ISCA Archive
http://www.isca-speech.org/archive

International Symposium on Chinese Spoken
Language Processing (ISCSLP 2004)
Hong Kong
December 15-18, 2004

# USE OF DIRECT MODELING IN NATURAL LANGUAGE GENERATION FOR CHINESE AND ENGLISH TRANSLATION

Fu-Hua Liu and Yuqing Gao

IBM T.J. Watson Research Center, Yorktown Heights
{fhl, yuqing}@us.ibm.com

## ABSTRACT

This paper proposes a new direct-modeling-based approach to improve a maximum entropy based natural language generation (NLG) in the IBM MASTOR system, an interlingua-based speech translation system. Due to the intrinsic disparity between Chinese and English sentences, the previous method employed only linguistic constituents from output language sentences to train the NLG model. The new algorithm exploits a direct-modeling scheme to admit linguistic constituent information from both source and target languages into the training process seamlessly when incorporating a concept padding scheme. When concept sequences from the top level of semantic parse trees are considered, the concept error rate (CER) is significantly reduced to 14.3%, compared to 23.9% in the baseline NLG. Similarly, when concept sequences from all levels of semantic parse trees are tested, the direct-modeling scheme yields a CER of 10.8% compared to 17.8% in the baseline. A sensible improvement on the overall translation is made when the direct-modeling scheme improves the BLEU score from 0.252 to 0.294.

## 1. INTRODUCTION

Automatic speech-to-speech translation can assist human communication using natural spoken languages for people who do not share a common language. To implement a spoken language translation system, there are some unique challenges to be addressed. First input speech is usually casual conversational speech, possibly non-grammatical with disfluencies. The output from conversational speech recognition often contains recognition errors, imperfect syntax, and no punctuations. Emotional expressions embedded in speech also need to be captured, understood, and re-generated in an appropriate manner for output sentences in the target language. While encouraging progress in speech recognition, natural language processing, and machine translation has been made over the past decades, multilingual speech-to-speech translation is still considered a grand challenge for human language technologies [1,2,3,4].

We developed a speech translation system employing a statistical framework in a DARPA force protection domain [5,6] between Chinese and English. This speech translation system consisted of modules to accomplish tasks of speech recognition, machine translation, and text-to-speech synthesis for English and Mandarin Chinese. For machine translation, an interlingua-based framework combining a natural language understanding (NLU) and natural language generation (NLG) was proposed. However, the baseline NLG was limited by an issue of concept inequality in linguistic representation due to language disparity in the semantic parse trees. Therefore, improvement of concept generation remains a crucial task. To this end, this paper presents a direct-modeling-based approach with two novel schemes in an attempt to improve the overall system performance.

This paper is organized as follows; we present in Section 2 a brief overview of IBM's speech-to-speech translation system, MASTOR. Knowledge representation and maximum entropy will be described in Section 2 as well. The new direct-modeling based NLG will be discussed in Section 3. Section 4 presents details of system setup, experiments and results. Finally, a summary will be given in Section5.

## 2. OVERVIEW OF SYSTEM

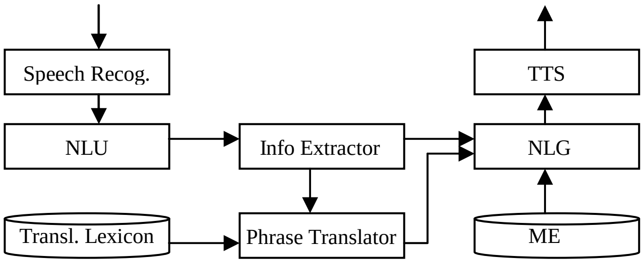

**Figure 1:** The architecture of MASTOR

Figure 1 depicts the architecture of MASTOR (Multilingual Automatic Speech-To-Speech TranslatOR) [5,6]. The speech input is processed and decoded by a large-vocabulary speech recognition system. Then the transcribed text is analyzed by a statistical parser [6,7] for semantic and syntactic features. A sentence-level natural language generator based on maximum entropy (ME) modeling [8] is used to generate a sentence in the target language from the parser output [6,9]. The produced text sentence in target language is synthesized into speech by a text-to-speech system.

### 2.1 Knowledge Representation - Annotation

In MASTOR, we design a tree-structured semantic/syntactic representation, referred to as "concept" in this paper, to express linguistic constituent information in sentences for both source and target languages. This is comparable to interlingua [1]. The design and selection of concepts is a lengthy but important step. The concepts have to be not only broad to cover all intended meanings in input sentences but also informative in order to re-generate output sentences with right word sense in a grammatically correct manner.

The concepts we use in MASTOR fall into two different categories. One category is “tag”. Its primary role is for word sense selection. The second category is “label”. Its primary use is to express phrase types. Coupled with tags, labels represent the intended expressions of a sentence in a specific order. In our current design each word has at least two concepts, one as a tag, the remaining as labels. Figure 2 shows a pair of parse trees for a sentence, “what is your primary language” for Chinese and English. Capitalized words such as “LANG” and “WHQ” denote sentence or phrase type, and words starting with “%” such as “%pron-poss” and “%whq” convey semantic/syntactic sense.

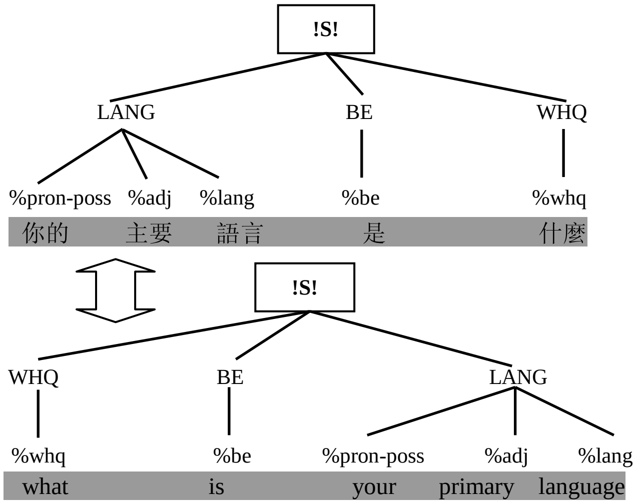

**Figure 2:** A parse tree representation of a sentence pair with unequal concepts between input and output languages

### 2.2 Maximum Entropy (ME) Model

The baseline statistical NLG system uses a maximum entropy probability model extended from the “NLG2” algorithm proposed by Ratnaparkhi [8]. Assume that $\boldsymbol{C} = \{c_1, c_2, c_{3,\ldots,} c_M\}$ is a complete concept sequence of the source language. We further assume that a concept output sequence $\boldsymbol{S_i} = \{s_1, s_2, s_{3,\ldots,} s_i\}$ has already been generated for the target language and that $\boldsymbol{C}_i = \{c_{i1}, c_{i2}, c_{i3,\ldots,} c_{iM}\}$ is a list of remaining concepts to be considered at time $i$.

To generate the next concept, $s_{i+1}$, at time $i+1$, the conditional probability of a concept candidate can be expressed as

$$p\left(s \mid c_{ij}, s_i, s_{i-1}\right) = \frac{\prod_k \alpha_k^{\,g(\bar{f}_k, s, c_{ij}, s_i, s_{i-1})}}{\sum_{s \in C_i} \prod_k \alpha_k^{\,g(\bar{f}_k, s, c_{ij}, s_i, s_{i-1})}} \qquad (1)$$

where $s$ is the concept candidate to be generated, $s_i$, and $s_{i-1}$ are the previous two concepts in $\boldsymbol{S_i}$. $\alpha_k$ is the probability weight corresponding to each feature $\bar{f}_k$ in the ME framework. g is a binary test function defined as

$$g(\bar{f}_k, s, c_{ij}, s_i, s_{i-1}) = \begin{cases} 1 \ \ if \ \ \bar{f}_k = (s, c_{ij}, s_i, s_{i-1}) \\ 0 \qquad\qquad otherwise \end{cases} \qquad (2)$$

where $\bar{f}_k$ represents the co-occurrence of the generated concept $s$ and its context information of $c_{ij}$, $s_i$, and $s_{i-1}$.

During the training session, $\alpha_k$ is estimated to maximize the overall logarithmic likelihood over the training corpus using the Improved Iterative Scaling (IIS) algorithm [9] as shown in Equation (3).

$$\alpha_k = \underset{\alpha}{\mathrm{argmax}} \sum_{l=1}^{L} \sum_{s \in q_l} \sum_j \log[p(s \mid c_{ij}, s_i, s_{i-1})] \qquad (3)$$

where $\boldsymbol{Q}=\{q_l,\ 1 \leq l \leq L\}$ is the total set of concept sequences from the training corpus.

Based on (1), (2), and (3), the next concept output, $s_{i+1}$, is generated by choosing the candidate with the highest probability as shown in Equation (4)

$$s_{i+1} = \underset{s \in C_i}{\mathrm{argmax}} \{\prod_{j=1}^{M} p(s \mid c_{ij}, s_i, s_{i-1})\} \qquad (4)$$

## 3. NATURAL LANGUAGE GENERATION

### 3.1 Baseline NLG and Feature Selection

The baseline NLG assumes that the concepts observed in the source language are sufficient to reconstruct a new concept sequence for the target language but in a different sequential order. For training purpose, the baseline NLG proposes to estimate the ME models using data in the target language only. Thus, the issue of concept inequality due to unequal concepts between the source and target sentence pairs exemplified in Figure 2 is alleviated.

For concept generation, the baseline NLG produces an output concept sequence using a recursive structural generation algorithm as follows,

1) Traverse an annotated parse tree in a bottom-up and left-to-right manner;
2) For each new concept sequence in the parse tree, generate the best concept sequence for the output language based on Equations (1), (2), and (4); after each non-terminal node is visited, register the output under the node name;
3) Repeat step 2) until all parse braches in the source language are processed;
4) Replace nodes with their corresponding output sequence to form a complete concept sequence for the sentence.

Based on the ME framework, the baseline NLG uses a set of features, $\bar{f}_k$ = ($s_i$, $s_{i-1}$, $T$, $c_{ij}$, $\boldsymbol{C}_i$) where $T$ is the local sentence/phrase type denoted by the parent node at time $i$.

### 3.2 Direct-Modeling NLG

A direct-modeling-based NLG is proposed in this paper to utilize the simultaneous concept sequences from annotated parallel corpora for the estimation of ME models. The proposed approach can accomplish two merits. First, it eliminates a mismatch condition in which the baseline NLG uses concept sequences of the target languages for training but uses concept sequences of the source language for generation. Second, inclusion of the temporal concept information into the feature design can be enabled for better performance.

Two novel schemes used with the direct-modeling NLG are devised as follows,

a) “Null” concept padding:

The direct-modeling approach expects the same number of concepts in the source and target sequences. To compensate for concept number inequality, we append “null” concepts as a padding to make both sequences equal in length. Let us assume that a source sequence has $M$ concepts and its target language sequence has $N$ concepts. Theoretically, at least $|M-N|$ "null" concepts are needed as a padding to the shorter sequence. However, during the generation phase only the source language sequence is available. It is difficult to know in advance whether the resultant target language sequence is longer than the input sequence. To eradicate the length uncertainty, "$L$" more "null" concepts are used in the training and generation phrases where $L$ is the number of sibling concepts to be considered individually.

b) Forming multiple-question features for temporal information: One big advantage to use source language sequences in the training phase is the ability to preserve temporal information in the linguistic constituents of the input concepts. To utilize the temporal information, extra features are created by designing binary questions to cover a long concept span.

### 3.3 Direct Modeling in a ME Framework

Assume that $\boldsymbol{C} = \{c_1, c_2, c_{3,\ldots,} c_M\}$ is the original source language concept sequence and that the original concept output sequence $\boldsymbol{S} = \{s_1, s_2, s_{3,\ldots,} s_N\}$. We can define $R=|M\text{-}N|+L$ the new length of both sequences after padding with "null" concepts. After padding, the new source concept sequence is $\boldsymbol{C} = \{c_1, c_2, c_{3,\ldots,} c_M ,c_{M+1},\ldots,c_R \}$ and the new output sequence is $\boldsymbol{S} = \{s_1, s_2, s_{3,\ldots,} s_N, s_{N+1},\ldots, s_R\}$. Note that $c_{M+1},\ldots,c_R$ and $s_{N+1},\ldots, s_R$ are "null" concepts. Furthermore, we define a new set of features as

$$f = (s_i, s_{i-1}, T, c_{i+1}, c_{i+2},\ldots, c_{i+L}, r_{i+1}) \qquad (5)$$

where $r_i$ is a long concept span covering more temporal sibling concepts and $L$ is the number of sibling concepts to be considered individually. In a ME framework, the binary test function for direct modeling is written as

$$g(\bar{f}_k, s, s_i, s_{i-1}, T, c_{i+1},.., c_{i+L}, r_{i+1}) = \begin{cases} 1 \;\; if \;\; \bar{f}_k = (s, s_i, s_{i-1}, T, c_{i+1},.., c_{i+L}, r_{i+1}) \\ 0 \qquad\qquad otherwise \end{cases}$$

(6).

Likewise, the probability weight can be re-written as

$$\alpha_k = \underset{\alpha}{\operatorname{argmax}} \sum_{l=1}^{L} \sum_{s \in q_l} \sum_{j} \log[p(s \mid s_i, s_{i-1}, T, c_{i+1},\ldots, c_{i+L}, r_{i+1})] \qquad (7)$$

The next concept output, $s_{i+1}$, is generated by choosing the candidate with the highest probability as

$$s_{i+1} = \underset{s \in C}{\operatorname{argmax}} \{\prod_{j=1}^{M} p(s \mid s_i, s_{i-1}, T, c_{i+1},\ldots, c_{i+L}, r_{i+1})\} \qquad (8)$$

## 4. EXPERIMENTS AND RESULTS

**Experiment Setup**. The experiments carried out to evaluate the NLG performance of IBM MASTOR are conducted in a DARPA domain covering medical assistance and force protection. In this paper, the experiments are carried out on the Mandarin Chinese to English translation in the two-way translation MASTOR system.

Two kinds of concept sequences will be used. The first type is the top-level concept sequences extracted only from the top level of the semantic parse trees. The second type is the all-level sequences obtained from all levels of depth in the parse trees. Totally, there are 10400 parallel top-level concept sequences and 44000 parallel all-level concept sequences. The in-domain word vocabulary size is about 3300 words for both English and Chinese. There are 207 concepts designed for NLU and NLG. All 207 concepts are used in the all-level concept generation while 67 of them are used in the top-level concept generation.

**Performance Measure:** For concept generation, we use a measure called concept error rate (CER), similar to word error rate (WER) used in speech recognition. We also compute the concept sequence error rate (CSER) in the same fashion as sentence error rate in speech recognition.

To measure the overall speech-to-speech translation, an objective measure, BLEU [10], proposed by IBM is used. BLEU measures the translation quality based on N-gram probabilities and brevity between hypothesis and reference sentences. 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. It is defined as

$$\text{BLEU} = \exp(\textstyle\sum_{\text{i}} \sum_{\text{n}} w_n * \log(P(w_i \mid w_{i-1},.., w_{i-n+1}))) * BP \qquad (9)$$

where $P(w_i|w_{i-1},\ldots,w_{i-n+1})$ is the n-gram probability, $w_n$ is the n-gram weight, and $BP$ is the brevity penalty.

### 4.1 Experiments on Top-Level Concept Sequences

Only the top-level concept sequences extracted from the top level of a semantic parse tree are considered in this section. 1486 out of 10400 concept sequences are randomly selected for testing and the remaining are used for training.

| NLG scheme | CER(%) | CSER(%) |
|---|---|---|
| Baseline | 23.9 | 46.9 |
| DM $L$=2 | 22.0 | 49.1 |
| DM $L$=3 | 19.2 | 44.3 |
| DM $L$=4 | 18.0 | 42.1 |
| DM $L$=5 | 16.7 | 40.4 |
| DM $L$=6 | 16.7 | 40.3 |
| DM $L$=7 | 16.6 | 40.4 |
| DM $L$=8 | 16.8 | 40.8 |

**Table 1:** Comparison of direct-modeling(DM) NLG and baseline NLG on top-level sequences; $L$ is length of "null" concept

As shown in Equation (5), first we need to determine an optimal value for $L$, the length of individual sibling concepts. Table 1 compares the concept generation between our new direct-modeling NLG and the baseline NLG on the 1486 test top-level concept sequences. The direct-modeling NLG outperforms the baseline by a significant margin. The direct-modeling NLG with $L$=7 yields a CER of 16.6% and a CSER of 40.4% while the baseline yields a CER of 23.9% and a CSER of 46.9%.

Next, we want to investigate the benefits of using a long concept span to incorporate the temporal information. In the experiments, the optimal length of the concept span, $F$, is to be determined while $L$ is set to be 7.

Table 2 shows that further improvement can be achieved by using a long concept span covering sibling concepts. The best performance is obtained with a CER of 14.3% and a CSER of 37.6% when the concept span length is 5. It represents a significant 40% of error reduction on CER generated by the direct-modeling NLG over the baseline NLG.

| NLG scheme | CER(%) | CSER(%) |
|---|---|---|
| Baseline | 23.9 | 46.9 |
| DM *F*=1 | 16.4 | 40.1 |
| DM *F*=2 | 15.2 | 38.6 |
| DM *F*=3 | 14.8 | 38.2 |
| DM *F*=4 | 14.3 | 37.6 |
| DM *F*=5 | 14.3 | 37.6 |
| DM *F*=6 | 14.5 | 37.6 |
| DM *F*=7 | 14.5 | 37.7 |

**Table 2:** Comparison of direct-modeling (DM) NLG and baseline NLG on top-level sequence when *L*=7; *F* is the concept span length

### 4.2 Experiments on All-Level Concept Sequences

Experiments are carried out on the all-level concept sequences in this section. To get an equal number of concept sequences for the direct-modeling approach, only the sequences corresponding to the same parent concept in each pair of sentence parse trees from both languages are considered. Altogether, there are 44000 parallel concept sequences, of which 6365 sequences are used as a test set and the remaining are used for training.

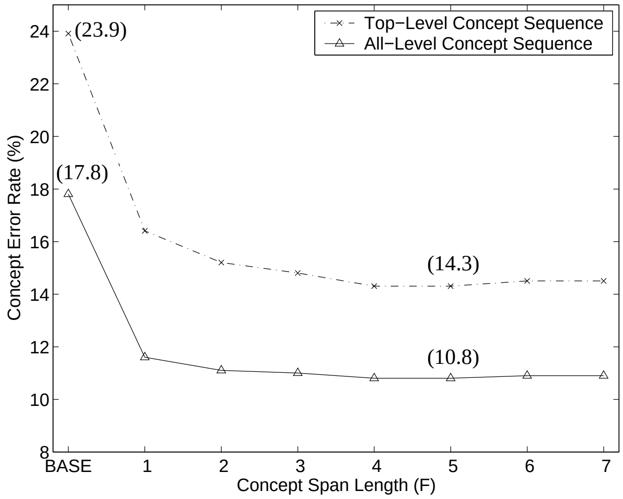

**Figure 3:** Comparison of direct-modeling (DM) NLG and baseline NLG on all-level sequences when *L*=7; *F* is the concept span length

Figure 3 compares the baseline NLG to the direct-modeling NLG of different length of concept span on the all-level sequences and top-level sequences. The best performance on the all-level concept sequences is achieved with a CER of 10.8% and a CSER of 37.6% when the concept span length is 5, while the baseline yields a CER of 17.8% and a CSER of 31.4%. It is worth noting that the results on the all-level sequences are better than those on the top-level sequences. This is due to the fact that the non-top level sequences have simpler parse tree structures with a shorter average sequence length than the top level sequences as illustrated in Figure 2.

### 4.3 Experiments of Direct-Modeling for Word Translation

With the success of direct-modeling approach on concept generation, preliminary experiments are conducted on the text translation. Table 3 shows the results of statistical interlingua-based text-to-text translation using BLEU in Equation (9).

| NLG type | BLEU | 1-grm / 2-grm / 3-grm / 4-grm / BP |
|---|---|---|
| Baseline NLG | 0.252 | 0.666 / 0.341 / 0.178 / 0.100 / 1.000 |
| DM(L=7,F=5) | 0.294 | 0.679 / 0.379 / 0.218 / 0.133 / 1.000 |

**Table 3:** Comparison of direct-modeling (DM) NLG and baseline NLG in the context of interlingua-based text-to-text translation

Both baseline NLG and direct-modeling NLG models are estimated using 68000 concept sequences for a word translation evaluation task. The experiments are conducted on 582 sentences from a development test set in a medical domain. Table 3 shows that the direct-modeling NLG generates a better BLEU score than the baseline with considerable improvement on higher N-gram probabilities. Further analysis also reveals that the direct-modeling tends to produce slightly shorter translation output. We believe that this is an artifact that current training scheme uses only concept sequence pairs with comparable concepts and that this can be improved with better NLU parser output. Overall, these preliminary results show that the direct-modeling scheme is a promising approach for interlingua-based speech translation systems.

## 5. SUMMARY

Statistical concept generation modeling is a crucial component in an interlingua-based speech translation system. A new direct-modeling-based scheme is proposed to improve natural language generation. To overcome the concept inequality issue, a "null" concept padding scheme is adopted to apply to both training and generation phrases. New feature designs to cover multiple concepts in a long concept span are investigated. Significant improvements are observed in concept generation for both top-level and all-level concept sequences. When it is incorporated into our system, the direct-modeling scheme achieves a sensible improvement in the preliminary translation experiments.

## 6. REFERENCES

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

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

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

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

[5] F.-H. Liu, et al, "Use of Statistical N-Gram Models in Natural Language Generation for Machine Translation", *ICASSP-2003*, 2003.

[6] Y. Gao, et al, "MARS: A Statistical Semantic Parsing and Generation Based Multilingual Automatic Translation System", to appear in Machine Translation

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

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

[9] A. Ratnaparkhi,"Trainable Methods for Surface Natural Language Generation", *NAACL-2000 Proc.*, Seattle, WA: NAACL, pp. 194-201

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