The CIST Summarization System at TAC 2011

Hongyan Liu, Ping'an Liu, Wei Heng, Lei Li
Center for Intelligence Science and Technology
Beijing University of Posts and Telecommunications
Beijing, 100876, China

Abstract

In this report, we present our extractive summarization system on both summarization and multiling tracks of TAC 2011. We introduce an extractive multi-document summarization method based on hierarchical topic model of hierarchical Latent Dirichlet Allocation (hLDA) and sentence compression. hLDA is a representative generative probabilistic model, which not only can mine latent topics from a large amount of discrete text data, but also can organize these topics into a hierarchy to achieve a deeper semantic analysis. We try to combine the hLDA model with some traditional features. The evaluation results showed some improvement compared with our own system in TAC 2010, which is based on sentence clustering. But there are still many problems needed to be studied in the future. As to the new multiling task of TAC 2011, we used the frame of the hLDA model again but deleting those knowledge base for English. We tried all the 7 languages, including Arabic, Czech, English, French, Hebrew, Hindi and Greek. The evaluations of human confirmed that our method has better performance than some other ones.

1 Introduction

With the increasingly wider use of the Internet and the improvement of social information, the amount of information in the public domain continues to increase and form rapid accumulation, rendering much of these information redundant. Companies, governments and research institutions are all facing the unprecedented challenges that how to process these information efficiently. In the other hand, with continuous increasing of the large amounts of valuable text data every day, it is very difficult to obtain needed information in the data resources for other technologies. Multi-document summarization is an essential technology to overcome this obstacle in technological environments. It aims to generate a brief and concise summary that retaining the main characteristic of the original set of documents, helping people to identify a summary of a set of documents without reading the entire texts.

With the explosive increase of documents on the Internet, this process can be used in many applications. At present, almost all of the science research results and news events appear on the Internet first and continuously updated. Multi-document summarization can extract important information from these topic related articles, and the most important content is presented to users in a condensed form and in a manner appropriate for the user's or application's needs. The short summaries for news groups in news services can facilitate users to understand the news articles in the group faster and better. With the rapid development of electronic commerce, product evaluation information can often be used as reference for customers, and also be used as a source of market feedback for companies. However, there are usually thousands of related evaluations on all aspects of products, summary that based on user's comments (Zhang Yanxing et al ,2009) can fuse various information and present in condensed form to the user. At the same time, mobile Internet is gradually coming into our daily lives. PDA, mobile phones and other intelligent terminals can access the Internet at anytime and anywhere. But mobile Internet has many restrictions, such as the transmission speed and the ease of browsing. Multi-document summary information is brief and can reduce data transmission. Thus users can access the most important content quickly. In short, multi-document summarization technology has important scientific significance and can be used in many applications.

The rest of this report is organized as follows. In Section 2 we will look at some of the algorithms and their derivations that researchers have proposed for multi-document summarization. Section 3 talks about hierarchical topic model of hierarchical Latent Dirichlet Allocation. Section 4 introduces our multi-document summarization algorithms. Section 5 gives the evaluation of our algorithms and Section 6 talks about the future work.

2 Related Works

In recent years, following the LSI(S. Deerwester et al, 1990) and pLSI(T. Hofmann,1999) models, complex probabilistic models are increasingly prevalent. The LDA (Latent Dirichlet Allocation)(D. M. Blei and M. Jordan , 2003) model is a representative and widespread topic model. It tries to reveal the deeper underlying topic information. It is a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. In the context of text modeling, the topic probabilities provide an explicit representation of a document. LDA has overcome the defects of pLSI and inherited the advantage of dimension reduction of pLSI. Firstly, the number of topics is fixed in LDA, and a model selection procedure is required to choose the number of topics. Secondly, a document or a sentence can have arbitrary probability distribution on the topics. Another important difference is that LDA is a model with flat structure; it can not seize the abstractive concept. Hierarchical Latent Dirichlet Allocation (hLDA) (D. Blei et al, 2003 and D.Blei et al, 2009) represents the future trend of unsupervised machine learning methods, which is a generalization of the Latent Dirichlet Allocation. The hLDA is a hierarchical model. It can successfully extract the potential connection between the levels of topics, and can automatically determine the number of topics, which not only helps us understanding, but also helps us obtaining the statistical structure of the documents or the sentences.

Rachit(2008) used the Latent Dirichlet Allocation to capture the events being covered by the documents and form the summary with sentences representing these different events. Their approach is distinguished from existing approaches in that they use mixture models to capture the topics and pick up the sentences without paying attention to the details of grammar and structure of the documents. After that, to reduce the common information content, the sentences of the summary need to be orthogonal to each other since orthogonal vectors have the lowest possible similarity and correlation between them. Rachit (2008) used Singular Value Decomposition to get the orthogonal representations of vectors and represent sentences as vectors, they can get the sentences that are orthogonal to each other in the LDA mixture model weighted term domain. Thus using LDA they find the different topics in the documents and using SVD they find the sentences that best represent these topics. Asli(2010) presented a novel approach that formulates MDS as a prediction problem based on a two-step hybrid model: a generative model for hierarchical topic discovery and a regression model for inference.

Wu Xiaofeng, Zong Chengqing(2009) proposed a new supervised method for the extraction-based single document summarization by adding LDA of the document as new features into a CRF summarization system. They study the power of LDA and analyze its different effects by changing the number of topics. Yang Xiao(2010) proposed a multi-document summarization method based on the LDA model in which the number of topics was determined by model perplexity. The probabilistic sentence distribution on topics and the probabilistic topic distribution on words were obtained by the Gibbs sampling method. The importance of topics was determined by the sum of topic weights on all sentences. They proposed two sentence scoring methods, one based on sentence distribution and the other on topic distribution.

Summarization Evaluation is an important driver that promoting the development of multi-document summarization system. Almost all authoritative international conferences about NLP have summarization task, leading the direction of research and development to a larger scale. In 2000, Document Understanding Conference (DUC) added the multi-document summarization system evaluation as one of the main tasks. DUC (now

TAC) also has kept multi-document summary evaluation as the main task since then. With the improvements of the tasks and the evaluation methods, participants can evaluate their systems on the large-scale and public corpus, indicating that the research is more standard and unified. TAC 2011 multi-document summarization task's objective is to make a deeper semantic analysis, which represents the trend of future research. There is a new task appearing in TAC 2011, which is the multiling multi-document summarization. There are 7 different languages for the first time. This is also a very important application.

In short, as the document itself is semi-structured or unstructured, without determining form and lack of machine-understandable semantics, the cross-field is related to information retrieval, natural language processing, machine learning and other fields. Presently a purely data-driven approach has obvious improvement. In paper(Asli Celikyilmaz. Dilek Hakkani-Tur,2010) a model is proposed based on hybrid hierarchical topic model for multi-document summarization, which used the ideal summaries, and achieved competitive results. However, this method has the disadvantage of needing ideal summaries. In practice the quality of the summary would be affected by the ideal summaries. The innovation of our proposed method is: using a simple approach, without the guide of ideal summaries, completely dependent on the document corpus, taking full advantage of hierarchy structure, including the concept of path and topic relationships between the levels of abstraction to extract summary sentences, and exploring sentence compression pruning technology to generate 100 words length summary. This method can be language independent theoretically. Thus we also tried it for the multiling task.

3 Hierarchical Latent Dirichlet Allocation (hLDA) Topic Model

Learning a topic hierarchy from data is an important instance of modeling challenge. Given a collection of "sentences", each of which contains a set of "words", we wish to discover common usage patterns or "topics" in the sentences, and to organize these topics into a hierarchy. We use Hierarchical Latent Dirichlet Allocation model, an efficient statistical method for constructing such a hierarchy which allows it to grow and change as the data accumulate.

In the approach, each node in the hierarchy is associated with a topic, where a topic is a distribution across words. A sentence is generated by choosing a path from the root to a leaf, repeatedly sampling topics along that path, and sampling the words from the selected topics. Note that all sentences share the topic distribution associated with the root node. Thus the organization of topics into a hierarchy aims to capture the breadth of usage of topics across the corpus reflecting underlying syntactic and semantic notions of generality and specificity.

The structure of tree is learnt along with the topics using a nested Chinese restaurant process (nCRP), which is used as a prior. The nCRP is a stochastic process, which assigns probability distributions to infinitely branching and infinitely deep trees. In our model, nCRP specifies a distribution of words into paths in an L-level tree. The assignments of sentences to paths are sampled sequentially: The first sentence takes the initial L-level path, starting with a single branch tree. Later, m_{th} subsequent sentence is assigned to a path drawn from the distribution:

$$p(path_{old}, c \mid m, m_c) = \frac{m_c}{\gamma + m - 1}$$

$$p(path_{new}, c \mid m, m_c) = \frac{\gamma}{\gamma + m - 1} \quad (1)$$

$path_{old}$ and $path_{new}$ represent an existing and novel path consecutively, m_c is the number of previous sentences assigned to path c, m is the total number of sentences seen so far, and γ is a hyper-parameter which controls the probability of creating new paths. Based on this probability each node can branch out a different number of child nodes proportional to. Small values of γ suppress the number of branches.

The following is the generative process for hLDA in our system:

(1) For each topic k ∈ T, sample a distribution β_k~Dirichlet(η)

(2) For each sentence s ∈ {S},

(a) draw a path c_s~ nCRP(γ)

(b) Sample L-vector θs mixing weights from θs ~ Dirichlet(α)

(c) For each word n, choose:

(i) $z_{s,n}|\theta s \sim Mult(\theta s)$,

(ii) word $w_{s,n}|\{z_{s,n}, c_s, \beta\}$

In this section, we describe a Gibbs sampling algorithm for sampling from the posterior nested CRP and corresponding topic distributions in the hLDA model. The Gibbs sampler provides a clean method of simultaneously exploring the parameter space and the model space. The variables needed by the sampling algorithm are: $w_{m,n}$, the n_{th} word in the m_{th} sententce. $c_{m,l,}$ the restaurant corresponding to the l_{th} topic distribution in sentence m; and $z_{m,n}$, the assignment of the n_{th} word in the m_{th} sentence to one of the L available topics. All other variables in the model-θ and β are integrated out. The Gibbs sampler thus assesses the values of $z_{m,n}$ and $c_{m,l}$.

The conditional distribution for c_m, the L topics associated with document m, is:

$$p(c_m \mid w, c_{-m}, z) \propto p(w_m \mid c, w_{-m}, z)\, p(c_m \mid c_{-m}) \tag{2}$$

where w_{-m} and c_{-m} denote the w and c variables for all sentences other than m. This expression is an instance of Bayes' rule with as the likelihood of the data given a particular choice of c_m and as the prior on c_m implied by the nested CRP. The likelihood is obtained by integrating over the parameters β, which gives:

$$p(w_m|c,w_{-m},z)=\prod_{l=1}^{L}\left(\frac{\Gamma(n^{(.)}_{c_{m,l},-m}+W\eta)}{\prod_w\Gamma(n^{(w)}_{c_{m,l},-m}+\eta)}\frac{\prod_w\Gamma(n^{(w)}_{c_{m,l},-m}+n^{(w)}_{c_{m,l},m}+\eta)}{\Gamma(n^{(.)}_{c_{m,l},-m}+n^{(.)}_{c_{m,l},m}+W\eta)}\right) \tag{3}$$

Where $n^{(w)}_{c_{m,l},-m}$ is the number of instances of word w that have been assigned to the topic indexed by $c_{m,l}$, not including those in the current sentence, W is the total vocabulary size, and (.) denotes the standard gamma function. When c contains a previously unvisited restaurant, is zero. Note that the c_m must be drawn as a block. Please refer to (D. Blei et al, 2003 and D.Blei et al, 2009) for more detailed description on Hierarchical topic model.

4 Our Proposed Method

4.1 Overview of our multi-document summarization

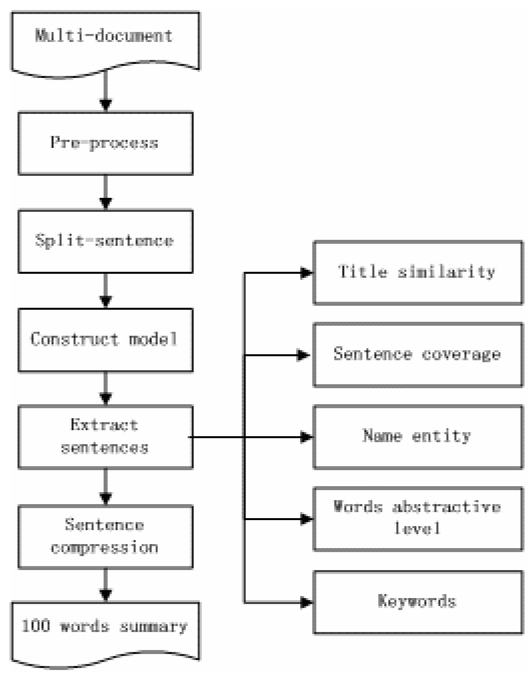

Figure 1. Overview of our multi-document summarization

4.2 Pre-processing of the data set

In our multi-document summarization system, we use single sentence as the basic processing unit and construct the model on sentence level. Firstly, we deal with the document corpus to conform to the needs of modeling. We use Beijing Foreign Studies University's automated tools to split English sentences, then remove stopwords, punctuations and other auxiliary modifiers. At the same time, the document is filtered by removing the sentences that contain quotation and the length of which is less than 10 words to improve the summary results. Because sentences that people said are not suitable for summary sentence, and shorter sentences usually contain a small amount of information relatively, so we choose to filter them in order to reduce their impact on final summary.

To generate a summary, we take full advantages of the hierarchical topic model, which includes the path information and the level of word abstractive information, and also added some features for weighting these sentences.

The hierarchical topic model assigns probability distributions to infinitely branching and infinitely deep trees. Our summarization system constructs a hierarchical tree structure of candidate sentences by positioning all sentences on a three-level tree. Each sentence is represented by a path in the tree, and each path can be shared by many sentences. The assumption is that sentences sharing the same path should be more similar to each other because they share the same topics. Moreover, if a path includes a title sentence, then candidate sentences on that path are more likely to be included in the generated summary.

4.3 Sentence Weighting

In the sentence extraction, the first feature is the similarity of title sentences with all candidate sentences. Title is an intention or an essence for one article, it's representative and indicative. In this system, we add the title sentence information. In one path if there is the title sentence, then all sentences assigned to that path have the value of s_t set to 1, otherwise they are 0. We will give priority to selecting sentences in that path to generate summary.

Secondly, we consider the number of sentences that assigned on each path, sorting these paths according to the number of sentences that they contain. We usually think that one path that contains more sentences is considered to be the main topic or hot topic, so we are apt to choose sentences in these paths.

Another feature we considered is the number of named entity that one sentence contains. We use Stanford University's named entity recognition toolkit. It can identify person names, address and institutional names in each sentence. If one sentence contains these particular entity names, it will have a high priority to be chosen as candidate summary sentence. S_n is the number of named entity categories in one sentence. For example, if one sentence has only one people name, then the s_n is set to 1; else if it also has address information, then s_n is set to 2.

The most important feature for weighting sentences is the frequency of word in sentences. We considered sentence coverage of each word in one sentence. The weight of the sentence is calculated by the following formula.

$$S_{tf} = \frac{\sum_{i=1}^{n} \frac{num_s(w_{ij})}{|si|}}{n} \tag{4}$$

W_{ij} is the j_{th} word in s_i, $num_s(w_{ij})$ is the number of sentences that w_{ij} covered, | s_i | is the length of the sentence, and n is the total number of all sentences.

In order to emphasize those required aspects of different topic categories by TAC 2011, we set higher weight for sentences containing more keywords in the knowledge base. We build a knowledge base for every required aspect listed in the guided summarization. We extract keywords from TAC sample document sets for all the five topics which contain tagged information units for the required aspects. We obtain an original version of the knowledge base for each topic with many required aspects and each aspect corresponds to a keyword list. Then we expand these keywords with thesaurus of Britannica Online Encyclopedia (http://www.britannica.com/bps/thesaurus?query=good) using a simple meta search engine. We expect that the knowledge base will help to select sentences which tightly cover the required aspects.

$$S_k = \frac{num(W_k)}{|s|} \tag{5}$$

s_k is the score of the keywords, $num(w_k)$ is the number of keywords that a sentence contain. |s| is the length of the sentence.

Hierarchical model can mine out abstract and specific topics, which will help identify candidate summary sentence. For a short summary, it should be more abstractive. Generally sentences are expressed through words; we evaluate the abstractive extent of the sentence by the following formula.

$$S_{abs} = \frac{num(W_{abs})}{|s|} \tag{6}$$

$num(W_{abs})$ is the number of abstractive words in one sentence. In our experiment, we use the level zero and level one, |s| is the length of the sentence.

At last, we calculate the sentence score by the following formula:

$$S = a*S_t + b*S_n + c*S_{tf} + d*S_k + e*S_{abs} \tag{7}$$

We use a linear additive manner, which Sabs is the score of sentence abstractive extent. St is the score of the similarity between the sentence and title. Sn is the score of the sentence containing named entities. In the experiment the parameters a,b,c,d,e are fixed at {0.3,0.1,2,1,1}.

Lastly, we consider the number of sentences that assigned on each path to select the number of candidate summary sentences. We look on each path as a cluster. If the number of the clusters is more than 10, we extract one sentence from each cluster in the top ten paths; else if it's less than 5, we choose two sentences from each cluster; otherwise, we choose one as candidate sentence to generate the summary.

4.4 Similarity Calculation

As candidate summary sentences are chosen from different articles, and may contain some duplicate information. So we work on the method that based on sentence similarity calculation to remove redundant information. If the similarity between a candidate sentence to be added to the summary and the sentences that have been added to the summary exceeds a certain threshold, we'll give up the sentence and re-select another sentence to compose the summary. The similarity of the two sentences is calculated as follows:

$$sim(s1, s2) = \frac{num(s1 \cap s2)}{1/2(|s1| + |s2|)} \tag{8}$$

Molecule num(s1∩s2) is the number of the same words that s1 and s2 contain, |s1|, |s2| represent the length of sentence s1, s2 respectively. We use the average value of |s1| and |s2| to balance sentences of different length.

4.5 Summary Sentence Compression

According to TAC 2011 Multi-document Summarization evaluation task, the goal is to generate a brief, fluent, coherent and readable 100 words length summary. To make our method fitful for TAC 2011's systems, we also use sentence compression techniques (Liu Hongyan et al,2010) that we participated in TAC 2010, which try to remove those unimportant or redundant elements in the sentence and retain important information to make the summary coherent, concise and readable.

4.6 Experiments and Discussions

We use the latest TAC 2011 multi-document summarization data sets to test our proposed method empirically. TAC is held by NIST (http://www.nist.gov). The basic objective of TAC 2011 is to guide researchers experiment on large-scale public data sets to promote the development of multi-document summarization technology. The data set consists of 46 topics, and each topic contains 10 articles related to that topic. It is designed for automatic multi-document summarization evaluation. TAC 2011 summarization task is creating a summary of no more than 100 words for each topic.

There are two kinds of summary evaluation methods: one is the manual evaluation, which is generated by the experts to rate the separate summary of the pros and cons of the various aspects; the other is the machine evaluation, the methods used now are mainly Recall-Oriented Understudy for Gisting Evaluation (ROUGE) metrics which is widely used by TAC for performance evaluation.

ROUGE-N is an n-gram recall computed as follows.

$$C_n = \frac{\sum_{C \in \{Model\_Units\}, n-gram \in C} \sum Count_{match}(n-gram)}{\sum_{C \in \{Model\_Units\}, n-gram \in C} \sum Count(n-gram)} \tag{9}$$

Where, n is the length of the n-gram, and Count match(n-gram) is the maximum number of n-grams co-occuring in an automatic machine summary and the human summaries, Count (n-gram)is the number of n-grams in human summaries. ROUGE-L uses the longest common subsequence (LCS) statistics, while ROUGE-W is based on weighted LCS and ROUGE-SU is based on skip-bigram plus unigram. Each of these evaluation methods in ROUGE can generate three scores (recall, precision and F-measure). As we have similar conclusions in terms of any of the three scores, for simplicity, in this paper, we only report the average F-measure scores generated by ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-4, ROUGE-L, ROUGE-W and ROUGE-SU to compare our proposed method with other implemented systems. TAC 2011 also used baseline1 and baseline2 as two comparative systems.

Baseline 1: leading sentences (up to 100 words) from the most recent document

Baseline 2: summary generated by publicly available summarizer MEAD with default settings

In our hLDA experimental system, we fixed the tree depth at 3. In the guided summarization track of TAC 2011 we submit two runs. The first one is mainly considering the feature of keyword coverage, and the second one is mainly considering the feature of aspect coverage. We compare our two summarization results with the other 48 runs, which contain 46 runs from 21 participants and two baseline runs.

We analyzed the data of evaluation results from three aspects. Firstly, we try to take a detailed analysis for the peer score tables with both initial summaries and update summaries. To compare the score of modified(pyramid) evaluation with the other peers, we selected both the first-five document sets which have a good performance and the last-five document sets which oppositely show a poor performance in our system. As showing in table1 and table2, the first column gives the rank of our system among all the peers with the indicator of the pyramid scores evaluation for each selected document sets.

With this method of comparison, we attempted to find the relationship between our system performance and the predefined category of topics corresponding to each document set. And the analysis results revealed that there is no obvious connection between the topics and performance.

Performance \ Description	Rank among all peers	setID	peerID	pyramid score
Five Good Performance	1	D1109-A	5	0.455
	2	D1113-A	5	0.364
	1	D1119-A	14	0.564
	5	D1132-A	14	0.529
	5	D1136-A	5	0.396
Five Poor Performance	48	D1106-A	14	0.118
	42	D1108-A	14	0.475
	39	D1111-A	5	0.340
	48	D1115-A	5	0.088
	41	D1129-A	5	0.154

Table1 Performance Through Initial Summaries

Performance \ Description	Rank among all peers	setID	peerID	pyramid score
Five Good Performance	3	D1117-B	5	0.478
	1	D1119-B	5	0.545
	1	D1123-B	5	0.571
	5	D1127-B	14	0.259
	2	D1133-B	5	0.667
Five Poor Performance	45	D1104-B	14	0.043
	38	D1135-B	5	0.091
	42	D1114-B	14	0.000
	39	D1130-B	14	0.185
	44	D1142-B	5	0.179

Table2 performance through update summaries

Secondly, we pay attention to the average scores per category per automatic summarizer for both initial summaries and update summaries. As depicted in figure 2 and figure 3, we chose one criteria of assessment for each category, and also for each run, which has a relatively high ranking compared to the other evaluation results of our system. From which we can see that there still exists a great space for improvement with the average scores of our system performance in each category.

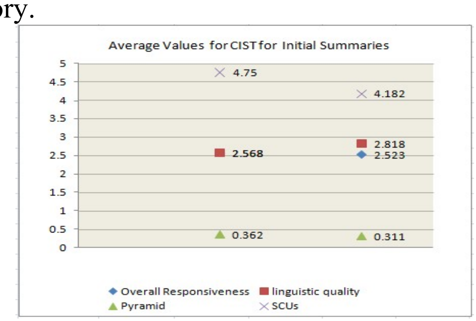

Figure 2. average values for initial summaries

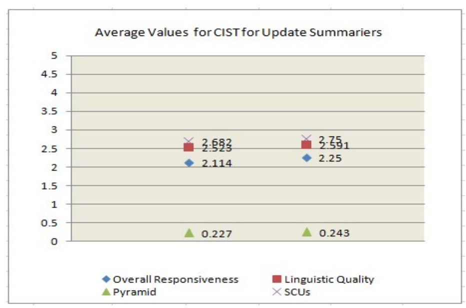

Figure 3. average values for update summaries

Finally, we analyzed the average scores per automatic summarizer for initial and update summaries. As displayed in figure 4 and figure 5, horizontal axis shows categories and vertical axis shows the ranking.

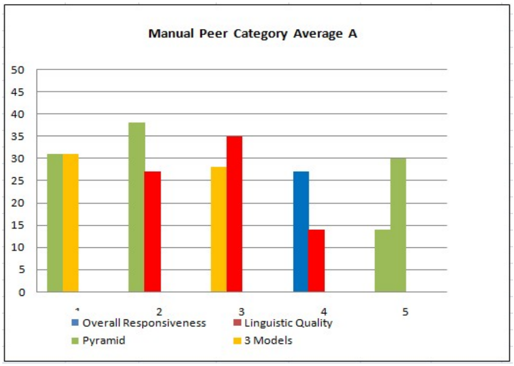

Figure 4. Manual Peer Category Average A

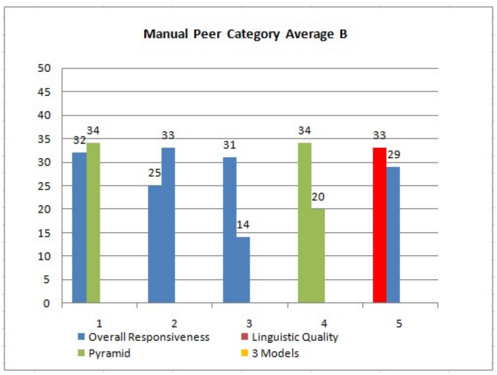

Figure 5. Manual Peer Category Average B

Furthermore，we also used the MultiLing Pilot in TAC 2011 Dataset to test the performance of our system. To adjust the system for multi-lingual environment, we used the frame of the hLDA model again but deleting those knowledge base for English. There are 7 languages in the source documents, which respectively are Arabic， Czech， English, French, Greek, Hebrew and Hindi. 700 files are contained in the dataset, 100 for each of the language. We tried to solve the problem of character coding and worked out all the 7 summaries. The evaluation results consist of AutomaticEvaluation and HumanEvaluation. AutomaticEvaluation Contains the evaluation results of AutoSummENG and ROUGE (LinChin-Yew， Edward Hovy,2003). HumanEvaluation Contains the evaluation results of human judges per language. In the 10 participants, we refer to the 2 baseline systems and rank them according to the grades. As depicted in Figure 6, our grades are excellent. There are 10 topics in each language. We count numbers of topics ranking in top three per language.

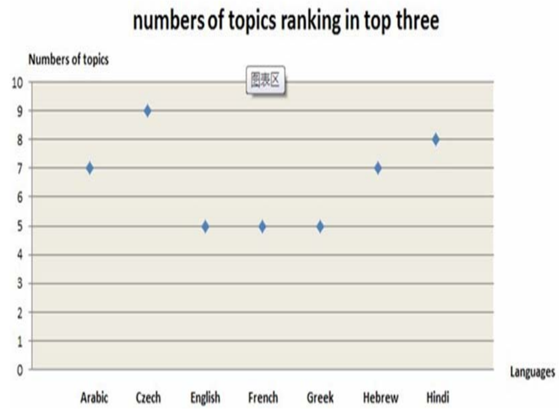

Figure 6. numbers of topics ranking in top three

5 Conclusions

This report introduces the details of a multi-document summarization system in TAC 2011 for both initial and update summaries. We describe the details of each step. In conclusion, our proposed multi-document summarization method, which is based on hierarchical Latent Dirichlet Allocation topic model, can improve the quality of the summary to a certain extent compared with our own system based on sentence clustering in TAC 2010. But compared with other systems, we still need to work hard.

There is good news for our new proposed method this year coming from the newly appeared MultiLing Pilot in TAC 2011. We can have more confidence that hLDA model can catch some good gist information for various topics in more than one languages without any added knowledge base designed for a special language. Next, we need to study more about its meaning combined with a language and more usage of the model. For instance, in our hierarchical model, we found out that some similar sentences will be assigned to different paths, and this is not we expected. We will go on to improve our model, and enhance the accuracy of it. At the same time, our sentence pruning method will remove some important content sometimes, we will also continue to research on sentence extraction and sentence compression method.

References

Asli Celikyilmaz. Dilek Hakkani-Tur. A hybrid hierarchical model for multi-document summarization[J].Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 815–824, Uppsala, Sweden, 11-16 July 2010.

D. M. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation[J]. In Jrnl. Machine Learning Research, 3:993-1022, 2003b.

D. Blei, T. Griffiths, M. Jordan, and J. Tenenbaum. Hierarchical topic models and the nested Chinese restaurant process[J]. In Neural Information Processing Systems [NIPS], 2003a.

D. Blei, T. Griffiths, and M. Jordan. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies[J]. In Journal of ACM, 2009.

H. Liu, Q. Zhao, Y. Xiong, L. Li, C. Yuan. The CIST Summarization System at TAC 2010. http://www.nist.gov/tac/publications/2010/participant.papers/CIST.proceedings.pdf

LinChin-Yew， Edward Hovy. Automatic Evaluation of Summaries Using N-gram Co-occurrences Statistics[J]. In Proceedings of 2003

R. Arora and R. Balaraman. Latent dirichlet allocation based multi-document summarization. Proceedings of the Second Workshop on Analysis for Noisy Unstructured Data AND,2008.

Rachit Arora and Balaraman Ravindran. Latent dirichl et allocation and singular value decomposition based multi-document summarization[J]. In ICDM'08: Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, pages 713-718, Washington, DC, USA, 2008.

S. Deerwester , S. Dumais , T. Landauer , G. Furnas , and R. Harshman. Indexing by latent semantic analysis[J] . Journal of the American Society of Information Science , 1990 , 41 (6) :3912407.

T. Hofmann. Probabilistic latent semantic indexing[C] SIGIR , 1999.

T. Griffiths and M. Steyvers. A probabilistic approach to semantic representation[J]. In Proceed-ings of the 24th Annual Conference of the Cognitive Science Society, 2002.

Wu Xiaofeng, Zong Chengqing. An Approach to Automatic Summarization by Integrating Latent Dirichlet Allocation in Conditional Random Field[J]. Journal of Chinese Information Processing. Vol.23, No.6.Nov.2009

Yang Xiao, Ma Jun, Yang Tongfeng. Automatic multi-document summarization based on the latent Dirichlet topic allocation model[J]. CAAI Transactions on Intelligent Systems. Vol 51.2 Apr.2010.

Zhang Yanxing, Zhang Ming, Deng Zhihong. Feature-Driven Summarization of Customer Reviews[J]. Journal of Computer Research and Development. 46(Suppl.):520-525.2009

Zhong Minghui, Wang Hongling,Zhou Guodong.Chinese multi-document summarization based on LDA Topic-Oriented method[J].The Fifth National Youth Conference on Computational Linguistics. Wuhan, China.2010.