# On Feature Selection in Maximum Entropy Approach to Statistical Concept-based Speech-to-Speech Translation

*Liang Gu* and *Yuqing Gao*

IBM T. J. Watson Research Center, Yorktown Heights, NY 10598
{lianggu, yuqing}@us.ibm.com

## ABSTRACT

Feature selection is critical to the performance of maximum-entropy-based statistical concept-based spoken language translation. The source language spoken message is first parsed into a structured conceptual tree, and then generated into the target language based on maximum entropy modeling. To improve feature selection in this maximum entropy approach, a new concept-word feature is proposed, which exploits both concept-level and word-level information. It thus enables the design of concise yet informative concept sets and easies both annotation and parsing efforts. The concept generation error rate is reduced by over 90% on training set and 7% on test set in our speech translation corpus within limited domains. To alleviate data sparseness problem, multiple feature sets are proposed and employed, which achieves 10%-14% further error rate reduction. Improvements are also achieved in our experiments on speech-to-speech translation.

## 1. INTRODUCTION

Automatic spoken language translation is crucial to speech-to-speech (S2S) translation systems that facilitate communication between people who speak different languages. While substantial progress has been made over the past decades in research areas of speech recognition and machine translation, multilingual natural speech translation remains a grand challenge for human speech and language technologies [1,2,3,4]. Compared to written-text messages, most conversational spoken messages are conveyed through casual spontaneous speech with strong disfluencies and imperfect syntax. In addition, the output from speech recognizers often contains recognition errors and no punctuations, which brings serious challenges to robust and accurate translation.

In our prior work [5], we presented a statistical spoken language translation framework based on tree-structured semantic/syntactic representations, or *concepts*, as illustrated in Figure 1. In this example, the source English sentence and the corresponding Chinese translation are represented by a set of concepts – {PLACE, SUBJECT, WELLNESS, QUERY, PREPPH, BODY-PART}. Some of the concepts (such as PLACE, WELLNESS and BODY-PART) are semantic representations while some of the concepts (such as PREPPH) are syntactic representations. There are also concepts (such as SUBJECT and QUERY) that represent both semantic and syntactic information. Note that although the source and target sentences share the same set of concepts, the tree structures are significantly different from each other because of the well-known distinct nature of these two languages (i.e., English and Chinese).

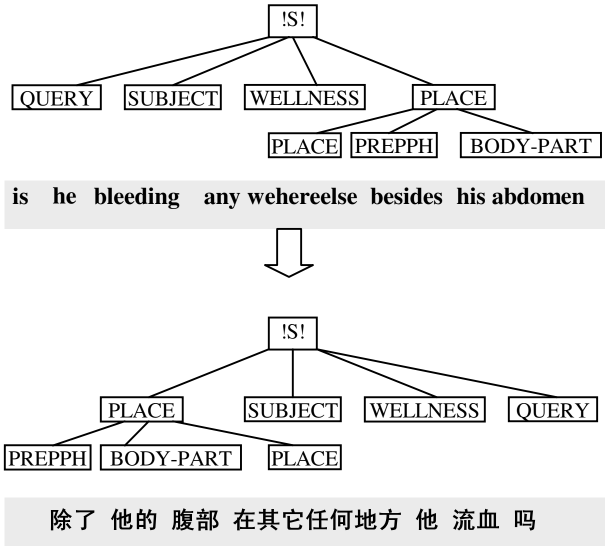


Figure 1. Example of Concept-based English-to-Chinese Translation

The above concept tree is comparable to interlingua [1] - a language-independent representation of intended meanings that is commonly used in modern spoken language translation systems. In our approach, the intended meanings are represented by a set of *language-independent* concepts (same as conventional interlingua approach) organized in a *language-dependent* tree-structure (different from conventional interlingua method). The process of this concept-based translation may be further divided into two cascaded sub-processes: a) the generation of conceptual tree structure, and b) the generation of words within each concept, in the target language. While the total number of concepts may usually be limited to alleviate data sparseness impacts (especially for new domains), there are no constraints on the structures of the conceptual trees. Therefore, compared to traditional interlingua-based speech translation approaches, our conceptual-tree-based approach could achieve more flexible meaning preservation with wider coverage and, hence, higher robustness and accuracy on translation tasks in limited domains, at the cost of additional challenges in the appropriate transformation of conceptual trees between source and target languages.

Two principal challenges remain open in the design of concept-based speech translation systems. One challenge is the design and selection of language-independent concepts, which usually depends on the domain in which the translation system is used.

This is a lengthy, tedious but very important task. The concepts have to be not only broad enough to cover all intended meanings in the source sentence but also informative so that a target sentence can be generated with right word sense and in a grammatically correct manner. The size of the concept set is also important as too many concepts may result in data sparseness for training, while too few concepts could degrade the translation accuracy.

Another challenge is the generation of concepts in the target language via a *natural concept generation* (NCG) process. The purpose of NCG is to generate the correct concept structure in the target language corresponding to the concept structure in the source language. As explained before, the concept structures are language-dependent. Errors in concept generation could greatly distort or even ruin the meaning to be expressed in the target language, particularly in conversational speech translations where in most cases only a few concepts are conveyed in the messages to be translated. Therefore, accurate and robust NCG is viewed as an essential step towards high-performance concept-based spoken language translation.

While NCG approaches can be rule-based or statistical, we prefer the latter because of its trainability, scalability and portability. One such approach based on maximum-entropy (ME) criterion was presented in our previous work [5]. It was then improved in [6] and [7] by the employment of a series of algorithms such as forward-backward modeling and confidence measurement.

One critical problem remain in our ME-based translation approach is feature selection. In theory, the principle of maximum entropy does not directly concern itself with the issue of feature selection [8]. It merely provides a framework to combine constraints of both source and target language into a translation model. In reality, however, the feature selection problem is crucial to the performance of ME-based approaches, since the universe of possible constraints (or features) is typically in thousands or even millions for natural language processing. Some of these impacts on ME-based speech translation were preliminarily described in our previous work [6].

In this paper, to address the above concerns, we analyze and discuss in greater detail the feature selection issue in the design of ME-based statistical concept-based speech translation systems. In particular, a novel feature is proposed to use the combination of concept and word information to achieve higher NCG accuracy while minimize the total number of distinct concepts and hence greatly reduce the concept annotation and natural language understanding effort. A multiple feature selection algorithm is further employed to handle data sparseness issues. Experiments with these new algorithms are performed and analyzed on both the NCG accuracy and the overall speech translation performance.

## 2. BASELINE STATISTICAL NATURAL CONCEPT GENERATION USING MAXIMIZING ENTROPY MODELS

### *A. Statistical Concept-based S2S Translation*

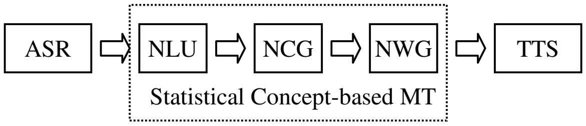


Figure 2 IBM MASTOR (**M**ultilingual **A**utomatic **S**peech-to-Speech **T**ranslat**OR**) System

Figure 2 shows a general framework of our MASTOR speech translation system for applications in limited domains. A cascaded scheme of large-vocabulary conversational automatic speech recognition (ASR), statistical concept-based machine translation and concatenative text-to-speech (TTS) synthesis is applied by using state-of-the-art speech and language processing techniques. While each of these three functional units is crucial to the overall speech-to-speech translation quality, we are only concerned with the performance of statistical concept-based translation here.

The baseline statistical concept-based translation further consists of three cascaded functional components: natural language understanding (NLU), natural concept generation (NCG) and natural word generation (NWG). In our MASTOR system, the NLU function is performed via a decision-tree-based statistical semantic parser pre-trained on an annotated text corpus [9]. The NWG process generates words in the target language based on the generated structural concepts from NCG as well as a tag-based word-to-word multilingual dictionary [10]. Although these two components are very important to our statistical interlingua-based translation, they are, again, beyond the scope of this paper.

The NCG process generates a set of structural concepts in the target language according to a concept-based semantic parse tree derived from the NLU process in the source language. The accuracy of the NCG process has a great impact on the final translation performance as any errors of inserted, missing, replaced or mistakenly ordered concepts may cause severe understanding problems or loss of meaning during multilingual speech communication. Therefore, highly accurate NCG is essential to our goal of meaning preservation in conversational speech translation. In this paper, we focus on improving the ME-based statistical NCG method, as explained next.

### *B. ME-based Statistical NCG on Sequence Level*

The baseline statistical NCG algorithm on sequence level was proposed in [5] as an extension from the “NLG2” algorithm described in [11]. During natural concept sequence generation, the concept sequences in the target language are generated sequentially according to the output of NLU parser. Each new concept is generated based on the local n-grams of the up-to-date generated concept sequence and the subset of the input concept sequence that has not yet appeared in the generated sequence.

Let us assume that the source language concept sequence produced from NLU parser is $C = \{c_1, c_2, \cdots, c_M\}$. Let us further assume that a concept sequence $S = \{s_1, s_2, \cdots, s_n\}$ containing *n* concepts has already been generated in target language. In order to generate the next new concept $s_{n+1}$, the conditional probability of a concept candidate is defined and computed as

$$p\left(s\middle|c_m, s_n, s_{n-1}\right) = \frac{\prod_k \alpha_k^{g\left(\vec{f}_k, s, c_m, s_n, s_{n-1}\right)}}{\sum_{s\in V}\prod_k \alpha_k^{g\left(\vec{f}_k, s, c_m, s_n, s_{n-1}\right)}} \quad , \qquad (1)$$

where $s$ is the concept candidate to be generated, $s_n$ and $s_{n-1}$ are the previous two concepts in $S$. $V$ is the set of all possible concepts that can be generated.

$\vec{f}_k = \left(s_{+1}^k, c^k, s_0^k, s_{-1}^k\right)$ is the *k*-th feature. The selection of $\vec{f}_k$ will be discussed in the next section.

$\alpha_k$ is a probability weight corresponding to each feature $\vec{f}_k$. The value of $\alpha_k$ is always positive and is optimized over a training corpus by maximizing the overall logarithmic likelihood, i.e.,

$$\alpha_k = \arg\max_{\alpha} \sum_{l=1}^{L}\sum_{s\in q_l}\sum_{m} \log\left[p\left(s\middle|c_m, s_n, s_{n-1}\right)\right], \qquad (2)$$

where $Q = \{q_l, 1 \le l \le L\}$ is the total set of concept sequences. The optimization process can be accomplished via the Improved Iterative Scaling algorithm using maximum entropy criterion described in [11].

$g$ is a binary test function defined as

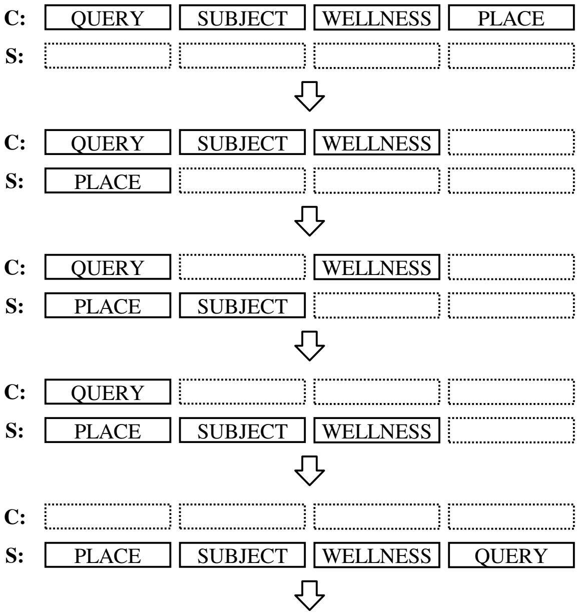
$$g\left(\vec{f}_k, s, c_m, s_n, s_{n-1}\right) = \begin{cases} 1 & \textit{if } \vec{f}_k = \left(s, c_m, s_n, s_{n-1}\right) \\ 0 & \textit{otherwise} \end{cases} \qquad (3)$$

where $\vec{f}_k$ represents the co-occurrence of the generated concept $s$ and its context information of $c_m, s_n$ and $s_{n-1}$.

Using (1), (2) and (3), $s_{n+1}$ is generated by selecting the concept candidate with highest probability, i.e.,

$$s_{n+1} = \arg\max_{s\in V}\left\{\prod_{m=1}^{M} p\left(s\middle|c_m, s_n, s_{n-1}\right)\right\} \quad . \qquad (4)$$

For an input concept sequence $C = \{c_1, c_2, \cdots, c_M\}$, the generation procedure is performed as follows:



C: QUERY SUBJECT WELLNESS PLACE
S:
C: QUERY SUBJECT WELLNESS
S: PLACE
C: QUERY WELLNESS
S: PLACE SUBJECT
C: QUERY
S: PLACE SUBJECT WELLNESS
C:
S: PLACE SUBJECT WELLNESS QUERY

Figure 3. Example of Concept Sequence Generation during translation of English sentence “is he bleeding anywhere else besides his abdomen” as illustrated in Figure 1.

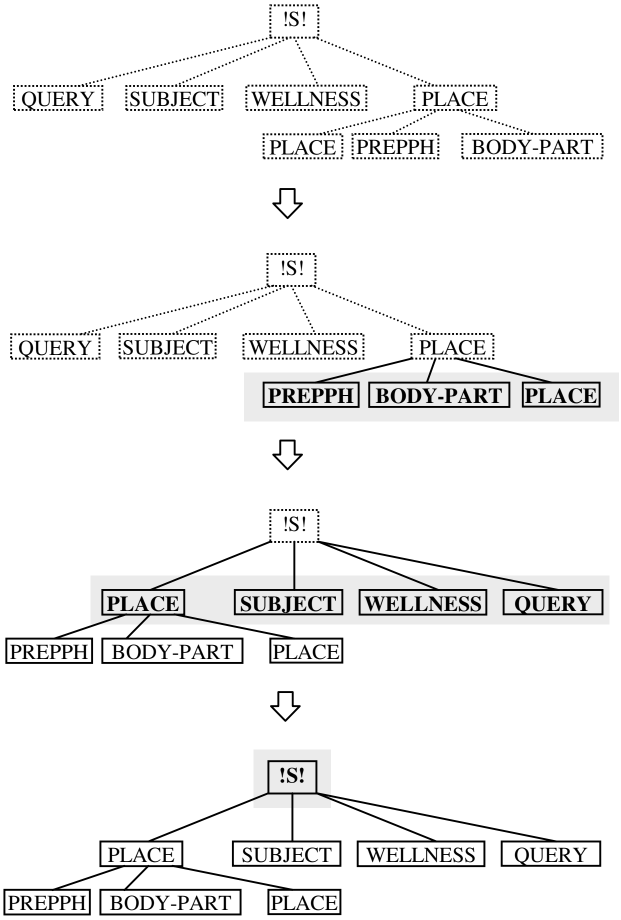


Figure 4. Example of Structural Concept Generation during translation of English sentence “is he bleeding anywhere else besides his abdomen” as illustrated in Figure 1 and Figure 3.

1) Set $s_0 = s_{-1} =$ START, where "START" is a pre-defined concept representing the start of the sequence; Set $n = 0$;

   Define initial set of generation sequence $S = \phi$;

2) For each $n$, generate $s_{n+1}$ according to equation (3) and set $S = \{s_1, \cdots, s_{n+1}\}$;

3) If $s_{n+1} \in C$, set $C = C - s_{n+1}$ (remove $s_{n+1}$ from $C$); accordingly, let $M \leftarrow M - 1$;

4) If $M \geq 1$ or $n + 1 \leq N$, repeat 2) and 3); Otherwise, stop and output generated concept sequence $S$.

Since the number of concepts generated in $S$ could be different from the number of concepts in the input sequence in the source language, only a maximum number (denoted as $N$) of concepts may be generated. In our experiments, $N = 11$.

An example of primary-level (or main level) concept sequence generation is depicted in Figure 3 when translating the English sentence in Figure 1 into Chinese.

### C. Structural Concept Sequence Generation

The algorithms described above only deal with the concept generation issue of a single sequence. To tackle the generation problem of multiple sequences at different structural levels, a recursive structural concept sequence generation algorithm is proposed in [2,3] as follows:

1) Traverse the semantic parse tree in a bottom-up left-to-right manner;

2) For each un-processed concept sequence in the parse tree, generate an optimal concept sequence in the target language based on the procedure described in sub-section 2.B; after each concept sequence is processed, mark the root-node of this sequence as visited;

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 tree for the output sentence.

An example of structural concept sequence generation is depicted in Figure 4 when translating the English sentence in Figure 1 and Figure 3 into Chinese.

## 3. FEATURE SELECTION IN MAXIMUM-ENTROPY-BASED STATISTICAL NCG

### A. Problem Statement and Baseline Features

Earlier we introduced two basic challenges in the design of statistical maximum-entropy-based models for natural concept generation: 1) finding appropriate facts or *features* about the observed data; 2) optimally incorporate these features into the target models. In the previous section, we solved the second problem by using maximum-entropy principle in equations (1-4). In this section, we will attack the first challenge and improve natural concept generation performance by augmenting feature dimensions and combining various feature sets, as explained next.

We begin with the basic four-dimensional feature set $\vec{f}_k^{(4)} = \left(s_{+1}^k, c^k, s_0^k, s_{-1}^k\right)$ defined in equation (1) and (2), which was first proposed in [5]. In this feature set, the order of concepts in the input sequence is discarded to alleviate performance degradation caused by sparse training data. However, there exist many cases in which the same set of concepts need to be generated into two different concept sequences depending on the order of the input sequence. For these typical concept sequences, generation errors are inevitable with the features of the specific form no matter how the statistical model is optimized.

To tackle this problem, we proposed in [6] an augmented five-dimensional feature as $\vec{f}_k^{(5)} = \left(s_{+1}^k, c_0^k, c_{+1}^k, s_0^k, s_{-1}^k\right)$, where $c_0^k$ and $c_{+1}^k$ are two adjacent concepts in the source concept sequence $C$. Accordingly, the conditional probability of a concept candidate and the probability weights are modified as

$$p\left(s \middle| c_m, c_{m+1}, s_n, s_{n-1}\right) = \frac{\prod_k \alpha_k^{g_k\left(\vec{f}_k^{(5)}, s, c_m, c_{m+1}, s_n, s_{n-1}\right)}}{\sum_{s \in V} \prod_k \alpha_k^{g_k\left(\vec{f}_k^{(5)}, s, c_m, c_{m+1}, s_n, s_{n-1}\right)}}, \quad (5)$$

$$\alpha_k = \arg\max_{\alpha} \sum_{l=1}^{L} \sum_{s \in q_l} \sum_{m=1}^{M-1} \log\left[p\left(s \middle| c_m, c_{m+1}, s_n, s_{n-1}\right)\right]. \quad (6)$$

Since the above features represents concept orders in both source and target sequences, $\vec{f}_k^{(5)}$ are extracted from pre-annotated parallel corpora during ME-based model training. Particularly, the optimization of (6) is performed upon a parallel tree-bank $QQ = \{u_l, v_l \mid 1 \leq l \leq L\}$, where $u_l$ and $v_l$ are the concept sequences in source and target language, respectively. For each feature $\left(s, c_m, c_{m+1}, s_n, s_{n-1}\right)$ during ME model training, $c_m$ and $c_{m+1}$ are derived from $u_l$, while $s_n$ and $s_{n+1}$ are derived from $v_l$. This augmented feature strengthens the link between sequences in source and target languages, and can thereby improve NCG accuracy as reported in [6].

### B. Conciseness versus Informativity of Concepts

So far we tried to extract features on the concept level. However, as explained earlier, the definition and detection of concept itself is a very challenging task. On the one hand, the concepts are defined as concise as possible, since the smaller the number of total distinct concepts, the less the effort will be endeavored in the labor-extensive and time-consuming annotation procedure, and the higher the accuracy and robustness will be of the statistical natural language understanding algorithms. On the other hand, the concepts should be as informative as possible, because the concept generation accuracy will largely rely on the sufficient information provided by each concept.

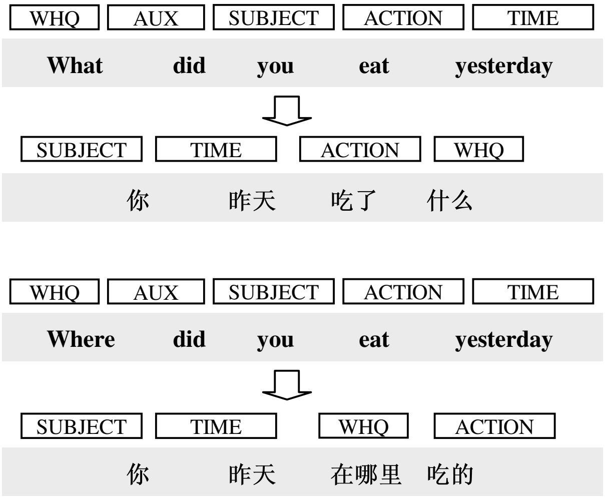


Figure 5(a) Example of Concept-based English-to-Chinese Translation with concept information

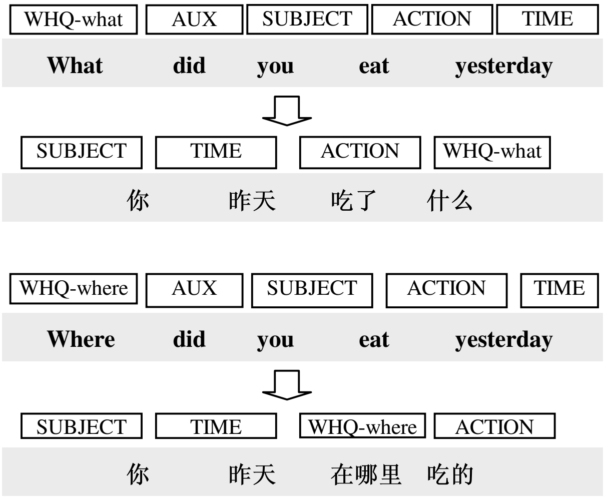


Figure 5(b) Example of Concept-based English-to-Chinese Translation with concept and sub-concept information

Figure 5(a) gives a real example in our medical speech translation domain (what did you eat yesterday vs. where did you eat yesterday) where insufficient information of concepts cause generation confusion. While the two input English sentences share exact the same set and order of concepts, the correct concept orders in Chinese are clearly different. Therefore, whether using feature $\vec{f}_k^{(4)}$ or $\vec{f}_k^{(5)}$, errors are inevitable no matter how well the ME models are optimized. For this specific example, the reason is quite obvious: the concept WHQ is too concise that it is not informative enough to discriminate the different generation behavior between (WHQ,what) and (WHQ,where).

One possible solution to the above problem is to expand current concept set and thus make it more informative. When concept WHQ is divided into two sub-concepts: WHQ-what and WHQ-where, the previous confusion is removed as depicted in Figure 5(b). Unfortunately, this also dramatically increase the total number of distinct, and therefore introduce more much burden on human annotation of these concepts in the training data. More importantly, the expansion of concepts may subject to much lower parsing accuracy during natural language understanding (NLU) due to the much worse data sparseness problem.

### *C. Features using both Concept and Word Information*

Instead of expanding concept sets with the above drawbacks, we propose a new approach based on a novel feature set that uses both concept and word level information. For any input concept sequence $C = \{c_1, c_2, \cdots, c_M\}$, assume its corresponding word sequence is $W = \{\overline{w}_1, \overline{w}_2, \cdots, \overline{w}_M\}$, a seven-dimensional feature is defined as $\vec{f}_k^{(7)} = \left(s_{+1}^k, c_0^k, c_{+1}^k, \overline{w}_0^k, \overline{w}_{+1}^k, s_0^k, s_{-1}^k\right)$, where $\overline{w}_0^k$ and $\overline{w}_{+1}^k$ are two word phrases belong to $c_0^k$ and $c_{+1}^k$ respectively. Accordingly, the conditional probability of a concept candidate and the probability weights are modified as

$$p(s|c_m, c_{m+1}, s_n, s_{n-1}) = \frac{\prod_k \alpha_k^{g_k\left(\vec{f}_k^{(7)}, s, c_m, c_{m+1}, \overline{w}_m, \overline{w}_{m+1}, s_n, s_{n-1}\right)}}{\sum_{s\in V}\prod_k \alpha_k^{g_k\left(\vec{f}_k^{(7)}, s, c_m, c_{m+1}, \overline{w}_m, \overline{w}_{m+1}, s_n, s_{n-1}\right)}} ,(7)$$

$$\alpha_k = \underset{\alpha}{\operatorname{argmax}} \sum_{l=1}^{L}\sum_{s\in q_l}\sum_{m=1}^{M-1}\log\left[p(s|c_m, c_{m+1}, \overline{w}_m, \overline{w}_{m+1}, s_n, s_{n-1})\right]. \quad (8)$$

The feature $\vec{f}_k^{(7)}$ removes the confusion in Figure 5(a) as now two different feature sets are extracted for sentence {what did you eat yesterday} and sentence {where did you eat yesterday}. Compared to the expansion approach discussed in the previous sub-section, this concept-word-feature approach keeps the original concept set intact, and therefore maintain both high conciseness and Informativity of concepts by taking into account the word-level information when building maximum-entropy-based statistical models.

### *D. Enhancing Robustness by Combining Multiple Feature Sets*

One concern to the above concept-word-feature is the much severe data sparseness problem. Given a typical concept vocabulary size of 70 and a word vocabulary of mere 3000, the total possible number of feature $\vec{f}_k^{(5)}$ in equation (5) is 1.7e+9, while the total possible number of feature $\vec{f}_k^{(7)}$ in equation (7) is 1.5e+16. That is a $10^7$–times bigger space! To solve the resulted data sparseness issue, a combination of feature sets is proposed in ME-based concept generation. Multiple feature sets are extracted with various dimensions and concept/word constraints. In particular, we combine features $\vec{f}_k^{(5)}$ in equation (5) and features $\vec{f}_k^{(7)}$ in equation (7). These two sets of features adopted in the optimization of ME models as

$$\alpha_k = \operatorname*{argmax}_{\alpha} \sum_{l=1}^{L} \sum_{s \in q_l} \sum_{m=1}^{M-1} \left\{ \begin{array}{l} \log \dfrac{\prod_k \alpha_k^{g_k\left(\vec{f}_k^{(5)}, s, c_m, c_{m+1}, s_n, s_{n-1}\right)}}{\sum_{s \in V} \prod_k \alpha_k^{g_k\left(\vec{f}_k^{5}, s, c_m, c_{m+1}, s_n, s_{n-1}\right)}} \\ + \log \dfrac{\prod_k \alpha_k^{g_k\left(\vec{f}_k^{(7)}, s, c_m, c_{m+1}, \overline{w}_m, \overline{w}_{m+1}, s_n, s_{n-1}\right)}}{\sum_{s \in V} \prod_k \alpha_k^{g_k\left(\vec{f}_k^{7}, s, c_m, c_{m+1}, \overline{w}_m, \overline{w}_{m+1}, s_n, s_{n-1}\right)}} \end{array} \right\} \quad (9)$$

During ME-based generation, when both feature sets are observed, the more informative feature $\vec{f}_k^{(7)}$ will dominate the generation results. In other cases, when only feature $\vec{f}_k^{(5)}$ is matched, the combined ME models will back off to the more robust models defined in (5).

# 4. EXPERIMENTS

The performance of our new algorithms in ME-NCG and statistical concept-based spoken language translation was evaluated on the English-to-Chinese speech translation task within a limited domain of emergency medical care. Altogether 10,000 conversational in-domain parallel sentences in both English and Chinese were collected and annotated as the data corpus for evaluation. The vocabulary size is about 3000 in each language. 68 concepts were designed and used for data annotation, NLU model training and NLU parsing.

### *A. Experiments on ME-based statistical NCG*

The first set of experiments is carried out on the concept level to evaluate the performance of ME-based statistical NCG. A primary concept sequence is extracted from each annotated sentence, which represents the top-layer concepts in a semantic parser tree. Concept sequences containing only one concept are removed as they are easy to generate. To further simplify the problem, we train and test on parallel concept sequences that contain the same set of concepts in English and Chinese. In this specific case, NCG is performed to generate the correct order of concepts in the sequences of target language. More general and complex experiments are performed and shown in the next subsection.

According to the above criterion, about 5600 concept sequences are selected as our experimental corpus. During experimentation, this corpus is randomly partitioned into training corpus containing 80% of the sequences and test corpus with the remaining 20%. This random process is repeated 50 times and the average performance is recorded. Two evaluation metrics were applied. A concept sequence is considered to have an error during measurement of sequence error rate if one or more errors occur in this sequence. Concept error rate, on the other hand, evaluates concept errors in concept sequences such as substitution, deletion and insertion.

In the first experiment, various feature types were implemented, tested and compared on both the training and test corpus with basic forward generation models. The results are shown in Table 1. As expected, the use of concept-word features dramatically reduced the sequence/concept generation error rate from 14.0% / 8.8% with baseline four-dimensional features, and 9.1% / 5.5 % with five-dimensional features on parallel corpora, to 0.7% / 0.4%, which represents a 95% and 92% error rate reduction, respectively. The improvement becomes smaller on the test-set error rate, which is 27.9% and 7.0%, respectively. After combining $\vec{f}_k^{(5)}$ and $\vec{f}_k^{(7)}$ according to equation (9), additional 13.9% error reduction was achieved on the test data. These experimental results clearly demonstrate that the concept-word features are superior to our previous proposed features, especially when the multiple feature set algorithm is employed.

| ME-NCG Methods | Training-set | Test-set |
|---|---|---|
| Baseline NCG with basic feature $\vec{f}_k^{(4)}$ | 14.0 % / 8.8 % | 28.0 % / 18.9 % |
| + feature on parallel corpora ( $\vec{f}_k^{(5)}$) | 6.2 % / 3.5 % | 21.7 % / 14.1 % |
| + concept-word features ( $\vec{f}_k^{(7)}$) | 0.7 % / 0.4 % | 20.2 % / 13.1 % |
| + multiple feature sets ( $\vec{f}_k^{(5)}$ + $\vec{f}_k^{(7)}$) | 0.7 % / 0.4 % | 17.4 % / 11.4 % |

*Table* 1. ME-NCG performance (sequence error rate / concept error rate) using different features with forward generation models.

| ME-NCG Methods | Training-set | Test-set |
|---|---|---|
| Baseline NCG with basic feature $\vec{f}_k^{(4)}$ | 9.1 % / 5.5 % | 24.4 % / 16.4 % |
| + feature on parallel corpora ( $\vec{f}_k^{(5)}$) | 5.7 % / 3.2 % | 17.8 % / 11.6 % |
| + concept-word features ( $\vec{f}_k^{(7)}$) | 0.5 % / 0.3 % | 17.7 % / 11.5 % |
| + multiple feature sets ( $\vec{f}_k^{(5)}$ + $\vec{f}_k^{(7)}$) | 0.5 % / 0.3 % | 15.8 % / 10.4 % |

*Table* 2. ME-NCG performance (sequence error rate / concept error rate) using different features with forward-backward generation models.

In the second experiments, the above features were evaluated with more advanced forward-backward generation models proposed in [7]. The results are listed in Table 2. While similar huge improvements were recorded on the training set, the concept-word features alone did not obtain significant accuracy improvement on the test set over previously proposed parallel features $\vec{f}_k^{(5)}$. Even so, 10.7% error rate reduction was achieved when multiple feature sets of $\vec{f}_k^{(5)}$ and $\vec{f}_k^{(7)}$ are used. The failure of significant improvement of $\vec{f}_k^{(7)}$ alone indicts strong overtraining because of much larger feature space and the resulted data sparseness problem. This problem is alleviated by the proposed multiple feature sets which lead to a decent improvement over our best performance previously proposed.

### *B. Experiments on statistical concept-based S2S translation*

Experimental results on statistical concept-based text-to-text and speech-to-text translation are shown in Table 4 based on the Bleu

| Translation Methods | $\vec{f}_k^{(4)}$ | $\vec{f}_k^{(5)}$ | ( $\vec{f}_k^{(5)}$ + $\vec{f}_k^{(7)}$ |
|---|---|---|---|
| Text-to-Text | 0.536 | 0.578 | 0.605 |
| Speech-to-Text | 0.437 | 0..469 | 0.489 |

*Table* 3. Improvement of Bleu score in S2S translation by using new algorithms in ME-NCG (the score may range from 0 .0 to 1.0, with 1.0 indicating best translation quality)

score described in [12], which measures MT performance by evaluating n-gram accuracy with a brevity penalty. It is now one of the most widely accepted evaluation metric in the machine translation society.

277 unseen speech sentences are tested. The new methods proposed achieved better performance compared to both baseline 1 (NCG process described in [6]) and baseline 2 (NCG methods proposed in [7]). From Table 3 we can see that, while the improvement is significant, the relatively smaller gains of overall S2S performance compared with NCG gains in Table 1 imply the importance of other S2S functional units, and the importance of further algorithmic improvement in all of these units.

## 5. CONCLUSION

Feature selection is a critical functional component in our maximum-entropy-based statistical natural concept generation. A new concept-word feature is proposed in this paper that exploits both the concept-level and word-level information during the training and decoding of maximum entropy models. It is then combined with our previous proposed features to alleviate the data-sparseness-caused over-training problem. Significant improvements are achieved in both concept sequence generation and speech translation experiments.

## 6. REFERENCES

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

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

[3] H. Hey, 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", *Proceedings of ICSLP*, 1998.

[5] Y. Gao, et al, "MARS: A statistical semantic parsing and generation based multilingual automatic translation system", *Machine Translation*, 2004.

[6] L. Gu, et al, "Improving Statistical Natural Concept Generation in Interlingua-based Speech-to-Speech Translation", *Proceedings of Eurospeech*, 2003.

[7] L. Gu, et al, "Forward-Backward Modeling in Statistical Natural Concept Generation for Interlingua-based Speech-to-Speech Translation", *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2003.

[8] A. Berger, et al, "A Maximum Entropy Approach to Natural Language Processing", *Computational Linguistics*, vol.22, no.1, 1996.

[9] D. Magerman. *Natural Language Parsing as Statistical Pattern Recognition*, Ph. D. thesis, Stanford Univ., 1994.

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

[11] A. Ratnaparkhi, "Trainable methods for surface natural language generation", *First Meeting of the North American Chapter of the Association for computational Linguistics (NAACL)*, Seattle, Washington, 2000.

[12] K. Papineni, et al, "Bleu: a Method for Automatic Evaluation of Machine Translation", *ACL 2002*.