Using Word Sense as a Latent Variable in LDA Can Improve Topic Modeling

Yunqing Xia[1], Guoyu Tang[1], Huan Zhao[1], Erik Cambria[2] and Thomas Fang Zheng[1]
[1]*Department of Computer Science and Technology, Tsinghua University, Beijing, China*
[2]*Temasek Laboratories, National University of Singapore, Singapore, Singapore*

Keywords: Topic Modeling, LDA, Latent Variable, Word Sense.

Abstract: Since proposed, LDA have been successfully used in modeling text documents. So far, words are the common features to induce latent topic, which are later used in document representation. Observation on documents indicates that the polysemous words can make the latent topics less discriminative, resulting in less accurate document representation. We thus argue that the semantically deterministic word senses can improve quality of the latent topics. In this work, we proposes a series of word sense aware LDA models which use word sense as an extra latent variable in topic induction. Preliminary experiments on document clustering on benchmark datasets show that word sense can indeed improve topic modeling.

1 INTRODUCTION

In the past decade, Latent Dirichlet Allocation (LDA) (Blei et al., 2003) has been proved an effective topic model for information retrieval (Dietz et al., 2007; Wang et al., 2007). So far, words are the common features to induce latent topic, which are later used in document representation. Observation on documents indicates that the polysemous words can make the latent topics less discriminative, resulting in less accurate document representation. For example, we all know that *apple* refers a kind of fruit in some cases but a computer company in other contexts. In the latent topics induced by LDA, the word feature "apple" is assigned to different topics with different probability. It is unknown which word sense plays a key role in these assignment: the fruit or the company? Our intuition is that topic models with ambiguous words can be less precise than that with word senses.

An empirical study has been conducted to confirm this intuition. With the word-topic probability matrix produced by LDA, we calculate average probability within the top N topics (*avgpr@N*) as follows. For each word, the topics that a word w is associated are ranked according to the probability $p(z|w)$[1]. *argpr@N* is calculated by averaging the probabilities $p(z|w)$ of all words on the top N topics. For each word sense, we calculate *argpr@N* based on $p(z|s)$. We use the senses in the last iteration of SLDA models(e.g. 2200) and the topics inferred by these word senses. We run the classic word based LDA and a word sense aware LDA (SLDA) proposed in this work on the same dataset (i.e., *Reuters*) and calculated the *avgpr@N* values with different N. The curve are presented in Figure 1.

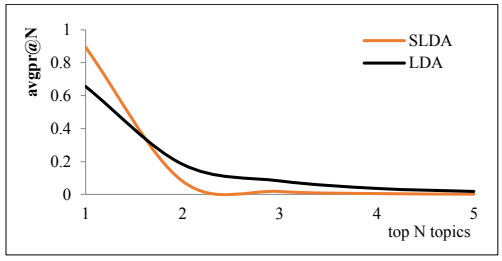

Figure 1: Word-topic distribution in LDA vs. sense-topic distribution in SLDA on *Reuters* dataset.

From Figure 1, we find the curve for sense-topic distribution is sharper than that for word-topic distribution. This indicates that word senses are more discriminative than words. This confirms with us that word sense can improve the LDA topic model. We argue that the semantically deterministic word senses can improve quality of the latent topics.

In this work, we proposes two word sense aware

[1] $p(z|w)$ can be calculated with $p(z|w) \propto p(w|z)\Sigma p(z|d)p(d)$ where $p(w|z)$ and $p(z|d)$ are parameters of the model thus can be estimated while we estimate $p(d)$ to be the proportion of ds document length to the length of the entire document collection .

Xia Y., Tang G., Zhao H., Cambria E. and Fang Zheng T..
Using Word Sense as a Latent Variable in LDA Can Improve Topic Modeling.
DOI: 10.5220/0004889705320537
In *Proceedings of the 6th International Conference on Agents and Artificial Intelligence* (ICAART-2014), pages 532-537
ISBN: 978-989-758-015-4

LDA models which use word sense as an extra latent variable in topic induction. The first SLDA model is standalone SLDA (SA-SLDA), in which word senses are first induced from a development dataset and then replace words in LDA. The second SLDA model is collaborative SLDA (CO-SLDA), in which the topic assigned to a word has a positive feedback on word sense induction. Preliminary experiments on document clustering on benchmark datasets show that word sense can indeed improve topic modeling.

The reminder of this paper is organized as follows. In Section 2, we elaborate the word sense aware LDA models. In Section 3, we present the experiments and discussions. We summarize related work in Section 4, and conclude this paper in Section 5.

2 WORD SENSE AWARE LDA MODELS

The classic LDA assigns each word in the document a topic and consider the surface words as the basic granularity for a document. Alternatively, our model emits a sense for each surface word and assigns each sense a topic. Therefore, the basic granularity for our model is the word sense. To address this motivation, we introduce a latent variable of word sense and induce it from the observed surface words. We design two approaches to implement this purpose as follows:

- Standalone SLDA (SA-SLDA): We isolate the Word Sense Induction (WSI) process as a standalone step. With the induced word sense in hand, we perform the word sense based LDA for document clustering.
- Collaborative SLDA (PCo-SLDA): We identify the generative story as two iteratively interchangeable steps. Given an observed topic, we generate the word sense from the topic. Given an observed word sense, we generate the topic for each word sense, where the word sense is a point estimate from the mode of the distribution.

We describe all our models in two perspectives. First, we perform word sense induction (WSI) on each word. Second, documents are represented (DR) as a collection of word senses, which are then used to infer topics. For presentation convenience, we first brief the classic LDA model (Blei et al., 2003).

2.1 The LDA Model

Given D documents and W word types, the generative story for LDA with Z topics is as follows:

1. For each topic z:
   (a) choose $\phi_z \sim Dir(\beta)$.
2. For each document d_i:
   (a) choose $\theta_{d_i} \sim Dir(\alpha)$.
   (b) for each word w_{ij} in document d_i:
      i. choose topic $z_{ij} \sim Mult(\theta_{d_i})$.
      ii. choose word $w_{ij} \sim Mult(\phi_{z_{ij}})$.

where d_i refers to i-th document in the corpus; w_{ij} refers to j-th word in document d_i; z_{ij} refers to the topic that word w_{ij} is assigned; α and β are hyperparameters of the model; $\phi_{z_{ij}}$ and θ_{d_i} are per topic word distributions and per document topic distributions respectively which are drawn from Dirichlet distributions.

For inference, Collapsed Gibb Sampling (Griffiths and Steyvers, 2004) is widely used to estimate the posterior distribution for latent variables in LDA. In this procedure, the distribution of a topic for the word $w_{ij} = w$ based on values of other data is computed as follows.

$$P(z_{ij} = z|\mathbf{z_{-ij}}, \mathbf{w}) \propto \frac{n^{d_i}_{-ij,z} + \alpha}{n^{d_i}_{-ij} + Z\alpha} \times \frac{n^{w}_{-ij,z} + \beta}{n_{-ij,z} + W\beta} \quad (1)$$

In Equation 1 $n^{d_i}_{-ij,z}$ is the number of words that are assigned topic z in document d_i; $n^{w}_{-ij,z}$ is the number of words ($= w$) that are assigned topic z; $n^{d_i}_{-ij}$ is the total number of words in document d_i and $n_{-ij,z}$ is the total number of words assigned topic z. $-ij$ in all the above variables refers to excluding the count for word w_{ij}.

2.2 SA-SLDA

In the SA-SLDA model, WSI and DR are considered as standalone modules, where DR takes the output (i.e., word senses) of WSI as input (see Figure 2).

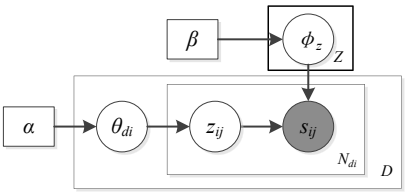

Figure 2: Illustration of the SA-SLDA model.

2.2.1 WSI with HDP

In SA-SLDA, we follow (Yao and Van Durme, 2011) to employ Hierarchical Dirichlet Processes (HDP) (Teh et al., 2004) for word sense induction. HDP

known as a nonparametric Bayesian method is often considered to be advantageous over the parametric methods like LDA, because LDA requires an external input to specify the number of topics while HDP does not. In our case, numbers of word senses differ amongst different words. Therefore, we favor to employ HDP for WSI so that we can equip each word with different number of senses according to their contexts.

We perform HDP on each word. In this paper, we define a word on which the WSI algorithm is performed as a target word and words in the context of a target word as context words of the target word. For each context v_{ij} of the target word w, the sense s_{ij} for each word c_{ij} in v_{ij} has a nonparametric prior G_{ij} which is sampled from a base distribution G_w. H is a Dirichlet distribution with hyper-parameter ε . The context word distribution η_s given a sense s is generated from H:$\eta_s \sim H$.

Then we simply take mode sense in the sense distribution as the sense of the target word.

2.2.2 DR with Word Senses

As shown in Figure 2, we replace word with its word sense in the gray plate. Then the formal procedure of document representation in SA-SLDA is given as follows:

1. For each topic z:
   (a) choose $\phi_z \sim Dir(\beta)$.
2. For each document d_i:
   (a) choose $\theta_{d_i} \sim Dir(\alpha)$.
   (b) for each word w_{ij} in document d_i:
      i. choose topic $z_{ij} \sim Mult(\theta_{d_i})$.
      ii. choose sense $s_{ij} \sim Mult(\phi_{z_{ij}})$.

Similar to LDA, we also use Collapse Gibbs Sampling (Griffiths and Steyvers, 2004) to do inference for SA-SLDA. In SA-SLDA, we replace the surface words with the induced word senses. Therefore, the topic inference is similar to the classic LDA, where the condition probability $P(z_{ij} = z|\boldsymbol{z_{-ij}}, \boldsymbol{s})$ is evaluated by

$$P(z_{ij} = z|\boldsymbol{z_{-ij}}, \boldsymbol{s}) \propto \frac{n^{d_i}_{-ij,z} + \alpha}{n^{d_i}_{-ij} + Z\alpha} \times \frac{n^{s}_{-ij,z} + \beta}{n_{-ij,z} + S\beta} \quad (2)$$

In Equation.2, $n^{s}_{-ij,z}$ is the number of senses with sense s that are assigned topic z , excluding the sense of the j-th word; S is the number of senses for the data set.

2.3 CO-SLDA

Alternatively, word senses can be incorporated into LDA in a collaborative manner. We are interested in whether the topic assigned to a word has a positive feedback on WSI, which then can be used to refine the topic distribution. Inspired by this motivation, we propose a Collaborative SLDA model which takes the topics of senses from SLDA as the pseudo feedback for WSI and iteratively infers both topics and word senses. Specifically, we achieve a point estimate for the target word in WSI and feed this estimated sense to DR.

In this model, a three-level HDP algorithm is used to capture the relationship between word senses and topics of a target word w (see Figure 3. In the three-level HDP, for each word type w, we choose for each topic a probability measure G_{wz} which is drawn from Dirichlet Process $DP(\rho_w, G_w)$. For each word w_{ij} in document d_i, given topic $z_{ij} = z$, we use G_{wz} as the base probability measure for the context of w_{ij} and draws its own G_{ij} from Dirichlet process $G_{ij} \sim DP(\kappa_{wz}, G_{wz})$. This means that word w may have different sense distributions in different topics.

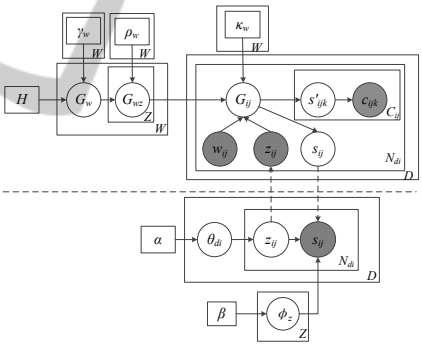

Figure 3: Illustration of the SA-SLDA model.

We show the graphical presentation for CO-SLDA in Figure 3. C_{ij} refers to the number of words in the context window v_j for word w_{ij} in document d_i. The above dotted line shows the WSI process while the below shows the DR process. Given observed topics $\{z_{ij}\}$, word senses $\{s_{ij}\}$ are inferred in WSI. Given observed senses $\{s_{ij}\}$, topics $\{z_{ij}\}$ are inferred in DR. The two processes are interchangeably performed. We provide the dashed arrows in Figure 4 to connect $\{s_{ij}\}$ and $\{z_{ij}\}$ that will change from hidden to observable during the alternation of two processes.

The word sense induction process is as follows:

1. For each word type w:
   (a) choose $G_w \sim DP(\gamma_w, H)$.
2. For each topic z:

(a) choose $G_{wz} \sim DP(\rho_w, G_w)$.

3. For each document d_i:

(a) each context v_j of word w_{ij}:

i. choose $G_{ij} \sim DP(\kappa_{wz}, G_{wz})$.

(b) For each context word c_k of target word w_{ij}:

i. choose $s'_{ijk} \sim G_{ij}$.

ii. choose $c_{ijk} \sim Mult(\eta_{s_{ijk}})$.

iii. set $s_{ij} = \arg\max_s P(s_{ij}|G_{ij})$.

The document representation process is the same as SA-SLDA. Intuitively, the CO-SLDA model is advantageous over the SA-SLDA model, because one word may carry different senses in different topics while in the same topic occurrences of one word often refer to the same sense.

For inference, we interchangeably infer two groups of hidden variables in CO-SLDA:

1. Given that the topic for each word sense z_{ij} is observed, we infer the sense distribution G_{ij} in the context window around a target word. This is achieved through the same scheme as (Teh et al., 2004). Then we estimate s_{ij} for the target word as sense with the highest probability in G_{ij}.
2. Given that the word sense z_{ij} is observed, we infer the topic z_{ij} for each word sense. This can be achieved using the same inference scheme as SA-SLDA.

3 EVALUATION

We use document clustering task to evaluate the accuracy of topics in our model.

3.1 Setup

Data Set

Three data sets used in our experiments are extracted from the following two corpora.

1. **TDT4**: Following (Kong and Graff, 2005), we use the English documents from TDT2002 and TDT2003, i.e., TDT41 and TDT42.
2. **Reuters**: Documents are extracted from Reuters-21578 (Lewis, 1997) with the most frequent 20 categories, i.e., Reuters20.

System Parameters

All hyper-parameters are tuned in the TDT42 dataset and the same ones are applied on the other two datasets as well. In all experiments, we let the Gibbs sampler burn in for 2000 iterations and subsequently take samples 20 iterations apart for another 200 iterations.

As we isolate the WSI process from the document representation process in SA-SLDA, we present the parameters accordingly. (1) In the WSI step, we set the HDP hyper-parameters γ_w, ρ_w, ε for every word type to be $\gamma_w \sim$ *Gamma* (1, 0.001), $\rho_w \sim$ *Gamma* (0.01, 0.028), ε=0.1; (2) In the Document representation step, we set α=1.5 and β=0.1. The topic number is set as cluster number in each dataset. In system CO-SLDA, (1) in the WSI step we set the hyper-parameters γ_w, ρ_w, ε for every word type to be $\gamma_w \sim$ *Gamma* (8,0.1), $\rho_w \sim$ *Gamma* (5,1), $\kappa_w \sim$ *Gamma* (0.1,0.028), ε=0.1; (2) in the DR step, we set α=1.5 and β=0.1. In LDA, We set α=1.5, β=0.1. The topic number is set to be equal to the cluster number in each data set.

Evaluation Metrics

In the experiments, we intend to evaluate the proposed topic models in document clustering task. Each topic in the test dataset is considered as a cluster and each document is clustered into the topic with the highest probability. We adopt the evaluation criteria proposed by (Steinbach et al., 2000). The calculation starts from maximum F-measure of each cluster. The general F-measure of a system is the micro-average of all the F-measures of the system-generated clusters.

3.2 Results and Discussions

Experimental results are presented in Table 1.

Table 1: F-measure values of the proposed models and the baseline LDA model.

Medel	TDT41	TDT42	Reuters20
LDA	0.744	0.867	0.496
SA-SLDA	0.792	0.870	0.512
CO-SLDA	0.825	0.874	0.597

According to the experimental results, we make two important observations:

Firstly, SA-SLDA outperforms the LDA baseline in all cases. This indicates that using word senses rather than surface words improves the document clustering results. The improvement ascribes to that words in LDA are viewed as independent and isolated strings, while in SA-SLDA they are facilitated with more information of word sense according to the context.

Secondly, CO-SLDA outperforms SA-SLDA in all data sets. This indicates that the joint inference process for topics of words and word senses makes a positive impact for each other. Two reasons are worthy of noting: (1) In common sense, instances of the

same word type in different topics often have different senses while instances in the same topic often refer to the same thing. Since CO-SLDA can jointly infer topics and word senses, instances of the same word in the same topic are more likely to be assigned the same sense while instances in different topics are likely to be assigned differently. As a result, word senses will be better identified. (2) Using topics as a pseudo feedback will facilitate the target words with topic-specific senses. For example, the word *election* only has one sense in general cases. However, in the TDT42 data set, topics are labeled in a more fine-grained perspective. For example, the following two sentences are labeled to be from two different topics as the countries of elections are different: *Ilyescu Wins Romanian Elections, Ghana Gets New Democratically Elected President*. With the joint inference of topic and sense, we can induce the word *election* with two senses, i.e., *election#1* and *election#2*, related to the electing process in Romania and Ghana respectively. By incorporating these topic-specific senses, *election* with context word *Romania* is identified as *election#1* and more likely to be assigned topic z_1 while election with context word *Ghana* is identified as *election#2* and more likely to be assigned z_2.

4 RELATED WORK

In Vector Space Model (VSM), it is assumed that terms are independent of each other and the semantic relations between terms are ignored.

Recently, models are proposed to represent documents in a semantic concept space using lexical ontologies, i.e. WordNet or Wikipedia (Hotho et al., 2003; Gabrilovich and Markovitch, 2007; Huang and Kuo, 2010). However, the lexical ontologies are difficult to be constructed and their coverage can be limited. In contrast, topic models are used as an alternative for discovering latent semantic space in corpora based on the per topic word distribution. LDA (Blei et al., 2003) as a classic topic model identifies topics of documents by evaluating word co-occurrences. Various topic models based on the LDA framework have been developed (Wang et al., 2007). However, those models all employ the surface word as the basic unit for document, which is lack of the word sense interpretation for topics. Some work attempt to integrate word semantics from lexical resources into topic models (Boyd-Graber et al., 2007; Chemudugunta et al., 2008; Guo and Diab, 2011). Alternatively, our models are fully unsupervised and do not rely on any external semantic resources, which will be extremely applicable for resource poor languages and domains.

5 CONCLUSIONS

In this paper, we propose to represent topics with distributions over word senses. In order to achieve this purpose in a fully unsupervised manner without relying on any external resources, we model the word sense as a latent variable and induced it from corpora via WSI. We design several models for this purpose. Empirical results verify that the word senses induced from corpora can facilitate the LDA model in document clustering. Specifically, we find the joint inference model (i.e., CO-SLDA) outperforms the standalone model (SA-SLDA) as they the estimation of sense and topic can be collaboratively improved.

In future, we will extend the proposed topic models for the cross-lingual information retrieval tasks. We believe that word senses induced from multilingual documents will be helpful in cross-lingual topic modeling.

ACKNOWLEDGEMENTS

This work is supported by NSFC (61272233). We thank the reviewers for the valuable comments.

REFERENCES

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Boyd-Graber, J. L., Blei, D. M., and Zhu, X. (2007). A topic model for word sense disambiguation. In *EMNLP-CoNLL*, pages 1024–1033. ACL.

Chemudugunta, C., Smyth, P., and Steyvers, M. (2008). Combining concept hierarchies and statistical topic models. In *Proceedings of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 1469–1470, New York, NY, USA. ACM.

Dietz, L., Bickel, S., and Scheffer, T. (2007). Unsupervised prediction of citation influences. In *In Proceedings of the 24th International Conference on Machine Learning*, pages 233–240.

Gabrilovich, E. and Markovitch, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th international joint conference on Artifical intelligence*, IJCAI'07, pages 1606–1611, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235.

Guo, W. and Diab, M. (2011). Semantic topic models: combining word distributional statistics and dictionary definitions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*,

EMNLP '11, pages 552–561, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hotho, A., Staab, S., and Stumme, G. (2003). Wordnet improves text document clustering. In *In Proc. of the SIGIR 2003 Semantic Web Workshop*, pages 541–544.

Huang, H.-H. and Kuo, Y.-H. (2010). Cross-lingual document representation and semantic similarity measure: a fuzzy set and rough set based approach. *Trans. Fuz Sys.*, 18(6):1098–1111.

Kong, J. and Graff, D. (2005). Tdt4 multilingual broadcast news speech corpus. *Linguistic Data Consortium, http://www. ldc. upenn. edu/Catalog/CatalogEntry. jsp*.

Lewis, D. D. (1997). Reuters-21578 text categorization test collection, distribution 1.0. *http://www. research. att. com/˜ lewis/reuters21578. html*.

Steinbach, M., Karypis, G., and Kumar, V. (2000). A comparison of document clustering techniques. In *In KDD Workshop on Text Mining*.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2004). Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101.

Wang, X., McCallum, A., and Wei, X. (2007). Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, ICDM '07, pages 697–702, Washington, DC, USA. IEEE Computer Society.

Yao, X. and Van Durme, B. (2011). Nonparametric bayesian word sense induction. In *Proceedings of TextGraphs-6: Graph-based Methods for Natural Language Processing*, pages 10–14. Association for Computational Linguistics.