# **Extended Topic Model for Word Dependency**

Tong Wang<sup>1</sup>, Vish Viswanath<sup>2</sup> and Ping Chen<sup>1</sup> <sup>1</sup>University of Massachusetts Boston, Boston, MA <sup>2</sup>Harvard School of Public Health, Boston, MA tongwang0001@gmail.com, Ping.Chen@umb.edu Vish\_Viswanath@dfci.harvard.edu

#### Abstract

Topic Model such as Latent Dirichlet Allocation(LDA) makes assumption that topic assignment of different words are conditionally independent. In this paper, we propose a new model Extended Global Topic Random Field (EGTRF) to model non-linear dependencies between words. Specifically, we parse sentences into dependency trees and represent them as a graph, and assume the topic assignment of a word is influenced by its adjacent words and distance-2 words. Word similarity information learned from large corpus is incorporated to enhance word topic assignment. Parameters are estimated efficiently by variational inference and experimental results on two datasets show EGTRF achieves lower perplexity and higher log predictive probability.

# 1 Introduction

Probabilistic topic model such as Latent Dirichlet Allocation(LDA) (Blei et al, 2003) has been widely used for discovering latent topics from document collections by capturing words' cooccuring relation. However, the "bag of words" assumption is employed in most existing topic models, it assumes the order of words can be ignored and topic assignment of each word is conditionally independent given the topic mixture of a document.

To relax the "bag of words" assumption, many extended topic models have been proposed to address the limitation of conditional independence. Wallach (Wallach, 2006) explores a hierarchical generative probabilistic model that incorporates both n-gram statistics and latent topic variables. Gruber (Gruber et al, 2007) models the topics of words in the document as a Markov chain, and assumes all words in the same sentence are more likely to have the same topic. Zhu (Zhu et al, 2010) incorporates Markov dependency between topic assignments of neighboring words, and employs a general structure of the GLM to define a conditional distribution of latent topic assignments over words. Most of the models above are limited to model linear topical dependencies between words, word topical dependencies can also be modeled by a non-linear way. In Syntactic topic models (Boyd-Graber et al, 2009), each word of a sentence is generated by a distribution that combines document-specific topic weights and parse-tree-specific syntactic transitions.

In Global Topic Random Field(GTRF) model (Li et al, 2014), sentences of a document are parsed into dependency trees (Marneffe et 2008) (Manning et al, 2014) (Marneffe et al. al, 2006). They show topics of semantically or syntactically dependent words achieve the highest similarity and are able to provide more useful information in topic modeling, which is also the basic assumption of our model. Then they propose GTRF to model non-linear topical dependencies, word topics are sampled based on graph structure instead of "bag of words" representation, the conditional independence of word topic assignment is thus relaxed.

However, GTRF assumes topic assignment of a word vertex depends on the topic mixture of the document and its neighboring word vertices, ignoring the fact that word vertex can also be influenced by the distance-2 or further word vertices. In this paper, we extend GTRF model and present a novel model Extended Global Topic Random Field (EGTRF) to exploit topical dependency between words. In EGTRF, the topic assignment of a word is assumed to depend on both distance-1 and distance-2 word vertices. An example of a simple document that has two sentences shows in Figure 1. The two sentences are parsed into dependency trees respectively, and then merged into a graph.



Example document: LDA stands for latent dirichlet allocation. It discovers latent topcis from corpus. Example word vertex: allocation

Distance-1 word vertics: {stands, latent, dirichlet} Distance-2 word vertics: {LDA, topics}

Figure 1: Dependency tree example

Some hidden dependency relations can also be extracted by merging dependency trees. For example, word "allocation" has a new distance-2 word "topics" after merging. Therefore, EGTRF can exploit more semantically or syntactically word dependencies. Theoretically, we can also model the distance further than 2, however, it leads to more complicated computation and small increase of performance.

Another advantage of EGTRF is it incorporates word features. The word vector representations are very interesting because the learned vectors explicitly encode many linguistic regularities and patterns (Mikolov et al, 2013). We use the pretrained model from Google News dataset(about 100 billion words) using word2vec<sup>1</sup> tool to represent each word as a 300-dimensional word vector, and apply normalized word similarity as a confidence score to indicate how possible two word vertices share same topic.

We organized the paper as below: EGTRF is presented in Section 2, variational inference and parameter estimation are derived in Section 3, experiments on two datasets are showed in Section 4, we conclude the paper in Section 5.

# 2 Extended Global Topic Random Field

In this section, we first present Extended Global Random Field(EGRF) in section 2.1, then show how to model topical dependencies using EGRF in section 2.2. We incorporate word similarity information into model in section 2.3.

#### 2.1 Extended Global Random Field

After representing document to undirected graph on previous section, we extend Global Random Field and give the definition of Extended Global Random Field to model the graph as below:

Given an undirected graph G, word vertex set is denoted as  $W = \{w_i | i = 1, 2, ..n\}$ , where  $w_i$  is a word vertex, and n is the number of unique words in a document.  $E_1$  is distance-1 edge set,  $E_1 =$  $\{(w_i, w_j) | \exists path between w_i, w_j that length is 1\}$ .  $E_2$  is distance-2 edge set,  $E_2 =$  $\{(w_i, w_j) | \exists path between w_i, w_j that length is 2\}$ . The state(topic assignment) of a word vertex w is generated from  $Z = \{z_i | i = 1, 2, ..., k\}$ , k is the number of topics.

$$P(G) = f_G(g) = \frac{1}{|E_1| + |E_2|} \prod_{w \in W} f(z_w) \times \left(\sum_{(w'_1, w''_1) \in E_1} f_{(1)}(z_{w'_1}, z_{w''_1}) + \sum_{(w'_2, w''_2) \in E_2} f_{(2)}(z_{w'_2}, z_{w''_2})\right)$$
(1)

s.t. 
$$1.f(z) > 0, f_{(1)}(z', z'') > 0, f_{(2)}(z', z'') > 0$$
  
 $2. \sum_{z \in Z} f(z) = 1$   
 $3. \sum_{z', z'' \in Z} f(z')f(z'')f_{(1)}(z', z'') = 1$   
 $4. \sum_{z', z'' \in Z} f(z')f(z'')f_{(2)}(z', z'') = 1$ 

In Equation (1), f(z) is the function defined on word vertex, which is a probability measure because of the constraints 1 and 2.  $f_{(1)}(z, z')$  and  $f_{(2)}(z, z')$  are the function defined on edge set  $E_1$ and  $E_2$ .  $f_{(1)}$  and  $f_{(2)}$  are not necessarily probability measure, however, summing over all possible states of the product of the edge and the linked word pair should equal to 1, which are from constraints 3 and 4. So  $f(z')f(z'')f_{(1)}(z', z'')$  and  $f(z')f(z'')f_{(2)}(z', z'')$  are probability measure. gis one sample of word topic assignments from graph G. If Equation (1) satisfies all the four constraints, it is easy to verify P(G) is also a probability measure since summing over all possible samples g equals to 1.

We define the random field as in Equation (1) a Extended Global Random Field (EGRF). And EGRF does not have normalization factor, which is much simplier than models with intractable normalizing factor.

#### 2.2 Topic Model Using EGRF

We define Extended Global Topic Random Field based on EGRF. EGTRF is a generative proba-

<sup>&</sup>lt;sup>1</sup>https://code.google.com/p/word2vec/

bilistic model, the basic idea is that documents are represented as mixtures of topics, words are generated depending on the topic mixtures and graph structure of current document. The generative process for word sequence of a document is described as below:

For each document d in corpus D: Transform document d into graph. Choose  $\theta \sim Dir(\alpha)$ . For each of the n words  $w_n$  in d: Choose topic  $z_n \sim P_{egrf}(z \mid \theta)$ , Choose word  $w_n \sim Multi(\beta_{z_n,w_n})$ .

Given Dirichlet prior  $\alpha$ , word distribution of topics  $\beta$ , topic mixture of document  $\theta$ , topic assignments z and words w. We obtain the marginal distribution of a document:

$$p(\mathbf{w} \mid \alpha, \beta) = \int P(\theta \mid \alpha) \sum_{z} P_{egrf}(z \mid \theta) \prod_{n} P(w_n \mid z_{w_n}, \beta) d\theta$$
(2)

We can see the marginal distribution is similar to LDA except topic assignment of word is sampled by Extended Global Random Field instead of Multinomial. So the word topic assignment is no longer conditionally independent. According to EGRF described in section 2.1, we define the probability of topic sequence z as below:

$$P_{egrf}(z \mid \theta) = \frac{1}{\mid E_1 \mid + \mid E_2 \mid} \prod_{w \in V} f(z_w) \times \left(\sum_{(w'_1, w''_1) \in E_1} f_{(1)}(z_{w'_1}, z_{w''_1}) + \sum_{(w'_2, w''_2) \in E_2} f_{(2)}(z_{w'_2}, z_{w''_2})\right)$$
(3)

where 
$$f(z_w) = Multi(z_w|\theta)$$
 (4)

$$f_{(1)}(z_{w_1'}, z_{w_1''}) = \sigma_{z_{w_1'} = z_{w_1''}} \lambda_1 + \sigma_{z_{w_1'} \neq z_{w_1''}} \lambda_2$$
(5)

$$f_{(2)}(z_{w_2'}, z_{w_2''}) = \sigma_{z_{w_2'} = z_{w_2''}} \lambda_3 + \sigma_{z_{w_2'} \neq z_{w_2''}} \lambda_4 \qquad (6)$$

 $\sigma$  is an indicator function and equals 1 if the topic assignments of two words on an edge are same. In order to model Equation (3) as an EGRF, it must satisfy all the four constraints in Equation (1). Equation (4) defines word vertex as multinomial distribution, and we assign  $\lambda_1$ ,  $\lambda_2$ ,  $\lambda_3$  and  $\lambda_4$  nonzero values, then it is clear to verify constraint 1 and 2 are satisfied. To satisfy the constraint 3 and 4, combine with (5), (6), we get the relation between  $\lambda_1$  and  $\lambda_2$ ,  $\lambda_3$  and  $\lambda_4$ .

$$\sum \theta_i^2 \lambda_1 + (1 - \sum \theta_i^2) \lambda_2 = 1 \quad i = 1, 2, .. |E_1|$$
(7)

$$\sum \theta_i^2 \lambda_3 + (1 - \sum \theta_i^2) \lambda_4 = 1 \quad i = 1, 2, .. |E_2|$$
(8)

Lower  $\lambda_2$ ,  $\lambda_4$  give higher reward to the edge that connects two word vertices with same topic. If (7) and (8) hold true, Equation (3) is an EGRF. And we define the topic model based on EGRF as Extended Global Topic Random Field(EGTRF). If  $|E_2| = 0$ ,  $|E_1| \neq 0$ , EGTRF is equivalent to GTRF. If  $|E_1| = 0$ ,  $|E_2| = 0$ , EGTRF is equivalent to LDA.

#### 2.3 Word Similarity Information

The *coherent edge* is the edge that the two linked words have same topic. In distance-i edge set, i = 1, 2.  $E_{C_i}$  includes all coherent edges,  $E_{NC_i}$  contains all non-coherent edges. Then equation (3) can be represented as below:

$$P_{egrf}(z \mid \theta) = \frac{1}{\mid E_{1} \mid + \mid E_{2} \mid} \prod_{w \in V} Multi(z_{w} \mid \theta) \times \\ (\mid E_{C_{1}} \mid \lambda_{1} + \mid E_{NC_{1}} \mid \lambda_{2} + \mid E_{C_{2}} \mid \lambda_{3} + \mid E_{NC_{2}} \mid \lambda_{4}) \\ = \frac{\prod_{w \in V} Multi(z_{w} \mid \theta)}{(\mid E_{1} \mid + \mid E_{2} \mid)\theta^{T}\theta} \times (\mid E_{C_{1}} \mid (1 - \lambda_{2}) + \mid E_{1} \mid \lambda_{2}\theta^{T}\theta + \\ \mid E_{C_{2}} \mid (1 - \lambda_{4}) + \mid E_{2} \mid \lambda_{4}\theta^{T}\theta)$$
(9)

From the second line to the third line of Equation (9), we represent  $\lambda_1, \lambda_3$  as the function of  $\lambda_2, \lambda_4$  based on (7) and (8). The expectation of the number of edges in  $E_{c_i}$  can be computed as:

F

$$E(\mid E_{C_i} \mid) = \sum_{(w_1, w_2) \in E_i} \phi_{w_1}^T \phi_{w_2} S_{w_1, w_2}$$
(10)

 $\phi$  is the K dimensional variational multinomial parameters and can be thought as the posterior probability of a word given the topic assignment.  $S_{w_1,w_2}$  is the similarity measure between word  $w_1$ and  $w_2$ .

As we discussed in section 1, word similarity information  $S_{w_1,w_2}$  works as a confidence score to model how likely two words on an edge have same topic. And we make assumption that two words are more likely to have same topic if they have a higher similarity score. To get the similarity score between words, we use word2vec tool to learn the word representation of each word from pre-trained model. The word representations are computed using neural networks, and the learned representations explicitly encode many linguistic regularities



Figure 2: Experimental results on NIPS(left) and 20 news(right) data

and patterns from the corpus. Normalized similarity between word vectors can be regarded as the confidence score of how possible two words have same topic. In this way, knowledge from large corpus other than current document collections is incorporated to guide topic modeling.

# **3** Posterior Inference and Parameter Estimation

We derive Variational Inference for posterior inference. The variational function q is same to the original LDA paper (Blei et al, 2003). All terms except  $P(z|\theta)$  in likelihood function are also same to LDA, Based on Equation (9), we obtain:

$$E_{q}[\log P_{egrf}(z \mid \theta)] \approx E_{q}[\log(\prod_{n} Multi(z_{w_{n}} \mid \theta))] + \frac{1 - \lambda_{2}}{\zeta_{1}} E_{q}(\mid E_{C_{1}} \mid) + \frac{1 - \lambda_{4}}{\zeta_{1}} E_{q}(\mid E_{C_{2}} \mid) + (11) + \frac{|E_{1} \mid \lambda_{2} + |E_{2} \mid \lambda_{4}}{\zeta_{1}} - \frac{|E_{1} \mid + |E_{2} \mid}{\zeta_{2}}) E_{q}(\theta^{T}\theta) + \log \zeta_{1} - \log \zeta_{2}$$

We get the approximation in Equation (11) from Taylor series, where  $\zeta_1$  and  $\zeta_2$  are Taylor approximation.  $E_q(|E_{C_i}|)$  is obtained directly from (10),  $E_q(\theta^T \theta)$  is from the property of Dirichlet distribution. The updating rule of  $\alpha$  and  $\beta$  are same to LDA,  $\gamma$  is updated using Newton method since we can not obtain the direct updating rule for  $\gamma$ .  $\phi$  can be approximated as:

$$\begin{split} \phi_{w_{n,i}} \propto \beta_{i,v} exp(\Psi(\gamma_i) + \frac{1 - \lambda_2}{\zeta_1} \times \sum_{(w_n, w_m) \in E_1} \phi_{w_{m,i}} S_{m,n} + \\ \frac{1 - \lambda_4}{\zeta_1} \times \sum_{(w_n, w_p) \in E_2} \phi_{w_{p,i}} S_{p,n}) \quad (12) \end{split}$$

EM algorithm is applied using above updating rules. At E-step, we estimate the best  $\gamma$  and  $\phi$ given current  $\alpha$  and  $\beta$ . At M-step, we update new  $\alpha$  and  $\beta$  based on obtained  $\gamma$  and  $\phi$ . We run such iterations until convergence.

# 4 Experiment

In this section we study the empirical performance of EGTRF on two datasets. For each dataset, we remove very short documents, and compute a vocabulary by removing stop words, rare words, frequent words. Eighty percent data are used for training, others for testing.

- 20 News Groups: After processing, it contains 13706 documents with a vocabulary of 5164 terms.
- *NIPS data* (Globerson et al, 2004): Spanning from 2000 to 2005. After processing, it contains 843 documents with a vocabulary of 6098 terms.

We evaluate how well a model fits the data with held-out perplexity (Blei et al, 2003) and predictive distribution (Hoffman et al, 2013). Lower perplexity, higher log predictive probability indicate better generalization performance. We implement GTRF without adding self defined edges from the original paper, and set  $\lambda_2 = 0.2$  to give higher reward to edges from  $E_1$  that the two word vertices have same topic. We set  $\lambda_4 = 1.2$  to give lower(even negative) reward to edges from  $E_2$  that the two word vertices have same topic in EGTRF, since the distance-1 words are expected to have greater topical affects than distance-2 words. Word is represented as vector from pretrained Google News dataset, we use the word vector learned from original corpus when the word does not exist in pre-trained Google News dataset.

We choose 10, 20, 30, 50 topics for 20 news dataset, 10, 15, 20, 25 topics for NIPS dataset. Figure 2 shows the experimental results of four models: Ida, gtrf, egtrf(EGTRF without word similarity information), and egtrf+s(EGTRF with word similarity information) on two datasets. The results show EGTRF outperforms LDA and GTRF in general, and EGTRF with word similarity information achieves best performance.

We believe modeling distance-2 word vertices can exploit more semantically or syntactically word dependencies from document, and word similarity information obtained from large corpus can make up the lack of sufficient information from the original corpus. Therefore, adding the influence of distance-2 word vertices and word similarity information can improve performance of topic modeling.

# 5 Conclusion

In this paper, we extended Global Topic Random Field(GTRF) and proposed a novel topic model Extended Global Topic Random Field(EGTRF) which can model dependency relation between adjacent words and distance-2 words. Word topics are drawed by Extended Global Random Field(EGRF) instead of Multinomial, the conditional independence of word topic assignment is thus relaxed. Word similarity information learned from large corpus is incorporated into the model. Experiments on two datasets show EGTRF achieves better performance than GTRF and LDA, which confirm our assumption that adding topical dependency of distance-2 words and incorporating word similarity information can improve model performance.

#### References

- Amir Globerson, Gal Chechik, Fernando Pereira, Naftali Tishby Euclidean Embedding of Co-occurrence Data. In *Advances in neural information processing systems*. pp. 497-504. 2004.
- Amit Gruber, Michal Rosen-Zvi and Yair Wei. Hidden Topic Markov Models. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistic.* pp. 163-170. 2007.

- David Blei, Andrew Ng., and Michael Jordan. Latent Dirichlet Allocation. *The Journal of Machine Learning Research.* 3:993-1022, 2003.
- Hanna M Wallach. Topic modeling: Beyond bag-ofwords. In *International Conference on Machine Learning*. pp. 977-984. ACM, 2006.
- Jordan Boyd-Graber and David Blei. Syntactic topic models. In *Neural Information Processing Systems*. pp. 185-192. 2009.
- Jun Zhu and Eric P. Xing. Conditional Topic Random Fields. In *Proceedings of the 27th International Conference on Machine Learning*. 2010.
- Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard and David Mc-Closky. The Stanford CoreNLP Natural Language Processing Toolkit. In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. pp. 55-60. 2014.
- Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*. Vol. 6, No. 2006, pp. 449-454. 2006.
- Marie-Catherine de Marneffe, Christopher D. Manning. The Stanford typed dependencies representation. In *COLING 2008 Workshop on Crossframework and Cross-domain Parser Evaluation*. pp. 1-8. 2008.
- Matthew Hoffman, David Blei, Chong Wang, John Paisley Stochastic Variational Inference *The Journal of Machine Learning Research*. 14(1), 1303-1347. 2013.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. In *Proceedings of Workshop at ICLR*. arXiv:1301.3781, 2013.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. In *Proceedings of NIPS*. pp. 3111-3119. 2013.
- Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of NAACL HLT*. pp. 746-751, 2013.
- Zhixing Li, Siqiang Wen, Juanzi Li, Peng Zhang and Jie Tang. On Modeling Non-linear Topical Dependencies. In Proceedings of the 31th International Conference on Machine Learning. pp. 458-466, 2014.