EXPERIMENTS IN SPEECH DRIVEN QUESTION ANSWERING

*C. González-Ferreras, V. Cardeñoso-Payo**

Dpto. Informática
Universidad de Valladolid
47011 Valladolid, Spain
{cesargf,valen}@infor.uva.es

E. Sanchis Arnal†

Dpto. Sistemas Informáticos y Computación
Universidad Politécnica de Valencia
46022 Valencia, Spain
esanchis@dsic.upv.es

ABSTRACT

In this paper we present a system that allows users to obtain the answer to a given spoken question expressed in natural language. A large vocabulary continuous speech recognizer is used to transcribe the spoken question into text. Then, a question answering engine is used to obtain the answer to the question. Some improvements over the baseline system were proposed in order to adapt the output of the speech recognizer to the question answering engine: capitalized output from the speech recognizer and a language model for questions. System performance was evaluated using a standard question answering test suite from CLEF. Results showed that the proposed approach outperforms the baseline system both in WER and in overall system accuracy.

***Index Terms*—** speech recognition, question answering, speech driven question answering, language model adaptation

1. INTRODUCTION

In the last years important advances have been achieved in several areas of man-machine communication and in the access to unstructured information repositories. Moreover, the development of systems that integrate question answering (QA) and automatic speech recognition (ASR) technologies will allow humans to communicate with computers from a more natural and very appealing perspective. However, there are some important and specific problems which need to be solved in order to obtain good results from the integration of both technologies.

The speech recognizer must be able to handle questions from the user, and thus, must be adapted to the task of open domain question answering. The main components of a ASR system are: the acoustic models, the vocabulary of the task and the language model (LM). The acoustic models, dependent of the language, are generally independent of the task. However, both the vocabulary and the language model are strongly dependent of the task.

The majority of the currently available QA systems are based on the detection of specific keywords, mostly Named Entities (NE). For instance, for the CLEF question "*What is the capital of Croatia?*", a failure in the detection of the NE "*Croatia*" would make impossible to find the answer. Then, the vocabulary of the ASR system must contain the set of NE that can appear in user questions. But the number of different NE in a standard QA task could be huge.

Related to the vocabulary of the ASR system, interrogative words play an important role. Errors in wh-words present in the questions as "*Who*", "*When*", ... can be very determinant in the question classification process. Then, the ASR system should provide good recognition rates on this set of words.

Another problem that affects these systems, like any other system which makes use of speech recognition, is the incorrect pronunciation of NEs (such as names of persons or places) when the NE is in a language different than the user's one. This problem does not allow the use of a typical phonetic/orthographic transcriptor. Another grapheme to phoneme mechanism is needed, considering alternative pronunciations of the same word or acronym.

Training the language model of the ASR system also poses some challenges. The language model provides constraints on the sequence of words that are allowed to be recognized and is a basic component of the ASR system. An important aspect is to determine how the language model has to be learned: user queries to a QA system present a specific syntax but, in general, a large enough number of samples of this kind of sentences from which obtaining robust models is not available.

In this work we present an approach to speech driven question answering in which we study the problem of adapting the output of the speech recognizer to the question answering engine. The system works for Spanish language.

The structure of the paper is as follows: section 2 presents some related work; section 3 explains the system in detail; section 4 describes the experiments and the analysis of results; conclusions and future work are presented in section 5.

2. RELATED WORK

There has been some work on speech driven question answering during the last years. Most of the work was based on the integration of automatic speech recognition and text QA systems and the main concern was how to reduce the effect of speech recognition errors on the QA performance.

In [1] a spoken interactive QA system was presented. They proposed two mechanisms to tackle with the difficulties inherent to the spoken input: a screening filter and a set of disambiguation questions. The screening filter tried to extract meaningful information from the recognized sentence. When the QA engine could not extract an appropriate answer to user's question, this question was considered ambiguous and an interaction with the user was carried out to ask him for additional information.

Another strategy for making a dialog with the user to avoid the effect of speech recognition errors on the QA was presented in [2]. After recognizing each query, keywords were automatically

*This work has been partially supported by *Consejería de Educación de la Junta de Castilla y León* under project number VA077A08.

†This work has been partially supported by the Spanish MEC and FEDER under contract TIN2005-08660-C04-02.

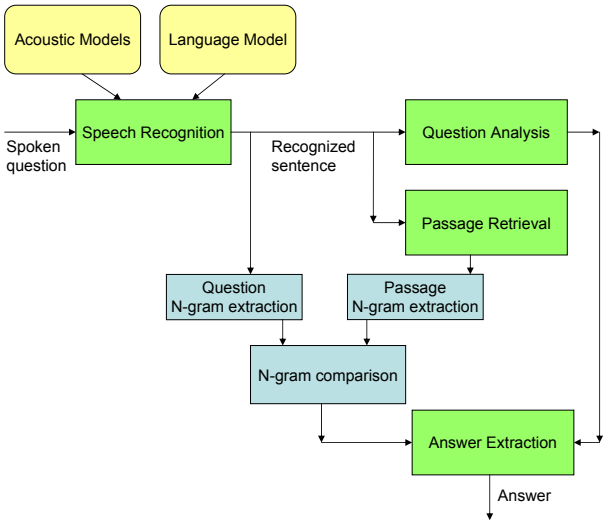

Fig. 1. *Architecture of the system.*

extracted and displayed on a graphical user interface. Multiple language models for recognizing spoken queries were trained using clustered newspaper articles; five domains were defined. Each domain-dependent language model was interpolated with a specific language model trained from a set of transcribed question utterances.

The Voice Activated QA (VAQA) was presented in [3]. First, the ASR generated not only the transcribed sentence but also a word lattice as the output of the ASR process. Then, a special filtering mechanism used both the question transcription and the word lattice to filter out words that could not be processed by a typical QA system due to syntactic, semantic or pragmatic inconsistencies. The result was a word lattice of smaller dimensions, which was used for generating an enhanced language model used to reprocess the spoken question. The output was processed by the QA system to obtain the final answer.

In [4] a method to deal with out of vocabulary (OOV) words in the ASR system was presented. The vocabulary defined for the language model included not only the 20,000 high-frequency words, but also a set of syllables. For speech recognition, a word could be formed from a word or a sequence of syllables. Then, the result of the recognition process was a sequence of words and sequences of syllables. The QA system worked in two steps. In the first step, the input was filtered so that only words were considered, and a specific number of top-ranked documents was obtained, which will be used for the search in the next step. In the second step, they replaced detected OOV words with index terms that were phonetically identical or similar and re-perform text retrieval.

3. SYSTEM OVERVIEW

The system consists of two components: a speech recognizer and a question answering engine. The objective of the system is to obtain an exact answer to a given spoken question. The architecture of the system is shown in figure 1. First, the user makes a question using speech. Next, the spoken question is transcribed into text by the speech recognizer. Finally, the answer to the question is obtained by the question answering engine.

3.1. Speech recognition

SONIC, the University of Colorado large vocabulary continuous speech recognizer was used for speech recognition [5]. It is based on continuous density hidden Markov model (CDHMM) technology and implements a two-pass search strategy. We trained acoustic and language models for Spanish language.

Acoustic models were triphone HMMs with associated gamma probability density functions to model state durations. Standard feature extraction was used: 12 Mel Frequency Cepstral Coefficients (MFCC) and normalized log energy, along with the first and second order derivatives. We used Albayzin corpus to train the acoustic models [6] (13,600 sentences read by 304 speakers).

We created a word based trigram language model using SRILM toolkit [7], with Katz backoff for smoothing. The target document collection was used to train the language model, because this can result in an adaptation of the LVCSR to the given task and provides better system performance. The document collection was composed of two years of newswire news from EFE news agency (1994 and 1995). We used a vocabulary of 60,000 words, that was created selecting the most frequent words found in the documents. The pronunciation lexicon was built using a rule based system for Spanish.

The effectiveness of this configuration has already been proved in our previous experiments on speech driven information retrieval [8].

3.2. Question answering

For question answering we used QUASAR [9] (QUestion AnSwering And information Retrieval). The system has been tested successfully in recent editions of CLEF QA track. Although the system was originally designed without a spoken interface, it has been used to study the impact of speech recognition errors on the search of answers [10].

The main advantage of using QUASAR is that the influence of speech recognition errors is minimized due to the fact that the passage retrieval and the answer extraction are based on keywords (or relevant sequences of words), and it is not necessary a deep syntax analysis that will be impossible in a sentence with some misrecognition errors.

3.2.1. Question analysis

Different question types were defined in order to classify the questions:

- NAME: Acronym, Person, Title, Firstname, Location (country, city, geographical).
- DEFINITION: Person, Organization, Object.
- DATE: Day, Month, Year, Weekday.
- QUANTITY: Money, Dimension, Age.

Each category was defined by one or more patterns written as regular expressions. The questions that do not match any defined pattern are labeled with OTHER. If a question matches more than one pattern, it is assigned the label of the longest matching pattern (i.e., we consider longest patterns to be less generic than shorter ones).

The question analyzer also identifies some constraints that are used in the answer extraction phase. These constraints are made by sequences of words extracted from the POS-tagged query by means of POS patterns and rules. For instance, any sequence of nouns (such as "*ozone hole*") is considered as a relevant pattern.

There are two classes of constraints. There is always one *target* constraint, which is the word of the question that should appear closest to the answer string in a passage, and zero or more *contextual* constraints, keeping the information that has to be included in the retrieved passage in order to have a chance of success in extracting the correct answer. For example, in the following question: "*Where did the Winter Olympic games of 1994 take place?*" "*Winter Olympic games*" is the target constraint, while "*1994*" is the contextual constraint.

3.2.2. Passage retrieval

For passage retrieval (PR) we used JIRS [9] (JAVA Information Retrieval System). The purpose of PR is to obtain the passages with the greatest probability of containing the correct answer.

The ranking of the passages is done by means of a ranking function based on an n-gram similitude measure between the question and the passages. This measure takes into account the fact that it is more likely to find the right answer in passages that share more and longer n-gram structures with the question. For instance, if we ask "*Who is the President of Venezuela?*" the system could retrieve two passages: one with the expression "*...Hugo Chávez is the President of Venezuela...*", and other with the expression "*...Nicholas Sarkozy is the President of France...*". Of course, the first passage must have more importance because it contains the 5-gram "*is the President of Venezuela*", whereas the second passage only contains the 4-gram "*is the President of* ".

This similarity value is calculated by:

$$Sim(p,q) = \frac{\sum_{\forall x \in Q} h(x,P) \cdot \frac{1}{d(x,x_{max})}}{\sum_{i=1}^{n} w_i} \quad (1)$$

where Q is the set of j-grams that are generated from the question q; P is the set of the greatest j-grams (i.e., those with the greatest weight) of the question q appeared in the passage p; x_{max} is the j-gram of Q with the greatest weight that appears in P; n is the number of question terms; w_i is the weight of the i-th question term.

The function $h(x,P)$ is defined by:

$$h(x,P) = \begin{cases} \sum_{k=1}^{|x|} w_k & \text{if } x \in P \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where w_1, w_2, ...,$w_{|x|}$ are the term weights of the n-gram x and are calculated as follows:

$$w_k = 1 - \frac{\log(n_k)}{1 + \log(N)} \quad (3)$$

where n_k is the number of passages in which the associated term appears and N is the number of system passages.

The function $d(x, x_{max})$ is the distance between the j-grams x and x_{max}, and is defined as the number of terms between them. In this way the j-grams that are close to x_{max} have more relevance.

3.2.3. Answer extraction

The input of this module is constituted by the passages returned by the passage retrieval module and the constraints (including the expected type of the answer) obtained through the question analysis module.

First, all the passage's substrings matching the expected answer pattern are located. Then, a weight is assigned to each found substring s, depending on the position of s with respect to the constraints, if s does not include any of the constraint words. If in the passage are present both the target constraint and one or more of the contextual constraints, then the product of the weights obtained for every constraint is used; otherwise, only the weight obtained for the constraints found in the passage is used.

A mini knowledge base is used in order to discard candidate answers which do not match an allowed pattern or that match a forbidden pattern. When a candidate is rejected, the next best-weighted candidate is used instead. Finally, the answer to be returned by the system is selected employing a most-frequent candidate strategy for names, and a most-weighted candidate strategy for other classes.

4. EXPERIMENTS

We measured the performance of the system using a standard QA evaluation test set, extended to include spoken queries. Cross-Language Evaluation Forum (CLEF) test sets were used [11]. CLEF organizes evaluation campaigns in a similar way to TREC. The general aim of the CLEF Multilingual Question Answering Track was to set up a common and replicable evaluation framework to test both monolingual and cross-language question answering systems that process queries and documents in several European languages.

In the following section the experimental set-up is described. Next, results of the baseline system are shown and errors affecting system performance are analyzed. Finally, we modified the vocabulary and the language model in order to improve performance.

4.1. Experimental set-up

We used CLEF 2005 Spanish monolingual QA test set which includes a document collection, a set of questions and the correct answers. The document collection has 454,045 documents of the years 1994 and 1995, from EFE newswire agency (1.06 GB). There are 200 questions, distributed in different question types: 118 factoid, 50 definition and 32 temporally restricted.

We expanded CLEF 2005 QA test set to include spoken questions. We recorded a speaker reading the questions. Headset microphone was used under office conditions, at 16 bit resolution and 16 kHz sampling frequency.

4.2. Baseline system

Spoken queries were processed by the speech recognizer and word error rate (WER) was calculated. The best hypothesis was used by the question answering engine. The same methodology of CLEF was used to evaluate the results [11]. For each question, an exact answer is obtained and the responses were judged as right, wrong, inexact or unsupported. Results are shown in Table 1. The OOV word rate was 1.87%.

We analyzed the result of each individual question and identified different error sources:

- Errors caused by the Named Entity detection system, which was not working properly because the hypothesis from the speech recognizer were in lowercase letters. The patterns used by the NE detection system were developed to work with written text in which named entities start with an uppercase letter.

- Question type identification errors, that happened because the interrogative pronoun at the beginning of each question was not properly recognized. This occurred because the LM was trained with declarative sentences instead of interrogative ones.
- Errors caused by out of vocabulary words.
- Errors caused by words in a foreign language, because our LVCSR was not able to understand them.
- Regular speech recognition errors.

4.3. Improved system

Some preliminary experiments were carried out to improve the performance of the baseline system. The objective was to adapt the output of the speech recognizer to a format suitable for the QA engine. The QA engine was developed to accept written sentences as input, and thus, is very sensitive to the format of the input sentence. Two improvements over the baseline system were proposed:

- **Capitalized output from the LVCSR**:

  In the baseline system the training text used by the speech recognizer was preprocessed and normalized. The normalization process removed case distinction. As the normalized text was used to build the vocabulary and the language model, the output of the speech recognizer was in lowercase letters.

  In order to obtain a capitalized output from the speech recognizer we modified our text normalization process. Case was maintained in the normalized text. We then created a case-sensitive vocabulary, in which the same word with different capitalization was considered different. Then, we trained the language model using mixed case texts.

  Using this approach, there were repeated words in the vocabulary, but with different case. From the 60,000 words in the vocabulary, only 52,867 were different words. The OOV word rate increased to 2.12%.

- **Language model adaptation**:

  The LM used in the baseline system was trained using declarative sentences, and thus, the speech recognizer was not optimized to process questions.

  We made an adaptation of the LM using a small corpus of questions: 600 questions from CLEF QA evaluation from the years 2003, 2004 and 2006. First, we trained a new LM using that questions. Then, we interpolated the general LM with this new LM. We used linear interpolation with an interpolation coefficient of 0.75.

Results using the improved system are shown in Table 1. Better results were obtained compared with the baseline system, because the output of the ASR was adapted to the QA engine.

	WER	R		W	X	U
		#	%	#	#	#
Text	—	61	30.5%	128	7	4
Baseline system	19.7%	42	21.0%	147	6	5
Improved system	15.2%	46	23.0%	141	7	6

Table 1. *Performance of different experiments using questions from CLEF-QA 2005 (WER: word error rate; R: right; W: wrong; X: inexact; U: unsupported).*

5. CONCLUSIONS

In this paper we have described a system that allows users to obtain an answer to a given spoken question. We used a standard QA test set to evaluate the performance of the system and made some preliminary experiments. We also proposed some improvements over the baseline system in order to adapt the output of the speech recognizer to the QA engine. Results showed that the proposed approach outperforms the baseline system both in WER and in overall system accuracy.

Each speech recognition error had big impact on system performance. The worst case happens with proper nouns, because they are essential to find the correct answer, but because of their low frequency in the training data, they have low probability in the LM or even are not included in the vocabulary of the speech recognizer. We analyzed the results of each individual question and identified different types of errors. There were some errors caused by OOV words. Speech driven question answering is a task with an open vocabulary, and thus, it is not possible to include all the words in the vocabulary of the speech recognizer. Words in a foreign language were also a problem, because the speech recognizer is not prepared to recognize them. Other errors were caused by the inaccuracy of speech recognition technology.

As future work we are working in a method to further integrate speech recognition and QA. Our objective is to obtain a better language modeling for questions using the classification patterns that the QA engine is using to classify the questions.

6. REFERENCES

[1] C. Hori, T. Hori, H. Isozaki, E. Maeda, S. Katagiri, and S. Furui, "Deriving Disambiguous Queries in a Spoken Interactive ODQA System," in *ICASSP*, 2003.

[2] D. Kim, S. Furui, and H. Isozaki, "Language Models and Dialogue Strategy for a Voice QA System," in *International Congress on Acoustics*, 2004.

[3] S. Harabagiu, D. Moldovan, and J. Picone, "Open-Domain Voice-Activated Question Answering," in *COLING*, 2002.

[4] A. Fujii, K. Itou, and T. Ishikawa, "A Method for Open-Vocabulary Speech-Driven Text Retrieval," in *EMNL*, 2002.

[5] B. Pellom and K. Hacioglu, "Recent Improvements in the CU SONIC ASR System for Noisy Speech: The SPINE Task," in *ICASSP*, 2003.

[6] A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J. B. Mariño, and C. Nadeu, "ALBAYZIN Speech Database: Design of the Phonetic Corpus," in *Eurospeech*, 1993.

[7] A. Stolcke, "SRILM – an Extensible Language Modeling Toolkit," in *ICSLP*, 2002.

[8] C. González-Ferreras and V. Cadeñoso-Payo, "A System for Speech Driven Information Retrieval," in *ASRU*, 2007.

[9] J. Gómez, M. Montes y Gómez, E. Sanchis, and P. Rosso, "A Passage Retrieval System for Multilingual Question Answering," in *TSD*, 2005.

[10] E. Sanchis, D. Buscaldi, S. Grau, Ll. Hurtado, and D. Griol, "Spoken QA based on a passage retrieval engine," in *SLT*, 2006.

[11] A. Vallin, B. Magnini, D. Giampiccolo, L. Aunimo, C. Ayache, P. Osenova, A. Peñas, M. de Rijke, B. Sacaleanu, D. Santos, and R. Sutcliffe, "Overview of the CLEF 2005 Multilingual Question Answering Track," in *CLEF*, 2005.