A Neural Topic Model based on Variational Auto-encoder for Aspect Extraction from Opinion Texts

Peng Cui, Yuanchao Liu* and Binqquan Liu

School of Computer Science and Technology, Harbin Institute of Technology,
Harbin, China, 150001
{pcui,lyc, liubq}@insun.hit.edu.cn
*: Corresponding author

Abstract. Aspect extraction is an important task in ABSA (Aspect Based Sentiment Analysis). To address this task, in this paper we propose a novel variant of neural topic model based on Variational Auto-encoder (VAE), which consists of an aspect encoder, an auxiliary encoder and a hierarchical decoder. The difference from previous neural topic model based approaches is that our proposed model builds latent variable in multiple vector spaces and it is able to learn latent semantic representation in better granularity. Additionally, it also provides a direct and effective solution for unsupervised aspect extraction, thus it is beneficial for low-resource processing. Experimental evaluation conducted on both a Chinese corpus and an English corpus have demonstrated that our model has better capacity of text modeling, and substantially outperforms previous state-of-the-art unsupervised approaches for aspect extraction.

Keywords: Aspect extraction, Neural topic model, VAE

1 Introduction

Aspect extraction is an important task in fine-gained sentiment analysis. It aims to extract target entities (or attributes of entities) that people have expressed in opinionated text. Aspect extraction involves two subtasks: (1) Aspect Term Extraction and (2) Aspect Category Detection. The former subtask aims to identify all the aspect terms present in the sentence, while the latter aims to identify the predefined aspect categories discussed in a given sentence. For example, given the sentence "the waiters were so rude and obnoxious", the first subtask should extract “waiters” as an aspect term and the second should identify “Staff” as aspect category.

Topic models have been widely applied for aspect extraction task because of some salient advantages, e.g., the ability of identifying aspect terms and grouping them into categories simultaneously, and the ability of domain adaption. Conventional topic models are mainly based on Latent Dirichlet Allocation (LDA) (Blei et al., 2003), which regard document as distribution over topics and topics as distribution over words. Many variants of LDA (Brody et al., 2010; Zhao et al., 2010; Chen et al., 2014) have been proposed for aspect extraction and achieved good results. Recently, there is a surge of research interest in neural topic models (Miao et al., 2016, Miao et al., 2017; Srivastava et al., 2017; Ding et al., 2018), which is based on Variational auto-encoder (VAE) (Kingma and Welling, 2015) and regard the latent variables as the topics of documents. VAE-based neural topic models make the most of the ability of auto-encoder structure in extracting features and have been proven by previous works to outperform conventional topic models in learning text representation (Srivastava et al., 2017).

In this paper, we proposed a multi-semantic neural topic model (called MS-NTM) based on variational auto-encoder for aspect extraction, which consists of an aspect encoder, an auxiliary encoder and a hierarchical decoder. In the encoding stage, we use two heterogeneous encoders, i.e., aspect encoder and auxiliary encoder, to build semantic representations in distinct vector spaces. Such structure is based on an intuitive assumption that variables with different priors and formalities (i.e. discrete and continuous) could capture distinct semantic. In the decoding process, we use a hierarchical decoder to decode the aspects and general semantic, which correlates better with the natural semantic structure of real-life reviews. In addition, we incorporate several empirical regularization terms to make two encoders work in a more collaborative manner.

To summarize, our contributions are three-fold: (1) We propose a neural topic model, which is able to learn semantic representation with better granularity by building latent variables in multiple spaces. (2) We apply the proposed neural topic model, which is based on VAE for the task of aspect extraction. and as far as we know, there are few similar work. (3) Experimental results of two domain datasets have demonstrated that our model achieves controllable semantic encoding, and outperforms previous state-of-the-art models.

2 Related Work

For aspect extraction, supervised approaches heavily depend on labelled data and suffer from domain adaption, so a number of unsupervised approaches have been proposed in recent years. Brody et al (2010) used a standard implementation of LDA to detect aspect from online reviews. Zhao et al (2010) proposed MaxEnt-LDA to jointly extract aspect and opinion words. Chen et al (2014) proposed to discover aspects by automatically learning prior knowledge from a large amount of online data. Wang et al (2015) proposed a modified restricted Boltzmann machine (RBM), which jointly learn aspects and sentiment of text by using prior knowledge. He et al (2017) proposed a neural approach based on word embedding and attention mechanism.

Recently, VAE-based neural topic model have been proved to perform well on text modelling. Miao et al (2016) proposed Neural Variational Document Model (NVDM) to use VAE framework for document modelling. Miao et al (2017) further proposed Gaussian Softmax Model (GSM), which modifies the NVDM by using a softmax function on Gaussian latent variables to endow model with the meaning of probability. Srivastava et al (2017) proposed ProdLDA to replace the Gaussian priors of latent variables with Dirichlet priors. Ding et al (2018) proposed several regular versions of NVDM, which leverage pre-trained word embedding to directly optimized coherence of inferred topics. Zeng et al (2018) proposed a hybrid model of neural topic model and memory networks for short text classification.

2.1 Overview of our model

Fig. 1. illustrates the overall architecture of our model. In comparison to previous neural topic models, the major superiority of proposed model is that it can learn underlying semantic representations in better granularity by our innovative modifications of architecture and effective regularization terms. Generally speaking, the whole model includes a discrete aspect encoder, a continuous auxiliary encoder and a hierarchical decoder. Let $r \in \mathbb{R}^V$ stand for bag-of-words (BOW) representation of a single review in dataset D, where V is the size of vocabulary. Two encoders both take r as input. Aspect encoder works like a discriminative model to learn the discrete latent vector z, which can be regarded as aspect label, while the goal of auxiliary encoder is to learn complementary semantic representation h, which are modelled by continuous values, to make aspect representation disentangled and lower the reconstruction loss of decoding process. Afterwards, hierarchical decoder jointly decodes these two types of latent variables in a two-step reconstruction process. We will explain each of the components in detail as follows.

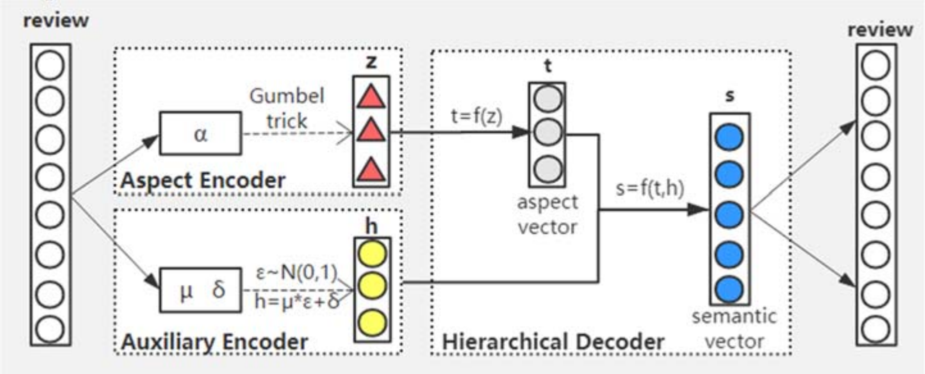

Fig. 1.. architecture of our model

2.2 Discrete Aspect Encoder

The goal of aspect encoder is to learn aspect representation z of reviews. It consists of K units and each of them represents an aspect. We choose Multinomial distribution as the prior of z. Given a review r, we compute its exclusive posterior parameters α_r with a softmax layer.

$$\alpha_r = softamax(W_a \cdot r + b_a) \tag{1}$$

Where $W_a \in \mathbb{R}^{V \times K}$ and $b_a \in \mathbb{R}^K$ are learnable parameters. α_r can be regarded as the aspect distribution of input review r.

To overcome the non-differentiable problem caused by sampling from non-parameterized distribution, we use Gumbel-Max trick (Gumbel and Lieblein, 1954) which is widely used by recent work (Zhou and Neubig, 2017, Zhao et al., 2018) to draw discrete samples from categorical distribution.

$$z_r \sim Gumbel\text{-}Softmax(\alpha_r) \quad (2)$$

Where $z_r \in \mathbb{R}^K$ can be regarded as the approximate one-hot representation of aspect label obtained by Gumbel-Softmax (1954).

According to Equation 1, the contribution of i-th word of review r towards k-th aspect is proportional to $W_a^{(i)(k)}$. Hence we could extract top representative words of each aspect by taking the most positive entries in each column of W_a.

2.3 Gaussian Auxiliary Encoder

Auxiliary encoder aims to learn complementary semantic representation h. Unlike aspect variables z, we model h with continuous values to flexibly capture implicit semantic besides aspects. We choose multivariate Gaussian p(h)=N(μ, σ) as the prior over h. Given a review r, its exclusive posterior parameters μ_r and σ_r are computed by a Multi-Layer Perceptron (MLP). We also use the re-parameterization trick (Kingma and Welling, 2015) to create a differentiable estimator for h.

$$\mu_r = \ tanh\ (W_m \cdot r + b_m) \quad (3)$$

$$log\sigma_r = \ tanh(W_s \cdot r + b_s) \quad (4)$$

$$h_r = \ \mu_r * \varepsilon \ + \sigma_r \quad (5)$$

where matrix $W_m,\ W_s \in \mathbb{R}^{V \times d_h}$ and bias $b_m,\ b_s \in \mathbb{R}^{d_h}$ are learnable parameters. h_r is the latent auxiliary variables of review r, ε is the random noise sampled from multivariate Gaussian with zero mean and unit variances.

Introducing this auxiliary encoder brings two salient advantages. Firstly, it is expected to capture other semantic representation besides aspect by choosing different priors, which enables aspect encoder obtain disentangled aspects representation. Secondly, it makes the whole model achieve better convergence. Clearly, aspect representation *z* is not adequate for decoding process because the number of aspect *K* is much smaller than the size of vocabulary *V*, which brings huge information loss and the high sparse form of *z* further also aggravates this problem. For instance, sentence "*Dessert can't be missed, so save room!*" and "*The beef is fabulous*!" express the same aspect (*Food*), implying that aspect encoder should learn the same representation of two reviews while the decoder is expected to generate totally different words, which increases the reconstruction error. Thus, an additional encoder is an essential component to improve model's overall performance.

2.4 Hierarchical Decoder

Now we have obtained discrete aspect vector z_r and continuous auxiliary vector h_r, Previous work (Serban et al., 2016) has demonstrated that hierarchical networks have better performance as data naturally possess a hierarchical structure. In our task, the aspects information is a higher-level abstract of the general semantic representation. To avoid the representations entangled in decoding process and lower the reconstruction error of decoder, we use a hierarchical decoder to reconstruct reviews in two-step generation process.

In the first step, we decode the aspect variables z_r by projecting it into a continuous space.

$$t_r = z_r \cdot E_a \tag{6}$$

where $E_a \in \mathbb{R}^{K\times d_a}$ can be regarded as a group of aspect vectors and is learned as a part of the training process. In the second step, we use aspect vector t_r and auxiliary vector h_r to get the general sematic representation S_r, then get reconstructed word distribution r' with a softmax layer.

$$S_r = \tanh\left(W_s \cdot [t_r; h_r] + b_s\right) \tag{7}$$

$$r' = softmax(S_r \cdot E_w) \tag{8}$$

Where W_s and b_s are learnable parameters, $E_w \in \mathbb{R}^{(d_a+d_h)\times V}$ can be regarded as a group of word vectors and is learned as a part of the training process.

Note that we build aspect and word vectors in distinct vector spaces. In comparison to that, previous works (Ding et al, 2018; He et al., 2017; Miao et al., 2017) simply put aspect vectors and word vectors within the same vector space. In fact, the aspect information should be a part of whole word semantic. Hence we build the aspect vectors in the subspace of word vector space to better learn these two different level semantic vectors.

2.5 Regularization Terms

We further introduced several intuitive regularization terms to enable the two encoders learn complementary semantic representation in a collaborative manner. Concretely, we first impose L1 normalization on W_a to help aspect encoder select aspect words and filter irrelevant words more effectively.

$$reg_w = ||W_m|| \tag{9}$$

In contrast to contractive auto-encoder (Rifai et al., 2011), our model is only local contractive since we do not use the such regularization on W_m. The reason is that the goal of auxiliary encoder is to provide non-explicit semantic for decoder, thus we do not restrict its ability of learning features.

Another regularization term is the distance between W_a and W_m, which make our two encoders focus on different regions of input text.

$$reg_d = ||T_a - T_m||^2 \tag{10}$$

Where T_a and T_m are the normalized vectors flattened by W_a and W_m, respectively.

At last, we incorporate minimum entropy regularization term (Grandvalet et al., 2004) .

$$reg_h = \sum_{r \in D} H(\alpha_r) \tag{11}$$

Where $H(\alpha_r)$ is the Shannon entropy of aspect distribution α_r. This regularization term enable the model to learn the labels with high confidence, which correlates well with the fact that aspect of real-life reviews should not be ambiguous. Additionally, it also lowers the variance brought by sampling operation.

2.6 Training Objective

There are two parts of the objective functions to optimise in the model. The first part is to maximize the log-likelihood of reviews, we derive the evidence lower bound (*ELBO*) of a single review r as followings:

$$\mathcal{L}_{ELBO}(r) = \mathbb{E}_{q_\theta(h,z)}\left[\sum_{w \in r} logp_\theta\,(w|z_r, h_r)\right] - D_{KL}[q_\theta(z_r|r)||p(z)] \\ -D_{KL}[q_\theta(h_r|r)||p(h)] \tag{12}$$

Where θ is total parameters of model. *p(w|z,h)* is the predictive word probability. $p(h) = N(0,\mathrm{I})$ is standard Gaussian prior, $p(\mathrm{z}) = M(\alpha)$ is Multinomial prior with hyper-parameter α.

The second part are the regularization terms i.e. Equation 9, 10 and 11. The final loss function is expressed as:

$$J(\theta) = \mathcal{L}_{reg} - \sum_{r \in D} \mathcal{L}_{ELBO}(r) \tag{13}$$

$$\mathcal{L}_{reg} = \tau \cdot reg_w + \varphi \cdot reg_d + \omega \cdot reg_h \tag{14}$$

Where τ,φ and ω are hyper-parameters to control the weight of regularization terms.

3 Experiment Setup

3.1 Dataset

We evaluate our model on two datasets of user-generated-reviews. Preprocessing of two datasets involves tokenization, removal of stop words, punctuation symbols and illegal characters. Besides, we only retain the reviews containing 10~100 words of both datasets as our training set. The vocabulary size of two datasets are truncated by word frequency. The detail statics of two datasets are summarized in table 1.

1) **Restaurant Corpus**: This is an English dataset composed of over 50,000 unlabelled restaurant reviews collected from CitySearch and 3400 labelled reviews, which is widely used by previous works (Ganu et al., 2009; Brody and Elhadad,

2010; Zhao et al., 2010; He et al., 2017). The six pre-defined aspect categories are: {*Food, Staff, Price, Ambience, Anecdotes, Miscellaneous*}

2) **Mobile Game Corpus**: This is a Chinese dataset composed of 128, 977 reviews of a popular mobile game *王者荣耀*(*Arena of Valor*) collected from social network and app. The five pre-defined aspect categories are: {*英雄(Hero), 皮肤(Skin), 装备(Item), 排位(Rank), 社交(Sociality)*}. We manually annotated 1500 reviews as its test set.

Table 1. statics of datasets

Datasets	#Reviews	#Vocab	#Aspects categories
Restaurant	52, 574	10, 000	6
Game	12, 8977	15, 000	5

3.2 Baseline methods

1) **LocLDA**: The standard implementation of LDA (Brody and Elhadad, 2010).
2) **BTM** (Yan et al., 2013): Biterm Topic Model directly models the generation process of word-pair to alleviate the sparsity problem.
3) **ABAE** (He et al., 2017): This approach uses Neural Bag of Words (NBOW) as sentence representation and learns aspect embedding in a reconstruction process, where attention mechanism is used to filter non-aspect words. This baseline has achieved state-of-the-art results on restaurant corpus.
4) **NVDM** (Miao et al., 2016): This is a general framework which employs the vanilla VAE framework with a Gaussian encoder to model document.

3.3 Experimental Settings

Hyper-parameters

The aspect numbers K is a part of our experiment and described in each evaluation task. Other hyper-parameters are described as follows. For LocLDA, we use the open-source implementation GibbsLDA++ and set Dirichlet priors α = 10/K and β = 0.1. For BTM, we use the implementation released by (Yan et al. 2013) and set α = 50/K and β = 0.1. Two topic models are run 1,000 iterations of Gibbs sampling. For ABAE, we use the code released by (He et al. 2017) and the settings in its reference. For NVDM, we re-implemented the model and set dimension of Gaussian variables to K. For MS-NTM, the dimension of aspect variables is equal to K, the dimension of the auxiliary variables and that of aspect vectors are both set to 128. The prior parameter $\alpha = 1/K$ representing the fully unsupervised version. τ, ω and φ are set to 1, 2 and 0.5 respectively. The Gumbel temperature is set to 0.1. The parameters of all neural models are randomly initialized and optimized using Adam (Kingma and Ba, 2014). Both neural models are trained with initial learning rate 0.001 for 50 epochs and batch size of 32.

Training Strategies

We separately trained two encoders of MS-NTM. In the first 25 epochs we fix the parameters of auxiliary encoder. In the second 25 epochs, we trained both encoders together with different learning rates, which are 0.0001 for aspect encoder and 0.001 for auxiliary encoder. Meanwhile the parameters of decoder are trained with the same learning rate 0.001 in all epochs. This training method is an empirical choice to alleviate the entanglement of two encoders during training process.

4 Evaluation Results

This section reports the experimental results of MS-NTM. We evaluated its capacity in modeling text and its performance on aspect extraction.

4.1 Text Modelling

Perplexity is a standard measure for evaluating topic models derived from the likelihood of unseen test data.

Fig. 2 presents the perplexity of each method on two datasets under different K value, which start from golden-standard aspect numbers of respective dataset. Note that perplexity of ABAE is intractable. Considering that approximate approaches may bring bias to evaluation., so we did not evaluate the performance on perplexity of ABAE.

We can have the following observations from Fig. 2: (1) MS-NTM outperforms previous models for all values of K. (2) LDA performs worst, it may be because that most of the reviews are relative short. (3) NVDM performs better than traditional topic models especially for higher K, implying the effectiveness of VAE framework for text modelling.

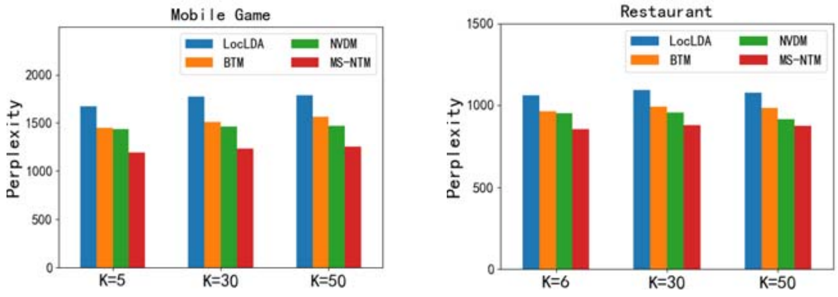

Fig. 2. results of Perplexity

4.2 Aspect Extraction

Aspect Coherence

We manually mapped each inferred aspect to gold-standard aspects as previous work (Brody and Elhadad, 2010; He et al., 2017). In order to compare with previous work, we set K=14 for subsequent evaluation tasks of aspect extraction.Table 2 presents representative words (k=6) selected from W_a. As can be seen from table 2, the aspects inferred by MS-NTM are quite coherent and informative.

To evaluate the quality of inferred aspects, one metric is coherence score, which has been used by previous work (Mimno et al., 2011; Chen et al., 2014; He et al., 2017) to judge whether an aspect is coherent. Given an aspect z and a set of top N

related words of z, $w^z = \{w_1^z, \ldots, w_N^z, \}$. The coherence score of aspect z is calculated as follows:

$$\mathrm{S}(\mathrm{z}, w^z) = \sum_{n=2}^{N} \sum_{l=2}^{n-1} \log \frac{T_2(w_n^z, w_l^z) + 1}{T_1(w_l^z)} \tag{15}$$

Where $T_1(w_l^z)$ is the document frequency of w_l^z and $T_2(w_n^z, w_l^z)$ is the co-document or co-review frequency of word w_n^z *and* w_l^z.

Table 2. representative words of inferred aspects.

Food	Staff	Ambience	Price	Anecdotes	Miscellaneous
beef	manager	atmosphere	charge	travel	location
pork	wait	music	bill	party	experience
Pancake	pleasure	room	dollar	evening	kid
duck	server	ambience	value	tourist	taxi
noodle	waitress	come	walk	sahara	weather
appetizer	waiter	lighting	waste	date	rock
steak	behavior	bar	tax	christmas	bravo
cocktail	rudeness	space	buck	wedding	reservation
salad	people	area	fee	alien	app

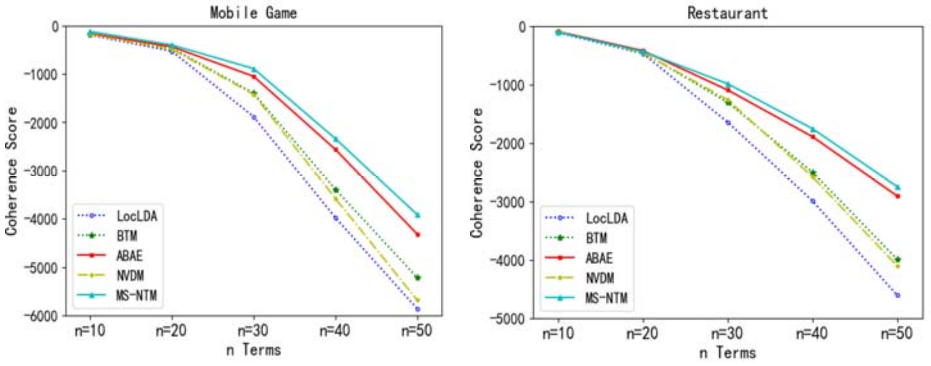

Fig. 3. Average coherence score under different terms N

Fig. 3 presents the coherence score on two datasets of all methods with different N, from which we can see that MS-NTM outperforms baseline models for all ranked buckets and the gap increases with the increasing of N. Besides, although NVDM performs well in perplexity, NVDM performs not well in coherence score. We raise the hidden size of NVDM to be the same as that of MS-NTM and our model still performs better and maintains good interpretability of latent representation, which prove the effectiveness of the collaborative learners of MS-NTM.

Aspect Identification

We then evaluated the performance of MS-NTM on aspect category identification task. Given a review, we assign it aspect label corresponding to the highest value of α_r in Eq. 4. In order to compare with previous work, we follow their settings (Brody and Elhadad, 2010; Wang et al., 2015; He et al., 2017) to remove the multi-labeled reviews and only evaluated on three major aspects for restaurant dataset, which are food, staff, and Ambience.

Table 3. Results of aspect identification, P, R and F1 represent precision, recall and F1 score, respectively. The results of LocLDA are from Zhao et al (2010), the results of BTM and ABAE are from He et al (2017), the results of SERBM are from Wang et al (2015)

	Restaurant Corpus				**Mobile Game Corpus**			
	Aspect	P	R	F1	Aspect	P	R	F1
LocLDA	Food	0.898	0.648	0.753	Hero	0.872	0.614	0.705
	Staff	0.804	0.585	0.677	Item	0.692	0.512	0.589
	Ambience	0.603	0.677	0.638	Social	0.694	0.504	0.584
BTM	Food	0.933	0.745	0.816	Hero	0.850	0.702	0.769
	Staff	0.828	0.579	0.677	Item	0.700	0.533	0.605
	Ambience	0.813	0.599	0.685	Social	0.670	0.475	0.599
ABAE	Food	**0.953**	0.741	0.828	Hero	**0.902**	0.701	**0.789**
	Staff	0.802	0.728	0.757	Item	0.711	0.517	0.599
	Ambience	0.815	0.698	0.740	Social	0.711	0.501	0.588
NVDM	Food	0.741	0.662	0.699	Hero	0.766	0.571	0.654
	Staff	0.701	0.497	0.582	Item	0.614	0.487	0.543
	Ambience	0.620	0.543	0.579	Social	0.652	0.448	0.531
SERBM	Food	0.891	**0.854**	0.872	Hero	-	-	-
	Staff	0.819	0.582	0.680	Item	-	-	-
	Ambience	0.805	0.592	0.682	Social	-	-	-
MS-NTM	Food	0.942	0.818	**0.873**	Hero	0.884	**0.709**	0.787
	Staff	**0.833**	**0.730**	**0.778**	Item	**0.719**	**0.594**	**0.651**
	Ambience	**0.819**	**0.701**	**0.755**	Social	**0.720**	**0.504**	**0.593**

Table 3 presents precision, recall and F1 score of two datasets. We also compare our model with SERBM (Wang et al., 2015) for restaurant corpus, one of the art-of-state models. Results show that MS-NTM substantially outperforms previous methods. For some frequent aspects (Food, Hero), ABAE slightly outperforms our model. A possible reason is that the word embedding quality of frequent words is higher (Bojanowski and Grave; 2016), which also indicates that the performance of ABAE may depend on the word embedding model, while our model has better robustness.

5 Conclusion

In this paper, we have presented a neural topic model based on variational autoencoder for aspect extraction from opinion texts. In comparison to previous models, it can learn multiple semantic representations by choosing appropriate priors and intuitive regularization terms. Experimental results have demonstrated that our model not only has better ability of text modelling, but also outperforms previous art-of-state unsupervised methods for aspect extraction task. Further explorations of VAE-based techniques for modelling contextual semantic, and neural topic model for solving more downstream NLP tasks will be addressed in our future research.

6 Acknowledgments

This study was supported by the National Natural Science Foundation of China (61672192 and 61572151).

References

1. Wayne Xin Zhao, Jing Jiang, Hongfei Yan, and Xiaoming Li. 2010. Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing.
2. Samuel Brody and Noemie Elhadad. 2010. An unsupervised aspect-sentiment model for online reviews. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics.
3. Yishu Miao, Lei Yu, and Phil Blunsom. 2016. Neural variational inference for text processing. In International Conference on Machine Learning, pages 1727–1736.
4. Ran Ding, Ramesh Nallapati and Bing Xiang 2018. Coherence-Aware Neural Topic Modeling. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 830–836.
5. Yishu Miao, Edward Grefenstette, and Phil Blunsom. 2017. Discovering discrete latent topics with neural variational inference. In International Conference on Machine Learning, pages 2410–2419.
6. Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. arXiv preprint arXiv:1703.01488.
7. David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. Journal of Machine Learning Research, 3(Jan):993–1022.
8. Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2017 An Unsupervised Neural Attention Model for Aspect Extraction. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pages 388–397
9. Linlin Wang, Kang Liu, Zhu Cao, Jun Zhao, and Gerard de Melo. 2015. Sentiment-aspect extraction based on restricted Boltzmann machines. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing.
10. Diederik P Kingma and Max Welling. 2013. Autoencoding variational bayes. arXiv preprint arXiv:1312.6114.
11. Grandvalet, Yves, Bengio, Yoshua, et al. Semi-supervised learning by entropy minimization. In NIPS, volume 17, pp. 529–536, 2004.
12. Tiancheng Zhao, Kyusong Lee and Maxine Eskenazi. 2018. Unsupervised Discrete Sentence Representation Learning for Interpretable Neural Dialog Generation. arXiv:1804.08069
13. Jichuan Zeng, Jing Li, Yan Song, Cuiyun Gao, Michael R. Lyu and Irwin King1. 2018. Topic Memory Networks for Short Text Classification. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3120–3131

14. Zhiyuan Chen, Arjun Mukherjee and Bing Liu. 2014. Aspect Extraction with Automated Prior Knowledge Learning. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pages 347–358.
15. Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
16. Salah Rifai, Pascal Vincent, Xavier Muller, Xavier Glorot and Yoshua Bengio. 2011. Contractive Auto-Encoders: Explicit Invariance During Feature Extraction. In Proceedings of the 28th International Conference on Machine Learning
17. Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville and Joelle Pineau. 2015. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models. In Association for the Advancement of Artificial Intelligence
18. Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2016. A hierarchical latent variable encoder-decoder model for generating dialogues. arXiv preprint arXiv:1605.06069.
19. Emil Julius Gumbel and Julius Lieblein. 1954. Statistical theory of extreme values and some practical applications: a series of lectures. US Government Printing Office Washington.
20. Chunting Zhou and Graham Neubig. 2017. Multi-space Variational Encoder-Decoders for Semi-supervised Labeled Sequence Transduction. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pages 310–320
21. Reed, Scott, Lee, Honglak, Anguelov, Dragomir, Szegedy, Christian, Erhan, Dumitru, and Rabinovich, Andrew. Training deep neural networks on noisy labels with bootstrapping. arXiv preprint arXiv:1412.6596, 2014.
22. Samuel R. Bowman and Luke Vilnis. 2016 Generating Sentences from a Continuous Space. In arXiv:1511.06349v4
23. Wayne Xin Zhao, Jing Jiang, Hongfei Yan, and Xiaoming Li. 2010. Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid. In Proceedings of EMNLP 2010, pages 56
24. Jonathan Chang, Jordan Boyd-Graber, Wang Chong, Sean Gerrish, and David Blei, M. 2009. Reading Tea Leaves: How Humans Interpret Topic Models. In Proceedings of NIPS, pages 288–296.
25. Shuming Ma, Xu Sun, Junyang Lin and Houfeng Wang. 2018. Autoencoder as Assistant Supervisor: Improving Text Representation for Chinese Social Media Text Summarization. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 725–731
26. Gayatree Ganu, Noemie Elhadad, and Amelie Marian. 2009. Beyond the stars: Improving rating predictions using review text content. In Proceedings of the 12th International Workshop on the Web and Databases.
27. David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing.
28. Piotr Bojanowski, Edouard Grave, Armand Joulin and Tomas Mikolov. 2016. Enriching Word Vectors with Subword Information. In arXiv:1607.04606
29. Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Controllable text generation. arXiv preprint arXiv:1703.00955.
30. Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. In Proceedings of the 2nd International Conference on Learning Representations.