PAPER

A Hybrid Topic Model for Multi-Document Summarization

JinAn XU[†a)], **JiangMing LIU**[†], ***Nonmembers, and*** **Kenji ARAKI**[††], ***Member***

SUMMARY Topic features are useful in improving text summarization. However, independency among topics is a strong restriction on most topic models, and alleviating this restriction can deeply capture text structure. This paper proposes a hybrid topic model to generate multi-document summaries using a combination of the Hidden Topic Markov Model (HTMM), the surface texture model and the topic transition model. Based on the topic transition model, regular topic transition probability is used during generating summary. This approach eliminates the topic independence assumption in the Latent Dirichlet Allocation (LDA) model. Meanwhile, the results of experiments show the advantage of the combination of the three kinds of models. This paper includes alleviating topic independency, and integrating surface texture and shallow semantic in documents to improve summarization. In short, this paper attempts to realize an advanced summarization system.

key words: *multi-document summarization, hybrid topic model, hidden topic Markov model (HTMM), surface texture model, topic transition model*

1. Introduction

Summarization is the process of extracting recapitulative information from numerous documents. The summary describes the main content of the documents. The main content of numerous documents can be quickly obtained through reading the summary, which can filter out redundant information and improve reading efficiency.

Two mainstream summarization approaches are extractive summarization [1] and generative summarization [2], [3]. The difference between the two kinds of summarization is that the summary of the former is extracted from documents and that of the latter is generated based on certain semantic representations. General methods of generating summarization are lacking in domain adaptation. The vast majority of extractive multi-document summary systems focus on identifying the importance of sentences exploiting unsupervised or supervised learning techniques. In supervised methods, summarization can be regarded as a classification task [4]–[6], which identifies the sentences belonging to a summary class. However, the necessity for a large number of annotated corpora is a major weakness. The disadvantage of supervised learning in the case of probabilistic methods is that they require a large amount of labelled training and test data, and it is not usually available due to properties of domain sensitivity.

In unsupervised methods, feature-based ranking algorithms have been widely used. Many document features have been proven to be useful [7], [8]. Recently, many summarization works are motivated to focus on user interest features [9], [10]. However, due to the sheer volume of varied information requiring updates, shallow semantic topic analysis is necessary and useful in the long run. Shallow semantic analysis brings about direction of summarization [11]. The topic is a useful shallow semantic feature [12], [13]. Under the unsupervised frame, this paper proposes a hybrid topic model, which combines the topic, surface texture, and topic transition models.

The topic model can effectively represent latent semantic information in documents. Topic can be described in the form of words probability distribution. Recently, many useful topic models have been presented to capture document information structures. Topic model is primarily originated from Latent Semantic Indexing (LSI). Based on LSI, probabilistic Latent Semantic Indexing (pLSI) is presented, which is regarded as a real topic model. Latent Dirichlet Allocation (LDA) topic model is presented by Blei et al. in a much general form [14]. After that, many extended models based on LDA-style are proposed using extra related information about documents [15]–[18]. It does not matter whether it is in LDA or extended LDA-style models; topic independence assumption limits the ability of topic representation and the structure of given documents is ignored.

To overcome these problems, Hidden Topic Markov Model (HTMM) is presented with an assumption that topics between sentences satisfy the Markov property [19]. Topic transition in HTMM is modeled as a binary relation at the sentence level. To regularize topic transition probability, structure Topic Model (strTM) is proposed [20]. However, strTM mainly focuses on topic transition and it does not directly model topic over the document.

A good summarization system should be able to identify the point content in an article and generate a summary that has good coverage of the ideas expressed in the article [21], [22]. For one sentence, there are many different topics assigned with different probabilities. Take the sentence "*The national funds support many different kinds of work in the field of mechanical engineering*" as an example, where it can be expressed by many topics with respective likelihoods, specifically economics, engineering and so on.

Manuscript received July 4, 2014.
Manuscript revised January 19, 2015.
Manuscript publicized February 9, 2015.

[†] The authors are with Beijing Jiaotong University, Haidian District, Beijing, 100044 China.

[††] The author is with Hokkaido University, Sapporo-shi, 060–0808 Japan.

a) E-mail: jaxu@bjtu.edu.cn

DOI: 10.1587/transinf.2014EDP7229

Suppose economics is more likely to be in one sentence, but not more likely in the whole set of documents which mainly describe engineering. Generally speaking, topic A can be a foreshadowing topic from which to draw topic B, which is the main topic in document sets. However, it is confusing in the sentence-level topic model.

To capture this feature, HTMM and regular topic transition model are mutual complementary. On the one hand, HTMM captures the topic distributions over sentences and the document. It means that we can select summary sentences by comparing topic distribution over sentences with that of the topic distribution over the document. On the other hand, regular topic transition model captures the main point topic using a graph-based algorithm. In our hybrid topic model, there are two main steps. One is extracting sentences with high information coverage through a combination of HTMM and surface texture model (Sect. 3). The other is re-ranking candidate sentences using the main topic expression (Sect. 4). The superiority of this approach is based on the valid combination of shallow semantic features and surface features, effectively using HTMM with regular topic transition to improve summary quality.

2. Summarization System

This paper proposes a summarization system with three main components and two actions as shown in Fig. 1. HTMM, Surface Texture Model and Topic Transition Model are implemented as components. Combination and Re-Ranking are implemented as actions.

In the HTMM component, topic distribution over document and topic distribution over sentence can be achieved. In the surface texture model component, surface features will be extracted and their respective scores will be computed. In the topic transition model component, the regular topic transition probability will be captured by the Markov chains, and a graph of topic transition will be generated. In the combination action, calculation of the summary candidate score consists of the computation of topic features (from the HTMM component) and the computation of surface features (from the surface texture model component). In the re-ranking action, candidates are then re-ranked to generate a final summary.

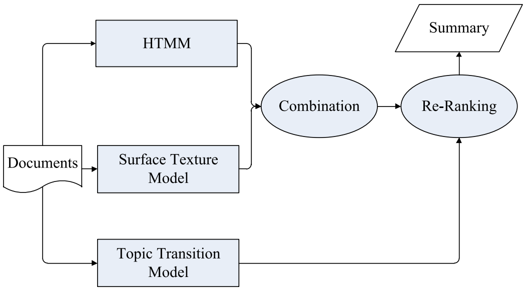

Fig. 1 The framework of proposed method.

3. Combination of HTMM and Surface Texture Model

LDA is a complete document generation model as shown in Fig. 2. In the process of generating the document, there are two assumptions in LDA: topic independence and the "bag-of-words" (ignoring the order of words in the document). HTMM loosely eliminates the topic independence assumption through modeling topic transitions as binary relations as shown in Fig. 3. It is clear that HTMM is a kind of extended LDA-style model; but topic assignment in the sentence level is particular. The main difference with LDA in the process of generating a document is importing extra parameters, which are used to identify whether or not to remain in the topic from the previous sentence or to generate a new one.

Due to its properties, the summary should cover different document contents, and documents points at different levels, so the summary has similar topic distribution to the document. This way, the similarity of topic distribution between each sentence and document becomes a main shallow semantic feature, which is calculated using the criterion of the Kullback-Leibler divergence (KL). On the one hand, topic distribution over sentences is computed through the following formula (1).

$$p(Z|S) = \frac{\sum_{w \in S} p(w|Z) \times p(Z|D)}{len(S)} \tag{1}$$

Where, Z denotes a given topic, w denotes a word in sentence S, which document D is composed of, $p(Z|D)$ comes

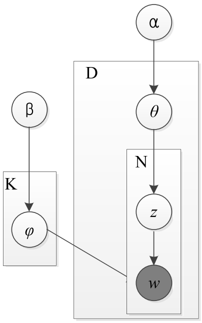

Fig. 2 LDA graphic structure.

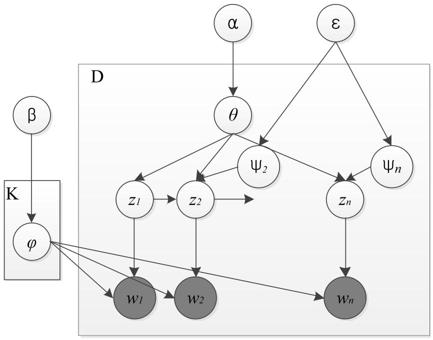

Fig. 3 HTMM graphic structure.

from φ, $p(w\,|\,Z)$ is drawn from θ, $len(S)$ denotes the number of words in sentence S, which is a penalty for long sentences. On the other hand, topic distribution over a given document is drawn from θ. φ and θ are estimated in HTMM, according to the work of (Gruber et al., 2007). Consider symmetry, we use average Kullback-Leibler divergence (avgKL) to compute distribution difference using formula (2).

$$D_{KL}(S\,\|\,D) = \sum\nolimits_i\left(P(i)\times\log\frac{P(i)}{Q(i)}\right)$$

$$D_{avgKL}(S\,\|\,D) = \frac{D_{KL}(P\|Q) + D_{KL}(Q\|P)}{2} \tag{2}$$

Where, P is the topic distribution over a sentence S, $p(Z|S)$ and Q is the topic distribution over documents D, $p(Z|D)$.

The surface texture model is represented by a tuple (S, D, F, λ), where S is a set of sentences in documents D and F is a set of surface feature functions, each of which returns a value ranging from 0 to 1 inclusive and λ is the weight of each feature ($|\lambda| = |F|$). The discriminative score on combination of HTMM and surface texture model is computed through the following formula (3).

$$score(S, D) = \sum\nolimits_i \lambda_i F_i(S, D) - D_{avgKL}(S\|D) \tag{3}$$

We choose a set of sentences with a high score as the candidate summary. If 250 words summary should be generated, we will generate double length (500 words) of sentences as candidate summary for the following components.

4. Topic Transition Model and Re-Ranking

HTMM models topic transitions as a simple binary relation and it hardly captures the complex structure. In order to generally represent topic relation, a topic transition model is proposed to capture probability transition between topics. As a result, the model with regular topic transitions contributes to capturing the main point in various documents.

In the topic transition model, the topic is a hidden state in a hidden Markov chain and observations are sentences as shown in Fig. 4. For certain topics Z, their word distribution $P(w\,|\,Z)$ is provided by previous component (HTMM). Assuming word independency in one sentence, the distribution of sentences given a certain topic is calculated using formula (4).

$$P(S\,|\,\theta_z, z) = \prod_{j=0}^{N-1} P(w_j\,|\,\theta_z, z) \tag{4}$$

Where, S denotes a sentence with N words, θ is the word distribution over topics in the HTMM component, θ_z is the word distribution given topic z, w_j is a word in S and $p(w\,|\,\theta_z, z)$ is drawn from θ_z. The probability of a sentence given a topic is obtained (hidden state emission probability), Regular probability of topic transition (hidden state transition probability) can be achieved by a forward-backward algorithm according to sentences (observation state) (Elliott et al., 1995). Each sentence respectively has a main topic, and therefore the relation between neighbouring sentences can be represented in the form of a topic transition. In other words, the structure of documents can be described by topic transition probability, as shown in Fig. 5. The edges with low transition probability are deleted (lower than 0.1). Zn denotes topic and it clear shows that $\{Z_1, Z_3, Z_4, Z_6\}$ is a cluster and $\{Z_2, Z_5\}$ is another one. This paper only uses the main topic in this graph.

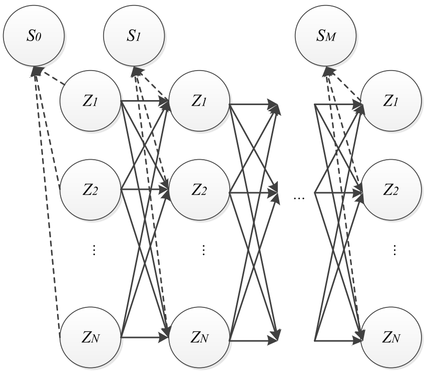

Fig. 4 Hidden Markov model framework for topic transition.

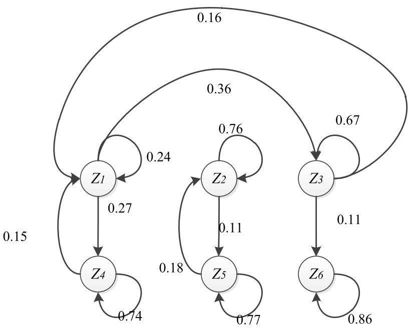

Fig. 5 An example of topic transition graph.

If a topic is transferred from other topics with a high probability, the topic is naturally one of the main topics in the documents. We use formula (5) to find the main topic.

$$\overline{z} = \operatorname*{argmax}_{z} \sum_{z_i \in Z} p(z|z_i) \tag{5}$$

Where, Z is a set of topics, z_i denote i^{th} topic and $p(z|z_i)$ is the topic transition probability achieved by the topic transition model. $\overline{z}$ is the main topic according to the topic transition graph.

In Re-Ranking action, sentences assignment over main topic is used to reassign the candidate summary sentences with scores, which can be calculated using formula (6).

$$score(S\,|\,\theta, \overline{z}) = \prod_{j=0}^{N-1} p(w_j\,|\,\theta_{\overline{z}}, \overline{z}) \tag{6}$$

5. Experiments

5.1 Data and Evaluation Method

The experimental corpus is provided by DUC 2007[†]. The corpus includes 45 document sets and in each document set, there are 25 documents with similar topics and 4 manual professional summaries of 250 words. We use the first document set (developing set) for parameter optimization and the rest of the document sets for testing. According to the corpus, we should also generate a summary of no more than 250 words for each document set.

ROUGE, an automatic evaluation metric [23], will be used to compare machine summary generated by the system with the manual professional summary. ROUGE is based on recall rate and this paper uses ROUGE values to evaluate the quality of summaries generated by the summarization system. ROUGE-n represents co-occurrence statistics based on n-gram. ROUGE-SU denotes skip-bigram plus unigram-based co-occurrence statistics.

5.2 Parameter Setting

In our experiments, parameters are divided into two groups, hyper parameters in topic model and surface feature weights in surface texture model. Hyper parameters are manually set, α is set to $\frac{50}{K}$, where K is the number of topics; β is set to the empirical value of 0.01 in HTMM. Through preliminary experiments, the number of topics is to be set to 60 with highest ROUGE-SU4 value in the developing data set. Moreover, we use 6 feature functions in the surface texture model as shown in Table 1, with feature values normalized into [0, 1], and the greedy algorithm is used on developing set to achieve optimal feature weights in surface texture model.

5.3 Experimental Results

We take the LDA based, surface texture model, and HTMM based systems as our baseline model. Meanwhile, we implement two types of system in our system framework: the one with only combination action and the other with combination and re-ranking actions.

The experimental results are shown in Table 2, ROUGE-unigram (R-1), ROUGE-bigram (R-2) and ROUGE-SU4 (R-SU4) values are presented in different systems. "*" denotes the best results among these systems ($p < 0.1$). Surface denotes when the system only applied the surface texture model component. HTMM denotes when the system only applied the HTMM component. Hybrid-TM #1 denotes the system applied the combination component of HTMM and surface texture model. Hybrid-TM #2 denotes the system applied all the components described in this paper.

Table 1 Features in surface texture model with their description.

Feature	Description
F_{SP}	the sentence position index in document
F_{SL}	the sentence length penalty
F_{ST}	the similarity between sentence and title
F_{NE}	the proportion of named entity in sentence
F_{WP}	the word position index in sentence
F_{WF}	the word frequency (TF-IDF)

Table 2 Experimental results.

System	R-1 (%)	R-2 (%)	R-SU4 (%)
LDA	35.7	8.9	12.1
Surface	32.3	5.0	10.1
HTMM	37.4	7.5	12.8
Hybrid-TM #1	37.6	8.5	13.4
Hybrid-TM #2	38.1*	8.9*	13.7*

Table 2 shows that our completed system outperforms LDA based systems (+2.4 on ROUGE-1 and +1.6 on ROUGE-SU4), outperforms the solo surface texture model (+5.8 on ROUGE-1 and +3.6 on ROUGE-SU4) and outperforms the solo HTMM (+0.7 on ROUGE-1 and +0.9 on ROUGE-SU4). Furthermore, our system with Re-Ranking step included outperforms our system without this stage included (+0.5 on ROUGE-1 and +0.3 on ROUGE-SU4).

In addition, this paper uses the SCU-marked summaries provided by DUC2007 to investigate feature functions used in the surface texture model. For one sentence in document, SCU provide the number of participants in the conference using this sentence for summary. So this paper divides all the sentences in corpus into summary and non-summary using formula (7).

$$\begin{cases} S \in \mathit{summaries} & \mathit{if}\ \mathit{count}(S) \geq \delta \\ S \in \mathit{non\text{–}summaries} & \mathit{if}\ \mathit{count}(S) < \delta \end{cases} \tag{7}$$

Where, $\mathit{count}(S)$ denotes the number of participants in the evaluation task using sentence S for summary and δ is set to 3 in all experiments.

The sum of all feature values is different between summary and non-summary as seen in Fig. 6. The horizontal axis represents different features and the vertical axis expresses content feature score. In summary, feature value of title similarity is nearly twice as high as non-summary; scores of sentence position, named entity and word frequency are much higher than non-summary. These features are also widely used. However, word-level features are influenced by the awareness and habits of authors, and therefore the information provided by word-level features is mixed. So to highlight the superiority of word-level features, we plan to consider relativeness and synonymy between words in our future work.

5.4 Analysis

In this subsection, we will present summary examples produced by our hybrid-TM about *Line-item veto (LIV)*, and the

[†] Http://duc.nist.gov/guidelines/2007.html

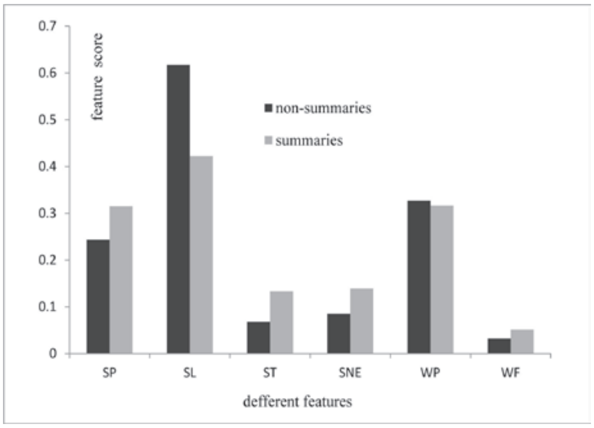

Fig. 6 Different features value in summaries and non-summaries.

professional summary on the same topic.

Produced summary

(1) The U.S. Justice Department Friday appealed the Supreme Court a federal judge's decision striking down the presidential line-item veto. The Supreme Court struck down as unconstitutional a law giving the president a line-item veto, which let him cancel specific items in tax and spending measures. ...... The Supreme Court agreed on Friday to hear argument and decide the constitutionality of the president line-item veto. A U.S. district judge Thursday declared the president's line-item veto authority unconstitutional. (3) Following are excerpts from Thursday's Supreme Court decision striking down the presidential line-item veto. The United States House of Representatives today voted 352-64 to reject President Bill Clinton's line-item veto of 38 military construction programs in fiscal 1998. ...... (2) "The line-item veto act was a vital force in restoring the appropriate balance of power, and eliminating wasteful, unnecessary pork-barrel spending," said Senator John McCain, Republican of Arizona.

Professional summary

(1) The line-item veto (LIV) has been sought by nearly every president this century as a tool to limit pork barrel spending which is traditionally reviled as the most cynically deployed and least utilitarian form of largess. (2) The 1998 budget included $300,000 for enhancing the flavor of peanuts, $150,000 for peanut competitiveness and $250,000 for pickle research. President Clinton said the LIV is an important tool for striking unnecessary spending, for preserving the integrity of federal spending and enlivening the public debate over how to make the best use of public funds. The Solicitor General contended that the LIV represents a presidential exercise of spending authority delegated by Congress. 110 years ago, Lord Bryce said the LIV was "desired by enlightened men and would save the nation millions of dollars a year" .... (3) President Clinton used the authority to veto 82 items in 11 bills, including money for New York hospitals, a tax break for Idaho potato growers, 38 projects worth $287M in military construction, $144M from a defense spending bill and $30M for intercepting asteroids.......

Comparing summary examples with professional summary, it shows that sentences (1) and (2) have the same meaning, because they all describe similar information. On the other hand, both sentence (3) describe the same thing; however, the focus is different. In the produced summary, sentence (3) is the result of this event, however, in the expert summary, sentence (3) describes the detail of the event itself. Then, while topic is selected correctly, we determine that it is difficult for our system (all statistic-based systems) to decide whether or not to focus on the influence, results, or contents of given events. With intuition, more syntax analysis for sentences or the whole document sets is not trivial.

6. Conclusion and Future Work

We propose a hybrid topic model for multi-document summarization using two stages, i.e. combination and reranking. We extract summaries with high content coverage and a main topic captured. We conclude from the experimental results:
1. Structure information of the document is necessary for a summarization. HTMM performs better than LDA in topic distribution representation.
2. Surface texture model directly describes sentences. Combining surface features and hidden features (topics) is an advanced approach of summarizing.
3. Topic transition information describes the structure of documents at the sentence level, and the main topic in documents can be captured in graphic topic transitions.

Furthermore, we will exploit graph algorithms to deeply capture topic transition information.

Acknowledgments

The authors were supported by National Nature Science Foundation of China (Contract 61370130 and 61473294).

References

[1] R. Barzilay and L. Lee, "Catching the Drift: Probabilistic Content Models, with Applications to Generation and Summarization," HLT-NAACL, pp.113–120, 2004.
[2] T. Liu and K. Wang, "Four main methods in summarization," J. China Society For Scientific and Technical Information, vol.18, no.1, pp.11–19, 1999.
[3] Y. Gong and X. Liu, "Generic text summarization using relevance measure and latent semantic analysis," Proc. 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.19–25, 2001.
[4] Y. Pei, W. Yin, Q. Fan, and L. Huang, "A supervised aggregation frame-work for multi-document summarization," Proc. COLING 2012, pp.2225–2242, Mumbai, India, 2012.
[5] Z. Yang, K. Cai, J. Tang, L. Zhang, Z. Su, and J. Li, "Social context summarization," Proc. 34th International ACM SIGIR Conference on Research and Development in Information Retrieval

(SIGIR 2011), pp.255–264, ACM, New York, NY, 2011.

[6] O. You, W.J. Li, S. Li, and Q. Liu, "Applying regression models to query-focused multi-document summarization," Information Processing and Management, vol.47, no.2, pp.227–237, 2011.

[7] G. Erkan and D.R. Radev, "Lex-PageRank: prestige in multi-document text summarization," Proc. 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004), pp.365–371, 2004.

[8] R. Mihalcea and P. Tarau, "TextRank: bringing order into texts," Proc. 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004), pp.404–411, 2004.

[9] H. Daumé III and D. Marcu, "Bayesian query-focused summarization," Proc. 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics, pp.305–312, 2006.

[10] P. Hu, D. Ji, C. Teng, and Y. Guo, "Context-enhanced personalized social summarization," Proc. COLING 2012, pp.1223–1238, 2012.

[11] A. Celikyilmaz and D. Hakkani-Tür, "Discovery of topically coherent sentences for extractive summarization," Proc. 49th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, pp.491–499, 2011.

[12] A. Haghighi and L. Vanderwende, "Exploring content models for multi-document summarization," Proc. Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2009, pp.362–370, 2009.

[13] Y. Chang and J. Chien, "Latent Dirichlet learning for document summarization," Acoustics, Speech and Signal Processing. (ICASSP 2009), pp.1689–1692, 2009.

[14] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent Dirichlet allocation," J. Machine Learning Research, vol.3, pp.993–1022, 2003.

[15] J. Boyd-Graber and D.M. Blei, "Syntactic topic models," arXiv preprint arXiv:1002.4665, 2010.

[16] M. Steyvers, P. Smyth, M. RosenZvi, and T. Grifths, "Probabilistic author-topic models for information discovery," Proc. tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.306–315, 2004.

[17] D.M. Blei and J.D. Lafferty, "Dynamic topic models," Proc. 23rd International Conference on Machine Learning, pp.113–120, 2006.

[18] Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai, "Topic sentiment mixture: modeling facets and opinions in weblogs," Proc. 16th International Conference on WWW, pp.171–180, 2007.

[19] A. Gruber, M. RosenZvi, and Y. Weiss, "Hidden topic Markov models," Artificial Intelligence and Statistics (AISTATS), pp.163–170, 2007.

[20] H. Wang, D. Zhang, and C. Zhai, "Structural Topic Model for Latent Topical Structure Analysis," ACL, pp.1526–1535, 2011.

[21] D. Contractor, Y. Guo, and A. Korhonen, "Using Argumentative Zones for Extractive Summarization of Scientific Articles," COLING, pp.663–678, 2012.

[22] R.J. Elliott, J.B. Moore, and L. Aggoun, "Hidden Markov Model," Stochastic Model-ling and Applied Probability, vol.29, 1995.

[23] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," Proc. Workshop on Text Summarization Branches Out, post-conference workshop of ACL 2004, pp.74–81, Barcelona, Spain, 2004.

JinAn Xu is currently an Associate Professor in the School of Computer Science and information technology, Beijing Jiaotong University, Beijing, China. He received the B.Eng. degree from Beijing Jiaotong University in 1992, the MS degree and Ph.D. degree in computer information from Hokkaido University, Sapporo, Japan, in 2003 and 2006, respectively. His research focuses on natural language processing, machine translation, information retrieve, text mining, and machine learning. He is a member of CCF, CIPSC, ACL and the ACM.

JiangMing Liu is a MS student in the School of Computer information and technology at Beijing Jiaotong University, China. His research focuses on summarization and semantic analysis for machine translation.

Kenji Araki was born in 1959 in Otaru in Japan. He received his B.E. and Ph.D. degrees from Hokkaido University in 1982 and 1988. He was an associate professor there from 1991 to 1998, and a professor in 1998 at Hokkai-Gakuen University. He then joined Hokkaido University as an associate professor. He is currently a professor at the Graduate School of Information Science and Technology. His research interests include natural language processing, morphological analysis, machine translation, and speech dialogue processing. He is a member of IEICE, IPSJ, JSAI, JCSS, ACL, IEEE, and the AAAI.