Use of Maximum Entropy in Natural Word Generation for Statistical Concept-based Speech-to-Speech Translation

Liang Gu and *Yuqing Gao*
IBM T. J. Watson Research Center, Yorktown Heights, NY 10598

ABSTRACT

Our statistical concept-based spoken language translation method consists of three cascaded components: natural language understanding, natural concept generation and natural word generation. In the previous approaches, statistical models are used only in the first two components. In this paper, a novel maximum-entropy-based statistical natural word generation algorithm is proposed that takes into account both the word level and concept level context information in the source and the target language. A recursive generation scheme is further devised to integrate this statistical generation algorithm with the previously proposed maximum-entropy-based natural concept generation algorithm. The translation error rate is reduced by 14%-20% in our speech-to-speech translation experiments.

1. INTRODUCTION

Automatic speech-to-speech machine translation (SSMT) involves several challenging research areas such as conversational spontaneous Automatic Speech Recognition (ASR), Spoken Language Translation (SLT), and Text-To-Speech synthesis (TTS). Among them, SLT is perhaps the most challenging issue, since the casual spontaneous messages to be translated often contain strong disfluencies, imperfect syntax, no punctuations and ASR errors. During the past decade, many approaches have been proposed to improve SLT and substantial progress has been made in both the SLT and the SSMT performance [1-3].

Recently, we proposed and presented a statistical SLT method based on tree-structured semantic representations, or *concepts* [4]. An example of English-to-Chinese translation using this methodology is illustrated in Figure 1. The source English sentence and the corresponding Chinese translation are represented by a set of concepts – {POLITE, ACTION, MEDICINE, TIME}. The conceptual parse trees are derived from natural language understanding (NLU) processing. Note that although the source-language and target-language sentences share the same set of concepts, their tree structures are different from each other because of the well-known distinct nature of these two languages.

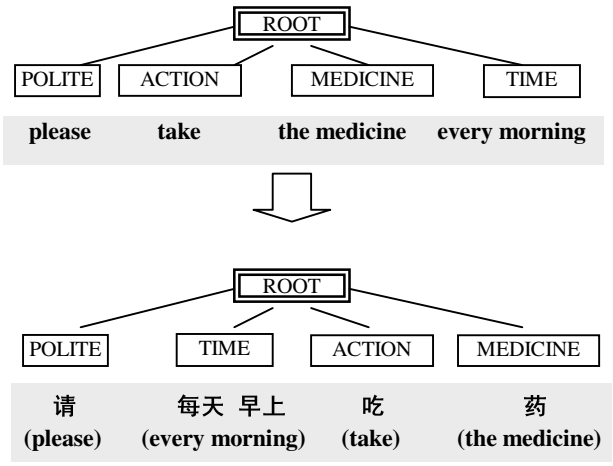

Figure 1. Example of Concept-based English-to-Chinese Translation

Given the parse tree in the source language, a concept-based natural language generation (NLG) approach is proposed tha consists of two closely-related procedures: 1) a natural concept generation (NCG) procedure that transforms the tree structures in the source language into appropriate the tree structures in the target language; 2) a natural word generation (NWG) procedure that translates source-language sentences into corresponding target-language sentences according to the words in the source-language sentences and the structural concepts in both source and target languages.

In our previous work [4], a NWG procedure is proposed that exploits a semantic-information-based multilingual dictionary by using the semantic information derived from a maximum-entropy-based statistical NCG algorithm. Several drawbacks exist in this preliminary NWG approach. The NCG and NWG procedures are not closely interconnected since only the semantic information for each word is utilized during word sequence translation, while the important concept structures in both source and target languages are ignored. Furthermore, the multilingual dictionary is often generic and not optimal for the restricted domains that a SSMT system usually aims at. Moreover, even if the semantic-information-based multilingual dictionary is designed for a specific domain, the NWG process is not trainable and is hence significantly difficult to be ported to another domain when new SSMT applications are needed.

In this paper, we propose a new NWG algorithm using maximum entropy criterion. After the NLU and NCG procedures are performed, a word sequence is generated in the target language according to the input word sequence in the source language as well as the structural concept information in both the source and the target languages. A word generation probability distribution is defined based on the maximum entropy criterion and estimated by maximizing the overall likelihood of the translated sentences in the training corpora. A feature set is designed and applied to utilize both the word and the concept context information. A recursive NWG algorithm is further proposed to integrate the ME-based NWG with the previously proposed ME-based NCG in the SLT process.

2. SYSTEM Overview

A. General Framework of MASTOR SSMT System

The general framework of our MASTOR SSMT system is depicted in Figure 2. The cascaded approach of ASR, MT and

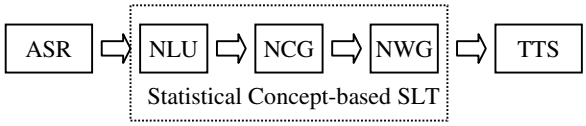

Figure 2 IBM MASTOR Speech-to-Speech Translation System

TTS allows us to deploy the power of the existing advanced speech and language processing techniques, while concentrating on the unique problems in SSMT.

The baseline statistical concept-based SLT further consists of three cascaded functional components: natural language understanding (NLU), natural concept generation (NCG) and natural word generation (NWG). The NLU process extracts the meaning of the sentence in the source language by evaluating a large set of potential parse trees based on pre-trained statistical models. 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 NWG process generates a word sequence in the target language based on the source word sequence and the generated structural concepts from NCG. The cascaded NCG and NWG processes constitute the Natural Language Generation (NLG) process.

B. Natural Language Understanding (NLU)

The NLU process in MASTOR is realized through a decision-tree based statistical semantic parser [5]. Through this method, the statistical parser incorporates semantic information into the parsing model via a decision tree algorithm, and uses a hidden derivational model to maximize the amount of semantic information available. A stack-decoding algorithm is further applied to efficiently search through the immense space of possible parses.

C. Semantic Annotation and Treebank

Our statistical SLT models are trained on treebanks, which are semantically annotated text corpora. The treebank in each language is utilized for the training of NLU, NCG and NWG models for the corresponding language. Current English and Chinese corpora include 10,000 sentences for each language in the domain of emergency medical care. 68 distinct labels and 144 distinct tags are used to capture the semantic information. An example of annotated English and Chinese sentences has been illustrated in Figure 1.

Current annotation process in MASTOR is performed mainly manually. Automatic annotation methods are under investigation and will be used for annotating the ever-expanding multilingual training corpora.

3. NATURAL WORD GENERATION USING MAXIMUM ENTROPY

A. Basic Formulation

Let $\mathbf{W}$ denote the word sequence in the source language and $\mathbf{A}$ denote the word sequence to be translated in the target language. Let $\mathbf{C}$ denote the structural concept set derived from a NLU parser in the source language such as the English concept tree illustrated in Figure 1. Let $\mathbf{S}$ denote the corresponding structural concept set in the target language such as the Chinese concept tree in Figure 1.

If $p(\mathbf{A}|\mathbf{W})$ denotes the probability that the translated sentence consists of word sequence $\mathbf{A}$, given that $\mathbf{W}$ is the input word sequence to be translated, then our proposed natural language generation (NLG) algorithm should select a word sequence $\hat{\mathbf{A}}$ as

$$\begin{aligned}\hat{A} &= \arg\max_A p(A|W) \\ &= \arg\max_A \left\{ \sum_S \sum_C p(C|W) p(S|C,W) p(A|S,C,W) \right\},\end{aligned} \quad (1)$$

where the conditional probabilities $p(C|W)$, $p(S|C,W)$ and $p(A|S,C,W)$ are estimated by the NLU, NCG and NWG procedures, respectively. In our approach, $p(C|W)$ is estimated by a decision-tree based statistical semantic parser as described in section 2.B, and $p(S|C,W)$ is estimated by maximizing the conditional entropy

$$H_{NCG}(p) \equiv - \sum_{(C,S,W)\in X} p(C,W) p(S|C,W) \log p(S|C,W), \quad (2)$$

where X is the training data that consists of semantically-annotated treebanks in the source and the target languages (see the description of annotation in section 2.C).

B. Previous NWG approach

In our previous NWG approach [4], $p(A|S,C,W)$ is approximated and estimated as

$$\begin{aligned}p(A|S,C,W) &\approx p(A|C,W) \\ &= h(A|C,W) = \begin{cases} 1, & if\ (A,W,C) \in D \\ 0, & otherwise \end{cases},\end{aligned} \quad (3)$$

where $h(A|C,W)$ is binary test function and D is a semantic-information-based multilingual dictionary.

Although equation (3) is easy to implement, it is over-simplified so that it cannot represent $p(A|S,C,W)$ adequately and accurately. In addition, due to the computational and storage constraints, only the semantic information for the word to be translated is used during the NWG procedure in our previous approach [4]. While the LM post-processing approach proposed in [6] can recoup some of generation accuracy, the critical correlation information between the target word sequence A and the structural concept sets S and C is still missing.

In this work, we propose a new NWG algorithm to estimate $p(A|S,C,W)$ based on maximum entropy. $p(A|S,C,W)$ is estimated by maximizing the conditional entropy

$$H_{NWG}(p) \equiv -\sum_{(A,C,S,W)\in X} p(S,C,W)p(A|S,C,W)\log p(A|S,C,W) \quad (4)$$

The principle of maximum entropy has been successfully used for statistical machine translation in several previous work [7][8]. Our ME-based NWG approach differs from the approach in [8] in that 1) it uses both word level and structural concept level information during probability estimation; 2) it is closely coupled with the previously proposed ME-based NCG algorithm in our statistical concept-based SLT method. Next, we will describe the ME-based NWG process in greater detail.

C. ME-based NWG

Let us assume $W=\{w_1, w_2, \cdots, w_L\}$ is the input word sequence and $C=\{c_1, c_2, \cdots, c_M\}$ is the structural concept set produced from NLU parser in the source language. Let $S=\{s_1, s_2, \cdots, s_K\}$ denote the corresponding concept sequence generated by NCG in the target language. Let $A=\{a_1, a_2, \cdots, a_N\}$ denote the word sequence to be translated. Using the chain rule of probability theory, $p(A|S,C,W)$ in equation (1) can be formally decomposed as

$$p(A|S,C,W)=\prod_{n=1}^{N} p(a_n|c_1,\ldots,c_M,s_1,\ldots,s_K,w_1,\ldots,w_L,a_1,\ldots,a_{n-1}) \quad (5)$$

where $p(a_n|c_1,\ldots,c_M,s_1,\ldots,s_K,w_1,\ldots,w_L,a_1,\ldots,a_{n-1})$ is the probability that word a_n will be generated if $W=\{w_1,w_2,\cdots,w_L\}$ is given and $C=\{c_1,c_2,\cdots,c_M\}$, $S=\{s_1,s_2,\cdots,s_K\}$ and word sequence $\{a_1,a_2,\cdots,a_{n-1}\}$ were generated previously.

In reality, it is almost impossible to estimate $p(a_n|c_1,\ldots,c_M,s_1,\ldots,s_K,w_1,\ldots,w_L,a_1,\ldots,a_{n-1})$ accurately since $(c_1,\ldots,c_M,s_1,\ldots,s_K,w_1,\ldots,w_L,a_1,\ldots,a_{n-1})$ may only occur for very limited times in the training corpora and hence involves serious data sparseness problem. Alternatively, equation (4) can be approximated as

$$p(A|S,C,W)=\prod_{n=1}^{N} p(a_n|\Phi(c_1,\ldots,c_M,s_1,\ldots,s_K,w_1,\ldots,w_L,a_1,\ldots,a_{n-1})) \quad (6)$$

where $\Phi(c_1,\ldots,c_M,s_1,\ldots,s_K,w_1,\ldots,w_L,a_1,\ldots,a_{n-1})$ represents a portion of $(c_1,\ldots,c_M,s_1,\ldots,s_K,w_1,\ldots,w_L,a_1,\ldots,a_{n-1})$. In this work, we select $\Phi(c_1,\ldots,c_M,s_1,\ldots,s_K,w_1,\ldots,w_L,a_1,\ldots,a_{n-1})$ as

$$\Phi(c_1,\ldots,c_M,s_1,\ldots,s_K,w_1,\ldots,w_L,a_1,\ldots,a_{n-1}) = (c_{n-1},c_n,c_{n+1},s_{n-1},s_{n+1},w_n,a_{n-1}) \quad (7)$$

where c_n is the concept information of the input word w_n, and $(c_{n-1},c_{n+1},s_{n-1},s_{n+1})$ are the concept context in the source and the target languages, respectively.

Under this approximation and according to equation (4), in order to generate the next new word a_n, the conditional probability of a word candidate is defined and computed as

$$p(A|S,C,W)=\prod_{n=1}^{N} p(a_n|c_{n-1},c_n,c_{n+1},s_{n-1},s_{n+1},w_n,a_{n-1}) \quad (8)$$

and

$$p(a|c_{n-1},c_n,c_{n+1},s_{n-1},s_{n+1},w_n,a_{n-1}) = \frac{\prod_k \alpha_k^{g(\bar{f}_k,c_{n-1},c_n,c_{n+1},s_{n-1},s_{n+1},w_n,a_{n-1},a)}}{\sum_{a'\in V}\prod_k \alpha_k^{g(\bar{f}_k,c_{n-1},c_n,c_{n+1},s_{n-1},s_{n+1},w_n,a_{n-1},a')}} \quad (9)$$

where V is the set of all possible words that can be generated.

$\bar{f}_k=(c_{-1}^k,c_0^k,c_{+1}^k,s_{-1}^k,s_{+1}^k,w_0^k,a_{-1}^k,a_0^k)$ is the k-th 8-dimensional feature consisting of uni-gram $\{w_0^k\}$ in the source word sequence, tri-gram $\{c_{-1}^k,c_0^k,c_{+1}^k\}$ in the source concept set, bi-gram $\{s_{+1}^k,s_{-1}^k\}$ in the target concept set, and bi-gram $\{a_{-1}^k,a_0^k\}$ in the target word sequence.

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

$$\alpha_k = \arg\max_{\alpha} \sum_{A\in Q}\sum_{n} \log[p(a_n|c_{n-1},c_n,c_{n+1},s_{n-1},s_{n+1},w_n,a_{n-1})] \quad (10)$$

where $A=(a_1,a_2,\ldots,a_N)$, and $Q=\{A_j, 1\le j\le J\}$ is the total set of word sequences in the target lanuage. The optimization process can be accomplished via the Improved Iterative Scaling algorithm using the maximum entropy criterion described in [9].

$g(\cdot)$ is a binary test function defined as

$$g(\bar{f}_k,c_{n-1},c_n,c_{n+1},s_{n-1},s_{n+1},w_n,a_{n-1},a) = \begin{cases} 1 & if\ \bar{f}_k=(c_{n-1},c_n,c_{n+1},s_{n-1},s_{n+1},w_n,a_{n-1},a) \\ 0 & otherwise \end{cases} \quad (11)$$

Using (9), (10) and (11), a_{n+1} is generated by selecting the word candidate with highest probability, i.e.,

$$a_n = \arg\max_{a\in V}\{p(a|c_{n-1},c_n,c_{n+1},s_{n-1},s_{n+1},w_n,a_{n-1})\} \quad (12)$$

D. Combination of ME-based NWG and ME-based NCG in NLG procedure

The above proposed ME-based NWG algorithm can be and should be combined with our previously proposed ME-based NCG algorithm in a NLG procedure to achieve optimal natural language generation performance. A recursive generation scheme is proposed as follows:

1) Traverse each un-processed concept sequence in the semantic parse tree in a top-down left-to-right manner;

2) For each un-processed concept sequence in the parse tree, generate an optimal concept sequence in the target language using the NCG procedure proposed in [4]; after each concept sequence is processed, mark the root-node of this sequence as visited;

3) Repeat step 2) until all parser branches in the source language are processed;

4) Traverse each un-processed word sequence in the generated semantic parse tree in a bottom-up left-to-right manner;

5) For each un-translated word sequence in the parse tree, generate an optimal word sequence in the target language using the NWG procedure described in equation (8) and (12); after word sequence is processed, mark the root-node of this sequence as visited;

6) Repeat step 2) until all the leaf nodes in the target-language conceptual parse tree are visited and the contained word sequences are translated.

4. EXPERIMENTS

The performance of our proposed NWG algorithm and the corresponding 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. During experimentation, the English-to-Chinese corpus is randomly partitioned into training set containing 80% of the sentences and test set with the remaining 20%. This random process is repeated 50 times and the average performance on the test set is recorded.

In the first experiment, various NWG algorithms were implemented, tested and compared on the 10000-sentence English-to-Chinese corpus. The average translation accuracy on the test set is shown in Table 1 in both the character error rate (CER) and the BLEU score. The BLEU score measures MT performance by evaluating n-gram accuracy with a brevity penalty [10]. As we can see, the proposed ME-based NWG algorithm reduced CER substantially from 45.44% with baseline NWG using semantic-information-based dictionaries, and 42.17% by adding LM-based post-processing, to 36.42%, which represents a 20% and 14% error rate reduction, respectively. The corresponding BLEU score was improved from 0.482 and 0.519 to 0.613. This clearly demonstrates the advantage of the proposed ME-based NWG compared to our previous proposed methods, mostly thanks to the powerful maximum-entropy-based statistical models and the rich feature set that contains both the word level and concept level context information.

In the second experiment, statistical concept-based text-to-text and speech-to-text translation performance is measured on 277 unseen spoken sentences using the baseline and new NWG algorithms. Table 2 shows consistent improvement was achieved in both text-to-text and speech-to-text translation experiments.

NWG Methods	Bleu score	Character Error Rate
Baseline NWG using semantic-information-based dictionaries (as proposed in [4])	0.482	45.44 %
+ LM Post-processing (as proposed in [6])	0.519	42.17 %
ME-based NWG with feature set in equation (11)	0.613	36.42 %

Table 1. Comparison of English-to-Chinese SLT using various NWG methods (BLEU score ranges from 0.0 to 1.0 with 1.0 indicating the best translation quality)

Input/Output	with Baseline NWG	with ME-based NWG
Text-to-Text	0.536	0.649
Speech-to-Text	0.437	0.505

Table 2. Improvement of Bleu score in SSMT by using the new NWG algorithm

5. CONCLUSION

Natural word generation is a critical functional component in our statistical concept-based speech-to-speech translation. In this paper, a new maximum-entropy-based statistical word generation algorithm is proposed that exploits both the word level and structural concept level context information during the training and decoding of maximum entropy based generation models. It is then combined with our previous proposed maximum-entropy-based natural concept generation models to achieve high performance spoken language translation. Significant improvements of translation accuracy are achieved in both text-to-speech and speech-to-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. Ney, et al, "Algorithms for Statistical Translation of Spoken Language", *IEEE Trans. on Speech and Audio Processing*, vol.8, no.1, January 2002.

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

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

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

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

[8] F. J. Och and H. Ney, "Discriminative Training and Maximum Entropy Models for Statistical Machine Translation," *ACL* 2002.

[9] 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.

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