DualSum: a Topic-Model based approach for update summarization

Jean-Yves Delort
Google Research
Brandschenkestrasse 110
8002 Zurich, Switzerland
jydelort@google.com

Enrique Alfonseca
Google Research
Brandschenkestrasse 110
8002 Zurich, Switzerland
ealfonseca@google.com

Abstract

Update summarization is a new challenge in multi-document summarization focusing on summarizing a set of recent documents relatively to another set of earlier documents. We present an unsupervised probabilistic approach to model novelty in a document collection and apply it to the generation of update summaries. The new model, called DUALSUM, results in the second or third position in terms of the ROUGE metrics when tuned for previous TAC competitions and tested on TAC-2011, being statistically indistinguishable from the winning system. A manual evaluation of the generated summaries shows state-of-the art results for DUALSUM with respect to focus, coherence and overall responsiveness.

1 Introduction

Update summarization is the problem of extracting and synthesizing novel information in a collection of documents with respect to a set of documents assumed to be known by the reader. This problem has received much attention in recent years, as can be observed in the number of participants to the special track on update summarization organized by DUC and TAC since 2007. The problem is usually formalized as follows: Given two collections $\mathcal{A}$ and $\mathcal{B}$, where the documents in $\mathcal{A}$ chronologically precede the documents in $\mathcal{B}$, generate a summary of $\mathcal{B}$ under the assumption that the user of the summary has already read the documents in $\mathcal{A}$.

Extractive techniques are the most common approaches in multi-document summarization. Summaries generated by such techniques consist of sentences extracted from the document collection. Extracts can have coherence and cohesion problems, but they generally offer a good tradeoff between linguistic quality and informativeness.

While numerous extractive summarization techniques have been proposed for multi-document summarization (Erkan and Radev, 2004; Radev et al., 2004; Shen and Li, 2010; Li et al., 2011), few techniques have been specifically designed for update summarization. Most existing approaches handle it as a redundancy removal problem, with the goal of producing a summary of collection $\mathcal{B}$ that is as dissimilar as possible from either collection $\mathcal{A}$ or from a summary of collection $\mathcal{A}$. A problem with this approach is that it can easily classify as redundant sentences in which novel information is mixed with existing information (from collection $\mathcal{A}$). Furthermore, while this approach can identify sentences that contain novel information, it cannot model explicitly what the novel information is.

Recently, Bayesian models have successfully been applied to multi-document summarization showing state-of-the-art results in summarization competitions (Haghighi and Vanderwende, 2009; Jin et al., 2010). These approaches offer clear and rigorous probabilistic interpretations that many other techniques lack. Furthermore, they have the advantage of operating in unsupervised settings, which can be used in real-world scenarios, across domains and languages. To our best knowledge, previous work has not used this approach for update summarization.

In this article, we propose a novel nonparametric Bayesian approach for update summarization. Our approach, which is a variation of Latent

Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 214–223,
Avignon, France, April 23 - 27 2012.

Dirichlet Allocation (LDA) (Blei et al., 2003), aims to learn to distinguish between common information and novel information. We have evaluated this approach on the ROUGE scores and demonstrate that it produces comparable results to the top system in TAC-2011. Furthermore, our approach improves over that system when evaluated manually in terms of linguistic quality and overall responsiveness.

2 Related work

2.1 Bayesian approaches in Summarization

Most Bayesian approaches to summarization are based on topic models. These generative models represent documents as mixtures of latent topics, where a topic is a probability distribution over words. In TOPICSUM (Haghighi and Vanderwende, 2009), each word is generated by a single topic which can be a corpus-wide background distribution over common words, a distribution of document-specific words or a distribution of the core content of a given cluster. BAYESSUM (Daumé and Marcu, 2006) and the Special Words and Background model (Chemudugunta et al., 2006) are very similar to TOPICSUM.

A commonality of all these models is the use of collection and document-specific distributions in order to distinguish between the general and specific topics in documents. In the context of summarization, this distinction helps to identify the important pieces of information in a collection.

Models that use more structure in the representation of documents have also been proposed for generating more coherent and less redundant summaries, such as HIERSUM (Haghighi and Vanderwende, 2009) and TTM (Celikyilmaz and Hakkani-Tur, 2011). For instance, HIERSUM models the intuitions that first sentences in documents should contain more general information, and that adjacent sentences are likely to share specic content vocabulary. However, HIERSUM, which builds upon TOPICSUM, does not show a statistically signicant improvement in ROUGE over TOPICSUM.

A number of techniques have been proposed to rank sentences of a collection given a word distribution (Carbonell and Goldstein, 1998; Goldstein et al., 1999). The Kullback-Leibler divergence (KL) is a widely used measure in summarization. Given a target distribution T that we want a summary S to approximate, KL is commonly used as the scoring function to select the subset of sentences S^* that minimizes the KL divergence with T:

$$S^* = \underset{S}{\mathrm{argmin}} KL(T, S) = \sum_{w \in \mathbf{V}} p_T(w) \log \frac{p_T(w)}{p_S(w)}$$

where w is a word from the vocabulary $\mathbf{V}$. This strategy is called KLSum. Usually, a smoothing factor τ is applied on the candidate distribution S in order to avoid the divergence to be undefined[1].

This objective function selects the most representative sentences of the collection, and at the same time it also diversifies the generated summary by penalizing redundancy. Since the problem of finding the subset of sentences from a collection that minimizes the KL divergence is NP-complete, a greedy algorithm is often used in practice[2]. Some variations of this objective function can be considered, such as penalizing sentences that contain document-specific topics (Mason and Charniak, 2011) or rewarding sentences appearing closer to the beginning of the document.

Wang et al. (2009) propose a Bayesian approach for summarization that does not use KL for reranking. In their model, Bayesian Sentence-based Topic Models, every sentence in a document is assumed to be associated to a unique latent topic. Once the model parameters have been calculated, a summary is generated by choosing the sentence with the highest probability for each topic.

While hierarchical topic modeling approaches have shown remarkable effectiveness in learning the latent topics of document collections, they are not designed to capture the novel information in a collection with respect to another one, which is the primary focus of update summarization.

2.2 Update Summarization

The goal of update summarization is to generate an *update summary* of a collection $\mathcal{B}$ of recent documents assuming that the users already read earlier documents from a collection $\mathcal{A}$. We refer

[1] In our experiments we set $\tau = 0.01$.

[2] In our experiments, we follow the same approach as in (Haghighi and Vanderwende, 2009) by greedily adding sentences to a summary so long as they decrease KL divergence.

to collection $\mathcal{A}$ as the *base collection* and to collection $\mathcal{B}$ as the *update collection*.

Update summarization is related to novelty detection which can be defined as the problem of determining whether a document contains new information given an existing collection (Soboroff and Harman, 2005). Thus, while the goal of novelty detection is to determine whether some information is new, the goal of update summarization is to extract and synthesize the novel information.

Update summarization is also related to contrastive summarization, i.e. the problem of jointly generating summaries for two entities in order to highlight their differences (Lerman and McDonald, 2009). The primary difference here is that update summarization aims to extract novel or updated information in the update collection with respect to the base collection.

The most common approach for update summarization is to apply a normal multi-document summarizer, with some added functionality to remove sentences that are redundant with respect to collection $\mathcal{A}$. This can be achieved using simple filtering rules (Fisher and Roark, 2008), Maximal Marginal Relevance (Boudin et al., 2008), or more complex graph-based algorithms (Shen and Li, 2010; Wenjie et al., 2008). The goal here is to boost sentences in $\mathcal{B}$ that bring out completely novel information. One problem with this approach is that it is likely to discard as redundant sentences in $\mathcal{B}$ containing novel information if it is mixed with known information from collection $\mathcal{A}$.

Another approach is to introduce specific features intended to capture the novelty in collection $\mathcal{B}$. For example, comparing collections $\mathcal{A}$ and $\mathcal{B}$, FastSum derives features for the collection $\mathcal{B}$ such as number of named entities in the sentence that already occurred in the old cluster or the number of new content words in the sentence not already mentioned in the old cluster that are subsequently used to train a Support Vector Machine classifier (Schilder et al., 2008). A limitation with this approach is there are no large training sets available and, the more features it has, the more it is affected by the sparsity of the training data.

3 DualSum

3.1 Model Formulation

The input for DUALSUM is a set of pairs of collections of documents $C = \{(\mathcal{A}_i, \mathcal{B}_i)\}_{i=1\ldots m}$, where $\mathcal{A}_i$ is a base document collection and $\mathcal{B}_i$ is an update document collection. We use c to refer to a collection pair $(\mathcal{A}_c, \mathcal{B}_c)$.

In DUALSUM, documents are modeled as a bag of words that are assumed to be sampled from a mixture of latent topics. Each word is associated with a latent variable that specifies which topic distribution is used to generate it. Words in a document are assumed to be conditionally independent given the hidden topic.

As in previous Bayesian works for summarization (Daumé and Marcu, 2006; Chemudugunta et al., 2006; Haghighi and Vanderwende, 2009), DUALSUM not only learns collection-specific distributions, but also a general background distribution over common words, ϕ^G and a document-specific distribution ϕ^{cd} for each document d in collection pair c, which is useful to separate the specific aspects from the general aspects of c. The main novelty is that DUALSUM introduces specific machinery for identifying novelty.

To capture the differences between the base and the update collection for each pair c, DUALSUM learns two topics for every collection pair. The joint topic, $\phi^{\mathcal{A}_c}$ captures the common information between the two collections in the pair, i.e. the main event that both collections are discussing. The update topic, $\phi^{\mathcal{B}_c}$ focuses on the specific aspects that are specific of the documents inside the update collection.

In the generative model,

- For a document d in a collection $\mathcal{A}_c$, words can be originated from one of three different topics: ϕ^G, ϕ^{cd} and $\phi^{\mathcal{A}_c}$, the last one of which captures the main topic described in the collection pair.

- For a document d in a collection $\mathcal{B}_c$, words can be originated from one of four different topics: ϕ^G, ϕ^{cd}, $\phi^{\mathcal{A}_c}$ and $\phi^{\mathcal{B}_c}$. The last one will capture the most important updates to the main topic.

To make this representation easier, we can also state that both collections are generated from the four topics, but we constrain the topic probability

```
1. Sample $\phi^G \sim Dir(\lambda_G)$
2. For each collection pair $c = (\mathcal{A}_c, \mathcal{B}_c)$:
   • Sample $\phi^{\mathcal{A}_c} \sim Dir(\lambda_\mathcal{A})$
   • Sample $\phi^{\mathcal{B}_c} \sim Dir(\lambda_\mathcal{B})$
   • For each document $d$ of type $u_{cd} \in \{\mathcal{A}, \mathcal{B}\}$:
       - Sample $\phi^{cd} \sim Dir(\lambda_D)$
       - If $(u_{cd} = \mathcal{A})$ sample $\psi^{cd} \sim Dir(\gamma^\mathcal{A})$
       - If $(u_{cd} = \mathcal{B})$ sample $\psi^{cd} \sim Dir(\gamma^\mathcal{B})$
       - For each word $w$ in document $d$:
         (a) Sample a topic $z \sim Mult(\psi^{cd})$, $z \in \{G, cd, \mathcal{A}_c, \mathcal{B}_c\}$
         (b) Sample a word $w \sim Mult(\phi^z)$
```

Figure 1: Generative model in DUALSUM.

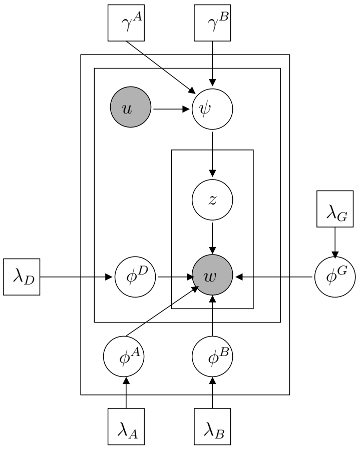

Figure 2: Graphical model representation of DUALSUM.

for $\phi^{\mathcal{B}_c}$ to be always zero when generating a base document.

We denote $u_{cd} \in \{A, B\}$ the type of a document d in pair c. This is an observed, Boolean variable stating whether the document d belongs to the base or the update collection inside the pair c.

The generation process of documents in DUALSUM is described in Figure 1, and the plate diagram corresponding to this generative story is shown in Figure 2. DUALSUM is an LDA-like model, where topic distributions are multinomial distributions over words and topics that are sampled from Dirichlet distributions. We use $\lambda = (\lambda_G, \lambda_D, \lambda_\mathcal{A}, \lambda_\mathcal{B})$ as symmetric priors for the Dirichlet distributions generating the word distributions. In our experiments, we set $\lambda_G = 0.1$ and $\lambda_D = \lambda_\mathcal{A} = \lambda_\mathcal{B} = 0.001$. A greater value is assigned to λ_G in order to reflect the intuition that there should be more words in the background than in the other distributions, so the mass is expected to be shared on a larger number of words.

Unlike for the word distributions, mixing probabilities are drawn from a Dirichlet distribution with asymmetric priors. The prior knowledge about the origin of words in the base and update collections is again encoded at the level the hyper-parameters. For example, if we set $\gamma^\mathcal{A} = (5, 3, 2, 0)$, this would reflect the intuition that, on average, in the base collections, 50% of the words originate from the background distribution, 30% from the document-specific distribution, and 20% from the joint topic. Similarly, if we set $\gamma^\mathcal{B} = (5, 2, 2, 1)$, the prior reflects the assumption that, on average, in the update collections, 50% of the words originate from the background distribution, 20% from the document-specific distribution, 20% from the joint topic, and 10% from the novel, update topic[3]. The priors we have actually used are reported in Section 4.

3.2 Learning and inference

In order to find the optimal model parameters, the following equation needs to be computed:

$$p(\mathbf{z}, \psi, \phi | \mathbf{w}, \mathbf{u}) = \frac{p(\mathbf{z}, \psi, \phi, \mathbf{w}, \mathbf{u})}{p(\mathbf{w}, \mathbf{u})}$$

Omitting hyper-parameters for notational simplicity, the joint distribution over the observed variables is:

$$\begin{aligned} p(\mathbf{w}, \mathbf{u}) = {} & p(\phi^G) \times \\ & \prod_c p(\phi^{\mathcal{A}_c}) p(\phi^{\mathcal{B}_c}) \times \\ & \prod_d p(u_{cd}) p(\phi^{cd}) \int_\Delta p(\psi^{cd} | u_{cd}) d\psi^{cd} \times \\ & \prod_n \sum_{cdn} p(w_{cdn} | z_{cdn}) p(z_{cdn} | \psi^{cd}) \end{aligned}$$

where Δ denotes the 4-dimensional simplex[4]. Since this equation is intractable, we need to perform approximate inference in order to estimate the model parameters. A number of Bayesian statistical inference techniques can be used to address this problem.

[3] To highlight the difference between asymmetric and symmetric priors we put the indices in superscript and subscript respectively.

[4] Remember that, for base documents, words cannot be generated by the update topic, so Δ denotes the 3-dimensional simplex for base documents.

Variational approaches (Blei et al., 2003) and collapsed Gibbs sampling (Griffiths and Steyvers, 2004) are common techniques for approximate inference in Bayesian models. They offer different advantages: the variational approach is arguably faster computationally, but the Gibbs sampling approach is in principal more accurate since it asymptotically approaches the correct distribution (Porteous et al., 2008). In this section, we provide details on a collapsed Gibbs sampling strategy to infer the model parameters of DUALSUM for a given dataset.

Collapsed Gibbs sampling is a particular case of Markov Chain Monte Carlo (MCMC) that involves repeatedly sampling a topic assignment for each word in the corpus. A single iteration of the Gibbs sampler is completed after sampling a new topic for each word based on the previous assignment. In a collapsed Gibbs sampler, the model parameters are integrated out (or collapsed), allowing to only sample $\mathbf{z}$. Let us call w_{cdn} the n-th word in document d in collection c, and z_{cdn} its topic assignment. For Gibbs sampling, we need to calculate $p(z_{cdn}|\mathbf{w}, \mathbf{u}, \mathbf{z_{-cdn}})$ where $\mathbf{z_{-cdn}}$ denotes the random vector of topic assignments except the assignment z_{cdn}.

$$p(z_{cdn} = j|\mathbf{w}, \mathbf{u}, \mathbf{z_{-cdn}}, \gamma^{\mathcal{A}}, \gamma^{\mathcal{B}}, \lambda) \propto \frac{n^{(w_{cdn})}_{-cdn,j} + \lambda_j}{\sum_{v=1}^{V} n^{(v)}_{-cdn,j} + V\lambda_j} \frac{n^{(cd)}_{-cdn,j} + \gamma_j^{u_{cd}}}{\sum_{k \in K}(n^{(cd)}_{-cdn,k} + \gamma_k^{u_{cd}})}$$

where $K = \{G, cd, \mathcal{A}_c, \mathcal{B}_c\}$, $n^{(v)}_{-cdn,j}$ denotes the number of times word v is assigned to topic j excluding current assignment of word w_{cdn} and $n^{(cd)}_{-cdn,k}$ denotes the number of words in document d of collection c that are assigned to topic j excluding current assignment of word w_{cdn}.

After each sampling iteration, the model parameters can be estimated using the following formulas[5].

$$\phi_w^k = \frac{n_k^{(w)} + \lambda_k}{\sum_{v=1}^{V} n_k^{(v)} + V\lambda_k}$$

$$\psi_k^{cd} = \frac{n_k^{(cd)} + \lambda_k}{\sum n_{\cdot}^{(cd)} + V\lambda_k}$$

where $k \in K$, $n_k^{(v)}$ denotes the number of times word v is assigned to topic k, and $n_k^{(cd)}$ denotes the number of words in document d of collection c that are assigned to topic k.

By the strong law of large numbers, the average of sample parameters should converge towards the true expected value of the model parameter. Therefore, good estimates of the model parameters can be obtained averaging over the sampled values. As suggested by Gamerman and Lopes (2006), we have set a lag (20 iterations) between samples in order to reduce auto-correlation between samples. Our sampler also discards the first 100 iterations as burn-in period in order to avoid averaging from samples that are still strongly influenced by the initial assignment.

4 Experiments in Update Summarization

The Bayesian graphical model described in the previous section can be run over a set of news collections to learn the background distribution, a joint distribution for each collection, an update distribution for each collection and the document-specific distributions. Once this is done, one of the learned collections can be used to generate the summary that best approximates this collection, using the greedy algorithm described by Haghighi and Vanderwende (2009). Still, there are some parameters that can be defined and which affects the results obtained:

- DUALSUM's choice of hyper-parameters affects how the topics are learned.
- The documents can be represented with n-grams of different lengths.
- It is possible to generate a summary that approximates the joint distribution, the update-only distribution, or a combination of both.

This section describes how these parameters have been tuned.

4.1 Parameter tuning

We use the TAC 2008 and 2009 update task datasets as training set for tuning the hyper-parameters for the model, namely the pseudo-counts for the two Dirichlet priors that affects the topic mix assignment for each document. By performing a grid search over a large set of possible hyper-parameters, these have been fixed to

[5] The interested reader is invited to consult (Wang, 2011) for more details on using Gibbs sampling for LDA-like models

$\gamma^{\mathcal{A}} = (90, 190, 50, 0)$ and $\gamma^{\mathcal{B}} = (90, 170, 45, 25)$ as the values that produced the best ROUGE-2 score on those two datasets.

Regarding the base collection, this can be interpreted as setting as prior knowledge that roughly 27% of the words in the original dataset originate from the background distribution, 58% from the document-specific distributions, and 15% from the topic of the original collection. We remind the reader that the last value in $\gamma^{\mathcal{A}}$ is set to zero because, due to the problem definition, the original collection must have no words generated from the update topic, which reflects the most recent developments that are still not present in the base collections $\mathcal{A}$.

Regarding the update set, 27% of the words are assumed to originate again from the background distribution, 51% from the document-specific distributions, 14% from an topic in common with the original collection, and 8% from the update-specific topic. One interesting fact to note from these settings is that most of the words belong to topics that are specific to single documents (58% and 51% respectively for both sets $\mathcal{A}$ and $\mathcal{B}$) and to the background distribution, whereas the joint and update topics generate a much smaller, limited set of words. This helps these two distributions to be more focused.

The other settings mentioned at the beginning of this section have been tuned using the TAC-2010 dataset, which we reserved as our development set. Once the different document-specific and collection-specific distributions have been obtained, we have to choose the target distribution T to with which the possible summaries will be compared using the KL metric. Usually, the human-generated update summaries not only include the terms that are very specific about the last developments, but they also include a little background regarding the developing event. Therefore, we try, for KLSum, a simple mixture between the joint topic ($\phi^{\mathcal{A}}$) and the update topic ($\phi^{\mathcal{B}}$).

Figure 3 shows the ROUGE-2 results obtained as we vary the mixture weight between the joint $\phi^{\mathcal{A}}$ distribution and the update-specific $\phi^{\mathcal{B}}$ distribution. As can be seen at the left of the curve, using only the update-specific model, which disregards the generic words about the topic described, produces much lower results. The results improve as the relative weight of the joined topic model increases until it plateaus at a maximum around roughly the interval [0.6, 0.8], and from that point performance slowly degrades as at the right part of the curve the update model is given very little importance in generating the summary. Based on these results, from this point onwards, the mixture weight has been set to 0.7. Note that using only the joint distribution (setting the mixture weight to 1.0) also produces reasonable results, hinting that it successfully incorporates the most important n-grams from across the base and the update collections at the same time.

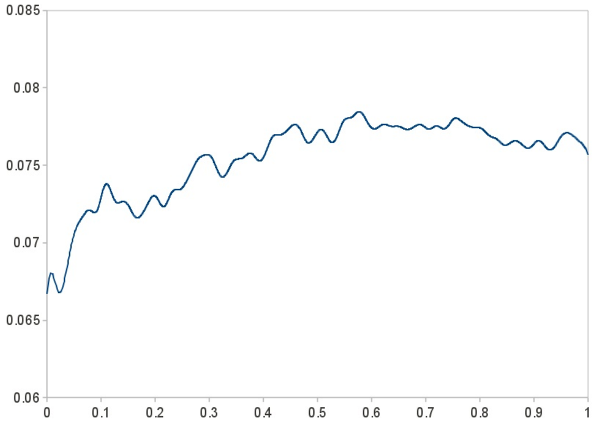

Figure 3: Variation in ROUGE-2 score in the TAC-2010 dataset as we change the mixture weight for the joined topic model between 0 and 1.

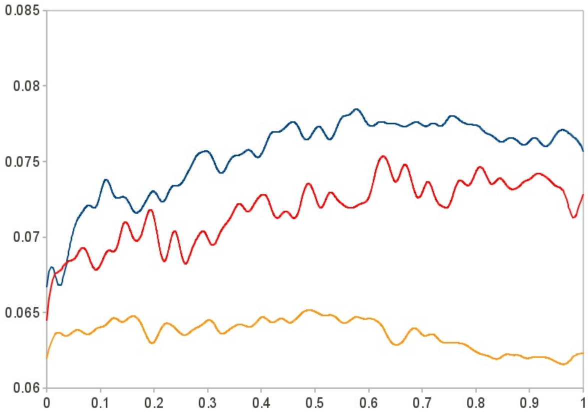

Figure 4: Effect of the mixture weight in ROUGE-2 scores (TAC-2010 dataset). Results are reported using bigrams (above, blue), unigrams (middle, red) and trigrams (below, yellow).

A second parameter is the size of the n-grams for representing the documents. The original implementations of SUMBASIC (Nenkova and Vanderwende, 2005) and TOPICSUM (Haghighi and Vanderwende, 2009) were defined over single words (unigrams). Still, Haghighi and Vanderwende (2009) report some improvements in the ROUGE-2 score when representing words as a bag of bigrams, and Darling (2010) mention similar improvements when running SUMBASIC with bigrams. Figure 4 shows the effect on the ROUGE-2 curve when we switch to using unigrams and trigrams. As stated in previous work, using bigrams has better results than using unigrams. Using trigrams was worse than either of them. This is probably because trigrams are too specific and the document collections are small, so the models are more likely to suffer from data sparseness.

4.2 Baselines

DUALSUM is a modification of TOPICSUM designed specifically for the case of update summarization, by modifying TOPICSUM's graphical model in a way that captures the dependency between the joint and the update collections. Still, it is important to discover whether the new graphical model actually improves over simpler applications of TOPICSUM to this task. The three baselines that we have considered are:

- Running TOPICSUM on the set of collections containing only the update documents. We call this run TOPICSUM$_{\mathcal{B}}$.
- Running TOPICSUM on the set of collections containing both the base and the update documents. Contrary to the previous run, the topic model for each collection in this run will contain information relevant to the base events. We call this run TOPICSUM$_{\mathcal{A}\cup\mathcal{B}}$.
- Running TOPICSUM twice, once on the set of collections containing the update documents, and the second time on the set of collections containing the base documents. Then, for each collection, the obtained base and update models are combined in a mixture model using a mixture weight between zero and one. The weight has been tuned using TAC-2010 as development set. We call this run TOPICSUM$_{\mathcal{A}}$+TOPICSUM$_{\mathcal{B}}$.

4.3 Automatic evaluation

DUALSUM and the three baselines[6] have been automatically evaluated using the TAC-2011 dataset. Table 1 shows the ROUGE results obtained. Because of the non-deterministic nature of Gibbs sampling, the results reported here are the average of five runs for all the baselines and for DUALSUM. DUALSUM outperforms two of the baselines in all three ROUGE metrics, and it also outperforms TOPICSUM$_{\mathcal{B}}$ on two of the three metrics.

The top three systems in TAC-2011 have been included for comparison. The results between these three systems, and between them and DUALSUM, are all indistinguishable at 95% confidence. Note that the best baseline, TOPICSUM$_{\mathcal{B}}$, is quite competitive, with results that are indistinguishable to the top participants in this year's evaluation. Note as well that, because we have five different runs for our algorithms, whereas we just have one output for the TAC participants, the confidence intervals in the second case were slightly bigger when checking for statistical significance, so it is slightly harder for these systems to assert that they outperform the baselines with 95% confidence. These results would have made DUALSUM the second best system for ROUGE-1 and ROUGE-SU4, and the third best system in terms of ROUGE-2.

The supplementary materials contain a detailed example of the the topic model obtained for the background in the TAC-2011 dataset, and the base and update models for collection D1110. As expected, the top unigrams and bigrams are all closed-class words and auxiliary verbs. Because trigrams are longer, background trigrams actually include some content words (e.g. *university* or *director*). Regarding the models for $\phi^{\mathcal{A}}$ and $\phi^{\mathcal{B}}$, the base distribution contains words related to the original event of an earthquake in Sichuan province (China), and the update distribution focuses more on the official (updated) death toll numbers. It can be noted here that the tokenizer we used is very simple (splitting tokens separated with white-spaces or punctuation) so that numbers such as 7.9 (the magnitude of the earthquake) and 12,000 or 14,000 are divided into two tokens. We thought this might be a for the bigram-based system to produce better results, but we ran the summarizers with a numbers-aware tokenizer and the statistical differences between versions still hold.

[6] Using the settings obtained in the previous section, having been optimized on the datasets from previous TAC competitions.

Method	**R-1**	**R-2**	**R-SU4**
TopicSum$_{\mathcal{B}}$	0.3442	0.0868	0.1194
TopicSum$_{\mathcal{A}\cup\mathcal{B}}$	0.3385	0.0809	0.1159
TopicSum$_{\mathcal{A}}$+TopicSum$_{\mathcal{B}}$	0.3328	0.0770	0.1125
DualSum	$0.3575^{\ddagger\dagger*}$	$0.0924^{\dagger*}$	$0.1285^{\ddagger\dagger*}$
TAC-2011 best system (Peer 43)	$0.3559^{\dagger*}$	$0.0958^{\dagger*}$	$0.1308^{\ddagger\dagger*}$
TAC-2011 2nd system (Peer 25)	$0.3582^{\dagger*}$	0.0926^{*}	$0.1276^{\dagger*}$
TAC-2011 3rd system (Peer 17)	$0.3558^{\dagger*}$	0.0886	$0.1279^{\dagger*}$

Table 1: Results on the TAC-2011 dataset. ‡, † and * indicate that a result is significantly better than TopicSum$_{\mathcal{B}}$, TopicSum$_{\mathcal{A}\cup\mathcal{B}}$ and TopicSum$_{\mathcal{A}}$+TopicSum$_{\mathcal{B}}$, respectively ($p < 0.05$).

4.4 Manual evaluation

While the ROUGE metrics provides an arguable estimate of the informativeness of a generated summary, it does not account for other important aspects such as the readability or the overall responsiveness. To evaluate such aspects, a manual evaluation is required. A fairly standard approach for manual evaluation is through pairwise comparison (Haghighi and Vanderwende, 2009; Celikyilmaz and Hakkani-Tur, 2011).

In this approach, raters are presented with pairs of summaries generated by two systems and they are asked to say which one is best with respect to some aspects. We followed a similar approach to compare DualSum with Peer 43 - the best system with respect to ROUGE-2, on the TAC 2011 dataset. For each collection, raters were presented with three summaries: a reference summary randomly chosen from the model summaries, and the summaries generated by Peer 43 and DualSum. They were asked to read the summaries and say which one of the two generated summaries is best with respect to: 1) Overall responsiveness: which summary is best overall (both in terms of content and fluency), 2) Focus: which summary contains less irrelevant details, 3) Coherence: which summary is more coherent and 4) Non-redundancy: which summary repeats less the same information. For each aspect, the rater could also reply that both summary were of the same quality.

For each of the 44 collections in TAC-2011, 3 ratings were collected from raters[7]. Results are reported in Table 2. DualSum outperforms Peer 43 in three aspects, including *Overall Responsiveness*, which aggregates all the other scores and can be considered the most important one. Regarding *Non-redundancy*, DualSum and Peer 43 obtain similar results but the majority of raters found no difference between the two systems. Fleiss κ has been used to measure the inter-rater agreement. For each aspect, we observe $\kappa \sim 0.2$ which corresponds to a slight agreement; but if we focus on tasks where the 3 ratings reflect a preference for either of the two systems, then $\kappa \sim 0.5$, which indicates moderate agreement.

	Best system		
Aspect	**Peer 43**	**Same**	**DualSum**
Overall Responsiveness	39	25	**68**
Focus	41	22	**69**
Coherence	39	30	**63**
Non-redundancy	40	**53**	39

Table 2: Results of the side-by-side manual evaluation.

4.5 Efficiency and applicability

The running time for summarizing the TAC collections with DualSum, averaged over a hundred runs, is 4.97 minutes, using one core (2.3 GHz). Memory consumption was 143 MB.

It is important to note as well that, while TopicSum incorporates an additional layer to model topic distributions at the sentence level, we noted early in our experiments that this did not improve the performance (as evaluated with ROUGE) and consequently relaxed that assumption in DualSum. This resulted in a simplification of the model and a reduction of the sampling time.

While five minutes is fast enough to be able to experiment and tune parameters with the TAC collections, it would be quite slow for a real-time summarization system able to generate summaries on request. As can be seen from the plate diagram in Figure 2, all the collections are generated independently from each other. The only exception, for which it is necessary to have all

[7] In total 132 raters participated to the task via our own crowdsourcing platform, not mentioned yet for blind review.

the collections available at the same time during Gibbs sampling, is the background distribution, which is estimated from all the collections simultaneously, roughly representing 27% of the words, that should appear distributed across all documents.

The good news is that this background distribution will contain closed-class words in the language, which are domain-independent (see supplementary material for examples). Therefore, we can generate this distribution from one of the TAC datasets only once, and then it can be reused. Fixing the background distribution to a pre-computed value requires a very simple modification of the Gibbs sampling implementation, which just needs to adjust at each iteration the collection and document-specific models, and the topic assignment for the words.

Using this modified implementation, it is now possible to summarize a single collection independently. The summarization of a single collection of the size of the TAC collections is reduced on average to only three seconds on the same hardware settings, allowing the use of this summarizer in an on-line application.

5 Conclusions

The main contribution of this paper is DUALSUM, a new topic model that is specifically designed to identify and extract novelty from pairs of collections.

It is inspired by TOPICSUM (Haghighi and Vanderwende, 2009), with two main changes: Firstly, while TOPICSUM can only learn the main topic of a collection, DUALSUM focuses on the differences between two collections. Secondly, while TOPICSUM incorporates an additional layer to model topic distributions at the sentence level, we have found that relaxing this assumption and modeling the topic distribution at document level does not decrease the ROUGE scores and reduces the sampling time.

The generated summaries, tested on the TAC-2011 collection, would have resulted on the second and third position in the last summarization competition according to the different ROUGE scores. This would make DUALSUM statistically indistinguishable from the top system with 0.95 confidence.

We also propose and evaluate the applicability of an alternative implementation of Gibbs sampling to on-line settings. By fixing the background distribution we are able to summarize a distribution in only three seconds, which seems reasonable for some on-line applications.

As future work, we plan to explore the use of DUALSUM to generate more general contrastive summaries, by identifying differences between collections whose differences are not of temporal nature.

Acknowledgments

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement number 257790. We would also like to thank Yasemin Altun and the anonymous reviewers for their useful comments on the draft of this paper.

References

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

Florian Boudin, Marc El-Bèze, and Juan-Manuel Torres-Moreno. 2008. A scalable MMR approach to sentence scoring for multi-document update summarization. In *Coling 2008: Companion volume: Posters*, pages 23–26, Manchester, UK, August. Coling 2008 Organizing Committee.

J. Carbonell and J. Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336. ACM.

Asli Celikyilmaz and Dilek Hakkani-Tur. 2011. Discovery of topically coherent sentences for extractive summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 491–499, Portland, Oregon, USA, June. Association for Computational Linguistics.

Chaitanya Chemudugunta, Padhraic Smyth, and Mark Steyvers. 2006. Modeling general and specific aspects of documents with a probabilistic topic model. In *NIPS*, pages 241–248.

W.M. Darling. 2010. Multi-document summarization from first principles. In *Proceedings of the third Text Analysis Conference, TAC-2010*. NIST.

Hal Daumé, III and Daniel Marcu. 2006. Bayesian query-focused summarization. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting* *of the Association for Computational Linguistics*, ACL-2006, pages 305–312, Stroudsburg, PA, USA. Association for Computational Linguistics.

Günes Erkan and Dragomir R. Radev. 2004. Lexrank: graph-based lexical centrality as salience in text summarization. *J. Artif. Int. Res.*, 22:457–479, December.

S. Fisher and B. Roark. 2008. Query-focused supervised sentence ranking for update summaries. In *Proceedings of the first Text Analysis Conference, TAC-2008*.

Dani Gamerman and Hedibert F. Lopes. 2006. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman and Hall/CRC.

Jade Goldstein, Mark Kantrowitz, Vibhu Mittal, and Jaime Carbonell. 1999. Summarizing text documents: sentence selection and evaluation metrics. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, pages 121–128, New York, NY, USA. ACM.

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

A. Haghighi and L. Vanderwende. 2009. Exploring content models for multi-document summarization. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 362–370. Association for Computational Linguistics.

Feng Jin, Minlie Huang, and Xiaoyan Zhu. 2010. The thu summarization systems at tac 2010. In *Proceedings of the third Text Analysis Conference, TAC-2010*.

Kevin Lerman and Ryan McDonald. 2009. Contrastive summarization: an experiment with consumer reviews. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, NAACL-Short '09, pages 113–116, Stroudsburg, PA, USA. Association for Computational Linguistics.

Xuan Li, Liang Du, and Yi-Dong Shen. 2011. Graph-based marginal ranking for update summarization. In *Proceedings of the Eleventh SIAM International Conference on Data Mining*. SIAM / Omnipress.

Rebecca Mason and Eugene Charniak. 2011. Extractive multi-document summaries should explicitly not contain document-specific content. In *Proceedings of the Workshop on Automatic Summarization for Different Genres, Media, and Languages*, WASDGML '11, pages 49–54, Stroudsburg, PA, USA. Association for Computational Linguistics.

A. Nenkova and L. Vanderwende. 2005. The impact of frequency on summarization. *Microsoft Research, Redmond, Washington, Tech. Rep. MSR-TR-2005-101*.

Ian Porteous, David Newman, Alexander Ihler, Arthur Asuncion, Padhraic Smyth, and Max Welling. 2008. Fast collapsed Gibbs sampling for latent Dirichlet allocation. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 569–577, New York, NY, USA, August. ACM.

Dragomir R. Radev, Hongyan Jing, Malgorzata Styś, and Daniel Tam. 2004. Centroid-based summarization of multiple documents. *Inf. Process. Manage.*, 40:919–938, November.

Frank Schilder, Ravikumar Kondadadi, Jochen L. Leidner, and Jack G. Conrad. 2008. Thomson reuters at tac 2008: Aggressive filtering with fastsum for update and opinion summarization. In *Proceedings of the first Text Analysis Conference, TAC-2008*.

Chao Shen and Tao Li. 2010. Multi-document summarization via the minimum dominating set. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 984–992, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ian Soboroff and Donna Harman. 2005. Novelty detection: the trec experience. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 105–112, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Yi Wang. 2011. Distributed gibbs sampling of latent dirichlet allocation: The gritty details.

Li Wenjie, Wei Furu, Lu Qin, and He Yanxiang. 2008. Pnr2: ranking sentences with positive and negative reinforcement for query-oriented update summarization. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 489–496, Stroudsburg, PA, USA. Association for Computational Linguistics.