Modeling Ambiguity in Text: A Corpus of Legal Literature

Roma Patel
University of Pennsylvania
romap@seas.upenn.edu

Ani Nenkova
University of Pennsylvania
nenkova@seas.upenn.edu

Abstract

Identifying instances of semantic ambiguity is fundamental to natural language understanding. While this research area has been studied extensively, there exists no dataset that is well suited to analysing and testing theories of ambiguity and misinterpretation in language. In this paper, we introduce a series of corpora i.e., a unique dataset of text of the U. S. Code, combined with text of Supreme and Circuit Court cases that together form a large-scale dataset that can serve as testing ground for the tasks and instances of ambiguity that we introduce. Our experiments show that our annotated corpus allows baseline neural and linear models to achieve measurable success in tasks focused on determining ambiguity at different levels of granularity e.g., words, paragraphs and documents. We hope that this introduction serves to enable further research in this area, by using the detailed instances outlined in the corpus.

1 Introduction

In law, statutory interpretation for legislation has long since been established as a necessary procedure for judicial authorities and the pragmatic effects of language i.e., its context-dependence are frequently brought up as matters of jurisprudential debate (Endicott, 2016). The words used in the text of the statute may be clear in certain cases, but often times result in multiple interpretations, thus giving rise to disputes that are brought to court. In such instances, it is up to the judiciary to make a final ruling by using traditional canons of statutory interpretation and legislative history. It is up to them to disambiguate the language used and make a decision as to which interpretation of the words is most correct, in that context.

Such court cases and their corresponding cited statutes serve as documented examples to help explore this area. While some cases that go to the Supreme Court are simply a result of the topical nature e.g., waterbodies around the U.S., treaties and conflicts with diplomats; others are instances where judicial opinion is necessary to *interpret* a vague statement in context and make a final judicial decision.

Consider an example statute in Title 29, §1383(d)(2) of the U. S. Code. The text of the statute is in italics, while the legal definition terms are bolded.

A ***plan*** *is described in this paragraph if substantially all of the contributions required under the* ***plan*** *are made by* ***employers*** *primarily engaged in the long and short haul trucking industry, the household goods moving industry, or the public warehousing industry.*

The use of the words *substantially all* lead to some degree of ambiguity. The phrase could be considered to mean simply an absolute majority, a significantly greater fraction, or some other measure and in instances like this, interpretative canons for judicial ruling are imperative. What is also important are the words in bold i.e. legal definition terms. These are words that the law requires to *not* be open to interpretation e.g., the word *plan* is a legal definition term and is defined as per its intended meaning, for that statute. Each section in the U. S. Code has a set of definition terms, and within the scope of that section, when used, the meaning of such words must be the one that is defined in that context only, to avoid differing interpretations.

The use of language is thus critical to any legal system. Research into this area, specifically for the legal domain, is limited by the fact that there exists no unified dataset that links the text of controversial cases to the text of the actual laws and statutes that are required to fully understand the case. There also does not exist, to the best of our knowledge, a detailed hierarchical outline of the statutes and legal definition terms contained in the U.S. Code. When these elements of the legal code are coupled with statistics relating to the number of times they were contested or have resulted in ambiguity in the past, they give rise to a new formalism of ambiguity in language and provide a corpus for further research in this area. In this paper we want to bring to attention both the importance and applicability of the data outlined here, to the NLP community and highlight the different aspects of ambiguity and interpretation that our experiments uncover.

2 Related Work

There has been an extensive study of ambiguity and meaning of utterances, when dealing with discourse and interactions between people and the implicatures that thus occur (Grice, 1975, 1981; Davis, 2014). These include *conversational implicatures* i.e., complex meanings that interacting agents create and *embedded implicatures* i.e., meaning derived from both semantics of the utterance and pragmatically enhanced inferences; and recent work has addressed predictive methods to identify and address such instances (Potts et al., 2016; Russell, 2012). Certain situations seem to demand more information that can be gained from only the utterances at hand, and are potential causes of ambiguity. Work in this area includes the compositional lexical uncertainty models of (Bergen et al., 2016) and (Bergen et al., 2012) that account for specificity implicatures and model how pragmatic effects of different utterances are driven by linguistic and social contexts. However, rather than considering documented statements e.g., codification of laws and defined terms, most of these models focus on how discourse participants coordinate with one another and seek to derive an understanding of the imparted meaning of utterances and possible ambiguity based on such interactions.

Traditional accounts of ambiguity for general text have dealt with structural ambiguity and lexical relations (Hindle and Rooth, 1993), which is different from the kind of conceptual ambiguity we outline in this paper. The canonical case for structural ambiguity is prepositional phrase attachment (Bailey et al., 2015; Hindle and Rooth, 1993) and resolving such attachment ambiguities has been studied extensively in syntactic analysis. Approaches for dealing with such ambiguities include memory-based learning (Zavrel et al., 1997), making use of semantically-tagged corpora (Stetina and Nagao, 1997) to improve modeling of semantic relations between prepositions and other constituents (Srikumar and Roth, 2013) and to predict prepositional phrase attachment (Ratnaparkhi et al., 1994; Collins and Brooks, 1995; Brill and Resnik, 1994).

Here we focus on a different aspect of ambiguity; solely considering documented text, rather than interactions between agents and more focused on semantic and pragmatic meaning, rather than structural and syntactic elements. The meaning of a word in different contexts can differ significantly, thus leading to different interpretations. Several aspects of word sense disambiguation have been studied, to create lexical resources (e.g., WordNet (Miller and Fellbaum, 1998)) that provided hierarchical information and distinct senses for words, and semantic networks (e.g., ConceptNet (Havasi et al., 2007)) that build on concepts i.e., aspects of the world that are related to corresponding words. Considering the variation of senses and concepts associated with words, it is then crucial in a legal system, to precisely define the meaning to be understood by a word in a context. When complex words that are associated with multiple concepts are used in legal literature, they must be defined appropriately, to allow reasoning about their interpretation in that context. There has been an extensive analysis of the use of language in legal systems (Endicott, 2016) and on the subject of legal reasoning (Dickson, 2016), to both avoid issues of ambiguity and misinterpretation and also correctly reason about contextual interpretations.

3 Corpus

Here we introduce the two parallel datasets: the text contained in the United States Code (U.S. Code) and more than 20,000 court cases from respective courts i.e., the Supreme Court and 11 Circuit Courts of Appeal.

3.1 United States Code

The U. S. Code is a comprehensive set of general and permanent laws, statutes and legal definition terms. After the President signs a bill into law (or the Congress enacts it over his veto), all such public laws are incorporated into the U. S. Code, each serving as an individual statute. Given the need for preciseness of language, when statutes refer to certain terms that need definition, these are defined in the text of the same section. Section 3.1.1 explains the overall structure of the U. S. Code, while sections 3.1.2 and 3.1.3 explain aspects of the individual elements in detail, in order to clearly outline their use in our corpus.

3.1.1 Structure

The entire Code is structured into 54 titles, each pertaining to a broad topic e.g., *Title 3 - The President*, *Title 8 - Aliens and Nationality* or *Title 31 - Money and Finance*. Each title is subcategorised into chapters, subchapters, parts and subparts, finally leading to a unique section that forms the fundamental unit of the U. S. Code. The text on each section page may be a single statute or definition term, or a number of statutes and terms categorised by their paragraphs. For example, §1101(2)(A) refers to the first subpart of the second part of the page corresponding to §1101. This kind of hierarchical structure is useful when we want to link to a precise sentence or textual unit e.g., a certain law or statement that was contested in court.

There exist 43041[1] (53061) sections of the U. S. Code across the 53 Titles, each of varying length. While the number of words per section ranges from 50 to 30,000, the average number of words per section is 496 with the average number of sentences being 12. The total number of words contained in the entire U. S. Code is greater than 19 million.

There is a distinct variation regarding the density of content contained in each title. For example, *Title 42 - The Public Health and Welfare* has by far the largest number of words, statutes and legal definition terms than any other title, while *Title 27 - Intoxicating Liquors* and *Title 9 - Arbitration* both have total number of words less than 1% that of Title 42. From Figure 1, we gain useful insight about the content of Titles in the U. S. Code. The blue line shows the distribution of total number of words i.e., the need for statutes, laws, definitions and content for the corresponding Titles on the x-axis. The red line shows the distribution of the vocabulary normalised by the number of legal definitions in that Title. This is intuitively, a measure of the kind of words used i.e., potentially ambiguous words that require definition, as opposed to general language. The red line shows a distribution of number of words normalised by the number of definitions. This shows us how often the legal definitions are required to be used, once defined. The plots portray the nature of Titles in the U. S. Code and give us useful information about their characteristics e.g., a larger number of laws and statutes signals that this involves possibly important or contentious issues. Similarly, a larger number of definitions (i.e., smaller number of unique words normalised by definitions) could imply more ambiguity i.e., these are words that require clear-cut definitions in context to avoid misinterpretation.

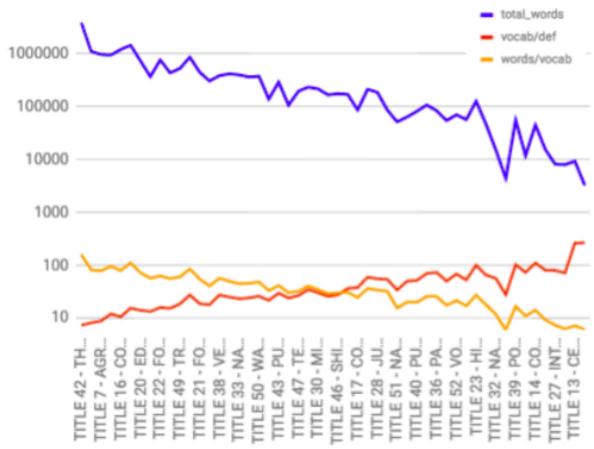


Figure 1: Figure shows three distributions across the largest 25 Titles of the U. S. Code. The blue line shows total number of words i.e., content in the title. The red line shows number of unique words normalised by definition terms. The yellow line shows words normalised by vocabulary. Note that the y-axis uses a log-scale, to note the general trend, rather than exact numbers, given the large number of words per Title.

3.1.2 Definitions

Certain words are explicitly defined by the U. S. Code, so as to not leave uncertainty about their meaning and intended use. Legal definition terms include proper nouns (e.g., *Secretary of State, China*) and acronyms (e.g., *AIDS, ANSI, LED*) as well as time periods of events (*1941, 1942*) as recognition of important terms. Apart from the above categories, there are definition terms that relate to concepts, actions and entities, and these are

[1] Of the 53061 sections, roughly 10,000 have been repealed, although section numbers exist on the U. S. Code website.

the terms often brought up or contested in court cases i.e. potentially lead to ambiguity. These include terms like *abuse, child, damages, royalties, predatory, violation*. What is also important is how these conceptual terms differ in their definitions in different contexts. These multiple definitions reflect the varied concepts that are associated with such words, thus adding to their vagueness in general language.

Section	Definition
Title 1, § 8	*every infant member of the species homo sapiens who is born alive at any stage of development*
Title 20, § 7801	*any person within the age limits for which the State provides free public education*
Title 42, § 1901	*a legitimate child, an adopted child, and, if designated as beneficiary by the insured, a stepchild or an illegitimate child*
Title 42, § 1397jj	*an individual under 19 years of age*
Title 42, §1769(d)	*a person under the age of 18*
Title 42, §5119(c)	*a person who is a child for purposes of the criminal child abuse law of a State*

Table 1: Examples of the how the legal definition *child* is defined across the U. S. Code.

The total number of unique defined terms in the U. S. Code is 15539, with 3210 words defined more than once and 136 words defined more than ten times. A large percentage (86.8%) of the definitions are multi-word terms. These sometimes include common expressions that appear normally in general text e.g., *homeless individual, urban area, natural disaster, combat zone* but also less common sequences such as *simplified acquisition threshold* and public safety answering point.

Only 10% (5341/53051) of the sections in the U. S. Code contain a list of legal terms defined in that section. Nearly all sections however, make use of at least one legal definition term. Given the importance of scoping language and multiple definitions, this is potentially a cause for ambiguity and gives rise to an important NLP application and test for pragmatic or contextual reasoning e.g., determining which definition term and corresponding gloss should best suit the context, to resolve a legal issue.

3.1.3 Statutes

A statute is an act of legislature that sets forth general propositions of law that courts apply to specific situations. In the U. S. Code, statutes appear within sections and the judiciary tends to follow a few general rules in determining the scope or meaning of a statute. The context or immediate scope is especially important e.g., the word *interest*, when used in a statute, can refer to either ownership of property or monetary charge. If the section that it appears in is related to real-estate or land ownership, the former meaning should be construed but in another context that refers to monetary policy, the latter meaning should take precedence.

Of all the sections of the U. S. Code that we have compiled, certain sections have been contested in court on several occasions. The proceedings of court cases cite relevant sections of the U. S. Code which we have extracted as one set of annotations. However, all cited sections are not necessarily under dispute and moreover, all cases do not necessarily have ambiguity issues. For the second type of annotation, we therefore consider *Chevron* cases (explained in Section 3.2) and the sections cited here, that were contested in court due to misinterpretation or ambiguity of the law. We moreover, have a (growing) set of manual annotations by law students, at varying levels of granularity within the section, that identify precise snippets of text that led to ambiguity, wherever possible.

3.2 Court Cases

Proceedings of all Circuit and Supreme Courts are documented by the judiciary. This includes the judicial opinion, the judge's recitation of the factual information surrounding the case, description of the case proceedings and if present, procedural history. Contested statutes and definition terms often appear throughout the text of the proceeding, along with citations to the section in the U. S. Code that they appear in. Certain cases require *Chevron Deference* (Bradley, 2000) that is particularly important to our study of ambiguity. The term *Chevron Deference* was coined in the 1900's after a landmark case, involving the company *Chevron U.S.A., Inc. v. Natural Resources Defense Council, Inc*. The reason that this case was brought to court was because the two parties differed in their interpretation of the law, and the judiciary had to make a decision as to whether to grant deference to the agency's interpretation of the law. Ever since, this term has been used by courts to determine whether to grant deference to a government agency's interpretation of a statute which it administers. Cases that require *Chevron Deference* thus indicate ambiguity and differing interpretations of statutes and definition terms contested in such cases.

This dataset contains cases from the Supreme Court and from the eleven Circuit Courts. All cases are annotated for *Chevron Deference* citations i.e. they can be clearly partitioned into cases that have resulted in multiple interpretations or not. All cases are also annotated and linked to the sections of the US Code that they refer to i.e. the precise statutes, laws and definitions that were contested in the case.

The total number of words contained in the court cases is greater than 12 million, while the number of words in the U. S. Code is greater than 19 million.

3.2.1 Annotations

Each case is annotated with several kinds of information as displayed in Table 2. These include the court that passed judicial ruling, the year that the case was brought to court, the parties involved in the case, the Titles and sections of the U. S. Code cited in the case, whether the case required *Chevron* deference or not and our manually annotated set of granular 5-level annotations.

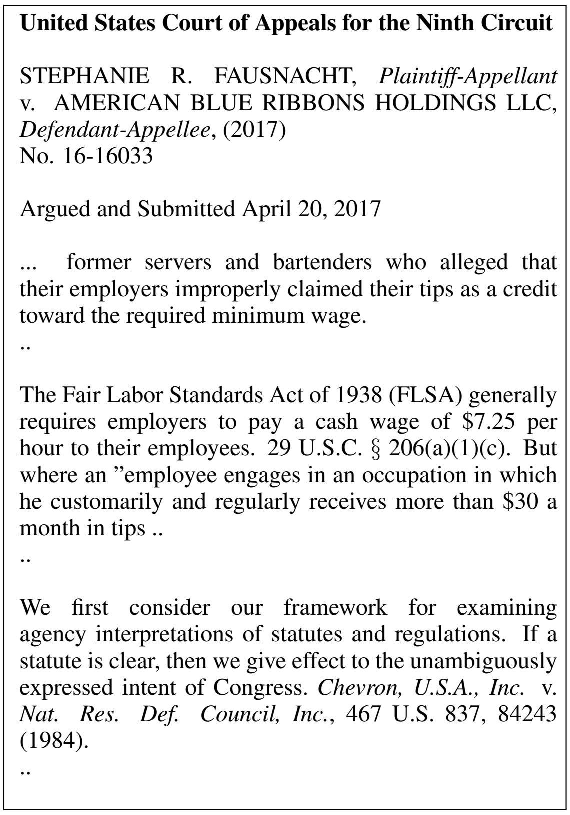
3.3 Data Collection

In recent years, a number of tools and applications have been proposed to help web scraping for building corpora for NLP applications. Here, we collect data from the web, referring only to several trusted web resources that are outlined below, for the different datasets we have built. We briefly describe the process undergone, tools used and websites scraped for the different datasets.

3.3.1 U. S. Code

The Legal Information Institute[2] is a not-for-profit institute, primarily funded by the Cornell Law School, that publishes legal resources online for no charge. Its website contains all the text of the U. S. Code.

We use the Python package BeautifulSoup (Richardson, 2013) in combination with LXML (Behnel et al., 2005) to extract all the text of the U. S. Code. We recursively traverse the links from Titles, to chapters, to subchapters and so on, until the final section page is reached, at which point the text for that section is assigned to the corresponding section number in our database. We thus have a queryable database for each Title of the U. S. Code, to allow efficient retrieval of laws, statutes and definition text for each section.



United States Court of Appeals for the Ninth Circuit

STEPHANIE R. FAUSNACHT, *Plaintiff-Appellant* v. AMERICAN BLUE RIBBONS HOLDINGS LLC, *Defendant-Appellee*, (2017)
No. 16-16033

Argued and Submitted April 20, 2017

... former servers and bartenders who alleged that their employers improperly claimed their tips as a credit toward the required minimum wage.
..

The Fair Labor Standards Act of 1938 (FLSA) generally requires employers to pay a cash wage of $7.25 per hour to their employees. 29 U.S.C. § 206(a)(1)(c). But where an "employee engages in an occupation in which he customarily and regularly receives more than $30 a month in tips ..
..

We first consider our framework for examining agency interpretations of statutes and regulations. If a statute is clear, then we give effect to the unambiguously expressed intent of Congress. *Chevron, U.S.A., Inc.* v. *Nat. Res. Def. Council, Inc.*, 467 U.S. 837, 84243 (1984).
..

Figure 2: Example of proceedings of a sample Ninth Circuit District Court case. The above case includes a Chevron citation and was a case of statutory misinterpretation.

3.3.2 Court Cases

There exist several websites that host legal resources e.g., codification of laws, proceedings of court cases, legal dictionaries etc. However, nearly all of these restrict access to the general public and some that allow viewing of documents restrict extensive automated scraping e.g., using the aforementioned web scraping tools and packages. We consult two websites that allow retrieval of cases for Supreme Court[3] and Circuit Court[4] cases respectively.

These websites only contained cases in PDF form, so we first batch downloaded PDFs and con-

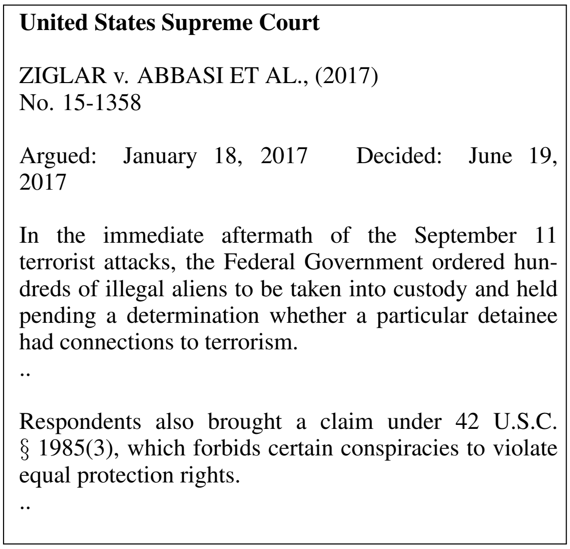
[2] https://www.law.cornell.edu/uscode/text

[3] http://caselaw.findlaw.com/court/us-supreme-court

[4] https://www.gpo.gov/fdsys/

Description	**Type**	**Detail**	**Examples**
Type of Court	Class	12 Court Classes	*Supreme Court* **or** *First Circuit* **or** *Ninth Circuit*
Year	Class	Years 2010-2017	2013 **or** 2015
Titles	List	54 Titles	[42, 8, 7] **or** [26, 21, 3, 8]
Sections	List	53061 Sections	[42 § 1397, 42 § 1340] **or** [15 § 1514(a)]
Chevron Sections	List	53061 Sections	[ ] **or** [8 § 1101(a)]
Granular Annotations	Sequences	Variable Length Snippets	*applicable increment* **or** *crimes involving moral turpitude*
Chevron	Class	2 Classes	True **or** False
Parties Involved	List	-Entity Names	[*Teva Pharmaceuticals U.S.A Inc., Sandoz*] **or** [*Bush, Gore*]

Table 2: Table shows the type of annotation for each case. The fourth column shows examples, where two possible case examples are separated by an **or** clause.



United States Supreme Court

ZIGLAR v. ABBASI ET AL., (2017)
No. 15-1358

Argued: January 18, 2017 Decided: June 19, 2017

In the immediate aftermath of the September 11 terrorist attacks, the Federal Government ordered hundreds of illegal aliens to be taken into custody and held pending a determination whether a particular detainee had connections to terrorism.
..

Respondents also brought a claim under 42 U.S.C. § 1985(3), which forbids certain conspiracies to violate equal protection rights.
..

Figure 3: Example of proceedings of a sample Supreme Court case. The above case does not include a Chevron citation.

verted to text, to create a dataset of court cases from years 2010-2017 for the Supreme Court and 11 Circuit Courts of Appeal.

Our annotations for the corpus of court cases is automated i.e., we extract section citations by searching for an identifying pattern e.g., **x** *U. S. C.* § **y**, where *x* corresponds to the Title of the U. S. Code and *y* corresponds to the section within that Title. We annotate for Chevron deference by searching for the identifying citation i.e. *Chevron, U.S.A., Inc. v. Nat. Res. Def. Council, Inc., (1984)* that every case is required to cite.

4 Tasks

4.1 Identifying Potentially Complex Words

The U.S. Code gives us a comprehensive list of the terms that are required by law, to be defined. Some of the words are defined multiple times, and this may be because they relate to multiple concepts and impart different meaning in different contexts. Here we seek to identify such words by training a classifier to recognise complex legal definition words, as opposed to other words used in general legal literature. Our aim is to then generalise this and predict e.g., in news articles and general literature, words and concepts that impart similarly complex and varied meaning.

We consider three classes of words from the text of the U. S. Code i.e., words that comprise legal definition terms, words contained in the gloss of the definition terms and all other words used in the U. S. Code. We construct these non-intersecting sets of 5727, 8488 and 21254 words respectively. In order to use pre-trained embeddings for classification, we remove OOV words for each of three types of embeddings i.e., GoogleNews (Mikolov et al., 2013), ConceptNet (Havasi et al., 2007) and Legal (embedding representations trained over our corpus). Considering only words that intersect the vocabulary of the pretrained embeddings, leaves us with 5513, 7964 and 15626 words for the three classes respectively.

	ConceptNet		**GoogleNews**		**Legal**	
Model	Train	Test	Train	Test	Train	Test
CNN	90.1	56.4	91.1	56.7	90.2	57.2
CNN-Char	94.9	55.5	86.2	55.4	71.5	56
SVM	-	59.9	-	52.3	-	51.2
Random	-	33.3	-	33.3	-	33.3

Table 3: 3-class accuracy on the train and test sets for neural models (CNN and CNN-Char), and on the test set for Linear SVM.

We compare classification results for 3 models; a one-layer CNN with character embeddings, a one-layer CNN without character embeddings and a Linear SVM model. The CNN architecture is similar to that of (Kim, 2014) and we do not optimise hyperparamaters for our classification task. For the SVM model, we use the implementation in (Pedregosa et al., 2011). For each model, we compare results across the three pre-trained word embeddings that we use as input to the model.

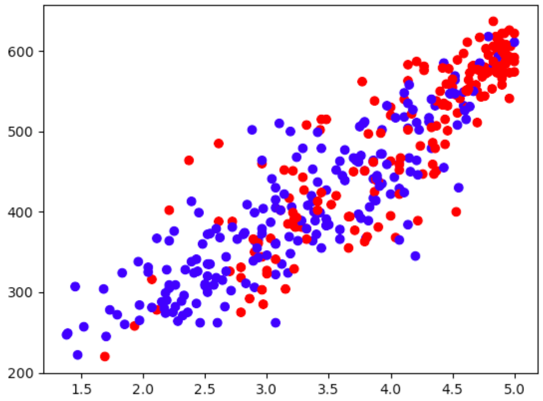
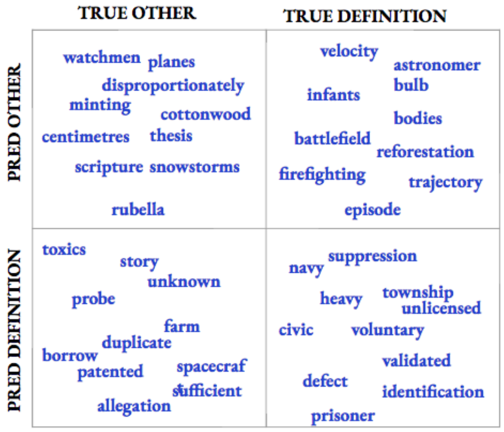


Figure 4: Figure shows a scatter-plot of classified words by the best-performing model by mapping them to MRC and Concreteness ratings. The y-axis shows the MRC scale from 200-600, while the x-axis shows Concreteness scale from 1-5. Blue data points correspond to words classified as definitions and red points correspond to words classified in the 2 other class i.e., words not requiring definition.



	TRUE OTHER	TRUE DEFINITION
PRED OTHER	watchmen planes disproportionately minting cottonwood centimetres thesis scripture snowstorms rubella	velocity astronomer infants bulb bodies battlefield reforestation firefighting trajectory episode
PRED DEFINITION	toxics story unknown probe farm duplicate borrow patented spacecraf sufficient allegation	suppression navy heavy township unlicensed civic voluntary validated defect identification prisoner

Figure 5: Figure shows a confusion matrix of classified words, that show samples of words classified e.g., true positives, false positives etc.

We compare the models' classification to standard ratings of vagueness, concreteness and ambiguity e.g., from the MRC Psycholinguistic Database (Coltheart, 1981) and Concreteness Ratings from (Brysbaert et al., 2014) to ascertain vagueness and ambiguity of words classified.

Figure 5 shows a scatter-plot of classified words for which we have coverage in both datasets. The x-axis shows Concreteness ratings on a scale of 0-5 while the y-axis shows MRC ratings for concreteness, on a scale of 200-600. We collapse the remaining two classes i.e., non-definition terms, to observe the difference in words that require definition by law, as compared to all other words. We observe that words not classified as legal definitions have higher concreteness scores on both scales and tend to cluster away from the origin. We report the mean Concreteness Database ratings as 3.9 and 3.2 and MRC Concreteness ratings as 476 and 401 for the red and blue classified data points. Thus, even with primitive baseline models, training on the set of legal definitions, gives a measure of word-level ambiguity and is likely to yield better results when incorporated into more sophisticated models.

4.2 Determining Importance of Cases

Here we address the difference in the nature of cases taken up by the Supreme Court as opposed to the cases that go to the Circuit Courts. Only a subset of the cases that are sent to the Supreme Court are taken up for judicial ruling. The precise reasoning behind this selection i.e., why only certain cases are chosen by the Supreme Court is unknown.

Our goal here is two-fold; we want to identify cases that go to the Supreme Court to ascertain not only the importance of an issue or the topical nature of the case, but also the need for superior authority or jurisdiction. When higher interpretative powers are required, these are often instances that require more reasoning for statutory interpretation by authorities. We thus seek to characterise the types of cases that go to different courts, to uncover possible instances of ambiguity.

Our positive and negative classes consist of Supreme and Circuit Court cases respectively, and for this task we filter cases that do not cite Titles and sections of the U. S. Code. There are still significantly more Circuit Cases than Supreme Court Cases and in order to address the class imbalance, we run 10 classifiers on samples of the data as outlined in the statistics, instead of losing important data by downsampling (Li and Nenkova, 2014) and report average results of the different classifiers on the test set. Our training set consists of cases in years 2010-2016, while we test on the most recent cases in 2017, thus testing how well the models do by learning from the past and generalising to the present.

ALL DATA	**Positive Class**	**Negative Class**
2010-2016	663	13505
2017	83	804
DOWNSAMPLE	**Positive Class**	**Negative Class**
2010-2016	464	6758
2017	58	548
TRAIN DATA	**Positive Class**	**Negative Class**
2010-2016	464	464
2017	58	548

Table 4: Tables shows statistics of the data we use for case-level classification. The top-most table shows number of cases in the entire dataset. The second table shows number of cases after filtering cases that do not reference the U. S. Code. The third table shows number of cases we sample for each set, therefore using all data in the second table.

The standard baseline models ascertain the representativeness of U. S. Code Titles and Sections, by considering only Title and section features. This is therefore a comparison of how much information we gain from adding more granularity i.e. if specifying specific sections that are contested within each Title gives us additional information about the nature of the case.

Model	**Features**	**Prec.**	**Recl.**	**F1**	**Acc.**
CNN	Titles	15.9	53.4	24.5	68.6
SVM		13.8	68.9	23.0	55.9
CNN	Titles+Secs	23.1	72.4	35.0	74.25
SVM		22.7	77.5	35.1	72.6
Baseline		9.6	100.0	17.5	9.6

Table 5: Precision, recall and accuracy of classification at the case (document) level. The first representation corresponds to considering only Title features. The second corresponds to considering both Titles and Sections. The last row is a baseline that classifies all cases as positive.

Table 9 shows the results of simple baseline models with different sets of features. While these primitive models do not predict with high evaluation metrics we gain useful insight from the error analysis about the importance of specific sections that are contested the most and the types of cases that are misclassified; both potential instances that possibly cause ambiguity.

We consider a confusion matrix of predictions and track the sections that the classified cases refer to the most e.g., for all false positive cases, we perform a comparison of the sections cited by populating a density matrix for Titles of the U. S. Code. We observe that (–)% of the sections in the Supreme Court class have definition terms, highlighting the importance of word-level complexity or ambiguity.

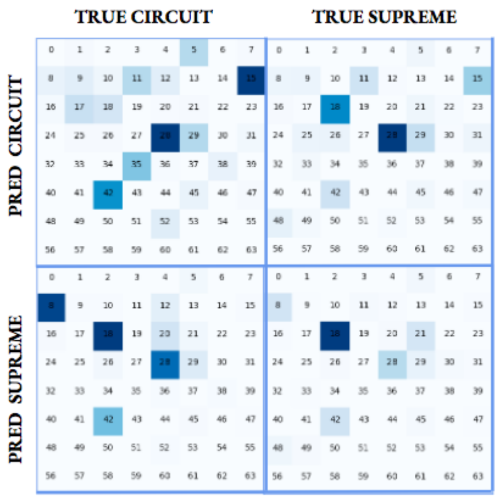


Figure 6: Confusion matrix with title densities for model classifications. The numbers indicate U. S. Code Title numbers e.g., block 8 corresponds to Title 8 - Aliens and Nationality and darker shades correspond to higher density of cited Titles. The matrix is padded with zeros to allow representation.

5 Conclusion

The legal domain serves as the ideal testing area for new models or theories of ambiguity and contextual interpretation. We have introduced a dataset to help further research in an area of natural language understanding that has been extensively studied in several different ways, but has never been applied to the situations and examples introduced by the data here. The baseline models for each of the tasks outlined above are standard architectures with minimal feature engineering and the results indicate that we can learn a considerable amount from the data as is. We hope that future work with lexicalised and unlexicalised feature-based models and neural networks with architectures that pay attention to critical identifying aspects will give considerable gain to help determine ambiguity and underspecification. We hope that this corpus serves as valuable evaluation and training data to learn to predict and learn instances of ambiguity and pragmatic reasoning and help further research in this domain.

References

Daniel Bailey, Yuliya Lierler, and Benjamin Susman. 2015. Prepositional phrase attachment problem revisited: how verbnet can help. In *Proceedings of the 11th International Conference on Computational Semantics*. pages 12–22.

Stefan Behnel, Martijn Faassen, and Ian Bicking. 2005. lxml: Xml and html with python.

Leon Bergen, Noah Goodman, and Roger Levy. 2012. That's what she (could have) said: How alternative utterances affect language use. In *Proceedings of the Annual Meeting of the Cognitive Science Society*. volume 34.

Leon Bergen, Roger Levy, and Noah Goodman. 2016. Pragmatic reasoning through semantic inference. *Semantics and Pragmatics* 9.

Curtis A Bradley. 2000. Chevron deference and foreign affairs. *Virginia Law Review* pages 649–726.

Eric Brill and Philip Resnik. 1994. A rule-based approach to prepositional phrase attachment disambiguation. In *Proceedings of the 15th conference on Computational linguistics-Volume 2*. Association for Computational Linguistics, pages 1198–1204.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods* 46(3):904–911.

Michael Collins and James Brooks. 1995. Prepositional phrase attachment through a backed-off model. *arXiv preprint cmp-lg/9506021* .

Max Coltheart. 1981. The mrc psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A* 33(4):497–505.

Wayne Davis. 2014. Implicature. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University. Fall 2014 edition.

Julie Dickson. 2016. Interpretation and coherence in legal reasoning. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University. Winter 2016 edition.

Timothy Endicott. 2016. Law and language. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University. Summer 2016 edition.

H Paul Grice. 1975. Logic and conversation. *1975* pages 41–58.

H Paul Grice. 1981. Presupposition and conversational implicature. *Radical pragmatics* 183.

Catherine Havasi, Robert Speer, and Jason Alonso. 2007. Conceptnet 3: a flexible, multilingual semantic network for common sense knowledge. In *Recent advances in natural language processing*. Citeseer, pages 27–29.

Donald Hindle and Mats Rooth. 1993. Structural ambiguity and lexical relations. *Computational linguistics* 19(1):103–120.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882* .

Junyi Jessy Li and Ani Nenkova. 2014. Addressing class imbalance for improved recognition of implicit discourse relations. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*. pages 142–150.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. pages 3111–3119.

George Miller and Christiane Fellbaum. 1998. Wordnet: An electronic lexical database.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research* 12(Oct):2825–2830.

Christopher Potts, Daniel Lassiter, Roger Levy, and Michael C Frank. 2016. Embedded implicatures as pragmatic inferences under compositional lexical uncertainty. *Journal of Semantics* 33(4):755–802.

Adwait Ratnaparkhi, Jeff Reynar, and Salim Roukos. 1994. A maximum entropy model for prepositional phrase attachment. In *Proceedings of the workshop on Human Language Technology*. Association for Computational Linguistics, pages 250–255.

Leonard Richardson. 2013. Beautiful soup. *Crummy: The Site* .

Benjamin Russell. 2012. Probabilistic reasoning and the computation of scalar implicatures. *Unpublished doctoral dissertation, Brown University* .

Vivek Srikumar and Dan Roth. 2013. Modeling semantic relations expressed by prepositions. *Transactions of the Association of Computational Linguistics* 1:231–242.

Jiri Stetina and Makoto Nagao. 1997. Corpus based pp attachment ambiguity resolution with a semantic dictionary. In *Fifth Workshop on Very Large Corpora*.

Jakub Zavrel, Walter Daelemans, and Jorn Veenstra. 1997. Resolving pp attachment ambiguities with memory-based learning. *CoNLL97: Computational Natural Language Learning* .