# Unsupervised Discovery of Biographical Structure from Text

**David Bamman**
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA
dbamman@cs.cmu.edu

**Noah A. Smith**
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA
nasmith@cs.cmu.edu



## Abstract

We present a method for discovering abstract event classes in biographies, based on a probabilistic latent-variable model. Taking as input timestamped text, we exploit latent correlations among events to learn a set of event classes (such as BORN, GRADUATES HIGH SCHOOL, and BECOMES CITIZEN), along with the typical times in a person's life when those events occur. In a quantitative evaluation at the task of predicting a person's age for a given event, we find that our generative model outperforms a strong linear regression baseline, along with simpler variants of the model that ablate some features. The abstract event classes that we learn allow us to perform a large-scale analysis of 242,970 Wikipedia biographies. 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)—we 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 Introduction

The written text that we interact with on an everyday basis—news articles, emails, social media, books—contains a vast amount of information centered on people: news (including common NLP corpora such as the *New York Times* and 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. This focus on people gives us an abundance of information on how the lives of those portrayed unfold; 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. The life events described in these texts 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



Transactions of the Association for Computational Linguistics, 2 (2014) 363–376. Action Editor: Jian Su.
Submitted 6/2014; Published 10/2014. 

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 paper 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. Data and code to support this work can be found at `http://www.ark.cs.cmu.edu/bio/`.

## 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 (Bamman et al., 2014), 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 random 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).

[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`).

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 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.

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.

## 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).

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

| 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 |

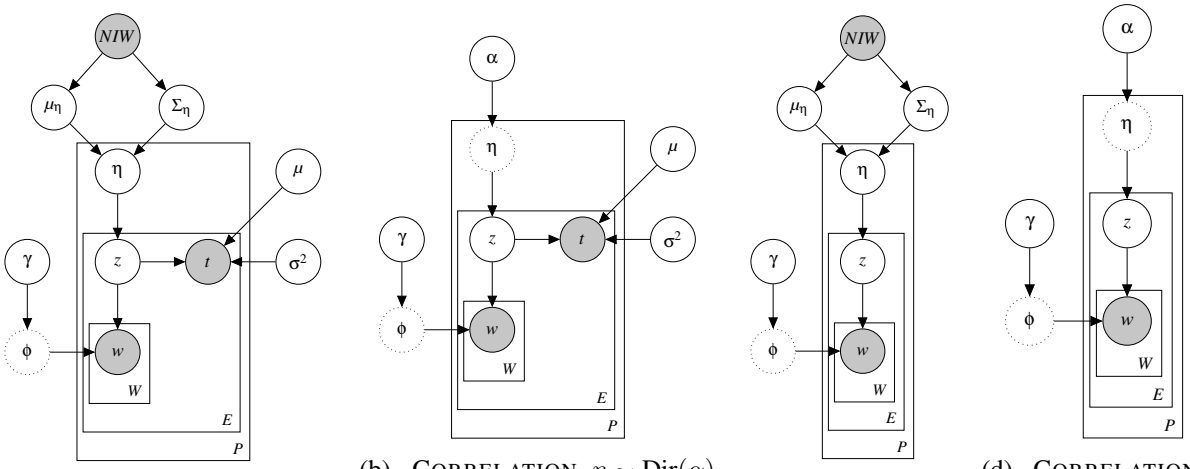
Table 1: A sample of 5 of the 64 sentences (original and converted) that constitute the data for Frank Lloyd Wright (born 1867). Each event is defined as one such temporally-scoped sentence.

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 pre-determined 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 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; Blei and Lafferty, 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 $\mu_\eta$ and covariance $\Sigma_\eta$, drawing them 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 $\eta_p$ into biography-event proportions $\beta_p$ through the softmax function: $\beta_{p,i} = \frac{\exp(\eta_{p,i})}{\sum_{k=1}^{K}\exp(\eta_{p,k})}$
  - For each event $e$ in biography $p$:
    - Draw event class index $z \sim \text{Mult}(\beta_p)$
    - Draw timestamp $t \sim \mathcal{N}(\mu_z, \sigma_z^2)$
    - For each token in event $e$:
      - Draw term $w \sim \text{Mult}(\phi_z)$




(a) FULL. (b) –CORRELATION. $\eta \sim \text{Dir}(\alpha)$ (c) –TIME (d) –CORRELATION, –TIME. $\eta \sim \text{Dir}(\alpha)$

Figure 1: Graphical form of the full model (described in §3) and models with ablations (described in §4).

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 $z$ using collapsed Gibbs sampling conditioned on current values for $\eta$ and all other $z$.
2. For each biography $p$, maximize likelihood with respect to $\eta_p$ via gradient ascent given the current samples of $z$ and priors $\mu_\eta$ and $\Sigma_\eta$.
3. Assign MAP estimates of $\mu_\eta$ and $\Sigma_\eta$ given current values of $\eta$ and the Normal-Inverse Wishart prior. Update $\mu$ and $\sigma^2$ according to its maximum likelihood estimate given $z$.

We describe the technical details of each step below.

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

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

Here $\mathbf{c}^-(k,v)$ is the count of the number of times vocabulary term $v$ shows up in all events whose current sample $z = k$ (excepting the current one being sampled), $\mathbf{c}^-(k,\star)$ is the total count of all terms in all events whose current $z = k$ (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. (Note the complexity of the last term is due to drawing multiple observations from a single collapsed multinomial; Carpenter, 2010.)

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

$$L(\eta) \propto \prod_{n=1}^{N} \frac{\exp(\eta_{z_n})}{\sum_{k=1}^{K} \exp(\eta_k)} \times \mathcal{N}(\eta \mid \mu_\eta, \Sigma_\eta) \quad (2)$$

The log likelihood is proportional to:

$$\ell(\eta) \propto \sum_{n=1}^{N} \eta_{z_n} - \sum_{n=1}^{N} \sum_{k=1}^{K} \exp(\eta_k) - \frac{1}{2} (\eta - \mu_\eta)^\top \Sigma_\eta^{-1} (\eta - \mu_\eta) \quad (3)$$

Given samples of the latent event class $z$ for all events in biography $p$, we maximize the value of $\eta_p$ using gradient ascent. We can think of this as maximizing the likelihood of the observations $z$ subject to $\ell_2$ (Gaussian) regularization, where the covariance

matrix in the regularizer encourages correlations in $\eta$: if a document contains many examples of $z = k$ and $z_k$ is highly correlated with $z_j$, then the optimal $\eta$ is encouraged to contain high weights at both $\eta_k$ and $\eta_j$ rather than simply $\eta_k$ alone.

**Maximizing $\mu_\eta, \Sigma_\eta, \mu, \sigma^2$.** Given values for $\eta$, we then find maximum *a posteriori* estimates of $\mu_\eta$ and $\Sigma_\eta$ conditioned on the Normal-Inverse Wishart (NIW) prior. The NIW is a conjugate prior to a multivariate Gaussian, parameterized by dimensionality $K$, initial mean $\mu_0$, positive-definite scale matrix $\Psi$, and scalars $\nu > K - 1$ and $\lambda > 0$. The prior parameters $\Psi$ and $\nu$ have an intuitive interpretation as the scatter matrix $\sum_{i=1}^{\nu} (x_i - \bar{x}) (x_i - \bar{x})^\top$ for $\nu$ pseudo-observations.

The expected value of the covariance matrix drawn from a NIW distribution parameterized by $\Psi$ and $\nu$ is $\frac{\Psi}{\nu - K - 1}$. To disprefer correlations among topics in the absence of strong evidence, we fix $\mu_0 = 0$ and set $\Psi$ so that this prior expectation over $\Sigma_\eta$ is the product of a scalar value $\rho$ and the identity matrix $\mathbf{I}$: $\Psi = (\nu - K - 1)\rho\mathbf{I}$; $\rho$ defines the expected variance, and the higher the value of $\nu$, the more strongly the prior dominates the posterior estimate of the covariance matrix (i.e., the more the covariance matrix is shrunk toward $\rho\mathbf{I}$). $\lambda$ 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 $\mu_\eta$ and $\Sigma_\eta$ have closed-form expressions given values of $\eta$ (here, $\bar{\eta}$ denotes the mean value of $\eta$ over all biographies).

$$\hat{\mu}_\eta = \frac{n}{\lambda + n} \bar{\eta} \tag{4}$$

$$\hat{\Sigma}_\eta = \frac{\Psi + \sum_{i=1}^{N} (\eta_i - \bar{\eta}) (\eta_i - \bar{\eta})^\top + \frac{\lambda n}{\lambda + n} \bar{\eta}\bar{\eta}^\top}{\nu + n + K + 1} \tag{5}$$

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

## 4 Evaluation

While the goal of this work is 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 $\ell_2$-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 1b. Rather than a logistic normal prior on the entity-specific distribution over event types ($\eta$), we draw $\eta$ from a symmetric Dirichlet distribution parameterized by a global $\alpha$. In a Dirichlet distribution, arbitrary correlations cannot be captured.
- –TIME, figure 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 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 multiword 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 burn-in period of 100 iterations, using slice sampling (Neal, 2003) to optimize the value of the Dirichlet hyperparameter $\gamma$ every 10 iterations; after inference, the parameters $\mu_\eta, \Sigma_\eta, \mu, \sigma^2$ and $\phi$ are es-

timated from samples drawn at the final iteration and held fixed. For test entities, we infer the MAP value of $\eta$ using development data, and predict the age of each test event as the mean time marginalizing over the event type indicator $z$. $\hat{t} = \mathbb{E}_z[\mu_z]$.

2. –CORRELATION. Here we perform collapsed Gibbs sampling for 100 iterations, using slice sampling to optimize the value of $\alpha$ and $\gamma$ every 10 iterations; after inference, the parameters $\mu, \sigma^2$ and $\phi$ are estimated from single final samples and held fixed. For development and test data, we run Gibbs sampling on event indicators $z$ for 10 iterations and predict the age of each test event as the mean time marginalizing over the event type indicator $z$. $\hat{t} = \mathbb{E}_z[\mu_z]$.
3. –TIME. Inference as above for 100 iterations, using slice sampling to optimize the value of $\gamma$ every 10 iterations; after inference, the parameters $\mu_\eta, \Sigma_\eta$ and $\phi$ 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 $\hat{\mu}_z$ as the empirical mean age of events sampled to event class $z$ using single samples for each event in the training data from the final sampling iteration. For test entities, we infer the MAP value of $\eta$ using development data, and predict the age of each test event as the average empirical age marginalizing over the event type indicator $z$. $\hat{t} = \mathbb{E}_z[\hat{\mu}_z]$.
4. –CORRELATION,–TIME. We perform inference as above for the –CORRELATION model, and time prediction as in the –TIME model. $\hat{t} = \mathbb{E}_z[\hat{\mu}_z]$.

To compare against a potentially more powerful discriminative model, we also evaluate linear regression with $\ell_2$ (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 $\lambda$ 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.

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{t} - t_i| \tag{6}$$

Figure 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.

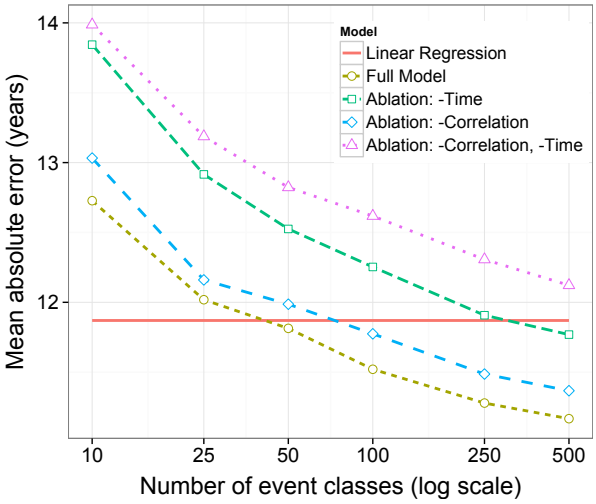



Figure 2: Mean average error (in years) for time prediction.

## 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

| Age $\mu$ | Age $\sigma$ | % 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 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`.

$K = 500$ event classes; as above, we run inference for a burn-in period of 100 iterations and collect 50 samples from the posterior distributions for $z$ (the event class indicator for each event).

Table 2 illustrates a sample of 20 event classes along with the mean time $\mu$ and standard deviation $\sigma$, the gender distribution (calculated from the posterior distribution over $z$ 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 3 and 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 [$p \ll 0.0001$] 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 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

[3]Using our method of gender inference described in §2, we are able to infer gender for 99.95% of biographies (242,858).

| $z$ | %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 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 4: Male-skewed event classes, ranked by z-score in a two-tailed binomial proportion test.

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.

## 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.

### 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 $\Sigma_\eta$, 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 $\eta$.

Table 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.

### 6.2 Historical distribution of events

Figure 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 dur-

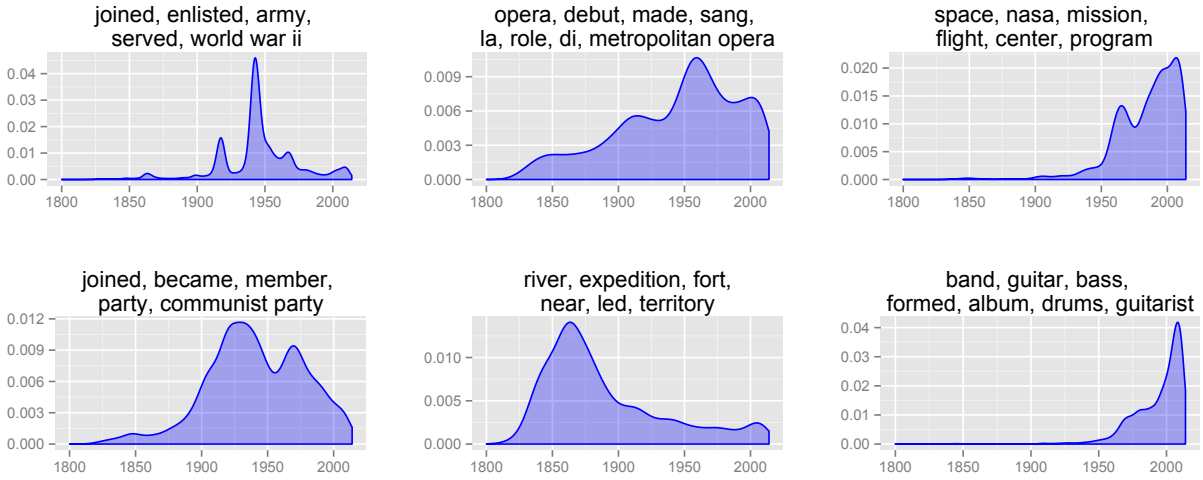



Figure 3: Historical distributions of event classes.

| $r$ | 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: Highest correlations between the *family, boss, murder, crime, mafia* class and other events.

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

## 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; Chambers and Jurafsky, 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, 2007a) 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.

## 8 Conclusion

We present a method for mining life events from biographies, leveraging the correlation structure of event descriptions. 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.

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. Latent topics have provided one way of organizing large document collections in the past (Mimno and McCallum, 2007b); 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. Data and code to support this work, including an interface to explore event classes in Wikipedia, can be found at `http://www.ark.cs.cmu.edu/bio/`.

## 9 Acknowledgments

We thank the anonymous reviewers, along with Dallas Card, Brendan O'Connor, Bryan Routledge, Yanchuan Sim and Ted Underwood, for their helpful comments. The research reported in this article was supported by U.S. National Science Foun-

dation grant CAREER IIS-1054319 to N.A.S. and Google's support of the Reading is Believing project at CMU. This work was made possible through the use of computing resources made available by the Open Science Data Cloud (OSDC), an Open Cloud Consortium (OCC)-sponsored project.

## References

David Bamman, Ted Underwood, and Noah A. Smith. 2014. A Bayesian mixed effects model of literary character. In *ACL*.

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

Fadi Biadsy, Julia Hirschberg, and Elena Filatova. 2008. An unsupervised approach to biography production using Wikipedia. In *ACL '08*, pages 807–815.

David M. Blei and John D. Lafferty. 2006a. Correlated topic models. In *NIPS '06*.

David M. Blei and John D. Lafferty. 2006b. Dynamic topic models. In *ICML '06*, pages 113–120.

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

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

Razvan Bunescu and Marius Pasca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *EACL '06*, pages 9–16, Trento, Italy.

Ewa S. Callahan and Susan C. Herring. 2011. Cultural bias in Wikipedia content on famous persons. *J. Am. Soc. Inf. Sci. Technol.*, 62(10):1899–1915, October.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr., and Tom M. Mitchell. 2010. Toward an architecture for never-ending language learning. In *AAAI '10*.

Bob Carpenter. 2010. Integrating Out Multinomial Parameters in Latent Dirichlet Allocation and Naive Bayes for Collapsed Gibbs Sampling. Technical report, LingPipe.

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

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *ACL '08*.

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *ACL '09*, pages 602–610.

Nathanael Chambers, Shan Wang, and Dan Jurafsky. 2007. 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. Association for Computational Linguistics.

Nathanael Chambers. 2013. Event schema induction with a probabilistic entity-driven model. In *EMNLP '13*, pages 1797–1807, Seattle, Washington, USA, October. Association for Computational Linguistics.

Jackie Chi Kit Cheung, Hoifung Poon, and Lucy Vanderwende. 2013. Probabilistic frame induction. In *NAACL '13*, pages 837–846, Atlanta, Georgia, June. Association for Computational Linguistics.

Benjamin Collier and Julia Bear. 2012. Conflict, criticism, or confidence: An empirical examination of the gender gap in Wikipedia contributions. In *CSCW '12*.

Mike Conway. 2010. Mining a corpus of biographical texts using keywords. *Literary and Linguistic Computing*, 25(1):23–35.

Silviu Cucerzan. 2007. Large-scale named entity disambiguation based on wikipedia data. In *EMNLP-CoNLL*, volume 7, pages 708–716.

Peter T. Davis, David K. Elson, and Judith L. Klavans. 2003. Methods for precise named entity matching in digital collections. In *JCDL '03*.

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

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

David K. Elson, Nicholas Dames, and Kathleen R. McKeown. 2010. Extracting social networks from literary fiction. In *ACL '10*, pages 138–147.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *EMNLP '11*, EMNLP '11, pages 1535–1545, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *ACL '05*, pages 363–370.

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

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

Brent Hecht and Darren Gergle. 2009. Measuring self-focus bias in community-maintained knowledge repositories. In *C&T '09*, pages 11–20.

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

Jerry R Hobbs, Douglas Appelt, John Bear, David Israel, Megumi Kameyama, and Mabry Tyson. 1993. Fastus: A system for extracting information from text.

In *Proceedings of the workshop on Human Language Technology*, pages 133–137. Association for Computational Linguistics.

Dennis P. Hogan and Nan Marie Astone. 1986. The transition to adulthood. *Annual Review of Sociology*, 12(1):109–130.

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

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

Josef Kolbitsch and Hermann A. Maurer. 2006. The transformation of the web: How emerging communities shape the information we consume. *J. UCS*, 12(2):187–213.

Lee A. Lillard and Linda J. Waite. 1993. A joint model of marital childbearing and marital disruption. *Demography*, 30(4):pp. 653–681.

Lee A. Lillard, Michael J. Brien, and Linda J. Waite. 1995. Premarital cohabitation and subsequent marital dissolution: A matter of self-selection? *Demography*, 32(3):pp. 437–457.

Inderjeet Mani, Marc Verhagen, Ben Wellner, Chong Min Lee, and James Pustejovsky. 2006. 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. Association for Computational Linguistics.

Sébastien Massoni, Madalina Olteanu, and Patrick Rousset. 2009. Career-path analysis using optimal matching and self-organizing maps. In *WSOM '09*.

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

Karl Ulrich Mayer. 1988. 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. Sage.

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

David Mimno and Andrew McCallum. 2007b. Organizing the OCA: Learning faceted subjects from a library of digital books. In *JCDL '07*, pages 376–385, New York, NY, USA. ACM.

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

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

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

John Modell. 1980. Normative aspects of american marriage timing since World War II. *Journal of Family History*, 5(2):210–234.

Ashutosh Modi, Ivan Titov, and Alexandre Klementiev. 2012. 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. Association for Computational Linguistics.

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

Pauline Ng. 2012. What Kobe Bryant and Britney Spears have in common: Mining Wikipedia for characteristics of notable individuals. In *ICWSM '12*.

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

Ulrike Pfeil, Panayiotis Zaphiris, and Chee Siang Ang. 2006. Cultural differences in collaborative authoring of Wikipedia. *Journal of Computer-Mediated Communication*, 12(1):88–113.

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

Joseph Reagle and Lauren Rhue. 2011. Gender bias in Wikipedia and Britannica. *International Journal of Communication*, 5(0).

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

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

Alan Ritter, Mausam, Oren Etzioni, and Sam Clark. 2012. Open domain event extraction from Twitter. In *KDD '12*, pages 1104–1112, New York, NY, USA. ACM.

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

Michael J. Shanahan. 2000. Pathways to adulthood in changing societies: Variability and mechanisms in

life course perspective. *Annual Review of Sociology*, 26(1):667–692.

Partha Pratim Talukdar, Derry Wijaya, and Tom Mitchell. 2012. Acquiring temporal constraints between relations. In *CIKM '12*, pages 992–1001, New York, NY, USA. ACM.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *NAACL '03*, pages 173–180.

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. 2007. 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.

Xuerui Wang and Andrew McCallum. 2006. Topics over time: a non-markov continuous-time model of topical trends. In *KDD '06*, pages 424–433.

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

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

Limin Yao, Aria Haghighi, Sebastian Riedel, and Andrew McCallum. 2011. Structured relation discovery using generative models. In *EMNLP '11*, pages 1456–1466, Stroudsburg, PA, USA. Association for Computational Linguistics.