Discovering Explanatory Sentences in Legal Case Decisions Using Pre-trained Language Modelss

Jaromir Savelka
School of Computer Science
Carnegie Mellon University
jsavelka@cs.cmu.edu

Kevin D. Ashley
School of Law
University of Pittsburgh
ashley@pitt.edu

Abstract

Legal texts routinely use concepts that are difficult to understand. Lawyers elaborate on the meaning of such concepts by, among other things, carefully investigating how have they been used in past. Finding text snippets that mention a particular concept in a useful way is tedious, time-consuming, and, hence, expensive. We assembled a data set of 26,959 sentences, coming from legal case decisions, and labeled them in terms of their usefulness for explaining selected legal concepts. Using the dataset we study the effectiveness of transformer-based models pre-trained on large language corpora to detect which of the sentences are useful. In light of models' predictions, we analyze various linguistic properties of the explanatory sentences as well as their relationship to the legal concept that needs to be explained. We show that the transformer-based models are capable of learning surprisingly sophisticated features and outperform the prior approaches to the task.

1 Introduction

Written laws enacted by legislative bodies set forth the collection of legally binding rules of conduct (e.g., rights, prohibitions, duties). Understanding written laws is difficult because the abstract rules must account for a variety of situations, even those not yet encountered. Written laws communicate general standards and refer to classes of persons, acts, things, and circumstances (Hart, 1994, p. 124). Therefore, legislators use vague (Endicott, 2000), open textured (Hart, 1994) terms, abstract standards (Endicott, 2014), principles, and values (Daci, 2010) to deal with the inherent uncertainty.

For example, let us focus on the two emphasized concepts from the following written provision of law (29 U.S. Code § 203):

> "Enterprise" means the *related activities* performed [...] for a *common business purpose* [...].

Understanding of the provision depends on understanding the meaning of the two emphasized concepts. Any doubts about the meaning may be removed by explanation or interpretation (MacCormick and Summers, 1991). Even small differences in understanding of a single concept may be crucial for determining how a provision applies and what are its effects in a particular context.

For example, the meaning of the concept *common business purpose* could be crucial in determining if two restaurants in different parts of the same city, sharing a single owner, constitute an "enterprise." The explanation of the concept would involve an investigation of how has it been referred to, explained, interpreted, or applied in the past. This is an important step that enables a lawyer to come up with arguments in support of or against particular accounts of meaning (Šavelka and Harašta, 2015; Savelka and Ashley, 2021).

Searching through a database of legal documents a lawyer may retrieve sentences such as the following:

1. Courts have held that a joint profit motive is insufficient to support a finding of *common business purpose*.
2. The fact of common ownership of the two businesses clearly is not sufficient to establish a *common business purpose*.
3. The third test is "*common business purpose*."

Some of these sentences are most likely useful for explaining the concept (1 and 2) but others appear to have very little value (3). Manually reviewing such sentences is labor intensive.

We would like to rank more highly the sentences the goal or effect of which is to elaborate upon the meaning of the selected concept. These include, but are not limited to, (i) definitional sentences (e.g., a sentence that provides a test for when the concept applies), (ii) sentences that state explicitly in a different way what the concept means or state what it does not mean, (iii) sentences that provide

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 4273–4283
November 7–11, 2021.

an example, instance, or counterexample of the concept, and (iv) sentences that show how a court determines whether something is such an example, instance, or counterexample.

2 Related and Prior Work

In prior work, we employed a variety of traditional information retrieval (IR) measures and their combinations, e.g., BM25, novelty, topic modeling (Savelka et al., 2019; Savelka, 2020; Šavelka and Ashley, 2021). These turned out to be remarkably successful in finding documents or their parts (e.g., paragraphs) that are likely to contain useful sentences. However, they fell short in performing finer-grained evaluation of the sentences contained in those document parts. Using a learning-to-rank approaches on hand-crafted features led to only moderate improvements (Šavelka and Ashley, 2016; Savelka and Ashley, 2020). In this work, we show that transformer-based pre-trained models (BERT family) are capable of such fine-grained evaluation by learning to detect sophisticated semantic features in sentences themselves as well as in their relationships to the explained concepts. Furthermore, we show that many of these features are sensible to humans.

The models based on the BERT architecture have been successfully used in a variety of IR tasks. A comprehensive survey of text ranking with transformers, such as BERT, is provided in (Lin et al., 2020). Several simple applications of BERT to ad hoc document retrieval are presented in (Yang et al., 2019). Successful applications of BERT for retrieval of short texts such as sentences are presented in Yilmaz et al. (2019) and Rao et al. (2019). Similar to the utilization of provisions of written law in this work, the authors of Mehrotra and Yates (2019) demonstrated the effectiveness of using a query context in a re-ranking component based on BERT. In Nogueira et al. (2019) BERT is fine-tuned on query-retrieved document pairs as is done in this work.

There are examples of successful applications of BERT on legal texts as well. A task of retrieving related case-law similar to a case decision a user provides is tackled in Rossi and Kanoulas (2019). BERT was also proposed as one of the approaches to predict court decision outcomes given the facts of a case (Chalkidis et al., 2019). BERT has been successfully used for classification of legal areas of Supreme Court judgments (Howe et al., 2019). BERT was used to tackle the challenging task of case law entailment (Rabelo et al., 2019; Westermann et al., 2020). BERT was also used in learning-to-rank settings, as is done in this work, for retrieval of legal news (Sanchez et al., 2020). Systematic investigation of BERT's adaptation to the legal domain, resulting in a release of several legal-BERT models, was performed in (Chalkidis et al., 2020). RoBERTa (variation of BERT) model was used for classification of legal principles applied in court case decisions (Gretok et al., 2020). The ability of pre-trained language models (RoBERTa) to generalize beyond the legal domain and dataset they were trained on was analyzed in (Šavelka et al., 2020).

3 Data Set

We downloaded the complete bulk data from the Caselaw access project[1] which includes all official, book-published U. S. cases from all federal and state courts as well as from a number of territorial courts (President and of Harvard University, 2018). The dataset comprises more than 6.7 million unique cases. For document indexing we used a lemmatizer based on the so-called induced ripple-down rules (Juršic et al., 2010).[2] Using the U.S. case law sentence segmenter (Savelka et al., 2017) we divided each case into individual sentences (0.8 billion).

We queried the system for sentences mentioning 42 selected legal concepts (i.e., terms/phrases, such as "audiovisual work," or "electronic signature") coming from provisions of the U.S. Code (the official collection of federal statutes).[3] Given the constraints imposed by available resources, we made the best effort to create a well-balanced dataset covering 20 different areas of legal regulation (26,959 retrieved sentences in total).

Eleven law students classified the sentences in terms of four categories with respect to their utility for explaining the legal concepts:

1. **High value** – This category is reserved for sentences the goal of which is to elaborate on the meaning of the concept.

2. **Certain value** – Sentences that provide

[1] A small portion of the dataset is available at `case.law`. The complete dataset could be obtained upon entering into research agreement with LexisNexis.

[2] `http://lemmatise.ijs.si`

[3] `https://www.law.cornell.edu/uscode/text/`

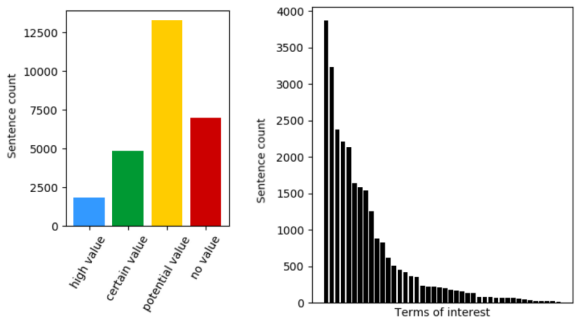

Figure 1: The graph on the left shows the distribution of the labels across all the sentences retrieved for the 42 selected concepts. The graph on the right presents the distribution of the number of sentences retrieved for each concept.

grounds to draw some conclusions about the meaning of the concept.

3. **Potential value** – Sentences that provide additional information over what is known from the provision of law.

4. **No value** – Sentences that do not provide any additional information over what is known from the provision.

Annotators needed to be properly trained to deal with this challenging task. We adopted multiple measures to ensure the annotations of the resulting dataset are of high-quality. The most important one was a second-pass annotation performed by two annotators with a completed law degree ($\alpha = 0.79$).[4]

Figure 1 shows the overall distribution of the labels and the number of sentences associated with each concept/query. The Figure shows that the less valuable categories, 'no value' and 'potential value,' are dominant. For all the "larger" queries and almost all the "small" queries it holds that either the 'no value' or the 'potential value' category is the most numerous one. No matter the size, it is still the case that some of the terms contain quite a considerable number of more valuable sentences (e.g., "audiovisual work" or "switchblade knife") while others are significantly more limited in this respect (e.g., "essential step" or "hazardous liquid"). As the dataset has not yet been released to the public we are making it available with this paper.[5]

[4]To measure inter-annotator agreement we used Krippendorff's α (Krippendorff, 2011).

[5]https://github.com/jsavelka/statutory_interpretation

4 Experiments

In this work, we use RoBERTa—a robustly optimized BERT pretraining approach (Liu et al., 2019)—as the starting point for the rankers.[6] Out of the available language models we chose to work with the smaller roberta.base model that has 125 million parameters. This choice was motivated by the ability to iterate the experiments faster when compared to working with roberta.large with 355 million parameters.

We experiment with three different setups. In the first setup we fine-tuned the base RoBERTa model on the task of classifying retrieved sentences in terms of their value for explaining the legal concepts. In prediction we then applied the model to classify the sentences, that were not seen during fine-tuning, in terms of the four value categories (see Section 3). By applying softmax to the final prediction layer we obtained the probability distribution over the four possible classes. To obtain the sentence's score we compute an inner product between the class probability distribution and value weight vector $(0, 1, 2, 3)^T$. The motivation to use the approach over considering only the predicted class is to take into account the confidence of the prediction. Henceforth, this model is referred to as *BERT snt* because it infers the usefulness of a sentence for explaining a legal concept from the sentence only.

In the second setup we fine-tune the base RoBERTa model on the sentence pair classification task. The model is provided a legal concept in place of the first sentence and a retrieved sentence as the second one. The task of predicting sentence value is thus recast as predicting the relationship between the concept and the retrieved sentence. The goal is still to predict one of the four sentence value labels. As in the previous setup, we applied softmax to the final prediction layer to obtain a probability distribution over the classes. Sentences' scores are determined in the same way as well. Henceforth, this model is referred to as *BERT qry2snt*.

The third setup is similar to the second one. Here, we again fine-tune the base RoBERTa model on the sentence pair classification task. Unlike in the second setup, the model is fine-tuned on the whole provision of written law as the first sentence and the retrieved sentence as the second one. There-

[6]https://github.com/pytorch/fairseq/tree/master/examples/roberta

fore, in this experiment the task is understood as prediction of the relationship between the provision of law and the retrieved sentence. As in the two previous experiments, softmax is applied to the final prediction layer and the probability distribution over the classes is obtained. Henceforth, this model is referred to as *BERT sp2snt*.

In all the experiments, we fine-tuned the base RoBERTa model, with a linear classification layer on top of the pooled output, for 10 epochs on the training splits of the selected datasets. We used the batch size of 8 which is the maximum allowed by our hardware setup (1080Ti with 11GB) given we set the length of a sequence to 512 (maximum). As optimizer we use the Adam algorithm (Kingma and Ba, 2014) with initial learning rate set to $4e^{-5}$. We stored models' checkpoints after the end of each training epoch. The checkpoints are evaluated on the validation set (see Section 4.1 for details). The model with the highest F_1 on the validation set was then selected as the one to make predictions on the test sets.

4.1 Evaluation

Since the notion of relevance in this work is non-binary, we use normalized discounted cumulative gain ($NDCG$) to evaluate the performance of different approaches. An output of the presented ranking algorithms for each concept/query q_j has the form of an ordered tuple of sentences $S_j = (s_1, s_2, \ldots, s_n)$. We chose to evaluate the rankings at $k = 10$ and 100 which means that the tuples produced by the algorithms are truncated to the respective lengths. Note that the chosen values of k are higher than typical. Measuring at $k = 100$ may even appear somewhat extreme. However, legal search differs from the general web search. Assuming a lawyer has confidence in the query (based on seeing several relevant hits towards the top of the results' list), he or she might be inclined to inspect the results way beyond what would a typical web search user do. For each query q_j the $NDCG$ at each k is then computed as:

$$NDCG(S_j, k) = \frac{1}{Z_{jk}} \sum_{i=1}^{k} \frac{rel(s_i)}{log_2(i+1)}$$

The function $rel(s_i)$ takes a sentence as an input and outputs its value in a numerical form. It is defined as follows:

$$rel(s_i) = \begin{cases} 3 & \text{if } s_i \text{ has high value} \\ 2 & \text{if } s_i \text{ has certain value} \\ 1 & \text{if } s_i \text{ has potential value} \\ 0 & \text{if } s_i \text{ has no value} \end{cases}$$

Z_{jk} is a normalizing quantity which is equal to $NDCG(S_j, k)$ where S_j is the ideal ranking. In our case this would mean that all the s_i with 'high value' labels are at the beginning positions of the tuple, followed by those with the 'certain value,' then 'potential value,' and finally 'no value' sentences.

We used stratified sampling to distribute the queries into six folds. There are many dimensions along which the result lists associated with the individual queries could be assessed. Two very important ones are the size of the list (i.e., the number of retrieved sentences) and its richness. Richness is a term often used in technology assisted review in eDiscovery. It refers to the prevalence of responsive documents in a collection (result list in case of this work). We adapted the idea for this work by defining a measure that describes the prevalence of valuable sentences in the dataset. First, we assigned a value to a sentence s_i depending on its label on a scale from 0 to 10:

$$val(s_i) = \begin{cases} 10 & \text{if } s_i \text{ has 'high value'} \\ 5 & \text{if } s_i \text{ has 'certain value'} \\ 1 & \text{if } s_i \text{ has 'potential value'} \\ 0 & \text{if } s_i \text{ has 'no value'} \end{cases}$$

The reason why we used the scale of 0 to 10 is to overcome the dominance of the less valuable sentences. It is important to emphasize that these scores do not reflect the value ratio among sentences with different labels. In order to determine the richness (R) of a concept/query q_j we simply computed an average value of the sentence within a results list associated with the concept/query:

$$R(q_j) = \frac{1}{n} \sum_{i=1}^{n} val(s_i)$$

Queries with over 550 retrieved sentences are deemed large whereas the rest is considered small. Figure 1 (right) shows that this is where the long tail starts. For richness, we chose 2.0 as a cut-off score. The sentences that fall below are dominated by low value sentences. The sentences that fall

above are quite rich in higher value sentences. This resulted in the four groups, i.e., small sparse (12 queries), small dense (18), large sparse (6), and large dense (6). Each of the six splits then contain 2 SmSp, 3 SmDs, 1 LgSp, and 1 LgDs sentences.

All the systems are then evaluated using 6-fold cross-validation. In each iteration four folds are used as a training set, one as a validation set, and one as a test set. We obtained two scores (NDCG@10 and NDCG@100) for each of the 42 concepts/queries. We report the unweighted means (i.e., the size of the result list is not taken into account) of each score for the four groups determined by the stratified sampling as well as the Overall performance.

For testing statistical significance we employ the strategy suggested by (Demšar, 2006) for testing k methods applied to N datasets. In our experiments, we use the NDCG@100 of the Overall group as the evaluation metric to determine statistical significance. Demšar (2006) recommends the Friedman test (Friedman, 1937), a non-parametric equivalent of the repeated-measures ANOVA. The null-hypothesis states that all the methods (i.e., the assessed ranker and the baselines) are equivalent. In case the null-hypothesis is rejected, we can draw a conclusion that some methods do differ. In order to learn which of them are different, a post-hoc test needs to be conducted. We use the Holm-Bonferroni step down method (Holm, 1979) where the comparisons are performed in sequential order from the most significant hypotheses until a null-hypothesis that cannot be rejected is encountered.

4.2 Baselines

As baselines we report the performance of a Random system on a large sample of repeated runs (for reference) as well as two simple methods based on BM25. The first method is the Okapi BM25 function (Robertson and Zaragoza, 2009) applied to query-sentence pairs. The second BM25-based baseline (BM25-c) is a linear interpolation of BM25 applied to the *query-sentence pair* (s) and to the *whole provision of written law-whole case decision pair* (c) it comes from (context). The two BM25 baselines are very close to what is typically used in many legal IR systems. Furthermore, they are very effective baselines that are often not easy to outperform. For comparison, we also report the performance of the best systems from the prior work. (Savelka et al., 2019; Savelka and Ashley, 2020; Šavelka and Ashley, 2021)

5 Results

The results of the experiments described in Section 4 are reported in Table 1 (group and overall means). The top section of the table presents the performance of the three baselines. The two BM25 baselines clearly outperform the Random system. Despite their similar performance they benefit from completely different phenomena. Intuitively, BM25 ranks high sentences that contain multiple mentions of the concept. In this work the method is optimized in such a way that the documents are not penalized for their length. Hence, the system would often prefer very long sentences. Obviously, such a simple approach works to a certain extent. BM25-c is a combination (linear) of the plain BM25 and another BM25 measure applied to the whole text of a case (i.e., sentence's context). Hence, this system can additionally use the fact of the concept appearing many times within the whole text. This is useful because a decision that mentions the term many times is more likely to contain useful sentences than a decision that mentions it just once. Apparently, the BM25-c is the most competitive of the three baselines.

The middle section of Table 1 shows the performance of the two best models from prior work (Savelka et al., 2019; Savelka, 2020). The BMp+NW+LDA is a linear combination of BM25 applied on a paragraph level, novelty measure, and topic similarity measure. The RF-PWT is the random forest model trained on the 161 hand-crafted features proposed in Savelka (2020); Savelka and Ashley (2020). These models appear to be an improvement over the two baselines.

The performance of the methods based on the pre-trained language models is very promising. Even the performance of the model that considers just the sentence itself (*BERT snt*) and completely ignores the legal concept or the source provision shows promise. The statistical evaluation corroborates the summary statistics reported in Table 1. The strongest conclusion as to outperforming the two BM25 and the Random baseline can be reached for the *BERT sp2snt* model ($p = 0.0002$). While for the *BERT qry2snt* ($p = 0.012$) and *BERT snt* ($p = 0.022$) models the conclusion is not as strong, it still solidly supports the finding (especially considering the relatively limited size of the dataset). The models also appear to improve over the prior

Table 1: The table shows the results of the experiments with pre-trained language models. The NDCG@10 and NDCG@100 are shown for the small sparse queries (SmSp), small dense queries (SmDs), large sparse queries (LgSp), large dense queries (LgDs), and all of them together (Overall).

Method	SmSp		SmDs		LgSp		LgDs		Overall	
	@10	@100	@10	@100	@10	@100	@10	@100	@10	@100
Random	.38 ± .10	.67 ± .15	.52 ± .07	.76 ± .09	.25 ± .16	.29 ± .18	.47 ± .09	.48 ± .09	.43 ± .13	.63 ± .21
BM25	.47 ± .13	.74 ± .11	.60 ± .18	.79 ± .11	.44 ± .21	.37 ± .22	.61 ± .17	.56 ± .12	.54 ± .18	.68 ± .20
BM25-c	.48 ± .12	.76 ± .09	.59 ± .17	.80 ± .11	.49 ± .14	.42 ± .17	.63 ± .19	.55 ± .13	.55 ± .16	.70 ± .18
BMp+NW+LDA	.55 ± .11	.78 ± .12	.64 ± .14	.82 ± .10	.58 ± .16	.56 ± .02	.65 ± .23	.62 ± .11	.61 ± .15	.74 ± .14
RF-PWT	.60 ± .16	.81 ± .11	.66 ± .12	.83 ± .10	.71 ± .17	.68 ± .08	.67 ± .10	.64 ± .09	.65 ± .14	.77 ± .12
BERT snt	.50 ± .18	.71 ± .17	.61 ± .14	.80 ± .11	.46 ± .24	.47 ± .21	.83 ± .15	.77 ± .12	.59 ± .20	**.72 ± .18**
BERT qry2snt	.59 ± .23	.76 ± .19	.72 ± .18	.85 ± .10	.64 ± .34	.50 ± .28	.86 ± .25	.77 ± .18	.69 ± .24	**.77 ± .20**
BERT sp2snt	.57 ± .19	.80 ± .12	.74 ± .15	.87 ± .07	.73 ± .12	.59 ± .18	.89 ± .16	.80 ± .14	.71 ± .19	**.80 ± .14**

state-of-the-art.

6 Discussion

While it should be apparent that the sentence detached from the legal concept (and the provision of law it is embedded in) does not provide reliable grounds for determining its value, it appears that the sentences themselves do carry some signal. This is evidenced by the performance of the *BERT snt* model that only considers sentences themselves. This model outperforms the BM25-based baselines. Interestingly, the system correctly recognized that very short pieces of text that do not form full grammatical sentences typically have very little value. For example, the following sentences have been placed at the bottom of their respective rankings (the explained context is highlighted in yellow):

> Communication & Navigation Equipment
> [No value]

> B. Non-Disclosure of PreExisting Works
> [No value]

Furthermore, it appears that the system relies on features such as the presence of numbering in the sentence, complicated sentence structures, abrupt endings or starts of the sentences, and references, to recognize quotations of written provisions of law and assign a low value to such sentences:

> Derives independent economic value, actual or potential, from not being generally known to the public or to other persons who can obtain economic value from its disclosure or use; and [f] (2)
> [No value]

This strategy makes sense in general. The quotation could either be the citation of the source provision ('no value') or a citation of a different provision (high chance of different meaning and hence lower value of a sentence). However, there are situations in which the strategy does not work well.

> The Electronic Communications Privacy Act of 1986 (ECPA), Pub. L. 99-508, §101(a)(6)(C), 100 Stat. 1848, 1849 (1986), codified, as amended, at 18 U.S.C. §2510(18) (1986), defines "aural transfer" to mean "a transfer containing the human voice at any point between and including the point of origin and the point of reception."
> [High value]

The 'aural transfer' is a rare example of a concept for which there is a legal definition. As a result *BERT snt* underperforms the Random baseline on this particular concept (NDGC@100 0.53 vs 0.62).

BERT snt also seems to have developed a certain tendency to rank high sentences where something is claimed to be something else:

> Screen output is considered an audiovisual work that falls within the subject matter of copyright.
> [High value]

This also appears to be a good strategy that works well many times but not always. For example, the following sentence is just 'potential value' because it uses "navigation equipment" in a different meaning (avionics instead of seafaring):

> Avionics are aircraft radios and navigation equipment.
> [Potential value]

The above examples demonstrate how the pre-trained deep architecture detects very complex features. It would be quite difficult for a human expert to hand-craft such features. While it is not difficult to come up with features such as sentence length, it is far more difficult to come up with features capturing complicated sentence structures, abrupt endings, or subsumption. It is even more difficult to ensure that all the relevant phenomena are considered.

BERT qry2snt models the relationship between the legal concept and the retrieved sentences. It appears to perform better than the base *BERT snt*

model. *BERT qry2snt* has access to the same kind of strategies as *BERT snt*, but since it does not ignore the concept it can go further. For example, there is a clear trend of ranking as very high sentences that contain the concept surrounded by quotation marks:

> The first subsection of that provision, entitled "Navigation Equipment," requires tankers to possess global positioning system ("GPS") receivers, as well as two separate radar systems.
> [High value]

> We believe the common meaning and general understanding of the term "switchblade knife" is a knife in which the blade extends and is securely locked open upon the pressing of a button or other mechanism.
> [High value]

This appears to be a viable strategy. However, there could be instances where it does not work perfectly.

BERT qry2snt appears to have the ability to recognize certain linguistic relationships between the term of interest and other parts of a sentence. The following sentences were not recognized as valuable by *BERT snt* but they are correctly ranked very high by *BERT qry2snt*:

> Airplanes need wings to fly, but that does not mean that all wing designs have independent economic value.
> [High value]

> As explained above, the duty titles in this case do not qualify as identifying particulars.
> [High value]

> And "motion pictures" are "audiovisual works consisting of a series of related images which, when shown in succession, impart an impression of motion, together with accompanying sounds, if any."
> [High value]

All these examples seem to exhibit certain higher level patterns that are intuitively very appealing. Rewriting the above sentences into such patterns could look like this:

> [. . .] NOUN_PHRASE have CONCEPT

> [. . .] qualify [. . .] NOUN_PHRASE [. . .] CONCEPT

> NOUN_PHRASE is defined to be CONCEPT [. . .]

> [. . .] "NOUN_PHRASE" are "CONCEPT [. . .]"

This is corroborated by the inspection of the model weights as applied to several sentences shown in Figure 2. The visualization was created using the tool published with (Vig, 2019). As mentioned earlier BERT is based on the transformer model from (Vaswani et al., 2017). An advantage of using the attention-based model is that it can be interpreted via inspection of the weights assigned to different input elements. As Vig (2019) warns one needs to be very conservative with respect to drawing any conclusions. The three diagrams in Figure 2 show how much attention the first special tokens pay to the individual words in the three input sequences. Note that the input sequences each consist of a term of interest and a retrieved sentence. The reason why the first special token is interesting is that this token stands for the vector representing the sequence which is then fed into a classifier. Hence, the visualization provides some indication of what influences the representation that is being used in the final classification step.

All three examples show that *BERT qry2snt* establishes the relationship between the term of interest (first part of the sequence) and its mention in the sentence. Additionally, the model attends to parts of the sentences that appear to be suggestive about the higher value of a sentence (i.e., "to satisfy the common business purpose requirement", the quotation marks surrounding the digital musical recording, or "that all . . . have independent economic value").

Finally, the *BERT sp2snt* model that focuses on the relationship between a written provision of law and a retrieved sentence appears to perform better than the *BERT qry2snt model*. This may seem somewhat surprising because *BERT sp2snt* does not have access to the focused legal concept. On the other hand, it is provided with the full provision of law. While *BERT sp2snt* appears to lack the ability of *BERT qry2snt* to detect the useful linguistic patterns attached to the legal concepts, it has the ability to recognize the sentences with 'no value' with a high level of accuracy. For example, *BERT qry2snt* ranked the following sentences high:

> In that article, a "wire communication" is defined as "an aural transfer made in whole or in part through the use of facilities for the transmission of communications by the aid of wire, cable, or other like connection between the point of origin and the point of reception." [No value]

> The semiconductor chip product in turn is defined as: the final or intermediate form of any product– [No value]

While these sentences appear to offer valuable definitions of the legal concepts, they merely quote the provision of law, and thus have 'no value.' Overall, it appears that with respect to the NDCG scores, it

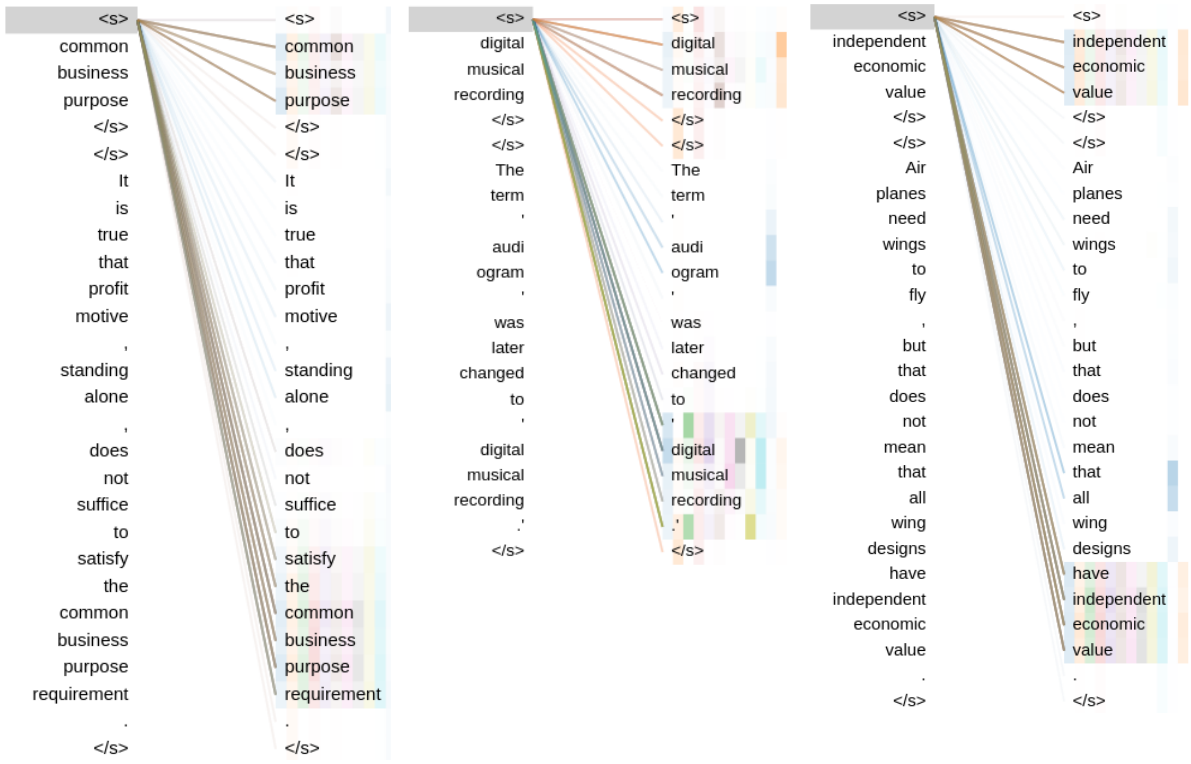

Figure 2: The figure provides some indication of what input elements influence the representation that is being used in the final classification step. The model attends to parts of the sentences that really appear to be suggestive about the higher value of a sentence (i.e., "to satisfy the common business purpose requirement", the quotation marks surrounding the digital musical recording, or "that all . . . have independent economic value").

is extremely important to make sure that sentences such as these do not appear at the top positions of the rankings.

Finally, to provide some concrete examples of the rankings produced by the assessed models Figure 3 shows the distributions of labels of the top 10 retrieved sentences as compared to the overall distribution for two selected concepts ("navigation equipment" and "common business purpose"). The changes in the distributions demonstrate how effective the models can be.

Figure 4 shows box and whisker plots augmented with swarm plots of per query performance for the evaluated systems. Interestingly, it appears that the progression starting from the BM25 method and ending with the RF-PWT (i.e., the prior work referenced above) mostly improves the performance by correcting the disastrous performance of the queries on the left tail of the swarm plots. Despite certain improvements happening at the right side as well, these are dwarfed by the events on the left.

The pre-trained language models fine-tuned on the task of sentence pair classification are interesting because they no longer focus on the improvement of the lowest performing queries only.

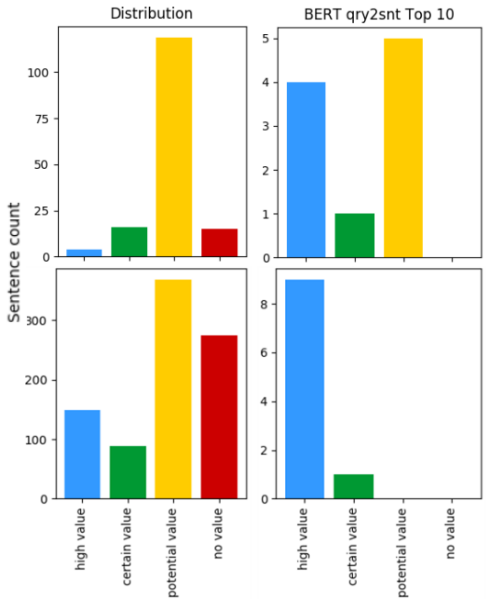

Figure 3: The top two graphs show the sentence value distribution for the concept "navigation equipment." The graph on the left shows the overall distribution while the graph on the right shows the distribution of the top ten sentences retrieved by *BERT qry2snt*. The bottom two graphs show the same for the concept of "common business purpose."

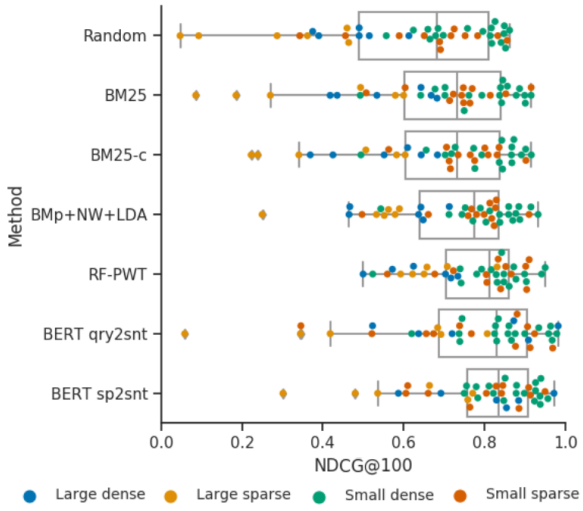

Figure 4: The figure shows scatter plots of the performance on the individual 42 queries measured in terms of NDCG@100 showing the progression from application of simpler similarity methods towards more complex learning-to-rank systems.

They also bring notable improvement to the queries where the performance had already been decent. This is especially true for the BERT qry2snt model that is completely oblivious to the source provision. Hence, this model cannot address the requirement of "providing additional information" as well as the requirement of "using the term of interest in the same meaning." Indeed, it appears that the model has similar issues with a number of queries that the BM25 method had. Yet, despite these notable issues the overall performance is comparable to (if not better than) the RF-PWT.

The BERT sp2snt method uses the source provision instead of the term of interest. On closer inspection one sees three large sparse queries that are not handled well by this method in Figure 4. There are two reasons why a sentence could have 'no value.' It either provides no additional information or it uses the term in a completely different meaning. The three mishandled queries have many sentences that use the term in a different meaning. It appears that BERT sp2snt learned to down-rank the sentences that do not provide additional information quite reliably whereas it completely failed to learn to down-rank the sentences that use the term in a different meaning. The data set may be too small for the method to capture this aspect.

7 Conclusions and Future Work

In this work, we showed that pre-trained language models based on transformers can be fine-tuned for the special task of retrieving sentences for explaining legal concepts. Specifically, we demonstrated that a pre-trained RoBERTa base model, fine-tuned on three variations of the task, resulted in effective ranking functions outperforming the BM25 baselines. The promising performance of *BERT snt* reveals the interesting fact that sentences themselves carry certain signal about their usefulness. The even stronger performance of *BERT qry2snt* and *BERT sp2snt* points to the important interactions among a legal concept, the provision of law in which it is embedded, and retrieved sentences, that both need to be accounted for in order to perform well in this challenging task. The whole work demonstrates the effectiveness of methods based on pre-trained language models applied to a legal domain task. This is important because advances in general NLP and ML do not always transfer in a straightforward manner to specialized domains such as automatic processing of legal or medical texts. Importantly, we fill the gap in prior work by showing that the transformer based methods are capable of fine-grained evaluations of the individual sentences as to their usefulness.

The application of pre-trained language models to the task of discovering sentences explaining legal concepts yielded promising results. At the same time, the work is subject to limitations and leaves much room for improvement. Hence, we suggest several directions for future work:

- Focus on *diversity* in addition to *relevance* to ensure that the top results do not repeat the same sentences.
- Account for all three constituents, i.e., the legal concept, the written provision of law, and retrieved sentences, simultaneously.
- Investigate the effects of including the context of a retrieved sentence, i.e., the full text of a case decision.
- Invest more resources in developing and *extending* the *dataset*.
- Investigate retrieving sentences from other types of legal documents beyond court case decisions (e.g., legislative histories, commentaries).
- Perform an *extrinsic evaluation* of the system in the context of an end-to-end legal project.

Acknowledgements

The first author would like to acknowledge the University of Pittsburgh as his home institution during the time this work was conducted. This work was supported in part by a National Institute of Justice Graduate Student Fellowship (Fellow: Jaromir Savelka) Award # 2016-R2-CX-0010, "Recommendation System for Statutory Interpretation in Cybercrime," and by a University of Pittsburgh Pitt Cyber Accelerator Grant entitled "Annotating Machine Learning Data for Interpreting Cyber-Crime Statutes."

References

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural legal judgment prediction in english. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4317–4323.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert:"preparing the muppets for court'". In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 2898–2904.

Jordan Daci. 2010. Legal principles, legal values and legal norms: are they the same or different? *Academicus International Scientific Journal*, 02:109–115.

Janez Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30.

Timothy Endicott. 2000. *Vagueness in Law*. Oxford University Press.

Timothy Endicott. 2014. Law and Language the stanford encyclopedia of philosophy. http://plato.stanford.edu/. Accessed: 2016-02-03.

Milton Friedman. 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association*, 32(200):675–701.

Evan Gretok, David Langerman, and Wesley M Oliver. 2020. Transformers for classifying fourth amendment elements and factors tests. *Legal Knowledge and Information Systems JURIX*, pages 63–72.

Herbert L. Hart. 1994. *The Concept of Law*, 2nd edition. Clarendon Press.

Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70.

Jerrold Soh Tsin Howe, Lim How Khang, and Ian Ernst Chai. 2019. Legal area classification: A comparative study of text classifiers on singapore supreme court judgments. In *Proceedings of the Natural Legal Language Processing Workshop 2019*, pages 67–77.

Matjaz Juršic, Igor Mozetic, Tomaz Erjavec, and Nada Lavrac. 2010. Lemmagen: Multilingual lemmatisation with induced ripple-down rules. *Journal of Universal Computer Science*, 16(9):1190–1214.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability. *Computing*, 1:25.

Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2020. Pretrained transformers for text ranking: Bert and beyond. *arXiv preprint arXiv:2010.06467*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

D. N. MacCormick and R. S. Summers. 1991. *Interpreting Statutes*. Darmouth.

Samarth Mehrotra and Andrew Yates. 2019. Mpii at trec cast 2019: Incoporating query context into a bert re-ranker.

Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy Lin. 2019. Multi-stage document ranking with bert. *arXiv preprint arXiv:1910.14424*.

The President and Fellows of Harvard University. 2018. Caselaw access project. https://case.law/. Accessed: 2018-12-21.

Juliano Rabelo, Mi-Young Kim, and Randy Goebel. 2019. Combining similarity and transformer methods for case law entailment. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, pages 290–296.

Jinfeng Rao, Linqing Liu, Yi Tay, Wei Yang, Peng Shi, and Jimmy Lin. 2019. Bridging the gap between relevance matching and semantic matching for short text similarity modeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5373–5384.

Stephen Robertson and Hugo Zaragoza. 2009. *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc.

Julien Rossi and Evangelos Kanoulas. 2019. Legal search in case law and statute law. In *Legal Knowledge and Information Systems: JURIX 2019: The Thirty-second Annual Conference*, volume 322, page 83. IOS Press.

Luis Sanchez, Jiyin He, Jarana Manotumruksa, Dyaa Albakour, Miguel Martinez, and Aldo Lipani. 2020. Easing legal news monitoring with learning to rank and bert. In *European Conference on Information Retrieval*, pages 336–343. Springer.

Jaromir Savelka. 2020. *Discovering sentences for argumentation about the meaning of statutory terms*. Ph.D. thesis, University of Pittsburgh.

Jaromír Šavelka and Kevin D Ashley. 2016. Extracting case law sentences for argumentation about the meaning of statutory terms. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 50–59.

Jaromír Savelka and Kevin D Ashley. 2020. Learning to rank sentences for explaining statutory terms. In *ASAIL@ JURIX*.

Jaromír Šavelka and Kevin D Ashley. 2021. Legal information retrieval for understanding statutory terms. *Artificial Intelligence and Law*, pages 1–45.

Jaromir Savelka and Kevin D. Ashley. 2021. On the role of past treatment of terms from written laws in legal reasoning. In *New Developments in Legal Reasoning and Logic*, pages 379–394. Springer.

Jaromir Šavelka and Jakub Harašta. 2015. Open texture in law, legal certainty and logical analysis of natural language. In *Logic in the Theory and Practice of Lawmaking*, pages 159–171. Springer.

Jaromir Savelka, Vern R Walker, Matthias Grabmair, and Kevin D Ashley. 2017. Sentence boundary detection in adjudicatory decisions in the united states. *Traitement automatique des langues*, 58(2):21–45.

Jaromır Šavelka, Hannes Westermann, and Karim Benyekhlef. 2020. Cross-domain generalization and knowledge transfer in transformers trained on legal data.

Jaromir Savelka, Huihui Xu, and Kevin D Ashley. 2019. Improving sentence retrieval from case law for statutory interpretation. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, pages 113–122.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Jesse Vig. 2019. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42.

Hannes Westermann, Jaromir Savelka, and Karim Benyekhlef. 2020. Paragraph similarity scoring and fine-tuned bert for legal information retrieval and entailment. In *JSAI International Symposium on Artificial Intelligence*, pages 269–285. Springer.

Wei Yang, Haotian Zhang, and Jimmy Lin. 2019. Simple applications of bert for ad hoc document retrieval. *arXiv preprint arXiv:1903.10972*.

Zeynep Akkalyoncu Yilmaz, Wei Yang, Haotian Zhang, and Jimmy Lin. 2019. Cross-domain modeling of sentence-level evidence for document retrieval. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3481–3487.