Coherence Regularization for Neural Topic Models

Katsiaryna Krasnashchok(✉) and Aymen Cherif

EURA NOVA, Rue Emilie Francqui 4, 1435 Mont-Saint-Guibert, Belgium
katherine.krasnoschok@euranova.eu
http://euranova.eu

Abstract. Neural topic models aim to predict the words of a document given the document itself. In such models perplexity is used as a training criterion, whereas the final quality measure is topic coherence. In this work we introduce a coherence regularization loss that penalizes incoherent topics during training of the model. We analyze our approach using coherence and an additional metric - exclusivity, responsible for the uniqueness of the terms in topics. We argue that this combination of metrics is an adequate indicator of the model quality. Our results indicate the effectiveness of our loss and the potential to be used in the future neural topic models.

Keywords: Topic modeling · Neural networks · NPMI · Topic coherence

1 Introduction

Topic Modeling is an established area of text mining focused on discovering topics in a collection of documents. Generative models like Latent Dirichlet Allocation (LDA) [1] have been long used as a standard in Topic Modeling. With the popularization of the Neural Networks and Deep Learning, several neural topic models have been suggested [2,3], implementing the same generative principles: generating a document given the same document, through the topic/word and document/topic distributions. The advantages of neural topic models include better resource management through the flexibility of training, and natural use of embeddings, a highly effective type of word representation, as opposed to bag-of-words model, traditionally used in LDA. Neural topic models, being unsupervised, use a metric like perplexity as a training criterion, to control the ability of the model to generate documents. However, perplexity does not reflect the human judgment [4] of the topic quality, unlike coherence, commonly used in evaluation of topic models. Therefore, such models cannot guarantee the best quality of the final output. Furthermore, coherence cannot be regarded as a single defining quality metric: being an average between all topics, the final coherence will favor repeating topics, coherent but containing the same words.

H. Lu et al. (Eds.): ISNN 2019, LNCS 11554, pp. 426–433, 2019.
https://doi.org/10.1007/978-3-030-22796-8_45

In this work we present a new regularization loss for a neural topic model, inspired by the coherence measure. Our contributions are the following: (i) we propose a simple and straightforward regularization function for a neural topic model, aimed to control the coherence of the intermediate topics, generated during training; (ii) we propose the usage of a new (to the best of our knowledge) composite metric, combining coherence and a uniqueness measure - *exclusivity* [15], thus solving the coherence's bottleneck of repeating topics; (iii) we introduce new multi-criteria training procedure – a combination of perplexity and aforementioned composite metric, to better control the topic model quality during training.

The paper is organized as follows: in Sect. 2 we outline the related work that has influenced our approach; Sect. 3 details the proposed regularization technique; implementation and experiments are described in Sect. 4, and their results are listed and analyzed in Sect. 5. Finally, Sect. 6 finalizes our findings and sets up a plan for the future work.

2 Related Work

In this section we outline the previous research that have shaped up our work.

2.1 Regularization for Standard Topic Models

Regularization of topic models is not a novel concept. [5] proposed to modify the LDA model by building a structured prior over words using a covariance matrix, enforcing co-occurring words to appear in the same topics. Another regularizer in form of a Markov Random Field, was presented in [6], with the same idea of incorporating word correlation knowledge into LDA. [7] offered a more efficient method called Sparse Constrained LDA. All of the aforementioned models carry the same idea of adding word co-occurrence information to the LDA algorithm. This concept is transferred to neural topic models in our work.

2.2 Neural Topic Models

The success of Neural Networks and Deep Learning in many NLP tasks has lead to the research in Neural Topic Modeling. The early methods use autoencoders [8], a Restricted Boltzmann Machine [9], autoregressive models [10] and Deep Boltzmann Machine [11]. [2] proposed a straightforward way to represent the topic model with a Neural Network. Their model uses word embeddings, which enables the input consisting of ngrams instead of single words. Topically Driven Neural Language Model (TDLM) [3] adopts a topic model to improve the language model. The authors report a state of the art coherence for several datasets, outperforming [2] and in most cases LDA. For this reason, TDLM serves in our work as the baseline and base model, where our regularization loss is applied.

2.3 Topic Quality Measures

Topic quality measures have been evolving in the recent years: from using perplexity [12] to various metrics for topic coherence [13]. Normalized Point-wise Mutual Information (NPMI) coherence is one of the most well performing and popular with researchers [13]. Whereas coherence is well-informative, it does not measure redundancy [1], e.g. a model containing N coherent, but identical topics would get a high score, despite being low quality. For this reason, many works conduct a qualitative analysis of the topics. The other way to handle redundancy problem is to add another metric, such as inter-topic similarity from [1], or generality in [14]. In this work we follow the latter approach and employ the exclusivity metric [15] for measuring the "uniqueness" of the topics. We then introduce a new metric – a combination of coherence and exclusivity, in order to give better quantitative assessment of the model quality.

3 Coherence Loss

Traditionally, in neural models dropout is used for regularization, to prevent overfitting. For neural topic models, other types of regularizers, besides dropout, are necessary to improve the coherence of the resulting topics. The main loss function of a neural topic model, such as [2,3], is designed for an autoencoder, where the training is led by metrics like perplexity, which, as we know, does not correlate with coherence [4]. From the need to explicitly control the coherence of the topics we have drawn the idea of adding a coherence loss as a new regularization technique.

In this work we use NPMI coherence in the evaluation, the reason being its popularity and superior performance. The definition for NPMI coherence is inferred from the framework by [13]:

$$C_{NPMI} = \sigma^a_{t=1\ldots N}(\sigma^a_{i,j\in\binom{N_t}{2}}(NPMI(w_i, w_j))) \tag{1}$$

where σ^a is an arithmetic mean [13], N is the number of topics, $w_i, w_j \in W_t$ – the set of top N_t terms of each topic and

$$NPMI(w_i, w_j) = \frac{\log \frac{P(w_i,w_j)+\epsilon}{P(w_i)\cdot P(w_j)}}{-\log P(w_i, w_j) + \epsilon} \tag{2}$$

NPMI values lie in $[-1; 1]$ interval. Probabilities P are estimated based on the word occurrence and co-occurrence matrices. It is assumed that the more the terms of a topic appear together in a corpus, the better they are related, which makes the topic more coherent. The co-occurrence matrix can be built on any corpus, including the training one, though the coherence calculated on a large external corpus (like Wikipedia) is known to perform better [16]. Our proposed coherence loss is inspired by the C_{NPMI} coherence formula (1) and defined as:

$$L_{NPMI} = \alpha * \sigma^a_{t=1\ldots N}(\sigma^a_{i,j\in\binom{N_t}{2}}(\phi_{t,i,j} * (1 - NPMI(w_i, w_j)))), \tag{3}$$

where α is the loss coefficient, to be decided empirically. The loss definition differs from the coherence formula in two aspects: firstly, we use $(1 - NPMI(w_i, w_j))$ instead of $NPMI(w_i, w_j)$, to represent the coherence penalty, and secondly, we add a topic-specific coefficient $\phi_{t,i,j}$, calculated from topic/word distribution ϕ of words w_i, w_j: $\phi_{t,i,j} = \phi_{t,i} * \phi_{t,j}$. The value of $\phi_{t,i,j}$ can be considered as the "importance" of an individual coherence loss, since it depends on the position of terms in the topic descriptor: the higher pair of terms in the list will be given more weight for its coherence penalty, while the last words in the descriptor are allowed to be slightly less coherent. We apply a $softmax$ over the top N_t ϕ values for each topic to avoid very small numbers in loss calculation.

Each step of the training process, our coherence cost is computed and added to the main cost, e.g. cross-entropy for [3], with a coefficient α.

4 Experiments

Seeing that our goal in this work is to improve coherence, we have chosen the state of the art neural topic model as both our baseline and base model: TDLM from [3], with a straightforward and elegant topic model neural representation. For our experiments we exclude the language model part of TDLM and focus only on the topic model, which also significantly accelerates the training time.

4.1 Implementation

For testing the proposed coherence loss, we have used the open-source tensorflow implementation[1] from [3], where we have disabled the language model part, leaving only the topic model. Occurrence and co-occurrence matrices for datasets were computed once and saved. For the lack of sufficient resources, we calculated the matrices based on the training data, instead of using an external bigger corpus. Since the goal is to demonstrate the work of the coherence loss in comparison with the base model, "local" co-occurrence information suits the task well. The α coefficient in (3) is a hyperparameter, of which several values were tested. In our results we report the best performing values.

4.2 Datasets

For the sake of comparison, three datasets from [3] have been chosen: IMDB reviews from [17], APNEWS[2] - collection of news from Associated Press, and British National Corpus (BNC)[3] [18]. The training and validation sets were taken directly from the open-source TDLM project, along with the default parameters. Due to difference in model and evaluation, TDLM's coherence values reported in [3] could not be used for comparison, thus all metrics were recalculated.

[1] https://github.com/jhlau/topically-driven-language-model.
[2] https://www.ap.org/en-gb/.
[3] http://www.natcorp.ox.ac.uk/.

4.3 Evaluation

As mention in the previous sections, using perplexity as a training criterion does not guarantee best coherence. Therefore, we modified the training and evaluation procedures in three different aspects, described below.

Quality Measures. We added a new metric, called *exclusivity*: $excl = \frac{|W_u|}{|W|}$, where $|W_u|$ is the number of unique terms and $|W|$ is the total number of terms in topic descriptors. It has been used in [15] as complimentary measure for coherence, to cover the problem of redundancy. Exclusivity is a simple and straightforward variation of the metrics offered by [1,14], and takes values from interval $(0; 1]$. The latter property allows to view exclusivity as a measure of "quality of coherence". Naturally, some redundancy is unavoidable in a topic model, yet, the less redundant model is the more coherent one [1]. Therefore, both metrics should be considered for evaluation. In this work we propose a composite measure $Q = C_{NPMI} * excl$, that captures both coherence and exclusivity, providing a fair assessment of the topics. This formula is only used with *positive* coherence values, as negative coherence already indicates low quality model. Eventually, more complex formulas may be used, depending on the importance of both measures, but for this work we focus on the basic definition.

Multi-criteria Training. After each epoch, we compute a tuple of metrics, consisting of validation perplexity *ppl* and Q. It is then checked in the following manner: if none of the metrics are improving compared to the previous epochs, the epoch is considered failed and the parameters are restored to the previous epoch, as in [3]. Otherwise, the training continues without changes. This way we make sure that good quality epochs do not get restarted because of worse perplexity.

Final Evaluation. We run the model for 20 epochs and average the results from several runs for each epoch. The best coherence and Q values may not be the final ones, therefore per-epoch analysis is performed. We compare three types of values: best coherence, best exclusivity, and finally best Q among epochs.

5 Results

We ran the experiments on IMDB, APNEWS and BNC for $N = 50, 100, 150$ topics. For each combination of dataset/N we found the value of α empirically. Table 1 shows the obtained results (with superior values **in bold**). Comparison of coherence indicates that the regularized version improves the metric in the majority of cases, while exclusivity values decrease, to different extents. This observation is justified, since our regularization is based on average coherence. An exclusivity regularizer may be beneficial, which we leave for the future work.

Table 1. Topic quality results, maximum values from 20 epochs

Dataset	IMDB			APNEWS			BNC		
C_{NPMI}									
N (α)	50 (1)	100 (1)	150 (1)	50 (1)	100 (2)	150 (2)	50 (2)	100 (2)	150 (2)
Baseline	0.026	0.044	**0.043**	0.150	**0.162**	0.160	**0.145**	0.140	0.137
Regularized	**0.035**	**0.045**	0.041	**0.151**	0.155	**0.163**	0.143	**0.142**	0.137
excl									
Baseline	**0.634**	**0.422**	**0.366**	0.868	0.659	**0.531**	0.885	**0.656**	**0.510**
Regularized	0.620	0.409	0.361	**0.869**	**0.674**	0.504	**0.905**	0.620	0.504
$Q = C_{NPMI} * excl$									
Baseline	0.016	0.018	0.014	0.129	**0.105**	0.082	0.128	**0.092**	0.067
Regularized	**0.021**	0.018	0.014	**0.130**	0.103	0.082	**0.129**	0.088	**0.069**

The analysis of the composite metric Q concludes that our regularization loss is mostly beneficial for smaller number of topics. The difference is especially significant with IMDB, our smallest dataset, where regularized model gained a big difference in C_{NPMI} and Q, without adding much redundancy to the topics. This indicates that the coherence loss is particularly useful for small and noisy corpora, where it is harder to obtain coherent topics. For a more detailed analysis, Fig. 1 shows the trends of the metrics for IMDB $N = 50$, $\alpha = 1.0$. It is evident that our proposed loss have improved the coherence at each iteration while only slightly decreasing the exclusivity. This trend is persistent through the training process. Moreover, the best quality topics do not correspond to the final epoch, which justifies our per-epoch analysis. It is also worth noticing that loss in perplexity for our regularized model is considerably small, compared to the gain in coherence.

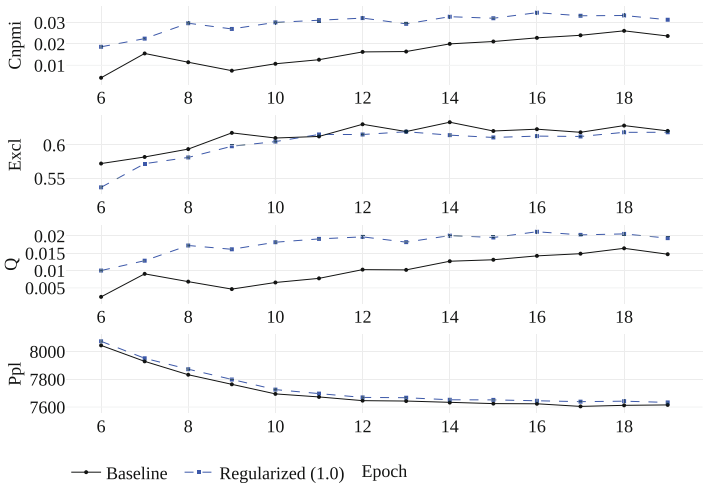

Fig. 1. Metrics per epoch of IMDB corpus for 50 topics (from epoch 6)

Other tests show the increase in coherence, but bigger decrease in exclusivity, which results in the trade-off Q metric being equal or (in two cases) worse than the baseline TDLM. We can account that on the resource limitations that allowed us to only test several values of the parameter α for large number of topics. We believe that more extensive experiments will reveal better values, with possible addition of an exclusivity regularization.

Based on the obtained results, we can observe that the proposed coherence loss does indeed improve the topic coherence metric. Moreover, our new composite quality metric Q proved to be a necessary addition to the evaluation process, showing that a better coherence value does not always indicate better model quality, as it may add much redundancy to the topics. Our new quality metric allows to estimate the trade-off between gain in coherence and loss in exclusivity, thus simplifying and improving the comparative analysis of the models.

6 Conclusion

In this work we presented new coherence loss, intended as a regularization loss for neural topic models. We have tested our loss on the state of the art model TDLM [3] and demonstrated its ability to increase the coherence of the topics (in most cases), and the overall quality of the topics (in few cases). Our regularization technique is flexible: the loss can be applied to any neural topic model, where a topic/word distribution can be computed during training. Moreover, we introduced a composite metric for the topic quality evaluation, representing the trade-off between topic coherence and the level of redundancy in topics (exclusivity). Finally, we proposed a multi-criteria training procedure, which allowed us to control both perplexity and topic quality metrics during training.

Future Work. Testing of our proposed loss revealed the need to add an exclusivity regularization to control the redundancy in topics. This can be achieved by adding an entropy-based loss that would ensure that each word in the vocabulary is assigned to maximum one topic. This addition we leave for the future work. Furthermore, the coherence loss should be tested on other neural topic models, such as NTM [2], where the main loss differs from TDLM. Additionally, future work will include testing more hyperparameter values for large topic numbers, with co-occurence matrices computed on a reference corpus.

Acknowledgements. The elaboration of this scientific paper was supported by the Ministry of Economy, Industry, Research, Innovation, IT, Employment and Education of the Region of Wallonia (Belgium), through the funding of the industrial research project Jericho (convention no. 7717).

References

1. Arora, S., et al.: A practical algorithm for topic modeling with provable guarantees. In: International Conference on Machine Learning, pp. 280–288 (2013)
2. Cao, Z., Li, S., Liu, Y., Li, W., Ji, H.: A novel neural topic model and its supervised extension. In: AAAI, pp. 2210–2216 (2015)

3. Lau, J.H., Baldwin, T., Cohn, T.: Topically driven neural language model. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, vol. 1: Long Papers. ACL (2017)
4. Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J.L., Blei, D.M.: Reading tea leaves: how humans interpret topic models. In: Advances in Neural Information Processing Systems, pp. 288–296 (2009)
5. Newman, D., Bonilla, E.V., Buntine, W.: Improving topic coherence with regularized topic models. In: Advances in Neural Information Processing Systems, pp. 496–504 (2011)
6. Xie, P., Yang, D., Xing, E.: Incorporating word correlation knowledge into topic modeling. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. ACL (2015)
7. Yang, Y., Downey, D., Boyd-Graber, J.: Efficient methods for incorporating knowledge into topic models. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. ACL (2015)
8. Ranzato, M.A., Szummer, M.: Semi-supervised learning of compact document representations with deep networks. In: Proceedings of the 25th International Conference on Machine Learning - ICML 2008. ACM Press (2008)
9. Hinton, G.E., Salakhutdinov, R.R.: Replicated softmax: an undirected topic model. In: Advances in Neural Information Processing Systems, pp. 1607–1614 (2009)
10. Larochelle, H., Lauly, S.: A neural autoregressive topic model. In: Advances in Neural Information Processing Systems, pp. 2708–2716 (2012)
11. Srivastava, N., Salakhutdinov, R., Hinton, G.: Modeling documents with a deep Boltzmann machine. In: Uncertainty in Artificial Intelligence, p. 616. Citeseer (2013)
12. Wallach, H.M., Murray, I., Salakhutdinov, R., Mimno, D.: Evaluation methods for topic models. In: Proceedings of the 26th Annual International Conference on Machine Learning, pp. 1105–1112. ACM (2009)
13. Röder, M., Both, A., Hinneburg, A.: Exploring the space of topic coherence measures. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM 2015. ACM Press (2015)
14. O'Callaghan, D., Greene, D., Carthy, J., Cunningham, P.: An analysis of the coherence of descriptors in topic modeling. Expert Syst. Appl. **42**(13), 5645–5657 (2015)
15. Krasnashchok, K., Jouili, S.: Improving topic quality by promoting named entities in topic modeling. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), vol. 2, pp. 247–253 (2018)
16. Newman, D., Lau, J.H., Grieser, K., Baldwin, T.: Automatic evaluation of topic coherence. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistic, pp. 100–108. ACL (2010)
17. Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C.: Learning word vectors for sentiment analysis. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 142–150. ACL (2011)
18. BNC Consortium. The British National Corpus, version 3 (BNC XML Edition). Distributed by Oxford University Computing Services on behalf of the BNC Consortium (2007)