Topic Models with Sparse and Group-Sparsity Inducing Priors

Christian Pölitz

TU Dortmund University, Otto Hahn Str. 12, 44227 Dortmund

Abstract. The quality of topic models highly depends on quality of used documents. Insufficient information may result in topics that are difficult to interpret or evaluate. Including external data to can help to increase the quality of topic models. We propose sparsity and grouped sparsity inducing priors on the meta parameters of word topic probabilities in fully Bayesian Latent Dirichlet Allocation (LDA). This enables controlled integration of information about words.

1 Introduction

Topic models have been used for text analysis in the last decade very successfully. Topic models assign a number of latent topics to documents and words from a given corpus. These topics can be interpreted as different meanings of words or semantic clusters of the documents in a text corpus. In text analysis the topics can be used in many ways.

The estimation of the topics highly depends on the amount of text data used. Considering the case when we have only very limited amounts of texts to estimate a topic model, the quality of the found topics can be quite poor. In such situation external information about the words can be quite beneficial. For instance prior word probabilities can help sampling word topic distributions from a Dirichlet distribution by adding prior weights on more likely words. In this sense, we try to align the topics with an external probability model like a language model $p(w)$ over some of the words. Structural external information like similarities of words can provide further help to align the topics. Hence, prior weights of whole groups of similar words can be used to estimate the topics.

To measure the quality of the found topics, intrinsic measures like the perplexity have been used in the past. Recently, coherence measures have been introduced as an evaluation measure for topics that agree well with human judgements, see [11]. These coherence measures use external information to evaluate how much related the most likeliest words in the topics are. To extract coherent topics by a topic model we must assume to have enough coherent documents. This is not always the case. In Word Sense Induction for instance, there may be rare words that appear only in a few documents. In such a case these documents might not be enough to generate coherent topics. Further, very sparse documents as in collections of Blog posts or Tweets might also lack enough information to extract coherent topics.

To increase the coherence, we propose to integrate external information like word probabilities or word similarities from external data sources. To control the influence from the external information we weight these information additionally. We integrate

In: P. Cellier, T. Charnois, A. Hotho, S. Matwin, M.-F. Moens, Y. Toussaint (Eds.): Proceedings of DMNLP, Workshop at ECML/PKDD, Riva del Garda, Italy, 2016.

external word probability information by appropriate prior distributions. We add a sparsity prior and a group sparsity prior on the log-likelihood of the topic model, see [16]. The sparsity inducing priors can now actively control the amount and the weight of the external information to be integrated in the estimation of a topic model. From the group sparsity we expect more coherence since whole groups of words are considered. These groups are expected to be more coherent since they are similar based on some external information.

2 Related Work

There are many previous approaches integrating external information into the generation of a topic model. [8] use a regression model on the hyperparameters of the Dirichlet prior for LDA. They use Dirichlet multinomial regression to make the prior probability of the document topic distribution dependent on document features. [14] integrate word features into LDA by adding a Logistic prior on the parameter of the Dirichlet prior of the word topic distribution. [9] integrate correlation information about words into a topic model. They propose regularized topic models that have structural priors instead of Dirichlet priors. These structural priors contain word co-occurrence statistics for instance. [7] propose a Pólya Urn Model to integrate co-occurrence statistics into a topic model. [4] use First Order Logic incorporated into LDA to leverage domain knowledge. [3] incorporate information about words that should or should not be together in a topic from topic model. [6] integrate lexical semantic relations like synonyms or antonyms derived from external dictionaries into a topic model.

In the last years many approaches have been proposed to evaluate topic models. [17] propose to estimate the probability of some held-out documents of the collection used for topic modelling. The authors propose several sampling techniques to efficiently approximate this probability. [10] propose to evaluate topic models based on external information. They use pointwise mutual information (PMI) based on co-occurrence statistics from external text sources to evaluate topics. [11] evaluate topic models by coherence of the topics. They authors showed that the coherence measure agrees with human evaluations of the topics. [1] evaluate topics based on distributional semantics. They find semantic spaces such that words that are semantically related based on statistics on Wikipedia are close in these spaces. [15] developed a framework for measuring coherence in topics. The authors performed large empirical experiments on standard data sets and possible coherence measures to evaluate the framework.

3 Topic Models with Prior Information of Words

We integrate external information into LDA via priors on the word-topic distributions. Similar to the approach by [8], we define an asymmetric Dirichlet prior with metaparameter β on the word topic distribution θ. β specifies the prior believe on the distribution of the words before we have seen any data. We make β dependent on the word distribution from the external information $p(w)$. We set $\beta_{w,t} = exp(\lambda_{w,t}) \cdot p(w)$ for a weight parameter $\lambda_{w,t}$ for the individual influence of the prior information in each topic. If $\lambda_{w,t}$ is zero, the prior believe of the probability of w is directly used. If $\lambda_{w,t}$ is

less than zero, the prior believe is weighted down. If $\lambda_{w,t}$ is greater than zero, the prior believe is weighted up.

The optimal parameters λ must be found by optimizing the likelihood of the topic model. We perform alternating optimization of the parameters with quasi Newton methods and Gibbs sampling of topics to find the optimal topic model.

For the optimization of the parameters we minimize the part of the negative log likelihood from standard LDA that depends on β:

$$L = \sum_t \log \Gamma(\tilde{\beta}_t + n_k) - \log \Gamma(\tilde{\beta}_t) + \sum_t \sum_{w:n_{w,t}>0} \log \Gamma(\beta_{w,t}) - \log \Gamma(\beta_{w,t} + n_{w,t})$$

with $\tilde{\beta}_t = \sum_w \beta_{w,t}$.

3.1 Sparsity Priors for LDA

We propose to use a sparsity inducting priors on the parameter $\lambda_{w,t}$ weights to influence the prior information about word w for topic t. We expect that some parts of the prior information play a bigger role than other parts in the estimated topic model. To find out which parts are important we impose sparsity to identify them.

We add a Laplace prior on the λ parameters to gain sparsity. This means, we aim at reducing the amount of adaptation of the external information. This has three advantages. First, we can easily read from the parameters which parts of the prior information influences the topics. Second, we get a simpler model that adapts the external prior information only for some words. Third, we gain control on the amount of external information to be integrated into the topic model.

The difference to standard LDA is that we have now an asymmetric prior β that is derived from the external information (the word probabilities) and the weight of this information has a Laplace prior. Adding the Laplace prior of the λ parameters of the DMR and optimizing for the negative log-likelihood is the same as putting a sparsity inducing penalty on them. Now, the negative log likelihood is simply extended by $\|\lambda_t\|_1$:

$$L_1 = L + \sigma^{-1} \sum_t \|\lambda_t\|_1.$$

Hence, the Laplace prior is integrated into the optimization via a sparse lasso penalty $\|\lambda\|_1$. We solve the optimization problem via Orthantwise Quasi Newton Optimization [2].

3.2 Group-Sparsity Priors for LDA

The previous idea of limiting the adaptation of the external prior information for some words does not consider that the information about similar words should also be treated similar. For instance, in case the prior information about the word “book” is not adapted, we should also not adapt the information about “author” or “books”.

We propose to add a group lasso penalty to the negative log likelihood to gain group sparsity:

$$L_2 = L + \sigma^{-1}\|\lambda\|_1 + \sum_g \gamma^{-1}\|\lambda_g\|_2$$

for the group lasso penalty $\sum_g \gamma^{-1} \cdot \|\lambda_g\|_2$ for the groups g and the variance γ. Conceptionally, this is the same as having a prior on the λ parameters that induces group sparsity.

Similar to above we solve the group lasso via Blockwise Coordinate Descent with Proximal Operators for the group penalty, see [5] for more details.

3.3 Finding Groups

To find the groups for the grouped sparsity priors on the weight parameters we use external information about similarities of words. From such similarities we can easily generate clusters that are used as groups. We divide the weight parameter $\lambda = (\lambda_1, \cdots, \lambda_G)$ with G partial weights $\lambda_g = (\lambda_{w_1,g}, \cdots \lambda_{w_k,g})$. The partial weights build a group g if the words $w_1, \cdots w_k$ build a cluster based on the similarities from the external information. The similarities we use are based on WordNet (see [13]). We generate a so called affinity matrix M such that $(M)_{ij} = exp(-(1 - sim(w_i, w_j))$ for sim the similarity derived from WordNet. Next, we perform a spectral clustering [12] to find the groups. Spectral clusterings performs a simple k-means clusterings on the words projected onto low-dimensional space spanned by the eigenvectors of the affinity matrix.

4 Experiments

In this section, we investigate the topics extracted by our proposed methods (**SparsePrior**) for LDA with sparsity prior, (**GroupPrior**) for LDA with group sparsity prior) and compare them with two standard state-of-the-art implementations of topic models that integrate external information about words: (**RegLDA**) by [9] and (**WordFeatures**) by [14]. Additionally, we also compare to the standard LDA with Gibbs sampling without external information. For each method, we use $T = 20$ topics, 1000 iterations and set $\alpha = 50/T$, $\beta = 0.1$ (for standard LDA and topic models with structural prior), $\gamma^{-1} = 0.1$, $\sigma^{-1} = 0.1$.

4.1 Data sets

We use two standard text data sets used in previous approaches of topic modelling. First, we use the 20 Newsgroups[1] data set. The data set contains about 20.000 text documents from 20 different newsgroups. Overall we have 1000 documents per newsgroup. We additional remove stop words and prune very infrequent and very frequent words. Second, we use the Senseval-3[2] data set of English lexical samples. The data set contains

[1] http://qwone.com/ jason/20Newsgroups/

[2] http://www.senseval.org/senseval3

Data	20 newsgroups				Wikipedia			
Method	NPMI	UCI	UMASS	nLL	NPMI	UCI	UMASS	nLL
LDA	-0.065	-2.268	-5.250	2332131	-0.065	-2.268	-5.250	2332131
WordFeatures	-0.061	-2.135	-4.825	**2330149**	-0.061	-2.135	-4.825	**2330149**
RegLDA	-0.069	-2.443	-5.520	2332699	-0.069	-2.443	-5.520	2332699
SparePrior	-0.070	-2.472	-5.359	2334633	-0.070	-2.472	-5.359	2334633
GroupPrior	**-0.055**	**-2.116**	**-4.796**	2333298	**-0.055**	**-2.116**	**-4.796**	2333298

Table 1. Results on the different data sets: 20 newsgroups data set and Wikipedia talk pages.

Data	SensEval			
Method	NPMI	UCI	UMASS	MI
LDA	-0.050	-1.712	-3.706	0.359
WordFeatures	-0.058	-1.744	-4.096	0.328
RegLDA	-0.056	-1.767	-3.693	0.323
SparePrior	-0.025	-0.747	-3.060	0.290
GroupPrior	**-0.021**	**-0.634**	**-3.056**	**0.360**

Table 2. Results on the SensEval data set.

texts from Penn Treebank II Wall Street Journal article. The sizes of the data sets range from 20 to 200 documents per word. Further, we use the wikipedia talk pages to apply the method to a more recent data source of internet based communication. As example, we extract 10.000 postings of discussions on wikipedia from 2002 to 2014 that contain the term ”cloud”.

4.2 Coherence Results

In the first experiments, we compare to the state-of-the-art LDA implementations with external information about words and standard LDA in terms of quality. We want to show that our model produces more coherent topics. To evaluate the coherence of

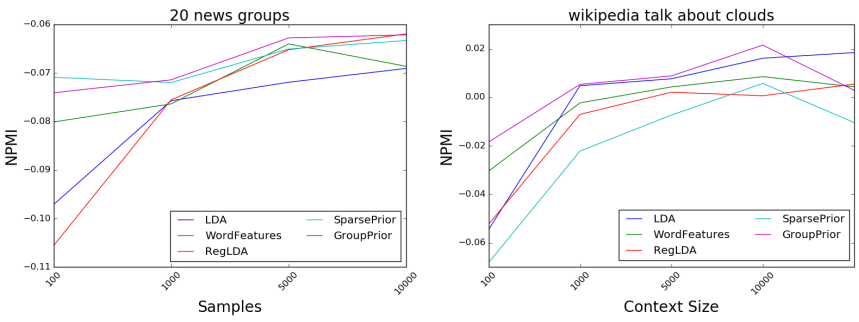

Fig. 1. NPMI for different sample sizes and document length used.

the found topics, we use Pointwise Mutual Information (UCI), normalized Pointwise Mutual Information (NPMI) and arithmetic mean of conditional probability (UMass), see [15]. Further, for the two larger data sets 20 news groups and the postings from wikipedia we also estimate the negative log-likelihood (nLL) on a held out data set. Finally, on the SensEval data set, we also estimate the Mutual Information (MI) of the found topics to the true sense.

The results on the 20 newsgroups data set on the left in Table 1 show that our proposed group sparsity prior results in topic with better coherence measures than the state-of-the art methods and the standard LDA. From the state-of-the-art competitors only **WordFeatures** performs comparably good. In terms of loglikelihood, **WordFeatures** performs best. For the wikipedia talk pages we get similar results as shown on the middle in Table 1.

Finally, we compare the different topic model methods on collection of very small data sets. In Table 2 shows the resulting coherence values on the SensEval data set. LDA with our proposed grouped sparsity prior performs better on all data samples compared to the competitor.

We are especially interested in how the different methods perform on very small data sets. To investigate this, we evaluate the NPMI for the different methods on different sample sizes and different document lengths of the samples. For the 20 news groups date, we sample 100, 100, 5000 and 10000 documents to extract topics. From the wikipedia talk pages we extract postings of different context sizes from 100 to 1000 characters. In Figure 1, we see that our propose sparsity and group sparsity priors results for small samples and small context sizes in the highest NMPI. In these situations our proposed methods of using the group sparsity pays of the most.

5 Conclusion

In this paper we propose to integrate external information about words into topic models to increase topic coherence. We use different priors on the metaparameters for LDA. To control the amount of the integration of the external information we weight them individually. Adding sparsity inducing priors on these weights enables active control on the how much we adapt the external information. By this we trade off topic coherences and likelihood of the topics. Our proposed group sparsity prior further enables integration of external similarity information about words. Now, we can influence the external information of whole groups of words that are similar. The results on large data collections showed the benefit of our proposed method in terms of topic coherence. Finally, we showed that on very small data sets, the group sparsity inducing prior results in better performance.

References

1. Nikolaos Aletras and Mark Stevenson. Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, pages 13–22, Potsdam, Germany, March 2013. Association for Computational Linguistics.
2. Galen Andrew and Jianfeng Gao. Scalable training of l1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pages 33–40, New York, NY, USA, 2007. ACM.
3. David Andrzejewski, Xiaojin Zhu, and Mark Craven. Incorporating domain knowledge into topic modeling via dirichlet forest priors. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 25–32, New York, NY, USA, 2009. ACM.
4. David Andrzejewski, Xiaojin Zhu, Mark Craven, and Benjamin Recht. A framework for incorporating general domain knowledge into latent dirichlet allocation using first-order logic. In *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011*, pages 1171–1177, 2011.
5. Francis Bach, Rodolphe Jenatton, and Julien Mairal. *Optimization with Sparsity-Inducing Penalties (Foundations and Trends(R) in Machine Learning)*. Now Publishers Inc., Hanover, MA, USA, 2011.
6. Zhiyuan Chen, Arjun Mukherjee, Bing Liu, Meichun Hsu, Malu Castellanos, and Riddhiman Ghosh. Discovering coherent topics using general knowledge. In *Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management*, CIKM, pages 209–218, New York, NY, USA, 2013. ACM.
7. David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 262–272, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.
8. David M. Mimno and Andrew McCallum. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. *CoRR*, abs/1206.3278, 2012.
9. David Newman, Edwin V. Bonilla, and Wray L. Buntine. Improving topic coherence with regularized topic models. In John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando C. N. Pereira, and Kilian Q. Weinberger, editors, *NIPS*, pages 496–504, 2011.
10. David Newman, Sarvnaz Karimi, and Lawrence Cavedon. External evaluation of topic models. In *Australasian Document Computing Symposium*, pages 11–18, Sydney, December 2009.
11. David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 100–108, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.
12. Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS*, pages 849–856. MIT Press, 2001.
13. Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. Wordnet::similarity: Measuring the relatedness of concepts. In *Demonstration Papers at HLT-NAACL 2004*, HLT-NAACL–Demonstrations '04, pages 38–41, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.
14. James Petterson, Alexander J. Smola, Tibrio S. Caetano, Wray L. Buntine, and Shravan M. Narayanamurthy. Word features for latent dirichlet allocation. In John D. Lafferty, Christopher K. I. Williams, John Shawe-Taylor, Richard S. Zemel, and Aron Culotta, editors, *NIPS*, pages 1921–1929. Curran Associates, Inc., 2010.

15. Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, pages 399–408, New York, NY, USA, 2015. ACM.
16. Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.
17. Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 1105–1112, New York, NY, USA, 2009. ACM.