People-Centric Natural Language Processing

David Bamman

CMU-LTI-15-007

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213
www.lti.cs.cmu.edu

Thesis committee

Noah Smith (chair), Carnegie Mellon University
Justine Cassell, Carnegie Mellon University
Tom Mitchell, Carnegie Mellon University
Jacob Eisenstein, Georgia Institute of Technology
Ted Underwood, University of Illinois

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
In Language and Information Technologies

Contents

Acknowledgments

The work presented in this thesis addresses the complexity of the interaction between people and text. The words on these pages naturally hold a similarly complex relationship; while I am their author, I count my words lucky to be influenced by many people: most immediately, my advisor, Noah Smith, and my collaborators Justine Cassell, Chris Dyer, Jacob Eisenstein, Tom Mitchell, Brendan O'Connor, Tyler Schnoebelen and Ted Underwood. In my time at Carnegie Mellon, you have each shaped my thinking profoundly; the work in this thesis has benefited not only from the clarity and depth of your thoughts on it, but also in the interests and ideals you have each inspired in me.

There are many others whose influence I am grateful to acknowledge: my colleagues Waleed Ammar, Adam Anderson, Dallas Card, Dipanjan Das, Jesse Dodge, Jeff Flanigan, Alona Fyshe, Kevin Gimpel, Dirk Hovy, Dan Jurafsky, David Kaufer, Charles Kemp, Lingpeng Kong, Fei Liu, Rohan Ramanath, Alan Ritter, Carolyn Rosé, Bryan Routledge, Nathan Schneider, Cosma Shalizi, Yanchuan Sim, Swabha Swayamdipta, Sam Thomson, Chris Warren, Tae Yano and Dani Yogatama; and those who have set me on this path and sustained me there—Greg Crane, David Smith and David Mimno especially (who have not only blazed the trail from Classics to computer science that I've followed, but continue to inspire by pushing it further beyond).

Above all, I am fortunate to be influenced by my family—Leo, you've only witnessed the last few months of this process, but I have written many of these words with you sleeping on my chest. And most of all, Kate: you have been a constant source of support throughout these years, and I am lucky to have you by my side. This thesis would simply not exist without you.

The material in this thesis was supported by U.S. National Science Foundation grants IIS-0915187, IIS-1251131, IIS-1211277 and CAREER IIS-1054319, Google's support of the Reading is Believing project at CMU, the ARCS Foundation, and IARPA via DoI/NBC contract number D12PC00337. This work was made possible through the use of computing resources made available by the Pittsburgh Supercomputing Center and the Open Science Data Cloud (OSDC), an Open Cloud Consortium (OCC)-sponsored project.

Chapter 1

Introduction

The written text that we interact with on an everyday basis—news articles, emails, social media, books—is the product of a profoundly social phenomenon with people at its core. With few exceptions, all of the text we see is written by people, and others constitute its audience. A vast amount of the content itself is centered on people: news (including classic NLP corpora such as the *Wall Street Journal*) details the roles of actors in current events, social media (including Twitter and Facebook) documents the actions and attitudes of friends, and books chronicle the stories of fictional characters and real people alike.

Robust text analysis methods provide us one way to understand or synthesize this volume of text without reading all of it; commercial and popular successes like IBM's Watson and Apple's Siri hinge on robust computational models of naturally occurring data. Computational models for linguistic analysis to date have largely focused on *events* as the organizing concept for representing text meaning. This is evident in many of the major trends in computational semantic analysis: frame semantics and semantic role labeling (Gildea and Jurafsky, 2002; Palmer et al., 2005; Das et al., 2010); information extraction into structured databases (Hobbs et al., 1993; Banko et al., 2007; Mitchell et al., 2015); and semantic parsing models based on truth-conditional semantics (Zelle and Mooney, 1996; Zettlemoyer and Collins, 2005). In such methods, *what* happens (or is true) is central, and *who* is involved is represented by a string, perhaps with a type,[1] and in a few cases by an identifier linking into a database. Considerable work has led to advances in resolving coreference of those strings (Bagga and Baldwin, 1998; Haghighi and Klein, 2009; Raghunathan et al., 2010, among others), and into resolving them to catalogs of real-world entities such as Freebase or Wikipedia

[1] For example, Person, Organization, Location, and Miscellaneous comprise a nominal taxonomy widely used in named entity recognition; knowledge bases also implicitly perform fine-grained typing when asserting IS-A relationships among people.

(Bunescu and Pasca, 2006; Cucerzan, 2007).

People, however, are not entries in relational databases. Their attributes, motivations, intentions, etc., cannot be stated with perfect objectivity, and so meaningful descriptions of *who* a person is must be qualified by the source of the description: who authors the description, as well as *their* attributes, motivations and intentions. This thesis explores a new approach to modeling and processing natural language that transforms the primitives of linguistic analysis—namely, from *events* to *people*—in anticipation of more "socially aware" language technologies.

In this thesis, I build NLP around people instead of events, developing methods that consider the interaction of author, audience and content in text analysis. In this work, I adopt a statistical point of view and consider the text we observe to be a **random variable** whose realization depends on several, often interacting, factors. When we read a sentence in the *New York Times* such as *Obama signed the legislation*, the action denoted by *signing legislation* is heavily dependent on the fact that the syntactic subject (and semantic agent) is *Obama*, a POLITICIAN. If *Obama* were a FIREFIGHTER instead, we would be much less likely to observe the phrase *signed legislation* with him, and more likely to observe *put out the fire*. These are not categorical decisions (we cannot rule out a firefighter signing legislation, or a president extinguishing a fire) but rather statistical regularities in the actions, attributes, and—at the most fundamental level—words associated with people. One source of variation we observe in text can in part be explained by knowing something of the category to which an individual described belongs.

A second source of variation in this random variable of text is due to the author. While the actions associated with *Obama* become more predictable once we know he is a POLITICIAN, the choice that individuals make in declaring that *Obama is a great president* as opposed to *Obama is a dictator* is of course dependent on their own political beliefs. If we know those beliefs, or can estimate them (for example, by observing that a sentence originates from Fox News or MSNBC), then we can, in part, offer an explanation for that variability.

A third source of variation in the text we see is due to its audience. Politicians are not the only individuals who change their message and even their regional accents (NPR, 2015) in response to a changing audience. While formal publications like the *New York Times* and the *Wall Street Journal* write with their readership in mind, so too do everyday users on Twitter, Facebook and other social media sites, often tailoring the content of their text to a specific audience (Goffman, 1959; Bell, 1984). Knowing something about who that audience is, and how they change, can help explain the variation we see.

Seeing text as a random variable gives us the machinery we need to reason about the mechanisms that govern it. While we may not directly observe the fact that *Obama* is a POLITICIAN or be able to measure the political beliefs of authors, we can reason about them by virtue of the statistical regularities they admit. These regularities between the text we observe with people in each of their roles as *content*, *author* and *audience* constitute the subject of this thesis.

1.1 Structure of this thesis

This thesis is structured around each of the axes on which people interact with text. Each chapter generally follows the same structure: I propose how a corpus of textual data can be used to tell us something about people in one of these roles; I build a new probabilistic model of that data that specifies the relationships between the observed evidence and the latent structure we are trying to learn; and I evaluate the fitness of that model in comparison to others. This work broadly falls into two main classes: exploring variation in how people are represented as content *within* text, and exploring variation in text as a function of variation in the authors and audience *outside* of it.

1.1.1 Variation in content

The first source of variation comes in how people are depicted in text. Chapters 3 and 4 consider representations of people through the lens of fine-grained entity types or **personas**, such as HERO, VILLAIN, VEGETARIAN, MUSICIAN and FIREFIGHTER. Modeling personas has the potential to tap into humans' natural tendencies to abstract and generalize about each other and our relationships and also—perhaps more importantly—to help bring those tendencies to light, supporting both literary studies and social-scientific research that uses text as data. In these chapters, the categories of variation are latent, and inferred through statistical regularities in the kinds of actions that people are described as performing in text; the latent structure of a persona is a generative variable that helps explain the data we see—bundles of actions bound together by people.

Chapter 5 leverages people as they are depicted in text as an organizing principle for a downstream application: learning event classes such as BIRTH, GRADUATING HIGH SCHOOL and MARRIAGE from biographies on Wikipedia, using similar statistical machinery as for persona inference above. This work also occasions social insight: though it is known that women are greatly underrepresented on Wikipedia—not only as editors (Wikipedia, 2011) but also as subjects of articles (Reagle and Rhue, 2011)—I find that there is a bias in their *characterization* as well, with biographies of women containing significantly more emphasis on events of marriage and divorce than biographies of men.

1.1.2 Variation in author and audience

The second source of variation we see reflected in text is due to authorship: different authors, with different attributes—such as gender, age, political preference—shape their text in ways that are predictable as a function of this variation. Indeed, this predictability has inspired a rich literature on inferring latent user attributes from the text they write (Herring and Paolillo, 2006; Koppel et al., 2006; Argamon et al., 2007; Mukherjee and Liu, 2010; Rosenthal and McKeown, 2011; Rao et al., 2010; Golbeck et al., 2011; Burger et al., 2011; Pennacchiotti and Popescu, 2011; Conover et al., 2011; Volkova et al., 2014)

In chapters 6 and 7, I leverage this variation for two downstream tasks: a.) estimating the political orientation of fine-grained opinions like OBAMA IS A SOCIALIST; and b.) improving word embeddings by making them sensitive to local geographical meanings of words (such as the different senses of *wicked* in Kansas and Massachusetts). In this work, I consider variation that is observed (such as geographical location of authors), and variation that is unobserved (the latent political leanings of authors) but still inferable through the statistical regularities of the data we see—bundles of propositions bound together by the people who assert them.

The third and last source of variation we see reflected in text is due to that of the audience: the same author can adapt their message in different ways as a function of who they're addressing—whether through the mechanism of self-presentation (Goffman, 1959), audience design (Bell, 1984) or the indexicality of linguistic variation (Eckert, 2008; Johnstone and Kiesling, 2008).

In chapter 8, I consider this kind of audience-based variation (in addition to authorial variation) through a case study of *sarcasm*: while most approaches to detecting this richly contextual phenomenon rely on lexical indicators (such as interjections and intensifiers) and other linguistic markers (such as nonveridicality and hyperbole), I consider information from a variety of sources, including not simply the content of the message, but properties of the interaction between an author and audience, leveraging Kreuz's (1996) "principle of inferability"—that speakers only use sarcasm if they can be sure it will be understood by the audience. I find that adding any kind of contextual information—information about the author, the recipient of the message, or their interaction—can help in predictive performance for this complex task.

1.2 Evaluation

Many tasks in natural language processing have well-established standard metrics for evaluation: part-of-speech tagging and phrase-structure syntactic parsing have a standard human-created reference corpus in the Penn Treebank (Marcus et al., 1993); other tasks like coreference resolution have set evaluation corpora such as OntoNotes (Hovy et al., 2006) but many evaluation metrics, such as B^3 (Bagga and Baldwin, 1998), MUC (Vilain et al., 1995) and CEAF (Luo, 2005); others—such as machine translation or summarization—have standard metrics but no fixed reference corpus.

In this thesis, I am often proposing new lines of research into questions that do not have pre-existing benchmarks for evaluation. Many of the sections below present models, often unsupervised, for exploratory data analysis. This requires constant assessment for how those models should be evaluated. While perplexity (a measure of how likely a sample of data is under a model) has been used for task-agnostic evaluation of unsupervised models in the past, recent work has shown that it is often uncorrelated with the ultimate goal of a model (or worse, at odds with it), such as the human interpretability of topics in topic models (Chang et al., 2009) or out-of-sample word error rate in speech recognition (Iyer et al., 1997; Chen et al., 1998). Part of the contribution this thesis makes is in giving careful assessment to what constitutes a good evaluation for each approach.

For each section in this thesis, I evaluate the models directly on real predictive tasks. In all cases, these evaluations judge the fitness of a model on its performance against a human standard or in a predictive task with gold standard data.

- Chapter 3 compares the performance of models on the task of measuring the overlap between automatically inferred clusters of fictional characters and human created clusters.

- Chapter 4 compares different model predictions of *a priori* hypotheses about the similarity between fictional characters to human expert opinions.

- Chapter 5 compares models' performance at predicting the time in a person's life when an event described in text takes place (leveraging a model's belief of its latent event class to do so).

- Chapter 6 compares the model predictions of the political import of propositions like OBAMA IS A SOCIALIST (learned from estimates of political forum users' latent political positions) to human judgments of the same.

- Chapter 7 compares the performance of different models on the predictive task of geographically-influenced term similarity.

- Chapter 8 compares the performance of different models on the predictive task of sarcasm detection in social media text.

In designing evaluations specific to each proposed task, one goal is to allow a range of other methods to be more directly comparable to each other. While assessments of perplexity are limited to generative models that assign probabilities to observations, the evaluations proposed throughout this thesis can equally be applied to discriminative models and to models that are not inherently probabilistic at all. Where applicable, I also release evaluation data for others to compare to.

1.3 Thesis statement

In this thesis, I advocate for a model of text analysis that focuses on people, leveraging ideas from machine learning, the humanities and the social sciences. People intersect with text in multiple ways: they are its authors, its audience, and often the subjects of its content. I argue that developing computational models that capture the complexity of their interaction will yield deeper, socio-culturally relevant descriptions of these actors, and that these deeper representations will open the door to new NLP and machine learning applications that have a more useful understanding of the world.

I explore this perspective by designing, implementing and evaluating computational models of three kinds: a.) unsupervised models of personas, which capture patterns of identity and behavior in the description of people as the **content** of text; b.) unsupervised models of author variation, which capture patterns in how latent and observed qualities of the **author** influence the text we see; and c.) models of **audience** variation, which capture patterns in how variation in the audience can influence the text we see. Each of these research fronts captures one dimension of how people interact with each other as mediated through text. Together, these three axes define a coordinate system for investigating written language in its socially embedded context. At a large scale, this thesis illustrates how organizing data around **people** and reasoning about the subtleties of their interaction with text can both generate new social insight and improve performance on practical tasks.

Chapter 2

Methods

2.1 Probabilistic graphical models

In viewing text as a random variable whose realization is dependent on many factors, the work presented in this thesis largely exploits the machinery of probabilistic graphical models, often in an unsupervised setting where our quantity of interest is never observed. Graphical models provide a powerful computational framework; by clearly delineating the exact relationships between all of the variables we consider (which include both observed data and presumed hidden structure), we clearly articulate our statistical assumptions and have access to a wide range of established inference techniques, including variational methods (Jordan et al., 1999) and Markov chain Monte Carlo (MCMC) techniques like Gibbs sampling (Geman and Geman, 1984; Casella and George, 1992; Griffiths and Steyvers, 2004) and Metropolis-Hastings (Metropolis et al., 1953; Hastings, 1970). For a broad overview of graphical models, see Koller and Friedman (2009).

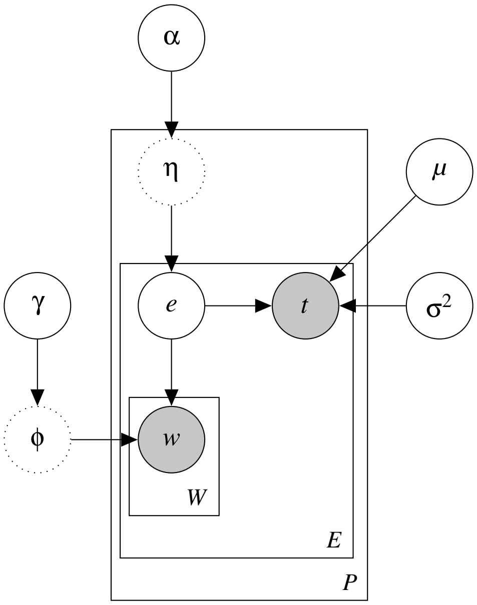
Text analysis using probabilistic generative models dates at least to Mosteller and Wallace (1963), who model document counts in the *Federalist Papers* as draws from a Poisson or negative binomial distribution in order to determine authorship. More recently, unsupervised models have become popular; topic models (Blei et al., 2003) treat documents as multinomial mixtures of latent topics (one per word), and have spawned a thriving industry: direct descendants of this work include models that add correlation among topics (Blei and Lafferty, 2006a), network structure (Chang and Blei, 2009), supervision (Mcauliffe and Blei, 2008; Ramage et al., 2010), time (Wang and McCallum, 2006) and sequence (Griffiths et al., 2005; Gruber et al., 2007). Topic models have been useful for organizing large digital libraries (Mimno and McCallum, 2007), tracing the history of a discipline through its journals (Goldstone and Underwood, 2014), identifying themes in 19th-century literature (Jockers and Mimno, 2013) and characterizing the agendas of politicians through their press releases (Grimmer, 2010).

The core idea behind graphical models is very simple: they articulate the relationships among variables, and specify, in their structure, the conditions under which one variable is able to exert influence on another. The art of designing models includes defining the variables of interest—both observational data (such as text and metadata) and latent variables (entity types, event classes, political beliefs), and positing their conditional dependencies in such a way as to allow interesting structure to be discovered and testable hypotheses to emerge.

As an example, figure 2.1 illustrates one of the simplest models described in this thesis: a Bayesian model for learning a set of latent event classes from words and timestamps divided among people—in this particular case, biographical subjects on Wikipedia (see §5 for more detail). Throughout this thesis, I use the standard convention of representing variables as circles (observed variables are shaded, latent variables are clear, and those that are collapsed out during inference are dotted) and plates to denote multiple variables. Arrows indicate direct statistical dependence.

In this particular model, the observed data we see are sets of "events"—defined as a set of terms w (such as the individual words in *Barack Obama became president at the age of*) and an observed timestamp t (*48*). What we don't see are the things we care about—a latent event class indicator e for each event (e.g., BECOMING PRESIDENT), a distribution over the these latent events for a particular person η (for Barack Obama, this might include BECOMING PRESIDENT, SERVING AS SENATOR, BIRTH, etc.), along with a characterization ϕ of each event class in terms of the words that are the most likely to appear in describing it. In this case, we mediate the relationship between all of the words in an event and the timestamp that appears with them through the latent structure of an event class. While we never directly see that *Barack Obama became president at the age of 48* is an instance of a BECOMING PRESIDENT event (which also appears in the biographies of Ronald Reagan, Theodore Roosevelt, François Hollande and others), knowing that it denotes such an event makes all of the observed data we see much more likely. What we seek in unsupervised models are the values of the latent variables we don't see that maximize the likelihood of the data we do.

Throughout this work, I rely primarily on the inference technique of Gibbs sampling (Geman and Geman, 1984), a Markov chain Monte Carlo method for approximating the joint distribution of a set of variables (both observed and latent) by sequentially sampling them. Gibbs sampling is an iterative technique in which the value of all unobserved variables



Var	Distribution	Description
e	~Cat	event class
t	~Norm	timestamp of event
w	~Cat	term in event
η	~Dir	subject's distribution over event class
ϕ	~Dir	Event class distribution over terms
μ	parameter	Event class mean
σ	parameter	Event class standard dev.
α, γ	parameter	Dirichlet concentration
W	observed	Number of words for event e
E	observed	Number of events for person p
P	observed	Number of people

Figure 2.1: Top: Example model. **Bottom:** Definition of variables.

in a model are sampled in sequence conditioned on the current values of all the others. Due to the conditional independencies that are inherent in the structure of the model, the conditional probability for a variable X is only dependent on the values of variables in its Markov blanket (Pearl, 1988)—the set of variables in the neighborhood of X that, when conditioned upon, make X independent of all *other* variables in the model; in the example above, a sample for a certain event class e is only dependent on the current samples of its associated timestamp t, the words associated with it w, the event distribution η and the event class characterization ϕ. All other words and timestamps associated with any *other* event description are independent of this variable when the others are conditioned upon (and thereby have no effect on the probability of any outcome). Throughout this work, I use *collapsed* Gibbs sampling Griffiths and Steyvers (2004), where sets of variables are integrated out of the model completely (as with ϕ and η above), yielding a different Markov blanket but the same principle: in practice, where Y is the set of variables that define the Markov blanket of X, this involves calculating the conditional probability of X from the full joint probability $P(X, Y)$, and then drawing a sample from that conditional distribution. In the case of *uncollapsed* Gibbs sampling for the model above, this reduces to an application of Bayes' rule:

$$P(e|\eta, t, w, \phi, \mu, \sigma^2) = \frac{P(e \mid \eta) P(t \mid e, \mu, \sigma^2) \prod P(w \mid e, \phi)}{\sum_{e'} P(e' \mid \eta) P(t \mid e', \mu, \sigma^2) \prod P(w \mid e', \phi)} \tag{2.1}$$

For *collapsed* sampling, the Markov blanket for a particular event class includes many other variables in the model (since integrating out ϕ and η induces dependencies on all other variables through α and γ). In this case, the full joint probability becomes more complicated (see chapter 5), but the fundamental principle remains the same: we can use this joint probability to calculate the conditional probability for a target variable and draw a sample from that conditional distribution. For more information about the details (and derivation) of Gibbs sampling in probabilistic models, see Carpenter (2010) and Resnik and Hardisty (2010).

2.2 Linguistic structure

Much work in using probabilistic models of text for exploratory data analysis uses a bag-of-words assumption in the representation of documents, treating words as simple and independent tokens; from a generative standpoint, this is a often well-motivated decision since words are the only thing we can directly observe.

But text does have structure, even if it is unobserved. At a most shallow level, words cluster together into non-compositional multiword expressions (Sag et al., 2002), so that a term like *white house* can denote something more significant than the sum of its parts. At a deeper, and unobserved, level, phrases bear syntactic relationships to each other (Chomsky, 1965), so that a sentence can be decomposed into a product of subjects, direct objects, attributes, and so on; under a dependency grammar (Sgall et al., 1986; Mel'čuk, 1988), these structural relationships hold between individual words. Beyond syntax, we can also see higher-order structural relations between words and phrases at the level of semantics and discourse.

The broad field of natural language processing deals with uncovering this kind of hid-

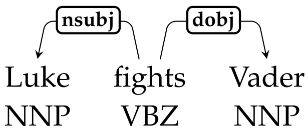


Figure 2.2: Sample syntactic dependency graph with part-of-speech tags for *Luke fights Vader.*

den structure in text, and has developed methods for part-of-speech-tagging (Ratnaparkhi et al., 1996; Toutanova et al., 2003; Søgaard, 2011), syntactic parsing (Collins, 2003; McDonald et al., 2005; McClosky et al., 2006; Petrov et al., 2006; Nivre et al., 2007b; Socher et al., 2013a), semantic parsing (Zettlemoyer and Collins, 2005; Das et al., 2014; Flanigan et al., 2014), and discourse parsing (Marcu, 2000; Baldridge and Lascarides, 2005; Sagae, 2009; Ji and Eisenstein, 2014) among many others. For a general overview, see Jurafsky and Martin (2009). The work presented here draws on this tradition, representing documents using linguistic structure. At its simplest, this amounts to uncovering multiword expressions (chapters 5 and 7) in the linear sequence of text; for the models presented in chapters 3 and 4, the observations we see are semantic dependencies (predicates along with their semantic agents and patients) associated with characters in movies and books (as illustrated in figure 2.2); this association of characters with their syntactic paths draws on rich techniques for clustering mentions of people into a single character and resolving coreference between them (e.g., resolving *Tom*, *Mr. Sawyer*, and *him* to the character known as TOM SAWYER). For the models presented in chapter 6, the observations we see are sentences that have been decomposed into propositional subjects and predicates (such as *Obama is a Socialist* → ⟨Obama, *is a Socialist*⟩).

In the framework of probabilistic models, we can think of this linguistic structure as amounting to meaningful prior information, accumulated through years of theoretical work and practical application. We know that words are not independent, and there exists a rich structure that binds them. By leveraging our best prediction as to what this structure is, we can reason over a finer and more nuanced representation of our data, and allow a kind of analysis that would not be possible if we treated all words as simple independent strings. By representing documents through their linguistic abstraction, we also gain statistical strength, requiring less data for accurate learning.

2.3 Conditioning on metadata

While much early work in the probabilistic modeling of text draws words as categorical variables from a flat multinomial distribution, one motif throughout this thesis is that language is profoundly situated—it does not arrive to us *ex nihilo* but is rather spoken or written at a particular time and place and by a particular person—and we can often condition on metadata associated with that situated context in order to influence the probabilities in a generative language model. This context can be either latent or observed. As one example, in §6, I parameterize the probability of a particular subject (like *gun rights*) used by an individual u as the exponentiated sum of a background log frequency of that subject in the corpus overall (m_{sbj}) and K real-valued additive effects, normalized over the space of S possible subjects (given a parameter matrix $\beta \in \mathbb{R}^{K \times S}$):

$$P(sbj \mid u, \eta, \beta, m_{sbj}) = \frac{\exp\left(m_{sbj} + \sum_{k=1}^{K} \eta_{u,k}\beta_{k,sbj}\right)}{\sum_{sbj'} \exp\left(m_{sbj'} + \sum_{k=1}^{K} \eta_{u,k}\beta_{k,sbj'}\right)} \tag{2.2}$$

In this, $\eta \in \mathbb{R}^K$ is a K-dimensional real-valued representation of the political preferences of a specific user—the situated context in which the text is embedded. They are latent (and so must be inferred), but our current estimate of their values influences the probability of observing the subjects we see; if η for a particular user falls into a "Republican" region of $\mathbb{R}^K$, the probability of that user saying the phrase *gun rights* in a proposition will likely go up.

In conditioning on metadata in this way, I draw most directly on work into sparse additive generative models (SAGE) of Eisenstein et al. (2011a) and also on a series of work that originates in the trigger n-gram language models of Rosenfeld (1996), which allowed the incorporation of long-distance information, such as previously-mentioned words, into maximum-entropy *n*-gram models. This work has since been extended to a Bayesian setting by applying both a Gaussian prior (Chen and Rosenfeld, 2000), which dampens the impact of any individual feature, and sparsity-inducing priors (Kazama and Tsujii, 2003; Goodman, 2004), which can drive many feature weights to 0. Other more recent work in this space includes the structural topic model (Roberts et al., 2014), inverse regression topic model (Rabinovich and Blei, 2014) and multinomial inverse text regression (Taddy, 2013). Much recent work in neural language modeling (Bengio et al., 2003; Mnih and Hinton, 2009; Mikolov et al., 2013) also falls in this class, by virtue of conditioning on words in context to influence language model probabilities.

As in the example above, the models in §6 condition on a user's latent political position to influence the words we see; the models in §7 condition on a user's geographical location to shape a language model by influencing low-dimensional word-representations; and the models in §4 condition on author identity in order to learn a set of character types that are able to discount the specific stylistic influence of the author. While the goals for including such metadata are different in each case, the principle is the same throughout.

2.4 Notation in this thesis

Throughout this thesis, I use the following notation consistently for all probabilistic graphical models:

- w represents an observed word in text (chapters 3, 4, 5 and 6)
- r represents an observed semantic role (chapters 3 and 4)
- z is a latent word-level topic (chapter 3)
- p represents a latent, single-membership entity type for an individual (chapters 3, 4 and 6)
- θ, ϕ and ψ represent latent multinomial distributions (chapters 3, 4, 5 and 6)
- β and ξ represent the parameters in a log-linear distribution (chapters 3, 4 and 6)
- μ and σ represent the mean and variance of Normal distributions (chapters 3, 5 and 6)
- α, γ and ν are Dirichlet hyperparameters (chapters 3, 4, 5 and 6)

While the precise meaning of these variables for each model is defined in more detail in the chapters below, these are consistent guidelines that govern their use throughout this work.

Part I

Variation in content

Overview

The first section of this thesis covers the depiction of people as *content* in text, and the statistical regularities in that depiction that allow us to infer higher-order information about them. Statistical regularities in how people are presented in text have been helpful in the past for two standard NLP tasks: named entity recognition and coreference resolution. In the former, statistical models are able to leverage informative contextual features in order to classify entities into one of a range of possible pre-existing categories; at the coarsest granularity, this includes the standard four-way CoNLL 2003 (Tjong Kim Sang and De Meulder, 2003) classification of PERSON, LOCATION, ORGANIZATION and MISC; finer-grained systems often draw supervision from Wikipedia and classify entities into an ontology containing anywhere from eight fine-grained types (Fleischman and Hovy, 2002) up to several thousand (Bunescu and Pasca, 2006). At its most extreme, fine-grained NER reaches its limit in named entity linking (Cucerzan, 2007), where mentions of individuals in text are resolved to the Wikipedia pages (or other gazetteers) of the individual they refer to.

State-of-the-art coreference systems incorporate gender information in order to link noun phrases and pronominal mentions together, either in the form of explicit rules (Raghunathan et al., 2010) or by learning surface contextual features that encode it (Durrett and Klein, 2013). Entity-centric approaches to coreference (Rahman and Ng, 2009; Haghighi and Klein, 2010; Durrett et al., 2013) also reason about document-level information about each entity being resolved, allowing the incorporation of regularities in fine-grained types. This kind of global, entity-centric modeling has been useful for other tasks as well, such probabilistic frame induction (Chambers, 2013).

The work presented here illustrates the value of pushing this exploration further. In chapter 3, I introduce the problem of *unsupervised persona inference*, the task of learning a set of latent entity types (or *personas*) from the actions and attributes associated with people (in this particular case, characters in plot descriptions of movies on Wikipedia). In chapter 4, I extend these models to enable persona inference in the presence of stylistic heterogeneity within a collection of 15,000 literary novels, giving the flexibility to draw on different theoretical assumptions that privilege learning different classes of entity types. In chapter 5, I show that modeling the variation that exists in descriptions of people online (in this case, Wikipedia biographies) both provides a means of learning fine-grained *events* (where we can exploit the structure that exists between events and the people who experience them) and also gives insight into the biases that exist in the characterization of people in this socially constructed space. Each of these three sections is only possible by privileging *people* as the central organizing principle of the data.

Chapter 3

Learning personas in movies

Work described in this chapter was undertaken in collaboration with Brendan O'Connor and Noah Smith, and published at ACL 2013 (Bamman et al., 2013)

3.1 Introduction

In the first two chapters of this thesis, I consider variation in the actions and attributes associated with fictional characters in movies (this chapter) and in books (chapter 4). Much computational work involving fictional texts of this kind has focused on narrative (Mani, 2012), attempting to learn the sequence of events by which a story is defined; in this tradition we might situate seminal work on learning procedural scripts (Schank and Abelson, 1977; Regneri et al., 2010), narrative chains (Chambers and Jurafsky, 2008), and plot structures (Finlayson, 2011; Elsner, 2012; McIntyre and Lapata, 2010; Goyal et al., 2010).

In this chapter, I introduce a complementary, people-centric perspective that addresses the importance of *character* in defining a story, leveraging the variation that emerges in descriptions of people in text to infer latent *character types* that they embody. The testbed for this chapter is film. This fictional domain provides an intuitive starting point for thinking about abstract character types. The notion that a fixed set of character types recur throughout narratives is common to structuralist theories of sociology, anthropology and literature (Campbell, 1949; Jung, 1981; Propp, 1968; Frye, 1957; Greimas, 1984); in this view, the role of a character in a story is less a depiction of an imagined person (with a real personality) and more a narrative *function*. A Campbellian view of narrative may see a protagonist depicted as the THE HERO, and other characters whose sole function is to offer guidance and training for them (the MENTOR) or to employ cunning as a way of advancing the plot (the TRICKSTER). A Proppian view of narrative likewise sees grand types such as the THE HERO and THE VILLAIN, and also many more specialized characters requisite in Russian folktales (such as THE DONOR, whose sole narrative function is to give THE HERO a magical object).[1] While such structuralist theories of narrative have tended to be eclipsed over the past fifty years by materialist theories that concentrate on the historical context in which a narrative in produced, they form a very living means by which contemporary audiences organize their perception across a variety of different media: the BYRONIC HERO, for example, represents a specialized type of the brooding, mysterious loner, and has been used to describe Heathcliff in *Wuthering Heights*, Edward Cullen in the *Twilight* books and movies, and Angel in the television series of the same name (Stein, 2004).

Under this perspective, a character's latent internal nature helps drive the action we observe. This leads to a natural generative story: we first decide that we're going to make a particular kind of movie (e.g., a romantic comedy), then decide on a set of character types, or *personas*, we want to see involved (the PROTAGONIST, the LOVE INTEREST, the BEST FRIEND). After picking this set, we fill out each of these roles with specific attributes (*female*, *28 years old*, *klutzy*); with this cast of characters, we then sketch out the set of events by which they interact with the world and with each other (*runs but just misses the train*, *spills coffee on their boss*) – through which they reveal to the viewer those inherent qualities about themselves. This work is inspired by past approaches that infer typed semantic arguments along with narrative schemas (Chambers and Jurafsky, 2009; Regneri et al., 2011), but seeks a more holistic view of character, one that learns from stereotypical attributes in addition to plot events. This work also naturally draws on earlier work on the unsupervised learning of verbal arguments and semantic roles (Pereira et al., 1993; Grenager and Manning, 2006; Titov and Klementiev, 2012) and unsupervised relation discovery (Yao et al., 2011).

This people-centric perspective leads to two natural questions. First, can we learn what those standard personas are by how individual characters (who instantiate those types) are portrayed? Second, can we learn the set of attributes and actions by which we recognize those common types? How do we, as viewers, recognize a VILLIAN?

3.2 Data

3.2.1 Text

Our primary source of data to answer these questions in this chapter comes from 42,306 movie plot summaries extracted from the November 2, 2012 dump of English-language

[1] Woloch (2003) provides one vivid example of characters defined by their structural relations in the "twelve young men" Achilles murders in *The Iliad* as revenge for his companion Patroklus' death (*Il.* 21.97–113); we know nothing of these characters outside of their function as THOSE KILLED IN REVENGE.

Wikipedia.[2] These summaries, which have a median length of approximately 176 words,[3] contain a concise synopsis of the movie's events, along with implicit descriptions of the characters (e.g., "rebel leader Princess Leia," "evil lord Darth Vader"). To extract structure from this data, we use the Stanford CoreNLP library[4] to tag and syntactically parse the text, extract entities, and resolve coreference within the document. With this structured representation, we extract linguistic features for each character, looking at immediate verb governors and attribute syntactic dependencies to all of the entity's mention headwords, extracted from the typed dependency tuples produced by the parser; we refer to "CCprocessed" syntactic relations described in de Marneffe and Manning (2008):

- **Agent verbs.** Verbs for which the entity is an agent argument (*nsubj* or *agent*).
- **Patient verbs.** Verbs for which the entity is the patient, theme or other argument (*dobj*, *nsubjpass*, *iobj*, or any prepositional argument *prep_**).
- **Attributes.** Adjectives and common noun words that relate to the mention as adjectival modifiers, noun-noun compounds, appositives, or copulas (*nsubj* or *appos* governors, or *nsubj, appos, amod, nn* dependents of an entity mention).

These three roles capture three different ways in which character personas are revealed: the actions they take on others, the actions done to them, and the attributes by which they are described. For every character we thus extract a bag of (r, w) tuples, where w is the word lemma and r is one of $\{\text{agent verb}, \text{patient verb}, \text{attribute}\}$ as identified by the above rules.

3.2.2 Metadata

Our second source of information consists of character and movie metadata drawn from the November 4, 2012 dump of Freebase.[5] At the movie level, this includes data on the language, country, release date and detailed genre (365 non-mutually exclusive categories, including "Epic Western," "Revenge," and "Hip Hop Movies"). Many of the characters in movies are also associated with the actors who play them; since many actors also have detailed biographical information, we can ground the characters in what we know of those real people – including their gender and estimated age at the time of the movie's release (the difference between the release date of the movie and the actor's date of birth).

[2] http://dumps.wikimedia.org/enwiki/

[3] More popular movies naturally attract more attention on Wikipedia and hence more detail: the top 1,000 movies by box office revenue have a median length of 715 words.

[4] http://nlp.stanford.edu/software/corenlp.shtml

[5] http://download.freebase.com/datadumps/

Across all 42,306 movies, entities average 3.4 agent events, 2.0 patient events, and 2.1 attributes. For all experiments described below, we restrict our dataset to only those events that are among the 1,000 most frequent overall, and only characters with at least 3 events. 120,345 characters meet this criterion; of these, 33,559 can be matched to Freebase actors with a specified gender, and 29,802 can be matched to actors with a given date of birth. Of all actors in the Freebase data whose age is given, the average age at the time of movie is 37.9 (standard deviation 14.1); of all actors whose gender is known, 66.7% are male.[6] The age distribution is strongly bimodal when conditioning on gender: the average age of a female actress at the time of a movie's release is 33.0 (s.d. 13.4), while that of a male actor is 40.5 (s.d. 13.7).

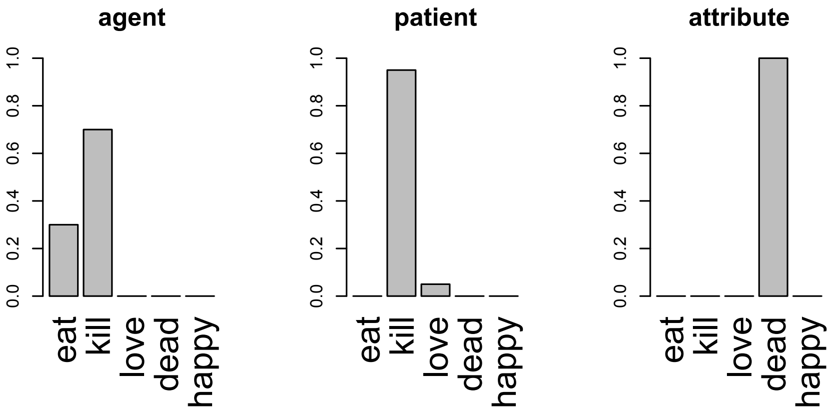


Figure 3.1: A *persona* is a set of three distributions over latent topics. In this toy example, the ZOMBIE persona is primarily characterized by being the agent of words from the *eat* and *kill* topics, the patient of *kill* words, and is attributively modified by words from the *dead* topic.

3.3 Personas

One way we recognize a character's latent type is by observing the stereotypical actions they perform (e.g., VILLAINS *strangle*), the actions done to them (e.g., VILLAINS are *foiled* and *arrested*) and the words by which they are described (VILLAINS are *evil*). To capture this intuition, we define a *persona* as a set of three typed distributions: one for the words for which the character is the agent, one for which it is the patient, and one for words by which the character is attributively modified. Each distribution ranges over a fixed set of latent word classes, or *topics*. Figure 3.1 illustrates this definition for a toy example: a ZOMBIE persona may be characterized as being the agent of primarily *eating* and *killing* actions, the

[6] Whether this extreme 2:1 male/female ratio reflects an inherent bias in film or a bias in attention on Freebase (or Wikipedia, on which it draws) is an interesting research question in itself.

patient of *killing* actions, and the object of *dead* attributes. The topic labeled *eat* may include words like *eat*, *drink*, and *devour*.

3.4 Models

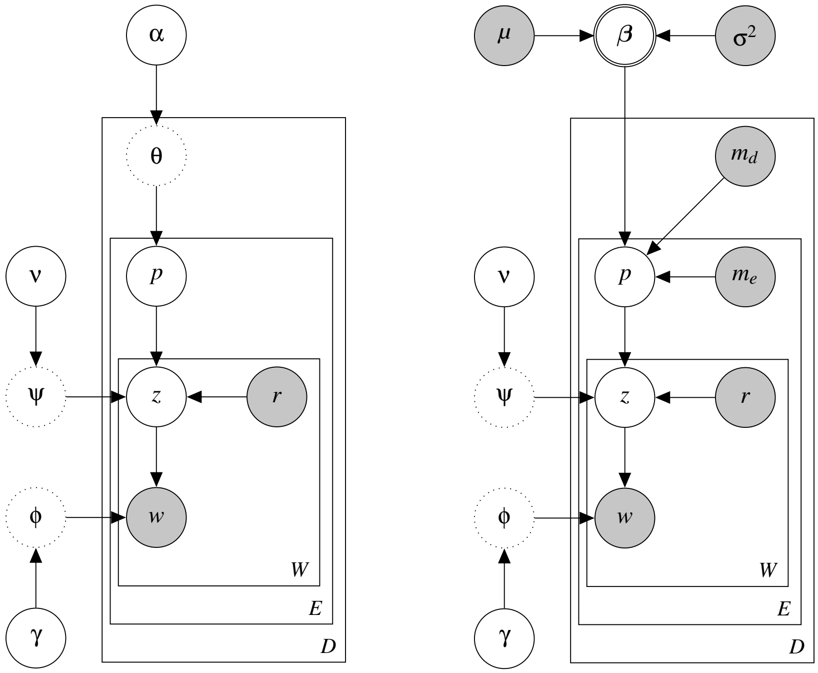
Both models that we present here simultaneously learn three things: 1.) a soft clustering over words to topics (e.g., the verb "strangle" is mostly a type of *Assault* word); 2.) a soft clustering over topics to personas (e.g., VILLIANS perform a lot of *Assault* actions); and 3.) a hard clustering over characters to personas (e.g., Darth Vader is a VILLAIN.) They each use different evidence: since our data includes not only textual features (in the form of actions and attributes of the characters) but also non-textual information (such as movie genre, age and gender), we design a model that exploits this additional source of information in discriminating between character types; since this extra-linguistic information may not always be available, we also design a model that learns only from the text itself. We present the text-only model first for simplicity. Throughout, V is the word vocabulary size, P is the number of personas, and K is the number of topics.

3.4.1 Dirichlet Persona Model

In the most basic model, we only use information from the structured text, which comes as a bag of (r, w) tuples for each character in a movie, where w is the word lemma and r is the relation of the word with respect to the character (one of agent verb, patient verb or attribute, as outlined in §3.2.1 above). The generative story runs as follows. First, let there be K latent word topics; as in LDA (Blei et al., 2003), these are words that will be soft-clustered together by virtue of appearing in similar contexts. Each latent word cluster $\phi_k \sim \mathrm{Dir}(\gamma)$ is a multinomial over the V words in the vocabulary, drawn from a Dirichlet parameterized by γ. Next, let a persona p be defined as a set of three multinomials ψ_p over these K topics, one for each typed role r, each drawn from a Dirichlet with a role-specific hyperparameter (ν_r).

Every document (a movie plot summary) contains a set of characters, each of which is associated with a single latent persona p; for every observed (r, w) tuple associated with the character, we sample a latent topic k from the role-specific $\psi_{p,r}$. Conditioned on this topic assignment, the observed word is drawn from ϕ_k. The distribution of these personas for a given document is determined by a document-specific multinomial θ, drawn from a Dirichlet parameterized by α.

Figure 3.2 (above left) illustrates the form of the model. To simplify inference, we collapse out the persona-topic distributions ψ, the topic-word distributions ϕ and the persona



(a) DIRICHLET PERSONA MODEL. **(b)** PERSONA REGRESSION.

P	Number of personas (hyperparameter)
K	Number of word topics (hyperparameter)
D	Number of movie plot summaries
E	Number of characters in movie d
W	Number of (role, word) tuples used by character e
ϕ_k	Topic k's distribution over V words.
r	Tuple role: agent verb, patient verb, attribute
$\psi_{p,r}$	Distribution over topics for persona p in role r
θ_d	Movie d's distribution over personas
p_e	Character e's persona (integer, $p \in \{1..P\}$)
j	A specific (r, w) tuple in the data
z_j	Word topic for tuple j
w_j	Word for tuple j
α	Concentration hyperparameter for Dirichlet model
β	Feature weights for regression model
μ, σ^2	Gaussian mean and variance (for regularizing β)
m_d	Movie features (from movie metadata)
m_e	Entity features (from movie actor metadata)
ν_r, γ	Dirichlet concentration hyperparameter

Figure 3.2: Models and definition of variables.

distribution θ for each document. Inference on the remaining latent variables – the persona p for each character type and the topic z for each word associated with that character – is conducted via collapsed Gibbs sampling (Griffiths and Steyvers, 2004).

We optimize the values of the Dirichlet hyperparameters α, ν and γ using slice sampling with a uniform prior every 20 iterations for the first 500 iterations, and every 100 iterations thereafter. After a burn-in phase of 10,000 iterations, we collect samples every 10 iterations (to lessen autocorrelation) until a total of 100 have been collected.

3.4.2 Persona Regression

To incorporate observed metadata in the form of movie genre, character age and character gender, we adopt an "upstream" modeling approach (Mimno and McCallum, 2008), letting those observed features influence the conditional probability with which a given character is expected to assume a particular persona, prior to observing any of their actions. This captures the increased likelihood, for example, that a 25-year-old male actor in an action movie will play an ACTION HERO than he will play a VALLEY GIRL.

To capture these effects, each character's latent persona is no longer drawn from a document-specific Dirichlet; instead, the P-dimensional simplex is the output of a multiclass logistic regression, where the document genre metadata m_d and the character age and gender metadata m_e together form a feature vector that combines with persona-specific feature weights to form the following log-linear distribution over personas, with the probability for persona k being:

$$P(p = k \mid m_d, m_e, \beta) = \frac{\exp([m_d; m_e]^\top \beta_k)}{1 + \sum_{j=1}^{P-1} \exp([m_d; m_e]^\top \beta_j)} \tag{3.1}$$

The persona-specific β coefficients are learned through stochastic EM (Wei and Tanner, 1990), in which we alternate between the following:

1. Given current values for β, for all characters e in all plot summaries, sample values of p_e and z_j for all associated tuples.

2. Given input metadata features m and the associated sampled values of p, find the values of β that maximize the standard multiclass logistic regression log likelihood, subject to squared ℓ_2 regularization.

Figure 3.2 (above right) illustrates this model. As with the Dirichlet persona model, inference on both p and z is conducted with collapsed Gibbs sampling. We optimize β every 1,000 iterations, until a burn-in phase of 10,000 iterations has been reached; at this point we following the same sampling regime as for the Dirichlet persona model.

3.5 Evaluation

We evaluate our methods in two quantitative ways by measuring the degree to which we recover two different sets of gold-standard clusterings. This evaluation also helps offer guidance for model selection (in choosing the number of latent topics and personas) by measuring performance on an objective task.

3.5.1 Character Names

First, we consider all character names that occur in at least two separate movies, generally as a consequence of remakes or sequels; this includes proper names such as "Rocky Balboa," "Oliver Twist," and "Indiana Jones," as well as generic type names such as "Gang Member" and "The Thief"; to minimize ambiguity, we only consider character names consisting of at least two tokens. Each of these names is used by at least two different characters; for example, a character named "Jason Bourne" is portrayed in *The Bourne Identity*, *The Bourne Supremacy*, and *The Bourne Ultimatum*. While these characters are certainly free to assume different roles in different movies, we believe that, in the aggregate, they should tend to embody the same character type and thus prove to be a natural clustering to recover. 970 character names occur at least twice in our data, and 2,666 individual characters use one of those names. Let those 970 character names define 970 unique gold clusters whose members include the individual characters who use that name.

3.5.2 TV Tropes

As a second external measure of validation, we consider a manually created clustering presented at the website TV Tropes,[7] a wiki that collects user-submitted examples of common tropes (narrative, character and plot devices) found in television, film, and fiction, among other media. While TV Tropes contains a wide range of such conventions, we manually identified a set of 72 tropes that could reasonably be labeled character types, including THE CORRUPT CORPORATE EXECUTIVE, THE HARDBOILED DETECTIVE, THE JERK JOCK, THE KLUTZ and THE SURFER DUDE.

We manually aligned user-submitted examples of characters embodying these 72 character types with the canonical references in Freebase to create a test set of 501 individual characters.

While the 72 character tropes represented here are a more subjective measure, we expect to be able to at least partially recover this clustering.

[7] `http://tvtropes.org`

		Character Names §3.5.1			TV Tropes §3.5.2		
K	Model	$P = 25$	$P = 50$	$P = 100$	$P = 25$	$P = 50$	$P = 100$
25	Persona regression	7.73	7.32	6.79	6.26	6.13	5.74
	Dirichlet persona	7.83	7.11	6.44	6.29	6.01	5.57
50	Persona regression	7.59	7.08	6.46	6.30	5.99	5.65
	Dirichlet persona	7.57	7.04	6.35	6.23	5.88	5.60
100	Persona regression	7.58	6.95	6.32	6.11	6.05	5.49
	Dirichlet persona	7.64	6.95	6.25	6.24	5.91	5.42

Table 3.1: Variation of information between learned personas and gold clusters for different numbers of topics K and personas P. Lower values are better. All values are reported in bits.

3.5.3 Variation of Information

To measure the similarity between the two clusterings of movie characters, gold clusters $\mathcal{G}$ and induced latent persona clusters $\mathcal{C}$, we calculate the variation of information (Meilă, 2007):

$$VI(\mathcal{G}, \mathcal{C}) = H(\mathcal{G}) + H(\mathcal{C}) - 2I(\mathcal{G}, \mathcal{C}) \tag{3.2}$$

$$= H(\mathcal{G}|\mathcal{C}) + H(\mathcal{C}|\mathcal{G}) \tag{3.3}$$

VI measures the information-theoretic distance between the two clusterings: a lower value means greater similarity, and VI $= 0$ if they are identical. Low VI indicates that (induced) clusters and (gold) clusters tend to overlap; i.e., knowing a character's (induced) cluster usually tells us their (gold) cluster, and vice versa. Variation of information is a metric (symmetric and obeys triangle inequality), and has a number of other desirable properties.

Table 3.1 presents the VI between the learned persona clusters and gold clusters, for varying numbers of personas ($P = \{25, 50, 100\}$) and topics ($K = \{25, 50, 100\}$). To determine significance with respect to a random baseline, we conduct a *permutation test* (Fisher, 1935; Pitman, 1937) in which we randomly shuffle the labels of the learned persona clusters and count the number of times in 1,000 such trials that the VI of the observed persona labels is lower than the VI of the permuted labels; this defines a nonparametric p-value. All results presented are significant at $p < 0.001$ (i.e. observed VI is never higher than the simulation VI).

Over all tests in comparison to both gold clusterings, we see VI improve as both P and, to a lesser extent, K increase. While this may be expected as the number of personas increase to match the number of distinct types in the gold clusters (970 and 72, respectively), the fact that VI improves as the number of latent topics increases suggests that more fine-grained

		Character Names §3.5.1		
K	Model	$P = 25$	$P = 50$	$P = 100$
25	Persona regression	62.8 (↑41%)	59.5 (↑40%)	53.7 (↑33%)
	Dirichlet persona	54.7 (↑27%)	50.5 (↑26%)	45.4 (↑17%)
50	Persona regression	63.1 (↑42%)	59.8 (↑42%)	53.6 (↑34%)
	Dirichlet persona	57.2 (↑34%)	49.0 (↑23%)	44.7 (↑16%)
100	Persona regression	63.1 (↑42%)	57.7 (↑39%)	53.0 (↑34%)
	Dirichlet persona	55.3 (↑30%)	49.5 (↑24%)	45.2 (↑18%)

		TV Tropes §3.5.2		
K	Model	$P = 25$	$P = 50$	$P = 100$
25	Persona regression	42.3 (↑31%)	38.5 (↑24%)	33.1 (↑25%)
	Dirichlet persona	39.5 (↑20%)	31.7 (↑28%)	25.1 (↑21%)
50	Persona regression	42.9 (↑30%)	39.1 (↑33%)	31.3 (↑20%)
	Dirichlet persona	39.7 (↑30%)	31.5 (↑32%)	24.6 (↑22%)
100	Persona regression	43.5 (↑33%)	32.1 (↑28%)	26.5 (↑22%)
	Dirichlet persona	39.7 (↑34%)	29.9 (↑24%)	23.6 (↑19%)

Table 3.2: Purity scores of recovering gold clusters. Higher values are better. Each absolute purity score is paired with its improvement over a controlled baseline of permuting the learned labels while keeping the cluster proportions the same.

topics are helpful for capturing nuanced character types.[8]

The difference between the persona regression model and the Dirichlet persona model here is not significant; while VI allows us to compare models with different numbers of latent clusters, its requirement that clusterings be mutually informative places a high overhead on models that are fundamentally unidirectional (in Table 3.1, for example, the room for improvement between two models of the same P and K is naturally smaller than the bigger difference between different P or K). While we would naturally prefer a text-only model to be as expressive as a model that requires potentially hard to acquire metadata, we tease apart whether a distinction actually does exist by evaluating the purity of the gold clusters with respect to the labels assigned them.

3.5.4 Purity

For gold clusters $\mathcal{G} = \{g_1 \ldots g_k\}$ and inferred clusters $\mathcal{C} = \{c_1 \ldots c_j\}$ we calculate purity as:

$$\text{Purity} = \frac{1}{N} \sum_k \max_j |g_k \cap c_j| \tag{3.4}$$

[8] This trend is robust to the choice of cluster metric: here VI and F-score have a correlation of -0.87; as more latent topics and personas are added, clustering improves (causing the F-score to go up and the VI distance to go down).

While purity cannot be used to compare models of different persona size P, it can help us distinguish between models of the same size. A model can attain perfect purity, however, by placing all characters into a single cluster; to control for this, we present a controlled baseline in which each character is assigned a latent character type label proportional to the size of the latent clusters we have learned (so that, for example, if one latent persona cluster contains 3.2% of the total characters, the probability of selecting that persona at random is 3.2%). Table 3.2 presents each model's absolute purity score paired with its improvement over its controlled permutation (e.g., ↑41%).

Within each fixed-size partition, the use of metadata yields a substantial improvement over the Dirichlet model, both in terms of absolute purity and in its relative improvement over its sized-controlled baseline. In practice, we find that while the Dirichlet model distinguishes between character personas in different movies, the persona regression model helps distinguish between different personas within the same movie.

3.6 Exploratory Data Analysis

As with other generative approaches, latent persona models enable exploratory data analysis. To illustrate this, we present results from the persona regression model learned above, with 50 latent lexical classes and 100 latent personas. Figure 3.3 visualizes this data by focusing on a single movie, *The Dark Knight* (2008); the movie's protagonist, Batman, belongs to the same latent persona as Detective Jim Gordon, as well as other action movie protagonists Jason Bourne and Tony Stark (*Iron Man*). The movie's antagonist, The Joker, belongs to the same latent persona as Dracula from *Van Helsing* and Colin Sullivan from *The Departed*, illustrating the ability of personas to be informed by, but still cut across, different genres. Table 3.3 (at the end of this chapter) presents an exhaustive list of all 50 topics, along with an assigned label that consists of the single word with the highest PMI for that class. Of note are topics relating to romance (*unite, marry, woo, elope, court*), commercial transactions (*purchase, sign, sell, owe, buy*), and the classic criminal schema from Chambers (2011) (*sentence, arrest, assign, convict, promote*).

Table 3.4 presents the most frequent 14 personas in our dataset, illustrated with characters from the 500 highest grossing movies. The personas learned are each three separate mixtures of the 50 latent topics (one for agent relations, one for patient relations, and one for attributes), as illustrated in figure 3.1 above. Rather than presenting a 3 × 50 histogram for each persona, we illustrate them by listing the most characteristic topics, movie characters, and metadata features associated with it. Characteristic actions and features are defined as those having the highest smoothed pointwise mutual information with that class; exem-

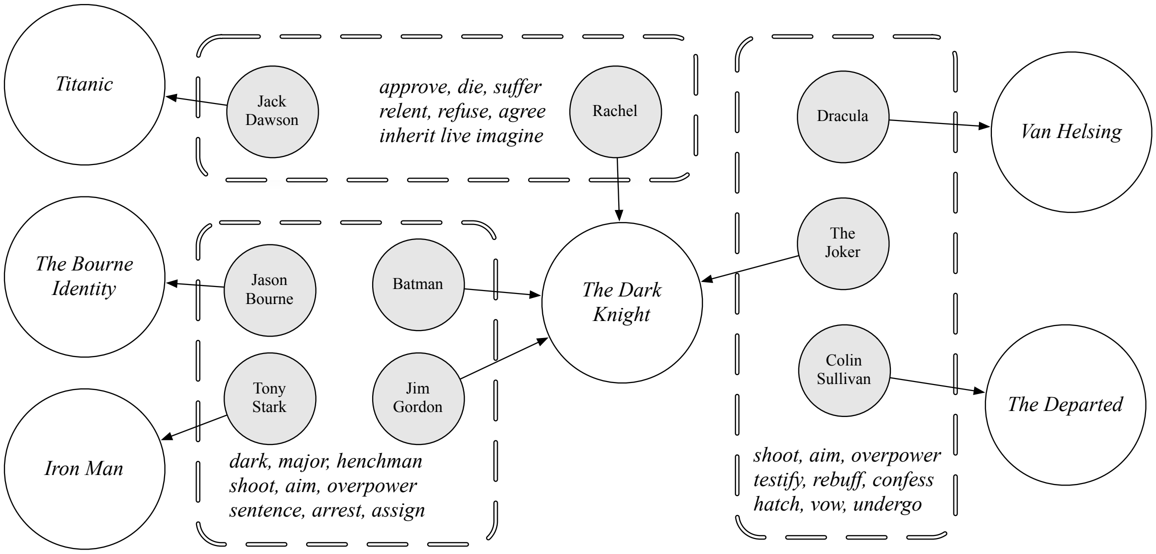


Figure 3.3: Dramatis personae of *The Dark Knight* (2008), illustrating 3 of the 100 character types learned by the persona regression model, along with links from other characters in those latent classes to other movies. Each character type is listed with the top three latent topics with which it is associated.

plary characters are those with the highest posterior probability of being drawn from that class. Among the personas learned are canonical male action heroes (exemplified by the protagonists of *The Bourne Supremacy*, *Speed*, and *Taken*), superheroes (*Hulk*, *Batman and Robin*, Hector of *Troy*) and several romantic comedy types, largely characterized by words drawn from the FLIRT topic, including *flirt*, *reconcile*, *date*, *dance* and *forgive*.

3.7 Conclusion and Future Work

This chapter introduces a method for automatically inferring latent character personas from text (and metadata, when available). In leveraging the statistical regularities in the actions and attributes with which characters are described in plot summaries, we are able to learn a set of coherent types through which those regularities can be explained. While the goal of this work has been to induce a set of latent character classes and partition all characters among them, there are several interesting questions that remain as directions for future work. One is in looking at how a specific character's actions may informatively be at odds with their inferred persona, given the choice of that persona as the single best fit to explain the actions we observe. By examining how any individual character deviates from the behavior indicative of their type, we might be able to paint a more nuanced picture of how a character can embody a specific persona while resisting it at the same time. Second, this work makes several strong modeling assumptions: we assign a single persona to a character that is valid throughout the entirety of a movie; a more realistic assumption may see personas as temporary masks that can be taken on and off, and a single character may transition through multiple roles (and there may be correlations that exist between those personas, and common pathways through them). Another modeling assumption we make in this work is that movies have multinomial distributions over latent character types, which allows multiple characters to embody the same persona; it may be more realistic to model personas through other distributions that strongly disprefer multiple HEROS, for instance. Rather than having personas drawn from a flat multinomial—where each persona is independent of each other—we may want to infer a *hierarchy* of personas, such as a single top-level PROTAGONIST class from which other, more fine-grained types inherit. Third, while we have adopted a people-centric perspective here in privileging individuals as an organizing principle for data, we only consider individuals in isolation from each other; a more realistic assumption here may see personas as partly defined by dyadic (or higher-order) relations among pairs (or sets) of people—so that a VILLAIN, for example, is only a seen as such through the lens of a PROTAGONIST. Fourth, while the testbed for this work is film, the textual data we draw on comes from a stylistically homogenous corpus (summaries of plots written on Wikipedia), where we can reasonably attribute the variation we see to latent factors other than individual stylistic variation; how we can extend this model to corpora that are stylistically *heterogenous* is the subject of the next chapter.

Label	Most characteristic words	Label	Most characteristic words
UNITE	unite marry woo elope court	SWITCH	switch confirm escort report instruct
PURCHASE	purchase sign sell owe buy	INFATUATE	infatuate obsess acquaint revolve concern
SHOOT	shoot aim overpower interrogate kill	ALIEN	alien child governor bandit priest
EXPLORE	explore investigate uncover deduce	CAPTURE	capture corner transport imprison trap
WOMAN	woman friend wife sister husband	MAYA	maya monster monk goon dragon
WITCH	witch villager kid boy mom	INHERIT	inherit live imagine experience share
INVADE	invade sail travel land explore	TESTIFY	testify rebuff confess admit deny
DEFEAT	defeat destroy transform battle inject	APPLY	apply struggle earn graduate develop
CHASE	chase scare hit punch eat	EXPEL	expel inspire humiliate bully grant
TALK	talk tell reassure assure calm	DIG	dig take welcome sink revolve
POP	pop lift crawl laugh shake	COMMAND	command abduct invade seize surrender
SING	sing perform cast produce dance	RELENT	relent refuse agree insist hope
APPROVE	approve die suffer forbid collapse	EMBARK	embark befriend enlist recall meet
WEREWOLF	werewolf mother parent killer father	MANIPULATE	manipulate conclude investigate conduct
DINER	diner grandfather brother terrorist	ELOPE	elope forget succumb pretend like
DECAPITATE	decapitate bite impale strangle stalk	FLEE	flee escape swim hide manage
REPLY	reply say mention answer shout	BABY	baby sheriff vampire knight spirit
DEMON	demon narrator mayor duck crime	BIND	bind select belong refer represent
CONGRATULATE	congratulate cheer thank recommend	REJOIN	rejoin fly recruit include disguise
INTRODUCE	introduce bring mock read hatch	DARK	dark major henchman warrior sergeant
HATCH	hatch don exist vow undergo	SENTENCE	sentence arrest assign convict promote
FLIRT	flirt reconcile date dance forgive	DISTURB	disturb frighten confuse tease scare
ADOPT	adopt raise bear punish feed	RIP	rip vanish crawl drive smash
FAIRY	fairy kidnapper soul slave president	INFILTRATE	infiltrate deduce leap evade obtain
BUG	bug zombie warden king princess	SCREAM	scream faint wake clean hear

Table 3.3: Latent topics learned for $K = 50$ and $P = 100$. The words shown for each class are those with the highest smoothed PMI, with the label being the single word with the highest PMI.

Freq	Actions	Characters	Features
0.109	DARK$_m$, SHOOT$_a$, SHOOT$_p$	Jason Bourne (*The Bourne Supremacy*), Jack Traven (*Speed*), Jean-Claude (*Taken*)	Action, Male, War film
0.079	CAPTURE$_p$, INFILTRATE$_a$, FLEE$_a$	Aang (*The Last Airbender*), Carly (*Transformers: Dark of the Moon*), Susan Murphy/Ginormica (*Monsters vs. Aliens*)	Female, Action, Adventure
0.067	DEFEAT$_a$, DEFEAT$_p$, INFILTRATE$_a$	Glenn Talbot (*Hulk*), Batman (*Batman and Robin*), Hector (*Troy*)	Action, Animation, Adventure
0.060	COMMAND$_a$, DEFEAT$_p$, CAPTURE$_p$	Zoe Neville (*I Am Legend*), Ursula (*The Little Mermaid*), Joker (*Batman*)	Action, Adventure, Male
0.046	INFILTRATE$_a$, EXPLORE$_a$, EMBARK$_a$	Peter Parker (*Spider-Man 3*), Ethan Hunt (*Mission: Impossible*), Jason Bourne (*The Bourne Ultimatum*)	Male, Action, Age 34-36
0.036	FLIRT$_a$, FLIRT$_p$, TESTIFY$_a$	Mark Darcy (*Bridget Jones: The Edge of Reason*), Jerry Maguire (*Jerry Maguire*), Donna (*Mamma Mia!*)	Female, Romance Film, Comedy
0.033	EMBARK$_a$, INFILTRATE$_a$, INVADE$_a$	Perseus (*Wrath of the Titans*), Maximus Decimus Meridius (*Gladiator*), Julius (*Twins*)	Male, Chinese Movies, Spy
0.027	CONGRATULATE$_a$, CONGRATULATE$_p$, SWITCH$_a$	Professor Albus Dumbledore (*Harry Potter and the Philosopher's Stone*), Magic Mirror (*Shrek*), Josephine Anwhistle (*Lemony Snicket's A Series of Unfortunate Events*)	Age 58+, Family Film, Age 51-57
0.025	SWITCH$_a$, SWITCH$_p$, MANIPULATE$_a$	Clarice Starling (*The Silence of the Lambs*), Hannibal Lecter (*The Silence of the Lambs*), Colonel Bagley (*The Last Samurai*)	Age 58+, Male, Age 45-50
0.022	REPLY$_a$, TALK$_p$, FLIRT$_p$	Graham (*The Holiday*), Abby Richter (*The Ugly Truth*), Anna Scott (*Notting Hill*)	Female, Comedy, Romance Film
0.020	EXPLORE$_a$, EMBARK$_a$, CAPTURE$_p$	Harry Potter (*Harry Potter and the Philosopher's Stone*), Harry Potter (*Harry Potter and the Chamber of Secrets*), Captain Leo Davidson (*Planet of the Apes*)	Adventure, Family Film, Horror
0.018	FAIRY$_m$, COMMAND$_a$, CAPTURE$_p$	Captain Jack Sparrow (*Pirates of the Caribbean: At World's End*), Shrek (*Shrek*), Shrek (*Shrek Forever After*)	Action, Family Film, Animation
0.018	DECAPITATE$_a$, DECAPITATE$_p$, RIP$_a$	Jericho Cane (*End of Days*), Martin Riggs (*Lethal Weapon 2*), Gabriel Van Helsing (*Van Helsing*)	Horror, Slasher, Teen
0.017	APPLY$_a$, EXPEL$_p$, PURCHASE$_p$	Oscar (*Shark Tale*), Elizabeth Halsey (*Bad Teacher*), Dre Parker (*The Karate Kid*)	Female, Teen, Under Age 22

Table 3.4: Of 100 latent personas learned, we present the top 14 by frequency. Actions index the latent topic classes presented in table 3.3; subscripts denote whether the character is predominantly the agent (a), patient (p) or is modified by an attribute (m).

Chapter 4

Learning personas in books

Work described in this chapter was undertaken in collaboration with Ted Underwood and Noah Smith, and published at ACL 2014 (Bamman et al., 2014d)

4.1 Introduction

In chapter 3, we learn a set of entity types for characters in movies based on statistical regularities in the actions and attributes with which they are associated in Wikipedia plot summaries. While Wikipedia is, of course, written by large community of contributors, it presents an advantage of a relatively uniform, encyclopedic style governed by strict conventions.[1] This stylistic homogeneity can help learning—the variation we see in how a character is described can, in part, reasonably be attributed to the latent entity they embody and not to simple variation in word choice, for example, among the authors.

As we generalize this people-centric technique to a broader set of domains, one additional complexity automatically crops up: in stylistically heterogeneous corpora, in which individual authors have their own unique styles that are often very different from each other, how can we learn statistical regularities in the entity types of characters in the presence of that overwhelming stylistic variation? Furthermore, would we want to?

In the work presented in chapter 3, the text we observe associated with an entity in a document is directly dependent on the class of entity—and only that class. This relationship between entity and text is a theoretical assumption, with important consequences for learning: entity types learned in this way will be increasingly similar the more similar the domain, author, and other extra-linguistic effects are between them. Many entities in Early Modern English texts, for example, may be judged to be more similar to each other than to

[1] `http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style`.

entities from later texts simply by virtue of using *hath* and other archaic verb forms. While in many cases the topically similar types learned under this assumption may be desirable, in this chapter we explore the alternative, in which entity types are learned in a way that controls for such effects. In introducing a model based on different assumptions, we provide a method that complements the persona model introduced in chapter 3 and provides researchers with more flexible tools to infer different kinds of character types.

In this chapter, we focus on the literary domain, exploring a large collection of 15,099 English novels published in the 18th and 19th centuries and significantly expanding on the models presented in chapter 3. By accounting for the influence of individual authors while inferring latent character types, we are able to learn personas that cut across different authors more effectively than if we learned types conditioned on the text alone. Modeling the language used to describe a character as the joint result of that character's latent type and of other formal variables allows us to test multiple models of character and assess their value for different interpretive problems. As a test case, we focus on separating character from authorial diction, but this approach can readily be generalized to produce models that provisionally distinguish character from other factors (such as period, genre, or point of view) as well.

4.2 Literary Background

Inferring character in novels is challenging from a literary perspective partly because scholars have not reached consensus about the meaning of the term. It may seem obvious that a "character" is a representation of a (real or imagined) person, and many humanists do use the term that way. But there is an equally strong critical tradition that treats character as a formal dimension of narrative. To describe a character as a "blocking figure" or "first-person narrator," for instance, is a statement less about the attributes of an imagined person than about a narrative function (Keen, 2003). Characters are in one sense collections of psychological or moral attributes, but in another sense "word-masses" (Forster, 1927). This tension between "referential" and "formalist" models of character has been a centrally "divisive question in . . . literary theory" (Woloch, 2003).

Considering primary source texts (as distinct from plot summaries) forces us to confront new theoretical questions about character. In the Wikipedia plot summaries used in chapter 3, a human reader may already have used implicit models of character to extract high-level features. To infer character types from raw narrative text, researchers need to explicitly model the relationship of character to narrative form. This is not a solved problem, even for human readers.

For instance, it has frequently been remarked that the characters of Charles Dickens share certain similarities—including a reliance on tag phrases and recurring tics. A referential model of character might try to distinguish this common stylistic element from underlying "personalities." A strictly formalist model might refuse to separate authorial diction from character at all. In practice, human readers can adopt either perspective: we recognize that characters have a "Dickensian" quality but also recognize that a Dickens villain is (in one sense) more like villains in other authors than like a Dickensian philanthropist. Our goal is to show that computational methods can support the same range of perspectives—allowing a provisional, flexible separation between the referential and formal dimensions of narrative.

4.3 Data

The dataset for this work consists of 15,099 distinct narratives drawn from HathiTrust Digital Library.[2] From an initial collection of 469,200 volumes written in English and published between 1700 and 1899 (including poetry, drama, and nonfiction as well as prose narrative), we extract 32,209 volumes of prose fiction, remove duplicates and fuse multi-volume works to create the final dataset. Since the original texts were produced by scanning and running OCR on physical books, we automatically correct common OCR errors and trim front and back matter from the volumes using the page-level classifiers and HMM of Underwood et al. (2013)

Many aspects of this process would be simpler if we used manually-corrected texts, such as those drawn from Project Gutenberg. But we hope to produce research that has historical as well as computational significance, and doing so depends on the provenance of a collection. Gutenberg's decentralized selection process tends to produce exceptionally good coverage of currently-popular genres like science fiction, whereas HathiTrust aggregates university libraries. Library collections are not guaranteed to represent the past perfectly, but they are larger, and less strongly shaped by contemporary preferences.

The goal of this work is to provide a method to infer a set of character types in an unsupervised fashion from the data. As in chapter 3, we define this target, a character *persona*, as a distribution over several categories of typed dependency relations:[3]

1. agent: the actions of which a character is the agent (i.e., verbs for which the character holds an nsubj or agent relation).

[2] `http://www.hathitrust.org`

[3] All categories are described using the Stanford typed dependencies (de Marneffe and Manning, 2008), but any syntactic formalism is equally applicable.

2. patient: the actions of which a character is the patient (i.e., verbs for which the character holds a dobj or nsubjpass relation).

3. possessive: the objects that a character possesses (i.e., all words for which the character holds a poss relation).

4. predicative: attributes predicated of a character (i.e., adjectives or nouns holding an nsubj relation to the character, with an inflection of *be* as a child).

This set captures the constellation of what a character *does* and *has done to them*, what they *possess*, and what they are described as *being*.

While the work described in chapter 3 uses the Stanford CoreNLP toolkit to identify characters and extract typed dependencies for them, we found this approach to be too slow for the scale of our data (a total of 1.8 billion tokens); in particular, syntactic parsing, with cubic complexity in sentence length, and out-of-the-box coreference resolution (with thousands of potential antecedents) prove to be the biggest bottlenecks.

Before addressing character inference, we present here a prerequisite NLP pipeline that scales well to book-length documents.[4] This pipeline uses the Stanford POS tagger (Toutanova et al., 2003), the linear-time MaltParser (Nivre et al., 2007a) for dependency parsing (trained on Stanford typed dependencies), and the Stanford named entity recognizer (Finkel et al., 2005). It includes the following components for clustering character name mentions, resolving pronominal coreference, and reducing vocabulary dimensionality.

4.3.1 Character Clustering

First, let us terminologically distinguish between a character *mention* in a text (e.g., the token *Tom* on page 141 of *The Adventures of Tom Sawyer*) and a character *entity* (e.g., TOM SAWYER the character, to which that token refers). To resolve the former to the latter, we largely follow Davis et al. (2003) and Elson et al. (2010): we define a set of initial characters corresponding to each unique character name that is not a subset of another (e.g., *Mr. Tom Sawyer*) and deterministically create a set of allowable variants for each one (*Mr. Tom Sawyer* → *Tom*, *Sawyer*, *Tom Sawyer*, *Mr. Sawyer*, and *Mr. Tom*); then, from the beginning of the book to the end, we greedily assign each mention to the most recently linked entity for whom it is a variant. The result constitutes our set of characters, with all mentions partitioned among them.

[4] All code is available at `http://www.ark.cs.cmu.edu/literaryCharacter`

4.3.2 Pronominal Coreference Resolution

While the character clustering stage is essentially performing proper noun coreference resolution, approximately 74% of references to characters in books come in the form of pronouns.[5] To resolve this more difficult class at the scale of an entire book, we train a log-linear discriminative classifier only on the task of resolving *pronominal* anaphora (i.e., ignoring generic noun phrases such as *the paint* or *the rascal*).

For this task, we annotated a set of 832 coreference links in 3 books (*Pride and Prejudice*, *The Turn of the Screw*, and *Heart of Darkness*) and featurized coreference/antecedent pairs with:

1. The syntactic dependency path from a pronoun to its potential antecedent (e.g., dobj↑pred→↓pred↓nsubj (where → denotes movement across sentence boundaries).

2. The salience of the antecedent character (defined as the count of that character's named mentions in the previous 500 words).

3. The antecedent part of speech.

4. Whether or not the pronoun and antecedent appear in the same quotation scope (false if one appears in a quotation and one outside).

5. Whether or not the two agree for gender.

6. The syntactic tree distance between the two.

7. The linear (word) distance between the two.

With this featurization and training data, we train a binary logistic regression classifier with ℓ_1 regularization (where negative examples are comprised of all character entities in the previous 100 words not labeled as the true antecedent). In a 10-fold cross-validation on predicting the true nearest antecedent for a pronominal anaphor, this method achieves an average accuracy of 82.7%.

With this trained model, we then select the highest-scoring antecedent within 100 words for each pronominal anaphor in our data.

[5] Over all 15,099 narratives, the average number of character proper name mentions is 1,673; the average number of gendered singular pronouns (*he, she, him, his, her*) is 4,641.

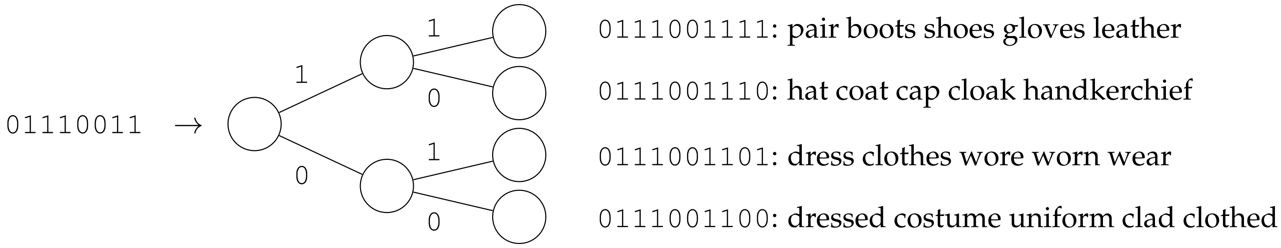

01110011 →
1
0
1
0
1
0
0111001111: pair boots shoes gloves leather
0111001110: hat coat cap cloak handkerchief
0111001101: dress clothes wore worn wear
0111001100: dressed costume uniform clad clothed

Figure 4.1: Bitstring representations of neural agglomerative clusters, illustrating the leaf nodes in a binary tree rooted in the prefix `01110011`. Bitstring encodings of intermediate nodes and terminal leaves result by following the left (`0`) and right (`1`) branches of the merge tree created through agglomerative clustering.

4.3.3 Dimensionality Reduction

To manage the degrees of freedom in the model described in §4.4, we perform dimensionality reduction on the vocabulary by learning word embeddings with a log-linear continuous skip-gram language model (Mikolov et al., 2013) on the entire collection of 15,099 books. This method learns a low-dimensional real-valued vector representation of each word to predict all of the words in a window around it; empirically, we find that with a sufficient window size (we use $n = 10$), these word embeddings capture semantic similarity (placing topically similar words near each other in vector space).[6] We learn a 100-dimensional embedding for each of the 512,344 words in our vocabulary.

To create a partition over the vocabulary, we use hard K-means clustering (with Euclidean distance) to group the 512,344 word types into 1,000 clusters. We then agglomeratively cluster those 1,000 groups to assign bitstring representations to each one, forming a balanced binary tree by only merging existing clusters at equal levels in the hierarchy. We use Euclidean distance as a fundamental metric and a group-average similarity function for calculating the distance between groups. Fig. 4.1 illustrates four of the 1,000 learned clusters.

4.4 Model

In order to separate out the effects that a character's persona has on the words that are associated with them (as opposed to other factors, such as time period, genre, or author), we adopt a hierarchical Bayesian approach in which the words we observe are generated conditional on a combination of different effects captured in a log-linear (or "maximum entropy") distribution. As noted in the Methods section above (§2.3), this parameterization of a language model that conditions on external metadata is a theme throughout this thesis.

[6] In comparison, Brown et al. (1992) clusters learned from the same data capture *syntactic* similarity (placing functionally similar words in the same cluster).

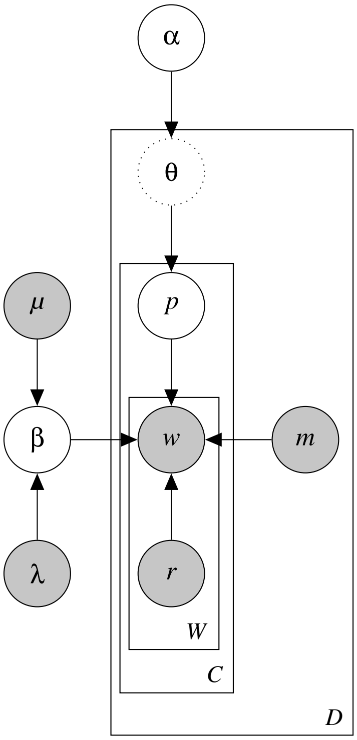
In contrast to chapter 3, where the probability of a word linked to a character is dependent entirely on the character's latent persona, in our model, we see the probability of a word as dependent on: (i) the **background** likelihood of the word, (ii) the **author**, so that a word becomes more probable if a particular author tends to use it more, and (iii) the character's **persona**, so that a word is more probable if appearing with a particular persona. Intuitively, if the author *Jane Austen* is associated with a high weight for the word *manners*, and all personas have little effect for this word, then *manners* will have little impact on deciding which persona a particular Austen character embodies, since its presence is explained largely by Austen having penned the word. While we address only the author as an observed effect, this model is easily extended to other features as well, including period, genre, point of view, and others.

The generative story runs as follows (Figure 4.2 depicts the full graphical model): Let there be M unique authors in the data, P latent personas (a hyperparameter to be set), and V words in the vocabulary (in the general setting these may be word types; in our data the vocabulary is the set of 1,000 unique cluster IDs). Each role type $r \in$ {agent, patient, possessive, predicative} and vocabulary word v (here, a cluster ID) is associated with a real-valued vector $\beta_{r,v} = [\beta^{meta}_{r,v}, \beta^{pers}_{r,v}, \beta^{0}_{r,v}]$ of length $M + P + 1$. The first $M + P$ elements are drawn from a Laplace prior with mean $\mu = 0$ and scale $\lambda = 1$; the last element $\beta^{0}_{r,v}$ is an unregularized bias term accounting for the background. Each element in this vector captures the log-additive effect of each author, persona, and the background distribution on the word's probability (Eq. 4.1, below).

Much like latent Dirichlet allocation (Blei et al., 2003), each document d in our dataset draws a multinomial distribution θ_d over personas from a shared Dirichlet prior α, which captures the proportion of each character type in that particular document. Every character c in the document draws its persona p from this document-specific multinomial. Given document metadata m (here, one of a set of M authors) and persona p, each tuple of a role r with word w is assumed to be drawn from Eq. 4.1 in Fig. 4.3. This model can be understood as a sparse additive generative (SAGE) model (Eisenstein et al., 2011a) with three kinds of features (metadata, persona, and background bias).

4.4.1 Hierarchical Softmax

The partition function in Eq. 4.1 can lead to slow inference for any reasonably-sized vocabulary. To address this, we reparameterize the model by exploiting the structure of the agglomerative clustering in §4.3.3 to perform a hierarchical softmax, following Goodman (2001), Morin and Bengio (2005) and Mikolov et al. (2013).



P	Number of personas (hyperparameter)
D	Number of documents
C_d	Number of characters in document d
$W_{d,c}$	Number of (cluster, role) tuples for character c
m_d	Metadata for document d (ranges over M authors)
θ_d	Document d's distribution over personas
$p_{d,c}$	Character c's persona (integer, $p \in \{1, \ldots, P\}$)
j	An index for a $\langle r, w \rangle$ tuple in the data
w_j	Word cluster ID for tuple j (integer, $w \in \{1, \ldots, V\}$)
r_j	Role for tuple $j \in \{\textsf{agent}, \textsf{patient}, \textsf{mod}, \textsf{poss}\}$
β	Coefficients for the log-linear language model
μ, λ	Laplace mean and scale (for regularizing β)
α	Dirichlet concentration hyperparameter

Figure 4.2: **Top:** Probabilistic graphical model. Observed variables are shaded, latent variables are clear, and collapsed variables are dotted. **Bottom:** Definition of variables.

The bitstring representations by which we encode each word in the vocabulary serve as natural, and inherently meaningful, intermediate classes that correspond to semantically related subsets of the vocabulary, with each bitstring prefix denoting one such class. Longer bitstrings correspond to more fine-grained classes. In the example shown in Figure 4.1, `011100111` is one such intermediate class, containing the union of *pair, boots, shoes, gloves*

$$
\begin{aligned}
P(w \mid m,p,r,\beta) &= \frac{\exp\left(\beta^{meta}_{r,w}[m] + \beta^{pers}_{r,w}[p] + \beta^{0}_{r,w}\right)}{\sum_{v=1}^{V}\exp\left(\beta^{meta}_{r,v}[m] + \beta^{pers}_{r,v}[p] + \beta^{0}_{r,v}\right)} && (4.1)\\
P(b \mid m,p,r,\beta) &= \prod_{j=0}^{n-1} \begin{cases} \operatorname{logit}^{-1}\left(\beta^{meta}_{r,b_{1:j}}[m] + \beta^{pers}_{r,b_{1:j}}[p] + \beta^{0}_{r,b_{1:j}}\right) & \text{if } b_{j+1} = 1 \\ 1 - \operatorname{logit}^{-1}\left(\beta^{meta}_{r,b_{1:j}}[m] + \beta^{pers}_{r,b_{1:j}}[p] + \beta^{0}_{r,b_{1:j}}\right) & \text{otherwise} \end{cases} && (4.2)
\end{aligned}
$$

Figure 4.3: Parameterizations of the SAGE word distribution. Eq. 4.1 is a "flat" multinomial logistic regression with one β-vector per role and word. Eq. 4.2 uses the hierarchical softmax formulation, with one β-vector per role and node in the binary tree of word clusters, giving a distribution over bit strings (b) with the same number of parameters as Eq. 4.1.

leather and *hat, coat, cap cloak, handkerchief*. Because these classes recursively partition the vocabulary, they offer a convenient way to reparameterize the model through the chain rule of probability.

Consider, for example, a word represented as the bitstring $c = \texttt{01011}$; calculating $P(c = \texttt{01011})$—we suppress conditioning variables for clarity—involves the product: $P(c_1 = \texttt{0}) \times P(c_2 = \texttt{1} \mid c_1 = \texttt{0}) \times P(c_3 = \texttt{0} \mid c_{1:2} = \texttt{01}) \times P(c_4 = \texttt{1} \mid c_{1:3} = \texttt{010}) \times P(c_5 = \texttt{1} \mid c_{1:4} = \texttt{0101})$.

Since each multiplicand involves a binary prediction, we can avoid partition functions and use the classic binary logistic regression.[7] We have converted the V-way multiclass logistic regression problem of Eq. 4.1 into a sequence of $\log V$ evaluations (assuming a perfectly balanced tree). Given m, p, and r (as above) we let $b = b_1 b_2 \cdots b_n$ denote the bitstring representation of a word cluster, and the distribution is given by Eq. 4.2 in Fig. 4.3.

In this paramaterization, rather than one β-vector for each role and vocabulary term, we have one β-vector for each role and conditional binary decision in the tree (each bitstring prefix). Since the tree is binary with V leaves, this yields the same total number of parameters. As Goodman (2001) points out, while this reparameterization is exact for true probabilities, it remains an approximation for estimated models (with generalization behavior dependent on how well the class hierarchy is supported by the data). In addition to enabling faster inference, one advantage of the bitstring representation and the hierarchical softmax parameterization is that we can easily calculate probabilities of clusters at different granularities.

[7] Recall that logistic regression lets $P_{LR}(y = 1 \mid x, \beta) = \operatorname{logit}^{-1}(x^\top \beta) = 1/(1 + \exp -x^\top \beta)$ for binary dependent variable y, independent variables x, and coefficients β.

4.4.2 Inference

Our primary quantities of interest in this model are p (the personas for each character) and β, the effects that each author and persona have on the probability of a word. Rather than adopting a fully Bayesian approach (e.g., sampling all variables), we infer these values using stochastic EM, alternating between collapsed Gibbs sampling for each p and maximizing with respect to β.

Collapsed Gibbs for personas. At each step, the required quantity is the probability that character c in document d has persona z, given everything else. This is proportional to the number of other characters in document d who also (currently) have that persona (plus the Dirichlet hyperparameter which acts as a smoother) times the probability (under $p_{d,c} = z$) of all of the words observed in each role r for that character:

$$(count(z; p_{d,-c}) + \alpha_z) \times \prod_{r=1}^{R} \prod_{j:r_j=r} P(b_j \mid m, p, r, \beta) \tag{4.3}$$

The metadata features (like author, etc.) influence this probability by being constant for all choices of z; e.g., if the coefficient learned for *Austen* for vocabulary term *manners* is high and all coefficients for all z are close to zero, then the probability of *manners* will change little under different choices of z. Eq. 4.3 contains one multiplicand for every word associated with a character, and only one term reflecting the influence of the shared document multinomial. The implication is that, for major characters with many observed words, the words will dominate the choice of persona; where the document influence would have a bigger effect is with characters for whom we don't have much data. In that case, it can act as a kind of informed background; given what little data we have for that character, it would nudge us toward the character types that the other characters in the book embody.

Given an assignment of all p, we choose β to maximize the conditional log-likelihood of the words, as represented by their bitstring cluster IDs, given the observed author and background effects and the sampled personas. This equates to solving $4V$ ℓ_1-regularized logistic regressions (see Eq. 4.2 in Figure 4.3), one for each role type and bitstring prefix, each with $M + P + 1$ parameters. We apply OWL-QN (Andrew and Gao, 2007) to minimize the ℓ_1-regularized objective with an absolute convergence threshold of 10^{-5}.

4.5 Evaluation

While standard NLP and machine learning practice is to evaluate the performance of an algorithm on a held-out gold standard, articulating what a true "persona" might be for a character is inherently problematic. Rather, we evaluate the performance and output of our model by preregistering a set of 29 hypotheses of varying scope and difficulty and comparing the performance of different models in either confirming, or failing to confirm, those hypotheses. This kind of evaluation was previously applied to a subjective text measurement problem by Sim et al. (2013).

All hypotheses were created by a literary scholar with specialization in the period to not only give an empirical measure of the strengths and weaknesses of different models, but also to help explore exactly what the different models may, or may not, be learning. All preregistered hypotheses establish the degrees of similarity among three characters, taking the form: "character X is more similar to character Y than either X or Y is to a distractor character Z"; for a given model and definition of distance under that model, each hypothesis yields two yes/no decisions that we can evaluate:

- $distance(X, Y) < distance(X, Z)$
- $distance(X, Y) < distance(Y, Z)$

To tease apart the different kinds of similarities we hope to explore, we divide the hypotheses into four classes:

A. This class constitutes **sanity checks**: character X and Y are more similar to each other in every way than to character Z. E.g.: Elizabeth Bennet in *Pride and Prejudice* resembles Elinor Dashwood in *Sense and Sensibility* (Jane Austen) more than either character resembles Allen Quatermain in *Allen Quatermain* (H. Rider Haggard). (Austenian protagonists should resemble each other more than they resemble a grizzled hunter.)

B. This class captures our ability to identify two characters in the same author as being more similar to each other than to a closely related character in a **different** author. E.g.: Wickham in *Pride and Prejudice* resembles Willoughby in *Sense and Sensibility* (Jane Austen) more than either character resembles Mr. Rochester in *Jane Eyre* (Charlotte Brontë).

C. This class captures our ability to discriminate among similar characters in the **same** author. In these hypotheses, two characters X and Y from the same author are more similar to each other than to a third character Z from that same author. E.g.: Wickham in *Pride and Prejudice* (Jane Austen) resembles Willoughby in *Sense and Sensibility* more than either character resembles Mr. Darcy in *Pride and Prejudice.*

D. This class constitutes more difficult, **exploratory** hypotheses, including differences among point of view. E.g.: Montoni in *Mysteries of Udolpho* (Radcliffe) resembles Heathcliff in *Wuthering Heights* (Emily Brontë) more than either resembles Mr. Bennet in *Pride and Prejudice.* (Testing our model's ability to discern similarities in spite of elapsed time.)

All 29 hypotheses can be found in Bamman et al. (2014c). We emphasize that the full set of hypotheses was locked *before* the model was estimated.

4.6 Experiments

Part of the motivation of the model presented here is to be able to tackle hypothesis class C—by factoring out the influence of a particular author on the learning of personas, we would like to be able to discriminate between characters that all have a common authorial voice. In contrast, the Persona Regression model of chapter 3, which uses metadata variables (like authorship) to encourage entities with similar covariates to have similar personas, reflects an assumption that makes it likely to perform well at class B.

To judge their respective strengths on different hypothesis classes, we evaluate three models:

1. The mixed-effects Author/Persona model (described above), which includes author information as a metadata effect; here, each β-vector (of length $M + P + 1$) contains a parameter for each of the distinct authors in our data, a parameter for each persona, and a background parameter.

2. A Basic persona model, which ablates author information but retains the same log-linear architecture; here, the β-vector is of size $P + 1$ and does not model author effects.

3. The Persona Regression model introduced in chapter 3.

All models are run with $P \in \{10, 25, 50, 100, 250\}$ personas; Persona Regression additionally uses $K = 25$ latent topics. All configurations use the full dataset of 15,099 novels, and all characters with at least 25 total roles (a total of 257,298 entities). All experiments are run for a total of 50 iterations, alternating between sampling personas p and maximizing β. The value of α is optimized using slice sampling (with a non-informative prior) every 5 iterations. The value of λ is held constant at 1. At the end of inference, we calculate the posterior distributions over personas for all characters as the sampling probability of the final iteration.

To formally evaluate "similarity" between two characters, we measure the Jensen-Shannon divergence between personas (calculated as the average JS distance over the cluster distributions for each role type), marginalizing over the characters' posterior distributions over personas; two characters with a lower JS divergence are judged to be more similar than two characters with a higher one.

As a Baseline, we also evaluate all hypotheses on a model with no latent variables whatsoever, which instead measures similarity as the average JS divergence between the empirical word distributions over each role type.

Table 4.1 presents the results of this comparison; for all models with latent variables, we report the average of 5 sampling runs with different random initializations. Figure 4.4 (below it) provides a synopsis of this table by illustrating the average accuracy across all choice of P. All models, including the baseline, perform well on the sanity checks (A). As expected, the Persona Regression model performs best at hypothesis class B (correctly judging two characters from the same author to be more similar to each other than to a character from a different author); this behavior is encouraged in this model by allowing an author (as an external metadata variable) to directly influence the persona choice, which has the effect of pushing characters from the same author to embody the same character type. Our mixed effects Author/Persona model, in contrast, outperforms the other models at hypothesis class C (correctly discriminating different character types present in the same author). By discounting author-specific lexical effects during persona inference, we are better able to detect variation among the characters of a single author that we are not able to capture otherwise. While these different models complement each other in this manner, we note that there is no absolute separation among them, which may be suggestive of the degree to which the formal and referential dimensions are fused in novels. Nevertheless, the strengths of these different models on these different hypothesis classes gives us flexible alternatives to use depending on the kinds of character types we are looking to infer.

4.7 Analysis

The latent personas inferred from this model will support further exploratory analysis of literary history. Table 4.2 illustrates this with a selection of three character types learned, displaying characteristic clusters for all role types, along with the distribution of that persona's use across time and the gender distribution of characters embodying that persona. In general, the personas learned so far do not align neatly with character types known to literary historians. But they do have legible associations both with literary genres and with social categories. Even though gender is not an observable variable known to the model

		Hypothesis Class			
P	Model	A	B	C	D
250	Author/Persona	1.00	0.58	0.75	0.42
	Basic Persona	1.00	0.73	0.58	0.53
	Persona Reg.	0.90	0.70	0.58	0.44
100	Author/Persona	0.98	0.68	0.70	0.46
	Basic Persona	0.95	0.73	0.53	0.47
	Persona Reg.	0.93	0.78	0.63	0.49
50	Author/Persona	0.95	0.73	0.63	0.50
	Basic Persona	0.98	0.75	0.48	0.53
	Persona Reg.	1.00	0.75	0.65	0.38
25	Author/Persona	1.00	0.63	0.65	0.50
	Basic Persona	1.00	0.63	0.50	0.50
	Persona Reg.	0.90	0.78	0.60	0.39
10	Author/Persona	0.95	0.63	0.70	0.51
	Basic Persona	0.78	0.80	0.48	0.46
	Persona Reg.	0.90	0.73	0.43	0.41
Baseline		1.00	0.63	0.58	0.37

Table 4.1: Agreement rates with preregistered hypotheses, averaged over 5 sampling runs with different initializations.

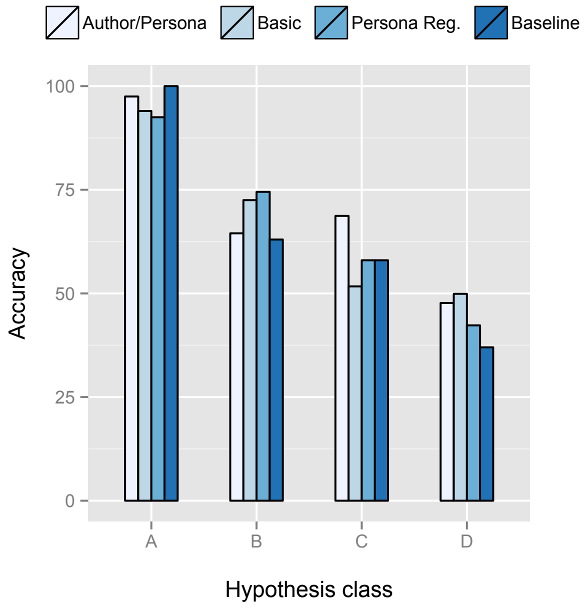
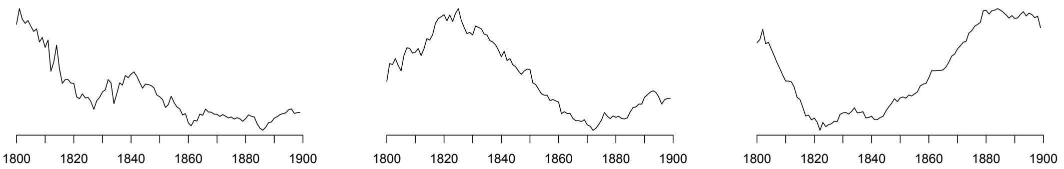


Figure 4.4: Synopsis of table 4.1: average accuracy across all *P*. Persona regression is best able to judge characters in one author to be more similar to each other than to characters in another (B), while our mixed-effects Author/Persona model outperforms other models at discriminating characters in the same author (C).

during inference, personas tend to be clearly gendered. This is not in itself surprising (since



Agent	carried ran threw rose fell suddenly is seems	sent received arrived appeared struck showed returned immediately waiting	turns begins returns thinks loves calls does knows comes
Patient	wounded killed murdered suffer yield acknowledge free saved unknown	wounded killed murdered destroy bind crush attend haste proceed	thinks loves calls love hope true turn hold show
Poss	death happiness future lips cheek brow mouth fingers tongue	army officers troops soldiers band armed party join camp	lips cheek brow eyes face eye table bed chair
Pred	crime guilty murder youth lover hers dead living died	king emperor throne general officer guard soldier knight hero	beautiful fair fine good kind ill dead living died
% Female	12.2	3.7	54.7

Table 4.2: Snapshots of three personas learned from the $P = 50$, Author/Persona model. Gender and time proportions are calculated by summing and normalizing the posterior distributions over all characters with that feature. We truncate time series at 1800 due to data sparsity before that date; the y-axis illustrates the frequency of its use in a given year, relative to its lifetime.

literary scholars know that assumptions about character are strongly gendered), but it does suggest that diachronic analysis of latent character types might cast new light on the history of gender in fiction. This is especially true since the distribution of personas across the time axis similarly reveals coherent trends.

Table 4.3 likewise illustrates what our model learns by presenting a sample of the fixed effects learned for a set of five major 19th-century authors. These are clusters that are conditionally more likely to appear associated with a character in a work by the given author than they are in the overall data; by factoring this information out of the inference process for learning character types (by attributing its presence in a text to the author rather than the persona), we are able to learn personas that cut across different topics more effectively than if a character type is responsible for explaining the presence of these terms as well.

4.8 Conclusion

This chapter extends the latent persona models introduced in chapter 3 to support inference over a wider range of entity types, including those that can be learned in the presence of stylistic heterogeneity. Postulating an interaction between authorial diction and character allows models that consider the effect of the author to more closely reproduce a human

Author	clusters
Jane Austen	praise gift consolation letter read write character natural taste
Charlotte Brontë	lips cheek brow book paper books hat coat cap
Charles Dickens	hat coat cap table bed chair hand head hands
Herman Melville	boat ship board hat coat cap feet ground foot
Jules Verne	journey travel voyage master company presence success plan progress

Table 4.3: Characteristic possessive clusters in a sample of major 19th-century authors.

reader's judgments, especially by learning to distinguish different character types within a single author's oeuvre. This opens the door to considering other structural and formal dimensions of narration. For instance, representation of character is notoriously complicated by narrative point of view (Booth, 1961); and indeed, comparisons between first-person narrators and other characters are a primary source of error for all models tested above. The strategy we have demonstrated suggests that it might be productive to address this by modeling the interaction of character and point of view as a separate effect analogous to authorship.

Like the models presented in chapter 3, the models tested above diverge from many structuralist theories of narrative (Propp, 1968) by allowing multiple instances of the same persona in a single work. As suggested in chapter 3, one way of addressing these structural constraints is to incorporate them into the model (e.g., by replacing a work's multinomial distribution over all entity types with a geometric distribution for each type). Another possibility is to empirically test those theories themselves, and learn structural limitations on the number of "protagonists" likely to coexist in a single story. In all cases, the machinery of hierarchical models gives us the flexibility to incorporate such effects at will, while also being explicit about the theoretical assumptions that attend them.

Chapter 5

Learning events through people

Work described in this chapter was undertaken in collaboration with Noah Smith and published in Transactions of the ACL *(Bamman and Smith, 2014)*

5.1 Introduction

The work described above sought to infer latent categories (or *personas*) of people as having intrinsic value of their own for exploratory data analysis of fictional works, learning those types through the variation in how fictional characters were described. In this chapter, I consider a different utility of exploring how textual depictions of how people vary: learning latent event classes.

As I highlight throughout this thesis, much of the text data that we interact with on an everyday basis describes *people* in various degrees. For corpora that include historically deep biographical information (such as Wikipedia, book-length biographies and autobiographies, and even newspaper obituaries) this data includes the actors involved in particular historical events and the times and places in which they occur, and provides an abundance of information on how the lives of those portrayed unfold. The life events described in these texts often have natural structure: event classes exhibit correlations with each other (e.g., those who DIVORCE must have been MARRIED), can occur at roughly similar times in the lives of different individuals (MARRIAGE is more likely to occur earlier in one's life than later), and can be bound to historical moments as well (FIGHTS IN WORLD WAR II peaks in the early 1940s).

Social scientists have long been interested in the structure of these events in investigating the role that individual agency and larger social forces play in shaping the course of an individual's life. Life stages marking "transitions to adulthood" (such as LEAVING SCHOOL, ENTERING THE WORKFORCE and MARRIAGE) have important correlates with demographic variables (Modell et al., 1976; Hogan and Astone, 1986; Shanahan, 2000); and researchers study the interactional effects that life events have on each other, such as the relationship between divorce and pre-marital cohabitation (Lillard et al., 1995; Reinhold, 2010) or having children (Lillard and Waite, 1993).

The data on which these studies draw, however, has largely been restricted to categorical surveys and observational data; we present here a latent-variable model that exploits the correlations of event descriptions in text to learn the structure of abstract events, grounded in time, from text alone. While our model can be estimated on any set of texts where the birth dates of a set of mentioned entities are known, we illustrate our method on a large-scale dataset of 242,970 biographies extracted from Wikipedia.

This chapter makes two contributions: first, we present a general unsupervised model for learning life event classes from biographical text, along with the structure that binds them; second, in using this method to learn event classes from Wikipedia, we uncover evidence of systematic bias in the presentation of male and female biographies (with biographies of women containing significantly disproportionate emphasis on the personal events of marriage and divorce). In addition to these contributions, we also present a range of other analyses that uncovering life events in text can make possible. This work illustrates not simply the value of orienting statistical inference on people even when the quantities of interest are, for example, *events*, but also the social insight that can be had in exploring the interaction of people as both the *content* and the *authors* of text.

5.2 Data

The data for this analysis originates in the January 2, 2014 dump of English-language Wikipedia.[1] We extract biographies by identifying all articles with `persondata` metadata[2] in which the `DATE OF BIRTH` field is known. This results in a set of 927,403 biographies.

For each biography, we perform part-of-speech tagging using the Stanford POS tagger (Toutanova et al., 2003) and named entity recognition using the Stanford named entity recognizer (Finkel et al., 2005), cluster all mentions of co-referring proper names (Davis et al., 2003; Elson et al., 2010) and resolve pronominal co-reference, aided by gender inference for each entity as the gender corresponding to the maximum number of gendered pronouns (i.e., *he* and *she*) mentioned in the article, as also used by Reagle and Rhue (2011). In a ran-

[1] http://dumps.wikimedia.org/enwiki/20140102/enwiki-20140102-pages-articles.xml.bz2

[2] "Persondata is a special set of metadata that can and should be added to biographical articles only" (http://en.wikipedia.org/wiki/Wikipedia:Persondata).

dom test set of 500 articles, this method of gender inference is overwhelmingly accurate, achieving 100% precision with 97.6% recall (12 articles had no pronominal mentions and so gender is not assigned).

As further preprocessing, we identify multiword expressions in all texts as maximal sequences of adjective + noun part of speech tags (yielding, for example, *New York*, *United States*, *early life* and *high school*), as first described in Justeson and Katz (1995). For each biographical article, we then extract all sentences in which the subject of the article is mentioned along with a single date and retain only the terms in each sentence that are among the most frequent 10,000 unigrams and multiword expressions in all documents, excluding stopwords such as *the* and all numbers (including dates). An "event" is the bag of these unigrams and multiword expressions extracted from one such sentence, along with a corresponding timestamp measured as the difference between the observed date in the sentence and the date of birth of the entity.

Table 5.1 illustrates the actual form of the data with a sample of extracted sentences from the biography of Frank Lloyd Wright, along with the data as input to the model. In the terminology of the model described below, each sentence constitutes one "event" in the subject's life.

Original sentence	Data as input to model	
	Terms (w)	Time (t)
He was admitted to the University of Wisconsin–Madison as a special student in 1886.	admitted university wisconsin madison special student	19
Wright first traveled to Japan in 1905, where he bought hundreds of prints.	wright first traveled japan bought hundreds prints	38
After Wright's return to the United States in October 1910, Wright persuaded his mother to buy land for him in Spring Green, Wisconsin.	wright return united_states wright persuaded mother buy land spring green wisconsin	43
This philosophy was best exemplified by his design for Fallingwater (1935), which has been called "the best all-time work of American architecture".	philosophy best design called best all-time work american architecture	68
Already well known during his lifetime, Wright was recognized in 1991 by the American Institute of Architects as " the greatest American architect of all time."	already well known lifetime wright recognized american institute architects greatest american architect time	124

Table 5.1: A sample of 5 of the 64 sentences (original and converted) that constitute the data for Frank Lloyd Wright (born 1867). Each event is defined as one such temporally-scoped sentence.

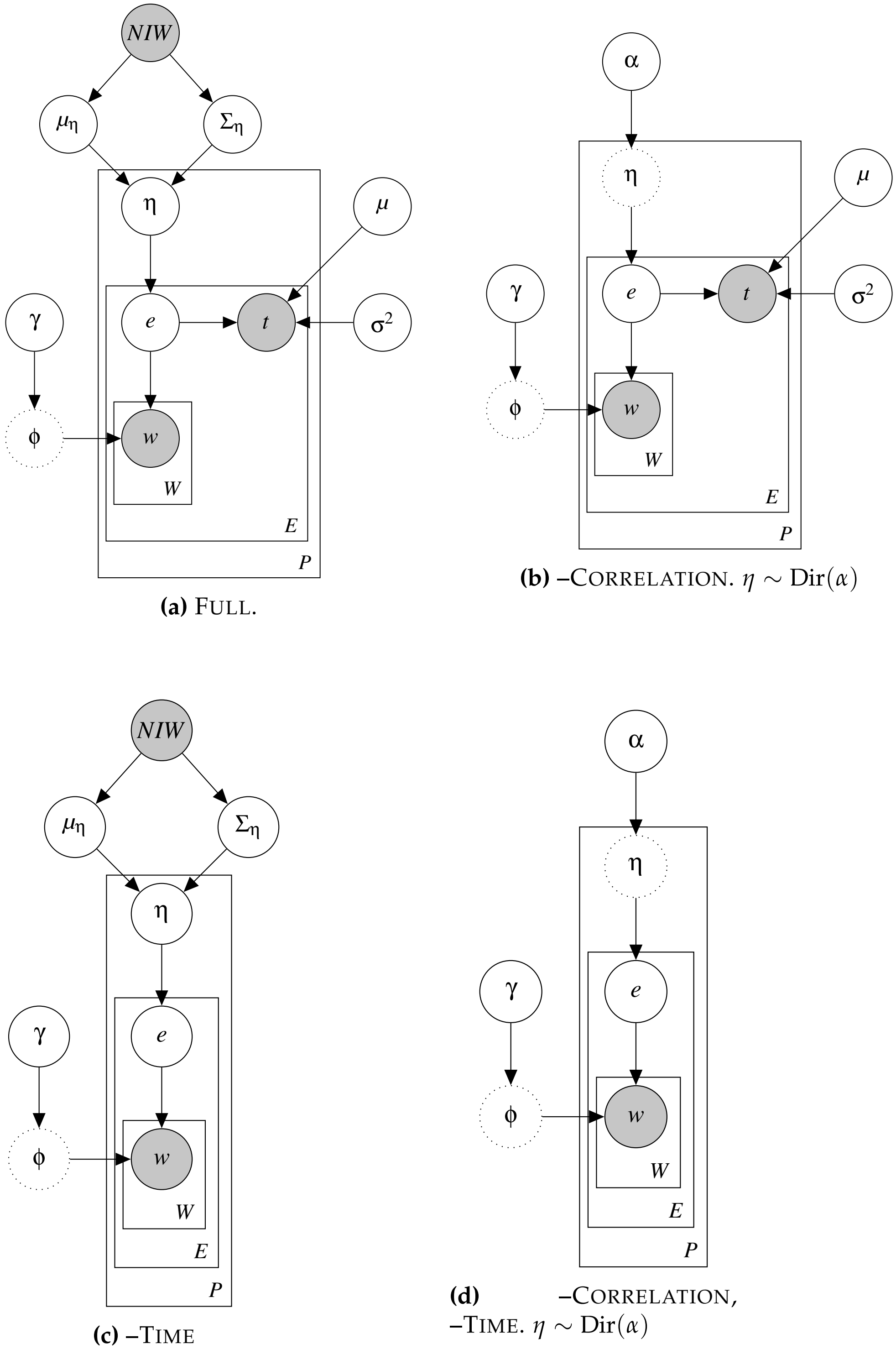
For the final dataset we retain all biographies where the subject of the article is born after the year 1800 and for which there exist at least 5 events (242,970 people). The complete data consists of 2,313,867 events across these 242,970 people.

5.3 Model

The quantities of interest that we want to learn from the data are: 1.) a broad set of major life events recorded in Wikipedia biographies that people experience at similar stages in their lives (such as BEING BORN, GRADUATING HIGH SCHOOL, SERVING IN THE ARMY, GETTING MARRIED, and so on); 2.) correlations among those life events (e.g., knowing that if an individual WINS A NOBEL PRIZE that they're more likely to RECEIVE AN HONORARY DOCTORATE); and 3.) an attribution of those classes of events to particular moments in a specific individual's life (e.g., John Nash RECEIVED AN HONORARY DOCTORATE in 1999). People in this case will provide the scaffolding to learn these quantities by providing a fixed point around which event classes can cluster.

We cast this problem as an unsupervised learning one; given no labeled instances, can we infer these quantities from text alone? One possible alternative approach would be to leverage the categorical information contained in Wikipedia biographies (or its derivatives, such as Freebase; Google, 2014) as a form of supervision (e.g., George Washington is a member of the categories *Presidents of the United States* and *American cartographers*, among others). These manual categories, however, are often sporadically annotated and have a long tail (with most categories appearing very few times); in learning event structure directly from text, we avoid relying on categories' accuracy and being constrained by a fixed ontology. One advantage of an unsupervised approach is that we eliminate the need to define a predetermined set of event classes *a priori*, allowing application across a variety of different domains and time periods, such as full-text books from the Internet Archive or Hathi Trust, or historical works like the *Oxford Dictionary of National Biography* (Matthew and Harrison, 2004).

Figure 5.1a illustrates the graphical form of our hierarchical Bayesian model, which articulates the relationship between an entity's set of *events* (where each event is an observation defined as the bag of terms in text and the difference between the year it was recorded as happening and the birth year), an abstract set of *event classes*, correlations among those abstract classes, and the distribution of vocabulary terms that defines each one. To capture correlations among different classes, we place a logistic normal prior on each biography's distribution over event classes (Blei and Lafferty, 2006a, 2007; Mimno et al., 2008); unlike a Dirichlet, a logistic normal is able to capture arbitrary correlations between elements through the structure of the covariance matrix of its underlying multivariate normal. We take a Bayesian approach to estimating the mean μ_η and covariance Σ_η, drawing them



(a) FULL.

(b) –CORRELATION. $\eta \sim \mathrm{Dir}(\alpha)$

(c) –TIME

(d) –CORRELATION, –TIME. $\eta \sim \mathrm{Dir}(\alpha)$

Figure 5.1: Graphical form of the full model (described in §5.3) and models with ablations (described in §5.4).

from a conjugate Normal-Inverse Wishart prior.

The generative story for the model runs as follows: let K be the number of latent event classes, P be the number of biographies, and E_p be the number of events in biography p.

- Draw event class means and covariances
  $\mu_\eta \in \mathbb{R}^K, \Sigma_\eta \in \mathbb{R}^{K \times K} \sim$ Normal-Inverse Wishart$(\mu_0, \lambda, \Psi, \nu)$
- For each event class $i \in \{1, \ldots, K\}$:
  - Draw event-term distribution $\phi_k \sim \text{Dir}(\gamma)$
- For each biography p:
  - Draw $\eta_p \sim \mathcal{N}(\mu_\eta, \Sigma_\eta)$
  - Convert η_p into biography-event proportions β_p through the softmax function: $\beta_{p,i} = \frac{\exp(\eta_{p,i})}{\sum_{k=1}^{K} \exp(\eta_{p,k})}$
  - For each event in biography p:
    - Draw event class index $e \sim \text{Mult}(\beta_p)$
    - Draw timestamp $t \sim \mathcal{N}(\mu_e, \sigma_e^2)$
    - For each token in event:
      - Draw term $w \sim \text{Mult}(\phi_e)$

Inference proceeds via stochastic EM: after initializing all variables to random values, we alternate between collapsed Gibbs sampling for the latent class indicators followed by maximization steps over all other parameters:

1. Sample all e using collapsed Gibbs sampling conditioned on current values for η and all other e.

2. For each biography p, maximize likelihood with respect to η_p via gradient ascent given the current samples of e and priors μ_η and Σ_η.

3. Assign MAP estimates of μ_η and Σ_η given current values of η and the Normal-Inverse Wishart prior. Update μ and σ^2 according to its maximum likelihood estimate given e.

We describe the technical details of each step below.

Sampling e. Given fixed biography-event class proportions η, observed tokens w, timestamp t, and current samples e^- for all other events, the probability of a given event belonging to event class k is as follows:

$$\begin{aligned} P(e=k \mid e^{-}, w, t, \eta, \gamma, \mu, \sigma^2) &\propto \exp(\eta_k) \\ &\times \sigma_k^{-1} \exp\left(-\frac{(t-\mu_k)^2}{2\sigma_k^2}\right) \\ &\times \frac{\prod_{v=1}^{V} \prod_{i=1}^{\mathbf{e}(v)} (\gamma + \mathbf{c}^{-}(k,v) + i - 1)}{\prod_{n=1}^{N_e} (V\gamma + \mathbf{c}^{-}(k,\star) + n - 1)} \end{aligned} \tag{5.1}$$

Here $\mathbf{c}^{-}(k,v)$ is the count of the number of times vocabulary term v shows up in all events whose current sample $e=k$ (excepting the current one being sampled), $\mathbf{c}^{-}(k,\star)$ is the total count of all terms in all events whose current $e=k$ (again excepting the current one), N_e is the number of terms in event e, and $\mathbf{e}(v)$ is the count of vocabulary term v in the current event.

Maximizing η. Under our model, the terms in the likelihood function that involve η include the likelihood of the samples drawn from it and its own probability given the multivariate Normal prior:

$$L(\eta) \propto \prod_{n=1}^{N} \frac{\exp(\eta_{e_n})}{\sum_{k=1}^{K} \exp(\eta_k)} \times \mathcal{N}(\eta \mid \mu_\eta, \Sigma_\eta) \tag{5.2}$$

The log likelihood is proportional to:

$$\begin{aligned} \ell(\eta) \propto & \sum_{n=1}^{N} \eta_{e_n} - \sum_{n=1}^{N} \sum_{k=1}^{K} \exp(\eta_k) \\ & - \frac{1}{2} (\eta - \mu_\eta)^\top \Sigma_\eta^{-1} (\eta - \mu_\eta) \end{aligned} \tag{5.3}$$

Given samples of the latent event class e for all events in biography p, we maximize the value of η_p using gradient ascent. We can think of this as maximizing the likelihood of the observations e subject to ℓ_2 (Gaussian) regularization, where the covariance matrix in the regularizer encourages correlations in η: if a document contains many examples of $e=k$ and e_k is highly correlated with e_j, then the optimal η is encouraged to contain high weights at both η_k and η_j rather than simply η_k alone.

Maximizing $\mu_\eta, \Sigma_\eta, \mu, \sigma^2$. Given values for $\jmath$, we then find maximum *a posteriori* estimates of μ_η and Σ_η conditioned on the Normal-Inverse Wishart (NIW) prior. The NIW is a conjugate prior to a multivariate Gaussian, parameterized by dimensionality K, initial mean μ_0, positive-definite scale matrix Ψ, and scalars $\nu > K-1$ and $\lambda > 0$. The prior parameters

Ψ and ν have an intuitive interpretation as the scatter matrix $\sum_{i=1}^{\nu} (x_i - \bar{x})(x_i - \bar{x})^\top$ for ν pseudo-observations.

The expected value of the covariance matrix drawn from a NIW distribution parameterized by Ψ and ν is $\frac{\Psi}{\nu - K - 1}$. To disprefer correlations among topics in the absence of strong evidence, we fix $\mu_0 = 0$ and set Ψ so that this prior expectation over Σ_η is the product of a scalar value ρ and the identity matrix $\mathbf{I}$: $\Psi = (\nu - K - 1)\rho\mathbf{I}$; ρ defines the expected variance, and the higher the value of ν, the more strongly the prior dominates the posterior estimate of the covariance matrix (i.e., the more the covariance matrix is shrunk toward $\rho\mathbf{I}$). λ likewise has an intuitive understanding as a dampening parameter: the higher its value, the more the posterior estimate of the mean $\hat{\mu}$ shrinks toward 0. For n data points, we set $\lambda = n/10$, $\nu = K + 2$, and $\rho = 1$.

Since the NIW is conjugate with the multivariate normal, posterior updates to μ_η and Σ_η have closed-form expressions given values of η (here, $\bar{\eta}$ denotes the mean value of η over all biographies).

$$\hat{\mu}_\eta = \frac{n}{\lambda + n}\bar{\eta} \tag{5.4}$$

$$\hat{\Sigma}_\eta = \frac{\Psi + \sum_{i=1}^{N} (\eta_i - \bar{\eta})(\eta_i - \bar{\eta})^\top + \frac{\lambda n}{\lambda + n}\bar{\eta}\bar{\eta}^\top}{\nu + n + K + 1} \tag{5.5}$$

Since we have no meaningful prior information on the values of μ and σ^2, we calculate their maximum likelihood estimate given current samples $\mathbf{e}$.

5.4 Evaluation

While the goal of this work is to leverage the organizing principle of people in order to learn qualitative categories of life events from text, we can quantitatively evaluate the performance of our model on the empirical task of predicting the age in a person's life when an event occurs.

For this task, we compare the full model described above with a strong baseline of ℓ_2-regularized linear regression and also with comparable models with feature ablations, in order to quantify the extent to which various aspects of the full model are contributing to its empirical performance. The comparable ablated models include the following:

- –CORRELATION, figure 5.1b. Rather than a logistic normal prior on the entity-specific distribution over event types (η), we draw η from a symmetric Dirichlet distribution parameterized by a global α. In a Dirichlet distribution, arbitrary correlations cannot be captured.

- –TIME, figure 5.1c. In the full model, the timestamps of the observed events influence the event classes we learn by encouraging them to be internally coherent and time-sensitive. To test this design choice, we ablate time as a feature during inference.

- –CORRELATION,–TIME, figure 5.1d. We also test a model that ablates both the correlation structure in the prior and the influence of time; this model corresponds to smoothed, unsupervised naïve Bayes.

As during inference, we define an event to be the set of terms, excluding stopwords and numbers, that are present in the vocabulary of the 10,000 most frequent words and multi-word expressions in the data overall. Each event is accompanied by the year of its occurrence, from which we calculate the gold target prediction (the age of the person at the time of the event) as the year minus the entity's year of birth. For all of the four models described above (the full model and three ablations), we train the model on 4/5 of the biographies (194,376 entities, on average 1,851,094 events); we split the remaining 1/5 of the biographies into development data (where t is observed) and test data (where t is predicted). The details of inference for each model are as follows:

1. FULL. Inference as above for a period of 100 iterations, using slice sampling (Neal, 2003) to optimize the value of the Dirichlet hyperparameter γ every 10 iterations; after inference, the parameters $\mu_\eta, \Sigma_\eta, \mu, \sigma^2$ and ϕ are estimated from samples drawn at the final iteration and held fixed. For test entities, we infer the MAP value of η using development data, and predict the age of each test event as the mean time marginalizing over the event type indicator e. $\hat{t} = \mathbb{E}_e[\mu_e]$.

2. –CORRELATION. Here we perform collapsed Gibbs sampling for 100 iterations, using slice sampling to optimize the value of α and γ every 10 iterations; after inference, the parameters μ, σ^2 and ϕ are estimated from single final samples and held fixed. For development and test data, we run Gibbs sampling on event indicators e for 10 iterations and predict the age of each test event as the mean time marginalizing over the event type indicator e. $\hat{t} = \mathbb{E}_e[\mu_e]$.

3. –TIME. Inference as above for 100 iterations, using slice sampling to optimize the value of γ every 10 iterations; after inference, the parameters μ_η, Σ_η and ϕ are estimated from single final samples and held fixed. Since time is not known to this model during inference, we create post hoc estimates of $\hat{\mu}_e$ as the empirical mean age of events sampled to event class e using single samples for each event in the training data from the final

sampling iteration. For test entities, we infer the MAP value of η using development data, and predict the age of each test event as the average empirical age marginalizing over the event type indicator e. $\hat{t} = \mathbb{E}_e[\hat{\mu}_e]$.

4. –CORRELATION,–TIME. We perform inference as above for the –CORRELATION model, and time prediction as in the –TIME model. $\hat{t} = \mathbb{E}_e[\hat{\mu}_e]$.

To compare against a potentially more powerful discriminative model, we also evaluate linear regression with squared ℓ_2 (ridge) regularization, using binary indicators of the same unigrams and multiword expressions available to the models above.

5. LINEAR REGRESSION. Train on training and development data, optimizing the regularization coefficient λ in three-fold cross-validation.

During training, linear regression learns that the terms most indicative of events that take place later in life are *stamp, descendant, commemorated, died, plaque, grandson*, and *lifetime achievement award*, while those that denote early events are *born, baptised, apprenticed*, and *acting debut*.

We evaluate all models on identical splits using 5-fold cross validation. For an interpretable error score, we use mean absolute error, which corresponds to the number of years, on average, by which each model is incorrect.

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} \left|\hat{t} - t_i\right| \tag{5.6}$$

Figure 5.2 presents the results of this evaluation for all models and different choices of the number of latent event classes $K \in \{10, 25, 50, 100, 250, 500\}$. Linear regression represents a powerful model, achieving a mean absolute error of 11.87 years across all folds, but is eclipsed by the latent variable model at $K \geq 50$. The correlations captured by the logistic normal prior make a clear difference, uniformly yielding improvements over otherwise equivalent Dirichlet models across all K. As expected, models trained without knowledge of time during inference perform less well than models that contain that information.

5.5 Analysis

To analyze the latent event classes in Wikipedia biographies, we train our full model (with a logistic normal prior and time as an observable variable) on the full dataset of 242,970 biographies with $K = 500$ event classes; as above, we run inference for a burn-in period of 100 iterations and collect 50 samples from the posterior distributions for e (the event class

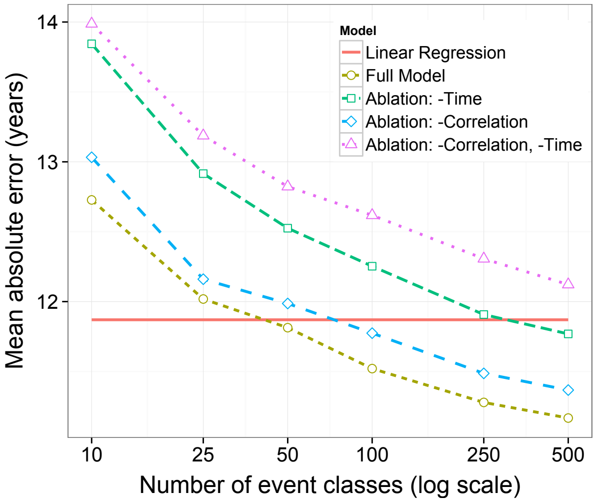


Figure 5.2: Mean absolute error (in years) for time prediction.

indicator for each event).

Table 5.2 illustrates a sample of 20 event classes along with the mean time μ and standard deviation σ, the gender distribution (calculated from the posterior distribution over e for all entities whose gender is known[3]) and the most probable terms in the class.

The latent classes that we learn span a mix of major life events of Wikipedia notable figures (including events that we might characterize as GRADUATING HIGH SCHOOL, BECOMING A CITIZEN, DIVORCE, BEING CONVICTED OF A CRIME, and DYING) and more fine-grained events (such as BEING DRAFTED BY A SPORTS TEAM and BEING INDUCTED INTO THE HALL OF FAME).

Emerging immediately from this summary is an imbalance in the gender distribution for many of these event classes. Among the 242,858 biographies whose gender is known, 14.8% are of women; we would therefore expect around 14.8% of the participants in most event classes to be female. Figures 5.3 and 5.4 illustrate five of the most highly skewed classes in both directions, ranked according to the z score of a two-tailed binomial proportion test ($H_0 = 14.8$).

While some of these classes reflect a biased world in which more men are drafted into sports teams, serve in the armed forces, and are ordained as priests, one latent class that calls out for explanation is that surrounding DIVORCE (*divorce, marriage, divorced, filed, married, wife, separated, years, ended, later*), whose female proportion of 39.4% is nearly triple that of the data overall (and whose z-score reveals it to be strongly statistically different [$p \ll$

[3] Using our method of gender inference described in §5.2, we are able to infer gender for 99.95% of biographies (242,858).

Age μ	Age σ	% Fem.	Most probable terms in class
18.00	0.67	15.6%	high school, graduated, attended, graduating, school, born, early life, class, grew
21.89	1.83	0.2%	drafted, nfl draft, round, professional career, draft, overall, selected
22.27	1.19	17.6%	graduated, bachelor, degree, university, received, college, attended, earned, b. a.
22.67	4.33	3.6%	joined, enlisted, army, served, world war ii, united states army, years, corps
25.81	3.47	11.1%	law, university, graduated, received, school, law school, degree, law degree
32.32	8.19	12.0%	thesis, received, university, phd, dissertation, doctorate, degree, ph. d., completed
38.24	15.29	17.0%	citizen, became, citizenship, united states, american, u. s., british, granted, since
39.33	12.53	39.4%	divorce, marriage, divorced, married, filed, wife, separated, years, ended, later
42.57	13.78	16.3%	university, teaching, professor, college, taught, faculty, school, department, joined
43.79	15.54	13.8%	trial, murder, case, court, charges, guilty, jury, judge, death, convicted
45.89	18.71	13.3%	died, accident, killed, death, near, crash, car, involved, car accident, injured
46.22	16.30	11.2%	prison, released, years, sentence, sentenced, months, parole, federal, serving
49.81	10.28	7.0%	governor, candidate, unsuccessful candidate, congress, ran, reelection
51.41	11.23	1.2%	bishop, appointed, archbishop, diocese, pope, consecrated, named, cathedral
54.91	12.04	7.9%	chairman, board, president, ceo, became, company, directors, appointed, position
59.06	14.17	16.9%	awarded, university, received, honorary doctorate, honorary degree, degree, doctor
62.81	24.16	11.1%	fame, inducted, hall, sports hall, elected, national, football hall, international
72.52	13.69	12.4%	died, hospital, age, death, complications, cancer, home, heart attack, washington
92.39	46.06	13.0%	national, historic, park, state, house, named, memorial, home, honor, museum
95.29	42.65	12.1%	statue, unveiled, memorial, plaque, anniversary, erected, monument, death, bronze

Table 5.2: Salient event classes learned from 242,970 Wikipedia biographies. All 500 event classes can be viewed at http://www.ark.cs.cmu.edu/bio.

0.0001] from the H_0 mean, even accounting for the Bonferroni correction we must make when considering the $K = 500$ tests we implicitly perform when ranking). While we did not approach this analysis with any *a priori* hypotheses to test, our unsupervised model reveals

z	%Fem.	Most frequent terms
60.46	76.9%	miss, pageant, title, usa, miss universe, beauty, held, teen, crowned, competed
57.21	49.9%	birth, gave, daughter, son, born, first child, named, wife, announced, baby
55.63	59.8%	fashion, model, show, campaign, week, appeared, face, career, became, modeling
37.89	39.4%	divorce, marriage, divorced, married, filed, wife, separated, years, ended, later
36.70	36.5%	summer olympics, competed, olympics, team, finished, event, final, world championships

Table 5.3: Female-skewed event classes, ranked by z-score in a two-tailed binomial proportion test.

z	%Fem.	Most frequent terms
-31.64	0.2%	drafted, nfl draft, round, professional career, draft, overall, selected, major league baseball
-23.81	2.1%	promoted, rank, captain, retired, army, lieutenant, colonel, major, brigadier general
-20.93	3.7%	bar, admitted, law, practice, called, commenced, studied, began, career, practiced
-20.48	1.0%	infantry, civil war, regiment, army, enlisted, served, company, colonel, captain
-20.30	1.7%	ordained, priest, seminary, priesthood, theology, theological, college, studies, rome

Table 5.4: Male-skewed event classes, ranked by z-score in a two-tailed binomial proportion test.

an interesting hypothesis to pursue with confirmatory analysis: biographies of women on Wikipedia disproportionately focus on marriage and divorce compared to those of men.

To test this hypothesis with more traditional means, we estimated the empirical gender proportions of biographies containing terms explicitly denoting divorce (*divorced, divorce, divorces* and *divorcing*). The result of this analysis confirms that of the model. Of the 4,608 biographies in which at least one of these terms appears, 38.8% are those of a woman, far more than the 14.8% we would expect (in a two-tailed binomial proportion test against $H_0 = 14.8$, this difference is significant at $p < 0.0001$); this corresponds to divorce being mentioned in 5.0% of all 35,932 women's biographies, and 1.4% of all 206,926 men's; on average, a woman's biography is 3.66 times more likely to mention divorce than a man's.

We repeat the gender proportion experiment with terms denoting marriage (*married, marry, marries, marrying* and *marriage*) and find a similar trend: of the 39,142 biographies where at least one of these terms is mentioned, 23.6% belong to women; again, in a two-tailed proportion test, this difference is significant at $p < 0.0001$. This corresponds to marriage appearing in 25.7% of all women's biographies, and 14.5% of men's; a woman's biography is 1.78 times more likely to mention marriage than a man's.

5.6 Additional Analyses

The analysis above represents one substantive result that mining life events from biographical data makes possible. To illustrate the range of other analyses that this method can occasion, we briefly present two other directions that can be pursued: investigating correlations among event classes and the distribution of event classes over historical time.

5.6.1 Correlations among events

In our full model with a logistic normal prior over a document's set of events, correlations among latent event classes are learned during inference. From the covariance matrix Σ_η, we can directly read off correlations among events; for other models (such as those with a Dirichlet prior), we can infer correlations using the posterior estimates for η.

r	Event class
1.000	family, boss, murder, crime, mafia, became, arrested, john, gang, chicago
0.031	killed, shot, police, home, two, car, arrested, murder, death, -year-old
0.028	trial, murder, case, guilty, court, jury, charges, convicted, death, judge
0.021	investigation, federal, charges, office, fraud, campaign, state, commission, former, corruption
0.019	arrested, sentenced, years, prison, trial, death, court, convicted, military, months

Table 5.5: Highest correlations between the *family, boss, murder, crime, mafia* class and other events.

Table 5.5 illustrates the event classes that have the highest correlations to the event class defined by *family, boss, murder, crime, mafia, became, arrested, john, gang, chicago*. The structure that we learn here neatly corresponds to a Criminal Action frame, with common events for Killing, Being subject to federal investigation, Being arrested and Being brought to trial.

5.6.2 Historical distribution of events

Figure 5.3 likewise illustrates the distribution over time for a set of learned event classes. While the only notion of time that our model has access to during inference is that of time relative to a person's birth, we can estimate the empirical distribution of event classes in historical time by charting the density plot of their observed absolute dates. Several historically relevant event classes are legible, including Serving in the army (with peaks during World War I and II, Vietnam and the later Iraq wars), Opera debut (with peaks in the 1950s), NASA (with peaks in 1960s and the turn of the millenium), Joining the Communist Party (with a rise in the early 20th century), Leading an expedition (with a slow historical decline) and Joining a band (with increasing historical presence). Grounding specific life events in history has the potential to enable analysis of how historical time affects the life histories of individuals—including both the influence of the general passage of time, as on transitions to adulthood (Modell et al., 1976; Hogan, 1981; Modell, 1980), and the influence of specific historical moments like the Great Depression (Elder, 1974) or World

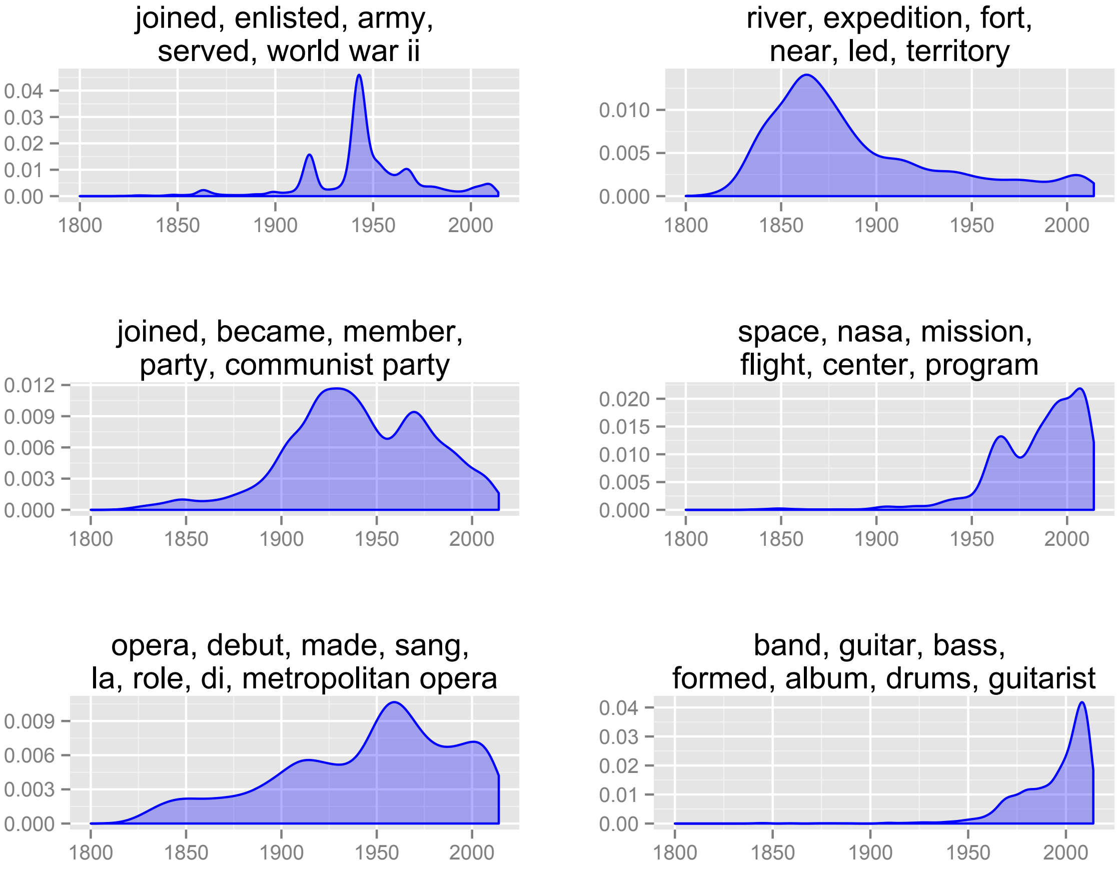


Figure 5.3: Historical distributions of event classes.

War II (Mayer, 1988; Elder, 1991).

5.7 Related Work

In learning general classes of events from text, our work draws on a rich background spanning several research traditions. By considering the structure that exists between event classes, we draw on the original work on procedural scripts and schemas (Minsky, 1974; Schank and Abelson, 1977) and narrative chains (Chambers and Jurafsky, 2008, 2009), including more recent advances in the unsupervised learning of frame semantic representations (Modi et al., 2012; O'Connor, 2013; Cheung et al., 2013; Chambers, 2013).

In learning latent classes from text, our work is also clearly related to research on topic modeling (Blei et al., 2003; Griffiths and Steyvers, 2004). This work differs from that tradition by scoping our data only over text that we have reason to believe describes events (by including absolute dates). While other topic models have leveraged temporal information in the learning of latent topics, such as the dynamic topic model (Blei and Lafferty, 2006b; Wang et al., 2012) and "topics over time" (Wang and McCallum, 2006), our model is the first to infer classes of events whose contours are shaped by the time in a person's life that they take place.

While the information extraction tasks of template filling (Hobbs et al., 1993) and relation detection (Banko et al., 2007; Fader et al., 2011; Carlson et al., 2010) generally fall into a paradigm of classifying text segments into a predetermined ontology, they too have been informed by unsupervised approaches to learning relation classes (Yao et al., 2011) and events (Ritter et al., 2012). Our work here differs from this past work in leveraging explicit absolute temporal information in the unsupervised learning of event classes (and their structure). Reasoning about the temporal ordering of events likewise has a long tradition of its own, both in NLP (Pustejovsky et al., 2003; Mani et al., 2006; Verhagen et al., 2007; Chambers et al., 2007) and information extraction (Talukdar et al., 2012). Rather than attempting to model the ordering of events relative to each other, we focus instead on their occurrence relative to the beginning of a person's life.

Wikipedia likewise has been used extensively in NLP; Wikipedia biographies in particular have been used for the task of training summarization models (Biadsy et al., 2008), recognizing biographical sentences (Conway, 2010), learning correlates of "success" (Ng, 2012), and disambiguating named entities (Bunescu and Pasca, 2006; Cucerzan, 2007). In our work in mining biographical structure from it, we draw on previous research into automatically uncovering latent structure in resumés (Mimno and Mccallum, 2007) and approaches to learning life path trajectories from categorical survey data (Massoni et al., 2009; Ritschard et al., 2013).

In using Wikipedia as a dataset for analysis, we must note that the subjects of biographies are not a representative sample of the population, nor are their contents unbiased representations. Nearly all encyclopedias necessarily prefer the historically notorious (if due to nothing else than inherent biases in the preservation of historical records); many, like Wikipedia, also have disproportionately low coverage of women, minorities, and other demographic groups, in part because of biases in community membership. Estimates of the percentage of female editors on Wikipedia, for example, ranges from 9% to 16.1% (Collier and Bear, 2012; Reagle and Rhue, 2011; Cassell, 2011; Hill and Shaw, 2013; Wikipedia, 2011). Different language editions of Wikipedia have a natural geographic bias in article selection (Hecht and Gergle, 2009), with each emphasizing their own "local heroes" (Kolbitsch and Maurer, 2006), and also differ in the kind of information they present (Pfeil et al., 2006; Callahan and Herring, 2011). This extends to selection of biographies as well, with one study finding approximately 16% of 1000 sampled biographies being those of women (Reagle and Rhue, 2011), a figure very close to the 14.8% we observe in our analysis here.

5.8 Conclusion

In this chapter, I present a method for mining life events from biographies, leveraging the correlation structure of events as they appear together under the same individuals. Unlike prior work that has focused on inferring "life trajectories" from categorical survey data, we learn relevant structure in an unsupervised manner directly from text, opening the door to applying this method to a broad set of biographies beyond Wikipedia (including full-text books from the Internet Archive or Hathi Trust, and other encyclopedic biographies as well). In a quantitative analysis, the model we present outperforms a strong baseline at the task of event time prediction, and surfaces a substantive qualitative distinction in the *content* of the biographies of men and women on Wikipedia: in contrast to previous work that uses computational methods to measure a difference in coverage, we show that such methods are able to tease apart differences in characterization as well.

There are several directions in which this work can be pushed. While the task of event time prediction provides a quantitative means to compare different models, we expect the real application of this work will lie in the latent event classes themselves, and the information they provide both about the subjects and authors of biographies. In addition to occasioning data analysis of the kind we describe here, we expect that personal event classes can have a practical application in helping to organize data describing people as well. While in this case, the events that typically appear with (e.g.) U.S. Presidents are sufficiently distinct from those that appear with (e.g.) American actors, and that difference is stark enough to learn crisp event classes, we might also draw on the persona models from chapters 3 and 4 to include fine-grained entity type information, either again in a fully unsupervised setting (as above), or perhaps with the partial supervision of Wikipedia category information. While this information has not been necessary for the models presented here, it may provide helpful guidance for more complex models, such as those directly learning distinct patterns in the *sequence* or trajectory of events)—having finer representations of people's roles may make learning divisions between common sequences more clear.

Part II

Variation in the author and audience

Overview

The first section of this thesis covers the variation that exists in depictions of people *within* text; we find that we can leverage people as an organizing principle of data to learn latent entity types (or *personas*) that regulate how people's actions and attributes cluster together into coherent groups, and show that centering our inference on people can enable other downstream applications, such as learning event classes for analysis that can occasion social insight (about the biases that exist in the characterization of women on Wikipedia).

This section, in contrast, considers people in the situated environment that surrounds text: as its *authors* and its *audience*. In these chapters, I model the variation that we see within text as a function of the differences—both observed and latent—in the people who write it and in the people to whom it is directed.

Authors. The connection between text and personal attributes of authors has been the focus of much work over the past few years into inferring latent qualities of individuals—such as age, gender, political affiliation—from the text they write. While this work can be thought to date back to the original task of authorship attribution (Mosteller and Wallace, 1964), where the hidden quality of interest is author identity, the rise of user-generated content on the web and (especially) social media has made this a thriving cottage industry. Prior to streaming social media, gender and age were common prediction targets in blogs (Herring and Paolillo, 2006; Koppel et al., 2006; Argamon et al., 2007; Mukherjee and Liu, 2010; Rosenthal and McKeown, 2011). With the rise of Twitter, these studies have expanded to encompass gender, age, political affiliation, place of birth, personality and ethnicity, among many others (Rao et al., 2010; Golbeck et al., 2011; Burger et al., 2011; Pennacchiotti and Popescu, 2011; Conover et al., 2011; Volkova et al., 2014). Recent work has expanded this attribute set even further, into the domain of fine-grained categories like MUSICIAN and ATHLETE (El-Arini et al., 2012, 2013; Bergsma and Van Durme, 2013; Beller et al., 2014). Unlike fine-grained named entity classification or relation extraction, the text for this task is not comprised of third-person descriptions of people; the input is first-person narrative.

Rather than focusing on inferring those properties directly, the work presented in this section considers downstream applications where people as *authors* can again provide an organizing principle for inference. Chapter 6 leverages the beliefs of individuals (even when those beliefs are unknown) in order to estimate the political import of propositions like OBAMA IS A SOCIALIST. In this work, we never observe whether an author is more liberal or conservative in their politics, but we do observe the bundle of propositions they assert. People, as authors, are the focal point around which we can infer information about language—here, propositional information—as it is used in the world.

Chapter 7 in contrast leverages observed metadata about the authors (their geographical location) in order to improve low-dimensional word embeddings by making them sensitive to geographic variation: rather than making an assumption that word types (like *wicked*) are a single point in some real-valued space, this work emphasizes that sense distinctions in words can be anchored in geography: knowing that an author using the word *wicked* lives in Boston shifts the predominant sense of that word away from being synonymous with *evil* and more toward an adverbial intensifier (*very*). It is information about people again, in their organizing space as authors, that enables these kinds of distinctions to be learned.

Audience. The connection between author and audience is of primary concern for several natural language processing tasks that are explicitly focused on discourse, such as conversational agents and other dialogue systems—voice dialogue systems can adapt their conversational strategies to the emotions (such as frustration) detected in their partners (Burkhardt et al., 2005), and virtual peers can be designed to match the dialect of their interlocutors (Finkelstein et al., 2013) and have strategies for developing rapport (Cassell and Bickmore, 2003; Gupta et al., 2007).

The connection between *written* text and personal attributes of audience, however, has not been so richly explored from a computational perspective, often in part because audience is typically unknown, or difficult to estimate, in the text we have available. This is especially true on social media, where the actual audience (the set of people who are addressed or overhear a conversation) is not necessarily the same as an imagined audience (the set of people the speaker believes to be addressing). As Bernstein et al. (2013) note for Facebook, users are often poor judges for estimating their actual audience, substantially underestimating the number of people who read their posts. Additionally, while Facebook has means for limiting the audience exposed to a given post, Twitter is a site of "context collapse" (boyd, 2008; Marwick and boyd, 2011), where a single message is broadcast to individuals from all walks of a person's life. Understanding the audience, however, is a necessary prerequisite for real text understanding. One common assumption throughout user attribute-prediction literature is that attributes are inherent, essential qualities of individuals; a person is at heart either a DEMOCRAT or a REPUBLICAN, a MAN or a WOMAN, and all of the text they write, in any circumstance, serves as evidence for these fundamental qualities about themselves. But we know that individuals project different aspects of themselves to different audiences, through sociological work on self-presentation (Goffman, 1959), linguistic work on style shifting and audience design (Bell, 1984) and third-wave studies in linguistic variation (Eckert, 2008; Johnstone and Kiesling, 2008). In my own work in collaboration with Jacob Eisenstein and Tyler Schnoebelen (Bamman et al., 2014b), we have seen how the presentation of gender on social media is much more complicated than a single binary distinction would imply, often involving complex interactions with audience. Context in all of these cases is crucial for measuring these aspects of identity.

In the final chapter of this thesis (§8), I present a case study that incorporates this kind of audience information along with other aspects about the situated context of an utterance on Twitter (such as information about the *content* of a message along with qualities of its *author*) for the complex task of sarcasm detection. This work provides a summary of several themes running throughout this work: that language is profoundly situated, spoken by people at particular times and places to particular audiences, and that incorporating any elements of this context can lead to improvements in predictive tasks.

Chapter 6

Learning ideal points of propositions through people

6.1 Introduction

The latent persona models outlined in chapters 3 and 4 represent people who are *described* in text (characters) as embodying one of a set number of entity types; in this chapter, we consider the analogous variation that people exhibit as the *authors* of text. Rather than learning these types for exploratory data analysis, in this chapter we have a fixed task in mind: estimating the political beliefs attached to propositions.

Over the past few years, much work has focussed on inferring political preferences of people from their behavior, both in unsupervised and supervised settings. Classical ideal point models (Poole and Rosenthal, 1985; Martin and Quinn, 2002) estimate the political ideologies of legislators through their observed voting behavior, possibly paired with the textual content of bills (Gerrish and Blei, 2012) and debate text (Nguyen et al., 2015); other unsupervised models estimate ideologies of politicians from their speeches alone (Sim et al., 2013). Twitter users have also been modeled in a similar framework, using their observed following behavior of political elites as evidence to be explained (Barberá, 2015). Supervised models, likewise, have not only been used for assessing the political stance of sentences (Iyyer et al., 2014) but are also very popular for predicting the holistic ideologies of everyday users on Twitter (Rao et al., 2010; Pennacchiotti and Popescu, 2011; Al Zamal et al., 2012; Cohen and Ruths, 2013; Volkova et al., 2014), Facebook (Bond and Messing, 2015) and blogs (Jiang and Argamon, 2008), where training data is relatively easy to obtain—either from user self-declarations, political following behavior, or third-party categorizations.

Aside from their intrinsic value, estimates of users' political ideologies have been useful for quantifying the orientation of news media sources (Park et al., 2011; Zhou et al., 2011). We consider in this work a different task: estimating the political import of propositions like OBAMA IS A SOCIALIST.

In focusing on propositional statements, we draw on a parallel, but largely independent, strand of research in open information extraction. IE systems, from early slot-filling models with predetermined ontologies (Hobbs et al., 1993) to the large-scale open-vocabulary systems in use today (Fader et al., 2011; Mitchell et al., 2015) have worked toward learning type-level propositional information from text, such as BARACK OBAMA IS PRESIDENT. To a large extent, the ability to learn these facts from text is dependent on having data sources that are either relatively factual in their presentation (e.g., news articles and Wikipedia) or are sufficiently diverse to average over conflicting opinions (e.g., broad, random samples of the web).

Many of the propositional statements that individuals make online are, of course, not objective descriptions of reality at all, but rather reflect their own beliefs, opinions and other private mental states (Wiebe et al., 2005). While much work has investigated methods for establishing the truth content of individual sentences — whether from the perspective of veridicality (de Marneffe et al., 2012), fact assessment (Nakashole and Mitchell, 2014), or subjectivity analysis (Wiebe et al., 2003; Wilson, 2008) — the structure that exists between users and their assertions gives us an opportunity to situate them both in the same political space: in this work we operate at the level of subject-predicate propositions, and present models that capture not only the variation in what subjects (e.g., OBAMA, ABORTION, GUN CONTROL) that individual communities are more likely to discuss, but also the variation in what predicates different communities assert of the same subject (e.g., GLOBAL WARMING IS A HOAX VS. IS A FACT). The contributions of this work are as follows:

- We present a new evaluation dataset of 766 propositions judged according to their point in a political spectrum.

- We present and evaluate several models for estimating the ideal points of subject-predicate propositions, and find that unsupervised methods perform best (on sufficiently partisan data).

6.2 Task and Data

The task that we propose in this work is assessing the political import of type-level propositions; on average, are liberals or conservatives more likely to claim that GLOBAL WARMING IS A HOAX? To support this task, we create a benchmark of political propositions, extracted from politically partisan data, paired with human judgments (details in §6.2.3). We define a **proposition** to be a tuple comprised of a subject and predicate, each consisting of one or more words, such as ⟨Global warming, *is a hoax*⟩.[1] We adopt an open vocabulary approach where each unique predicate defines a unary relation.

6.2.1 Data

In order to first extract propositions that are likely to be political in nature and exhibit variability according to ideology, we collect data from a politically volatile source: comments on partisan blogs.

We draw data from NPR,[2] Mother Jones[3] and Politico[4], all listed by Pew Research (Mitchell et al., 2014) as news sources most trusted by those with consistently liberal views; Breitbart,[5] most trusted by those with consistently conservative views; and the Daily Caller,[6] Young Conservatives[7] and the Independent Journal Review,[8] all popular among conservatives (Kaufman, 2014). All data comes from articles published between 2012–2015.

Source	Articles	Posts	Tokens	Users
Politico	10,305	9.8M	348.4M	173,519
Breitbart	46,068	8.8M	336.4M	165,607
Daily Caller	46,114	5.4M	240.4M	228,696
Mother Jones	16,830	1.9M	119.2M	138,995
NPR	14993	1.6M	82.6M	62,600
IJ Review	3,396	278K	13.1M	51,589
Young Cons.	4,948	222K	10.6M	34,434
Total	142,654	28.0M	1.15B	621,231

Table 6.1: Data.

We gather comments using the Disqus API;[9] as a comment hosting service, Disqus allows users to post to different blogs using a single identity. Table 6.1 lists the total number of articles, user comments, unique users and tokens extracted from each blog source. In total, we extract 28 million comments (1.2 billion tokens) posted by 621,231 unique users.[10]

[1] We use these typographical conventions throughout this chapter: Subjects are in sans serif, predicates in *italics*.

[2] `http://www.npr.org`

[3] `http://www.motherjones.com`

[4] `http://www.politico.com`

[5] `http://www.breitbart.com`

[6] `http://dailycaller.com`

[7] `http://www.youngcons.com`

[8] `https://www.ijreview.com`

[9] `https://disqus.com/api/`

[10] While terms of service prohibit our release of this data, we will make available tools to allow others to collect similar data from Disqus for these blogs.

6.2.2 Extracting Propositions

The blog comments in table 6.1 provide raw data from which to mine propositional assertions. In order to extract structured ⟨subject, *predicate*⟩ propositions from text, we first parse all comments using the collapsed dependencies (de Marneffe and Manning, 2008) of the Stanford parser (Manning et al., 2014), and identify all subjects as those that hold an `nsubj` or `nsubjpass` relation to their head. In order to balance the tradeoff between generality and specificity in the representation of assertions, we extract three representations of each predicate.

1. Exact strings, which capture verbatim the specific nuance of the assertion. This includes all subjects paired with their heads and all descendants of that head. All tense and number are preserved.

   Example: ⟨Reagan, *gave amnesty to 3 million undocumented immigrants*⟩

2. Reduced syntactic tuples, which provide a level of abstraction by lemmatizing word forms and including only specific syntactic relationships. This includes propositions defined as nominal subjects paired with their heads and children of that head that are negators, modal auxiliaries (*can, may, might, shall, could, would*), particles and direct objects. All word forms are lemmatized, removing tense information on verbs and number on nouns.

   Example: ⟨Reagan, *give amnesty*⟩

3. Subject-verb tuples, which provide a more general layer of abstraction by only encoding the relationship between a subject and its main action. In this case, a proposition is defined as the nominal subject and its lemmatized head.

   Example: ⟨Reagan, *give*⟩

The human benchmark defined in §6.2.3 below considers only verbatim predicates, while all models proposed in §6.3 and all baselines in §6.4 include the union of all three representations as data.

Here, syntactic structure not only provides information in the representation of propositions, but also allows us to define criteria by which to exclude predicates — since we are looking to extract propositions that are directly asserted by an author of a blog comment (and not second-order reporting), we exclude all propositions dominated by an attitude predicate (*Republicans think that Obama should be impeached*) and all those contained within a conditional clause[11] (*If Obama were impeached...*). We also exclude all assertions drawn from questions (i.e., sentences containing a question mark) and all assertions extracted from quoted text (i.e., surrounded by quotation marks).

In total, from all 28 million comments across all seven blogs, we extract all propositions defined by the criteria above, yielding a total of 61 million propositions (45 million unique).

6.2.3 Human Benchmark

From all propositions with a verbatim predicate extracted from the entire dataset, we rank the most frequent subjects and manually filter out non-content terms (like *that*, *one*, *someone*, *anyone*, etc.) to yield a set of 138 target topics, the most frequent of which are *obama*, *democrats*, *bush*, *hillary*, and *america*.

For each proposition containing one of these topics as its subject and mentioned by at least 5 different people across all blogs, we randomly sampled 1,000 in proportion to their frequency of use (so that sentences that appear more frequently in the data are more likely to be sampled); the sentences selected in this random way contain a variety of politically charged viewpoints. We then presented them to workers on Amazon Mechanical Turk for judgments on the extent to which they reflect a liberal vs. conservative political worldview.

For each sentence, we paid 7 annotators in the United States to a.) confirm that the extracted sentence was a well-formed assertion and b.) to rate "the most likely political belief of the person who would say it" on a five-point scale: very conservative/Republican (-2), slightly conservative/Republican (-1), neutral (0), slightly liberal/Democrat (1), and very liberal/Democrat (2).

We keep all sentences that at least six annotators have marked as meaningful (those excluded by this criterion include sentence fragments like *bush wasn't* and those that are difficult to understand without context, such as *romney is obama*) and where the standard deviation of the responses is under 1 (which excludes sentences with flat distributions such as *government does nothing well* and those with bimodal distributions, such as *christie is done*). After this quality control, we average the responses to create a dataset of 766 propositions paired with their political judgments. Table 6.2 presents a random sample of annotations from this dataset.

[11] Narayanan et al. (2009) estimate conditionals account for about approximately 8% of sentences across a variety of domains.

proposition	mean	s.d.
obama lied and people died	-2.000	0.000
gay marriage is not a civil right	-1.857	0.350
obama can't be trusted	-1.714	0.452
hillary lied	-0.857	0.990
hillary won't run	-0.714	0.452
bush was just as bad	0.857	0.639
obama would win	1.429	0.495
rand paul is a phony	1.429	0.495
abortion is not murder	1.571	0.495
hillary will win in 2016	1.857	0.350

Table 6.2: Random sample of AMT annotations.

6.3 Models

The models we introduce to assess the political import of propositions are based on two fundamental ideas. First, users' latent political preferences, while unobserved, can provide an organizing principle for inference about propositions in an unsupervised setting. Second, by decoupling the variation in *subjects* discussed by different communities (e.g., liberals may talk more about global warming while conservatives may talk more about gun rights) from variation in what statements are *predicated* of those subjects (e.g., liberals may assert that ⟨global warming, *is a fact*⟩ while conservatives may be more likely to assert that it *is a hoax*), we are able to have a more flexible and interpretable parameterization of observed textual behavior that allows us to directly measure both.

We present two models below: one that represents users and propositions as real-valued points, and another that represents each as categorical variables. For both models, the input is a set of users paired with a list of ⟨subject, *predicate*⟩ tuples they author; the variables of interest we seek are representations of those users, subjects, and predicates that explain the coupling between users and propositions we see.

6.3.1 Additive Model

The first model we present (fig. 6.1) represents each user, subject, and predicate as a real-valued point in K-dimensional space. In the experiments that follow, we consider the simple case where $K = 1$ but present the model in more general terms below.

In this model, we again recapitulate a theme in this thesis by conditioning on metadata (here, the latent political position of an individual) in parameterizing a language model. Here, we parameterize the generative probability of a subject (like Obama) as used by an individual u as the exponentiated sum of a background log frequency of that subject in the corpus overall (m_{sbj}) and K additive effects, normalized over the space of S possible subjects, as a real-valued analogue to the SAGE model of (Eisenstein et al., 2011a). While the background term controls the overall frequency of a subject in the corpus, $\beta \in \mathbb{R}^{K \times S}$ mediates the relative increase or decrease in probability of a subject for each latent dimension. Intuitively, when both $\eta_{u,k}$ and $\beta_{k,sbj}$ (for a given user u, dimension k, and subject sbj) are the same sign (either both positive or both negative), the probability of that subject under that user increases; when they differ, it decreases. $\beta_{\cdot,sbj}$ is a K-dimensional representation of subject sbj, and $\eta_{u,\cdot}$ is a K-dimensional representation of user u.

$$P(sbj \mid u, \eta, \beta, m_{sbj}) = \\ \frac{\exp\left(m_{sbj} + \sum_{k=1}^{K} \eta_{u,k}\beta_{k,sbj}\right)}{\sum_{sbj'} \exp\left(m_{sbj'} + \sum_{k=1}^{K} \eta_{u,k}\beta_{k,sbj'}\right)} \tag{6.1}$$

Likewise, we parameterize the generative probability of a predicate (conditioned on a subject) in the same way; for S subjects, each of which contains (up to) P predicates, $\xi \in \mathbb{R}^{S \times K \times P}$ captures the relative increase or decrease in probability for a given predicate conditioned on its subject, relative to its background frequency in the corpus overall, $m_{pred|sbj}$.

$$P(pred \mid sbj, u, \eta, \xi, m_{pred|sbj}) = \\ \frac{\exp\left(m_{pred|sbj} + \sum_{k=1}^{K} \eta_k \xi_{sbj,k,pred}\right)}{\sum_{pred'} \exp\left(m_{pred'|sbj} + \sum_{k=1}^{K} \eta_k \xi_{sbj,k,pred'}\right)} \tag{6.2}$$

The full generative story for this model runs as follows. For a vocabulary of subjects of size S, where each subject s has P predicates:

- For each dimension k, draw subject coefficients $\beta_k \in \mathbb{R}^S \sim \text{Norm}(\mu_s, \sigma_s \mathbf{I})$
- For each subject s:
    - For each dimension k, draw subject-specific predicate coefficients $\xi_{s,k} \in \mathbb{R}^P \sim \text{Norm}(\mu_p, \sigma_p \mathbf{I})$
- For each user u:
    - Draw user representation $\eta \in \mathbb{R}^K \sim \text{Norm}(\mu, \sigma \mathbf{I})$
    - For each proposition $\langle sbj, pred \rangle$ made by u:
        - Draw sbj according to eq. 6.1
        - Draw $pred$ according to eq. 6.2

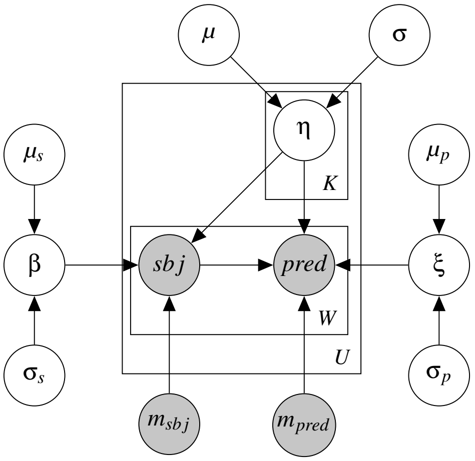


Figure 6.1: Additive model with decoupled subjects and predicates. η contains a K-dimensional representation of each user; β captures the variation in observed subjects, and ξ captures the variation in predicates for a fixed subject.

The unobserved quantities of interest in this model are η, β and ξ. In the experiments reported below, we set the prior distributions on η, β and ξ to be standard normals ($\mu = 0, \sigma = 1$) and perform maximum *a posteriori* inference with respect to η, β and ξ in turn for a total of 25 iterations.

While β and ξ provide scores for the political import of subjects and of predicates conditioned on fixed subjects, respectively, we can recover a single ideological score for both a subject and its predicate by adding their effects together. In the evaluation given in §6.5, let the PREDICATE SCORE for $\langle \textsf{subject}, \textit{predicate} \rangle$ be that given by $\xi_{\textsf{subject},\cdot,\textit{predicate}}$, and let the PROPOSITION SCORE be $\beta_{\cdot,\textsf{subject}} + \xi_{\textsf{subject},\cdot,\textit{predicate}}$.

6.3.2 Single Membership Model

While the additive model above represents each user and proposition as a real-valued point in K-dimensional space, we can also represent those values as categorical variables in an unsupervised naïve Bayes style parameterization; in this case, a user is not defined as a mixture of different effects, but rather has a single unique community to which they belong. The generative story for this model (shown in fig. 6.2) is as follows:

- Draw population distribution over categories $\theta \sim \mathrm{Dir}(\alpha)$
- For each category k, draw distribution over subjects $\phi_k \sim \mathrm{Dir}(\gamma)$
- For each category k and subject s:
  - Draw distribution over subject-specific predicates $\psi_{k,s} \sim \mathrm{Dir}(\gamma_s)$
- For each user u:

- Draw user type index $p \sim \text{Cat}(\theta)$
- For each proposition $\langle sbj, pred \rangle$ made by u:
  - Draw subject $sbj \sim \text{Cat}(\phi_p)$
  - Draw predicate $pred \sim \text{Cat}(\psi_{p,sbj})$

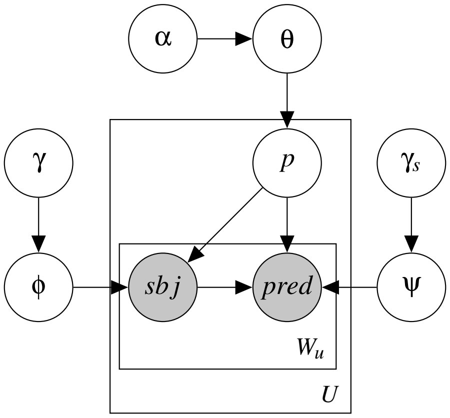


Figure 6.2: Single membership model with decoupled subjects and predicates. p is the latent category identity of a user (e.g., liberal or conservative); ϕ is a distribution over subjects for each category; and ψ is a distribution of predicates given subject s.

We set $K = 2$ in an attempt to recover a distinction between liberal and conservative users. For the experiments reported below, we run inference using collapsed Gibbs sampling (Griffiths and Steyvers, 2004) for 100 iterations, performing hyperparameter optimization on α, γ and γ_s (all asymmetric) every 10 using the fixed-point method of Minka (2003).

In order to compare the subject-specific predicate distributions across categories, we first calculate the posterior predictive distribution by taking a single sample of all latent variables p to estimate the following (Asuncion et al., 2009):

$$\hat{\zeta}_{p,v} = \frac{\mathbf{c}(p,v) + \gamma_v}{\sum_{v'} \mathbf{c}(p,v') + \gamma_{v'}} \tag{6.3}$$

Where $\hat{\zeta}_{p,v}$ is the vth element of the pth multinomial being estimated, $\mathbf{c}(p,v)$ is the count of element v associated with category p and γ_v is the associated Dirichlet hyperparameter for that element. Given this smoothed distribution, for each proposition we assign it a real valued score, the log-likelihood ratio between its value in these two distributions. In the evaluation that follows, let the PREDICATE SCORE for a given ⟨subject, *predicate*⟩ under this model be:

$$\log\left(\frac{\hat{\psi}_{0,\text{subject},\textit{predicate}}}{\hat{\psi}_{1,\text{subject},\textit{predicate}}}\right) \tag{6.4}$$

Let the PROPOSITION SCORE be:

$$\log\left(\frac{\hat{\phi}_{0,\text{subject}} \times \hat{\psi}_{0,\text{subject},\mathit{predicate}}}{\hat{\phi}_{1,\text{subject}} \times \hat{\psi}_{1,\text{subject},\mathit{predicate}}}\right) \tag{6.5}$$

6.4 Comparison

The two models described in §6.3 are unsupervised methods for estimating the latent political positions of users along with propositional assertions. We compare with three otther models, a mixture of unsupervised, supervised, and semi-supervised methods. Unlike our models, these were not designed for the task described in §6.2.

6.4.1 Principal Component Analysis

To compare against another purely unsupervised model, we evaluate against principal component analysis (PCA), a latent linear model that minimizes the average reconstruction error between an original data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ and a low-dimensional approximation $\mathbf{Z}\mathbf{W}^\top$, where $\mathbf{Z} \in \mathbb{R}^{n \times k}$ can be thought of as a k-dimensional latent representation of the input and $\mathbf{W} \in \mathbb{R}^{p \times k}$ contains the eigenvectors of the k largest eigenvalues of the covariance matrix $\mathbf{X}\mathbf{X}^\top$, providing a k-dimensional representation for each feature. We perform PCA with $k = 1$ on two representations of our data: a.) **counts**, where the input data matrix contains the counts for each feature for each user, and b.) **frequencies**, where we normalize those counts for each user to unit length. While the input data is sparse, we must center each column to have a 0 mean (resulting in a dense matrix) and perform PCA through a singular value decomposition of that column-centered data using the method of Halko et al. (2011); in using SVD for PCA, the right singular vectors correspond to the principal directions; from these we directly read off a $k = 1$ dimensional score for each proposition in our data.

6.4.2 ℓ_2-Regularized Logistic Regression

While unsupervised methods potentially allow us to learn interesting structure in data, they are often eclipsed in prediction tasks by the addition of any form of supervision. While purely supervised models give more control over the exact decision boundary being learned, they can suffer by learning from a much smaller training set than unsupervised methods have access to. To evaluate this tradeoff, we compare against a supervised model trained using naturally occurring data – users who self-declare themselves in their profiles to be *liberal*, *conservative*, *democrat*, or *republican*. We randomly sampled 150 users who self-identify as liberals and 150 who identify as conservatives. We do not expect these users to be a truly random sample of the population — those who self-declare their political affiliation are more likely to engage with political content differently from those who do not (Sandvig, 2015; Hargittai, 2015) — but is a method that has been used for political prediction tasks in the past (Cohen and Ruths, 2013).

We build a predictive model using two classes of features: a.) binary indicators of the most frequent 25,000 unigrams and multiword expressions[12] in the corpus overall; and b.) features derived from user posting activity to the seven blogs shown in table 6.1 (binary indicators of the blogs posted to, and the identity of the most frequent blog). In a tenfold cross-validation (using squared ℓ_2-regularized logistic regression), this classifier attains an accuracy rate of 76.7% (with a standard error of ±1.7 across the ten folds).

In order to establish real-valued scores for propositions, we follow the same method as for the single membership model described above, using the log likelihood ratio of the probability of the proposition under each condition, where that probability is given as the count of the proposition among users classified as (e.g.) liberals (plus some small smoothing factor) divided by the total number of propositions used by them overall.

$$score(prop) = \log \frac{P(prop \mid p = conservative)}{P(prop \mid p = liberal)} \tag{6.6}$$

6.4.3 Co-Training

Since the features we use for the supervised model provide two roughly independent views of the data, we also evaluate against the semi-supervised method of co-training (Blum and Mitchell, 1998). Here, we train two different logistic regression classifiers, each with access to only the unigrams and multiword expressions employed by the user (h_{words}) or to binary indicators of the blogs posted to and the identity of the most frequent blog (h_{blogs}). For ten iterations, we pick a random sample U' of 1,000 data points from the full dataset U and classify each using the two classifiers; each classifier then adds up to 100 of the highest-confidence predictions to the training set, retaining the class distribution balance of the initial training set. In a tenfold cross-validation, co-training yielded a slightly higher (but not statistically significant) accuracy over pure supervision (77.0% ±1.8). We calculate scores for propositions in the same way as for the fully supervised case above.

6.5 Evaluation

For the experiments that follow, we limit the input data available to all models to only those propositions whose subject falls within the evaluation benchmark; and include only propositions used by at least five different users, and only users who make at least five dif-

[12] Multiword expressions were found using the method of Justeson and Katz (1995).

ferent assertions, yielding a total dataset of 40,803 users and 1.9 million propositions (81,728 unique), containing the union of all three kinds of extracted propositions (§6.2.2).

Each of the automatic methods that we discuss above assigns a real-valued score to propositions like OBAMA IS A SOCIALIST. Our goal in evaluation is to judge how well those model scores recover those assigned by humans in our benchmark. Since each method may make different assumptions about the distribution of scores (and normalizing them may be sensitive to outliers), we do not attempt to model them directly, but rather use two nonparametric tests: Spearman's rank correlation coefficient and cluster purity.

Spearman's rank correlation coefficient. The set of scores in the human benchmark and as output by a model each defines a ranked list of propositions; Spearman's rank correlation coefficient (ρ) is a nonparametric test of the Pearson correlation coefficient measured over the ranks of items in two lists (rather than their values). We use the absolute value of ρ to compare the degree to which the ranked propositions of two lists are linearly correlated; a perfect correlation would have $\rho = 1.0$; no correlation would have $\rho = 0.0$.

Purity. While Spearman's rank correlation coefficient gives us a nonparametric estimate of the degree to which the exact order of two sequences are the same, we can also soften the exact ordering assumption and evaluate the degree to which a ranked proposition falls on the correct side of the political continuum (i.e., not considering whether OBAMA IS A SOCIALIST is more or less conservative than OBAMA IS A DICTATOR but rather that it is more conservative than liberal). For each ranked list, we form two clusters of propositions, split at the midpoint: all scores below the midpoint define one cluster, and all scores above or equal define a second. For $N = 766$ propositions, given gold clusters $\mathcal{G} = \{g_1, g_2\}$ and model clusters $\mathcal{C}_n = \{c_1, c_2\}$ (each containing 383 propositions), we calculate purity as the average overlap for the best alignment between the two gold clusters and their model counterparts.[13]

$$\text{Purity} = \frac{1}{N}\left(\max_j |g_1 \cap c_j| + \max_j |g_2 \cap c_j|\right) \tag{6.7}$$

A perfect purity score (in which all items from each cluster in $\mathcal{C}$ are matched to the same cluster in $\mathcal{G}$) is 1.0; given that all clusters are identically sized (being defined as the set falling on each half of a midpoint), a random assignment would yield a score of 0.50 in expectation.

Table 6.3 presents the results of this evaluation. For both of the models described in §6.3, we present results for scoring a proposition like OBAMA IS A SOCIALIST based only on the conditional predicate score (PRED.) and on a score that includes variation in the subject

[13] In this case, with two clusters on each side, the best alignment in maximal in that $g_{n,i} \rightarrow c_{n,j} \Rightarrow g_{n,\neg i} \rightarrow c_{n,\neg j}$.

Model	Purity	Spearman's ρ
Additive (PROP.)	0.757 ±0.020	0.639 $[0.595, 0.679]$
Single mem. (PROP.)	0.754 ±0.019	0.619 $[0.573, 0.661]$
Single mem. (PRED.)	0.702 ±0.018	0.548 $[0.496, 0.596]$
Additive (PRED.)	0.705 ±0.018	0.484 $[0.428, 0.536]$
Co-training	0.695 ±0.018	0.444 $[0.385, 0.499]$
LR	0.619 ±0.016	0.274 $[0.207, 0.338]$
PCA (frequency)	0.518 ±0.014	0.097 $[0.026, 0.167]$
PCA (counts)	0.514 ±0.014	0.065 $[0, 0.135]$

Table 6.3: Evaluation. Higher is better.

as well (PROP.). Since both models are fit using approximate inference with a non-convex objective function, we run five models with different random initializations and present the average across all five (the largest standard error across tasks is ±0.003, indicating that none of the models are very sensitive to initial conditions).

We estimate confidence intervals for Spearman's rank correlation coefficient using the Fisher transformation of ρ (Fieller et al., 1957):

$$F(\rho) = \frac{1}{2} \ln \frac{1+\rho}{1-\rho} \tag{6.8}$$

This value is approximately normally distributed in this transformed space, yielding a 95% confidence interval of $[\tanh(F(\rho) - 1.96\sqrt{n-3}), \tanh(F(\rho) + 1.96\sqrt{n-3})]$ (note the hyperbolic tangent function transforms this range back to its original space).

As a clustering metric, purity has no closed-form expression for confidence sets, and since its evaluation requires its elements to be unique (in order to be matched across clusters), we cannot use common resampling-with-replacement techniques such as the bootstrap (Efron, 1979). Instead, we estimate confidence intervals using the block jackknife (Quenouille, 1956; Efron and Stein, 1981), calculating purity over 76 resampled subsets of the full 766 elements, each leaving out 10. In both cases, the two best performing models show statistically significant improvement over all other models, but are not significantly different from each other.

We draw two messages from these results:

For heavily partisan data, unsupervised methods are sufficient. In drawing on comments on politically partisan blogs, we are able to match human judgments of the political import of propositions quite well (both of the unsupervised models described in §6.3 outperform their supervised and semi-supervised counterparts by a large margin), which suggests that the easiest structure to find in this particular data is the affiliation of users with their political ideologies. Both unsupervised models are able to exploit the natural structure without being constrained by a small amount of training data that may be more biased (e.g., in its class balance) than helpful. The two generative models also widely outperform PCA, which may reflect a mismatch between its underlying assumptions and the textual data we observe; PCA treats data sparsity as structural zeros (not simply missing data) and so must model not only the variation that exists between users, but also the variation that exists in their frequency of use; other latent component models may be a better fit for this kind of data.

Joint information is important. For both models, including information about the full joint probability of a subject and predicate together yields substantial improvements for both purity and the Spearman correlation coefficient compared to scores calculated from variation in the conditional predicate alone. While we might have considered variation in the predicate to be sufficient in distinguishing between political parties, we see that this is simply not the case; variation in the subject may help anchor propositions in the spectrum relative to each other.

6.6 Convergent Validity

The primary quantity of interest that we are trying to estimate in the models described above is the political position of an *assertion*; a user's latent political affiliation is only a helpful auxiliary variable in reaching this goal. We can, however, also measure the correlation of those variables themselves with other variables of interest, such as users' self-declarations of political affiliation and audience participation on the different blogs. Both provide measures of convergent validity that confirm the distinction being made in our models is indeed one of political ideology.

6.6.1 Correlation with Self-declarations

One form of data not exploited by the unsupervised models described above are users' self-declarations; we omit these above in order to make the models as general as possible (requiring only text and not metadata), but they can provide an independent measure of the distinctions our unsupervised models are learning. (The supervised baselines in contrast are able to draw on this profile information for training data.)

Approximately 12% of the users in the data input to our models (4,718 of 40,804) have affiliated self-declared profile information; the most frequent of these include *retired*, *businessman*, *student*, and *patriot*. For all of these users, we regress binary indicators of the top 25,000 unigrams in their profiles against the MAP estimate of their political affiliation in the single-membership model. Across all 5 folds, the features with the highest predictive weights for one class were *patriot*, *conservative*, *obama*, and *god* while the highest predictive weights for the other are *progressive*, *voter*, *liberal*, and *science*.

6.6.2 Estimating Media Audience

We can also use users' latent political ideologies to estimate the overall ideological makeup of a blog's active audience. If we assign each post to our estimate of the political ideology of its author, we find that Mother Jones has the highest fraction of comments by estimated liberals at 80.4%, while Breitbart has the highest percentage of comments by conservatives (79.5%).

Blog	% Liberal by post
Mother Jones	80.4%
NPR	67.4%
Politico	51.6%
Young Conservatives	38.0%
Daily Caller	28.4%
IJ Review	28.0%
Breitbart	20.5%

Table 6.4: Media audience.

This broadly accords with Mitchell et al. (2014), which finds that among the blogs in our dataset, consistently liberal respondents trust NPR and Mother Jones most, while consistent conservatives trust Breitbart most and NPR and Mother Jones the least.

6.7 Conclusion

In this chapter, we introduce the task of estimating the political import of propositions such as OBAMA IS A SOCIALIST; while much work in open information extraction has focused on learning facts such as OBAMA IS PRESIDENT from text, we are able to exploit the structure between such propositions and the people who assert them in order to align them all within the same political space. Given sufficiently partisan data (here, comments on political blogs), we find that the unsupervised generative models presented here are able to outperform other models, including those given access to supervision. Even when the latent attributes are not the intrinsic quantities of interests, they can still provide informative structure for meaningful downstream applications.

In this work, we represent users either by a single categorical variable or with a more expressive parameterization that sees users as points in continuous space. In both cases, however, the cardinality of this space only allows distinctions to be made on a single dimension: a user is either a member of class A or B (which we interpret to be *liberal* or *conservative*) or occupies a single point on a line. There is much room to improve on this representation; for the categorical case, we can again leverage insight from the persona models in chapters 3 and 4 to represent users not as members of one of two political parties, but as an individual with more fine-grained political beliefs—to take an example from contemporary politics, the Tea Party may tend often caucus with neoconservatives or libertarians, but each represents a very different set of political beliefs, and the beliefs of individuals are certainly even more varied. By capturing the more fine-grained distinctions that exist within our contemporary political landscape, we can elicit even more fine-grained associations between the propositions and the communities who assert them.

Chapter 7

Learning word representations through people

Work described in this chapter was undertaken in collaboration with Chris Dyer and Noah Smith and published at ACL 2014 (Bamman et al., 2014a)

7.1 Introduction

The models we have presented in the previous four chapters have all been unsupervised probabilistic generative models, where we articulate a relationship between variables that offers an explanation for the observed data we see, and where the target quantity of interest—the personas of characters in movies and books, the event class associated with event descriptions in text, and the political beliefs of individuals—are never observed. This chapter represents a break from that paradigm; here we consider metadata associated with people (their US state-level geographical location), and illustrate how incorporating this kind of data into a discriminative model of representation learning can again lead to a downstream application that is improved by considering qualities of people as authors in explaining the variation we see in text.

As this thesis illustrates, language is a profoundly *situated* phenomenon: it is spoken by a particular person in a particular place and time. Nowhere is this more apparent than in social media sources such as Twitter and Facebook, where a large volume of streaming text is paired with explicit information about its author and social-historical context. The coupling of text with demographic information has enabled computational modeling of linguistic variation, including uncovering words and topics that are characteristic of geographical regions (Eisenstein et al., 2010; O'Connor et al., 2010; Hong et al., 2012; Doyle, 2014), learning correlations between words and socioeconomic variables (Rao et al., 2010; Eisenstein et al., 2011b; Pennacchiotti and Popescu, 2011; Bamman et al., 2014b); and charting how new terms spread geographically (Eisenstein et al., 2014). These models can tell us that *hella* was (at one time) used most often by a particular demographic group in northern California, echoing earlier linguistic studies (Bucholtz, 2006), and that *wicked* is used most often in New England (Ravindranath, 2011); and they have practical applications, facilitating tasks like text-based geolocation (Eisenstein et al., 2010; Wing and Baldridge, 2011; Roller et al., 2012; Ikawa et al., 2012). One desideratum that remains, however, is how the *meaning* of these terms is shaped by geographical influences – while *wicked* is used throughout the United States to mean *bad* or *evil* ("he is a wicked man"), in New England it is often used as an adverbial intensifier ("my boy's wicked smart"). In leveraging grounded social media to uncover linguistic variation, what we want to learn is how a word's meaning is shaped by its geography.

In this chapter, we introduce a method that extends vector-space lexical semantic models to learn representations of geographically situated language. Vector-space models of lexical semantics have been a popular and effective approach to learning representations of word meaning (Lin, 1998; Turney and Pantel, 2010; Reisinger and Mooney, 2010; Socher et al., 2013a; Mikolov et al., 2013, *inter alia*). In bringing in extra-linguistic information to learn word representations, our work falls into the general domain of multimodal learning; while other work has used visual information to improve distributed representations (Andrews et al., 2009; Feng and Lapata, 2010; Bruni et al., 2011, 2012a,b; Roller and im Walde, 2013), this work generally exploits information about the object being described (e.g., *strawberry* and a picture of a strawberry); in contrast, we use information about the *author* to learn representations that vary according to contextual variables from the author's perspective. Unlike classic multimodal systems that incorporate multiple active modalities (such as gesture) from a user (Oviatt, 2003; Yu and Ballard, 2004), our primary input is textual data, supplemented with metadata about the author and the moment of authorship. This information enables learning models of word meaning that are sensitive to such factors, allowing us to distinguish, for example, between the usage of *wicked* in Massachusetts from the usage of that word elsewhere, and letting us better associate geographically grounded named entities (e.g, *Boston*) with their hypernyms (*city*) in their respective regions.

7.2 Model

The model we introduce is grounded in the distributional hypothesis (Harris, 1954), that two words are similar by appearing in the same kinds of contexts (where "context" itself can be variously defined as the bag or sequence of tokens around a target word, either by linear distance or dependency path). We can invoke the distributional hypothesis for many instances of regional variation by observing that such variants often appear in similar contexts. For example:

- my boy's *wicked* smart
- my boy's *hella* smart
- my boy's *very* smart

Here, all three variants can often be seen in an immediately pre-adjectival position (as is common with intensifying adverbs).

Given the empirical success of vector-space representations in capturing semantic properties and their success at a variety of NLP tasks (Turian et al., 2010; Socher et al., 2011; Collobert et al., 2011; Socher et al., 2013a), we use a simple, but state-of-the-art neural architecture (Mikolov et al., 2013) to learn low-dimensional real-valued representations of words. The graphical form of this model is illustrated in figure 7.1.

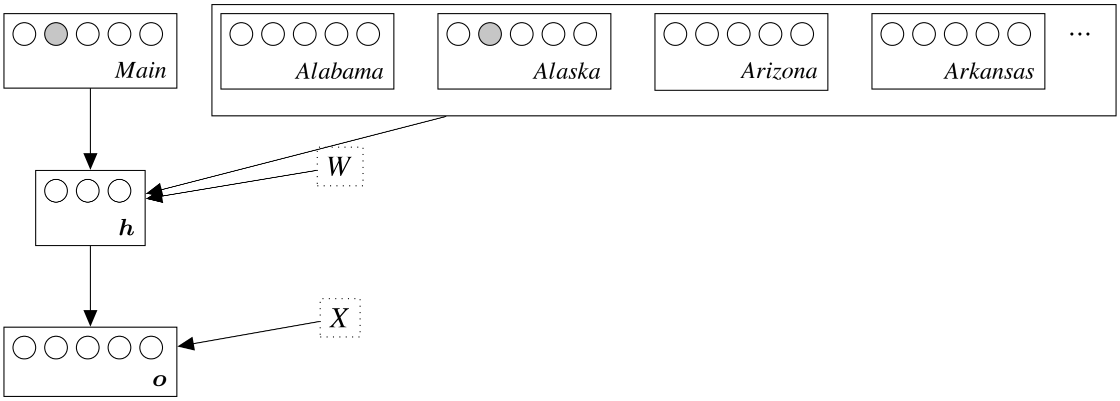


Figure 7.1: Model. Illustrated are the input dimensions that fire for a single sample, reflecting a particular word (vocabulary item #2) spoken in Alaska, along with a single output. Parameter matrix W consists of the learned low-dimensional embeddings.

This model corresponds to an extension of the "skip-gram" language model (Mikolov et al., 2013) (hereafter SGLM). Given an input sentence $\boldsymbol{s}$ and a context window of size t, each word s_i is conditioned on in turn to predict the identities of all of the tokens within t words around it. For a vocabulary V, each input word s_i is represented as a one-hot vector $\boldsymbol{w}_i$ of length $|V|$. The SGLM has two sets of parameters. The first is the representation matrix $W \in \mathbb{R}^{|V|\times k}$, which encodes the real-valued embeddings for each word in the vocabulary. A matrix multiply $\boldsymbol{h} = \boldsymbol{w}^\top W, \in \mathbb{R}^k$ serves to index the particular embedding for word $\boldsymbol{w}$, which constitutes the model's hidden layer. To predict the value of the context word y (again, a one-hot vector of dimensionality $|V|$), this hidden representation $\boldsymbol{h}$ is then multiplied by a second parameter matrix $X \in \mathbb{R}^{|V| \times k}$. The final prediction over the output vocabulary is then found by passing this resulting vector through the softmax function $\boldsymbol{o} = \text{softmax}(X\boldsymbol{h}^\top)$, giving a vector in the $|V|$-dimensional unit simplex. Backpropagation using (input x, output y) word tuples learns the values of W (the embeddings) and X (the output parameter matrix) that maximize the likelihood of y (i.e., the context words) conditioned on x (i.e., the s_i's). During backpropagation, the errors propagated are the difference between $\boldsymbol{o}$ (a probability distribution with k outcomes) and the true (one-hot) output y.

Let us define a set of contextual variables $\mathcal{C}$; in the experiments that follow, $\mathcal{C}$ is comprised solely of geographical state $\mathcal{C}_{state} = \{\text{AK}, \text{AL}, \ldots, \text{WY}\})$ but could in principle include any number of features, such as calendar month, day of week, or other demographic variables of the speaker. Let $|\mathcal{C}|$ denote the sum of the cardinalities of all variables in $\mathcal{C}$ (i.e., 51 states, including the District of Columbia). Rather than using a single embedding matrix W that contains low-dimensional representations for every word in the vocabulary, we define a global embedding matrix $W_{main} \in \mathbb{R}^{|V| \times k}$ and an additional $|\mathcal{C}|$ such matrices (each again of size $|V| \times k$, which capture the effect that each variable value has on each word in the vocabulary. Given an input word $\boldsymbol{w}$ and set of active variable values $\mathcal{A}$ (e.g., $\mathcal{A} = \{state = \text{MA}\}$), we calculate the hidden layer $\boldsymbol{h}$ as the sum of these independent embeddings: $\boldsymbol{h} = \boldsymbol{w}^\top W_{main} + \sum_{a \in \mathcal{A}} \boldsymbol{w}^\top W_a$. While the word *wicked* has a common low-dimensional representation in $W_{main,wicked}$ that is invoked for every instance of its use (regardless of the place), the corresponding vector $W_{\text{MA},wicked}$ indicates how that common representation should shift in k-dimensional space when used in Massachusetts. Backpropagation functions as in standard SGLM, with gradient updates for each training example $\{x, y\}$ touching not only W_{main} (as in SGLM), but all active $W_{\mathcal{A}}$ as well.

The additional W embeddings we add lead to an increase in the number of total parameters by a factor of $|\mathcal{C}|$. To control for the extra degrees of freedom this entails, we add squared ℓ_2 regularization to all parameters, using stochastic gradient descent for backpropagation with minibatch updates for the regularization term. As in Mikolov et al. (2013) and our work in chapter §4, we speed up computation using the hierarchical softmax (Morin and Bengio, 2005) on the output matrix X.

This model defines a joint parameterization over all variable values in the data, where information from data originating in California, for instance, can influence the representations learned for Wisconsin; a naive alternative would be to simply train individual models on each variable value (a "California" model using data only from California, etc.). A joint model has three *a priori* advantages over independent models: (i) sharing data across variable values encourages representations across those values to be similar; e.g., while *city* may be closer to *Boston* in Massachusetts and *Chicago* in Illinois, in both places it still generally connotes a *municipality*; (ii) such sharing can mitigate data sparseness for less-witnessed areas; and (iii) with a joint model, all representations are guaranteed to be in the same vector space and can therefore be compared to each other; with individual models (each with different initializations), word vectors across different states may not be directly compared.

7.3 Evaluation

We evaluate our model by confirming its face validity in a qualitative analysis and estimating its accuracy at the quantitative task of judging geographically-informed semantic similarity. We use 1.1 billion tokens from 93 million geolocated tweets gathered between September 1, 2011 and August 30, 2013 (approximately 127,000 tweets per day evenly sampled over those two years). This data only includes tweets that have been geolocated to state-level granularity in the United States using high-precision pattern matching on the user-specified location field (e.g., "new york ny" $\rightarrow$ NY, "chicago" $\rightarrow$ IL, etc.). As a preprocessing step, we identify a set of target multiword expressions in this corpus as the maximal sequence of adjectives + nouns with the highest pointwise mutual information; in all experiments described below, we define the vocabulary V as the most frequent 100,000 terms (either unigrams or multiword expressions) in the total data, and set the dimensionality of the embedding $k = 100$. In all experiments, the contextual variable is the observed US state (including DC), so that $|\mathcal{C}| = 51$; the vector space representation of word $\boldsymbol{w}$ in state s is $\boldsymbol{w}^\top W_{main} + \boldsymbol{w}^\top W_s$.

7.3.1 Qualitative Evaluation

To illustrate how the model described above can learn geographically-informed semantic representations of words, table 7.1 displays the terms with the highest cosine similarity to *wicked* in Kansas and Massachusetts after running our joint model on the full 1.1 billion words of Twitter data; while *wicked* in Kansas is close to other evaluative terms like *evil* and *pure* and religious terms like *gods* and *spirit*, in Massachusetts it is most similar to other intensifiers like *super*, *ridiculously* and *insanely*.

Table 7.2 likewise presents the terms with the highest cosine similarity to *city* in both California and New York; while the terms most evoked by *city* in California include regional locations like Chinatown, Los Angeles' South Bay and San Francisco's East Bay, in New York the most similar terms include *hamptons*, *upstate* and *borough* (New York City's term of

Kansas		Massachusetts	
term	cosine	term	cosine
wicked	1.000	wicked	1.000
evil	0.884	super	0.855
pure	0.841	ridiculously	0.851
gods	0.841	insanely	0.820
mystery	0.830	extremely	0.793
spirit	0.830	goddamn	0.781
king	0.828	surprisingly	0.774
above	0.825	kinda	0.772
righteous	0.823	#sarcasm	0.772
magic	0.822	sooooooo	0.770

Table 7.1: Terms with the highest cosine similarity to *wicked* in Kansas and Massachusetts.

California		New York	
term	cosine	term	cosine
city	1.000	city	1.000
valley	0.880	suburbs	0.866
bay	0.874	town	0.855
downtown	0.873	hamptons	0.852
chinatown	0.854	big city	0.842
south bay	0.854	borough	0.837
area	0.851	neighborhood	0.835
east bay	0.845	downtown	0.827
neighborhood	0.843	upstate	0.826
peninsula	0.840	big apple	0.825

Table 7.2: Terms with the highest cosine similarity to *city* in California and New York.

administrative division).

7.3.2 Quantitative Evaluation

As a quantitative measure of our model's performance, we consider the task of judging semantic similarity among words whose meanings are likely to evoke strong geographical correlations. In the absence of a sizable number of linguistically interesting terms (like *wicked*) that are known to be geographically variable, we consider the proxy of estimating the named entities evoked by specific terms in different geographical regions. As noted above, geographic terms like *city* provide one such example: in Massachusetts we expect the term *city* to be more strongly connected to grounded named entities like *Boston* than to other US cities. We consider seven categories for which we can reasonably expect the connotations of each term to vary by geography; in each case, we calculate the distance between two terms x and y using representations learned for a given state ($\delta_{state}(x, y)$). Note that some terms may be multiword expressions (such as *New York* or *Red Sox*); these are all contained within the vocabulary described above.

1. *city*. For each state, we measure the distance between the word *city* and the state's most populous city; e.g., $\delta_{\text{AZ}}(\textit{city}, \textit{phoenix})$.

2. *state*. For each state, the distance between the word *state* and the state's name; e.g., $\delta_{\text{WI}}(\textit{state}, \textit{wisconsin})$.

3. *football*. For all NFL teams, the distance between the word *football* and the team name; e.g., $\delta_{\text{IL}}(\textit{football}, \textit{bears})$.

4. *basketball*. For all NBA teams from a US state, the distance between the word *basketball* and the team name; e.g., $\delta_{\text{FL}}(\textit{basketball}, \textit{heat})$.

5. *baseball*. For all MLB teams from a US state, the distance between the word *baseball* and the team name; e.g., $\delta_{\text{IL}}(\textit{baseball}, \textit{cubs})$, $\delta_{\text{IL}}(\textit{baseball}, \textit{white sox})$.

6. *hockey*. For all NHL teams from a US state, the distance between the word *hockey* and the team name; e.g., $\delta_{\text{PA}}(\textit{hockey}, \textit{penguins})$.

7. *park*. For all US national parks, the distance between the word *park* and the park name; e.g., $\delta_{\text{AK}}(\textit{park}, \textit{denali})$.

Each of these questions asks the following: what words are evoked for a given target word (like *football*)? While *football* may everywhere evoke similar sports like *baseball* or *soccer* or more specific football-related terms like *touchdown* or *field goal*, we expect that particular sports teams will be evoked more strongly by the word *football* in their particular geographical region: in Wisconsin, *football* should evoke *packers*, while in Pennsylvania, *football* evokes *steelers*. Note that this is not the same as simply asking which sports team is most frequently (or most characteristically) mentioned in a given area; by measuring the distance to a target word (*football*), we are attempting to estimate the varying strengths of association between concepts in different regions.

For each category, we measure similarity as the average cosine similarity between the vector for the target word for that category (e.g., *city*) and the corresponding vector for each state-specific answer (e.g., *chicago* for IL; *boston* for MA). We compare three different models:

1. JOINT. The full model described in section 7.2, in which we learn a global representation for each word along with deviations from that common representation for each state.

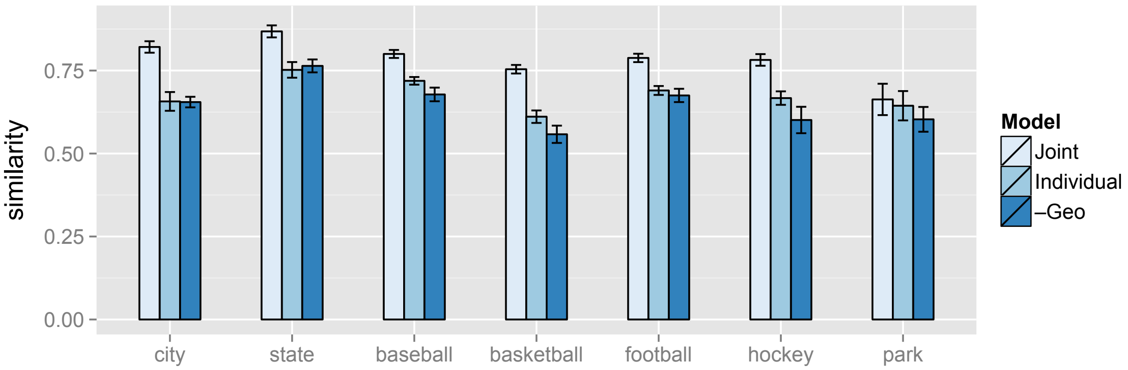


Figure 7.2: Average cosine similarity for all models across all categories, with 95% confidence intervals on the mean.

2. Individual. For comparison, we also partition the data among all 51 states, and train a single model for each state using only data from that state. In this model, there is no sharing among states; California has the most data with 11,604,637 tweets; Wyoming has the least with 47,503 tweets.

3. –Geo. We also train a single model on all of the training data, but ignore any state metadata. In this case the distance δ between two terms is their overall distance within the entire United States.

As one concrete example of these differences between individual data points, the cosine similarity between *city* and *seattle* in the –Geo model is 0.728 (*seattle* is ranked as the 188th most similar term to *city* overall); in the Individual model using only tweets from Washington state, $\delta_{WA}(city, seattle) = 0.780$ (rank #32); and in the Joint model, using information from the entire United States with deviations for Washington, $\delta_{WA}(city, seattle) = 0.858$ (rank #6). The overall similarity for the city category of each model is the average of 51 such tests (one for each city).

Figure 7.2 present the results of the full evaluation, including 95% confidence intervals for each mean. While the two models that include geographical information naturally outperform the model that does not, the Joint model generally far outperforms the Individual models trained on state-specific subsets of the data.[1] A model that can exploit all of the information in the data, learning core vector-space representations for all words along

[1] This result is robust to the choice of distance metric; an evaluation measuring the Euclidean distance between vectors shows the Joint model to outperform the Individual and –Geo models across all seven categories.

with deviations for each contextual variable, is able to learn more geographically-informed representations for this task than strict geographical models alone.

7.4 Conclusion

This chapter introduces a model for leveraging situational information in learning vector-space representations of words that are sensitive to the author's social context. While our results use geographical information in learning low-dimensional representations, other contextual variables are straightforward to include as well; incorporating effects for time – such as time of day, month of year and absolute year – may be a powerful tool for revealing periodic and historical influences on lexical semantics.

Our approach explores the degree to which geography, and other contextual factors, influence word *meaning* in addition to frequency of usage. By allowing all words in different regions (or more generally, with different metadata factors) to exist in the same vector space, we are able compare different points in that space – for example, to ask what terms used in Chicago are most similar to *hot dog* in New York, or what word groups shift together in the same region in comparison to the background (indicating the shift of an entire semantic field). For all of these directions, this work is only made possible by leveraging contextual information (an author's geographic location); by incorporating this information into models of lexical semantics, we are able to paint a more nuanced picture of word sense variability than if we were to consider text as a disembodied phenomenon.

Chapter 8

Improving sarcasm detection with situated features

Work described in this chapter was undertaken in collaboration with Noah Smith and published at ICWSM 2015 (Bamman and Smith, 2015)

8.1 Introduction

In the work considered in this thesis so far, we have modeled the interaction of people as the *content* and *authors* of text. This chapter presents a case study incorporating a third axis of interaction: people as the *audience* of text.

For many NLP tasks that primarily operate on non-dialogic, written text, it is easy to forget that language is primarily a communicative act involving at least two interlocutors; while the medium of written text can distance this relationship between a speaker/author and their interlocutor/reader[1] the audience still has the power to shape the text we see by being a guiding principle for the author. This relationship between authors and audience is crucial for many natural language understanding tasks that require shared knowledge between interlocutors. Humor (Mihalcea and Strapparava, 2005; Petrovic and Matthews, 2013) is heavily dependent on the interaction between an author and audience; so too is sarcasm.

Most approaches to sarcasm detection to date have treated the task primarily as a text categorization problem, relying on the insights of Kreuz and Caucci (2007) that sarcastic utterances often contain lexical indicators (such as interjections and intensifiers) and other

[1] As pointed out Plato, *Phaedrus* 275d: "And so it is with written words; you might think they spoke as if they had intelligence, but if you question them, wishing to know about their sayings, they always say only one and the same thing." (Fowler, 1925; Derrida, 1972)

linguistic markers (such as nonveridicality and hyperbole) that signal their irony. These purely text-based approaches can be surprisingly accurate across different domains (Carvalho et al., 2009; Davidov et al., 2010; González-Ibáñez et al., 2011; Riloff et al., 2013; Lukin and Walker, 2013; Reyes et al., 2013), but are divorced from any notion of their potentially useful *context*. Yet context seems to matter. For example, humans require access to the surrounding context in which a Reddit post was written (such as the thread it appears in) in order to judge its tone (Wallace et al., 2014). On Twitter, modeling the relationship between a tweet and an author's past tweets can improve accuracy on this task (Rajadesingan et al., 2015).

This kind of contextual information is only one small part of the shared common ground that must be present between a speaker and their audience in order for sarcasm to be available for use between them. Kreuz (1996) calls this the "principle of inferability" – speakers only use sarcasm if they can be sure it will be understood by the audience – and finds in surveys that sarcasm is more likely to be used between two people who know each other well than between those who do not.

In all of these cases, the relationship between author and audience is central for understanding the sarcasm phenomenon. While the notion of an "audience" is relatively well defined for face-to-face conversations between two people, it becomes more complex when multiple people are present (Bell, 1984), and especially so on social media, when a user's "audience" is often unknown, underspecified or "collapsed" (boyd, 2008; Marwick and boyd, 2011), making it difficult to fully establish the shared ground required for sarcasm to be detected, and understood, by its intended (or imagined) audience.

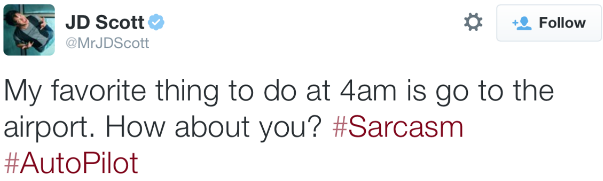
We present here a series of experiments to discern the effect of extra-linguistic information on the detection of sarcasm, reasoning about features derived not only from the local context of the message itself (as in past work), but also using information about the author, their relationship to their audience and the immediate communicative context they both share. Our main findings are:

- Including any aspect of the environment (features derived from the communicative context, the author, or the audience) leads to improvements in prediction accuracy.

- Users are more likely to tag their message with the explicit hashtag `#sarcasm` when they are *less* familiar with their audience. Following Kreuz (1996), we argue that this is a means of ensuring inferability in the face of uncertainty.

8.2 Data

Prior work on sarcasm detection on Twitter (González-Ibáñez et al., 2011) found low agreement rates between human annotators at the task of judging the sarcasm of *others'* tweets; consequently, recent research exploits users' self-declarations of sarcasm in the form of `#sarcasm` or `#sarcastic` tags of their own tweets. This design choice does not capture the likely more common varieties of sarcasm expressed without an explicit hashtag, but does yield positive examples with high precision. Figure 8.1 gives one such example.



My favorite thing to do at 4am is go to the airport. How about you? #Sarcasm #AutoPilot

Figure 8.1: User self-reporting of sarcasm.

We follow the same methodology here as well, identifying the authors of all tweets mentioning `#sarcasm` or `#sarcastic` in the Gardenhose sample of tweets from August 2013–July 2014, and crawling up to the most recent 3,200 tweets of those authors. As in past work, we label a tweet as SARCASTIC if it contains the hashtag `#sarcasm` or `#sarcastic` as its final term, is written in English, is not a retweet, and contains at least three words. To explore the influence of the communicative context on our perception of sarcasm, we further subsample this set to include only tweets that are responses to another tweet. This yields a positive training set of 9,767 tweets; for negative data, we select an equal number of tweets from users over the same time period who have *not* mentioned `#sarcasm` or `#sarcastic` in their messages. The total dataset is evenly balanced at 19,534 tweets. Since the hashtags `#sarcasm` and `#sarcastic` are used to define the positive examples, we remove those tags from all tweets for the prediction task.

8.3 Experimental Setup

For the classification task of deciding whether a tweet is SARCASTIC or NOT SARCASTIC, we adopt binary logistic regression with ℓ_2 regularization using tenfold cross-validation, split on authors (so that tweets by the same author do not appear in multiple splits). We tune the ℓ_2 regularization parameter on development data (train on $^8/_{10}$, tune on $^1/_{10}$, test on the remaining held-out $^1/_{10}$) and repeat across ten folds. We perform this cross-validation and parameter tuning for *every* feature combination reported below, since different feature sets (with different cardinalities) will result in different optimal parameter settings.

8.4 Features

We divide the features used in our models into four classes: those scoped only over the immediate tweet being predicted (§8.4.1); those that reason over the author of that tweet, including historical data by that author (§8.4.2); those that reason over the addressee of the tweet (the person to whom the target tweet under consideration is responding), including historical data for that individual and the author's history of interaction with them (§8.4.3); and features that consider the interaction between the tweet being predicted and the tweet that it is responding to (§8.4.4).

The baseline accuracy, using only the majority class in each training fold, is 47.4% (this is lower than an even 50% since the folds are split by author and contain varying numbers of tweets). In describing each feature below, we also report in parentheses the tenfold cross-validated accuracy of a model trained *only* on that feature type.

8.4.1 Tweet Features

- **Word unigrams (72.4%) and bigrams (69.5%).** We create binary indicators of lowercased word unigrams and bigrams. The most indicative unigrams include *dare, shocked, clearly, #lol* and *gasp*, and the most indicative bigrams include *you mean*, *how dare*, *i'm shocked*, *i'm sure* and *at all*.

- **Brown cluster unigrams (72.0%) bigrams (69.1%).** For dimensionality reduction, we map each word in our vocabulary to one of 1000 non-overlapping clusters using the Brown clusters of Owoputi et al. (2013), which group words used in similar contexts into the same cluster. We compute unigrams and bigrams over terms in this reduced space.

- **Unlabeled dependency bigrams, lexicalized (70.3%) and Brown clusters (70.2%).** We create binary features from unlabeled dependency arcs between a.) two words and b.) their corresponding Brown clusters after parsing the tweet with TweeboParser (Kong et al., 2014).

- **Part of speech features (66.0%).** Past work has shown that part of speech information (such as the density of hashtags and emoticons) is among the most informative for this task. We apply the POS tagger of Owoputi et al. (2013) and include features based on the absolute count and ratio of each of the 25 tags, along with the "lexical density" of the tweet, which models the ratio of nouns, verbs, adjectives and adverbs to all words (Rajadesingan et al., 2015).

- **Pronunciation features (57.5%)** To model the use of Twitter-specific writing style (as in Rajadesingan et al., 2015), we include the number of words with only alphabetic characters but no vowels (e.g., *btw*) and the number of words with more than three syllables.

- **Capitalization features (57.5%).** We include the number of words with initial caps and all caps and the number of POS tags with at least initial caps.

- **Tweet whole sentiment (55.0%).** We include several types of tweet-level sentiment features. The first is a feature containing the numeric value of the entire tweet's sentiment as determined by the Stanford Sentiment Analyzer (Socher et al., 2013b); since this phrase-based analyzer also determines the sentiment value of each non-terminal node in its syntactic parse tree, we also include the fraction of nonterminals with each sentiment score as a feature (which allows us to capture differences in sentiments across the tree).

- **Tweet word sentiment (53.7–54.7%).** As in much past work, we also include word-level sentiment features, modeling the maximum word sentiment score, minimum word sentiment score, and distance between the max and min. As in Rajadesingan et al. (Rajadesingan et al., 2015), we use the dictionaries of Warriner et al. (Warriner et al., 2013) (54.7%) and the emotion scores of Thelwall et al. (Thelwall et al., 2010) (53.7%).

- **Intensifiers (50.1%).** Since prior theoretical work has stressed the importance of hyperbole for sarcasm (Kreuz and Roberts, 1995), we include a binary indicator for whether the tweet contains a word in a list of 50 intensifiers (*so, too, very, really*) drawn from Wikipedia (`http://en.wikipedia.org/wiki/Intensifier`).

8.4.2 Author Features

- **Author historical salient terms (81.2%).** We create one feature for each term in a vocabulary shared among all authors; for a given tweet t containing word w, the feature $f(t, w) = 1$ if w is among the 100 highest-scoring TF-IDF terms used by the tweet's author in the past, and 0 otherwise. The terms with the highest weight predicting sarcasm are *#fail, govt, humor, lol, fact, excited, ff, truth, :-P* and *#gameofthrones*. This is the single most informative feature of all those we evaluated.

- **Author historical topics (77.4%).** We create broader topic-based features by inferring a user's topic proportions under LDA (Blei et al., 2003) with 100 topics over all tweets, where each document consists of up to 1,000 words of a user's tweets (excluding all messages in the test dataset). The topics most indicative of sarcasm include those relating to art and television shows.

- **Profile information (73.7%).** We create features for the author of the tweet drawn from their user profile information, including gender (as inferred by their first name, compared to trends in U.S. Social Security records), number of friends, followers and statuses, their duration on Twitter, the average number of posts per day, their timezone, and whether or not they are verified by Twitter (designating a kind of celebrity status). Being unverified, male, and from time zones in the United States are all strong markers of sarcasm.

- **Author historical sentiment (70.8%).** As in Rajadesingan et al. (2015), we model the distribution over sentiment in the user's historical tweets (excluding the test dataset), using the same word-level dictionaries applied to tweet-level sentiment described above. Users with historically negative sentiments have higher likelihoods of sarcasm.

- **Profile unigrams (66.2%).** We create binary indicators for all unigrams in the author's profile. The most indicative terms include *sarcasm, chemistry, #atheist* and *humor*.

8.4.3 Audience Features

- **Author historical topics (71.2%), Author historical salient terms (70.0%), Profile unigrams (68.6%), Profile information (66.3%).** As above, but for the author of the tweet to which the target tweet being predicted responds.

- **Author/Addressee interactional topics (73.9%).** To capture the similarity in interests between the author and addressee, we include features defined by the elementwise product of the author and addressee's historical topic distribution (resulting in a feature that it high if the two have both together tweeted about the same topics).

- **Historical communication between author and addressee (61.7%).** To model Kreuz's finding that sarcasm is more likely to take place between two people who are more familiar with each other, we include features that model that the degree of interaction between two users, including the number of previous messages sent from the author to the addressee, the rank of the addressee among the user's @-mention recipients and whether or not there have been at least one (and two) mutual @-messages exchanged between the author and the addressee (i.e., not simply unrequited messages sent from the author).

8.4.4 Environment Features

- **Pairwise Brown features between the original message and the response (71.7%).** To model the interaction between a target tweet and the tweet to which it is responding, we include binary indicators of pairwise Brown features between all terms in the two tweets.

- **Unigram features of the original message (68.8%).** To capture the original linguistic context a tweet is responding to, we include binary indicators of all unigrams in the original tweet as features. The most indicative terms in the original tweet include clear markers that already define a sarcastic environment, including *#sarcasm, sarcastic* and *sarcasm* as well as *worry, defense, advice, vote* and *kidding*.

8.5 Evaluation

To compare performance across different features, we consider five feature combinations: those with access to a.) tweet-level features; b.) tweet-level features and response features; c.) tweet features and audience features; d.) tweet features and author features; and e.) a combined model that includes all feature sets.

Figure 8.2 illustrates the relative gains in accuracy that result from including contextual information outside the immediate scope of the tweet being predicted: while tweet-only information yields an average accuracy of 75.4% across all ten folds, adding response features pushes this to 77.3%, audience features to 79.0% and author features to 84.9%. Including all features together yields the best performance at 85.1%, but most of these gains come simply from the addition of author information.

While the individual features above all report the accuracy of a model trained *only* on that feature, an ablation test (training the model on the full feature set excluding one feature) reveals that no feature is crucial for model performance: the most critical features are AUTHOR HISTORICAL SALIENT TERMS (–0.011), AUTHOR PROFILE FEATURES (–0.008), PAIRWISE BROWN FEATURES between the original message and the response (–0.008), PART OF SPEECH FEATURES (–0.002) and RESPONSE UNIGRAMS (–0.001). Training a model on these five features alone yields an accuracy of 84.3%, less than a point behind the full feature set.

8.6 Analysis

While features derived from the author yield the greatest improvements in accuracy over the tweet alone, all feature classes—including those drawn from the response, as well

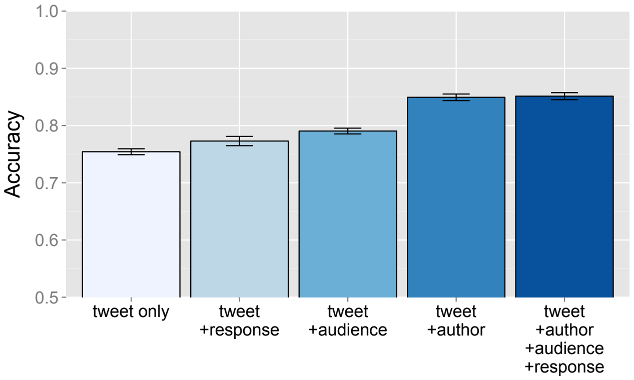


Figure 8.2: Accuracy across different feature sets, with 95% confidence intervals on the mean across 10 folds.

as those that consider the interaction of people in the form of audience and authors—display statistically significant improvements over the tweet-only features that ignore the communicative context. This confirms an effect on the interaction of the author and audience in the recognition of sarcasm, which can lead us to ask: who is this audience, and what about them is predictive of sarcasm across users? While Kreuz (1996) shows that sarcasm is primarily available between people who know each other well, we find that the strongest audience-based features that act as markers of sarcasm in this dataset are not those that suggest intimacy between the author and audience; the strongest audience predictors of sarcasm are the *absence* of mutual mentions (at least one mutual mention is a contraindicator, and at least two is more so); living in different time zones (i.e., not being geographically proximate) and features of celebrity (being verified and having many followers). In total, these features suggest that the `#sarcasm` hashtag is not a natural indicator of sarcasm expressed between friends, but rather serves an important communicative function of signaling the author's intent to an audience who may not otherwise be able to draw the correct inference about their message (as distinct from close friends who may be able to infer sarcasm without such labels).

8.7 Conclusion

With this chapter, we illustrate the important effect that people have in their roles as authors and audience of text—by leveraging any kind of situational information about people in these roles, we see improvements in predictive accuracy for sarcasm detection, a linguistic phenomenon that requires shared common ground between interlocutors. While sarcasm represents one such task that heavily relies on context, there are many others that lean on real contextual knowledge for understanding. As mentioned above, computational models of humor that not only model internal properties of text (such as surprising word combinations) but also the specific kind of relationship that holds between interlocutors, including models of mental states (X is aware that Y knows Z) may lead to similar improvements. At a broader level, the same may equally be true of classic NLP tasks like word sense disambiguation or semantic parsing: knowing something of the history of interaction between an author and their audience may help in discriminating which sense of *bank* is intended in a given utterance—where in addition to the maxims of "one sense per discourse" (Gale et al., 1992) and "one sense per Tweeter" (Gella et al., 2014), we might find the same to hold of edges in a social graph. In either case, this work highlights how useful contextual information can be—we know language is situated, and considering the depth of this context can lead not only to more realistic models of the world (from which we can gain social insight) but also to real gains in practical downstream tasks.

Chapter 9

Conclusion

9.1 Summary of contributions

In this thesis, I explore the utility of text analysis from a perspective centered on people, investigating the role that people play in their threefold interaction with text: as **content** of that text itself, as the **authors** of text, and as the **audience** of text. Each of these dimensions defines a research axis along which we can analyze text as it is socially embedded; while it is tempting to view the text we have available to us as existing in isolation from its material, historical context, it is, in contrast, a profoundly situated phenomenon, involving people in each of these dimensions. Detailed contributions of this work include the following.

- I present several models for learning latent personas of characters in movies (chapter 3) and books (chapter 4) by exploiting statistical regularities in the actions they perform and they attributes by which they are described, both in stylistically homogeneous and heterogeneous texts.

- I present a model (chapter 5) for learning latent event classes from a set of timestamped sentences in biographies; in an analysis of 242,970 biographies on Wikipedia, I uncover evidence of bias in the characterization of women, with up to four times the emphasis on events of marriage and divorce compared to men.

- I present a model (chapter 6) for learning the political import of type-level propositions like ⟨Obama, *is a Socialist*⟩, leveraging latent user political preferences in their estimation. The ideal points of propositions learned in this way outperform those learned from a variety of other methods (unsupervised, supervised and semi-supervised) when evaluated against human judgments.

- I present a model (chapter 7) for incorporating context-level metadata in learning low-dimensional word embeddings. By incorporating observed author-level geographical location, I show that we can learn embeddings that are sensitive to geographic variation and reflect a more realistic view of situated lexical semantics.
- I develop a predictive model for sarcasm detection (chapter 8) that considers not simply the text of a message (as in past work), but also *contextual* features, including those scoped over the author and the audience; all contextual features lead to improvements over features that only consider the message in isolation, showing that author and audience information is important for solving this richly contextual problem.

At a high level, the first section of this thesis (chapters 3, 4 and 5) illustrates the value of using *people* as an organizing principle in text: by leveraging the statistical regularities with which they are described, we can both learn interesting commonalities among them (such as common entity types they represent) and gain insight into where those commonalities defy expectations (as with the representation of women on Wikipedia). The second section (chapters 6, 7 and 8), in contrast, illustrates the value of incorporating more *contextual* information about people into text processing—in this case, information about the author and audience. Reasoning about this wider context in which the text we observe is situated can both open up new socially-informed analysis (allowing us to learn geographic sense distinctions and the political force of textual propositions) while also improving our performance on existing tasks (such as sarcasm detection).

9.2 Future directions

The work presented in this thesis outlines the three axes with which people interact with text. There are several directions that deserve further pursuit.

9.2.1 Richer models

The personas we inferred in chapters 3 and 4 are essentially clusters of characters that share some commonality in the actions they perform and the attributes with which they are described. The work that we have presented here is only the tip of the iceberg in investigating how character in represented (and how it varies) in text. The models we present can be directly extended to incorporate other varieties of contextual metadata, both at the author level (such as historical date of composition) and at the character level (such as gender) and to include more realistic assumptions of how a single character can evolve over the course of a book or movie, and how characters' types are dependent not simply on unary actions but also on relational information with other characters (and character's types). On a bigger scale, there is much more to be done in investigating what these (and other) models of character can reveal about literary history. These models are useful for exploratory data analysis, and there is much work to be done to explore what substantive disciplinary questions—if any—this exploration can help address.

9.2.2 Supervision

Most of the models presented in this thesis have been used in an unsupervised setting, where the goal is to learn interesting structure that exists in data, whether that data is comprised of movie summaries, the full text of books, or political blog comments. However, in many cases these models can be most useful to practitioners when they are allowed to be guided by supervision. In this case, this could be an ontology of character types drawn from specific works of literary criticism, particular types that characters in certain books have been argued to embody, or even adjusting the model output to meet human expectations, as in Hu et al. (2014). How can these models best enable exploratory analysis that is supported by human-directed guidance?

9.2.3 Fine-grained opinion mining

The models in chapter 6 learn the political import of propositions like ⟨Obama, *is a Socialist*⟩ in an unsupervised setting, leveraging users' latent political beliefs in their estimation. This work draws inspiration from open information extraction (Fader et al., 2011; Mitchell et al., 2015), which has been primarily concerned with learning facts from open data sources (such as broad samples from the web). There is an opportunity to bring these two lines of research together to build a similarly large-scale knowledge base of fine-grained propositions paired with the metadata (such as political affiliation) of the populations that assert them. While such a knowledge base may be useful for problems of fact assessment (Nakashole and Mitchell, 2014), it also presents an opportunity for large-scale, fine-grained opinion polling. Much current work has used social media sources like Twitter for correlating user micropublications with political polls (O'Connor et al., 2010) and flu trends (Paul and Dredze, 2011); this work presents a direction for estimating what fraction of those users feel *Obama* is a *socialist*, *dictator*, *great president*, and so on, along with the likely political affiliations of their communities.

9.2.4 Stereotyping

Chapter 6 also models the variability that exists in the predicates that people assert of the same fixed subject—⟨Obama, *is a Socialist*⟩ is a more conservative proposition, while ⟨Obama, *is a genius*⟩ is more liberal. This principle—modeling variability in text around a fixed target—is useful not only for broad applications like fine-grained opinion mining, but also for other tasks as well. At a social level, the question of how individuals differently describe the same fixed phenomenon has been richly studied in the context of framing (Entman, 1993). By restricting our scope of those phenomena to people, we enable study of how different individuals characterize social *groups* in different ways; this paves the way for the computational measurement of such phenomena as stereotyping.

9.2.5 Richer context

The approach taken in chapter 8 to identifying sarcasm in text—considering not simply the text itself for lexical indicators, but also the broader context including the author and audience—touches on the very interesting problem of estimating the shared context that interlocutors hold: sarcasm can often be successful because an author can exploit their audience's understanding of their own belief system (Kreuz, 1996). To what extent can we measure the degree to which two interlocutors have a shared background context, either by being close friends or immersed in the same cultural environment? Are there linguistic indicators that help reveal this?

At a high level, the development of models that can exploit **richer context** is the main future research direction to which this entire thesis points. While many classical NLP tasks (like part-of-speech tagging or syntactic parsing) only need to reason about the internal structure of a sentence (and of language more broadly), this work points toward tasks that rely on understanding language as it is used in the world—not handed to us as a corpus *ex nihilo*, but as spoken by people, to others, all of whom have their own beliefs and intentions, who are immersed in a wider, shared context, and who differ from each other in meaningful (and occasionally predictable) ways. The more we can incorporate this rich variation into our models of text, the more realistic they will be, with potential both for improved performance on practical tasks and for greater social insight into the circumstances of its production.

Chapter 10

Bibliography

Faiyaz Al Zamal, Wendy Liu, and Derek Ruths. Homophily and latent attribute inference: Inferring latent attributes of Twitter users from neighbors. In *ICWSM*, 2012.

Galen Andrew and Jianfeng Gao. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pages 33–40, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-793-3. doi: 10.1145/1273496.1273501. URL `http://doi.acm.org/10.1145/1273496.1273501`.

Mark Andrews, Gabriella Vigliocco, and David Vinson. Integrating experiential and distributional data to learn semantic representations. *Psychological Review*, 116(3):463–498, 2009.

Shlomo Argamon, Moshe Koppel, James W Pennebaker, and Jonathan Schler. Mining the blogosphere: Age, gender and the varieties of self-expression. *First Monday*, 12(9), 2007.

Arthur U. Asuncion, Max Welling, Padhraic Smyth, and Yee Whye Teh. On smoothing and inference for topic models. In Jeff Bilmes and Andrew Y. Ng, editors, *UAI*, pages 27–34. AUAI Press, 2009.

Amit Bagga and Breck Baldwin. Entity-based cross-document coreferencing using the vector space model. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pages 79–85, Montréal, Québec, Canada, August 1998.

Jason Baldridge and Alex Lascarides. Probabilistic head-driven parsing for discourse structure. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, pages 96–103. Association for Computational Linguistics, 2005.

David Bamman and Noah A. Smith. Unsupervised discovery of biographical structure from text. Transactions of the ACL, 2014.

David Bamman and Noah A. Smith. Contextulized sarcasm detection on Twitter. In *ICWSM*, 2015.

David Bamman, Brendan O'Connor, and Noah A. Smith. Learning latent personas of film characters. *Proc. of ACL*, 2013.

David Bamman, Chris Dyer, and Noah A. Smith. Distributed representations of geographically situated language. In *ACL*, 2014a.

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2), 2014b.

David Bamman, Ted Underwood, and Noah A. Smith. Appendix to 'A Bayesian mixed effects model of literary character'. Technical report, Carnegie Mellon University, University of Illinois-Urbana Champaign, 2014c.

David Bamman, Ted Underwood, and Noah A. Smith. A Bayesian mixed effects model of literary character. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 370–379, Baltimore, Maryland, June 2014d. Association for Computational Linguistics.

Michele Banko, Michael J Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. Open information extraction for the web. In *IJCAI*, volume 7, pages 2670–2676, 2007.

Pablo Barberá. Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political Analysis*, 23(1):76–91, 2015.

Allan Bell. Language style as audience design. *Language in Society*, 13:145–204, 1984.

Charley Beller, Rebecca Knowles, Craig Harman, Shane Bergsma, Margaret Mitchell, and Benjamin Van Durme. I'm a belieber: Social roles via self-identification and conceptual attributes. In *Association for Computational Linguistics (ACL), Short Papers*, 2014.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *The Journal of Machine Learning Research*, 3:1137–1155, 2003.

Shane Bergsma and Benjamin Van Durme. Using conceptual class attributes to characterize social media users. In *ACL*, pages 710–720, 2013.

Michael S. Bernstein, Eytan Bakshy, Moira Burke, and Brian Karrer. Quantifying the invisible audience in social networks. In *CHI '13*, 2013.

Fadi Biadsy, Julia Hirschberg, and Elena Filatova. An unsupervised approach to biography production using Wikipedia. In *Proceedings of ACL-08: HLT*, pages 807–815, Columbus, Ohio, June 2008. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P/P08/P08-1092`.

David M. Blei and John D. Lafferty. Correlated topic models. In *In Proceedings of the 23rd International Conference on Machine Learning*, pages 113–120. MIT Press, 2006a.

David M. Blei and John D. Lafferty. Dynamic topic models. In *ICML '06*, pages 113–120, 2006b. ISBN 1-59593-383-2. doi: 10.1145/1143844.1143859. URL `http://doi.acm.org/10.1145/1143844.1143859`.

David M. Blei and John D. Lafferty. A correlated topic model of science. *AAS*, 1(1):17–35, 2007.

David M. Blei, Andrew Ng, and Michael Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.

Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, COLT' 98, pages 92–100, New York, NY, USA, 1998. ACM. ISBN 1-58113-057-0. doi: 10.1145/279943.279962. URL `http://doi.acm.org/10.1145/279943.279962`.

Robert Bond and Solomon Messing. Quantifying social media's political space: Estimating ideology from publicly revealed preferences on Facebook. *American Political Science Review*, 109(01):62–78, 2015.

Wayne Booth. *The Rhetoric of Fiction*. University of Chicago Press, Chicago, 1961.

danah boyd. *Taken Out of Context: American Teen Sociality in Networked Publics*. PhD thesis, University of California-Berkeley, School of Information, 2008.

Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. Class-based n-gram models of natural language. *Comput. Linguist.*, 18(4):467–479, December 1992. ISSN 0891-2017. URL `http://dl.acm.org/citation.cfm?id=176313.176316`.

Elia Bruni, Giang Binh Tran, and Marco Baroni. Distributional semantics from text and images. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, GEMS '11, pages 22–32, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-937284-16-9. URL `http://dl.acm.org/citation.cfm?id=2140490.2140493`.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 136–145, Stroudsburg, PA, USA, 2012a. Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=2390524.2390544`.

Elia Bruni, Jasper Uijlings, Marco Baroni, and Nicu Sebe. Distributional semantics with eyes: Using image analysis to improve computational representations of word meaning. In *Proceedings of the 20th ACM International Conference on Multimedia*, MM '12, pages 1219–1228, New York, NY, USA, 2012b. ACM. ISBN 978-1-4503-1089-5. doi: 10.1145/2393347.2396422. URL `http://doi.acm.org/10.1145/2393347.2396422`.

Mary Bucholtz. Word up: Social meanings of slang in California youth culture. In Jane Goodman and Leila Monaghan, editors, *A Cultural Approach to Interpersonal Communication: Essential Readings*, Malden, MA, 2006. Blackwell.

Razvan Bunescu and Marius Pasca. Using encyclopedic knowledge for named entity disambiguation. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06)*, pages 9–16, Trento, Italy, 2006.

John D. Burger, John Henderson, George Kim, and Guido Zarrella. Discriminating gender on Twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1301–1309, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics.

Felix Burkhardt, Markus van Ballegooy, Roman Englert, and Richard Huber. An emotion-aware voice portal. *Proc. Electronic Speech Signal Processing ESSP*, pages 123–131, 2005.

Ewa S. Callahan and Susan C. Herring. Cultural bias in Wikipedia content on famous persons. *J. Am. Soc. Inf. Sci. Technol.*, 62(10):1899–1915, October 2011. ISSN 1532-2882. doi: 10.1002/asi.21577. URL `http://dx.doi.org/10.1002/asi.21577`.

Joseph Campbell. *The Hero with a Thousand Faces*. Pantheon Books, 1949.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr., and Tom M. Mitchell. Toward an architecture for never-ending language learning. In *Proceedings of the Twenty-Fourth Conference on Artificial Intelligence (AAAI 2010)*, 2010.

Bob Carpenter. Integrating Out Multinomial Parameters in Latent Dirichlet Allocation and Naive Bayes for Collapsed Gibbs Sampling. Technical report, LingPipe, 2010. URL `http://lingpipe.files.wordpress.com/2010/07/lda3.pdf`.

Paula Carvalho, Luís Sarmento, Mário J. Silva, and Eugénio de Oliveira. Clues for detecting irony in user-generated contents: Oh...!! it's "so easy" ;-). In *1st International CIKM Workshop on Topic-sentiment Analysis for Mass Opinion*, pages 53–56, 2009. ISBN 978-1-60558-805-6. doi: 10.1145/1651461.1651471. URL `http://doi.acm.org/10.1145/1651461.1651471`.

George Casella and Edward I. George. Explaining the Gibbs sampler. *The American Statistician*, 46(3): 167–174, 1992.

Justine Cassell. Editing wars behind the scenes. *New York Times*, February 4 2011.

Justine Cassell and Timothy Bickmore. Negotiated collusion: Modeling social language and its relationship effects in intelligent agents. *User Modeling and User-Adapted Interaction*, 13(1-2):89–132, 2003.

Nathanael Chambers. *Inducing Event Schemas and their Participants from Unlabeled Text*. PhD thesis, Stanford University, 2011.

Nathanael Chambers. Event schema induction with a probabilistic entity-driven model. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/D13-1185`.

Nathanael Chambers and Dan Jurafsky. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pages 789–797, Columbus, Ohio, June 2008. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P/P08/P08-1090`.

Nathanael Chambers and Dan Jurafsky. Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL '09, pages 602–610, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-46-6. URL `http://dl.acm.org/citation.cfm?id=1690219.1690231`.

Nathanael Chambers, Shan Wang, and Dan Jurafsky. Classifying temporal relations between events. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 173–176, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=1557769.1557820`.

J. Chang and D. Blei. Relational topic models for document networks. In *Artificial Intelligence and Statistics*, pages 81–88, 2009.

Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. Reading tea leaves: How humans interpret topic models. In *Neural Information Processing Systems (NIPS)*, 2009.

S.F. Chen and R. Rosenfeld. A survey of smoothing techniques for ME models. *Speech and Audio Processing, IEEE Transactions on*, 8(1):37–50, 2000. ISSN 1063-6676. doi: 10.1109/89.817452.

Stanley Chen, Douglas Beeferman, and Roni Rosenfeld. Evaluation metrics for language models, 1998.

Jackie Chi Kit Cheung, Hoifung Poon, and Lucy Vanderwende. Probabilistic frame induction. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 837–846, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/N13-1104`.

Noam Chomsky. *Aspects of the Theory of Syntax*. MIT Press, 1965.

Raviv Cohen and Derek Ruths. Classifying political orientation on Twitter: It's not easy! In *International AAAI Conference on Weblogs and Social Media*, 2013.

Benjamin Collier and Julia Bear. Conflict, criticism, or confidence: An empirical examination of the gender gap in Wikipedia contributions. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, pages 383–392, New York, NY, USA, 2012. ACM.

Michael Collins. Head-driven statistical models for natural language parsing. *Computational linguistics*, 29(4):589–637, 2003.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 12:2493–2537, November 2011. ISSN 1532-4435. URL `http://dl.acm.org/citation.cfm?id=1953048.2078186`.

M.D. Conover, B. Goncalves, J. Ratkiewicz, A Flammini, and F. Menczer. Predicting the political alignment of Twitter users. In *Proceedings of 3rd IEEE Conference on Social Computing SocialCom*, 2011.

Mike Conway. Mining a corpus of biographical texts using keywords. *Literary and Linguistic Computing*, 25(1):23–35, 2010. doi: 10.1093/llc/fqp035. URL `http://llc.oxfordjournals.org/content/25/1/23.abstract`.

Silviu Cucerzan. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 708–716, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

Dipanjan Das, Nathan Schneider, Desai Chen, and Noah A. Smith. Probabilistic frame-semantic parsing. In *Proceedings of the North American Chapter of the Association for Computational Linguistics Human Language Technologies Conference*, Los Angeles, CA, June 2010.

Dipanjan Das, Desai Chen, André F. T. Martins, Nathan Schneider, and Noah A. Smith. Frame-semantic parsing. *Computational Linguistics*, 40(1), March 2014. URL `http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00163`.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. Semi-supervised recognition of sarcastic sentences in Twitter and Amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, CoNLL '10, pages 107–116, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. ISBN 978-1-932432-83-1. URL `http://dl.acm.org/citation.cfm?id=1870568.1870582`.

Peter T. Davis, David K. Elson, and Judith L. Klavans. Methods for precise named entity matching in digital collections. In *Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries*, JCDL '03, pages 125–127, Washington, DC, USA, 2003. IEEE Computer Society. ISBN 0-7695-1939-3. URL `http://dl.acm.org/citation.cfm?id=827140.827158`.

Marie-Catherine de Marneffe and Christopher D. Manning. Stanford typed dependencies manual. Technical report, Stanford University, 2008.

Marie-Catherine de Marneffe, Christopher D. Manning, and Christopher Potts. Did it happen? the pragmatic complexity of veridicality assessment. *Comput. Linguist.*, 38(2):301–333, June 2012. ISSN 0891-2017. doi: 10.1162/COLI_a_00097. URL `http://dx.doi.org/10.1162/COLI_a_00097`.

Jacques Derrida. La pharmacie de Platon. In *La dissémination.* Éditions du Seuil, 1972.

Gabriel Doyle. Mapping dialectal variation by querying social media. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 98–106, Gothenburg, Sweden, April 2014. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/E14-1011.

Greg Durrett and Dan Klein. Easy victories and uphill battles in coreference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Seattle, Washington, October 2013. Association for Computational Linguistics.

Greg Durrett, David Hall, and Dan Klein. Decentralized entity-level modeling for coreference resolution. In *ACL*, pages 114–124, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

Penelope Eckert. Variation and the indexical field. *Journal of Sociolinguistics*, 12(4):453–476, 2008.

Bradley Efron. Bootstrap methods: another look at the jackknife. *The annals of Statistics*, pages 1–26, 1979.

Bradley Efron and Charles Stein. The jackknife estimate of variance. *The Annals of Statistics*, pages 586–596, 1981.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1277–1287, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL http://portal.acm.org/citation.cfm?id=1870658.1870782.

Jacob Eisenstein, Amr Ahmed, and Eric P. Xing. Sparse additive generative models of text. In *ICML*, 2011a.

Jacob Eisenstein, Noah A. Smith, and Eric P. Xing. Discovering sociolinguistic associations with structured sparsity. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1365–1374, Stroudsburg, PA, USA, 2011b. Association for Computational Linguistics. ISBN 978-1-932432-87-9. URL http://dl.acm.org/citation.cfm?id=2002472.2002641.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. Diffusion of lexical change in social media. *PLoS ONE*, 9(11):e113114, 11 2014. doi: 10.1371/journal.pone.0113114. URL http://dx.doi.org/10.1371%2Fjournal.pone.0113114.

Khalid El-Arini, Ulrich Paquet, Ralf Herbrich, Jurgen Van Gael, and Blaise Agüera y Arcas. Transparent user models for personalization. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '12, pages 678–686, New York, NY, USA, 2012. ACM.

Khalid El-Arini, Min Xu, Emily B. Fox, and Carlos Guestrin. Representing documents through their readers. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 14–22, New York, NY, USA, 2013. ACM.

Glen Elder. *Children of the Great Depression*. University of Chicago Press, 1974.

Glen Elder. Talent, history, and the fulfillment of promise. *Psychiatry*, 54(3):251–267, 1991.

Micha Elsner. Character-based kernels for novelistic plot structure. In *Proceedings of the 13th Conference of the EACL*, 2012.

David K. Elson, Nicholas Dames, and Kathleen R. McKeown. Extracting social networks from literary fiction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 138–147, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=1858681.1858696`.

Robert M. Entman. Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4):51–58, 1993.

Anthony Fader, Stephen Soderland, and Oren Etzioni. Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1535–1545, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

Yansong Feng and Mirella Lapata. Visual information in semantic representation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 91–99, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. ISBN 1-932432-65-5. URL `http://dl.acm.org/citation.cfm?id=1857999.1858010`.

E. C. Fieller, H. O. Hartley, and E. S Pearson. Tests for rank correlation coefficients. *Biometrika*, 1957.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 363–370, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics. doi: 10.3115/1219840.1219885. URL `http://dx.doi.org/10.3115/1219840.1219885`.

Samantha Finkelstein, Amy Ogan, Caroline Vaughn, and Justine Cassell. Alex: A virtual peer that identifies student dialect. In *Proc. Workshop on Culturally-aware Technology Enhanced Learning in conjuction with EC-TEL 2013, Paphos, Cyprus, September 17*, 2013.

Mark Alan Finlayson. *Learning Narrative Structure from Annotated Folktales*. PhD thesis, MIT, 2011.

R. A. Fisher. *The Design of Experiments*. Oliver and Boyde, Edinburgh and London, 1935.

Jeffrey Flanigan, Sam Thomson, Jaime Carbonell, Chris Dyer, and Noah A Smith. A discriminative graph-based parser for the abstract meaning representation. In *ACL*. Citeseer, 2014.

Michael Fleischman and Eduard Hovy. Fine grained classification of named entities. In *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1*, COLING '02, pages 1–7, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

E. M. Forster. *Aspects of the Novel*. Harcourt, Brace & Co., 1927.

Harold N. Fowler. *Plato in Twelve Volumes*. Harvard University Press, Cambridge, 1925.

Northrup Frye. *Anatomy of Criticism*. Princeton University Press, 1957.

William A Gale, Kenneth W Church, and David Yarowsky. One sense per discourse. In *Proceedings of the workshop on Speech and Natural Language*, pages 233–237. Association for Computational Linguistics, 1992.

Spandana Gella, Paul Cook, and Timothy Baldwin. One sense per tweeter... and other lexical semantic tales of twitter. *EACL 2014*, page 215, 2014.

Stuart Geman and Donald Geman. Stochastic relaxation, Gibbs distributions, and the bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 6(6):721–741, November 1984. ISSN 0162-8828. doi: 10.1109/TPAMI.1984.4767596. URL `http://dx.doi.org/10.1109/TPAMI.1984.4767596`.

Sean Gerrish and David M. Blei. How they vote: Issue-adjusted models of legislative behavior. In F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 2753–2761. Curran Associates, Inc., 2012. URL `http://papers.nips.cc/paper/4715-how-they-vote-issue-adjusted-models-of-legislative-behavior.pdf`.

Daniel Gildea and Daniel Jurafsky. Automatic labeling of semantic roles. *Computational Linguistics*, 24(3):245–288, 2002.

Erving Goffman. *The Presentation of the Self in Everyday Life*. Anchor, 1959.

Jennifer Golbeck, Cristina Robles, and Karen Turner. Predicting personality with social media. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '11, pages 253–262, New York, NY, USA, 2011. ACM.

Andrew Goldstone and Ted Underwood. The quiet transformations of literary studies: What thirteen thousand scholars could tell us. *New Literary History*, 2014.

Roberto González-Ibáñez, Smaranda Muresan, and Nina Wacholder. Identifying sarcasm in Twitter: A closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational*

Linguistics: Human Language Technologies: Short Papers - Volume 2, HLT '11, pages 581–586, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-932432-88-6. URL `http://dl.acm.org/citation.cfm?id=2002736.2002850`.

Joshua Goodman. Classes for fast maximum entropy training. In *Proc. of ICASSP*, 2001.

Joshua Goodman. Exponential priors for maximum entropy models. In Daniel Marcu Susan Dumais and Salim Roukos, editors, *HLT-NAACL 2004: Main Proceedings*, pages 305–312, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for Computational Linguistics.

Google. Freebase data dumps. `https://developers.google.com/freebase/data`, 2014.

Amit Goyal, Ellen Riloff, and Hal Daumé, III. Automatically producing plot unit representations for narrative text. In *Proceedings of the 2010 Conference on EMNLP*, 2010.

Algirdas Julien Greimas. *Structural Semantics: An Attempt at a Method*. University of Nebraska Press, Lincoln, 1984.

Trond Grenager and Christopher D. Manning. Unsupervised discovery of a statistical verb lexicon. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 1–8, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics. ISBN 1-932432-73-6. URL `http://dl.acm.org/citation.cfm?id=1610075.1610077`.

Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5228–5235, 2004.

Thomas L. Griffiths, Mark Steyvers, David M. Blei, and Joshua B. Tenenbaum. Integrating topics and syntax. In L.K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 537–544. MIT Press, 2005. URL `http://papers.nips.cc/paper/2587-integrating-topics-and-syntax.pdf`.

Justin Grimmer. A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis*, 18(1):1–35, 2010. doi: 10.1093/pan/mpp034. URL `http://pan.oxfordjournals.org/content/18/1/1.abstract`.

Amit Gruber, Michal Rosen-zvi, and Yair Weiss. Hidden topic markov models. In *In Proceedings of Artificial Intelligence and Statistics*, 2007.

Swati Gupta, Marilyn A. Walker, and Daniela M. Romano. Generating politeness in task based interaction: An evaluation of the effect of linguistic form and culture. In *Proceedings of the Eleventh European Workshop on Natural Language Generation*, ENLG '07, pages 57–64, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=1610163.1610173`.

Aria Haghighi and Dan Klein. Simple coreference resolution with rich syntactic and semantic features. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1152–1161, Singapore, August 2009.

Aria Haghighi and Dan Klein. Coreference resolution in a modular, entity-centered model. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 385–393, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

Nathan Halko, Per-Gunnar Martinsson, Yoel Shkolnisky, and Mark Tygert. An algorithm for the principal component analysis of large data sets. *SIAM Journal on Scientific Computing*, 33(5):2580–2594, 2011.

Eszter Hargittai. Why doesn't Science publish important methods info prominently? http://crookedtimber.org/2015/05/07/why-doesnt-science-publish-important-methods-info-prominently/, May 2015.

Zellig Harris. Distributional structure. *Word*, 10(23):146–162, 1954.

W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):pp. 97–109, 1970.

Brent Hecht and Darren Gergle. Measuring self-focus bias in community-maintained knowledge repositories. In *Proceedings of the Fourth International Conference on Communities and Technologies*, pages 11–20, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-713-4. doi: 10.1145/1556460.1556463. URL `http://doi.acm.org/10.1145/1556460.1556463`.

Susan C. Herring and John C. Paolillo. Gender and genre variation in weblogs. *Journal of Sociolinguistics*, 10(4):439–459, 2006.

Benjamin Mako Hill and Aaron Shaw. The Wikipedia gender gap revisited: Characterizing survey response bias with propensity score estimation. *PLoS ONE*, 8(6), 2013.

Jerry R Hobbs, Douglas Appelt, John Bear, David Israel, Megumi Kameyama, and Mabry Tyson. Fastus: A system for extracting information from text. In *Proceedings of the workshop on Human Language Technology*, pages 133–137. Association for Computational Linguistics, 1993.

Dennis Hogan. *Transitions and Social Change: The Early Lives of American Men*. Academic, New York, 1981.

Dennis P. Hogan and Nan Marie Astone. The transition to adulthood. *Annual Review of Sociology*, 12(1):109–130, 1986. doi: 10.1146/annurev.so.12.080186.000545. URL `http://www.annualreviews.org/doi/abs/10.1146/annurev.so.12.080186.000545`.

Liangjie Hong, Amr Ahmed, Siva Gurumurthy, Alexander J. Smola, and Kostas Tsioutsiouliklis. Discovering geographical topics in the Twitter stream. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 769–778, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1229-5. doi: 10.1145/2187836.2187940. URL `http://doi.acm.org/10.1145/2187836.2187940`.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. OntoNotes: the 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, NAACL-Short '06, pages 57–60, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

Yuening Hu, Jordan Boyd-Graber, Brianna Satinoff, and Alison Smith. Interactive topic modeling. *Machine learning*, 95(3):423–469, 2014.

Yohei Ikawa, Miki Enoki, and Michiaki Tatsubori. Location inference using microblog messages. In *Proceedings of the 21st International Conference Companion on World Wide Web*, WWW '12 Companion, pages 687–690, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1230-1. doi: 10.1145/2187980.2188181. URL `http://doi.acm.org/10.1145/2187980.2188181`.

R. Iyer, M. Ostendorf, and M. Meteer. Analyzing and predicting language model improvements. In *Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on*, pages 254–261, Dec 1997. doi: 10.1109/ASRU.1997.659013.

Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. Political ideology detection using recursive neural networks. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1113–1122, Baltimore, Maryland, June 2014. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P14-1105`.

Yangfeng Ji and Jacob Eisenstein. Representation learning for text-level discourse parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 13–24, 2014.

Maojin Jiang and Shlomo Argamon. Exploiting subjectivity analysis in blogs to improve political leaning categorization. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 725–726. ACM, 2008.

Matthew L. Jockers and David Mimno. Significant themes in 19th-century literature. *Poetics*, 41 (6):750 – 769, 2013. ISSN 0304-422X. doi: http://dx.doi.org/10.1016/j.poetic.2013.08.005. URL `http://www.sciencedirect.com/science/article/pii/S0304422X13000673`. Topic Models and the Cultural Sciences.

Barbara Johnstone and Scott F. Kiesling. Indexicality and experience: Exploring the meanings of /aw/-monophthongization in Pittsburgh. *Journal of Sociolinguistics*, 12(1):5–33, 2008.

Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. *Mach. Learn.*, 37(2):183–233, November 1999. ISSN 0885-6125. doi: 10.1023/A:1007665907178. URL `http://dx.doi.org/10.1023/A:1007665907178`.

Carl Jung. *The Archetypes and The Collective Unconscious*, volume 9 of *Collected Works*. Bollingen, Princeton, NJ, 2nd edition, 1981.

Dan Jurafsky and James H Martin. *Speech and language processing*. Pearson, 2009.

John S Justeson and Slava M Katz. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural language engineering*, 1(1):9–27, 1995.

Leslie Kaufman. Independent Journal Review website becomes a draw for conservatives. *New York Times*, Nov. 2, 2014 2014.

Jun'ichi Kazama and Jun'ichi Tsujii. Evaluation and extension of maximum entropy models with inequality constraints. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, EMNLP '03, pages 137–144, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics. doi: 10.3115/1119355.1119373. URL `http://dx.doi.org/10.3115/1119355.1119373`.

Suzanne Keen. *Narrative Form.* Palgrave Macmillan, Basingstoke., 2003.

Josef Kolbitsch and Hermann A. Maurer. The transformation of the web: How emerging communities shape the information we consume. *J. UCS*, 12(2):187–213, 2006. URL `http://dblp.uni-trier.de/db/journals/jucs/jucs12.html#KolbitschM06`.

Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

Lingpeng Kong, Nathan Schneider, Swabha Swayamdipta, Archna Bhatia, Chris Dyer, and Noah A. Smith. A dependency parser for tweets. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1001–1012, Doha, Qatar, October 2014. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/D14-1108`.

Moshe Koppel, J. Schler, Shlomo Argamon, and J. Pennebaker. Effects of age and gender on blogging. In *In AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs*, 2006.

Roger J. Kreuz. The use of verbal irony: Cues and constraints. In Jeffrey Scott Mio and Albert N. Katz, editors, *Metaphor: Implications and Applications*, pages 23–38, 1996.

Roger J. Kreuz and Gina M. Caucci. Lexical influences on the perception of sarcasm. In *Proceedings of the Workshop on Computational Approaches to Figurative Language*, FigLanguages '07, pages 1–4, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=1611528.1611529`.

Roger J Kreuz and Richard M Roberts. Two cues for verbal irony: Hyperbole and the ironic tone of voice. *Metaphor and Symbol*, 10(1):21–31, 1995.

Lee A. Lillard and Linda J. Waite. A joint model of marital childbearing and marital disruption. *Demography*, 30(4):pp. 653–681, 1993. ISSN 00703370. URL `http://www.jstor.org/stable/2061812`.

Lee A. Lillard, Michael J. Brien, and Linda J. Waite. Premarital cohabitation and subsequent marital dissolution: A matter of self-selection? *Demography*, 32(3):pp. 437–457, 1995. ISSN 00703370. URL `http://www.jstor.org/stable/2061690`.

Dekang Lin. Automatic retrieval and clustering of similar words. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 2*, ACL '98, pages 768–774, Stroudsburg, PA, USA, 1998. Association for Computational Linguistics. doi: 10.3115/980691.980696. URL `http://dx.doi.org/10.3115/980691.980696`.

Stephanie Lukin and Marilyn Walker. Really? well. apparently bootstrapping improves the performance of sarcasm and nastiness classifiers for online dialogue. In *Proceedings of the Workshop on Language Analysis in Social Media*, 2013. URL `http://aclweb.org/anthology/W13-1104`.

Xiaoqiang Luo. On coreference resolution performance metrics. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 25–32. Association for Computational Linguistics, 2005.

Inderjeet Mani. Computational modeling of narrative. *Synthesis Lectures on Human Language Technologies*, 5(3):1–142, 2012.

Inderjeet Mani, Marc Verhagen, Ben Wellner, Chong Min Lee, and James Pustejovsky. Machine learning of temporal relations. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, pages 753–760, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics. doi: 10.3115/1220175.1220270. URL `http://dx.doi.org/10.3115/1220175.1220270`.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, 2014. URL `http://www.aclweb.org/anthology/P/P14/P14-5010`.

Daniel Marcu. The rhetorical parsing of unrestricted texts: A surface-based approach. *Computational linguistics*, 26(3):395–448, 2000.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993.

Andrew D. Martin and Kevin M. Quinn. Dynamic ideal point estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953–1999. *Political Analysis*, 10(2):134–153, 2002. doi: 10.1093/pan/10.2.134. URL `http://pan.oxfordjournals.org/content/10/2/134.abstract`.

Alice E Marwick and danah boyd. I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New media & Society*, 13(1):114–133, 2011.

Sébastien Massoni, Madalina Olteanu, and Patrick Rousset. Career-path analysis using optimal matching and self-organizing maps. In *Advances in Self-Organizing Maps: 7th International Workshop, WSOM 2009, St. Augustine, FL, USA, June 8-10, 2009*, volume 5629 of *Lecture Notes in Computer Science*, pages 154–162. Springer, Berlin, 2009. doi: http://dx.doi.org/10.1007/978-3-642-02397-2_18.

Henry Colin Gray Matthew and Brian Harrison. *The Oxford dictionary of national biography*. Oxford University Press, 2004.

Karl Ulrich Mayer. German survivors of World War II: The impact on the life course of the collective experience of birth cohorts. In *Social Structure and Human Lives*, Newbury Park, 1988. Sage.

Jon D. Mcauliffe and David M. Blei. Supervised topic models. In J.C. Platt, D. Koller, Y. Singer, and S.T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 121–128. Curran Associates, Inc., 2008. URL `http://papers.nips.cc/paper/3328-supervised-topic-models.pdf`.

David McClosky, Eugene Charniak, and Mark Johnson. Effective self-training for parsing. In *Proceedings of the main conference on human language technology conference of the North American Chapter of the Association of Computational Linguistics*, pages 152–159. Association for Computational Linguistics, 2006.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajič. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 523–530. Association for Computational Linguistics, 2005.

Neil McIntyre and Mirella Lapata. Plot induction and evolutionary search for story generation. In *Proceedings of the 48th Annual Meeting of the ACL*. Association for Computational Linguistics, 2010.

Marina Meilă. Comparing clusterings—an information based distance. *Journal of Multivariate Analysis*, 98(5):873–895, 2007.

Igor Mel'čuk. *Dependency Syntax: Theory and Practice*. University of New York Press, Albany, 1988.

N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. Equations of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087–1091, 1953.

Rada Mihalcea and Carlo Strapparava. Making computers laugh: Investigations in automatic humor recognition. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 531–538. Association for Computational Linguistics, 2005.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *ICLR*, 2013.

David Mimno and Andrew Mccallum. Modeling career path trajectories. Technical Report 2007-69, University of Massachusetts, Amherst, 2007.

David Mimno and Andrew McCallum. Organizing the OCA: Learning faceted subjects from a library of digital books. In *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '07, pages 376–385, New York, NY, USA, 2007. ACM.

David Mimno and Andrew McCallum. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. In *UAI*, 2008. URL `http://www.cs.umass.edu/~mimno/papers/dmr-uai.pdf`.

David Mimno, Hanna M. Wallach, and Andrew Mccallum. Gibbs sampling for logistic normal topic models with graph-based priors. In *NIPS Workshop on Analyzing Graphs*, 2008.

Thomas P. Minka. Estimating a dirichlet distribution. `http://research.microsoft.com/en-us/um/people/minka/papers/dirichlet/`, 2003.

Marvin Minsky. A framework for representing knowledge. Technical report, MIT-AI Laboratory Memo 306, 1974.

Amy Mitchell, Jeffrey Gottfried, Jocelyn Kiley, and Katerina Eva Matsa. Political polarization and media habits: From Fox News to Facebook, how liberals and conservatives keep up with politics. Technical report, Pew Research Center, 2014.

T. Mitchell, W. Cohen, E. Hruschka, P. Talukdar, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K. Mazaitis, T. Mohamed, N. Nakashole, E. Platanios, A. Ritter, M. Samadi, B. Settles, R. Wang, D. Wijaya, A. Gupta, X. Chen, A. Saparov, M. Greaves, and J. Welling. Never-ending learning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15)*, 2015.

Andriy Mnih and Geoffrey E Hinton. A scalable hierarchical distributed language model. In *Advances in neural information processing systems*, pages 1081–1088, 2009.

John Modell. Normative aspects of american marriage timing since World War II. *Journal of Family History*, 5(2):210–234, 1980. doi: 10.1177/036319908000500206. URL `http://jfh.sagepub.com/content/5/2/210.short`.

John Modell, Frank F. Furstenberg Jr., and Theodore Hershberg. Social change and transitions to adulthood in historical perspective. *Journal of Family History*, 1(1):7–32, 1976.

Ashutosh Modi, Ivan Titov, and Alexandre Klementiev. Unsupervised induction of frame-semantic representations. In *Proceedings of the NAACL-HLT Workshop on the Induction of Linguistic Structure*, WILS '12, pages 1–7, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=2390426.2390428.

Frederic Morin and Yoshua Bengio. Hierarchical probabilistic neural network language model. In Robert G. Cowell and Zoubin Ghahramani, editors, *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, pages 246–252. Society for Artificial Intelligence and Statistics, 2005. URL http://www.iro.umontreal.ca/~lisa/pointeurs/hierarchical-nnlm-aistats05.pdf.

F. Mosteller and D. Wallace. *Inference and Disputed Authorship: The Federalist*. Addison-Wesley, 1964.

Frederick Mosteller and David L Wallace. Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed federalist papers. *Journal of the American Statistical Association*, 58(302):275–309, 1963.

Arjun Mukherjee and Bing Liu. Improving gender classification of blog authors. In *Proceedings of the 2010 conference on Empirical Methods in natural Language Processing*, pages 207–217. Association for Computational Linguistics, 2010.

Ndapandula Nakashole and Tom M Mitchell. Language-aware truth assessment of fact candidates. In *ACL*, 2014.

Ramanathan Narayanan, Bing Liu, and Alok Choudhary. Sentiment analysis of conditional sentences. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 180–189, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL http://dl.acm.org/citation.cfm?id=1699510.1699534.

Radford M Neal. Slice sampling. *Annals of Statistics*, pages 705–741, 2003.

Pauline Ng. What Kobe Bryant and Britney Spears have in common: Mining Wikipedia for characteristics of notable individuals. In *ICWSM*, 2012. URL https://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/view/4556.

Viet-An Nguyen, Jordan Boyd-Graber, Philip Resnik, and Kristina Miler. Tea party in the house: A hierarchical ideal point topic model and its application to Republican legislators in the 112th congress. In *ACL*, 2015.

Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13:95–135, 5 2007a. ISSN 1469-8110. doi: 10.1017/S1351324906004505. URL http://journals.cambridge.org/article_S1351324906004505.

Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(02):95–135, 2007b.

NPR. When politicians lose their accents. http://www.npr.org/2015/04/18/400658679/when-politicians-lose-their-accents, 2015.

Brendan O'Connor. Learning frames from text with an unsupervised latent variable model. *ArXiv*, abs/1307.7382, 2013.

Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *ICWSM*, 2010.

Brendan O'Connor, Jacob Eisenstein, Eric P. Xing, and Noah A. Smith. Discovering demographic language variation. In *NIPS-2010 Workshop on Machine Learning and Social Computing*, 2010.

Sharon Oviatt. Multimodal interfaces. In Julie A. Jacko and Andrew Sears, editors, *The Human-computer Interaction Handbook*, pages 286–304, Hillsdale, NJ, USA, 2003. L. Erlbaum Associates Inc. ISBN 0-8058-3838-4. URL `http://dl.acm.org/citation.cfm?id=772072.772093`.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of NAACL*, 2013.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–105, 2005.

Souneil Park, Minsam Ko, Jungwoo Kim, Ying Liu, and Junehwa Song. The politics of comments: Predicting political orientation of news stories with commenters' sentiment patterns. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work*, CSCW '11, pages 113–122, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0556-3. doi: 10.1145/1958824.1958842. URL `http://doi.acm.org/10.1145/1958824.1958842`.

Michael J Paul and Mark Dredze. You are what you tweet: Analyzing Twitter for public health. In *ICWSM*, pages 265–272, 2011.

Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, 1988.

Marco Pennacchiotti and Ana-Maria Popescu. Democrats, Republicans and Starbucks afficionados: User classification in Twitter. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 430–438, New York, NY, USA, 2011. ACM.

Fernando Pereira, Naftali Tishby, and Lillian Lee. Distributional clustering of english words. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, ACL '93, pages

183–190, Stroudsburg, PA, USA, 1993. Association for Computational Linguistics. doi: 10.3115/981574.981598. URL `http://dx.doi.org/10.3115/981574.981598`.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 433–440. Association for Computational Linguistics, 2006.

Sasa Petrovic and David Matthews. Unsupervised joke generation from big data. In *ACL (2)*, pages 228–232. Citeseer, 2013.

Ulrike Pfeil, Panayiotis Zaphiris, and Chee Siang Ang. Cultural differences in collaborative authoring of wikipedia. *Journal of Computer-Mediated Communication*, 12(1):88–113, 2006. ISSN 1083-6101. doi: 10.1111/j.1083-6101.2006.00316.x. URL `http://dx.doi.org/10.1111/j.1083-6101.2006.00316.x`.

E. J. G. Pitman. Significance tests which may be applied to samples from any population. *Supplement to the Journal of the Royal Statistical Society*, 4(1):119–130, 1937.

Keith T. Poole and Howard Rosenthal. A spatial model for legislative roll call analysis. *American Journal of Political Science*, 29(2):pp. 357–384, 1985. ISSN 00925853. URL `http://www.jstor.org/stable/2111172`.

Vladimir Propp. *Morphology of the Folktale*. University of Texas Press, Austin, 2nd edition, 1968.

James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, et al. The timebank corpus. In *Corpus linguistics*, volume 2003, page 40, 2003.

Maurice H Quenouille. Notes on bias in estimation. *Biometrika*, pages 353–360, 1956.

Maxim Rabinovich and David Blei. The inverse regression topic model. In Tony Jebara and Eric P. Xing, editors, *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 199–207. JMLR Workshop and Conference Proceedings, 2014. URL `http://jmlr.org/proceedings/papers/v32/rabinovich14.pdf`.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 492–501, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

Altaf Rahman and Vincent Ng. Supervised models for coreference resolution. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, EMNLP '09, pages 968–977, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-62-6. URL `http://dl.acm.org/citation.cfm?id=1699571.1699639`.

Ashwin Rajadesingan, Reza Zafarani, and Huan Liu. Sarcasm detection on Twitter: A behavioral modeling approach. In *WSDM*, 2015.

Daniel Ramage, Susan Dumais, and Dan Liebling. Characterizing microblogs with topic models. In *ICWSM*, 2010.

Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. Classifying latent user attributes in Twitter. In *Proceedings of the 2nd International Workshop on Search and Mining User-generated Contents*, SMUC '10, pages 37–44, New York, NY, USA, 2010. ACM.

Adwait Ratnaparkhi et al. A maximum entropy model for part-of-speech tagging. In *Proceedings of the conference on empirical methods in natural language processing*, volume 1, pages 133–142. Philadelphia, PA, 1996.

Maya Ravindranath. A wicked good reason to study intensifiers in New Hampshire. In *NWAV 40*, 2011.

Joseph Reagle and Lauren Rhue. Gender bias in Wikipedia and Britannica. *International Journal of Communication*, 5, 2011.

Michaela Regneri, Alexander Koller, and Manfred Pinkal. Learning script knowledge with web experiments. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 979–988, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=1858681.1858781`.

Michaela Regneri, Alexander Koller, Josef Ruppenhofer, and Manfred Pinkal. Learning script participants from unlabeled data. In *Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP)*, Hissar, 2011.

Steffen Reinhold. Reassessing the link between premarital cohabitation and marital instability. *Demography*, 47(3):719–733, 2010. ISSN 0070-3370.

Joseph Reisinger and Raymond J. Mooney. Multi-prototype vector-space models of word meaning. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 109–117, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. ISBN 1-932432-65-5. URL `http://dl.acm.org/citation.cfm?id=1857999.1858012`.

Philip Resnik and Eric Hardisty. Gibbs sampling for the uninitiated. Technical report, DTIC Document, 2010.

Antonio Reyes, Paolo Rosso, and Tony Veale. A multidimensional approach for detecting irony in Twitter. *Lang. Resour. Eval.*, 47(1):239–268, March 2013. ISSN 1574-020X. doi: 10.1007/s10579-012-9196-x. URL `http://dx.doi.org/10.1007/s10579-012-9196-x`.

Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 704–714. Association for Computational Linguistics, 2013. URL `http://aclweb.org/anthology/D13-1066`.

Gilbert Ritschard, Reto Bürgin, and Matthias Studer. Exploratory mining of life event histories. In J. J. McArdle and G. Ritschard, editors, *Contemporary Issues in Exploratory Data Mining in the Behavioral Sciences*, Quantitative Methodology, pages 221–253. Routledge, New York, 2013.

Alan Ritter, Mausam, Oren Etzioni, and Sam Clark. Open domain event extraction from Twitter. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 1104–1112, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1462-6. doi: 10.1145/2339530.2339704. URL `http://doi.acm.org/10.1145/2339530.2339704`.

Margaret E Roberts, Brandon M Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G Rand. Structural topic models for open-ended survey responses. *American Journal of Political Science*, 58(4):1064–1082, 2014.

Stephen Roller and Sabine Schulte im Walde. A multimodal LDA model integrating textual, cognitive and visual modalities. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*, pages 1146–1157, Seattle, WA, October 2013. URL `http://www.cs.utexas.edu/users/ai-lab/pub-view.php?PubID=127403`.

Stephen Roller, Michael Speriosu, Sarat Rallapalli, Benjamin Wing, and Jason Baldridge. Supervised text-based geolocation using language models on an adaptive grid. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 1500–1510, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=2390948.2391120`.

Roni Rosenfeld. A maximum entropy approach to adaptive statistical language modelling. *Computer Speech and Language*, 10(3):187–228, 1996.

Sara Rosenthal and Kathleen McKeown. Age prediction in blogs: A study of style, content, and online behavior in pre-and post-social media generations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 763–772. Association for Computational Linguistics, 2011.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. Multiword expressions: A pain in the neck for NLP. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 2276 of *Lecture Notes in Computer Science*, pages 1–15. Springer Berlin Heidelberg, 2002. ISBN 978-3-540-43219-7. doi: 10.1007/3-540-45715-1_1. URL `http://dx.doi.org/10.1007/3-540-45715-1_1`.

Kenji Sagae. Analysis of discourse structure with syntactic dependencies and data-driven shift-reduce parsing. In *Proceedings of the 11th International Conference on Parsing Technologies*, pages 81–84. Association for Computational Linguistics, 2009.

Christian Sandvig. The Facebook "it's not our fault" study. http://blogs.law.harvard.edu/niftyc/archives/1062, May 2015.

Roger C. Schank and Robert P. Abelson. *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Lawrence Erlbaum, Hillsdale, NJ, 1977.

Petr Sgall, Eva Hajičová, and Jarmila Panevová. *The Meaning of the Sentence in its Semantic and Pragmatic Aspects*. Dordrecht: Reidel Publishing Company and Prague: Academia, 1986.

Michael J. Shanahan. Pathways to adulthood in changing societies: Variability and mechanisms in life course perspective. *Annual Review of Sociology*, 26(1):667–692, 2000. doi: 10.1146/annurev.soc.26.1.667. URL `http://www.annualreviews.org/doi/abs/10.1146/annurev.soc.26.1.667`.

Yanchuan Sim, Brice D. L. Acree, Justin H. Gross, and Noah A. Smith. Measuring ideological proportions in political speeches. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 91–101, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/D13-1010`.

Richard Socher, Jeffrey Pennington, Eric H. Huang, Andrew Y. Ng, and Christopher D. Manning. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 151–161, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-937284-11-4. URL `http://dl.acm.org/citation.cfm?id=2145432.2145450`.

Richard Socher, John Bauer, Christopher D. Manning, and Ng Andrew Y. Parsing with compositional vector grammars. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 455–465, Sofia, Bulgaria, August 2013a. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P13-1045`.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*, 2013b.

Anders Søgaard. Semisupervised condensed nearest neighbor for part-of-speech tagging. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 48–52. Association for Computational Linguistics, 2011.

Atara Stein. *The Byronic Hero in Film, Fiction and Television*. Southern Illinois University, 2004.

Matt Taddy. Multinomial inverse regression for text analysis. *Journal of the American Statistical Association*, 108(503):755–770, 2013.

Partha Pratim Talukdar, Derry Wijaya, and Tom Mitchell. Acquiring temporal constraints between relations. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 992–1001, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1156-4. doi: 10.1145/2396761.2396886. URL http://doi.acm.org/10.1145/2396761.2396886.

Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558, 2010. ISSN 1532-2890. doi: 10.1002/asi.21416. URL http://dx.doi.org/10.1002/asi.21416.

Ivan Titov and Alexandre Klementiev. A Bayesian approach to unsupervised semantic role induction. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 12–22, Avignon, France, April 2012. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/E12-1003.

Erik F Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 142–147. Association for Computational Linguistics, 2003.

Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180. Association for Computational Linguistics, 2003.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 384–394, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1858681.1858721.

Peter D. Turney and Patrick Pantel. From frequency to meaning: Vector space models of semantics. *J. Artif. Int. Res.*, 37(1):141–188, January 2010. ISSN 1076-9757. URL http://dl.acm.org/citation.cfm?id=1861751.1861756.

Ted Underwood, Michael L Black, Loretta Auvil, and Boris Capitanu. Mapping mutable genres in structurally complex volumes. In *Big Data, 2013 IEEE International Conference on*, pages 95–103. IEEE, 2013.

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. Semeval-2007 task 15: Tempeval temporal relation identification. In *Proceedings of the 4th*

International Workshop on Semantic Evaluations, pages 75–80. Association for Computational Linguistics, 2007.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. A model-theoretic coreference scoring scheme. In *Proceedings of the 6th Conference on Message Understanding*, MUC6 '95, pages 45–52, Stroudsburg, PA, USA, 1995. Association for Computational Linguistics.

Svitlana Volkova, Glen Coppersmith, and Benjamin Van Durme. Inferring user political preferences from streaming communications. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 186–196, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

C. Byron Wallace, Kook Do Choe, Laura Kertz, and Eugene Charniak. Humans require context to infer ironic intent (so computers probably do, too). In *ACL*, pages 512–516, 2014. URL `http://aclweb.org/anthology/P14-2084`.

Chong Wang, David M. Blei, and David Heckerman. Continuous time dynamic topic models. *ArXiv*, 2012.

Xuerui Wang and Andrew McCallum. Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '06, pages 424–433, New York, NY, USA, 2006. ACM. ISBN 1-59593-339-5. doi: 10.1145/1150402.1150450. URL `http://doi.acm.org/10.1145/1150402.1150450`.

AmyBeth Warriner, Victor Kuperman, and Marc Brysbaert. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, 45(4):1191–1207, 2013. doi: 10.3758/s13428-012-0314-x. URL `http://dx.doi.org/10.3758/s13428-012-0314-x`.

Greg C. G. Wei and Martin A. Tanner. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 1990.

Janyce Wiebe, Eric Breck, Chris Buckley, Claire Cardie, Paul Davis, Bruce Fraser, Diane J Litman, David R Pierce, Ellen Riloff, Theresa Wilson, et al. Recognizing and organizing opinions expressed in the world press. In *Proceedings of the 2003 AAAI Spring Symposium on New Directions in Question Answering*, pages 12–19, 2003.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3):165–210, 2005. ISSN 1574-020X. doi: 10.1007/s10579-005-7880-9. URL `http://dx.doi.org/10.1007/s10579-005-7880-9`.

Wikipedia. Wikipedia editors study: Results from the editor survey, April 2011.

Theresa Ann Wilson. *Fine-grained subjectivity and sentiment analysis: recognizing the intensity, polarity, and attitudes of private states*. PhD thesis, University of Pittsburgh, 2008.

Benjamin P. Wing and Jason Baldridge. Simple supervised document geolocation with geodesic grids. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 955–964, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-932432-87-9. URL `http://dl.acm.org/citation.cfm?id=2002472.2002593`.

Alex Woloch. *The One vs. the Many: Minor Characters and the Space of the Protagonist in the Novel*. Princeton University Press, Princeton NJ, 2003.

Limin Yao, Aria Haghighi, Sebastian Riedel, and Andrew McCallum. Structured relation discovery using generative models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1456–1466, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

Chen Yu and Dana H. Ballard. A multimodal learning interface for grounding spoken language in sensory perceptions. *ACM Trans. Appl. Percept.*, 1(1):57–80, July 2004. ISSN 1544-3558. doi: 10.1145/1008722.1008727. URL `http://doi.acm.org/10.1145/1008722.1008727`.

John M. Zelle and Raymond J. Mooney. Learning to parse database queries using inductive logic programming. In *Proceedings of the 13th National Conference on Artificial Intelligence*, pages 1050–1055, Portland, Oregon, USA, August 1996.

Luke S. Zettlemoyer and Michael Collins. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. In *Proceedings of the 21st Conference in Uncertainty in Artificial Intelligence*, pages 658–666, Edinburgh, UK, July 2005.

Daniel Xiaodan Zhou, Paul Resnick, and Qiaozhu Mei. Classifying the political leaning of news articles and users from user votes. In *ICWSM*, 2011.