Rapid Development of Web-based Monolingual Question Answering Systems

E.W.D. Whittaker, J. Harmonic, D. Yang, T. Klingberg, and S. Furui

Dept. of Computer Science
Tokyo Institute of Technology
2-12-1, Ookayama, Meguro-ku
Tokyo 152-8552 Japan
{edw,yuuki,raymond,tor,furui}@furui.cs.titech.ac.jp

Abstract. In this paper we describe the application of our statistical pattern classification approach to question answering (QA) to the rapid development of monolingual QA systems. We show how the approach has been applied successfully to QA in English, Japanese, Chinese, Russian and Swedish to form the basis of our publicly accessible web-based multilingual QA system at `http://asked.jp`.

1 Introduction

The approach to question answering (QA) that we adopt has previously been described in [3–5] where the details of the mathematical model and how it was trained for English and Japanese were given. In this paper we demonstrate how this new statistical pattern classification approach to QA has been successfully applied to monolingual QA for the five distinct languages of English, Japanese, Chinese, Russian and Swedish. Using our approach and given appropriate training data it is found that a proficient developer can build a QA system in a new language in approximately 10 hours. The systems, built using this method, form the basis of our web demo which is publicly available at `http://asked.jp`.

Our approach to QA is significantly different to that commonly employed in contemporary QA systems. Specifically, our approach was designed to exploit the vast amounts of data available on the web, to require an absolute minimum of linguistic knowledge about the language to be encoded in the system and to be robust to the kinds of input errors that might come from a spoken interface to the system. For example, in our English-language system we only use capitalised word tokens in our system and do not use WordNet, named-entity (NE) extraction, regular expressions or any other linguistic information e.g. from semantic analysis or from question parsing. We do, however, rely heavily on the web and a conventional web search engine as a source of data for answering questions, and also require large collections of example questions and answers (q-and-a). Nonetheless, our approach is still very different to other purely web-based approaches such as askMSR and Aranea. For example, we use entire documents rather than the snippets of text returned by web search engines; we do not use

structured document sources or databases and we do not transform the query in any way neither by term re-ordering nor by modifying the tense of verbs. These basic principles apply to each of our language-specific QA systems thus simplifying and accelerating development.

Our approach has been successfully evaluated in the 2005 text retrieval conference (TREC) question answering track evaluations [1] where our group placed eleventh out of thirty participants [3]. Although the TREC QA task is substantially different to web-based QA this evaluation showed that the approach works and provides an objective assessment of its quality. Similarly, for our Japanese language system we have evaluated the performance of our approach on the NTCIR-3 QAC-1 task [5]. Although our Japanese experiments were applied retrospectively, the results would have placed us in the mid-range of participating systems.

We briefly describe our statistical pattern classification approach to QA in Section 2. In Section 3 we describe the basic building blocks of our QA system and how they can typically be trained. We also give a breakdown of the data used to train each language specific QA system and the approximate number of hours required for building each system.

2 Statistical pattern classification approach to QA

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. For simplicity, we choose to consider only the dependence of an answer A on the question Q. In particular, we hypothesize that the answer A depends on two sets of features extracted from Q: $W = \mathcal{W}(Q)$ and $X = \mathcal{X}(Q)$ as follows:

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

where W can be thought of as a set of l_W features describing the "question-type" part of Q such as *who*, *when*, *where*, *which*, etc. and X is a set of 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 is Mount Everest?* and *How high is Mount Everest?* 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. Making various conditional independence assumptions to simplify modelling we obtain the final optimisation criterion:

$$\arg\max_{A} \underbrace{P(A \mid X)}_{\substack{retrieval \\ model}} \cdot \underbrace{P(W \mid A)}_{\substack{filter \\ model}}. \quad (3)$$

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. It models the proximity of A to features in X. This model is referred to as the *retrieval model*.

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 names of people or companies with *who*-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. Consequently, we call it the *filter model*.

3 System components

There are four basic ingredients to building a QA systen using our approach: (1) a collection of example question-and-answer (*q-and-a*) pairs used for answer-typing (answers need not necessarily be correct but must be of the correct answer type); (2) a classification of words (or word-like units cf. Japanese/Chinese) into classes of similar words (*classes*) e.g. a class of country names, of given names, of numbers etc.; (3) a list of question words (*qlist*) such as *"Who"*, *"Where"*, *"When"* etc.; and (4) a stop list of words that should be ignored by the retrieval model (*stoplist*).

The q-and-a for different languages can often be found on the web or in commercial quiz software that is relatively cheap to acquire. To obtain the classes C for each language a fast automatic clustering algorithm taken from the statistical language modelling literature was applied [2]. To obtain word classes in this manner only a large source of training text T comprising $|T|$ words in the target language is required. Typically, the vocabulary V is taken to be the most frequent $|V|$ word tokens in T which are then clustered into $|C|$ classes. The qlist is generated by taking the most frequently occurring terms in the q-and-a examples and the stoplist is formed from the 50 or so most frequently occurring words in T.

At run time Google is used to select web documents related to the question being asked. The question is passed as-is to Google after the removal of stop words. In our web demos the top 100 documents are downloaded in their entirety, HTML markup removed, the text cleaned and upper-cased. We have found that the more documents used the better the performance with no observed performance degradation even up to 10000 documents in Japanese, for example. For consistency, all data in our system in encoded using UTF-8.

The data and relevant system details for each language-specific QA system are given in Table 1 where the estimated number of man-hours to build each

of the new systems is also shown. For the Japanese system Chasen[1] is used to segment character sequences into word-like units. For Chinese each sentence is mapped to a sequence of space-separated characters.

Language	# q-and-a examples	T (corpus name)	$\|T\|$	$\|V\|$	$\|C\|$	# hours
English	290k	AQUAINT	300M	300k	5k	—
Japanese	270k	MAINICHI	150M	215k	500	—
Chinese	7k	TREC Mandarin	68M	33k	1k	10
Russian	98k	LUB [2]	100M	500k	1k	10
Swedish	5k	PAROLE	19M	367k	1k	10

Table 1. System description and number of hours to build each new language's QA system.

4 Conclusion and Further work

In this paper we have shown how our recently introduced statistical pattern classification approach to QA can be applied successfully to create with minimal effort monolingual web-based QA systems in many languages. In the official TREC2005 QA evaluation our approach was shown to be comparable to the state-of-the-art for English language QA. On the NTCIR-3 QAC-1 Japanese-language QA task comparable performance with the state-of-the-art was also obtained. Although no official results are available for Chinese, Russian and Swedish QA systems our subjective evaluations show that performance is lower but competitive with the English and Japanese systems. In future we aim to develop QA systems in many more languages and evaluate performance objectively for example by participating in the annual CLEF evaluations.

5 Acknowledgments

This research was supported by JSPS and the Japanese government 21st century COE programme. The authors also wish to thank Dietrich Klakow for all his contributions.

References

1. E. Voorhees and H. Trang Dang. Overview of the TREC 2005 Question Answering Track. In *Proceedings of the TREC 2005 Conference*, 2005.
2. E. Whittaker. *Statistical Language Modelling for Automatic Speech Recognition of Russian and English*. PhD thesis, Cambridge University, 2000.
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.

[1] http://chasen.naist.jp/hiki/ChaSen