社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2006－SLP－64（44）
2006／12／22

東京工業大学における質問応答システムの研究概要

ウィッタッカー エドワード† 古井 貞熙†

† 東京工業大学大学院 情報理工学研究科 計算工学専攻
〒 152–8552 東京都目黒区大岡山 2–12–1–W8–77
E-mail: †{edw,furui}@furui.cs.titech.ac.jp

あらまし 本論文では，これまで4年間にわたって東京工業大学で開発されてきた，オープンドメイン，ファクトイド質問を対象とした，データ駆動，非言語的アプローチによる質問応答（QA）システムの概要を述べる．本システムは，本年，TREC（英語），CLEF（スペイン語とフランス語），およびNTCIR（日本語）の3つのQA技術の国際評価に参加して高い成績を得ている．TREC2005では，参加者30チーム中11位，TREC2006では正解精度25.1%で，27チーム中9位であった．CLEFの公式QA評価では，サポートドキュメントが存在しない場合が多かったためによい結果が得られなかったが，非公式「実時間」スペイン語QA試験の性能では，トップクラスに入った．本原稿執筆時点で，日本語に関する結果はまだ公開されていない．

キーワード 質問応答（QA）システム，東京工業大学，オープンドメイン，ファクトイド質問，データ駆動，非言語的アプローチ，国際評価

An Overview of Question Answering at Tokyo Institute of Technology

Edward WHITTAKER† and Sadaoki FURUI†

† Department of Computer Science, Tokyo Institute of Technology
2–12–1–W8–77 Ookayama, Meguro-ku, Tokyo, 152–8552 Japan
E-mail: †{edw,furui}@furui.cs.titech.ac.jp

Abstract In this paper, we give an overview of the data-driven and non-linguistic approach to open-domain factoid question answering (QA) that has been developed over the past 4 years at Tokyo Institute of Technology and which culminated this year in our participation in three international evaluations of QA technology at TREC (for English), CLEF (for Spanish and French) and NTCIR (for Japanese). In TREC2005 we placed 11th out of a total of 30 participants in the factoid QA task and in TREC2006 we came 9th out of 27 participants with an accuracy of 25.1%. While our performance in the official CLEF QA tracks was poor due to a large number of unsupported answers, the performance on an informal "real-time" Spanish QA exercise was one of the best. At the time of writing no results have yet been released for Japanese.

Key words question answering (QA) system, Tokyo Institute of Technology, open domain, factoid question, data-driven approach, non-linguistic approach, international evaluation

1. Introduction

In this paper, we give an overview of the data-driven and non-linguistic approach to open-domain factoid question answering (QA) that has been developed over the past 4 years at Tokyo Institute of Technology and which culminated this year in our participation in three international evaluations of QA technology at TREC, CLEF and NTCIR.

From the outset our approach to QA was designed with language independence and portability in mind. Indeed, in [6], [7] we demonstrated how a baseline QA performance could be achieved for several very different languages with only several tens of hours of data preparation and system development. To this end we developed a probabilistic model of the QA process that captures dependencies between features in the question and features in the answer. We describe our

mathematical model for question answering in detail in Section 2.. For training, this model requires only a few different sets of readily available data, the most important of which is a corpus of example question-and-answer (q-and-a) pairs which is used by the system for what in a traditional QA system would be question-typing and answer-typing. Other data sets include a list of words that are used for question-typing such as *"What"*, *"When"*, *"Who"* and *"Where"* etc. and a set of stop-words, typically frequently occurring non-informative words that should not be used for candidate answer retrieval. For all language systems only the surface forms of words are used and no linguistic processing is performed, for example to determine parts-of-speech or grammatical dependencies between words. In fact, all words are converted to upper-case (where appropriate for the language) and all punctuation is removed except for the insertion of sentence boundaries. Given the relative simplicity of training the QA system it is easy to see how the method can be rapidly applied to new languages without the need for specialized expert knowledge about the language or about QA technology itself.

While most contemporary QA systems employ some form of query expansion so as to be able to find answers that co-occur with grammatical or semantic variants of terms in the question, we instead perform what might be termed data-expansion and use the web to retrieve a large number of documents that match the unmodified query terms. For the tasks in evaluations such as TREC, NTCIR and CLEF where a supporting document is required for each supplied answer it is necessary to project the answer we have found using web data back on to the supplied document collection. Typically this incurs a loss of around 20% relative although on some languages we have found it to be much higher (up to 80%). Since this is more an artifact of evaluation we do not concern ourselves too much with this problem.

Another major difference between our approach and contemporary approaches is in the use of named entities (NE). Most systems use NE-tagging of both the question and the document collection for question and answer typing, typically classifying a question as requiring a particular NE type as the answer. In our approach we perform no such tagging and consider instead all word sequences between one and five words long and how well each word sequence matches the answers in our q-and-a database. Consequently, at no stage is a hard decision about an expected type made; instead all types are assigned a probability and the decision about a final answer postponed until all knowledge sources have been considered. Moreover, we do not require any linguistically motivated labels to be attached to any of our data. This minimizes the effort and complexity of data preparation and also minimizes errors from making hard decisions about a word sequence's identity at too early a stage.

In the rest of this paper we describe our mathematical model of the QA process in Section 2. and present the results on a series of international QA evaluations in Section 3.. A discussion and conclusion are given in Sections 4. and 5., respectively.

2. QA as statistical classification

This section is re-produced verbatim from the paper "TREC2005 Question Answering Experiments at Tokyo Institute of Technology" [3].

It is clear that the answer to a question depends primarily on the question itself but also on many other factors such as the person asking the question, the location of the person, what questions the person has asked before, and so on. Although such factors are clearly relevant in a real-world scenario they are difficult to model and also to test in an off-line mode, for example, in the context of the TREC evaluations. We therefore choose to consider only the dependence of an answer A on the question Q, where each is considered to be a string of l_A words $A = a_1, \ldots, a_{l_A}$ and l_Q words $Q = q_1, \ldots, q_{l_Q}$, respectively. In particular, we hypothesize that the answer A depends on two sets of features $W = \mathcal{W}(Q)$ and $X = \mathcal{X}(Q)$ as follows:

$$P(A \mid Q) = P(A \mid W, X), \tag{1}$$

where $W = w_1, \ldots, w_{l_W}$ can be thought of as a set of l_W features describing the "question-type" part of Q such as *when*, *why*, *how*, etc. and $X = x_1, \ldots, x_{l_X}$ is a set of l_X features comprising the "information-bearing" part of Q i.e. what the question is actually about and what it refers to. For example, in the questions, *Where was Tom Cruise married?* and *When was Tom Cruise married?* the information-bearing component is identical in both cases whereas the question-type component is different.

Finding the best answer $\hat{A}$ involves a search over all A for the one which maximizes the probability of the above model:

$$\hat{A} = \arg\max_A P(A \mid W, X). \tag{2}$$

This is guaranteed to give us the optimal answer in a maximum likelihood sense if the probability distribution is the correct one. We don't know this and it's still difficult to model so we make various modeling assumptions to simplify things. Using Bayes' rule this can be rearranged as

$$\arg\max_A \frac{P(W, X \mid A) \cdot P(A)}{P(W, X)}. \quad (3)$$

The denominator can be ignored since it is common to all possible answer sequences and does not change. Further, to facilitate modeling we make the assumption that X is conditionally independent of W given A to obtain:

$$\arg\max_A P(X \mid A) \cdot P(W \mid A) \cdot P(A). \quad (4)$$

Using Bayes rule, making further conditional independence assumptions and assuming uniform prior probabilities, which therefore do not affect the optimization criterion, we obtain the final optimization criterion:

$$\arg\max_A \underbrace{P(A \mid X)}_{\substack{retrieval \\ model}} \cdot \underbrace{P(W \mid A)}_{\substack{filter \\ model}}. \quad (5)$$

The $P(A \mid X)$ model is essentially a language model which models the probability of an answer sequence A given a set of information-bearing features X, similar to the work of [2]. It models the proximity of A to features in X. We call this model the *retrieval model* and examine it further in Section 2.1.

The $P(W \mid A)$ model matches an answer A with features in the question-type set W. Roughly speaking this model relates ways of asking a question with classes of valid answers. For example, it associates dates, or days of the week with *when*-type questions. In general, there are many valid and equiprobable A for a given W so this component can only re-rank candidate answers retrieved by the retrieval model. If the filter model were perfect and the retrieval model were to assign the correct answer a higher probability than any other answers of the same type the correct answer should always be ranked first. Conversely, if an incorrect answer, in the same class of answers as the correct answer, is assigned a higher probability by the retrieval model we cannot recover from this error. Consequently, we call it the *filter model* and examine it further in Section 2.2.

2.1 Retrieval model

The retrieval model essentially models the proximity of A to features in X. Since $A = a_1, \ldots, a_{l_A}$ we are actually modeling the distribution of multi-word sequences. This should be borne in mind in the following discussion whenever A is used. As mentioned above, we currently use a deterministic information-feature mapping function $X = \mathcal{X}(Q)$. This mapping only generates word m-tuples ($m = 1, 2, \ldots$) from single words in Q that are not present in a *stop-list* of around 50 high-frequency words. In principle the function could of course extract deeper linguistic features but we leave this for future work.

We first assume that a corpus of text data S is available for searching for answers comprising $|S|$ sentences $S_1, \ldots, S_{|S|}$ and $|U|$ documents and a vocabulary V of $|V|$ unique words. We use the notation X_i to define an active set of the features $x_1, \ldots, x_{l_X}$ such that $X_i = x_1 \cdot \delta(d_1), x_2 \cdot \delta(d_2), \ldots, x_{l_X} \cdot \delta(d_{l_X})$ where $\delta(\cdot)$ is a discrete indicator function which equals 1 if its argument evaluates true (i.e. its argument(s) are equal, is not an empty set, or is a positive number) and 0 if false (i.e. its argument(s) are not equal, is an empty set, is 0 or is a negative number) and $\vec{d} = [d_1, \ldots, d_{l_X}]$ is the solution(注1) to $i = \sum_{j=1}^{l_X} 2^{j-1} d_j$.

The probability $P(A \mid X)$ is modeled as a linear interpolation of the 2^{l_X} distributions(注2):

$$P(A \mid X) = \sum_{i=0}^{2^{l_X}-1} \lambda_{X_i} \cdot P(A \mid X_i), \quad (6)$$

where $\lambda_{X_i} = 1/2^{l_X}$ for all i, $P(A \mid X_0)$ is a zerogram distribution, and $P(A \mid X_i)$ is the conditional probability of A given the feature set X_i and is computed as the maximum likelihood estimate from the corpus S:

$$P(A \mid X_i) = \frac{N(A, X_i)}{N(X_i)}, \quad (7)$$

where

$$N(A, X_i) = \sum_{j=1}^{|S|} \delta(X_i \in \mathcal{X}(S_j)) \cdot \delta(A \in S_j), \quad (8)$$

$$N(X_i) = \sum_{v \in V} N(v, X_i). \quad (9)$$

We modify Equation (8) to include contributions from adjacent sentences weighted by λ_{adj} which typically has a value $\leqq 1$:

$$N(A, X_i) = \sum_{j=1}^{|S|} \delta(X_i \in \mathcal{X}(S_j)) \cdot \max\{\delta(A \in S_j), \lambda_{adj} \cdot \delta(A \in S_{j-1}), \lambda_{adj} \cdot \delta(A \in S_{j+1})\}. \quad (10)$$

(注1) : Note that the value of i is simply the base10 number that represents the binary encoding of the active features in X_i.

(注2) : A linear interpolation of models, which borrows directly from statistical language modeling techniques for speech recognition, was found to give retrieval performance approximately twice that of a naive-Bayes or log-linear formulation.

It turns out that smoothing the maximum likelihood estimates from each component distribution has little effect on performance so none is performed. This is partly because of the inherent smoothing effect achieved by interpolating all the distributions together and partly since there is no need to smooth for non-occurring events since such zerotons are never likely to be selected as answers.

One clear deficiency, however, is the use of equal-valued interpolation weights for all distributions. One might expect a dependence on the number of active features or on $N(X_i)$, however, no such reliable relationship has so far been determined although investigations continue.

2.2 Filter model

The question-type mapping function $\mathcal{W}(Q)$ extracts n-tuples ($n = 1, 2, \ldots$) of question-type features from the question Q, such as *How*, *How many* and *When were*. A set of $|V_{\mathcal{W}}| = 2522$ single-word features is extracted based on frequency of occurrence in questions in previous TREC question sets. Some examples include: *when*, *where*, *who*, *whose*, *how*, *many*, *high*, *deep*, *long* etc.

Modeling the complex relationship between W and A directly is non-trivial. We therefore introduce an intermediate variable representing classes of example questions-and-answers (q-and-a) c_e for $e = 1 \ldots |C_E|$ drawn from the set C_E, and to facilitate modeling we say that W is conditionally independent of c_e given A as follows:

$$P(W \mid A) = \sum_{e=1}^{|C_E|} P(W, c_e \mid A) \tag{11}$$

$$= \sum_{e=1}^{|C_E|} P(W \mid c_e) \cdot P(c_e \mid A). \tag{12}$$

Given a set E of example q-and-a t_j for $j = 1 \ldots |E|$ where $t_j = (q_1^j, \ldots, q_{l_{Qj}}^j, a_1^j, \ldots, a_{l_{Aj}}^j)$ we define a mapping function $f : E \mapsto C_E$ by $f(t_j) = e$. Each class $c_e = (w_1^e, \ldots, w_{l_{We}}^e, a_1^e, \ldots, a_{l_{Ae}}^e)$ is then obtained by $c_e = \bigcup_{j:f(t_j)=e} \mathcal{W}(t_j) \bigcup_{i=1}^{l_{Aj}} a_i^j$, so that:

$$P(W \mid A) = \sum_{e=1}^{|C_E|} P(W \mid w_1^e, \ldots, w_{l_{We}}^e) \cdot P(a_1^e, \ldots, a_{l_{Ae}}^e \mid A). \tag{13}$$

Assuming conditional independence of the answer words in class c_e given A, and making the modeling assumption that the jth answer word a_j^e in the example class c_e is dependent only on the jth answer word in A we obtain:

$$P(W \mid A) = \sum_{e=1}^{|C_E|} P(W \mid c_e) \cdot \prod_{j=1}^{l_{Ae}} P(a_j^e \mid a_j). \tag{14}$$

Since our set of example q-and-a cannot be expected to cover all the possible answers to questions that may be asked we perform a similar operation to that above to give us the following:

$$P(W \mid A) = \sum_{e=1}^{|C_E|} P(W \mid c_e) \prod_{j=1}^{l_{Ae}} \sum_{a=1}^{|C_A|} P(a_j^e \mid c_a) P(c_a \mid a_j), \tag{15}$$

where c_a is a concrete class in the set of $|C_A|$ answer classes C_A. The independence assumption leads to underestimating the probabilities of multi-word answers so we take the geometric mean of the length of the answer (not shown in Equation (15)) and normalize $P(W \mid A)$ accordingly.

The system using the above formulation of filter model given by Equation (15) is referred to as model ONE. Systems using the model given by Equation (13) are referred to as model TWO. The training of Model ONE has been described in detail in [4].

2.3 Reconciling $P(A \mid X)$ and $P(W \mid A)$

The approach to QA that has been presented is similar in essence to that of approaches to automatic speech recognition (ASR) where there are separate acoustic and language models. In ASR, it is necessary to include a *language model weight*, α, which raises the probabilities given by the language model to the power α, otherwise performance is very poor:

$$\hat{A} = \arg\max_A \frac{P(A \mid X)^\alpha \cdot P(W \mid A)}{\sum_{A'} P(A' \mid X)^\alpha \cdot P(W \mid A')}.$$

Several, possibly related, explanations have been given for this requirement including compensation for the independence assumption. In any case, the dynamic range of the models is typically very different and needs compensating somehow. α can be optimized easily once the individual models have been optimized separately.

3. Results

In this section we present the results on the factoid question task of several different evaluations that we undertook over the last year. Each evaluation has its own specification of correctness or accuracy so the results are not necessarily directly comparable with each other although the variation across languages and across questions from year to year

makes such direct comparisons difficult anyway.

In Tables 1 and 2 we give the results from our participation in CLEF in June 2006 that were presented in [8]. For the CLEF evaluation two new systems were built to handle questions in Spanish and French. The results on the monolingual Spanish and French tasks are shown in Table 1.

Table 1 Breakdown of performance on the French and Spanish mono-lingual tasks by type of question and assessment of answer.

Task	Right	ineXact	Unsupp.
Spanish-Spanish	26 (13.7%)	1	29
French-French	27 (14.2%)	12	12

Since the focus of the CLEF QA track is cross-lingual QA we interfaced the two new QA systems with publicly, web-accessible machine translation systems. The results for different combinations of source and target languages are shown in Table 2.

Table 2 Breakdown of performance on the English, French and Spanish cross-lingual combinations by type of question and assessment of answer.

Task	Right	ineXact	Unsupp.
English-French	19 (10.0%)	6	8
English-Spanish	11 (5.8%)	0	10
French-English	7 (3.7%)	10	37
Spanish-English	10 (5.3%)	11	34
French-Spanish	22 (11.6%)	0	15

At the CLEF2006 workshop in Alicante a novel evaluation was performed to assess the speed and performance of systems in a Spanish language real-time QA task. Each participant was given the same 20 questions to answer and the emphasis was on speed and accuracy of results. The mean reciprocal rank (MRR) of both exact and inexact answers for the five participants are shown in Table 3.

Table 3 Mean reciprocal rank (MRR) of both exact and inexact answers and timings of systems that participated in the Spanish real-time QA task. (Participants anonymized since results are not official.)

Group	MRR (Exact)	MRR (Inexact)	Time (s)
Group A	0.41	0.41	549
Group B	0.35	0.35	56
Group C	0.30	0.30	1966
TokyoTech	**0.38**	**0.48**	**5141**
Group E	0.24	0.28	76

In July 2006 we participated in the annual TREC QA evaluation. We submitted three different systems which were each based on different combinations of individual systems using different model formulations (ONE and TWO as described in Section 2. 2 and the open-source Aranea system [1]) and different languages. The three different systems are described in Table 4 and the results for each system given in Table 5. Where a different language system was employed questions were first translated from English into the language of the QA system, then its answers translated back into English.

Table 4 Brief descriptions of the three runs asked06a,b,c submitted to TREC2006.

System	Which model	Languages
asked06a	ONE	English,French,Spanish
asked06b	ONE+TWO	English
asked06c	ONE+TWO	English

Table 5 Performance on the factoid task of the 3 runs submitted to TREC2006.

System	Right	Unsupp.	ineXact
asked06a	62 (15.4%)	12 (3.0%)	24 (6.0%)
asked06b	95 (23.6%)	22 (5.5%)	27 (6.7%)
asked06c	101 (25.1%)	26 (6.5%)	27 (6.7%)

4. Discussion

Our results in the CLEF evaluation were certainly not the best but nonetheless very competitive when compared against the other participants particularly when it is considered that the French and Spanish systems were developed from scratch in the two months prior to the evaluation. Moreover, the official score for all evaluations only considers supported answers and since we project all our answers on to the appropriate corpus we tend to lose many correct answers due to their lack of support. In particular, on the CLEF Spanish and French monolingual tasks our accuracy was 28.9% and 20.5%, respectively, when unsupported answers were included. On some of the cross-language tasks e.g. French to English we gain 19.5% absolute if unsupported answers are taken into consideration, i.e. we lost 84% of our correct answers due to lack of support on that task.

Although our system was the slowest in the Spanish QA real-time task, we achieved the second best MRR for exact answers and by far the highest MRR for inexact answers. Since system speed is more of an implementation issue (and dominated by retrieval and processing of web documents in the current system) we know this can easily be improved. These results clearly show our system's potential for web-based factoid QA especially when support information is not being assessed.

In TREC2006 we placed 9th out of 27 participating groups on the factoid task with 25.1% correct and supported, a score

that is well above the median but still substantially lower than the best participating system's performance (57.8%). (Since our focus is still primarily on the factoid task we did not expend much effort on answering the list and "other" questions in the task and consequently our overall, combined score was below the median—results not shown.) Again, as with the CLEF results, our performance increases to 31.8% when unsupported answers are also included.

The current model of QA has been shown to be remarkably effective particularly when using web documents to find answers and exploiting the web's inherent redundancy. However, the performance is significantly reduced when a much smaller document collection is used in which the correct answer may only occur several times and in contexts which share very few words in common with the question. This was demonstrated with our participation in TREC2005 where we compared the same model using web data and using the supplied document corpus—accuracies of 17.7% and 14.3% were obtained, respectively.

Run **asked06b** used a similar combination of component runs as our best run in TREC2005 giving a performance on the factoid task this year of 23.6%, compared to 21.3% last year. This demonstrates the large variation in absolute accuracy that comes from using different questions. The inclusion of the translated French and Spanish runs and also a run from a modified version of the open-source Aranea system [1] improved system performance by 1.5% absolute to 25.1%. Most of this increase probably comes from the inclusion of the Aranea answers rather than the translated multi-lingual runs.

5. Conclusion

In this paper we have given an overview of the statistical, data-driven and language independent approach to question answering (QA) adopted at Tokyo Institute of Technology and presented the results on a variety of different evaluations involving factoid QA in different languages. It was shown that because the web is used as the source of data for finding answers many answers are unsupported when assessed in the evaluations. This means that the actual performance on monolingual QA tasks is often around 30% and is largely independent of the language. While good, compared to the best linguistic-based systems this performance still falls somewhat short. However, our model is currently still extremely simple and shows great potential for improvement via the inclusion of more discriminative features for question and answer typing and improved candidate answer retrieval through query expansion techniques taken from the language modelling for IR literature. We have many ideas for future work and aim to implement the most promising in time for next year's evaluations.

6. Online demonstration

A demonstration of the system using model ONE supporting questions in English, Japanese, Chinese, French, Spanish, Russian and Swedish can be found online at `http://asked.jp/`

7. Acknowledgments

The authors wish to thank P. Chatain, P. Dixon, Y. Dong, J. Hamonic, M. Heie, D. Klakow, T. Klingberg and J. Novak for their contributions to this work. This research was supported by JSPS and the Japanese government 21st Century COE Programme.

References

[1] J. Lin and B. Katz. Question Answering from the Web Using Knowledge Annotation and Knowledge Mining Techniques. In *Proceedings of Twelfth International Conference on Information and Knowledge Management (CIKM 2003)*, 2003.

[2] J. Ponte and W. Croft. A Language Modeling Approach to Information Retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval*, Melbourne, Australia, 1998.

[3] E. Whittaker, P. Chatain, S. Furui, and D. Klakow. TREC2005 Question Answering Experiments at Tokyo Institute of Technology. In *Proceedings of the 14th Text Retrieval Conference*, 2005.

[4] E. Whittaker, S. Furui, and D. Klakow. A Statistical Pattern Recognition Approach to Question Answering using Web Data. In *Proceedings of Cyberworlds*, 2005.

[5] E. Whittaker, J. Hamonic, and S. Furui. A Unified Approach to Japanese and English Question Answering. In *Proceedings of NTCIR-5*, 2005.

[6] E. Whittaker, J. Hamonic, T. Klingberg, Y. Dong, and S. Furui. Monolingual Web-based Factoid Question Answering in Chinese, Swedish, English and Japanese. In *Proceedings of the Workshop on Multilanguage Question Answering, EACL*, 2006.

[7] E. Whittaker, J. Hamonic, T. Klingberg, Y. Dong, and S. Furui. Rapid Development of Web-based Monolingual Question Answering Systems. In *Proceedings of ECIR2006*, 2006.

[8] E. Whittaker, J. Novak, P. Chatain, P. Dixon, M. Heie, and S. Furui. CLEF2006 Question Answering Experiments at Tokyo Institute of Technology. In *Proceedings of the CLEF2006 Workshop*, 2006.