Supervised Deep Polylingual Topic Modeling for Scholarly Information Recommendations

Pannawit Samatthiyadikun[1] and Atsuhiro Takasu[1,2]

[1]*Department of Informatics, The Graduate University for Advanced Studies (SOKENDAI), Kanagawa, Japan*

[2]*National Institute of Informatics (NII), Tokyo, Japan*

Keywords: Deep Generative Topic Model, Recommender Systems, Multiple Information Sources.

Abstract: Polylingual text processing is important for content-based and hybrid recommender systems. It helps recommender systems extract content information from broader sources. It also enables systems to recommend items in a user's native language. We propose a cross-lingual keyword recommendation method based on a polylingual topic model. The model is further extended with a popular deep learning architecture, the CNN–RNN model. With this model, keywords can be recommended from text written in different languages; model parameters are very meaningful, and we can interpret them. We evaluate the proposed method using cross-lingual bibliographic databases that contain both English and Japanese abstracts and keywords.

1 INTRODUCTION

Recommender systems are important for e-commerce, as they can improve customers' satisfaction and increase company revenues. The systems encode items' characteristics into a representation vector. They then recommend strongly relevant items to users.

The feature vector is usually high-dimensional and sparse, which results in ineffective and costly recommendations. There are many ways to solve the dimensionality reduction problem. Topic models are a statistical method for discovering clusters from a collection of specific objects, where the characteristics are represented as the mixture of classes in proportion.

One well-known topic model is latent Dirichlet allocation (LDA) (Blei et al., 2003). This model reveals clusters, called topics, from a collection of documents based on the frequency of word occurrences. It extracts a probabilistic relation between topic and word, so that a document becomes a mixture of topics based on the words appearing in the document.

Many models extended from LDA have been introduced. The LDA–dual model is used for solving the entity resolution problem, where two types of information—a document's content and its authors—are used to categorize it into a mixture of topics. The model relies on the co-occurrence of authors and words appearing in each document across the collection (Shu et al., 2009).

Another more interesting application is topic modeling for polylingual documents, where additional polylingual information is required. By using more information in the learning model, more meaningful topics can be found. It also leads to polylingual keyword recommendation, where keywords and document content are used for modeling. This approach is very useful for the foreigner who wants to publish a document in a nonnative language, as it helps authors to choose suitable keywords when writing papers in a nonnative language (Takasu, 2010).

Scholarly information usually consists of multiple entities such as titles, abstracts, and authors. When developing topic models for scholarly information, the model should exploit these multiple types of entities. One way to handle those entities is to merge words, i.e., words in titles and abstracts and authors' names, into one vocabulary. Then, each instance of scholarly information is represented as a bag-of-words, and a topic model such as LDA can be applied. However, the role of a word in a title and abstract may be different. For example, the word "method" in an abstract may have a more general meaning than in a title, where it would mean that the article may propose some specific method for theoretical analysis. The LDA–dual model can handle two types of information. We develop a model that can deal with multiple types of information in this paper.

Many topic models are learned in an unsupervised

Samatthiyadikun, P. and Takasu, A.
Supervised Deep Polylingual Topic Modeling for Scholarly Information Recommendations.
DOI: 10.5220/0006654901960201
In *Proceedings of the 7th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2018)*, pages 196-201
ISBN: 978-989-758-276-9

manner. However, we can use small amounts of training data for recommender systems. In this paper we develop a supervised topic model. We introduce the deep learning technique into our polylingual topic model, while the parameters can still be probabilistically inferred.

2 BACKGROUND

2.1 Probabilistic Generative Topic Model

Generative models are more powerful than discriminative models as they can move beyond associations between inputs and outputs. They can recognize and represent hidden structures in the data and invariants: for instance, the concepts of rotation, light intensity, brightness, or shape in 3-dimensional objects. They can image the world "as it could be" rather than as "as it is presented," and concepts useful for decision-making and reasoning can be established. Lastly, they can detect surprising events in the world and can "plan the future" by generating plausible scenarios.

LDA (Blei et al., 2003) is the best-known, most popular, and simplest topic model. In addition to the brief introduction in Section 1, its generative process is as follows:

1. Draw a per topic-word distribution $\vec{\beta}_k \sim \mathrm{Dir}(\vec{\eta})$, for each topic $k \in \mathbb{K}$
2. For each document $d \in D$:
   (a) Draw a per document-topic distribution $\vec{\theta}_d \sim \mathrm{Dir}(\vec{\alpha})$
   (b) For each word position $i^{th} \in \{1, \ldots, |\vec{x}_d|\}$:
      i. Draw a topic $z_{di} = k \sim \mathrm{Multi}\left(\vec{\theta}_d\right)$
      ii. Draw a word $x_{di} \sim \mathrm{Multi}\left(\vec{\beta}_k\right)$ corresponding to the drawn topic $z_{di} = k$

LDA has been further developed into online learning for dealing with huge datasets by splitting the dataset into small chunks; then, the modeling parameters learned using the variational Bayesian technique from each chunk are combined (Hoffman et al., 2010).

2.2 Handling Parallel Corpus

For a polylingual document, each language of document should be treated separately, while sharing the same topic space. Thus, the only major difference from LDA is that there are parallel contents in a document, i.e., polylingual contents with similar meanings. Because of this characteristic, a per document–topic distribution $\left(\vec{\theta}_x\right)$ is shared among the parallel contents. The generative process must generate a per topic–word distribution $\left(\vec{\beta}_k^{(s)}\right)$ for each language $s \in \mathbb{S}$ with the Dirichlet distribution parameter $\vec{\eta}^{(l)}$ as in process step (1). Another part that must be changed is the one related to the parallel contents, step (2b). It loops through all word positions $\left(1, \ldots, |\vec{x}_d^{(m)}|\right)$ for each $m \in M$ content language (Krstovski and Smith, 2013)

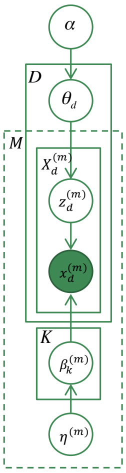

Figure 1: Probabilistic graphical model of polylingual topic model (PLTM), where the only difference from LDA is the M plate for each language.

2.3 Incorporating with Multi-label

To model an annotated document, labeled LDA (L–LDA) assumes that those annotations provide a document concept. That is, a set of document topics corresponds to the set of observed annotations in a one-to-one relation, which is different from the content that is generated from a document's topics (Ramage et al., 2009). Online learning with Gibbs sampling and variational Bayes are described in (Zhou et al., 2015; Jaradat et al., 2015), respectively.

However, L–LDA does not assume the existence of any latent topics (neither global nor within a label): only the documents' distributions over their observed labels, as well as those labels' distributions over words, are inferred. As a result, L–LDA does not support latent subtopics within a given label nor any global latent topics. In this sense, L–LDA is not so latent. Partially labeled Dirichlet allocation (PLDA) extends L–LDA by incorporating classes of latent topics, and providing per-label latent topics (Ramage et al., 2011).

2.4 Deep Generative Model (DGM)

Recently, deep neural networks (DNNs) have become very popular and widely used in various fields of machine learning, although they were first introduced in the 1980s. They can adapt to any problem and achieve excellent results compared with other approaches, but suffer from high computation costs and the learned parameters are hard to interpret. Thanks to huge improvements in computation technology, amounts of data populated, and mathematical knowledge, these models take less time to be optimized and can be used in practical applications.

The most common deep neural network-based generative models are the variational autoencoder (VAE), and generative adversarial network (GAN). To train these models, we rely on Bayesian deep learning, variational approximations, and Monte Carlo Markov Chain (MCMC) estimation, or the old faithful: stochastic gradient estimation (SGD).

The DNN version of LDA has been proposed and tested using VAE (LDA–VAE) as the inference method with a special technique for modeling Dirichlet beliefs. Surprisingly, it achieves more meaningful topics and takes less time than the usual LDA, but has many hyperparameters to be determined when constructing the model (Srivastava and Sutton, 2017). One way to infer stick-breaking construction for VAE (SB–VAE) has been presented and tested in an image classification task. The results illustrate that this approach has greater discriminative quality than the usual VAE and is also supported by t-SNE projections (Nalisnick and Smyth, 2017).

Although DNNs are the best at most supervised problems, they cannot directly output lists. One way to provide label recommendations is to treat the problem as a multilabel classification. The main approach to using DGN s is to embed items and their related labels separately, followed by learning the joint representation for the multilabel classification process. This approach has been tested using image data with multilabels, where images and labels are embedded via a convolution neural network (CNN) and a recurrent neural network (RNN) respectively. The network is called CNN–RNN for this reason. Labels are recommended via predicted probabilities. The experimental results indicate that this approach outperforms the competitors, including a previous DGN approach (Wang et al., 2016).

In our problem, one network may be used for document embedding, like CNN, and another network embeds the keywords corresponding to each document. We will explain how we can define the network for our problem in the next section.

3 PROPOSED MODEL

We now build an improved model for polylingual data. As we have seen from the LDA–VAE model, VAE has a strong emphasis on improving topical quality compared with the normal LDA model. We expect a similar emphasis by applying VAE to PLTM, but the main challenge is how to apply this inference method to a polylingual document.

Topic models for polylingual data like PLDA are usually more complicated and difficult to learn than those for unilingual data, because polylingual documents can be considered as data with multiple sources, where each source provides documents or even parts of documents in a different language. Moreover, there are relationships among those sources to be modeled. The model for multiple data sources is called a "multimodal model." Fortunately, there is a variational autoencoder for multimodal learning that involves relating information from multiple sources, which can model joint relationships among them. It is called a joint multimodal variational autoencoder (JMVAE) (Suzuki et al., 2016), where a general evidence lower bound (ELBO) of the model is written as:

$$\mathcal{L}_{\text{JMVAE}} = \underbrace{-D_{KL}\left(q_{\Phi}\left(\mathbf{z}|\{\mathbf{x}^{(s)}\}_{s\in\mathbb{S}}\right)||p\left(\mathbf{z}\right)\right)}_{\text{Regularization term}} + \underbrace{\sum_{s\in\mathbb{S}}\mathbb{E}_{q_{\Phi}\left(\mathbf{z}|\{\mathbf{x}^{(s')}\}_{s'\in\mathbb{S}}\right)}\left[\log p_{\Theta_{\mathbf{x}^{(s)}}}\left(\mathbf{x}^{(s)}|\mathbf{z}\right)\right]}_{\text{Expected reconstruction error}}. \tag{1}$$

where $\Theta_{\mathbf{x}^{(s)}}$ is a set of model parameters relating to observed data from source s, $(\mathbf{x}^{(s)})$, and Φ is a set of variational distribution parameters. The only difference from the original VAE is the summation over all data sources $\mathbb{S}$.

We can simply apply JMVAE for polylingual documents as used in LDA–VAE (Srivastava and Sutton, 2017) and write the ELBO function as:

$$\begin{aligned} &D_{KL}\left(q_{\Phi}\left(\mathbf{z}|\{\mathbf{x}^{(s)}\}_{s\in\mathbb{S}}\right)||p\left(\mathbf{z}\right)\right) \\ &= \frac{1}{2}\sum_{s\in\mathbb{S}}\left\{tr\left(\Sigma_1^{-1}\Sigma_0^{(s)}\right) - K + \log\frac{|\Sigma_1|}{|\Sigma_0^{(s)}|} + \right. \\ &\qquad \left.\left(\mu_1 - \mu_0^{(s)}\right)^{\top}\Sigma_1^{-1}\left(\mu_1 - \mu_0^{(s)}\right)\right\} \end{aligned} \tag{2}$$

Table 1: Comparison amongs related models with their features.

Features	PLDA	PLTM	LDA-VAE	SB-VAE	JMVAE	Proposed
Annotated Data	✓	✓				✓
Multimodal		✓			✓	✓
Deep Generative Model			✓	✓	✓	✓

[ht]

Table 2: Summary of symbols definitions for the proposed model.

Symbols	Definition
$s \in \mathbb{S}; \lvert\mathbb{S}\rvert = S$	Data source s from the set of all data sources $\mathbb{S}$, where S is the number of data sources.
$\mathbf{x}^{(s)} = \{\vec{x}_1, \ldots, \vec{x}_X\}$	Counting matrix from data source s, where each row is a vector $\vec{x}$ corresponding to an observed datum for all X data.
$\mathbf{y} = \{\vec{y}_1, \ldots, \vec{y}_X\}$	Counting matrix of possible labels, where each row is a vector $\vec{y}$ corresponding to an observed datum for all X data.
$\mathbf{z} = \{\vec{z}1, \ldots, \vec{z}X\}$	Latent variable matrix, where each row is a vector $\vec{z}$ corresponding to an observed datum for all X data.
$\Theta = \{\theta, \{\beta^{(s)}\}_{s\in\mathbb{S}}\}$	Set of model probability distribution parameters.
$\theta = \{\vec{\theta}_1, \ldots, \vec{\theta}_X\}$	Joint representation for all sources of observed data, where each row is a latent distribution vector $\vec{\theta}$ corresponding to an observed datum for all X data.
$\beta^{(s)} = \{\vec{\beta}_1^{(s)}, \ldots, \vec{\beta}_K^{(s)}\}$	Observed data distribution of source s, where each row is an observed data distribution vector $\vec{\beta}^{(s)}$ corresponding to a class for all K classes.
$\Phi = \{\mu_1, \Sigma_1, \{\mu_0^{(s)}, \Sigma_0^{(s)}\}_{s\in\mathbb{S}}\}$	Set of variational distribution parameters.
μ_1, Σ_1	Mean and variance of logistic normal distribution; variational distribution of $p(\theta\|\vec{\alpha})$, respectively.
$\{\mu_0^{(s)}, \Sigma_0^{(s)}\}_{s\in\mathbb{S}}$	Mean and variance of logistic normal distribution; variational distribution of $p(\mathbf{x}^{(s)}\|\beta^{(s)}, \theta)$, respectively.

$$\mathbb{E}_{q_\Phi\left(\mathbf{z}|\{\mathbf{x}^{(s')}\}_{s'\in\mathbb{S}}\right)}\left[\log p_{\Theta_{\mathbf{x}^{(s)}}}\left(\mathbf{x}^{(s)}|\mathbf{z}\right)\right]$$
$$= \sum_{\vec{x}\in\mathbf{x}^{(s)}} \mathbb{E}_{\varepsilon\sim\mathcal{N}(0,1)}\left[\vec{x}^\top \log\left(\sigma(\beta^{(s)})\sigma(\mu_0^{(s)} + \Sigma_0^{(s)1/2}\varepsilon)\right)\right] \quad (3)$$

where σ is the softmax function.

4 RECOMMENDATION MODEL

Our objectives are not only to make recommendations but also to obtain meaningful document representations, which is the main reason JMVAE is used in place of CNN in the CNN–RNN model. The joint representation here is also used in the same way as a CNN image representation is made through the predicted probability. However, the loss function must be changed accordingly to cover those objectives. We simply combine two loss functions as:

$$\mathcal{L}_{\text{JMVAE-RNN}} = \underbrace{\mathcal{L}_{\text{JMVAE}}}_{\text{Eq.(1)}} + \mathcal{L}_{\text{RNN}} \quad (4)$$

$$\mathcal{L}_{\text{RNN}} = \underbrace{-\sum_{\vec{y}\in\mathbf{y}}\left(I(\vec{y})\log(\sigma(\vec{y}')) + (1 - I(\vec{y}))\log(1 - \sigma(\vec{y}'))\right)}_{\text{Softmax cross-entropy}}. \quad (5)$$

$I(\vec{y})$ is a vector with one nonzero value for each element, corresponding, in this work, to a label. $\sigma(\vec{y}')$ is a label probability distribution predicted by softmax, which is used for making the ranked recommendation.

5 EXPERIMENT

5.1 Datasets and Preprocessing

In our experiment to evaluate the proposed model and its recommendation methods, we used two sets of bibliographic data for academic papers on computer science gathered from Microsoft Academic Search (MAS)[1] and from CiNii[2]. Those from MAS are English papers on scientific computing, from the networks and communications subdomains, while the CiNiis papers are bilingual English–Japanese papers.

These papers have many parts we can use, but in this work, we focused on only three: abstract, authors, and keywords. The words in those parts were processed by removing stopwords, tokenizing and stemming using the Porter stemming algorithm (Porter,

[1] https://www.microsoft.com/en-us/research/project/academic/

[2] http://ci.nii.ac.jp/

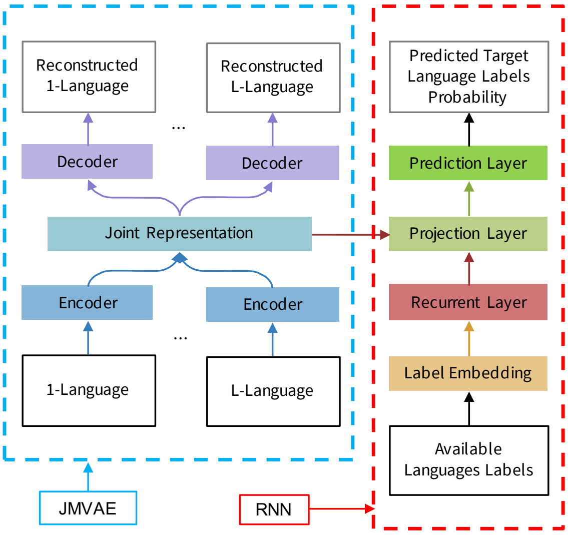

Figure 2: The architecture of the proposed joint multimodal variational autoencoder (JMVAE) combined with an RNN. The JMVAE is employed as the polylingual text representation, and the recurrent layer captures the information about the previously predicted labels (keywords). The output label probability is computed according to the polylingual text representation and the output of the recurrent layer.

Table 3: Size of vocabulary each dataset and its sources.

Variables	CiNii (57,257 papers)		MAS (113,247 papers)
	Japanese	English	
Title	2,634	4,945	
Abstract	6,686	15,410	
Keyword	924	944	8,008

1997). For Japanese words, a morphological analyzer (MeCab[3]) was applied for word segmentation. We then chose those words that appeared in more than 50 papers but not more than 80% of the whole dataset.

We obtained 924 Japanese and 944 English keywords to form the bilingual science database, and 8008 English keywords to form the English computer science database. There were 2634 and 6686 words for Japanese title and abstract variables, and 4945 and 15,410 words for English title and abstract variables.

We chose papers that contained at least one keyword from the chosen sets of keywords. As a result, 57,257 papers from the bilingual dataset and 113,247 papers from the English dataset were used in our experiments. The results of the preprocessing are shown in Table 3.

We evaluated our model by conducting a five-fold cross validation in which the 57,257 and 113,247 papers from the datasets are randomly split into five groups of almost equal size. The model parameters were estimated from four groups of the split datasets, and the remaining one was used to evaluate the accuracy of keyword recommendations.

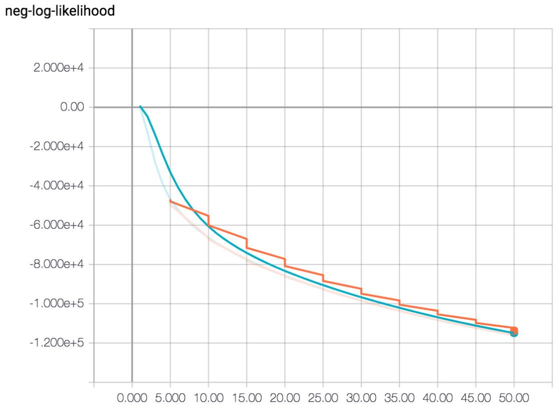

Figure 3: Negative log likelihood of JMVAE–RNN with bilingual corpus CiNii by iteration with blue and orange lines as training and test sets, respectively, from a division of the data set. The values should be positive, but because of high computation requirements, this plot does not include corrections from JMVAE's regularization term, so the likelihoods are negative.

Because of high computation requirements, we only have preliminary results, which show negative log likelihoods for both training and test sets. The deep learning can approximate the parameters effectively as the log likelihood is monotonic.

We plan further evaluations of our method's accuracy with precision and recall and comparisons with related models to show the extent of improvement of accuracy and topic quality by using deep learning techniques.

6 CONCLUSION

We have proposed a new deep generative topic model for recommending keywords from polylingual documents by combining JMVAE and RNN in the same way as CNN–RNN. The model also uses a special trick to approximate the Dirichlet distribution in the form of a Gaussian distribution for easier derivation of JMVAE. We believe that the model parameters can be interpreted easily as they are probabilistically derived while maintaining a level of effectiveness equivalent to that of other deep learning models.

The experimental evaluation of the model is still in its early stages, and preliminary results of negative log likelihood have been shown. We plan on further evaluation of the recommendation accuracy, including comparisons with related models.

[3] http://mecab.sourceforge.net

REFERENCES

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Hoffman, M. D., Blei, D. M., and Bach, F. R. (2010). Online learning for latent dirichlet allocation. In Lafferty, J. D., Williams, C. K. I., Shawe-Taylor, J., Zemel, R. S., and Culotta, A., editors, *NIPS*, pages 856–864. Curran Associates, Inc.

Jaradat, S., Dokoohaki, N., and Matskin, M. (2015). Ollda: A supervised and dynamic topic mining framework in twitter. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pages 1354–1359.

Krstovski, K. and Smith, D. A. (2013). Online polylingual topic models for fast document translation detection. In *In Proc. Workshop on Statistical MT*.

Nalisnick, E. and Smyth, P. (2017). Stick-breaking variational autoencoders. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Porter, M. F. (1997). Readings in information retrieval. chapter An Algorithm for Suffix Stripping, pages 313–316. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Ramage, D., Hall, D., Nallapati, R., and Manning, C. D. (2009). Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ramage, D., Manning, C. D., and Dumais, S. (2011). Partially labeled topic models for interpretable text mining. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 457–465, New York, NY, USA. ACM.

Shu, L., Long, B., and Meng, W. (2009). A latent topic model for complete entity resolution. In *Proceedings of the 2009 IEEE International Conference on Data Engineering*, ICDE '09, pages 880–891, Washington, DC, USA. IEEE Computer Society.

Srivastava, A. and Sutton, C. (2017). Neural variational inference for topic models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Suzuki, M., Nakayama, K., and Matsuo, Y. (2016). Joint multimodal learning with deep generative models. *arXiv preprint arXiv:1611.01891*.

Takasu, A. (2010). Cross-lingual keyword recommendation using latent topics. In *Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems*, HetRec '10, pages 52–56, New York, NY, USA. ACM.

Wang, J., Yang, Y., Mao, J., Huang, Z., Huang, C., and Xu, W. (2016). Cnn-rnn: A unified framework for multi-label image classification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2285–2294.

Zhou, Q., Huang, H., and Mao, X.-L. (2015). *An Online Inference Algorithm for Labeled Latent Dirichlet Allocation*, pages 17–28. Springer International Publishing, Cham.