A Maximum Entropy model for Automatic Summarization

Wei-Jiang Li[1] Zhen-Zhen Wang[1] Dong-Jun Li[2] Zheng-Tao Yu[1] Tie-Jun Zhao[3]

[1] Faculty of Information Engineering and Automation, Kunming University of Science and Technology

Kunming 650500, China

{hrbrichard, zhen_kl}@126.com, ztyu@hotmail.com

[2] R&D Department, Jinan Qingqi Peugeot motorcycle company limited

Jinan 250104, China

Ldj1972@163.com

[3] School of Computer Science and Technology, Harbin Institute of Technology

Harbin 150001, China

tjzhao@hit.edu.cn

Abstract. Nowadays, the number of electronic information has grown largely. Research on summarization is particularly important. Automatic summarization can accelerate the speed of access to resources. This paper studied the automatic summarization existing technical methods and the basic principle of maximum entropy model. With the principle of maximum entropy and automatic summarization technical characteristics, the automatic summarization method based on maximum entropy model is designed. Summary sentences multifaceted characteristics, design the automatic summarization sentences characteristics of maximum entropy extraction rules, different digest the results of different combinations of features. The results of experiments and examples show that the new method has a good practical effect.

Keywords: Automatic Summarization, Maximum Entropy, Feature Extraction

1 Introduction

Nowadays, with the amount of web information resource increasing dramatically, we are confronted with the problem of "information overload". One of the important carriers of information is language and characters, so it would be particularly important to manage text information. Automatic summarization is considered as one of the effective means to process information resource, because it can automatically extract essential contents which can cover the core of original text from the electronic text in the form of natural language, and express original lengthy text with condensed abstract. Ultimately it can provide people with efficient and fast way to browse information quickly and lock their interest exactly [1].

People have started to study automatic summarizing by the time computers emerged. Summary is the concentration, similar to abstract, of original document. It concisely explains some information such as the subject and scope of the original document. Automatic summarization is the mean which makes use of computer to extract summary from original document automatically, and the goal of generating summary is to describe the core content of original text accurately and comprehensively. After reading and summarizing relative literature, the existing automatic summarization method can be divided into five categories: automatic excerpt method, automatic summarization method based on understanding, information extraction method, automatic summarization method based on discourse structure and automatic abstract method biased toward query.

Originally, maximum entropy model was established by Jaynes[8] in 1957.And he based this model on information entropy theory. His basic principle is to build statistical computing model for known information factors, excluding the effect of unknown factors. In this way, can we obtain uniform probability distribution of known information. Without any hypothesis of unknown factors, while reserving all uncertainty, deviation can be decreased to minimum [2].

The application of maximum entropy model has been very successful in many fields of natural language processing[3,4], such as part-of-speech labeling, phrase recognition, syntactic analysis, semantic role labeling, automatic intelligent question answering, machine translation, etc. The use of maximum entropy model almost reached the best level in these fields. As far as I am concerned, at present, there is no study report on the application of maximum entropy model in automatic digest. So I hope to apply the maximum entropy model to automatic summarization effectively and improve the effect of automatic.

This paper is divided into four chapters; the second chapter introduced automatic summarization technology research based on the maximum entropy method. It introduced the application of Maximum entropy model, the application of is characteristic function, the application of rules of feature extraction, and the application of pa-

rameter calculation in automatic summarization, etc. The third chapter introduced the main method of digest evaluation, and explained the selection group experiment corpus and the generation method of standard digest. Moreover, it determined the parameters of experimental evaluation. Finally, the third chapter did 5 kinds of feature combination experiments for the problem of expository, argumentative and narrative, and analyzed the experimental results .The fourth chapter summarized the main work of this paper.

2 The Automatic Summarization Based on Maximum Entropy Method

2.1 The Characteristic Function of Automatic Summarization Technology Based on Maximum Entropy Method

The process of predicting whether a sentence is digest will involve various factors. Supposing that X is an attribute set composed by these factors, variable Y is the result of whether a sentence is digest, X is called the characteristics of the sentence. P (y | X) is the probability of system predicting some sentences are digest sentences , and this probability can be refers to a system of a certain sentence probability forecast for abstract words. This probability can be estimated by thought above. Maximum entropy model requires that the p (y | X) must make use of the following defined entropy to get maximum, under the condition of meeting certain constraints.

$$H(P) = -\sum_{X,y} p(\mathrm{y|X})\mathrm{logp(y|X)} \quad . \tag{1}$$

Here the constraint is actually refers to all the known facts or conditions, for example, (x0, y0) can be expressed in the following ways:

$$f_i(\mathrm{X,y}) = \begin{cases} 1: \mathrm{if}\ y = y_0 \ \mathrm{and}\ x = x_0 \\ 0: else \end{cases} ,\mathrm{i} = 1,2,3,\ldots,\mathrm{n} \quad . \tag{2}$$

fi (X, y) is called the characteristic function of maximum entropy model, n is the total of all the characteristics. As you can see, these features describe the contact between characteristic X and digest result y. The final output of probability is:

$$\begin{cases} \mathrm{p(y|X)} = \dfrac{1}{\mathrm{Z(X)}} \exp\left(\sum_{\mathrm{i}} \lambda_{\mathrm{i}} f_i(\mathrm{X,y})\right) \\ \mathrm{Z(X)} = \sum_{y} \exp\left(\sum_{\mathrm{i}} \lambda_{\mathrm{i}} f_i(\mathrm{X,y})\right) \end{cases} \quad . \tag{3}$$

λ_i is the weight parameter of each feature (i.e., the importance of characteristics or credibility.) Z (X) is a normalized factor.

2.2 The Feature Extraction and Expression of Automatic Summarization Technology Based on Maximum Entropy Method

The general selection method of digest sentence is to select, accordifng to sentence frequency (the frequency characteristic of the word in sentence, called F features), title (the title characteristic in sentence, called T features), position (the position characteristic in sentence, called L features), syntactic structure, (the syntactic characteristic in sentence, called S features), indicative phrases (indicative phrase in a sentence, called I characteristics) and other characteristics. To a certain extent, each character described the importance that a sentence works as a digest sentence. In the maximum entropy model, we tried to change the sentence characteristic above into sentence features, building bridge between automatic summarization and maximum entropy model.

The Extraction Principle of F Features in Sentences. The computational formula of F features in sentences is following:

$$WordsScore = \frac{(NumWords)^2}{TotalWords} \quad . \tag{4}$$

NumWords means the quantity of key words in the sentences. TotalWords means the number of words in sentences. Numerator is made by quadratic, increasing the difference value of WordsScore, which can distinguish F features of the sentences better. While denominator is invariable, numerator becomes biger . That means the main key words is more in sentences and WordScore is biger. Of course, we are supposed to consider the length of sentence. The longer the sentence is, the larger the quantity of content is. If the quantity of key words is equal which a sentence contains, the content is relatively concentrated in short sentences. However, long sentences describe not only some contents related to keywords but more other contents. So in the formula 3-9, the sentence length work as denominator, simple processing normalization. Test on DUC2002 dataset, formula 3-9 is effective.

A simple division of F features is: when the value of WordsScore is more than a specific threshold, expressing the sentence with high right of sentence F, the F for F0; Otherwise, F for F1.In order to refine the partition degree of F features in sentences, this paper uses the following distinguishing methods.

According to the corresponding value of WordsScore in sentences, the sentences are sorted in descending order. Then in groups of 5% of the total number of sentences, the sentences in an article are divided into 20 categories successively. And the 20 parts are named after F features separately, the top 5% in sentences for F1, the 6% - 10% of the sentences for F2. So all the sentences in which the value of WordsScore is 0 are marked F20. For example, in an original text containing 100 sentences, F features with WordsScore ranking in the top 5% of the sentences is marked as F1.

The Extraction Principle of T Features in Sentences. The computational formula of T features in sentences is following:

$$TitleScore = \frac{TitleTerms}{TotalTerms} . \quad (5)$$

Title Terms means the quantity of title words in the sentences. TotalWords means the number of words in sentences.

A simple division of T features is: when the value of TitleScore is more than a specific threshold, expressing the sentence is related to the core content of article, the T for T1; Otherwise, T for T0.

In order to refine the partition degree of T features in sentences, this paper uses the following distinguishing methods.

According to the corresponding value of TitleScore in sentences, the sentences are sorted in descending order. Then in groups of 5% of the total number of sentences, the sentences in original article are divided into 20 categories successively. And the 20 parts are named after T features separately, the top 5% in sentences for T1, the 6% - 10% of the sentences for T2. So all the sentences in which the value of TitleScore is 0 are marked T20. For example, in an original text containing 100 sentences, T features with WordsScore ranking in the top 5% of the sentences is marked as T1.

The Extraction Principle of L Features in Sentences. The description of L features is as follows: L features of sentences use binary group (L_s L_p) to express, L_p shows that the sentences are in the first paragraph, the last paragraph or the middle segment. When sentences are in the first paragraph, L_p is labeled as L_p0. When sentences are in the middle paragraph, L_p is labeled as L_p1. When sentences are in the last paragraph, L_p is labeled as L_p2. L_s shows that a sentence is in the first place of a paragraph and the middle or in the last part of a sentence. When the sentence is in the first place of a paragraph, L_s is marked with L_s0. When the sentence is in the middle of a paragraph, L_s is marked with L_s1. When the sentence is in the last part of a paragraph, L_s is marked with L_s2.In this way, can the positional relationships be changed into characteristics in form of binary group. And the positional relationships include that sentences are in the first paragraph, sentences are in the middle paragraph, and sentences are in the last paragraph. Meanwhile, they contain a sentence is in the first place of a paragraph, and the last or the middle of a paragraph.

The Extraction Principle of S Features in Sentences. The description of S features is as follows:

According to the punctuation, sentences can be divided into declarative sentences and non-declarative sentences. When a sentence is a declarative sentence, S characteristics of a sentence is marked with s0. Otherwise, it is marked with s1.

The Extraction Principle of I Features in Sentences. In this paper, when we make indicative glossary, indicative word can be divided into two types. The demonstrative of Class A contains important contents, such as "the purpose of the article" "the main purpose of the study" "In summary", etc. The demonstrative of Class B does not contain essential contents, such as "such as" "For example", etc.

According to the indicative glossary, we divided sentences into three categories. Sentences containing indicative words of Class A is called indicative - strengthening statements, and I features of sentences is marked as I0;Sentence containing indicative words of class B is called indicative-weakening statements, and the I features of sentences is marked as I1; The rest of the statement is called non-directive sentence, I features for I2.

The Character Representation of Sentences – Features Template. Synthesizing the description above, we can get five features for every sentence in the text : F, L, S, T, I. We will combine and superimpose features at random which are potentially helpful for generating digest ,getting a variety of expression forms to show a sentence. For example, a sentence contains more keywords and its WordScore ranks in the top 5%.It contains many title words, and its TitleScore ranks in the top 5%.The sentence is the first sentence of the first paragraph ,being declarative sentence ,and doesn't contain indicative words. Then the formal characteristic of this sentence is expressed as: (F1, T1, S1, (L_p1, L_S1), I2).

According to the number of value of F, L, S, TandI, we can calculate the totality of candidate features. Of course, the characteristics of the sentence can be more. Such as considering the length of a sentence and latent semantic analysis of sentences, and so on. In order to improve the efficiency of the experiment, this paper chose the five characteristics above.

There are two kinds of analysis results for every sentence in document, with Y (yes) expressing digest sentence and N (no) expressing non-digest. So there are a corresponding characteristic and a analysis result for sentence S¬i, and corresponding characteristic vector is expressed as ((T, F, L, S, I), P), as is shown below:

(Fi、Ti、Li、Si、Ii) $\rightarrow$ Y expresses the sentence is a digest sentence

(Fi、Ti、Li、Si、Ii) $\rightarrow$ N expresses the sentence is not a digest sentence

Characteristics (T, F, L, S, I) also can combine at random according to need, this paper mainly study the combinations: (F, T), (F, T, S), (F, T, S, L), (F, T, S, L, I).

Thus, recognition extraction of digest sentence based on maximum entropy becomes a tagging problem that whether a sentence is digest sentence (abstract sentence is tagged as Y, or for N).

2.3 Feature Set Selection

Feature selection is a laborious thing, if we list all the characteristics; the workload is heavy .And it will lead to choosing some useless characteristics and wasting processing time. General feature selection has two steps: firstly, extract candidate set from training instance, and then choose the final feature set from the candidates.

There are three commonly used methods of extracting feature set from a candidate set:

The commonly used methods of extracting feature set from a candidate set have three types:

(1) to retain all the candidate set.

(2)only keep the feature set whose characteristic frequency is greater than a certain experience value (for example 5).

(3) to use the incremental method to choose the most differentiation feature set.

The first method is simple, but it kept too much characteristic of no value. The second method is not complicated, but there is no standard basis to determine the experience value .And only on the condition of characteristics candidate set being too much, can we achieve a great result. This paper chose the third method. ts advantage is that it compared to characteristics distinction ,using information increment as measure standard, which can use less characteristics to gain good result. But, the disadvantage is that the training time is longer.

The following is basic process：

Input: $\tilde{p}(x), \tilde{p}(y \mid x)$ and candidate feature group $f_i(x, y)$, $1 \le i \le n$.ar.

Output: to choose a group of feature :$f^*_i (x, y)$,$1 \le i \le n$ and corresponding p*(y| x).

Arithmetic:

Initialization : P*(y| x)= uniform distribution;

To calculate the gain of every selected characteristic, and get the feature that increment is greatest;

To calculate the maximum entropy distribution as a new distribution after increasing characteristics above;

To repeat the two steps for several times, until the improvement is no significant.

The increment calculation of characters:

$$G_{s,f} \equiv \Delta L(S, f) \equiv L(ps \cup f) - L(ps) \quad . \tag{6}$$

The work of characteristics increment is very hard. In order to reduce the work, this paper reference the features gain approximation algorithm to select quickly form literature [7].Approximately, we believes, the model joining a feature relies on original model and parameters α ,namely,

$$p^{\alpha}_{p,s} = \frac{1}{Z_{\alpha}(x)} ps(y \mid x) e^{\alpha f(x,y)}$$

, Z_{α}（x）is normalizing factor

$$G_{s,f}(\alpha) \equiv L(p^{\alpha} s \cup f) - L(p^{\alpha} s) = -\sum_{X} \tilde{p}(x) \log Z_{\alpha}(x) + \alpha \tilde{p}(f),$$

$$\tilde{\Delta} L(S, f) \equiv \max_{\alpha} G_{S,f}(\alpha), \tag{7}$$

$$\tilde{p}_{s \cup f} \equiv \arg \max_{p^{\alpha}_{s,f}} G_{S,f}(\alpha). \tag{8}$$

To calculate (8), the characteristics gain, we should use (7) to derivative, and the value is equal to zero, getting equation. Then to calculate α with Newton method, and make use of (7) to calculate each characteristic increment. Finally find characteristics whose characteristic increment is greatest, cycle (8) in turn.

3 Results Analysis and Evaluation

3.1 Evaluation method

At present, the evaluation method of automatic abstract based on sentence extraction can be roughly divided into two categories: one is called internal evaluation (Intrinsic) method, and it evaluate the abstract system [5] by directly analyzing the quality of digest. The second is called external evaluation (Extrinsic) method; it is an indirect method, through testing the effect of abstract system accomplishing a specific task to evaluate abstract system [6].

Internal evaluation method. The basic thinking of internal evaluation method is comparing the achieved digest with "ideal or "standard abstract", and according to the similarity of the two to evaluate. The more closer the achieved digest is to "ideal digest", the higher the quality of digest is**.** For the "ideal digest", you can refer to the information which the author provided, and adopt the way of professor-making. The following introduces several common internal measuring indexes:

1) recall rate and accuracy

The idea is to evaluate the quality of the abstract by information coverage and accuracy .Generally, comparing automatic digest with ideal digest though the indicators: recall rate (recall is marked as R), accuracy (precision is marked as P). Hypothesizing that Nk is the sentence number of “ideal digest” extraction, Nm is the sentence number of mechanical digest extraction, and Nkm is the number of sentences that are extracted by mechanical digest and “ideal digest”. Then

$$\text{R}=\frac{N_{km}}{N_{k}} . \tag{9}$$

$$\text{P}=\frac{N_{km}}{N_{m}} . \tag{10}$$

2)F-Measure

Recall rate and accuracy are two interrelated evaluation criterion. Usually a system accuracy is enhanced, and the recall rate will drop; the recall rate is enhanced, the accuracy will drop .Therefore it may be biased to evaluate

by one of the two evaluation criterion. F - Measure (marked as F) is a indicator comprehensively considering the recall rate and accuracy .Its definition is as follows:

$$F=\frac{2\times P\times R}{P+R} \quad . \tag{11}$$

External Evaluation Method. Usually, we use external evaluation method to evaluate a digest system in a particular task. Compared with internal evaluation, external evaluation has less subjective advantages. It is easy to evaluate more digest system objectively. However, this method also has some disadvantages: each evaluation is only for a particular task, its pertinence is strong, it has more limitations, and it isn't beneficial to evaluate and improve the system performance.

This paper studies the automatic abstract technology, not for a single area. So it isn't proper to use external evaluation to do experiment, and experimental evaluation adopts internal evaluation method.

3.2 Experiment Corpus and Performance Evaluation Indexes

Experiment Data Set. In this paper, we select 100 copies of expository articles, argumentative articles and narrative articles from DUC2002 data set, and these 300 copies of articles constituted corpus data set. To produce expert digest with better quality, we recruited three experimenter to compose expert group whose score of College English Test -6six is more than 500,and often read English literature (the three students are called A, B, C).Every document needs the three experimenter to accomplish digest together. A selects digest sentence through reading and referencing to artificial digest provided by DUC; B makes expert digest according to his own reading and referencing to automatic writing paper in the office 2003; C accomplishes final expert digest through his own reading and integrating A and B students' digest. The corporation purpose three experimenters is try to reduce the influence of subjective factors to the quality of the experts digest.

To train Maximum entropy model better, the training set can be as much as possible. This paper selects 240 copy of articles as training set from corpus (80 expository articles ,80 argumentative articles ,80 narrative articles).The rest 60 articles is test set (20 expository articles ,20 argumentative articles ,20 narrative articles) .

The Evaluation Indicator of Experiment performance. Because what this paper study is automatic digest sentence extraction based on the maximum entropy, and it doesn't apply to particular system .So this paper adopt internal evaluation method. According to Recall rate, Precision and F-Measure, we evaluate experiment result. To analysis the performance of automatic abstract system based on maximum entropy model (in this paper, in order to describe the system conveniently, we make the maximum entropy system be short for this system), this paper add the automatic abstract system based on bayesian (in this paper, in order to describe the system conveniently, we make bayesian system be short for this system), which benefits performance comparison. The system firstly does the word segmentation and part-of-speech tagging for document .Secondly, it establishes vector space model for the text .Thirdly ,this system makes feature extraction and combines this features automatically by the simple bayesian machine learning method, transforming digest into classification .Finally it will extract document digest and does lubrication processing.

3.3 Experiment Result and Analysis

We train maximum entropy system and bayesian system by using four characteristics combination ((F, T), P), ((F, T, S, P), ((F, T, S, L), P), ((F, T, S, L, I), P) in a same sentence, at the same time. And we test maximum entropy system and bayesian system by corresponding characteristics combination .The following is the experimental result of expository article, argumentative article and narrative article.

Test 1 – Expository Article. Train maximum entropy system and bayesian system with the expository article training set .Test result tested by expository article test set, and results are as the following Table 1.

Table 1. Test result of expository article

characteristics combination	maximum entropy system (%)			bayesian system (%)		
	F	P	R	F	P	R
(F,T)	38	37	40	36	39	33
(F,T,S)	51	48	55	39	43	36
(F,T,S,L)	83	79	88	80	75	87
(F,T,S,L,I)	91	90	93	87	80	95

From Table 1, we can know that whether in maximum entropy system or bayesian system, the more child characteristics, the higher the parameter value is .Which illustrates child characteristics have more or less contributions to digest selection in expository article .And we can see that maximum entropy system is superior to bayesian system in the aspect of accuracy, recall rate and F values in expository article .

Test 2 – Argumentative Article. Train maximum entropy system and bayesian system with the argumentative article training set .Test result tested by argumentative article test set, and results are as the following Table 2.

Table 2. Test result of argumentative article

characteristics combination	maximum entropy system (%)			bayesian system (%)		
	F	P	R	F	P	R
(F,T)	44	45	44	41	45	38
(F,T,S)	46	46	46	55	50	61
(F,T,S,L)	91	90	93	87	86	89
(F,T,S,L,I)	97	98	97	94	91	98

From Table 2, we can see that generally the accuracy, recall rate, and F values in maximum entropy system is superior to those in bayesian system .

Test 3 – Narrative Article. Train maximum entropy system and bayesian system with the narrative article training set. Test result tested by narrative article test set, and results are as the following Table 3.

Table 3. Test result of narrative article

characteristics combination	maximum entropy system (%)			bayesian system (%)		
	F	P	R	F	P	R
(F,T)	34	30	38	33	31	35
(F,T,S)	57	62	53	60	58	63
(F,T,S,L)	87	85	89	75	79	71
(F,T,S,L,I)	92	90	94	85	84	87

From Table 3, we can see that generally the accuracy, recall rate, and F values in maximum entropy system is superior to those in bayesian system .

From the three experiments above, as far as we can see whether for expository article, argumentative article or narrative article, maximum entropy system is superior to bayesian system in the aspect of accuracy, recall rate and F values .So we believe automatic digest system based on maximum entropy model is better than that based on Bayesian. The use of maximum entropy model is effective in automatic digest.

However, from the experiment above, we also can see the maximum entropy system performance of same characteristic combination in three problems. For example (F, T, S, L, I),Its performance is as shown in Fig. 1.

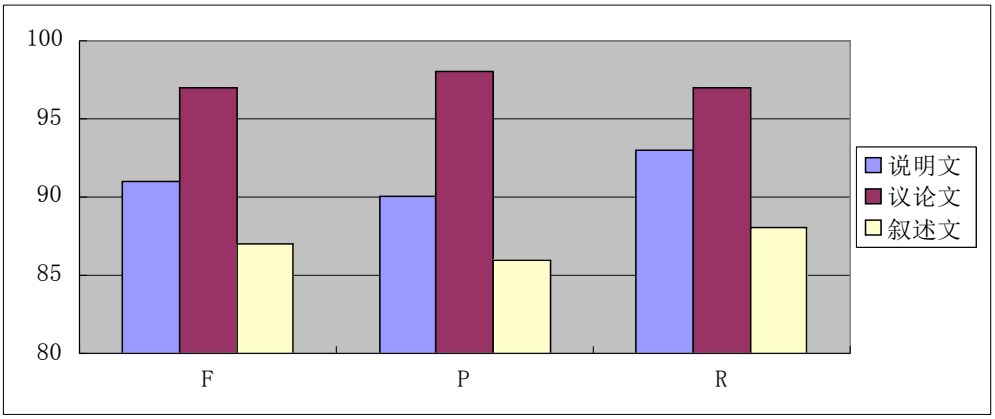

Fig. 1. FPR value contrast of (F, T, S, L, I) in maximum entropy system on the three problems

From the Fig. 1, we can see that characteristics (F, T, S, L, I) performed best in the argumentative article, argument, but performed worst in narrative, which shows that the selection in maximum entropy model have different effects on different styles.

4 Conclusion

The main purpose of this article is to study the use effect of the maximum entropy model in automatic digest. This paper implemented the automatic digest experiment system based on maximum entropy. With referring to the research results of scholars both at home and abroad, and aiming at characteristics of the maximum entropy model and the need of automatic abstract technology, this paper studied basic principle, mathematical function representation, feature extraction, parameter estimation of maximum entropy model, etc. Moreover, this paper realized the application of maximum entropy model in automatic digest by combining with the characteristics of automatic digest technology. Finally, It turns out that maximum entropy model is superior to bayesian model in test parameter, which verified the correctness of original thought and reached expected effect.

Acknowledgement

This work is supported by the National Natural Science Foundation of China (61363045); The Key Project of Yunnan Nature Science Foundation(No.2013FA130) ; Science and technology innovation talents fund projects of Ministry of Science and Technology(No.2014HE001).

References

[1] A. Nenkova, S. Maskey, Y. Liu, "Automatic Summarization," in *Proceedings of the 49th Annual Meeting of the Asso ciation for Computational Linguistics: Tutorial Abstracts of ACL 2011*, No. 3, 2011.

[2] X. Chen, W. Dai. "Maximum Entropy Principle for Uncertain Variables," *International Journal of Fuzzy Systems*, Vol.13, No.3, pp. 232-236, 2011.

[3] L. Zitnick, T. Kanade, "Maximum entropy for Collaborative Filtering," *Eprint Arxiv*, pp. 636-643, 2012.

[4] F.L. Huang, C.J. Hsieh, K.W. Chan, C.J. Lin, "Iterative Scaling and Coordinate Descent Methods for Maximum Entropy Models," *Journal of Machine Learning Research*, Vol. 11, No. 3, pp. 815-848, 2010.

[5] F.C.T. Chua, S. Asur, "Automatic Summarization of Events from Social Media," Technical Report, HP Labs, 2012.

[6] B. Sharifi, M.A. Hutton, J. Kalita, "Automatic summarization of twitter topics," in *Proceedings of National Workshop on Design and Analysis of Algorithm*, 2010.

[7] Y.Q. Zhou, Y.K. Guo, X.J. Huang, A.M.L. De, "Chinese and English BaseNP Recognition Based on a Maximum Entropy Model, " *Journal of Computer Research and Development*, Vol. 40, No. 3, pp. 440-446, 2003.

[8] E.T. Jaynes, " Information theory and statistical mechanics," *Physical Review*, Vol.106, No. 4, pp. 620–630, 1957.