Answer Extraction Based on System Similarity Model and Stratified Sampling Logistic Regression in Rare Date

Peng Li, Yi Guan, Xiaolong Wang and Yongdong Xu,
pli@insun.hit.edu.cn
School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China

Summary
This paper provides a novel and efficient method for extracting exact textual answers from the returned documents that are retrieved by traditional IR system in large-scale collection of texts. The main intended contribution of this paper is to propose System Similarity Model (SSM), which can be considered as an extension of vector space model (VSM) to rank passages, and to apply binary logistic regression model (LRM), which seldom be used in IE to extract special information from candidate data sets. The parameters estimated for the data gathered with serious problem of data sparse, therefore we take stratified sampling method, and improve traditional logistic regression model parameters estimated methods. The series of experimental results show that the overall performance of our system is good and our approach is effective. Our system, Insun05QA1, participated in QA track of TREC 2005 and obtained excellent results.

Key words:
Answer extraction, System Similarity Model, Stratified sampling, Logistic regression model

1. Introduction

Open-domain question answering has recently received more and more attention by reason of advances in the areas of information retrieval (IR), information extraction (IE), and natural language processing (NLP). The goal of a question answering system is to retrieve exact answers to question rather than full documents or passages containing answers as most information retrieval systems usually do. Question answering systems combine IR and IE technology and focus on fact-based, short-answer natural language questions such as "Who invented the paper clip? ". The state-of-the-arts in QA research has been represented in the Text Retrieval Evaluation Conference (TREC) question answering track evaluation [1]. From the first QA mission that the TREC 8 executed in 1999 to TREC 2005, more and more organizations have joined in evaluation which is much more complicated and develops rapidly. The TREC Question Answering Track has motivated much of the advancement in the open-domain QA system. Our system, Insun05QA1, participated in QA track of TREC 2005 for the first time, and obtained satisfied results. Especially, our system ranks fifth for factoid questions.

Answer extraction can be considered as a type of information retrieval (IR) and an important component of QA system [2]. With the constant development and ripe of IR technology, AE has already become the critical factor of influencing the ultimate result of QA system. Therefore, various types of clues have been used to extract answer. For example, Steven Abney extracted answers according to their frequency and position in the passages [3]. Diego et al utilized dependency theory and semantic interpreter for answer extracting [4]. Marius and Sandra employed perceptron-based machine learning approach to answer ranking [5] [6]. Abraham Ittycheriah applied Maximum Entropy and decision tree for answer tagging [7]. It can be concluded from the related work, many researchers deem the problem of AE as a very important problem and adopt many kinds of methods. However, another people still couldn't be full satisfied up to now. People still explore new avenue constantly and make every effort to reach better result.

In this article, we only recommend two innovative methods emphatically, namely, System Similarity Model and stratified sampling logistic regression model (LRM) as well as their application in passage retrieval and answer ranking respectively. System Similarity Model (SSM) can be regarded as an extension of vector space model (VSM), it overcomes a great deal of deficiency in VSM. Answer extraction is typical two-category case, and it is proved in a lot of fields that logistic regression model is a valid tool to solve two-category problems, but because of the serious sparse problem of data that are used for training has existed, make traditional logistic regression can't carry through the parameter estimate well, so we have improved the traditional algorithm by means of stratified sampling, make it can estimate better under the circumstances of multi-feature sparse data.

The rest of the article is organized as follows. We present the architecture of our system in the next section. Then we describe the system similarity model and passage retrieval in section 3. Answer ranking based on stratified sampling logistic regression is introduced in section 4. Finally, experiments and conclusions are described in section 5 and 6.

Manuscript received March 25, 2006.
Manuscript revised March 30 , 2006.

2. Brief Description of Our System

Our system can be divided into three main components. The architecture structure of system is shown in Figure 1. Query preprocessing is comprised of keyword extraction and expansion, answer type prediction. IR component consists of document retrieval, Web retrieval and passage retrieval. Answer extraction includes Named Entity recognition, shallow parser, answer ranking, and answer selection. Question sentences will be changed to a series of keywords and their expansion with non-stop word list and WordNet [8]. It has been apparent for a long time that the most common words in English provide no benefit to searching for a topic [9]. So, we discard the non-stop words in question sentences, and the rest of words will be taken as keywords. For a question, the amount of keyword information that is contained in original sentence is not enough. In order to improve the amount of information, we use WordNet to add synonyms of the keywords. Since this is not the emphasis of this article and space reason, we only give a simple introduction. The detail contents of system can refer to our companion paper [10].

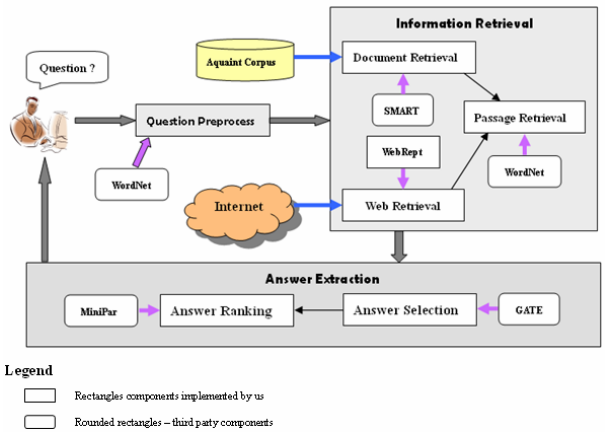

Fig.1. Architecture of our system

3. System Similarity Model and Similarity Calculation

Since our aim is to retrieve exact answer of a natural language question from a large-scale (3GBytes) document collection of texts, documents are too large as a target to retrieve. Therefore documents are segmented into a set of passages based on surface clues such as punctuation symbols. The main action of this step is similarity degree calculation in order to further diminish the extent of extracting. We use a novel information retrieval model, System Similarity Model (SSM), which can be considered as an extension of vector space model (VSM) to calculate the similarity degree [11].

3.1 Brief Review of Vector Space Model

Vector space model (VSM) has been widely applied in information processing fields. Nowadays, most search engines use similarity measures based on this model to rank web documents. In VSM, a document is represented by a set of index terms with weight (vector of terms), and so is user query; a weight expresses the relative importance of the term with respect to the document. Let $\vec{A}$ denotes user query vector $(x_1, x_2, ..., x_N)$, $\vec{B}$ denotes document vector $(y_1, y_2, ..., y_N)$, their relevance score (similarity) is computed by:

$$Similarity(\vec{A}, \vec{B}) = \frac{\sum_{i=1}^{N} x_i y_i}{\sqrt{\sum_{i=1}^{N} x_i^2} \sqrt{\sum_{i=1}^{N} y_i^2}} \quad (1)$$

That is, the cosine of the angle between the two vectors. The ranking formula of VSM not only sorts the results according to their similarity to user query, but also has advantages of simple and fast; hence more and more people admit that VSM is the best balance of quality and simplicity. The criticism to this model concentrates on the following aspects:

(1) Terms are presumed to be independent one another, which is not true of reality for both document texts and user queries.
(2) Order of terms is ignored, it is still in its assumption that word order in text is not important for information retrieval.
(3) The problem of polysemy and synonymy, which is one reason for poor performance of IR systems, is completely overlooked.

Although many attempts have been made to remedy these shortcomings [12], few works have been done to revise the generic framework of VSM itself that solves above-mentioned problems integrally.

3.2 Passage Retrieval with System Similarity Model

We define passages as overlapping sets consisting of a sentence and its two immediate neighbors. The main action of this step is similarity degree calculation in order to further diminish the extent of extracting. System Similarity Model (SSM) is a novel information retrieval model, which can be considered as an extension of vector space mode (VSM) to calculate the similarity degree. It

overcomes a great deal of deficiency in vector space model (VSM). The score for passage i is the score of the central sentence i , and the score of central sentences were calculated with System Similarity Model (SSM).

Given a system $A=\{\alpha_1,\alpha_2,...,\alpha_m\}$, $m=|A|$ and a system $B=\{b_1,b_2,...,b_n\}$, $n=|B|$ .Let $x_i>0$ denote the importance of component a_i $(1\le i\le m)$; Let $y_j>0$ denote the importance of component b_j $(1\le j\le n)$, assuming the number of the most similar binary tuple (MSBTs) is p $\left(p\le \min\{m,n\}\right)$, denoted by $s_1,s_2,...,s_p\in A\times B$, without loss of generality, supposing they are, with similarity degrees $\mu_i(1\le i\le p)$ respectively. The system similarity degree is $Q(A,B)$, then:

$$Q(A,B)=\frac{\sum_{i=1}^{p}\mu_i x_i^2}{\sqrt{\sum_{i=1}^{m}x_i^2}\sqrt{\sum_{i=1}^{p}\mu_i^2x_i^2+\sum_{i=p+1}^{n}y_i^2}} \qquad (2)$$

Here, A denote the user query and the $\alpha_1,\alpha_2,...,\alpha_m$ means keywords of query. In the same way, B denote the every sentence of document set or page set retrieved from prior step. $b_1,b_2,...,b_n$ mean keywords of sentence. x_i and y_j compute by means of variations of standard TF-IDF term weighting scheme [13]. Similarity degrees μ_i are computed with WordNet. The top 50 passages are passed on as input to the answer extraction component.

To compute the relevance score between user query and passage, SSM uses system similarity function (2) to compute the systematic similarity between query vector and passage return the result. To describe it in more detail, we present a system similarity-computing algorithm as Figure 2.

The algorithm is a recursive algorithm. Three input parameters are: Thesaurus, Set representation of system A and system B. thesaurus is global system resource contains all elements with importance and similarity degree between two elements. System A and system B are two systems under computation that are represented by set already, the importance of all level of components has been determined.

Systematic similarity computation algorithm
System Similarity (Thesaurus, A, B)
Input:
1. Thesaurus：Storage of similarity degree between elements
2. Set representation of system A
3. Set representation of system B

Output: systematic similarity degree
Begin
1 **if** both A and B are element
2 **return** Element Similarity (Thesaurus, A, B) ;
3 **else**
4 **begin**
5 initialize weights and element similarities of A and B;
6 **for each** i<=number of components in system A
7 **for each** j<=number of components in system B
8 Constructing systematic similarity matrix: similarity matrix[i] [j] SystemSimilarity(Thesaurus, the ith component of A, the jth component of B);
9 **end_for each**
10 **end_for each**
11 **for each** i<=number of components in system A
12 finding the component in system B that has the maximum similarity with the ith component of system A, checking order constraints and weight variance if necessary, constructing the MSCP;
13 updating the matched vector $\left(x_1,x_2,...,x_p\right)$ of system A;
14 updating the matched vector $\left(\mu_1x_1,\mu_2x_2,...,\mu_px_p\right)$ of system B;
15 **end_for each**
16 constructing unmatched vector of system A $\left(x_{p+1},x_{p+2},...,x_m\right)$;
17 constructing unmatched vector of system B $\left(y_{p+1},y_{p+2},...,y_n\right)$;
18 computing systematic similarity using formula (2);
19 **return** systematic similarity degree;
20 **end**
end

Fig.2. System similarity computation algorithm

4. Answers Ranking Based on Stratified Sampling Logistic Regression Model

Logistic regression model (LRM) is a regular and effective method of statistical analysis for two-category regression analysis. It has extensive application in such fields as economics [14], sociology [15], medicine and so on, but it is less in the field of information processing. Recently, some researchers begin to focus their interest on it such as Xu et al utilized logistic regression to calculate the text units' similarity [16]. Logistic regression is a nonlinear model ， therefore the parameters of the model

are estimated by maximum likelihood generally. It is proved that maximum-likelihood estimation of logistic regression has the characteristics of consistency, asymptotic validity and asymptotic normality [17]. Maximum-likelihood estimation methods have a number of attractive attributes. First, they nearly always have good convergence properties as the number of training samples increases. Furthermore, maximum-likelihood estimation often can be simpler than alternative methods, such as Bayesian techniques or other methods.

Answer extraction is typical two-category case, because one candidate answer only has two kinds of situations, that it is an answer or not. Therefore, this kind of problem is suitable for the method of logistic regression for analyzing. But in the actual conditions, the positive instance (correct answer) far less than negative instance (interference answer), it brings about serious data sparse. In this case, if you directly adopt maximum-likelihood estimation, it will result in the model parameter and probability estimate deviation. This paper brings forward a method of parameter estimation, which can diminish the deviation of estimation.

4.1 Binary Logistic Regression: Model and Parameter Estimated

In logistic regression, a single outcome variable Y_i $(i = 1, ..., n)$ follows a Bernoulli probability function that takes on the value 1 with probability P_i and 0 with probability $1 - P_i$. $P_i / 1 - P_i$ is referred to as the *odds* of an event occurring. Then P_i varies over the observations as an inverse logistic function of a vector X_i, which includes a constant and K explanatory variables :

$$Y_i \sim Bernoulli(Y_i / P_i) \tag{3}$$

$$\ln \frac{P(Y_i = 1)}{1 - P(Y_i = 1)} = \ln(odds) = \alpha_0 + \sum_{k=1}^{K} \beta_k X_{ik} \tag{4}$$

The above is referred to as the log odds and also the logit. By taking the antilog of both sides, the model can also be expressed in odds rather than log odds, i.e.

$$odds = \frac{P(Y_i = 1)}{1 - P(Y_i = 1)} = \exp(\alpha_0 + \sum_{k=1}^{K} \beta_k X_{ik}) \tag{5}$$

$$= e^{\alpha_0 + \sum_{k=1}^{K} \beta_k X_{ik}} = e^{\alpha_0} * \prod_{k=1}^{K} e^{\beta_k X_k} = e^{\alpha_0} * \prod_{k=1}^{K} (e^{\beta_k})^{X_k} \tag{6}$$

As Aldrich and Nelson note, there are several alternatives to the LRM that might be just as plausible or more plausible in a particular case. However,

- the LRM is comparatively easy from a computational standpoint
- there are many tools available which can estimate logistic regression models
- the LRM tends to work fairly well in practice

Note that, if we know either the odds or the log odds, it is easy to figure out the corresponding probability:

$$P_{x_i} = \frac{odds}{1 + odds} = \frac{\exp(\alpha_0 + \beta' X)}{1 + \exp(\alpha_0 + \beta' X)} \tag{7}$$

The unknown parameter α_0 is a scalar constant term and β' is a $k \times 1$ vector with elements corresponding to the explanatory variables. The parameters of the model are estimated by maximum likelihood. That is, the coefficients that make our observed results most "likely" are selected. The likelihood function formed by assuming independence over the observations:

$$L(\alpha_0, \beta) = \prod_{i=1}^{n} P_{x_i}^{Y_i} (1 - P_{x_i})^{1 - Y_i} \tag{8}$$

To random sample (x_i, y_i) , $i = 1, 2, ..., n$, By taking logs and using formula (4), the log-likelihood simplifies to

$$\ln(L(\alpha_0, \beta)) = \sum_{i=1}^{n} [y_i (\alpha_0 + \beta' x_i) - \ln(1 + \exp(\alpha_0 + \beta' x_i))] \tag{9}$$

The estimator of unknown parameter α_0 and β' can be gained from following equations by means of maximum-likelihood estimation.

$$\begin{cases} \dfrac{\partial \ln[L(\alpha_0, \beta)]}{\partial \alpha_0} = \sum_{i=1}^{n} \left| y_i - \dfrac{\exp(\alpha_0 + \beta' x)}{1 + \exp(\alpha_0 + \beta' x)} \right| = 0 \\ \dfrac{\partial \ln[L(\alpha_0, \beta)]}{\partial \beta_j} = \sum_{i=1}^{n} \left| y_i - \dfrac{\exp(\alpha_0 + \beta' x)}{1 + \exp(\alpha_0 + \beta' x)} \right| x_{ij} = 0 \\ j = 1, 2, 3, ..., m. \end{cases} \tag{10}$$

4.2 Stratified Sampling Logistic Regression in Rare Data

In actual application, it often have lager gap between the positive instance and the negative instance, and the positive instance far less than negative instance, so such data have serious data sparse problem. If we adopt general logistic regression to estimate parameters in such data, usually the results are not good or even the wrong. Therefore, we utilized the method of stratified sampling to take full advantage of the resource of positive instances. The concrete process is: random extract some examples from positive instances and negative instances and merge the training samples to parameter estimation.

Under the condition of stratified sampling, sample distribution and population distribution doesn't have identity. Then even though we know the x, the observed value $y=1$ isn't equal to P_x. Of course, the observed value $y=0$ isn't equal to $1-P_x$. In other word, the conditional probability of a sample observed value $y=k$ can't be expressed by formula (8) and formula (10) can't be found naturally.

Assuming that positive instances and negative instances have P_0N and $(1-P_0)N$ respectively among the population, the positive instances of independent variable x divided by total positive instances is γ_x, then the positive instances of independent variable x is $P_0N\gamma_x$. We assume that the negative instances of independent variable x is κ_x, namely,

$$P_x = P_0N\gamma_x / (P_0N\gamma_x + \kappa_x) \tag{11}$$

Then, $\kappa_x = (1-P_x)P_0N\gamma_x / P_x$ and the negative instances of independent variable x divided by total negative instances is λ_x.

$$\lambda_x = (1-P_x)\gamma_x P_0 / (1-P_0)P_x \tag{12}$$

Adopting the method of stratified sampling, we randomly extract r_1 positive instances and r_2 negative instances as sample. The probability of the observed value $y=1$, $y=0$ is:

$$P_x(1) = \frac{r_1\gamma_x}{r_1\gamma_x + r_2\lambda_x} = \frac{r_1(1-P_0)P_x}{r_1(1-P_0)P_x + r_2(1-P_x)P_0} \tag{13}$$

$$P_x(0) = \frac{r_2\lambda_x}{r_1\gamma_x + r_2\lambda_x} = \frac{r_2(1-P_x)P_0}{r_1P_x(1-P_0) + r_2(1-P_x)P_0} \tag{14}$$

Assuming $\omega_0 = P_0N/(1-P_0)N$, $\omega_1 = r_1/r_2$, namely, ω_0 is the ratio of the positive instances and the negative instances in population; ω_1 is the ratio of the positive instances and the negative instances in sample. As to stratified sample (x_i, y_i), $i=1,2,...,n$, the logarithmic likelihood function is:

$$\ln[L(\alpha_0,\beta)] = \sum_{i=1}^{n}\left\{\begin{array}{l} y_i(\ln\omega_1 + \ln P_{x_i}) + \\ (1-y_i)[\ln\omega_0 + \ln(1-P_{x_i})] - \\ \ln[\omega_1 P_{x_i} + (1-P_{x_i})\omega_0] \end{array}\right\}$$

$$= \sum_{i=1}^{n}\left| y_i \ln\frac{\omega_1}{\omega_0}\right| + \sum_{i=1}^{n} y_i \ln\frac{P_{x_i}}{1-P_{x_i}} - \sum_{i-1}^{n}\left|\frac{\omega_1}{\omega_0}\frac{P_{x_i}}{1+P_{x_i}} + 1\right| \tag{15}$$

Utilizing formula (4), the log-likelihood simplifies to

$$\ln[L(\alpha_0,\beta)] = \Omega + \sum_{i=1}^{n}\left\{\begin{array}{l} y_i(\alpha_0 + \beta' x_i) - \\ \ln[1+\exp(\alpha_0 + \omega + \beta' x_i)] \end{array}\right\} \tag{16}$$

Here, $\Omega = \omega\sum_{i=1}^{n} y_i$ and $\omega = \ln\omega_1/\omega_0$ are nothing to estimated parameters. If we assume that $\alpha_1 = \alpha_0 + \omega$, then The estimator of unknown parameter α_1 and β' can be gained from following equations by means of maximum-likelihood estimation.

$$\left\{\begin{array}{l} \dfrac{\partial\ln[L(\alpha_0,\beta)]}{\partial\alpha_0} = \sum_{i=1}^{n}\left| y_i - \dfrac{\exp(\alpha_1 + \beta' x)}{1+\exp(\alpha_1 + \beta' x)}\right| = 0 \\ \dfrac{\partial\ln[L(\alpha_0,\beta)]}{\partial\beta_j} = \sum_{i=1}^{n}\left| y_i - \dfrac{\exp(\alpha_1 + \beta' x)}{1+\exp(\alpha_1 + \beta' x)}\right| x_{ij} = 0 \\ j = 1, 2, 3, ..., m. \end{array}\right. \tag{17}$$

Formula (17) is the parameter estimation formula of stratified sampling logistic regression model. Under the condition of random sampling, sample distribution is identical to population distribution, $\omega_1 = \omega_0$, then $\omega = 0$, $\alpha_1 = \alpha_0$, formula (17) is equal to formula (10). Therefore, formula (17) can be considered as an expansion of formula (10) under the condition of stratified sampling.

5. Experiments and Evaluation

At QA Track of TREC 2005, the Q/A system we developed, in which system similarity model and stratified sampling logistic regression model are adopted as core component of passage retrieval and answer ranking, Insun05QA, participated in the Main Task, which submitted answers to three types of questions: factoid questions, list questions and other questions. The evaluation result is described in Table1.

Table 1: Performance of Insun05QA1 in TREC 2005

		Insun05QA1
Average per-series score		0.187
Factoid questions	Number of correct	106
	Number of unsupported	15
	Number of inexact	16
	Number of wrong	225
	Accuracy	0.293 (median accuracy scores 0.153)
	Precision of NIL	0.057
	Recall of NIL	0.176
List questions	Average F	0.085 (median average F score 0.053)
Others questions	Average F	0.079 (median average F score 0.156)

Among seventy-one participants, our system ranks fifth for factoid questions, seventh for list questions, and eighth in synthetic average score [18]. Although there are other auxiliary technologies such as formalization templates, Web information supporting etc. to make the ultimate system upgrade.

Run Tag	Submitter	Accuracy	NIL Prec	NIL Recall
lcc05	Language Computer Corp.	0.713	0.643	0.529
NUSCHUA1	National Univ. of Singapore	0.666	0.148	0.529
IBM05L3P	IBM T.J. Watson Research	0.326	0.200	0.118
ILQUA2	Univ. of Albany	0.309	0.075	0.235
Insun05QA1	Harbin Inst. of Technology	0.293	0.057	0.176
csail2	MIT	0.273	0.098	0.294
FDUQA14B	Fudan University	0.260	0.082	0.412
QACTIS05v2	National Security Agency (NSA)	0.257	0.045	0.176
mk2005qar2	Saarland University	0.235	0.071	0.353
Edin2005b	Univ. of Edinburgh	0.215	0.068	0.176

Fig.3. Performance of Insun05QA1 in factoid questions

5.1 The Result and Analysis of Passage Retrieval

We use the traditional information retrieval system, SMART, as information extracting component, will get the document related with the question from TREC data sets. Obviously, the precision of document retrieval has determined the final upper limit of our system. Figure 4 shows IR part with increase that file counting of feedback, the variation tendency of precision. We can find out the rate of accuracy of increase with the document quantity is a trend increased progressively, but the rate of accuracy rises very low after the document feedback counts over 50, the rate of accuracies from 50 to 70 only geared up 3%. This proves that such a document counting under the circumstances, it doesn't already have much point to the improvement of the rate of accuracy to depend on increasing the document quantity, the increase of document counting will exert a negative influence on the rate of accuracy of the follow-up step, so we choose file counting of feedback to be 70. The precision of document retrieval is 59%. The rest processing are all to do on the basis of this rate of accuracy, such a rate of accuracy is not very ideal and this is restricted by ability of SMART system. So, it is one of the main research directions in our future work to develop own high-level IR system.

The next step is to calculate passage retrieval through similar degree. Every loss through the rate of accuracy of a step is unavoidable. We can see that generally drop by 6 percentage points or so through the rate of accuracy of passage retrieval step, such loss can be accepted. It proved that the System Similarity Model is effective to calculate the similarity degree. Ultimately, the top 50 passages are passed on as input to the answer extraction component.

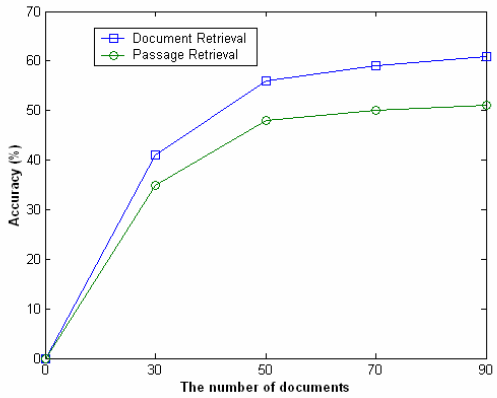

Fig.4. Relation of accuracy and document quantity

5.2 The Result and Analysis of Answer Ranking

The component of answer ranking is a nonlinear classifier using a set of ten features with weights developed by a machine-learning algorithm employing stratified sampling logistic regression. The features used were: the number of keywords and their expansion, the number of different keywords and their expansion, number of named entities in the passage, formalization, subject or

object, the average distance in words between the beginning of the candidate and the keyword and their expansion that also appear in the passage and so on.

The training data set is selected from TREC 8 to TREC 2004 factoid questions set. We adopted 150 questions that have correct answers and their corresponding passages set as training data set to estimate parameter by stratified sampling logistic regression. We devised several experiments and appraised the effect of different scheme. We exam the effect of three methods in TREC 2005 factoid questions set and the total is 362 questions.

- Experiment 1: Answer ranking with empirical formula as baseline method
- Experiment 2: Answer ranking with traditional logistic regress
- Experiment 3: Answer ranking with stratified sampling logistic regression

The results of three experiments can see the following:

Table 2: Results of three experiments

	Experiment 1	Experiment 2	Experiment 3
Total number	362	362	362
Number of correct	66	42	77
Accuracy	0.182	0.116	0.213

From the results, we can see that experiment 1 as baseline method gains a satisfied result because of the features selection strategy. We will introduce the detail content about feature selection method in following paper. The result of experiment 2 is not good by reason of serious data sparse. So, the deviation of parameters estimation is too big. Since we adopt stratified sampling logistic regression to estimate parameters in rare data and the effect is improved obviously. It proved that the method of stratified sampling logistic regression is effective to solve the problem of data sparse.

5. Conclusions and Future

We have described some core technologies of a Q/A system to automatic answer extraction from large-scale text collections, in response to open-domain, natural language questions. The main intended contribution of this paper is to propose System Similarity Model and stratified sampling logistic regression model are adopted as core technologies to apply in passage retrieval and answer ranking respectively. Evaluations indicate that the effect of answer extraction is effective and the method improved the overall performance of system obviously. At QA Track of TREC 2005, our system ranks fifth for factoid questions and seventh for list questions among seventy-one participants. Although there is some other auxiliary technologies making the ultimate system upgrade. However, the satisfied performances of core methods play the crucial role in system.

There are several possible areas for future work. It is potential for us to improved performance through more up-to-date machine learning methods and sophisticated use of NLP techniques. In particular, the semantic and syntactic information of texts may provide significant help. Another field of our future work is to further discover new and effective method to solve the problem of serious data sparse. Developing a high level IR system is also an important factor in the development of an excellent Q/A system.

Acknowledgments

This paper is supported by National Natural Science Foundation of China (60504021, 60425020) and Key Project of Chinese Ministry of Education & Microsoft Asia Research Centre (01307620). And author give thanks to Zhu Kun-peng Ph.D candidate, Jia Wen-Jie and Yu Hong-Xia for their work of this system.

References

[1] Voorhees, E. Overview of the TREC 2004 Question Answering Track. The Thirteenth Text Retrieval Conference, 2004.
[2] John O'Connor. Retrieval of answer sentences and answer-figures from papers by text searching. Information Processing & Management, 11(5/7):155-164. 1975
[3] Abney, S., and Collins, M. Answer Extraction. In Proceedings of the Applied Natural Language Processing Conference (ANLP-2000). 296-301, Seattle, Washington, 2000.
[4] Diego Molla and Ben Hutchinson. Dependency-Based Semantic Interpretation for Answer Extraction. ????????
[5] Marius A. Pasca and Sandra M. Harabagiu. High Performance Question/Answering. Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. 366-374, New Orleans, Louisiana, United States, 2001.
[6] Marius A. Pasca. High-Performance, Open-Domain Question Answering From Large Text Collections. PhD Thesis, University of Southern Methodist, 2001.
[7] Abraham Ittycheriah. Trainable Question Answering Systems. PhD Thesis, The State University of New Jersey, 2001.
[8] Miller, G. WordNet: An on-line lexical database. International Journal of Lexicography 3, 4, 235-312. 1991.

[9] H.P. Luhn. A Statistical Approach to Mechanized Encoding and Searching of Literary Information. IBM Journal of Research and Development, 1(4):309-317, 1957.
[10] Peng Li. XiaoLong Wang and Yi Guan. Extracting Answers to Natural Language Questions from Large-Scale Corpus. Proceedings of 2005 IEEE International Conference on Natural Language Processing and Knowledge Engineering (IEEE NLP-KE'05). 690-694, Wuhan, China. 2005.
[11] GUAN-Yi, WANG Xiao-long, WANG-Qiang. Measurement of System Similarity. JSCL-2005, Nanjing, Aug.2005.
[12] S.K.M. Wong, W. Ziarko, and P.C.N. Wong. Generalized Vector Space Model in Information Retrieval. ACM SIGIR 18-25, 1985.
[13] G. Salton. Automatic Text Processing: The transformation, analysis, and retrieval of information by computer. Addison-Wesley, 1989.
[14] Liang Qi. Distress Prediction: Application of the PCA in Logistic Regression. Journal of Industrial Engineering and Engineering Management. Vol.19, No.1. 100-104. 2005
[15] Gary King, Michael Tomz, and Langche Zeng. ReLogit: Rare Events Logistic Regression. Journal of Statistical Software. Vol.8, 2003.
[16] Yong-Dong Xu, Zhi-Ming Xu, Xiao-Long Wang. Using Multiple Features and Statistical Model to Calculate Text Units Similarity. In Proceedings of the fourth International Conference on Machine Learning and Cybernetics (ICMLC 2005), Guangzhou, China, Aug.2005
[17] Richard O. Duda, Peter E. Hart, and David G. Stork. Pattern Classification, Second Edition. John Wiley & Sons, Inc.84-113, 2001
[18] Ellen M. Voorhees, Hoa Trang Dang. Overview of the TREC 2005 Question Answering Track, TREC 2005. 2005

Peng Li received the B.S. and M.S. degrees in computer science and technology from ShannXi University of Science & Technology in 2000 and 2003, respectively. Currently, he is pursuing his PH.D. Degree in Intelligent Technology & Natural Language Processing Lab of Harbin Institute of Technology. His research interests include information Retrieval, Web information Processing and Question Answering system.

Xiaolong Wang received the B.E. degree in computer science from Harbin Institute of Electrical Technology, China, and the M.E. degree in Computer Architecture from Tianjin University, China, in 1982, and 1984, respectively, and the Ph.D. degree in Computer Science and Engineering from Harbin Institute of Technology, China, in 1989. He joined Harbin Institute of Technology, China, as an Assistant Lecture in 1984 and was an Associate Professor in 1990. He was a senior research fellow at the polytechnic University from 1998 to 2000. Currently, he is a Professor of computer Science at Harbin Institute of Technology, China. His research interest includes artificial intelligence, machine learning, computational linguistics, and Chinese information processing.

GUAN YI is presently an associate professor of the School of Computer Science and Technology at Harbin Institute of Technology. He holds a B.Sc. degree in Computer Science and Technology from Tianjin University in 1992, and a Ph.D. degree in Computer Science and Technology from Harbin Institute of Technology in 1999. In 1996, Dr. GUAN was an invited visiting scholar in Canotec Co.,Japan. In 2000, Dr. GUAN was research associate in Human Language Technology Center at Hong Kong University of Science and Technology, and he was a research scientist in Weniwen.com (Hong Kong) limited in 2001. In October 2001, he became an associate professor in School of Computer Science and Technology in Harbin Institute of Technology. Dr. GUAN's research interests include: question answering, statistical language processing, parsing, text mining.