The THU Summarization Systems at TAC 2010

Feng Jin, Minlie Huang and Xiaoyan Zhu
State Key Laboratory of Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology
Dept. of Computer Science and Technology, Tsinghua University, Beijing, China
jinfengfeng@gmail.com, {aihuang, zxy-dcs}@tsinghua.edu.cn

Abstract

The TAC 2010 Guided Summarization task requires participants to generate coherent summaries with the guidance of predefined categories and aspects. In this paper, we present our two extractive summarization systems. In the first system, we employ a topic model - Labeled LDA to model the aspects. The correspondence between the aspects and the topics in Labeled LDA is established through identifying indicative words for each aspect. After training and inference of Labeled LDA, we get the salience scores of concepts (named entities and bigrams) from topic concept distributions. Then we use an Integer Linear Programming (ILP) based maximal coverage method to generate summaries. In the other system which also uses ILP and maximal coverage during sentence extraction, the salience of concepts is obtained using a pairwise learning to rank algorithm - RankNet. The training samples are constructed based on the human annotated Pyramid data.

1 Introduction

Automatic document summarization has attracted a lot of researchers since 1950s (Luhn, 1958). And in recent years, Text Analysis Conference (TAC) has organized a series of summarization tracks by providing dataset, conducting both automatic and manual evaluations for fair comparison of systems, etc. In TAC 2010 Guided Summarization Task, participants are required to generate summaries with the guidance of predefined categories and aspects. For each TAC [1]topic which is specified into a category, the resulting summaries should cover important information of all the category's aspects as well as other content.

We participated the Guided Summarization Task and submitted two systems. In both systems, the Integer Linear Programming (ILP) based maximal coverage (Gillick and Favre, 2009; Takamura and Okumura, 2009) approach is used to extract a set of sentences that include important concepts of the given document collection to form a summary, where the concepts refer to word bigrams or named entities such as location and human names.

How to measure the salience of concepts is critical to the quality of the generated summaries. In our first system, we choose to exploit a supervised topic model - Labeled LDA (Ramage et al., 2009) to score the concepts, in order to make use of the information conveyed by the aspects. The correspondence between aspect and topic (in topic models) is established through indicative words.

In the other system we submitted, the specification of the aspects is just treated as the *narrative* in previous TAC tasks, and summaries are generated as in TAC 2009. The importance of each concept is computed by RankNet (Burges et al., 2005), a pairwise learning to rank algorithm. The training data is constructed using the Pyramid (Passonneau et al., 2005) annotations in previous years TAC dataset.

The rest of this paper is organized as follows. We

[1] The term *topic* may refer to two different meanings in this paper. One means a topic in topic models and the other denotes a given topic to be summarized in TAC. We believe these can be differentiated in the context.

introduce related works in section 2. In section 3 our systems are described in detail. And the experimental results are presented in section 4. Finally, conclusions and future works are outlined in section 5.

2 Related Work

Concept ranking based sentence selection methods have been frequently applied in document summarization and different concepts such as words, bigrams and atomic events (Filatova and Hatzivassiloglou, 2004) have been defined. And in performance evaluation of summarization systems, metrics are usually based on measuring the coverage of concepts in computer generated summaries, for example, n-grams in ROUGE, Basic Elements, and Semantic Content Units in Pyramid evaluation, etc. In recent years, Integer Linear Programming has been introduced for modeling sentence selection (Gillick and Favre, 2009; Takamura and Okumura, 2009). Although these systems usually consider bigrams as concepts and use document frequency to weigh the bigrams, they have achieved promising results.

Topic models have already been employed in document summarization (Haghighi and Vanderwende, 2009). However, there are usually no definition of aspect in previous summarization studies. In this paper, we used Labeled LDA to take into account the information provided by aspects.

Learning to rank is a hot topic in information retrieval society nowadays. The pairwise algorithm of RankNet is well-known and has been used in summarization of CNN news (Svore et al., 2007). Different from their work which ranks sentences, we use RankNet to rank concepts in this paper.

3 Our Summarization Approaches

Our summarization process contains two primary stages. First each concept is scored using Labeled LDA or RankNet, and then ILP is used to select a set of sentences that cover concepts with high scores. A concept (named entity or bigram) is a much smaller unit than a sentence and this can benefit redundancy removal, which is of great importance in multi-document summarization.

In the guided summarization task, each topic in TAC 2010 falls into a single category which has a list of predefined aspects. Although the automatically generated summaries are required to cover all these aspects as well as other important information in the original document collection, the use of these aspects in summarization systems is optional. And our second system which use RankNet for concept scoring ignores these aspects.

Here, we roughly divide the aspects into two types. The first type includes *WHEN*, *WHERE* and *WHO*, and these aspects may be identified through named entity recognition. All others belong to the second type, and we believe that each of them can be characterized by a language model which is a probabilistic distribution over terms (bigrams).

Suppose for each topic, the given document cluster to be summarized is $D = d_1, d_2, \cdots, d_M$. Note that in our first system which employs Labeled LDA, each sentence is considered as a document. Each document d_i is considered as a sequence of N_d concepts $\boldsymbol{c_d} = c_1, c_2, \cdots, c_{N_d}$. All the concepts belong to a vocabulary of size V. Each concept is either a named entity representing *time*, *location* and *person names*, or a bigram. In Labeled LDA, each document is treated as a bag of concepts and the concept order is not taken into account. As for RankNet, all the concepts in the document cluster is considered together.

3.1 Labeled LDA based concept ranking

Given the topic and its associated category, each aspect of that category corresponds to exactly one topic in Labeled LDA. In addition, a background topic is added in order to represent common information of the text which belongs to no aspect. So if there are K aspects in the category, the number of topic in Labeled LDA is $K + 1$. Then, similar to most existing work on topic models, each topic k is modeled as a multinomial distribution over concepts $\{c_i|\boldsymbol{\beta}_k\}_{c_i \in V}$ where $\boldsymbol{\beta}_k$ is the multinomial parameter of topic k.

Each document (each sentence is considered as a document as mentioned above) usually covers only a subset of all the topics (aspects). In Labeled LDA a set of labels $\boldsymbol{\Lambda} = \{l_1, l_2, \cdots, l_{k+1}\}$ is used where each label indicate the occurrence of a topic. Then for each document d, a subset of labels $\boldsymbol{\Lambda}_d \subseteq \boldsymbol{\Lambda}$ is associated in order to show that this document refers to just these topics. That is to say, the topic mixture distribution $\boldsymbol{\theta}_d$ of this document is defined only over the topics corresponds to labels in $\boldsymbol{\Lambda}_d$. It is clear that by incorporating supervision of document specific labels, Labeled LDA is an extension of LDA which addresses each document as a mixture of all topics.

Following the author of Labeled LDA, we used collapsed Gibbs sampling for learning and inference. After the learning and inference process, we obtain the topic mixture proportions $\boldsymbol{\theta}_d$ of each document d and the concept distribution $\boldsymbol{\beta}_z$ for each topic z. And according to correspondence between aspects and topics, we now know the probability distribution of concepts for each aspect. And for each concept c_i, the occurrence probability $P(c_i|\boldsymbol{\beta}_k)$ is used as its score. If c_i occurs in multiple aspects, the largest probability in the aspects is used as its final score.

An important issue not mentioned so far is that supervision information incorporated by Labeled LDA is the label set $\boldsymbol{\Lambda}_d$ of each document d. The best and most accurate annotations may be obtained through human labeling, but this is not applicable in the setting of our summarization task. We need to establish the labels which indicates topic occurrence in documents automatically.

For those aspects of *WHEN*, *WHO* and *WHERE*, it is straightforward to determine whether a document refers to these aspects through named entity recognition. As for other aspects, firstly frequent words for this aspect are obtained from the sample data provided by TAC 2010, in which each aspect is labeled clearly in the human written summaries. Then we calculate the Point Wise Mutual Information (PMI) between these words and words in TAC (DUC) test data from year 2006 to 2009. Then those with high PMI values are also included as indicative words. And finally, if and only if a document contains one of the aspect specific indicative words, we assume that the document refers to this aspect, and so the document is labeled to this aspect.

3.2 RankNet based concept ranking

In our second system THU_HUANG2, the specification of categories and aspects is ignored and the summarization task is handled in the same way as in previous years (such as the TAC 2009 summarization track). The aspects descriptions are used as narratives which specified user information needs in previous TAC tasks. Here, we choose to use learning to rank approach and concepts salience are measured in a quite different way.

RankNet is a pairwise learning to rank method (Burges et al., 2005) which has been adopted in commercial search engine (Liu, 2009). This algorithm learns a ranking function from a list of training examples where each training example is a pair of objects associated with the gold standard label.

Each concept c_i is first represented by a feature vector x_i. Given a pair of feature vectors (x_i, x_j), the gold standard probability $\overline{P}_{ij}$ is set to be 1 if the label of the pair is 1, which means x_i ranks ahead of x_j. Otherwise, the gold standard probability is 0 when the label of the pair is 0. Then the predicted probability P_{ij}, which is determined by the ranking scores of x_i and x_j, is modeled as a logistic function:

$$P_{ij} = \frac{exp(f(x_i) - f(x_j))}{1 + exp(f(x_i) - f(x_j))} \tag{1}$$

where $f(x)$ is the ranking function. The objective of the algorithm is to minimize the cross entropy between the gold standard probability and the predicted probability, which is defined as follows:

$$C_{ij}(f) = -\overline{P}_{ij} log P_{ij} - (1 - \overline{P}_{ij}) log(1 - P_{ij}) \tag{2}$$

In RankNet, a three-layer (one hidden layer) neural network where the third layer contains a single node is used as the ranking function, and the output of the third layer is the ranking score for the input feature vector, as follows:

$$f(x_n) = g^3(\sum_j w_{ij}^{32} g^2(\sum_k w_{jk}^{21} x_{nk} + b_j^2) + b_i^3) \tag{3}$$

where for weights w and bias b, the superscripts indicate the node layer while the subscripts indicate the node indexes within each layer. And x_{nk} is the k-th component of input feature vector x_n. Then a gradient descent method based on back propagation is used to learn the parameters. For more details, please refer to the original paper (Burges et al., 2005).

The features we exploit to characterize each concept are listed in the following:

1. Cluster frequency: $tf_D(c_i)$ which is the frequency of c_i in the given document cluster for this topic;

2. Title frequency: the occurrence frequency of a concept c_i in the titles of the given documents;

3. Query frequency: the frequency of c_i occurring in the topic title and aspect descriptions;

4. Average term frequency: $\sum_{d \in D} tf_d(c_i)/|D|$, the term frequency of c_i averaged by the size of the document cluster;

5. Document frequency: the document frequency of c_i;

6. Minimal position: the first occurrence position of c_i;

7. Average position: the average occurrence position of c_i in the cluster D.

3.3 ILP based summarization

After we have salience scores for concepts, either through Labeled LDA or RankNet, we need to select an optimal group of sentences that contains as many high scored concepts as possible. We choose to use the ILP approach (Gillick and Favre, 2009; Takamura and Okumura, 2009; Jin et al., 2010) which addresses sentence selection as a global optimization problem and generally outperforms greedy search based selection methods. In ILP both the optimization objective and constraints are linear, with some variables restricted to be integers.

The ILP selection approach is formally presented as follows:

$$
\begin{aligned}
& \max \sum_i s_i \cdot z_i^x \\
& s.t. \\
& \sum_j z_j^u \cdot |u_j| \leq L \\
& \sum_j z_j^u \cdot I(i,j) \geq z_i^x, \forall i \\
& z_i^x, z_j^u \in \{0,1\}, \forall i, j
\end{aligned}
\tag{4}
$$

where:

- $\boldsymbol{s_i}$ - the score (weight) of a concept x_i calculated by Labeled LDA or the ranking function $f(x_i)$ in RankNet;
- $\boldsymbol{u_j}$ - the selection unit which represents a sentence in this paper;
- $|\boldsymbol{u_j}|$ - the number of words in u_j;
- $\boldsymbol{z_i^x}$ - the indicator variable which denotes the presence or absence of x_i in the summary;
- $\boldsymbol{z_j^u}$ - the indicator variable denoting whether u_j is included in the summary;
- $\boldsymbol{I(i,j)}$ - a binary constant indicating whether x_i appears in u_j;
- $\boldsymbol{L}$ - the length limit of the resulting summary.

The first constraint in the above ILP addresses the length limit. And the second constraint takes consistency into account - if a concept c_i is included in the summary, one or more sentences that contain c_i must be selected. Redundancy removal and content diversity of the generated summary are handled implicitly where each concept is considered once in the objective function. And duplicate inclusion of concepts will not benefit the objective. In implementation, the [2]LpSolve package is adopted to solve the above ILP problem.

4 Experimental Results

In our systems, the given documents are first segmented into sentences using the [3]LingPipe toolkit. Then we conducted named entity recognition using the [4]Stanford NER package. It can identifies *person*, *organization* and *location* from text, where *person* and *location* are closely related to the *WHO* and *WHERE* aspects, repectively. In addition, for the *WHEN* aspect we use regular expression to find date and time in sentences. The same entity may have different names, for example *Barack Hussein Obama* may also be written as *Barack Obama* or *Obama*. So we have also performed coreference resolution and normalized the named entities as in our TAC 2009 system (Long et al., 2009). Words are

[2] http://lpsolve.sourceforge.net/5.5/

[3] http://alias-i.com/lingpipe/index.html

[4] http://nlp.stanford.edu/ner/index.shtml

Table 1: Evaluation results of THU_HUANG1 system

Cluster	Main task			Update task		
Evaluation Metric	Best	Ours	Rank	Best	Ours	Rank
ROUGE-2	0.096	0.095	2	0.080	0.067	12
ROUGE-SU4	0.130	0.124	4	0.120	0.108	12
Average Modified Pyramid Score	0.425	0.371	13	0.321	0.243	13
Average Linguistic Quality	3.652	3.152	10	3.739	3.283	4
Average Overall Responsiveness	3.174	2.978	6	2.717	2.304	10

Table 2: Evaluation results of THU_HUANG2 system

Cluster	Main task			Update task		
Evaluation Metric	Best	Ours	Rank	Best	Ours	Rank
ROUGE-2	0.096	0.092	4	0.080	0.073	3
ROUGE-SU4	0.130	0.123	6	0.120	0.108	11
Average Modified Pyramid Score	0.425	0.397	5	0.321	0.274	6
Average Linguistic Quality	3.652	3.304	5	3.739	3.130	6
Average Overall Responsiveness	3.174	3.022	5	2.717	2.457	6

stemmed using the Porter Stemmer and bigrams are extracted. And a bigram is discarded if both of the two words it contains are stopwords.

For RankNet, the training samples are constructed from data in TAC 2008 summarization task. The gold standard label of each concept pair is obtained with the help of human annotated Pyramid evaluation data. The weight of each Summary Content Unit (SCU) is the number of human written reference summaries that contain it. And we assume that the weight of each concept is the largest weight of the SCUs that it occurs in. For concepts that do not occur in any SCU, their weight is 0. So the weight of each concept belongs to $\{0, 1, \cdots, 4\}$ since there are 4 reference summaries for each topic. Then the label of concept pair (x_i, x_j) is 1 if the weight of x_i is larger than that of x_j, and 0 otherwise.

Both our systems are submitted to TAC for evaluation and the ID of the Labeled LDA based system is 18 (THU_HUANG1) while the ID of RankNet based system is 36 (THU_HUANG2). The experimental results of the two system are presented in table 1 and Table 2, respectively.

We can see that our systems achieved promising results. The application of the aspects should be further considered and aspects need to be modeled more carefully, as the second system which makes less use of the aspect performs even better. Our systems get better results in the main task than in the update task, since we have handled the update task in the same way as in the main task, and no special efforts have been made.

5 Conclusion and Future Work

In this paper, we described our two systems submitted to TAC 2010 Guided Summarization task. The first system uses Labeled LDA to model the predefined aspects while the other one just uses the aspect descriptions as narrative in previous years' TAC tracks. In both systems, the salience of concepts is first measured and then an ILP based selection approach is applied for sentence selection and summary generation. Our systems have achieved promising results in TAC 2010 evaluations.

For guided summarization, the provided categories and aspects convey important information that can guide summary generation. And in the future, we would like to study the modeling of predefined aspects more sophisticatedly and how to make full use of the aspects.

Acknowledgments

This work was partly supported by the Chinese Natural Science Foundation under grant No. 60973104 and No. 60803075, and with the aid of a grant from the International Development Research Center, Ottawa, Canada IRCI project from the International Development.

References

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton and Greg Hullender. 2005. Learning to Rank Using Gradient Descent. In *Proceedings of the 22nd International Conference on Machine Learning*.

Elena Filatova and Vasileios Hatzivassiloglou. 2004. Event-based Extractive Summarization. In *Proceedings of ACL Workshop on Summarization*, volume 111.

Dan Gillick and Benoit Favre. 2009. A Scalable Global Model for Summarization. In *Proceedings of the Workshop on Integer Linear Programming for Natural Language Processing*.

Aria Haghighi and Lucy Vanderwende. 2009. Exploring Content Models for Multi-document Summarization. In *The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Feng Jin, Minlie Huang and Xiaoyan Zhu. 2010. A Comparative Study on Ranking and Selection Strategies for Multi-Document Summarization. In *Proceedings of COLING 2010.*

Tie-Yan Liu. 2009. Learning to Rank for Information Retrieval, Foundation and Trends on Information Retrieval. *Now Publishers*.

Chong Long, Minlie Huang and Xiaoyan Zhu. 2009. Tsinghua University at TAC 2009: Summarizing Multi-documents by Information Distance. In Proceedings of TAC 2009.

Luhn. 1958. The Automatic Creation of Literature Abstracts. *IBM J. of R. and D., 2(2), 1958.*

Rebecca J. Passonneau, Ani Nenkova, Kathleen McKeown and Sergey Sigelman. 2005. Applying the Pyramid Method in DUC 2005. *DUC 2005 Workshop.*

Daniel Ramage, David Hall, Ramesh Nallapati and Christopher D. Manning. 2009. Labeled LDA: a Supervised Topic Model for Credit Attribution in Multi-labeled Corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*.

Krysta Svore, Lucy Vanderwende, and Chris Burges. 2007. Enhancing Single-Document Summarization by Combining RankNet and Third-Party Sources. In *Proceedings of EMNLP-CoNLL (2007).*

Hiroya Takamura and Manabu Okumura. 2009. Text Summarization Model Based on Maximum Coverage Problem and its Variant. In *Proceedings of EACL, 2009.*