A Trainable Approach for Multi-Lingual Speech-To-Speech Translation System

Y. Gao, J. Sorensen, H. Erdogan, R. Sarikaya, F. Liu, M. Picheny
IBM T. J. Watson Research Center
yuqing@us.ibm.com

B. Zhou
Center for Spoken Language Research
Univ. of Colorado at Boulder
zhoub@cslr.colorado.edu

Z. Diao
Department of Mathematics
Texas A&M University

ABSTRACT

This paper presents a statistical speech-to-speech machine translation (MT) system for limited domain applications using a cascaded approach. This architecture allows for the creation of multilingual applications. In this paper, the system architecture and its components, including the speech recognition, parsing, information extraction, translation, natural language generation (NLG) and text-to-speech (TTS) components are described. We have implemented the described system for translating speech between Mandarin and English language pair in an air travel application domain. We are current porting the system to the military domain. Encouraging experimental results have been observed and are presented.

Keywords

speech-to-speech translation, statistical natural language understanding based translation, statistical natural language generation

1. INTRODUCTION

Commerce and travel have created an ever increasing need for translation between languages. Recently, progress in the fields of speech and language processing have begun to allow the creation of automated systems to accomplish this task. However, the technical challenges of creating a useful speech-to-speech translation device pushes against the limitations of current technologies such as speech recognition, natural language understanding, machine translation, natural language generation, and text-to-speech synthesis. In this paper, we present a speech translation system employing a statistical framework appropriate for use in language restricted domains. In our cascaded approach, the recognition results obtained in the speaker's language are analyzed and then, through a series of distinct abstract representations, corresponding sentences are generated in the language of the listener.

Compared to other speech translation systems [1], [2], [3], [4] developed by other researchers, our system has several distinguished features. First, all the components in our system, including parser and language generation, are statistics based, and therefore are trainable from speech or text corpus. No handcrafted rules, grammars, or templates are needed. This makes our system flexible, scalable, and robust as the application domain changes. We have built the system for the air travel domain and are currently porting it to the military domain to verify the robustness. Another feature is that our translation is based on the understanding of the meaning of the sentence. It avoids the literal translation and the correct grammar assumption of the transcribed text from the input speech, as it most likely includes recognition errors, and emphasizes meaning preservation. When the automatic transcribed speech script is parsed, the statistical parser focuses on the meaningful phrases and automatically ignores irrelevant portions or out-of-domain concepts in the script. Another innovative characteristic of our system is that we are trying to use the natural language understanding component to help the speech recognizer, as we believe the language structure and conversation structure can be utilized to improve speech recognition performance.

This paper is organized as follows: In Sec. 2, we describe the system architecture and its components, including the speech recognition, parsing, information extraction, translation, natural language generation (NLG) and text-to-speech (TTS) components. Sec. 3 presents the evaluation results and discussions and Sec. 4 contains concluding remarks.

2. SYSTEM OVERVIEW

Figure 1 shows the architecture of our speech translation system. The input speech is recognized through an automatic speech recognizer and parsed by a statistical natural language understanding (NLU) model. An information extraction component is responsible for analyzing the semantic tree that was obtained from the NLU, and extracting two kinds of information from the tree. The first kind of information is a language independent "interlingua" representation. This, combined with a canonical representation of the language dependent attributes instantiated within the semantic model, is sent to a natural language generation (NLG) engine to render in the target language. Both types of information are translated using distinct models, with the specific attributes of items, such as times and dates, using conventional techniques familiar to the machine translation community. The interlingua translation, however, takes place at a semantic level and can result in considerable surface changes in the final result. Finally, when a text representation of the utterance in the target language is completed, a text-to-speech synthesizer is used to produce spoken output.

For a cascaded approach to machine translation to work, the hierarchical information represented in the semantic tree for a sentence

Proceedings of HLT 2002, Second International Conference on Human Language Technology Research, M. Marcus, ed., Morgan Kaufmann, San Francisco, 2002.

must be invariant across translations, at least in the application domain of interest. This invariance has been validated, in part, by work involving translation between English and certain European languages [4]. A more complete test of this assumption would require dissimilar languages that have quite different phrase order conventions, such as Chinese and English, as in the work presented here. Part of the design of our system is reflected in the annotation of the training corpus, which forms the embodiment of a particular interlingua style.

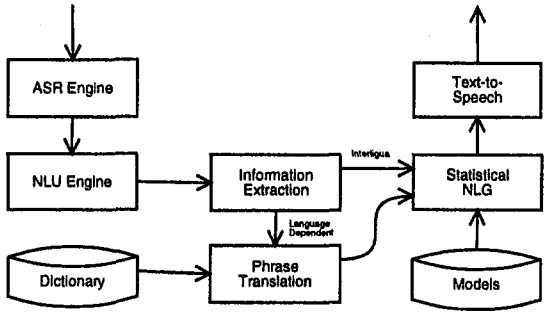


Figure 1: Statistical Speech-to-Speech Translation System Architecture

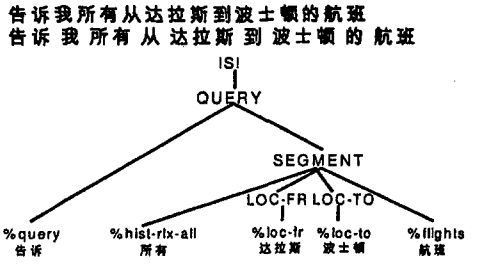
2.1 Statistical NLU Classer and Parser

The natural language understanding component is one of the most important parts of a machine translation system. The NLU component includes a statistical classer as well as a semantic parser [5], [6], both are decision-tree based. The "classer" is used to replace specific categories of phrases in a sentence that have high word variability with a single token identifying the phrase class. Typical classes in the air travel domain include locations, dates, and times. Following classing, the separate "parser" is used to determine the meaning and structure of the classed sentence by assigning a specific hierarchical tree structure to the sentence as predicted by a statistical model. The classer and parser are trained from hand-annotated sentences from the source language.

For the translation from English to Mandarin, a classer and a parser for English are trained using over 100,000 well-annotated sentences in the air travel domain. This allows our system to train a high accuracy classer that achieves a 96% recall and 95% precision, and an accurate parser that achieves a 87% recall and 88% precision. However, for the translation from Mandarin to English, only 2000 Chinese sentences have been collected and manually annotated for the training of both classer and parser. 1800 sentences were used for training the statistical models and another 200 sentences were used for smoothing the resulting decision tree models. The parser and classer performances are yet to be improved. Figure 2 shows a typical parsing example for a Chinese sentence, where the words "QUERY" or "SEGMENT" denote sentence or phrase type, and words that start with "%" such as "%loc-fr" or "%loc-to" represent specific annotation attributes.

2.2 Information Extraction

Information extraction is the task of analyzing the classer and parser outputs to extract appropriate information required in subsequent processing. Two associative arrays are used to store both the classer and parser results. The first array maps the class tags to the phrases in the original sentence for which they were substituted, such as locations and times. The second array maps the concepts contained in the semantic tree to the specific values that appeared in



{QUERY 告诉_query {SEGMENT 所有_hist-rlx-all {LOC-FR LOC_loc-fr LOC-FR} {LOC-TO LOC_loc-to LOC-TO} 航班_flights SEGMENT} QUERY}

Figure 2: Example for Chinese sentence - The Semantic Tree Representation

the sentence, including speech actions, for example. The semantic concept representation, along with the class constituents, captures the information contained in the parsed sentence. This representation allows direct phrase translation to be applied to the class constituents, but a more general conceptual translation should occur at the second level. Thus this design should allow for the rearrangements that occur when concepts are presented in different order in different languages.

2.3 Attribute Translation

Only the named attributes at the leaf nodes in the semantic tree need to be translated in the traditional sense of word-for-word translation. This is currently performed using language to language dictionaries. In cases where phrases or words may be ambiguous, a semantic tag specific phrase translation dictionary is created. That is, a phrase or word may have different translations when considered generally, but usually not within a specific semantic context. Our sets of Chinese-to-English and English-to-Chinese tag-dependent dictionaries were created semi-automatically from general Chinese-to-English and Chinese-to-English dictionaries which were subsequently audited to remove erroneous translations.

2.4 Automatic Speech Recognition and Synthesis

The English and Mandarin speech recognizers were developed for large vocabulary (over 64K words) continuous dictation speech. The baseline speaker-independent systems are trained on over 200 hours of speech collected from over 2000 speakers for each language [7]. The English recognizer uses an alphabet of 52 phones, while the Mandarin system uses 162 phones, including some tone-dependent phones [8]. Each phone is modeled with a 3-state left-to-right HMM. Both systems have approximately 3000 context-dependent states modeled using 40K Gaussian distributions. The acoustic front-end uses a 24-dimension cepstral feature vector and transformed using LDA.

To improve the recognition accuracy, a LM trained for the DARPA Communicator [9] air travel task was substituted for the general English LM. Unfortunately, for Mandarin system, we do not have a similar domain specific LM; therefore the general dictation Chinese LM is used.

One distinguished feature our system has is that our speech recognizer is coupled with the NLU analyzer, rather than separated as in other systems [2, 10, 1]. The NLU classer and parser results are sent back to the recognizer in order to activate a dialog-state-dependent LM [9], or a turn-based LM [11], to futher improve the recognition accuracy. This is motivated by the fact that task ori-

ented conversations often have clear conversation structures which can help to limit the search space for speech recognition. We are also exploring the use of semantic classer and parser information for a better LM that utilizes the language structure.

For speech synthesis, a trainable, phrase-splicing nad variable substitution system [12] is used to synthssize Mandarin or English speech from the translated Chinese or English sentence. Because of the fact that sometimes the translation output is in the form of mixed input and output languages, for example, when there are untranslatable names of locations, our text-to-speech system has the unusual ability to generate speech across language boundaries seamlessly.

2.5 Statistical Natural Language Generation

The high level semantic translation is accomplished by natural language generation of the semantic representation in the target language. More specifically, statistical NLG is used to discover the preferred concept ordering and to assign the lexical form of a grammatical sentence in the target language. Our statistical NLG models [13] are directly learned from a training corpus, using no manually designed grammars or knowledge base.

This work uses a maximum entropy [14] probability model extended from the "NLG2" model described in [13], [15]. It uses a conditional distribution over $V \cup$ *STOP* for the current symbol to be generated, where V is the vocabulary of all possible symbols and *STOP* is an artificial symbol to mark the end of an independent generation. In this context, *symbols* refer to the introduction of semantic concepts or individual target words into the output word sequence. Output begins with a particular sentence type, as identified in the parsed input. Examples of the sentence type in the air travel domain include *Query*, *Action*, *Book* and *Define*. By including the sentence or phrase type in our generation scheme, we can narrow the probability space. The details of the NLG model training and search are in [15].

3. EVALUATION

Defining a useful performance metric for a language to language translation system is a challenging problem in itself. One such metric, *Bleu* [16], recently proposed by our colleagues for use in text-to-text machine translation was used for evaluation. The *Bleu* metric requires human translated scripts as reference, and for our tests, the modified unigram precision (MUP) measure was used. However, the domain specific translation task with spoken word input has many characteristics that present particular challenges. One issue relates to the problem of spontaneous speech recognition, errors due to disfluencies are considered beyond the scope of this paper.

Because the algorithm presented here is trained on a domain specific corpus, the system is more likely to insert domain-specific words such as words pertaining to flights and booking (in our example) even in the case where these words were not spoken explicitly. For this reason, we have also compared the performance against the online *Babel Fish* system [17]. More complete evaluation requires human judges and, for illustration, we include two examples of translations in Fig 3.

4. CONCLUSIONS

The new statistical translation model presented in this paper shows great promise due to its ability to build semantic representations of spoken phrases and to allow general transformations on these representations when creating statements in the target language. This ability has been demonstrated by choosing the highly dissimilar language pair, English and Chinese. In addition, because the presented translation system is based on statistical models trained using corpora, developing translations systems for new languages should require only a fraction of the effort normally invested in building language to language translation systems.

Source	Text	MUP
Chinese	星期六从波士顿飞往匹兹堡的航班都有哪些	
Human 1	Which flights are available from Boston to Pittsburgh on Saturday?	
Human 2	What are the flights from Boston to Pittsburgh this Saturday?	
Babel Fish	Which Saturday flies to Pittsburgh from Boston the scheduled flight all has?	58%
New System	Which ones flight from Boston to Pittsburgh on Saturday?	83%
Chinese	我想了解星期五从费城去加州奥克兰的航班信息	
Human 1	Show me all the flights from Philadelphia to Oakland, California on Friday.	
Human 2	Give me a list of flights from Philadelphia to Oakland, California on this Friday.	
Babel Fish	I want to understand Friday goes to the the California Oakland from the Philadelphia the scheduled flight information.	47%
New System	I want to know flights from Philadelphia to Oakland California on Friday.	67%

Figure 3: Example Translations

5. ACKNOWLEDGEMENT

Thanks to Drs J. Hansen and W. Ward at CSLR, Univ. of Colorado for their valuable comments and discussions. Thanks to Drs. A. Ratnaparkhi, W. Zhu and X. Luo from IBM Research for their help with this work.

6. REFERENCES

[1] Federal Ministry of Education, Science, and Technology, "http://verbmobil.dfki.de," website.

[2] Carnegie Mellon University, "http://www.is.cs.cmu.edu/js/janus.html," website.

[3] F. Casscuberta et al, "Speech-to-speech translation based on finite-state transducers," in *ICASSP2001*, Salt Lake City, Utah, May 2001.

[4] L.M. Tomokiyo, M. Gavaldà, W. Ward, and A. Waibel, "Concept based speech translation," in *ICASSP*, May 1995.

[5] D. Magerman, *Natural language parsing as statistical pattern recognition*, Ph d desertation, Stanford University, 1994.

[6] K. Davies et al, "The ibm conversational telephony system for financial applications," in *Eurospeech1999*, Budapest, Hangary, Sept 1999, vol. 1, pp. 275–278.

[7] Y. Gao et al, "New adaptation techniques for large vocabulary continuos speech recognition," in *ASR workshop 2000*, Paris, France, September 2000, pp. 107–111.

[8] J. Chen et al, "New methods in continuous mandarin speech recognition," in *EuroSpeech1997*, Rhodes, Greece, September 1997, pp. 1543–1546.

[9] Y. Gao et al, "Recent advances in speech system for ibm darpa communicator," in *EuroSpeech2001*, Aalborg, Denmark, September 2001, pp. 503–506.

[10] Carnegie Mellon University, "http://www.c-star.org," website.

[11] R. Sarikaya et al, "Turn-based language modeling for spoken dialog systems," in *submitted to ICASSP2002*.

[12] R. Donovan et al, "Phrase splicing and variable substitution using the ibm trainable speech synthesis system," in *ICASSP1999*, Phoenix, Arizona, March 1999, pp. 373–376.

[13] A. Ratnaparkhi, "Trainable methods for surface natural language generation," in *Proc. of the 1st Meeting of the North American Chapter of the Association of Computational Liguistics*, Seattle, WA, 2000, pp. 194–201.

[14] A. Berger, S. A. Della Pietra, and V. J. Della Pietra, "A maximum entropy approach to natural language processing," *Computational Linguistics*, vol. 22, no. 1, pp. 39–71, 1996.

[15] B. Zhou et al, "Statistical natural language generation for speech-to-speech machine translation," in *ICSLP2002*, Denver, Colorado, Sept. 2002, pp. 1897–1900.

[16] K. Papineni, "Blue: A method for automatic evaluation of machine translation," Research Report RC22176, IBM, Sept 2001.

[17] Systran Technologies, "http://world.altavista.com," website.