InsunQA06 on QA track of TREC2006

Yuming Zhao, ZhiMing Xu, Peng Li, Yi Guan
School of Computer Science and Technology, Harbin Institute of Technology,
Harbin, China,
Email :{ ymzhao, xuzm, pli, guanyi} @insun.hit.edu.cn

1 Introduction

This is the second time that our group takes part in the QA track of TREC. We developed a question-answering system, named InsunQA06, based on our Insun05QA system, and with InsunQA06 we participated in the Main Task, which submitted answers to three types of questions: factoid questions, list questions and others questions.

The structure of InsunQA06 is similar with the structure of Insun05QA. Towards Insun05QA, the main difference of InsunQA06 is that new methods are developed and used in answer extraction module, for factoid and “others” questions. And external knowledge such as knowledge from Internet plays more important role in answer extraction. Besides that, we accomplished our documents retrieval module based on Indri, instead of SMART in InsunQA06.

In Section 2, the structure of our InsunQA06 system will be describe, the details of the new methods that we adopted to process the factoid, and “others” questions will separately be described in Section 3, and our results in TREC2006 will be presented in Section 4.

2 Architecture of InsunQA06 system

InsunQA06 system is composed of 4 parts: preprocessing, including questions preprocess and documents preprocess, question analysis, composed of keywords generation and answer type prediction, retrieval, including documents retrieval and passages retrieval, and the last part, answer extraction.

The architecture of InsunQA06 system can be described by Figure.1.

3 Main Components

3.1 Questions Analysis

The question analysis module has two functions: keyword generation and answer type prediction. We accomplished keywords generation just in the same way that used in Insun05QA. And there are slight different in answer type prediction in InsunQA06.

In InsunQA06, we still adopt a rule-based algorithm to predict the answer type of input questions. We define a new answer type classification system which containing seventeen answer types. It is shown in Table 1.

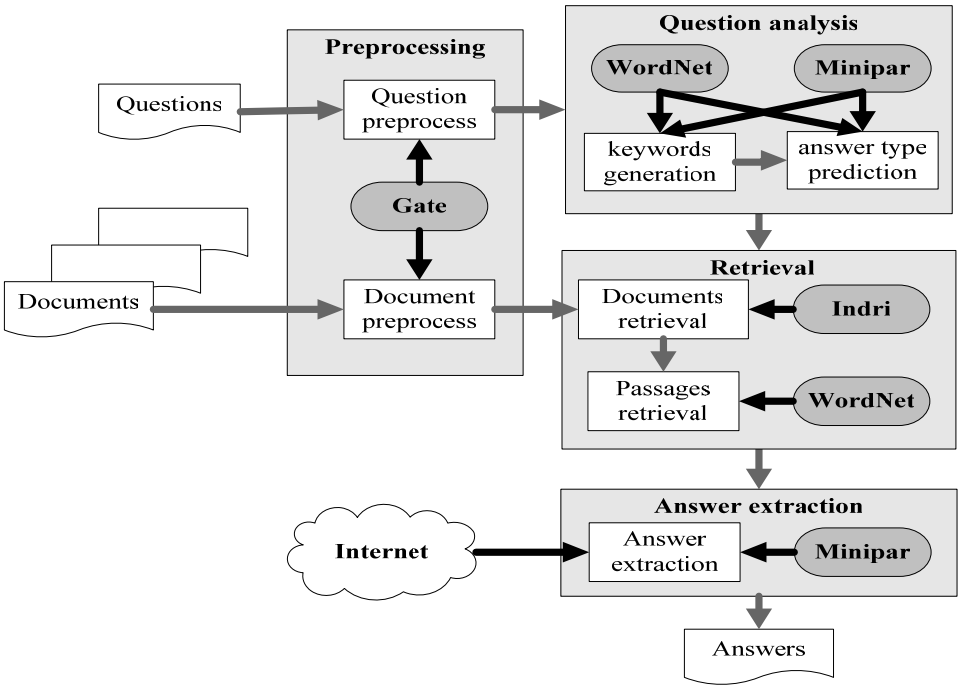

Figure 1. Architecture of InsunQA06

Table 1. Answer Types of InsunQA06

Person	I-PER
Location	I-LOC
Organization	I-ORG
Address	I-ADD
Money	I-MON
Date	I-DAT
Duration	I-DUR
Measurement	I-MEA
Quantity	I-QUA
Ratio	I-RAT
NNP	I-NNP
NN	I-NN
Ill	I-ILL
Song	I-SON
Book	I-BOK
Film	I-FIM
Other	OTHER

3.2 Documents Retrieval

In InsunQA06, we developed our documents retrieval module based on Indri, instead of SMART. Indri is a tool that accomplishes retrieval based on statistical model of natural language, while SMART using the VSM model.

During calculating the documents relativity, the influence of documents frequency isn't taken into account. And the experiments showed that the accuracy didn't decrease for this.

And in documents retrieval module of InsunQA06, no stop words had been got rid of, for our former experiments showed that getting rid of stop words made the results worse than not adopting this processing step.

3.3 Answer Extraction

3.3.1 Factoid Questions

The method we adopted in InsunQA06 is the combination of Stratified Sampling Logistic Regression Model and formalization answer extraction.

Logistic regression model (LRM) is a regular and effective method of statistical analysis for two-category regression analysis. 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. 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. We bring forward a method of parameter estimation, which can diminish the deviation of estimation.

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:

$$Y_i : Bernoulli \quad (Y_i / P_i) \tag{1}$$

$$\ln \frac{P(Y_i = 1)}{1 - P(Y_i = 1)} = \ln(odds) = \alpha_0 + \sum_{k=1}^{K} \beta_k X_{ik} \tag{2}$$

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.

$$odds = \frac{P(Y_i = 1)}{1 - P(Y_i = 1)} = \exp(\alpha_0 + \sum_{k=1}^{K} \beta_k X_{ik}) \tag{3}$$

$$= e^{\alpha_0 + \sum_{k=1}^{K} \beta_k X_{ik}} = e^{\alpha_0} * \prod_{k=1}^{K} e^{\beta_k X_k} = e^{\alpha_0} * \prod_{k=1}^{K} (e^{\beta_k})^{X_k} \tag{4}$$

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:

$$P_{x_i} = \frac{odds}{1+odds} = \frac{\exp(\alpha_0 + \beta' X)}{1+\exp(\alpha_0 + \beta' X)} \qquad (5)$$

The unknown parameter α_0 is a scalar constant term and β' is a k×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:

$$L(\alpha_0, \beta) = \prod_{i=1}^{n} P_{x_i}^{Y_i} (1 - P_{x_i})^{1-Y_i} \qquad (6)$$

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

$$\ln(L(\alpha_0, \beta)) = \sum_{i=1}^{n} [y_i(\alpha_0 + \beta' x_i) - \ln(1 + \exp(\alpha_0 + \beta' x_i))] \qquad (7)$$

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

$$\begin{cases} \dfrac{\partial \ln[L(\alpha_0, \beta)]}{\partial \alpha_0} = \sum_{i=1}^{n} \left| y_i - \dfrac{\exp(\alpha_0 + \beta' x)}{1 + \exp(\alpha_0 + \beta' x)} \right| = 0 \\ \dfrac{\partial \ln[L(\alpha_0, \beta)]}{\partial \beta_j} = \sum_{i=1}^{n} \left| y_i - \dfrac{\exp(\alpha_0 + \beta' x)}{1 + \exp(\alpha_0 + \beta' x)} \right| x_{ij} = 0 \\ j = 1, 2, 3, ..., m. \end{cases} \qquad (8)$$

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 (6) and formula (8) can't be found naturally.

Assuming that positive instances and negative instances have $P_0 N$ 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 $P_0 N\gamma_x$. We assume that the negative instances of independent variable x is κ_x, namely,

$$P_x = P_0 N\gamma_x \big/ \left(P_0 N\gamma_x + \kappa_x\right) \tag{9}$$

Then, $\kappa_x = \left(1-P_x\right)P_0 N\gamma_x / P_x$ and the negative instances of independent variable x divided by total negative instances is λ_x.

$$\lambda_x = \left(1-P_x\right)\gamma_x P_0 \big/ \left(1-P_0\right)P_x \tag{10}$$

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:

$$P_x(1) = \frac{r_1\gamma_x}{r_1\gamma_x + r_2\lambda_x} = \frac{r_1\left(1-P_0\right)P_x}{r_1\left(1-P_0\right)P_x + r_2\left(1-P_x\right)P_0} \tag{11}$$

$$P_x(0) = \frac{r_2\lambda_x}{r_1\gamma_x + r_2\lambda_x} = \frac{r_2\left(1-P_x\right)P_0}{r_1P_x\left(1-P_0\right) + r_2\left(1-P_x\right)P_0} \tag{12}$$

Assuming $\omega_0 = P_0 N \big/ \left(1-P_0\right)N$, $\omega_1 = r_1 / r_2$, 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 $\left(x_i, y_i\right)$, $i=1,2,...,n$, the logarithmic likelihood function is:

$$\begin{aligned}\ln\left[L\left(\alpha_0,\beta\right)\right] &= \sum_{i=1}^{n}\left\{\begin{array}{l} y_i\left(\ln\omega_1 + \ln P_{x_i}\right) + \\ \left(1-y_i\right)\left[\ln\omega_0 + \ln\left(1-P_{x_i}\right)\right] - \\ \ln\left[\omega_1 P_{x_i} + \left(1-P_{x_i}\right)\omega_0\right]\end{array}\right\} \\ &= \sum_{i=1}^{n}\left| y_i \ln\frac{\omega_1}{\omega_0}\right| + \sum_{i=1}^{n} y_i \ln\frac{P_{x_i}}{1-P_{x_i}} - \sum_{i-1}^{n}\left|\frac{\omega_1}{\omega_0}\frac{P_{x_i}}{1+P_{x_i}} + 1\right|\end{aligned} \tag{13}$$

Utilizing formula (2), the log-likelihood simplifies to

$$\ln\left[L\left(\alpha_0,\beta\right)\right] = \Omega + \sum_{i=1}^{n}\left\{\begin{array}{l} y_i\left(\alpha_0 + \beta' x_i\right) - \\ \ln\left[1+\exp\left(\alpha_0 + \omega + \beta' x_i\right)\right]\end{array}\right\} \tag{14}$$

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.

$$\left\{\begin{array}{l} \dfrac{\partial\ln[L(\alpha_0,\beta)]}{\partial\alpha_0} = \displaystyle\sum_{i=1}^{n}\left| y_i - \frac{\exp(\alpha_1 + \beta' x)}{1+\exp(\alpha_1 + \beta' x)}\right| = 0 \\ \dfrac{\partial\ln[L(\alpha_0,\beta)]}{\partial\beta_j} = \displaystyle\sum_{i=1}^{n}\left| y_i - \frac{\exp(\alpha_1 + \beta' x)}{1+\exp(\alpha_1 + \beta' x)}\right| x_{ij} = 0 \\ j = 1, 2, 3, ..., m. \end{array}\right. \tag{15}$$

Formula (15) 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 (15) is equal to formula (8). Therefore, formula (15) can be considered as an expansion of formula (8) under the condition of stratified sampling.

Our formalization answer extraction method mainly orients to Web resources. We developed a method which can be described as "from strict to loose". We defined "hard pattern" and "general pattern". "Hard pattern" was used to extract answers directly and "general pattern" was processed as a feature.

3.3.2 Others Questions

For "Others" questions, we use patterns and abstract method to extract answers from documents. The processing flow is described as Figure 2.

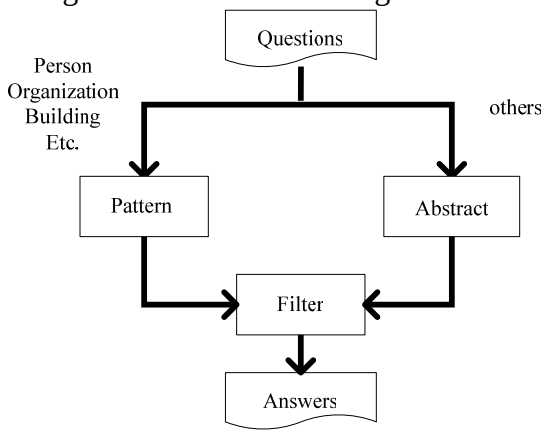

Figure 2. Processing flow of answer extraction for "Others" questions

Taking the factoid and "list" questions as the context of every target, we recognize the types of targets by automatic Name Entity Recognition. According to the type of a target, the system judges which processing step will be chosen. For targets which types are something like "Person", "Organization", and "Building", etc, the pattern method is adopted, since these kinds of targets have obvious and relatively steady features and it is easy and useful to get the patterns of these kinds of targets. An example of the patterns is shown as:

%s was born in |country
%s was born in |time
%s was a |profession
%s has played
%s win the title
%s win the prize in |time
%s win the award in |time
%s graduated from |time
%s was died
%s first
%s's nickname

%s's wife
%s's husband
%s was |number years old when
%s was most known for
%s joined
%s's primary career is
%s married
%s buried in |location
%s lived in |time for
%s rules
…

While the targets are not the types which can be described by steady patterns, such as targets of events, we process them just in the way of Insun05QA for “others” questions. That means based on the ranking of relevant documents, our InsunAbs system, an automatic document abstract generating system, is used on the relevant documents.

For every result obtained from either “pattern” method or abstract method, a filter is used to wipe off the information having already appeared in the answers of the former questions which for the same target.

4 Results

This year, we have submitted only one run, InsunQA06, for the main task of question answering track. The evaluation result is described in Table 2.

Table 2**.** Performance of InsunQA06 in TREC2006

		InsunQA06
Average per-series score		0.157
Factoid questions	Number globally right	120
	Number of unsupported	11
	Number of inexact	18
	Number locally correct	4
	Number of wrong	250
	Accuracy	0.298
	Precision of NIL	0.118
	Recall of NIL	0.353
List questions	Average F	0.118
Others questions	Average F	0.050
	Average F(pyramid evaluation)	0.067

From Table 2, we can see that the performance of our new methods for factoid questions is similar with the method we used in TREC2005. The precision and recall of NIL were obviously raised. For “others” questions, though we adopt new algorithm, it is

obviously that it didn't improve the performance. Maybe it made the result even worse. So, further research should be done and effective method should be developed to improve the performance of our system on dealing with the "others" questions. And for the decrease of average per-series score according to Insun05QA, we think it is mainly induced by the change of the weights in its calculating formular.

Reference

[1] Yuming Zhao, ZhiMing Xu,Yi Guan, Peng Li,Insun05QA on QA Track of TREC 2005, in Proceedings of TREC 2005, NIST, Gettysburg, Pennsylvania

[2] GUAN-Yi, WANG Xiao-long, WANG-Qiang. Measurement of System Similarity. JSCL-2005, Nanjing, Aug.2005

[3] Rohini Srihari and Wei Li. Question answering supported by information extraction. TREC-8 Proceedings, pages 75-85, 1999

[4] H. Cunningham. GATE, a General Architecture for Text Engineering. Computers and the Humanities, volume 36, pp. 223-254, 2002

[5] D. Lin. A Dependency-based Method for Evaluating Broad-Coverage Parsers. Proceedings of IJCAI-95. 1995

[6] Buckley, C., Singhal, A., and Mitra, M. Using Query Zoning and Correlation with SMART: TREC-5. In Proceeding of the 5th Text REtrieval Conference, NIST. 1996

[7] G. Salton. Automatic Text Processing: The transformation, analysis, and retrieval of information by computer. Addison-Wesley, 1989