CONCEPT-BASED CLASSIFICATION FOR MULTIDOCUMENT SUMMARIZATION

Asli Celikyilmaz
University of California, Berkeley
Berkeley, USA
asli@eecs.berkeley.edu

Dilek Hakkani-Tür∗
Speech at Microsoft — Microsoft Research
Mountain View, CA, USA
dilek@ieee.org

ABSTRACT
Documents often contain inherently many concepts reflecting specific and generic aspects. To automatically generate a short summary text of documents on similar topics, it is imperative that we discover general aspects in documents because summaries usually contain general rather than specific concepts. This paper presents a semi-supervised extractive summarization model based upon latent concept classification that can differentiate between the two types of aspects as hidden concepts being mentioned in documents. A classifier is trained on hidden concepts discovered from documents and their corresponding human-generated summaries using a probabilistic Bayesian model: the summary-focused topic model. Experimental results based on ROUGE evaluations indicate that ranking sentences to be included in summary text based on the latent summary concept classification has improvements on the quality of the generated summaries.

Index Terms— Automatic document summarization, latent topic classification.

1. INTRODUCTION
Multi-document summarization is the task of generating a short summary of documents retrieved from a search engine in response to a user formulated query. Similarly in a speech summarization task, such documents are generated by automatic speech recognition. In such tasks, either the most important aspects of documents are reflected in a summarization task or the most relevant aspects related to user’s information need is the basis in generating a short summary text. Document Understanding Conferences (DUC) fosters such tasks [?] by providing data sets and organizing public evaluations.

There are two streams of previous research directions on multi-document summarization for identifying the most important sentences to be included in a short summary text: One line of research builds language models or discriminative classifier models on sentences using common features such as term-frequency or sentence-length [?, ?]. These models lack the semantic content information in documents and their summaries. Other line of research is based upon generative models, whose aim is to capture hidden concepts in documents and reflect them on the generated summaries. Most of these studies are based on Latent Dirichlet Allocation (LDA) [?] style topic models or several extensions of LDA to encode the lexical variability of documents in summarization modeling. However one issue with these models is that they can yield too general results [?]. To the best of our knowledge, there is very little interest on probabilistic topic models to predict summaries for new documents via previously trained models.

In [?] a new summary focused topic model (sLDA) is presented to tackle with the problems above. sLDA discovers explicitly the summary and non-summary related topics on sentence level under the assumption that sentences convey a particular content. They build a discriminative model to directly classify each sentence in the documents as summary (non-)/related by the discovery of hidden summary concepts contained in them. Then sentences are ranked based on their posterior probability according to a classification model so as to choose salient sentences to be included in summary text.

Our purposes in this paper are not very different. We implement the sLDA model to discover two separate sets of concepts: generic and specific, based upon the gold-summaries used as supervision. An important distinction with the previous model and the one presented in this paper is that: in the previous summarization model the concepts discovered were not interpretable. More specifically, in this paper we would like to build a classifier for hidden concepts to characterize which ones are summary or non-summary related in the new documents. For instance, we will be able to represent a hidden concept extracted from documents via the shared terms in the documents and gold-summaries (summaries generated by humans from the given set of documents), and hence they will be interpretable by the domain experts and it will be easier to cluster sentences as (non-)salient based on the ngrams.

After we discover the summary and non-summary related hidden concepts, i.e., specific and generic concepts, for each individual document cluster, we combine all the information on discovered concepts, i.e, the expected posterior sentence-topic (θ) and sentence-term latices (φ) from sLDA, and build a classifier on concepts via the meta-information we extract from the documents as input. The classifier model enables to infer the hidden concepts we capture in testing documents as

∗This research was done while the first author was working at the International Computer Science Institute of Berkeley.
summary related or not. The framework of the concept-based classification model for automatic document summarization is given in Fig. ??.

In the next we present the concept based classification of hidden topics extracted from sLDA and the features used for the model. In § 3 the inference and our sentence scoring methods for ranking salient sentences to generate summary text is presented. Later in § 4, we present the experiments on DUC evaluations and finally we draw conclusions.

Fig. 1. Framework of the extractive summary generation process using concept-based classification model.

2. CONCEPT-BASED CLASSIFICATION - CBC - FOR MULTI-DOCUMENT SUMMARIZATION

In the DUC evaluations, a set of documents, namely a document cluster, $d \in D$, each containing documents on similar topics based upon a user-query, as well as a set of (human-generated) gold-summaries are provided for training purposes. The aim is to generate a word-limited summary text for each document-cluster $d$. DUC fosters the automatic generation of summaries of multiple-documents, which are compiles of newswire articles on the same topic. The task is to extract salient sentences from documents in each $d$ to generate coherent and non-redundant short summaries that would also have high ROUGE measure, a measure based on word n-gram recall performance of generated summaries of a set of document clusters against gold summaries.

Summary-focused LDA [?] is an extension of LDA [?] which is based on two assumptions: (1) sentences in summaries represent generic topics (concepts) which are discovered from document clusters; (2) the rest of the concepts are usually specific to the document clusters. sLDA generative process and graphical model is shown in Fig. ??.

Fig. 2. Summary sentence identification through sLDA.

The sentences $M = \{o_m \}_{m=1}^M$ in a document cluster $D = \{d_i \}_{i=1}^D$ are represented by a list of tokens $o_m = \{w_1, ..., w_{N_m}\}$, where $w_i \in W = \{w_1, ..., w_V\}$. $K$ topics are represented by multinomial distributions over $V$ unique terms, where $P(w_i|z_k) = \phi_k^{(w)}$, $k = 1, ..., K$ is the probability of a word given topic $z_k$. Each sentence $s$ is associated with a multi-nomial distribution representing the sparse-topic distributions as a mixture of components of sentences. The

Parameter $y$, known a priori determines from which part of the sentence-topic mixing distribution the topics should be sampled. Specifically, if a word $w_j$ in a sentence exists in a summary ($y_i = 1$), then the sentence-topic mixing variables $\theta$ for topics $k = 1, ..., S$– allocated for generic topics– are updated, otherwise, only the specific topics $k = S+1, ..., (K=S+R)$ are sampled. Hence the first $S$ topics are designated for generic (summary-related) and the rest $R=K-S$ are designated to be samples from specific topics and the mixing distributions are updated accordingly. The update equations for the Gibbs sampling are given in detail in [7].

In our concept-based classification model, we initially build one sLDA model on each document cluster along with their human generated summaries to extract the expected posterior probabilities from topic-word $(\phi)$ and sentence-topic distribution $(\theta)$ matrices (lattices). Note that, based on the sLDA output, out of $K$ topics, the first $S$ topics are designated for generic concepts and the rest $R=K-S$ are for specific concepts. This distinction is captured via a semi-supervised sLDA model, where the gold-summaries are used to shape the distribution and extract two separate concepts. At testing time, because we won’t have the gold summaries, we propose to build a discriminative classifier model, with which we can discover certain characteristics of the two types of concepts. In addition, since we use a linear classifier, i.e., Maximum Entropy (Max-Ent) model [7], the discovered concepts are more interpretable compared to the black-box generative models, where only topics are inferred but not explicitly characterized [?]. In summary, using the extracted hidden topics in document clusters via sLDA, we train a concept-based classifier (CBC) that can classify topics (concepts) as generic or specific in unseen(test) documents.

CBC Model: We use the Max-Ent model, also known as log-linear and exponential models, with which we can integrate features from many heterogeneous information sources for classification. In our models we use meta-features to characterize hidden topics, which are organized according to document and sentence clusters. Given a training set $\{x, y\}$, where $x$ is a multi-dimensional vector of data points (in our models for each hidden topic) comprised of set of features and $y \in \{0, 1\}$ is a set of class labels, Max-Ent model maximizes
the log-likelihood:

\[
\log P(y = 1 | x, \beta) = \sum_{(x_i, y_i) \in (x, y)} \log \frac{\sum_{j} \beta_j f_j(x_i, y_i)}{\sum_{j} \beta_j f_j(x_i, y_i)}
\]

where \(f_i(x_i, y_i)\) are feature indicator functions and \(\beta_i\) are the parameters needed to be estimated which reflects the importance of each feature \(f_i(x_i, y_i)\) in prediction.

**CBC Features:** Our training dataset consists of the generic and specific topics discovered from each document clusters. Hence, each sLDA model generates \(K\) hidden-concepts on sentences \(z_k^{mt}\) individually from the \(D\) document clusters \(D\), a total of \(N=K^D\) number of concepts(topics) are extracted to compile our training set of \(N\) data points, i.e., hidden concepts. Each concept has a latent binary output variable representing the type of the topic (i.e., generic or specific), also predicted from the sLDA model. Since the classifier is on topic level, we extract concept related meta-features based on term-frequencies as follows:

- **Given a document cluster:** we first identify the most frequent (non-stop word) unigrams and as bi-grams, i.e.,
  \(v_i^u = \{w_i^u\}_{i=1}^{f_1}, v_i^{b} = \{w_i^{*}\}_{i=1}^{f_2}\)
  where \(f_1\) and \(f_2\) are model parameters for the number of most frequent uni/bi-gram features. Given a sentence \(o_m\), let \(d\) represent the document that \(o_m\) belongs to, i.e., \(o_m \in d\). We measure unigram probabilities for each \(w_i\) in \(o_m\) by \(p(w_i) = n_d(w_i)/n_d(w_i)\), where \(n_d(w_i)\) is the frequency of \(w_i\) in document \(d\) and \(n_d(w_i)\) is the frequency of \(w_i\) in the document cluster \(D\).

- **Given a hidden concept \(z_k^{mt}\) in a sentence \(o_m\), we measure:**

\[
x_{im} = \hat{p}(z_k^{mt} | w_i^*) * p(w_i^*)
\]

The first term in (2) is the posterior probability of a hidden topic \(z_k\) given a word \(w_i^*\) that is among the most frequent \(f_1\) words. We generate similar features for bi-grams as well.

- **We also introduce another count variable for each concept to characterize the percentage of the most frequent uni-grams \(w^*\) that are highly likely to be generated:**

\[
x_{im} = \frac{1}{T} \sum_{i=1}^{f_1} I[\hat{p}(z_k^{mt} | w_i^*) \geq T] \quad (2)
\]

\(I(\cdot) \in \{0, 1\}\) is an indicator variable, and \(T\) is a user defined threshold value, e.g., \(T=0.1\) is used in the experiments. The same feature is used in the model for bi-grams, i.e., \(w_i^{**}\).

### 3. INFERENCEx AND SENTENCE SCORING

- **Inference:** Once we train the Max-Ent to classify the hidden concepts discovered from sLDA model, we turn to test set to infer the summary sentences so as to generate summary text. Since we do not have the gold-standard summaries at testing time, we build standard LDA [2] model on each test document cluster and predict the sentence-topic and topic-word mixing proportions for expected posterior probabilities.

Each sentence in test document cluster is represented as \(MT = \{o_m\}_{m=1}^{|MT|}\). We discover \(K\) number of hidden concepts in each sentence of testing documents using LDA. The idea is to infer the class of each the hidden topics, either generic or specific, using the CBC model. Thus, we use the CBC model to predict the labels of \(K\) topics, i.e., generic or specific and obtain \(\hat{p}(y_{zk}^{mt} = 1)\), the likelihood of a topic \(z_k^{mt}\) of testing sentence \(o_m\) representing a generic topic. Using the estimated likelihood information of each topic in a given sentence, we calculate sentence rank scores based on the frequency of predicted hidden generic topics they contain.

- **Sentence Scoring:** From the CBC model we predict the class for each hidden topic of each sentence. Thus, we construct function to measure the score of each sentence reflecting the predicted likelihood of \(K\) hidden topics onto the sentences. However, it should be noted that hidden topic-based scores can be too coarse to generate high quality summaries. Thus, to improve the salient sentence score performance, we interpolate the scores obtained from the previous sLDA model onto the CBC topic classification scores and assign predicted probability rank scores of each sentence as follows:

\[
r(s_m^{\text{CBC}}) = \gamma_1 + \frac{1}{K} \sum_{k=1}^{K} \hat{p}(y_{zk}^{mt} = 1) + \gamma_2 * r(s_m^{\text{sLDA}})
\]

where \(m = 1...M\) and \(r(s_m^{\text{sLDA}})\) is the estimated score for sentence \(s_m\) predicted via sLDA [2]. It’s also been shown that the term frequencies play an important role in determining the saliency of sentences to generate summary text [?]. Therefore, we incorporate term frequency score to the predicted sentence score in (3). We use SumBasic [2], which assigns a score to each sentence \(s_m\) based upon the count of high frequency words it contains as:

\[
r(s_m^{\text{uni}}) = \sum_{w \in s_m} \frac{1}{|s_m|} p_d(w)
\]

where \(p_d(w)\) is the observed unigram probabilities in the document cluster \(d\), and \(|s_m|\) is the total number of words in sentence \(s_m\). In addition to unigrams, we also measure the bigram probabilities \(r(s_m^{bi})\) and interpolate each score as:

\[
r(s_m) = A * r(s_m^{\text{CBC}}) + B * r(s_m^{\text{uni}}) + C * r(s_m^{bi})
\]

The score in (3) is used as the final sentence score in ranking salient sentences for CBC model.

- **Redundancy Elimination:** To eliminate the redundant sentences from the generated summaries, we incrementally add onto the summary the highest ranked sentence \(o_m\), and check if the newly added sentence significantly repeats the information already included in the summary until the algorithm reaches word count limit. We us a word overlap measure between sentences normalized to sentence length. A sentence is discarded if its similarity to any of the previously selected sentences is greater than a threshold identified by a greedy search on the training dataset.

### 4. EXPERIMENTS

In this section we describe a number of experiments using our CBC model on 100 document clusters each containing 25 news articles from DUC2005-2006 tasks. We evaluate
Table 1. ROUGE results of the best systems on DUC2007 dataset (best results are bolded.)

<table>
<thead>
<tr>
<th>ROUGE</th>
<th>w/o stop words</th>
<th>w/ stop words</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R-1  R-2  R-4</td>
<td>R-1  R-2  R-4</td>
</tr>
<tr>
<td>Baseline</td>
<td>32.4  7.4  10.6</td>
<td>41.0  9.3  15.2</td>
</tr>
<tr>
<td>PYTHY</td>
<td>35.7  8.9  12.1</td>
<td>42.6  11.9 16.8</td>
</tr>
<tr>
<td>HIERSUM</td>
<td>33.8  9.3  11.6</td>
<td>42.4  11.8 16.7</td>
</tr>
<tr>
<td>sLDA</td>
<td>36.2  9.0  12.5</td>
<td>45.4  11.5 17.3</td>
</tr>
<tr>
<td>CBC</td>
<td>35.1  9.2  12.1</td>
<td>46.1  11.7 17.5</td>
</tr>
</tbody>
</table>

our performance using 45 document clusters each containing 25 news articles from DUC2007 task. From these sets, we collected ~80K and ~25K sentences to compile training and testing data respectively. The task is to create max. 250 word long summary for each document cluster.

We use Gibbs sampling for inference in sLDA and LDA models. The sLDA is used to capture generality and specificity of topics in sentences based on gold-summaries. Thus we set different parameters while sampling the designated topics for sentence-topic mixing updates. Dirichlet hyper-parameters generally have smoothing effect on multinomial parameters. Reducing this smoothing effect in sLDA by setting lower values for $\alpha$, $\alpha^s$ and $\beta$ parameters can result in more decisive topic associations. Sparsity of topic-word distributions is controlled by the $\alpha$ and $\alpha^s$ parameters, hence, we set $\alpha^s$, $\alpha \leq 1$, so as to generate more sparse topic-term distributions rather than uniform distributions.

ROUGE is used for performance measure [8, 9], to evaluate summaries based on the maximum number of overlapping units between generated summary text and a set of human summaries. We use R-1 (recall against unigrams), R-2 (recall against bigrams), and R-SU4 (recall against skip-4 bigrams). Note that these measures favor inclusion for stop-words and hence the more word overlap the better the result.

We use the following models for benchmark purposes:

* **PYTHY** : [1] A state-of-the-art supervised summarizer that ranked first in overall ROUGE evaluations in DUC2007. Similar to HybHSum, human generated summaries are used to train a sentence ranking system using a classifier model.

* **HIERSUM** : [1] A generative method based on topic models, which uses sentences as an additional level. Using an approximation for inference, sentences are greedily added to a summary so long as they decrease KL-divergence.

* **sLDA** : [2] A generative summarization model that our CBC model of this paper extends. It assigns a score to each sentence via direct sentence classification using concept based features, e.g., the total number of summary related topics a sentence includes, etc. The concepts a sentence contains are not explicitly classified as in CBC model of this paper.

We also compared results to a Baseline method, which simply replaces the scoring algorithm of our sLDA with a simple scoring function, which is measured by its lexical similarity (cosine distance) to maximum matching summary sentence. Then we build a regression model with the same variables as our sLDA to create a summary. As depicted in Table ??, the R-2 results of the topic classification based summary ranking model improves the results compared to the sLDA and other state-of-the-art summarization models on this dataset. In summary, the hidden topic classification can be used to classify salient sentences on a higher level of abstraction via Bayesian network models.

5. CONCLUSIONS

In this paper, we present a latent concept classification algorithm based upon the output of a probabilistic topic model for the problem of multi-document summarization. Our approach is based on learning summary content distributions from document sets using provided summary texts as supervision. We find that a simple discriminative learning method trained explicitly the generated latent concepts from the topic model can produce comparable results to the state-of-the-art multi-document summarization methods.

One prominent feature of our proposed modeling tool is that it can be used not only for summarization tasks, but also for any other problem related to information extraction, where hidden topics can be extracted from unlabeled collections. In this sense, our approach is generic and forms a baseline. One area of direction is making the approach robust to errors in automatic speech recognition and spontaneous speech.

6. REFERENCES


