Towards the Evaluation of Voice-Activated Question Answering Systems: Spontaneous Questions for QAST 2009

Davide Buscaldi¹ and Paolo Rosso¹ and Jordi Turmo² and Pere R. Comas²

¹ Natural Language Engineering Lab., RFIA
Dpto. de Sistemas Informáticos y Computación (DSIC),
Universidad Politénica de Valencia, Spain,
{dbuscaldi,prosso}@dsic.upv.es
² TALP Research Center,
Universitat Politècnica de Catalunya
{turmo,pcomas}@lsi.upc.edu

Abstract. This is a preliminary report of the work carried out in order to introduce “spontaneous” questions into QAST at CLEF 2009. QAST (Question Answering in Speech Transcripts) is a track of the CLEF campaign. The aim of this report is to show how difficult can be to generate “spontaneous” questions and the importance to take into account the real information needs of users for the evaluation of question answering systems.

1 Introduction

In the Question Answering (QA) task, search engines have to extract concise and precise fragments of texts that contain an answer to a question posed by the user in natural language. This task is very close to what is usually considered as “automatic text understanding”.

The availability of effective QA systems may change the type of interaction between humans and machines, mainly thanks to the fact that in QA the user obtains an answer and not a list of documents to be browsed. Competitions like CLEF [http://www.clef-campaign.org/] have been created in order to develop and improve existing systems and to evaluate and compare their behavior.

A potentially interesting evolution of QA systems is their application to spoken documents. Until now, most QA research has focused on mining document collections containing written texts to answer written questions [2]. They usually share a decent writing quality, at least grammar-wise. In addition to these written sources, a lot of potentially interesting information appears in spoken documents, such as broadcast news, speeches, seminars, meetings or telephone conversations. The QAST track has been introduced with the objective of investigating the problem of question answering in such audio documents [4].

A major step towards a more natural type of interaction between machines and users will be represented by the introduction of Speech Language Technologies (SLT) into QA systems [1]. Automatic Speech Recognition (ASR) and the
development of spoken human-machine interfaces are currently considered mature enough to be used in most common applications. Some examples of these systems are represented by spoken QA systems that could be used by means of mobile devices.

The realization of this kind of systems will meet with a series of issues, due to the nature of the input medium, that is particularly sensitive to errors. We propose to introduce spontaneous oral questions into QAST 2009. In most QA evaluations, questions have been posed in good quality, without errors and with a clear definition of focus and topic. Voice-activated systems will have to face with questions that include misspelled names, pauses, hesitations.

Unfortunately, until now, a corpus of questions of this kind has not been released. This is not only due to the usual problems for the production of corpora, especially for the time necessary to collect the questions or the money needed to develop the resource. In this report we will give an overview of the issues of producing spontaneous oral questions.

2 QAST 2009 proposal

The objective of this pilot track is to develop a framework in which QA systems can be evaluated when the answers have to be found in speech transcripts. There are three main objectives to this evaluation:

1. Motivating and driving the design of novel and robust QA architectures for speech transcripts and voice-activated systems;
2. Measuring the loss due to the inaccuracies in state-of-the-art ASR technology;
3. Measuring this loss at different ASR performance levels given by the ASR word error rate;

The collections will be composed by the 2005 and 2006 TC-STAR European Parliament Plenary Sessions (EPPS) corpora, in English and Spanish, and the ESTER French broadcast news corpus, with automatic and manual transcriptions.

The types of definitional questions for the task will be 4: Person, Organisation, Object and Other. There are 10 types of factual questions: Person, Location, Organisation, Time (including dates), Measure, System, Language, Colour, Shape and Material. Questions will be both “standard” written questions and “spontaneous” oral questions, manually transcribed.

The organisation of the track will be carried out by the UPC (Universitat Politècnica de Catalunya), together with the LIMSI (Laboratoire d’Informatique pour la Mécanique et les Sciences de l’Ingénieur) and the UPV (Universidad Politécnica de Valencia).

3 Issues in the production of spontaneous questions

Two tests were carried out with 4 Spanish speakers, who were requested to pick one or more text fragments from the EPPS transcriptions of days 15-18
Nov. 2004 and formulate two or more questions over each text. In the first test they were instructed to ask for something that they felt it was missing for the understanding of the text. In the second test they were requested to formulate their questions following the task guidelines.

In the first test, the users produced 17 questions, an average of 4.25 per user, using 6 text fragments (2.83 questions per text fragment), employing an average of 1 : 15 minutes for each questions.

In Table 1 we show a sample of the produced questions.

<table>
<thead>
<tr>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>3. Quién es Annetta Flanigan?</td>
</tr>
<tr>
<td>4. Permiten a los rehenes establecer comunicación con el exterior?</td>
</tr>
<tr>
<td>6. Porqué acusan de omicidio aaa ... Dow Chemical?</td>
</tr>
<tr>
<td>7. Cuánta gente ha muerto ... a causa de los sucesos relacionados con esta empresa?</td>
</tr>
<tr>
<td>8. Er... quién les ha hecho el embargo a China?</td>
</tr>
<tr>
<td>12. Quién lee este texto?</td>
</tr>
<tr>
<td>15. Cuál es la política del parlamento europeo sobre ... el cambio climático?</td>
</tr>
<tr>
<td>17. A quién va dirigido el texto?</td>
</tr>
</tbody>
</table>

As it can be observed from this sample, most questions do not respect guidelines, contain anaphoras (4 and 7), ask about the speaker or the audience (12 and 17) or contain hesitations, pauses and misspellings (6,7,8 and 15).

In the second test, the users were able to produce questions that better fit the guidelines, although in many cases there were still anaphoras (see Table 2 questions 10 and 15) and “meta”-questions (i.e., questions about the speech and not the content, such as question 5). The main problem, in this case, is due to the fact that most questions did not have an answer in the collection.

<table>
<thead>
<tr>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Cómo se llama el presidente de China?</td>
</tr>
<tr>
<td>5. En qué fecha se sitúa el texto?</td>
</tr>
<tr>
<td>6. Quién es el presidente de la delegación de Estados Unidos?</td>
</tr>
<tr>
<td>9. Quién es el señor Buttiglione?</td>
</tr>
<tr>
<td>10. Cuál es su postura?</td>
</tr>
<tr>
<td>14. Quién es ... el nombre del defensor del pueblo?</td>
</tr>
<tr>
<td>15. Hay uno en el parlamento europeo para toda Europa?</td>
</tr>
</tbody>
</table>
4 Conclusions

The evaluation showed that it is possible to produce a set of questions with the methodology proposed, even if three main problems emerged:

1. Users do not stick to the guidelines even if they were told to, with the result that a number of questions cannot be used;
2. The formulated questions contain a high percentage of NIL questions;
3. The production of the text fragments could be very demanding if it is not possible to guide users to produce more “good” questions.

Acknowledgements

The introduction of spontaneous questions in QAST 2009 is the result of the collaboration between UPV and UPC in the context of the TextMESS research project (TIN2006-15265-C06).

References