Automatic Labeling of Topic Models Using Text Summaries

Xiaojun Wan and Tianming Wang
Institute of Computer Science and Technology, The MOE Key Laboratory of Computational
Linguistics, Peking University, Beijing 100871, China
{wanxiaojun, wangtm}@pku.edu.cn

Abstract

Labeling topics learned by topic models is a challenging problem. Previous studies have used words, phrases and images to label topics. In this paper, we propose to use text summaries for topic labeling. Several sentences are extracted from the most related documents to form the summary for each topic. In order to obtain summaries with both high relevance, coverage and discrimination for all the topics, we propose an algorithm based on submodular optimization. Both automatic and manual analysis have been conducted on two real document collections, and we find 1) the summaries extracted by our proposed algorithm are superior over the summaries extracted by existing popular summarization methods; 2) the use of summaries as labels has obvious advantages over the use of words and phrases.

1 Introduction

Statistical topic modelling plays very important roles in many research areas, such as text mining, natural language processing and information retrieval. Popular topic modeling techniques include Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and Probabilistic Latent Semantic Analysis (pLSA) (Hofmann, 1999). These techniques can automatically discover the abstract "topics" that occur in a collection of documents. They model the documents as a mixture of topics, and each topic is modeled as a probability distribution over words.

Although the discovered topics' word distributions are sometimes intuitively meaningful, a major challenge shared by all such topic models is to accurately interpret the meaning of each topic (Mei et al., 2007). The interpretation of each topic is very important when people want to browse, understand and leverage the topic. However, it is usually very hard for a user to understand the discovered topics based only on the multinomial distribution of words. For example, here are the top terms for a discovered topic: {*fire miles area north southern people coast homes south damage northern river state friday central water rain high california weather*}. It is not easy for a user to fully understand this topic if the user is not very familiar with the document collection. The situation may become worse when the user faces with a number of discovered topics and the sets of top terms of the topics are often overlapping with each other on many practical document collections.

In order to address the above challenge, a few previous studies have proposed to use phrases, concepts and even images for labeling the discovered topics (Mei et al., 2007; Lau et al., 2011; Hulpus et al., 2013; Aletras and Stevenson, 2013). For example, we may automatically extract the phrase "*southern california*" to represent the example topic mentioned earlier. These topic labels can help the user to understand the topics to some extent. However, the use of phrases or concepts as topic labels are not very satisfactory in practice, because the phrases or concepts are still very short, and the information expressed in these short labels is not adequate for user's understanding. The case will become worse when some ambiguous phrase is used or multiple discrete phrases with poor coherence are used for a topic. To address the drawbacks of the above short labels, we need to provide more contextual information and consider using long text descriptions to represent the topics. The long text descriptions can be used independently or used as beneficial complement to the short labels. For example, below is part of the summary label produced by our proposed method and it provides much more contextual information for understanding the topic.

> *Showers and thunderstorms developed in parched areas of the southeast , from western north carolina into south central alabama , north central and northeast texas and the central and southern gulf coast . ... The quake was felt over a*

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 2297–2305,
Berlin, Germany, August 7-12, 2016. ©2016 Association for Computational Linguistics

> *large area , extending from santa rosa , about 60 miles north of san francisco , to the santa cruz area 70 miles to the south .... Fourteen homes were destroyed in baldwin park 20 miles northeast of downtown los angeles and five were damaged along with five commercial buildings when 75 mph gusts snapped power lines , igniting a fire at allan paper co. , fire officials said . ...*

The contributions of this paper are summarized as follows:

1) We are the first to invesitage using text summaries for topic labeling;

2) We propose a summarization algorithm based on submodular optimization to extract summaries with both high relevance, coverage and discrimination for all topics.

3) Automatic and manual analysis reveals the usefulness and advantages of the summaries produced by our algorithm.

2 Related Work

2.1 Topic Labeling

After topics are discovered by topic modeling techniques, these topics are conventionally represented by their top N words or terms (Blei et al., 2003; Griffiths and Steyvers, 2004). The words or terms in a topic are ranked based on the conditional probability $\mathrm{p}(w_i|t_j)$ in that topic. It is sometimes not easy for users to understand each topic based on the terms. Sometimes topics are presented with manual labeling for exploring research publications (Wang and McCallum, 2006; Mei et al., 2006), and the labeling process is time consuming.

In order to make the topic representations more interpretable and make the topics easier to understand, there are a few studies proposing to automatically find phrases, concepts or even images for topic labeling. Mei et al. (2007) proposed to use phrases (chunks or ngrams) for topic labeling and cast the labeling problem as an optimization problem involving minimizing Kullback-Leibler (KL) divergence between word distributions and maximizing mutual information between a label and a topic model. Lau et al. (2011) also used phrases as topic labels and they proposed to use supervised learning techniques for ranking candidate labels. In their work, candidate labels include the top-5 topic terms and a few noun chunks extracted from related Wikipedia articles. Mao et al. (2012) proposed two effective algorithms that automatically assign concise labels to each topic in a hierarchy by exploiting sibling and parent-child relations among topics. Kou et al. (2015) proposed to map topics and candidate labels (phrases) to word vectors and letter trigram vectors in order to find which candidate label is more semantically related to that topic. Hulpus et al. (2013) took a new approach based on graph centrality measures to topic labelling by making use of structured data exposed by DBpedia. Different from the above works, Aletras and Stevenson (2013) proposed to use images for representing topics, where candidate images for each topic are retrieved from the web and the most suitable image is selected by using a graph-based algorithm. In a very recent study (Aletras et al., 2015), 3 different topic representations (lists of terms, textual phrase labels and images labels) are compared in a document retrieval task, and results show that textual phrase labels are easier for users to interpret than term lists and image labels.

The phrase-based labels in the above works are still very short and are sometimes not adequate for interpreting the topics. Unfortunately, none of previous works has investigated using textual summaries for representing topics yet.

2.2 Document Summarization

The task of document summarization aims to produce a summary with a length limit for a given document or document set. The task has been extensively investigated in the natural language processing and information retrieval fields, and most previous works focus on directly extracting sentences from a news document or collection to form the summary. The summary can be used for helping users quickly browse and understand a document or document collection.

Typical multi-document summarization methods include the centroid-based method (Radev et al., 2004), integer linear programming (ILP) (Gillick et al., 2008), sentence-based LDA (Chang and Chien, 2009), submodular function maximization (Lin and Bilmes, 2010; Lin and Bilmes, 2011), graph based methods (Erkan and Radev, 2004; Wan et al., 2007; Wan and Yang, 2008), and supervised learning based methods (Ouyang et al., 2007; Shen et al., 2007). Though different summarization methods have been proposed in recent years, the submodular function maximization method is still one of the state-of-the-art summarization methods. Moreover, the method is easy to follow and its framework is very flexible. One can design specific submodular functions for addressing special summarization tasks, without altering the overall greedy selection framework.

Though various summarization methods have been proposed, none of existing works has investigated or tried to adapt document summarization techniques for the task of automatic labeling of topic models.

3 Problem Formulation

Given a set of latent topics extracted from a text collection and each topic is represented by a multinomial distribution over words, our goal is to produce understandable text summaries as labels for interpreting all the topics. We now give two useful definitions for later use.

Topic: Each topic θ is a probability distribution of words $\{p_\theta(w)\}_{w\in V}$, where V is the vocabulary set, and we have $\sum_{w\in V} p_\theta(w) = 1$.

Topic Summary: In this study, a summary for each topic θ is a set of sentences extracted from the document collection and it can be used as a label to represent the latent meaning of θ. Typically, the length of the summary is limited to 250 words, as defined in recent DUC and TAC conferences.

Like the criteria for the topic labels in (Mei et al., 2007), the topic summary for each topic needs to meet the following two criteria:

High Relevance: The summary needs to be semantically relevant to the topic, i.e., the summary needs to be closely relevant to all representative documents of the topic. The higher the relevance is, the better the summary is. This criterion is intuitive because we do not expect to obtain a summary unrelated to the topic.

High Coverage: The summary needs to cover as much semantic information of the topic as possible. The summary usually consists of several sentences, and we do not expect all the sentences to focus on the same piece of semantic information. A summary with high coverage will certainly not contain redundant information. This criterion is very similar to the diversity requirement of multi-document summarization.

Since we usually produce a set of summaries for all the topics discovered in a document collection. In order to facilitate users to understand all the topics, the summaries need to meet the following additional criterion:

High Discrimination: The summaries for different topics need to have inter-topic discrimination. If the summaries for two or more topics are very similar with each other, users can hardly understand each topic appropriately. The higher the inter-topic discrimination is, the better the summaries are.

4 Our Method

Our proposed method is based on submodular optimization, and it can extract summaries with both high relevance, coverage and discrimination for all topics. We choose the framework of submodular optimization because the framework is very flexible and different objectives can be easily incorporated into the framework. The overall framework of our method consists of two phases: candidate sentence selection, and topic summary extraction. The two phrases are described in the next two subsections, respectively.

4.1 Candidate Sentence Selection

There are usually many thousands of sentences in a document collection for topic modelling, and all the sentences are more or less correlated with each topic. If we use all the sentences for summary extraction, the summarization efficiency will be very low. Moreover, many sentences are not suitable for summarization because of their low relevance with the topic. Therefore, we filter out the large number of unrelated sentences and treat the remaining sentences as candidates for summary extraction.

For each topic θ, we compute the Kullback-Leibler (KL) divergence between the word distributions of the topic and each sentence s in the whole document collection as follows:

$$KL(\theta, s) = \sum_{w\in TW\cup SW} p_\theta(w) * log\frac{p_\theta(w)}{tf(w,s)/len(s)}$$

where $p_\theta(w)$ is the probability of word w in topic θ. TW denotes the set of top 500 words in topic θ according to the probability distribution. SW denotes the set of words in sentence s after removing stop words. $tf(w,s)$ denotes the frequency of word w in sentence s, and $len(s)$ denotes the length of sentence s after removing stop words. For a word w which does not appear in SW, we set $tf(w,s)/len(s)$ to a very small value (0.00001 in this study).

Then we rank the sentences by an increasing order of the divergence scores and keep the top 500 sentences which are most related to the topic. These 500 sentences are treated as candidate sentences for the subsequent summarization step for each topic. Note that different topics have different candidate sentence sets.

4.2 Topic Summary Extraction

Our method for topic summary extraction is based on submodular optimization. For each topic θ associated with the candidate sentence set V, our method aims to find an optimal summary $\tilde{E}$ from all possible summaries by maximizing a score function under budget constraint:

$$\tilde{E} = argmax_{E \subseteq V}\{f(E)\} \quad \text{s.t. } len(E) \le L$$

where $len(E)$ denotes the length of summary E. Here E is also used to denote the set of sentences in the summary. L is a predefined length limit, i.e. 250 words in this study.

$f(E)$ is the score function to evaluate the overall quality of summary E. Usually, $f(E)$ is required to be a submodular function, so that we can use a simple greedy algorithm to find the near-optimal summary with theoretical guarantee. Formally, for any $A \subseteq B \subseteq V \backslash v$, we have

$$f(A+v) - f(A) \ge f(B+v) - f(B)$$

which means that the incremental "value" of *v* decreases as the context in which v is considered grows from *A* to *B*.

In this study, the score function $f(E)$ is decomposed into three parts and each part evaluates one aspect of the summary:

$$f(E) = REL(E) + COV(E) + DIS(E)$$

where $REL(E)$, $COV(E)$ and $DIS(E)$ evaluate the relevance, coverage and discrimination of summary *E* respectively. We will describe them in details respectively.

4.2.1 Relevance Function

Instead of intuitively measuring relevance between the summary and the topic via the KL divergence between the word distributions of them, we consider to measure the relevance of summary *E* for topic θ by the relevance of the sentences in the summary to all the candidate sentences for the topic as follows:

$$REL(E) = \sum_{s' \in V} \min\{\sum_{s \in E} sim(s',s), \alpha \sum_{s \in V} sim(s',s)\}$$

where *V* represents the candidate sentence set for topic θ, and E is used to represent the sentence set of the summary. $sim(s',s)$ is the standard cosine similarity between sentences s' and s. $\alpha \in [0,1]$ is a threshold co-efficient.

The above function is a monotone submodular function because $f(x) = min(x,a)$ where $a \ge 0$ is a concave non-decreasing function. $\sum_{s \in E} sim(s',s)$ measures how similar *E* is to sentence s' and then $\sum_{s \in V} sim(s',s)$ is the largest value that $\sum_{s \in E} sim(s',s)$ can achieve. Therefore, s' is saturated by E when $\sum_{s \in E} sim(s',s) \ge \alpha \sum_{s \in V} sim(s',s)$. When $s'$is already saturated by *E* in this way, any new sentence very similar to s' cannot further improve the overall relevance of *E*, and this sentence is less possible to be added to the summary.

4.2.2 Coverage Function

We want the summary to cover as many topic words as possible and contain as many different sentences as possible. The coverage function is thus defined as follows:

$$COV(E) = \beta * \sum_{w \in TW} \left\{ p_\theta(w) * \sqrt{\sum_{s \in E} tf(w,s)} \right\}$$

where $\beta \ge 0$ is a combination co-efficient.

The above function is a monotone submodular function and it encourages the summary *E* to contain many different words, rather than a small set of words. Because $f(x) = \sqrt{x}$ where $x \ge 0$ is a concave non-decreasing function, we have $f(x + y) \le f(x) + f(y)$. The value of the function will be larger when we use x and y to represent two frequency values of two different words respectively than that when we use $(x+y)$ to represent the frequency value of a single word. Therefore, the use of this function encourages the coverage of more different words in the summary. In other words, the diversity of the summary is enhanced.

4.2.3 Discrimination Function

The function for measuring the discrimination between the summary *E* of topic θ and all other topics $\{\theta'\}$ is defined as follows:

$$DIS(E) = -\gamma \sum_{\theta'} \sum_{s \in E} \sum_{w \in TW} p_{\theta'}(w) * tf(w,s)$$

where $\gamma \ge 0$ is a combination co-efficient.

The above function is still a monotone submodular function. The negative sign indicates that the summary E of topic θ needs to be as irrelevant with any other topic as possible, and thus making different topic summaries have much differences.

4.2.4 Greedy Selection

Since $REL(E)$, $COV(E)$ and $DIS(E)$ are all submodular functions, $f(E)$ is also a submodular function. In order to find a good approximation to the optimal summary, we use a greedy algorithm similar to (Lin and Bilmes, 2010) to select sentence one by one and produce the final summary, as shown in Algorithm 1.

Algorithm 1 Greedy algorithm for summary extraction

```
1: $E \leftarrow \emptyset$
2: $U \leftarrow V$
3: **while** $U \neq \emptyset$ **do**
4:     $\hat{s} \leftarrow argmax_{s \in U} \frac{f(E \cup \{s\}) - f(E)}{len(s)^{\varepsilon}}$
5:     $E \leftarrow E \cup \{\hat{s}\}$ if $\sum_{s \in E} len(s) + len(\hat{s}) \leq L$ and
       $f(E \cup \{s\}) - f(E) \geq 0$
6:     $U \leftarrow U \setminus \{\hat{s}\}$
7: **end while**
8: return $E$
```

In the algorithm, $len(s)$ denotes the length of sentence s and $\varepsilon > 0$ is the scaling factor. At each iteration, the sentence with the largest ratio of objective function gain to scaled cost is found in step 4, and if adding the sentence can increase the objective function value while not violating the length constraint, it is then selected into the summary and otherwise bypassed.

5 Evaluation and Results

5.1 Evaluation Setup

We used two document collections as evaluation datasets, as in (Mei et al. 2007): AP news and SIGMOD proceedings. The AP news dataset contains a set of 2250 AP news articles, which are provided by TREC. There is a total of 43803 sentences in the AP news dataset and the vocabulary size is 37547 (after removing stop words). The SIGMOD proceeding dataset contains a set of 2128 abstracts of SIGMOD proceedings between the year 1976 and 2015, downloaded from the ACM digital library. There is a total of 15211sentences in the SIGMOD proceeding dataset and the vocabulary size is 13688.

For topic modeling, we adopted the most popular LDA to discover topics in the two datasets, respectively. Particularly, we used the LDA module implemented in the MALLET toolkit[1]. Without loss of generality, we extracted 25 topics from the AP news dataset and 25 topics from the SIGMOD proceeding dataset.

The parameter values of our proposed summarization method is either directly borrowed from previous works or empirically set as follows: $\alpha = 0.05$, $\beta = 250$, $\gamma = 300$ and $\varepsilon = 0.15$.

We have two goals in the evaluation: comparison of different summarization methods for topic labeling, and comparison of different kinds of labels (summaries, words, and phrases).

In particular, we compare our proposed summarization method (denoted as **Our Method**) with the following typical summarization methods and all of them extract summaries from the same candidate sentence set for each topic:

MEAD: It uses a heuristic way to obtain each sentence's score by summing the scores based on different features (Radev et al., 2004): centroid-based weight, position and similarity with first sentence.

LexRank: It constructs a graph based on the sentences and their similarity relationships and then applies the PageRank algorithm for sentence ranking (Erkan and Radev, 2004).

TopicLexRank: It is an improved version of LexRank by considering the probability distribution of top 500 words in a topic as a prior vector, and then applies the topic-sensitive PageRank algorithm for sentence ranking, similar to (Wan 2008).

Submodular(REL): It is based on submodular function maximization but only the relevance function is considered.

Submodular(REL+COV): It is based on submodular function maximization and combines two functions: the relevance function and the coverage function.

We also compare the following three different kinds of labels:

Word label: It shows ten topic words as labels for each topic, which is the most intuitive interpretation of the topic.

Phrase label: It uses three phrases as labels for each topic, and the phrase labels are extracted by using the method proposed in (Mei et al., 2007), which is very closely related to our work and considered a strong baseline in this study.

Summary Label: It uses a topic summary with a length of 250 words to label each topic and the summary is produced by our proposed method.

5.2 Evaluation Results

5.2.1 Automatic Comparison of Summarization Methods

In this section, we compare different summarization methods with the following automatic measures:

[1] http://mallet.cs.umass.edu/

KL divergence between word distributions of summary and topic: For each summarization method, we compute the KL divergence between the word distributions of each topic and the summary for the topic, then average the KL divergence across all topics. Table 1 shows the results. We can see that our method and Submodular(REL+COV) have the lowest KL divergence with the topic, which means our method can produce summaries relevant to the topic representation.

Topic word coverage: For each summarization method, we compute the ratio of the words covered by the summary out of top 20 words for each topic, and then average the ratio across all topics. We use top 20 words instead of 500 words because we want to focus on the most important words. The results are shown in Table 2. We can see that our method has almost the best coverage ratio and the produced summary can cover most important words in a topic.

	AP	**SIGMOD**
MEAD	0.832503	1.470307
LexRank	0.420137	1.153163
TopicLexRank	0.377587	1.112623
Submodular(REL)	0.43264	1.002964
Submodular(REL+COV)	0.349807	0.991071
Our Method	0.360306	0.907193

Table 1. Comparison of KL divergence between word distributions of summary and topic

	AP	**SIGMOD**
MEAD	0.422246	0.611355
LexRank	0.651217	0.681728
TopicLexRank	0.678515	0.692066
Submodular(REL)	0.62815	0.713159
Submodular(REL+COV)	0.683998	0.723228
Our Method	0.673585	0.74572

Table 2. Comparison of the ratio of the covered words out of top 20 topic words

	AP		**SIGMOD**	
	average	max	average	max
MEAD	0.026961	0.546618	0.078826	0.580055
LexRank	0.019466	0.252074	0.05635	0.357491
TopicLexRank	0.022548	0.283742	0.062034	0.536886
Submodular(REL)	0.028035	0.47012	0.07522	0.52629
Submodular (REL+COV)	0.023206	0.362795	0.048872	0.524863
Our Method	0.010304	0.093017	0.024551	0.116905

Table 3. Comparison of the average and max similarity between different topic summaries

Similarity between topic summaries: For each summarization method, we compute the cosine similarity between the summaries of any two topics, and then obtain the average similarity and the maximum similarity. Seen from Table 3, the topic summaries produced by our method has the lowest average and maximum similarity with each other, and thus the summaries for different topics have much difference.

5.2.2 Manual Comparison of Summarization Methods

In this section, we compare our summarization method with three typical summarization methods (MEAD, TopicLexRank and Submodular(REL)) manually. We employed three human judges to read and rank the four summaries produced for each topic by the four methods in three aspects: relevance between the summary and the topic with the corresponding sentence set, the content coverage (or diversity) in the summary and the discrimination between different summaries. The human judges were encouraged to read a few closely related documents for better understanding each topic. Note that the judges did not know which summary was generated by our method and which summaries were generated by the baseline methods. The rank k for each summary ranges from 1 to 4 (1 means the best, and 4 means the worst; we allow equal ranks), and the score is thus (4-k). We average the scores across all summaries and all judges and the results on the two datasets are shown in Tables 4 and 5, respectively. In the table, the higher the score is, the better the corresponding summaries are. We can see that our proposed method outperforms all the three baselines over almost all metrics.

	relevance	**coverage**	**discrimination**
MEAD	1.03	0.8	1.13
TopicLexRank	1.9	1.6	1.83
Submodular(REL)	2.23	2	2.07
Our Method	2.33	2.4	2.33

Table 4. Manual comparison of different summarization methods on AP news dataset

	relevance	**coverage**	**discrimination**
MEAD	1.6	1.4	1.83
TopicLexRank	1.77	2.1	2.1
Submodular(REL)	2.07	2.1	2.03
Our Method	2.43	2.17	2.1

Table 5. Manual comparison of different summarization methods on SIGMOD proceeding dataset

5.2.3 Manual Comparison of Different Kinds of Labels

In this section, we manually compare the three kinds of labels: words, phrases and summary, as mentioned in Section 5.1. Similarly, the three human judges were asked to read and rank the three kinds of labels in the same three aspects: relevance between the label and the topic with the corresponding sentence set, the content coverage (or diversity) in the label and the discrimination between different labels. The rank *k* for each kind of labels ranges from 1 to 3 (1 means the best, and 3 means the worst; we allow equal ranks), and the score is thus (3-*k*). We average the scores across all labels and all judges and the results on the two datasets are shown in Tables 6 and 7, respectively. It is clear that the summary labels produced by our proposed method have obvious advantages over the conventional word labels and phrase labels. The summary labels have better evaluation results on relevance, coverage and discrimination.

	relevance	**coverage**	**discrimination**
Word label	0.67	0.67	1.11
Phrase label	1	0.87	1.4
Summary label	1.83	1.87	1.9

Table 6. Manual comparison of different kinds of labels on AP news dataset

	relevance	**coverage**	**discrimination**
Word label	0.87	0.877	1.27
Phrase label	1.4	1.53	1.43
Summary label	1.8	1.97	1.9

Table 7. Manual comparison of different kinds of labels on AP news dataset

5.2.4 Example Analysis

In this section, we demonstrate some running examples on the SIGMOD proceeding dataset. Two topics and the three kinds of labels are shown below. For brevity, we only show the first 100 words of the summaries to users unless they want to see more. We can see that the word labels are very confusing, and the phrase labels for the two topics are totally overlapping with each other and have no discrimination. Therefore, it is hard to understand the two topics by looking at the word or phrase labels. Fortunately, by carefully reading the topic summaries, we can understand what the two topics are really about. In this example, the first topic is about data analysis and data integration, while the second topic is about data privacy. Though the summary labels are much longer than the word labels or phrase labels, users can obtain more reliable information after reading the summary labels and the summaries can help users to better understand each topic and also know the difference between different topics.

In practice, the different kinds of labels can be used together to allow users to browse topic models in a level-wise matter, as described in next section.

Topic 1 on SIGMOD proceeding dataset:

word label: *data analysis scientific set process analyze tool insight interest scenario*

phrase label: *data analysis ; data integration ; data set*

summary label: *The field of data analysis seek to extract value from data for either business or scientific benefit . ... Nowadays data analytic application are accessing more and more data from distributed data store , creating a large amount of data traffic on the network . ...these service will access data from different data source type and potentially need to aggregate data from different data source type with different data format ....Various data model will be discussed , including relational data , xml data , graph-structured data , data stream , and workflow ....*

Topic 2 on SIGMOD proceeding dataset:

word label: *user information attribute model privacy quality record result individual provide*

phrase label: *data set ; data analysis ; data integration*

summary label: *An essential element for privacy metric is the measure of how much adversaries can know about an individual ' sensitive attribute (sa) if they know the individual ' quasi-identifier (qi) ....We present an automated solution that elicit user preference on attribute and value , employing different disambiguation technique ranging from simple keyword matching , to more sophisticated probabilistic model ....Privgene need significantly less perturbation than previous method , and it achieve higher overall result quality , even for model fitting task where ga is not the first choice without privacy consideration ....*

5.2.5 Discussion of Practical Use

Although the summary labels produced by our method have higher relevance, coverage and discrimination than the word labels and the phrase labels, the summary labels have one obvious shortcoming of consuming more reading time of users, because the summaries are much longer than the words and phrases. The feedback from the human judges also reveals the above problem and all the three human judges said they need to take more than five times longer to read the summaries. Therefore, we want to find a better way to make use of the summary label in practice.

In order to consider both the shorter reading time of the phrase labels and the better quality of the summary labels, we can use both of the two kinds of labels in the following hierarchical way:

For each topic, we first present only the phrase label to users, and if they can easily know about the topic after they read the phrase label, the summary label will not be shown to them. Whereas, if users cannot know well about the topic based on the phrase label, or they need more information about the topic, they may choose to read the summary label for better understanding the topic. Only the first 100 words of the summary label are shown to users, and the rest words will be shown upon request. In this way, the summary label is used as an important complement to the phrase label, and the burden of reading the longer summary label can be greatly alleviated.

6 Conclusions and Future Work

In this study, we addressed the problem of topic labeling by using text summaries. We propose a summarization algorithm based on submodular optimization to extract representative summaries for all the topics. Evaluation results demonstrate that the summaries produced by our proposed algorithm have high relevance, coverage and discrimination, and the use of summaries as labels has obvious advantages over the use of words and phrases.

In future work, we will explore to make use of all the three kinds of labels together to improve the users' experience when they want to browse, understand and leverage the topics.

In this study, we do not consider the coherence of the topic summaries because it is really very challenging to get a coherent summary by extracting different sentences from a large set of different documents. In future work, we will try to make the summary label more coherent by considering the discourse structure of the summary and leveraging sentence ordering techniques.

Acknowledgments

The work was supported by National Natural Science Foundation of China (61331011), National Hi-Tech Research and Development Program (863 Program) of China (2015AA015403) and IBM Global Faculty Award Program. We thank the anonymous reviewers and mentor for their helpful comments.

References

Nikolaos Aletras, and Mark Stevenson. 2013. Representing topics using images. *HLT-NAACL.*

Nikolaos Aletras, Timothy Baldwin, Jey Han Lau, and Mark Stevenson. 2015. Evaluating topic representations for exploring document collections. *Journal of the Association for Information Science and Technology* (2015).

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

Ying-Lang Chang and Jen-Tzung Chien. 2009. Latent Dirichlet learning for document summarization. *Proccedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2009).*

Güneş Erkan and Dragomir R. Radev. 2004. LexPageRank: Prestige in multi-document text summarization. In *Proceedings of EMNLP*.

Dan Gillick, Benoit Favre, and Dilek Hakkani-Tur. 2008. The ICSI summarization system at TAC 2008. In *Proceedings of the Text Understanding Conference*.

Thomas Hofmann. 1999. Probabilistic latent semantic indexing. *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM.

Ioana Hulpus, Conor Hayes, Marcel Karnstedt, and Derek Greene. 2013. Unsupervised graph-based topic labelling using dbpedia. *Proceedings of the sixth ACM international conference on Web search and data mining*. ACM.

Wanqiu Kou, Fang Li, and Timothy Baldwin. 2015. Automatic labelling of topic models using word vectors and letter trigram vectors. *Information Retrieval Technology*. Springer International Publishing, 253-264.

Jey Han Lau, Karl Grieser, David Newman, and Timothy Baldwin. 2011. Automatic labelling of topic models. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics.

Hui Lin and Jeff Bilmes. 2010. Multi-document summarization via budgeted maximization of submodular functions. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Hui Lin and Jeff Bilmes. 2011. A class of submodular functions for document summarization. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics.

Qiaozhu Mei, Chao Liu, Hang Su, and ChengXiang Zhai. 2006. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *Proceedings of the 15th international conference on World Wide Web*, pp. 533-542. ACM.

Qiaozhu Mei, Xuehua Shen, and ChengXiang Zhai. 2007. Automatic labeling of multinomial topic models. *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM.

Xian-Ling Mao, Zhao-Yan Ming, Zheng-Jun Zha, Tat-Seng Chua, Hongfei Yan, and Xiaoming Li. 2012. Automatic labeling hierarchical topics. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pp. 2383-2386. ACM.

You Ouyang, Sujian Li, and Wenjie Li. 2007. Developing learning strategies for topic-based summarization. *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. ACM.

Dragomir R. Radev, Hongyan Jing, Małgorzata Styś, and Daniel Tam. 2004. Centroid-based summarization of multiple documents. *Information Processing & Management* 40, no. 6: 919-938.

Dou Shen, Jian-Tao Sun, Hua Li, Qiang Yang, and Zheng Chen. 2007. Document summarization using Conditional Random Fields. In *IJCAI*, vol. 7, pp. 2862-2867.

Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences* 101.suppl 1: 5228-5235.

Xiaojun Wan. 2008. Using only cross-document relationships for both generic and topic-focused multi-document summarizations. *Information Retrieval* 11.1: 25-49.

Xiaojun Wan, Jianwu Yang, and Jianguo Xiao. 2007. Manifold-ranking based topic-focused multi-document summarization. In *IJCAI*, vol. 7, pp. 2903-2908.

Xiaojun Wan and Jianwu Yang. 2008. Multi-document summarization using cluster-based link analysis. *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM.

Xuerui Wang and Andrew McCallum. 2006. Topics over time: a non-Markov continuous-time model of topical trends. *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM.