DISCOVERING SENTENCES FOR ARGUMENTATION ABOUT MEANING OF STATUTORY AND REGULATORY TERMS

by

Jaromír Šavelka

JUDr. in Law and Legal Science, Masaryk University, 2013

M.S. in Intelligent Systems, University of Pittsburgh, 2016

Submitted to the Graduate Faculty of

the Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2016

TABLE OF CONTENTS

1.0 INTRODUCTION

In this work I propose to study, design, and evaluate computational methods to support interpretation of statutory and regulatory terms. The result of this work will be a proof of concept prototype system for retrieval of useful example sentences. The assumption is that a carefully selected set of sentences that mention or use a statutory or regulatory term of interest may reveal some fine-grained contours of the term's meaning. As such these sentences could be useful in argumentation about the exact meaning of the terms. Argumentation about the meaning of statutory and regulatory terms is foundational to the interpretation of the terms.

In legal argumentation a lawyer must often defend a specific account of the meaning of one or more terms (i.e., words, phrases). The persuasiveness and validity of a complex argument may hinge on a particular account of the meaning. Argumentation about the meaning of a term may even be the crux of an overall argument. One class of terms that regularly lends itself to interpretation are the terms from statutory and regulatory texts. Statutes and regulations are legally binding sets of rules enacted by legislative or (authorized) executive bodies. Understanding statutes and regulations is difficult because the abstract rules they express must account for diverse situations, even those not yet encountered. The legislators use vague, open textured terms, abstract standards, principles and values in order to achieve generality at the cost of uncertainty. Consider the following (abridged) excerpt from 29 U.S. Code 203:

> "Enterprise" means the *related activities* performed [...] by any person or persons for a *common business purpose* [...]

A lawyer wishing to argue that two restaurants located in different parts of a city owned by a single person do not constitute an enterprise may, e.g., argue that they cannot be considered related activities or that their operation is not performed for a common business purpose. This effectively amounts to defending an account of *common business purpose* where the ownership of the two restaurants could not be subsumed under it (similarly with respect to *related activities*).

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 lawyer to then construct arguments in support of or against particular interpretations. Searching through a database of statutory and regulatory texts, case law, legislative history, or law review articles one may stumble upon sentences such as these:

i. [...] the fact of common ownership of the two businesses clearly is not sufficient to establish a *common business purpose*.
ii. [...] the profit motive is a *common business purpose* if shared.
iii. Were the buildings managed by their owners, the Government would not attempt to link them together as an enterprise bound together by a '*common business purpose*.'
iv. Because the activities of the two businesses are not related and there is no *common business purpose*, the question of common control is not determinative.
v. The defendants weakly challenge the *common business purpose* conclusion [...]

Some of the sentences are useful for the interpretation of the term *common business purpose* 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 if any value (iv. and v.). Going through the sentences manually is labor intensive. A response to a search query may consist of hundreds or thousands of documents. Usually most of the sentences that contain the term of interest would be useless and redundancy would be high. In this work I propose to develop methods to retrieve the set of useful sentences automatically. I am interested in finding sentences that are useful for interpretation of a specific term. In other words, the objective is to discover sentences that could be helpful in arguing about the meaning of the term. In case of the *common business purpose* a lawyer could use sentence i. to argue that the two restaurants cannot be considered an enterprise:

> Since *the common ownership of the two businesses is not sufficient to establish a common business purpose* (sentence i.) it is not possible to conclude that the two restaurants share the common business purpose. Therefore, the two restaurants cannot be considered an enterprise.

An opposing lawyer could use sentence ii. to argue the opposite:

> The two restaurants share the profit. Since *the profit motive is a common business purpose if shared* (sentence ii.) it is possible to conclude that the two restaurants share the common business purpose. Therefore, the two restaurants may be considered an enterprise.

The proposed system should facilitate this type of argumentation by retrieving a specified number of sentences such as these. For example, given the provision defining 'enterprise' and the user's interest in the term 'common business purpose', the system could retrieve the following five sentences:

i. The "*common business purpose*" requirement is not defined in the Act.
ii. The utilization of a common service does not by itself establish a *common business purpose* shared by the owners of separate businesses.
iii. Activities are performed for a *common business purpose* if they are "directed toward the same business objective or to similar objectives in which the group has an interest."
iv. In a situation such as this, in which the Court has concluded that there are no related activities, the fact of common ownership of the two businesses clearly is not sufficient to establish a *common business purpose*.
v. The Fifth Circuit has held that the profit motive is a *common business purpose* if shared.

Or consider the following provision where the user might be interested in the meaning of the term 'treatment' (45 CFR 164.502(a)(1)):

Covered entities: Permitted uses and disclosures.
A covered entity is permitted to use or disclose protected health information as follows:
(i) To the individual;
(ii) For *treatment*, payment, or health care operations, as permitted by and in compliance with §164.506; [...]

The system could respond with the following five sentences:

i. Activities involving patient care information, such as peer review, quality assurance, mortality and morbidity studies and medical education do not involve patient *treatment* directly and, therefore, will require that a minimum necessary determination be made for each use and disclosure of protected health information involved in those complicated processes.
ii. The definition of "*treatment*" for example would include cost containment mechanisms such as case and disease management that go to managing the costs of populations, rather than the health care of an individual.
iii. For example, while the regulations allow for drug formulary management as part of "health care operations," the definitions of "marketing," "*treatment*", and "health care operations" overlap in many places and are unclear.
iv. HIPAA also defines "*treatment*," as follows: *Treatment* means the provision, coordination, or management of health care and related services by one or more health care providers, including the coordination or management of health care by a health care provider with a third party; consultation between health care providers relating to a patient; or the referral of a patient for health care from one health care provider to another.
v. These cases are inapposite if a reasonable person would expect the giving of medical *treatment* to include prescriptions for medication.

The proposed work is based on three hypotheses that I would like to test. First, a sentence may be objectively evaluated in terms of its utility for argumentation about the meaning of a statutory or regulatory term of interest. Second, it is possible to perform the evaluation automatically. Third, the information about the usefulness of sentences allows one to select better example sentences than if one relies solely on existing state-of-the-art IR and query-focused summarization techniques.

Chapter 2 provides the introductory background. It explains why the discovering of sentences for argumentation about the meaning of statutory and regulatory terms is an interesting problem,

why it is important to tackle it, and what benefits could deeper understanding of the problem and its potential solution offer. In Chapter 3 the utility of a sentence for argumentation about meaning of terms is explained and defined. Results of preliminary experiments with human experts labeling sentences in terms of the utility measure are presented as well. Chapter 4 provides an overview of the proposed framework for automatic selection of the most useful sentences as well as results of preliminary experiments from implementation of some of the parts of the framework. The hypotheses are laid down in Chapter 5 while the experimental plan for their testing is described in Chapter 6. Finally, related work (Chapter 7), contributions of this work (Chapter 8), and the expected timeline are presented.

2.0 BACKGROUND

The purpose of this chapter is to set the proposed work into context. The aim is to explain to readers of different backgrounds why the discovering of sentences for argumentation about the meaning of statutory and regulatory terms is an interesting problem, why it is important to tackle it, and what benefits could deeper understanding of the problem and its potential solution offer. Statutes and regulations could be understood as sets of legally binding rules. The rules usually are abstract statements about what type of conduct is forbidden, required or permitted by law (see Section 2.1). Statutes and regulations are written in natural language. It has been recognized that natural language communications are inherently vague (see Section 2.2). The rules are applied to specific circumstances of a situation to determine if there is a breach of law. The application of statutes and regulations is often contingent on dealing with vagueness (see Section 2.3). The goal of legal regulation is to establish certain expectations as recognized and protected by law (i.e., backed by a state). Therefore, it is important that the application is done in a consistent and predictable manner (see Section 2.4). This in turn is why past treatment of specific words, phrases, or whole sentences (terms for short) needs to be taken into account in the application of law to a specific case (see Section 2.5). The required investigation is often labor intensive 2.6. Thus, its automation would be very valuable. The aim of this work is to perform the investigation automatically by means of a system that selects a small number of sentences that mention or use a term of interest in a particularly helpful way.

2.1 STATUTES AND REGULATIONS

Statutes and regulations are legally binding sets of rules enacted by legislative bodies (statutes) or authorized executive/administrative bodies (regulations). They consist of legal norms, i.e., legally

binding rules of conduct. A single statute or regulation is usually concerned with a specific area. It consists of provisions which express the individual legal rules (e.g., rights, prohibitions, duties). An example provision may look like this (15 U.S. Code § 1644(f)):

> Whoever in a transaction affecting interstate or foreign commerce furnishes money, property, services, or anything else of value, which within any one-year period has a value aggregating $1,000 or more, through the use of any counterfeit, fictitious, altered, forged, lost, stolen, or fraudulently obtained credit card knowing the same to be counterfeit, fictitious, altered, forged, lost, stolen, or fraudulently obtained [...] shall be fined not more than $10,000 or imprisoned not more than ten years, or both.

The contents of statutes and regulations do not need to be rules exclusively. In exceptional cases (e.g., in preambles of statutes) it is possible to include declarations of certain goals or values. An example is the preamble to the Constitution of the United States:

> We the people of the United States, in order to form a more perfect union, establish justice, insure domestic tranquility, provide for the common defense, promote the general welfare, and secure the blessings of liberty to ourselves and our posterity, do ordain and establish this Constitution for the United States of America.

Definitions such as the following one are quite common as well (29 U.S. Code § 203):

> "Enterprise" means the related activities performed (either through unified operation or common control) by any person or persons for a common business purpose, and includes all such activities whether performed in one or more establishments or by one or more corporate or other organizational units including departments of an establishment operated through leasing arrangements, but shall not include the related activities performed for such enterprise by an independent contractor. [...]

Provisions of law are difficult to understand because the rules they express must account for diverse situations, even those not yet encountered. This means the rules need to be abstract and general. In the words of Herbert L. A. Hart, provisions of law need to communicate general standards and refer to classes of persons, and to classes of acts, things, and circumstances. [20, p. 124] In order to achieve the required generality legislators use vague [13] open textured [20] terms, abstract standards [14], principles, and values [9]. This comes at the cost of increase in vagueness. The successful use of the rules depends on a capacity to recognize particular acts, things, and circumstances as instances of the general classifications which the law makes. [20, pp. 124–126] In other words, in order to use the rules successfully it is necessary to map the general norms onto specific factual circumstances. This may often prove to be a considerable challenge.

2.2 VAGUENESS IN NATURAL LANGUAGE

Natural language terms are inherently vague. This applies even to common terms such as "red," "bald," and "young." [12, p. 1] Claiming that a term is vague usually amounts to ascribing it with three related features:

1. the existence of borderline cases,
2. the lack of a sharp boundary along the transition from clear cases to clear counter-instances,
3. susceptibility to sorites arguments (see below).

Considering the example of the term "young" almost everyone could agree that a 16-year-old man is "young" while a 90-year-old man is not. It is much less clear if a 30, 40, or 50-year-old man could be considered "young." Instances that are unclear with respect to their membership in a category are the *borderline cases*. Assuming a 16-year-old man is "young" while a 90-year-old man is not, it would be interesting to identify a boundary age separating instances that belong to the category from those that do not. Although it is clear that the boundary is an age greater than 16 and less than 90 ($16 < boundary < 90$), it turns out that it is not really possible to select a single age as the *boundary*. If almost everyone can agree that a man of a certain age is "young" then almost everyone should also agree that a man who is one year older is "young" as well. Although, this seems reasonable a repeated application of this logic could lead to an obviously absurd conclusions. For example, starting with a 16-year-old man as "young" it is possible to conclude that a 90-year-old man is "young" too. This phenomenon is called *sorites paradox*. [12, pp. 1–2]

Vagueness is not restricted to adjectives. It is found in many other lexical categories for which some notion of grading can be relevant. Nouns such as "heap," or even "chair" or "apple" can be vague. Verbs such as "run" or "walk" could be vague too. We can even consider determiners such as "many," "few," "much," or "little" as well as adverbs (e.g., "quickly," "surprisingly," "clearly") and modifiers (e.g., "very," "somewhat," "completely"). [12, pp. 3]

There are numerous properties that are similar or related to vagueness. Some examples are *imprecision*, *underdetermination*, *openness of meaning*, *contextual variability*, *inexactness*, *overdetermination*, and *overlap* or *ambivalence* between categories. [12, pp. 7–8] Vagueness is often contrasted to *ambiguity* and *generality*. [46] The fine-grained distinctions are not crucial for this work. The main point is that in most of the cases a communication in natural language does not have a precise meaning everyone could agree on. This does not appear to be a flaw in communication

but it rather appears to be its feature.

Vagueness does not seem to cause major problems in everyday communication. Consider the following example utterance:

> It is cold outside. Wear warm clothes.

This is a perfectly meaningful utterance. Any reasonable person could understand it and, moreover, a person could act upon it (e.g., by wearing a sweater). One could object that the utterance may in fact be precise given the context in which it is uttered (e.g., the specific weather conditions, the available clothes, the clothing habits of the person to whom the utterance is directed). This is certainly true and in many circumstances the utterance may be as precise as pointing to a specific piece of clothing. For example, there might be a mutual understanding between the persons involved in the discussion that the utterance requests a specific sweater to be worn. However, this merely means that the utterance is specific enough that it conveys a clear message that can be acted upon. By no means it would be possible to claim that the utterance is free of vague terms. Consider the term "cold"—assuming it is 38°F outside we can easily show the term manifests the three features described above. While almost everyone would agree that 90°F is not cold, 45°F or 50°F are examples of the boundary cases. The lack of a sharp boundary as well as the susceptibility to the sorites argument are quite clear too. The same applies to other terms such as "warm," "clothes," or even "outside." The main point is that a perfectly valid, useful, and clear communication could be achieved with terms that are vague.

2.3 APPLICATION OF LAW AND VAGUENESS

The use of the rules from statutes and regulations subsists in their application to specific factual circumstances. This is often much less straightforward than it would seem at the first sight. One of the chief contributors to the difficulty is the inherent vagueness of natural language. When the application of a general rule is not straightforward a lawyer must present arguments as to why a provision should be applied in a particular way. In doing so the lawyer must often defend a specific account of the meaning of one or more terms. The persuasiveness and validity of a complex argument may hinge on a particular account of the meaning. Argumentation about the meaning of a term may even be the crux of an overall argument.

It is interesting to point out the difference in the level of scrutiny an everyday communication and a communication of general legal rules are usually subjected to. Considering the example utterance from the previous section (Section 2.2) it would be bizarre to respond with a series of utterances like these:

> What is meant by being "cold?"
> What is meant by being "cold outside?"
> Do shoes qualify as "clothes?"
> Does a person "wear" a sweater if it is fasten to his waist?

Yet, this is exactly the kind of scrutiny general legal rules undergo regularly in the course of being applied to specific factual circumstances. As an example consider the following rule posted at the entrance to a park:

> No vehicles in the park.[1]

Focusing on the term "vehicle" there could be little doubt as to whether a car, a bus, or a motorcycle are "vehicles." It is much more challenging to decide about in-line skates, or a bike. This means that there are objects which are clearly prohibited from entering the park and no doubt with respect to this can be entertained by any reasonable person. But there are also objects in case of which it cannot be easily, if at all, determined (borderline cases). Interestingly, it should not strike anyone as bizarre, when thinking about the rule, to ask questions like these:

> What is meant by "vehicle?"
> What is meant by "the park?"
> What does it mean to be "in" the park?

Unlike most of the terms in everyday communications the terms from general legal rules, embodied in statutes and regulations, are expected to be examined on their exact meaning.

In this work it is assumed that terms used in legal rules could often be vague as described above. This has serious implications for the examination of the meaning of terms. Although, considerations of these issues are far beyond the scope of this work, it makes sense to provide an introductory overview here. A theory of judicial decision making which is sometimes called the 'standard view of adjudication' presents the view that the judge's task is to simply give effect to the legal rights and duties of the parties. [13, p. 1] Such a view does not allow for vagueness as described above because it assumes the meaning of terms is determined—perhaps it is not immediately obvious but it could be discovered through careful analysis. The view is usually challenged with the well-known

[1] The example is an adaptation of the rule from the classic 1958 Hart-Fuller debate over the interpretation of rules.

"indeterminacy claim." The claim states that the requirements of the law in particular cases are frequently indeterminate. [13, pp. 1–2] This view allows for vague terms.

Consider an example of a cyclist entering a park who accidentally rides into a jogger. As a result both persons are injured. Although it is not clear who caused the situation, the jogger claims that bikes are not allowed in the park. Because the cyclist was in wrong the jogger asks him to pay all the medical expenses. The cyclist does not agree that bikes are forbidden from entering the park and believes that each should pay for his own expenses. If the jogger's claim is brought in front of a court of justice the investigations into the meaning of the term 'vehicle' from the example rule would play the central role. If such an interpretation where a bike is to be considered 'vehicle' prevails the jogger wins and the cyclist should pay for the expenses. The argument could run as follows:

Legal rule prohibits vehicles in the park.
A bike is to be considered a vehicle.
=>
The cyclist broke the rule by entering the park.
A person breaking a legal rule is responsible for harm occurring as a consequence of the violation.
=>
The cyclist is responsible for the harm occurring as a consequence of him entering the park.
The harm to the jogger occurred in a direct consequence of the cyclist entering the park.
=>
The cyclist is responsible for the harm caused to the jogger.

An interpretation excluding bikes from the 'vehicle' category would have an opposite effect. Swapping the italicized part of the argument with its negation would not allow the above chain of reasoning to be instantiated. The final conclusion of the cyclist's liability would be impossible to reach. Therefore, the result of the case depends on the prevailing account of the meaning of the term 'vehicle'.

In the presented example the judge could decide either way depending on whether he concludes a bike is to be considered 'vehicle' or not for the purpose of the rule forbidding vehicles from entering the park. Such a discretion implies considerable uncertainty concerning legal rules. In Harts opinion this uncertainty is the price to be paid for the use of general classifying terms in any form of communication concerning matters of fact. And it is indeed the price that must be paid because human legislators cannot have knowledge of all the possible combinations of circumstances which the future may bring. [20, pp. 126–128]

2.4 FUNCTION OF LAW AND CONSISTENT APPLICATION OF LAW

The cases in which it is not immediately clear if a specific legal norm should apply and what exact effects should it have are common. This is not a flaw in legal regulation but its inherent feature. [20, pp. 124–135] Niklas Luhmann claims that the main function of law subsists in securing certain expectations of individuals as stable over time. [26, pp. 147–148] This relates the main function of law to certainty concerning legal rules. Indeed, legal certainty is an important value that has been traditionally recognized as crucial to the rule of law. Therefore any threat to legal certainty should be taken very seriously. In the previous section (Section 2.3) it was established that vagueness and indeterminacy give rise to considerable uncertainty concerning legal rules. It is of key importance to reflect on proper approaches to tackle the phenomenon. This essentially means finding approaches that facilitate communication of general standards and reference to classes of persons, to classes of acts, and circumstances, [20] in a way that is understandable, transparent and open to rigorous assessment that leads to persuasive interpretations of that communication. [40]

Future is uncertain. But a person wants to be certain about the future because operations in society take time. [26, pp. 143 and 146] For example, there may be an individual who could immediately use a fixed sum of money to generate a profit of 10% of the sum. There may be another individual who has the sum. It would be desirable if the second individual (creditor) could temporarily transfer the money to the first individual (debtor) and then both of them could share the profit. This could be done only if there is a guarantee that the debtor is going to return the money (with the agreed share of profit) to the creditor. It is the main goal of law to provide such a guarantee. [40]

Legal norms can be understood as structure of symbolically generalized expectations. By stabilized usage of this symbolization society produces specific stabilities and specific sensibilities. [26, pp. 142–146] In case of the above described example the creditor can temporarily transfer the possession of the money to the debtor. It is the case because he can reliably expect that the society will acknowledge his entitlement to get the money back with the agreed upon share of profit. If necessary the society will help him in enforcing the legitimate claim. [40] Law makes it possible to know which expectations will meet with social approval. Given this certainty of expectations one can take on the disappointments of everyday life with a higher degree of composure. This means that one can live in a more complex society. [26, pp. 147–148]

Consider an example where the creditor lends the money to the debtor. They both agree that

after a fixed period of time the money will be returned together with the half of the profit. When the time comes the debtor refuses to provide both, the money he borrowed as well as the profit he promised to deliver. Since the entitlement of the creditor to receive the money and the profit is acknowledged by the society (enforceable by law) he can turn to the society for help. This would usually mean that he can file a claim with a court of justice. [40]

How the court addresses the claim is of vital importance with respect to legal certainty. Brian Bix offers an interesting example of a judge deciding cases on the basis of a coin-flip. [5, p. 106] One could imagine how much trust in law would be generated if the court dismisses the creditor's claim on the basis of a coin flip. Luhmann claims that where law is no longer respected, or is no longer enforced as far as it is possible so to do, the consequences extend much further than what amounts to breach of law. The system has to retreat to much more basic forms of securing confidence. [26, p. 148] If law fails to provide members of the society with sufficient amount of legal certainty, i.e., fails to persuade them that their legitimate expectations will be acknowledged, it fails to perform its function altogether. [40]

The example with a coin-flipping is extreme. Such procedure would be immediately recognized as unacceptable. However, there can be more subtle forms of coin-flipping that are more difficult to recognize. One of them is closely connected with vagueness in law. If vagueness is misunderstood in a way that in certain cases it gives a judge total freedom to decide a case that appears to be unclear, a kind of coin-flipping is being introduced in law. As has been argued above, this may have far reaching consequences outgrowing breaches of law and dismissal of legitimate claims in individual cases. [40]

2.5 PAST TREATMENT OF TERMS AND ITS ROLE IN ARGUMENTATION ABOUT THEIR MEANING

In Section 2.3 I have concluded that quite often there is room for competing interpretations of statutory and regulatory terms. This in turn provides a space for competing (even conflicting) solutions to specific legal issues. Recall an example in which there was a claim against the cyclist causing a harm to a jogger in the park. Depending on the interpretation of the term 'vehicle' from the example legal rule it was possible to conclude that the cyclist should either pay for the jogger's medical expenses or that he should not pay for the expenses. Although conflicting, both

conclusions seem possible. In Section 2.4 I have warned against understanding the existence of a room for competing interpretations as a blank permission to pick whichever understanding of the term one might prefer. In this section I would like to explain how do past mentions and uses of terms function as constraints on the freedom to choose an interpretation.

Consider a scenario in which there are two different cases involving an a cyclist in the park. In a single legal system cases must be considered in the light of the similar cases that were decided before them. Assuming the cases come in sequence there are four different possible outcomes:

1. First case is decided in such a way that a bike is considered to be a vehicle while the second one in such a way that a bike is not considered to be a vehicle.
2. First case is decided in such a way that a bike is not considered to be a vehicle while the second one in such a way that a bike is considered to be a vehicle.
3. Both cases are decided in such a way that a bike is considered to be a vehicle.
4. Both cases are decided in such a way that a bike is not considered to be a vehicle.

It appears that the outcomes 1 and 2 are less desirable than the outcomes 3 and 4. It is the case because in light of the outcomes 1 or 2 it would be very difficult (if at all possible) for anyone to hold any legitimate expectations related to the presence of bikes in the park.

Suppose that the rule was enacted after repeated complaints by the public. The merit of these complaints was that vehicles in the park cause a lot of noise as well as many other externalities having negative impact on the possibility to spend a pleasant time in the park. Indeed, a written document related to the enactment of the rule states, among other things, that:

> The goal of the rule is to secure serenity in the park.

Also suppose that in the first case the cyclist was behaving very noisily to the point where he was certainly disturbing serenity in the park. Whereas in the second case the cyclist was behaving orderly. Given the additional information the outcome 2 clearly appears to be the least desirable. In this case the orderly behaving cyclist would be considered to fall under the 'vehicle' category for the purpose of the rule promoting serenity in the park. At the same time the disturbing cyclist would not be considered 'vehicle' for the purpose of the rule. Interestingly, the outcome 4 appears less desirable than before. There clearly is an object that could with a bit of a stretch be regarded as 'vehicle' and this object interferes with serenity in the park. Yet, the court concludes that the object is not forbidden from entering the park. An argument for why the outcome 3 is less desirable than before could be made in a similar fashion. Probably, the most interesting is that the outcome

1 appears much more desirable than before. At least there is an explanation for treating the first and the second cyclist differently. However, this is not to say that the outcome 1 would be the most desirable out of the four. While the outcome 2 is clearly the least preferable one, it is not possible to decide among the other three outcomes on the basis of the available information.

Difficulties in formation of legitimate expectations have very damaging effect within the legal system (Section 2.4). Therefore the outcome that best promotes the formation of legitimate expectations should be preferred. For the purpose of this work I propose to understand the problem of argumentation about the meaning of a term as a problem similar to hypothesis testing. Specifically, the particular account of the meaning of a term is the hypothesis to be tested against the available evidence. The evidence consists of sentences that mention or use the term of interest. These sentences may come from many different sources—typically they come from past court decisions, journal articles, conference papers, or legislative histories.

In the simple example of the two cyclists the situation when the first case appears in front of a court of justice is the following. The outcome of the case depends on the meaning of the term 'vehicle' from the example rule:

> No vehicles in the park.

The available evidence consists of the sentence describing the goal of the rule:

> The goal of the rule is to secure serenity in the park.

In such a setup I claim that the decision could go either way. However, a line of reasoning that aligns the hypothesis with the evidence is preferable. Consider the two sets of arguments in case where the cyclist was behaving noisily. First, see the arguments supporting the hypothesis that a bike should not be considered 'vehicle.'

1. In ordinary language it appears to be a very long stretch to subsume a bike under the umbrella term 'vehicle.' Therefore, we conclude that the cyclist skater did not violate the rule by entering the park.
2. In ordinary language it appears to be a very long stretch to subsume a bike under the umbrella term 'vehicle.' *Despite disturbance of serenity in the park in this particular case we would like to point out that in most cases cyclists are not noisy. The presence of cyclists does not interfere with the goal of the rule.* Therefore, we conclude that the cyclist did not violate the rule by entering the park.

Second, consider the arguments supporting the hypothesis that a bike should be considered 'vehicle:'

1. In ordinary language it seems possible to subsume a bike under the umbrella term 'vehicle.' Therefore, we conclude that the cyclist did violate the rule by entering the park.

2. In ordinary language it seems possible to subsume a bike under the umbrella term 'vehicle.' *In addition, the cyclist was disturbing serenity in the park. Since, the goal of the rule is to prevent such a disturbance*, we conclude that the cyclist did violate the rule by entering the park.

Setting aside a preference for a particular conclusion, I think it is reasonable to assume that most of the people would agree that the second argument in each set is superior to the first argument. In the first set, the first argument does not go all the way to promote the formation of legitimate expectations. A citizen could be puzzled that there is an entity that could be considered 'vehicle' and that is disturbing the serenity, while not being forbidden from entering the park. The second argument does not have this weakness. In the second set, the second argument offers an extra reason why should a bike be considered 'vehicle.' In comparison, the first argument looks a little bit arbitrary.

The situation when the second case appears in front of a court of justice is much more interesting. Lets assume that the first case was decided in such a way that a bike is to be considered 'vehicle.' The available evidence now consists of two sentences:

The goal of the rule is to secure serenity in the park.
In ordinary language it seems possible to subsume a bike under the umbrella term 'vehicle.'

In light of this evidence it is much more difficult to defend the hypothesis that a bike should not be considered 'vehicle:'

1. In ordinary language it appears to be a very long stretch to subsume a bike under the umbrella term 'vehicle.' Therefore, we conclude that the cyclist did not violate the rule by entering the park.
2. In ordinary language it appears to be a very long stretch to subsume a bike under the umbrella term 'vehicle.' *We would like to point out that in most cases cyclists are not noisy. That is precisely the state of affairs in this particular case. Since the presence of cyclists does not interfere with the goal of the rule,* we conclude that the cyclist did not violate the rule by entering the park.
3. *In the previous case the court stated that: In ordinary language it seems possible to subsume a bike under the umbrella term 'vehicle.' We would like to argue* that in ordinary language it appears to be a very long stretch to subsume a bike under the umbrella term 'vehicle.' *We are under the impression that in that case the court was carried away by the fact that the particular cyclist was a cause of disturbance in the park.* We would like to point out that in most cases cyclists are not noisy. That is precisely the state of affairs in this particular case. Since the presence of cyclists does not interfere with the goal of the rule, we conclude that the cyclist did not violate the rule by entering the park.

Although, not impossible it could be clearly seen that it has become more difficult to defend the hypothesis that a bike should not be considered 'vehicle.' It is the case because there is a contradicting evidence that needs to be dealt with. Thus, the second argument that appeared

sufficient in the previous case is clearly inadequate in the present case. It was necessary to come up with the third argument which could be adequate. Although, the sequence of events where at one point a bike is 'vehicle' and at other moment a bike is not 'vehicle' is certainly not very conducive to the formation of related legitimate expectations.

Lets consider the arguments supporting the hypothesis that a bike should be considered 'vehicle:'

1. In ordinary language it seems possible to subsume a bike under the umbrella term 'vehicle.' Therefore, we conclude that the cyclist did violate the rule by entering the park.
2. In ordinary language it seems possible to subsume a bike under the umbrella term 'vehicle.' *In addition, in the previous case the cyclist was disturbing serenity in the park. Since, the goal of the rule is to prevent such a disturbance, we agree with the past ruling of the court and* conclude that the cyclist did violate the rule by entering the park.

Even the first argument, which had to be considered as weak in the first case, could be considered as adequate in light of the new evidence. The reason is that after the first case happened the expectation should be that a bike is 'vehicle.' Since the first argument conforms to the expectation it appears to be sufficient. Although, one could argue that in the present case the cyclist is not disturbing serenity in the park. Thus, it is rather confusing that there is an entity which is not causing any disturbance and yet it is forbidden to enter the park. In my opinion this is just a minor issue. It could be fixed as shown in the second argument.

Dealing with legal issues that hinge on a particular account of a meaning of a statutory or regulatory term (or more terms) is common. A thorough analysis of the past treatment of the term of interest is contingent for formation of an adequate argument leading to a solution of the issue. It is of crucial importance that the proposed (hypothesized) account of the meaning of the term can withstand a scrutiny of the available evidence consisting of past mentions and uses of the term in sentences from documents such as court decisions, legislative histories, or journal articles. Adoption of such accounts of meaning that do not play well with the available evidence undermines the formation of related legitimate expectations.

2.6 ANALYSIS OF THE PAST TREATMENT OF TERMS

A lawyer that needs to argue about a meaning of a term will most likely use some of the available commercial legal IR systems to analyze the past treatment of the term. The most likely source of useful sentences are court decisions. Thus, a lawyer would most likely search through the data

base of court decisions and inspect mentions and uses of a term of interest. Although, this strategy works it is labor intensive and often clearly not very effective. A lawyer may need to go through many decisions before putting together a reasonable set of useful sentences or before concluding that such a set most likely does not exist.

Two existing tools deserve our special attention—Westlaw's Words and Phrases module[2] and LexisNexis' Legal Issue Trail service.[3] These two services are the state of the art technology supporting legal interpretation. Words and Phrases lists sentences from court decisions if they contain the word or the phrase entered by a user and at the same time the verb in the sentence is "be" in any form. The Legal Issue Trail service is concerned with the interpretation of court decisions. By offering the user other decisions cited by the decision of interest and citing the decision the service attempts to provide users with additional information.

[2] https://lawschool.westlaw.com/marketing/display/RE/217

[3] http://www.lexisnexis.com/legalnewsroom/lexis-hub/b/lexis-advance/archive/2011/11/30/lexis-advance-legal-issue-trail.aspx

3.0 SENTENCE UTILITY FOR INTERPRETATION OF TERMS

In the preceding chapter I have proposed to understand argumentation about the meaning of a statutory or regulatory term as hypothesis testing. In this understanding sentences that mention or use the term of interest are evidence against which the hypothesis is tested (see Chapter 2). It should be self-evident that not all of the sentences are created equal, i.e., some sentences are more useful for the proposed kind of hypothesis testing than others. Consider the following (abridged) excerpt from 29 U.S. Code 203:

> "Enterprise" means the *related activities* performed [...] by any person or persons for a *common business purpose* [...]

Searching through a database of statutes, regulations, court decisions, legislative histories, or law review articles one may stumble upon sentences such as these:

i. [...] the fact of common ownership of the two businesses clearly is not sufficient to establish a *common business purpose*.
ii. [...] the profit motive is a *common business purpose* if shared.
iii. Were the buildings managed by their owners, the Government would not attempt to link them together as an enterprise bound together by a '*common business purpose*.'
iv. Because the activities of the two businesses are not related and there is no *common business purpose*, the question of common control is not determinative.
v. The defendants weakly challenge the *common business purpose* conclusion [...]

Some of the sentences are useful for the interpretation of the term *common business purpose* 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 if any value (iv. and v.). As explained in Section 2.6 going through the sentences manually is labor intensive. In this work I investigate if it is possible to retrieve the set of useful sentences automatically. Specifically, I test the hypotheses laid down in Chapter 5.

I propose the general interpretive utility u_g that measures how useful a sentence is for argumentation about a meaning of a statutory or regulatory term. The measure u_g is defined by the general interpretive utility function $\mathcal{U}_g$ as follows:

$$\mathcal{U}_{g,\theta}(s, t, \boldsymbol{E}) \rightarrow \boldsymbol{\mathcal{L}}, \text{ where}$$

t is a statutory or regulatory term of interest (sequence of words),

s is a sentence,

$\boldsymbol{E}$ is known evidence which is a subset of the available evidence $\boldsymbol{E}^{+}$,

$\boldsymbol{\mathcal{L}}$ is an output space, either discrete (low, high) or continuous ($\{i|0 \leq i \leq 1, i \in \mathbb{R})\}$),

θ is a set of parameters of $\mathcal{U}_g$, and

$\mathcal{U}_g$ is the general interpretive utility function mapping t, s, and $\boldsymbol{E}$ to $\boldsymbol{\mathcal{L}}$.

The purpose of $\mathcal{U}_g$ is to evaluate sentence s_i with respect to its potential utility for argumentation about the meaning of term t given the known evidence (i.e., what is already known to a user). Initially, $\boldsymbol{E}$ consists of the sentences from the provision in which t resides. $\boldsymbol{E}$ grows as additional sentences are presented to the user. Since the user wishes to test some (unknown) hypothesis about the meaning of a selected term a useful sentence should meet at least some of the following criteria:

1. It provides some extra information to what is already known (i.e., what is already part of $\boldsymbol{E}$).
2. It defines t.
3. It elaborates on the meaning of t.
4. It uses or mentions t in a particularly informative way.

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 useful sentences. Legislative history and legal commentaries tentatively appear to be promising sources as well. For efficiency reasons we focused on sentences from court decisions in [39]. I will investigate the usefulness of other types of documents in future work.

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.[1] 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

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

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. For future work I 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.[2] 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.

Authors of [39] (each has a law degree) classified the sentences into categories according to their utility for the interpretation of the corresponding term. The sentences were grouped according to the decisions they come from and they were listed in the order in which they appeared in the decisions. The annotators' task was to evaluate each sentence in terms of its utility for argumentation about the meaning of the term of interest. For the purpose of this exercise $\boldsymbol{E}$ was fixed to the provision t comes from. $\boldsymbol{\mathcal{L}}$ was defined in the following way:

$$\boldsymbol{\mathcal{L}} = \{\text{high value, certain value, potential value, no value}\}$$

The categories were defined as follows:

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.

For the sake of clarity let us give a couple of examples using the example rule from the previous sections:

No vehicles in the park.

[2] 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.

The sentences that directly elaborate on the meaning of the "vehicle" belong to the "high value" category.

> Any mechanical device used for transportation of people or goods is a vehicle.
> A vehicle usually has wheels, engine and controls.

The sentences that assign or contrast the term of interest to some other term also belong to the "high value" category.

> A car is a vehicle.
> Not every vehicle is a man-made object.

The sentences that can be used to elaborate on the meaning of the "vehicle" but do not directly elaborate on the meaning themselves belong to the "certain value" category.

> Today I took my horse for a ride in that park where no vehicles are allowed.
> The main reason why no vehicles are allowed in that park is to secure a tranquil environment there.

The sentences that do not seem to be useful for elaboration on the meaning of the "vehicle" but at the same time provide additional information over what is known from the source provision belong to the "potential value" category.

> The park where no vehicles are allowed was closed during the last month.
> The courts often need to analyze the provision stating that "No vehicles are allowed in the park."

If the sentence does not provide any additional information over what is already known from the source provision it belongs to the "no value" category.

> The provision states that: "No vehicles are allowed in the park."
> A vehicle is forbidden from entering the park.

Finally, there are three possible scenarios that fall between the cracks of the above categorization. If they occur it is necessary to reassign the sentence to a lower label than would otherwise be assigned to the sentence according to the above rules. These scenarios include:

1. The sentence uses the phrase of interest from a statutory provision or case law from a different jurisdiction.

2. The sentence is attributed to a person who has a personal interest in interpreting the phrase of interest in a certain way.
3. The phrase of interest in the source provision and the phrase of interest in the sentence have different meanings.

If the sentence uses the phrase of interest from a statutory provision or case law from a different jurisdiction its usefulness could be discounted. If using the standard rules the sentence is assigned to the "potential value" or the "no value" category it is not necessary to do the discounting. However, if the sentence is assigned to the "high value" or the "certain value" category it should be re-assigned to a category one step lower.

The Indiana law states that: "Vehicle is anything which serves as a means of transport."

If the sentence is attributed to a person who has a personal interest in interpreting the term of interest in a certain way its usefulness should be discounted. If using the standard rules the sentence is assigned to the "potential value" or the "no value" category it is not necessary to do the discounting. If the sentence is assigned to the "high value" or the "certain value" category it should be re-assigned to a category one step lower.

The defendant claimed he did not break the rule since roller skates cannot be considered a vehicle.

If the phrase of interest in the source provision and the phrase of interest in the sentence have different meanings the usefulness of the sentence should be discounted. If the meanings are significantly different the sentence should be labeled as "no value".

A body is a vehicle for a soul.

If the meanings are different but strongly related the sentence should be re-assigned to a category one step lower if it would normally be labeled with the "high value" or the "certain value" category. If it would normally be labeled with the "potential value" or the "no value" category the usefulness of the sentence should not be discounted.

Any autonomous vehicle is subject to the approval of the executive committee.

Eventually, I would like the come up with a more sophisticated account of u_g, where the score comes from a continuous interval. Since we could not ask the human annotators to do the same,

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

we discretized the interval into the four categories for the purpose of the evaluation. There was no time limit imposed on the annotation process.

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.

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.

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 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.

4.0 COMPUTATIONAL SUPPORT FOR INTERPRETATION OF STATUTORY TERMS

In this chapter we present a tentative (minimalistic) framework capable of retrieving a small set of useful sentences automatically. Some parts of the framework have been implemented and tested in [39]. The purpose of this chapter is to establish feasibility of the proposed work. The processing performed by the framework for each query can be divided into four rather self-contained stages: (i) sentence retrieval (Section 4.1), (ii) sentence annotation (Section 4.2), (iii) sentence scoring (Section 4.3), and (iv) sentence selection (Section 4.4). The input to the process are:

1. t (the term of interest),
2. $\boldsymbol{E} = \{s_i | s_i \in \text{source provision}\}$ (known evidence), and
3. $\boldsymbol{DB}$ (database as described in Section 4.1).

4.1 SENTENCE RETRIEVAL

At this stage the available evidence $\boldsymbol{E}^+$ is retrieved from $\boldsymbol{DB}$. The minimalistic version of the sentence retrieval component would consist of an indexed database of sentences. The sentences are generated by a sentence segmentation tool applied to the original documents (i.e., court decisions, legislative histories, journal articles). Because general segmentation tools are known to be insufficient for legal texts it will be an interesting exercise to develop a domain specific tool. The described minimalistic solution should be sufficient for the purpose of this work.

Depending on the time and available resources it could be interesting to go beyond the above described minimalistic solution. For example a co-reference resolution could improve the retrieval. Some of the sentences that do not contain t directly may contain its co-referent. Such sentences should be considered for the inclusion in $\boldsymbol{E}^+$. If t consists of more than one word a more complex

mechanism than simple phrase matching could be helpful, especially in situations when $\boldsymbol{E}^+$ would be small. Considering other text segments apart from sentences could be useful. Some of the sentences may be related to adjacent sentences so closely that it is quite difficult to understand them without the context (e.g., some sentences starting with "therefore"). Grouping such sentences together could yield more informative results. On the other hand, some of the sentences may be very long and their division could be helpful. In legal text it is not rare to see sentences spanning many lines in printed documents.

4.2 SENTENCE ANNOTATION

The goal of this stage is to annotate sentences in terms of features that could be helpful in predicting how useful a sentence is for the argumentation about the meaning of the term of interest. Note that some of the features do not depend on the query context. The related annotations could be produced in advance for performance and scalability reasons. I present them here for the sake of clarity.

In [39] we came up with a tentative list of features that could be helpful in predicting the interpretive usefulness of a sentence. I plan the refinement of this list for future work. In addition, many features were generated with very simple models which leaves space for significant improvements. I briefly describe each of the features in the following subsections.

4.2.1 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.[1] 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.

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

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 5. 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 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 provision' label.

4.2.2 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 or the cosine similarity is higher than 0.7.
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 5. As we have predicted the 'same provision' label in most of the cases, this feature did not prove as very helpful in our experiments (see Section 4.3). I plan to refine the notion of this feature in future work. For example, I 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.'

4.2.3 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 [8]. 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 5. 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:

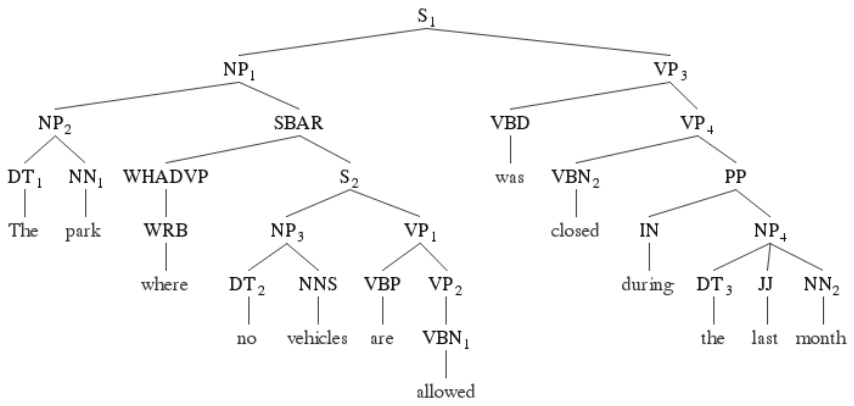

Figure 1: A syntactic tree where only one token ("allowed") is deeper in the structure than the term of interest ("vehicle").

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

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.'

4.2.4 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 5. 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 [...]

4.2.5 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 [**?**, **?**, **?**] 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 5. Most of the sentences were labeled as the 'statement of law,' the 'application of law,' or the 'interpretation of law.'

4.2.6 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 5. 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 [...].

4.2.7 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 5. 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 data set 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.

4.2.8 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 5. 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.

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.

4.3 SENTENCE SCORING

At this stage the goal is to score each sentence $s^{(i)} \in \boldsymbol{E}^{+}$ by its general interpretive utility measure $u_g^{(i)}$. The initial steps towards this goal have been already taken during the previous stage where for each $s^{(i)}$ a set of annotations was generated (see Section 4.2). Therefore, the preceding stage and this stage together form the general interpretive utility function $\mathcal{U}_g$ defined in Chapter 3. At the end of this stage each sentence is assigned the score with respect to t and $\boldsymbol{E}$ (sentences from the source provision).

In [39] we have implemented and evaluated a minimalistic version of this component by training a couple of general classification algorithms. We worked with the data set described in Chapter 3. The goal was to classify the sentences into the four categories (high, certain, potential, and no value) reflecting their utility for the argumentation about the meaning of the terms of interest. As features we use the categories described in Section 4.2.

The experiment started with a random division of the sentences into a training set (2/3) and a test set. The resulting training set consisted of 162 sentences while there were 81 sentences in the test set. As classification models we trained 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 [36]. We used a simple classifier always predicting the most frequent label as the baseline.

Because our data set was small and the division into the training and test set influenced the performance we repeated 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 3.

All the three classifiers outperformed the most frequent class baseline. However, due to the large variance of the results from the 100 runs the improvement was statistically significant ($\alpha = .05$) only for the Random Forest which was 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 was quite close to the inter-annotator agreement between the two human expert annotators (.746).

We also tested which features were the most important for the predictions with the Random

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 3: 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)

Forest. We ran the 100-batches of the experiments leaving out one feature in each batch. The results reported in Table 4 show that the source and the syntactic importance were the most important.

The results of the experiments are promising even though we used extremely simplistic (sometimes clearly inadequate) approaches to generate the sentence annotations automatically. We have every reason to expect that improvements in the quality of the annotation will improve the quality of the interpretive utility assessment. I 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 used would most likely lead to an improvement in the quality of the interpretive utility 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. I would caution against the conclusion that those features are not useful. It may very well be the case that our simplistic techniques for

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

the automatic generation of those features did not model them adequately. As already mentioned, I plan on improving the means by which the features are generated in future work.

4.4 SENTENCE SELECTION

At this stage the goal is to select the best $\boldsymbol{E}$ from $\boldsymbol{E}^+$. By best I mean such $\boldsymbol{E}$ that is the most useful for the argumentation about the meaning of a statutory or regulatory term of interest. This could be similar to the problem of selecting best sentences for a summary. The focus will be on selecting the highest scoring sentences while maximizing the difference among the sentences selected for $\boldsymbol{E}$. The selection criteria could include but not be limited to the following:

- the sentence is similar to many other sentences in $\boldsymbol{E}^+$
- joint informativeness of $\boldsymbol{E}$
- sum of the relevance scores of the documents the sentences in $\boldsymbol{E}$ come from

	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 5: 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).

5.0 HYPOTHESES

5.1 SENTENCES' GENERAL INTERPRETIVE UTILITY FOR ARGUMENTATION ABOUT MEANING OF STATUTORY OR REGULATORY TERMS IS AN OBJECTIVE MEASURE

The first question I would like to investigate is if the sentence's general interpretive utility u_g for argumentation about the meaning of a statutory or regulatory term is a well-defined measure in a sense that it is rather objective than subjective. I am interested if different people tend to evaluate sentences similarly or if the evaluation is more a matter of individual taste. For example, when people rate movies it can be expected that individual preferences have a great impact on the assigned score. On the other hand, rating of a daily weather in terms of being hot or cold should be more influenced by the actual temperature than individual preferences. This is not to say that the individual preferences would not play a role here—this is merely to say that they carry less weight in case of assessing the temperature than in case of rating a movie.

It is crucial for this work to understand if u_g is more similar to rating a movie than assessing a temperature. If individual preferences would be the dominant determiner of the utility then it would make sense to focus on user modeling and subsequent use of the user-specific models for sentence evaluation. If the sentence itself is the chief influencer then it is reasonable to concentrate on the distinctive features of the highly useful sentences. Although I do not deny that there is a certain amount of subjectivity to the utility assessment, I expect the task is more on the objective side. In order to confirm my intuition I would like to test two closely related hypotheses.

At first I would like to analyze how different people judge the utility of a sentence as compared to a utility of other sentences. This investigation aims at confirming that, even though different people may disagree about the actual utility of a sentence, they would still be able to agree on one sentence being more useful than another sentence. Specifically, I am going to test the following null

hypothesis ($\mathcal{H}0_{1a}$) and if rejected I am going to accept the following alternative ($\mathcal{H}A_{1a}$):

> $\mathcal{H}0_{1a}$: Given two sentences $s^{(i)}, s^{(j)}$, a term t, a fixed evidence $\boldsymbol{E}$, and a human expert assessment of the sentences' utility $u_g^{(i)}$, $u_g^{(j)}$ such that $u_g^{(i)} < u_g^{(j)}$, the likelihood that a different human expert assesses $u_g^{(i)}$, $u_g^{(j)}$ such that $u_g^{(i)} < u_g^{(j)}$ is equal to chance.
>
> $\mathcal{H}A_{1a}$: Given two sentences $s^{(i)}, s^{(j)}$, a term t, a fixed evidence $\boldsymbol{E}$, and a human expert assessment of the sentences' utility $u_g^{(i)}$, $u_g^{(j)}$ such that $u_g^{(i)} < u_g^{(j)}$, the likelihood that a different human expert assesses $u_g^{(i)}$, $u_g^{(j)}$ such that $u_g^{(i)} < u_g^{(j)}$ is not equal to chance.

If the experiments (described in Section 6.1.1) reject $\mathcal{H}0_{1a}$, thereby confirming $\mathcal{H}A_{1a}$, I am going to conclude that the utility assessment is objective in a sense that different people agree on the relative utility of one sentence as compared to another sentence. The firmness of this conclusion would naturally depend on the strength of the correlation.

Moreover, I would like to investigate to what extent do different people agree on the utility of sentences in absolute terms. Specifically, I am going to test the following null hypothesis ($\mathcal{H}0_{1b}$) and if rejected I am going to accept the following alternative ($\mathcal{H}A_{1a}$):

> $\mathcal{H}0_{1b}$: Given a set of sentences $\boldsymbol{S} = \{s_1, s_2, \dots s_n\}$, a term t, a fixed evidence $\boldsymbol{E}$, and two human experts' assessment of the sentences' utility $u_{g,e1}^{(i)}$, $u_{g,e2}^{(i)}$ for all $s_i \in \boldsymbol{S}$, the rate of agreement on the utility between the experts is equal to chance.
>
> $\mathcal{H}A_{1b}$: Given a set of sentences $\boldsymbol{S} = \{s_1, s_2, \dots s_n\}$, a term t, a fixed evidence $\boldsymbol{E}$, and two human experts' assessment of the sentences' utility $u_{g,e1}^{(i)}$, $u_{g,e2}^{(i)}$ for all $s_i \in \boldsymbol{S}$, the rate of agreement on the utility between the experts is not equal to chance.

If $\mathcal{H}0_{1b}$ is rejected I am going to conclude that the utility assessment is objective in a sense that different people agree on the utility of a sentence in absolute terms. The strength of the conclusion is going to depend on the rate of the p-value with which $\mathcal{H}0_{1b}$ would be rejected as well as on the actual rate of agreement among the human experts.

On the basis of rejecting both $\mathcal{H}0_{1a}$ and $\mathcal{H}0_{1b}$ I would like to discuss how objective versus subjective a measure the utility appears to be. I expect it would be possible to conclude that the utility is a rather objective measure in a sense that it warrants further investigations as described

in the following Section (5.2). The expectation is based on the results of experiments from [39] described in 3.

5.2 SENTENCES' GENERAL INTERPRETIVE UTILITY FOR ARGUMENTATION ABOUT MEANING OF STATUTORY OR REGULATORY TERMS CAN BE PREDICTED AUTOMATICALLY

Assuming the investigation of questions laid down in the previous Section (5.1) confirms that the sentence utility for interpretation of statutory terms is a sufficiently objective measure I would like to explore the possibility of determining the utility of a sentence automatically. In order to achieve this goal I would like to design a system described in Chapter 4. Such a system would allow for a similar analysis setup as the one employed in previous Section (5.1). At first I would like to determine how does the system judge the utility of a sentence as compared to a utility of other sentences with respect to gold standard created by multiple human experts. Specifically, I am going to test the following null hypothesis ($\mathcal{H}0_{2a}$) and if rejected I am going to accept the following alternative ($\mathcal{H}A_{2a}$):

$\mathcal{H}0_{2a}$: Given two sentences $s^{(i)}, s^{(j)}$, a term t, a fixed evidence $\boldsymbol{E}$, and a gold standard assessment of the sentences' utility $u_g^{(i)}$, $u_g^{(j)}$ such that $u_g^{(i)} < u_g^{(j)}$, the likelihood that the automatic utility assessment system predicts $u_g^{(i)}$, $u_g^{(j)}$ such that $u_g^{(i)} < u_g^{(j)}$ is equal to chance.

$\mathcal{H}A_{2a}$: Given two sentences $s^{(i)}, s^{(j)}$, a term t, a fixed evidence $\boldsymbol{E}$, and a gold standard assessment of the sentences' utility $u_g^{(i)}$, $u_g^{(j)}$ such that $u_g^{(i)} < u_g^{(j)}$, the likelihood that the automatic utility assessment system predicts $u_g^{(i)}$, $u_g^{(j)}$ such that $u_g^{(i)} < u_g^{(j)}$ is not equal to chance.

If the experiments (described in Section 6.2.1) reject $\mathcal{H}0_{2a}$, thereby confirming $\mathcal{H}A_{2a}$, I am going to conclude that the it is possible to evaluate sentences automatically in a sense that the computer system can agree with the human experts on the relative utility of one sentence as compared to another sentence. The firmness of this conclusion would depend on the p-value with which $\mathcal{H}0_{2a}$ would be rejected.

Furthermore, I would like to investigate to what extent the system and human experts agree on the utility of sentences in absolute terms. Specifically, I am going to test the following null

hypothesis ($\mathcal{H}0_{2b}$) and if rejected I am going to accept the following alternative ($\mathcal{H}A_{2a}$):

> $\mathcal{H}0_{2b}$: Given a set of sentences $\boldsymbol{S} = \{s_1, s_2, \ldots s_n\}$, a term t, a fixed evidence $\boldsymbol{E}$, and the computer system's assessments of the sentences' utility for interpretation $u^{(i)}_{g,sys}$ as well as the gold standard $u^{(i)}_{g,gs}$ for all $s_i \in \boldsymbol{S}$, the rate of agreement on the utility between the system and the gold standard is equal to chance.
>
> $\mathcal{H}A_{2b}$: Given a set of sentences $\boldsymbol{S} = \{s_1, s_2, \ldots s_n\}$, a term t, a fixed evidence $\boldsymbol{E}$, and the computer system's assessments of the sentences' utility for interpretation $u^{(i)}_{g,sys}$ as well as the gold standard $u^{(i)}_{g,gs}$ for all $s_i \in \boldsymbol{S}$, the rate of agreement on the utility between the system and the gold standard is not equal to chance.

If $\mathcal{H}0_{2b}$ is rejected I am going to conclude that the utility assessment is objective in a sense that the computer system and human experts agree on the utility of a sentence in absolute terms. The strength of the conclusion is going to depend on the rate of the p-value with which $\mathcal{H}0_{2b}$ would be rejected as well as on the actual rate of agreement among the human experts.

On the basis of rejecting both $\mathcal{H}0_{2a}$ and $\mathcal{H}0_{2b}$ I would like to discuss how successful the designed system is in predicting the utility of a sentence for interpretation. The preliminary results presented in Section 4.3 are quite promising. They provide a strong basis for expectation that it would be possible to conclude that the system is capable to predict the utility with reasonable success. By reasonable success I mean such an agreement with the gold standard that further investigations as described in the following Section (5.3) would be warranted.

5.3 SENTENCES' GENERAL INTERPRETIVE UTILITY IMPROVES SELECTION OF EXAMPLE SENTENCES FOR ARGUMENTATION ABOUT MEANING OF STATUTORY OR REGULATORY TERMS

Provided the analysis of questions posed in the previous Sections (5.1 and 5.2) establishes that (i) the sentence's general interpretive utility for argumentation about meaning of statutory or regulatory terms is an objective measure and that (ii) the sentence's general interpretive utility for argumentation about meaning of statutory or regulatory terms can be predicted automatically, I would like to explore the potential of the utility measure to enhance the selection of example

sentences for argumentation about meaning of statutory or regulatory terms. To achieve this goal I would like to implement a prototype system using the sentence utility to select the best example sentences. I would like to compare the system's performance to that of the baselines (standard IR and summarization techniques). Specifically, I am going to test the following null hypothesis ($\mathcal{H}0_3$) and if rejected I am going to accept the following alternative ($\mathcal{H}A_3$):

$\mathcal{H}0_3$: Given a set of sentences $\boldsymbol{S} = \{s_1, s_2, \ldots s_n\}$, a fixed evidence $\boldsymbol{E}$, the computer system's assessments of the sentences' utility for interpretation $u^{(i)}_{g,sys}$, the computer system's ranked selection of k most useful sentences, the baseline systems' ranked selection of k most useful sentences, as well as the gold standard selection of k most useful sentences, the rate of agreement between the computer system's ranking and the gold standard ranking is the same as the rates of agreement between the baseline systems' rankings and the gold standard ranking.

$\mathcal{H}A_3$: Given a set of sentences $\boldsymbol{S} = \{s_1, s_2, \ldots s_n\}$, a fixed evidence $\boldsymbol{E}$, the computer system's assessments of the sentences' utility for interpretation $u^{(i)}_{g,sys}$, the computer system's ranked selection of k most useful sentences, the baseline systems' ranked selection of k most useful sentences, as well as the gold standard selection of k most useful sentences, the rate of agreement between the computer system's ranking and the gold standard ranking is not the same as the rates of agreement between the baseline systems' rankings and the gold standard ranking.

On the basis of rejecting (or not rejecting) $\mathcal{H}0_3$ I would like to discuss how successful the designed system is in selecting the k most useful sentences and their ranking.

6.0 EVALUATION

6.1 OBJECTIVITY OF SENTENCES' GENERAL INTERPRETIVE UTILITY FOR ARGUMENTATION ABOUT MEANING OF STATUTORY OR REGULATORY TERMS

The hypotheses $\mathcal{H}_{1a}$ and $\mathcal{H}_{1b}$ (see Section 5.1) will be tested by means of the labeling effort the purpose of which is to create the data set to test hypotheses $\mathcal{H}_{2a}$ and $\mathcal{H}_{2b}$ (see Section 5.2). For this effort I will first compile a short list of approximately 10–20 statutory and regulatory terms. For each term I will retrieve a set of sentences mentioning the term. The sentences will come from court decisions, legislative histories, and journal articles. The number of sentences for each term may vary but I expect it to be between 100 and 200. Hence, there will be a data set of a couple of thousands sentences.

The data set will be focused on the area of cybercrime. At a minimum it will include cybercrime-related: (i) federal statutory provisions; (ii) statutory provisions from selected states (iii); federal case-law; (iv) case-law of selected states; (v) doctrinal texts; (vi) legislative histories, and (vi) available materials from other countries. The required documents will be obtained from publicly available free sources. The statutory law at the federal level can be freely accessed at Legal Information Institute.[1] State governments usually maintain publicly available databases for their respective states.[2] The Court Listener free service offers almost 3 million opinions from both federal and state courts.[3] Doctrinal texts may be obtained from publicly available databases such as the Social Science Research Network.[4] Federal legislative information, including full-text access to public laws and congressional bills (103rd Congress forward), House and Senate reports (104th

[1] https://www.law.cornell.edu/

[2] E.g., New York (http://public.leginfo.state.ny.us/navigate.cgi), Pennsylvania (http://www.legis.state.pa.us/cfdocs/legis/LI/Public/cons_index.cfm), or California (http://leginfo.legislature.ca.gov/faces/codes.xhtml).

[3] https://www.courtlistener.com/

[4] http://ssrn.com/en/

Congress forward), nominations (97th Congress forward), and the Congressional Record (104th Congress forward), are available at Congress.gov website.[5]

We propose to evaluate the system for the domain of cybercrime because statutory interpretation is probably the most challenging in rapidly developing areas with strong element of cutting edge information technology (IT). One such area is cybercrime, which concerns criminal offenses in cyberspace. [7, p. 3] In dealing with cybercrime, there are challenges to cope with in addition to the usual challenges involved in working with statutes. There is a demand to understand the relevant concepts and terminology of IT. Often it is necessary to apply statutory provisions enacted before the advent of IT to address a cybercrime offense. We deliberately select such a challenging domain because it is our goal that the system will generalize well to other domains. We are convinced that if the system could reasonably address statutory interpretation in the domain of cybercrime it could handle any other domain where statutory interpretation plays an important role. In addition, the fight against cybercrime is now based on an extensive international cooperation, which would allow us to explore the possibilities of working with resources from multiple countries. If time and resources permit I plan to include some resources from other countries into a data set.

I will create a labeling scheme for sentences and labeling instructions for human expert annotators. These could be similar to the scheme and the instructions used in [39] and described in Chapter 3. However, I expect there will be certain changes based on the experience gained from those preliminary experiments. I will also consider a possibility of creating a training module for the annotators to ensure that they provide high quality labels from the beginning of their work. The preliminary experiments show that the agreement between annotators was lower at the beginning.

I will design a labeling procedure on the basis of which at least two persons with a law degree and possibly a number of law students will label the sentences. Each sentence will be labeled by at least two human experts. If resources allow I would like to obtain more than two labels for some of the sentences. There is a law librarian's course taught at the University of Pittsburgh's Law School. This kind of annotation could be used as a task for extra credit.

[5] https://www.congress.gov/

6.1.1 Relative Sentences' Utility for Argumentation about Meaning of Statutory or Regulatory Terms

To analyze how different people judge the utility of a sentence as compared to a utility of other sentences I would like to test the hypothesis $\mathcal{H}_{1a}$ by means of a rank test. Computation of the Kendall rank correlation coefficient seems appropriate for this case:

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{n(n-1)/2}$$

In addition to rejecting the null hypothesis I would be looking for a high correlation coefficient. Based on the p-value with which the null hypothesis is rejected and the strength of the correlation I will discuss how different people judge the utility of a sentence as compared to a utility of other sentences. I expect that it would be possible to conclude that human experts can reasonably agree on the relative utility of sentences.

6.1.2 Absolute Sentences' Utility for Argumentation about Meaning of Statutory or Regulatory Terms

To assess to what extent do different people agree on the utility of sentences in absolute terms I would like to test the hypothesis $\mathcal{H}_{1b}$ by means of measuring inter-annotator agreement. Cohen's kappa and its weighted version appear appropriate for this test:

$$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e}$$

$$\kappa = 1 - \frac{\sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} x_{ij}}{\sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} m_{ij}}$$

In addition to rejecting the null hypothesis I would be looking for a high correlation coefficient. Based on the p-value with which the null hypothesis is rejected and the strength of the correlation I will discuss to what extent do different people agree on the utility of sentences in absolute terms. I expect that it would be possible to conclude that human experts can reasonably agree on the utility of sentences in absolute terms.

6.2 AUTOMATIC PREDICTION OF SENTENCES' GENERAL INTERPRETIVE UTILITY FOR ARGUMENTATION ABOUT MEANING OF STATUTORY OR REGULATORY TERMS

The hypotheses $\mathcal{H}_{2a}$ and $\mathcal{H}_{2b}$ (see Section 5.2) will be tested by means of a prototype sentence evaluation system. A rudimentary version of such a system was implemented in [39] and described in 4. For the purpose of this work I plan to go beyond the system from [39]. Specifically, I plan to work on the features the system uses. Additionally, going beyond the simple classification algorithms could improve the performance significantly. I will use the data set described in the preceding section (6.1) to train and evaluate the system.

I will design a procedure to generate gold standard labels for the data set. I could use the procedure that was used in [39]. There the human annotators met and discussed the data points where their labels differed. The consensus labels were then treated as the gold standard. The experiments will be similar to those described in the preceding section (6.1). Instead of labels from multiple human annotators I will use the gold standard labels and the predictions of the implemented prototype system.

6.2.1 Automatic Prediction in Relative Terms

To analyze how the prototype system predicts the utility of a sentence as compared to a utility of other sentences with respect to the gold standard labels I would like to test the hypothesis $\mathcal{H}_{1a}$ by means of a rank test. Computation of the Kendall rank correlation coefficient seems appropriate for this case.

In addition to rejecting the null hypothesis I would be looking for a high correlation coefficient. Based on the p-value with which the null hypothesis is rejected and the strength of the correlation I will discuss how the prototype system predicts the utility of a sentence as compared to a utility of other sentences with respect to the gold standard labels. I expect that it would be possible to conclude that the prototype system can reasonably agree on the relative utility of sentences with the gold standard labels.

6.2.2 Automatic Prediction in Absolute Terms

To assess to what extent do the prototype system's prediction on the utility of sentences agree with the gold standard labels in absolute terms I would like to test the hypothesis $\mathcal{H}_{1b}$ by means of measuring inter-annotator agreement. Cohen's kappa and its weighted version appear appropriate for this test.

In addition to rejecting the null hypothesis I would be looking for a high correlation coefficient. Based on the p-value with which the null hypothesis is rejected and the strength of the correlation I will discuss to what extent do the prototype system's prediction on the utility of sentences agree with the gold standard labels in absolute terms. I expect that it would be possible to conclude that the prototype system's predictions on the utility of sentences can reasonably agree with the gold standard labels in absolute terms.

6.3 AUTOMATIC SELECTION AND RANKING OF THE MOST USEFUL SENTENCES BASED ON THE GENERAL INTERPRETIVE UTILITY MEASURE

The hypothesis $\mathcal{H}_3$ (see Section 5.3) will be tested by means of a prototype sentence selection and ranking system. In order to facilitate training and evaluation of the system I plan to use the data set of sentences from court decisions, legislative histories, and journal articles (described in 6.1). For each term multiple human experts will select and rank approximately 20 best sentences. The gold standard ranking will be computed from rankings of different experts. As the baseline systems I plan to use some variant(s) of a standard IR system based on a bag of words representation and some variant(s) of a general purpose document summarization system.

To analyze how the prototype system performs in selecting and ranking k most useful sentences for argumentation about meaning of a term as compared to the baseline systems I would like to test the hypotheses by means of traditional IR measures such as recall, precision, F1-measure, mean average precision, or normalized discounted cumulative gain. I expect that it would be possible to conclude that the prototype system significantly outperforms the baseline systems. Precision, Recall, and F_1-measure are defined in the following way:

$$P = \frac{TP}{TP + FP}$$

$$R = \frac{TP}{TP + FN}$$

$$F_1 = \frac{2 * P * R}{P + R}$$

TP stands for true positives (sentences that were correctly retrieved as useful), FP for false positives (sentences that were incorrectly retrieved as useful) and FN for false negatives (sentences that were incorrectly retrieved as useful).

Mean average precision (MAP) is defined as follows:

$$MAP(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} P(R_{jk})$$

For a single information need, Average Precision is the average of the precision value obtained for the set of top k documents existing after each relevant document is retrieved, and this value is then averaged over information needs. That is, the set of relevant documents for an information need $q_j \in Q$ is $\{d_1, \ldots d_{m_j}\}$ and R_{jk} is the set of ranked retrieval results from the top result until you get to document d_k. [43]

Normalized discounted cumulative gain is defined in the following way:

$$NDCG(Q, k) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} Z_{kj} \sum_{m=1}^{k} \frac{2^{R(j,m)} - 1}{log_2(1 + m)}$$

Like precision at k, it is evaluated over some number k of top search results. For a set of queries Q, $R(j, d)$ is the relevance score assessors gave to document d for query j. Z_{kj} is a normalization factor calculated to make it so that a perfect ranking's NDCG at k for query j is 1. For queries for which $k' < k$ documents are retrieved, the last summation is done up to k'. [43]

7.0 RELATED WORK

Many legal philosophers focus their research on the relationship between law and language. [9, 14, 28] Some legal scholars are approaching the topic from a more practical point of view focusing on interpretation [27, 28, 45] and reading [15, 41] of legal texts, especially the statutes.

The current tool of choice for support of statutory interpretation are traditional legal IR systems. Their major limitation is that they return too many documents to read and assess. [33] A typical response to this challenge is enhancing IR systems with techniques used in case-based reasoning [33], dynamic document classification [29], keyword extraction [44], recommendation systems [37, 49], or conceptual retrieval [18].

Recommendation systems are applications that specialize in predicting user responses to options. [23] A survey of recommendation systems and techniques used can be found in [1, 2]. Especially the content-based recommendation systems, which base the recommendations on the properties of the recommended items, are relevant for our work. In [49] a legal recommendation system is proposed using references among legal documents as a basis for the suggestions. In our work we also plan to use the references but as described above we plan to go well beyond that.

The work is closely related to the area of question answering. The task of selecting an answer sentence is to choose the correct sentence containing an answer given a question and a set of candidate sentences. The most notable success in question answering is the IBM Watson, the computer that won Jeopardy! [16] Watson is based on UIMA, the architecture we plan to use as well. In our work we do not aim to provide answers directly. Instead we aim to provide a list of statements that most effectively reduce user's uncertainty given the term of interest and the corpus evidence. In this respect the work is similar to the claim detection [24] used in IBM's Debater project.

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 [34, 30, 32, 31]. In their work on vaccine injury decisions Walker, Ashley, Grabmair and other researchers focus on extraction of evidential reasoning [47, 3, 18]. Bruninghaus and Ashley [6] and Kubosawa et al. [22] 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 [21] or public participation in rulemaking [35].

The task we deal with is close to the traditional NLP task of query-focused summarization of multiple documents as described in Gupta [19]. Fisher and Roark [17] presented a system based on supervised sentence ranking. Daumé and Marcu [10] tackled the situation in which the retrieved pool of documents is large. Schiffman and McKeown [42] cast the task into a question answering problem. An extension introducing interactivity was proposed by Lin et al. [25].

A number of interesting applications deal with similar tasks in different domains. Sauper and Barzilay [38] proposed an approach to automatic generation of Wikipedia articles. Demner-Fushman and Lin [11] described an extractive summarization system for clinical QA. Wang et al. [48] presented a system for recommending relevant information to the users of Internet forums and blogs. Yu et al. [50] mine important product aspects from online consumer reviews.

8.0 CONTRIBUTIONS

The main contributions of the work will pertain to the following areas:

a) **Legal IR** – The work focuses on a very important type of a user information need in legal IR. An original definition of the task is provided. A special set of techniques to solve the task is put together. A prototype system capable of handling the task is implemented. All these pieces are unique contributions to general research on legal IR.

b) **Summarization** – The work could generate some interesting results in summarization. It presents a unique notion of sentence importance that could be very relevant in query-focused summarization.

c) **Corpus-based Lexicography** – In corpus-based lexicography the availability of good example sentences plays a key role in formulation of a definition. The selection of particularly helpful example sentences for the inclusion into definition is also important. [4] This work contributes to a body of work on the automatic selection of example sentences.

d) **Applications** – The resulting prototype system could find many valuable applications in different areas. Apart from facilitating easier access to law for lawyers it could lower the barrier for public officials and other users who need to work with legal texts. In addition, we believe the system could support dialog between lawyers and experts from other fields. There could be a great impact on legal education as well. The resulting system would significantly improve the effectiveness of statutory analysis. The technology could become a valuable part of existing legal IR systems, or it can be deployed as a standalone system.

e) **Resources** – The resources created in the course of the work (e.g., corpus, ranking data) would make it easy for other researchers to improve the system or extend it. We plan to make the resources publicly available.

9.0 TENTATIVE TIMELINE

Following the proposal defense in November 2016 a work on the data set will commence. First, the required documents such as court decisions, legislative histories, or journal articles, will be downloaded and organized into a coherent data base. Then, a list of example terms of interest will be assembled and a framework for sentence labeling will be put together. This stage of the work should be finished by the end of February 2017. The labeling of sentences with respect to their utility for argumentation about the meaning of the terms of interest will take place through March and April 2017. The rankings of top sentences for each term will be put together at the same time. The prototype system will be implemented and tested during the summer months of 2017. Starting in the summer the dissertation should be finished in early fall 2017. I would like to defend the dissertation in October or November 2017.

BIBLIOGRAPHY

[1] Gediminas Adomavicius and Alexander Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE transactions on knowledge and data engineering*, 17(6):734–749, 2005.

[2] Chris Anderson. *The long tail: Why the future of business is selling more for less.* Hyperion, 2006.

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

[4] BT Sue Atkins and Michael Rundell. *The Oxford guide to practical lexicography.* Oxford University Press, 2008.

[5] Brian Bix. Law, language, and legal determinacy. 1993.

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

[7] Mohamed Chawki, Ashraf Darwish, Mohammad Ayoub Khan, and Sapna Tyagi. *Cybercrime, digital forensics and jurisdiction*, volume 593. Springer, 2015.

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

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

[10] Hal Daumé III and Daniel Marcu. 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, 2006.

[11] Dina Demner-Fushman and Jimmy Lin. 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, 2006.

[12] Paul Égré and Nathan Klinedinst. Vagueness and language use. 2010.

[13] Timothy Endicott. *Vagueness in Law*. Oxford University Press, 2000.

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

[15] William N Eskridge, Philip P Frickey, and Elizabeth Garrett. *Cases and Materials on Legislation: Statutes and the Creation of Public Policy*. West Academic Publishing, 2001.

[16] David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A Kalyanpur, Adam Lally, J William Murdock, Eric Nyberg, John Prager, et al. Building watson: An overview of the deepqa project. *AI magazine*, 31(3):59–79, 2010.

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

[18] Matthias Grabmair, Kevin D Ashley, Ran Chen, Preethi Sureshkumar, Chen Wang, Eric Nyberg, and Vern R Walker. 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, 2015.

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

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

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

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

[23] Jure Leskovec, Anand Rajaraman, and Jeffrey David Ullman. *Mining of massive datasets*. Cambridge University Press, 2014.

[24] Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. Context dependent claim detection. 2014.

[25] Jimmy Lin, Nitin Madnani, and Bonnie J Dorr. 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, 2010.

[26] Niklas Luhmann, Klaus A Ziegert, and Fatima Kastner. *Law as a social system*. Oxford University Press on Demand, 2004.

[27] Neil MacCormick and Robert S Summers. *Interpreting statutes: a comparative study*, volume 23. Dartmouth Publishing Company, 1991.

[28] Andrei Marmor. *The language of law*. OUP Oxford, 2014.

[29] Dieter Merkl and Erich Schweighofer. The exploration of legal text corpora with hierarchical neural networks: a guided tour in public international law. In *Proceedings of the 6th international conference on Artificial intelligence and law*, pages 98–105. ACM, 1997.

[30] Raquel Mochales and Aagje Ieven. 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, 2009.

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

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

[33] Marie-Francine Moens. What information retrieval can learn from case-based reasoning. In *Proceedings JURIX*, pages 83–91, 2002.

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

[35] Joonsuk Park, Cheryl Blake, and Claire Cardie. 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, 2015.

[36] 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. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[37] Paulo Quaresma and Irene Pimenta Rodrigues. A collaborative legal information retrieval system using dynamic logic programming. In *Proceedings of the 7th international conference on Artificial intelligence and law*, pages 190–191. ACM, 1999.

[38] Christina Sauper and Regina Barzilay. 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, 2009.

[39] Jaromır Šavelka and Kevin D Ashley. Extracting case law sentences for argumentation about the meaning of statutory terms. *ACL 2016*, page 50, 2016.

[40] Jaromir Šavelka and Jakub Harašta. 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, 2015.

[41] Antonin Scalia and Bryan A Garner. *Reading law: The interpretation of legal texts*. Thomson/West, 2012.

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

[43] Hinrich Schütze. Introduction to information retrieval. In *Proceedings of the international communication of association for computing machinery conference*, 2008.

[44] Erich Schweighofer and Dieter Merkl. A learning technique for legal document analysis. In *Proceedings of the 7th international conference on Artificial intelligence and law*, pages 156–163. ACM, 1999.

[45] Lawrence Solan. *The language of statutes: Laws and their interpretation*. University of Chicago Press, 2010.

[46] Roy Sorensen. Vagueness. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Winter 2016 edition, 2016.

[47] Vern R Walker, Nathaniel Carie, Courtney C DeWitt, and Eric Lesh. 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, 2011.

[48] Jia Wang, Qing Li, Yuanzhu Peter Chen, and Zhangxi Lin. 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, 2010.

[49] Radboud Winkels, Alexander Boer, Bart Vredebregt, and Alexander van SOMEREN. Towards a legal recommender system. In *JURIX*, pages 169–178, 2014.

[50] Jianxing Yu, Zheng-Jun Zha, Meng Wang, and Tat-Seng Chua. 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, 2011. Association for Computational Linguistics.