Understanding Legal Meaning Through Word Embeddings

September 17, 2019

Abstract

For judges and legal scholars, the quest for meaning and the identification of methods appropriate for understanding word meaning animates volumes upon volumes of debate. Likewise, social scientists interested in studying the law have increasingly recognized the variations in choices over words as important barometers for understanding the law. In this paper, I suggest recent advances in computational linguistics – notably, the efficient estimation of distributed representations of word meanings, or word embeddings – offer a potentially transformative avenue by which to assess the strategy, choice, and impact of judicial language. Utilizing a corpus of more than one million federal and state appellate court decisions, I estimate word embeddings for the more than 400,000 most common words found in legal opinions. In a series of simple illustrations, I demonstrate the value of this treatment of word meaning for new avenues of law and politics research.

Word Count: 7,778
Language is central to the study of the law. For everyone from judges to citizens, understandings of and appeals to the law are shaped by our interpretations of language. Through carefully crafted judicial opinions, courts explicitly construct, present, and defend interpretations, with those choices intended to shape legal development and subsequent interpretation. Thus, interpretation of word meaning lies at the heart of the judicial enterprise. From the perspective of some, meaning is entrenched, defined, and discoverable. For others, the meaning of a word is nothing of the sort. Indeed, the very ambiguity of language generates the necessity for judicial interpretation. The absence of definitive conclusions is conspicuous in the law, where “when words are clear, there is not even likely to be litigation” (Easterbrook, 1984, 87).

In many respects, this division of interpretive opinion in the legal sphere is analogous to the state of computational linguistics research. Specifically, standard text-as-data approaches treat terms as monosemous; that is, having only a single meaning. In a term-document matrix, for instance, a term appears only once, and meaning is codified through counts of occurrences. Yet this treatment was always a matter of pragmatism rather than a realistic appraisal of language. Linguists recognize that most words are polysemous, containing many meanings. Capturing multiple meanings computationally, however, was in large measure intractable until recently. To wit, the idea of distributed representations of meaning was first suggested at least as far back as Rumelhart, Hintont and Williams (1986), and the foundations for statistical modeling of that distribution were laid in 2003 (Bengio et al., 2003). Yet only in 2013 were robust methods developed for the more general estimation of the distributed representations (Mikolov, Chen, Corrado and Dean, 2013; Mikolov, Sutskever, Chen, Corrado and Dean, 2013). The result has been transfor-

---

Note that analyses using n-gram type approaches will disambiguate the senses to some degree.
mative for computation linguistics and research implicating language more generally.² Though at heart relying on counts themselves, these distributed representations – each word represented by a vector of weights over some number of dimensions, thus termed word embeddings or word vectors – reflect the sophistication of linguistic meaning at a level not generally observed in the monosemous, word count based approaches.

This distributed representation of word meaning matches closely with the intuition of the justices in Towne v. Eisner (1918). There, the Court states “[a] word is not a crystal, transparent and unchanged, it is the skin of a living thought and may vary greatly in color and content according to the circumstances and the time in which it is used.”

The variety of weights across the word suggests a host of potential meanings, from which the judge must derive the specific meaning of the term as utilized. In so doing, the judge might consider any number of other concepts, including other words used nearby, related terms not used, the speakers own proclivity for using that sort of language, and so on. Many volumes have been written on how one should go about this task. That debate is sure to continue. Yet for researchers, word embeddings – distributed word representations – offer a method by which to systematically analyze these choices in interpretation, and the implications of these choices.

In this paper, I introduce a massive, original dataset of federal and state appellate court decisions appropriate for training word embeddings in the legal context. This corpus of U.S. appellate case law – totaling nearly 3 billion tokens in all – is comparable to the large web corpora where approaches for estimating word embeddings have been developed. Using the estimated embeddings, I provide three illustrations of areas in which the trained embeddings offer promising avenues for future research. First, I demonstrate the utility of word embeddings for studying the relationships between terms, a potential

avenue for future work on judicial interpretation and strategic judicial behavior. Second, I demonstrate how the relationships between terms can be generalized to the study of relationships between concepts. I illustrate by expanding recent work in computational social science for the study implicit gender bias in text, providing evidence of it’s manifestation in judicial opinions. Third, and finally, I demonstrate how word embeddings can be utilized to expand our ability to use computational text analysis to create new and robust measures of important judicial constructs. Here, I illustrate by generating domain-specific dictionaries of ideological dimensions which can be used in the analysis of the ideological dimensions of individual judicial opinions.

These initial analyses offer a glimpse of the promise of word embeddings for research on law, society, and politics. Language is indeterminate and ambiguous. The study and practice of the law, then, turns on deriving meaning, with a plethora of approaches debated. The intent of this paper is not to suggest an avenue by which to resolve the debate over interpreting meaning. That discussion is already taking place among legal scholars as they consider the value of corpus linguistics. However, the derivation of meaning through word vectors does offer an important and potentially transformative method for empirical legal research advancing our understanding of questions of import to social scientists and legal scholars.

What’s in a (legal) word?

As Easterbrook notes, “[w]ords are tricky to start with; words in legal documents are worse” (Easterbrook, 1984, 90). Consider former Supreme Court justice Antonin Scalia’s

---

declaration in dissent in *King v. Burwell* (2015), that “[w]ords no longer have meaning if an Exchange that is not established by a State is ‘established by the State.’” Scalia’s lamentation was in dissent, however, with six justices instead holding that the term ‘State’ must be understood within the broader context of the Affordable Care Act. What seemed clear in isolation, that is, was less clear when understood within context. Similar indeterminacies of language are manifest in law.

This is well-reflected in the state of research on law and politics. Consider the classic debate surrounding models of judicial behavior. Here, the attitudinal model specifically leverages the ambiguity of language as one reason to suspect justices on the Supreme Court are free to vote their policy preferences (Segal and Spaeth, 2002). Likewise, research invoking the strategic model has proposed, among other things, that judges strategically manipulate language to avoid review (Staton and Vanberg, 2008; Hinkle et al., 2012; Owens, Wedeking and Wohlfarth, 2013; Black et al., 2016) and that opinion drafting is influenced by panel effects (Rice, 2016; Hinkle, 2017). This scholarship has importantly moved the field forward by considering behavior beyond the vote outcome decision of justices or the Court. Understanding how language is employed – and in particular differentially employed – thus offers important avenues to address fundamental questions in the study of the law. What words does a justice use to address a controversy, and what words are theoretically related but never invoked? How do these patterns manifest across a host of variables of interest? When and how do legal concepts, constructions, and understandings arise and become embedded in the law? How does the understanding of key legal concepts change over time in the language of the law? These are but a few of the questions that can be addressed by beginning to deconstruct legal meaning using new approaches from computational linguistics.
Computing Meaning

Scalia’s view of meaning suggests a single, discoverable definition. The quest for that meaning is of course at times difficult. Indeed, Scalia’s own search frequently led him to dictionaries as he cited among others the *Random House College Dictionary*, Noah Webster’s *American Dictionary* (1828), Samuel Johnson’s *Dictionary of the English Language* (1773), and Timothy Cunningham’s *A New and Complete Law Dictionary* (1771). Yet once that meaning was identified as Scalia deemed correct, it was treated as definitive; in the words of one critic, Scalia’s “approach [was always] to reach for a dictionary; find, in one edition or other, a definition that drives toward his predetermined decision; and express, eyes wide with disbelief, utter amazement that anyone could even think of seeing it any other way.”

Such a treatment is consistent with the bulk of computational text analysis, where counts of raw word occurrences are used to capture meaning. In using counts, a word is treated as monosemous, with the appearance of a word treated as having a singular (or, in computational terms, “one-hot”) meaning. Summing over occurrences (then potentially adding some form of document-frequency weighting) yields a measure of the construct of interest. Looking back, Scalia’s lamentation regarding the meaning of “State” is instructive; like Scalia, the approach treats “State” as captured by a single definition or meaning which is identical across usages, such that each usage indicates that identical definition.

Recent work in computational linguistics instead comes closer to more flexible constructions of meaning in language, constructions which incorporate multiple potential meanings. Though one observes only the occurrence of the word, that occurrence instead

---


indicates a host of associated concepts. Computationally, this means that rather than the identity matrix of the “one-hot” formulation – where each word indicates only that no other word occurred but the indicated word – in this new (or word vector approach) each word is a distribution of weights across some number of dimensions. In re-visiting Scalia’s quote, the word vector approach suggests that “State” may have some loading on the dimension associated with the political entities composing the United States, but it may also have a loading on a dimension related to the broad concept of “governmental polity” as well as a loading on a dimension related to “speech.” Moreover, each word or token in the corpus has a distribution across the same dimensions. The word, then, is represented by the distribution of weights across the vector, while the precise intended meaning could be the general distribution, or could instead be related to one of those individual dimensions. Importantly, the meaning of a term is defined by the distribution across weights.

This understanding of linguistic meaning – which maps much closer to language as understood by linguists – is best seen in recent computational work on word embedding (Mikolov, Chen, Corrado and Dean, 2013; Mikolov, Sutskever, Chen, Corrado and Dean, 2013). Though the idea of distributed meaning has been around for some time in linguistics, the computational framework for estimating the distribution is quite recent. The fundamental underlying idea permitting the estimation of the distribution is that words appearing in the same context are semantically related. In other words, the estimation relies on the fact that similar words appear in similar contexts, and can be predicted by similar context words and weightings.⁶

⁶Though there are numerous approaches to constructing the vector space, the original approach is through neural networks. One estimates word vectors by using the context of each word in the corpus, then for each word one obtains a layer – a distribution of weights across k dimensions where k is defined by the researcher – which best predicts the context
To estimate the word embeddings, I utilize GloVe (Pennington, Socher and Manning, 2014). GloVe leverages the statistical information available in word co-occurrences. Begin by specifying a window of size $x$, then count all of the times in a corpus two words co-occur within that window across the corpus. Doing so for all terms across the corpus yield a term co-occurrence matrix. Then, factorize the log of the co-occurrence matrix but incorporate some weighting functions to address very rare or very frequent co-occurrences; GloVe does so by including a weighted least squares regression model. The result of the factorization of the log of the co-occurrence matrix with the weighting is – for each unique word in the corpus – a vector of $d$ dimensions (with $d$ specified by the user) where each $d_i$ represents the loading of that term on that dimension.

Operations on these vectors reveal a variety of substantively meaningful relationships, each of which demonstrates the value of word embeddings for uncovering meaning. For instance, one classic example is obtaining the vector from estimating $\text{vec(} \text{king} \text{)} - \text{vec(} \text{man} \text{)} + \text{vec(} \text{woman} \text{)}$. In other words, take the vector of estimated weights across “king” then subtract the vector of weights associated with “man” and add the vector of weights for “woman.” Comparing the resultant vector to all other estimated embeddings in the corpus, one regularly finds the resulting vector is generally most similar to $\text{vec(} \text{queen} \text{)}$, the vector of weights associated with “queen.”

These word embeddings – and the meaning embedded within the distribution of weights – have been employed extensively in computer science applications. But for research on law and politics, they provide an avenue by which to begin addressing a variety of classic questions. I turn to three such illustrations now, then discuss other areas of potential utility in the final section.

words (skip-gram formulations) or the target word (continuous bag of words). For every word, one thus has a vector of weights of length $k$. 
Data

The data come from Justia, a free and open online repository of legal information. The website includes broad coverage of judicial opinions from both state and federal courts. From Justia, I acquired all U.S. Supreme Court opinions, all available U.S. courts of appeals opinions, and all available state courts of last resort opinions. Table 1 offers an overview of coverage. As would be expected, between them the U.S. courts of appeals and state courts of last resort comprise the overwhelming bulk of activity. Coverage of the courts of appeals covers nearly 600,000 opinions and dates as far back as 1935 and as recent as 2007. Nearly the entire history of the U.S. Supreme Court is included, with a corpus that consists of 42,691 opinions from 1793 through 2012. Finally, within the area of state courts, Oklahoma (40,814 from 1890-2015) and Pennsylvania (33,740 from 1950 to 2015) have the highest volume of opinions. The number for Pennsylvania is inflated as separate opinions (dissenting, concurring, etc.) are included as separate opinion files. Three states – Maryland, New York, and Ohio – are excluded, as the retrieved opinion sets for each proved unreliable.

Taken together, the corpus contains 1,134,383 judicial opinions covering 47 states, the U.S. courts of appeals, and the U.S. Supreme Court over an extended period of the nation’s history. More importantly for the estimation of distributed representations, the corpus features 1,630,175 unique tokens, and more than 2.950 billion tokens in all. As points of comparison, recent work in computational linguistics evaluating methods for extracting word embeddings utilized “a 2010 Wikipedia dump with 1 billion tokens; a 2014 Wikipedia dump with 1.6 billion tokens; Gigaword 5 which has 4.3 billion tokens;”

Note that for estimating word vectors the separate opinion files are essentially irrelevant. This raises problems only for downstream approaches that would look to identify, say, differences in behavior across these two opinions.
<table>
<thead>
<tr>
<th>Court</th>
<th>Coverage</th>
<th>Number</th>
<th>Court</th>
<th>Coverage</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>U.S. Supreme Court</td>
<td>1793-2012</td>
<td>42,691</td>
<td>Missouri</td>
<td>1950-2015</td>
<td>8,137</td>
</tr>
<tr>
<td>Florida</td>
<td>1950-2015</td>
<td>14,674</td>
<td>Ohio</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Georgia</td>
<td>1950-2015</td>
<td>18,055</td>
<td>Oklahoma</td>
<td>1890-2015</td>
<td>40,814</td>
</tr>
<tr>
<td>Hawaii</td>
<td>1957-2015</td>
<td>8,957</td>
<td>Oregon</td>
<td>1949-2015</td>
<td>8,597</td>
</tr>
<tr>
<td>Maine</td>
<td>1950-2015</td>
<td>8,448</td>
<td>Vermont</td>
<td>1950-2015</td>
<td>6,675</td>
</tr>
<tr>
<td>Maryland</td>
<td>–</td>
<td>–</td>
<td>Virginia</td>
<td>1925-2015</td>
<td>20,906</td>
</tr>
</tbody>
</table>

Table 1: Opinions Included in the Corpus By Court
the combination Gigaword5 + Wikipedia2014, which has 6 billion tokens; and 42 billion tokens of web data, from Common Crawl” (Pennington, Socher and Manning, 2014, 7). Taken together, the corpus of judicial opinions compares favorably to the work in computational linguistics employing these tools, with two of the common corpora actually featuring fewer tokens than the compiled judicial corpus.

To estimate the word vectors, I retain only tokens appearing more than 10 times in the corpus; doing so reduces the unique tokens to 422,220, but the total number of tokens in the corpus is only reduced to 2.947 billion. Again, it is worth emphasizing the scale of the data is comparable to that utilized in prominent computer science applications, including the original GloVe article. Capitalization and punctuation were removed, and I constructed a matrix of co-occurrence counts for words appearing within a window of 20 words. I utilize GloVe (Pennington, Socher and Manning, 2014), and estimate 300-dimensional vectors, training the model using a 50-iteration AdaGrad (Duchi, Hazan and Singer, 2011). I turn now to a series of illustrations of the value of these legally-derived word vectors for empirical legal research.

Illustration One: Meaning in Corpus

With the vectors I begin by looking at whether or not the recovered word vectors capture substantive meaning in ways that match with legal and scholarly understandings of terms. Employing the word vectors offers an avenue by which to explore related concepts, or the semantic relatedness of words. In the words of Budanitsky and Hirst (2006),

[S]imilar entities are semantically related by virtue of their similarity (bank - trust company), but dissimilar entities may also be semantically related by lexical relationships such as meronymy (car - wheel) and antonymy (hot - cold)
or just by any kind of functional relationship or frequent association (pencil-paper, penguin-Antartica, rain-food).

Thus, where one observes a specific word, then a related word should be more likely to appear than an unrelated word. To start, take a basic legal concept which tracks across federal and state courts, justiciability. In the top row of Table 2, I present the most similar vectors to vec(justiciability), computed using cosine similarity. Importantly, each concept is central to the defined vector. As a comparison, Wex Legal Dictionary\(^8\) defines “Justiciability” as:

To be **justiciable**, the court must not be offering an advisory opinion, the plaintiff must have standing, and the issues but must be **ripe** but neither **moot** nor violative of the political question doctrine.

The words in bold are – save for some variation in suffix – precisely the terms which are considered central to the definition of justiciability. In fact, the only component missing is the consideration of advisory opinions, a consideration that rarely actually confronts the courts due to the clarity of the law in the area.

Another way to examine the validity of the vectors at capturing the meaning of concepts is to work in reverse. That is, one can consider taking separate vectors related to concepts, combining those vectors, and examining the most similar vector. Consider the definition of “murder” as “killing with malice aforethought or intent.” Thus, one could construct a vector from \( \text{vec(killing)} + \text{vec(malice)} + \text{vec(intent)} \). The most similar vectors to that vector within the corpus are presented in the the second row of Table 2; importantly, the most similar is “kill”, followed by “murder”, and then a variety

\(^8\)The Cornell Legal Information Institute’s online legal dictionary available at https://www.law.cornell.edu/wex/justiciability
Table 2: Examples of Similarity: The above presents the most similar vectors (right column) relative to the vector defined in the left column. Similarity is computed as the cosine similarity of the vectors.

The third and fourth rows explore the concept of race and racial discrimination. Beginning with the third row, one sees that the “discrimination” vector is highly related to the term “racial” as well as “ethnic” and “gender.” That is, the term “racial” reflects a more general concept of discrimination that matches other forms of discrimination. Subtracting one other form (vec(racial) - vec(gender)) yields a dimension much more closely related to forms and assessments of racial discrimination, including a variety of different linguistic forms of segregation.

Finally, consider framing effects. In the fifth and sixth rows, I include the most similar vectors to vec(abortion) + vec(liberty) and vec(abortion) + vec(life), respectively. Though coarse, the first vector is intended to carry an interpretation which tracks more closely to substantive due process claims related to a woman’s freedom to choose, while the second is intended to carry an interpretation which tracks more closely to the characterization of pro-life advocates. The most related words differ across the vectors in substantively interesting and important ways. In the case of “liberty”, the word “fundamental” is identified as highly similar, while in the case of “life”, the word “death” and the “health” of the mother stick out as highly related. The different substantive em-
phases reflected by the choices suggest different words one is likely to find employed in an opinion.

While these relationships may be employed in myriad ways, one avenue is in the area of using corpus linguistics to address disputes over substantive meaning. As an example, in King v. Burwell, Scalia particularly laments the treatment of the word “State” as potentially including the federal government. Scalia, that is, treats the meaning as exclusive. However, by cosine similarity the terms are closely related; “State” is – across this corpus of opinions – the tenth most closely related term to “Federal”, while inversely “Federal” is the seventh most closely related term to “State” across the corpus. That is, the two distributed representations of the two terms are closely intertwined, such that the juxtaposition is perhaps not entirely without merit. Moving from meaning within the context of the courts, one can easily see the utility of such an approach for capturing usages within – for instance – Congress, and the relative interchangeability of the terms, or the ignorability of particular word choices.

**Illustration Two: Measuring Concept Associations**

Beyond the relations between individual words, one can also utilize the trained word embeddings to measure the associations between concepts. Consider sets of words that all tap into dimensions of a broader concept; race, for instance, broken down into sets of words that tap into African American or Caucasian identities, respectively. Or, consider legal interpretation, with sets of words that tap into different modes of statutory interpretation, or different sets of words that tap into different approaches to a more general concept like federalism. Word embeddings offer an important opportunity to measure the association between these different dimensions of general concepts as reflected by sets of words.
A recent article appearing in *Science* offers an important, theoretically and empirically rigorous approach to understanding and measuring the association between concepts using word embeddings. Specifically, Caliskan, Bryson and Narayanan (2017) introduce the Word Embedding Association Test (WEAT) for testing for associations between key concepts in the context of large, web-based corpora. They leverage the WEAT to, among other things, identify implicit racial and gender bias in web corpora including Wikipedia and a general crawl of the internet. To do so, they identify sets of terms related to particular concepts (i.e., a set of stereotypically female or stereotypically male names, a set of negative terms, etc.) and test for the relative associations between the concepts.

To illustrate, I turn to consideration of gender bias in the law, echoing recent work by Rice, Rhodes and Nteta (2019) on racial bias. A rich academic literature establishes the prevalence of gender bias throughout different areas relevant to legal studies, including settings as diverse as ABA qualification ratings (Sen, 2014), nomination hearings (Boyd, Collins and Ringhand, 2018), and decision outcomes (Boyd, Epstein and Martin, 2010; Gleason, Jones and McBean, 2017). Yet harder to detect is the extent to which it implicitly manifests in the language of the law, judicial opinions (Tiller and Cross, 2006). The potential manifestation of gender bias in judicial opinions would have particularly pernicious effects, as it would extend and reinforce systemic bias against the out-group among future interpreters of the law.

Outside of the courts, implicit gender bias has been captured in an experimental setting through implicit association tests wherein respondents are prompted to quickly sort male and female names or words as well as areas of potential bias. In Nosek, Banaji and Greenwald (2002), for instance, the authors find evidence of implicit cultural biases over how appropriate it is for women to have careers, and conversely for men to take on domestic roles. Likewise, Nosek, Banaji and Greenwald (2002) also find that male terms are more prevalently associated with concepts related to mathematics, while female terms are
more frequently associated with arts terms. Extending this work, in their recent *Science* article, Caliskan, Bryson and Narayanan (2017) propose measuring implicit gender bias in text by using word embeddings in association with word sets taken directly from these implicit association tests. In both analyses, they find strong evidence that the trained word embeddings offer evidence of gender bias.

Given the implicit nature of the gender bias noted above, as well as the observation of other forms of gender bias in the legal system, one might reasonably expect to observe implicit gender bias manifest in the language of the law, judicial opinions. Thus, to test whether judicial opinions do contain such implicit biases, I estimate word embedding association tests [WEATs] across a series of comparison groupings using the word embeddings trained above. I treat male/female as the attribute category, and measure associations with a series of different target concepts.\(^9\) To calculate WEAT, I follow Caliskan, Bryson and Narayanan (2017) and compute word-specific association scores as:

\[
s(w, A, B) = \text{mean}_{a \in A} \cos(\bar{w}, \bar{a}) - \text{mean}_{b \in B} \cos(\bar{w}, \bar{b})
\]

(1)

where \(A\) and \(B\) represent words in the target categories (i.e., positive terms, negative terms, etc.), \(\bar{w}\) represents the the word embedding for a name or term \(w\), and \(\cos\) represents cosine similarity of the word \(w\) with a word in the target category. By comparing the relative distance between each individual word in the attribute list and each individual word in the target list, the approach minimizes the influence of relative prevalence. Then, again following Caliskan, Bryson and Narayanan (2017), I derive a test statistic from the associations as:

\[
s(w, X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B)
\]

(2)

\(^9\)The full sets of terms are available in the supplementary material.
where \((X,Y)\) indicates the respective attribute vectors (here, male and female).

The test statistic is particularly valuable in order to understand the significance of the measured associations between concepts in that it permits comparison to a distribution of random alternatives. For each WEAT, I take 10,000 random permutations of the target concepts, randomly shuffling the labels applied to each then re-estimating the test statistic. In doing so, I generate a contextually-appropriate distribution of the test statistic in order to make some inference as to the significance of the association.

The results of the tests for gender bias appear in Figure 1. The first and second rows are for comparisons with particular forms of gender bias – the first, career versus family words; the second, math versus arts words – while the third and fourth rows are for comparisons with negative and positive terms. The left column offers a comparison between male and female names, while the right column offers the comparison between male and female terms. The distribution is derived from 10,000 test statistics estimated from random permutations of the data. The comparison exceeds the 5% level of significance where the black line (the estimated test statistic for the un-permuted data) appears to the left of the vertical, gray dashed line (the 5% significance level as derived from the distribution).

Strikingly, in three of the four areas, the results indicate strong evidence of implicit gender bias. Consider the first row, which compares the average distance between male and female names (or terms) with career-related terms and family-related terms. The observed test statistic is at the extreme left of the distribution, indicating overwhelming evidence that female names or terms are more likely to be associated with family terms in legal opinions than male names or terms, and correspondingly less likely to be associated with career terms in legal opinions than male names or terms.

One may perhaps suggest that this observed bias is reflective of historical realities of differential societal treatments of gender. In *Muller v. Oregon*, for instance, Justice Brewer wrote for the Court in upholding limitations on a woman’s right to contract that “her
Figure 1: Distribution of Word Embedding Association Test Statistics from 10,000 Iteration Permutation Test. These plots provide a distribution of test statistics estimated across random permutations of the target characteristic groupings. The vertical grey dashed line indicates a 5% significance level (one-sided), while the vertical black line indicates the observed test statistic in the real data.
physical structure and a proper discharge of her maternal functions – having in view not merely her own health, but the wellbeing of the race – justify legislation to protect her from the greed, as well as the passion, of man.” Yet the results reported in rows three and four – reporting comparisons to negative concepts as compared to positive concepts – suggest the problem is yet broader. Specifically, in comparison to male names and terms, female names and terms are more frequently associated with negative words and less frequently associated positive words. Again, the evidence is overwhelming, with the test statistic at the farthest reaches of the distribution.

Yet evidence of bias does not extend to each of the explored dimensions. Instead, the second row of Figure 1 reports the results from a comparison of the male/female vectors with concepts relating to math terms and arts terms. This form of bias is particularly unlikely to manifest in judicial opinions, as such discussion is so rare in legal venues. This stands in contrast to more general corpora (like those derived from web crawls or from Wikipedia), where language may incorporate this sort of bias. Indeed, in both names (left column) and terms (right column) analyses, the observed test statistic is almost precisely at the midpoint of the permuted distribution. In all, we have evidence implicit gender bias in judicial opinions manifests in theoretically appropriate ways.

Illustration Three: Leveraging Associations

Finally, beyond measuring associations as the outcome of interest, one can leverage the associations between terms and concepts for downstream approaches to empirical legal studies. As an example, I turn to a prominent and widely-debated area of research on courts, the estimation of ideology. A great host of approaches have been conceived of for measuring the ideology of a justice (e.g., Segal and Cover, 1989; Martin and Quinn, 208 U.S. at 423.)
2002; Lauderdale and Clark, 2012), but less prominent have been efforts to identify the ideological placement of an individual opinion. Recently, Robinson and Swedlow (2018) have suggested one avenue forward in research on ideology would be by shifting analysis away from unidimensional, liberal-conservative understandings of ideology and towards a more multifaceted “cultural theory” of ideology. This “cultural theory” identifies varying emphases on different values – equality, order, liberty – as instead a way of capturing ideological preferences across diverse groups.

This research mimics, in many ways, the growing body of work in political and social psychology in which differing values – or moral foundations – are cast as structuring political ideologies (see, e.g., Haidt and Graham, 2011; Haidt, 2012). Here, five moral foundations – Care/Harm, Fairness/Cheating, Loyalty/Betrayal, Authority/Subversion, and Sanctity/Degradation – and the relative weight a person applies across those dimensions are said to underlie judgments. The underlying idea is that the observed variance in ideology – and the difficulty of overcoming distances between ideologies – is a function of differing emphases on different moral foundations. Evidence in support of the theory has largely been gathered through large-scale surveys (Haidt and Graham, 2011; Haidt, 2012), and particularly analysis of text responses explaining, say, political preferences.

One avenue of development – completed by Haidt and Graham, two of the original proponents of Moral Foundations Theory – has been in the construction of a dictionary of terms which fall into each of the foundations, thereby permitting forms of computational text analysis through software like Linguistic Inquiry and Word Court [LIWC] (Pennebaker, Francis and Booth, 2001; Pennebaker et al., 2007). Though dictionaries carry a host of potential concerns (Grimmer and Stewart, 2013), they offer value in their ease of use.

---

11A sixth moral foundation – Liberty – is regularly hypothesized but with no publicly available dictionary to leverage for the below analyses it is not included here.

12The full dictionary is available http:\www.moralfoundations.org
of deployment. Here, the hope of the researchers is that scholars might employ them to count the relative frequency of terms used in each of the categories within some text, and thus be able to identify the presence or absence of different moral foundations. When a term is included in the dictionary, we would count the occurrence; all other terms in the text would be ignored. Yet the dictionaries are limited in size, and developed in a domain (survey responses) that differs markedly from the measured language of judicial opinions.

To address this and construct more robust, opinion-specific measures of ideology, consider calculating the distance of a term to each of the foundations. Understanding how closely a particular term matches to a foundation provides a weight for each term in the corpus, which helps to address both differences in the loading of each term on the dimension (Hansen et al., 2011), as well as problems with domain relevance.

To that end, I create enhanced Moral Foundations dictionaries for use with legal texts. There are ten total dictionaries provided by the Moral Foundations website, two for each of the five Foundations. Each of the Foundations features what the authors refer to as “Virtue” and “Vice” dictionaries for the relevant Foundation. For instance, in the Authority Foundation, virtue words include “obedient”, “authority”, and “command”, while vice words include “defy”, “lawless”, and “unfaithful.” Using the word vectors previously estimated, I create a series of vectors that are the sum of the word vectors for terms in each Moral Foundation category. Then, for each of the Foundation vectors, I estimate the most similar 200 terms according to cosine similarity. The result are expanded, domain-appropriate dictionaries for each of the Moral Foundations.

How might these be employed? Let us return to King v. Burwell. I acquired the text of both Chief Justice Roberts’ majority opinion and Justice Scalia’s dissenting opinion. For each, I count the number of occurrences of a term in the enhanced dictionary, weighted by

13 Note that I remove a limited set of stop words from the resulting sets.
Figure 2: *Plot of Differences in Moral Foundations Emphasis in Opinions in King v. Burwell.* Values are the difference (Scalia - Roberts) in the relative emphases on different moral foundations. For measurement details, see text.
the cosine similarity scores. I then sum the weighted word counts within each Foundation dictionary, and calculate the proportion each Foundation area comprises of the total amount of attention by that author to all Foundations.

The result is a measure of the relative weight each justice placed on the Moral Foundations. In Figure 2, I plot the difference in these emphases, with more positive values indicating Justice Scalia placed greater emphasis on the indicated foundation, and negative values indicating Chief Justice Roberts placed greater emphasis on the indicated foundation. A number of interesting differences are immediately apparent. Most noticeably, Justice Scalia places far greater emphasis on positive treatments of Authority, and slightly greater emphasis on treatments of Purity. As Sinn and Hayes (2017) document, these two foundations are generally more emphasized by conservatives. On the other hand, the majority opinion by Roberts – who joined by Justice Kennedy and the Court’s four liberal justices in the majority – places greater emphasis on Fairness and Care, two areas which are regularly hypothesized as of greater importance to liberals.

**Discussion and Conclusion**

Though interesting on its own, this is only scratching the surface of the potential value of word embeddings for law and courts research. One could, for instance, look for related words that are not employed, or unrelated words employed, to understand a much finer grain of judicial behavior and strategy. Moreover, partitioning the corpus into separate corpora defined by any number of metrics – perhaps ideological characteristics – offers an opportunity to better understand variation in the terms related to central concepts. Tracking, for instance, dimensions related to federalism across liberal and conservative justices offers an important avenue to understanding when and how the concept is employed towards potentially strategic ends (e.g., Baird and Jacobi, 2009) and how that might shape
future behavior (e.g., Rice, 2014; Cranmer, Rice and Siverson, 2016). Alternatively, one might consider employing similar methods to identify judicial compliance, partitioning datasets into before and after sets that track changes in related terms before and after, say, the Supreme Court intervenes. Similar work is ongoing in computational linguistics (Rosin, Adar and Radinsky, 2017) and holds tremendous promise for law and politics research.

In all, for scholars interested in studying the law, word embeddings hold tremendous promise. In this paper, I have presented three implementations of word vectors as ways to deconstruct meaning in the law. Each opens the door to exciting new avenues for research on law, courts, and politics. As but one example, future work might look to disentangle the meaning of specific terms, or partition corpora in such a way as to identify changes in semantic meaning over time. Or one might look to any of a number of extensions – including some of those discussed above – to estimate ideology.

While the emphasis here has been on estimation of word embeddings within the judicial domain, the utility of moving beyond judicial opinions to other historical documents is perhaps even more exciting. For instance, word embeddings offer an avenue by which to capture historical understandings of terms within particular time frames, and words that might reasonably be used interchangeably with the identified term. For scholars interested in identifying the original meaning of historical terms, this offers a clear avenue for constructing such an understanding. Acquiring sufficient texts over an extended period of time – for instance, congressional records, newspaper articles, and scholarly commentary – is a clear challenge, but with the rapid expansion in the availability of digitized text the possibility grows ever closer of creating such a system.

Another clear avenue to expand this research is to move into the realm of substantively meaningful n-grams. As an example, the phrases “rational-basis”, “strict-scrutiny”, and “heightened-scrutiny” all offer important avenues for analyzing variations in judicial
behavior but are treated coarsely in this iteration of the research. The additional detail comes at significant computational cost, but recent work by Handler et al. (2016) suggest methods for deriving these substantively meaningful n-grams that helps to mitigate these concerns. I leave this to future research. For now, it is clear that the distributed representation of terms offers stark improvements over existing approaches to computational text analysis in the context of legal institutions, and important new possibilities for empirical legal research.
References


URL: http://dirichlet.net/pdf/handler16bag.pdf

URL: https://doi.org/10.1007/978-3-642-22309-9_5


