A Correlated Topic Model Using Word Embeddings

Guangxu Xun[1], Yaliang Li[1], Wayne Xin Zhao[2,3], Jing Gao[1], Aidong Zhang[1]
[1]Department of Computer Science and Engineering, SUNY at Buffalo, NY, USA
[2]School of Information, Renmin University of China, Beijing, China
[3]Beijing Key Laboratory of Big Data Management and Analysis Methods, Beijing, China
[1]{guangxux, yaliangl, jing, azhang}@buffalo.edu, [2]batmanfly@gmail.com

Abstract

Conventional correlated topic models are able to capture correlation structure among latent topics by replacing the Dirichlet prior with the logistic normal distribution. Word embeddings have been proven to be able to capture semantic regularities in language. Therefore, the semantic relatedness and correlations between words can be directly calculated in the word embedding space, for example, via cosine values. In this paper, we propose a novel correlated topic model using word embeddings. The proposed model enables us to exploit the additional word-level correlation information in word embeddings and directly model topic correlation in the continuous word embedding space. In the model, words in documents are replaced with meaningful word embeddings, topics are modeled as multivariate Gaussian distributions over the word embeddings and topic correlations are learned among the continuous Gaussian topics. A Gibbs sampling solution with data augmentation is given to perform inference. We evaluate our model on the 20 Newsgroups dataset and the Reuters-21578 dataset qualitatively and quantitatively. The experimental results show the effectiveness of our proposed model.

1 Introduction

Conventional topic models, such as Probabilistic Latent Semantic Analysis (PLSA) [Hofmann, 1999] and Latent Dirichlet Allocation (LDA) [Blei *et al.*, 2003], have proven to be a powerful unsupervised tool for the statistical analysis of document collections. Those methods [Zhu *et al.*, 2012], [Zhu *et al.*, 2014] follow the bag-of-word assumption and model each document as an admixture of latent topics, which are multinomial distributions over words.

A limitation of the conventional topic models is the inability to directly model correlations between topics, for instances, a document about autos is more likely to be related to motorcycles than to politics. In reality, it is natural to expect correlated latent topics in most text corpora. In order to address this limitation, the Correlated Topic Model (CTM) [Blei and Lafferty, 2006a] replaces the Dirichlet prior with logistic normal distribution which allows for covariance structure among the topics.

Nowadays, the rapidly developing technique in natural language processing – word embeddings [Bengio *et al.*, 2003], [Mikolov and Dean, 2013] – provides us with the possibility to model topics and topic correlations in the continuous semantic space. Word embeddings, also known as word vectors and distributed representations of words, are real-valued continuous vectors for words, which have proven to be effective at capturing semantic regularities in language. Words with similar semantic and syntactic properties tend to be projected into nearby area in the vector space. By replacing the original discrete word types in LDA with continuous word embeddings, Gaussian-LDA [Das *et al.*, 2015] has shown that the additional semantics in word embeddings can be incorporated into topic models and further enhance the performance.

The main goal of correlated topic models is to model and discover correlation between topics. And now we know that word embeddings are able to capture semantic regularities in language, and the correlations between words can be directly measured by the Euclidean distances or cosine values between the corresponding word embeddings. Moreover, semantically related words are close to each other in space and should be more likely to be grouped into the same topic. Since Gaussian distributions depict a notion of centrality in continuous space, it is a natural choice to model topics as Gaussian distributions over word embeddings in space. Therefore, the motivation of this paper is to model topics in the word embedding space, exploit the known correlation information at word level and further improve the correlation discovery at topic level.

In this paper, we propose the Correlated Gaussian Topic Model (CGTM) to model both topics and topic correlations in the word embedding space. More specifically, first we learn word embeddings with the help of external large unstructured text corpora to obtain additional word-level correlation information; Second, in the vector space of word embeddings, we model topics and topic correlations to exploit useful additional semantics in word embeddings, wherein each topic is represented as a Gaussian distribution over the word embeddings and topic correlations are learned among those Gaussian topics. Third, we develop a Gibbs sampling algorithm for CGTM.

To validate the efficacy of our proposed model, we evaluate our model on the 20 Newsgroups dataset and the Reuters-21578 dataset, which are well-known dataset for experiments in text mining domain. The experimental results show that our model can discover more reasonable topics and topic correlations than the baseline models.

2 Related Works

Correlation is an inherent property in many text corpora, for example, [Blei and Lafferty, 2006b] explores the time evolution of topics and [Mei *et al.*, 2008] analyzes the locational correlation among topics. However, due to the use of the Dirichlet prior, traditional topic models are not able to model the topic correlation directly. CTM [Blei and Lafferty, 2006a] proposes to use logistic normal distribution to model the variability among topic proportions and thus learn the covariance structure of topics.

Word embeddings can capture the semantic meanings of words via low-dimensional real-valued vectors [Mikolov and Dean, 2013], for example, vector operation vector('king') - vector('man') + vector('woman') results in a vector which is very close to vector('queen'). The concept of word embeddings was first introduced into natural language processing by Neural Probabilistic Language Model (NPLM) [Bengio *et al.*, 2003]. Due to its effectiveness and wide variety of application domains, word embeddings have garnered a great deal of attention and development [Mikolov *et al.*, 2013], [Pennington *et al.*, 2014], [Morin and Bengio, 2005], [Collobert and Weston, 2008], [Mnih and Hinton, 2009], [Huang *et al.*, 2012].

Since word embeddings carry additional semantics, many researchers have tried to incorporate them into topic models to improve the performance [Das *et al.*, 2015], [Li *et al.*, 2016], [Liu *et al.*, 2015], [Li *et al.*, 2017]. [Liu *et al.*, 2015] proposed Topical Word Embeddings (TWE) which combines word embeddings and topic models in a simple and effective way to achieve topical embeddings for each word. [Das *et al.*, 2015] uses Gaussian distributions to model topics in the word embedding space.

The aforementioned models either fail to directly model correlation among topics or fail to leverage the word-level semantics and correlations. We propose to leverage the word-level semantics and correlations within word embeddings to aid us in learning the topic-level correlations.

3 Learning Word Embeddings

We begin our topic discovery process with learning the word embeddings with semantic regularities. Unlike the traditional one-hot representations of words which encode each word as a binary vector of N (the size of vocabulary) digits with only one digit being 1 the others 0, the distributed representations of words encode each word as a unique real-valued vector. By mapping words into this vector space, word embeddings are able to overcome several drawbacks of the one-hot representations such as the curse of dimensionality, the out-of-vocabulary words and the lack of semantics.

In this paper, we adopt a recently developed, very effective and efficient distributed representations of words based model called Word2Vec [Mikolov and Dean, 2013] to train word embeddings. In the learning process of Word2Vec, words with similar meanings gradually converge to nearby areas in the vector space. In this model, words in the form of word embeddings are used as input to a softmax classifier and each word is predicted based on its neighbourhood words within a certain context window.

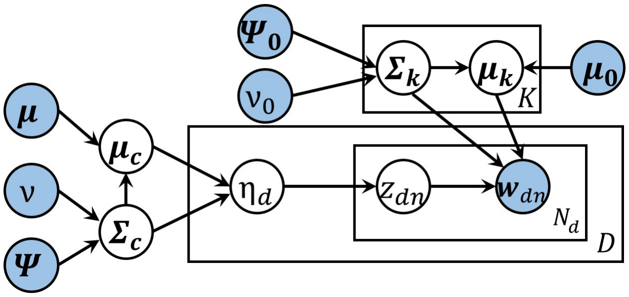

Figure 1: Schematic illustration of the CGTM framework.

Having learnt the word embeddings, given a word w_{dn}, which denotes the n^{th} word in d^{th} document, we can enrich that word by replacing it with the corresponding word embedding. The following section describes how this enrichment is used in a generative process to model topics and topic correlations.

4 Generative Process

Trained word embeddings give us useful additional semantics, which helps us discover reasonable topics and topic correlations in the vector space. However, each document now is a sequence of continuous word embeddings instead of a sequence of discrete word types. Therefore, conventional topic models no longer are applicable. Since the word embeddings are located in space based on their semantics and syntax, inspired by [Hu *et al.*, 2012] and [Das *et al.*, 2015], we consider them as draws from several Gaussian distributions. Hence, each topic is characterized as a multivariate Gaussian distribution in the vector space. The choice of Gaussian distribution is justified by the observations that Euclidean distances between word embeddings are consistent with their semantic similarities.

The graphical model of CGTM is shown in Figure 1. More formally, there are K topics and each topic is represented by a multivariate Gaussian distribution over the word embeddings in the word vector space. Let $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ denote the mean and covariance for the k^{th} Gaussian topic. Each document is a admixture of K Gaussian topics. $\boldsymbol{\eta_d}$ is a K dimensional vector where each dimension represents the weight of each topic in document d. Then the document-specific topic distribution $\boldsymbol{\theta_d}$ can be computed based on $\boldsymbol{\eta_d}$. $\boldsymbol{\mu}_c$ is the mean of $\boldsymbol{\eta}$ and $\boldsymbol{\Sigma}_c$ is the covariance of $\boldsymbol{\eta}$. By replacing the Dirichlet priors in conventional LDA with logistic normal priors, the topic correlation information is integrated into the model. $\boldsymbol{\mu}_0$, $\boldsymbol{\Sigma}_0$, ν_0, $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and ν are hyper parameters for Gaussian topics and logistic normal priors.

Note that variables in bold font mean they are either vectors or matrices, for example, $\boldsymbol{w}_{dn}$. The generative process is as follows:

1. Draw $\boldsymbol{\Sigma}_c \sim \mathcal{W}^{-1}(\boldsymbol{\Psi}, \nu)$.
2. Draw $\boldsymbol{\mu}_c \sim \mathcal{N}(\boldsymbol{\mu}, \frac{1}{\tau_c}\boldsymbol{\Sigma}_c)$.
3. For each Gaussian topic $k = 1, 2, \cdots, K$:
   (a) Draw topic covariance $\boldsymbol{\Sigma}_k \sim \mathcal{W}^{-1}(\boldsymbol{\Psi}_0, \nu_0)$.
   (b) Draw topic mean $\boldsymbol{\mu}_k \sim \mathcal{N}(\boldsymbol{\mu}_0, \frac{1}{\tau}\boldsymbol{\Sigma}_k)$.
4. For each document $d = 1, 2, \cdots, D$:
   (a) Draw $\boldsymbol{\eta}_d \sim \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$.
   (b) For each word index $n = 1, 2, \cdots, N_d$:
      i. Draw a topic $z_{dn} \sim Multinomial(f(\boldsymbol{\eta}_d))$.
      ii. Draw a word $\boldsymbol{w}_{dn} \sim \mathcal{N}(\boldsymbol{\mu}_{z_{dn}}, \boldsymbol{\Sigma}_{z_{dn}})$.

where τ and τ_c are constant factors; and $f(\boldsymbol{\eta})$ is the logistic transformation:

$$f(\eta_d^k) = \theta_d^k = \frac{\exp(\eta_d^k)}{\sum_i \exp(\eta_d^i)}. \tag{1}$$

The following conjugate priors are utilized for topic parameters: a Gaussian distribution $\mathcal{N}$ for the mean and an inverse Wishart distribution $\mathcal{W}^{-1}$ for the covariance. However, note that there is still a non-conjugacy problem between the logistic normal distribution and multinomial distribution, and we will solve this with data augmentation technique in the following section.

5 Parameter Inference

The observed variables are documents consisting of word embeddings, and our goal is to infer the posterior Gaussian distribution of each topic, topic assignment of each word, and topic correlations. Given D documents and the corresponding word embeddings $\boldsymbol{w}$, the joint distribution of topic assignments $\boldsymbol{z}$ and logistic normal parameters $\boldsymbol{\eta}$ is:

$$\begin{aligned} p(\boldsymbol{z}, \{\boldsymbol{\eta}_d\}_{d=1}^D | \boldsymbol{w}) &\propto p(\boldsymbol{w}|\boldsymbol{z}) \prod_{d=1}^{D} (\prod_{n=1}^{N_d} \theta_d^{z_{dn}}) \mathcal{N}(\boldsymbol{\eta}_d | \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) \\ &\propto p(\boldsymbol{w}|\boldsymbol{z}) \prod_{d=1}^{D} (\prod_{n=1}^{N_d} \frac{\exp(\eta_d^{z_{dn}})}{\sum_i^K \exp(\eta_d^i)}) \mathcal{N}(\boldsymbol{\eta}_d | \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c), \end{aligned} \tag{2}$$

where $p(\boldsymbol{w}|\boldsymbol{z})$ is the Gaussian probability of words $\boldsymbol{w}$ under topic assignments $\boldsymbol{z}$. Because of the choice of conjugate priors for topic parameters, those variables can be integrated out and we can efficiently re-sample topic assignment for each word. However, due to the non-conjugacy between the logistic normal and multinomial distributions, regular Gibbs sampling scheme doesn't work for the logistic normal parameters. Thus we adopt Gibbs sampling with data augmentation technique to solve this non-conjugacy problem.

5.1 Sampling Topic Assignments

Since the topic parameters have conjugate priors, the sampling process of topic assignments is similar to the Gibbs sampling scheme for LDA [Griffiths, 2002]. Given $\boldsymbol{\eta}$ and $\boldsymbol{z}_{-dn}$ which is the topic assignment scheme without considering the current word w_{dn}, the topic of each word is drawn iteratively as:

$$\begin{aligned} p(z_{dn} = k | \boldsymbol{z}_{-dn}, \boldsymbol{w}) &\propto p(z_{dn} = k | \boldsymbol{z}_{-dn}) p(\boldsymbol{w}_{dn} | z_{dn} = k) \\ &\propto \frac{\exp(\eta_d^k)}{\sum_i \exp(\eta_d^i)} \cdot T_r(\boldsymbol{w}_{dn} | \boldsymbol{\mu}_k, \frac{\tau_k + 1}{\tau_k} \boldsymbol{\Sigma}_k), \end{aligned} \tag{3}$$

where $T_r(\boldsymbol{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the multivariate Student's t-distribution for Gaussian sampling with $(r = \nu_k - dim + 1)$ being its degrees of freedom and dim being the dimensionality of word embeddings. $(\nu_k = \nu + N_k)$ and $(\tau_k = \tau + N_k)$ are the parameters of topic k, where N_k denotes the total number of words that are assigned to topic k.

5.2 Updating Gaussian Topics

Every time we re-sample topic assignment z_{dn}, we need to update the two involved Gaussian topics because the current word w_{dn} is either leaving or joining this Gaussian topic. Following [Das *et al.*, 2015], we derive the updates for $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ of the posterior Gaussian distributions for topic k:

$$\begin{aligned} \boldsymbol{\mu}_k &= \frac{\tau \boldsymbol{\mu}_0 + N_k \bar{\boldsymbol{w}}_k}{\tau_k}, \\ \boldsymbol{\Sigma}_k &= \frac{\boldsymbol{\Psi}_0 + \boldsymbol{C}_k + \tau N_k (\bar{\boldsymbol{w}}_k - \boldsymbol{\mu}_0)(\bar{\boldsymbol{w}}_k - \boldsymbol{\mu}_0)^T / \tau_k}{\nu_k - dim + 1}, \end{aligned} \tag{4}$$

where $\bar{\boldsymbol{w}}_k$ is the sample mean of all the word embeddings assigned to topic k, and $\boldsymbol{C}_k$ is the scaled form of sample covariance of all the word embeddings assigned to topic k. These two intermediate variables are calculated as follows:

$$\begin{aligned} \bar{\boldsymbol{w}}_k &= \frac{\sum_{d=1}^{D} \sum_{n=1}^{N_d} \delta(z_{dn}, k) \boldsymbol{w}_{dn}}{N_k}, \\ \boldsymbol{C}_k &= \sum_{d=1}^{D} \sum_{n=1}^{N_d} \delta(z_{dn}, k)(\boldsymbol{w}_{dn} - \bar{\boldsymbol{w}}_k)(\boldsymbol{w}_{dn} - \bar{\boldsymbol{w}}_k)^T, \end{aligned} \tag{5}$$

where $\delta(z_{dn}, k)$ is the Kronecker delta function that $\delta(z_{dn}, k) = 1$ if $z_{dn} = k$, $\delta(z_{dn}, k) = 0$ otherwise.

5.3 Sampling Logistic Normal Parameters

Given topic assignments, directly sampling logistic normal parameters $\boldsymbol{\eta}$ is difficult due to non-conjugacy. To address the non-conjugacy problem between the logistic normal distribution and multinomial distribution, following [Holmes *et al.*, 2006], [Polson *et al.*, 2013] and [Chen *et al.*, 2013], we sample the logistic normal parameters $\boldsymbol{\eta}$ based on $\boldsymbol{z}$ with auxiliary variables. For document d, the likelihood for η_d^k conditioned on $\boldsymbol{\eta}_d^{-k}$ is:

$$\begin{aligned} & l(\eta_d^k | \boldsymbol{\eta}_d^{-k}) \\ &= \prod_{n=1}^{N_d} \left(\frac{\exp(\eta_d^k)}{\sum_i \exp(\eta_d^i)} \right)^{z_{dn}^k} \left(1 - \frac{\exp(\eta_d^k)}{\sum_i \exp(\eta_d^i)} \right)^{1 - z_{dn}^k} \\ &= \frac{(\exp(\rho_d^k))^{C_d^k}}{(1 + \exp(\rho_d^k))^{N_d}}, \end{aligned} \tag{6}$$

where z_{dn}^k is the topic indicator that $z_{dn}^k = 1$ if word w_{dn} is assigned to k^{th} topic, $z_{dn}^k = 0$ otherwise. $\rho_d^k = \eta_d^k - \zeta_d^k$, $\zeta_d^k = \log(\sum_{j \neq k} \exp(\eta_d^j))$ and C_d^k is the number of words assigned to topic k in document d. Therefore, we obtain the posterior distribution of η_d^k proportional to multiplying the likelihood by the prior:

$$p(\eta_d^k | \boldsymbol{\eta}_d^{-k}, \boldsymbol{z}, \boldsymbol{w}) \propto l(\eta_d^k | \boldsymbol{\eta}_d^{-k}) \mathcal{N}(\eta_d^k | \mu_d^k, \sigma_k^2). \quad (7)$$

For the prior part, it is a univariate Gaussian distribution conditioned on the other logistic normal parameters in the current document $\boldsymbol{\eta}_d^{-k}$. Thus, given $\boldsymbol{\eta}_d^{-k}$ and $\boldsymbol{\mu}_c$, $\boldsymbol{\Sigma}_c$ of the multivariate Gaussian distribution over $\boldsymbol{\eta}$, we have:

$$\begin{aligned} \mu_d^k &= \mu_k - \boldsymbol{\Lambda}_{kk}^{-1} \boldsymbol{\Lambda}_{k-k} (\boldsymbol{\eta}_d^{-k} - \boldsymbol{\mu}_{-k}), \\ \sigma_k^2 &= \boldsymbol{\Lambda}_{kk}^{-1}, \end{aligned} \quad (8)$$

where $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}_c^{-1}$ is the precision matrix. However, the non-conjugacy makes it difficult to directly calculate the likelihood $l(\eta_d^k | \boldsymbol{\eta}_d^{-k})$ and thus unable to directly sample η_d^k.

By introducing auxiliary Polya-Gamma variable λ_d^k [Polson *et al.*, 2013], we are able to get around the non-conjugacy problem and the likelihood $l(\eta_d^k | \boldsymbol{\eta}_d^{-k})$ can now be expressed as:

$$\begin{aligned} & l(\eta_d^k | \boldsymbol{\eta}_d^{-k}) = \\ & \frac{1}{2^{N_d}} \exp(\kappa_d^k \rho_d^k) \int_0^\infty \exp(-\frac{\lambda_d^k (\rho_d^k)^2}{2}) p(\lambda_d^k | N_d, 0) d\lambda_d^k, \end{aligned} \quad (9)$$

where $\kappa_d^k = C_d^k - N_d/2$ and $p(\lambda_d^k | N_d, 0)$ is the Polya-Gamma distribution $\mathcal{PG}(N_d, 0)$. As one can observe, Equation 9 implies that $p(\eta_d^k | \boldsymbol{\eta}_d^{-k}, \boldsymbol{z}, \boldsymbol{w})$ is the marginal distribution of the joint distribution:

$$\begin{aligned} & p(\eta_d^k, \lambda_d^k | \boldsymbol{\eta}_d^{-k}, \boldsymbol{z}, \boldsymbol{w}) \propto \\ & \frac{1}{2^{N_d}} \exp(\kappa_d^k \rho_d^k - \frac{\lambda_d^k (\rho_d^k)^2}{2}) p(\lambda_d^k | N_d, 0) \mathcal{N}(\eta_d^k | \mu_d^k, \sigma_k^2). \end{aligned} \quad (10)$$

Therefore we can sample η_d^k based on the auxiliary variable λ_d^k. The sampling procedure is as follows:

- Sampling λ_d^k: according to Equation 10 and [Polson *et al.*, 2013], we have the conditional distribution $p(\lambda_d^k | \boldsymbol{z}, \boldsymbol{w}, \boldsymbol{\eta}) \propto \exp(-\frac{\lambda_d^k (\rho_d^k)^2}{2}) p(\lambda_d^k | N_d, 0)$, which results in a Polya-Gamma distribution $\mathcal{PG}(N_d, \rho_d^k)$. Following the ideas in [Polson *et al.*, 2013] and [Chen *et al.*, 2013], Polya-Gamma variables can be drawn in O(1) time, and so a sample of λ_d^k is obtained.
- Sampling η_d^k: according to Equation 10, we can sample η_d^k with posterior probability:

  $$p(\eta_d^k | \boldsymbol{\eta}_d^{-k}, \boldsymbol{z}, \boldsymbol{w}, \lambda) \propto \exp(\kappa_d^k \eta_d^k - \frac{\lambda_d^k (\eta_d^k)^2}{2}) \mathcal{N}(\eta_d^k | \mu_d^k, \sigma_k^2). \quad (11)$$

  This results in a univariate Gaussian distribution $\mathcal{N}(\gamma_d^k, (\tau_d^k)^2)$ conditioned on the auxiliary variable λ_d^k, where $\gamma_d^k = (\tau_d^k)^2 (\sigma_d^{-2} \mu_d^k + \kappa_d^k + \lambda_d^k \zeta_d^k)$ and $(\tau_d^k)^2 = (\sigma_d^{-2} + \lambda_d^k)^{-1}$. Thus, given the auxiliary variable λ_d^k, η_d^k can be easily drawn from a univariate Gaussian distribution.

5.4 Updating Topic Correlation

Given $\{\boldsymbol{\eta}_d\}_{d=1}^D$, the logistic normal parameters $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$ are updated as:

$$\begin{aligned} \boldsymbol{\mu}_c &= \frac{\tau_c}{\tau_c + D} \boldsymbol{\mu} + \frac{D}{\tau_c + D} \bar{\boldsymbol{\eta}}, \\ \boldsymbol{\Sigma}_c &= \boldsymbol{\Psi} + Q + \frac{\tau_c D}{\tau_c + D} (\bar{\boldsymbol{\eta}} - \boldsymbol{\mu})(\bar{\boldsymbol{\eta}} - \boldsymbol{\mu})^T, \end{aligned} \quad (12)$$

where $\bar{\boldsymbol{\eta}}$ is the mean of $\{\boldsymbol{\eta}_d\}_{d=1}^D$, and $Q = \frac{1}{D}(\boldsymbol{\eta}_d - \bar{\boldsymbol{\eta}})(\boldsymbol{\eta}_d - \bar{\boldsymbol{\eta}})^T$.

6 Experiments

In this section, we carry out experiments on two real-world text collections – the 20 Newsgroups dataset[1] and the Reuters-21578 dataset[2] to demonstrate the efficacy of our proposed model. 20 Newsgroups contains approximately 20,000 text documents partitioned evenly across 20 different newsgroups. Reuters contains about 10,000 documents, but due to the imbalance of each category, only the largest 8 categories are selected in Reuters, leaving us with 7,674 documents in total. Both datasets have become popular datasets for experiments in many data mining tasks, such as text classification. Each document is associated with one single category label. For 20 Newsgroups, correlation is exhibited across different newsgroups (e.g. rec.sport.baseball and rec.sport.hockey), which makes this dataset a suitable choice to verify the effectiveness of topic correlation discovery for CGTM.

We compare CGTM with three topic modeling methods: LDA [Blei *et al.*, 2003], CTM [Blei and Lafferty, 2006a] and Gaussian-LDA [Das *et al.*, 2015]. CTM replaces the dirichlet prior in LDA with logistic normal distribution to capture the correlation among topic proportions. Gaussian-LDA was first proposed for audio retrieval [Hu *et al.*, 2012] and then used to leverage word embeddings in the continuous vector space [Das *et al.*, 2015].

To learn high quality word embeddings, we combine the current dataset with Wikipedia as the knowledge source. The motivation of using Wikipedia as the supplemental source lies in the sheer range of topics and subjects that are covered and it allows us to enhance the semantics of word embeddings extracted from 20 Newsgroups and Reuters. In the experiment, we set the dimensionality of word embeddings to 100, and the context window size to 12. We train word embeddings for 100 epochs.

We are interested to see if the learned topics can reveal a similar mixture and correlation with the ground truth text categories. Hence we set the number of topics K to the number of categories. For uniformity, all the models are implemented with Gibbs sampling and run for 100 iterations. The Gaussian topic hyper parameter $\boldsymbol{\mu}_0$ is set to the sample mean of all the word vectors, the initial degree of freedom ν_0 to the dimensionality of word embeddings, and $\boldsymbol{\Psi}_0$ to an identity matrix.

[1] www.qwone.com/ jason/20Newsgroups/

[2] www.daviddlewis.com/resources/testcollections/reuters21578/

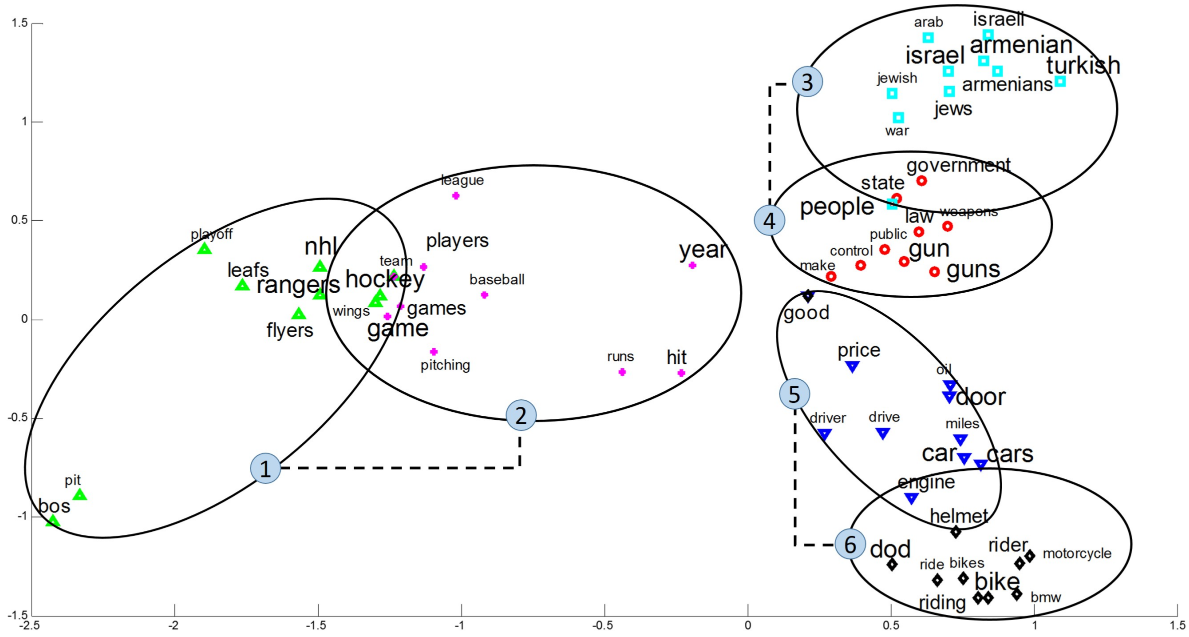

Figure 2: Topic Words and Correlations.

6.1 Topic Words and Correlations

To investigate the quality of topics and the topic correlations discovered by CGTM, we visualize each topic with their top words as well as the topic correlations. To make the visualization clearer, we select only 6 categories from the 20 Newsgroups dataset whose topic words and correlations can be easily recognized and defined. The selected newsgroups are “rec.autos”, “rec.motorcycles”, “rec.sport.baseball”, “rec.sport.hockey”, “talk.politics.guns”, and “talk.politics.mideast”. Thus, in this experiment, we set the number of topics K to 6. As Figure 2 shows, we display top 10 words for each topic discovered by CGTM and map the corresponding word embeddings into a two-dimensional space via Principal Component Analysis (PCA). The size of each word varies with its relative frequency in the corresponding topic. The different colors and shapes of words indicate they are from 6 different topics. Each circle depicts the Gaussian distribution for each topic. The detected topic correlation is represented as a dashed line between topics. As one can observe, all the newsgroups, Hockey (topic 1), Baseball (topic 2), Mideast (topic 3), Guns (topic 4), Autos (topic 5) and Motorcycles (topic 6), are successfully discovered with reasonable topic words.

As the ground truth labels indicate, one can easily figure that Autos is correlated with Motorcycles, Baseball is correlated with Hockey, and Guns is correlated with Mideast. The dashed lines in the figure denote the automatically detected topic correlations by CGTM. With the help of word embeddings and Gaussian topics, topic correlations are also correctly detected, as the dashed lines show. We can see that, since word embeddings can capture the regularities in language such as synonyms, two topics tend to be correlated if their topic word embeddings overlap in the continuous vector space. This demonstrates how the known word-level correlation information can aid us in discovering the topic-level correlations.

In this subsection, we qualitatively exhibit the effectiveness of discovering topics and topic correlations of CGTM. In the following subsections, we will quantitatively evaluate CGTM on topic coherence and topic correlation discovery.

6.2 Topic Coherence

In order to quantitatively assess the topic coherence, we adopt a metric called coherence score of topics proposed by [Mimno *et al.*, 2011] which is able to automatically evaluate the coherence of each discovered topic. Given a topic z and its top T words $V^z = \{v_1^z, v_2^z, ..., v_T^z\}$, the coherence score of this topic is defined as:

$$C(z; V^z) = \sum_{t=2}^{T} \sum_{l=1}^{t} \log \frac{D(v_t^z, v_l^z) + 1}{D(v_l^z)}, \qquad (13)$$

where $D(v_l^z)$ is the document frequency of word v_l^z and $D(v_t^z, v_l^z)$ is the number of documents in which words v_t^z and v_l^z co-occurred. The coherence score follows the intuition that words from the same topic tend to co-occur in documents. This topic coherence score has been proven to be highly consistent with human coherence judgements [Mimno *et al.*, 2011].

The topic coherence result on the 20 Newsgroups dataset is reported in Table 1. In order to investigate the overall quality of all the discovered topics, the average coherence score is reported, which is calculated as $\bar{C} = \frac{1}{K} \sum_z C(z; V^z)$. To make this evaluation more comprehensive, the number of topic words T ranges from 5 to 50. For all the models, the topic words are ordered by word counts in each topic. Though for Gaussian-LDA and CGTM, topic words can also

Top T words	5	10	20	50
LDA	-13.86	-64.11	-322.07	-2384.68
CTM	-13.77	-64.49	-323.71	-2395.58
Gaussian-LDA	-14.83	-66.31	-323.91	-2505.33
CGTM	**-12.37**	**-60.48**	**-317.43**	**-2362.75**

Table 1: Comparison of topic coherence scores.

be ordered by word probabilities under each Gaussian topic, we still order them by word counts, since first, the Gaussian posterior probability information has already been fully utilized in the training phase and second, this coherence score is more appropriate to measure frequent words in a topic. The result shows that the topic words discovered by our model are more coherent than the topic words discovered by the baseline models.

6.3 Document Topics and Topic Correlation

We see that the topics discovered by CGTM qualitatively exhibit good topic words and reasonable correlations, and CGTM also outperforms the baseline models in terms of coherence score. But are the topics discovered by our model really corresponding to the coherent news categories? If yes, it would be very convenient for us to assess the quality of the detected topic correlations, because the correlations among the ground truth newsgroups labels are well defined. For example, 20 Newsgroups categories "rec.autos" and "rec.motorcycles" are clearly correlated, and Reuters categories "money" and "trade" should also exhibit correlations. To answer this question, we compare the ground truth document labels with the document-topic labels discovered by the models to see if they are consistent. The label of each document comes from the dataset and is used as the ground truth. The document-topic label of each document is assigned by the models. More specifically, for each model, we can assign one single topic to document d according to:

$$z_d = argmax_z p(z|d).$$

So this is a clustering evaluation problem where each document is a sample. To solve the cluster matching problem, e.g., ground truth label 1 may correspond to topic 5 instead of topic 1, we adopt pairwise comparison [Menestrina *et al.*, 2010] to measure the consistency between the ground truth document labels and the learned topic representation of documents. The pairwise comparison is defined as:

$$precision(E, G) = \frac{||pair_E \cap pair_G||}{||pair_E||},$$
$$recall(E, G) = \frac{||pair_E \cap pair_G||}{||pair_G||},$$
$$F1(E, G) = \frac{2 \times precision \times recall}{precision + recall},$$

where E and G are two clustering solutions corresponding to the document-topic clusters and the ground truth document labels respectively in our case, and $pair_E$ denotes the set of pairs in clustering result E, and $||pair_E||$ represents the number of instances in $pair_E$. The experimental results of document clustering on 20 Newsgroups and Reuters are reported in Table 2 and Table 3 respectively. We can see that, with respect to the consistency between ground truth document labels and discovered topics, CGTM outperforms the other baselines on both datasets.

Model	Precision	Recall	F1
LDA	0.438	0.507	0.470
CTM	0.447	**0.634**	0.524
Gaussian-LDA	0.438	0.496	0.465
CGTM	**0.523**	0.623	**0.568**

Table 2: Comparison of document-topic distribution on the 20 Newsgroups dataset.

Model	Precision	Recall	F1
LDA	0.844	0.392	0.535
CTM	0.796	**0.433**	0.561
Gaussian-LDA	0.865	0.405	0.552
CGTM	**0.870**	0.431	**0.576**

Table 3: Comparison of document-topic distribution on the Reuters dataset.

7 Conclusions

In this paper, we have proposed a correlated topic model using word embeddings. Word embeddings learnt from large, unstructured corpora, such as Wikipedia, can aid us in modeling topics and topic correlation by bringing in additional useful semantics. The known word-level correlation information in word embeddings is passed to topic-level correlation discovery task via Gaussian topics. In our case, the word embeddings are trained on the combined collections of Wikipedia and the 20 newsgroups dataset. We model each topic as a Gaussian distribution over word embeddings and directly learn topic correlations in the vector space. The experiments qualitatively show CGTM is able to learn meaningful topics and topic correlation, and quantitatively validate the effectiveness of our model in terms of topic coherence score and document clustering on two real-world datasets.

Acknowledgements

This work was supported in part by the US National Science Foundation under grants NSF IIS-1218393, IIS-1514204 and IIS 1319973. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[Bengio *et al.*, 2003] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *journal of machine learning research*, 3(Feb):1137–1155, 2003.

[Blei and Lafferty, 2006a] David Blei and John Lafferty. Correlated topic models. *Advances in neural information processing systems*, 18:147, 2006.

[Blei and Lafferty, 2006b] David M Blei and John D Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120. ACM, 2006.

[Blei *et al.*, 2003] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[Chen *et al.*, 2013] Jianfei Chen, Jun Zhu, Zi Wang, Xun Zheng, and Bo Zhang. Scalable inference for logistic-normal topic models. In *Advances in Neural Information Processing Systems*, pages 2445–2453, 2013.

[Collobert and Weston, 2008] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM, 2008.

[Das *et al.*, 2015] Rajarshi Das, Manzil Zaheer, and Chris Dyer. Gaussian lda for topic models with word embeddings. In *Proceedings of the 53nd Annual Meeting of the Association for Computational Linguistics*, 2015.

[Griffiths, 2002] Tom Griffiths. Gibbs sampling in the generative model of latent dirichlet allocation. 2002.

[Hofmann, 1999] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM, 1999.

[Holmes *et al.*, 2006] Chris C Holmes, Leonhard Held, et al. Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian analysis*, 1(1):145–168, 2006.

[Hu *et al.*, 2012] Pengfei Hu, Wenju Liu, Wei Jiang, and Zhanlei Yang. Latent topic model based on gaussian-lda for audio retrieval. In *Chinese Conference on Pattern Recognition*, pages 556–563. Springer, 2012.

[Huang *et al.*, 2012] Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 873–882. Association for Computational Linguistics, 2012.

[Li *et al.*, 2016] Shaohua Li, Tat-Seng Chua, Jun Zhu, and Chunyan Miao. Generative topic embedding: A continuous representation of documents. In *Proceedings of The 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2016.

[Li *et al.*, 2017] Shaohua Li, Jun Zhu, and Chunyan Miao. Psdvec: A toolbox for incremental and scalable word embedding. *Neurocomputing*, 237:405–409, 2017.

[Liu *et al.*, 2015] Yang Liu, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Topical word embeddings. In *AAAI*, pages 2418–2424, 2015.

[Mei *et al.*, 2008] Qiaozhu Mei, Deng Cai, Duo Zhang, and ChengXiang Zhai. Topic modeling with network regularization. In *Proceedings of the 17th international conference on World Wide Web*, pages 101–110. ACM, 2008.

[Menestrina *et al.*, 2010] David Menestrina, Steven Euijong Whang, and Hector Garcia-Molina. Evaluating entity resolution results. *Proceedings of the VLDB Endowment*, 3(1-2):208–219, 2010.

[Mikolov and Dean, 2013] T Mikolov and J Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 2013.

[Mikolov *et al.*, 2013] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[Mimno *et al.*, 2011] David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 262–272. Association for Computational Linguistics, 2011.

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

[Morin and Bengio, 2005] Frederic Morin and Yoshua Bengio. Hierarchical probabilistic neural network language model. In *Aistats*, volume 5, pages 246–252. Citeseer, 2005.

[Pennington *et al.*, 2014] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–43, 2014.

[Polson *et al.*, 2013] Nicholas G Polson, James G Scott, and Jesse Windle. Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American statistical Association*, 108(504):1339–1349, 2013.

[Zhu *et al.*, 2012] Jun Zhu, Amr Ahmed, and Eric P. Xing. Medlda: maximum margin supervised topic models. *Journal of Machine Learning Research*, 13:2237–2278, 2012.

[Zhu *et al.*, 2014] Jun Zhu, Ning Chen, Hugh Perkins, and Bo Zhang. Gibbs max-margin topic models with data augmentation. *Journal of Machine Learning Research*, 15(1):1073–1110, 2014.