diachronic terminology from japanese statutory corpora, J. Open Access L. 4 (2016) 1.

[9] E. de Maat, K. Krabben, R. Winkels, Machine learning versus knowledge based classification of legal texts, in: Legal Knowledge and Information Systems, IOS Press, 2010, pp. 87–96.
[10] E. Francesconi, S. Montemagni, W. Peters, D. Tiscornia, Integrating a bottom–up and top–down methodology for building semantic resources for the multilingual legal domain, Springer, 2010.
[11] C. Biagioli, E. Francesconi, A. Passerini, S. Montemagni, C. Soria, Automatic semantics extraction in law documents, in: Proceedings of the 10th international conference on Artificial intelligence and law, 2005, pp. 133–140.
[12] S. Horvát, S. Krajči, L. Antoni, Semantic representation of slovak words, CEUR Workshop Proceedings Vol-2718 (2020).
[13] S. Krajči, R. Novotný, Tvaroslovník–databáza tvarov sl'ov slovenského jazyka., Proceedings of international conference ITAT 2012, SAIA (2012) 57–61.
[14] M. Pikuliak, Štefan Grivalský, M. Konôpka, M. Blšták, M. Tamajka, V. Bachratý, M. Šimko, P. Balážik, M. Trnka, F. Uhlárik, Slovakbert: Slovak masked language model, 2021. arXiv:2109.15254.
[15] Slovak law thesaurus, Legislative and information portal, Ministry of Justice of the Slovak Republic (2022). URL: https://www.slov-lex.sk/tezaury/terminy.