CLEF2006 Question Answering Experiments at Tokyo Institute of Technology

E.W.D. Whittaker, J.R. Novak, P. Chatain, P.R. Dixon, M.H. Heie and S. Furui
Dept. of Computer Science,
Tokyo Institute of Technology,
2-12-1, Ookayama, Meguro-ku,
Tokyo 152-8552 Japan
{edw,novakj,pierre,dixonp,heie,furui}@furui.cs.titech.ac.jp

Abstract

In this paper we present the experiments performed at Tokyo Institute of Technology for the CLEF2006 Multiple Language Question Answering (QA@CLEF) track. Our approach to question answering centres on a non-linguistic, data-driven, statistical classification model that uses the redundancy of the web to find correct answers. Using this approach a system can be trained in a matter of days to perform question answering in each of the target languages we considered—English, French and Spanish. For the cross-language aspect we employed publicly available web-based text translation tools to translate the question from the source into the corresponding target language, then used the corresponding mono-lingual QA system to find the answers. The hypothesised correct answers were then projected back on to the appropriate closed-domain corpus. Correct and supported answer performance on the mono-lingual tasks was around 14% for both Spanish and French. Performance on the cross-lingual tasks ranged from 5% for Spanish-English, to 12% for French-Spanish. Our projection method was shown not to work well: in the worst case on the French-English task we lost 84% of our otherwise correct answers. Ignoring the need for correct support information the exact answer accuracy increased to 29% and 21% correct on the Spanish and French mono-lingual tasks, respectively.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software;

General Terms

Measurement, Performance, Experimentation

Keywords

Question answering, Statistical classification, Cross-language, Spanish, French

1 Introduction

In this paper we describe how we applied our recently developed statistical, data-driven approach to question answering (QA) to the task of multiple language question answering in the CLEF2006 QA evaluation. The approach that we used, described in detail in previous publications [14, 15, 16],

uses a noisy-channel formulation of the question answering problem and the redundancy of data on the web to answer questions. Our approach permits the rapid development of factoid QA systems in new languages given the availability of suitable question and answer training examples and a large corpus of text data such as web-pages or newspaper text as described in [17].

Although we had previously developed systems in five different languages using the same method none of these languages except English were included in the CLEF2006 evaluation. We therefore chose to build French and Spanish systems from scratch and participate in the French and Spanish mono-lingual tasks and the cross-language combinations of both languages together with English. Using the procedure applied successfully in [17] we developed first-cut French and Spanish systems in a couple of days and used the remaining time before the actual evaluation for system optimisation on previous years' CLEF evaluation questions[1].

Our approach is substantially different to conventional approaches to QA though it shares elements of other statistical, data-driven approaches to factoid question answering found in the literature [1, 2, 3, 7, 11, 12, 13]. More recently, similar approaches to the answer typing employed in our system have appeared [9] although they still use linguistic notions that our approach eschews in favour of data-driven classifications. While this approach results in a small number of parameters that must be optimised to minimise the effects of data sparsity and to make the search space tractable, we largely remove the need for numerous ad-hoc weights and heuristics that are an inevitable feature of many rule-based systems. The software for each language-specific QA system is identical; only the training data differs. All model parameters are determined when the data is loaded at system initialisation; this typically only takes a couple of minutes to compute and they do not change in between questions. New data or system settings can therefore be easily applied without the need for time-consuming model re-training.

Many contemporary approaches to QA require the specialised skills of native speaking linguistic experts for the construction of rules and databases that are often used by other QA systems. In contrast, our method allows us to include all kinds of dependencies in a consistent manner and has the important benefits that it is fully trainable and requires minimal human intervention once sufficient data is collected. This was particularly important in the case of our participation in CLEF this year since although our French QA system was developed by a native French-speaker, our Spanish system was built by a student with a conversational level of Spanish learnt in school.

Our QA systems tend to work well when there are numerous (redundant) sentences that contain the correct answer which is why a web search engine is used to obtain nominally relevant documents. In particular, it is advantageous for good performance that the correct answer co-occurs more frequently (roughly speaking) with words from the question than other candidate answers of the same answer type do. If this is not the case, the QA system has no other information with which to differentiate the correct answer from competing alternatives of the same answer type. By using a large amount of text data that contain the question words we are essentially replacing query expansion (as performed by most QA systems) with what might be called document expansion: the documents are expanded to match the query rather than expanding the query to match the documents. Due to the evaluation requirement that support from a fixed document collection be provided for each question, our answers must subsequently be projected on to the appropriate collection. Inevitably this is a lossy operation as will be discussed in Section 5 and also means we never attempt to predict "unanswerable" questions by giving "NIL" as an answer.

The rest of this paper is organised as follows: we first present a summary in Section 2 of the mathematical framework for factoid QA as a classification task that was presented in [15]. We describe the experimental setup specific to the CLEF2006 evaluation in Section 3 and present the results on each task that were obtained in the evaluation in Section 4. A discussion and conclusion are given in Sections 5 and 6.

[1]Development of the QA system itself is relatively fast and straightforward—by far the most time-consuming part is the development of robust text download, extraction and text normalisation tools for any given language.

2 QA as statistical classification with non-linguistic features

This section is re-produced verbatim from the paper "TREC2005 Question Answering Experiments at Tokyo Institute of Technology" [14].

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 hypothesise 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 maximises 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 modelling assumptions to simplify things. Using Bayes' rule this can be rearranged as

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

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

$$\arg\max_A P(X \mid A) \cdot P(W \mid A) \cdot P(A). \tag{4}$$

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

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

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 [10]. It models the proximity of A to features in X. We call this model the *retrieval model* and do not examine it further—please refer to [14, 15, 16] for more details.

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 the next section.

2.1 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.

Modelling 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 modelling we say that W is conditionally independent of c_e given A as follows:

$$P(W \mid A) = \sum_{e=1}^{|C_E|} P(W, c_e \mid A) \tag{6}$$

$$= \sum_{e=1}^{|C_E|} P(W \mid c_e) \cdot P(c_e \mid A). \tag{7}$$

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:

$$P(W \mid A) = \sum_{e=1}^{|C_E|} P(W \mid w_1^e, \ldots, w_{l_{We}}^e) \cdot P(a_1^e, \ldots, a_{l_{Ae}}^e \mid A). \tag{8}$$

Assuming conditional independence of the answer words in class c_e given A, and making the modelling 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:

$$P(W \mid A) = \sum_{e=1}^{|C_E|} P(W \mid c_e) \cdot \prod_{j=1}^{l_{Ae}} P(a_j^e \mid a_j). \tag{9}$$

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:

$$P(W \mid A) = \sum_{e=1}^{|C_E|} P(W \mid c_e) \prod_{j=1}^{l_{Ae}} \sum_{a=1}^{|C_A|} P(a_j^e \mid c_a) P(c_a \mid a_j), \tag{10}$$

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 (10)) and normalise $P(W \mid A)$ accordingly.

The system using the above formulation of filter model given by Equation (10) is referred to as model ONE. Systems using the model given by Equation (8) are referred to as model TWO. Only systems based on model ONE were used in the CLEF2006 evaluation systems described in this paper.

3 Experimental setup for CLEF2006

The CLEF2006 tasks we took part in are as follows: F-F, S-S, F-S, E-S, E-F, F-E and S-E where F=French, S=Spanish and E=English. For each task we submitted only one run and up to ten ranked answers for each question. No classification as to whether a question was more likely to be a factoid, definition or list question was performed prior to answering a question. Therefore all questions were treated as factoid questions since the QA systems we developed were trained using only factoid questions and the features extracted from factoid questions.

For the two mono-lingual tasks (French and Spanish) questions were passed as-is to the appropriate mono-lingual system after minimal query normalisation and upper-casing of all question terms. For the cross-lingual tasks questions were first translated into the target language using web-based text translation tools: Altavista's Babelfish [5] for French-Spanish and Google Translate [6] for all other combinations. The translated question was then normalised and upper-cased and passed to the appropriate mono-lingual system[2]

For each question input to our QA system the question was passed to Google after removing stop words and the (up to) top 500 documents were downloaded for each question. For answering a specific question only that question's downloaded data was used. Document processing involved the removal of any document markup, conversion to UTF-8, and the same language-specific normalisation and upper-casing as applied to the questions.

For answering questions in a given language the corpus in which answers were to be located was not used. However, once a set of answers to a question had been obtained the final step was to project the answers on to the appropriate corpus. Due to a lack of time and resources for a full development of the projection system we relied on using Lucene [4] to determine the document which had the highest ranking when the answer and question were used as a Boolean query. Snippets were likewise determined using Lucene's summary function. If an answer could be found somewhere in the document collection the snippets were further filtered to ensure that the snippet always included the answer string (though possibly none of the question words).

Due to time limitations we chose only to implement the French and Spanish system using model ONE, given by Equation (10). Although we have implementations for English using both models ONE and TWO, for consistency we used the English system that implemented model ONE only. In the TREC2005 QA evaluation model TWO outperformed model ONE by approximately 50% relative—our aim is therefore to implement model TWO for the Spanish real-time task and in other languages in time for future CLEF evaluations.

4 Results

All tasks were composed of a set of 190 factoid and definition questions and 10 list questions. For all tasks up to 10 answers were given by our QA systems to all questions. In general the top 3 answers for the factoid/definition questions were assessed and all list answers were assessed for exactness and support.

In Table 1 a breakdown is given of the results obtained on the two mono-lingual (French and Spanish) tasks for factoid/definition questions with answers in first place and for all answers to list questions.

[2]One alternative would have been to use the mono-lingual system of the source language to obtain answers then translate its answers into the target language. A combination of these two approaches could also have been used to try to minimise the effects of poor automatic translation performance.

Task	Factoid/definition questions					List questions			
	Right		ineXact	Unsupp.	CWS	Right	ineXact	Unsupp.	P@N
S-S	26	(13.7%)	1	29	0.035	3	0	0	0.03
F-F	27	(14.2%)	12	12	0.142	9	2	0	0.09

Table 1: Breakdown of performance on the French and Spanish mono-lingual tasks by type of question and assessment of answer, where *right* means exactly correct and supported.

A similar breakdown for the five[3] cross-lingual tasks is given in Table 2 for factoid/definition questions with answers in first place and for all answers to list questions.

Task	Factoid/definition questions					List questions			
	Right		ineXact	Unsupp.	CWS	Right	ineXact	Unsupp.	P@N
E-F	19	(10.0%)	6	8	0.017	4	1	1	0.06
E-S	11	(5.8%)	0	10	0.005	1	0	0	0.01
F-E	7	(3.7%)	10	37	0.003	1	3	15	0.01
S-E	10	(5.3%)	11	34	0.008	0	1	18	0.00
F-S	22	(11.6%)	0	15	0.037	2	0	0	0.02

Table 2: Breakdown of performance on the English, French and Spanish cross-lingual combinations by type of question and assessment of answer, where *right* means exactly correct and supported.

5 Discussion

There were four main factors in our submissions to CLEF2006 that were expected to have a large impact on performance: (1) the mis-match in time period between the document collection and the web documents used for answering questions; (2) the use of factoid QA systems to also answer definition and list questions; (3) the effect of the machine translation tools for the cross-language tasks; and (4) the projection method of mapping answers back on to the appropriate document collection.

Since all our QA systems relied on web data to answer questions for all languages there was an inevitable mis-match in the time period of documents used for answering questions and the time-frame that was meant to be used i.e. 1994-1995. Although web documents exist which cover the same period, web search engines typically return more recent documents. However, it turned out that this was not a major problem although there were inconsistencies for questions such as *"¿Quién es el presidente de Letonia?"/"Qui est le président de la Létonie*[4] *?"/"Who is the president of Latvia?"* and *"¿Quién es el secretario general de la Interpol?"/"Qui est le secrétaire général d'Interpol?"/"Who is the secretary general of Interpol?"* the answers to which have changed in the intervening period.

It was observed during our participation in the TREC2005 evaluations that simply using a factoid QA system to output the top so many answers for list questions was not a very promising approach, even when list questions were used for training. Part of the problem was due to a paucity of list question training examples compared to the number of factoid questions available. Another problem lay in how to determine the threshold for outputting answers: whether simply to output a fixed number of answers each time, or to base it on some function of the answer score. In the CLEF2006 evaluations the problem was further compounded by not knowing in advance which questions would be factoid, definition and list questions. We therefore decided to assume

[3]Note that the Spanish-French cross-lingual task was not run in CLEF2006.

[4] *"Létonie"* should have been written as *"Lettonie"* in the French mono-lingual test set; it was, however, written correctly in the French-Spanish and French-English test sets.

all questions were factoids and output ten answers in all cases. Our poor performance on all list questions for all tasks can be attributed mostly to there being very few list question examples in our training data and very few list question features (such as plurals) used in the filter model. As a consequence, answer typing for list questions was not very effective. For definition questions the independence assumptions made by model ONE render very poor answer typing of definition questions unless an answer is able to be defined in one or two words.

The substantially lower answer accuracies (between 3.7% and 12.0%) obtained on the cross-language tasks where Babelfish and Google Translate were used for question translation were generally expected due to the well documented quality of such translation tools. It was deemed unlikely that the highest result that was obtained, for French-Spanish, was due to using Babelfish rather than Google Translate and was instead due more to the relative similarity of the two languages (see Section 5.1). In any case, further improvements in machine-translation techniques will almost certainly result in considerable improvements in our cross-language QA performance and multi-language combination experiments.

We were far more surprised and disappointed by the loss incurred by our projection method which reduced our set of correct answers by 47% and 31% on the Spanish and French mono-lingual tasks, respectively. If we were to ignore the need for correct support information the performance would increase to 29% and 21% correct on the Spanish and French mono-lingual tasks, respectively. In the worst case on the French-English task we lost 84% of our otherwise correct answers; equivalently we would have obtained an exact answer accuracy of 23% if the support requirement were ignored. Our previous experience with projection onto the AQUAINT document collection for English language answers on the TREC2005 QA task using the algorithm included in the open-source Aranea system [8] had shown fairly consistent losses of around 20%. While the algorithm that we applied in CLEF2006 was far simpler than that employed by Aranea, it did have access to the full document collection for finding documents containing answers whereas for TREC we relied only on the (up to) top 1000 documents supplied by NIST that were obtained using the PRISE retrieval system. This prevented any errors from only retrieving documents that were selected using only question features however the increased recall of documents containing the answers might have been offset by lower precision.

The Spanish system, like the French system, was developed in a very short period of time. Making further refinements, increasing the amount of training data used, and implementing model TWO are expected to bring accuracy into line with the English system. The possible advantages of applying a refined cross-language approach to mono-lingual tasks, e.g. using the combined results of multiple mono-lingual systems to answer questions in a particular language, are also being investigated. This will provide a means of further exploiting the redundancy of the web, as well as a method to improve the results for languages which are still under-represented on the web.

In the next two sections we present brief language-specific discussions of the results that were obtained.

5.1 Spanish

As indicated in Table 2 results for the mono-lingual Spanish task were considerably better than those obtained for the cross-lingual tasks (English-Spanish, French-Spanish). The discrepancy between the results for the mono-lingual and cross-language Spanish test sets can be almost entirely explained in terms of the relative accuracy of the automatic translation tools used as an intermediate step to obtaining results for the latter. Furthermore, the difference between the results for the French-Spanish task and those for the English-Spanish task is almost certainly due to the relative closeness of the language pair, with Spanish and French both being members of the Romance family of languages, rather than the use of different automatic translation tools. These differences aside, results for all three Spanish tasks exhibited similar characteristics.

The results on all three tasks that included Spanish as a source or target language were by far the best for factoid questions, especially those whose answers could be categorised as names or dates. Of the 26 exactly correct and supported answers obtained for the Spanish mono-lingual

task, a total of 20 consisted of proper names or dates, (11 dates, 9 proper names). If the 29 correct but unsupported answers are also taken into account, this total rises to 41, and accounts for approximately 75% of all correct answers obtained for this particular task.

Definition questions, in addition to being ambiguous in the evaluation sense, are much more difficult than factoids for our QA system to answer. Yet, despite treating all questions as inherently factoid, some interesting results were obtained for the Spanish mono-lingual task. In particular, these results included 2 exactly correct and supported answers, and 8 correct but unsupported answers in the definition category. A cursory analysis of the data revealed that each of these correct answers could be construed as the result of a categorisation process, whereby the subject of the question had been classified into a larger category, and this category was then returned as the answer, e.g. *"¿Qué es la Quinua?" ("What is Quinua?")* answer: [a] cereal, and *"¿Quién fue Alexander Graham Bell?" ("Who was Alexander Graham Bell?")* answer: [an] inventor. The system gives these 'category' words high scores due to the fact that they often appear in the context of proper nouns, where they are used as definitions, or as noun-qualifiers. However, because the system uses no explicit linguistic typology or categories, this results in occasional mismatches such as: *"¿Quién es Nick Leeson?" ("Who is Nick Leeson?")* (answer: Barings). This answer would be categorised as a retrieval error since *'Barings'* is a valid answer type for a Who-question but its high co-occurrence with the subject of the question results in an overly high retrieval model score.

5.2 French

For the French mono-lingual task, unsupported answers were not as much an issue as for the Spanish mono-lingual task, although there were still 12 unsupported answers for the factoid questions, 10 or 11 of which would have also been exact. For the English-French task, there were 8 unsupported answers for factoid questions, almost all of which were also exact. Projection onto French documents, however imperfect, seems to have been less of a problem than for the other languages though it is unlikely that the differences are significant.

Out of our 27 correct and supported answers on the mono-lingual task, 23 were places, dates or names. For those types of questions the answer types that were returned were usually correct. Questions involving numbers, however, were a serious problem: out of 15 *"How many..."* questions we got only one correct, and the answer types which were given were not consistent instead being dates, numbers, names or nouns. The same observation holds for *"How old was X when..."* questions, which were all incorrectly answered, with varying answer types for the answers given. With a rule-based or regular-expressions-based system it is difficult to make such errors. However, with our probabilistic approach, in which no hard decisions are made and all types of answers are valid but with varying probabilities, it is entirely possible to incur such *filter model errors*. Although some cases would be trivially remedied with a simple regular-expression this is against our philosophy; instead we feel the problem should be solved through better parameter estimation and better modelling of the training data, rather than ad-hoc heuristics.

Another interesting observation on the mono-lingual task was that for 19 questions where the first answer was inexact, wrong, or unsupported we got an exact and supported answer in second place. For answers in third place the number of exact and supported answers was only 3. In most of these cases, the answer types were the same. This is untypical of our results obtained previously on English and Japanese where there is typically a significant drop in the number of correct answers at each increase in the rank of answers considered.

As with Spanish, automatic translation of the questions into other languages was far from perfect. One common problem was words with several meanings which were (correctly) translated into French using the wrong meaning, thus radically changing the meaning of keywords in the question. For example, in the English question *"In which settlement was a mass slaughter of Muslims committed in 1995?"* "settlement", is translated into "règlement". Consequently, answers given for this question related to the French legal system rather than a location. Moreover, it was apparent that Google Translate was far from optimal for translating questions presumably because source sentences are expected to be in the affirmative. Thus, *"What is..."* and *"Which*

is..." became *"Ce qui..."* which our QA system tended to interpret as *"Qui..."* thus favouring a person or company as the answer type. Similarly *"How old..."* often became *"Comment vieux..."* rather than *"Quel âge..."* and so was answered as if it were a regular *"How..."* question.

6 Conclusion

With the results obtained in the CLEF2006 QA evaluation we feel we have proven the language independence and generality of our statistical data-driven approach. Comparable performance using model ONE has been obtained under evaluation conditions on the three languages of English, French and Spanish in both this evaluation and TREC2005. In addition, post-evaluation experimentation with Japanese [16] has confirmed the efficacy of the approach for an Asian language as well.

While the absolute performance of our QA systems falls short of that obtained by state-of-the-art linguistics-based systems both the French and Spanish systems were developed only over the two months prior to the evaluation and use an absolute minimum of linguistic knowledge to answer questions in favour of using the redundancy of the web.

Further work will concentrate on how to answer questions using less redundant data through data-driven query expansion methods and also look at removing the independence assumptions made in the formulation of the filter model to improve question and answer typing accuracy. We expect that improvements made on language-specific systems will feed through to improvements in all systems and we hope to be able to compete in more and different language combinations in CLEF evaluations in the future.

7 Online demonstration

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

8 Acknowledgements

This research was supported by the Japanese government's 21st century COE programme: "Framework for Systematization and Application of Large-scale Knowledge Resources".

References

[1] A. Berger, R. Caruana, D. Cohn, D. Freitag, and V. Mittal. Bridging the Lexical Chasm: Statistical Approaches to Answer-Finding. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, Athens, Greece, 2000.

[2] E. Brill, S. Dumais, and M. Banko. An Analysis of the AskMSR Question-answering System. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2002.

[3] A. Echihabi and D. Marcu. A Noisy-Channel Approach to Question Answering. In *Proceedings of the 41st Annual Meeting of the ACL*, 2003.

[4] O. Gospodnetic and E. Hatcher. *Lucene in Action*. Manning, 2005.

[5] `http://babelfish.altavista.com`.

[6] `http://translate.google.com`.

[7] A. Ittycheriah and S. Roukos. IBM's Statistical Question Answering System—TREC-11. In *Proceedings of the TREC 2002 Conference*, 2002.

[8] 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.

[9] C. Pinchak and D. Lin. A Probabilistic Answer Type Model. In *European Chapter of the ACL*, Trento, Italy, 2006.

[10] 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.

[11] D. Radev, W. Fan, H. Qi, H. Wu, and A. Grewal. Probabilistic Question Answering on the Web. In *Proc. of the 11th international conference on WWW*, Hawaii, US, 2002.

[12] D. Ravichandran, E. Hovy, and F. Josef Och. Statistical QA—Classifier vs. Re-ranker: What's the difference? In *Proceedings of the ACL Workshop on Multilingual Summarization and Question Answering*, 2003.

[13] R. Soricut and E. Brill. Automatic Question Answering: Beyond the Factoid. In *Proceedings of the HLT/NAACL 2004: Main Conference*, 2004.

[14] 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.

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

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

[17] E. Whittaker, J. Hamonic, D. Yang, T. Klingberg, and S. Furui. Monolingual Web-based Factoid Question Answering in Chinese, Swedish, English and Japanese. In *Workshop on Multi-lingual Question Answering (EACL)*, Trento, Italy, 2006.