Toward Computational Literature Reviews:

Applying Expert-Built Dictionaries for Automated Analysis of Complex Texts

Jaren Haber, Heather A. Haveman, and Yoon Sung Hong

Abstract

The social sciences face growing demand for reproducible tools for processing massive troves of often-complex text data (political speeches, medical notes, etc.). In response, we aim toward computational literature review by developing an inductive method of applying expert-built dictionaries for automated analysis of complex texts. Our workflow begins with developing dictionaries from foundational texts and domain expertise. Next, we apply text-analytic methods of differential domain-specificity and complexity to create vector-space representations of texts. Finally, we compare the validity of these methods by using regression models to evaluate relationships between their representations and ground truth. Taking as our use case a large corpus of academic articles in organizational science, we find that domain-specific, relatively simple embeddings were most valid--while the more sophisticated models were very weak. Thus, we suggest that social science workflows for learning from complex texts incorporate embeddings that are domain-specific and straightforward--rather than convoluted.

Research in many academic fields requires reviewing the literature to document what we know about a phenomenon and what gaps exist in our knowledge. In this paper, we outline a computational workflow for reviewing academic literature that takes advantage of online collections containing nearly all articles published in academic journals (e.g., [JSTOR](#), [MathSciNet](#), [Web of Science](#), [MEDLINE](#)).

Our computational review of academic literature stands in sharp contrast to the time-honored practice of reading a small number of texts, usually far less than 1,000. The latter is prone to biases stemming from researchers' training and social networks and less reliable than computational coding--which in contrast is capable of coding the same text the same way every time (check veracity of this; cites on reliability of human vs. computer coding). We thus provide an analytical framework for reliable, cross-disciplinary, computational literature review with the potential to accelerate the research process as well as open new avenues for the science of science (cites in this area; name any other prominent examples of computational literature review, e.g. JSTOR's tools). Indeed, computational text analytic methods are increasingly important in the era of massive online text availability.

Our methodological framework also provides researchers with robust, theoretically generative tools for analysis based on dictionaries (lists of conceptually related terms). Dictionary methods are simple and flexible and their results relatively easy to comprehend (Grimmer and Stewart 2013), and most research using dictionaries relies on pre-existing, validated sentiment dictionaries (e.g., LIWC: Pennebaker et al. 2001; ANEW: Bradley and Lang 1999). Our goal is to expand the reach of this method through publicly available tools for generating and validating more complex, domain-specific dictionaries. Combined with widely

available sources of text data--such as social media (e.g., Twitter, Reddit), popular news (e.g., Google News), consumer reviews (e.g., Yelp, IMDb), or digitized academic journals (e.g., JSTOR)--accessible, beginner-friendly methods for developing and applying dictionaries represent an exciting frontier for reproducible, large-scale social scientific research.

Previous work in social sciences has modeled conceptually simple dictionaries in corpora of comparatively short texts, such as predicting positive-negative sentiment in Twitter posts (e.g., Tumasjan et al. 2010). Contributing to this stream of research, the present study models conceptually complex, abstract dictionaries (relational, cultural, and demographic perspectives) in longer, denser texts (academic journal articles). This study specifically addresses two questions: Do domain-specificity or model complexity improve the fidelity of word and sentence embeddings when used to apply complex dictionaries? In addressing these questions, we contribute to emerging research in the domain-specificity and complexity of word embeddings. The need for such research is especially pressing given that pre-trained word embeddings form the foundation for the current generation of highly complex, universal language encoders from transformers (e.g., BERT, ELMo; Peters et al. 2018; Devlin et al. 2019). While such encoders have many potential applications, the benefits of such added complexity for social science research is little studied (but see, e.g., Vicinanza 2019).

Our workflow proceeds in the following steps. First, we draw on expert knowledge to generate theoretically distinctive dictionaries from a set of foundational texts in sociology and management/organizational behavior. We then use supervised (count-based dictionary methods) and unsupervised (word embeddings, both pre- and custom-trained) text-analytic methods to create vector-space representations of each academic article in our corpus. Finally,

we use regression models to assess associations between these representations, on the one hand, and articles' citations of the foundational texts, on the other. Such citations represent a ground truth in organizational science, because they indicate direct reliance on the theoretical tradition such works established. Through such context-specific linkages, we answer calls for dictionary validation (e.g., Grimmer and Stewart 2013; Nelson, Burk, Knudsen, and McCall 2018) in addition to providing a sound metric through which to compare the efficacy of word counts against word and sentence embedding algorithms.

Our various embedding methods each rely on a distinct architecture to create linguistic representations; such architectures are consequential and continue to compel innovations in natural language processing. Nonetheless, it is standard practice in applied studies to compare the ability of multiple algorithms to create representations, which are validated through comparison with a ground truth, e.g. hand-coding (e.g., Nelson et al. 2018; Garten et al. 2018). Thus, while in this paper we broadly consider the relative complexity of these algorithms' architectures, we adopt the approach of previous studies and emphasize the association of the various models' representations with a valid outcome--in our case, citations of theoretically foundational texts.

Thus, extending work on distributed dictionary representations (Garten et al. 2016, 2018), we provide best practices for applying complex dictionaries to large-scale digital text corpora. For instance, if we discover that the custom-trained `word2vec` model performs best in the modeling step, then future research using complex dictionaries would do well to privilege domain-specific word embedding models over word counts or more complex--but less contextually relevant--sentence embedding models. Reliability is built into our computational approach--and by making our code publicly available, we also ensure reproducibility.

The rest of this paper proceeds as follows. First, we outline prior research on dictionary methods and vector space models. We then describe our methodological approach, including dictionary construction and application and several word and sentence embedding models. Finally, we outline and discuss the results of our models, stressing the comparison of word counts with pre-trained and custom-trained embedding models.

Prior Work

We draw on two computational text analytic methods--with social-scientific traditions--that are each domain-specific: dictionary methods and vector space models.

Dictionary Methods

Dictionary-based approaches to text analysis represent documents by counting theoretically-derived word lists (dictionaries; Stone et al. 1966). In dictionary methods, researchers develop a list of words connected to a category or concept of interest, then count instances of these words (either weighted or unweighted) in a sample of texts for purposes of categorization (Grimmer and Stewart 2013; Jurafsky and Martin 2018). Dictionary methods have been used extensively in the social sciences, not just in sociology (e.g., Goldberg et al. 2016), but also in political science (e.g., Laver and Garry 2000), finance (e.g., Tetlock 2007), and psychology (e.g., Snefjella and Kuperman 2015).

Dictionary methods are based on the notion that language reflects culture: the cognitive categories through which individuals attend to the world are embedded in the words they use (Whorf 1956; Sapir 1958). Words that are used frequently are cognitively central and reflect what is most on the speaker's (or writer's) mind. Words that are used infrequently or not at all

are at the speaker's cognitive periphery, perhaps even representing uncomfortable or alien concepts.

Analysis of dictionaries with closed-class terms--that is, word lists that are rarely extended or modified (e.g., pronouns, articles, and prepositions)--has garnered many insights in computational linguistics, including understanding of emotional experience (Pennebaker 1997), authorship identification (Boyd and Pennebaker 2015), and personality and social identity (Pennebaker et al. 2003). However, closed-class dictionaries are limited in their ability to capture phenomena of theoretical interest to social science (e.g., educational models, organizational theories, authoritarianism), the terms of which are less grammatically circumscribed. Instead, it is standard for social researchers to develop dictionaries of open-class terms (i.e. nouns, verbs, adjectives) suited to a specific measurement task and analytical domain, such as economic interest (Enns et al. 2016), political ideology (Graham, Haidt, and Nosek 2009), anti-establishment rhetoric (Bartels, Oliver, and Rahn 2016), or ideological zeal (Haveman and Khaire 2004). Such dictionaries may be more hand-driven (Schwartz and Ungar 2015), that is, developed by experts from domain knowledge, literature review, and/or identifying synonyms and syntactic variants through tools like WordNet (Miller 1985); or more data-driven, or developed by means like crowd-sourcing (e.g., Dodds et al. 2011; Benoit et al. 2016) or identifying distinguishing words from pre-labeled corpora (Monroe et al. 2008).

Despite the rise of data-driven approaches to developing open-class dictionaries and adapting them to particular analytical domains, this work must contend with an inherent limitation: no dictionary captures a concept universally, but only within the specific context in which it was generated and validated (Grimmer and Stewart 2013; Nelson et al. 2018). Indeed,

because both researchers and text data are socially embedded, the significance of any set of terms shifts across sociolinguistic (Louwerse 2004) and sociocultural contexts (Henrich et al. 2010). For instance, applying the Harvard General Inquirer (Stone et al. 1966) to classify negative tone in corporate earnings reports leads to serious misclassifications (Loughran and McDonald 2011). A classic example of applying a dictionary beyond its target domain, consequences include both many false positives (e.g., the term *vice* characterizes an executive rather than immorality, while *tire* is not negative in the context of automobile industry reports) and false negatives (e.g., *litigation* and *unanticipated* don't appear in the Harvard list; Loughran and McDonald 2011). Moreover, language use changes with communication technology: for instance, social media incubates many context-specific expressions of tone (e.g., emoticons, hashtags), and as a result formal linguistic features (e.g., parts of speech) poorly predict sentiment (Kouloumpis et al. 2011).

Despite the importance of domain-specific dictionary development and validation, many scholars continue to subjectively derive dictionaries from content knowledge or inspection of data (Egami et al. 2018) or rely on (supposedly) universal sentiment dictionaries developed and validated using large, non-specific text corpora (especially LIWC: Pennebaker et al. 2001; but also PANAS-X: Watson & Clark 1999). Scholars do so because dictionary development and validation is time- and resource-intensive--difficulties compounded by the absence of universal dictionary validation procedures (but see Grimmer and Stewart 2013) and the inability of manual content analysis to reproduce granular dictionary-derived measures (Krosnick 1999). And yet foregoing dictionary refinement and validation can obscure causal relationships: it's often

unclear which words drive statistical relationships, and relying on fixed (or even black-box) categories can conceal unexpected associations (Schwartz and Ungar 2015:81).

Our workflow provides an accessible solution to domain-specific dictionary development and validation. In a modified form of computational grounded theory (Nelson 2017), we begin with an expert-generated seed dictionary, iteratively expand on these seed terms using vector space models, and confirm sociological patterns through dictionary application to a specific corpus (as others have done with, e.g., movie reviews or social media posts: Garten et al. 2018; Sivak and Smirnov 2019). By assessing term frequencies, we also identify each dictionary's most central concepts and clusters--an important step in development of more complex dictionaries.

To further set the stage for our contributions, we now discuss how our approach builds on previous research with vector space models.

Vector Space Models

Computers are essential tools for working with today's text data, which is increasingly available in volumes beyond the capacity of humans to read (Bail 2014). However, because computers outperform humans only on narrowly defined tasks (Geiger, Cases, Karttunen, and Potts 2019), computational text analysis necessarily involves reducing the complexity of texts through preprocessing.¹ Such dimensionality reduction typically entails translating words into numbers--or more specifically, *vectors* defined by relationship with other words and/or with documents.

¹ As a consequence of this dimensionality reduction and the inherent complexities of language, all text analytic methods are "wrong" as models as language (Grimmer and Stewart 2013:3-4) and also often qualitatively "weird" when interpreted by humans (Shane 2019).

The classic (and still common) form of text vectorization is the term-document matrix: words are represented as rows, documents as columns, and each entry represents the occurrence, frequency, or weight of a given term in a given document (Salton, Wong, and Yang 1975). Information retrieval can be enhanced by weighting words by rarity within and across documents, often through the *term frequency by inverse document frequency* metric (Salton and Buckley 1988). Whether frequencies or weights are used, the term-document matrix represents text using $N \times D$ dimensions, where N is vocabulary size (number of words) and D is corpus size (number of documents). As such, even for relatively small corpora of less than 10,000 article-length documents, the term-document matrix is a massive matrix with billions or even trillions of cells. As a result, matrix operations on term-document matrices pose formidable costs in time and resources. Such expense is exacerbated when using still larger text matrix types: term-term (words as rows, words as columns), term-context (words as rows, phrases, paragraphs, or other length text as columns) or pair-pattern (word pairs as rows, i.e. X:Y; semantic patterns as columns, e.g. “X makes Y”) (Turney and Pantel 2010).

To enable efficient, sophisticated operations on vectorized texts, over the 20th and early 21st century there emerged models known as *vector space models*. Unlike the term-document matrix or its analogues, for purposes of efficiency and scalability vector space models represent words as numerical vectors of dimensionality far smaller than N or D . In principle, such models are a form of *distributed representation* in that each word is “understood as the sum of all its environments” or co-occurents (Harris 1954:146), i.e. that text elements similar in distribution will have similar meanings. This model of semantic meaning structure is known as the *distributional hypothesis*, expressed by the axiom “you shall know a word by the company it

keeps” (Firth 1957). Built on the distributional hypothesis is the popular method called Latent Semantic Analysis (LSA), which applies the dimensionality reduction technique of singular value decomposition to the term-document matrix and compares document vectors using cosine similarities (more on cosine similarities below; Deerwester et al. 1990). A more recent example is global word vectors (or GloVe), which are developed by applying weighted least squares regression to the term-term matrix (Pennington, Socher, and Manning 2014). Such models encode within words’ vector relations both attributional similarity (which measures synonymy) and relational similarity (which measures analogy) (Medin, Goldstone, and Gentner 1990; Turney, 2006).

Training on large, unlabeled text corpora, many other vector space models learn word representations with neural network architectures aiming to predict either words or their contexts within word-context windows (Bengio 2003, Collobert and Weston 2008, Collobert et al. 2011). Because they embed structural, semantic patterns in numerical word vectors, such models are known as *word embeddings*. Examples of recovering semantic and syntactic structure from word embeddings include finding the nearest terms of “frog” to be “toad” and “litoria” (a species of frog; Pennington et al. 2014) and successfully solving analogies like “man is to woman as king is to (queen)” (Mikolov, Yih, and Zweig 2013). Word embeddings are frequently used in digital humanities research; for example, to trace the shifting meaning of “gay” over the twentieth century from a position near “dapper” and “cheerful” to one close to “lesbian” and “homosexual” (Kulkarni et al. 2015). And word embeddings are becoming more common in the social sciences; for example, to analyze associations between basic cultural categories such as gender, race, and status/wealth (Kozlowski, Taddy, and Evans 2019). Probably the most

influential word embedding model is `word2vec`, which uses a simple, single-layer neural network architecture to predict context or focal words (Mikolov, Yih, and Zweig 2013).

We draw on vector space models to refine and apply dictionaries. Thus, we turn now to discuss the intersection of these approaches.

Word Embeddings and Dictionaries

Approaches such as Distributed Dictionary Representations offer the potential to apply small, conceptually tight sets of seed words to effectively capture concepts in textual corpora of variable length and complexity (Garten et al. 2016, 2018). While the context-dependence of dictionaries themselves has been well studied (e.g., Loughran and McDonald 2011; Nelson et al. 2018), studies of dictionary generation and implementation using word embeddings have only begun to explore how dependent are such embeddings on the specific context in which they were trained. And yet the domain-specialization of embeddings explains the increasingly common use of word embeddings to build and refine specialized dictionaries across disciplines including sociology (e.g., Haber 2019), finance (e.g., Tsai and Wang 2014), and medicine (e.g., Gianfrancesco 2019). It also provides the impetus behind techniques for customizing pre-trained embeddings to produce domain-specialized word representations even with small corpora (e.g., Dingwall and Potts 2018; Faruqui et al. 2015).

Despite such efforts, there remains pressing uncertainty around how portable are embeddings--that is, with how much loss of accuracy--to linguistic contexts other than those on which they were trained. As such, much work remains to be done to clarify the scope conditions of embeddings pre-trained on large but not task-specific corpora. The limited evidence available suggests that while domain-specific embeddings produce superior results when

conceptually simple, (supposedly) universal sentiment dictionaries are used, this tendency is contradicted when more complex dictionaries such as the five-fold Moral Foundations Dictionary are at hand (Garten et al. 2018). The present study directly builds on this research by comparing the performance of custom-trained and pre-trained embeddings for complex dictionaries applied to large text corpora.

Having discussed relevant prior work, we now describe our two methods: dictionary analysis and word/sentence embeddings.

Dictionary Methods

Because no pre-existing, validated dictionaries have been generated for our analytical domain (organizational science), we construct our own dictionaries through a rigorous process rooted in expert knowledge of organizational theories and their specialized concepts. In the following, we briefly describe the rationale for this approach and then outline its application to our domain of application: organizational science.

Specialized Concepts in Organizational Science

One common way language reflects culture (Whorf 1956; Sapir 1958) is through specialized vocabulary, words and phrases coined by cultural subgroups to capture important concepts – in other words, jargon. Jargon makes it possible to communicate key concepts quickly and precisely among members of a cultural subgroup, including scientists: “Every science requires a special language, because every science has its own ideas” (de Condillac 1782; quoted in Braudel 1982 [1992]: 234). Every scientific community possesses a long list of concepts that are central to its theories – these are de Condillac’s “ideas.” Through their use of jargon, scholars align their writing with their particular intellectual community (Vilhena et al.

2014) and with their own epistemic culture (Knorr-Cetina 1999). For example, when an organizational demographer refers to *resource partitioning* (Carroll 1985), this is an efficient shorthand for the idea that increasing concentration (reduced numbers of entities coupled with increased size) of organizations that serve a broad customer base can create opportunities for small niche players. Similarly, when a cultural scholar of organizations refers to *decoupling* (Meyer and Rowan 1977), this communicates a disconnect between an organization's policies and actual practices. And when a relational scholar of organizations mentions *structural holes* (Burt 1992), this conveys the idea that actors benefit from creating bridges between otherwise unconnected subgroups, either by transferring information or gatekeeping. The lists of concepts developed by different scientific communities often do not overlap; for example, network scholars are less likely to mention resource partitioning or decoupling than they are to talk about structural holes. Accordingly, to capture the theoretical concepts that are important to three intellectual communities in the field of organizations, we focus on the specialized vocabulary used by each of them.

The focus of this analysis is the field of organizations, which spans multiple disciplines (e.g., sociology, political science, management, public health) and which is largely published in articles, rather than books. Moreover, we emphasize macro-organizational theory – the study of firms, industries, populations, and fields. This literature has been dominated for the last four decades by three perspectives: demographic, relational, and cultural (Haveman and Wetts 2018). These perspectives have different conceptions of social structure, and thus different conceptions of what creates opportunities for and constraints on the actions of individuals, groups, and organizations. The *demographic perspective* holds that social structure is

constituted by distributions of social actors along salient dimensions of social and physical space: individual age, race, and gender; workgroup size and composition; and organizational strategy, size, and location (Blau 1977; Kanter 1977; Baron and Bielby 1980; Pfeffer 1983; Hannan and Freeman 1984, 1989). The *relational perspective* holds that social structure is constituted by webs of social connections among people, groups, and organizations; e.g., friendship networks, workgroup interactions, and buyer-supplier ties between organizations (Emerson 1962; Granovetter 1973; Pfeffer and Salancik 1978; Pfeffer 1981; Burt 1983). The *cultural perspective* holds that social structure is constituted by widely shared norms, values, expectations, roles, and rituals, which generate understandings of what is possible and reasonable (Meyer and Rowan 1977; DiMaggio and Powell 1983; Scott 1995 [2008]; Scully and Creed 1997).

These perspectives also have different conceptions of identity (“Who am I?”, “What do I stand for?”, “What do I want?”) and therefore different motivations for action. For demographers, identity and motivation derive from position, absolute or relative, along salient dimensions of social life. For relational scholars, identity and motivation derive from ties among individuals, groups, and organizations. For cultural scholars, identity and motivation derive from social interaction. All three perspectives have been applied to explain both *micro-level behavior* (of individuals like professors and students; small groups like campus clubs, fraternities, and sororities; or organizational subunits like academic departments), and *macro-level behavior* (of entire organizations like UC Berkeley or Penn State, organizational populations like American public universities, and fields like American higher education). This analysis focuses on macro-level behavior.

From Foundational Texts to Dictionaries

To begin the dictionary construction process, the second author read foundational texts to generate a dictionary of concepts for each perspective. The foundational texts were as follows: for the *demographic* perspective, Hannan and Freeman (1977, 1984); for the *relational* perspective, Pfeffer and Salancik (1978 chapters 1-2 and 5-8; she focused on the theoretical sections and not the reports of their empirical analyses); and for the *cultural* perspective, Meyer and Rowan (1977) and DiMaggio and Powell (1983).

From these texts, the second author recorded key terms and their cognates (e.g., imitate, imitation, imitating) and removed the same stop words as were removed in the creation of the JSTOR text data (see Appendix 2). She then compared the three dictionaries and eliminated overlapping terms. For example, “organization,” “uncertain,” and “uncertainty” were deleted from all three dictionaries, “homogenous” was deleted from the demographic dictionary because it was more central to the cultural perspective, “relational network” was deleted from the cultural dictionary because it was more central to the relational perspective, and “survival” was deleted from the relational dictionary because it was more central to the demographic perspective. The only overlapping terms she retained were “legitimate” and its cognates in the demographic and cultural dictionaries because these concepts were central to both perspectives. In the final manual filtering step, compound terms with shared roots were removed (e.g., “ritual commitment” and “ritual confidence” were reduced to the word “ritual”), leaving sharp, distinctive lists of 36 words each (see Appendix 1 for dictionary terms). These filters ensured that new terms were theoretically relevant; for this reason, manual checks are a

common means of validating dictionaries built with word embeddings (e.g., Sivak and Smirnov 2019; Haber 2019).

Calculating engagement with dictionaries

To assess how much each theoretical perspective has been employed in journal articles, we first use dictionary methods. Dictionary-based analysis is appropriate when the categories (here, theoretical perspectives) and the textual features (words and phrases that are associated with each category) are known and fixed (Quinn et al. 2010). Each dictionary covers the language that situates scholarly work within its perspective: the concepts in a perspective's dictionary express the ideas in that perspective. To the extent that a published article uses the concepts in a perspective's dictionary, the article engages intellectually with that perspective. For example, if an article focuses on sources of *structural inertia*, it is engaged with the demographic perspective. But if an article focuses on the development of *interlocking directorates*, it is engaged with the relational perspective. And if an article focuses on *institutional isomorphism*, it is engaged with the cultural perspective.

For each perspective, we measure how much each article published in academic journals use the focal perspective's concepts, scaled by article length. We label this measure *engagement*. Engagement with perspective p by article a is as follows:

$$engagement_{pa} = \frac{\sum_{i=1}^m perspective\ concept_{ipa}}{\sum_{j=1}^q article\ concept_{ja}}$$

where $engagement_{pa}$ is average engagement by article a with perspective p . The numerator in the equation above is the total number of times the concepts in perspective p are used in article a ; i.e., we counted each instance of each dictionary word or phrase in article a separately. The

denominator is the total number of unigrams in article a , i.e., the sum of the frequencies for all the unigrams in the focal article. $Engagement_{pa}$ is thus a continuous variable ranging from 0 to 1, where 0 indicates that an article uses none of the perspective's concepts and 1 indicates that all of the articles' words are that perspective's concepts. Since no article consists entirely and exclusively of a perspective's concepts, the observed maximum for $engagement_{pa}$ is likely to be a very small fraction.²

We now describe the various word and sentence embeddings we use as comparisons to count-based dictionary analysis.

Word and sentence embeddings

Word embeddings map words onto a high-dimensional vector space and represent semantic relations between words as geometric relations in space (Mikolov, Sutskever, et al. 2013; Kozłowski et al. 2019). A pair of words that share many contexts (i.e., they are collocated with the same other words) are positioned near each other in vector space, and a pair of words that have very different contexts (i.e., they are collocated with different other words) are positioned far apart. Importantly, this relational mapping captures commonalities in words' local contexts, rather than collocation alone. This large-scale mapping of contexts encodes word embeddings with underlying cultural meanings, rather than strictly on-the-ground observable patterns.

Applied over many sentences, word embeddings efficiently and accurately predict semantic similarity—that is, the extent to which any given pair of words in the corpus are used in similar contexts and, thus, presumably have similar meanings. For example, in the text, “King

² Even the texts I used to create the dictionary for each perspective will not yield an engagement score of 1 because each dictionary captures only the text concepts most relevant to that perspective and ignores text concepts that are common across perspectives, such as “organization” and “uncertain.”

rides the horse. Prince rides the donkey”, the intuitively similar words “King” and “Prince” are also semantically similar (because they both precede “ride the”), as are the intuitively similar words “horse” and “donkey” (because they both come after “ride the”). Moreover, this large-scale mapping of contexts encodes word embeddings with underlying cultural meanings, rather than strictly on-the-ground observable patterns.

Euclidean or straight-line distances are inappropriate given the high number of dimensions in vector space models (typically 100-300; Kozlowski et al. 2019:9). Instead, cosine similarities establish the distances between vectors, measuring the angle between vectors such that a score of 0 indicates perfect independence (orthogonality or 90 degrees between vectors) and 1 indicates perfect similarity or synonymy (parallelity or 0 degrees between vectors). Compared to similar metrics like Manhattan distance, cosine similarities offer the advantage of not being sensitive to document length (Zhang, Callan, and Minka 2002).³

We use four different algorithms to generate word and sentence embeddings. To generate custom-trained embeddings from our corpus, we use word2vec for word embeddings and doc2vec for sentence embeddings. To provide pre-trained embeddings, we use GloVe for word embeddings and InferSent to produce sentence embeddings. Word and sentence embeddings differ in how they represent articles: while the latter has a built-in representation for each text, the former requires the additional step of deriving a document vector by computing the arithmetic mean of its component word vectors. One goal of our study is to discover whether such differences in model complexity as native document (sentence) embeddings amount to improved resolution when tracking focal qualities of texts--in this case, the degree to which each article invokes a given organizational perspective.

³ Cosine distances achieve this by first normalizing word vectors, rendering them a consistent unit length. This means that all word vectors lie along the surface of a hypersphere (i.e., the generalization of a sphere into >3 dimensions) (Levy, Goldberg, and Dagan 2015; Xing, Wang, Liu, and Lin 2015).

As such, this assortment of methods facilitates two important comparisons when applying expert-developed dictionaries to complex texts. First, to evaluate domain-specificity in embeddings, we compare the accuracy of custom- and pre-trained embeddings. Second, to assess the impact of model complexity, we compare the performance of more complex architectures (sentence embeddings, deeper neural nets) with that of simpler ones (word embeddings, word counts).

We now briefly describe the custom- and pre-trained algorithms we use to create embeddings.

Custom-trained Embeddings: word2vec and doc2vec

To custom-train word and sentence embeddings on our corpus of academic articles, we use the `word2vec` and `doc2vec` implementations of the `gensim` library (Řehůřek and Sojka 2010). The `doc2vec` technique builds directly on `word2vec`; thus, we describe the latter first.

Mechanically, `word2vec` assesses word associations using ‘word context windows’, indicating the number of words (typically 6-12) on either side of focal word w to consider as connected to w . Computationally, `word2vec` learns word vectors using a simple, single-layer ‘neural network’, a machine learning architecture that tries to predict (in this case) the connections between words using a sample of observations of word connections or ‘training words’ (e.g., Turian, Ratinov, and Bengio 2010). Formally, for a series of training words $w_1, w_2, w_3, \dots, w_T$, the goal of `word2vec` is to maximize the average log probability of predicting w_{t+j} given w_t (Mikolov, Sutskever, et al. 2013:2):

$$\frac{1}{T} \sum_{t=1}^T \sum_{-c \leq j \leq c, j \neq 0} \log p(w_{t+j} | w_t)$$

where w indicates a word in the sentence, t is the iterator over T training words in the sample sentence, c is the word context window size, and j is a number between $-c$ and c (excluding zero) that indicates the distance in words from focal word w_t to a word within its context, w_{t+j} .

Much like `word2vec`, `doc2vec` uses word context windows and the same neural net architecture--but for an important difference: it derives not only word but also sentence vectors when computing embeddings, improving performance on some classification and information retrieval tasks compared to other document representations such as vector averaging (Le and Mikolov 2014). A “sentence” may be any length of text, including paragraphs and documents; for this paper, each sentence vector represents an academic article. See Appendix 3 for details on `word2vec` and `doc2vec`.

Pre-trained Embeddings: GloVe and InferSent

Word embeddings pre-trained using the GloVe algorithm (see description above) on the massive Common Crawl web-crawled corpus⁴ are widely used in natural language processing. Considering many embeddings are pre-trained on news or social media corpora (e.g., Google News, Gigaword, Twitter), the diverse lexicon of the Common Crawl is out of all options probably the least alien to academic discourse. For this reason, for our pre-trained embeddings we chose the Common Crawl GloVe embeddings as the basis. For commensurability, the custom embeddings above were trained using parameters as similar as possible to GloVe, the relevant parameters of which are: 300 vector space dimensions, 50 iterations through the data, and $\alpha = 0.75$ (compare to negative sampling exponent used in `word2vec`) (Pennington et al.

⁴ This GloVe model was trained on 840 billion tokens and 2.2 million vocabulary words of web crawl data (see <https://commoncrawl.org/>). The model is freely available online at <https://nlp.stanford.edu/projects/glove/>.

2014). We also used context window size 10, as did the creators of GloVe in their comparison word2vec models (Pennington et al. 2014:1538).

Our most complex text vectorization method, known as InferSent, begins with the pre-trained GloVe embeddings but uses the much more sophisticated Recurrent Neural Network or RNN (see Appendix 4) architecture to synthesize these into sentence embeddings that are holistic and sequential representation of all inputs (Conneau et al. 2017). In another layer of pre-training, InferSent uses the bidirectional LSTM architecture with max-pooling to classify labels for the Stanford Natural Language Inference (SNLI) (consisting of 570,000 human-labeled sentence pairs) in order to train and optimize an RNN. It thereby produces a high-dimensional 4096-dimensional vector to encode each text input. We don't use the multilayer perceptron architecture that follows the RNN, given that our objective with InferSent is not (yet) prediction but rather a homogeneous semantic representation at the article (more generally, sentence) level (see Appendix 5 for more details on InferSent).

Having outlined our various text vectorization methods, we now describe our data source of academic articles before analyzing results.

Data

The data come from the JSTOR Data-for-Research⁵ corpus from 1970 to 2015. The data included all articles in the two JSTOR subject areas (academic disciplines) central to the sociological study of organizations: (1) sociology and (2) management and organizational behavior (N = 398,703). (See Appendix 2 for additional notes on raw JSTOR data.) However, the full downloaded corpus contained many articles in journals not in English or only secondarily

⁵ See <https://www.jstor.org/dfr/>.

related to sociology or management and organizational behavior, as well as many compositions other than full-length journal articles and some themselves in a language other than English.

Thus, prior to analysis, we extensively filtered our corpus as follows.

To filter articles to those primarily in our two subjects of interest, we first identified the single most important subject for each journal to which any article may belong. To do so, we used the JSTOR Complete Title History List (JSTOR 2018) to map the title of each journal onto the JSTOR subject(s) under which it is classified. We then forced each journal into a single primary subject area through rigorous manual checks informed by the second author's expert knowledge of the field (for details, see Appendix 2). Definitive journal subjects in hand, we removed 160,202 articles not linked to known English-language journals⁶ and 82,948 additional articles not from journals in sociology or management and organizational behavior, leaving 155,553 compositions in relevant journals.

Given our theoretical focus on full-length scholarly articles, we also excluded from the analysis 85,471 remaining book reviews, books received, front matter, and any article shorter than 1,000 words.⁷ We also excluded 423 articles written in languages other than English.⁸ This final filter left a total of 69,659 articles to analyze.

Variables

We used a binary scheme to operationalize citations of foundational texts (the dependent variable): each article was coded as 1 if it cites one or more foundational texts for a given perspective, 0 if not. The foundational texts are: for the demographic perspective, Hannan

⁶ A tiny number of journals that publish articles in multiple languages, including English, were excluded (e.g., *Acta Historica Academiae Scientiarum Hungaricae* and *Anabases*).

⁷ Article lengths were determined using word counts provided by JSTOR.

⁸ Article language was identified by article metadata and (as a validity check) machine learning algorithms--specifically, pre-trained *InferSent* embeddings.

and Freeman (1977, 1984); for the relational perspective, Pfeffer and Salancik (1978); and for the cultural perspective, Meyer and Rowan (1977) and DiMaggio and Powell (1983).

Positive cases of the dependent variable are rare at 5.0% overall. Broken down by perspective, this includes 1,039 (1.5%) demographic citations, 916 (1.3%) relational citations, and 1,526 (2.2%) cultural citations.

Our analysis uses 15 independent variables: the five text vectorization methods (word counts, `word2vec`, `doc2vec`, `GloVe`, and `InferSent`) juxtaposed across the three organizational perspectives (demographic, relational, and cultural).

To ensure reproducibility, we analyzed these data in Jupyter notebooks in the flexible, open-source Python 3 scripting language (Van Rossum and Drake 2011) using Jetstream--a service provided by the Extreme Science and Engineering Discovery Environment (XSEDE; Towns et al. 2014).⁹ All code is extensively commented and openly available online to the research community.¹⁰

Having described the data at hand, we now describe the results of our analyses.

Results

We now compare the accuracy of five text vectorization methods: word counts, word embeddings with `word2vec` (custom-trained) and `GloVe` (pre-trained), and sentence embeddings with `doc2vec` (custom-trained) and `InferSent` (pre-trained). Each method is judged against the ground truth of citations of foundational texts, for in scholarship this indicates direct reliance on the theoretical tradition such texts established. We first compare the Pearson's

⁹ In particular, we used the Jetstream resource at Indiana University through allocation TG-SES180020.

¹⁰ View and contribute to our code at <http://bit.ly/comp-lit-review>.

Correlation Coefficients between the vectorized text representations and foundational text citations before discussing the results of logistic regression models.

Correlations

As shown in Tables 2-4, the highest correlations between text representations and foundational text citations are achieved by two methods: raw word counting and relatively simple, custom-trained word embeddings (`word2vec`). The `GloVe` algorithm places third--and performs much better in the management and organizational behavior discipline than in sociology. Perhaps surprisingly, the sentence embedding algorithms perform far worse: not only are the scores for `doc2vec` and `InferSent` practically zero, but they are entirely negative for `InferSent` and statistically insignificant for `doc2vec`. That simpler models outperform more complex ones suggests that with complex dictionaries and texts, sophisticated modeling architecture can backfire. However, given that these correlations don't account for the role of time or discipline, we turn now to logistic regression models that do.

[TABLES 2-4 ABOUT HERE]

Logistic Regression Models

To provide a more sophisticated measure of the relationships between variables, we use logistic regression (Berkson 1944)--a simple, widely used algorithm in social science that allows us to compare the five text vectorization methods by standardized means. To capture effects accurately, in each model we include a single text representation from one of the five methods (`word counts`, `word2vec`, `doc2vec`, `GloVe`, and `InferSent`) along with controls for document year published and a dummy variable indicating whether an article is in management and organizational behavior (as opposed to sociology). These models therefore account for the influence of time and discipline. Figure 1 shows the results in terms of McFadden's pseudo- R^2 ,

or the proportion of variance in the outcome explained by the model. This measure indicates how closely tracked is the ground truth through the lens of the vectorization approach in hand.

[FIGURE 1 ABOUT HERE]

As Figure 1 shows, logistic models replicate the pattern established by the correlation tables just above: in terms of explained variance, text representations derived from custom-trained embeddings are more valid than those from more complex models. In fact, according to these models, simple, domain-specific embeddings (`word2vec`) outperform word counts--replicating a key finding from previous research (e.g., Garten et al. 2018)--while pre-trained word embeddings (`GloVe`) do just as well as word counts. Moreover, the pseudo- R^2 values achieve relatively high levels for `word2vec` in particular, reaching 0.438 when predicting citations of the foundational demographic texts.

The log-likelihoods and beta coefficients in these logistic regressions tell the same story: models with custom-trained embeddings have higher log-likelihoods (indicating better fit) and higher betas (indicating stronger effect size) than do pre-trained and word count models.¹¹ As such, logistic regression confirms the superiority of simple, domain-specific embeddings models when applying custom dictionaries to complex texts.

Discussion and Conclusion

Responsible, reproducible social science in the digital age is reinforced by accessible computational tools and the research community to support using them both ethically and effectively (Salganik 2017). Hoping to contribute to this effort, we offer tools designed for a

¹¹ Results available on request from first author.

widely available digital text data source (JSTOR Data-for-Research)--which could reshape how scholars approach academic literature in the modern era.

We particularly aim to provide a reproducible workflow for computational literature review by evaluating the validity of methods of complex text representation used in automated analysis. We investigate a large corpus of academic articles in organizational science, for which case we use citations of foundational theoretical texts as a gold standard against which various models--both pre- and custom-trained, with varying degrees of complexity--can be judged. In doing so, we find that relatively shallow, domain-specific word embeddings offer more valid text representations than other approaches--and the most complex approaches (sentence embeddings and recurrent neural networks) had practically zero relationship with ground truth. Our findings extend previous research (e.g., Garten et al. 2018) by exploring complex dictionaries and texts amidst embeddings of varying sophistication and domain-specificity.

As such, we respond to the pressing needs among social scientists to develop and implement tools for dictionary validation (e.g., Grimmer and Stewart 2013; Nelson, Burk, Knudsen, and McCall 2018) and for learning from massive troves of culturally relevant text data (e.g., Bail 2014). In providing tools for comprehensive (computer-assisted) reading of academic literature, we also maximize the potential for discovery of new findings beyond the theoretical scope of the literature itself. The ability to reveal unexpected findings could make our approach useful to scholars in multiple fields--especially those that span multiple disciplines.

Our results also offer a cautionary tale amidst growing excitement toward manifold applications of linguistic transformer models: complex architectures built around embeddings pre-trained on huge text corpora (e.g., online news sources, Wikipedia) with impressive

predictive performance on standardized natural language tasks (e.g., Devlin et al. 2019).

Adversarial research shows that such models are surprisingly brittle and prone to inaccuracies given shifts in usage or informational cues (Geiger et al. 2019). Contributing to this line of research, we show that complex models may lack the sensitivity to provide valid representations in domains with complex local lexica such as academic discourse. Much more research is needed to provide theoretical and empirical scope conditions on the domain specificity of linguistic models based on embeddings--and whether the additional complexity of approaches such as transformers pays off within domains with specialized language (jargon).

Limitations and Future Work

Given that foundational texts for each perspective are cited in only 1-2% of articles in our dataset, the imbalanced dependent variable could limit the predictive performance of our logistic models. To improve results and facilitate more complex approaches (e.g., predicting citation counts rather than a binary measure), future studies could use citation metrics to broaden the range of “foundational” texts--while retaining theoretical pedigree--and increase the proportion of true positives in citation measures. Alternatively, hand-coding selected articles--with upsampling of likely positive cases for each perspective--could also provide a granular and balanced dependent measure.

Moreover, because our study uses dictionaries of a fixed size (36 terms), it does not account for the possibility that weak conceptual coherence in the dictionaries places a ceiling on our models’ performance. Future studies could address this issue by drawing on additional subject-area experts in the initial dictionary development stage. Given a set of foundational texts, these experts could create from scratch their own lists of theoretically distinctive terms.

Alternatively, they could be delivered an already-processed list and be asked to rank, filter, and append to it to maximize theoretical fidelity. The resulting lists could be compared and integrated to validate and refine the core dictionaries (like those used here).

A complementary, data-driven approach to dictionary development would involve using word embeddings to inductively expand the original terms into dictionaries of multiple lengths. Comparing cosine similarities would allow measurement of conceptual coherence within and distinctiveness between dictionaries, and inductively added terms could be checked using word counts and associations with ground truth measures. Coupled with prediction of citation measures, this would provide a valid and reproducible workflow for generating dictionaries from core terms and testing which dictionary length is optimal for each vectorization method.

Finally, while we have released our code to the public (<http://bit.ly/comp-lit-review>), we aim also to provide a well-documented, accessible replication repository that includes step-by-step instructions for downloading original datasets from JSTOR Data-for-Research as well as code to parse and preprocess the resulting raw data. Such tools would extend our work here by making computational literature review available open-source, with no cost (except those imposed by JSTOR) and minimum coding skill requirements.

References

- Aizawa, Akiko. 2003. "An Information-Theoretic Perspective of Tf-Idf Measures." *Information Processing & Management* 39(1):45–65.
- Allen, David M. 1974. The relationship between variable selection and data augmentation and a method for prediction. *Technometrics* 16, 125-127. (<https://doi.org/10.1080/00401706.1974.10489157>)
- Bail, Christopher A. 2014. "The Cultural Environment: Measuring Culture with Big Data." *Theory and Society* 43(3–4):465–82.
- Bamman, David, Chris Dyer, and Noah A. Smith. 2014. Distributed representations of geographically situated language. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*: 828-834. Baltimore, MD, June 23-25.
- Baron, James N., and William T. Bielby. 1980. Bringing the firms back in: Stratification, segmentation, and the organization of work. *American Sociological Review*, 45: 737-765.
- Bartels, Larry M., J. Eric Oliver, and Wendy M. Rahn. 2016. "Rise of the Trumpenvolk: Populism in the 2016 Election." *The ANNALS of the American Academy of Political and Social Science* 667(1):189–206.
- Benoit, Kenneth, Drew Conway, Benjamin E. Lauderdale, Michael Laver, and Slava Mikhaylov. 2016. "Crowd-Sourced Text Analysis: Reproducible and Agile Production of Political Data." *American Political Science Review* 110(2):278–95.
- Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. "A Neural Probabilistic Language Model." *Journal of Machine Learning Research* 3(Feb):1137–55.
- Berkson, Joseph. 1944. "Application of the Logistic Function to Bio-Assay." *Journal of the American Statistical Association* 39(227):357–65.
- Blau, Peter M. 1977. *Inequality and Heterogeneity: A Primitive Theory of Social Structure*. New York: Free Press.
- Boyd, Ryan L. and James W. Pennebaker. 2015. "Did Shakespeare Write Double Falsehood? Identifying Individuals by Creating Psychological Signatures With Text Analysis." *Psychological Science* 26(5):570–82.
- Braudel, Fernand. 1982 [1992]. *Civilization and Capitalism, 15th-18th Century, Vol. 2: The Wheels of Commerce*. (Translated by Siân Reynolds.) Berkeley: University of California Press.
- Burt, Ronald S. 1983. *Corporate Profits and Co-optation: Networks of Market Constraints and Directorate Ties in the American Economy*. New York: Academic Press.
- Burt, Ronald S. 1992. *Structural Holes: The Social Structure of Competition*. Cambridge, MA: Harvard University Press.
- Carroll, Glenn R. 1985. Concentration and specialization: Dynamics of niche width in populations of organizations. *American Journal of Sociology*, 90: 1262-1283.
- Collobert, Ronan, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. "Natural Language Processing (Almost) from Scratch." *Journal of Machine Learning Research* 12(Aug):2493–2537.
- Collobert, Ronan and Jason Weston. 2008. "A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning." Pp. 160–167 in *Proceedings of the 25th International Conference on Machine Learning*. ACM.
- Conneau, Alexis, Douwe Kiela, Holger Shwenk, Loïc Barrault, Antoine Bordes. 2017. "Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. Conference on *Empirical Methods on Natural Language Processing (EMNLP)*.
- Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. "Indexing by Latent Semantic Analysis." *Journal of the American Society for Information Science* 41(6):391–407.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." *ArXiv:1810.04805 [Cs]*.
- Diesner, Jana, Terrill L. Frantz, and Kathleen M. Carley. 2005. "Communication Networks from the Enron Email Corpus 'It's Always about the People. Enron Is No Different.'" *Computational & Mathematical Organization Theory* 11(3):201–228.
- DiMaggio, Paul, Manish Nag, and David Blei. 2013. "Exploiting Affinities between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of U.S. Government Arts Funding." *Poetics* 41(6):570–606.
- DiMaggio, Paul J., and Walter W. Powell. 1983. The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. *American Sociological Review*, 48: 147-160.

- Dingwall, Nicholas and Christopher Potts. 2018. "Mittens: An Extension of GloVe for Learning Domain-Specialized Representations." Pp. 212–217 in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 2*. New Orleans, Louisiana.
- Dodds, Peter Sheridan, Kameron Decker Harris, Isabel M. Kloumann, Catherine A. Bliss, and Christopher M. Danforth. 2011. "Temporal Patterns of Happiness and Information in a Global Social Network: Hedonometrics and Twitter." *PloS One* 6(12):e26752.
- Egami, Naoki, Christian J. Fong, Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart. 2018. "How to Make Causal Inferences Using Texts." Retrieved April 16, 2018 (<https://arxiv.org/abs/1802.02163>).
- Emerson, Richard M. 1962. Power-dependence relations. *American Sociological Review*: 31-41.
- Enns, Peter, Nathan Kelly, Jana Morgan, and Christopher Witko. 2016. "The Power of Economic Interests and the Congressional Economic Policy Agenda." *Scholars Strategic Network*, June 2.
- Faruqui, Manaal, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2015. "Retrofitting Word Vectors to Semantic Lexicons." Pp. 1606–1615 in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics*.
- Firth, John R. 1957. "A Synopsis of Linguistic Theory, 1930-1955." Pp. 1–32 in *Studies in Linguistic Analysis*. Oxford: Philological Society.
- Garten, Justin, Reihane Boghrati, Joe Hoover, Kate M. Johnson, and Morteza Dehghani. 2016. "Morality between the Lines: Detecting Moral Sentiment in Text." in *Proceedings of IJCAI 2016 workshop on Computational Modeling of Attitudes*. New York, NY.
- Garten, Justin, Joe Hoover, Kate M. Johnson, Reihane Boghrati, Carol Iskiwitch, and Morteza Dehghani. 2018. "Dictionaries and Distributions: Combining Expert Knowledge and Large Scale Textual Data Content Analysis." *Behavior Research Methods* 50(1):344–61.
- Geisser, Seymour. 1975. The predictive sample reuse method with applications. *Journal of the American Statistical Association*, 70: 320-328. (<https://doi.org/10.1080/01621459.1975.10479865>)
- Geiger, Atticus, Ignacio Cases, Lauri Karttunen, and Christopher Potts. 2019. "Posing Fair Generalization Tasks for Natural Language Inference - ACL Anthology." Pp. 4484–94 in. Hong Kong, China: Association for Computational Linguistics.
- Gianfrancesco, Milena. 2019. "Application of Text Mining Methods to Identify Lupus Nephritis from Electronic Health Records." Presented at the Text Analysis Across Domains (TextXD) Symposium, December 5, University of California, Berkeley.
- Goldberg, Amir, Sameer B. Srivastava, V. Govind Manian, William Monroe, and Christopher Potts. 2016. Fitting in or standing out? The tradeoffs of structural and cultural embeddedness. *American Sociological Review*, 81 (6): 1190-1222.
- Graham, Jesse, Jonathan Haidt, and Brian A. Nosek. 2009. "Liberals and Conservatives Rely on Different Sets of Moral Foundations." *Journal of Personality and Social Psychology* 96(5):1029–46.
- Granovetter, Mark. 1973. The strength of weak ties. *American Journal of Sociology*, 78: 1360–1380.
- Grimmer, Justin, and Brandon Stewart. 2013. Text as data: The promises and pitfalls of content analysis methods for political texts. *Political Analysis*, 21 (3): 267-297.
- Haber, Jaren. 2019. Sorting Schools: A computational analysis of charter school identities and stratification. Unpublished manuscript.
- Hall, David, Daniel Jurafsky, and Christopher D. Manning. 2008. Studying the history of ideas using topic models. *Proceedings of the EMNLP [Empirical Methods in Natural Language Processing]*.
- Hannan, Michael T., and John Freeman. 1977. The population ecology of organizations. *American Journal of Sociology*, 82: 929-964.
- Hannan, Michael T., and John Freeman. 1984. Structural inertia and organizational change. *American Sociological Review*, 49: 149-164.
- Hannan, Michael T., and John Freeman. 1989. *Organizational Ecology*. Cambridge, MA: Harvard University Press.
- Harris, Zellig S. 1954. "Distributional Structure." *Word* 10(2–3):146–162.
- Haveman, Heather A. and Mukti V. Khair. 2004. "Survival beyond Succession? The Contingent Impact of Founder Succession on Organizational Failure." *Journal of Business Venturing* 19(3):437–63.
- Haveman, Heather A. and Rachel Wetts. 2019. "Contemporary Organizational Theory: The Demographic, Relational, and Cultural Perspectives." *Sociology Compass* 13(3):e12664.
- JSTOR. 2018. "JSTOR Title Lists." Retrieved October 11, 2018 (<https://support.jstor.org/hc/en-us/articles/115007466248-JSTOR-Title-Lists>).

- Jurafsky, Daniel and James H. Martin. 2018. "19. Lexicons for Sentiment, Affect, and Connotation." *Speech and Language Processing (3rd Edition)*. Retrieved April 1, 2019 (<https://web.stanford.edu/~jurafsky/slp3/19.pdf>).
- Henrich, Joseph, Steven J. Heine, and Ara Norenzayan. 2010. "The Weirdest People in the World?" *Behavioral and Brain Sciences* 33(2-3):61-83.
- Iliev, Rumen, Morteza Dehghani, and Eyal Sagi. 2015. "Automated Text Analysis in Psychology: Methods, Applications, and Future Developments." *Language and Cognition* 7(2):265-90.
- Kanter, Rosabeth Moss. 1977. *Men and Women of the Corporation*. New York: Basic Books.
- Knorr-Cetina, Karen. 1999. *Epistemic Cultures: How the Sciences Make Knowledge*. Cambridge, MA: Harvard University Press.
- Kouloumpis, Efthymios, Theresa Wilson, and Johanna Moore. 2011. "Twitter Sentiment Analysis: The Good the Bad and the OMG!" in *Fifth International AAAI Conference on Weblogs and Social Media*.
- Kozlowski, Austin C., Matt Taddy, and James A. Evans. 2019. "The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings." *American Sociological Review* 84(5):905-49.
- Krosnick, Jon A. 1999. "Survey Research." *Annual Review of Psychology* 50(1):537-67.
- Kulkarni, Vivek, Rami Al-Rfou, Brian Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. *Proceedings of the 24th International Conference on World Wide Web*: 625-635. Geneva, Switzerland.
- Lareau, Annette. 2011. *Unequal Childhoods: Class, Race, and Family Life*. 2nd ed. Berkeley: University of California Press.
- Laver, Michael, and John Garry. 2000. Estimating policy positions from political texts. *American Journal of Political Science*, 44 (3): 619-634.
- Le, Quoc, and Tomas Mikolov. 2014. Distributed representations of sentences and documents. *Proceedings of the 31st International Conference on Machine Learning*, Beijing, China, June 22-24. *Journal of Machine Learning Research: Workshop and Conference Proceedings*, 32 (2): 1188-1196. (<https://arxiv.org/abs/1405.4053>)
- Levy, Omer, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the ACL*, 3: 211-225.
- Louwerse, Max M. 2004. "Semantic Variation in Idiolect and Sociolect: Corpus Linguistic Evidence from Literary Texts." *Computers and the Humanities* 38(2):207-21.
- Loughran, Tim and Bill McDonald. 2011. "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks." *The Journal of Finance* 66(1):35-65.
- Lowe, Will. 2008. Understanding wordscores. *Political Analysis*, 16: 356-371.
- Medin, Douglas L., Robert L. Goldstone, and Dedre Gentner. 1990. "Similarity Involving Attributes and Relations: Judgments of Similarity and Difference Are Not Inverses." *Psychological Science* 1(1):64-69.
- Meyer, John W., and Brian Rowan. 1977. Institutionalized organizations: Formal structure as myth and ceremony. *American Journal of Sociology*, 83: 340-363.
- Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. Paper presented at the *International Conference on Learning Representations*, Scottsdale, AZ, May 2-4. (<https://arxiv.org/abs/1301.3781>)
- Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, Lake Tahoe, CA (December 2013): 3111-3119.
- Miller, George A. 1985. "Wordnet: A Dictionary Browser." in *Proceedings of the First Conference of the UW Centre for the New Oxford Dictionary*. Waterloo, Canada: University of Waterloo.
- Monroe, Burt L., Michael P. Colaresi, and Kevin M. Quinn. 2009. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16: 372-403.
- Moody, James, and Ryan Light. 2006. A view from above: The evolving sociological landscape. *American Sociologist*, 37 (2): 67-86.
- Nectoux, François, and Fleur Thomése. 1999. Editorial. *European Journal of Social Quality*, 1: 3-11.
- Nelson, Laura K., Derek Burk, Marcel Knudsen, and Leslie McCall. 2018. The future of coding: A comparison of hand-coding and three types of computer-assisted text analysis methods. *Sociological Methods and Research*, forthcoming.
- Pennebaker, James W., Cindy K. Chung, Molly Ireland, Amy Gonzalez, and Roger J. Booth. 2007. *The Development and Psychometric Properties of LIWC2007*. (www.LIWC.net)
- Pennebaker, James W., Martha E. Francis, and Roger J. Booth. 2001. "Linguistic Inquiry and Word Count: LIWC 2001." *Mahway: Lawrence Erlbaum Associates* 71.

- Pennebaker, James W., Matthias R. Mehl, and Kate G. Niederhoffer. 2003. "Psychological Aspects of Natural Language Use: Our Words, Our Selves." *Annual Review of Psychology* 54(1):547–77.
- Pennebaker, James W. 1997. "Writing About Emotional Experiences as a Therapeutic Process." *Psychological Science* 8(3):162–66.
- Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. 2014. "GloVe: Global Vectors for Word Representation." Pp. 1532–1543 in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. "Deep Contextualized Word Representations." *North American Chapter of the Association for Computational Linguistics (NAACL)*.
- Pfeffer, Jeffrey. 1981. *Power in Organizations*. Cambridge, MA: Ballinger.
- Pfeffer, Jeffrey. 1983. Organizational demography. *Research in Organizational Behavior*, 5: 299-357.
- Pfeffer, Jeffrey, and Gerald R. Salancik. 1978. *The External Control of Organizations: A Resource Dependence Perspective*. New York: Harper and Row.
- Quinn, Kevin M., Burt L. Monroe, Michael Colaresi, Michael H. Crespin, and Dragomir R. Radev. 2010. How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, 54 (1): 209-228.
- Řehůřek, Radim and Petr Sojka. 2010. "Software Framework for Topic Modelling with Large Corpora." Pp. 45–50 in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*.
- Van Rossum, Guido and Fred L. Drake. 2011. *Python Language Reference Manual*. United Kingdom: Network Theory Limited.
- Salganik, Matthew J. 2017. *Bit by Bit: Social Research in the Digital Age*. Princeton University Press.
- Salton, Gerard and Christopher Buckley. 1988. "Term-Weighting Approaches in Automatic Text Retrieval." *Information Processing & Management* 24(5):513–523.
- Salton, G., A. Wong, and C. S. Yang. 1975. "A Vector Space Model for Automatic Indexing." *Communications of the ACM* 18(11):613–620.
- Sapir, Edward. 1958. *Culture, Language and Personality*. Berkeley: University of California Press.
- Schakel, Adriaan M. J. and Benjamin J. Wilson. 2015. "Measuring Word Significance Using Distributed Representations of Words." *ArXiv:1508.02297 [Cs]*.
- Schwartz, H. Andrew and Lyle H. Ungar. 2015. "Data-Driven Content Analysis of Social Media: A Systematic Overview of Automated Methods." *The ANNALS of the American Academy of Political and Social Science* 659(1):78–94.
- Scott, W. Richard. 1995 [2008]. *Institutions and Organizations*, 3rd Ed. Thousand Oaks, CA: Sage.
- Scully, Maureen, and W.E. Douglas Creed. 1997. Stealth legitimacy: Employee activism and corporate response during the diffusion of domestic partner benefits. Paper presented at the Academy of Management Meetings, Boston, MA.
- Shane, Janelle. 2019. *You Look Like a Thing and I Love You: How Artificial Intelligence Works and Why It's Making the World a Weirder Place*. Little, Brown.
- Sivak, Elizaveta and Ivan Smirnov. 2019. "Parents Mention Sons More Often than Daughters on Social Media." *Proceedings of the National Academy of Sciences* 116(6):2039–41.
- Sneffjella, Bryor, and Victor Kuperman. 2015. Concreteness and psychological distance in natural language use. *Psychological Science*, 26 (9): 1449-1460.
- Stone, Philip J., Dexter C. Dunphy, and Marshall S. Smith. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. Oxford, England: M.I.T. Press.
- Spirling, Arthur and Pedro L. Rodriguez. 2019. "Word Embeddings: What Works, What Doesn't, and How to Tell the Difference for Applied Research."
- Stone, M. 1974. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B (Methodological)*, 36 (2): 111-147. (<https://www.jstor.org/stable/2984809>)
- Tetlock, Paul C. 2007. Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance*, 62 (3): 1139-1168.
- TensorFlow. 2018. "Vector Representations of Words." *TensorFlow*. Retrieved September 21, 2018 (<https://www.tensorflow.org/tutorials/representation/word2vec>).

- Towns, John, Timothy Cockerill, Maytal Dahan, Ian Foster, Kelly Gaither, Andrew Grimshaw, Victor Hazlewood, Scott Lathrop, Dave Lifka, Gregory D. Peterson, Ralph Roskies, J. Ray Scott, and Nancy Wilkins-Diehr. 2014. "XSEDE: Accelerating Scientific Discovery." *Computing in Science & Engineering* 16(5):62–74.
- Tsai, Ming-Feng and Chuan-Ju Wang. 2014. "Financial Keyword Expansion via Continuous Word Vector Representations." Pp. 1453–1458 in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*.
- Tumasjan, Andranik, Timm Oliver Sprenger, Philipp G. Sandner, and Isabell M. Welp. 2010. "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment." *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media* 10(1):178–185.
- Turian, Joseph, Lev-Arie Ratinov, and Yoshua Bengio. 2010. "Word Representations: A Simple and General Method for Semi-Supervised Learning." Pp. 384–394 in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden: Association for Computational Linguistics.
- Turney, Peter D. 2006. "Similarity of Semantic Relations." *Computational Linguistics* 32(3):379–416.
- Turney, Peter D. and Patrick Pantel. 2010. "From Frequency to Meaning: Vector Space Models of Semantics." *Journal of Artificial Intelligence Research* 37:141–88.
- Vicinanza, Paul. 2019. "Quantifying Innovation with BERT: Linguistic Prescience and Firm Stock Returns." Presented at the Text Analysis Across Domains (TextXD) Symposium, December 5, University of California, Berkeley.
- Vilhena, Daril A., Jacob G. Foster, Martin Rosvall, Jevin D. West, James Evans, and Carl T. Bergstrom. 2014. Finding cultural holes: How structure and culture diverge in networks of scholarly communication. *Sociological Science*, 1: 221-238.
- Vogel, Rick. 2013. Visible colleges of management and organization studies: A bibliometric analysis of academic journals. *Organization Studies*, 33 (8): 1015-1043.
- Weber, Max. 1946. "Class, Status, Party." Pp. 180–95 in *From Max Weber: Essays in Sociology*. New York, NY: Oxford University Press.
- Watson, David and Lee Anna Clark. 1999. *The PANAS-X: Manual for the Positive and Negative Affect Schedule-Expanded Form*. University of Iowa.
- Whorf, Benjamin L. 1956. Science and linguistics. In John B. Carroll, ed., *Language, Thought and Reality: Selected Writings of Benjamin Lee Whorf*. Cambridge, MA: MIT Press.
- Wilson, Benjamin J. and Adriaan M. J. Schakel. 2015. "Controlled Experiments for Word Embeddings." *ArXiv:1510.02675 [Cs]*.
- Xing, Chao, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*, Denver, CO, May 31 – June 5: 1006-1011.
- Zhang, Yi, Jamie Callan, and Thomas Minka. 2002. Novelty and redundancy detection in adaptive filtering. *Proceedings of the 25th Annual International Association for Computing Machinery SIGIR Conference on Research and Development in Information Retrieval*: 81–88.

TABLES 2-4: Heatmaps of Pearson's Correlation Coefficients showing associations between scores from five text vectorization methods and citations of foundational texts for three organizational perspectives: demographic, relational, and cultural. All associations shown are statistically significant at the $p < 0.001$ level, except for `doc2vec` (not significant). Texts are coded with binary scheme: 1 if they at least once cite one or more foundational texts for a given perspective, 0 if not. The foundational texts are: for the demographic perspective, Hannan and Freeman (1977, 1984); for the relational perspective, Pfeffer and Salancik (1978); and for the cultural perspective, Meyer and Rowan (1977) and DiMaggio and Powell (1983). *Data source*: JSTOR Data-for-Research corpus from 1970 to 2015, filtered to 69,659 articles from journals in two disciplines: sociology and management/organizational behavior.

FIGURE 1: Proportion of variance that five text vectorization methods explain in citations of foundational texts for three organizational perspectives: demographic, relational, and cultural. Each logistic regression model is derived separately and includes controls for year and discipline together with the score for a given perspective from a single vectorization method (word count, etc.). Texts are coded with binary scheme: 1 if they at least once cite one or more foundational texts for a given perspective, 0 if not. The foundational texts are: for the demographic perspective, Hannan and Freeman (1977, 1984); for the relational perspective, Pfeffer and Salancik (1978); and for the cultural perspective, Meyer and Rowan (1977) and DiMaggio and Powell (1983). *Data source*: JSTOR Data-for-Research corpus from 1970 to 2015, filtered to 69,659 articles from journals in two disciplines: sociology and management/organizational behavior.

Appendix 1: Concept Dictionaries for Three Perspectives on Organizations

<i>Demographic</i>	<i>Relational</i>	<i>Cultural</i>
age dependence	board directors	ceremonial
birth rate	buffer	coercion
carrying capacity	coalition	coercive
chance survival	constrain	conform
competition	constraint	conformity
death rate	control	cultural
density	cooperate	decouple
ecological	cooperation	decoupled
ecology	coopt	diffuse
evolution	cooptation	diffusion
evolutionary	dependence	imitate
failure rate	dependent	imitation
fitness	director interlock	innovation
founding rate	director interlocks	institutional
generalism	diversification	institutionalize
generalist	diversify	institutionalize
inertia	dominance	interorganizational field
inertia	exchange	isomorphic
legitimacy	external	isomorphism
legitimate	horizontal	legitimacy
legitimation	influence	legitimate
liability newness	interdependence	legitimation
natural selection	interdependent	loosely coupled
niche	interlock	mimetic
organizational form	interlocking	norm
population	joint venture	normative
population ecology	merge	norms
reliability	merged	organizational field
resistance change	merger	profession
selection	network	professional
selection	network	rationalize
size dependence	power	rationalized
specialism	pressure	ritual
specialist	resource dependence	socially constructed
structural inertia	sanction	structuration
survival chance	vertical	taken granted

Appendix 2: Notes on Raw Data from JSTOR Data-for-Research

As future researchers following our workflow may wish to know, the raw data from JSTOR came in the form of 14 zip files, each about 1.5GB in size or 4.5 GB when unzipped. Thus, about 400,000 compositions from academic journals downloaded as 21 GB and extracted as 63 GB. Making up these raw JSTOR files is article metadata (e.g., publication date, author name, journal title), counts of n-grams (single words and phrases of two or three words), and a full-text file created through Optimal Character Recognition or OCR (in *.txt format). While the OCR files are in raw format, JSTOR preprocessed the text data prior to counting n-grams. In our word counts we used bigrams and trigrams as well as unigrams because bigrams and trigrams capture complex ideas better than single words (Vilhena et al. 2014).

Specifically, they turned upper-case letters into lower-case letters, removed punctuation marks (including hyphens at the end of a line), replaced symbols such as hyphens and parentheses with white spaces, concatenated word sections separated by apostrophes, discarded common or stop words, and tokenized (but did not stem or lemmatize) the text using the Apache Lucene Standard Tokenizer according to standard unicode text segmentation practices.¹² The specific stop words removed were: “a”, “an”, “and”, “are”, “as”, “at”, “be”, “but”, “by”, “for”, “if”, “in”, “into”, “is”, “it”, “no”, “not”, “of”, “on”, “or”, “such”, “that”, “the”, “their”, “then”, “there”, “these”, “they”, “this”, “to”, “was”, “will”, and “with”. Words were left in their original form rather than being stemmed (usually this means removal of word endings to leave a common base, e.g. “walking” and “walks” both reduce to “walk”) or lemmatized (using morphology to

¹² For more details on preprocessing, see <https://www.jstor.org/dfr/about/technical-specifications>; https://lucene.apache.org/solr/guide/6_6/understanding-analyzers-tokenizers-and-filters.html; http://unicode.org/reports/tr29/#Word_Boundaries.

index a word to an entry in a detailed dictionary, e.g. “studies” and “studying” are both linked to “study”).

Regarding identifying the primary subject of JSTOR compositions, we primarily relied on the second author’s expert knowledge of the field. Some journals are listed under two or more subjects; for example, *Administrative Science Quarterly* is listed under both sociology and management and organizational behavior. Due to her familiarity with this publication, however, the second author resolved this case by categorizing *Administrative Science Quarterly* as having management and organizational behavior as its primary subject. Sometimes journal titles clearly indicated the primary subject: e.g., the primary subject for *Industrial and Labor Relations Review* is labor and employment relations—even though the journal is also listed under sociology and economics. A few titles required the second author to peruse journal contents or founding editorial statements. For instance, we categorized the *Journal of European Social Quality* under sociology rather than political science because its founding editors focused on “contemporary social issues in Europe,” a combination of “the wellbeing of the individual person on the one side, and social cohesion, integration, and participation on the other” (Nectoux and Thomése 1999: 3), topics that are closer to sociology than political science. Finally, if a journal spanned subjects covered by this analysis and subjects outside this analysis (e.g., history and sociology)—and if it had a significant footprint in one of the two subjects covered by this analysis—then we categorized it as being under a subject within this analysis.

Appendix 3: Notes on word2vec and doc2vec

While a few alternative model architectures have been proposed for `word2vec`, in this paper I use the “continuous skip-gram” model, which is the most accurate option for semantic comparisons (Mikolov, Chen, et al. 2013)—the focus of this paper—and is best suited for large data sets (TensorFlow 2018) like our academic articles corpus. In technical terms, continuous skip-gram seeks to classify a word given each other word in its context, allowing each word vector to have a separate impact on its neighbors. This is in contrast to the other principal word embedding architecture, the “continuous bag-of-words” model, which predicts target words based on their context, averaging all context word vectors, which treats the entire context as a single observation.

Again like `word2vec`, `doc2vec` also offers alternative models: the “distributed bag of words” model, which is similar to continuous skip-gram except it uses the sentence vector to predict the words around a given focal word; and the “distributed memory model”, which is similar to continuous bag of words except that it incorporates the sentence vector when predicting the focal word in a given context (Le and Mikolov 2014). While the creators of this technique suggest combining these modeling approaches, in preliminary studies we found that performing this more complex combination gave less accurate results than using the distributed memory model alone. Thus, in this paper we use the distributed memory model.

In our implementation of both `word2vec` and `doc2vec`, we take advantage of several extensions to the original continuous skip-gram model: noise reduction, undersampling of frequent words, and detection of common phrases in the corpus (Mikolov, Sutskever, et al. 2013). Common phrases, also called multi-word expressions, often possess unique meanings compared to their individual words; for instance, compare the meaning of the phrase “age dependent” to the individual terms “age” and “dependent”. Our approach accepts phrases so

long as their ‘collocation score’—the ratio of a given word pair’s collocations divided by the product of each word’s individual appearances—exceeds some threshold. Mathematically, this score is derived as follows:

$$\text{score}(w_i, w_j) = \frac{\text{count}(w_i w_j) - \delta}{\text{count}(w_i) \times \text{count}(w_j)}$$

where w_i and w_j are two words in the corpus, $w_i w_j$ is their collocation, and δ is a “discounting coefficient” that makes infrequent phrases less likely (Mikolov, Sutskever, et al. 2013:6).

For purposes of efficiency, for both `word2vec` and `doc2vec` we use the `gensim` defaults for some parameters: 10000 words per sample, 5 noise words (for noise reduction), an initial learning rate of 0.025, a discounting coefficient of 3, and a phrase detection threshold of 10. To improve the model’s ability to capture semantic nuances—without inducing unnecessary computational burden—we expand the defaults for a few other parameters: word context windows of size 10 (to better capture syntactic relations: Mikolov, Chen, et al. 2013; Spirling and Rodriguez 2019; also found optimal for `doc2vec` in Le and Mikolov 2014), 50 iterations through the data and a hidden layer of size 300 (the most common choice). As suggested in Mikolov et al. (2013:4), we use a negative sampling exponent of 0.75 to subsample frequent words.

Appendix 4: Notes on Recurrent Neural Networks

A recurrent neural network (RNN) is a neural network architecture that specializes in processing sequential data. In language, the order of the words can significantly alter meaning. For example, “I had cleaned my house” has a different meaning from “I had my house cleaned”. Using the RNN framework allows us to extract the difference in the meanings of these two sentences by processing the sequence of text segments: phrases, sentences, or documents. RNNs perform the same task for every element of an inputted sequence, with each step’s output dependent on the previous computations and steps. This allows the network to compute a final semantic output which contextualizes each of inputted words’ semantic meanings.

This diagram displays a simplified RNN structure:¹³

Box A represents the neural network. The input goes through the neural network and is passed onto the next iteration to an awaiting neural network, where the next input is added, and so on. The model is iterative in that it takes a new input each time; at the final iteration, it produces an output which references all previous inputs and therefore is a holistic representation of the sequence of text inputs.

An RNN goes through the following steps to create its output:

¹³ Image source: <http://colah.github.io/posts/2015-08-Understanding-LSTMs/>

- 1) The model initializes the weight matrices using random distribution and bias with zeros.
- 2) The function of $z = W_{(1)}x$, where W is the weight matrix and x is the input matrix, is used to compute predictions. An activation function is used to calculate the output, z . Finally, output values are generated using $o = W_{(2)}h$, where h is the previously outputted values from the activation function (softmax, sigmoid, etc). This whole process is often referred to as forward propagation.
- 3) The predictions are then passed into a loss function to measure the accuracy of the predicted output in relation to the category or label the model seeks to predict. This can be denoted as $L = l(\hat{y}, y)$, where $\hat{y}$ is the generated output from step 2 (denoted as o) and y is the label being predicted. The function l is the loss function itself, and cross-entropy loss function is one of the most commonly used functions, denoted by
$$L = -\sum_i y_i \log(\hat{y}_i),$$
 where $\hat{y}$ and y represent the predicted value via output and the actual label of the inputted data (phrases).
- 4) The focus of each epoch becomes the minimization of the loss function, $L = l(\hat{y}, y)$. The minimization of the loss function will be in the context of our weights, W_1 and W_2 , the core components of the functions in step 2 used to predict the outputs.
- 5) Using the two calculated derivatives at step 4 and predetermined learning rate value η , weights get updated as $W_i = W_i - \eta \frac{\partial L}{\partial W_i}$. This step is known as back-propagation.
- 6) Steps 2-5 are repeated until the predetermined number of iterations/epochs are reached. The number of epochs are usually above $n = 50$ in order to ensure that the constantly updated weight values converge to the representative, relative accurate values.

Appendix 5: Inference

Facebook's `InferSent` research focuses on developing a pre-trained recurrent neural network that can be used with any input of words, sentences, or documents (Conneau et al., 2017). The Stanford Natural Language Inference (SNLI) corpus is used to train and optimize the recurrent neural network. The SNLI corpus consists of 570,000 human-written sentence pairs, with each sentence pair containing a premise sentence (the contextual sentence itself) and a hypothesis sentence aiming to predict/summarize the sentence. These pairs are then classified under 3 different labels - Contradiction, Neutral, and Entailment.

There are many standard RNN architectures available, and the `InferSent` team (Conneau et al. 2017) tests many of these architectures as well as other neural network architectures, including Long Term Short Memory (LSTM), Gated Recurrent Units (GRU), and Hierarchical Convolutional Neural Network. They find that the bidirectional LSTM architecture with max-pooling creates the best accuracy results for classifying the SNLI corpus labels, and makes the pre-trained encoder for this architecture available. The structure of bidirectional LSTM architecture with max-pooling is similar to that of the recurrent neural network discussed in Appendix 3. However, this LSTM architecture is bidirectional, meaning that the encoder will first take the word embeddings in the sequence of the words and output embeddings then take the reverse sequence of the words and output another set of embeddings, ultimately concatenating the two output embeddings to ensure that the output from the encoder gets a full understanding of the context of the words in the inputted document, similar to the way in which skip-gram uses $w_{t-2}, w_{t-1}, w_{t+1}, w_{t+2}$ to predict w_t , given a window parameter of $n = 2$.

Additionally, this LSTM architecture uses max-pooling, choosing the maximum value for each dimension of the hidden units in the LSTM layer.

`InferSent` specifies certain parameters they used for training. They start with the learning rate of 0.1, and learning rate is divided by 5 for the next training iteration if the training accuracy decreases. Training stops when learning rate falls below the threshold of 10^{-5} . Weight decay of 0.99 is used as a regularization measure.

The training optimizes the encoder and decoder to perform best in semantically differentiating sentences in context of premise-hypothesis sentence pairing and identifying three specific semantic categories: entailment, neutral, and contradiction. The encoder is then made available as a pre-trained encoder optimized to take any sentences and output homogeneous embeddings which accurately represents the positive, negative, and neutral contexts of the sentences. (The decoder was used for back-propagation and weight updates purposes only and serves no functionality in generating sentence semantics.)

The training of the encoder follows the 4 higher level steps, repeated until learning rate falls under the specified threshold:

- 1) *Sentence preprocessing and standardization*. The premise sentence and hypothesis sentence are prepared and inputted into separate encoders. For each of the sentences, the underlying words are standardized using word embeddings (`GloVe` vectors trained on Common Crawl 840B with 300 dimensions). As recurrent neural network takes numeric matrix as an input, the standardization ensures that homogeneous scale of values are inputted into the encoders.

- 2) *Embeddings generation.* Standardized word embeddings are inputted into the respective encoder and its hidden layers. The hidden layers are then max-pooled, producing a final embedding/vector of size 1×4096 for each of the two sentences.
- 3) *Combination of the premise and hypothesis sentences.* It is important to create a vector representative of the premise sentence, the hypothesis sentence, and the contextual similarity of the two sentences. With u being the premise sentence vector and v being the hypothesis sentence vector generated from step 2, the two vectors are concatenated, denoted as (u, v) , multiplied as element-wise product, denoted as $u * v$, and subtracted as an absolute element-wise difference, denoted as $|u - v|$. These three components are then concatenated into a single vector.
- 4) The resulting vector from step 3 is inputted into a multilayer perceptron (MLP) with 1 hidden layer of 512 hidden units, with 3-way softmax layer for activation, producing $\hat{y}$. Steps 4 and 5 from Appendix 5 are followed: calculating the loss value, finding derivatives (both for the encoder and the MLP decoder? Need to clarify - go to ML OH at wayfair), and updating weights through back-propagation.