論文 / 著書情報
Article / Book Information

論題(和文)	
Title(English)	Question Answering Experiments on Speech Transcripts at Tokyo Institute of Technology
著者(和文)	JOSEF R NOVAK, ウィッタッカー エドワード, ﾍｲｴ ﾏﾃｨｱｽ, 古井 貞熙
Authors(English)	Josef Novak, Edward Whittaker, Matthias Heie, Sadaoki Furui
出典(和文)	日本音響学会 2007年秋季講演論文集, Vol. , No. 1-3-7, pp. 25-26
Citation(English)	, Vol. , No. 1-3-7, pp. 25-26
発行日 / Pub. date	2007, 9

Question Answering Experiments on Speech Transcripts at Tokyo Institute of Technology. *

◎ Josef Novak, Edward Whittaker, Matthias Heie, Sadaoki Furui (Tokyo Institute of Technology)

1 Introduction

This year the Cross Language Evaluation Forum (CLEF) initiated a new pilot track, "Question Answering on speech transcripts" (QAst), which asks participants to answer questions about spoken documents. The general aim of this new track is to promote QA research on spontaneous speech corpora, and other spoken document collections[1, 2].

CLEF has offered a Multiple Language Question Answering track (QA@CLEF) since 2003, which has traditionally focused on restricted-domain factoid-based question-answering for various European source and target languages. This new track breaks from that tradition in a couple of interesting ways. Below we describe the new track, and the results of our official runs achieving 0.21 MRR on manual, and 0.12 MRR on ASR transcripts.

2 Track Motivation

QA research has traditionally focused on extracting answers to questions from written-text document collections. However most human-human interactions occur through spontaneous speech, e.g. meetings, lectures, telephone conversations, etc. Furthermore, there is a large and growing wealth of such information resources. Research into how best to exploit these resources is an important step towards creating useful, real-world QA solutions and applications, and towards further expanding the universality and availability of knowledge resources.

With this in mind, the aim of the QAst track is to provide a QA evaluation framework which targets spoken documents and corresponding transcripts, and more generally to promote QA research on similar knowledge resources.

2.1 Track Objective

The practical objective of the QAst pilot track was to return token pointers to answer locations in audio streams, based on corresponding transcripts. Thus **a:** questions were formulated on the basis of the actual audio, not the ASR-based or manual transcripts and **b:** an acceptable answer was defined as a minimal sequence of word tokens pointing to the correct answer in the audio stream. Thus, in the following example,

Question: Where did Jacques Chiraq go?
Relevant audio: "Jacques Chiraq went to Berlin."
ASR transcription: "Jacques Chiraq ate to Barcelona."

the correct answer would be "Barcelona", as it refers to the location of the correct answer in the actual audio stream. This is an interesting variation on traditional QA evaluations which usually require participants to find answers in written-text corpora which do not tend to contain the filled pauses, repetitions or other ASR errors characteristic of spontaneous speech corpora.

3 System Outline

We apply a statistically driven architecture to QA[3]. In general we treat the problem as an instance of the noisy channel model and define it,

$$\arg\max_{A} \underbrace{P(A \mid X)}_{\substack{retrieval \\ model}} \cdot \underbrace{P(W \mid A)}_{\substack{filter \\ model}}. \tag{1}$$

where A refers to the answer, X refers to the information bearing part of the question i.e., what the question is really about, and W refers to the question-type part of the question e.g., words such as how, or why. We call $P(A \mid X)$ the retrieval model and call $P(W \mid A)$ the filter model and refer the reader to [3] for an in-depth treatment of these components. Details on new sentence selection component of the retrieval model are also described in another paper currently under review.

*東工大における音声書き起こし/認識結果テキストを用いた QA 実験
ノバック・ジョセフ、ウィッタッカー・エドワード、ヘイー・マティアス、古井貞熙 (東工大)

Team	Top5	MRR	%Top1
clt1	16	0.09	0.06
dfki1	19	0.16	0.15
limsi2	56	0.45	0.39
tokyo2	36	0.22	0.16
upc1	56	0.41	0.52

Table 1 Out of 100, number of correct answers in the top 5, MRR, and top 1 answer accuracy for all participants on the manual transcription task, T1. Our team name is tokyo2.

4 Evaluation Description

For the evaluation we were asked to extract answers from manual and ASR-based transcripts. Our team participated in task 1 (QA in manual transcriptions of lectures), and task 2 (QA in automatic transcriptions of lectures). Prior to the actual evaluation, we were provided with development data for the individual subtasks in which we were participating. In the case of the lecture-based tasks, these development document collections consisted of 10 seminars and 50 questions. For the actual evaluation we were provided with an additional 15 seminars and 100 questions. There was no overlap between the development and evaluation data. Participants were allowed to submit 2 runs per task, and to provide a maximum of 5 possible answers for each question in a given run. LIMSI provided recognition lattices (not used) and the 1-best pass for each lecture. The WER for the ASR transcripts was 28%.

4.1 Setup

Prior to use, both manual and ASR transcription data were pre-processed by removing fillers, pauses, and non-character data, and normalizing abbreviations and numerical expressions.

4.2 Results

In addition to our group, four other teams participated in the pilot track. Table 1 describes the official results for each participant's best run from task 1.

Our system was able to correctly answer 36 of 100 questions on the manual data set, placing us third overall in the field. Results for the ASR transcripts were lower, at 18 correct answers for 100 questions, however we placed third here as well. Table 2 shows a comparison of our manual versus ASR results.

tokyo2	MAN	ASR	Perf.Loss
Top5	36	18	50%
Top1	16	8	50%
MRR	0.22	0.12	N/A

Table 2 Comparison of our manual and ASR results, and relative performance loss caused by ASR.

There is a performance drop of approximately 50% for results based on both the top 5 answers and the top 1 answer. Similar drops were reflected in other participants results however, we believe that these drops primarily reflect ASR errors. In several cases the answers to questions in the transcripts were represented by mis-recognized tokens, which considerably magnifies the difficulty level.

5 Conclusion

In this paper we have described the new CLEF pilot track, "Question Answering on speech transcripts", and given a brief overview of our system architecture and participation. We think this new track represents a useful and promising addition to the QA and speech research domains and hope to see similar evaluations emerge for other language sets in the near future.

Finally, in future it may be interesting to focus on natural language queries, and/or spoken queries. It may be useful to provide original audio data from the lectures, so as to better stress the real aim of the QAst track: finding answers to questions in audio streams.

6 Acknowledgments

This research was supported in part by JSPS and the 21st Century COE Programme

References

[1] http://www.lsi.upc.edu/~qast/files/Guidelines_QAst_2007_v1.2.pdf
[2] http://www.lsi.upc.edu/~qast/
[3] Whittaker et al., "CLEF 2006 Question Answering Experiments at Tokyo Institute of Technology", LNCS CLEF, 351-361, 2007