

Improving Statistical Natural Concept Generation in Interlingua-based Speech-to-Speech Translation

Liang Gu, Yuqing Gao, and Michael Picheny

IBM T. J. Watson Research Center, Yorktown Heights, NY 10598

ABSTRACT

Natural concept generation is critical to statistical interlingua-based speech translation performance. To improve maximum-entropy-based concept generation, a set of novel features and algorithms are proposed including features enabling model training on parallel corpora, employment of confidence thresholds and multiple sets of features. The concept generation error rate is reduced by 43%-50% in our speech translation corpus within limited domains. Improvements are also achieved in our experiments on speech-to-speech translation.

1. INTRODUCTION

Accurate automatic translation of conversational messages between different languages is critical for speech-to-speech (S2S) translation systems. One common approach is to employ interlingua-based methods [1,2], in which a language-independent representation of intended meanings - interlingua (referred to as *concepts* in this paper) is obtained by parsing the source language, and then used to generate an equivalent expression in the target language. This method was originally proposed to ease language portability by maintaining language-independence during translation. The translated expression in new languages can be generated from language-independent interlingua rather than from one of the source languages. Moreover, the interlingua or concepts can be utilized to enhance translation performance as the speech messages are generated based on concept understanding rather than blind statistical learning or rule-based searching.

While the interlingua-based translation method brings benefits of higher portability and accuracy, two challenges remain open in the design of interlingua-based S2S systems. One challenge is the appropriate design and selection of concepts, which usually depend on the domain in which the translation system is used. 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. However, this challenge is beyond our focus in this paper.

Another challenge is the appropriate generation of the concepts in the target language through a *natural concept generation* (NCG) process. The goal of NCG is not only to generate the correct set of concepts in the target language, but also to produce them in an appropriate order. Compared to the conventional generation of words in non-interlingua speech translation approaches, the NCG process in the interlingua approach has much higher impact on the overall S2S performance. In interlingua-based approaches, the accuracy of concepts in target language is critical to the speech translation quality, particularly in conversational speech translations where in most cases only a few concepts are contained in the

messages to be translated. Therefore, accurate and robust NCG methods are of great importance to S2S translation systems.

Current NCG methods include rule-based approaches and statistical approaches. While the former method is easy to understand and implement, it requires extensive linguistic knowledge and is questionable in terms of scalability and portability between different tasks and domains. Statistical NCG, on the other hand, is trainable and hence has the inherent nature of high scalability and portability, such as the maximum-entropy (ME)-based statistical NCG (ME-NCG) method presented in our previous work [2]. Despite the promising performance of the ME-NCG method, the appropriate design of a generation procedure and the selection of features remain two open challenges in improving the accuracy and robustness of current ME-NCG method and, thereby, the overall S2S systems.

In this paper, we present a set of new algorithms to improve our previously proposed ME-NCG method by addressing the above issues. A new feature set is proposed to train ME models on pre-annotated parallel corpora. A confidence threshold is then introduced to enhance NCG robustness when data sparseness problem exists. A multiple feature selection algorithm is further designed to reduce the generation error rate for unseen concept sequences. Experiments are performed to evaluate the benefits of these new algorithms on both the NCG accuracy and the overall speech translation performance.

2. MAXIMUM-ENTROPY-BASED STATISTICAL NATURAL CONCEPT GENERATION

A. Statistical Interlingua-based S2S Translation

The purpose of this work is to improve the performance of statistical interlingua-based translation via new and superior NCG algorithms. Before we move on to the NCG issues, let us first describe the S2S translation system that we are targeting at.

Figure 1 is a general framework of our previously developed S2S translation system for applications in limited domains. A cascaded scheme of automatic speech recognition (ASR), statistical interlingua-based machine translation and text-to-speech (TTS) synthesis is applied by using existing advanced speech and language processing techniques. While each of these three functional units is crucial to the overall speech

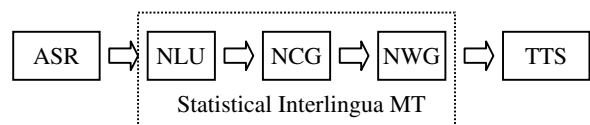


Figure 1 IBM Speech-to-Speech Translation System

translation quality, we are only concerned with the performance of statistical interlingua-based translation here.

The baseline statistical interlingua-based translation further consists of three cascaded functional components: natural language understanding (NLU), natural concept generation (NCG) and natural word generation (NWG). The NLU function is performed via a decision-tree-based statistical semantic parser, which extracts the meaning of the sentence in source language by evaluating a large set of potential trees based on pre-trained statistical models [3]. The NWG process generates words in target language based on the generated structural concepts from NCG and a tag-based word-to-word multilingual dictionary. The word-sense disambiguation problem can be alleviated by using N-gram language models proposed in [4]. 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 source language. The accuracy of the NCG process is critical to the final speech translation quality as any errors of inserted, missing, replaced, or mistakenly ordered concepts may cause severe understanding problems during multilingual speech communications. Therefore, the enhancement of NCG is an essential step towards high-performance S2S translation systems. In this paper, we focus on improving the ME-based statistical NCG method, as explained next.

B. Statistical NCG on Sequence Level

The baseline statistical NCG algorithm was proposed in [2] as an extension from the “NLG2” algorithm described in [5]. 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 NLU parser produces a concept sequence $C = \{c_1, c_2, \dots, c_M\}$ for the source sentence. Let us further assume that a concept sequence $S = \{s_1, s_2, \dots, s_n\}$ 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

$$p(s|c_m, s_n, s_{n-1}) = \frac{\prod_k \alpha_k^{g_k(\bar{f}_k, s, c_m, s_n, s_{n-1})}}{\sum_{s \in V} \prod_k \alpha_k^{g_k(\bar{f}_k, s, c_m, s_n, s_{n-1})}}, \quad (1)$$

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. α_k is a probability weight corresponding to each feature \bar{f}_k . The value of α_k is always positive and is optimized via a maximum-entropy criterion

described in next sub-section. g_k is a binary test function related to \bar{f}_k , which is defined as

$$g_k(\bar{f}_k, s, c_m, s_n, s_{n-1}) = \begin{cases} 1 & \text{if } \bar{f}_k = (s, c_m, s_n, s_{n-1}) \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

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

$$s_{n+1} = \arg \max_{s \in V} \left\{ \prod_{m=1}^M p(s|c_m, s_n, s_{n-1}) \right\}. \quad (3)$$

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

- 1) Set $s_0 = s_{-1} = \text{START}$, where “START” is a pre-defined concept representing the start of the sequence; Set $n = 0$;
Set initial set of generation sequence $S = \emptyset$;
- 2) For each n , generate s_{n+1} according to equation (3) and set $S = \{s_1, \dots, s_{n+1}\}$;
- 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 .

Note that the number of concepts generated in S may be different from the number of concepts in the input sequence in source language, and N in step 4) is the maximum number of concepts that can be generated. In our experiments, $N = 10$.

The above generation algorithm is greedy and sub-optimal as the concepts are generated sequentially. An improved search algorithm via concept lattices is under investigation.

C. Model Training based on Maximum-Entropy Criterion

The probability weights α_k in (1) can be optimized over a training corpus by maximizing the overall logarithmic likelihood, i.e.

$$\alpha_k = \arg \max_{\alpha} \sum_{l=1}^L \sum_{s \in q_l} \sum_m \log [p(s|c_m, s_n, s_{n-1})], \quad (4)$$

where $Q = \{q_l, 1 \leq l \leq L\}$ is the total set of concept sequences and $p(s|c_m, s_n, s_{n-1})$ is the conditional probability defined in equation (1).

The optimization in Equation (4) can be accomplished by the Improved Iterative Scaling algorithm using the maximum-entropy criterion described in [5].

D. Structural Concept Sequence Generation

In the previous two sub-sections, we described the ME-based statistical concept generation on the sequence level. However, a structural concept sequence generation algorithm is required when a concept-based parse tree needs to be generated in the target language. A recursive NCG algorithm is thus proposed as follows:

- 1) Traverse the semantic parse tree in a bottom-up left-to-right breath-first mode;
- 2) For each un-processed concept sequence on a parse tree, generate an optimal concept sequence in target language via the procedure described in 2.B;
- 3) Repeat step 2) until all parser units in the source language are processed.

3. IMPROVING ME-BASED NATURAL CONCEPT GENERATION

A. Feature Selection and Model Training on Parallel Corpora

A remaining problem in current ME-based NCG is to improve selection of features during ME model training and concept generation. In particular, in the feature set defined in [2], 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 augment the feature set in [2] with a new set of features proposed as $(s, c_m, c_{m+1}, s_n, s_{n-1})$, where c_m and c_{m+1} are two sequential concepts in \mathcal{C} . Accordingly, the conditional probability of a concept candidate and the probability weights are modified as

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

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

Different from the method proposed in [4] which obtains features entirely within the target language during model training, the features in (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 $QQ = \{u_l, v_l | 1 \leq l \leq L\}$, where u_l and v_l are the concept sequences in source and target language, respectively. For each feature $(s, c_m, c_{m+1}, s_n, s_{n-1})$ 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 . The proposed feature on annotated parallel

corpora strengthens the link between sequences in source and target languages, and can greatly improve the NCG accuracy (as illustrated in the experiments of the next section), when, of course, the annotated parallel training corpora are available.

B. Confidence Threshold

During the optimal concept generation in equation (3), the concept with highest conditional probability in equation (1) is chosen. Although the parameters of the statistical models are optimized in (4) via ME criterion, the accuracy of statistical generation could be greatly degraded if the data sparseness problem is severe in the training corpus. In practice, this problem is very common when the S2S translation system is ported and designed for a new task or in a new domain.

To reduce the performance degradation caused by data sparseness, a new confidence threshold parameter is introduced in the statistical concept generation procedure. One phenomenon we observed from our English-Chinese S2S experiments is that about 50% of the time the concept sequences tend to keep the original order rather than change to a new set of concepts or the same set of concepts with different order. Based on this prior knowledge, we propose that, instead of generating the concept with highest conditional probability, the generation procedure is controlled by a confidence threshold β upon conditional probability ratio defined as

$$r(s|c_m, c_{m+1}, s_n, s_{n-1}) = \frac{\max_{s \in V} \left\{ \prod_{m=1}^{M-1} p(s|c_m, c_{m+1}, s_n, s_{n-1}) \right\}}{\prod_{m=1}^{M-1} p(c_1|c_m, c_{m+1}, s_n, s_{n-1})}, \quad (7)$$

where c_1 is the first concept in the remaining input concept sequence in source language.

The generation procedure in (3) is modified as

$$s_{n+1} = \begin{cases} s & \text{if } r(s|c_m, c_{m+1}, s_n, s_{n-1}) > \beta \\ c_1 & \text{otherwise} \end{cases}, \quad (8)$$

i.e., when the conditional probability of the best candidate s is below a confidence threshold, the concept order in the source language is kept during generation by selecting s_{n+1} as c_1 .

In our experiments, $1 \leq \beta \leq e^{3.5}$.

C. Multiple Feature Selection

As explained in the previous sections, the selection of features is critical to the success of an ME-based statistical NCG approach. While the new features proposed in sub-section 3.A enable ME model training on annotated parallel corpora and hence increase NCG accuracy, the training data may become sparser in the new larger feature space. To further improve the NCG performance, we propose to use additional sets of features in ME-based concept generation. Multiple sets of features are extracted containing context information in both the source and the target language at different levels. A typical example is to use two sets of features, one as in equation (5), the other as $(s, c_m, c_{m+1}, *, *)$, where $*$ is a symbol representing

all kinds of concepts. These two sets of features can be both employed in ME model training as

$$\alpha_k = \arg \max_{\alpha} \sum_{l=1}^L \sum_{s \in q_l} \sum_{m=1}^{M-1} \left\{ \begin{array}{l} \log \frac{\prod_k \alpha_k^{g_k(\bar{f}_k, s, c_m, c_{m+1}, s_n, s_{n-1})}}{\sum_{s \in V} \prod_k \alpha_k^{g_k(\bar{f}_k, s, c_m, c_{m+1}, s_n, s_{n-1})}} \\ + \log \frac{\prod_k \alpha_k^{g_k(\bar{f}_k, s, c_m, c_{m+1}, *, *)}}{\sum_{s \in V} \prod_k \alpha_k^{g_k(\bar{f}_k, s, c_m, c_{m+1}, *, *)}} \end{array} \right\}. \quad (9)$$

A multiple feature selection is performed during statistical concept generation, where feature $(s, c_m, c_{m+1}, s_n, s_{n-1})$ is first selected as in (5). If $g_k(s, c_m, c_{m+1}, s_n, s_{n-1})$ is zero, the second set of features $(s, c_m, c_{m+1}, *, *)$ is applied with less constraint of context information in the target language.

4. EXPERIMENTS

The performance of our new algorithms in ME-NCG and statistical interlingua-based S2S translation was evaluated on the English-to-Chinese speech translation task within a limited domain of force protection and medical triage. About 6000 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 2000 in each language. 64 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 simply 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 sub-section.

According to the above criterion, about 2500 concept sequences are selected as our experimental corpus. During experimentation, this corpus is randomly partitioned into training corpus containing 80% sequences and test corpus with the remaining 20% sequences. This random process is repeated 100 times and the average performance is recorded. In experiment I, sequences appear in the training corpus are not allowed to appear in the test corpus, which is thus referred as the worst-case test. In experiment II, sequences may exist in both the training corpus and test corpus, which is referred as the normal-case test. We carry out the normal-case test as it simulates the situation in S2S translation where an unseen sentence in test set may contain the known concept sequence in training. 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.

The experimental results on test corpus are shown in Table 1.

ME-NCG Methods	Worst-case Test	Normal-case Test
Mono-gram feature	40.7% / 21.4%	26.7% / 13.9%
Bi-gram feature	33.0% / 17.4%	25.1% / 13.0%
Bi-gram parallel feature	32.3% / 16.7%	15.8% / 8.3%
+ confidence threshold	28.0% / 14.4%	14.3% / 7.4%
+ multiple feature selection	23.9% / 12.3%	13.5% / 7.0%

Table 1. ME-NCG performance (sequence error rate / concept error rate) using different features and generation criteria.

Translation Methods	Baseline	New Algorithms
Text-to-Text	0.514	0.565
Speech-to-Text	0.429	0.448

Table 2. Improvement of Bleu score on 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)

Both sequence error rate and concept error rate decrease consistently from baseline mono-gram feature in (1), through bi-gram feature in [4], to bi-gram parallel feature in (5). Additional improvement is achieved by using confidence threshold and multiple feature selection. Compare with the baseline ME-NCG, the proposed new algorithms reduce the concept error rate by 43% in worst-case test and 50% in normal-case test.

B. Experiments on statistical interlingua-based S2S translation

Experimental results on statistical interlingua-based text-to-text and speech-to-text translation are illustrated in Table 2 based on Bleu score described in [6]. 277 unseen speech sentences are tested. While the improvement is significant, the relative smaller gains of overall S2S performance compared with NCG gains imply the importance of other S2S functional units and the high demand for the algorithmic improvement in all of these units.

5. CONCLUSION

The statistical natural concept generation algorithms presented in this paper attack the problems of feature selection and generation control during maximum-entropy-based model training and concept generation. A new feature set is proposed to train models on pre-annotated parallel corpora. Confidence threshold and multiple feature selection are introduced to enhance generation robustness when data sparseness problem exists. Significant improvements are achieved in both concept sequence generation test and speech translation experiments.

6. REFERENCES

- [1] S. Nirenburg, et al, *Machine Translation: A Knowledge-Based Approach*, Morgan Kaufmann, 1992.
- [2] Y. Gao, et al, "MARS: A statistical semantic parsing and generation based multilingual automatic translation system", to appear in *Machine Translation*.
- [3] D. Magerman. *Natural Language Parsing as Statistical Pattern Recognition*, Ph. D. thesis, Stanford Univ., 1994.
- [4] F. Liu, et al, "Use of statistical n-gram models in natural language generation for machine translation", *ICASSP*, 2003.
- [5] 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.
- [6] K. Papineni, et al, "Bleu: a Method for Automatic Evaluation of Machine Translation", *ACL* 2002.