Anchored Speech Recognition for Question Answering

Sibel Yaman[1], **Gokhan Tur**[2], **Dimitra Vergyri**[2], **Dilek Hakkani-Tur**[1],
Mary Harper[3] **and Wen Wang**[2]
[1] International Computer Science Institute
[2] SRI International
[3] Hopkins HLT Center of Excellence, University of Maryland
{sibel,dilek}@icsi.berkeley.edu,{gokhan,dverg,wwang}@speech.sri.com,mharper@casl.umd.edu

Abstract

In this paper, we propose a novel question answering system that searches for responses from spoken documents such as broadcast news stories and conversations. We propose a novel two-step approach, which we refer to as *anchored speech recognition*, to improve the speech recognition of the sentence that supports the answer. In the first step, the sentence that is highly likely to contain the answer is retrieved among the spoken data that has been transcribed using a generic automatic speech recognition (ASR) system. This candidate sentence is then re-recognized in the second step by constraining the ASR search space using the lexical information in the question. Our analysis showed that ASR errors caused a 35% degradation in the performance of the question answering system. Experiments with the proposed anchored recognition approach indicated a significant improvement in the performance of the question answering module, recovering 30% of the answers erroneous due to ASR.

1 Introduction

In this paper, we focus on finding answers to user questions from spoken documents, such as broadcast news stories and conversations. In a typical question answering system, the user query is first processed by an information retrieval (IR) system, which finds out the most relevant documents among massive document collections. Each sentence in these relevant documents is processed to determine whether or not it answers user questions. Once a candidate sentence is determined, it is further processed to extract the exact answer.

Answering factoid questions (i.e., questions like *"What is the capital of France?"*) using web makes use of the redundancy of information (Whittaker et al., 2006). However, when the document collection is not large and when the queries are complex, as in the task we focus on in this paper, more sophisticated syntactic, semantic, and contextual processing of documents and queries is performed to extract or construct the answer. Although much of the work on question answering has been focused on written texts, many emerging systems also enable either spoken queries or spoken document collections (Lamel et al., 2008). The work we describe in this paper also uses spoken data collections to answer user questions but our focus is on improving speech recognition quality of the documents by making use of the wording in the queries. Consider the following example:

Manual transcription: We understand from Greek officials here that it was a Russian-made rocket which is available in many countries but certainly not a weapon used by the Greek military

ASR transcription: to stand firm greek officials here that he was a a russian made rocket uh which is available in many countries but certainly not a weapon used by he great moments

Question: What is certainly not a weapon used by the Greek military?

Answer: a Russian-made rocket

Answering such questions requires as good ASR transcriptions as possible. In many cases, though, there is one generic ASR system and a generic language model to use. The approach proposed in this paper attempts to improve the ASR performance by re-recognizing the candidate sentence using lexical information from the given question. The motiva-

Proceedings of NAACL HLT 2009: Short Papers, pages 265–268,
Boulder, Colorado, June 2009.

tion is that the question and the candidate sentence should share some common words, and therefore the words of the answer sentence can be estimated from the given question. For example, given a factoid question such as: "*What is the tallest building in the world?*", the sentence containing its answer is highly likely to include word sequences such as: "*The tallest building in the world is NAME*" or "*NAME, the highest building in the world, ...*", where NAME is the exact answer.

Once the sentence supporting the answer is located, it is re-recognized such that the candidate answer is constrained to include parts of the question word sequence. To achieve this, a word network is formed to match the answer sentence to the given question. Since the question words are taken as a basis to re-recognize the best-candidate sentence, the question acts as an *anchor*, and therefore, we call this approach *anchored recognition*.

In this work, we restrict out attention to questions about the subject, the object and the locative, temporal, and causative arguments. For instance, the followings are the questions of interest for the sentence *Obama invited Clinton to the White House to discuss the recent developments*:
Who invited Clinton to the White House?
Who did Obama invite to the White House?
Why did Obama invite Clinton to the White House?

2 Sentence Extraction

The goal in sentence extraction is determining the sentence that is most likely to contain the answer to the given question. Our sentence extractor relies on non-stop word n-gram match between the question and the candidate sentence, and returns the sentence with the largest weighted average. Since not all word n-grams have the same importance (e.g. function vs. content words), we perform a weighted sum as typically done in the IR literature, i.e., the matching n-grams are weighted with respect to their inverse document frequency (IDF) and length.

A major concern for accurate sentence extraction is the robustness to speech recognition errors. Another concern is dealing with alternative word sequences for expressing the same meaning. To tackle the second challenge, one can also include synonyms, and compare paraphrases of the question and the candidate answer. Since our main focus is on robustness to speech recognition errors, our data set is limited to those questions that are worded very similarly to the candidate answers. However, the approach is more general, and can be extended to tackle both challenges.

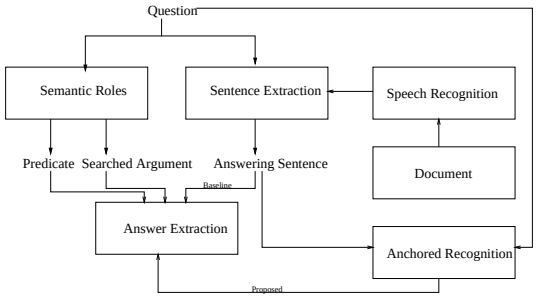

Figure 1: Conceptual scheme of the baseline and proposed information distillation system.

3 Answer Extraction

When the answer is to be extracted from ASR output, the exact answers can be erroneous because (1) the exact answer phrase might be misrecognized, (2) other parts of the sentence might be misrecognized, so the exact answer cannot be extracted either because parser fails or because the sentence cannot match the query.

The question in the example in the *Introduction* section is concerned with the object of the predicate "*is*" rather than of the other predicates "*understand*" or "*was*". Therefore, a pre-processing step is needed to correctly identify the object (in this example) that is being asked, which is described next.

Once the best candidate sentence is estimated, a syntactic parser (Harper and Huang,) that also outputs *function tags* is used to parse both the question and candidate answering sentence. The parser is trained on Fisher, Switchboard, and speechified Broadcast News, Brown, and Wall Street Journal treebanks without punctuation and case to match input the evaluation conditions.

An example of such a syntactic parse is given in Figure 2. As shown there, the "SBJ" marks the surface subject of a given predicate, and the "TMP" tag marks the temporal argument. There are also the "DIR" and "LOC" tags indicating the locative argument and the "PRP" tag indicating the causal argument. Such parses not only provide a mechanism to extract information relating to the subject of the predicate of interest, but also to extract the part of

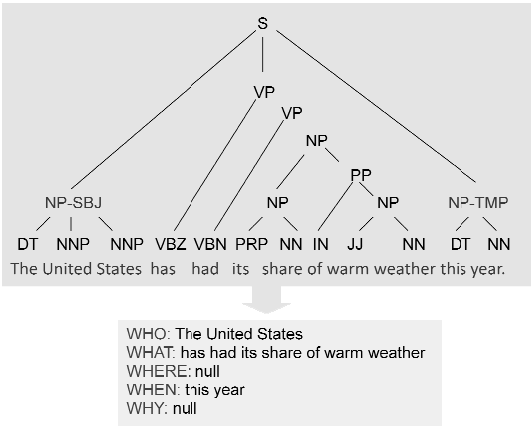

Figure 2: The function tags assist in finding the subject, object, and arguments of a given predicate.

the sentence that the question is about, in this example "*a Russian-made rocket [which] is certainly not a weapon used by the Greek military*". The extraction of the relevant part is achieved by matching the predicate of the question to the predicates of the subsentences in the best candidate sentence. Once such syntactic parses are obtained for the part of the best-candidate sentence that matches the question, a set of rules are used to extract the argument that can answer the question.

4 Anchored Speech Recognition

In this study we employed a state-of-the-art broadcast news and conversations speech recognition system (Stolcke et al., 2006). The recognizer performs a total of seven decoding passes with alternating acoustic front-ends: one based on Mel frequency cepstral coefficients (MFCCs) augmented with discriminatively estimated multilayer-perceptron (MLP) features, and one based on perceptual linear prediction (PLP) features. Acoustic models are cross-adapted during recognition to output from previous recognition stages, and the output of the three final decoding steps are combined via confusion networks.

Given a question whose answer we expect to find in a given sentence, we construct a re-decoding network to match that question. We call this process *anchored speech recognition*, where the anchor is the question text. Note that this is different than forced alignment, which enforces the recognition of an audio stream to align with some given sentence. It is used for detecting the start times of individual words or for language learning applications to exploit the acoustic model scores, since there is no need for a language model.

Our approach is also different than the so-called *flexible alignment* (Finke and Waibel, 1997), which is basically forced alignment that allows skipping any part of the given sentence, replacing it with a reject token, or inserting hesitations in between words. In our task, we require all the words in the question to be in the best-candidate sentence without any skips or insertions. If we allow flexible alignment, then any part of the question could be deleted. In the proposed anchored speech recognition scheme, we allow only pauses and rejects between words, but do not allow any deletions or skips.

The algorithm for extracting anchored recognition hypotheses is as follows: (i) Construct new recognition and rescoring language models (LMs) by interpolating the baseline LMs with those trained from only the question sentences and use the new LM to generate lattices - this aims to bias the recognition towards word phrases that are included in the questions. (ii) Construct for each question an "anchored" word network that matches the word sequence of the question, allowing any other word sequence around it. For example if the question is *WHAT did Bruce Gordon say?*, we construct a word network to match *Bruce Gordon said ANYTHING* where "ANYTHING" is a filler that allows any word (a word loop). (iii) Intersect the recognition lattices from step (i) with the anchored network for each question in (ii), thus extracting from the lattice only the paths that match as answers to the question. Then rescore that new lattice with higher order LM and cross-word adapted acoustic models to get the best path. (iv) If the intersection part in (iii) fails then we use a more constrained recognition network: Starting with the anchored network in (ii) we first limit the vocabulary in the ANYTHING word-loop sub-network to only the words that were included in the recognition lattice from step (i). Then we compose this network with the bigram LM (from step (i)) to add bigram probabilities to the network. Vocabulary limitation is done for efficiency reasons. We also allow optional filler words and pauses to this network to allow for hesitations, non-speech events and pauses within the utterance we are trying to match. This may limit somewhat the potential improvement from this approach and we are working

Question Type	ASR Output	Manual Trans.
Subject	85%	98%
Object	75%	90%
Locative Arg.	81%	93%
Temporal Arg.	94%	98%
Reason	86%	100%
Total	83%	95%

Table 1: Performance figures for the sentence extraction system using automatic and manual transcriptions.

Question Type	ASR Output	Manual Trans.	Anchored Output
Subject	51%	77%	61%
Object	41%	73%	51%
Locative Arg.	18%	22%	22%
Temporal Arg.	55%	73%	63%
Reason	26%	47%	26%
Total	44%	68%	52%

Table 2: Performance figures for the answer extraction system using automatic and manual transcriptions compared with anchored recognition outputs.

towards enhancing the vocabulary with more candidate words that could contain the spoken words in the region. (v) Then we perform recognition with the new anchored network and extract the best path through it. Thus we enforce partial alignment of the audio with the question given, while the regular recognition LM is still used for the parts outside the question.

5 Experiments and Results

We performed experiments using a set of questions and broadcast audio documents released by LDC for the DARPA-funded GALE project Phase 3. In this dataset we have 482 questions (177 subject, 160 object, 73 temporal argument, 49 locative argument, and 23 reason) from 90 documents. The ASR word error rate (WER) for the sentences from which the questions are constructed is 37% with respect to noisy closed captions. To factor out IR noise we assumed that the target document is given.

Table 1 presents the performance of the sentence extraction system using manual and automatic transcriptions. As seen, the system is almost perfect when there is no noise, however performance degrades about 12% with the ASR output.

The next set of experiments demonstrate the performance of the answer extraction system when the correct sentence is given using both automatic and manual transcriptions. As seen from Table 2, the answer extraction performance degrades by about 35% relative using the ASR output. However, using the anchored recognition approach, this improves to 23%, reducing the effect of the ASR noise significantly[1] by more than 30% relative. This is shown in the last column of this table, demonstrating the use of the proposed approach. We observe that the WER of the sentences for which we now get corrected answers is reduced from 45% to 28% with this approach, a reduction of 37% relative.

6 Conclusions

We have presented a question answering system for querying spoken documents with a novel anchored speech recognition approach, which aims to re-decode an utterance given the question. The proposed approach significantly lowers the error rate for answer extraction. Our future work involves handling audio in foreign languages, that is robust to both ASR and machine translation noise.

Acknowledgments: This work was funded by DARPA under contract No. HR0011-06-C-0023. Any conclusions or recommendations are those of the authors and do not necessarily reflect the views of DARPA.

References

M. Finke and A. Waibel. 1997. Flexible transcription alignment. In *Proceedings of the IEEE ASRU Workshop*, Santa Barbara, CA.

M. Harper and Z. Huang. *Chinese Statistical Parsing*, chapter To appear.

L. Lamel, S. Rosset, C. Ayache, D. Mostefa, J. Turmo, and P. Comas. 2008. Question answering on speech transcriptions: the qast evaluation in clef. In *Proceedings of the LREC*, Marrakech, Morocco.

A. Stolcke, B. Chen, H. Franco, V. R. R. Gadde, M. Graciarena, M.-Y. Hwang, K. Kirchhoff, N. Morgan, X. Lin, T. Ng, M. Ostendorf, K. Sonmez, A. Venkataraman, D. Vergyri, W. Wang, J. Zheng, and Q. Zhu. 2006. Recent innovations in speech-to-text transcription at SRI-ICSI-UW. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1729–1744, September.

E. W. D. Whittaker, J. Mrozinski, and S. Furui. 2006. Factoid question answering with web, mobile and speech interfaces. In *Proceedings of the NAACL/HLT*, Morristown, NJ.

[1] according to the Z-test with 0.95 confidence interval