Measuring the Similarity between Automatically Generated Topics

Nikolaos Aletras and **Mark Stevenson**
Department of Computer Science,
University of Sheffield,
Regent Court, 211 Portobello,
Sheffield,
United Kingdom S1 4DP
{n.aletras, m.stevenson}@dcs.shef.ac.uk

Abstract

Previous approaches to the problem of measuring similarity between automatically generated topics have been based on comparison of the topics' word probability distributions. This paper presents alternative approaches, including ones based on distributional semantics and knowledge-based measures, evaluated by comparison with human judgements. The best performing methods provide reliable estimates of topic similarity comparable with human performance and should be used in preference to the word probability distribution measures used previously.

1 Introduction

Topic models (Blei et al., 2010) have proved to be useful for interpreting and organising the contents of large document collections. It seems intuitively plausible that some automatically generated topics will be similar while others are dis-similar. For example, a topic about basketball (*team game james season player nba play knicks coach league)* is more similar to a topic about football (*world cup team soccer africa player south game match goal*) than one about the global finance (*fed financial banks federal reserve bank bernanke rule crisis credit*). Methods for automatically determining the similarity between topics have several potential applications, such as analysis of corpora to determine topics being discussed (Hall et al., 2008) or within topic browsers to decide which topics should be shown together (Chaney and Blei, 2012; Gretarsson et al., 2012; Hinneburg et al., 2012).

Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is a popular type of topic model but cannot capture such correlations unless the semantic similarity between topics is measured. Other topic models, such as the Correlated Topic Model (CTM) (Blei and Lafferty, 2006), overcome this limitation and identify correlations between topics.

Approaches to identifying similar topics for a range of tasks have been described in the literature but they have been restricted to using information from the word probability distribution to compare topics and have not been directly evaluated. Word distributions have been compared using a variety of measures such as KL-divergence (Li and McCallum, 2006; Wang et al., 2009; Newman et al., 2009), cosine measure (He et al., 2009; Ramage et al., 2009) and the average Log Odds Ratio (Chaney and Blei, 2012). Kim and Oh (2011) also applied the cosine measure and KL-Divergence which were compared with four other measures: Jaccard's Coefficient, Kendall's τ coefficient, Discount Cumulative Gain and Jensen Shannon Divergence (JSD).

This paper compares a wider range of approaches to measuring topic similarity than previous work. In addition these measures are evaluated directly by comparing them against human judgements.

2 Measuring Topic Similarity

We compare measures based on word probability distributions (Section 2.1), distributional semantic methods (Sections 2.2-2.4), knowledge-based approaches (Section 2.5) and their combination (Section 2.6).

2.1 Topic Word Probability Distribution

We first experimented with measures based on comparison of the topics' word distributions (see Section 1), by applying the JSD, KL-divergence and Cosine approaches and the Log Odds Ratio (Chaney and Blei, 2012).

2.2 Topic Model Semantic Space

The semantic space generated by the topic model can be used to represent the topics and the topic words. By definition each topic is a probability distribution over the words in the training corpus. For a corpus with D documents and V words, a topic model learns a relation between words and topics, T, as a $T \times V$ matrix, $\mathbf{W}$, that indicates the probability of each word in each topic. $\mathbf{W}$ is the topic model semantic space and each topic word can be represented as a vector, V_i, with topics as features weighted by the probability of the word in each topic. The similarity between two topics is computed as the average pairwise cosine similarity between their top-10 most probable words (**TS-Cos**).

2.3 Reference Corpus Semantic Space

Topic words can also be represented as vectors in a semantic space constructed from an external source. We adapt the method proposed by Aletras and Stevenson (2013) for measuring topic coherence using distributional semantics[1].

Top-N Features A semantic space is constructed considering only the top n most frequent words in Wikipedia (excluding stop words) as context features. Each topic word is represented as a vector of n features weighted by computing the Pointwise Mutual Information (PMI) (Church and Hanks, 1989) between the topic word and each context feature, $\text{PMI}(w_i, w_j)^\gamma$. γ is a variable for assigning more importance to higher PMI values. In our experiments, we set $\gamma = 3$ and found that the best performance is obtained for $n = 5000$. Similarity between two topics is defined as the average cosine similarity of the topic word vectors (**RCS-Cos-N**).

Topic Word Space Alternatively, we consider only the top-10 topic words from the two topics as context features to generate topic word vectors. Then, topic similarity is computed as the pairwise cosine similarity of the topic word vectors (**RCS-Cos-TWS**).

Word Association Topic similarity can also be computed by applying word association measures directly. Newman et al. (2010) measure topic coherence as the average PMI between the topic words. This approach can be adapted to measure topic similarity by computing the average pairwise PMI between the topic words in two topics (**PMI**).

2.4 Training Corpus Semantic Space

Term-Document Space A matrix $\mathbf{X}$ can be created using the training corpus. Each term (row) represents a topic word vector. Element x_{ij} in $\mathbf{X}$ is the tf.idf of the term i in document j. Topic similarity is computed as the pairwise cosine similarity of the topic word vectors (**TCS-Cos-TD**).

Word Co-occurrence in Training Documents Alternatively, we generate a matrix $\mathbf{Z}$ of co-document frequencies. The matrix $\mathbf{Z}$ consists of V rows and columns representing the V vocabulary words. Element z_{ij} is the log of the number of documents that contains the words i and j normalised by the document frequency, DF, of the word j. Mimno et al. (2011) introduced that metric to measure topic coherence. We adapted it to estimate topic similarity by aggregating the co-document frequency of the words between two topics (**Doc-Co-occ**).

2.5 Knowledge-based Methods

UKB (Agirre et al., 2009) is used to generate a probability distribution over WordNet synsets for each word in the vocabulary V of the topic model using the Personalized PageRank algorithm. The similarity between two topic words is calculated by transforming these distributions into vectors and computing the cosine metric. The similarity between two topics is computed by measuring pairwise similarity between their top-10 topic words and selecting the highest score.

Explicit Semantic Analysis (ESA) proposed by Gabrilovich and Markovitch (2007) transforms the topic keywords into vectors that consist of Wikipedia article titles weighted by their relevance to the keyword. For each topic, the centroid is computed from the keyword vectors. Similarity between topics is computed as the cosine similarity of the ESA centroid vectors.

2.6 Feature Combination Using SVR

We also evaluate the performance of a support vector regression system (**SVR**) (Vapnik, 1998) with a linear kernel using a combination of approaches described above as features[2]. The system is trained and tested using 10-fold cross validation.

[1] Wikipedia is used as a reference corpus to count word co-occurrences and frequencies using a context window of ± 10 words centred on a topic word.

[2] With the exception of JSD, features based on the topics' word probability distributions were not used by SVR since it was found that including them reduced performance.

3 Evaluation

Data We created a data set consisting of pairs of topics generated by two topic models (LDA and CTM) over two document collections using different numbers of topics. The first consists of 47,229 news articles from New York Times (NYT) in the GigaWord corpus and the second contains 50,000 articles from ukWAC (Baroni et al., 2009). Each article is tokenised then stop words and words appearing fewer than five times in the corpora removed. This results in a total of 57,651 unique tokens for the NYT corpus and 72,672 for ukWAC.

LDA Topics are learned by training LDA models over the two corpora using *gensim*[3]. The number of topics is set to $T = 50, 100, 200$ and hyperparameters, α and β, are set to $\frac{1}{T}$. Randomly selecting pairs of topics will result to a data set in which the majority of pairs would not be similar. We overcome that problem by assuming that the JSD between likely relevant pairs will be low while it will be higher for less relevant pairs of topics. We selected 800 pairs of topics. 600 pairs represent topics with similar word distributions (in the top 6 most relevant topics ranked by JSD). The remaining 200 pairs were selected randomly.

CTM is trained using the EM algorithm[4]. The number of topics to learn is set to $T = 50, 100, 200$ and the rest of the settings are set to their default values. The topic graph generated by CTM was used to create all the possible pairs between topics that are connected. This results in a total of 70, 468 and 695 pairs in NYT, and a total of 80, 246 and 258 pairs in ukWAC for the 50, 100 and 200 topics respectively.

Incoherent topics are removed using an approach based on distributional semantics (Aletras and Stevenson, 2013). Each topic is represented using the top 10 words with the highest marginal probability.

Human Judgements of Topic Similarity were obtained using an online crowdsourcing platform, Crowdflower. Annotators were provided with pairs of topics and were asked to judge how similar the topics are by providing a rating on a scale of 0 (completely unrelated) to 5 (identical). The average response for each pair was calculated in order to create the final similarity judgement for use as a gold-standard. The average Inter-Annotator agreement (IAA) across all pairs for all of the collections is in the range of 0.53-0.68. The data set together with gold-standard annotations is freely available[5].

4 Results

Table 1 shows the correlation (Spearman) between the topic similarity metrics described in Section 2 and average human judgements for the LDA and CTM topic pairs. It also shows the performance of a **Word Overlap** baseline which measures the number of terms that two topics have in common normalised by the total number of topic terms.

The correlations obtained using the topics' word probability distributions (Section 2.1), i.e. JSD, KL-divergence and Cos, are comparable with the baseline for all of the topic collections and topic models. The metric proposed by Chaney and Blei (2012) also compares probability distributions and fails to perform well on either data set. These results suggest that these metrics may be sensitive to the high dimensionality of the vocabulary. They also assign high similarity to topics that contain ambiguous words, resulting in low correlations with human judgements.

Performance of the cosine of the word vector (TS-Cos) in the Topic Model Semantic Space (Section 2.2) varies implying that the quality of the latent space generated by LDA and CTM is sensitive to the number of topics.

The similarity metrics that use the reference corpus (Section 2.3) consistently produce good correlations for topic pairs generated using both LDA and CTM. The best overall correlation for a single feature in most cases is obtained using average PMI (in a range of 0.43-0.74). The performance of the distributional semantic metric using the Topic Word Space (RCS-Cos-TWS) is comparable and slightly lower for the top-N features (RCS-Cos-N). This indicates that the reference corpus covers a broader range of semantic subjects than the latent space produced by the topic model.

When the term-document matrix from the training corpus is used as a vector space (Section 2.4) performance is worse than when the reference corpus is used. In addition, using co-document frequency derived from the training corpus does not correlate particularly well with human judgements. These methods are sensitive to the size of the corpus, which may be too small to gener-

[3] http://radimrehurek.com/gensim

[4] http://www.cs.princeton.edu/~blei/ctm-c/index.html

[5] http://staffwww.dcs.shef.ac.uk/people/N.Aletras/resources/topicSim.tar.gz

	Spearman's r											
	LDA						CTM					
	NYT			ukWAC			NYT			ukWAC		
Method	50	100	200	50	100	200	50	100	200	50	100	200
Baseline												
Word Overlap	0.32	0.40	0.51	0.22	0.32	0.41	0.56	0.45	0.49	0.35	0.33	0.53
Topic Word Probability Distribution												
JSD	0.37	0.44	0.53	0.29	0.30	0.34	0.59	0.43	0.49	0.38	0.34	0.60
KL-Divergence	0.29	0.29	0.41	0.20	0.24	0.33	0.54	0.39	0.56	0.31	0.29	0.47
Cos	0.31	0.37	0.59	0.30	0.30	0.36	0.58	0.45	0.52	0.50	0.40	0.58
Chaney and Blei (2012)	0.16	0.26	0.18	0.29	0.21	0.25	0.29	0.40	0.31	-0.23	0.12	0.61
Topic Model Semantic Space												
TS-Cos	0.35	0.41	0.67	0.29	0.35	0.42	0.67	0.51	0.49	0.51	0.42	0.42
Reference Corpus Semantic Space												
RCS-Cos-N	0.37	0.46	0.61	0.35	0.32	0.39	0.60	0.47	0.61	0.57	0.42	0.41
RCS-Cos-TWS	0.40	0.54	0.70	0.38	0.43	0.51	0.63	0.59	0.62	0.60	0.55	0.54
PMI	0.43	0.63	0.74	0.43	0.53	0.64	0.68	0.70	**0.64**	0.58	0.62	0.64
Training Corpus Semantic Space												
TCS-Cos-TD	0.36	0.42	0.67	0.29	0.31	0.40	0.64	0.54	0.58	0.49	0.43	0.43
Doc-Co-occ	0.28	0.29	0.45	0.28	0.22	0.30	0.65	0.36	0.57	0.31	0.26	0.34
Knowledge-based												
UKB	0.25	0.38	0.56	0.22	0.35	0.41	0.52	0.41	0.40	0.41	0.43	0.42
ESA	0.43	0.58	0.71	**0.46**	0.55	0.61	0.69	0.67	**0.64**	**0.70**	0.62	0.61
Feature Combination												
SVR	**0.46**	**0.64**	**0.75**	**0.46**	**0.58**	**0.66**	**0.72**	**0.71**	0.62	0.60	**0.65**	**0.66**
IAA	0.54	0.58	0.61	0.53	0.56	0.60	0.68	0.68	0.64	0.67	0.63	0.64

Table 1: Results for various approaches to topic similarity. All correlations are significant $p < 0.001$. Underlined scores denote best performance of a single feature. Bold denotes best overall performance.

ate reliable estimates of tf.idf or co-document frequency.

ESA, one of the knowledge-based methods (Section 2.5), performs well and is comparable to (or in some cases better than) PMI. UKB does not perform particularly well because the topics often contain named entities that do not exist in WordNet. ESA is based on Wikipedia and does not suffer from this problem. Overall, metrics for computing topic similarity based on rich semantic resources (e.g. Wikipedia) are more appropriate than metrics based on the topic model itself because of the limited size of the training corpus.

Combining the features using SVR gives the best overall result for LDA (in the range 0.46-0.75) and CTM (0.60-0.72). However, the feature combination performs slightly lower than the best single feature in two cases when CTM is used (T=200, NYT and T=50, ukWAC). Analysis of the coefficients produced by the SVR in each fold demonstrated that including JSD and the Word Overlap reduce SVR performance. We repeated the experiments by removing these features[6] which resulted in higher correlations (0.64 and 0.65 respectively).

Another interesting observation is that using LDA the correlations of the various similarity metrics with human judgements increase with the number of topics for both corpora. This result is consistent with the findings of Stevens et al. (2012) that topic model coherence increases with the number of topics. Fewer topics makes the task of identifying similar topics more difficult because it is likely that they will contain some terms that do not relate to the topic's main subject. Correlations in CTM are more stable for different number of topics because of the nature of the model, the pairs have been generated using the topic graph which by definition contains correlated topics.

5 Conclusions

We explored the task of determining the similarity between pairs of automatically generated topics and described a range of approaches to the problem. We constructed a data set of pairs of topics generated by two topic models, LDA and CTM, together with human judgements of similarity. The data set was used to evaluate a wide range of approaches. The most interesting finding is the poor performance of the metrics based on word probability distributions previously used for this task. Our results demonstrate that word association measures, such as PMI, and state-of-the-art textual similarity metrics, such as ESA, are more appropriate.

[6]These features are useful for the other experiments since performance drops when they are removed.

References

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Pasca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT '09)*, pages 19–27, Boulder, Colorado.

Nikolaos Aletras and Mark Stevenson. 2013. 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.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226.

David Blei and John Lafferty. 2006. Correlated topic models. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 147–154. MIT Press, Cambridge, MA.

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

David Blei, Lawrence Carin, and David Dunson. 2010. Probabilistic topic models. *Signal Processing Magazine, IEEE*, 27(6):55–65.

Allison June-Barlow Chaney and David M. Blei. 2012. Visualizing topic models. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, Dublin, Ireland.

Kenneth Ward Church and Patrick Hanks. 1989. Word association norms, mutual information, and lexicography. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, pages 76–83, Vancouver, British Columbia, Canada.

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI '07)*, pages 1606–1611.

Brynjar Gretarsson, John O'Donovan, Svetlin Bostandjiev, Tobias Höllerer, Arthur Asuncion, David Newman, and Padhraic Smyth. 2012. TopicNets: Visual analysis of large text corpora with topic modeling. *ACM Trans. Intell. Syst. Technol.*, 3(2):23:1–23:26.

David Hall, Daniel Jurafsky, and Christopher D. Manning. 2008. Studying the history of ideas using topic models. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 363–371, Honolulu, Hawaii.

Qi He, Bi Chen, Jian Pei, Baojun Qiu, Prasenjit Mitra, and Lee Giles. 2009. Detecting topic evolution in scientific literature: how can citations help? In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM '09)*, pages 957–966, Hong Kong, China.

Alexander Hinneburg, Rico Preiss, and René Schröder. 2012. TopicExplorer: Exploring document collections with topic models. In Peter A. Flach, Tijl Bie, and Nello Cristianini, editors, *Machine Learning and Knowledge Discovery in Databases*, volume 7524 of *Lecture Notes in Computer Science*, pages 838–841. Springer Berlin Heidelberg.

Dongwoo Kim and Alice Oh. 2011. Topic chains for understanding a news corpus. In *Computational Linguistics and Intelligent Text Processing*, pages 163–176. Springer.

Wei Li and Andrew McCallum. 2006. Pachinko allocation: Dag-structured mixture models of topic correlations. In *Proceedings of the 23rd International Conference on Machine Learning (ICML '06)*, pages 577–584.

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*, pages 262–272, Edinburgh, Scotland, UK.

David Newman, Arthur Asuncion, Padhraic Smyth, and Max Welling. 2009. Distributed algorithms for topic models. *J. Mach. Learn. Res.*, 10:1801–1828.

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 (NAACL-HLT '10)*, pages 100–108, Los Angeles, California.

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. 2009. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP '09)*, pages 248–256, Singapore.

Keith Stevens, Philip Kegelmeyer, David Andrzejewski, and David Buttler. 2012. Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP '12)*, pages 952–961, Jeju Island, Korea.

Vladimir N Vapnik. 1998. *Statistical learning theory*. Wiley, New York.

Xiang Wang, Kai Zhang, Xiaoming Jin, and Dou Shen. 2009. Mining common topics from multiple asynchronous text streams. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining (WSDM '09)*, pages 192–201, Barcelona, Spain.