Re-Ranking Words to Improve Interpretability of Automatically Generated Topics

Areej Alokaili[1,2], **Nikolaos Aletras**[1] and **Mark Stevenson**[1]
[1]University of Sheffield, United Kingdom
[2]King Saud University, Saudi Arabia
{areej.okaili,n.aletras,mark.stevenson}@sheffield.ac.uk

Abstract

Topics models, such as LDA, are widely used in Natural Language Processing. Making their output interpretable is an important area of research with applications to areas such as the enhancement of exploratory search interfaces and the development of interpretable machine learning models. Conventionally, topics are represented by their n most probable words, however, these representations are often difficult for humans to interpret. This paper explores the re-ranking of topic words to generate more interpretable topic representations. A range of approaches are compared and evaluated in two experiments. The first uses crowdworkers to associate topics represented by different word rankings with related documents. The second experiment is an automatic approach based on a document retrieval task applied on multiple domains. Results in both experiments demonstrate that re-ranking words improves topic interpretability and that the most effective re-ranking schemes were those which combine information about the importance of words both within topics and their relative frequency in the entire corpus. In addition, close correlation between the results of the two evaluation approaches suggests that the automatic method proposed here could be used to evaluate re-ranking methods without the need for human judgements.

1 Introduction

Probabilistic topic modelling (Blei, 2012) is a widely used approach in Natural Language Processing (Boyd-Graber et al., 2017) with applications to areas such as enhancing exploratory search interfaces (Chaney and Blei, 2012; Aletras et al., 2014; Smith et al., 2017; Aletras et al., 2017) and developing interpretable machine learning models (Paul, 2016). A topic model, e.g. Latent Dirichlet Allocation (LDA) (Blei et al., 2003) learns a low-dimensional representation of documents as a mixture of latent variables called topics. Topics are multinomial distributions over a predefined vocabulary of words.

Traditionally, topics have been represented by lists of the topic's n most probable words, however it is not always straightforward to interpret them due to noisy or domain specific data, spurious word co-occurrences and highly-frequent/low-informative words assigned with high probability (Chang et al., 2009).

Improving the interpretability of topic models is an important area of research. A range of approaches have been developed including computing topic coherence (Newman et al., 2010; Mimno et al., 2011; Aletras and Stevenson, 2013a; Lau et al., 2014), determining optimal topic cardinality (Lau and Baldwin, 2016), labelling topics text and/or images (Lau et al., 2011; Aletras and Stevenson, 2014, 2013b; Aletras and Mittal, 2017; Sorodoc et al., 2017) and corpus pre-processing (Schofield et al., 2017). However, methods for re-ranking topic words to improve topic interpretability have not been systematically evaluated yet. We hypothesise that some words relevant to a particular topic have not been assigned with a high probability due to data sparseness or low frequency in the training corpus. Our goal is to identify these words and re-rank the list representing the topic to make it more comprehensible. Table 1 shows topics represented by the 30 most probable words. Words displayed in bold font are more general (less informative, e.g. with high document frequency) while the remaining words are more likely to represent

Table 1: Examples of topics represented by 30 most probable words from New York Times. Less informative words are shown in **bold**.

Topic
space museum **years** history science earth mission **could** art shuttle universe flight **people theory world** radar crew **site** pincus plane **three** scientists **day century** pilot exhibit **back anniversary** landing **project**
percent million market company stock **billion sales** bank shares **price business** investors **money** share **companies rates** fund **interest rate quarter prices** investment funds financial amp analysts **growth industry york** banks
film **even** movie **world** stars **man much little story good way** star **best** show **see well seems american people love** hollywood director **big ever rating though great seem** production **makes**
officials agency office report department investigation government **former** federal **charges secret information card** cia law agents security **documents case** investigators **official** fraud intelligence illegal **commission service** police **cards** enforcement attorney

a coherent thematic subject. For example in the second topic, relevant words (e.g. investment, fund) have been assigned with lower probability compared to less informative words (e.g. percent, million). As a result, these words will not appear in the top 10 words.

This paper compares several word ranking methods and evaluate them using two approaches. The first approach is based on a crowdsourcing task in which participants are provided with a document and a list of topics then asked to identify the correct one, i.e. the topic that is most closely associated with the document. Topics are represented by word lists ranked using different methods. The effectiveness of the re-ranking approaches is evaluated by computing the accuracy of the participants on identifying the correct topic. The second evaluation approach is based on an information retrieval (IR) task and does not rely on human judgements. The re-ranked words are used to form a query and retrieve a set of documents from the collection. The effectiveness of the word re-ranking is then evaluated in terms of how well it can retrieve documents in the collection related to the topic. Results show that re-ranking topic words improves performance in both experiments.

The paper makes the following contributions. It highlights the problem of re-ranking topic words and demonstrates that it can improve topic interpretability. It introduces the first systematic evaluation of topic word re-ranking methods using two approaches: one based on crowdsourcing and another based on an IR task. The latter evaluation is an automated approach and does not rely on human judgments. Experiments demonstrate strong agreement between the results produced by these approaches which indicates that the IR-based approach could be used as an automated evaluation method in future studies. The paper also compares multiple approaches to word re-ranking and concludes that the most effective ones are those which combine information about the importance of words within topics and their relative frequency across the entire corpus. Code used in the experiments described in this paper can be downloaded from `https://github.com/areejokaili/topic_reranking`.

2   Background

The standard approach to representing topics has been to show the top n words with the highest probability given the topic, e.g. (Blei and Lafferty, 2009a,b). However, these words may not be the ones that are most informative about the topic and a range of approaches to re-ranking them has been proposed in the literature.

Blei and Lafferty (2009a) proposed a re-ranking method inspired by the tf-idf word weighting which includes two types of information: firstly, the probability of a word given a topic of interest and, secondly, the same probability normalised by the average probability across all topics. The intuition behind this approach is that good words for representing a topic will be those which have both high probability for a given topic and low probability across all topics. Blei and Lafferty (2009a) did not describe any empirical evaluation of the effectiveness of their approach.

Other word re-ranking methods have also combined information about the overall probability of a word and its relative probability in one topic compared to others. Chuang et al. (2012) describe a word re-

ranking method applied within a topic model visualisation system. Their approach combines information about the word's overall probability within the corpus and its distinctiveness for a particular topic which is computed as the Kullback-Leibler divergence between the distribution of topics given the word and the distribution of topics. Sievert and Shirley (2014) also combine both types of information within a topic visualisation system. Bischof and Airoldi (2012) developed an approach for hierarchical topic models which balances information about the word frequency in a topic and the exclusivity of that word to that topic relative to a set of similar topics within the hierarchy.

Others have proposed approaches that only take into account the relative probability of each word in a topic compared to the others. Song et al. (2009) introduced a word ranking method based on normalising the probability of a word in a topic with the sum of the probabilities of that word across all topics. They evaluated their method against two other methods, the topic model's default ranking and the approach proposed by Blei and Lafferty (2009a), and found that it performed better than either. A similar method was proposed by Taddy (2012) who used the ratio of the probability of a word given a topic and the word's probability across the entire document collection.

Recently, Xing and Paul (2018) proposed to use information gathered while fitting the topic model. They made use of topic parameters from posterior samples generated during Gibbs sampling and re-weighted words based on their variability. Words with high uncertainty (i.e. their probabilities fluctuate relatively highly) are less likely to be representative of the topic than those with more stable probability estimates.

Topic re-ranking has also been explored within the context of measuring topic quality (Gollapalli and Li, 2018). A main claim of that work is that word importance should not only depend on its probability within a topic but also on its association with relevant neighbour words in the corpus. This information is incorporated by constructing topic-specific graphs capturing neighborhood words in a corpus. The PageRank (Brin and Page, 1998) algorithm is used to assign word importance scores based on centrality and then re-rank words based on their importance. The top n words with the highest PageRank values are used to compute the topics quality.

A common characteristic of previous work on topic word re-ranking is that it has been carried out within the context of an application of topic models (e.g. topic visualisation) and approaches have been evaluated in terms of these applications, if at all. The fact that word re-ranking methods have been considered in previous studies demonstrates their importance. The lack of direct and systematic evaluation is addressed in this work.

3 Word Re-ranking Methods

This paper explores a range of methods for word re-ranking based around the main approaches that have been applied to the problem (see Section 2). Let $\hat{\varphi}_{w,t}$ be the probability of a word w given a topic t produced by a topic model, e.g. LDA.[1] The following methods are used to re-rank topic words.

Original LDA Ranking (R_{Orig}) The most obvious and commonly used method for ranking words associated with a topic is to use $\hat{\varphi}_{w,t}$ to score each word, i.e. $score_{w,t} = \hat{\varphi}_{w,t}$. The ranking generated by this scoring function is equivalent to choosing the n most probable words for the topic and is referred to as R_{Orig}.

Normalised LDA Ranking (R_{Norm}) The first re-ranking method is a simple extension of R_{Orig} that represents approaches that normalise the probability of a word given a particular topic by the sum of probabilities for that word across all topics (Song et al., 2009; Taddy, 2012). This measure is computed as:

---

[1]We use the topic-word posterior distribution of LDA, but re-ranking can be applied to any topic model that estimates probabilities for words given topics.

$$score_{w,t} = \frac{\hat{\varphi}_{w,t}}{\sum_{j=1}^{T} \hat{\varphi}_{w,j}} \quad (1)$$

where T denotes the number of topics in the model. This approach scales the importance of words based on their overall occurrence within all topics in the model and downweights those that occur frequently.

Tf-idf Ranking ($\mathbf{R_{TFIDF}}$)   The second re-ranking method was proposed by Blei and Lafferty (2009a) and represents methods that combine information about the probability of a word in a single topic with information about its probability across all topics (Bischof and Airoldi, 2012; Chuang et al., 2012; Sievert and Shirley, 2014). Blei and Lafferty re-rank each word as:

$$score_{w,t} = \hat{\varphi}_{w,t} \, log \frac{\hat{\varphi}_{w,t}}{\left(\prod_{j=1}^{T} \hat{\varphi}_{w,j}\right)^{\frac{1}{T}}} \quad (2)$$

Inverse Document Frequency (IDF) Ranking ($\mathbf{R_{IDF}}$)   The final word re-ranking method explored in this paper is a variant on the previous method that takes account of a word's distribution across documents rather than topics. This method has not been explored in previous literature. In this approach each word is weighted by the Inverse Document Frequency (IDF) score across the corpus used to train the topic model:

$$score_{w,t} = \hat{\varphi}_{w,t} \, log \frac{|D|}{|D_w|} \quad (3)$$

where D is the entire document collection and D_w the documents within D containing the word w.

To better understand the effect of re-ranking words, consider the various representations of two topics shown in Table 2. The first row for each topic represents the baseline rank produced by topic model (R_{Orig}), while the other rows show the topic after re-ranking using Equations 1, 2 and 3, respectively. The bold words included in the original ranking (R_{Orig}) are down weighted and removed by at least two methods. Underlined words are those weighted higher by a re-ranking method and included in the topic representation.

Table 2: Examples of topic representations produced using various ranking approaches. Words in the R_{Orig} representation that were removed by at least two methods are shown in **bold**. Words that are ranked higher by the other approaches and included in the topic representation are shown underlined.

Method	Topic
R_{Orig}	space museum **years** history science earth mission **could** art shuttle
R_{Norm}	pincus abrams downey gettysburg particles sims emery landers lillian alamo
R_{TFIDF}	museum space earth shuttle pincus science universe radar exhibit art
R_{IDF}	museum space pincus science earth shuttle universe radar history mission
R_{Orig}	film **even** movie **world** stars **man much little story good**
R_{Norm}	vampire que winchell tomei westin swain marisa laughlin faye beatty
R_{TFIDF}	film movie stars vampire rating spielberg hollywood star characters actors
R_{IDF}	film movie stars vampire rating star spielberg even hollywood story

4 Experiment 1: Human Evaluation of Topic Interpretability

The first experiment compares the effectiveness of different topic representations (i.e. word re-rankings) by asking humans to choose the correct topic for a given document. We hypothesize that humans would be able to find the correct topic more easily when the representation is more interpretable.

4.1 Dataset and Preprocessing

We randomly sampled approximately 33,000 news articles from the New York Times included in the English GigaWord corpus fifth edition.[2] Documents were tokenized and stopwords removed. Words occurring in fewer than five or more than half of the documents were also removed to control for rare and common words. The size of the resulting vocabulary is approximately 52,000 words.

4.2 Topic Generation

Topics were generated using LDA's implementation in Gensim[3] fitted with online variational Bayes (Hoffman et al., 2010). The most important tuning parameter for LDA models is the number of topics and it was set to 50 after experimenting with varying number of topics optimised for coherence. To assess the quality of the resulting LDA models, topic coherence was computed[4] using: (1) C_V (Röder et al., 2015); (2) C_{UCI} (Newman et al., 2010); and (3) C_{NPMI} (Bouma, 2009).

4.3 Crowdsourcing Task

A job was created on the Figure Eight[5] crowdsourcing platform (previously known as CrowdFlower) in which participants were presented with ten micro-tasks per page[6]. Each micro-task consists of a text followed by six topics represented by a list of n words selected using one of the re-ranking methods presented in Section 3. Participants were asked to select the topic that was most closely associated with the text.

Micro-tasks were created using 48 New York Times articles (see Section 4.1). The correct answer is the topic with the highest probability given the article and incorrect answers (i.e. distractors) are five topics with low probabilities. The probability of the correct topic was at least 0.6 and the probability of the five distractors lower than 0.3.[7] Each article micro-tasks were created using each of the four ranking methods (Section 3) generated from topics created using three cardinalities (5, 10, 20). Five assessments were obtained for each micro-task and consequently 60 judgments were obtained for each document.[8]

Figure 1 shows an example of the micro-task presented to participants were they are asked to choose one of the topics. Participants are first provided with a brief description and an example to help them understand the task, followed by a quiz of ten micro-tasks to ensure their reliability and to eliminate random answers (Kazai, 2011). Participants who fail to answer seven out of ten micro-tasks correctly are eliminated from the job. If they qualify and proceed, a further quality micro-task is added per page and they need to maintain an accuracy of responses above 70%. To ensure non-redundant results, participants were always shown questions using the same topic word re-ranking method and the same number of words per topic. Also, participants can only answer a single page of 10 micro-tasks.

---

[2] https://catalog.ldc.upenn.edu/LDC2011T07
[3] https://radimrehurek.com/gensim
[4] The implementations available in Gensim were used.
[5] https://www.figure-eight.com/
[6] One of the ten micro-tasks on each page is reserved for quality assessment.
[7] Alternative values for these parameters were explored but it was found that lowering the probability of the correct answer and/or raising the probability of the distractors made the task too difficult.
[8] 4 (ranking methods) × 3 (cardinalities) × 5 (judgments per document)

"ground controllers grew increasingly worried Thursday about a malfunctioning navigational unit aboard the shuttle Columbia as the ship's busy crew packed for Friday's trip home. The trouble with one of the shuttle's three Inertial Measurement Units, or IMUs, began days ago, but erratic readings from the device have worsened, said Jeff Bantle, National Aeronautics and Space Administration entry flight director. Bantle informed the seven-member crew that additional degradation could alter the shuttle's landing plan if there is also poor weather at the Kennedy Space Center in Florida. The shuttle is scheduled to land at Kennedy on Friday at 6:47 a.m. EDT. ``We have brought up Edwards Air Force Base just in case,'' said Bantle, referring to NASA's alternative landing site in California. Three IMUs are not needed for landing, however, and flight controllers were confident Thursday that the spacecraft would land as scheduled. An IMU that had been powered down to preserve energy was brought into operation Thursday as a backup, Bantle said. IMUs are important because they send data to an on-flight computer about the shuttle's position relative to the Earth. Spikes "

water	city	space	players	race	medical
river	police	museum	baseball	cup	hospital
fish	york	years	owners	day	doctors
boat	mayor	history	league	racing	patients
sea	street	science	strike	old	treatment
beer	people	earth	season	run	doctor
ship	officials	mission	union	back	hospitals
fishing	officers	could	cap	horse	health
lake	giuliani	art	salary	mile	medicine
boats	fire	shuttle	labor	three	patient
A	B	C	D	E	F

Figure 1: Example of the crowdsourcing micro-task interface.

4.4 Results and Discussion

Results for R_{Orig}, R_{Norm}, R_{TFIDF} and R_{IDF} when topics are represented by the 5, 10 and 20 highest scoring words are shown in Table 3. *Accuracy* represents the percentage of questions for which participants were able to identify the correct topic (i.e. topic with the highest probability given the document). *Time/page* is the mean time taken for participants to complete a page of 10 questions. *Coherence* is the average coherence of the topics, computed using NPMI (Aletras and Stevenson, 2013a)[9].

Table 3: Results of experiment comparing re-ranking methods in which crowdsourcing participants were asked to associate topic representations with documents. Topics are represented with their top 5, 10 or 20 probable words.

#words		Ranking Methods			
		R_{Orig}	R_{Norm}	R_{TFIDF}	R_{IDF}
5	Accuracy (%)	64	55	70	**73**
	Time/page	11:46	13:13	13:00	11:28
	Coherence (NPMI)	0.092	0.035	0.112	0.100
10	Accuracy (%)	67	48	**76**	70
	Time/page	12:40	15:23	12:15	12:31
	Coherence (NPMI)	0.072	0.038	0.091	0.084
20	Accuracy (%)	69	64	**74**	72
	Time/page	14:18	14:30	11:47	11:13
	Coherence (NPMI)	0.050	0.029	0.071	0.062

Results show a variation in performance which indicates that re-ranking words affects individual's ability to interpret topics. Performance when the words are ranked using R_{TFIDF} and R_{IDF} outperform the default ranking (R_{Orig}). Performance when words are ranked using R_{Norm} is considerably lower than their word re-ranking methods, both in terms of accuracy and time taken to complete the task.

These results show that the improvement obtained by using R_{TFIDF} and R_{IDF} is consistent when the number of words in the representation is varied. Results using R_{Orig} improve as the number of words

[9]The implementation provided by https://github.com/jhlau/topic_interpretability was used.

increases but never achieve the same performance as the re-ranking methods (except R_{Norm}), even when 20 words are included. This demonstrates that choosing the most appropriate words to represent a topic is more useful than simply increasing the number of words shown to the user. In fact, increasing the number of words shown for the default ranking appears to come at the cost of slowing down the time taken for a user to interpret the topic. The same increase in task completion time is not observed for R_{TFIDF} and R_{IDF} and this may be down to the fact that more useful words appear earlier in the ranking, allowing participants to interpret the topic more quickly.

The R_{TFIDF} and R_{IDF} approaches both combine information about the word's importance within an individual topic and across the entire document collection which results into more effective rankings than R_{Orig}. On the other hand, R_{Norm} only considers the relative importance of a word across topics and it would be possible for a word with a relatively low probability given the topic to be ranked highly if that word also had low probability across all the other topics.

Our results contrast with those reported by Song et al. (2009) who concluded that R_{Norm} was more effective for word re-ranking than R_{Orig} and R_{TFIDF} (see Section 3). However, their evaluation methodology used a single annotator per-task and asked them to judge whether words included within topic representations were important or not. Our approach measures a participants ability to interpret topic representations more directly and makes use of multiple annotations. The low results for R_{Norm} suggest that crowdworkers were simply unable to interpret many of the topics and, in those cases, their judgments about which words are important are unlikely to be reliable.

Overall R_{TFIDF} appears to be the most effective of the re-ranking approaches evaluated. This method achieves the best performance for 10 and 20 words, although not as well as R_{IDF} for 5 words.

5 Experiment 2: Automatic Evaluation of Topic Interpretability via Document Retrieval

In this second experiment, we automate the evaluation of the different topic representations obtained by re-ranking the topic words. The automated evaluation is based on an IR task in which the re-ranked topic words are used to form a query and retrieve documents relevant to the topic. The motivation behind this approach is that the most effective re-rankings are the ones that can retrieve documents related to the topic, while ineffective re-rankings will not be able to distinguish these from other documents in the collection. This evaluation method does not rely on human judgments, unlike the crowdsourcing approach presented in the previous section.

5.1 Evaluation Pipeline

The evaluation approach assumes that given a document collection in which each document is mapped to a label (or labels) indicating its topic. We refer to these labels as *gold standard topics* (to distinguish them from the automatically generated topics created by the topic model).

First, a set of automatically generated topics are created by running a topic model over a document collection. For each gold standard topic, a set of all documents labelled with that topic is created. The document-topic distribution created by the topic model is then used to identify the most probable automatically generated topic within that set of documents. This is achieved by summing the document-topic distributions and choosing the automatically generated topic with the highest value. A query is then created by selecting the re-ranked top n words from that automatically generated topic and use it to retrieve a set of documents from the collection. The set of retrieved documents is then compared against the set of all documents labelled with the gold standard label.

5.2 Datasets

Evaluation was carried out using datasets representing documents from a wide range of domains: news articles, scientific literature and online reviews.

Table 4: Datasets statistics.

Dataset	Documents	Distinct Words
NYT Annotated	39,218	60,339
MEDLINE	23,640	18,571
Amazon	40,000	24,943

5.2.1 New York Times

A subset of the NYT annotated dataset[10] consisting of approximately approximately 39,000 articles was used for this experiment[11]. This collection contains news articles from the *New York Times* labelled with 1,746 topics which we use as gold standard labels. These labels, which we refer to as NYT_topics, belong to a controlled set of topic categories and have been manually verified by NYTimes.com production staff. Each article has at least one NYT_topic, and articles are organised into a topic hierarchy. Examples of NYT_topics include:

- *Top/Features/Travel/Guides/Destinations/North America/United States*

- *Top/News/New York and Region*

- *Top/News/Technology*

The hierarchy into which the topics are organised is quite deep in some places and consequently we truncated each topic to the top most four levels of the hierarchy to control the number of topics. For example, the topic *Top/Features/Travel/Guides/ Destinations/North America/United States* is truncated to *Top/Features/Travel/Guides*. This produces a total of 132 truncated NYT_topics. The number of articles associated with each of the 132 NYT_topics ranges from 1 to 18,489. To avoid NYT_topics that are associated with small numbers of documents, we used the 50 NYT_topics that are associated with the most documents which resulted in NYT_topic that are each associated with at least 560 documents.

5.2.2 MEDLINE

MEDLINE contains abstracts of more than 25 million scientific publications in medicine and related fields. These abstracts are labelled with *Medical Subject Headings* (MeSH) codes which index publications into a hierarchy structure. Each publication is associated with a set of MeSH codes to describe the content of the publication.

The 50 most frequently used MeSH codes with the most publications from a subset of MEDLINE containing publications from 2017. This set of code are referred to as *MeSH_topics*.

5.2.3 Amazon Product Reviews

The Amazon Product Reviews dataset (McAuley et al., 2015)[12] contains reviews of products purchased from the Amazon website. Reviews are organized into 24 top-level categories each of which is divided into subcategories. The number of subcategories ranges from 1 to 1961. We chose eight main categories (*Cell Phones and Accessories, Electronics, Movies and TV, Musical Instrument, Office Products, Pet Supplies, Tools and Home Improvement and Automotive*) and extracted the 10 sub-categories with the most reviews from each which yielded 76 distinct sub-categories. This set of categories is referred to as *AMZ_topics*. 5,000 product reviews are extracted from each main category and reviews must belong to at least one category from the 50 most frequent in the AMZ_topics, resulting in a total of 40,000 reviews.

---

[10]https://catalog.ldc.upenn.edu/LDC2008T19

[11]Note that this is a different dataset to the one used for experiment 1 and contains the gold standard topics required for this evaluation

[12]http://jmcauley.ucsd.edu/data/amazon

Table 5: Results of experiment in which top 5, 10 and 20 ranked words are used to form query.

#Words	Ranking Method			
	R_{Orig}	R_{Norm}	R_{TFIDF}	R_{IDF}
New York Times Dataset				
5	0.0945	0.0463	**0.1363**	0.1187
10	0.1161	0.0608	**0.1417**	0.1291
20	0.1256	0.0721	**0.1392**	0.1321
Medline Dataset				
5	0.1420	0.0202	**0.1738**	0.1578
10	0.1518	0.0289	**0.1612**	0.1575
20	0.1498	0.0372	**0.1662**	0.1642
Amazon Product Reviews Dataset				
5	0.0231	0.0202	0.0208	**0.0244**
10	0.0195	0.0154	**0.0244**	0.0236
20	0.0258	0.0137	**0.0279**	0.0266

5.3 Experimental Settings

Each of the datasets was indexed using Apache Lucene.[13] The same preprocessing steps used in Experiment 1 were applied to the datasets and the statistics of the datasets are shown in Table 4.

For each dataset, LDA was used to generate topics and the number of topics for each dataset was set based on optimising for coherence which yielded 35 for NYT, 45 for MEDLINE and 35 for Amazon. The automatically generated topic that was most closely associated with each of the gold topics (i.e. 50 NYT_topics, 50 MeSH_topics and 50 AMZ_topics) were identified by applying the process outlined above in Section 5.1. The top 5, 10 and 20 words from this topic is used to form a query which is submitted to Lucene. The BM25 retrieval model (Robertson, 2004) was used to measure the similarity between the document to a given query. The documents retrieved by applying these queries are compared against the entire set of documents labelled with the dataset topics (i.e. NYT_topic, MeSH_topics, or AMZ_topics) by computing Mean Average Precision (MAP)[14] which is commonly used as a single metric to summarise IR system performance.

5.4 Results and Discussion

Queries were created using the top 5, 10, and 20 topic words using the R_{Orig}, R_{Norm}, R_{TFIDF}, and R_{IDF} re-rankings and applied to each of the three datasets (Section 5.2). Results are shown in Table 5.

Re-ranking words using R_{TFIDF} and R_{IDF} consistently enhances retrieval performance compared to the default ranking (R_{Orig}). R_{TFIDF} produces the best results in the majority of configurations, the exception being when 5 words are used with the Amazon corpus where R_{IDF} outperforms the other re-ranking methods. Re-ranking using R_{Norm} is less effective than all the other rankings, including the default ranking. The relative performance of the four approaches is generally stable when the number of words used to form the query is varied and across the three datasets representing very different genres of text that were used in this experiment.

These results demonstrate that topic word re-ranking can produce words which are more effective for discriminating documents describing a particular topic from those which do not. The pattern of results is very similar to the crowdsourcing experiments suggesting that the re-rankings preferred by human subjects are those which are also useful within applications such as document retrieval.

---

[13]http://lucene.apache.org/

[14]MAP is computed using the *trec_eval* tool: http://trec.nist.gov/trec_eval/

6 Evaluating Topic Representations

This paper presented two novel methods for the evaluation of topic representations: a crowdsourcing experiment that relied on human judgments (Section 4) and an automated evaluation based on an IR task (Section 5). Although there are some differences between results using the two methods, the relative performance of the re-ranking methods explored in this paper are very similar. The correlations between results of the crowdsourcing experiment and IR evaluations are statistically significant for all three datasets (Pearson's r varies between 0.81 and 0.90, $p < 0.05$). This suggests that the automated evaluation approach presented in Section 5 is a useful tool for assessing the effectiveness of methods for word re-ranking with the advantage that results can be obtained more rapidly than methods that require human judgments. However, human judgments are recommended when performance is similar and automated evaluation should not be relied upon to make fine-grained distinctions between approaches, as is common for some tasks (e.g. Machine Translation (Papineni et al., 2002)).

7 Conclusion

We presented a study on word re-ranking methods designed to improve topic interpretability. Four methods were presented and assessed through two experiments. In the first experiment, participants on a crowdsourcing platform were asked to associate documents with related topics. In the second experiment, automated evaluation was based on a document retrieval task.

Re-ranking the topic words was found to improve the interpretability of topics and therefore should be used as a post-processing step to improve topic representation. The most effective re-ranking schemes were those which combined information about the importance of words both within topics and their relative frequency in the entire corpus, thereby ensuring that less informative words are not used.

References

Aletras, N., T. Baldwin, J. Lau, and M. Stevenson (2014). Representing Topics Labels for Exploring Digital Libraries. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '14)*, pp. 239–248.

Aletras, N., T. Baldwin, J. H. Lau, and M. Stevenson (2017). Evaluating Topic Representations for Exploring Document Collections. *Journal of the Association for Information Science and Technology 68*(1), 154–167.

Aletras, N. and A. Mittal (2017). Labeling Topics with Images using Neural Networks. In *European Conference on Information Retrieval (ECIR '17)*, pp. 500–505.

Aletras, N. and M. Stevenson (2013a). Evaluating Topic Coherence using Distributional Semantics. In *Proceedings of the 10th international conference on Computational Semantics (IWCS '13)*, Volume 13, , Potsdam, Germany, pp. 13–22.

Aletras, N. and M. Stevenson (2013b). Representing Topics Using Images. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT '13)*, pp. 158–167.

Aletras, N. and M. Stevenson (2014). Labelling Topics using Unsupervised Graph-based Methods. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL '14)*, pp. 631–636.

Bischof, J. and E. Airoldi (2012). Summarizing Topical Content with Word Frequency and Exclusivity. In *Proceedings of the 29 th International Conference on Machine Learning (ICML '12)*, pp. 201–208.

Blei, D. (2012). Probabilistic topic models. *Communications of the ACM 4*(55), 77–84.

Blei, D. and J. Lafferty (2009a). Topic Models. *Text mining: classification, clustering, and applications 10*(71), 34.

Blei, D. and J. Lafferty (2009b). Visualizing Topics with Multi-Word Expressions. *arXiv preprint arXiv:0907.1013*.

Blei, D., A. Ng, and M. Jordan (2003). Latent Dirichlet Allocation. *Journal of machine Learning research 3*, 993–1022.

Bouma, G. (2009). Normalized (Pointwise) Mutual Information in Collocation Extraction. In *In Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology (GSCL 09)*, pp. 31–40.

Boyd-Graber, J., Y. Hu Google, and D. Mimno (2017). Applications of Topic Models. *Foundations and Trends R in Information Retrieval 11*(2-3), 143–296.

Brin, S. and L. Page (1998, 4). The Anatomy of a Large-scale Hypertextual Web search engine. *Computer Networks and ISDN Systems 30*(1-7), 107–117.

Chaney, A. and D. Blei (2012). Visualizing Topic Models. In *Proceedings of the 6th International AAAI Conference Weblogs and Social Media (AAAI ICWSM '12)*, pp. 419–422.

Chang, J., J. Boyd-Graber, S. Gerrish, C. Wang, and D. Blei (2009). Reading Tea Leaves: How Humans Interpret Topic Models. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems (NIPS '09)*, pp. 288–296.

Chuang, J., C. Manning, and J. Heer (2012). Termite: Visualization Techniques for Assessing Textual Topic Models. In *Proceeding of the International Working Conference on Advanced Visual Interfaces (AVI '12)*, pp. 74–77.

Gollapalli, S. and X.-l. Li (2018). Using PageRank for Characterizing Topic Quality in LDA. In *Proceedings of the ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '18)*, pp. 115–122.

Hoffman, M., F. Bach, and D. Blei (2010). Online Learning for Latent Dirichlet Allocation. In *Proceedings of the 23rd Advances in Neural Information Processing Systems (NIPS '10)*, pp. 856–864.

Kazai, G. (2011). In Search of Quality in Crowdsourcing for Search Engine Evaluation. In *Proceedings of the 33rd European Conference on Advances in Information Retrieval (ECIR '11)*, pp. 165–176.

Lau, J. and T. Baldwin (2016). The Sensitivity of Topic Coherence Evaluation to Topic Cardinality. In *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT '16)*, pp. 483–487.

Lau, J., K. Grieser, D. Newman, and T. Baldwin (2011). Automatic Labelling of Topic Models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT '11)*, pp. 1536–1545.

Lau, J., D. Newman, and T. Baldwin (2014). Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL '14)*, pp. 530–539.

McAuley, J., C. Targett, Q. Shi, and A. van den Hengel (2015). Image-Based Recommendations on Styles and Substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '15)*, pp. 43–52.

Mimno, D., H. Wallach, E. Talley, M. Leenders, and A. Mccallum (2011). Optimizing Semantic Coherence in Topic Models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11)*, pp. 262–272.

Newman, D., H. Lau, K. Grieser, and T. Baldwin (2010). Automatic Evaluation of Topic Coherence. In *Proceedings of Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT10)*, pp. 100–108.

Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL '02)*, pp. 311–318.

Paul, M. (2016). Interpretable Machine Learning: Lessons from Topic Modeling. In *CHI Workshop on Human-Centered Machine Learning*.

Robertson, S. (2004, 10). Understanding Inverse Document Frequency: on Theoretical Arguments for IDF. *Journal of Documentation 60*(5), 503–520.

Röder, M., A. Both, and A. Hinneburg (2015). Exploring the Space of Topic Coherence Measures. In *Proceedings of the eight ACM International Conference on Web Search and Data Mining (WSDM '15)*, pp. 399–408. ACM.

Schofield, A., M. Magnusson, and D. Mimno (2017). Pulling Out the Stops: Rethinking Stopword Removal for Topic Models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL '17)*, Volume 2, pp. 432–436.

Sievert, C. and K. Shirley (2014). LDAvis: A Method for Visualizing and Interpreting Topics. In *Preceedings of the Workshop on Interactive Learning Visualization, and Interfaces*, pp. 63–70.

Smith, A., T. Lee, F. Poursabzi-Sangdeh, J. Boyd-Graber, N. Elmqvist, and L. Findlater (2017). Evaluating Visual Representations for Topic Understanding and Their Effects on Manually Generated Topic Labels. *Transactions of the Association for Computational Linguistics 5*, 1–15.

Song, Y., S. Pan, S. Liu, M. Zhou, and W. Qian (2009). Topic and Keyword Re-ranking for LDA-based Topic Modeling. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM '09)*, pp. 1757–1760.

Sorodoc, I., J. H. Lau, N. Aletras, and T. Baldwin (2017). Multimodal Topic Labelling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL '17)*, pp. 701–706.

Taddy, M. (2012). On Estimation and Selection for Topic Models. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS '12)*, pp. 1184–1193.

Xing, L. and M. J. Paul (2018). Diagnosing and Improving Topic Models by Analyzing Posterior Variability. In *Proceedings of the Advancement of Artificial Intelligence (AAAI '18)*, pp. 6005–6012.