CLEF2007 Question Answering Experiments at Tokyo Institute of Technology

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

Abstract

In this paper we describe the experiments carried out at Tokyo Institute of Technology for the CLEF 2007 QAst (Question Answering in speech transcripts) pilot task, as well as our results from the official evaluation. We apply a non-linguistic, data-driven approach to Question Answering (QA), based a noisy channel model. The system we use for the QAst evaluation comprises an Information Retrieval (IR) module which uses an LM-based approach to sentence retrieval, and an Answer Extraction (AE) module which identifies and ranks the exact answer candidates in the retrieved sentences. Our team participated in the CLEF 2007 QAst pilot track, task T1: QA in manual transcriptions of lectures, and task T2: QA in automatic transcriptions of lectures. On the official evaluation our system achieved a best run MRR of 0.20 and a top1 score of 0.14 on task T1, and a best run MRR of 0.12 and a top1 score of 0.08 on task T2, placing us 3rd in a field of 5 teams that submitted results for these tasks. All experiments and evaluations descibed in this paper were conducted using the CHIL corpus (transcriptions of lectures) which was supplied to all track participants by the QAst track coordinators. ASR lattices were also provided by LIMSI, however we did not use these during the official evaluation.

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; H.3.7 Digital Libraries; H.2.3 [**Database Managment**]: Languages—*Query Languages*

General Terms

Measurement, Performance, Experimentation

Keywords

Question answering, Language modeling, Speech recognition, Spoken document retrieval

1 Introduction

In this paper we explain our experimental setup and general approach to automatic Question Answering (QA), and report our official evaluation results for the CLEF 2007 QAst (Question Answering in speech transcripts) pilot track. We employed an entirely data-driven, non-linguistic and largely language independent QA framework for the QAst track, which was similar but not identical to that which we used in previous QA evaluations such as TREC 2006, CLEF 2006, NTCIR 2006, etc. This approach, which is detailed in [11, 12, 13] centers on a noisy-channel model of the QA problem and generally speaking relies on the redundancy of answer data in the target corpus in order to identify and extract correct answers.

Our QAst system comprises two major components, an Information Retrieval (IR) module used to identify and retrieve relevant sentences from a large corpus, and an Answer Extraction module which is used to identify and rank exact answers in the sentences returned by the IR module. Our approach, which is data-driven and does not require human-guided interaction except for the development of a short list of frequent stop words and common question words, makes it possible to rapidly develop new systems for a wide variety of different languages. Furthermore performance accuracy is roughly comparable even across very disparate languages such as English and Japanese, and developers need not have more than a perfunctory acquaintance with the language [6, 14] in order to build and deploy a new system.

Our data-driven approach differs substantially from conventional rule-based approaches, yet it does share certain features with other approaches in the literature [1, 2, 3, 4, 8, 9, 10]. Systems which employ similar answer-typing approaches have lately begun to appear [7], however most of these systems still utilize some form of specific linguistic knowledge in contrast to our all-data driven, non-linguistic, classification approach. Although our approach requires that a small number of parameters be optimized to minimize the effects of data sparsity, these parameters are all determined at system initialization time and are invariant across different questions. This means that new data or system settings can be applied without the need for wearisome model re-training.

Due to its data-driven nature our QA system performs best when there are numerous redundant sentences containing the correct answer and question words. This reliance on data redundancy to help identify correct answers has seldom been a source of difficulty in past evaluations, however the QAst pilot track presented a unique challenge due to the relatively small size of the CHIL lectures target corpus. In other closed domain evaluations with medium-sized corpora we have opted to utilize web data, however this did not seem entirely appropriate for the QAst track due to the spoken nature of the data and very small corpus size. In part to help combat the resulting data sparsity, we employed a new language-modeling based sentence retrieval IR module as a precursor to the Answer Extraction (AE) stage. This sentence retrieval module acts as an intermediate filter and helps to eliminate noise usually contained in the larger original documents.

The rest of the paper is structured as follows. Section 2 describes our QA architecture in detail. In section 3 we detail our experimental setup. Section 4 describes our results and Section 5 presents a brief discussion of the results. Finally, section 6 concludes the paper.

2 QA Architecture for QAst

The answer to a question depends primarily on the question itself but also on many other factors such as the identity and location of the questioner, previous questions, social context and so on. Although such factors are clearly relevant in many situations, they are difficult to model and also to test. In our approach to QA we therefore limit ourselves to modeling the most straightforward dependence, the probability of an answer A given the question Q. In the system used for the QAst evaluation, we divide the work of identifying answers between two major modules, the Information Retrieval (IR) module which employs an LM-based approach to sentence retrieval, and the Answer Extraction (AE) module. We briefly describe the IR module, the AE module and the Query Expansion process below.

2.1 Information Retrieval module

The general approach to IR for QA is to treat the question as a standard search query, but discard question-type words such as *"what"*, *"when"*, *"who"*, etc., and possibly also a set of stop words.

We employ a language modeling approach to this problem where an individual LM is estimated for each document. The documents are then ranked according to the conditional probability $P(Q|D)$, the probability of generating the query Q given the document D.

In our system we employ a sentence-based retrieval approach similar to that described in [5], where each document comprises only one sentence. Due to lack of data to train the sentence specific LMs, all words are treated as independent, and a unigram model is applied,

$$P(Q|S) = \prod_{i=1}^{|Q|} P(q_i|S), \tag{1}$$

where q_i is the ith query term in the query $Q = (q_1 ... q_{|Q|})$ composed of $|Q|$ query terms. Throughout this paper we calculate the probability of a query term q given a sentence S in three different ways: $P_1(q|S)$, $P_2(q|S)$ and $P_3(q|S)$, as explained below.

We use absolute discounting in order to smooth the otherwise sparse LMs, where the probability of a query term q given a sentence S is calculated as:

$$P_1(q|S) \quad = \quad \frac{\max\{tf(q,S) - \delta, 0\}}{l(S)} + \frac{\delta \cdot h(S,\delta)}{l(S)} \cdot P(q|B), \tag{2}$$

where $tf(q,S)$ is the term frequency of q in S, $l(S)$ is the length (number of words) of S, δ is the discount parameter, $h(S,\delta)$ is the count of how many unique words in S have a term frequency higher than δ, and $P(q|B)$ is the unigram probability of the query term q according to the background collection model. Note that if $\delta < 1$ then $h(S,\delta)$ is equal to the number of unique words in S.

A problem with the model presented in [5] is that words relevant to the sentence might not occur in the sentence itself, but in the surrounding text. For example, for the question *"Who is Tom Cruise married to?"*, the sentence *"He is married to Katie Holmes"* in an article about Tom Cruise should ideally be assigned a high probability, despite the sentence missing the words *"Tom"* and *"Cruise"*. To account for this, we train document LMs, $P_1(q|D)$, in the same manner as for $P_1(q|S)$ in Eq. (2), and perform a linear interpolation between $P_1(q|S)$ and $P_1(q|D)$:

$$P_2(q|S) = (1-\alpha) \cdot P_1(q|S) + \alpha \cdot P_1(q|D), \tag{3}$$

where $0 \leq \alpha \leq 1$ is an interpolation parameter.

2.2 Query expansion

In order to help further improve QA performance we experiment with a global query expansion method in which words are grouped into a set $C = \{c_1 ... c_{|C|}\}$ of $|C|$ overlapping classes beforehand, and calculate the unigram class model probability of a query term q given a sentence S as follows:

$$P_C(q|S) = \sum_{j=1}^{|C|} P(q|c_j) \cdot P(c_j|S), \tag{4}$$

where $P(q|c_j) = 1/|c_j|$ if $q \in c_j$, else $P(q|c_j) = 0$, where $|c_j|$ is the number of words in c_j. $P(c_j|S)$ can be re-written as a sum over the $|V|$ words in the vocabulary $V = \{w_1 ... w_{|V|}\}$:

$$P(c_j|S) = \sum_{k=1}^{|V|} P(c_j|w_k) \cdot P(w_k|S), \tag{5}$$

where $P(c_j|w_k) = 1/N(w_k, C)$ if $w_k \in c_j$, else $P(c_j|w_k) = 0$ where $N(w_k, C)$ is the number of classes in C where w_k occurs. $P(w_k|S)$ is the unigram probability of the word w_k given the sentence S.

The word LM in Eq.(2) and the class LM in Eq.(4) are combined using linear interpolation:

$$P_{int}(q|S) = (1 - \beta) \cdot P_1(q|S) + \beta \cdot P_C(q|S), \tag{6}$$

where $0 \leq \beta \leq 1$ is an interpolation parameter.

$P_{int}(q|D)$ is calculated in a similar manner. Eq.(2) is then adjusted to give $P_3(q|S)$ as follows:

$$P_3(q|S) = (1 - \gamma) \cdot P_{int}(q|S) + \gamma \cdot P_{int}(q|D), \tag{7}$$

where $0 \leq \gamma \leq 1$ is an interpolation parameter. For all QAst evaluation runs, either P_2 or P_3 were used.

2.3 Answer Extraction

The AE module models the probability of an answer A given a question Q as:

$$P(A|Q) = P(A|W, X), \tag{8}$$

where W is a set of features describing the question-type part of Q, such as *"when"*, *"why"* and *"how"*, etc., while X is a set of features describing the information-bearing part of of Q, i.e. what the question is 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 parts are identical while the question-type parts differ. Finding the best answer $\hat{A}$ involves a search over all A for the one which maximizes the probability of the above model:

$$\hat{A} = \arg\max_A P(A|W, X). \tag{9}$$

Using Bayes' rule and making various conditional independence and uniform prior distribution assumptions, Eq. (9) can be rearranged to give:

$$\arg\max_A P(A|X) \cdot P(W|A), \tag{10}$$

where $P(A|X)$ is termed the answer retrieval model and $P(W|A)$ the answer filter model. $P(A|X)$ essentially models the proximity of A to features in X. $P(W|A)$ can be viewed as a LM that models the probability of the question-type features W given a candidate answer A.

We will not examine the answer retrieval model and the answer filter model further, see [15] for further details.

3 Experimental Setup for QAst

We participated in task T1: QA in manual transcriptions of lectures, and task T2: QA in automatic transcriptions of lectures. For the official evaluation we used the data released for the QAst evaluation task T1 and task T2. This data comprised a development set and an evaluation set with characteristics described in Table 1. The development set consisted of manual transcripts (MAN) and ASR-based transcripts (ASR) for 10 lectures, a set of questions, and a set of answers for each transcript set. The evaluation set consisted of MAN and ASR for 15 lectures, and a set of 100 questions. The development and evaluation data did not overlap. All questions were of one of the following answer types: *person*, *location*, *organization*, *language*, *system/method*, *measure*, *time*, *color*, *shape*, and *material.* Word lattices were also made available however, after preliminary experiments with the development data revealed minor inconsistencies between the lattices and ASR, we chose not to use any of the lattices in the actual evaluation. No audio was provided.

We cleaned the data by automatically removing fillers and pauses, and performed simple text processing of abbreviations and numerical expressions using perl's Lingua CPAN module to ensure consistency between ASR, MAN, questions and answers. ASR documents were sentence segmented

Data Set	#Lect.	#Sent.	#Words	WER	#Quest.
Dev. Set	10	2966	54633	32%	45
Eval. Set	15	2917	50986	28%	86

Table 1: Number of lectures, number of sentences, number of words, word error rate and number of questions for each data set after preprocessing.

according to the sentence boundaries provided, and MAN was sentence segmented using an in-house segmenter developed by one of the authors. Our system is not able to identify whether the answer to a question can be found in the corpus, thus we chose never to return a *"nil"* response for any question.

For retrieval purposes we filtered out question-type words and stop words (in total 28 words) from the questions. Using the remaining words as query terms, we ranked sentences according to either $P_2(q|S)$ or $P_3(q|S)$, depending on the run. We optimized weights on the development set and used these weights for the official evaluation.

Classes for query expansion were generated based on the overlap in features, which are computed using standard mutual information techniques, for each word in the vocabulary based on a large text corpus.

4 QAst Evaluation Results

Question sets for both task T1 and task T2 comprised the same 100 factoid questions, however 2 of these questions were deemed faulty by the coordinators following submissions and were removed prior to making assessments, resulting in a total of 98 evaluation questions. Our system returned a maximum of 5 answer candidates per question per run. We submitted two (2) runs each for task T1 and task T2. For both tasks, $P_2(q|S)$ was used for the first run and $P_3(q|S)$ was used for the second run. In addition to our group, four other teams participated. Table 2 details the official best-run results for the entire field for task T1.

Team ID	Questions Returned	Top5	MRR	%Top1
clt1	98	16	0.09	0.06
dfki1	98	19	0.17	0.15
limsi2	98	56	0.46	0.39
tokyo2	98	34	0.20	0.14
upc1	98	54	0.53	0.51

Table 2: Out of 98, number of correct answers in the top 5, MRR, and top1 answer accuracy for all participants on the manual transcription task, T1. Our team name is tokyo2.

As can be seen in table 2 our system achieved a best run MRR of 0.20, and was able to correctly answer 34 of 98 questions on the manual data set, placing us third overall. Results for the ASR transcripts were lower, as expected, at 18 correct answers for 98 questions, however other systems showed similar losses on the ASR data. Table 3 shows a comparison of our group's manual versus ASR results by submission. $P_2(q|S)$ was used for runs tokyo1_t1 and tokyo1_t2, while $P_3(q|S)$ was used for runs tokyo2_t1 and tokyo2_t2. As can be seen, query expansion employed by $P_3(q|S)$ slightly improved our Top5 scores, but had no effect on Top1 accuracy. There was a performance drop of approximately 44% for results based on the top 5 answers using $P_3(q|S)$ and a drop of approximately 43% for results based on the top 1 answer for both $P_2(q|S)$ and $P_3(q|S)$. Similar drops were reflected in other participants results however, and we suspect that this primarily reflects ASR errors.

tokyo2	MAN	ASR	Perf.Loss
Top5(P_2)	32	17	47%
Top5(P_3)	34	18	44%
Top1(P_2)	14	8	43%
Top1(P_3)	14	8	43%
MRR	0.20	0.12	N/A

Table 3: Comparison of our manual and ASR results, and relative performance loss due to ASR using P_2 (tokyo1) and P_3 (tokyo2).

Table 4 gives an accuracy break-down by answer type for task T1 and T2 for both of our submissions. Table 4 also shows that performance loss is generally consistent regardless of answer type, which serves as further evidence that the loss is due primarily to ASR errors. In several of these cases the answers to questions in the ASR transcripts were represented by mis-recognized tokens, this considerably magnified the difficulty of extracting the proper token for the question.

run tokyo2	**org**	**per**	**loc**	**tim**	**mea**	**met**	**lan**	total
totals	20	9	9	10	28	18	4	98
task T1	6	5	4	0	12	5	2	34
task T2	2	3	1	0	8	2	2	18

Table 4: Break-down of results by answer type and task. **org**=organization, **per**=person, **loc**=location, **tim**=time, **mea**=measure, **met**=method, **lan**=language.

5 Discussion and Analysis

Our results from task T1 compare favorably with results from previous CLEF and TREC evaluations, despite the size and relative lack of redundancy in the target CHIL lectures corpus. Additional experiments on this corpus which are documented in a paper that is currently pending publication show that our system is able to correctly select the sentence containing the answer over 50% of the time, indicating that there is upwards of a 20% performance loss between the sentence retrieval and answer extraction stages.

While performance across different answer types was fairly consistent, there was a conspicuous gap for the **time** type, where we did not answer any of the related questions correctly. Analysis of the data indicates that this was caused by multiple factors. There were two **time** questions for which there was no appropriate answer in the document corpus. There was also a problem with automatically normalizing complex dates which the perl Lingua module was not particular consistent, and as our system generally performs better when times and dates are represented as digits, this made it difficult to correctly extract answers such as “nineteen ninety-eight”. Finally, there was at least one **time** question for which the question itself did not clearly specify the type.

Finally, we observed a considerable drop in performance between task T1 and task T2, which was similarly mirrored in all other participants’ results. We surmise that in our case this was mainly due to answer typing issues resulting from ASR errors since answer words of the correct answer type are crucial for good AE performance in our system. This can be explained by the way the answer filter model (Section 2.3) works: if the answer words in ASR are of the wrong answer type, then $P(W|A)$ will assign a low probability to the correct answer candidate.

6 Conclusion

In this paper we have presented our results from the CLEF 2007 QAst pilot track for task T1 and T2, and described our system and experimental setup for the evaluation. In general our results compare favorably with past evaluations, and place us in the middle of the field for this evaluation. We noticed considerable performance drops between the manual transcripts and ASR transcripts, but because these drops were consistent across submissions and participants we are led to believe that this is mainly a result of ASR errors. In future evaluations we think it would be preferable to supply both recognition lattices which consistently match the ASR transcripts, and to be able to use the actual audio. Given that the real aim of this track is to find answers to natural language, factoid questions in spoken documents, having access to these resources might provide greater opportunities for teams to directly exploit the source data in more interesting way.

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

[5] A. Merkel and D. Klakow. Comparing Improved Language Models for Sentence Retrieval in Question Answering. In *Proceedings of CLIN*, 2007.

[6] J. Novak, E. Whittaker, M. Heie, S. Imai, and S. Furui. NTCIR-6 CLQA Question Answering Experiments at the Tokyo Institute of Technology. In *Proceedings of the NTCIR-6 Conference*, 2006.

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

[8] 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.

[9] 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.

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

[11] 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.

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

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

[14] E. Whittaker, J. Novak, P. Chatain, P. Dixon, M. Heie, and S. Furui. CLEF2006 Question Answering Experiments at Tokyo Institute of Technology. In *CLEF 2006, LNCS 4730 proceedings*, 2006.

[15] E. Whittaker, J. Novak, P. Chatain, and S. Furui. TREC 2006 Question Answering Experiments at Tokyo Institute of Technology. In *Proceedings of TREC-15*, 2006.