Comparison of models based on summaries or documents towards extraction of update summaries

Maheedhar Kolla
University of Waterloo
Canada. N2L 3G1
mkolla@uwaterloo.ca

Olga Vechtomova
University of Waterloo
Canada. N2L 3G1
ovechtom@uwaterloo.ca

Charles L.A Clarke
University of Waterloo
Canada. N2L 3G1
claclarke@uwaterloo.ca

ABSTRACT

In this paper, we experiment with methods to extract sentences as summaries for a given cluster of documents. We construct background model(s) to capture information seen so far and use these model(s) to extract summary for current cluster. We propose two methods to construct these background models using documents seen so far and using summaries of previously seen document(s). We then compare the performance of these methods in context of DUC 2007 update task.

1. INTRODUCTION

Users looking for information about a series of related events face a daunting task of filtering out redundant information. In current work, we aim to extract sentences as summary for a given cluster under the assumption that user has already gone through previous document cluster(s). In current paper, we create a background model using document sets, already read by the user, and use that model to find an *updated probability of user selecting this word in the given document, given the background model* as follows:

$$P(w|d) = (1-\alpha)P_{mle}(w|d) - (\alpha)(P(w|\theta_b)) \quad (1)$$

where $P_{mle}(w|d)$ is the probability of the word in the current document and $P(w|\theta_b)$ represents the probability of the term in the background model and α with values between [0,1], could be used to control the effect of the background model. Using this framework, we wish to explore the following questions:

- Do background model constructed from documents better in estimating the probabilities?
- Could we obtain a better background model by only considering the *gist* of the previous document sets?

Although *gist* of previous documents could be determined in several ways, we consider the single document summary of each previously seen document to build the background model in our experiments. In the following sections, we explain the methods to compute the term probabilities and use them to rank the sentences in a given document set. We then evaluate the methods using the test set provided by NIST and automatic evaluation software, ROUGE.

2. CLUSTER BASED LANGUAGE MODEL

Probability of user selecting the word, while attempting to extract a summary, could be computed as follows:

$$P_{mle}(w|d) = \frac{tf(w,d)}{\sum_{w_j} C(w_j;d)} \quad (2)$$

where $tf(w,d)$ represent the raw term frequency of the word in the given document. As mentioned in Zhai.et.al[3], this sort of probability would under-estimate the unseen words in the documents. One possible solution to this problem is *smoothing*, which distributes the weights of the terms that are more frequent in the topic set but have relatively lower count in the current document. This method of smoothing, cluster based smoothing model (CBDM), was proposed by [1] who compute the probability of a word in the given document by:

$$P(w|d) = \lambda(P_{mle}(w|d)+(1-\lambda)[\beta(P(w|cl)+(1-\beta)(P(W|coll))] \quad (3)$$

where λ and β are smoothing parameters that could be determined through training. The maximum likelihood estimate of the word in the given cluster could be calculated as

$$P(w|cl) = \lambda(\frac{tf(w,cl)}{\sum_{w^j \epsilon cl} tf(w^j;cl)})+((1-\lambda)\frac{tf(w,coll}{\sum_{w^j \epsilon V} tf(w^j,coll)} \quad (4)$$

where $tf(w,cl)$ and $tf(w,coll)$ are the frequencies of a term in the given topic cluster and general collection. In our current experiments, we used the document sets of documents given by NIST as clusters and AQUAINT corpus for the collection model.

We then score the sentences in a given document based on the probabilities of the terms present in the sentence(s). This step is straightforward in cases where there is no background model, i.e. the first set. In the following sets, the term probabilities are interpolated with the background model probabilities as shown in Eq(1). In both cases, we then score the sentences using the term weights and extract the sentences until the summary of desired length is obtained.

3. BACKGROUND MODELS

The purpose of the background models constructed from the previously seen documents is to be able to update the give relative more weight to the terms that are novel to the current document. This could have a negative effect in cases where we could lower the weights of the terms that

Table 1: ROUGE-SU4 measures for various values of α on the two different background models [λ and β constant in this case]

Model	0.0	0.1	0.3	0.4
θ_{doc}	0.10632	0.11024	0.10830	0.10514
θ_{summ}	0.10632	0.11310	0.10630	0.10192

Table 2: ROUGE-2 measures for various values of α on the two different background models [λ and β constant in this case]

Model	0.0	0.1	0.3	0.4
θ_{doc}	0.06930	0.07221	0.07073	0.06938
θ_{summ}	0.06930	0.07400	0.06752	0.06433

are essentially topic central and give weight to terms that are not so important. To better understand the effects of what should the background models be constructed from, we conduct the following experiments:

- Models constructed from all previous documents.
- Models constructed from only summaries of previous documents.

3.1 Documents based background model

In order to estimate the probability of a given term in the background model, we compute the average probability [4]of the term across all previously seen document(s).

3.2 Summaries based background model

In this method of model construction, we first extract a generic summary for each document in previously seen set. The hypothesis behind using generic summaries is to be able to estimate the key words that are representative of documents. To extract generic summaries for each document, we rank sentences within the document based on the number of lexical links the sentence forms with the rest of the sentences as proposed by [2]. We then extract the top 2 sentences for each document as their generic summary. We then concatenate all generic summaries and use this concatenated text as background model to compute term probabilities.

4. EXPERIMENTS

In order to evaluate effects of background models in extracting summaries, we used the topics and document sets provided by NIST in context of DUC 2007 update task. The task is defined as to create summaries (length <= 100 words) for each of the given document set under the assumption that the user has gone through the documents in the previous set(s). In order to evaluate, we used the ROUGE package, which is based on the *n-gram* overlap between system generated summaries and the human generated model summaries.

To start with, we used the first set of documents, ones that do not have a background model, to find the optimal values for the parameters λ and β. By varying those parameters, 0we computed the weight of each term in a document and then extract the top ranked sentences as summaries. We narrowed down the values of λ and β to 0.4 and 0.6 respectively, for which the ROUGE scores (ROUGE-SU4 and ROUGE-2) are maximized.

Now with the values of λ and β fixed, we studied the effect of combining different background models (with constant α) and also the effect of the parameter α on the system's performance.

In order to save some table space, we have not shown the performance with respect to all parameter settings. In the tables above, we could observe that as the value of α increases towards 1, the performance of the system using documents as background model experiences a steeper drop than the performance of the system using summaries for background model. Also, it should be noted that both models show some improvment in performance at $\alpha = 0.1$ when compared to $\alpha = 0$, which is equivalent of not using any background model.

5. CONCLUSION AND FUTURE WORK

In this paper, we propose methods to construct background models to capture the information already seen by user. We then used these models to update the probability of a user selecting a term from current document, to extract a summary. We compared the performance of these models using the data set provided by NIST in context of DUC 2007 update task.

One drawback of this set of experiments is the nature of data. Document sets in these experiments were handpicked by humans for the task and it would be interesting to see the performance of methods using data closer to real world problem. Further experiments are to be carried out to combine query-likelihood models to attain a better estimate of the word probabilities.

6. REFERENCES

[1] X. Liu and W. B. Croft. Cluster-based retrieval using language models. In *Proceedings of the 27th SIGIR conference*, pages 186–193, NY, USA, 2004.

[2] O. Vechtomova, M. Karamuftuoglu, and S. Robertson. A study of document relevance and lexical cohesion between query terms. In *ELECTRA 2005*, pages 18–25, Brazil, (2005).

[3] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th SIGIR conference*, pages 334–342, NY, USA, 2001.

[4] C. X. Zhai, W. W. Cohen, and J. Lafferty. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *Proceedings of the 26th annual international ACM SIGIR conference*, pages 10–17, NY, USA, 2003.