On the Voice-activated Question Answering

Paolo Rosso, *Member*, IEEE, Lluís-F. Hurtado,
Encarna Segarra, and Emilio Sanchis, *Member*, IEEE

Abstract

Question Answering is probably one of the most challenging tasks in the field of Natural Language Processing. It requires search engines that are capable of extracting concise, precise fragments of texts that contain an answer to a question posed by the user. The incorporation of voice interfaces to Question Answering systems adds a more natural and very appealing perspective for these systems. This paper provides a comprehensive description of current state-of-the-art Voice-activated Question Answering systems. Finally, the scenarios that will emerge from the introduction of speech recognition in Question Answering will be discussed.

Index Terms

Question Answering, Speech Input, Information Extraction, Automatic Speech Recognition, Language Model, Named Entity Recognition.

I. Introduction

The understanding of human language by means of machines is probably the most challenging and investigated issue in the field of Artificial Intelligence and Human Language Technology. The continuous growth of the information available in digital format is making old Information Retrieval (IR) methods obsolete, demanding more precision from search engines, which can only be obtained

Paolo Rosso, Lluís F. Hurtado, Encarna Segarra and Emilio Sanchis are members of the ELiRF research group of the Departament de Sistemes Informàtics i Computació (DSIC), Universitat Politècnica de València, Camí de Vera s/n, 46022 València, Spain.

e-mail: {prosso, lhurtado, esegarra, esanchis}@dsic.upv.es

We would like to thank the TIN2009-13391-C04-03 and the TIN2008-06856-C05-02 research projects for partially supporting this work.

by giving them the ability to (at least partially) deal with the document semantics and not only to consider documents as bag-of-words.

In the Question Answering (QA) task, search engines have to extract concise, precise fragments of texts that contain an answer to a question posed by the user. This task is very close to what humans usually consider "text understanding". The availability of QA systems that can provide us with the specific information related to a query without the need to read all the documents that an IR system would return, or the possibility of extracting the information contained in the documents in an automatic way, are two examples of how this kind of research may change the type of interaction between humans and machines. Competitions like TREC (http://trec.nist.gov) or CLEF (http://www.clef-campaign.org/) have been created in order to develop and improve existing systems and to evaluate and compare their behavior. These competitions now have a long history of demonstrating the success and the good results of some systems. Another step towards a more natural, "human-like" communication between machines and users in need of information is represented by the introduction of Speech Language Technologies into QA systems. Automatic Speech Recognition (ASR) and the development of spoken human-machine interfaces are currently considered to be mature enough to be used in most common applications. There are some recognition systems that can process more than $60,000$ words; some prototypes of mixed-initiative dialog systems, which are capable of an acceptable robust behavior, are also in an advanced stage of development.

The development of systems that are the result of the integration of the two technologies (QA and ASR) permits humans to communicate with computers from a more natural and very promising perspective. Some examples of these systems are represented by spoken QA systems that are used by mobile devices. Although more improvements need to be integrated by means of multilinguality or multimodality, combining both oral and visual communication, these systems are a first concrete example of this new "human-like" communication that we are starting to have with machines.

In this paper, first we will briefly present the main characteristics and achievements of the classical ASR and QA systems. Then, we introduce the reader to the Voice-activated QA systems that have been developed so far. Finally, we briefly discuss the specific issues that are derived from this new scenario from the perspective of users as well as from that of QA and ASR researchers.

II. Automatic Speech Recognition

The general problem with automatic speech recognition by any speaker in any environment in any language is still far from being completely solved. Despite this, ASR technology is currently considered to be mature enough to be used in limited domains. In order to analyze its viability there are several aspects to be considered in the different speech recognition applications: the vocabulary size, the type of speech, and the environment conditions. Speech recognition is easier if the size of the vocabulary is small. For example, tasks with a small vocabulary, such as recognizing sequences of digits, are relatively easy. However, tasks with large vocabularies (60,000 words or more), such as transcribing broadcast news, are much harder. Another aspect is the type of speech. Isolated word recognition, where each word is uttered in isolation (with pauses), is easier than continuous speech recognition, where the boundaries of the words are not clearly delimited. Also, read speech is easier than spontaneous speech, even though both are continuous speech recognition. A third aspect is the environment conditions. In a low-noise laboratory environment using close-talk microphones the recognition performance is higher than in a noisy environment were telephone is used for communication with the ASR system.

The performance of ASR systems for different tasks can be very different. The evaluation of this process is usually measured by the Word Error Rate (WER), which is defined as the average number of substitutions, insertions, and deletions of words that are needed to transform the recognized sentence into the correct transcription of the speech input. Some recognition results for different systems are reported in [34]: a WER of 0.5% for a task of isolated digits recognition; a WER of 3% for a task of read speech with a vocabulary of 5,000 words; a WER of 10% for a task of broadcast news transcription with a vocabulary of 64,000 words, and a WER of 20% for a task of conversational telephone speech, with a vocabulary of 64,000 words.

The Automatic Speech Recognition process consists of obtaining a sequence of words from an audio signal. The mathematical formulation of this process from a statistical point of view is as follows [9]:

$$\hat{W} = \underset{W \in \Sigma^*}{\operatorname{argmax}} P(W|X)$$

where, $\hat{W}$ is the best of all possible sentences (sequences of words over a vocabulary Σ) according to the sequence of acoustic features (X) obtained from the audio signal by means of a parametrization process.

Applying the Bayes' theorem and assuming that $P(X)$ is the same for all sentence, the maximization process can be formulated as follows:

$$\hat{W} = \underset{W \in \Sigma^*}{\operatorname{argmax}} \frac{P(X|W)P(W)}{P(X)} = \underset{W \in \Sigma^*}{\operatorname{argmax}} P(X|W)P(W)$$

Thus, the recognition process consists of finding the sentence $\hat{W}$ that maximizes the product of two different probabilities: the $P(X|W)$ provided by the so-called acoustic model, and, $P(W)$ provided by the so-called language model.

The usual acoustic model employed in speech recognizers is based on Hidden Markov Models (HMM) [34]. A set of acoustic units is selected (e.g. phonemes) and represented by HMM. Then the acoustic model for each word in the vocabulary is built by a suitable combination of acoustic unit models (e.g. the sequence of phonemes corresponding to that word).

The language model applies the constraints on the sequences of words that are allowed to be recognized, and, therefore, represents the syntactic restrictions for large vocabulary tasks. The most successful and widely used language models are N-gram models, which model the frequency of appearance of sequences of N words. This frequency is estimated from training corpora of written sentences.

In an N-gram language model, the probability of each word does not depend on all the previous words in the sentence, but only on the $N-1$ previous ones. Therefore, the N-gram probability assigned to a sequence of words W is:

$$P(W) = \prod_{k=1}^{n} P(w_k | w_{k-N+1}^{k-1})$$

where $W = w_1\ w_2\ \ldots\ w_n = w_1^n$, $w_i \in \Sigma$, $1 \leq i \leq n$.

The performance of a language model component is evaluated by the perplexity, which measures both the task complexity and the adequacy of the language model [34].

Once the acoustic and language models are estimated from a training corpus, the recognition process consists of maximizing the product $P(X|W)P(W)$, which is performed through search algorithms. The most widely used algorithm is the Viterbi algorithm [22]. Due to the difficulty of including all the information provided by the different knowledge sources during the recognition process, some ASR systems provide both the most probable recognized sentence and the N-best hypotheses (N-best sentences). This way, all the hypotheses can be analyzed by a posterior process, and errors that might appear in a 1-best approach can be corrected.

III. Question Answering

In this section, we describe the main characteristics of current state-of-the-art QA systems. Most of them have been presented at the TREC and CLEF competitions. The basic functionality of a QA system is to allow a user to question a non-structured document collection in natural language in order to find the correct answer. An alternative view of QA is to consider it as a specialization of Information Retrieval, where the amount of information to be retrieved is minimal. This can be very important in a future scenario where the growth of the quantity of available information will make current IR technologies obsolete.

Question Answering is usually divided into three main phases: Question Analysis and Classification, Document or Passage Retrieval, and Answer Extraction. A further phase for Answer Validation is often added at the end.

A. Question Analysis and Classification

Question classification assigns a class to each question. Its main objective is to obtain the expected answer type from the question. This first phase is probably the most crucial step of a QA system, since, in the last phase, the extraction strategy of the answer completely depends on the correct classification of the question. Some examples of factoid questions are shown in Fig. 1. For instance, extracting the answer to "What is Aspirin?", which is looking for a definition, is not the same as extracting the answer to "Who invented the radio?", which is asking for the name of a person. More complicated questions could be formulated, thereby increasing the difficulty of the problem. This is the case of the questions shown in Fig. 2, which involve compiling lists, reasoning, or anaphora resolving. In [46], the authors reported that more than 36% of errors in QA are due to mistakes of question classification. Therefore, a proper classification of the question will allow the candidate answers to be restricted.

The design of the question classification module for a QA system always starts by determining what the number of classes is and how to arrange them. In [33], the authors introduced a QA typology made up of 94 question types. Most systems being presented at the TREC and CLEF QA competitions use no more than 20 question types. The approaches to question classification can be divided into two categories: (i) pattern-based and (ii) machine learning-based. The latter has to face the problem of the lack of training data [38], [27]. Those QA systems whose question classification makes use of patterns and heuristic rules are based on the detection of interrogative keyword pronouns ("wh-words" such as "When", for date or time, "Where", for places, "Who",

```
Where was the America's Cup celebrated in 2007?
What is Aspirin?
What is BMW?
Who invented the radio?
What is the capital of France?
When did J.F. Kennedy die?
What musicals did Kurt Weill write?
What museums have displayed Chanel clothing?
What are the names of the Baltic republics?
What is the distance from the Earth to the Sun?
```

Fig. 1. Examples of factoid questions

```
265 target: Mahmud (or Mahmood, Mahmoud) Ahmadinejad
265.1 What country is Ahmadinejad president of?
265.2 In what town was he born?
265.3 On what date was he born?
265.4 He holds a Ph.D. in what field?
265.5 What foreign countries has he visited since his election to the
presidency?
265.6 What other positions has he held in government?
265.7 Other (that is, ''tell me something important about him that I did
not ask you'')
```

Fig. 2. Example of a TREC question topic (from the TREC question set) that includes list questions and questions that need anaphora resolution

for people) and specific trigger words [30], [62]. The main problem with these QA systems is the amount of work needed for pattern formulation and definition because those patterns must capture any (or almost any) possible query reformulation.

A further task performed in the question classification and analysis phase is the extraction of the *focus* and the *target* (or *topic*) of the question. The focus is the property or entity sought by the question, whereas the target or topic is the event or object the question is about. For instance, in the question "What is the capital of France?", the focus is *capital* and the topic is *France*. The extraction of the focus and topic of the question can be performed in several ways: looking for specific keywords, using classification methods (for instance, MIRACLE [18] and TALP-QA [21]), using syntactic parsing and/or Natural Language Processing (NLP) tools ([65], [11], [7]), exploiting world knowledge by means of wikipedia [35] or other ontologies such as WordNet [49], the web [39] or on the basis of patterns of words or Part-Of-Speech labels [25], or combinations of some

of these.

Most of the approaches use Named Entities (NEs) tagging for question and answer typing, typically classifying a question as requiring a particular NE type as the answer. A completely different approach is the one proposed by the Tokyo Institute of Technology [63] where all word sequences (between one and five words long) are used to match the answer in a QA database: all types are initially assigned a probability, and the decision about the final answer is postponed until all knowledge sources have been considered.

B. Document or Passage Retrieval

The second phase in a QA system is document or passage retrieval, that is, the selection of fragments of documents where it is possible to find the answer. Document retrieval systems supply a set of ranked documents based on a distance function between the question and the documents and use classical weighting schemes that are based on term frequency, such as $tf \cdot idf$ [53], or on statistical analysis, such as BM25 [52]. Most QA systems are based on IR methods that have been adapted to work on passages instead of the whole document [43], [8], [61], [48]. The main problems with these QA systems come from the use of methods which are adaptations of classical document retrieval systems, that are not specifically oriented to the QA task. [33], [51] show that off-the-shelf IR engines (MG and Okapi, respectively) often fail to find documents containing the answer when presented with natural language questions.

Passage Retrieval (PR) systems return pieces of texts (passages) which are relevant to the user questions instead of returning a ranked-list of documents like IR systems do. QA-oriented PR systems present some technical challenges that require an improvement of existing standard IR methods or the definition of new ones. First of all, the answer to a question may be unrelated to the terms used in the question itself, making classical term-based search methods useless. These methods usually look for documents characterized by a high frequency of query terms. For instance, in the question “What is Aspirin?”, the only non-stopword term is “Aspirin”, and a document that contains the term “Aspirin” many times probably does not contain a definition of the drug. Another problem is to determine the optimal size of the passage: if it is too small, the answer may not be contained in the passage; if it is too long, it may bring in some information that is not related to the answer, requiring a more accurate answer extraction.

In [15], [16] the authors investigate the impact that NE recognition, as well as relation extraction, may have on document and/or passage results. In order to study the significant degradation in search

performance with imperfect NEs, the authors applied different error models to the gold standard annotations in order to simulate errors made by automatic recognizers. A significant document retrieval gain was achieved when adopting NE recognizers (15.7% improvement in precision).

It is important to consider the differences between traditional document retrieval and QA-oriented passage retrieval. In the first case, the greatest effort is done to retrieve documents about the same topic of the query, while in the second case, the aim is to retrieve pieces of text that contain the answer to a given question. Various methods have been proposed to determine the similarity between the passage and the question [59]. There are passage retrieval approaches that are based on NLP techniques, such as those in [26], [4], [31], [40]. In [14], the authors argue that the use of NLP improves the results in the case of questions that are related to a specific domain. Others include semantics to allow QA systems to answer specific types of questions [47]. The main disadvantage of these approaches is that they are very difficult to adapt to other languages or to multilingual tasks. Moreover, they are usually slower than bag-of-words approaches [10]. This issue is critical for commercial systems that can be accessed on the web, where short wait times are a key for success. Of particular interest are those QA-oriented PR systems that are based on the overlapping between the question and the passage terms, or on the density of question terms in the passage such as the one presented in [25]. Pattern-matching approaches [24] perform well only if it is possible to achieve great answer redundancy, such as in the web or in large document collections.

The use of the web as corpus for QA has been investigated in [19], [12], [13]. The user question is put into a search engine (e.g. Yahoo or Google) with the expectation of getting a passage that contains the same expression as the question or a similar one. To increase the possibility of finding relevant passages, some reformulations of the question are made (i.e., terms are moved or deleted to search for other structures with the same question terms). A model that has been shown to be remarkably effective when using web documents to find answers and to exploit the web's inherent redundancy is the statistical and data-driven QA system of the Tokyo Institute of Technology [63]. The approach they adopted for QA is very interesting, especially in the context of this survey because it is somewhat similar to those approaches used in ASR where there are separate acoustic and language models. Although its performance still "falls somewhat short if compared to the best linguistic-based systems" [63] of TREC and CLEF, its model is extremely simple and shows great potential for improvement. Another interesting QA system that uses ASR-style techniques is described in [20], where a noisy-channel approach is applied to standard QA.

C. Answer Extraction

The last main phase is the answer extraction which is responsible for extracting the final answer from the retrieved passages or documents. Every piece of information extracted during the previous phases is important in order to determine the right answer. The answer is searched for within the retrieved passages, using the information obtained from the question analysis in the first phase where the focus and the topic of the question are extracted. The main problem that can be found in the answer extraction phase is determining which of the possible answers is the right one, or the most informative one. For instance, an answer for "What is BMW?" can be "A car manufacturer"; however, better answers could be "A German car manufacturer", or "A producer of luxury and sport cars based in Munich, Germany". Another problem that is similar to the previous one is related to the normalization of quantities: the answer to the question "What is the distance from the Earth to the Sun?" might be "149,597,871 km", "one AU", "92,955,807 miles" or "almost 150 million kilometers". These are descriptions of the same distance, and the answer extraction module should take this into account in order to exploit redundancy. In fact, redundancy can help in the answer extraction process due to the fact that if the same answer is obtained many times for different passages, it should be a good candidate. In other words, the good behavior of the document or passage retrieval phase is a critical aspect [1], [17], [2]: a syntactic or semantic analysis of the most promising sentences or fragments can be performed, or regular expressions that represent the possible structure of the answer can be used.

From a general viewpoint, the answer extraction strategies used by most QA systems can be grouped into three major categories:

- Collection-based strategies: they search only in a static textual collection in order to find the answer and return it with a justification passage.
- Web-based strategies: they use the web to find the answer to the question, then return a justification extracted from a static collection, if necessary (notably TREC and CLEF tasks)
- Database approaches: the possible answers to a question are stored in a database; the system only needs to pick up the answer and return it, eventually with a justification passage.

A typical collection-based approach such as the one of INAOE for factoid questions [50] is based on the detection of Named Entities (supervised) or on lexical patterns (pattern-matching). Web-based strategies can use the same collection-based method on the large, dynamic web (see for instance [39]) or also write the question in a search engine and analyze the returned snippets. Another system

that makes use of the information available on the web is the one of the University of Amsterdam [3] which searches for answers in Wikipedia. Finally, in database approaches, the effort is made on the extraction of answers, on the basis of automatic or semi-automatic text processing of static text. In [50], a database approach is used for definition questions. This system identifies NEs that can potentially be the object of questions and stores the answers with them. However, in the last few years, most systems have evolved into more complex ones that use a combination of the above strategies, depending on the type of question (such as in the INAOE QA system [50], where a collection-based approach is employed for factoid questions and a database one for definitions), or using a strategy for the extraction and another one for the validation.

D. Answer Validation

In order to check whether or not the extracted answer from a given corpus is correct, before returning it, a further answer validation phase is often added at the end of the three main phases of question classification and analysis, passage or document retrieval, and answer extraction. There are two different approaches for handling answer validation by current QA systems: (i) corpus-based and (ii) redundancy-based. Corpus-based methods such as [28] rely on a deep linguistic analysis of the question and the answer candidates, while redundancy-based ones use the web as a lexical resource for validating the extracted answer [42], [12], [17], [37]. In order not to favor short or unspecific answers (e.g. ”1928” over ”July 26, 1928” for the factoid question of Fig. 1) some heuristics may be applied (e.g. [12], [37]) to boost the score of specific answers (”July 26, 1928” over ”1928”). In the case of partial redundancy among candidate answers, they can be represented by fuzzy sets and their feasibility can be estimated by means of frequency counts and co-occurrence statistics as well as by asking additional questions [55]. This approach seems to be particularly useful when asking, for instance, for the creation date of a work of art, which, in practice, is often stated by means of an interval or a fuzzy description instead of an exact date.

In this section, we have presented an overview of current state-of-the-art approaches for the main phases for open-domain QA (question analysis and classification, document or passage retrieval and answer extraction) as well as for the further optional answer validation phase. A description of question answering in restricted domains goes beyond the scope of this survey on Voice-activated QA and can be found in [60]. In the next section, we describe the QA systems where an initial speech input phase has been added.

IV. VOICE-ACTIVATED QUESTION ANSWERING

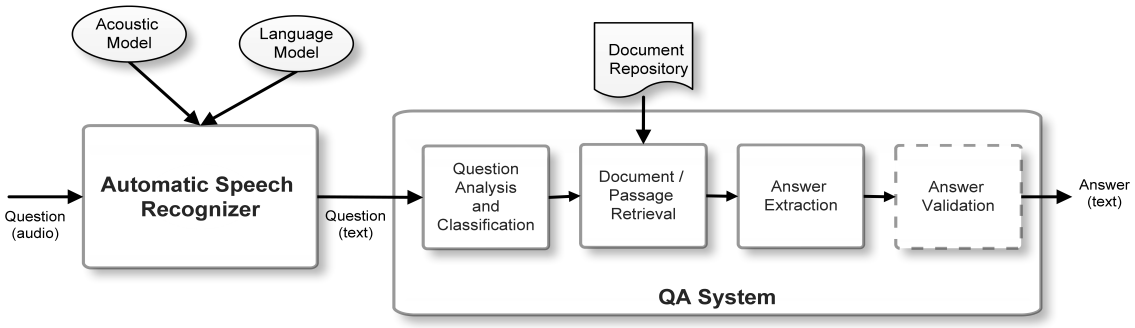

Fig. 3. Diagram of a Voice-activated Question Answering system

As we mentioned in the introduction, the development of systems that are the result of the integration of an automatic speech recognition system as input to a QA system will permit humans to communicate with computers in a more natural way. Figure 3 shows the main components of a Voice-activated Question Answering system.

There are some important and specific problems that need to be solved in order to obtain good results from the integration of the two technologies, Question Answering and Automatic Speech Recognition. It has been widely demonstrated that this integration degrades the behavior of the complete system. In the Voice-activated QA presented in [29], the authors show the results of an experiment in which the best QA system at TREC [46] and an ASR system were employed. They investigated the impact that the ASR system had on their open-domain system. The baseline performance of the QA system with text input was 76%, whereas the same QA system with speech input (the output of the speech recognizer operated at a word error rate of 30%) obtained only 7% accuracy.

Some experiments were also carried out with the QUASAR QA system [54] with the aim of studying the effect on the QA system accuracy of the word error rate introduced by the ASR system, especially from the perspective of the recognition of NEs. The preliminary experiments with simulated speech input (i.e., errors in the input questions -200 questions of the CLEF 2005 Spanish monolingual test set- were introduced) showed that when recognition errors do not affect NEs, the QA system performance is still good, even with a WER of 25%. Error rates greater than 30% made the system behavior deteriorate quickly.

The main components of an ASR system which are strongly dependent on the task are: (i) the vocabulary, i.e., the set of words (e.g. function words and content words) in terms of which the recognition system constructs the output; (ii) and the language model, i.e., the set of sequences of these words that the recognition system can give as output. We analyze below the effects of recognition errors due to both vocabulary and language models.

The majority of the currently available QA systems are based on the detection of specific keywords, mostly NEs. For instance, for the individuation of the answer to the CLEF question "What is the capital of Croatia?", a failure in the detection of the NE "Croatia" would make it impossible to find the answer. Thus, the vocabulary of the ASR system must contain the set of NEs that can appear in the user questions. Unfortunately, the number of different NEs in a standard QA task could be huge, and state-of-the-art automatic speech recognition systems still need to limit the vocabulary size. Therefore, the vocabulary size of the ASR is much smaller than the one in a standard QA task. Thus, the presence of words in the user questions that were not in the limited vocabulary of the system (out-of-vocabulary (OOV) words) is a crucial problem in this context. Throughout this section, we will summarize the different solutions suggested in the literature.

With regard to the vocabulary of the ASR system, and in addition to the problem of NEs mentioned above, there are some special words that must be taken into account: the interrogative pronouns which determine the type of question. Recognition errors in keywords that are present in the queries such as Who, When, Where ..., can be very determinant in the question classification process. Therefore, the ASR system should provide very 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 mispronunciation of NEs (such as names of persons or places) especially when the NE is in a language different than the user's. To manage this problem a mechanism that considers alternative pronunciations of the same word or acronym must be implemented.

With regard to the language model of the ASR system, there are some issues that must be considered. Typically, a language model provides constraints on the sequences 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. The queries to a QA system have a specific syntax but, in general, a sample of sentences of this kind that is large enough to learn robust models (usually bigrams or trigrams) is not available. If a large written general corpus is used, then the advantage of knowing the strength of syntactic restrictions of the queries is lost. Different solutions have been suggested in the literature, as described below.

Apart from increasing the robustness of the speech recognizer as well as the QA system, the incorporation of a certain dialog mechanism (a short interaction between system and user) would help to improve the performance of the complete system. This dialog mechanism should focus on the confirmation of the keywords of the query. The use of confidence measures associated to the words in the query can be used by the dialog manager to decide what confirmations or new requests must be generated.

In the last few years, a few Voice-activated QA systems have been developed on the basis of the integration of automatic speech recognition and text QA systems. For example, in [64], the authors (who work in classical QA [63]) describe the web and mobile-phone interface where they have incorporated their multilingual factoid QA system and a prototype speech interface. The recognition of the user's questions is performed on a separate speech recognizer that is dedicated to recognizing questions.

The main concern of all these Voice-activated QA works was how to reduce the effect of speech recognition errors introduced by the recognition process of the input questions on the QA performance. In the following subsections, we describe the state-of-the-art research on Voice-activated QA systems. We present the proposals from different perspectives that are related to improvements in the vocabulary and in the language models. Finally, we describe those interactive Voice-activated QA systems that make use of a dialog mechanism.

A. Improvements in the vocabulary

As previously stated, recognition errors in NEs and in interrogative pronouns that are present in the user query can degrade system performance considerably. In this subsection, we review some works that aim to provide solutions to these problems.

In [23], the authors describe an approach to deal with open vocabulary in the context of spoken text retrieval. The adopted solution to the problem of the effect of recognition errors on the system is based on the use of OOV words in the ASR system. The vocabulary defined for the language model includes not only the 20,000 most frequent words, but also a set of syllables. For the recognition process, a word can be formed by a word or a sequence of syllables. The result of the recognition process is a sequence of words and sequences of syllables. The system is based on a two-step approach. In the first step, the input is filtered so that only the words are considered. A specific number of top-ranked documents is obtained which will then be used for the search in the next step. In the second step, the detected OOV words are replaced with index terms that are phonetically

identical or similar, and the text retrieval is performed again. The preliminary results presented in the paper show a precision of 23% in detecting and completing the OOV words process.

In [6], the authors investigate an open vocabulary recognizer that makes use of a static language model that allows the addition of new words without the need for language model retraining or adaptation. Three steps are necessary to add a new word in an ASR system. First, the word must be added to the vocabulary; then, a phonetic transcription must be associated with it which, thirdly, must then be included in the language model distribution with non-zero probabilities. To introduce a word in the vocabulary without retraining the language model, the authors propose using special forms called back-off word classes (e.g. proper name, singular noun, plural noun, conjugated verb, and geographic name). During the static language model training, one of these forms replaces one or more words, which are considered as unknown words, by discounting a mass of probability from the OOV words. Several classes of back-off words are defined, and a certain probability is assigned to each of them in the training phase. Before the recognition phase, when a new word is added, its orthographic form is also added to its corresponding back-off word class, allowing the language model to assign a probability to those sentences that include this word. The experiments show a 30% reduction in the OOV rate, with 80% of the occurrences of the newly introduced words being correctly recognized. Moreover, about 84% of the introduced NEs are correctly recognized.

In [58], the authors propose a method for improving the speech recognition of questions by allowing interaction during the question specification phase. In this interactive system, a user may first specify the NE of interest: a person's name, an organization, and so on. A grammar for the NE is created from a database of NEs existing in the corpus of documents. When the NE is recognized, a search engine is employed to extract the documents matching the NE; these documents are used to train a specific language model. The language model used in the ASR for the recognition of the question is a linear combination of this specific one and a language model trained with question-type sentences. The results of the experimentation show a dramatic reduction in word error rate (a 32% in the best experiment), and the ratio of missed NE recognition is halved.

These three works, in which different approaches are presented to improve the results related to OOV and NE recognition, show important improvements in the ASR system results. As we commented in the introduction of this section, the improvements in the ASR system results strongly depend on an adequate treatment of the OOV words, especially NEs. As we also previously mentioned in the section devoted to Document or Passage Retrieval, better results in NE recognition [16] imply better performance of the Voice-activated QA system.

B. Improvements in the language model estimation

Determining how the language model has to be estimated when a Voice-activated QA system is designed is also a very important aspect to take into account. In this subsection, we review some works that describe different approaches for combininig specific and more general language models.

In [29], the authors propose an approach based on several interactions, called iterative refinement, between the ASR and the QA system in order to improve the behavior of both systems. First, the ASR generates not only the transcribed sentence but also a word lattice as the output of the ASR process. A special filtering mechanism then uses both the question transcription and the word lattice to filter out words that cannot be processed by a typical QA system due to syntactic, semantic or pragmatic inconsistencies. The result is a word lattice of smaller dimensions. This reduced lattice is used for generating an enhanced language model that is used to reprocess the spoken question. The result of the application of this language model is a semantic transformation of the query rather than a string of words. This output is processed by a QA system that is capable of using lexical and semantic alternations of the question keywords when searching for the answer. The results of the experimentation show that the interactions enabled by the system both improve the accuracy of the spoken QA (three times better) and provide a better word error rate of the ASR (a 50% of reduction in WER).

An interesting study on specific language models for QA is presented in [56]. A set of 280,000 unique user questions was collected from various online resources. The work compares the perplexity of several N-gram language models that were learned using this corpus. Different lexicon sizes, N values, discounting techniques, and cutoff thresholds were used to construct the models. According to the experimental results, (without taking into account OOV words which are less than 4% in a 32,000-word lexicon) the complexity of the language model is simple enough to be used in a predictive question recognition scene. In [57], the authors also present a multimodal QA system called AnswerBus, which is specifically designed for mobile devices where the question recognition is performed by means of a predictive interface. Moreover, multiple domain-dependent language models were used for the voice queries recognition. Although the presented system is very interesting, it does not take into account some important aspects of Voice-activated QA such as the difficulty in the recognition of NEs or the differences between written text and spontaneous speech. These problems are bypassed allowing the user to correct the incorrect prediction of the system using a stylus or keypad.

The problem of constructing an accurate language model is tackled in [36]. Multiple language models for recognizing voice queries are trained using clustered newspaper articles; five domains are defined: economy, entertainment, international affairs, general public and sports. Each domain-dependent language model is interpolated with a class language model that is trained from a set of transcribed query utterances. The experimental results (a 5% of reduction in WER) show that every domain-dependent language model achieves better recognition performance than the domain-independent model. Even if the domain that each question belongs to is unknown, the best performance can be obtained by choosing the domain-dependent language model that maximizes the likelihood.

The structure of most of the questions to a QA system can be divided into two different parts: (i) one that is related to the query topic and (ii) another that is related to the query itself (the wh-questions interrogative structure described in the previous section). In [5], the authors investigate how to learn a language model that models both parts efficiently. The method proposed in this paper consists of adapting a N-gram language model learned using the document corpus by amplifying the N-gram sequences from a list of interrogative expressions defined by hand. The experimental results show that this method increases the recognition rate of the interrogative structure, compared with a language model that is learned only with the document corpus, without decreasing the performances in the topic recognition.

As the above works conclude, improvements in language model estimation also entail important improvements in ASR system performance. In these works, the authors presented different approaches that combine specific language models that are estimated from samples of queries with language models that are estimated from more general documents.

C. Interactive Voice-activated QA systems

As described in the previous subsections, interaction with the user may help for improvements both in the vocabulary [58] (especially, in the OOV Named Entities) as well in the language model estimation [29].

In [36], the authors present a strategy that makes use of an interaction with the user to avoid the effect of recognition errors on the QA system. After recognizing each query, the keywords are automatically extracted and displayed in a graphical user interface. Then, the user is requested to select the incorrect keywords; the selected incorrect keywords are automatically replaced by other candidate keywords from the N-best list obtained as a result of query utterance recognition.

A spoken interactive QA system is presented in [32]. The authors propose two mechanisms to tackle with the difficulties that are inherent to the spoken input: a screening filter and a set of disambiguous queries. A screening filter tries to extract meaningful information from the recognized sentence in two steps. The ASR system provides a transcribed sentence with a confidence measure for each word. Then the acoustically and linguistically unreliable words that are based on the threshold for the confidence measure are removed from the transcribed sentence. In a second step, a meaningful sentence from this last filtered sentence is constructed using a speech summarization technique. When the QA engine cannot extract an appropriate answer to a user's question, this question is considered ambiguous and a mechanism of interaction with the user is triggered asking him/her for additional information. Disambiguous queries are automatically generated to disambiguate the question. These disambiguous queries are generated using templates of interrogative sentences. The experimentation shows the potential of generated disambiguation queries in requiring indispensable information that is lacking to extract answers.

In [44], a dialog strategy to correct and to clarify spoken queries to a document retrieval system is presented. This strategy is based on the use of two statistical measures: the relevance score and the significance score. The relevance score, which is defined as being inversely proportional to the perplexity, allows the system to detect phrases with important recognition errors before those errors affect the document retrieval process. The significance score detects vagueness in the user query. The vagueness appears as differences between the N-best hypotheses that actually have an impact on the retrieval result. These measures allow the system to ask the user for confirmation of those query parts that may contain recognition errors or ambiguities. The confirmation requests are generated before and after the document retrieval process. The algorithm proposed in [44] is as follows: first, the user query is recognized and the N-best hypothesis are generated, then the relevance score of each phrase is calculated (in this context, a phrase is a basic unit of Japanese grammar). A confirmation to the user of those phrases with a low relevance score is required. Then, using the N-best hypothesis generated by the ASR, a retrieval from the document base is performed, the significance scores are calculated and, if necessary (i.e., if the significance score is higher than a threshold) another confirmation to the user is generated. This second confirmation consists in presenting the difference in the N-best list of the ASR to the user and letting him/her choose one. Finally, the document retrieval process result is presented. Some experimental results are presented in the paper using a software support knowledge base with 40,000 entries. The results show that the system is able to generate more efficient confirmations than those generated just using

the confidence scores or the ASR.

An interesting application of Voice-activated document retrieval systems of the same authors is presented in [45]. The authors have introduced spoken QA capabilities in the ”Dialog Navigator for Kyoto City”. The system accepts questions about Kyoto, searching for the answers in Wikipedia and the official tourist information guide of Kyoto. Typical questions are about sightseeing places and monuments, e.g. ”Who built this shrine?”. The dialog is simple, since it is based only on confirmations by the user. Moreover, a summarization process was introduced in order to provide the user with a compact document representation.

All these works have presented different approaches to voice-activated QA systems in which the interaction with the user helped the system to improve its performance. These approaches differ mainly in two aspects: (i) the point in time that the system interacts with the user and (ii) how the information supplied by the user is used by the system (mainly to select the next word in the N-best list of words and to delete erroneous recognized words).

V. Voice-activated Question Answering systems and beyond

Voice-activated QA applications have a promising future. Even though today, mainly only research prototypes have been developed, we think that in the next few years more accurate systems will be developed. This development will take place due to the interest in facilitating access to certain services, such as providing mobile devices with more interesting functionalities. These advances in Voice-activated QA have been achieved due to previous and current efforts in many areas of spoken human-machine interaction. These efforts have been oriented in different lines of research, which can add their particular results to the QA problem, such as dialog systems, machine translation, or information retrieval. Therefore, progress in QA and ASR technologies will foster the construction of systems that will be capable of accessing non-structured information in both restricted-domain and open domains.

A. Examples of Voice-activated Question Answering systems

In addition to the Voice-activated QA systems presented in the previous section, there are also other systems developed by teams worldwide within the framework of research projects such as QALL-ME and GeoVAQA.

The European research project QALL-ME (http://qallme.itc.it/) shows the increasing interest by the research community in improving access to information systems, particularly in the way

this information can be accessed. Its general objective is to establish a shared infrastructure for multilingual and multimodal, open-domain Question Answering for mobile phones. Researchers have developed QA systems in mobile phones to search for information in the fields of tourist information and local events in a city, integrating spontaneous speech as input and textual answers with maps, images, or videos as output.

The GeoVAQA prototype described in [41] is a restricted-domain spoken Question Answering system. It has been designed to work in the scope of Spanish geography. The system supports spoken questions about Spanish geography as input and returns a concise textual answer and a set of relevant snippets. GeoVAQA uses a Geographical Knowledge Base and Google to obtain the snippets. This application can be used as a specific information service as well as a learning tool.

B. New scenarios for Voice-activated Question Answering systems

Adding speech input to the existing systems of Information Retrieval and Question Answering will be a topic of interest in the coming years within the human language technology community.

E-learning is a suitable field for applying this kind of technology since, frequently, only short answers are required for learning purposes (names or dates in history, names in medicine, definitions or formulae in technical studies,..). A lot of this information is not structured and is not available in a database, but it is provided as raw texts. Therefore, Voice-activated QA systems can be the multimodal interfaces that would improve their accessibility. An example of a spoken dialog tutoring system in the field of learning is the ITspoke project (http://www.cs.pitt.edu/ litman/itspoke.html). One of the goals of this research project is to integrate spoken language technology with instructional technology in order to promote learning gains by enhancing communication richness. Although this project is focused on dialog management and knowledge representation for interaction with students, it shows that e-learning can be one field of application of Voice-activated QA.

The goal of the DARPA GALE program (http://www.darpa.mil/ipto/programs/gale/gale.asp) is to develop a system that is able to automatically take multilingual newscasts, text documents, and other forms of communication and make their information available to human queries. It has three major technical challenges: automatic speech recognition, which processes audio data; machine translation; and distillation, which extracts the most useful pieces of information related to a given query. In particular, one of the goals of the GALE project is the information distillation task that attempts to extract exact answers to 5-W questions, i.e., who, what, when, where, and why. It is expected that if a system can isolate these pieces of information successfully, then it can obtain the basic meaning

of the sentence. In this project, different corpora that represent possible fields of applications were used: broadcast news from radio or TV, broadcast conversations, phone conversations in different languages, web newsgroups, and weblogs. In this project, the problem of QA is tackled considering different types of data resources, that is, text and speech in different languages.

From the viewpoint of commercial efforts, there is increasing interest in QA. For instance, within the DeepQA project (http://www.research.ibm.com/deepqa/deepqa.shtml), IBM is working to build a computing system named Watson that can understand and answer complex questions with enough precision and speed to compete against some of the best human contestants of the popular TV game show Jeopardy. Jeopardy is a game that requires knowledge from a broad range of topics including history, literature, politics, film, pop culture, and science. Although at the present time only text input is allowed, providing Watson with the ability to answer oral questions would undoubtedly enormously extend the potential of the system in the way it is accessed, the kind of people that access it, and the situations in which the system is used.

In the last few years, interest in augmented reality applications (especially those that provide information such as augmented reality systems for providing sightseeing information on mobile devices) has grown considerably. An example of applications of this kind is the European project Archeoguide (http://archeoguide.intranet.gr). Within the framework of informational augmented reality applications, the use of VAQA interfaces would allow these systems to answer specific oral questions instead of providing generic contextual information.

Finally, we would like to point out the interest of Information Retrieval, Human Language Technology, and Spoken Language Translation researchers in the Voice-activated QA topic. In 2009, the CLEF QA track included a Voice-activated QA task in different languages, although, in this first attempt, the proposed evaluation was on manual transcriptions of a set of spontaneous speech questions. The Question Answering in Speech Transcripts (QAST) evaluation showed no significant difference between the use of written and spoken (although transcribed) questions, indicating that the noise introduced in spontaneous questions does not represent a major issue for Voice-activated QA systems.

VI. Concluding remarks

In this paper, we have focused on the potential offered by Question Answering in the creation of systems that are capable of starting to understand human language. Although progress has been made, we still have not reached full understanding of language even with current human

language technology. Therefore, Voice-activated QA systems currently do not have the "human-like" communication that we would like to have in a machine. Nevertheless, Voice-activated QA systems can indeed be useful, specifically for restricted-domain QA tasks where more limited knowledge is needed.

The next generation of Voiced-activated QA systems will allow users to have a deeper speech interaction in order to obtain answers from different kinds of repositories of unstructured information (e.g. web pages, newspapers and books in a digital format). Nowadays, this ambitious goal can be attempted with the help of advances in the principal fields involved in these processes: speech processing and IR/QA systems. However, there are important and specific difficulties that make this task still a challenge.

AUTHORS

Paolo Rosso received his Ph.D. degree in Computer Science (1999) from the Trinity College Dublin, University of Ireland. He is currently an Associate Professor at the Universitat Politècnica de València, Spain, where he leads the Natural Language Engineering Laboratory of the Natural Language Engineering and Pattern Recognition (ELiRF) research group. He has published over 200 papers in different conferences, workshops and journals being involved in many national and international research projects. His main research interests are mainly focused on question answering, text categorization, geographical information retrieval and automatic plagiarism detection topics on what has organized tracks at CLEF.

Lluís F. Hurtado received his Ph.D. degree in Computer Science from the Universitat Politècnica de València in 2004. He is currently a permanent lecturer in the Departament de Sistemes Informàtics i Computació of the Universitat Politècnica de València. He is member of the Natural Language Engineering and Pattern Recognition (ELiRF) research group at the same institution. His current research interests are mainly focused on language modeling and spoken dialog systems.

Encarna Segarra received her Ph.D. degree in Computer Science from the Universitat Politècnica de València, in 1993. In 1986 she joined the Departament de Sistemes Informàtics i Computació of the Universitat Politècnica de València, where she is now an Associate Professor. She is an active member of the Natural Language Engineering and Pattern Recognition (ELiRF) research group at the same institution. She has published over 80 papers in different conferences, workshops and journals being involved in many research projects. Her current research interests are mainly focused on the automatic learning of language models and its application to spoken dialog systems

and lexical disambiguation.

Emilio Sanchis received his Ph.D. degree in Computer Science from the Universitat Politècnica de València in 1994. He is currently a Full Professor in the Departament de Sistemes Informàtics i Computació of the Universitat Politècnica de València. He is the head of the Natural Language Engineering and Pattern Recognition (ELiRF) research group at the same institution. He has published over 100 papers in different conferences, workshops and journals being involved in many research projects. His main research interests are focused on dialog systems, question answering and automatic learning.

REFERENCES

[1] S. Abney, M. Collins, and A. Singhal, "Answer extraction," in *Proceedings of the Sixth Applied Natural Language Processing Conference*, Seattle, Washington, 2000, pp. 296–301.

[2] R. Aceves, L. Villaseñor, and M. , Montes, "Towards a Multilingual QA System Based on the Web Data Redundancy," in *Advances in Web Intelligence*, ser. Lecture Notes in Computer Science. Berlin / Heidelberg: Springer, 2005, vol. 3528, pp. 32–37.

[3] D. Ahn, V. Jijkoun, G. Mishne, K. Mller, M. de Rijke, and S. Schlobach, "Using Wikipedia at the TREC QA Track," in *Proceedings of the Thirteenth Text Retrieval Conference (TREC 2004)*. Gaithersburg, MD, USA: NIST, 2004.

[4] K. Ahn, B. Alex, J. Bos, T. Dalmas, J. Leidner, and M. Smillie, "Cross-lingual question answering using off-the-shelf machine translation," in *Multilingual Information Access for Text, Speech and Images*, ser. Lecture Notes in Computer Science, vol. 3491. Berlin / Heidelberg: Springer, 2005, pp. 446–457.

[5] T. Akiba, K. Itou, and A. Fujii, "Language model adaptation for fixed phrases by amplifying partial n-gram sequences," *Systems and Computers in Japan*, vol. 38, no. 4, pp. 63–73, 2007.

[6] A. Allauzen and J. Gauvain, "Open vocabulary asr for audiovisual document indexation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05)*, 2005, pp. 1013–1016.

[7] C. Amaral, H. Figueira, A. Martins, A. Mendes, P. Mendes, and C. Pinto, "Priberam's Question Answering System for Portuguese," in *Accessing Multilingual Information Repositories*, ser. Lecture Notes in Computer Science. Berlin / Heidelberg: Springer, 2006, vol. 4022, pp. 410–419.

[8] L. Aunimo, R. Kuuskoski, and J. Makkonen, "Finnish as source language in bilingual question answering," in *Multilingual Information Access for Text, Speech and Images*, ser. Lecture Notes in Computer Science. Berlin / Heidelberg: Springer, 2005, vol. 3491, pp. 482–493.

[9] L. Bahl, F. Jelinek, and R. Mercer, "A Maximum Likelihood Approach to Continuous Speech Recognition," *IEEE Journal of Pattern Analysis and Machine Intelligence*, vol. PAMI-5, no. 2, pp. 179–190, 1983.

[10] M. W. Bilotti, P. Ogilvie, J. Callan, and E. Nyberg, "Structured retrieval for question answering," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. Lecture Notes in Computer Science, Amsterdam, The Netherlands, 2007, pp. 351–358.

[11] G. Bouma, J. Mur, G. van Noord, L. van der Plas, and J. Tiedemann, "Question Answering for Dutch Using Dependency Relations," in *Accessing Multilingual Information Repositories*, ser. Lecture Notes in Computer Science. Berlin / Heidelberg: Springer, 2006, vol. 4022, pp. 370–379.

[12] E. Brill, J. Lin, M. Banko, S. T. Dumais, and A. Y. Ng, "Data-intensive question answering," in *Text REtrieval Conference*, 2001.

[13] S. Buchholz, "Using grammatical relations, answer frequencies and the world wide web for trec question answering," in *Proceedings of the 10th Text REtrieval Conference (TREC-10)*, Gaithersburg, Maryland, 2001, pp. 502–506.

[14] J. Cao, D. Roussinov, J. A. Robles-Flores, and J. F. Nunamaker Jr., "Automated question answering from lecture videos: Nlp vs. pattern matching," in *Proceedings of the 38th Hawaii International Conference on System Sciences (HICSS 2005), IEEE Computer Society*, Big Island, Hawaii, 2005.

[15] J. Chu-Carroll, J. Prager, K. Czuba, D. Ferrucci, and P. Duboue, "Semantic search via XML fragments: A high-precision approach to IR," in *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2006, pp. 445–452.

[16] J. Chu-Carroll and J. Prager, "An experimental study of the impact of information extraction accuracy on semantic search performance," in *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. ACM, 2007, pp. 505–514.

[17] C. Clarke, G. Cormack, and T. Lynam, "Exploiting redundancy in question answering," in *Proceedings of the 24th ACM SIGIR Conference*. ACM Press, 2001, pp. 358–365.

[18] C. de Pablo, A. González, J. L. Martínez, J. M. Guirao, P. Martínez, and A. Moreno, "MIRACLE's Cross-Lingual Question Answering Experiments with Spanish as a Target Language," in *Accessing Multilingual Information Repositories*, ser. Lecture Notes in Computer Science. Berlin / Heidelberg: Springer, 2006, vol. 4022, pp. 488–491.

[19] A. Del Castillo, M. Gómez, and L. Villaseñor-Pineda, "QA on the web: a preliminary study for Spanish language," in *Proceedings of the Fifth Mexican International Conference in Computer Science (ENC'04)*, Colima, Mexico, 2004, pp. 322–328.

[20] A. Echihabi and D. Marcu, "A Noisy-Channel Approach to Question Answering." in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL'03)*, 2003, pp. 16–23.

[21] D. Ferrés, S. Kanaan, A. Ageno, E. González, H. Rodríguez, M. Surdeanu, and J. Turmo, "The TALP-QA System for Spanish at CLEF 2004: Structural and Hierarchical Relaxing of Semantic Constraints," in *Multilingual Information Access for Text, Speech and Images*, ser. Lecture Notes in Computer Science. Berlin / Heidelberg: Springer, 2005, vol. 3491, pp. 557–568.

[22] D. Forney, "The Viterbi Algorithm," *Proceedings of the IEEE*, vol. 61, no. 3, pp. 268–278, 1973.

[23] A. Fujii, K. Itou, and T. Ishikawa, "A method for open-vocabulary speech-driven text retrieval," in *2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, Philadelphia, PA, USA, 2002, p. 188.

[24] G. Geleijnse and J. Korst, "Learning effective surface patterns," in *Workshop on Adaptive Text Extraction and Mining (ATEM)*, 2006.

[25] J. M. Gómez, D. Buscaldi, E. Bisbal, P. Rosso, and E. Sanchis, "QUASAR: The Question Answering System of the Universidad Politécnica de Valencia," in *Accessing Multilingual Information Repositories*, ser. Lecture Notes in Computer Science. Berlin / Heidelberg: Springer, 2006, vol. 4022, pp. 439–448.

[26] M. A. Greenwood, "Using pertainyms to improve passage retrieval for questions requesting information about a location," in *Workshop on Information Retrieval for Question Answering (SIGIR 2004)*, Sheffield, UK, 2004.

[27] K. Hacioglu and W. Ward, "Question classification with support vector machines and error correcting codes," in *HLT-NAACL*, 2003.

[28] S. Harabagiu, D. Moldovan, M. Pasca, R. Mihalcea, M. Surdeanu, R. Bunescu, R. Girju, V. Rus, and P. Morarescu, ""FALCON: Boosting Knowledge for Answer Engines," in *Proceedings of the Text REtrieval Conference (TREC-9)*, 2000.

[29] S. Harabagiu, D. Moldovan, and J. Picone, "Open-Domain Voice-Activated Question Answering," in *COLING2002*, 2002.

[30] U. Hermjakob, "Parsing and question classification for question answering," in *Proceedings of the ACL 2001 Workshop on Open-Domain Question Answering*, 2001, pp. 17–22.

[31] M. Hess, "The 1996 international conference on tools with artificial intelligence (tai 96)," in *Proc. Conference on Research and Development in Information Retrieval (SIGIR 1996)*, Zurich, Switzerland, 1996.

[32] C. Hori, T. Hori, H. Isozaki, E. Maeda, S. Katagiri, and S. Furui, "Study on Spoken interactive open domain Question Answering," in *ISCA / IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR03)*, Tokyo, Japan, 2003, pp. 111–114.

[33] E. Hovy, L. Gerber, U. Hermjakob, M. Junk, and C. Lin, "Question answering in webclopedia," in *The Ninth Text REtrieval Conference*, 2000. [Online]. Available: citeseer.ist.psu.edu/hovy00question.html

[34] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 2nd ed. Prentice Hall, 2009.

[35] K. Kan-Lo and W. Lam, "Using semantic relations with world knowledge for question answering," in *The Fifteenth Text REtrieval Conference (TREC 2006)*. Gaithersburg, MD, USA: NIST, 2006, pp. 403–408.

[36] D. Kim, S. Furui, and H. Isozaki, "Language models and dialogue strategy for a voice QA system," in *18th international congress on acoustics*, Kyoto, Japan, 2004, pp. 3705–3708.

[37] C. Kwok, O. Etzioni, and D. Weld, "Scaling question answering to the web," *ACM Trans. Inf. Syst.*, vol. 19, no. 3, pp. 242–262, 2001.

[38] X. Li and D. Roth, "Learning question classifiers," in *COLING*, 2002.

[39] J. Lin and B. Katz, "Question answering from the web using knowledge annotation and knowledge mining techniques," in *CIKM '03: Proceedings of the twelfth international conference on Information and knowledge management*. New York, NY, USA: ACM Press, 2003, pp. 116–123.

[40] X. Liu and W. Croft, "Passage retrieval based on language models," in *Proceedings of the eleventh international conference on Information and knowledge management*, 2002. [Online]. Available: citeseer.ist.psu.edu/liu02passage.html

[41] J. Luque, D. Ferrés, J. Hern, J. B. Mariño, and H. Rodríguez, "GEOVAQA: A voice activated geographical question answering system," in *Proceedings of JTH 2006*, Zaragoza, Spain, 2006, pp. 309–314.

[42] B. Magnini, M. Negri, R. Prevete, and H. Tanev, "Is it the Right Answer? Exploiting Web Redundancy for Answer Validation," in *Proceedings of the ACL*, 2002, pp. 425–432.

[43] ——, "Multilingual question/answering: the DIOGENE system," in *The 10th Text REtrieval Conference*, 2001. [Online]. Available: citeseer.ist.psu.edu/555536.html

[44] T. Misu and T. Kawahara, "Dialogue strategy to clarify users queries for document retrieval system with speech interface," *Speech Communication*, vol. 48, no. 9, pp. 1137–1150, 2006.

[45] ——, "Bayes risk-based optimization of dialogue management for document retrieval system with speech interface," in *Interspeech 2007*, Antwerpen, Belgium, 2007, pp. 2705–2708.

[46] D. Moldovan, M. Pasca, S. Harabagiu, and M. Surdeanu, "Performance issues and error analysis in an open-domain question answering system," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, New York, USA, 2003.

[47] S. Narayanan and S. Harabagiu, "Question answering based on semantic structures," in *International Conference on Computational Linguistics (COLING 2004)*, Geneva, Switzerland, 2004, pp. 693–702.

[48] G. Neumann and B. Sacaleanu, "Experiments on robust nl question interpretation and multi-layered document annotation for a cross-language question/answering system," in *Multilingual Information Access for Text, Speech and Images*, ser. Lecture Notes in Computer Science, vol. 3491. Berlin / Heidelberg: Springer, 2005, pp. 411–422.

[49] M. Pasca and S. Harabagiu, "The informative role of wordnet in open-domain question answering," in *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*. Pittsburgh, PA, USA: Carnegie Mellon University, 2001, pp. 138–143.

[50] M. Pérez Coutiño, M. Montes y Gómez, A. López López, and L. Villaseñor Pineda, "The Role of Lexical Features in Question Answering for Spanish," in *Accessing Multilingual Information Repositories*, ser. Lecture Notes in Computer Science. Berlin / Heidelberg: Springer, 2006, vol. 4022, pp. 492–501.

[51] I. Roberts and R. J. Gaizauskas, "Data-intensive question answering," in *ECIR*, ser. Lecture Notes in Computer Science, vol. 2997. Berlin / Heidelberg: Springer, 2004.

[52] E. Robertson, S. Walker, and M. Beaulieu, "Experimentation as a way of life: Okapi at trec," *Information Processing and Management*, vol. 36, no. 1, pp. 95–108, 2000.

[53] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manage.*, vol. 24, no. 5, pp. 513–523, 1988.

[54] E. Sanchis, D. Buscaldi, S. Grau, L. Hurtado, and D. Griol, "Spoken QA based on a Passage Retrieval engine," in *IEEE-ACL Workshop on Spoken Language Technology*, Aruba, 2006, pp. 62–65.

[55] S. Schockaert, M. D. Cock, and E. E. Kerre, "Fuzzy constraint based answer validation," in *AWIC*, 2005, pp. 394–400.

[56] E. J. Schofield, "Language models for questions," in *Proceedings of EACL Workshop on Language Modeling for Text Entry Methods*, 2003.

[57] E. J. Schofield and Z. Zeng, "A speech interface for open-domain question answering," in *Proceedings of the ACL*, Sapporo, Japan, 2003, pp. 177–180.

[58] S. Stoyanchev, G. Tur, and D. Hakkani-Tur, "Name-aware speech recognition for interactive question answering," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'08)*, 2008, pp. 5113–5116.

[59] S. Tellex, B. Katz, J. Lin, A. Fernandes, and G. Marton, "Quantitative evaluation of passage retrieval algorithms for question answering," in *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. New York, NY, USA: ACM Press, 2003, pp. 41–47.

[60] J. L. Vicedo and D. Mollá, "Question answering in restricted domains: An overview," *Computational Linguistics*, vol. 33, no. 1, pp. 41–62, 2007.

[61] J. Vicedo, R. Izquierdo, F. Llopis, and R. Munoz, "Question answering in Spanish," in *Working notes of the Cross-Lingual Evaluation Forum (CLEF 2003)*, Trondheim, Norway, 2003.

[62] E. Voorhees, "Overview of trec 2001," in *Proceedings of the tenth Text REtrieval Conference (TREC-10)*, Gaithersburg, Maryland, 2001.

[63] E. Whittaker and S. Furui, "An overview of question answering at Tokyo Institute of Technology," in *Proceedings of the Symposium on Large-scale Knowledge Resources LKR2007*, Tokio, Japan, 2007, pp. 169–174.

[64] E. Whittaker, J. Mrozinski, and S. Furui, "Factoid question answering with web, mobile and speech interfaces," in *Proceedings of the HLT-NAACL conference*, New York, USA, 2006, pp. 288–291.

[65] J. Yousefi and L. Kosseim, "Using Semantic Constraints to Improve Question Answering," in *Natural Language Processing and Information Systems*, ser. Lecture Notes in Computer Science. Berlin / Heidelberg: Springer, 2006, vol. 3999, pp. 118–128.