CLEF 2009 Question Answering Experiments at Tokyo Institute of Technology

Matthias H. Heie, Josef R. Novak, Edward W. D. Whittaker and Sadaoki Furui
Tokyo Institute of Technology
{heie,novakj,edw,furui}@furui.cs.titech.ac.jp

Abstract

In this paper we describe the experiments carried out at Tokyo Institute of Technology for the CLEF 2009 Question Answering on Speech Transcriptions (QAST) task, where we participated in the English track. We apply a non-linguistic, data-driven approach to Question Answering (QA). Relevant sentences are first retrieved from the supplied corpus, using a language model based sentence retrieval module. Our probabilistic answer extraction module then pinpoints exact answers in these sentences. In this year's QAST task the question set contains both factoid and non-factoid questions, where the non-factoid questions ask for definitions of given named entities. We do not make any adjustments of our factoid QA system to account for non-factoid questions. Moreover, we are presented with the challenge of searching for the right answer in a relatively small corpus. Our system is built to take advantage of redundant information in large corpora, however, in this task such redundancy is not available. The results show that our QA framework does not perform well on this task: we end last of four participating teams in seven out of eight runs. However, our performance does not regress as automatic transcriptions of speeches or questions are used instead of manual transcriptions. Thus the only run in which we are not placed last, is the most difficult task, where spoken questions and ASR transcriptions with high WER are used.

Categories and Subject Descriptors

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

General Terms

Measurement, Performance, Experimentation

Keywords

Question answering, Questions beyond factoids

1 Introduction

In this paper we describe the application of our data-driven and non-linguistic framework for the CLEF 2009 QAST task. Two sets of questions were given: written questions and manual transcriptions of the corresponding spoken questions. The corpus consisted of transcriptions of European Parliament Plenary sessions in English. The question set contained *Definition* questions and *Factoid* questions. 4 versions of the corpus were available: manual transcriptions and 3 ASR transcriptions. We submitted one answer set for each combination of question sets and corpora, in total 8 submissions.

Our system is a factoid QA system that we previously have participated with in other QA evaluations [1][2][3][4]. 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, as well as simple rules for pre-processing of the data, makes it possible to rapidly develop new systems for a wide variety of different languages and domains. Due to its data-driven nature, our QA system performs best when there is a large corpus available, containing several co-occurrences of question words and the correct answer. Thus the QAST task presented a challenge to us due to the small size of the corpus. Moreover, we made no adjustments to our factoid QA system to account for *Definition* questions, i.e. we treated *Definition* questions as *Factoid* questions.

The system comprises two main components, an Information Retrieval (IR) module used to retrieve relevant sentences from a corpus, and an Answer Extraction (AE) module which is used to identify and rank exact answers in the sentences returned by the IR module. For IR we used language model based sentence retrieval. In this approach, a language model (LM) is generated for each sentence and these models are combined with document LMs to take advantage of contextual information. From the retrieved information, we extract rigid answers using our answer filter model.

2 Sentence retrieval

Language modeling for IR has gained in popularity over the last decade since the approach was proposed [5]. Under this approach a LM is estimated for each document. The documents are then ranked according to the conditional probability $P(Q \mid D)$, the probability of generating the query Q given the document D.

We rank sentences correspondingly [6]. Due to lack of data to train the sentence specific LM, it is assumed that all words are independent, hence unigrams are used:

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

where q_i is the ith query term in the query $Q = (q_1 ... q_{|Q|})$ composed of $|Q|$ query terms.

Smoothing methods are normally employed with LMs to avoid the problem of zero probabilities when one of the query terms does not occur in the document. This is typically achieved by redistributing probability mass from the document model to a background collection model $P(Q \mid C)$. We use Dirichlet prior, where the probability of a query term q given a sentence S is calculated as:

$$P_1(q \mid S) \quad = \quad \frac{c(q;S) + \mu \cdot p(q \mid C)}{\sum_w c(w;S) + \mu}, \tag{2}$$

where $c(q;S)$ is the count of query term q in sentence S, μ is a smoothing parameter, $p(q \mid C)$ is the unigram probability of q according to the background collection model and $\sum_w c(w;S)$ is the count of all words in S.

A problem with this model is that words relevant to the sentence might not occur in the sentence itself, but in the surrounding text. For example, for the question *Where was George Bush born?*, the sentence *He was born in Connecticut* in an article about George Bush should ideally be assigned a high probability, despite the sentence missing important query terms. To account for this, we train document LMs, $P_1(q \mid D)$, in the same manner as for $P_1(q \mid S)$ in Eq. (2), and perform a linear interpolation between $P_1(q \mid S)$ and $P_1(q \mid D)$:

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

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

3 Answer extraction

For answer extraction we use the framework described in detail in [7]. We model the most straightforward and obvious dependence of the probability of an answer A depending on a question Q:

$$P(A \mid Q) = P(A \mid W, X), \tag{4}$$

where A and Q are considered to be strings of l_A words $A = a_1, \ldots, a_{l_A}$ and l_Q words $Q = q_1, \ldots, q_{l_Q}$, respectively. Here $W = w_1, \ldots, w_{l_W}$ represents a set of features describing the "question-type" part of Q such as *when*, *why*, *how*, etc. while $X = x_1, \ldots, x_{l_X}$ represents a set of features that describe 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 available A for the one which maximizes the probability of the above model, i.e.,

$$\hat{A} = \arg\max_A P(A \mid W, X). \tag{5}$$

Given the correct probability distribution, this is guaranteed to give us the optimal answer in a maximum likelihood sense. We don't know this distribution and it is still difficult to model but, using Bayes' rule and making various simplifying, modeling and conditional independence assumptions (as described in detail in [7]) Eq. (5) can be rearranged to give

$$\arg\max_A \underbrace{P(A \mid X)}_{\substack{answer \\ retrieval \\ model}} \cdot \underbrace{P(W \mid A)}_{\substack{answer \\ filter \\ model}}. \tag{6}$$

The $P(A \mid X)$ model we call the *answer retrieval model*. In this year's evaluation we didn't use the answer retrieval model, i.e. $P(A \mid X)$ is uniform.

The $P(W \mid A)$ model matches a potential answer A with features in the question-type set W. For example, it relates place names with *where*-type questions. We call this component the *answer filter model* and it is structured as follows.

The question-type feature set $W = w_1, \ldots, w_{l_W}$ is constructed by extracting n-tuples ($n = 1, 2, \ldots$) such as *Who*, *Where* and *In what* from the input question Q. A set of single-word features is extracted based on frequency of occurrence in our collection of example questions.

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. In order to construct these classes, given a set E of example q-and-a, we then define a mapping function $f : E \mapsto C_E$ which maps each example q-and-a t_j for $j = 1 \ldots |E|$ into a particular class $f(t_j) = e$. Thus each class c_e may be defined as the union of all component q-and-a features from each t_j satisfying $f(t_j) = e$. Finally, to facilitate modeling we say that W is conditionally independent of c_e given A so that

$$P(W \mid A) \quad = \quad \sum_{e=1}^{|C_E|} P(W \mid c_W^e) \cdot P(c_A^e \mid A), \tag{7}$$

where c_W^e and c_A^e refer respectively to the subsets of question-type features and example answers for the class c_e.

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:

Transcriptions			Run ID	
			Written	Spoken
ID	Type	WER	questions	questions
m	Manual	-	a_m	b_m
a	ASR	10.6%	a_a	b_a
b	ASR	14.0%	a_b	b_b
c	ASR	24.1%	a_c	b_c

Table 1: Details of the 4 transcriptions and 8 runs.

Type	Subtype	#questions
Factoid	Person	17
	Organisation	17
	Location	14
	Time	25
	Measure	2
Definition	Person	12
	Organisation	3
	Other	10

Table 2: Number of questions of each question type, 100 in total. Of these, 19 have no answer in the corpus and should be answered *NIL*.

$$P(W \mid A) = \sum_{e=1}^{|C_E|} P(W \mid c_e) \cdot \prod_{j=1}^{l_{A^e}} P(a_j^e \mid a_j). \tag{8}$$

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:

$$P(W \mid A) = \sum_{e=1}^{|C_E|} P(W \mid c_e) \prod_{j=1}^{l_{A^e}} \sum_{k=1}^{|C_A|} P(a_j^e \mid c_k) P(c_k \mid a_j), \tag{9}$$

where c_k 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 Eq. (9)) and normalize $P(W \mid A)$ accordingly.

4 Experimental work

For QAST 2009, two sets of questions were given: 100 written questions and manual transcriptions of the corresponding spoken questions. The answers were to be extracted from transcriptions of European Parliament Plenary sessions in English (TC-STAR05 EPPS English corpus), which consists of 6 spoken documents, transcribed from 3 hours of recordings. 4 versions of the corpus were available: manual transcriptions and 3 ASR transcriptions. There were one run for each of the possible combinations of question sets and transcriptions, thus there were $2 \times 4 = 8$ runs, as shown in in Table 1.

Two main types of questions were considered: *Factoid* questions and *Definition* questions. The *Factoid* questions were further divided into the following types: *Person*, *Organisation*, *Location*, *Time* and *Measure*. The *Definition* questions were of the following types: *Person*, *Organisation* and *Other*. Questions where an answer cannot be found in the corpus, were to be answered by *NIL*. Details are given in Table 2.

We cleaned the data by automatically removing fillers and pauses, and performed simple text processing of abbreviations and numerical expressions to ensure consistency between the different

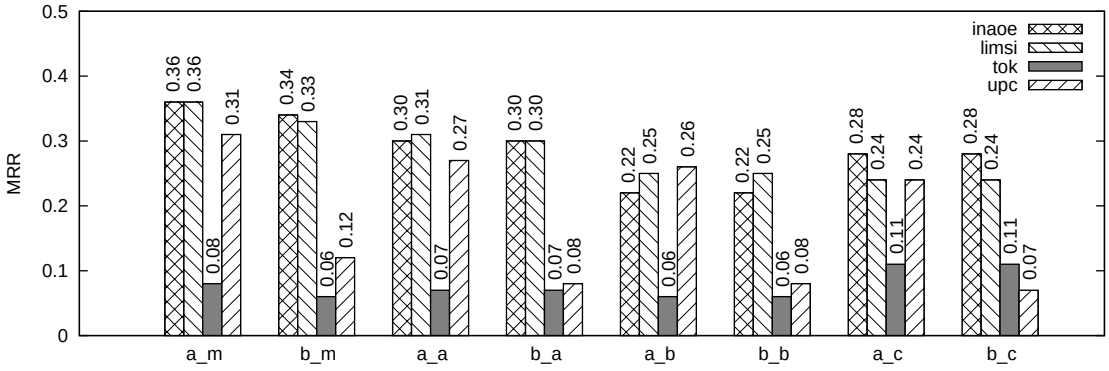

Figure 1: Results of each run for all teams. Our team id is `tok`

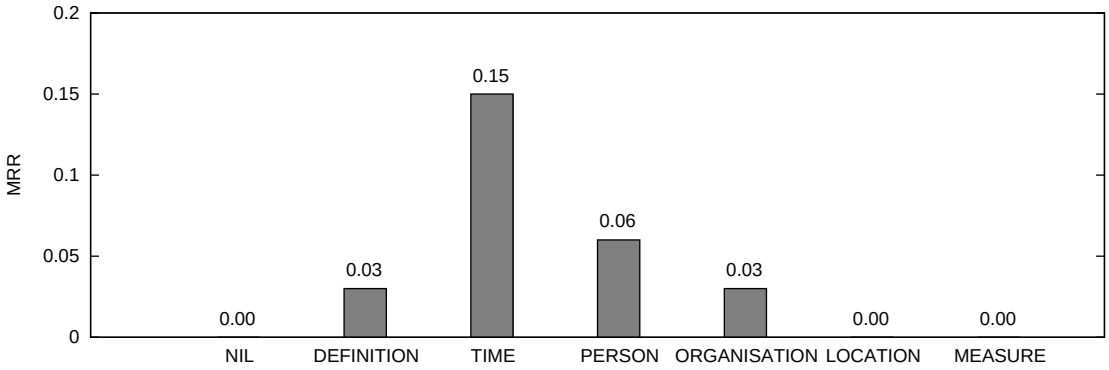

Figure 2: Results of our `a_m` run, by type. In this figure *NIL* and *Definition* are not further divided into subtypes. The other types are subtypes of *Factoid*.

question sets and transcriptions. The ASR transcriptions lacked sentence boundaries, unlike the manual transcriptions, where punctuation was provided. We sentence segmented the ASR transcriptions by automatically aligning the text with the manual transcriptions using the `GNU sdiff` tool. A set of stop words was also used. The top 100 sentences and their contexts (the immediately preceding and succeeding sentence), were passed to the answer extraction module. The top 5 answer candidates, as ranked by the AE module, were submitted for evaluation. Although each team was allowed to submit two answer sets for each run, we decided to submit only one per run.

5 Results and Discussion

The results of each team's best submission for each run are plotted in Figure 1. These results show that we end up last of the four teams in all but one run. Our performance does not regress as automatic transcriptions of speeches or questions are used instead of manual transcriptions. Thus the only run in which we are not placed last, is the most difficult task: `b_c`.

Since our system does not treat automatic transcriptions differently from manual transcriptions (except in the pre-processing stage), we restrict ourselves to further analyzing run `a_m`, in which written questions and manual transcriptions are used. Figure 2 shows the break-down of the results by answer type for this run.

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. Therefore the score for *NIL* questions is zero. Furthermore, since the system is a factoid QA system, we could in advance predict a low

score for *Definition* questions. For *Factoid* questions we achieve the highest performance on *Time* questions, which has an MRR which is more than double that of the MRR for the *Person* type, the second best question type. This might be explained by the fairly restricted format of *Time* answers and the efforts we made on date normalization. *Organisation* questions are difficult to answer since there is little restrictions on the format of organisation names. The zero score for *Location* score is more disappointing, since those answers mostly consist of a single geographical term. The score for *Measure* questions yields little information, since there were only 2 such questions.

Normally our QA system utilizes a large corpus, such as the Web, and the more often an answer candidate occurs in the context of query terms, the more likely it is to be considered a correct answer. However, in this task the corpus was small, thus we are not able to benefit from such redundancy, which might be an explanatory factor for our low performance.

6 Conclusion

In this paper we have given an overview of our methods and results for the CLEF 2009 Question Answering on Speech Transcriptions evaluation. The results show that, using our QA system, we are not able to achieve good performance on this task. Obvious explanations are the presence of non-factoid questions, which our system is not built to answer, in addition to our inability to identify questions which have no answer in the given corpus. Another possible reason is the small size of the corpus, which means our system cannot take advantage of redundant answer information.

References

[1] Whittaker, E., Chatain, P., Furui, S. and Klakow, D., "TREC2005 Question Answering Experiments at Tokyo Institute of Technology", *Proc. TREC-14*, 2005.

[2] Whittaker, E., Novak, J., Chatain, P. and Furui, S., "TREC2006 Question Answering Experiments at Tokyo Institute of Technology", *Proc. TREC-15*, 2006.

[3] Whittaker, Heie, M., Novak, J. and Furui, S., "TREC2007 Question Answering Experiments at Tokyo Institute of Technology", *Proc. TREC-16*, 2007.

[4] Heie, M., Whittaker, E., Novak, J., Mrozinski, J. and Furui, S., "TAC2008 Question Answering Experiments at Tokyo Institute of Technology", *Proc. TAC*, 2008.

[5] Ponte, J. and Croft, W., "A Language Modeling Approach to Information Retrieval", *Proc. SIGIR*, 1998, pp. 275-281.

[6] Heie, M., Whittaker, E., Novak, J. and Furui, S., "A Language Modeling Approach to Question Answering on Speech transcriptions", *Proc. ASRU*, 2007, pp. 219-224.

[7] Whittaker, E., Furui, S. and Klakow, D., "A Statistical Pattern Recognition Approach to Question Answering using Web Data", *Proc. Cyberworlds*, 2005, pp. 421-428.