

Personalized Multi-Document Summarization using N-Gram Topic Model Fusion

Leonhard Hennig, Sahin Albayrak

DAI Labor, TU Berlin
Berlin, Germany
firstname.lastname@dai-labor.de

Abstract

We consider the problem of probabilistic topic modeling for query-focused multi-document summarization. Rather than modeling topics as distributions over a vocabulary of terms, we extend the probabilistic latent semantic analysis (PLSA) approach with a bigram language model. This allows us to relax the conditional independence assumption between words made by standard topic models. We present a unified topic model which evolves from sentence-term and sentence-bigram co-occurrences in parallel. Sentences and queries are represented as probability distributions over latent topics to compute thematic and query-focused sentence features in the topic space. We find that the inclusion of bigrams improves the descriptive quality of the latent topics, and substantially reduces the number of latent topics required for representing document content. Experimental results on DUC 2007 data show an improved performance compared to a standard term-based topic model. We further find that our method performs at the level of current state-of-the-art summarizers, while being built on a considerably simpler model than previous topic modeling approaches to summarization.

1. Introduction

Automatically producing summaries from input documents is an extensively studied problem in Information Retrieval and Natural Language Processing (Jones, 2007). In this paper, we investigate the problem of multi-document summarization (MDS), where the task is to “synthesize from a set of related documents a well-organized, fluent answer to a complex question”¹. In particular, we focus on generating an extractive summary by selecting relevant sentences from a set of related documents (Goldstein et al., 2000). The condensation of information from different sources into an informative summary is an increasingly important task, since it helps to reduce information overload.

A major challenge in identifying relevant information is to model document content. A document will generally contain a variety of information centered around a main topic, and covering different aspects (subtopics) of this main theme (Barzilay and Lee, 2004). Human summaries also tend to cover different aspects of the original source text to increase the informative content of the summary. In addition, in query-focused multi-document summarization tasks, the user query often explicitly requests information about different aspects of the main theme of the document cluster (see Table 1). An ideal summary should therefore aim to include information for each of the “subquestions” of the complex user query.

Various summarization approaches have exploited observable features based on the identification of topics (or thematic foci) to construct summaries. Often, such features rely on the identification of important keywords (Yih et al., 2007; Nenkova et al., 2006), or on the creation of term-based topic signatures (Lin and Hovy, 2000; Conroy et al., 2007). However, it is well known that term matching has severe drawbacks due to the ambivalence of words and to differences in word usage across authors (Manning

and Schütze, 2001). This is especially important for automatic summarization, as summaries produced by humans may differ significantly. (Lin and Hovy, 2003b).

Topic models such as Latent Semantic Indexing (LSI) (Deerwester et al., 1990), Probabilistic Latent Semantic Analysis (PLSA) (Hofmann, 1999), and Latent Dirichlet Allocation (LDA) (Blei et al., 2003) provide a means to overcome the problem of term matching, and furthermore allow for the modeling of inter- and intradocument statistical structure. These models introduce hidden variables as the origin of the observed term co-occurrences. Whereas LSI is a mapping of the original term-document vector space onto a linear subspace based on singular value decomposition, PLSA and LDA model documents as a distribution of mixture components, where each mixture component is a multinomial distribution over words. The mixture components can be interpreted as “topics”, and the corresponding word distributions characterize the semantics of the topics.

The reduced description will typically capture some aspects of synonymy and polysemy, since words with similar meanings tend to occur in similar contexts. Furthermore, semantically similar words are clustered based on the assumption that the co-occurrence of terms signals semantic relatedness. However, words are considered independent given the topics, resulting in the standard bag-of-words assumption (Blei et al., 2003). N-Gram language models (Ponte and Croft, 1998) allow us to relax this assumption in order to capture multi-word concepts, where word order plays a critical role (Wang et al., 2007).

1.1. Our contribution

Our approach extends the standard topic modeling approach such that the topic model is estimated from the term co-occurrence as well as bigram co-occurrence observations in parallel. The integration of a bigram language model allows us to represent the mixture components as multinomial distributions over terms and bigrams, leading

¹Document Understanding Conference summarization track task description, <http://www.nist.gov/tac>

Table 1: A complex user query from DUC 2006.

ID	D0631D
Title	Crash of the Air France Concorde
Query	Discuss the Concorde jet, its crash in 2000, and aftermaths of this crash.

to an improved representation of the components. Each document’s distribution over the mixture components is estimated based on maximizing the likelihood of the observed data given both the term co-occurrence and the bigram co-occurrence distributions.

Furthermore, the integration of the bigram language model allows us to relax the independence assumption of terms made by the standard topic model, since bigrams encode syntactic dependencies between consecutive terms. Even though one can consider a bigram simply to be a co-occurrence of two terms, and as such captured well enough by a standard topic model, our assumption is that bigram co-occurrence patterns will reinforce the observed term co-occurrence patterns. We show that this results in more descriptive latent topics, and considerably reduces the number of latent topics required for a good model.

We evaluate the modified topic model on the task of query-focused multi-document summarization by modeling sentences and queries in this novel latent topic space. This allows us to compute thematic and query-focused sentence similarity features for extractive summarization.

The rest of this paper is structured as follows: We start with an overview of related work in Section 2. In Section 3. we describe our approach for integrating a language model into the PLSA algorithm. Next, in Section 4., we give details of our summarization system, the sentence-level features we use, and of our sentence ranking and selection approach. In Section 5., we describe and analyze the data sets we use to verify the assumptions of our approach, and we present experimental results. Finally, Section 6. concludes the paper.

2. Related work

Probabilistic topic models for the representation of document content have also been explored by Barzilay and Lee (Barzilay and Lee, 2004). They use Hidden Markov Models (HMM) to model topics and topic change in text, albeit only for generic single-document summarization. The model assumes that a topic is formed by clustering sentences based on vector space similarity, and bigram distribution patterns are learned from these topical clusters. Each sentence is assigned to exactly one topic cluster, corresponding to a HMM state. Documents are modeled as sequences of topics, representing the typical discourse structuring of texts in specific domains. In contrast, our approach models each sentence as a distribution over multiple topics, and also models queries in the latent topic space for query-focused multi-document summarization.

More related to our method is the approach of Daumé and Marcu (Daumé and Marcu, 2006), who utilize a model similar in style to LDA. However, the latent classes are chosen to capture general language background vocabulary, document- and query-specific vocabularies. Each sentence is modeled as a distribution over these three mixture components, e.g. consisting of 60% query information, 30%

background document information, and 10% general English (Daumé and Marcu, 2006). Topical information is not considered, and neither are the subtopics contained in a document.

The method proposed by Haghighi and Vanderwende takes up this approach, but constructs summaries by optimizing the KL-divergence between the summary topic distribution and the topic distribution of the source document set (Haghighi and Vanderwende, 2009). Subtopics are modeled by introducing a hierarchical LDA process. Instead of drawing words only from a generic “content” distribution they allow for either generic or topic-specific word distributions for each sentence. However, for each sentence only one distribution is selected, and all content words of that sentence are drawn from this distribution. Topic-specific distributions are ordered sequentially over sentences similar to the approach of Barzilay and Lee. The proposed approach does not address query-focused summarization.

In previous work, we showed that a term-sentence co-occurrence based PLSA model can be effectively used for query-focused multi-document summarization (Hennig, 2009). The proposed model outperformed existing systems on DUC 2006 data, and performed comparable to state-of-the-art summarization systems on the DUC 2007 dataset.

All of the above methods consider either unigram or bigram distributions for representing topics, but not the combination of both. In our approach, we combine unigram and bigram observations to create topic representations that consist of multinomial distributions over both unigrams and bigrams.

In the area of topic modeling, Wallach proposed an approach to relax the bag-of-words assumption in (Wallach, 2006). The LDA model she discusses incorporates, in a fashion similar to typical n-gram language models, the conditional probability of a word at position t given the word at position $t - 1$, such that $p(w_t) = p(w_t|w_{t-1})$. Each topic is represented as a set of W distributions – contrasting with the single word distribution per topic typically used – where W is the size of the vocabulary. Each of the W word distributions per topic is conditioned on the context of a previous word w_{t-1} . The number of parameters to be estimated is hence $WT(W - 1)$, whereas our model has $(W - 1)T + (B - 1)T$ free parameters (B is the number of distinct bigrams).

3. Topic and Language Model Combination using PLSA

For simplicity, we utilize and adapt the PLSA algorithm to test the validity of our approach, but for all purposes this can be considered equivalent to using a more complex topic model such as LDA.

PLSA is a latent variable model for co-occurrence data

that associates an unobserved class variable $z \in \mathcal{Z} = \{z_1, \dots, z_k\}$ with each observation (d, w) , where word $w \in \mathcal{W} = \{w_1, \dots, w_i\}$ occurs in document $d \in \mathcal{D} = \{d_1, \dots, d_j\}$. Documents and words are assumed independent given the topic variable Z . The probability that a word occurs in a document can be calculated by summing over all latent variables Z :

$$P(w_i|d_j) = \sum_k P(w_i|z_k)P(z_k|d_j). \quad (1)$$

Similarly, we can associate each observation (d, b) of a bigram $b = (ww')$, where bigram $b \in \mathcal{B} = \{b_1, \dots, b_l\}$ occurs in document d , with the same unobserved class variable z . We assume the same hidden topics of the term-sentence co-occurrences (d, w) as the origin of the bigram-sentence co-occurrence observations (d, b) :

$$P(b_l|d_j) = \sum_k P(b_l|z_k)P(z_k|d_j). \quad (2)$$

Notice that both decompositions share the same document-specific mixing proportions $P(z_k|d_j)$. This couples the conditional probabilities for terms and bigrams: each ‘‘topic’’ has some probability $P(b_l|z_k)$ of generating bigram b_l as well as some probability $P(w_i|z_k)$ of generating an occurrence of term w_i . The advantage of this joint modeling approach is that it integrates term and bigram information in a principled manner. This coupling allows the model to take evidence about bigram co-occurrences into account when making predictions about terms and vice versa. Following the procedure in Cohn and Hofmann (Cohn and Hofmann, 2000), we can now combine both models based on the common factor $P(z|d)$ by maximizing the log-likelihood function

$$L = \sum_j \left[\alpha \sum_i n(d_j, w_i) \log P(w_i, d_j) + (1 - \alpha) \sum_l n(d_j, b_l) \log P(b_l, d_j) \right] \quad (3)$$

where α is a predefined weight for the influence of each two-mode model. Using the Expectation-Maximization (EM) algorithm we then perform maximum likelihood parameter estimation for the aspect model. During the expectation (E) step we first calculate the posterior probabilities:

$$P(z_k|w_i, d_j) = \frac{P(w_i|z_k)P(z_k|d_j)}{P(w_i|d_j)} \quad (4)$$

$$P(z_k|b_l, d_j) = \frac{P(b_l|z_k)P(z_k|d_j)}{P(b_l|d_j)}, \quad (5)$$

and then re-estimate parameters in the maximization (M) step as follows:

$$P(w_i|z_k) = \sum_j \frac{n(w_i, d_j)}{\sum_{i'} n(w_{i'}, d_j)} P(z_k|w_i, d_j) \quad (6)$$

$$P(b_l|z_k) = \sum_j \frac{n(b_l, d_j)}{\sum_{l'} n(b_{l'}, d_j)} P(z_k|b_l, d_j) \quad (7)$$

The class-conditional distributions are recomputed in the M-step as

$$p(z_k|d_j) \propto \alpha \sum_i \frac{n(w_i, d_j)}{\sum_{i'} n(w_{i'}, d_j)} P(z_k|w_i, d_j) + (1 - \alpha) \sum_l \frac{n(b_l, d_j)}{\sum_{l'} n(b_{l'}, d_j)} P(z_k|b_l, d_j) \quad (8)$$

Based on the iterative computation of the above E and M steps, the EM algorithm monotonically increases the likelihood of the combined model on the observed data. Using the α parameter, our new model can be easily reduced to a term co-occurrence based model by setting $\alpha = 1.0$.

4. Topic-based summarization

Our approach for producing a summary consists of three steps: First, we represent sentences and queries in the latent topic space of the combined PLSA model by estimating their mixing proportions $P(z|d)$. We then compute several sentence-level features based on the similarity of sentence and query distributions over latent topics. Finally, we combine individual feature scores linearly into an overall sentence score to create a ranking, which we use to extract sentences for the summary. We follow a greedy approach for selecting sentences, and penalize candidate sentences based on their similarity to the partial summary. These steps are described in detail below.

4.1. Data Set

We conduct our analysis and evaluate our model based on the multi-document summarization data sets provided by DUC². Specifically, we use the DUC 2007 data set for evaluation. The data set consists of 45 document clusters, with each cluster containing 25 news articles related to the same general topic. Participants are asked to generate summaries of at most 250 words for each cluster. For each cluster, a title and a narrative describing a user’s information need are provided. The narrative (query) is usually composed of a set of questions or a multi-sentence task description.

4.2. Sentence representation in the latent topic space

Given a corpus \mathcal{D} of topic-related documents, we perform sentence splitting on each document using the NLTK toolkit³. Each sentence is represented as a bag-of-words $\mathbf{w} = (w_1, \dots, w_m)$. We remove stop words for the unigram model, and apply stemming using Porter’s stemmer (Porter, 1980). We create a term-sentence matrix TS containing all sentences of the corpus, where each entry $TS(i, j)$ is given by the frequency of term i in sentence j , and a bigram-sentence matrix BS , where each entry $BS(l, j)$ is given by the frequency of bigram l in sentence j . We then train the combined PLSA model on the matrices TS and BS .

After the model has been trained, it provides a representation of the sentences as probability distributions $P(z|s)$ over the latent topics Z , and we arrive at a representation of sentences as a vector in the ‘‘topic space’’:

$$\mathbf{s} = (p(z_1|s), p(z_2|s), \dots, p(z_K|s)), \quad (9)$$

²<http://www.nist.gov/tac>

³<http://www.nltk.org>

where $p(z_k|s)$ is the conditional probability of topic k given the sentence s .

In order to produce a query-focused summary, we also need to represent the query in the latent topic space. This is achieved by folding the query into the trained model. The folding is performed by EM iterations, where the factors $P(w|z)$ and $P(b|z)$ are kept fixed, and only the mixing proportions $P(z|q)$ are adapted in each M-step (Hofmann, 1999). We fold the title and the query of the document clusters, the document titles, and document and cluster vectors into the trained PLSA model. Query vectors are preprocessed in the same way as training sentences, except that no sentence splitting is performed. Document and document cluster term vectors are computed by aggregating sentence term vectors.

4.3. Computing query- and topic-focused sentence features

Since we are interested in producing a summary that covers the main topics of a document set and is also focused on satisfying a user’s information need, specified by a query, we create sentence-level features that attempt to capture these different aspects in the form of per-sentence scores. We then combine the feature scores to arrive at an overall sentence score. Each feature is defined as a similarity $r(S, Q)$ of a sentence topic distribution $S = P(z|s)$ compared to a “query” topic distribution $Q = P(z|q)$:

- $r(S, CT)$ - similarity to the cluster title
- $r(S, N)$ - similarity to the cluster narrative (query)
- $r(S, T)$ - similarity to the document title
- $r(S, D)$ - similarity to the document centroid
- $r(S, C)$ - similarity to the cluster centroid

Since measures for comparing two probability distributions are typically defined as divergences, not similarities, we invert the computed divergence. In our approach, we employ the Jensen-Shannon (JS) divergence, but a variety of other similarity measures can be utilized towards this end. The JS-divergence is a symmetrized and smoothed version of the Kullback-Leibler divergence:

$$r_{JS}(S, Q) = 1 - \left[\frac{1}{2} D_{KL}(S||M) + \frac{1}{2} D_{KL}(Q||M) \right], \quad (10)$$

where $M = 1/2(S + Q)$.

As the training of a PLSA model using the EM algorithm with random initialization converges on a local maximum of the likelihood of the observed data, different initializations will result in different locally optimal models. As Brants et al. (Brants et al., 2002) have shown, the effect of different initializations can be reduced by generating several PLSA models, then computing features according to the different models, and finally averaging the feature values. We have implemented this model averaging using 5 iterations of training the PLSA model.

4.4. Sentence scoring

The system described so far assigns a vector of feature values to each sentence. The overall score of a sentence consisting of the features (r_1, \dots, r_P) is then defined as:

$$score(s) = \sum_p w_p r_p, \quad (11)$$

where w_p is a feature-specific weight. We optimized the features weights on the DUC 2006 data set. We initialized all feature weights to a default value of 1, and then optimized one feature weight at a time while keeping the others fixed. The most dominant features in our experiments are the sentence-narrative similarity $r(S, N)$ and the sentence-document similarity $r(S, D)$, which confirms previous research. Sentences are ranked by this score, and the highest-scoring sentences are selected for the summary.

We model redundancy similar to the maximum marginal relevance framework (MMR) (Carbonell and Goldstein, 1998). MMR is a greedy approach that iteratively selects the best-scoring sentence for the summary, and then updates sentence scores by computing a penalty based on the similarity of each sentence with the current summary:

$$score_{mmr}(s) = \lambda(score(s)) - (1 - \lambda)r(S, SUM), \quad (12)$$

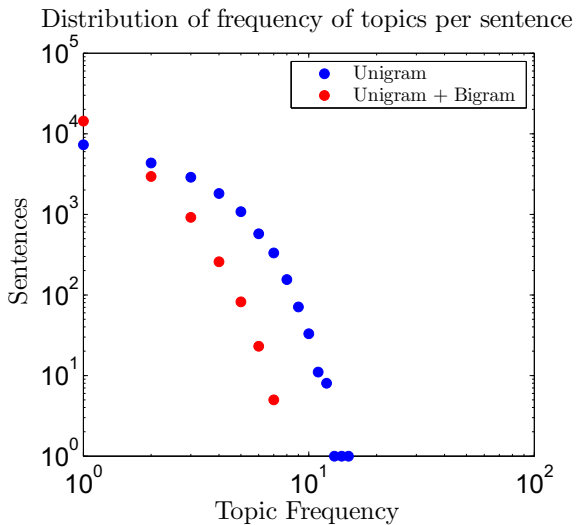
where the score of sentence s is scaled to $[0, 1]$ and $r(S, SUM)$ is the cosine similarity of the sentence and the summary centroid vector, which is based on the averaged distribution over topics of sentences selected for the summary. We optimized λ on DUC 2006 data, with the best value $\lambda = 0.4$ used in our experimental evaluation.

4.5. Topic distribution over sentences

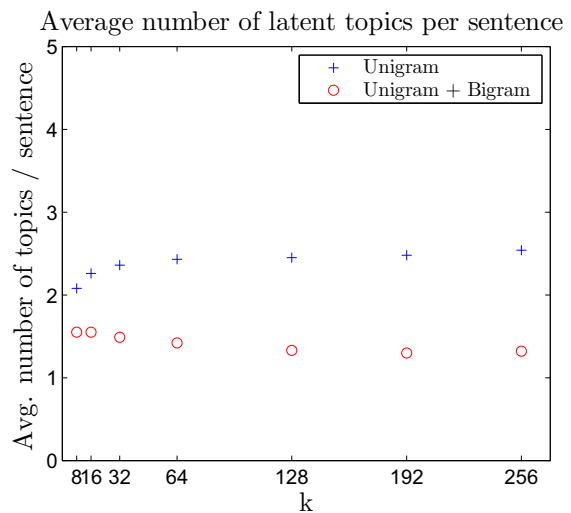
It is well known that documents cover multiple subtopics related to the main theme of the document (Barzilay and Lee, 2004). Standard topic models such as LDA therefore represent a document as a distribution over a set of latent topics. In our approach, we extend this notion and treat each sentence as a document, thus assuming that a sentence covers one or more topics of the document set. For example, a sentence of a news article related to a meeting of government leaders may provide information on the people who have met as well as on the location of the meeting. Our intuition is that the number of topics that a sentence covers should be rather low, but larger than one.

Figure 1(a) shows the distribution of the number of topics per sentence for a PLSA model based on terms only and for the PLSA model combining unigrams and bigrams. We only consider topics with a probability greater than some small value ϵ ($\epsilon > 0.01$). We see that the distributions follow a power law: There are very many sentences which are assigned a single dominant topic, and very few sentences which are assigned many topics. We note that the combined model assigns less topics to a sentence than the term-based model.

From Figure 1(b) we see that the average number of topics assigned to a sentence is relatively robust to varying the value of k (the free parameter specifying the number of latent topics for the PLSA algorithm). Even for $k \leq 16$,



(a) DUC 2007: Topic distribution



(b) DUC 2007: Average number of topics

Figure 1: (a) Distribution of number of topics per sentence ($p(z|d) > 0.01$) for a $k = 128$ factor decomposition of the DUC 2007 document sets, using terms only or the combined model; and (b) Average number of topics per sentence ($p(z|d) > 0.01$) for different values of k , using terms only or the combined model

where k is actually smaller than the number of input documents, on average more than one topic is assigned to a sentence. This confirms our intuition that sentences may cover multiple subtopics. Again we see that the combined model on average assigns less topics to a sentence, which suggests that the descriptive quality of the topics better fits the available data.

5. Experiments

For the evaluation of our system, we use the data set from the multi-document summarization task in DUC 2007. For all our evaluations, we use ROUGE metrics⁴. ROUGE metrics are recall-oriented and based on n-gram overlap. ROUGE-1 has been shown to correlate well with human judgements (Lin and Hovy, 2003a). In addition, we also report the performance on ROUGE-2 (bigram overlap) and ROUGE-SU4 (skip bigram) metrics.

5.1. Results

We present the results of our system in Table 2. We compare our results to the best peer (*peer 15*) and to a *Lead* sentence baseline system. The *Lead* system uses the first n sentences from the most recent news article in the document cluster to create a summary. In the table, system *PLSA* uses a standard term co-occurrence based model, and system *PLSA-F* combines term and bigram co-occurrences, based on the best value for parameter $\alpha = 0.6$. The *PLSA-F* system outperforms the standard *PLSA* model on ROUGE-1, ROUGE-2 and ROUGE-SU4 scores, although the improvements are not significant. More interestingly, the *PLSA-F* achieves its best score using only $k = 32$ latent classes, compared to $k = 256$ for the *PLSA* system. This suggests

Table 2: DUC-07: ROUGE recall scores for best number of latent topics k . The *PLSA* system uses term co-occurrences only, the *PLSA-F* combines term and bigram co-occurrence information, with $\alpha = 0.6$. The *PLSA-F* variant outperforms the best participating system (*peer 15*) on ROUGE-1.

System	k	Rouge-1	Rouge-2	Rouge-SU4
peer 15	-	0.44508	0.12448	0.17711
PLSA-F	32	0.45400	0.11951	0.17573
PLSA	256	0.44885	0.11774	0.17552
Lead	-	0.31250	0.06039	0.10507

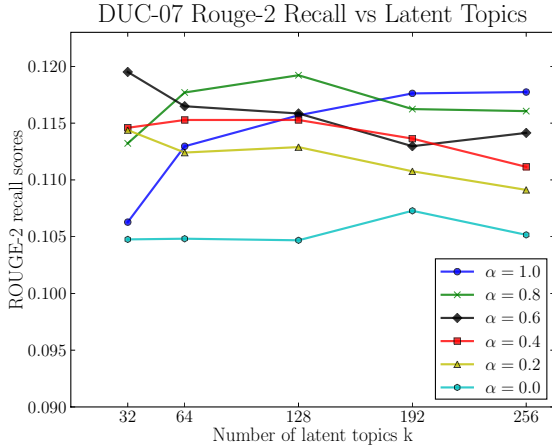
that the information supplied by the bigram co-occurrence observations indeed reinforces the term co-occurrence observations, such that the model can better represent the different latent topics contained in the document cluster.

Our combined approach outperforms *peer 15* on ROUGE-1 recall, and is not significantly worse on ROUGE-SU4 recall. For ROUGE-2, our system’s performance is only slightly lower than the 95%-confidence interval of the top system’s performance (0.11961–0.12925). The results of our system are also comparable to the topic modeling approach of Haghghi and Vanderwende (Haghghi and Vanderwende, 2009), who report a ROUGE-2 score of 0.118 for a model based on bigram distributions, but are significantly better than the 0.097 they report for an unigram-based model.

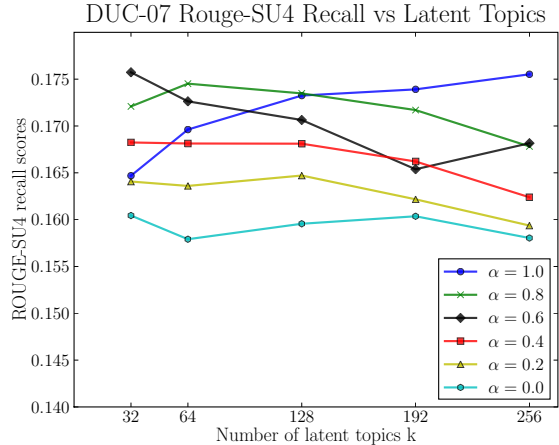
5.2. System variations

To verify the experimental observation that the combined model allows for a better representation of the latent topics, we conducted a series of experiments varying the number of latent classes and the weight of the parameter α . The results of these experiments are shown in Figure 2. We have

⁴ROUGE version 1.5.5, with arguments -n 4 -w 1.2 -m -2 4 -u -c 95 -r 1000 -f A -p 0.5 -t 0



(a) DUC 2007 Rouge-2



(b) DUC 2007 Rouge-SU4

Figure 2: Summarization performance on DUC 2007 data in terms of ROUGE-2 (a) and ROUGE-SU4 (b) recall for different values of latent topics k and parameter α .

omitted results for $k < 32$, where none of the models can cope with the complexity of the data. We also do not show results for $k > 256$, since the performance of all models either stabilizes with respect to their performance at $k = 256$, or the models start to overfit, resulting in lower ROUGE scores. We observe that the models combining term and bigram co-occurrence information outperform the models based only on term co-occurrence ($\alpha = 1.0$) respectively bigram co-occurrence ($\alpha = 0.0$) for small numbers of latent classes k . As k increases, the performance of the combined models decreases, or exhibits only small variations (e.g. $\alpha = 0.6$ for $k = 256$). This suggests that the quality of the learned latent topics is starting to decrease, as the algorithm creates topics with idiosyncratic word combinations (Steyvers and Griffiths, 2006). The performance of the term-based model, however, increases until $k = 256$, reaching a maximum ROUGE-2 recall of 0.11776, before also overfitting (not shown here).

Our observations therefore indicate that the information obtained from the combined model allows for a more descriptive representation of the latent topics contained in the document collection. The most interesting observation shown in Figure 2 is that adding bigram-sentence co-occurrence observations to a standard PLSA model can substantially improve ROUGE-2 scores and significantly reduce the number of latent classes required for a good model. The effect is less pronounced for ROUGE-SU4 scores, but still recognizable. All combined models outperform the term and bigram baseline models on ROUGE-2 for $k = 32$ latent classes.

We further note that the term-based model ($\alpha = 1.0$) consistently outperforms the bigram-based model ($\alpha = 0.0$), indicating that bigram co-occurrence information alone captures less of the topical relations that exist in the document collection.

We also find that the effect of varying the parameter α follows an expectable pattern: For $\alpha = 0.8$, the term observations dominate the combined topic model, and the ROUGE-2 score curve follows that of the model with $\alpha = 1.0$

(for $k \leq 128$). With decreasing α ($\alpha \geq 0.6$) a model reaches its best performance for a smaller value of k (Since we omit scores for $k < 32$, this ‘peak’ is not as clearly visible for $\alpha = 0.6$).

The experimentally optimal value of $\alpha = 0.6$ weights term and bigram co-occurrences almost equally, with ROUGE-2 scores similar for $\alpha = 0.4$. For lower values of α , i.e. models where bigram observations contribute more prominently during parameter estimation, the summarization performance of the model decreases substantially. ROUGE-SU4 scores are consistently lower than for the other models.

6. Conclusion

We introduced a novel approach for query-focused multi-document summarization that combines term and bigram co-occurrence observations into a single probabilistic latent topic model. The integration of a bigram language model into a standard topic model results in a system that outperforms models which are based on term respectively bigram co-occurrence observations only. Furthermore, it requires fewer latent classes for optimal summarization performance.

We observe that the distribution of topic frequencies across sentences follows a power law. On average, sentences are assigned more than two latent topics for a standard topic model, but only between one and two topics for our combined model. This suggests that the combined model results in a better representation of the underlying subtopics of a document set. We also find that the average number of topics assigned to a sentence is relatively robust with respect to variations in the number of latent classes.

Our results are among the best reported on the DUC-2007 multi-document summarization tasks for ROUGE-1, ROUGE-2 and ROUGE-SU4 scores. We have achieved these excellent results with a system that utilizes a considerably simpler model than previous topic modeling approaches to multi-document summarization.

In future work, we plan to implement our approach using

LDA instead of PLSA to address shortcomings of PLSA such as overfitting and the lack of generative modeling at the document level.

7. References

- Regina Barzilay and Lillian Lee. 2004. Catching the drift: Probabilistic content models, with applications to generation and summarization. In *Proc. of HLT-NAACL*.
- David M. Blei, Andrew Y. Ng, Michael I. Jordan, and John Lafferty. 2003. Latent dirichlet allocation. *J. of Machine Learning Research*, 3:2003.
- Thorsten Brants, Francine Chen, and Ioannis Tsochantridis. 2002. Topic-based document segmentation with probabilistic latent semantic analysis. In *Proc. of CIKM*, pages 211–218.
- Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proc. of SIGIR '98*, pages 335–336.
- David Cohn and Thomas Hofmann. 2000. The missing link - a probabilistic model of document content and hypertext connectivity. In *Neural Information Processing Systems 13*, pages 430–436.
- J. M. Conroy, J. D. Schlesinger, and D.P. Leary. 2007. CLASSY 2007 at DUC 2007. In *Proc. of DUC 2007*.
- Hal Daumé, III and Daniel Marcu. 2006. Bayesian query-focused summarization. In *Proc. Int. Conf. on Computational Linguistics (ACL)*, pages 305–312.
- Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *J. of the American Society for Information Science*, 41:391–407.
- Jade Goldstein, Vibhu Mittal, Jaime Carbonell, and Mark Kantrowitz. 2000. Multi-document summarization by sentence extraction. In *NAACL-ANLP 2000 Workshop on Automatic summarization*, pages 40–48.
- Aria Haghighi and Lucy Vanderwende. 2009. Exploring content models for multi-document summarization. In *Proc. of NAACL-HLT*.
- Leonhard Hennig. 2009. Topic-based multi-document summarization with probabilistic latent semantic analysis. In *Recent Advances in Natural Language Processing, RANLP 2009*.
- Thomas Hofmann. 1999. Probabilistic latent semantic indexing. In *Proc. of SIGIR '99*, pages 50–57.
- Karen Spärck Jones. 2007. Automatic summarising: The state of the art. *Inf. Process. Manage.*, 43(6):1449–1481.
- Chin-Yew Lin and Eduard Hovy. 2000. The automated acquisition of topic signatures for text summarization. In *Proc. Int. Conf. on Computational Linguistics (ACL)*, pages 495–501.
- Chin-Yew Lin and Eduard Hovy. 2003a. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proc. of NAACL-HLT 2003*, pages 71–78.
- Chin-Yew Lin and Eduard Hovy. 2003b. The potential and limitations of automatic sentence extraction for summarization. In *Proc. of the HLT-NAACL 2003 Workshop on Text Summarization*, pages 73–80.
- Christopher D. Manning and Hinrich Schütze. 2001. *Foundations of Statistical Natural Language Processing*. MIT Press.
- Ani Nenkova, Lucy Vanderwende, and Kathleen McKeown. 2006. A compositional context sensitive multi-document summarizer: exploring the factors that influence summarization. In *Proc. of SIGIR '06*, pages 573–580.
- Jay M. Ponte and W. Bruce Croft. 1998. A language modeling approach to information retrieval. In *Proc. of SIGIR '98*, pages 275–281.
- M. F. Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.
- Mark Steyvers and Tom Griffiths. 2006. Probabilistic topic models. In S. Dennis T. Landauer, Mcnamara and W. Kintsch, editors, *Latent Semantic Analysis: A Road to Meaning*. Laurence Erlbaum.
- Hanna M. Wallach. 2006. Topic modeling: beyond bag-of-words. pages 977–984, Pittsburgh, Pennsylvania. ACM.
- Xuerui Wang, Andrew McCallum, and Xing Wei. 2007. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *Proc. of ICDM '07*, pages 697–702.
- W. Yih, J. Goodman, L. Vanderwende, and H. Suzuki. 2007. Multi-document summarization by maximizing informative content-words. In *Proc. of IJCAI 2007*, pages 1776–1782.