Amith Ananthram (aa4461)

Discovering Events in Multilingual News Documents
with Word Embedding Based Topic Modeling
December 17, 2020

1 Introduction

Unsupervised and self-supervised learning techniques, applied to large text corpora, have dominated natural language processing (NLP) in recent years. These approaches are underpinned by Harris' distributional hypothesis **?**: words, sentences, or even documents that appear in similar contexts are likely similar themselves. You can recognize them by the company they keep.

Latent Dirichlet Allocation **?**, an early success in applying probabilistic modeling to NLP, can be viewed as an extension of this simple idea. Documents that contain the same sets of words likely express the same topic. By balancing sparsity in the distribution of topics over words against sparsity in the distribution of documents over topics, LDA is able to find sets of words (topics) and distributions over those sets of words (document proportions) that meaningfully explain a text corpus. More impressively, it uncovers this latent structure in a completely unsupervised manner.

While LDA is powerful, it is limited in a few important ways. First, because topics are modeled as distributions over tokens in a vocabulary, they are regularly heterogeneous, sometimes in ways that lack semantic coherence. While a topic that places mass on “Kuwait", “Saddam Hussein", and “oil" can reasonably be interpreted as referencing Iraq's invasion of its neighbor, a topic which favors “Hollywood", “explain", and “referendum" (discovered during my homework assignment) demands a good deal of imagination. Second, because LDA operates on a fixed vocabulary, it is difficult to adapt a fully trained model to new documents which contain (possibly important) words outside of its vocabulary. Finally, because LDA operates on the discrete tokens of its vocabulary, it is unable to find shared latent structure across languages in multilingual corpora.

In this work, I present a method for discovering events in multilingual news articles by addressing some of these limitations, arrived at through several iterations over Box's loop. It does so by 1) modeling the semantic senses present in the corpus in a multilingual continuous embedding space and 2) the named entities involved in the corpus in a discrete normalized space. I evaluate two separate methods for approximating the relevant posteriors, Gibbs sampling and autoencoding variational Bayes, and share findings for each.

2 Motivation

Clustering multilingual documents in meaningful ways enables a large variety of downstream NLP tasks. Such clusters regularly exhibit a distributional structure that is more predictable than the structure exhibited by the full corpus, affording distant supervision techniques in machine translation and information extraction less noisy training examples. In my own research in relation extraction, I currently rely on descriptions of date-marked events to find sentences in news articles that describe the same relation **?**. By leveraging an unsupervised topic model to discover these clusters instead of the limited set of labeled events on Wikipedia, I hope to greatly expand my training corpus and enable code-switched training.

Such a technique could also benefit other disciplines. Clusters of multilingual news articles describing the same event could give historians and sociologists an easier way to study how the same set of events are covered in different parts of the world. One possible extension could be automatically monitoring the freedom of the press in different nations based on these clusters.

3 Modeling Events

In broad strokes, there exists a fixed set of semantic senses whose union describe all news articles. This is evidenced in part by the topic tags that regularly accompany such articles. But, two articles about the market more likely than not describe different events – what disambiguates them are the specific entities involved. Building on these intuitions, I present a generative process designed to reveal the latent structure we seek: events.

Equation 1 describes the first model, a generative process that models the semantic senses present in a multilingual corpus as K multivariate Gaussian distributions with mean μ_k and covariance Σ_k in the space of a set of word embeddings $\boldsymbol{w}$.

$$p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\theta}, \boldsymbol{\zeta}, \boldsymbol{w}) = \prod_{k}^{K} p(\mu_k; m, \Sigma_k) p(\Sigma_k; \nu, \Psi) \prod_{i}^{N} (p(\theta_i; t) \prod_{j}^{M_i} (p(\zeta_{ij}|\theta_i) p(w_{ij}|\zeta_{ij}, \mu_{\zeta_{ij}}, \Sigma_{\zeta_{ij}})) \tag{1}$$

Equation 2 describes the second model, a generative process that models the normalized named entities (persons, places, etc) present in a multlingual corpus as H categorical distributions β over the shared vocabulary. While this model is simply LDA, the novelty lies in the pre-processing which normalizes named entities across different character sets.

$$p(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\xi}, \boldsymbol{e}) = \prod_{h}^{H} p(\beta_h; b) \prod_{i}^{N} (p(\phi_i; f) \prod_{j}^{M_i} (p(\xi_{ij}|\phi_i) p(e_{ij}|\xi_{ij}, \beta_{\xi_{ij}})) \tag{2}$$

By modeling the nouns and verbs in an article as independent from the normalized named entities in an article, I can compose equations 1 and 2 into a more complicated model simply by taking their products, resulting in the complete generative model described in Algorithm 1.

Algorithm 1: Complete generative model

```
for each semantic sense $k$ do
    Draw semantic sense mean, covariance
      $\mu_k, \Sigma_k \sim$ Normal-Inverse-Wishart$(m_k, \lambda_k, \nu_k, \Psi_k)$;
end
for each named entity topic $h$ do
    Draw named entity distribution $\beta_h \sim$ Dirichlet$(b_h)$;
end
for each document $w_i, e_i$ do
    Draw semantic sense proportions $\theta_i \sim$ Dirichlet$(t_i)$;
    for each noun or verb embedding $w_{ij}$ do
        Sample sense $\zeta_{ij} \sim$ Multinomial$(\theta_i)$;
        Sample noun or verb embedding $w_{ij} \sim$ Normal$(\mu_{\zeta_{ij}}, \Sigma_{\zeta_{ij}})$;
    end
    Draw named entity proportions $\phi_i \sim$ Dirichlet$(f_i)$;
    for each named entity $e_{ij}$ do
        Sample sense $\xi_{ij} \sim$ Multinomial$(\phi_i)$;
        Sample named entity $e_{ij} \sim$ Multinomial$(\beta_{\xi ij})$;
    end
end
```

4 Inference

To approximate the posterior distributions of my models, I evaluate both Gibbs sampling and autoencoding variational Bayes, improving naive implementations with techniques adapted from the literature.

For Gibbs sampling, I began by deriving the complete conditionals for latent variables in my model (presented in the Appendix). A quick inspection of these complete conditionals reveals a computationally expensive sampler, requiring the inversion of multiple large matrices at every iteration $O(d^3)$. **?** derive a collapsed sampler that avoids these expensive inversions by leveraging the Cholesky decomposition and confining mutations to rank one operations. I adopt their technique to improve my sampler.

For autoencoding variational Bayes, I adapt the methodology presented in **?**. The greatly speed up inference by 1) collapsing the individual document-word assignment terms via amortization and 2) enabling re-parameterization of the Dirichlet prior on topic proportions with a log-normal distribution derived via Laplace approximation in the soft-max basis (allowing unconstrained optimization). While this enables extremely fast inference, the global latent variables of my model (the topic distributions) lose their Bayesian treatment, becoming learned parameters in the decoder network, providing an interesting point of comparison to the Gibbs sampling approach.

5 Experiments

Dataset

I use the Reuters RCV1 and RCV2 corpora **?**. While this corpus contains millions of articles in many languages from 1996 and 1997, I select only the $100k$ English, Spanish and Russian articles from the last 3 months of 1996, allowing me to evaluate more experimental variants in languages with relatively more (English and Spanish) and relatively less (Russian) vocabulary overlap.

Feature extraction

Using SpaCy[1] and Stanza[2] language models, I extract lemmatized forms of all nouns, verbs and named entities, producing two bag of words representations for each article (one of all noun and verb lemmas, another of all named entities).

I replace all noun and verb lemmas with their corresponding dense representation as provided by Facebook's MUSE aligned multilingual embeddings[3]. I normalize all non-English named entities by translating them into English with Helsinki NLP's machine translation models[4], producing a shared vocabulary of named entities across all the articles in the corpus.

Inference

While I implemented my Gibbs sampler from scratch in Python, I relied on many of the primitives in Pyro[5] to implement my variational autoencoder (in addition to a very helpful tutorial[6]).

6 Results

[TODO once I have results]

7 Conclusion

[TODO once I have results]

8 Appendix

Derivation of complete conditionals for μ_k and Σ_k

Below is the derivation of the complete conditional distributions for μ_k and Σ_k, the latent variables that govern our multivariate Gaussian distributions which model semantic sense (with Gaussian and inverse-Wishart priors, respectively). Note that these are the same for both variants 1 and 2 of our model as the bag of word embeddings is modeled independently from the bag of words.

W_k is the set of words across all documents assigned to topic k with size n_{W_k}. $\bar{w}_k = (1/n_{W_k}) \sum_l^{n_{W_k}} w_l$. Updated parameters for the conjugate complete conditionals are highlighted in green. Normalizing factors (that is, factors that are constant with respect to the random variable being modeled) which are dropped from the exponents (or in rare cases raised into the exponent) are highlighted in yellow.

[1] https://spacy.io/

[2] https://stanfordnlp.github.io/stanza/

[3] https://ai.facebook.com/tools/muse/

[4] https://huggingface.co/Helsinki-NLP

[5] https://pyro.ai/

[6] https://github.com/pyro-ppl/pyro/pull/2655

$$
\begin{aligned}
p(\mu_k|\mu_{-k}, \boldsymbol{\Sigma}, \boldsymbol{\theta}, \boldsymbol{\zeta}, \boldsymbol{w}) &\propto p(\mu_k; m_k, S_k) \prod_{w \in W_k} p(w|\mu_k, \Sigma_k) \\
&\propto \exp -\frac{1}{2}(\mu_k - m_k)^T S_k^{-1}(\mu_k - m_k) \prod_{w \in W_k} \exp -\frac{1}{2}(w - \mu_k)^T \Sigma_k^{-1}(w - \mu_k) \\
&\propto \exp -\frac{1}{2}(\mu_k - m_k)^T S_k^{-1}(\mu_k - m_k) \exp \sum_{w \in W_k} -\frac{1}{2}(w - \mu_k)^T \Sigma_k^{-1}(w - \mu_k) \\
&\propto \exp -\frac{1}{2}((\mu_k - m_k)^T S_k^{-1}(\mu_k - m_k) + \sum_{w \in W_k} (w - \mu_k)^T \Sigma_k^{-1}(w - \mu_k)) \\
&\propto \exp -\frac{1}{2}((\mu_k - m_k)^T S_k^{-1}(\mu_k - m_k) + \sum_{w \in W_k} (w^T \Sigma_k^{-1} w - 2\mu_k^T \Sigma_k^{-1} w + \mu_k^T \Sigma_k^{-1} \mu_k)) \\
&\propto \exp -\frac{1}{2}((\mu_k - m_k)^T S_k^{-1}(\mu_k - m_k) - 2\mu_k^T \Sigma_k^{-1} n_{W_k} \bar{w}_k + n_{W_k} \mu_k^T \Sigma_k^{-1} \mu_k) \\
&\propto \exp -\frac{1}{2}(\mu_k^T (S_k^{-1} + n_{W_k} \Sigma_k^{-1})\mu_k - 2\mu_k^T (S_k^{-1} m_k + \Sigma_k^{-1} n_{W_k} \bar{w}_k) + m_k^T S_k^{-1} m_k) \\
&\propto \exp -\frac{1}{2}(\mu_k^T \hat{S}_k^{-1} \mu_k - 2\mu_k^T \hat{S}_k^{-1} \hat{S}_k(S_k^{-1} m_k + \Sigma_k^{-1} n_{W_k} \bar{w}_k)) \\
&\propto \exp -\frac{1}{2}(\mu_k^T \hat{S}_k^{-1} \mu_k - 2\mu_k^T \hat{S}_k^{-1} \hat{m}_k + \hat{m}_k^T \hat{S}_k^{-1} \hat{m}_k) \\
&= \mathcal{N}(\hat{m}_k, \hat{S}_k), \hat{m}_k = \hat{S}_k(S_k^{-1} m_k + \Sigma_k^{-1} n_{W_k} \bar{w}_k), \hat{S}_k^{-1} = S_k^{-1} + n_{W_k} \Sigma_k^{-1}
\end{aligned}
$$

Note that, when our conjugate prior is the normal inverse-Wishart distribution with mean covariance scaling factor λ, $S_k = \frac{1}{\lambda}\Sigma_k$.

$$
\begin{aligned}
p(\Sigma_k|\boldsymbol{\mu}, \Sigma_{-k}, \boldsymbol{\theta}, \boldsymbol{\zeta}, \boldsymbol{w}) &\propto p(\Sigma_k) \prod_{w \in W_k} p(w|\mu_k, \Sigma_k) \\
&\propto |\Sigma_k|^{-(\nu_k + d + 1)/2} \exp -\frac{1}{2} \operatorname{Tr}(\Psi_k \Sigma_k^{-1}) \prod_{w \in W_k} |\Sigma_k|^{-1/2} \exp -\frac{1}{2}(w - \mu_k)^T \Sigma_k^{-1}(w - \mu_k) \\
&\propto |\Sigma_k|^{-(n_{W_k} + \nu_k + d + 1)/2} \exp -\frac{1}{2}(\operatorname{Tr}(\Psi_k \Sigma_k^{-1}) + \sum_{w \in W_k} (w - \mu_k)^T \Sigma_k^{-1}(w - \mu_k)) \\
&\propto |\Sigma_k|^{-(n_{W_k} + \nu_k + d + 1)/2} \exp -\frac{1}{2}(\operatorname{Tr}(\Psi_k \Sigma_k^{-1}) + \sum_{w \in W_k} \operatorname{Tr}((w - \mu_k)^T \Sigma_k^{-1}(w - \mu_k))) \\
&\propto |\Sigma_k|^{-(n_{W_k} + \nu_k + d + 1)/2} \exp -\frac{1}{2}(\operatorname{Tr}(\Psi_k \Sigma_k^{-1}) + \sum_{w \in W_k} \operatorname{Tr}((w - \mu_k)(w - \mu_k)^T \Sigma_k^{-1})) \\
&\propto |\Sigma_k|^{-(n_{W_k} + \nu_k + d + 1)/2} \exp -\frac{1}{2}(\operatorname{Tr}(\Psi_k \Sigma_k^{-1} + \sum_{w \in W_k} (w - \mu_k)(w - \mu_k)^T \Sigma_k^{-1})) \\
&\propto |\Sigma_k|^{-(n_{W_k} + \nu_k + d + 1)/2} \exp -\frac{1}{2}(\operatorname{Tr}((\Psi_k + \sum_{w \in W_k} (w - \mu_k)(w - \mu_k)^T)\Sigma_k^{-1}))
\end{aligned}
$$

$$= \mathcal{W}^{-1}(\hat{\Psi}_k, \hat{\nu}_k), \hat{\Psi}_k = \Psi_k + \sum_{w \in W_k} (w - \mu_k)(w - \mu_k)^T, \hat{\nu}_k = n_{W_k} + \nu_k$$

Derivation of collapsed sampler for ζ_i

To support faster posterior inference, we can collapse our sampler by integrating out $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ and $\boldsymbol{\theta}$. This saves us the trouble of needing to resample all the topic means, covariances and document proportions at each iteration.

$$\begin{aligned}
p(\zeta_{ij}|\zeta_{-(ij)}, \boldsymbol{w}) &= \frac{p(\boldsymbol{\zeta}, \boldsymbol{w})}{p(\zeta_{-(ij)}, \boldsymbol{w})} \\
&\propto p(\boldsymbol{\zeta}, \boldsymbol{w}) \\
&\propto \int \int \int p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\theta}, \boldsymbol{\zeta}, \boldsymbol{w}) \, d\theta \, d\Sigma \, d\mu \\
&\propto \int \int \int p(\boldsymbol{\mu}) p(\boldsymbol{\Sigma}) p(\boldsymbol{\theta}) p(\boldsymbol{\zeta}|\theta) p(\boldsymbol{w}|\boldsymbol{\zeta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \, d\theta \, d\Sigma \, d\mu \\
&\propto \int \int p(\boldsymbol{\mu}) p(\boldsymbol{\Sigma}) p(\boldsymbol{w}|\boldsymbol{\zeta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \, d\Sigma \, d\mu \int p(\boldsymbol{\theta}) p(\boldsymbol{\zeta}|\theta) \, d\theta
\end{aligned}$$

We can simplify these integrals separately, dropping terms that are constant with respect to ζ_{ij}.

$$\begin{aligned}
&\int p(\boldsymbol{\mu}) p(\boldsymbol{\Sigma}) p(\boldsymbol{w}|\boldsymbol{\zeta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \, d\Sigma \, d\mu \\
&= \int \prod_k^K p(\mu_k; \vec{0}, (1/\lambda)\Sigma_k) p(\Sigma_k; \nu, \Psi) \prod_i^N \prod_j^{M_i} p(w_{ij}|\zeta_{ij}, \mu_{\zeta_{ij}}, \Sigma_{\zeta_{ij}}) \, d\Sigma \, d\mu \\
&= \prod_k^K \int p(\mu_k; \vec{0}, (1/\lambda)\Sigma_k) p(\Sigma_k; \nu, \Psi) \prod_i^N \prod_j^{M_i} p(w_{ij}|\zeta_{ij}, \mu_{\zeta_{ij}}, \Sigma_{\zeta_{ij}}) \, d\Sigma_k \, d\mu_k \\
&= \prod_k^K \int (2\pi)^{-d/2} |(1/\lambda)\Sigma_k|^{-1/2} \exp\left((-1/2)\mu_k^T ((1/\lambda)\Sigma_k)^{-1} \mu_k\right) \\
&\times \frac{|\Psi|^{\nu/2}}{2^{\nu d/2}\Gamma_d(\nu/2)} |\Sigma_k|^{-(\nu+d+1)/2} \exp\left((-1/2) \operatorname{Tr} \Psi \Sigma_k^{-1}\right) \\
&\times \prod_i^N \prod_j^{M_i} (2\pi)^{-d/2} |\Sigma_{\zeta_{ij}}|^{-1/2} \exp\left((-1/2)(w_{ij} - \mu_{\zeta_{ij}})^T \Sigma_{\zeta_{ij}}^{-1} (w_{ij} - \mu_{\zeta_{ij}})\right) d\Sigma_k \, d\mu_k \\
&= \prod_k^K \int (2\pi)^{-d/2} |(1/\lambda)\Sigma_k|^{-1/2} \exp\left((-1/2)\mu_k^T ((1/\lambda)\Sigma_k)^{-1} \mu_k\right) \\
&\times \frac{|\Psi|^{\nu/2}}{2^{\nu d/2}\Gamma_d(\nu/2)} |\Sigma_k|^{-(\nu+d+1)/2} \exp\left((-1/2) \operatorname{Tr} \Psi \Sigma_k^{-1}\right)
\end{aligned}$$

$$
\begin{aligned}
&\times \prod_{w \in W_k} (2\pi)^{-d/2} |\Sigma_k|^{-1/2} \exp\left((-1/2)(w-\mu_k)^T \Sigma_k^{-1} (w-\mu_k)\right) d\Sigma_k \, d\mu_k \\
&= \prod_k^K \int (2\pi)^{-d/2} |(1/\lambda)\Sigma_k|^{-1/2} \exp\left((-1/2)\mu_k^T ((1/\lambda)\Sigma_k)^{-1} \mu_k\right) \\
&\times \frac{|\Psi|^{\nu/2}}{2^{\nu d/2}\Gamma_d(\nu/2)} |\Sigma_k|^{-(\nu+d+1)/2} \exp\left((-1/2) \operatorname{Tr} \Psi \Sigma_k^{-1}\right) \\
&\times (2\pi)^{-(n_{W_k} d)/2} |\Sigma_k|^{-n_{W_k}/2} \exp \sum_{w \in W_k} \left((-1/2)(w-\mu_k)^T \Sigma_k^{-1} (w-\mu_k)\right) d\Sigma_k \, d\mu_k \\
&= \prod_k^K (2\pi)^{-(n_{W_k} d)/2} \frac{|\Psi|^{\nu/2}}{2^{\nu d/2}\Gamma_d(\nu/2)} \int (2\pi)^{-d/2} |(1/\lambda)\Sigma_k|^{-1/2} \exp\left((-1/2)\mu_k^T ((1/\lambda)\Sigma_k)^{-1} \mu_k\right) \\
&\times |\Sigma_k|^{-(\nu+n_{W_k}+d+1)/2} \exp\left((-1/2) \operatorname{Tr} (\Psi \Sigma_k^{-1}) + \sum_{w \in W_k} ((-1/2)(w-\mu_k)^T \Sigma_k^{-1} (w-\mu_k))\right)) \, d\Sigma_k \, d\mu_k \\
&= \prod_k^K (2\pi)^{-(n_{W_k} d)/2} \frac{|\Psi|^{\nu/2}}{2^{\nu d/2}\Gamma_d(\nu/2)} \int (2\pi)^{-d/2} |(1/\lambda)\Sigma_k|^{-1/2} \exp\left((-1/2)\mu_k^T ((1/\lambda)\Sigma_k)^{-1} \mu_k\right) \\
&\times |\Sigma_k|^{-(\nu+n_{W_k}+d+1)/2} \exp\left((-1/2) \operatorname{Tr} (\Psi \Sigma_k^{-1} + \sum_{w \in W_k} ((-1/2)(w-\mu_k)(w-\mu_k)^T \Sigma_k^{-1}))\right)) \, d\Sigma_k \, d\mu_k \\
&= \prod_k^K (2\pi)^{-(n_{W_k} d)/2} \frac{|\Psi|^{\nu/2}}{2^{\nu d/2}\Gamma_d(\nu/2)} \int (2\pi)^{-d/2} |(1/\lambda)\Sigma_k|^{-1/2} \exp\left((-1/2)\mu_k^T ((1/\lambda)\Sigma_k)^{-1} \mu_k\right) \\
&\times |\Sigma_k|^{-(\nu+n_{W_k}+d+1)/2} \exp\left((-1/2) \operatorname{Tr} ((\Psi + \sum_{w \in W_k} ((-1/2)(w-\mu_k)(w-\mu_k)^T) \Sigma_k^{-1}))\right)) \, d\Sigma_k \, d\mu_k
\end{aligned}
$$

$$
\begin{aligned}
\int p(\boldsymbol{\theta}) p(\boldsymbol{\zeta}|\theta) \, d\theta &= \int \prod_i^N p(\theta_i; t) \prod_j^M p(\zeta_{ij}|\theta_i) \, d\theta \\
&= \prod_i^N \int p(\theta_i; t) \prod_j^M p(\zeta_{ij}|\theta_i) \, d\theta_i \\
&= \prod_i^N \frac{\Gamma(\sum_k^K t_k)}{\prod_k^K \Gamma(t_k)} \int \prod_k^K \theta_{ik}^{t_k-1} \prod_j^{M_i} \theta_{i,\zeta_{ij}} \, d\theta_i \\
&= \prod_i^N \frac{\Gamma(\sum_k^K t_k)}{\prod_k^K \Gamma(t_k)} \int \prod_k^K \theta_{ik}^{t_k-1} \prod_k^K \theta_{ik}^{n_{ik}} \, d\theta_i \\
&= \prod_i^N \frac{\Gamma(\sum_k^K t_k)}{\prod_k^K \Gamma(t_k)} \int \prod_k^K \theta_{ik}^{t_k+n_{ik}-1} \, d\theta_i \\
&= \prod_i^N \frac{\Gamma(\sum_k^K t_k)}{\prod_k^K \Gamma(t_k)} \frac{\prod_k^K \Gamma(t_k+n_{ik})}{\Gamma(\sum_k^K t_k+n_{ik})} \int \frac{\Gamma(\sum_k^K t_k+n_{ik})}{\prod_k^K \Gamma(t_k+n_{ik})} \prod_k^K \theta_{ik}^{t_k+n_{ik}-1} \, d\theta_i
\end{aligned}
$$

$$= \prod_{i}^{N} \frac{\Gamma(\sum_{k}^{K} t_k)}{\prod_{k}^{K} \Gamma(t_k)} \frac{\prod_{k}^{K} \Gamma(t_k + n_{ik})}{\Gamma(\sum_{k}^{K} t_k + n_{ik})}$$