# USE OF STATISTICAL N-GRAM MODELS IN NATURAL LANGUAGE GENERATION FOR MACHINE TRANSLATION

*Fu-Hua Liu, Liang Gu, Yuqing Gao and Michael Picheny*

IBM T.J. Watson Research Center
Yorktown Heights, NY 10598, U.S.A.
fhl@us.ibm.com, lianggu@us.ibm.com, yuqing@us.ibm.com, picheny@us.ibm.com

## ABSTRACT

Various language modeling issues in a speech-to-speech translation system are described in this paper. First, the language models for the speech recognizer need to be adapted to the specific domain to improve the recognition performance for in-domain utterances, while keeping the domain coverage as broad as possible. Second, when a maximum entropy based statistical natural language generation model is used to generate target language sentence as the translation output, serious inflection and synonym issues arise, because the compromised solution is used in semantic representation to avoid data sparseness problem. We use N-gram models as a post-processing step to enhance the generation performance. When an interpolated language model is applied to a Chinese-to-English translation task, the translation performance, measured by an objective metric of BLEU, improves substantially to 0.514 from 0.318 when we use the correct transcription as input. Similarly, the BLEU score is improved to 0.300 from 0.194 for the same task when the input is speech data.

## 1. INTRODUCTION

The need to develop technologies to accomplish useful and satisfactory translation between languages is increasingly appreciated with rapid growth of internet applications and globalization of economy development. Many approaches, statistical and/or rule-based, have been proposed and experimented to overcome technical barriers [e.g. 4,5,7,10]. The task becomes even more challenging when the source input switches from written text to spoken speech. Nevertheless, most systems are designed to work with the speech translation in some restricted domains, such as air travel information, meeting scheduling, and financial transaction [e.g. 1,4,5].

Recently, we presented a speech translation system employing a statistical framework in an air travel information domain [1]. The relevant semantic information was extracted from the source sentence using a statistical parser. The extracted information was then passed to the natural language generation process to yield the translated target sentence. Because the phrase type determined the basic output sentence structure, word inflection was not relevant. However, in general it is not the case when another translation application is concerned. The degree of word inflection differs from language to language. For a very dissimilar language pair such as English and Chinese, this discrepancy is palpable and needs to be addressed to ensure a good quality of translation. To this end, a statistical N-gram modeling approach is proposed to tackle this issue.

This paper is organized as follows, section 2 presents a brief overview of IBM's speech-to-speech translation system, MASTOR. Section 3 describes the natural language generation process, the challenges needed to be addressed and our statistical approach to tackle these problems. Then, details of system setup, experiments and results will be given in section 4. Finally, a conclusion and summary will be presented in Section5.

## 2. OVERVIEW OF SYSTEM

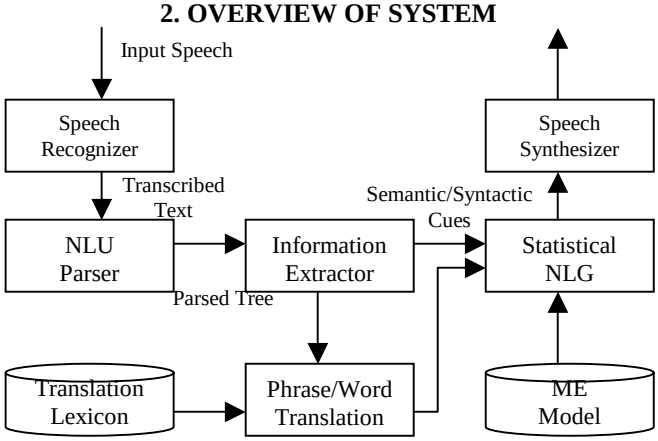


**Figure 1:** The architecture of MASTOR

MASTOR (Multilingual Automatic Speech-To-Speech TranslatOR) is IBM's highly trainable speech-to-speech translation system, targeting conversational spoken language translation between English and Mandarin Chinese for limited domains. Figure 1 depicts the architecture of MASTOR. The speech input is processed and decoded by a large-vocabulary speech recognition system. Then the transcribed text is analyzed by a statistical parser [9] for semantic and syntactic features. A sentence-level natural language generator based on maximum entropy (ME) modeling [3,8] is used to generate sentences in the target language from the parser output. The produced

sentence in target language is synthesized into speech by a high quality text-to-speech system [6].

### 2.1. NLU Parser

The NLU module includes a statistical, decision-tree based parser. This parser utilizes statistical models, originally developed for natural language understanding applications [9]. It does not rely on any handcrafted grammars or rules. The semantic and syntactic information is denoted in a tree-structured semantic/syntactic representation, which is somewhat comparable to interlingua [4]. It assigns each word at least two tags for semantic and syntactic cues. Figure 2 shows a parse tree for a sentence, "Next bridge is five miles away". Capitalized words such as "PLACE" and "MEASURE" denote sentence or phrase type, and words starting with "%" such as "%place" "%length" convey semantic sense. Typical domain-specific classes include "DIRECTION", "PLACE", and "%place" while general semantic classes include "BE", "%num", and "%pron". The parser is trained from pre-annotated sentences, similar to other tree-bank corpora, in the source language [11].

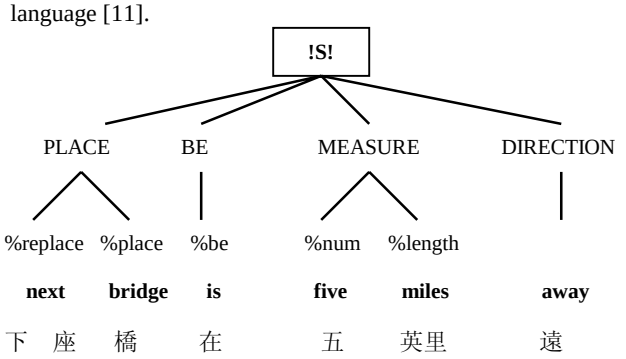


**Figure 2:** A parsed example of an English sentence

### 2.2. Information Extractor

The purpose of information extractor is twofold. First, it analyzes the output from parser and identifies the candidates for phrase or larger semantic constituents. Only words of the same semantic class are candidates to be evaluated for phrase. Second, it forges semantic and syntactic information in a proper format for NLG.

### 2.3. Lexicon Translation

As shown in Figure 2, all words in the input sentence are at the leave nodes in the semantic tree. A word at the leave node is to be translated into the translated token according its corresponding semantic sense tag. A phrase translation is also considered for each word along with its neighboring words in the same semantic context.

### 2.4. Statistical Natural Language Generation

A maximum entropy probability model extended from the "NLG2" model proposed by Ratnaparkhi [8] is used in the NLG system. Both the sentence level and concept level classes are used as constituents. The features used in the ME modeling include the previous symbols, local sentence or phrase type in the semantic tree, and the concept list that remains to be generated before current symbol. During the translation, a recursive search is performed on the parsed tree of the input sentence in a bottom-up manner to generate the output word sequence in the target language. After each non-terminal node is traversed, the resultant symbol string is appended to the output. At the end of search, the concepts are substituted with their variables.

### 2.5. Speech Synthesis

The output from NLG is input to a trainable, phrase-slicing and variable substitution speech synthesis subsystem [6] to synthesize target language speech. Our text-to-speech has the ability to generate speech across different languages. Currently, MASTOR has one male and one female voices for both English and Mandarin Chinese TTS systems.

## 3. ISSUES IN NATURAL LANGUAGE GENERATION

After the ME model produces the most likely order sequence for concept constituents, choices of word are to be decided for each concept. The issue of word inflection arises.

### 3.1. Word Inflection

The degree of word inflection varies with language. In our Chinese-to-English translation, the word inflection is quite palpable. The input Chinese words are relaxed in inflection compared with its counterparts in English. For example, given the desired word order and word sense, the following two illustrations manifest the inflection issue by comparing Chinese words and the corresponding English words.

- A Chinese sentence, "我/他 **看見** 一個/二個 **學生**", and its English translation, "I/he **see/sees** a/two **student/students**"
  - the Chinese verb "**看見**" in "我 **看見**"(I **see**) and "他 **看見**"(he **sees**) corresponding to two English words, "see" and "sees"
  - the Chinese noun "**學生**" in "一個 **學生**"(one **student**), and "二個 **學生**"(two **students**) corresponding to two English words, "student" and "students"
- A Chinese sentence, "你 ''/想要/喜歡 **說話**", and its English translation, "you ''/like/enjoy **talk/to talk/talking**"
  - the Chinese verb "**說話**" in "你 **說話**" (you **talk**) , "你 想要 **說話**" (you like **to talk**), and "你 喜歡 **說話**"(you enjoy **talking**) corresponding to three English words, "talk" , "to talk", and "talking"

For a Chinese to English translation, the typical choices for a verb can have up to 6 different forms and those for a noun can have possible 2 variations, let alone the choice for synonym words and the possible associate propositions.

### 3.2. Lexicon Design and Translation

In our Chinese to English translation lexicon, a list of translation output is provided for each semantic constituent to account for inflection as well as synonyms, if any. Table 1 illustrates some entries from the lexicon. It is worth noting that the lexicon chooses output depending on the semantic tag produced by the parser for word sense. The output also contains word list accounting for possible inflectional forms.

| Chinese | tag | English |
|---|---|---|
| 說 | **speak** | speak/speaks/spoke/spoken/speaking/to speak |
| 說 | **verb** | say/says/said/saying/to say |
| 語言 | **lang** | language/languages |
| 你的 | **pron-poss** | your/yours |
| 表 | **doc** | form/forms |
| 表 | **tool** | watch/watches |

**Table 1**: semantic-based lexicon with word inflection

## 4. LANGUAGE MODELS FOR S2S SYSTEM

Language models have been used successfully in the speech recognition to improve performance for many years. In particular, simplicity and efficiency of N-gram models make them a favorite choice for large-vocabulary speech recognition applications. The N-gram probability can be expressed as

$$P(w_i \mid w_{i-1},....,w_{i-N+1}) = \frac{C(w_i, w_{i-1},....,w_{i-N+1})}{C(w_{i-1},...,w_{i-N+1})} \qquad (1)$$

where $C(w_i, w_{i-1},\ldots,w_{i-N+1})$ is the number of counts for the word sequence string $w_i, w_{i-1},\ldots,w_{i-N+1}$. To solve the problem of data sparseness in training process, various smoothing techniques are proposed for unseen word string pattern. Backing off to the (N-1)-gram is a common and effective method.

### 4.1. LM's in Speech Recognition

It is expected that the system performance of a speech recognizer will improve in a specific application using a domain-specific LM for that domain. In the case where only a small amount of domain-specific training data is available, the technique of interpolating the domain-specific LM with a domain-independent LM is commonly used for better performance. Since MASTOR is designed and developed for the DARPA Babylon project with very limited data, how to design a LM for this specific domain is crucial to the speech recognition module, and therefore, the overall system performance.

### 4.2. LM's in Natural Language Generation

As described in Section 3, when the variable substitution is to be performed after ME modeling in NLG, a lexicon is consulted to replace the concept constituents. There may exist multiple forms for the substitution word. In theory, the word inflection could be resolved by ME modeling if every word sense is regarded as a distinct concept element. This requires the availability of an enormous amount of data to train both NLU parser and NLG ME model, which is impractical for our application. In this paper, we propose the use of statistical language models as a solution to the word inflection problem. Trigram language models are used in our system for two reasons. First, a powerful domain-independent trigram model designed for IBM ViaVoice recognition system can be used directly. Second, in the case of domain-specific applications, adapted N-gram models can be derived handily from existing trigram models by using available domain-specific text data. Language models are used to re-score all inflection forms in a post-processing manner and generate the best hypothesis as the generation output.

## 5. EXPERIMENTS AND RESULTS

**Experiment Setup**. To evaluate the effectiveness of statistical N-gram models for NLG, the following experiments are conducted in a DARPA force protection domain. The source language for translation is Mandarin Chinese and the target language is English. The sentences from the DARPA force protection domain are relatively conversational, interactive, and relaxed in sentence syntax compared with dictation applications.

The translation component is evaluated separately when the correct transcription and the recognized transcription are used as input text. For simplicity and fast turnaround time, the translation performance is evaluated by using an objective measure, BLEU[2]. BLEU measures the translation quality based on N-gram and brevity between hypothesis and reference. The BLEU score is in the range of 0 and 1, where 1 represents a perfect matched translation and 0 means an entirely mismatched translation.

### 5.1. Language Models in Speech Recognition

With the limited domain-specific data, we are to evaluate the quality of different language model configuration. First, we train a domain-specific LM using the text data from the application domain. Because the in-domain data is quite limited with about 33 thousand words, the data sparseness issue will be addressed by interpolating with a general-domain LM, developed for a dictation task. Equation (2) expresses the probability for the interpolated LM, where $LM_d$ is the domain-specific LM, $LM_g$ is the general domain dictation LM, and $\lambda$ is the interpolation weights for domain-specific LM.

$$P(w \mid LM) = \lambda P(w \mid LM_d) + (1-\lambda) P(w \mid LM_g) \qquad (2)$$

Table 2 lists the speech recognition results using various interpolated LM's with different interpolation weights in a speaker-independent mode evaluation on a test corpus consisting of 295 test utterances from two speakers, one male and one female, respectively.

| Interpolation weight for domain-specific LM | Word Error Rate (%) |
|---|---|
| 0.0 | 53.19 |
| 0.2 | 34.56 |
| 0.4 | 32.20 |
| 0.6 | 32.44 |
| 0.7 | 29.09 |
| 0.8 | 28.45 |
| 1.0 | 27.31 |

**Table 2:** Speech recognition error rates for interpolated LM's where $\lambda$ is the interpolation weight for domain-specific LM

Not surprisingly, the recognition error rate decreases with more weight on the in-domain LM. The general domain LM has an error rate of 53.19% while the domain-specific LM has 27.84%. We also observe that the improvement appears to saturate after $\lambda$ exceeds 0.7. To account for possible out-of-domain sentences, we choose an interpolation weight of 0.7 for the interpolated LM to be used in the speech recognition module. It

is worth noting some errors are simply due to word boundary inconsistency, a well-known text segmentation issue for some languages such as Chinese. The error rate is reduced to 24.73% after the recognition output is re-segmented properly.

**5.2. N-Gram Models With Speech Input**

First, we use the output from the speech recognizer, obtained from Section 4.1, as the input to machine translation. It provides a fair evaluation on speech-to-speech translation. Without any N-gram model, the baseline result is obtained using the baseline setup where the phrase/word translation module selects the first entry, i.e., the most frequently used form from the dictionary, as the translation output. There are three references for each sentence for BLEU generated by two human translators.

Two N-gram models for the target language are evaluated, one being the general-domain LM (LM1), and the other the domain-specific LM (LM2) used in [11]. In Table 3, the BLEU score is 0.267 for the general-domain LM and 0.300 for the interpolated LM, substantially better than the baseline's 0.194.

| Configuration | BLEU |
|---|---|
| B (Baseline) | 0.194 |
| B+LM1 | 0.267 |
| B+LM2 | 0.300 |

**Table 3:** BLEU scores, when the input is from speech recognizer, for baseline, NLG with LM1, and NLG with LM2

**5.3. N-Gram Models With Text Input**

Upon examining the results, we find that one speaker has a poor speech recognition result with an error rate of 38.2% while the other speak has a better error rate of 11.26%. This uneven performance and non-trivial recognition errors make it difficult to see how the N-gram models alone perform in NLG. Therefore, we need to study the N-gram models for the translation task using the "correct" text input instead.

Table 4 lists the results obtained using the text input. The BLEU score is 0.467 for the general domain N-gram model (LM1) and 0.477 for interpolated N-gram model (LM2), compared with 0.311 from the baseline. As we expect, many inflection-related errors have been corrected. For example, the baseline output for the input "有 三 輛 箱型車" is "there are three van". And the N-gram models produce "there are three vans".

| Configuration | BLEU | BLEU0 |
|---|---|---|
| B (Baseline) | 0.311 | 0.192 |
| B+LM1 | 0.467 | 0.271 |
| B+LM2 | 0.477 | 0.301 |

**Table 4:** BLEU scores for text input using three references (denoted as BLUE) and single reference (denoted as BLEU0)

The last column (BLEU0) of Table 4 compares the same translation hypothesis output using only one reference in calculating BLEU. It is clear that the effectiveness of N-gram models still preserves even when the number of available references changes the absolute BLEU scores.

**5.5. Feature Selection For Maximum Entropy**

One of the important design considerations in ME modeling is the selection of input features. As described in Section 2.4, the ME model uses the features $f_i(s_{i-1}, s_{i-2}, T_i, c_i, C_i)$ where $\{s_{i-1}, s_{i-2}\}$ are the previous symbols in the output, $T_i$ is the local sentence type in the semantic tree, and $c_i$ is a concept in the remaining concept list, $C_i$. In this section, a preliminary experiment is conducted using a new feature $f_i(s_{i-1}, s_{i-2}, T_i, c_i, c_i', C_i)$ in the ME modeling by adding an additional sibling symbol from the remaining concept list, $C_i$, where $\{c_i, c_i'\}$ are the sibling symbols in $C_i$. Table 5 shows that the best result, 0.514, is obtained using an interpolated LM and the sibling-included feature.

| Configuration | BLEU |
|---|---|
| B2 (Baseline) | 0.318 |
| B2+LM1 | 0.460 |
| B2+LM2 | 0.514 |

**Table 5:** Results of N-gram models on systems derived from ME using sibling symbols (Baseline: B2)

## 6. SUMMARY

Severe translation performance degradation is observed due to word inflection when a speech-to-speech system is used for Chinese-to-English translation. In NLG, the selected word is chosen mainly based on semantic information while semantic and syntactic cues facilitate the word re-ordering functionality. The statistical N-gram models are proposed to address the issue of word inflection for better grammatical agreement. In the context of a DARPA force protection domain, this approach improves the BLEU score from 0.192 to 0.301 when a single reference is used and from 0.318 to 0.514 when sibling symbols are included and multiple references are used

## 7. REFERENCES


[1] B. Zhou, et al, "Statistical Natural Language Generation for Speech-to-Speech Machine Translation Systems", *ICSLP-2002*, pp. 1897-1900, Sept. 2002.

[2] K. Papineni, et al, "Bleu: A Method for Automatic Evaluation of Machine Translation", *Research Report RC22176*, IBM, Sept. 2001.

[3] A. Berger, et al, "A Maximum Entropy Approach to Natural Language Processing", *Computer Linguistics*, Vol. 22, No. 1, pp. 39-71, 1996.

[4] A. Lavie, et al, "Janus-III: Speech-to-Speech Translation in Multiple Languages", *Proceedings of ICASSP-97*, 1997.

[5] W. Wahlster, ed., *Verbmobile: Foundation of Speech-to-Speech Translation*, Springer, 2000.

[6] R. Donovan, et al, "Phrase Splicing and Variable Substitution Using the IBM Trainable Speech Synthesis System", *ICASSP-1999*, pp. 373-376, 1999.

[7] H. Ney, et al, "Algorithms for Statistical Translation of Spoken Language", *IEEE Trans. on Speech And Audio Processing*, vol.8, no. 1, January 2002.

[8] A. Ratnaparkhi, "Trainable Methods for Surface Natural Language Generation", *NAACL-2000*, April 2000.

[9] K. Davies, et al, "The IBM Conversational Telephony System for Financial Applications", *EuroSpeech-1999*, pp. 275-278, 1999.

[10] T. Takezawa, et al, "A Japanese-to-English Speech Translation System: ART-MATRIX", *ICSLP-1998*, pp. 2779-2782, 1998

[11] L. Gu, et al, "IBM S2S Translation System for DARPA Babylon Program", *submitted to ICASSP-2003*, 2002