Topic Intrusion for Automatic Topic Model Evaluation

Shraey Bhatia[1] **Jey Han Lau**[1,2] **Timothy Baldwin**[1]

[1] School of Computing and Information Systems,
The University of Melbourne

[2] IBM Research Australia

shraeybhatia@gmail.com, jeyhan.lau@gmail.com, tb@ldwin.net

Abstract

Topic coherence is increasingly being used to evaluate topic models and filter topics for end-user applications. Topic coherence measures how well topic words relate to each other, but offers little insight into the utility of the topics in describing the documents. In this paper, we explore the topic intrusion task — the task of guessing an outlier topic given a document and a set of topics — and propose a method to automate it. We improve upon the state-of-the-art substantially, demonstrating its viability as an alternative method for topic model evaluation.

1 Introduction

Topic models have traditionally been evaluated using model perplexity, but there is an increasing trend to use topic coherence as a task-independent evaluation (Newman et al., 2010; Mimno et al., 2011; Aletras and Stevenson, 2013; Lau et al., 2014; Röder et al., 2015). In earlier work (Bhatia et al., 2017), we showed that topic coherence as a standalone evaluation can be misleading, which we illustrated with an adversarial topic model that produces highly coherent topics that collectively tell us little about the content of the document collection.

We went on to explore an alternative approach to assessing topics using topic intrusion, based on the manual task of Chang et al. (2009). In the original topic intrusion setup, users are presented with a document, a set of associated topics (from a topic model) and an intruder topic, and are tasked to find the intruder. Success in the task demonstrates that the topics learnt by the topic model are relevant to the document. We proposed a method to automate this (Bhatia et al., 2017), by training a support vector regression model based on information retrieval (IR) and word co-occurrence features to predict the intruder topic.

Although our earlier method is able to distinguish between good and bad topic models (at the model-level), we provided no evaluation at the document level other than the observation that "there are still slight disparities between human annotators and the automated method in intruder topic selection". Additionally, the method involves a number of dependencies on complex external systems such as Indri, and no implementation of the method was ever released. In this paper, we extend our earlier work (Bhatia et al., 2017) as follows: (1) we improve the results based on a novel neural model and provide additional analysis of document-level evaluation via mean-absolute-error; (2) we propose a new metric to measure the performance of the system; and (3) we release an open source implementation of our system.[1]

2 Related Work

Chang et al. (2009) introduced the word and topic intrusion tasks to assess topic models. Since then, various automatic measures to assess topics have been proposed (Newman et al., 2010; Mimno et al., 2011; Aletras and Stevenson, 2013). Lau et al. (2014) compared and contrasted these approaches, and proposed a variant method based on normalised pointwise mutual information. Röder et al. (2015) conducted a systematic search using a framework that combines various existing measures.

In Bhatia et al. (2017), we revisited the topic intrusion task of Chang et al. (2009), and explored its viability as an alternative task-independent approach for topic model evaluation. We tested a number of topic models and found that there can be large discrepancies between conventional topic coherence measures and topic intrusion results, suggesting that topics can be individually coherent but

[1] Source code and dataset can be downloaded at: https://github.com/sb1992/Topic-Intrusion-for-Automatic-Topic-Model-Evaluation.

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 844–849
Brussels, Belgium, October 31 - November 4, 2018.

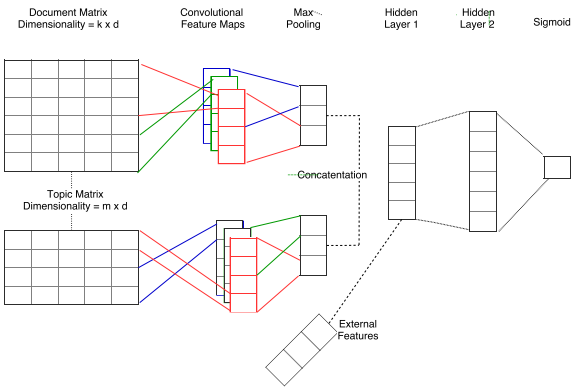


Figure 1: Architecture diagram of our method

poor descriptors of the documents. In addition, we proposed a method to automate the topic intrusion task and reported encouraging correlation levels with human judgements for model-level evaluation.

3 Datasets and Topic Models

We conduct our experiments using the datasets of Bhatia et al. (2017): (1) APNEWS, a collection of Associated Press news articles; and (2) the British National Corpus ("BNC": Burnard (1995)), made up of excerpts from diverse sources such as journals, books, letters, and articles. For the topic models we experiment with the following: standard LDA (`lda`: Blei et al. (2003)), correlated topic model (`ctm`: Blei and Lafferty (2006)), non-parametric topic model (`hca`: Buntine and Mishra (2014)), neural topic model (`ntm`: Cao et al. (2015)), and an adversarial topic model (`cluster`: Bhatia et al. (2017)). `cluster` is adversarial in the sense that it is designed to produce topics that are coherent but poor descriptors of documents.

4 Methodology

In this section, we briefly describe the topic intrusion task and propose an improved methodology to automate it.

4.1 Task

Chang et al. (2009) first proposed the topic intrusion task with the aim of assessing whether topics associated with a document capture its content. In this task, an annotator is presented with a document along with its top-3 highest probability topics and a low probability intruder topic, and are asked to identify the outlier intruder topic. Bhatia et al. (2017) incorporate an additional constraint: the intruder topic has to have high probability for at least one other document. Their rationale is to ensure that the intruder is interpretable. We follow the approach of our earlier work (Bhatia et al., 2017) when generating intruder topics.

4.2 Human Judgements

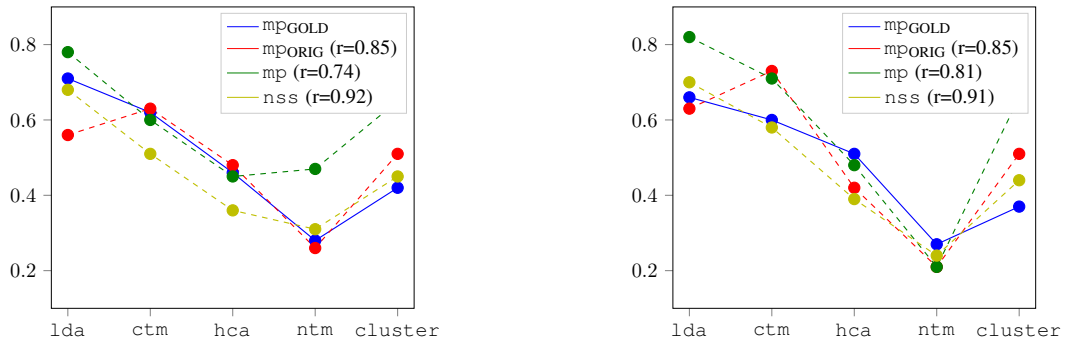
To assess our methodology, we need human annotations for the topic intrusion task. We collect human judgements using Amazon Mechanical Turk. Each HIT is comprised of 5 documents, and each document is paired with 4 topics (3 real and 1 intruder). To control for annotation quality, an additional document–topics pair is inserted as part of the HIT. The control item's intruder topic is generated by randomly sampling words from the corpus vocabulary. To pass the quality control, an annotator has to select the correct intruder topic; they are filtered out if their pass rate over all controls is < 0.6.[2]

Each HIT is judged by 10 workers. We collect additional annotations by releasing the task internally to a small number of local workers. We needed to carry out some annotations internally to make sure that each HIT had at least 4 annotations. The average number of internal annotations was approximately 1.6. For each topic model, we collected annotations for 100 documents on 2 corpora (5 topic models $\times$ 100 documents $\times$ 2 collections $=$ 1000). After filtering and including internal judgements we have an average of 6.7 and 6.9 annotations for APNEWS and BNC, respectively.

4.3 Intruder Topic Detection

We propose a neural network model to automatically predict intruder topics. Our model is inspired by Severyn and Moschitti (2015), where they combine a learn-to-rank deep learning architecture in an IR setting to rank the documents for a given query. We adapt it to our topic intrusion task by ranking topics for a given document. Our task takes the form of a document d_i with corresponding topics $T_i = \{t_i^1, t_i^2, t_i^3, t_i^4\}$, where 3 topics are real and 1 is the intruder. The topic set T_i has labels $Y_i = \{y_i^1, y_i^2, y_i^3, y_i^4\}$ ("1" denotes the intruder topic, or "0" otherwise). We train using a pointwise ranking approach, where training examples

[2] We fixed the threshold to 0.6 based on preliminary experiments. We found that it was a challenging task, and this value provides quality without filtering out most of the workers.



(a) APNEWS (b) BNC

Figure 2: mpGOLD vs. System Scores at the model level

are triples of (d_i, t_i^j, y_i^j) — essentially the task is formulated as a binary classification problem.

The architecture of our network is given in Figure 1. The input to our model is a document–topic pair, with each represented as a sequence of words. These words are mapped to embeddings, via embedding matrix $W \in \mathbb{R}^{|V| \times d}$, where V is the vocabulary and d the dimensionality of the embeddings. The document embeddings $E_d \in \mathbb{R}^{k \times d}$ (k = document length) and topic embeddings $E_t \in \mathbb{R}^{m \times d}$ (m = number of topic words) are processed via convolutional layers (Kim, 2014; Severyn and Moschitti, 2015) to produce two hidden representations for the document and topic. The convolution operation is performed using feature maps of varying size followed by a max-pooling operation to produce a constant-length vector. The document and topic hidden representations are concatenated and fed to 2 dense layers and ultimately reduced to a sigmoid-activated score.

4.3.1 External IR Feature

A good topic model learns common themes in the document collection. A limitation of our network is the lack of global- or collection-level information (as the input consists of only a document and topic). To incorporate collection-level information, we include an IR feature where we query document d_i using the topic words of t_i^j. We use Okapi BM25 (Robertson and Walker, 1994) to compute the relevance score of the document with respect to its N topic words independently, thereby constructing an N-dimensional feature vector.[3] This external feature vector is incorporated into the network after the convolutional layers (see Figure 1).

[3] $N = 5$ in our experiments.

Model	BNC → APNEWS			APNEWS → BNC		
	mpORIG	mp	nss	mpORIG	mp	nss
lda	0.47	0.31	**0.21**	0.40	0.32	**0.22**
ctm	0.44	0.34	**0.20**	0.41	0.31	**0.19**
hca	0.48	0.37	**0.21**	0.42	0.35	**0.20**
ntm	0.40	0.43	**0.19**	0.37	0.32	**0.18**
cluster	0.48	0.42	**0.19**	0.51	0.47	**0.22**
Overall	0.46	0.37	**0.20**	0.42	0.36	**0.21**

Table 1: mae between mpGOLD and nss/mp. "BNC → APNEWS" means the model is trained on BNC and tested on APNEWS. Boldface indicates optimal performance for each dataset.

4.4 Aggregating Human and System Scores for a Document

For each document we have a number of workers identifying the intruder topic. To aggregate the results, Chang et al. (2009) define model precision (mpGOLD), which is the proportion of workers who correctly identified the intruder, as a proxy for how clearly the intruder topic is inappropriate for the document.

Our system and that of Bhatia et al. (2017) compute several scores for a document (one for each topic). Bhatia et al. (2017) select the topic with the maximum score as the intruder, and compute model precision (mp) based on that. This yields binary precision scores (i.e. the model either predicts the intruder correctly or not) and ignores the relative magnitude of the system score. We additionally propose using the normalised sigmoid score (nss) as a means of scoring the intruder topic for a given document, which is computed by normalising the raw sigmoid scores over all topics.

Model	Best/Worst Topics	nss
lda	share revenue cents billion quarter earnings analysts net rose income	0.001
	european greece europe billion debt country crisis minister french france	0.002
ctm	building lodge bauer buildings fee part stephens hall property council	0.007
	military army afghanistan killed soldiers forces troops iraq war attacks	0.013
hca	shares earnings keywords insights profit thomson cents reuters premarket net	0.011
	upheld ruling appeals justices appellate supreme injunction plaintiffs unconstitutional rulings	0.051
ntm	rose shrank pct decliners quadrupled exhibitors parade spectrum index outperform	0.110
	arraigned burglarizing arrested bigamy detectives motorcyclist arraignment coroner accomplice fondled	0.141
cluster	soared plummeted climbed surged dipped tumbled dropped fell slipped rose	0.005
	students teachers kindergarten tutors elementary coursework curriculum teaching tutoring education	0.013
lda	lot good things long put start number making kind place	0.291
	political issue called issues policy decision long change statement support	0.271
ctm	online information internet book video media facebook phone computer technology	0.263
	show music film movie won festival tickets game band play	0.233
hca	richter riverboat sheppard lander plazas tam mandarin amarillo colosseum nassau	0.376
	deplorable interaction foresee envelope handwriting knot quickest scrambled alarmed mum	0.368
ntm	aboard spacewalks bushels budget lifeboats flotilla lifeboat spacewalk millage spaceflight	0.364
	evacuated evacuations evacuate evacuating airlifted twisters aftershocks evacuation driest barricaded	0.323
cluster	accord delegations accords cooperation consultations negotiators negotiation committees intergovernmental negotiations	0.323
	summaries summary critiques excerpts articles responses quotes references descriptions critique	0.309

Table 2: Examples of best topics (top-half) and worst topics (bottom-half) based on nss.

4.4.1 Implementation Details

For our experiments, we train the model on outputs from all topics models over one dataset, and test it on the other (cross-domain training). We use a single channel for the convolutional networks, pad the documents as necessary ($k = 200$), and use the top-10 words to represent a topic (i.e. $m = 10$). Word embeddings are initialised using pre-trained GloVe (Pennington et al., 2014) vectors ($d = 100$), and their weights are fixed during training. We use kernel windows of width $= \{3, 5, 7\}$ with 100 feature maps each and two (fully-connected) hidden layers, with dimensionality of 50 and 10. We use a dropout rate of 0.5, 0.5 and 0.25 after the document, topic and first hidden layer, respectively. We set the batch size to 100, and use Adam as the optimizer with a learning rate of 0.001. For activation functions, we use ReLU for the fully-connected layers and sigmoid for the final layer. To reduce variance, we run the models with 8 different seeds for initialisation and take the average for a topic's sigmoid score.

5 Results

By taking the mean of mp_{GOLD} and mp over documents, Bhatia et al. (2017) compute a single human/system score for each topic model. Although this resulted in a strong correlation between mp_{GOLD} and mp, the evaluation is limited to model-level comparison: it separates good topic models from bad topic models, but does not provide any insights into the performance of each top model over individual documents. We aim to improve model-level correlation in this work, in addition to analysing document-level evaluation, i.e. investigating how well the system predicts mp_{GOLD} for each individual document.

We present plots of human and system scores in Figure 2. There are 3 system scores: mp of Bhatia et al. (2017) (mp_{ORIG}), and mp and nss of our proposed system. In general, we found strong correlation for all systems, but nss of our proposed system performs substantially better than mp_{ORIG}, though our mp is lower than mp_{ORIG}.

To compare the performance of our system with human judgements at the document level, we compute mean absolute error (mae) between mp_{GOLD} and nss/mp, as summarised in Table 1. We find for both datasets nss consistently outperforms mp_{ORIG} and mp by a substantial margin, and also has a score close to human judgements. We can attribute this to the fact that nss provides more nuanced system predictions (over the full range $[0, 1]$), whereas mp tends to be binary.[4]

6 Discussion

One motivation we have in this paper is to test whether topic intrusion can be used as an alternative means for assessing topics. Given the encouraging

[4]Strictly speaking, it is continuous as it is averaged over the runs for the multiple random seeds, but in general, it tends to be (close to) 0 or 1.

mae results, we attempt to use nss to rank topics produced by a topic model.

To accomplish this, we first filter out the topics that occur in less than 5 documents as top 1-topic: these topics tend to be noisy, and as such do not appear with significant weight in any documents. For each of the filtered topics we randomly select 5–10 documents for which it is a top topic and calculate its mean nss over these documents. We then use the topics' mean nss to rank them; in Table 2 we show some selected best and worst topics for different topic models. Overall, the top-ranked topics appear to be more descriptive than the bottom-ranked topics. Having said that, we found instances where coherent topics have low nss ranking (e.g. ctm topics in the bottom half of Table 2), but stress that ultimately the topic intrusion approach to assessing topics is very different to topic coherence. We include a more comprehensive list of best/worst topics in the supplementary material.

7 Conclusion

We explore an alternative approach to evaluate topic models based on topic allocations in documents, i.e. via topic intrusion. We propose an automated method that improves upon the state-of-the-art substantially at the model- and document-levels, and demonstrate that it can be used to rank/filter topics. As future work we intend to explore ways that combine both the topic coherence and topic intrusion for topic model evaluation.

References

Nikos Aletras and Mark Stevenson. 2013. Evaluating topic coherence using distributional semantics. In *Proceedings of the Tenth International Workshop on Computational Semantics (IWCS-10)*, pages 13–22, Potsdam, Germany.

Shraey Bhatia, Jey Han Lau, and Timothy Baldwin. 2017. An automatic approach for document-level topic model evaluation. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 206–215. Association for Computational Linguistics.

David Blei and John Lafferty. 2006. Correlated topic models. *Advances in Neural Information Processing Systems*, 18.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Wray L Buntine and Swapnil Mishra. 2014. Experiments with non-parametric topic models. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 881–890.

Lou Burnard. 1995. User guide for the British National Corpus.

Ziqiang Cao, Sujian Li, Yang Liu, Wenjie Li, and Heng Ji. 2015. A novel neural topic model and its supervised extension. In *Proceedings of AAAI 2015*, pages 2210–2216.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L. Boyd-Graber, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems 21 (NIPS-09)*, pages 288–296, Vancouver, Canada.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *EMNLP*.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of EACL 2014*, pages 530–539.

David Mimno, Hanna 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 (EMNLP 2011)*, pages 262–272, Edinburgh, UK.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

S.E. Robertson and S. Walker. 1994. Some simple effective approximations to the 2-Poisson Model for probabilistic weighted retrieval. In *Proceedings of 17th International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR'94)*, pages 232–241.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM 2015)*, pages 399–408, Shanghai, China.

Aliaksei Severyn and Alessandro Moschitti. 2015. Learning to rank short text pairs with convolutional deep neural networks. In *Proceedings of the 38th International ACM SIGIR Conference on Research*

and Development in Information Retrieval, pages 373–382. ACM.