Extracting Case Law Sentences for Argumentation about the Meaning of Statutory Terms

Jaromír Šavelka
University of Pittsburgh
4200 Fifth Avenue
Pittsburgh, PA 15260, USA
jas438@pitt.edu

Kevin D. Ashley
University of Pittsburgh
4200 Fifth Avenue
Pittsburgh, PA 15260, USA
ashley@pitt.edu

Abstract

Legal argumentation often centers on the interpretation and understanding of terminology. Statutory texts are known for a frequent use of vague terms that are difficult to understand. Arguments about the meaning of statutory terms are an inseparable part of applying statutory law to a specific factual context. In this work we investigate the possibility of supporting this type of argumentation by automatic extraction of sentences that deal with the meaning of a term. We focus on case law because court decisions often contain sentences elaborating on the meaning of one or more terms. We show that human annotators can reasonably agree on the usefulness of a sentence for an argument about the meaning (interpretive usefulness) of a specific statutory term (kappa>0.66). We specify a list of features that could be used to predict the interpretive usefulness of a sentence automatically. We work with off-the-shelf classification algorithms to confirm the hypothesis (accuracy>0.69).

1 Introduction

Statutory law is written law enacted by an official legislative body. A single statute is usually concerned with a specific area of regulation. It consists of provisions which express the individual legal rules (e.g., rights, prohibitions, duties).

Understanding statutory provisions is difficult because the abstract rules they express must account for diverse situations, even those not yet encountered. The legislators use vague (Endicott, 2000) open textured (Hart, 1994) terms, abstract standards (Endicott, 2014), principles, and values (Daci, 2010) in order to deal with this uncertainty. When there are doubts about the meaning of the provision they may be removed by interpretation (MacCormick and Summers, 1991). Even a single word may be crucial for the understanding of the provision as applied in a particular context.

Let us consider the example rule: "No vehicles in the park."[1] While it is clear that automobiles or trucks are not allowed in the park it may be unclear if the prohibition extends to bicycles. In order to decide if a bicycle is allowed in the park it is necessary to interpret the term 'vehicle'.

The interpretation involves an investigation of how the term has been referred to, explained, interpreted or applied in the past. This is an important step that enables a user to then construct arguments in support of or against particular interpretations. Searching through a database of statutory law, court decisions, or law review articles one may stumble upon sentences such as these:

i. Any mechanical device used for transportation of people or goods is a *vehicle*.
ii. A golf cart is to be considered a *vehicle*.
iii. To secure a tranquil environment in the park no *vehicles* are allowed.
iv. The park where no *vehicles* are allowed was closed during the last month.
v. The rule states: "No *vehicles* in the park."

Some of the sentences are useful for the interpretation of the term 'vehicle' from the example provision (i. and ii.). Some of them look like they may be useful (iii.) but the rest appears to have very little (iv.) if any (v.) value. Going through the sentences manually is labor intensive. The large number of useless sentences is not the only problem. Perhaps, even more problematic is the large redundancy of the sentences.

[1] The example comes from the classic 1958 Hart-Fuller debate over the interpretation of rules.

Proceedings of the 3rd Workshop on Argument Mining, pages 50–59,
Berlin, Germany, August 7-12, 2016. ©2016 Association for Computational Linguistics

In this paper we investigate if it is possible to retrieve the set of useful sentences automatically. Specifically, we test the hypothesis that by using a set of automatically generated linguistic features about/in the sentence it is possible to evaluate how useful the sentence is for an interpretation of the term from a specific statutory provision.

In Section 2 we describe the new statutory term interpretation corpus that we created for this work. Section 3 describes the tentative set of the features for the evaluation of the sentences' interpretive usefulness. In Section 4 we confirm our hypothesis by presenting and evaluating a rudimentary version of the system (using stock ML algorithms) capable of determining how useful a sentence is for term interpretation.

2 Statutory Term Interpretation Data

Court decisions apply statutory provisions to specific cases. To apply a provision correctly a judge usually needs to clarify the meaning of one or more terms. This makes court decisions an ideal source of sentences that possibly interpret statutory terms. Legislative history and legal commentaries tentatively appear to be promising sources as well. We will investigate the usefulness of these types of documents in future work. Here we focus on sentences from court decisions only.

In order to create the corpus we selected three terms from different provisions of the United States Code, which is the official collection of the federal statutes of the United States.[2] The selected terms were 'independent economic value' from 18 U.S. Code § 1839(3)(B), an 'identifying particular' from 5 U.S. Code § 552a(a)(4), and 'common business purpose' from 29 U.S. Code § 203(r)(1). We specifically selected terms that are vague and come from different areas of regulation. We are aware that the number of terms we work with is low. We did not specify additional terms because the cost of subsequent labeling is high. Three terms are sufficient for the purpose of this paper. For future work we plan to extend the corpus.

For each term we have collected a small set of sentences by extracting all the sentences mentioning the term from the top 20 court decisions retrieved from Court Listener.[3] The focus on the top 20 decisions only reflected the high cost of the labeling. In total we assembled a small corpus of 243 sentences.

	# HV	# CV	# PV	# NV
# HV	**19** (1/4/14)	**1** (0/0/1)	**1** (0/1/0)	**0** (0/0/0)
# CV	**15** (1/6/8)	**12** (2/0/10)	**9** (1/4/4)	**1** (0/1/0)
# PV	**2** (0/0/2)	**27** (11/1/15)	**105** (29/36/40)	**11** (0/3/8)
# NV	**0** (0/0/0)	**0** (0/0/0)	**4** (2/2/0)	**36** (5/13/18)

Table 1: Confusion matrix of the labels assigned by the two annotators (HV: high value, CV: certain value, PV: potential value, NV: no value; the number in bold is the total count and the numbers in the brackets are the counts for the individual terms: ('independent economic value'/'identifying particular'/'common business purpose')).

Two expert annotators, each with a law degree, classified the sentences into four categories according to their usefulness for the interpretation of the corresponding term:

1. **high value** - This category is reserved for sentences the goal of which is to elaborate on the meaning of the term. By definition, these sentences are those the user is looking for.
2. **certain value** - Sentences that provide grounds to draw some (even modest) conclusions about the meaning of the term. Some of these sentences may turn out to be very useful.
3. **potential value** - Sentences that provide additional information beyond what is known from the provision the term comes from. Most of the sentences from this category are not useful.
4. **no value** - This category is used for sentences that do not provide any additional information over what is known from the provision. By definition, these sentences are not useful for the interpretation of the term.

Eventually, we would like the system to assign a sentence with a score from a continuous interval. Since we cannot ask the human annotators to do the same, we discretized the interval into the four categories for the purpose of the evaluation. There was no time limit imposed on the annotation process.

[2] Available at https://www.law.cornell.edu/uscode/text/

[3] Available at https://www.courtlistener.com/. The search query was formulated as the phrase search for the term and it was limited to the 120 federal jurisdictions. The corpus corresponds to the state of Court Listener database on February 16, 2016, which is the last time we updated the corpus.

Term	# HV	# CV	# PV	# NV	# Total
Ind. economic val.	2	5	40	5	52
Identifying part.	6	8	40	17	71
C. business purp.	20	26	51	23	120
Total	28	39	131	45	243

Table 2: Distribution of sentences with respect to their interpretive value (HV: high value, CV: certain value, PV: potential value, NV: no value).

Table 1 shows the confusion matrix of the labels as assigned by the two expert annotators. The average inter-annotator agreement was 0.75 with weighted kappa at 0.66. For the 'independent economic value' the agreement was 0.71 and the kappa 0.51, for the 'identifying particular' 0.75 and 0.67, and for the 'common business purpose' 0.75 and 0.68 respectively. The lower agreement in case of the 'independent economic value' could be explained by the fact that this term was the first the annotators were dealing with. Although, we provided a detailed explanation of the annotation task we did not provide the annotators with an opportunity to practice before they started with the annotation. The practice could be helpful and we plan to use it in future additions to the corpus.

After the annotation was finished the annotators met and discussed the sentences for which their labels differed. In the end they were supposed to agree on consensus labels for all of those sentences. For example, the following sentence from the 'identifying particular' part of the corpus was assigned with different labels:

> Here, the district court found that the duty titles were not numbers, symbols, or other *identifying particulars*.

One of the reviewers opted for the 'certain value' label while the other one picked the 'high value' label. In the end the reviewers agreed that the goal of the sentence is not to elaborate on the meaning of the 'identifying particular' and that it provides grounds to conclude that, e.g., duty titles are not identifying particulars. Therefore, the 'certain value' label is more appropriate.

Table 2 reports counts for the consensus labels. The most frequent label (53.9%) is the 'potential value.' The least frequent (11.5%) is the 'high value' label. The distribution varies slightly for the different terms.

3 Features for Predicting Interpretive Usefulness of Sentences

For testing the hypothesis we came up with a tentative list of features that could be helpful in predicting the interpretive usefulness of a sentence. We reserve the refinement of this list for future work. In addition, many features were generated with very simple models which leaves space for significant improvements. We briefly describe each of the features in the following subsections.

3.1 Source

This category models the relation between the source of the term of interest (i.e., the statutory provision it comes from) and the source of the term as used in the retrieved sentence. To automatically generate this feature we used a legal citation extractor.[4] Each sentence can be assigned with one of the following labels:

1. *Same provision*: This label is predicted if we detect a citation of the provision the term of interest comes from in any of the 10 sentences preceding or following the sentence mentioning the term of interest.
2. *Same section*: We predict this label if we detect a citation of the provision from the same section of the United States Code in the window of 10 sentences around the sentence mentioning the term of interest.
3. *Different section*: This label is predicted if we detect any other citation to the United States Code anywhere in the decision's text.
4. *Different jurisdiction*: We predict this label if we are not able to detect any citation to the United States Code.

The distribution of the labels in this category is summarized in the top left corner of Table 3. We can see that the distribution wildly differs across the terms we work with. For the 'independent economic value' the 'different jurisdiction' (DJR) label is clearly dominant whereas for the 'common business purpose' we predict the 'same provision' (SPR) almost exclusively.

As an example let us consider the following sentence retrieved from one of the decisions:

> The full text of § *1839(3)(B)* is: "[...]". [...] Every firm other than the original

[4] https://github.com/unitedstates/citation

	Source				Semantic Similarity				Structural Placement				
	SPR	SSC	DSC	DJR	SAM	SIM	REL	DIF	STS	CIT	QEX	HD	FT
Ind. economic val.	9	0	0	43	37	1	14	0	9	29	11	0	3
Identifying part.	39	28	0	4	67	0	0	4	29	33	5	0	4
C. business purp.	118	0	0	2	118	2	0	0	65	29	24	2	0
Total	166	28	0	49	224	3	14	4	103	91	40	2	7

	Syntactic Importance			Rhetorical Role								
	DOM	IMP	NOT	STL	APL	APA	STF	INL	EXP	RES	HLD	OTH
Ind. economic val.	5	25	22	23	13	0	3	3	2	7	1	0
Identifying part.	3	21	47	32	7	1	6	9	5	6	5	0
C. business purp.	22	64	34	32	27	1	8	23	14	6	5	4
Total	30	110	103	87	47	2	17	35	21	19	11	4

	Attribution					Assignment/Contrast					Feature		
	JUD	LEG	PTY	WIT	EXP	NA	ASC	TSC	TSA	TNA	NA	AF	TF
Ind. economic val.	20	25	7	0	0	52	0	0	0	0	37	0	15
Identifying part.	36	32	3	0	0	15	49	0	0	7	28	0	43
C. business purp.	87	25	7	0	1	107	8	0	3	2	98	11	11
Total	143	82	17	0	1	177	57	0	3	9	163	11	69

Table 3: The table shows distribution of the features generated for the prediction of sentences' interpretive usefulness.
Source: Same provision (SPR), same section (SSC), different section (DSC), different jurisdiction (DJR).
Semantic similarity: same (SAM), similar (SIM), related (REL), different (DIF).
Structural placement: quoted expression (QEX), citation (CIT), heading (HD), footnote (FT), standard sentence (STS).
Syntactic importance: dominant (DOM), important (IMP), not important (NOT).
Rhetorical role: application of law to factual context (APL), applicability assessment (APA), statement of fact (STF), interpretation of law (INL), statement of law (STL), general explanation or elaboration (EXP), reasoning statement (RES), holding (HLD), other (OTH).
Attribution: legislator (LEG), party to the dispute (PTY), witness (WIT), expert (EXP), judge (JUD).
Assignment/Contrast: another term is a specific case of the term of interest (ASC), the term of interest is a specific case of another term (TSC), the term of interest is the same as another term (TSA), the term of interest is not the same as another term (TNA), no assignment (NA).
Feature assignment: the term of interest is a feature of another term (TF), another term is a feature of the term of interest (AF), no feature assignment (NA).

> equipment manufacturer and RAPCO had to pay dearly to devise, test, and win approval of similar parts; the details unknown to the rivals, and not discoverable with tape measures, had considerable "*independent economic value* ... from not being generally known".

Here we detect the citation to the same provision in the sentence mentioning the term of interest. We predict the 'same source' label.

3.2 Semantic Similarity

This category is auxiliary to the 'source' discussed in the preceding subsection. Here we model the semantic relationship between the term of interest as used in the statutory provision and in the retrieved sentence. Essentially, we ask if the meaning of the terms is the same and if not how much do the meanings differ. We partially model this feature based on the label in the 'source' category as well as on the cosine similarity between the bag-of-words (TFIDF) representations of the source provision and the retrieved sentence. Each sentence can be assigned with one of the following labels:

1. *Same*: We predict this label if the 'same provision' label was predicted in the source category.
2. *Similar*: We predict this label if the cosine similarity is higher than 0.5.
3. *Related*: We predict this label if the cosine similarity is between 0.25 and 0.5.
4. *Different*: We predict this label if the cosine similarity is lower than 0.25.

By definition this feature is useful only in case the 'same provision' label is not predicted in the 'source' category. The distribution of the labels in this category can be seen in the middle component of the top row in Table 3. As we have predicted the 'same' label in most of the cases, this feature did not prove as very helpful in our experiments (see Section 4). We plan to refine the notion of this feature in future work. For example, we would like to use a more sophisticated representation of the term of interest such as word2vec.

The two following examples show sentences that use the same term with different meaning:

> [...] the information derives independent economic value, actual or potential, from not being generally known to, and not being readily ascertainable through proper means by, the *public*;

> [...] posted in the establishment in a prominent position where it can be readily examined by the *public*;

The first sentence mentions the term 'public' for the purpose of the trade secret protection. The term refers to customers, competitors and the general group of experts on a specific topic. The second sentence uses the term to refer to a general 'public.'

3.3 Syntactic Importance

In this category we are interested in how dominant the term is in the retrieved sentence. To model the feature we use syntactic parsing (Chen and Manning, 2014). Specifically, we base our decision on the ratio of the tokens that are deeper in the tree structure (further from the root) than the tokens standing for the term of interest divided by the count of all the tokens. Each sentence can be assigned with one of the following labels:

1. *Dominant*: We predict this label if the ratio is greater than 0.5.
2. *Important*: This label is predicted if the ratio is less than 0.5 but greater than 0.2.
3. *Not important*: We predict this label if the ratio is less than 0.2.

The distribution of the labels in this category is summarized in the left section of the middle row in Table 3. We labeled most sentences as either 'important' or 'not important' (around the same proportion). Only a small number of sentences were labeled with the 'dominant' label.

As an example let us consider the following example sentence with its syntactic tree shown in Figure 1:

> The park where no *vehicles* are allowed was closed during the last month.

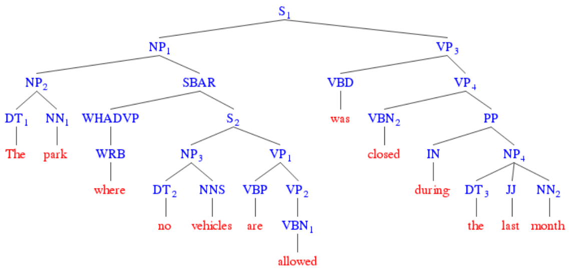

Figure 1:

The syntactic tree contains only one token which is deeper in the structure than the 'vehicle' (the term of interest). Therefore, the ratio is $1/13$ and this sentence is labeled as 'not important.'

3.4 Structural Placement

This category describes the place of the retrieved sentence and the term of interest in the structure of the document it comes from. To model this feature we use simple pattern matching. Each sentence can be assigned with one of the following labels:

1. *Quoted expression*: We predict this label for a sentence that contains the term of interest in a sequence of characters enclosed by double or single quotes if the sequence starts with a lower case letter.
2. *Citation*: This label is predicted if all the conditions for the 'quoted expression' label are met except that the starting character of the sequence is in upper case.
3. *Heading*: This label is predicted if we detect an alphanumeric numbering token at the beginning of the retrieved sentence.
4. *Footnote*: We predict this label for a sentence that starts a line with a digits enclosed in square brackets.
5. *Standard sentence*: We predict this label if none of the patterns for other labels matches.

The distribution of the labels in this category is shown in the top right corner of Table 3. Almost all the sentences were labeled as the 'standard sentence', the 'citation', or the 'quoted expression.'

Only a very small number of sentences was recognized as the 'heading' or the 'footnote.'

Two examples below show a heading and a footnote correctly recognized in the retrieved sentences:

> A. Related Activities and *Common Business Purpose*
>
> [5] [...] However, in view of the '*common business purpose*' requirement of the Act, we think [...]

3.5 Rhetorical Role

In this category we are interested in the rhetorical role that the retrieved sentence has in the document it comes from. Although, some more sophisticated approaches to automatic generation of this feature have been proposed (Saravanan and Ravindran, 2010; Ravindran and others, 2008; Grabmair et al., 2015) we model it as a simple sentence classification task. We used bag of words (TFIDF weights) representation as features and manually assigned labels for training. Each sentence can be assigned with one of the following labels:

1. *Application of law to factual context*
2. *Applicability assessment*
3. *Statement of fact*
4. *Interpretation of law*
5. *Statement of law*
6. *General explanation or elaboration*
7. *Reasoning statement*
8. *Holding*
9. *Other*

The distribution of the labels in this category is shown in the right part of the middle row in Table 3. Most of the sentences were labeled as the 'statement of law,' the 'application of law,' or the 'interpretation of law.'

3.6 Attribution

This category models who has uttered the retrieved sentence. For the purpose of this paper we rely on pattern matching with the assumption that the judge utters the sentence if none of the patterns matches. Each sentence can be assigned with one of the following labels:

1. *Legislator*: We predict this label if we detect a citation to US statutory law followed by a pattern corresponding to citation described in the earlier category.

2. *Party to the Dispute*: We predict this category if we detect a mention of the party (either its name or its role such as plaintiff) followed by one of the specifically prepared list of verbs such as 'contend', 'claim', etc.

3. *Witness*: This label is predicted if we match the word 'witness' followed by one of the verbs from the same set as in case of the preceding label.

4. *Expert*: This label is predicted in the same way as the 'witness' label but instead of the word 'witness' we match 'expert'.

5. *Judge*: We predict this label if none of the patterns for other labels matches.

The distribution of the labels in this category is shown in the bottom left corner of Table 3. We were able to recognize a reasonable number of the 'legislator' labels but apart from that we almost always used the catch-all 'judge' label.

The following example shows a sentence for which we predict the 'party to the dispute' label:

> In support of his contention that Gold Star Chili and Caruso's Ristorante constitute an enterprise, *plaintiff alleges* that Caruso's Ristorante and Gold Star Chili were engaged in the related business activity [...].

3.7 Assignment/Contrast

Here we are interested if the term of interest in the retrieved sentence is said to be (or not to be) some other term. To model this category we use pattern matching on the verb phrase of which the term of interest is part (if there is such a phrase in the sentence). Each sentence can be assigned with one of the following labels:

1. *Another term is a specific case of the term of interest*: This label is predicted if one of the specified set of verbs (e.g., may be, can be) is preceded by a noun and followed by a term of interest within a verb phrase.

2. *The term of interest is a specific case of another term*: In case of this label we proceed in the same way as in case of the preceding label but the noun and the term of interest are swapped.

3. *The term of interest is the same as another term*: In case of this label we use a different set of verbs (e.g., is, equals) and we do not care about the order of the term of interest and the noun.

4. *The term of interest is not the same as another term*: We proceed in the same way as in the case of the preceding label but we also require a negation token to occur (e.g., not).

5. *No assignment*: We predict this label if none of the patterns for other labels matches.

The distribution of the labels in this category is shown in the middle part of the bottom row in Table 3. A certain amount of the 'another term is a specific case of the term of interest' was predicted in the 'identifying particular' part of the data set. For the rest of the dataset the catch-all 'no assignment' label was used in most of the cases.

The following example shows a sentence that we labeled with the 'the term of interest is the same as another term' label:

> The Fifth Circuit has held that the *profit motive* is a *common business purpose* if shared.

3.8 Feature Assignment

In this category we analyze if the term of interest in the retrieved sentence is said to be a feature of another term (or vice versa). We model this category by pattern matching on the verb phrase of which the term of interest is part. Each sentence can be assigned with one of the following labels:

1. *The term of interest is a feature of another term*: This label is predicted if one of the specified set of verbs (e.g., have) is followed by a term of interest within a verb phrase.

2. *Another term is a feature of the term of interest*: This label is predicted if the term of interest precedes one of the verbs.

3. *No feature assignment*: We predict this label if none of the patterns for other labels matches.

The distribution of the labels in this category is shown in the bottom left corner of Table 3. The 'no feature assignment' label was predicted in approximately 2/3 of the cases and the 'term of interest is a feature of another term' in the rest.

Classifier	CV	STD	TEST	STD	SIG
Most frequent	.545	.025	.531	.049	–
Naïve Bayes	.544	.037	.611	.066	no
SVM	.633	.044	.657	.066	no
Random Forest	**.677**	.033	**.696**	.042	yes

Table 4: Mean results from 100 runs of a classification experiment (CV: 10-fold cross validation on the training set, TEST: validation on the test set, SIG: statistical significance)

Features	CV	STD	TEST	STD
-source	**.519**	.05	.586	.046
-semantic relationship	.675	.031	.694	.049
-syntactic importance	.532	.028	**.521**	.047
-structural placement	.695	.033	.708	.047
-rhetorical role	.687	.033	.695	.049
-attribution	.657	.034	.671	.048
-assignment/contrast	.668	.032	.669	.045
-feature assignment	.662	.032	.684	.047

Table 5: Mean results of classification experiment where each line reports the performance when the respective feature was removed.

The following example shows a sentence that we labeled with the 'the term of interest is a feature of another term' label:

> However, Reiser concedes in its brief that the *process* has *independent economic value*.

Here, the independent economic value is said to be an attribute of the process.

4 Predicting Usefulness of Sentences for Interpretation of the Terms of Interest

We work with the dataset described in Section 2. The goal is to classify the sentences into the four categories reflecting their usefulness for the interpretation of the terms of interest. As features we use the categories described in Section 3.

The experiment starts with a random division of the sentences into a training set (2/3) and a test set. The resulting training set consists of 162 sentences while there are 81 sentences in the test set. As classification models we train a Naïve Bayes, an SVM (with linear kernel and L2 regularization), and a Random Forest (with 10 estimators and Gini impurity as a measure of the quality of a split) using the scikit-learn library (Pedregosa et al., 2011). We use a simple classifier always predicting the most frequent label as the baseline.

Because our data set is small and the division into the training and test set influences the performance we repeat the experiment 100 times. We report the mean results of 10-fold cross validation on the training set and evaluation on the test set as well as the standard deviations in Table 4.

All the three classifiers outperform the most frequent class baseline. However, due to the large variance of the results from the 100 runs the improvement is statistically significant ($\alpha = .05$) only for the Random Forest which is the best performing classifier overall. With the accuracy of .696 on the test set the agreement of the Random Forest classifier with the consensus labels is quite close to the inter-annotator agreement between the two human expert annotators (.746).

We also tested which features are the most important for the predictions with the Random Forest. We ran the 100-batches of the experiments leaving out one feature in each batch. The results reported in Table 5 show that the source and the syntactic importance were the most important.

5 Related Work

Because argumentation plays an essential role in law, the extraction of arguments from legal texts has been an active area of research for some time. Mochales and Moens detect arguments consisting of premises and conclusions and, using different techniques, they organize the individual arguments extracted from the decisions of the European Court of Human Rights into an overall structure (Moens et al., 2007; Mochales and Ieven, 2009; Mochales-Palau and Moens, 2007; Mochales and Moens, 2011). In their work on vaccine injury decisions Walker, Ashley, Grabmair and other researchers focus on extraction of evidential reasoning (Walker et al., 2011; Ashley and Walker, 2013; Grabmair et al., 2015). Bruninghaus and Ashley (2005) and Kubosawa et al. (2012) extract case factors that could be used in arguing about an outcome of the case. In addition, argumentation mining has been applied in a study of diverse areas such as parliamentary debates (Hirst et al., 2014) or public participation in rulemaking (Park et al., 2015).

The task we deal with is close to the traditional NLP task of query-focused summarization of multiple documents as described in Gupta (2010). Fisher and Roark (2006) presented a system based on supervised sentence ranking. Daumé and Marcu (2006) tackled the situation in which the retrieved pool of documents is large. Schiffman and McKeown (2007) cast the task into a question answering problem. An extension introducing interactivity was proposed by Lin et al. (2010).

A number of interesting applications deal with similar tasks in different domains. Sauper and Barzilay (2009) proposed an approach to automatic generation of Wikipedia articles. Demner-Fushman and Lin (2006) described an extractive summarization system for clinical QA. Wang et al. (2010) presented a system for recommending relevant information to the users of Internet forums and blogs. Yu et al. (2011) mine important product aspects from online consumer reviews.

6 Discussion and Future Work

The results of the experiments are promising. They confirm the hypothesis even though we used extremely simplistic (sometimes clearly inadequate) approaches to generate the features automatically. We have every reason to expect that improvements in the quality of the feature generation will improve the quality of the interpretive usefulness assessment. We would like to investigate this assumption in future work.

It is also worth mentioning that we used only simple off-the-shelf classification algorithms that we did not tweak or optimize for the task. As in the case of the features, improvements in the algorithms we use would most likely lead to an improvement in the quality of the interpretive usefulness assessment. We plan to focus on this aspect in future work.

The analysis of the importance of the individual features for the success in our task showed that contribution of some of the features was quite limited. We would caution against the conclusion that those features are not useful. It may very well be the case that our simplistic techniques for the automatic generation of those features did not model them adequately. As already mentioned, we plan on improving the means by which the features are generated in future work.

We are well aware of the limitations of the work stemming from the small size of the corpus. This is largely due to the fact that getting the labels is very expensive. Since the nature of this work is exploratory in the sense of showing that the task is (a) interesting and (b) can be automatized, we could not afford a corpus of more adequate size. However, since the results of the experiments are promising we plan to extend the corpus.

This work is meant as the first step towards a fully functional and well described framework supporting argumentation about the meaning of statutory terms. Apart from facilitating easier access to law for lawyers, it is our goal to lower the barrier for public officials and other users who need to work with legal texts. In addition, we believe such a framework could support dialogue between lawyers and experts from other fields. There could be a great impact on legal education as well.

7 Conclusion

We investigated the possibility of automatic extraction of case law sentences that deal with the meaning of statutory terms. We showed that human annotators can reasonably agree on the interpretive usefulness of a sentence for argumentation about the meaning of a specific statutory term. We specified the list of features that could be useful for a prediction of the interpretive usefulness of a sentence. We used stock classification algorithms to confirm the hypothesis that by using a set of automatically generated linguistic features about/in the sentence it is possible to evaluate how useful the sentence is for an argumentation about the meaning of a term from a specific statutory provision.

References

Kevin D Ashley and Vern R Walker. 2013. From informanon retrieval (ir) to argument retrieval (ar) for legal cases: Report on a baseline study.

Stefanie Brüninghaus and Kevin D Ashley. 2005. Generating legal arguments and predictions from case texts. In *Proceedings of the 10th international conference on Artificial intelligence and law*, pages 65–74. ACM.

Danqi Chen and Christopher D Manning. 2014. A fast and accurate dependency parser using neural networks. In *EMNLP*, pages 740–750.

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

Hal Daumé III and Daniel Marcu. 2006. Bayesian query-focused summarization. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 305–312. Association for Computational Linguistics.

Dina Demner-Fushman and Jimmy Lin. 2006. Answer extraction, semantic clustering, and extractive summarization for clinical question answering. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 841–848. Association for Computational Linguistics.

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.

Seeger Fisher and Brian Roark. 2006. Query-focused summarization by supervised sentence ranking and skewed word distributions. In *Proceedings of the Document Understanding Conference, DUC-2006, New York, USA*. Citeseer.

Matthias Grabmair, Kevin D Ashley, Ran Chen, Preethi Sureshkumar, Chen Wang, Eric Nyberg, and Vern R Walker. 2015. Introducing luima: an experiment in legal conceptual retrieval of vaccine injury decisions using a uima type system and tools. In *Proceedings of the 15th International Conference on Artificial Intelligence and Law*, pages 69–78. ACM.

Vishal Gupta and Gurpreet Singh Lehal. 2010. A survey of text summarization extractive techniques. *Journal of Emerging Technologies in Web Intelligence*, 2(3):258–268.

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

Graeme Hirst, Vanessa Wei Feng, Christopher Cochrane, and Nona Naderi. 2014. Argumentation, ideology, and issue framing in parliamentary discourse. In *ArgNLP*.

Shumpei Kubosawa, Youwei Lu, Shogo Okada, and Katsumi Nitta. 2012. Argument analysis. In *Legal Knowledge and Information Systems: JURIX 2012, the Twenty-fifth Annual Conference*, volume 250, page 61. IOS Press.

Jimmy Lin, Nitin Madnani, and Bonnie J Dorr. 2010. Putting the user in the loop: interactive maximal marginal relevance for query-focused summarization. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 305–308. Association for Computational Linguistics.

D. Neil MacCormick and Robert S. Summers. 1991. *Interpreting Statutes*. Darmouth.

Raquel Mochales and Aagje Ieven. 2009. Creating an argumentation corpus: do theories apply to real arguments?: a case study on the legal argumentation of the echr. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, pages 21–30. ACM.

Raquel Mochales and Marie-Francine Moens. 2011. Argumentation mining. *Artificial Intelligence and Law*, 19(1):1–22.

Raquel Mochales-Palau and M Moens. 2007. Study on sentence relations in the automatic detection of argumentation in legal cases. *FRONTIERS IN ARTIFICIAL INTELLIGENCE AND APPLICATIONS*, 165:89.

Marie-Francine Moens, Erik Boiy, Raquel Mochales Palau, and Chris Reed. 2007. Automatic detection of arguments in legal texts. In *Proceedings of the 11th international conference on Artificial intelligence and law*, pages 225–230. ACM.

Joonsuk Park, Cheryl Blake, and Claire Cardie. 2015. Toward machine-assisted participation in erulemaking: an argumentation model of evaluability. In *Proceedings of the 15th International Conference on Artificial Intelligence and Law*, pages 206–210. ACM.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Balaraman Ravindran et al. 2008. Automatic identification of rhetorical roles using conditional random fields for legal document summarization.

M Saravanan and Balaraman Ravindran. 2010. Identification of rhetorical roles for segmentation and summarization of a legal judgment. *Artificial Intelligence and Law*, 18(1):45–76.

Christina Sauper and Regina Barzilay. 2009. Automatically generating wikipedia articles: A structure-aware approach. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 208–216. Association for Computational Linguistics.

Barry Schiffman, Kathleen McKeown, Ralph Grishman, and James Allan. 2007. Question answering using integrated information retrieval and information extraction. In *HLT-NAACL*, pages 532–539.

Vern R Walker, Nathaniel Carie, Courtney C DeWitt, and Eric Lesh. 2011. A framework for the extraction and modeling of fact-finding reasoning from legal decisions: lessons from the vaccine/injury project corpus. *Artificial Intelligence and Law*, 19(4):291–331.

Jia Wang, Qing Li, Yuanzhu Peter Chen, and Zhangxi Lin. 2010. Recommendation in internet forums and blogs. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 257–265. Association for Computational Linguistics.

Jianxing Yu, Zheng-Jun Zha, Meng Wang, and Tat-Seng Chua. 2011. Aspect ranking: Identifying important product aspects from online consumer reviews. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1496–1505, Stroudsburg, PA, USA. Association for Computational Linguistics.