Statistical Natural Language Generation for Speech-to-Speech Machine Translation

Bowen Zhou, Yuqing Gao, Jeffrey Sorensen, Zijian Diao, Michael Picheny

Robust Speech Processing Group
Center for Spoken Language Research
University of Colorado Boulder, Campus Box 594
(Bus Mail: 3215 Marine Street, Room E-265)
Boulder, Colorado  80309-0594
303-735-5148 (Phone) 303-735-5072 (Fax)
http://cslr.colorado.edu/
John.Hansen@colorado.edu. (email)

ICSLP-2002: Inter. Conf. on Spoken Language Processing
INTERSPEECH 2002
Denver, Colorado USA
September 16-20, 2002
ABSTRACT

This paper presents a statistical natural language generation scheme for trainable speech-to-speech machine translation (MT) systems for limited domain applications using a cascaded approach. The natural language generation scheme in the translation systems is based on a maximum entropy (ME) statistical model fully trained from a corpus, allowing flexible translation outputs. In this paper, the system architecture and some of its components, including the parsing, information extraction, and translation etc are briefly overviewed, followed by the descriptions of training and search algorithms for ME based sentence level NLG within the MT context. Details of NLG including feature selection and robustness are also addressed. We have implemented the described system for translating between Chinese speech and English speech in an air travel application domain. Encouraging experimental results have been observed and are presented.

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. There have been numerous efforts to create such a device in recent years.

Many technological frameworks have been proposed for the task of speech translation, ranging from a cascaded approach [1] to finite state transducers [2]. Recently, we presented a speech translation system [3] employing a statistical framework appropriate for use in language restricted domains. In a 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. Other cascaded speech translation systems have been proposed in the past few years. However, most of the generation components in such systems are based on fixed templates, which produce translated sentences lacking in variability. Template based systems are difficult to maintain, scale, and lack robustness as application domains change, often requiring complete redesign.

In our system, the generation of target sentences is based on a maximum entropy (ME) statistical model [4]. ME-based generation is fully trainable from a corpus. This modeling process is largely domain independent, speeding the development of systems for different application tasks. In addition, the translation is more flexible, allowing slight modifications in source language to be reflected with grammatical changes in the target language. To evaluate this ability, two very dissimilar languages (Chinese and English) were selected as the source and the target languages in an air travel task.

This paper is organized as follows: Sec. 2 describes the translation system architecture and the building of each component of the system; Sec. 3 provides a detailed description of the Maximum Entropy based sentence level NLG from the interlingua representations within the MT context. Sec. 4 presents the evaluation results and discussions and Sec. 5 contains concluding remarks.

2. SYSTEM OVERVIEW

Fig. 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 an natural language generation (NLG) engine to render in the target language. Both types of information are translated using distinct models, with the 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 complete, 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 must be invariant across translations, at least in the application domains of interest. This invariance has been validated, in part, by work involving translation between English and certain European languages [1]. 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 interlingual style. In the following parts of this section, we briefly discuss some system components. More system information and improvements in speech recognition can be found in [3]. For simplifying illustrations, we treat the Chinese as the source language and the English as the target language in the remainder of this paper, though the system is bi-directional.

2.1. Building the NLU engine

The natural language understanding component is one of the most important parts of a machine translation system. The NLU engine used in our system is based on IBM’s ViaVoice Telephony Toolkit [5], which includes a statistical, decision-tree based recognizer that is used to identify instances of specific semantic classes as well as general semantic parsing. A “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 classeeing, a separate “parser” is used to determine the meaning and structure of the classified 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. Fig. 2 shows a typical parsing example for a Chinese sentences, where the words “QUERY” or “SEGMENT” denote sentence or phrase type, and words start that with “%” like “%loc-fr” or “%loc-to” represent specific 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 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 to occur at the second level. Thus this design should allow for the rearrangements that occur when concepts are represented in differently 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.

3. 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. The statistical models are directly learned from a training corpus, using no manually designed grammars or knowledge base. In our speech translation system, the statistical NLG component has three kinds of inputs: a set of tree-annotated language independent (interlingual) concepts as demonstrated in Fig. 2, a set of unordered translated words in the target language, and probability models for word generation.

During translation, the source sentence is parsed, yielding the constituent structure of semantic tree which is kept, while the concept ordering information is discarded. The word generation model is a maximum likelihood prediction based on maximum entropy modeling [6, 4].

3.1. Probability Model

This work uses a maximum entropy probability model extended from the “NLG2” model described in [6]. It describes 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 the context of natural language generation in this paper, 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 in the air travel domain include Query, Action, Book and Define etc. By including the sentence or phrase type in our generation scheme, we can narrow the probability space.
The sentence generation is further conditioned based on local \(n\)-grams and the set of concepts included in the sentence type that have not yet been included in the sentence being generated, using

\[
P(s_i | T_e, C_i, s_{i-2}, s_{i-1}) = \frac{\prod_{j=1}^{K} f_j \left( s_i, s_{i-1}, s_{i-2}, T_e, C_i \right)}{\sum_{j' \in V} \prod_{j=1}^{K} f_{j'} \left( s_i, s_{i-1}, s_{i-2}, T_e, C_i \right)}
\]

where \( \{s_{i-1}, s_{i-2}\} \) are the previous symbols in the generated sequence, and \( s_i \) is the current symbol, \( T_e \) is the local sentence or phrase type in corresponding portion of the semantic tree, and \( C_i \) is the concept list that remains to be generated before symbol \( i \). \( f_i \) is a binary feature that captures the co-occurrence evidence of current symbol and its contexts:

\[
f_i \left( s_i, s_{i-1}, s_{i-2}, T_e, C_i \right) = \begin{cases} 
1 & \text{if } s_i \text{ appears with } s_{i-1}, s_{i-2}, T_e, C_i \\
0 & \text{otherwise} 
\end{cases}
\]

the feature weight \( \alpha_i \) captures the influence of each feature.

### 3.2. Model Training

The training corpus used in our system is the corpus collected during IBM’s DARPA Communicator project, a dialog system that provides information on air travel. This corpus has been manually annotated with both classer or parser tags used in training the NLU models. To reuse this corpus for generation, we employ a model training procedure that is similar to the one described in [6]. Through a left-to-right breadth-first search, the symbol sequences that end with \(*STOP*\) and mention all concepts exactly once are obtained. More specifically, at each time \( n \) for each active symbol \( s_n \), we score its symbol path \( s_1s_2\ldots s_{n-1} \) with

\[
P(s_1s_2\ldots s_n | \text{tree}) = \prod_{i=1}^{n} P(s_i | T_e, C_i, s_{1-i}, s_{i-2})
\]

As with \(N\)-best search in Viterbi decoding, only the top \( N \) best scored symbols are kept active at each time, with new symbols expanded from them. The use of \(N\)-best search adds additional linguistic constraints. The use of a “symbolic” trigram assumes that words (symbols) generated from different parents are independent. As a result, the search may generate following pseudo sentence: “...book a [SEGMENT flights from Boston to Denver SEGMENT].” To address the mismatch “a flights” we may re-score the \(N\)-best generated sentences with additional linguistic knowledge. One such simple re-score can be achieved by using pure word \(n\)-grams to re-rank the generated \(N\)-best sentences after attribute substitution. However, more complex processing, like re-ranking the sentences using the parsing score through a target language parser may also be used.

### 3.3. Generation Search

A recursive search based on the semantic tree structure of the input sentence is used to generate the word sequence in the target language. The generation procedure consists of the following steps:

1. Traverse the semantic tree in a bottom-up fashion;
2. For each non-terminal, search for the qualified symbol sequence according to the \textit{macro} model;
3. Go to 2 unless all non-terminal nodes are traversed;
4. Link phrases to a sentence by another traversal;
5. Apply the \textit{micro} model to the terminal concepts to predict inclusion of one or more sub-concepts, if necessary;
6. Substitute concepts with their variables.

Each single round of search in step 2 of this procedure is similar to the one described in [6]. Through a left-to-right breadth-first search, the symbol sequences that end with \(*STOP*\) and mention all concepts exactly once are obtained. More specifically, at each time \( n \) for each active symbol \( s_n \), we score its symbol path \( s_1s_2\ldots s_{n-1} \) with

\[
P(s_1s_2\ldots s_n | \text{tree}) = \prod_{i=1}^{n} P(s_i | T_e, C_i, s_{1-i}, s_{i-2})
\]

### 3.4. Robustness to OOT Questions

As with many other applications of statistical models, we may encounter the out-of-training (OOT) questions in searching stage due to the mismatch between training and decoding. OOT happens whenever the tuple \( \{T_e, C_i, s_{1-i}, s_{i-2}\} \) has never been observed in the training data, which usually results from the mismatch from \( \{T_e, C_i\} \). If some remaining concepts mismatch with \( T_e \), the generation result may lose some information; if all remaining concepts mismatch with current \( T_e \), then no phrase could be generated! OOT can occur for the following reasons: inconsistent annotations used in source and target language training data, incorrect parser results, radical source/target language differences and scarce training data.
we take following steps whenever we encounter OOT questions:

1. If there is a single concept remaining, remove its parent (since which may indicate that the parser result is wrong!) and promote this concept to higher level;
2. If a concept in a multi-concept list is OOT, promote this OOT concept to higher level, and generate local phrase after removing the OOT concept;
3. If a concept reaches the root, a local sentence is generated by ignoring the OOT concepts, and another parallel sentence is generated using the OOT concepts;

This constitutes a method for generating sentences under most conditions, but does not eliminate the need for a reasonably large training corpus in both the source and target language.

4. EVALUATION

Defining a useful performance metric for a language to language translation system is a challenging problem in itself. One such metric, Blue [7], recently proposed by our colleagues for use in text to text machine translation was used for evaluation. The Blue 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, error 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 [8]. More complete evaluation requires human judges and, for illustration, we include two examples of translations in Fig. 3. With the similar setting, we also evaluate the performance speech translation from English to Chinese over a test corpus of 30 English spoken sentences. The average ASR Word Error Rate (WER) over the test data is 10.5%. For a better comparison, both automatic and human transcripts are used as the translation input. Similarly, two human translators provide the reference translations for evaluating the Blue metric. Fig. 4 shows the average performance compared with BabelFish.

5. CONCLUSIONS

The statistical language generation and 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 languages English and Chinese. In addition, because the presented generation and translation systems are based on statistical models trained using corpora, developing translation systems for new languages should require only a fraction of the effort normally invested in building language to language translation systems.

6. REFERENCES