Topic Quality Metrics Based on Distributed Word Representations

Sergey I. Nikolenko
Laboratory for Internet Studies,
National Research University Higher School of Economics, St. Petersburg,
Steklov Mathematical Institute at St. Petersburg
sergey@logic.pdmi.ras.ru

ABSTRACT

Automated evaluation of topic quality remains an important unsolved problem in topic modeling and represents a major obstacle for development and evaluation of new topic models. Previous attempts at the problem have been formulated as variations on the coherence and/or mutual information of top words in a topic. In this work, we propose several new metrics for evaluating topic quality with the help of distributed word representations; our experiments suggest that the new metrics are a better match for human judgement, which is the gold standard in this case, than previously developed approaches.

Keywords

topic quality; topic modeling; text mining

1. INTRODUCTION

Evaluating topic quality has been an important problem in topic modeling since its very inception. The problem here is that while it is usually immediately evident for a human whether a topic is "good" or not, i.e., whether it is easily interpretable and can serve to draw conclusions regarding the dataset, it is hard to evaluate it automatically. This problem is especially prominent in real-life applications of topic modeling to social sciences, where the goal is usually to get an overview of what the dataset is about and which documents to actually read, which is impossible without interpretable topics [5, 15]. However, it it still an open and interesting problem to devise metrics that would give good approximations to human interpretability; actual human evaluation remains the gold standard here.

In this work, we propose several candidate metrics based on distributed word representations. Word representations have been extensively used in natural language processing; the idea is to map words into some kind of semantic space where geometric relations between vectors will supposedly correspond to semantic relations between the original words.

SIGIR '16, July 17-21, 2016, Pisa, Italy

DOI: http://dx.doi.org/10.1145/2911451.2914720

In this work, we posit that such embeddings and the semantic information that they capture can be leveraged to evaluate topic models, with results significantly improving upon previously known techniques. We perform quantitative evaluation by comparing the ranking produced by automatic quality metrics with rankings produced by human experts asked specifically to evaluate topic interpretability. In other words, we measure interpretability directly by a consensus of human experts and try to approximate it by automated metrics. We show that even very simple metrics that measure how close top words in a topic are in the semantic space still significantly outperform previously used metrics.

The paper is organized as follows. In Section 2, we describe the problem setting in detail and survey related work, including previously known approaches to topic evaluation that we compare with. In Section 3, we describe the new metrics based on distributed representations. Section 4 describes our experimental setup and presents the results, and Section 5 concludes the paper.

2. RELATED WORK

2.1 Topic modeling

We begin by surveying (very briefly due to space constraints) the topic models whose results we try to evaluate. Let D be a finite set (collection) of texts and W a finite set (vocabulary) of all terms from these texts. Probabilistic topic models represent the text collection as a sequence of triples (d_i, w_i, z_i), where d_i is a document, w_i is a word, and z_i is the topic from which w_i has been drawn in this instance; d_i and w_i are observed from the data, while z_i are latent variables. Introducing the word-topic distributions Φ, $\phi_{wt} = p(w \mid t)$, and document-topic distributions Θ, $\theta_{td} = p(t \mid d)$, we get that $p(w \mid d) = \sum_{t \in T} p(w \mid t)\, p(t \mid d) = \sum_{t \in T} \phi_{wt}\theta_{td}$, and the generative process is similar in all topic models: for every word instance, we first sample the topic t_i from distribution $p(t \mid d)$ and then sample the word w_i from distribution $p(w \mid t_i)$.

In the basic *probabilistic latent semantic analysis* (pLSA) model [10], Φ and Θ matrices are learned by directly optimizing the log-likelihood of the training dataset $L(\Phi, \Theta) = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt}\theta_{td}$. In the recently developed approach known as *additive regularization of topic models* (ARTM) [19], the basic pLSA model is augmented with additive regularizers, and Φ and Θ matrices are learned by maximizing a linear combination of $L(\Phi, \Theta)$ and r regulariz-

ers $R_i(\Phi, \Theta)$, $i = 1, \ldots, r$ with *regularization coefficients* τ_i:

$$R(\Phi, \Theta) = \sum_{i=1}^{r} \tau_i R_i(\Phi, \Theta), \quad L(\Phi, \Theta) + R(\Phi, \Theta) \to \max_{\Phi, \Theta}.$$

This makes it easy to devise new regularizers, adding desired properties to the topic model [19, 20].

The latent Dirichlet allocation (LDA) model [3, 4, 8] introduces prior Dirichlet distributions for the vectors of term probabilities in topics $\phi_t \sim \mathrm{Dir}(\beta)$ as well as for the vectors of topic probabilities in documents $\theta_d \sim \mathrm{Dir}(\alpha)$ with parameters β and α respectively. Inference in LDA is usually done via either variational approximations or Gibbs sampling. Over the last decade, LDA has been subject to many extensions (too many to survey them here), each of them presenting either a variational of a Gibbs sampling algorithm for a model that builds upon LDA to incorporate some additional information or additional presumed dependencies.

In each case, the result of learning a topic model can be represented as the Φ and Θ matrices. In this work, we concentrate on evaluating the quality of the Φ word-topic distributions, usually presented to a user as an ordered list of top words for every topic, i.e., words with the largest $\phi_{wt} = p(w \mid t)$.

2.2 Distributed word representations for natural language processing

The modern neural network approaches to natural language processing can be roughly divided into two subproblems: constructing and training new models for individual words (this field is known as word embeddings or distributed word representations) and developing subsequent layers of deep architectures to find syntactic and semantic features while taking into account the context of a word in a sentence and specific problems that a system attempts to solve.

To train distributed word representations, one first constructs a vocabulary with one-hot representations of individual words (where each word is represented with a vector of size equal to vocabulary size with a single 1) and then trains representations for individual words starting from there, basically as a dimensionality reduction problem. For this purpose, researchers have usually employed a model with one hidden layer that attempts to predict the next word based on a window of several preceding words. Then representations learned at the hidden layer are taken to be the word's features; this approach has been applied, for instance in the Polyglot system developed in 2013 [1] and in other methods of learning distributed word representations [17]. A recent study on the performance of various vector space models for word semantic similarity evaluation [16] demostrates that compositions of models such as GloVe and Word2Vec as well as unsupervised one-model approaches show reasonable results for the Russian language (which we use in evaluations since we have expert evaluations available for Russian-language topics).

2.3 Topic quality metrics

Next, we survey the topic quality metrics that we build upon in this work. We begin with *coherence*, proposed as a topic quality metric in [7, 13]. For a topic t characterized by its set of top words W_t, coherence is defined as

$$c(t, W_t) = \sum_{w_1, w_2 \in W_t} \log \frac{d(w_1, w_2) + \epsilon}{d(w_1)},$$

where $d(w_i)$ is the number of documents that contain w_i, $d(w_i, w_j)$ is the number of documents where w_i and w_j cooccur, and ϵ is a smoothing count usually set to either 1 or 0.01. Coherence and word cooccurrence statistics in general have also been used to initialize LDA parameters [18]. However, in a recent work [15] coherence was criticized for a number of shortcomings, primarily because it was found to be too heavily reliant on common words that cooccur often but do not define interpretable topics. To alleviate this, the work [15] proposed a modification of the coherence metric called tf-idf coherence defined as

$$c_{\mathrm{tfidf}}(t, W_t) = \sum_{w_1, w_2 \in W_t} \log \frac{\sum_{d: w_1, w_2 \in d} \mathrm{tfidf}(w_1, d)\mathrm{tfidf}(w_2, d) + \epsilon}{\sum_{d: w_1 \in d} \mathrm{tfidf}(w_1, d)},$$

where the tfidf metric is computed with augmented frequency,

$$\mathrm{tfidf}(w, d) = \mathrm{tf}(w, d) \times \mathrm{idf}(w) = \left(\frac{1}{2} + \frac{f(w, d)}{\max_{w' \in d} f(w', d)}\right) \log \frac{|D|}{|\{d \in D : w \in d\}|},$$

where $f(w, d)$ is the number of occurrences of term w in document d. This skews the metric towards topics with high tfidf scores in top words, since the numerator of the coherence fraction has quadratic dependence on the tfidf scores and the denominator only linear.

Another class of topic quality metrics is based on the notion of *pairwise pointwise mutual information* (PMI) between the top words in a topic. Following a recent work [11], we compute three variations on this idea. For a given ordered set of top words $W_t = (w_1, \ldots, w_N)$ in a topic,

(1) the basic *pairwise PMI* metric [14] is computed as

$$\mathrm{PMI}_t = \sum_{i<j} \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)};$$

(2) the *normalized PMI* variation [6] is computed as

$$\mathrm{NPMI}_t = \sum_{i<j} \frac{\log \frac{p(w_i, w_j)}{p(w_i)p(w_j)}}{-\log p(w_i, w_j)};$$

(3) the *pairwise log conditional probability* (LCP) metric [13] is computed as

$$\mathrm{LCP}_t = \sum_{i<j} \log \frac{p(w_i, w_j)}{p(w_i)}.$$

3. TOPIC QUALITY METRICS BASED ON WORD VECTORS

In this section, we propose a number of metrics for evaluating topic quality based on distributed word representations. Suppose that a vector space model has been trained, and every vocabulary word $w \in W$ has been assigned with a vector $v_w \in \mathbb{R}^d$. The basic assumption in our metrics is that a good metric should be well localized in the semantic space: specifically, top words in a topic should be close to each other in the semantic space.

Hence, we use the following general scheme for topic quality metrics: given a set of top words W_t for a topic t with

weights (word probabilities) ϕ_{wt}, $w \in W_t$, their distributed representations $c_w \in \mathbb{R}^d$, and a distance function $d : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, we define topic quality as the average distance between the top words in the topic:

$$\mathrm{Q}_t = \frac{1}{|W_t|(|W_t|-1)} \sum_{w_1 \neq w_2 \in W_t} d(v_{w_1}, v_{w_2}).$$

If $d(w_1, w_2)$ is a distance function in the space $\mathbb{R}^d$, larger results correspond to, supposedly, worse topics (with words not as localized as in topics with smaller average distances). We have compared four different distance-like metrics:

(1) cosine distance $d_{\cos}(x, y) = 1 - x^\top y$ (no normalization here since we have used normalized word vectors);

(2) L_1-distance $d_{L_1}(x, y) = \sum_{i=1}^d |x_i - y_i|$;

(3) L_2-distance $d_{L_2}(x, y) = \sum_{i=1}^d (x_i - y_i)^2$;

(4) coordinate distance $d_{\text{coord}}(x, y) = \sum_{i=1}^d [x_i - y_i > \eta]$, where $[A] = 1$ if A holds and 0 otherwise; the intuition here was that different dimensions of the semantic space may represent different aspects of a word's semantics and may be incomparable directly, so it may make sense to simply count in how many coordinates the two vectors differ significantly; the threshold η was tuned by hand to give reasonable results.

4. EVALUATION

4.1 Datasets and experimental setup

The core of all proposed metrics are word vector representations. We used the skip-gram word2vec model of dimension 500 trained on a large Russian language corpus [2, 16]; the corpus consisted of:

- Russian Wikipedia: 1.15M documents, 238M tokens;
- web crawl data: 890K documents, 568M tokens;
- *lib.rus.ec* library: 234K documents, 12.9G tokens.

A large collection of general-purpose texts ensured good resulting distributed representations; for an in-depth description of the model we refer to [2, 16].

To evaluate the proposed approach, we have used a dataset with approximately 1.58 million lemmatized posts from the top *LiveJournal* bloggers, all in Russian; the Russian language was chosen since we had experts estimates available for topic interpretability only in Russian. The complete vocabulary amounted to 860K words, but after preprocessing it was reduced to 90K words; after dictionary reduction and filtering, there remained about 1.38 million nonempty documents. We have trained all models with $T = 400$ topics, a number chosen by training pLSA models with 100, 300, and 400 topics and evaluating the results.

We have evaluated the quality of six different topic models; since the human coding results were obtained as part of a case study for mining ethnic-related content, two models work specifically with ethnonyms, but in each case the assessors simply evaluated top words in every topic:

(1) probabilistic latent semantic analysis model (pLSA);

(2) latent Dirichlet allocation model (LDA);

(3) ARTM model with smoothing and sparsity regularizers;

(4) ARTM model with a decorrelation regularizer;

(5) ARTM model with a separate modality for ethnonyms, small dictionary of ethnonyms;

(6) ARTM model with a separate modality for ethnonyms, large dictionary of ethnonyms.

Human assessors were asked to interpret the topics based on 20 most probable words in every topic of each model. For each topic, assessors answered the following question: "Do you understand why these words are collected together in this topic?". They were also told that the idea was to see whether the topic was generally understandable and given three options: (a) absolutely not; (b) partially; (c) yes.

The "correct" (human-generated) ranking of topics was produced according to the total number of positive answers, counting (b) as 1 and (c) as 2. For comparison, we have computed five previously known metrics: coherence, tf-idf coherence, PMI, NPMI, and LPC[1], and four word2vec-based metrics with four different distances shown in Section 3.

To evaluate how well a metric matches this ranking, we have computed, for each subject and each metric, the *area-under-curve* (AUC) measure [9, 12]. AUC is a popular quality metric for classifiers that produce ranking results; by definition it represents the probability that for a uniformly selected pair consisting of a positive and a negative example the classifier ranks the positive one higher. Thus, the optimal AUC is 1 (all positive examples come before negative ones), the worst possible AUC is 0, and a random classifier would get an AUC of 0.5. In our case AUC is the perfect fit since the actual values of a metric are mostly irrelevant, and the users are interested in the ranking (to view the best topics from a dataset).

Results of our experiments are shown in Table 1. We see that the new word2vec-based metrics outperform, often significantly, previously known approaches. As expected, tf-idf coherence is better than regular coherence and NPMI is better than PMI, but all of them lose to vector space metrics in most cases. There is little difference between the new metrics themselves, but we can recommend to use L_1 and L_2 metrics.

5. CONCLUSION

In this work, we have proposed a number of simple metrics based on distributed word representations for the purpose of evaluating topic quality in topic modeling results. We have shown that the new metrics outperform previously used topic quality metrics in terms of agreeing with human interpretation. The metrics introduced in this work are rather straightforward; in further work we expect to significantly improve upon our results shown here by learning distributions in the semantic space that are better aligned with actual topics. For instance, it is perfectly plausible for a topic to have several clusters of closely matching words that describe the same issue from different sides; in this case, it would probably be beneficial to consider a clustering model. We expect new exciting developments along these lines, although even the basic metrics proposed here already outperform existing state of the art topic evaluation approaches.

[1]To compute PMI, NPMI, and LPC we have used the companion software for [11] available at https://github.com/jhlau/topic_interpretability.

Model	Quality metrics								
	Coherence		Mutual information			Word2vec metrics			
	regular	tf-idf	PMI	NPMI	LPC	cos	L_1	L_2	coord
1 (pLSA)	0.7720	0.8910	0.8675	0.8707	0.8811	0.8954	0.8957	**0.8965**	0.8832
2 (LDA)	0.7817	0.8748	0.8469	0.8372	0.8541	0.8786	0.8797	**0.8806**	0.8692
3 (ARTM smoothing + sparsity)	0.7513	0.8439	0.6637	0.6973	0.7738	0.8543	**0.8616**	0.8589	0.8597
4 (3 + decorrelation)	0.6783	0.8635	0.7571	0.7821	0.8408	0.8675	**0.8718**	0.8704	0.8605
5 (4 + ethnic modality, small vocab.)	0.7425	0.8924	0.8791	0.8865	0.8885	0.8889	**0.8920**	0.8906	0.8848
6 (5 + ethnic modality, large vocab.)	0.7857	0.8767	0.8526	0.8430	0.8672	0.8812	**0.8827**	0.8815	0.8588

Figure 1: Experimental comparison between LDA topic quality metrics: area under curve (AUC) comparison metrics between human-generated interpretability evaluation and automatic quality metrics.

Acknowledgements

This work was supported by the Russian Science Foundation grant no. 15-18-00091. I thank Olessia Koltsova and Sergei Koltcov for the experimental human coding data, Konstantin Vorontsov, Anna Potapenko, Murat Apishev, and Oleksander Frei for the topics themselves, and Alexander Panchenko and Nikolay Arefyev for the Russian-language training corpora used for the *word2vec* models.

6. REFERENCES

[1] R. Al-Rfou, B. Perozzi, and S. Skiena. Polyglot: Distributed word representations for multilingual NLP. In *Proc. 17th Conference on Computational Natural Language Learning*, pages 183–192, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[2] N. Arefyev, A. Panchenko, A. Lukanin, O. Lesota, and P. Romanov. Evaluating three corpus-based semantic similarity systems for russian. In *Proc. International Conference on Computational Linguistics Dialogue*, 2015.

[3] D. M. Blei. Introduction to probabilistic topic models. *Communications of the ACM*, 2011.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(4–5):993–1022, 2003.

[5] S. Bodrunova, S. Koltsov, O. Koltsova, S. I. Nikolenko, and A. Shimorina. Interval semi-supervised LDA: Classifying needles in a haystack. In *Proc. 12th Mexican International Conference on Artificial Intelligence*, volume 8625 of *Lecture Notes in Computer Science*, pages 265–274. Springer, 2013.

[6] G. Bouma. Normalized (pointwise) mutual information in collocation extraction. In *Proc. Biennial GSCL Conference*, pages 31–40, 2013.

[7] J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang, and D. M. Blei. Reading tea leaves: How humans interpret topic models. *Advances in Neural Information Processing Systems*, 20, 2009.

[8] T. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101 (Suppl. 1):5228–5335, 2004.

[9] D. J. Hand and R. J. Till. A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learning*, 45:171–186, 2001.

[10] T. Hoffmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1):177–196, 2001.

[11] J. H. Lau, D. Newman, and T. Baldwin. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *EACL*, pages 530–539, 2014.

[12] C. X. Ling, J. Huang, and H. Zhang. AUC: a statistically consistent and more discriminating measure than accuracy. In *Proc. International Joint Conference on Artificial Intelligence 2003*, pages 519–526, 2003.

[13] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum. Optimizing semantic coherence in topic models. In *Proc. Conference on Empirical Methods in Natural Language Processing*, pages 262–272, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[14] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin. Automatic evaluation of topic coherence. In *Human Language Technologies 2010*, HLT '10, pages 100–108, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[15] S. I. Nikolenko, O. Koltsova, and S. Koltsov. Topic modelling for qualitative studies. *Journal of Information Science*, 2015.

[16] A. Panchenko, N. Loukachevitch, D. Ustalov, D. Paperno, C. M. Meyer, and N. Konstantinova. Russe: The first workshop on Russian semantic similarity. In *Proc. International Conference on Computational Linguistics and Intellectual Technologies (Dialogue)*, pages 89–105, May 2015.

[17] J. Pennington, R. Socher, and C. Manning. GloVe: Global vectors for word representation. In *Proc. 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.

[18] A. S. Rathore and D. Roy. Performance of LDA and DCT models. *Journal of Information Science*, 40(3):281–292, 2014.

[19] K. Vorontsov. Additive regularization for topic models of text collections. *Doklady Mathematics*, 89(3):301–304, 2014.

[20] K. V. Vorontsov and A. A. Potapenko. Additive regularization of topic models. *Machine Learning, Special Issue on Data Analysis and Intelligent Optimization with Applications*, 101(1):303–323, 2015.