THE NTNU SUMMARIZATION SYSTEM AT TAC 2009

Shih-Hsiang Lin and Berlin Chen

Department of Computer Science & Information Engineering
National Taiwan Normal University, Taipei, Taiwan
{shlin, berlin}@csie.ntnu.edu.tw

ABSTRACT

In this paper, we presents the results obtained by using a probabilistic summarization framework for the TAC 2009 update summarization task, which has the merits of combining the sentence generative probability and the sentence prior probability for sentence ranking systematically. Especially, each sentence of a document to be summarized is treated as a probabilistic generative model for predicting the documents. Nevertheless, the results of our first participation in the TAC evaluation seem to have room for further improvement.

1. INTRODUCTION

Automated summarization systems which enable user to quickly digest the important information conveyed by either a single or a cluster of documents are indispensible for managing the rapidly growing amount of textual information and multimedia content. A summary can be either abstractive or extractive [1]. In abstractive summarization, a fluent and concise abstract that reflects the key concepts of a document is generated, whereas in extractive summarization, the summary is usually formed by selecting salient sentences from the original document. The former requires highly sophisticated natural language processing techniques, including semantic representation and inference, as well as natural language generation, while this would make abstractive approaches difficult to replicate or extend from constrained domains to more general domains.

In addition to being extractive or abstractive, a summary may also be generated by considering several other aspects like being generic or query-oriented summarization, single-document or multi-document summarization, and so forth. The readers may refer to Mani and Maybury (1999) for a comprehensive overview of automatic text summarization. In this work, we focus exclusively on extractive summarization which also forms the building block for many other summarization tasks.

Aside from traditional ad-hoc extractive summarization methods [1], machine-learning approaches with either supervised or unsupervised learning strategies have gained much attention and been applied with empirical success to many summarization tasks [2]. For supervised learning strategies, the summarization task is usually cast as a two-class (summary and non-summary) sentence-classification problem [3]: A sentence with a set of indicative features is input to the classifier (or summarizer) and a decision is then returned from it on the basis of these features. In general, they usually require a training set, comprised of several documents and their corresponding handcrafted summaries (or labeled data), to train the classifiers. However, manual labeling is expensive in terms of time and personnel. The other potential problem is the so-called "*bag-of-sentences*" assumption implicitly made by most of these summarizers. In other words, sentences are classified independently of each other without leveraging the dependence relationships among the sentences or the global structure of the document [2].

Another line of work attempts to conduct document summarization using unsupervised machine-learning approaches, getting around the need for manually labeled training data. Most previous studies conducted along this line have their roots in the concept of sentence *centrality* [4-5]. Put simply, sentences more similar to others are deemed more salient to the main theme of the document; such sentences thus will be selected as part of the summary. Even though the performance of unsupervised summarizers is usually worse than that of supervised summarizers, their domain-independent and easy-to-implement property still makes them attractive.

Building on these observations, we expect that researches conducted along the above-mentioned two directions could complement each other, and it might

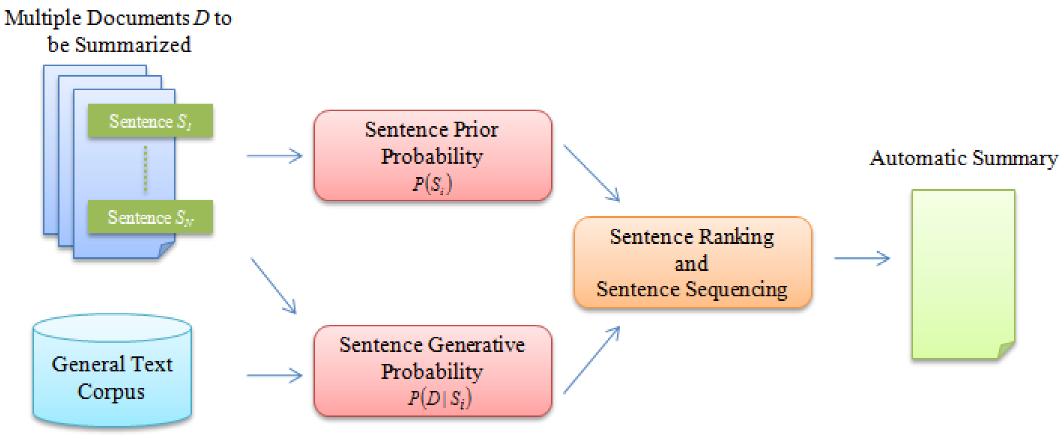
Multiple Documents D to be Summarized
Sentence S_1
Sentence S_N
General Text Corpus
Sentence Prior Probability $P(S_i)$
Sentence Generative Probability $P(D|S_i)$
Sentence Ranking and Sentence Sequencing
Automatic Summary

Figure 1: A schematic description of the probabilistic summarization framework.

be possible to inherit their individual merits to overcome their inherent limitations [6]. Therefore, we propose a probabilistic summarization framework that naturally integrates the above-mentioned two modeling paradigms for the TAC 2009 update summarization task.

2. PROBABILISTIC GENERATIVE FRAMEWORK

We have recently presented an unsupervised probabilistic framework for extractive summarization recently [6-7], where each sentence S_i of a document D is treated as a language model for generating D, and the sentences are ranked and selected according to their posterior probability $P(S_i|D)$ expressed by

$$P(S_i|D)=\frac{P(D|S_i)P(S_i)}{P(D)}, \qquad (1)$$

where $P(D|S_i)$ is the sentence generative probability, i.e., the likelihood that D is generated by S_i; $P(S_i)$ is the prior probability of S_i being important; and the evidence $P(D)$ is the marginal probability of D, it can be eliminated because it is identical for all sentences. A remarkable feature of this framework is that that a sentence to be considered as part of the summary is evaluated from two different perspectives: (1) $P(S_i)$ addresses the importance of sentence S_i itself; (2) $P(D|S_i)$ captures the degree of relevance between the document D and sentence $P(S_i)$. Fig. 1 illustrates extractive document summarization using the probabilistic generative framework.

2.1. Sentence Generative Model

In order to estimate the sentence generative probability, we explore the language modeling (LM) approach, which has been introduced to a wide spectrum of IR tasks and demonstrated with good empirical success, to predict the sentence generative probability [8]. In the LM approach, each sentence in a document can be simply regarded as a probabilistic generative model consisting of a unigram distribution (the so-called "*bag-of-words*" assumption) for generating the document:

$$P(D|S_i)=\prod_{w_j\in D}P(w_j|S_i)^{c(w_j,D)}, \qquad (2)$$

where $c(w_j,D)$ is the number of times that index term (or word) w_j occurs in D, reflecting that w_j will contribute more in the calculation of $P(D|S_i)$, or the document likelihood. Note that the sentence model $P(w_j|S_i)$ is simply estimated on the basis of the frequency of index term w_j occurring in the sentence S_i with the maximum likelihood (ML) criterion. In a sense, it belongs to a kind of literal term matching strategy and may suffer the problem of unreliable model estimation owing particularly to only a few sampled index terms present in the sentence [9]. To mitigate this potential defect, a unigram probability estimated from a general collection, which models the general distribution of words in the target language, is often used to smooth the sentence model. On the other hand, there probably would be word usage mismatch between a document and one of its sentences even if they are topically related to each other. Therefore, instead of constructing the sentence models based on literal term information, we can exploit the probabilistic topic models to represent a sentence through a latent topic space [8]:

TABLE I: THE MANUAL RESULTS OF OUR PROPOSED UPDATED SUMMARIZER

RUN	DATA	READABILITY (RANK)	RESPONSIVENESS (RANK)	PYRAMID (RANK)
RUN 33	SET A	4.705 (33)	3.841 (42)	0.232 (41)
	SET B	4.727 (29)	3.682 (36)	0.181 (39)
RUN 36	SET A	4.864 (25)	4.000 (34)	0.248 (34)
	SET B	4.591 (32)	4.000 (25)	0.203 (25)

TABLE II: THE ROUGE RESULTS OF OUR PROPOSED UPDATED SUMMARIZER

RUN	DATA	ROUGE-2	ROUGE-SU4
RUN 33	SET A	0.08470 (0.07288 - 0.09633)	0.12314 (0.11240 - 0.13425)
	SET B	0.07002 (0.05937 - 0.08092)	0.11617 (0.10662 - 0.12592)
RUN 36	SET A	0.08821 (0.07650 - 0.10029)	0.12647 (0.11630 - 0.13775)
	SET B	0.08260 (0.07180 - 0.09351)	0.12415 (0.11407 - 0.13428)

$$P(w_j \mid S_i) = \sum_k P(w_j \mid T_k) P(T_k \mid S_i) \tag{3}$$

where $P(w_j|T_k)$ and $P(T_k|S_i)$, respectively, are the probability of a word w_j occurring in a specific latent topic T_k and the probability of the topic T_k conditioned on S_i. This computation, in fact, exhibits some sort of concept matching. The probabilistic latent semantic analysis (PLSA) [3] and the latent Dirichlet allocation (LDA) [6] are often considered two basic representatives of this category and hence can be leverage to calculate the document likelihood $P(D|S_i)$.

Moreover, in order to avoid the redundancy in the updated summary, we modified Eq. (2) as

$$\tilde{P}(D|S_i) = P(D|S_i) - \beta \times P(D|\mathbf{Summ}) \tag{4}$$

where **Summ** represents the set of sentences that have already been included into the summary and the novelty factor β is used to trade off between relevance and redundancy.

2.2. Sentence Prior Probability

The sentence prior can be regarded as the likelihood of a sentence being important without seeing the whole document. It could be assumed uniformly distributed over sentences or estimated from a wide variety of factors like the lexical information, the structural information, to name a few. A straightforward way is to assume that the sentence prior probability $P(S_i)$ is in proportion to the posterior probability of a sentence $P(S_i)$ being included in the summary class when observing a set of indicative features derived from such factors or other sentence importance measures. These features can be integrated in a systematic way by taking the advantage of the learning capability of various supervised machine-learning methods.

However, due to the lack of document-summary reference pairs, in this work, the sentence prior probability $P(S_i)$ is instead estimated on the basis of the *centrality* of S_i among all sentences in a document to be summarized. More specifically, if a sentence S_i is more similar to other sentences in a document, it might be a representative sentence and can be used to depict the main theme of the document. For this idea to work, we adopt the LexRank [5] algorithm to estimate the sentence prior probability. LexRank conceptualizes the document to be summarized as a network of sentences, where each node represents a sentence and the associated weight of each link represents the lexical or topical similarity relationship between a pair of nodes. After the LexRank algorithm has been conducted on the conceptualized network of a document, the associated normalized similarity score of each sentence can be taken as its sentence prior probability.

3. EXPERIMENTS AND EVALUATION RESULTS

The TAC 2009 update summarization task is to generate short fluent multi-document summaries of news articles. For each topic, participants are given a topic statement expressing the information need of a user, and two chronologically ordered batches of articles (SET A and SET B) about the topic. Participants are asked to generate a 100-word summary for each batch of articles addressing the information need of the user. The summary of the second batch of articles (SET B) should be generated under the assumptions that the user has already read the earlier batch of articles (SET A) and the summary should provide the user with new information about the topic.

For our participation to TAC 2009, we were allowed to submit two runs, which are identified as runs 33 and 36. In what follow, we describe each of the above-mentioned runs and their associated performance evaluations.

Run33 We constructed the sentence model $P(w_j|S_i)$ based on the unigram language model (ULM), where each sentence of the document can respectively offer a unigram distribution for observing words, which is estimated on the basis of the words occurring in the sentence and is further smoothed by a unigram distribution estimated from a general collection.

Run36 We took LDA as an example for modeling the sentence model since it has exhibited the better performance among various probabilistic topic models in the literature. It is also worth mentioning that LDA was trained without supervision with a set of 16 latent topics.

Table I and II illustrate the manual evaluations and automatic evaluations of our submitted runs. As can be seen, the results demonstrate the superiority of utilizing the concept matching strategy (i.e., RUN 36) over the simple literal term matching (i.e., RUN 33), which confirm the utility of using the topical information for summarization. However, we found that our proposed two systems did not outperform other competing systems. One possible reason is that we do not leverage topic statements, which express the information needs, in our proposed probabilistic summarization framework. How to integrate the information need into our proposed framework will be one of our work items. The other possible reason is that the parameters used in this work were not optimally tuned. This will be one of our other considerations for future work

4. CONCLUSIONS

In this work, we have presented a probabilistic summarization framework, combining the sentence generative probability and the sentence prior probability, for TAC 2009 update summarization task. Each sentence of a document to be summarized is treated as a probabilistic generative model for predicting the document. Two modeling approaches, i.e., the unigram language model (ULM) and the sentence topic model (LDA), have been investigated to model the document-likelihoods. We believe that this initial attempt could provide a new avenue for future research on text summarization.

5. REFERENCES

1. Mani, I. and M.T. Maybury, *Advances in automatic text summarization*. 1999, Cambridge: MIT Press.
2. Lin, S. H., B. Chen, and H. M. Wang, *A comparative study of probabilistic ranking models for Chinese spoken document summarization.* ACM Transactions on Asian Language Information Processing, 2009. **8**(1): p. 3:1 - 3:23.
3. Kupiec, J., J. Pedersen, and F. Chen. *A trainable document summarizer*. in *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1999.
4. Gong, Y. and X. Liu. *Generic text summarization using relevance measure and latent semantic analysis*. in *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2001.
5. Erkan, G. and D. R. Radev, *LexRank: graph-based lexical centrality as salience in text summarization.* Journal or Artificial Intelligence Research, 2004. **22**: p. 457 - 479.
6. Lin, S. H., et al. *Hybrids of supervised and unsupervised models for extractive speech summarization*. in *Annual Conference of the International Speech Communication Association*. 2009.

7. Chen, Y. T., B. Chen, and H.-M. Wang, *A probabilistic generative framework for extractive broadcast news speech summarization.* IEEE Transactions on Audio, Speech and Language Processing, 2009. **17**(1): p. 95 - 106.

8. Zhai, C. X., *Statistical language models for information retrieval*. Synthesis lectures series on human language technologies. 2008: Morgan & Claypool Publishers.

9. Lee, L. S. and B. Chen, *Spoken document understanding and organization.* IEEE Signal Processing Magazine, 2005. **22**(5): p. 42 - 60.