DESIGN AND EVALUATION OF A LIMITED TWO-WAY SPEECH TRANSLATOR

David Stallard, John Makhoul, Frederick Choi, Ehry Macrostie, Premkumar Natarajan, Richard Schwartz, Bushra Zawaydeh

BBN Technologies, Cambridge, MA 02138
{stallard,makhoul,fchoi,emacrost,pnatarajan,schwartz,bushra}@bbn.com

ABSTRACT

We present a limited speech translation system for English and colloquial Levantine Arabic, which we are currently developing as part of the DARPA Babylon program. The system is intended for question/answer communication between an English-speaking operator and an Arabic-speaking subject. It uses speech recognition to convert a spoken English question into text, and plays out a pre-recorded speech file corresponding to the Arabic translation of this text. It then uses speech recognition to convert the Arabic reply into Arabic text, and does information extraction on this text to find the answer content, which is rendered into English. A novel aspect of our work is its use of a statistical classifier to extract information content from the Arabic text. We present evaluation results for both individual components and the end-to-end system.

1. INTRODUCTION

English-speaking field workers often need to communicate with residents of a host country who do not speak English. In a crisis situation, there will be little time to train personnel in the host country language, and human interpreters will often be in short supply. Portable devices for speech-to-speech language translation would therefore be very useful in such environments.

A partial solution to these problems is the recently developed DARPA One-Way translator [1]. This device consists of a PDA with a microphone. The English speaker, whom we shall call the *operator*, uses the one-way device by speaking into it one of a fixed set of utterances in English. The device uses a speech recognizer to determine which of the utterances he has spoken, and then uses table lookup to find a pre-recorded audio file containing the translation of that utterance into the language of the other party, whom we shall call the *subject*.

The DARPA One Way avoids having to translate in the reverse direction by designing all questions so that the subject can reply to them with a gesture (nodding the head, holding up the number of fingers, etc), rather than by speaking. While this approach has proven useful in field trials, it has obvious limitations in terms of efficiency and expressive power.

Our goal in the present work is to overcome these limitations by building a two-way translator, in which the subject answers questions verbally in his own language, and his answers are rendered into spoken or textual English for the operator. While the work here is intended to be generally useful across languages, the current target language for this effort is colloquial Levantine Arabic, which is the dialect of Arabic spoken by natives of Lebanon, Jordan, Syria, and Palestine. The target domain is refugee/medical processing

Special challenges are presented by such an application. First, for Levantine Arabic, as for many languages of the developing world, large corpora of transcribed speech data do not exist. Levantine Arabic is quite different from the Modern Standard Arabic (MSA) that is used in broadcasts and written communication. Not only does Levantine Arabic have radically different word pronunciations from MSA, it also has different words, and for the latter there is no agreed-upon spelling. Furthermore, Levantine Arabic is itself divided into different national sub-dialects, each with their own differences in pronunciation and vocabulary. Quite apart from dialectal issues, there are other challenges to be expected in the actual use of such a system, including outdoor interviewing conditions, background noise, and operator and subject stress.

All of these issues mean that the speech recognition part of the task will be difficult, and word error rates will be high. For this reason, a more restricted and limited approach seems appropriate, at least for an initial system. An analogy can be made with the many system-directed telephone dialogue systems that are in use today. In these systems, the problem of recognizing and understanding the user's speech is kept to a minimum because the interaction is largely controlled by the system. In our system, it is the operator who controls the interaction by asking the questions, which in turn provides the context for interpreting the answers.

Our system does not attempt to do translation as such. Rather, we view the problem as one of information extraction, in which we attempt to extract just the sought-after information from the foreign language utterance, and render this information into English. A novel aspect of our work is that we employ statistical techniques to extract this information. The advantage of such techniques is that they are more robust to speech errors, as well as to lexical variations in the way different subjects express the same information, than are rule-based approaches

In the remainder of the paper, we present the design and overall approach of our system, with special emphasis on the information extraction problem.

2. ARCHITECTURE

A block diagram of the system is shown in Figure 1. The operator's speech is captured by a microphone and sent to an English-language recognizer, which produces a text version of his question, such as "What is your occupation?" or "What is your date of birth?". The resulting English string is then passed to a table lookup module, which maps the English string to a prerecorded utterance in the subject's language. It also maps the utterance to its semantic type, such as OCCUPATION or DATE. (There are approximately 40 semantic types in the system.) The prerecorded utterance is then played out to the subject.

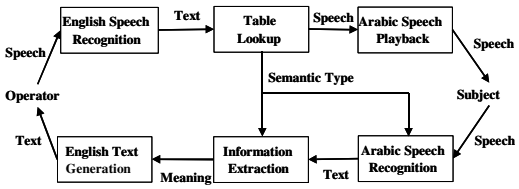

Figure 1: System Architecture

The subject replies into the microphone, and his speech is sent to a foreign-language recognizer. The foreign language recognizer produces a text string in the subject's language, using the semantic type to select an appropriate language model. This text string is then passed to an information extraction module, which computes from it a representation of the relevant meaning. The meaning representation is then passed to an English-language text generator, which generates an English text string from it, using straightforward template-based generation. The resulting English is then displayed to the operator on a small screen, or spoken via a speech synthesizer.

3. DATA COLLECTION

Data collection is obviously vital to the success of a system such as the one described here. To collect the data, a set of scenarios was generated, organized around various themes relevant to the refugee/medical domain. Each scenario consisted of a series of questions that were asked of the subject. For example, one scenario dealt with the collection of biographical information ("What is your name?", "How old are you?"), while another dealt with the problem of finding a missing person ("When did you last see him?", "What color eyes does he have?"). Still others dealt with gathering medical information. In some cases, the subjects were given a written narrative describing the person that they were to pretend to be, including that person's name, age, medical problems, etc. These narratives were written in English so as not to prime the subject with particular Arabic forms. In other cases, subjects were allowed to give their own personal information, or to make up answers to the question, using a picture of a person as a cue, for instance.

Native Levantine speakers (Lebanese, Palestinian, Jordanian, and Syrians) were recruited to play the part of subjects. Some of the data was collected in the Boston area, but most was collected in Beirut, Lebanon, in collaboration with the American University of Beirut (AUB). A noise-canceling headset microphone was used to record the data. Table 1 gives total statistics for the collection:

Number of subjects	167
Number of utterances	77K
Vocabulary	15K
Total words	287K
Hours of speech	45 hrs

Table 1: Data Collection

4. TRANSCRIPTION AND ANNOTATION

A key issue in transcription for this project was deciding how to spell many of the words. There are many words in colloquial Arabic that do not appear in MSA, and there is thus no agreed-upon spelling for them. To ensure transcription consistency, a set of guidelines was produced specifying the correct spelling for various words. To minimize the number of differences between regional sub-dialects, allophones were transcribed as their underlying phonemes. For non-colloquial words, the MSA spelling was used as much as possible

Once subject responses were transcribed, they also had to be annotated with the answer content to be extracted from them. This annotation consisted of marking the answer content portion of the utterance, and translating just that portion into English. An example of an annotated response to the question "What is your job?", rendered in English for readability, would be "I work as a nurse in Homs". In this annotated utterance, the response "nurse" is the answer, and is therefore underlined and marked with its English translation.

This annotation yielded a corpus in which each utterance was partitioned into a sequence of sub-strings. Each sub-string was either marked as answer content, and associated with its English translation, or marked as general Arabic, and not translated. Multiple answers ("I am a nurse but I am unemployed now") were allowed.

5. SPEECH RECOGNITION

Because only a fixed set of English questions can be asked with this system, a finite-state grammar can be employed for recognizing the English question. This grammar includes variant forms of the question (e.g. "What is your weight?" vs. "How much do you weigh?"), but is otherwise straightforward.

The Arabic-speaking subject has a good deal more freedom in formulating his response, so the speech recognition of the Arabic response requires a statistical language model. For this recognition task, BBN's Byblos system is used [2]. A 35-hour subset of our Arabic data was the sole source of training. A single Arabic acoustic model was trained for the entire system, while separate Arabic language models were trained for each semantic type.

6. INFORMATION EXTRACTION

6.1 Overview

Our system employs two different information extraction strategies, depending upon the semantic type of the answer to be extracted. The first strategy is bottom-up partial parsing with a semantic grammar, used for types such as DATE or AGE. The other strategy is a statistical extraction technique that uses a Hidden Markov Model (HMM). We devote a subsection to each of these approaches in what follows, with emphasis on the HMM approach. Evaluation results will be given in section 7.

6.2 Partial Parsing for Structured-Value Answers

Partial parsing is used for semantic types such as DATE and TIME, and for amounts, such as AGE or WEIGHT. We call such answer-types "structured values" because they have constituent parts, such as the month, day, and year for a date, or the quantity and unit for an amount, which must be determined as part of the extraction and translation process. Natural languages typically have well-defined sub-grammars for these expressions, so it makes sense to employ these grammars as part of the translation process.

Our system includes a semantic grammar for each structured value type, in which attached semantic rules produce a meaning representation of the sought-after answer content. This meaning representation is a slot/filler meaning frame like the following:

[AGE qty: 20 unit: YEAR]

To produce such structures, a bottom-up, all-paths chart parser is run on the input, and produces a filled-in chart. The system extracts from the chart the longest parsed constituent that has a meaning of the desired type. This constituent need not span the entire input string, but may be only a substring of it, as in "My name is Samir and I am 20 years old".

6.4 HMM-based Text Classification

There are other semantic types for which the semantic grammar approach is not suitable. An example is the semantic type OCCUPATION. Suppose the subject has given a response for which the English answer "student" is the correct translation. There are many ways to convey this information. A person might say the Arabic word for "student", or he might say, "I'm in school" or "I'm going for a Master's", among many other possibilities. It is clearly not feasible to write a grammar to generate all of these.

To cope with this problem, we define a generative, probabilistic model in which the observed sequence of Arabic words **W** is produced by a sequence of language-states **S**. The language states correspond to either one of the answers ("student", "nurse", "unemployed", etc.), or to general Arabic. We thus seek the sequence of states **S** that maximizes the likelihood P(**W**|**S**).

We represent this model with an HMM as shown in Figure 2. The PREAMBLE and POSTAMBLE states correspond to the general Arabic parts of the utterance that precede and follow the answer content, respectively. The model starts out in either PREAMBLE or one of the answer states, and can go through any number of answer states or through PREAMBLE again before finally terminating with POSTAMBLE. To determine the English answers for a given Arabic utterance, one can simply decode using the Viterbi procedure, find the list of answer states traversed, and present these as the translation of the utterance. (The model topology reflects the assumption that some answer will be given; this assumption could be relaxed by adding a NO_ANSWER state.)

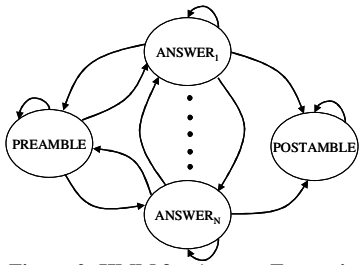

Figure 2: HMM for Answer Extraction

So far, we have spoken of the language states as if they were just ordinary states in an HMM, without any internal structure of their own, and with only the familiar unigram word output probabilities. Such a model is indeed possible, but we have found that it does not perform well. The reason is that Arabic answer content typically has some phrasal structure, and this is not adequately modeled by unigram word probabilities. As a simple example, consider the answer state SINGLE of the semantic type MARITAL_STATUS. One of the Arabic forms for this answer is a two-word utterance exactly like the English form "not married". Because the Arabic word for "married" has a high output probability in the answer state MARRIED, a response containing "not married" can, depending upon other probabilities, be wrongly classified.

To overcome this problem, we define a model in which the language states are sub-models that contain a bigram language model for the language state. A simplified example for the language state SINGLE is shown in Figure 3. Each such sub-model has a null state BEGIN, a state for each word of the vocabulary, and a null state END. The bigram word probabilities are represented by the transition probabilities between states, while each state has a unit output probability for the word it represents. This model is similar to that of Nymble [3], a statistical name-finding system.

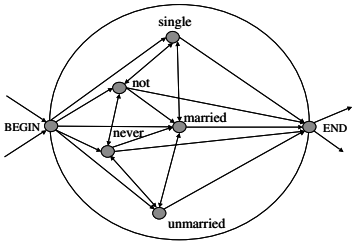

Figure 3: Sub-Model for Language State SINGLE

As implemented, our system does not actually have a separate state for each word in the 15K word vocabulary. Instead, it contains just one HMM state for each language state, and the Viterbi algorithm is modified to use bigram probabilities directly. When computing the transition probability from one language-state S_{i-1} to another language state S_i, there are two cases. If S_{i-1} and S_i are different language states, the model uses the following formula, representing the compound transition from the previous word-node w_{i-1} to the END node of S_{i-1}, across to the BEGIN node of S_i, and from the BEGIN node of S_i to the current word-node w_i:

$$P(\text{END} \mid w_{i-1},S_{i-1})P(S_i \mid S_{i-1})P(w_i \mid \text{BEGIN},S_i)$$

If S_{i-1} and S_i are the same language state, the transition probability is just

$$P(w_i \mid w_{i-1},S_i)$$

The word bigram probabilities for each language state are smoothed with the word unigram probabilities for that state, which is in turn smoothed with the uniform distribution for the entire semantic type. The Witten-Bell method [4] is used.

7. EVALUATION RESULTS

At this point in the project, evaluation results are necessarily preliminary. We are able to present results for a subset of the semantic types of the system, however.

The criterion for success is accuracy of translation, relative to the corpus of annotations we have developed. Because more than one answer can be produced for each utterance (there is an average of 1.2 answers per utterance), we need an error metric that can cope with multiple answers. We define for this purpose a Translation Error Rate (TER), equivalent to the Slot Error Rate metric defined in [5]. This metric is the sum of the insertions, substitutions, and deletions of answers, divided by the total number of correct answers in the test corpus.

A 10-hour test set of speech was held out to test both the speech recognition and translation components. To test the speech recognition component, separate language models were trained for each type, tested on the relevant subset of the test set, and the results combined. The total Word Error Rate (WER) achieved on this test set was 26.2%.

The translation component was separately tested on both the transcriptions and the speech recognizer output for the test set. We first report results for two structured-value semantic types, AGE and WEIGHT, using the partial-parsing method. We give WER and TER for both the transcribed text and the speech recognizer output, in Table 2.

	AGE	WEIGHT
WER	7.8	13.8
TEXT TER	4.7	2.9
SPEECH TER	7.9	12.2

Table 2: Parsing Translation Results

These results show the degradation with speech error that one would expect with a hand-written grammar-based approach.

We next give results for the HMM-based method, tested on the types MARITAL_STATUS, GENDER, RELIGION, EDUCATION, and OCCUPATION (results for OCCUPATION are given only for the most common 30 occupations in the corpus) in Table 3.

	MAR	GEN	REL	EDU	OCC
WER	15.7	18.7	13.7	24.2	27.5
TEXT TER	0.8	0.4	0.2	5.4	6.3
SPEECH TER	3.2	4.9	2.5	7.2	11.6

Table 3: HMM Translation Results

These figures show that the HMM-based method is quite successful, giving near-zero translation error rates on text for some of the simpler answer types, and low overall translation error rates, even in the presence of relatively high WER. Even for the more difficult OCCUPATION class, whose WER is quite high, the performance is still fairly good.

8. CONCLUSIONS

We have presented a limited, task-directed speech translation system for English and Levantine colloquial Arabic. While still in an initial stage, this system shows promise in terms of its overall performance. The work presented here has two novel and interesting aspects. First, it incorporates speech recognition work on a colloquial Arabic dialect, as distinct from Modern Standard Arabic. Second, our work makes use of a novel HMM-based method to extract the desired information from the recognized text, which works well even in the presence of speech recognition errors.

9. ACKNOWLEDGEMENTS

This work was sponsored, in part, by DARPA and monitored by SPAWAR Systems Center under Contract No. N66001-99-D-8615.

10. REFERENCES

[1] Ace Sarich. (2000) The DARPA One-Way Phrase Translation System. *URL: http://www.sarich.com/translator/*

[2] Nguyen L., Anastasakos T., Kubala F., LaPre C., Makhoul J., Schwartz R., Yuan N., Zavaliagkos G., and Zhao Y. (1995) "The 1994 BBN/BYBLOS Speech Recognition System", In *Proc of ARPA Spoken Language Systems Technology Workshop*, Austin, Texas, pp. 77-81.

[3] Bikel, D., Schwartz, R., and Weischedel, R. "An algorithm that learns what's in a Name," (1999) *Machine Learning 34*, pp. 211-231.

[4] Witten, I., and Bell, T. (1991) "The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *IEEE Transactions on Information Theory, 37(4), 1085-1094.*

[5] Makhoul, J., Kubala, F., Schwartz, R., Weischedel, R. (1999) "Performance Measures for Information Extraction" In *Proc. DARPA Broadcast News Workshop*, Herndon, Virginia, Morgan Kaufmann Publishers, pp. 249-252, Feb. 28-March 3, 1999.