NAME-AWARE SPEECH RECOGNITION FOR INTERACTIVE QUESTION ANSWERING

Svetlana Stoyanchev[1], *Gokhan Tur*[2], *Dilek Hakkani-Tür,* [3]

[1] University of Stony Brook, SUNY, Stony Brook, NY 11794
[2] SRI International, Menlo Park, CA 94025
[3] International Computer Science Institute (ICSI), Berkeley, CA 94704
sveta@cs.sunysb.edu, gokhan@speech.sri.com, dilek@icsi.berkeley.edu

ABSTRACT

In this work we show how interactivity in a voice-enabled question answering application may improve speech recognition. We allow the user to provide a target named entity before asking the question. Then we build a named entity specific language model using the documents containing the named entity. The question-specific model is obtained by merging the named entity specific model with the model built on a set of questions. We present a set of experiments using the TREC question set on the AQUAINT corpus. The question-specific language model is compared with the baseline model built by merging a model of the AQUAINT corpus and past TREC questions. The question-specific model achieves 32.2% reduction in word error rate from the baseline using the questions where pronominal references are resolved.

***Index Terms*—** Spoken question answering, speech recognition, spoken dialog systems

1. INTRODUCTION

Question answering (QA) is the task of automatic retrieval of an answer given a question. Typically the question is linguistically processed, and search phrases are automatically extracted. The search phrases are then used to extract the candidate documents and sentences for the answer.

Question answering provides a natural language interface for information retrieval. This interface also opens the possibility of access to the system by using voice. User of a spoken question answering system may be a reporter on a job who needs to check a fact, a driver on the go, a researcher in the field, or a visually disabled person. Spoken question answering can be seen as a more sophisticated version of spoken information access systems such as phone-based directory assistance [1, among others] or weather/restaurant/flight/hotel information systems [2, among others].

In this work we are addressing speech recognition performance for the spoken question answering task. The word error rates of the state-of-the-art open-domain speech recognition technology are around 25%-30% [3]. Performance is known to be even lower for names and rare words. If a question is asked about a person, an organization, or another named entity, the recognition of this named entity is essential for finding a correct answer.

In this work we propose a method for improving speech recognition of the question by allowing interaction during the question specification phase. The interactivity allows the system to dynamically change the language models based on the dialog state.

We design a voice interface to an open-domain interactive question answering system. We show that the interactivity feature improves speech recognition performance of the open-domain system. In an interactive system a user may first specify the named entity of interest: a person's name, an organization, and so on. A grammar for the named entities is created from a database of named entities existing in the target corpus. If this named entity is recognized, a model specific to the name is used by the speech recognizer. In this study we create models using matching documents from the dataset.

The main idea is that named entities are strongly associated with the content words. For example for the target name *Gordon Gekko*, one question used in TREC 2004 evaluations is *In what film is Gordon Gekko the main character?*, including non-function words related to the movie industry, such as *film* or *character*. Our goal is capturing these content words using the documents where this name appears frequently.

Note that in most question answering evaluations, such as TREC or GALE Distillation, the named entity in consideration is provided in an explicit way. For example in TREC, first the target named entity is given and then several questions are asked about the target. Similarly in the GALE Distillation task, the questions are organized in templates such as *Describe attacks in [LOCATION]* where the variable portion is the named entity. This is in parallel to our design of first getting the name in question.

In the next section we review related work in the speech recognition literature. We present our modeling approach in detail in Section 3. Section 2 describes the related work. Then Section 4 describes the experiments we performed using the TREC benchmark evaluation questions and the AQUAINT corpus along with the results. We conclude in Section 5.

1-4244-1484-9/08/$25.00 ©2008 IEEE

2. RELATED WORK

To the best of our knowledge, no prior study attempted to use named entities for better speech recognition. The closest study in speech recognition literature is the topic-based language adaptation. Iyer and Ostendorf [4] used mixture models for broadcast news recognition. The documents in the training data are clustered using the expectation maximization (EM) algorithm to obtain topically coherent sentences. For each topic, separate language models are created. Then either offline (static) or online (dynamic) mixture language model adaptation is proposed. They obtain 21% reduction in perplexity and 4.5% lower word error rate on the Wall Street Journal corpus using static mixture modeling. They also applied this approach for conversational telephone speech recognition using the Switchboard corpus [5]. The conversations in this corpus have already been marked with 71 topics. They still employed the clustering approach using this annotation as the seed and came up with five clusters. They obtained a humble 1.2% relative reduction in the word error rate. Gildea and Hofmann proposed combining the topic language model with a generic model during runtime in a dynamic fashion [6]. Similar to the previous work the EM algorithm is used for clustering the documents. Model combination is done using a linear or log-linear interpolation. While they obtain 16% lower perplexity using the resulting language model, the word error rate increased by 2.5% relative on the TDT-pilot corpus. In spoken dialog systems, on the other hand, it is a usual practice to use dialog state specific language models [7]. For example after the confirmation prompt it is more likely the user will say *yes* or *no*. This study is different in that in addition to the dialog state, we also exploit the information (i.e. the target name) gathered from the user in the earlier turns.

3. APPROACH

We simulate the interactive system where the user first specifies a target named entity. The named entity concept is grounded: the user confirms that the named entity is recognized correctly. In the case of continuous misrecognition, a named entity may be spelled. This task has been widely studied in the framework of directory assistance systems [1, among others]. The idea is limiting the language model using the names in consideration and such systems perform with very high accuracy. A keypad aided spelling correction may be used as a back-off mechanism [8] where the user the phone keypad while spelling the name.

Figure 1 shows the control flow of the simulated system. First, a user is asked to specify the target named entity. The recognition uses grammar generated from the AQUAINT Named Entity database [9]. The system then asks the user to specify the question about the given named entity. Meanwhile, a question-specific language model is built.

In this paper we focus on the experiments that show improvement in speech recognition of the question by employing interactivity. Once the target named entity is recognized by the system, a target-specific model is built. To this end, we employ a search engine to extract the documents matching the named entity in the target corpus and use these documents to build the name-specific language model. We hypothesize that these documents will be likely to contain the lexicon of the question resulting in a more relevant model for speech recognition. For example, a question *On what date did Michael Brown resign as head of FEMA*, the words *resign* and *FEMA* may have relatively low probability in a generic model, but higher probability in the top matching documents. The documents are extracted from the AQUAINT corpus indexed by the Lucene information retrieval engine [10]. We match the string pattern of the target named entity using the Lucene API.

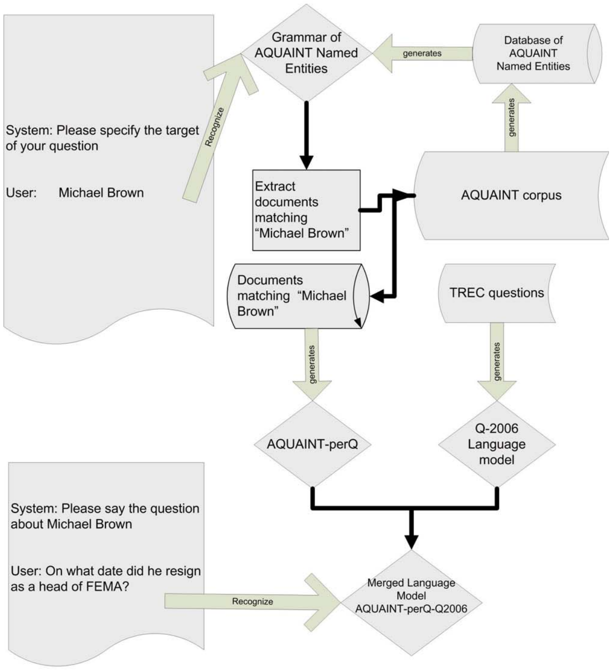
System: Please specify the target of your question
User: Michael Brown
Recognize
Grammar of AQUAINT Named Entities
generates
Database of AQUAINT Named Entities
generates
Extract documents matching "Michael Brown"
AQUAINT corpus
Documents matching "Michael Brown"
TREC questions
generates
generates
AQUAINT-perQ
Q-2006 Language model
System: Please say the question about Michael Brown
User: On what date did he resign as a head of FEMA?
Recognize
Merged Language Model AQUAINT-perQ-Q2006

Fig. 1. Dialog flow example.

While the name-specific language model, LM_{AperQ}, provides the context words, the questions from the earlier TREC evaluations provide the typical characteristics of questions, such as the Wh- words at the sentence initial position. So we train a separate model using these questions, named LM_{TREC}. These two models are then merged using linear interpolation. The interpolation weight, λ, is kept constant as optimized on a couple held-out spoken questions.

$$P_{LM}(W) = \lambda \times P_{LM_{AperQ}}(W) + (1 - \lambda) \times P_{LM_{TREC}}(W)$$

Note that this approach is only for using name-specific language models, and in all experiments we kept the acoustic model fixed. Using this approach while we have a better

Original TREC question	How many times has **Limbaugh** been married?
Target NE	Rush Limbaugh
Modified with NE (set 3)	How many times has **Rush Limbaugh** been married?
Modified without NE (set 4)	How many times has **he** been married?

Table 1. Example of the question in the test set

Model	type	vocab size	description
Q-2007	general	5,337	TREC questions containing test set (total 4158 questions)
Q-2006	general	5,012	TREC questions **not** containing test set (total 3713 questions)
AQUAINT	general	3,000	all AQUAINT documents
AQUAINT-Q2006	general	6,344	all AQUAINT documents merged with the TREC questions
AQUAINT-perQ	per target name	7,211	up-to-100 top matches for the target of the question
AQUAINT-perQ-Q2006	per target name	10,210	up-to-100 top matches for the target of the question merged with the TREC questions

Table 2. Models used in the experiment

language model, its size is also smaller than the one obtained using the whole target corpus. This is very important for efficiency of a real-time recognizer.

4. EXPERIMENTS AND RESULTS

In this study we use the TREC [11] annual benchmark evaluation questions targeting the AQUAINT corpus consisting of 3 GB of written news. The corpus is indexed using the Lucene [10] information retrieval engine.

We have selected 40 questions from the TREC 2007 evaluations. For 18 of the selected questions in the test set the target is a person, for 17 of the questions the target is an organization, and 5 of the questions have other type of target.

The questions are modified for the experiments. In *set3*, all questions are modified to contain the named entity. That is, if the original question contains a pronoun referring to the target, it is replaced with an appropriate form of the target. In *set4*, all questions are modified to *not* contain the named entity by replacing it with an appropriate pronoun. Table 1 describes how the questions are modified for the experiment. 40 questions with resolved and 40 questions with unresolved named entities are read and recorded by three subjects.

We compare using target-specific language models and a generic language model for the recognition of questions. All models in this experiment are built using the the SRILM language modeling toolkit [12]. The speech recognition experiments are performed using SRI's Dynaspeek$^{\text{TM}}$ [13] speech recognition system.

The models used in the experiment are summarized in Table 2. We report the average number of named entities missed from *set3*, the average word error rate over 40 questions, and the relative error rate reduction for the test model AQUAINT-perQ-Q2006 from each other model in Table 3. The TREC-2007 model is a "cheating model" that contains the questions used in the experiment. This model expectedly achieves the lowest error rate of 19.77% on *set3* (with named entities) and 17.27% on *set4* (without named entities).

The TREC-2006 model is the first baseline model built from 3713 TREC questions not containing the test set. This model has a relatively high error rate of 45.65% on the *set3* and 32.13% on the *set4*. Notice that although *set4* does not contain the target named entities, its error rate on the TREC-2006 model is relatively 45.9% higher than on the TREC-2007 "cheating" set. This shows the importance of the content words associated with the target names.

The second baseline model is built using the 3 GB AQUAINT corpus, pruning the vocabulary to 3000 words (guided by a system constraint). The AQUAINT model has the highest error rate of 58.36% on *set3* and 46.64% on *set4*. Although the AQUAINT corpus has large vocabulary coverage, the form of the questions differs from the form of the sentences in the corpus (such as sentence starting with Wh- words). We merge the AQUAINT model with the model trained with only questions, reducing the error rate by 25.4% on *set3* and 38.9% relatively on *set4*. The higher error rate reduction on the *set4* shows that the recognition improvement is not due to the better recognition of named entities.

Next, we create a per-question model. The AQUAINT corpus is indexed with Lucene and queried using Lucene search API to extract as many as 100 documents matching the target named entity. These documents are used to build a question-specific language model AQUAINT-perQ. This model achieves 42.51% on *set3* and 42.59% on *set4*. Our final model AQUAINT-perQ-Q2006 is a merger of the AQUAINT model with the Q-2006 model. This model achieves the lowest WER among all tested models (except the "cheating" Q-2007 model) of 32.4% on *set3* and 28.65% on *set4*. This is a relative reduction of 32.2% on *set3* compared to the best generic model performance.

Note that, in addition to the dramatic reduction in word

Training/testing	num ne missed	40 unres set 3 avg	40 res set 4 avg	% err reduction set3	% err reduction set4
Q-2007	14.67	19.77	17.27		
Q-2006	35.67	45.65	32.12	35.3	19.0
AQUAINT	37.33	58.36	46.64	49.4	44.2
AQUAINT-Q2006	36	43.55	28.49	32.2	8.7
AQUAINT-perQ	14.67	42.51	42.59	30.5	38.9
AQUAINT-perQ-Q2006	14.67	**29.55**	**26.02**		

Table 3. Results averaged between the 3 speakers: number of named entities missed, error rate, relative error reduction for the AQUAINT-perQ-Q2006 model

error rate, the ratio of missed named entity recognitions are halved coming down to levels which can be obtained using the cheating experiment.

We would like to point out the difference in speech recognition results between the speakers. Speakers 1 and 2 have higher WER on *set3* than on *set4* for all the models; however, speaker 4 achieves higher WER on *set4* for the AQUAINT-perQ model. It is possible that speaker 4 was very clear in pronouncing the target named entities and was able to achieve lower WER on the models that contain target named entities. We would like to perform this experiment on a larger subject pool to study the variability between the subjects.

5. CONCLUSION AND FUTURE WORK

We have presented an approach for improving the speech recognition accuracy for spoken question answering systems. The idea is using name specific language models where the target name in question is asked to the user beforehand. This approach has been effective in our experiments using the TREC benchmark questions on the AQUAINT corpus.

Future work includes the grounding process of the named entity. It may involve asking a user for the type of the named entity (e.g. person, organization, location, movie) and associations and building a focused grammar. For example, a user may be asking a question about *Orhan Pamuk*, a Turkish writer. If the user specifies that the target is a person, a writer, and of Turkish descent, a more focused grammar may be built that would allow the recognition of the named entity.

Acknowledgments: We would like thank Dr. Amanda Stent for her encouragement and discussion of the ideas.

6. REFERENCES

[1] A. Kellner, B. Rueber, and H. Schramm, "Strategies for name recognition in automatic directory assistance systems," in *Proceedings of IEEE IVTTA*, Torino, Italy, 1998.

[2] V. Zue, S. Seneff, J. R. Glass, J. Polifroni, C. Pao, T. J. Hazen, and L. Hetherington, "Jupiter: a telephone-based conversational interface for weather information," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 1, pp. 85–96, 2000.

[3] G. Riccardi and D. Hakkani-Tür, "Active and unsupervised learning for automatic speech recognition," in *Proceedings of the EUROSPEECH*, Geneva, Switzerland, September 2003.

[4] R. Iyer and M. Ostendorf, "Modeling long distance dependencies in language: Topic mixtures versus dynamic cache model," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 1, pp. 30–39, 1999.

[5] J.J. Godfrey, E.C. Holliman, and J. McDaniel, "Switchboard: telephone speech corpus for research and development," in *Proceedings of ICASSP*, San Francisco, CA, 1992.

[6] D. Gildea and T. Hofmann, "Topic-based language models using em," in *Proceedings of Eurospeech*, Budapest, Hungary, September 1999.

[7] F. Bechet, G. Riccardi, and D. Hakkani-Tür, "Mining spoken dialog corpora for system evaluation and modeling," in *Proceedings of EMNLP*, Barcelona, Spain, July 2004.

[8] S. Parthasarathy, "Experiments in keypad-aided spelling recognition," in *Proceedings of ICASSP*, Montreal, 2004.

[9] Levon Lloyd, Dimitrios Kechagias, and Steven Skiena, "Lydia: A system for large-scale news analysis," in *SPIRE*, 2005, pp. 161–166.

[10] "Lucene search engine," http://lucene.apache.org/.

[11] H. T. Dang1, J. Lin2, and D. Kelly, "Overview of the TREC 2006 question answering track," in *NIST Special Publication: SP 500-272*, 2006.

[12] A. Stolcke, "SRILM – An Extensible Language Modeling Toolkit," in *Proceedings of the ICSLP*, Denver, CO, September 2002.

[13] Horacio Franco, J. Zheng, J. Butzberger, F. Cesari, M. Frandse, J. Arnold, V. Ramana, R. Gadde, A. Stolcke, and V. Abrash, "Dynaspeak: Sri's scalable speech recognizer for embedded and mobile systems," in *Proceedings of HLT*, San Diego, CA, 2002.