Learning Summary Content Units with Topic Modeling

Leonhard Hennig **Ernesto William De Luca** **Sahin Albayrak**
Distributed Artificial Intelligence Laboratory (DAI-Lab)
Technische Universität Berlin
{leonhard.hennig,ernesto.deluca,sahin.albayrak}@dai-labor.de

Abstract

In the field of multi-document summarization, the Pyramid method has become an important approach for evaluating machine-generated summaries. The method is based on the manual annotation of text spans with the same meaning in a set of human model summaries. In this paper, we present an unsupervised, probabilistic topic modeling approach for automatically identifying such semantically similar text spans. Our approach reveals some of the structure of model summaries and identifies topics that are good approximations of the Summary Content Units (SCU) used in the Pyramid method. Our results show that the topic model identifies topic-sentence associations that correspond to the contributors of SCUs, suggesting that the topic modeling approach can generate a viable set of candidate SCUs for facilitating the creation of Pyramids.

1 Introduction

In the field of multi-document summarization (MDS), the Pyramid method has become an important approach for evaluating machine-generated summaries (Nenkova and Passonneau, 2004; Passonneau et al., 2005; Nenkova et al., 2007). The method rewards automatic summaries for conveying content that has the same meaning as content represented in a set of human model summaries. This approach allows for variation in the way the content is expressed, which contrasts the Pyramid method with other evaluation methods such as ROUGE that measure word n-gram overlap (Lin and Hovy, 2003).

The Pyramid method groups content with the same meaning into Summary Content Units (SCU). Shared content needs to be identified manually by human inspection of summaries, adding yet another level of human effort (on top of creating model summaries) to the task of summary evaluation. However, Nenkova and Passonneau (2004) as well as Harnly et al. (2005) observe that semantically similar text spans written by different human summarizers are often expressed with a similar choice of words, albeit with differences e.g. in word variants, word order and paraphrasing (Section 2).

In this paper, we present an approach for automatically identifying semantically similar text spans in human model summaries on the basis of such re-occurring word patterns. We utilize a method known as probabilistic topic modeling (Steyvers and Griffiths, 2007). Topic models are claimed to derive semantic information from text in an unsupervised fashion, using only the observed word distributions (Section 3).

- We train a probabilistic topic model based on Latent Dirichlet Allocation (Blei et al., 2003) on the term-sentence matrix of human model summaries used in the Document Understanding Conference (DUC) 2007 Pyramid evaluation[1]. We analyze the resulting model to evaluate whether a topic model captures useful structures of these summaries (Section 4.1).

[1] http://duc.nist.gov

Coling 2010: Poster Volume, pages 391–399,
Beijing, August 2010

- Given the model, we compare the automatically identified topics with SCUs on the basis of their word distributions (Sections 4.2 and 4.3). We discover a clear correspondence between topics and SCUs, which suggests that many automatically identified topics are good approximations of manually annotated SCUs.

- We analyze the distribution of topics over summary sentences in Section 4.4, and compare the topic-sentence associations computed by our model with the SCU-sentence associations given by the Pyramid annotation. Our results suggest that the topic model finds many SCU-like topics, and associates a given topic with the same summary sentences in which a human annotator identifies the corresponding SCU.

Automatically identifying topics that approximate SCUs has clear practical applications: The topics can be used as a candidate set of SCUs for human annotators to speed up the process of SCU creation. Topics can also be identified in machine-generated summaries using standard statistical inference techniques (Asuncion et al., 2009).

2 Summary Content Units

In this section, we briefly introduce the Pyramid method and the properties of Summary Content Units that we intend to exploit in our approach.

A Pyramid is a model predicting the distribution of information content in summaries, as reflected in the summaries humans write (Passonneau et al., 2005; Nenkova et al., 2007). Similar information content is identified by inspection of similar sentences, and parts of these, in different human model summaries. Typically, the text spans which express the same semantic content are not longer than a clause. An SCU consists of a collection of text spans with the same meaning (contributors) and a defining label specified by the annotator.

Each SCU is weighted by the number of human model summaries it occurs in (i.e. the number of contributors). The Pyramid metric assumes that an SCU with a high number of contributors is more informative than an SCU with few contributors. An optimal summary, in terms of content selection, is obtained by maximizing the sum of SCU weights, given a maximum number of SCUs that can be included for a predefined summary length (Nenkova and Passonneau, 2004).

Two example SCUs are given in Table 1. SCU 18 has a weight of 3, since three model summaries contribute to it, SCU 21 has a weight of 2. SCU 18 aggregates contributors which share some key phrases such as "Air National Guard" and "search", but otherwise exhibit a quite heterogeneous word usage. Contributor 3 gives details on the aircraft type, and specifies a time when the first sea vessel was launched to search for the missing plane. Only contributor 1 gives information about the location of the search. In SCU 21, the first contributor contains additional information about communication with the Kennedy family, which is not expressed in the SCU label and therefore not part of the meaning of the SCU. Both contributors contain key terms such as "officials", "search" and "recovery", but vary in word order and verb usage. Passonneau et al. (2005) discuss this observation, and argue that SCUs emerge from the judgment of annotators, and are thus independent of what words are used, or how many.

However, an analysis of typical SCUs shows that contributors written by different human summarizers are often expressed with a similar choice of words or even phrases. Contributors vary in using different forms of the same words (inflectional or derivational variants), different word order, syntactic structure, and paraphrases (Harnly et al., 2005; Nenkova et al., 2007).

3 Probabilistic Topic Models

Our approach for discovering semantically similar text spans makes use of a statistical method known as topic modeling. Probabilistic topic models can derive semantic information from text automatically, on the basis of the observed word patterns (Hofmann, 1999; Blei et al., 2003; Steyvers and Griffiths, 2007). The main assumption of these models is that a latent set of variables – the topics – can be utilized to explain the observed patterns in the data. Documents are represented as mixtures of topics, and each topic is a distribution

SCU 18	The US Coast Guard with help from the Air National Guard then began a massive search-and-rescue mission, searching waters along the presumed flight path
Contributor 1:	The US Coast Guard with help from the Air National Guard then began a massive search-and-rescue mission, searching waters along the presumed flight path
Contributor 2:	A multi-agency search and rescue mission began at 3:28 a.m., with the Coast Guard and Air National Guard participating
Contributor 3:	The first search vessel was launched at about 4:30am. An Air National Guard C-130 and many Civil Air Patrol aircraft joined the search
SCU 21	Federal officials shifted the mission to search and recovery
Contributor 1:	Federal officials shifted the mission to search and recovery and communicated the Kennedy and Bessette families
Contributor 2:	federal officials ended the search for survivors and began a search-and-recovery mission

Table 1: Example SCUs from topic D0742 of DUC 2007.

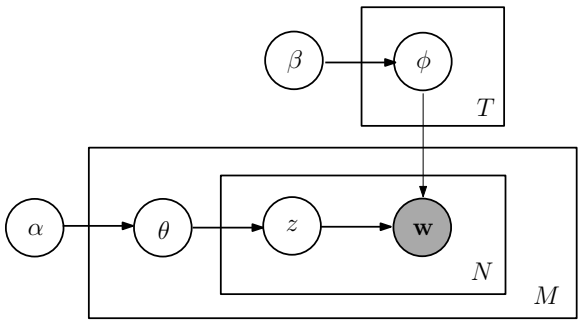


Figure 1: Graphical model representation of LDA for N words, T topics and a corpus of M documents.

over words. For example, a news article describing a meeting of the International Monetary Fund may in equal parts discuss economic and political issues. Topic models discover in a completely unsupervised fashion meaningful topics as well as intra- and inter-document statistical structure using no information except the distribution of the words themselves (Griffiths and Steyvers, 2004).

For our analysis, we use the Latent Dirichlet Allocation (LDA) model introduced by Blei et al. (2003). In this model, each document is generated by first choosing a distribution over topics $\theta^{(d)}$, parametrized by a conjugate Dirichlet prior α. Subsequently, each word of this document is generated by drawing a topic z_k from $\theta^{(d)}$, and then drawing a word w_i from topic z_k's distribution over words $\phi^{(k)}$. We follow Griffiths et al. (2004) and place a conjugate Dirichlet prior β over $\phi^{(k)}$ as well. Figure 1 shows the graphical model representation of LDA.

For T topics, the matrix $\mathbf{\Phi}$ specifies the probability $p(w|z)$ of words given topics, and $\mathbf{\Theta}$ specifies the probability $p(z|d)$ of topics given documents. $p(w|z)$ indicates which words are important in a topic, and $p(z|d)$ tells us which topics are dominant in a document. We employ Gibbs sampling (Griffiths and Steyvers, 2004) to estimate the posterior distribution over z (the assignment of word tokens to topics), given the observed words $\mathbf{w}$ of the document set. From this estimate we can approximate the distributions for the matrices $\mathbf{\Phi}$ and $\mathbf{\Theta}$.

4 Experiments

Can a topic model reveal some of the structure of human model summaries and learn topics that are approximations of manually annotated SCUs? To answer these questions, we train a topic model on the human model summaries of each of the 23 document clusters of the DUC 2007 dataset that were used in Pyramid evaluation[2]. There are 4 human model summaries available for each document cluster. On average, the summary sets contain 52.4 sentences, with a vocabulary of 260.5 terms, which occur a total of 549.7 times. The Pyramids of these summary sets consist of 68.8 SCUs on average. The number of SCUs per SCU

[2]http://www-nlpir.nist.gov/projects/duc/data.html

weight follows a Zipfian distribution, i.e. there are typically very few SCUs of weight 4, and very many SCUs of weight 1 (see also Passonneau et al. (2005)).

4.1 Topic model training

Since we are interested in modeling topics for sentences, we treat each sentence as a document[3]. We construct a matrix $\mathbf{A}$ of term-sentence co-occurrence observations for each set of human model summaries S. Each entry $\mathbf{A}_{ij}$ corresponds to the frequency of word i in sentence j, and j ranges over the union of the sentences contained in S. We preprocess terms using stemming and removing a standard list of stop words with the NLTK toolkit[4].

We run the Gibbs sampling algorithm on $\mathbf{A}$, setting the parameter T, the number of latent topics to learn, equal to the number of SCUs contained in the Pyramid of S. We use this particular value for T since we want to learn a topic model with a structure that reflects the SCUs and the distribution of SCUs of the corresponding Pyramid. For an unannotated set of summaries, determining an optimal value for T is a Bayesian model selection problem (Kass and Raftery, 1995).

The topic distribution for each sentence should be peaked toward a single or only very few topics. To ensure that the topic-specific word distributions $p(w|z)$ as well as the sentence-specific topic distributions $p(z|d)$ behave as intended, we set the Dirichlet priors $\alpha = 0.01$ and $\beta = 0.01$. This enforces a bias toward sparsity, resulting in distributions that are more peaked (Steyvers and Griffiths, 2007). A low value of β also favors more fine-grained topics (Griffiths and Steyvers, 2004). We run the Gibbs sampler for 2000 iterations, and collect a single sample from the resulting posterior distribution over topic assignments for words. From this sample we compute the conditional distributions $p(w|z)$ and $p(z|d)$.

During our experiments, we observed that the Gibbs Sampler did not always use all the topics available. Instead, some topics had a uniform distribution over words, i.e. no words were assigned to these topics during the sampling process. We assume this is due to the relatively low prior $\alpha = 0.01$ we use in our experiments. We explore the consequences of varying the LDA priors and T in Section 4.4.

This observation indicates that the topic model cannot learn as many distinct topics from a given set of summaries as there are SCUs in the Pyramid of these summaries. On average, 24.4% ($\sigma = 17.4$) of the sampled topics had a uniform word distribution, but the fraction of such topics varied. For some summary sets, it was very low (D0701, D0706 with 0%), whereas for others it was very high (D0704, D0728 with 52%). Both of the latter summary sets contain many SCUs with very similar labels and often only a single contributor, e.g. about 'Amnesty International':

- AI criticism frequently involves genocide
- AI criticism frequently involves intimidation
- AI criticism frequently involves police violence

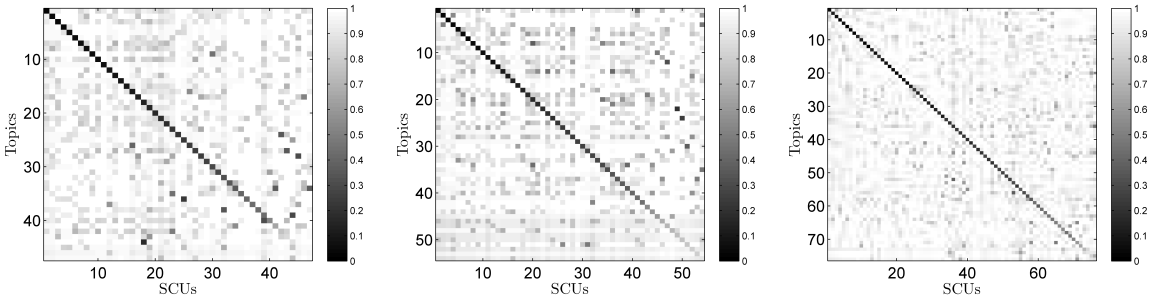
These SCUs are derived from summary sentences that contain enumerations: "AI criticism frequently involves political prisoners, torture, intimidation, police violence, the death penalty, no alternative service for conscientious objectors, and interference with the judiciary." A topic model is based on word-document co-occurrence data, and cannot distinguish between the different grammatical objects in this case. Instead, it treats these phrases as semantically similar since they occur in the same sentence.

4.2 SCU word distributions and SCU-sentence associations

In order to evaluate the quality of the LDA topics, we compare their word distributions to the word distributions of SCUs. This allows us to analyze if the LDA topics capture similar word patterns as SCUs. We approximate the distribution over words $p(w|s_l)$ for each SCU s_l as the relative frequency of word w_i in the bag-of-words constructed from the texts of s_l's label and contributors. We denote the resulting matrix of for a set of SCUs as $\widehat{\mathbf{\Phi}}$.

[3] We will use the words document and sentence interchangeably from here on.

[4] http://www.nltk.org



(a) DUC Topic D0706 (b) DUC Topic D0742 (c) DUC Topic D0743

Figure 2: Pairwise Jensen-Shannon divergence of word distributions of LDA topics and Summary Content Units (SCUs), for 3 DUC 2007 Pyramids. Topic-SCU matches are ordered by increasing divergence along the diagonal, using a simple greedy algorithm. The examples suggest that many of the automatically identified LDA topics correspond to manually annotated SCUs.

Topic 17	SCU 31	Topic 5	SCU 32	Topic 9	SCU 25	Topic 8	SCU 36
pilot	pilot	analysi	analysi	bodi	bodi	kennedi	kennedi
kennedi	condit	control	control	diver	diver	edward	edward
condit	conduc	corkscrew	corkscrew	entomb	entomb	recoveri	recoveri
conduc	dark	descent	descent	floor	floor	son	son
dark	disorient	fall	fall	found	found	wit	wit

Table 2: Top terms of best matching LDA topics and SCUs for summary set D0742

In addition, we can compare the topic-sentence associations computed by the model to the SCU-sentence associations given by the Pyramid annotation. If the probability of a given topic is high in those sentences which contribute to a particular SCU, this would suggest that the topic model can automatically learn topics which not only have a word distribution similar to a specific SCU, but also a similar distribution over contributing sentences.

SCU contributors are typically annotated as a set of contiguous sequences of words within a single sentence. In the DUC 2007 data, there are only a few cases where a contributor spans more than one sentence. The DUCView annotation tool[5] stores the start and end character position of the phrases marked as contributors of an SCU. We can utilize this information to define which sentences an SCU is associated with. We store the associations in a matrix $\widehat{\mathbf{\Theta}}$, where $\widehat{\mathbf{\Theta}}_{ij} = 1$ if SCU i is associated with sentence j. Sentences may contain multiple SCUs, and SCUs are associated with as many sentences as their number of contributors.

[5]http://www1.cs.columbia.edu/~becky/DUC2006/2006-pyramid-guidelines.html

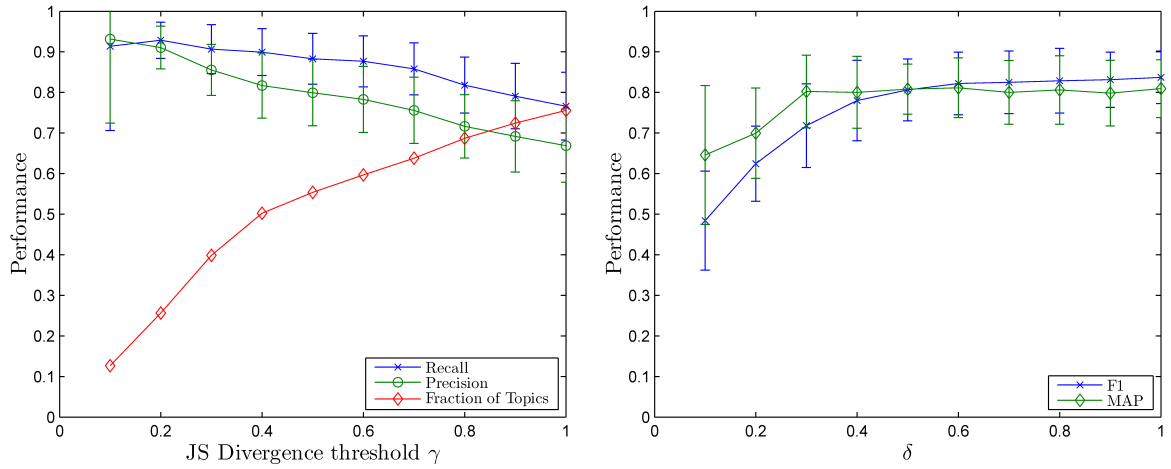
4.3 Matching SCUs and LDA topics

Before we can compare the topic-sentence associations computed by the LDA topic model with the SCU-sentence associations, we need to match SCUs to LDA topics. We consider a topic to be similar to an SCU if their word distributions are similar. We discard all LDA topics with a uniform word distribution (see Section 4.1) before the matching step.

We then compute the pair-wise Jensen-Shannon (JS) divergence between columns j of $\mathbf{\Phi}$ and k of $\widehat{\mathbf{\Phi}}$:

$$JS(\mathbf{\Phi}_j, \widehat{\mathbf{\Phi}}_k) = \left[\frac{1}{2} D_{KL}(\mathbf{\Phi}_j || M) + \frac{1}{2} D_{KL}(\widehat{\mathbf{\Phi}}_k || M)\right], \quad (1)$$

where $M = 1/2(\mathbf{\Phi}_j + \widehat{\mathbf{\Phi}}_k)$. SCUs from $\widehat{\mathbf{\Phi}}$ are matched to topics of $\mathbf{\Phi}$ on the basis of this dissimilarity using a simple greedy approach, i.e. by iteratively selecting the current most similar SCU-



(a) Precision, recall and fraction of Topic-SCU matches for different settings of γ

(b) F1 and MAP for different values of T as a fraction δ of the number of SCUs

Figure 3: (a) Precision, Recall and the fraction of LDA topics matched to SCUs for different settings of parameter γ, averaged over all summary sets with Pyramid annotations from DUC 2007. Error bars show the standard deviation. Only topic-SCU matches with $JS(\mathbf{\Phi}_j, \widehat{\mathbf{\Phi}}_k) \leq \gamma$ are considered when computing precision and recall. Both are very high, suggesting that the model identifies topics that are very similar to SCUs. (b) F_1 measure and Mean Average Precision (MAP) for different settings of the number of latent topics T as a fraction of the number of SCUs in the corresponding Pyramid ($\gamma = 0.5$).

topic pair. We reorder the rows of $\mathbf{\Theta}$ according to the computed matching.

Figure 2 shows some example SCU-topic matches for three different DUC 2007 summary sets. Each cell displays the JS divergence of the word distributions of an LDA topic (rows) compared to an SCU (columns). On the diagonal, the best matches of LDA topics and SCUs are ordered by increasing JS divergence. Multiple points with low JS divergence in a single column indicate that more than one LDA topic was very similar to this SCU. Overall, the graphs show a clear correspondence of LDA topics to the SCUs. The plots suggest that a large percentage of topics have similar distributions over words as the corresponding SCUs. Table 2 shows the most likely terms for some example topic-SCU matches. For each of these matches, the top terms are almost identical.

4.4 Evaluation

To compare the topic distributions $\mathbf{\Theta}$ with the SCU-sentence assignments $\widehat{\mathbf{\Theta}}$, we binarize $\mathbf{\Theta}$ to give $\mathbf{\Theta}'$ by setting all entries $\mathbf{\Theta}'_{ij} = 1$ if $\mathbf{\Theta}_{ij} > \epsilon$, and 0 otherwise. We set $\epsilon = 0.1$ in our experiments[6]. $\mathbf{\Theta}'_{ij}$ is therefore equal to 1 if a topic i has a high probability sentence j. We can now evaluate if a given topic occurs in the same sentences as the corresponding SCU (recall), and if it occurs in no other sentences (precision).

We compute precision and recall for each topic-SCU match with $JS(\mathbf{\Phi}_j, \widehat{\mathbf{\Phi}}_k) \leq \gamma$. Averaged over matches, these measures give us an indication of how well the LDA model approximates the set of SCUs. The parameter γ allows us to tune the performance of the model with respect to the quality and number of topic-SCU matches. Setting γ to a low value will consider only topic-SCU matches with a low JS divergence, which generally results in higher precision and recall. Increasing γ will include more topic-SCU matches, namely those with a larger JS divergence, which will therefore introduce some noise.

Figure 3(a) shows the precision and recall

[6]Since the LDA algorithm learns very peaked distributions, the actual value of this threshold does not have a large impact on the resulting binary matrix and subsequent evaluation results. We evaluated a range of settings for ϵ in $[0.001 - 0.5]$, all with similar performance. This observation is confirmed by the threshold-less Mean Average Precision results in Figure 3(b).

curves for different values of the parameter γ, averaged over all summary sets. The plots show that both the precision and recall of the discovered topic-sentence associations are quite high, suggesting that the model automatically identifies topics which are very similar to manually annotated SCUs. With increasing γ, precision and recall scores decrease: The word distributions of the topic-SCU pairs are increasingly dissimilar, and hence the sentences associated with a topic do not necessarily overlap anymore with the sentences of the paired SCU. The figure also shows the fraction of topic matches that are considered in the evaluation of precision and recall. There is a clear trade off between performance and the number of matches retrieved. However, many of the topic-SCU matches ($\approx 50\%$) have a JS divergence ≤ 0.4, suggesting that the word distributions of many LDA topics are very similar to SCU word distributions.

Since we observed that the Gibbs sampling does not always utilize the full set of topics, we repeat our experiments to evaluate how the performance of the model changes when varying the LDA priors and T. Figure 3(b) shows F_1 and Mean Average Precision (MAP)[7] results of the topic model for different values of the parameter δ, where $T = \delta * |SCU|$. For example, a value of 0.6 means that for each summary set, T was set to 60% of the number of SCUs in the corresponding Pyramid. We see that the MAP score increases quickly, and reaches a plateau for $\delta \geq 0.3$. The F_1 score increases more slowly, and levels out for $\delta \geq 0.6$. The model's performance is relatively robust with respect to δ. This observation can be helpful when training models for new summary sets without an existing Pyramid, and which therefore consider T as a parameter to be optimized.

When varying the LDA priors, we observe that for $0.01 \leq \alpha \leq 0.05$, F_1 and MAP scores are consistently high, whereas for other settings, performance decreases significantly. Similarly, $\beta \geq 0.05$ results in lower F_1 and MAP scores. The fraction of uniform topics decreases with higher α, e.g. for $\alpha = 0.1$ it is close to zero. In contrast, higher settings of β increase the fraction of uniform topics.[8]

Finally, Figure 4 shows separate precision and recall curves for SCUs of different weights, and for different settings of parameter γ. Results are again averaged over all summary sets. In 4(a), we see that the recall of topic-sentence associations is very similar for all SCUs, with SCUs of higher weight exhibiting a slightly better recall. However, as Figure 4(b) shows, the average precision of SCUs with lower weight is much higher. Intuitively, this is expectable as SCUs of higher weight tend to have a larger vocabulary due to the higher number of contributors. This results in a larger word overlap with non-relevant sentences. The fraction of topic-SCU matches retrieved for SCUs of different weight is similar for all types of SCUs (not shown here).

5 Related Work

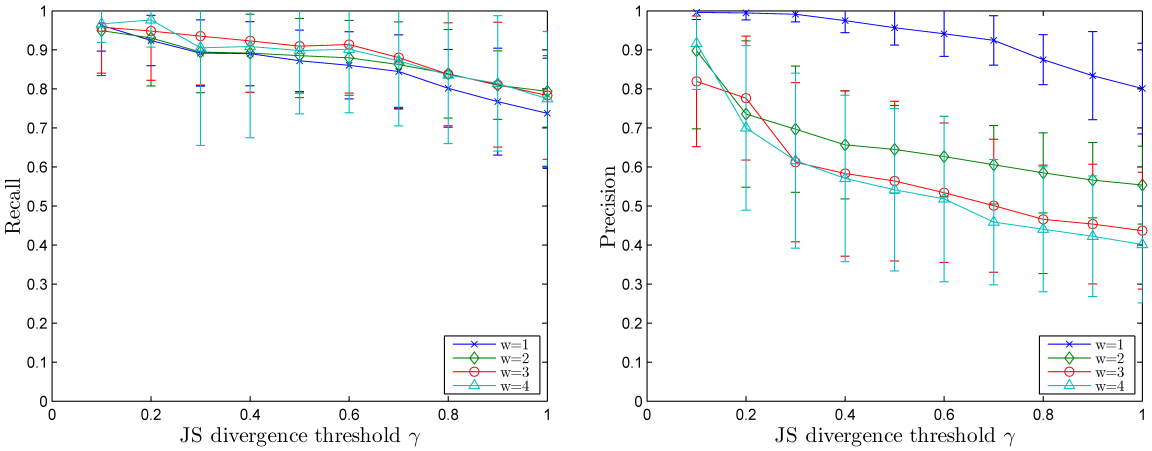
The Pyramid approach was introduced by Nenkova and Passonneau (2004) as a method for evaluating machine-generated summaries based on a set of human model summaries. The authors address a number of shortcomings of manual and automatic summary evaluation methods such as ROUGE (Lin and Hovy, 2003), and argue that the Pyramid method is reliable, diagnostic and predictive.

Passonneau and et al. (2005) give an account of the results of applying the Pyramid method during the DUC 2005 summarization evaluation, and discuss the annotation process. In subsequent work, Nenkova et al. (2007) describe in more detail the incorporation of human variation in the Pyramid method, the reliability of content annotation, and the correlation of Pyramid scores with other evaluation measures.

Harnly et al. (2005) present an approach for automatically scoring a machine summary given an existing Pyramid. Their method searches for an optimal set of candidate contributors created automatically from the machine summary and matches candidates to SCUs using a clustering approach.

[7] MAP is a rank-based measure, which avoids the need for introducing a threshold to binarize Θ (Baeza-Yates and Ribeiro-Neto, 1999). For each topic, we create a ranked list of sentences according to the transposed matrix $\mathbf{\Theta}^T$. This gives high ranks to sentences for which a particular topic has a high probability.

[8] Results are not shown due to space constraints.



(a) Recall of Topic-SCU matches for SCUs by weight

(b) Precision of Topic-SCU matches for SCUs by weight

Figure 4: (a) Recall of topic-SCU matches for SCUs of different weights, and settings of parameter γ, averaged over all summary sets. Recall is similar for SCUs of all weights. (b) Precision of the same topic-SCU matches. SCUs with a lower weight have a higher average precision. (Error bars show the standard deviation.)

The method assumes the existence of a Pyramid, whereas our approach aims to discover candidate SCUs from a set of human model summaries in an unsupervised fashion.

Recently, Louis and Nenkova (2009) presented an approach for fully automatic, model-free evaluation of machine-generated summaries. The method assumes that the distribution of words in the input and an informative summary should be similar. We think that it could be an interesting idea to combine the proposed method with our approach, in an attempt to exploit both the model-free evaluation and the shallow semantics of latent topics.

Probabilistic topic models have been successfully applied to a variety of tasks (Hofmann, 1999; Blei et al., 2003; Griffiths and Steyvers, 2004; Hall et al., 2008). In text summarization, most topic modeling approaches utilize a term-sentence co-occurrence matrix to discover topics in the set of input documents. Each sentence is typically assigned to a single topic, and a topic is a cluster of multiple sentences (Wang et al., 2009; Tang et al., 2009; Hennig, 2009).

6 Conclusions and future work

We presented a probabilistic topic modeling approach that reveals some of the structure of human model summaries. The topic model is trained on the term-sentence matrix of a set of human summaries, and discovers semantic topics in a completely unsupervised fashion. Many of the topics identified by our model for a given set of summaries show a similar distribution over words as the manually annotated Summary Content Units of the summaries' Pyramid.

We utilized the word distributions of SCUs and topics to match topics to similar SCUs, and showed that the topics identified by the model often occur in the same sentences as the contributors of the corresponding SCU. Precision and recall of these topic-sentence assignments are very high when compared to the SCU-sentence associations, indicating that many of the automatically acquired topics are good approximations of SCUs. Our results suggest that a topic model can be used to learn a candidate set of SCUs to facilitate the process of Pyramid creation.

We note that the topic model that we applied is one of the simplest latent variable models. A more complex model could integrate syntax to relax the bag-of-words assumption (Wallach, 2006), or combine the statistical model with more linguistically-grounded methods to handle linguistic features such as enumerations or negation.

References

Asuncion, Arthur, Max Welling, Padhraic Smyth, and Yee Whye Teh. 2009. On smoothing and inference for topic models. In *UAI '09: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 27–34.

Baeza-Yates, Ricardo A. and Berthier Ribeiro-Neto. 1999. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Griffiths, T. L. and M. Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl. 1):5228–5235.

Hall, David, Daniel Jurafsky, and Christopher D. Manning. 2008. Studying the history of ideas using topic models. In *EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 363–371.

Harnly, A., A. Nenkova, R. Passonneau, and O. Rambow. 2005. Automation of summary evaluation by the Pyramid method. In *Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP)*.

Hennig, Leonhard. 2009. Topic-based multi-document summarization with probabilistic latent semantic analysis. In *International Conference on Recent Advances in Natural Language Processing (RANLP)*.

Hofmann, Thomas. 1999. Probabilistic latent semantic indexing. In *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57.

Kass, R. E. and A. E. Raftery. 1995. Bayes factors. *Journal of the American Statistical Association*, 90:773–795.

Lin, Chin-Yew and Eduard Hovy. 2003. Automatic evaluation of summaries using N-gram co-occurrence statistics. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 71–78.

Louis, Annie and Ani Nenkova. 2009. Automatically evaluating content selection in summarization without human models. In *EMNLP '09: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 306–314.

Nenkova, Ani and Rebecca Passonneau. 2004. Evaluating Content Selection in Summarization: The Pyramid Method. In Susan Dumais, Daniel Marcu and Salim Roukos, editors, *HLT-NAACL 2004: Main Proceedings*, pages 145–152.

Nenkova, Ani, Rebecca Passonneau, and Kathleen McKeown. 2007. The Pyramid Method: Incorporating human content selection variation in summarization evaluation. *ACM Trans. Speech Lang. Process.*, 4(2):4.

Passonneau, R. J, A. Nenkova, K. McKeown, and S. Sigelman. 2005. Applying the Pyramid method in DUC 2005. In *Proceedings of the Document Understanding Conference (DUC'05)*.

Steyvers, Mark and Tom Griffiths. 2007. Probabilistic topic models. In Landauer, T., S. Dennis McNamara, and W. Kintsch, editors, *Latent Semantic Analysis: A Road to Meaning*. Laurence Erlbaum.

Tang, J., L. Yao, and D. Chen. 2009. Multi-topic based query-oriented summarization. In *Proceedings of the Siam International Conference on Data Mining*.

Wallach, Hanna M. 2006. Topic modeling: beyond bag-of-words. In *ICML '06: Proceedings of the 23rd international conference on Machine learning*, pages 977–984.

Wang, Dingding, Shenghuo Zhu, Tao Li, and Yihong Gong. 2009. Multi-document summarization using sentence-based topic models. In *ACL-IJCNLP '09: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 297–300.