Proceedings of the Workshop on Speech-to-Speech Translation:
Algorithms and Systems, Philadelphia, July 2002, pp. 61-68.
Association for Computational Linguistics.

Interactive Chinese-to-English Speech Translation Based on Dialogue Management

Chengqing Zong, Bo Xu, and Taiyi Huang
National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences
P. O. Box 2728, Beijing 100080, China
{cqzong, xubo, huang}@nlpr.ia.ac.cn

Abstract

In this paper, we propose a novel paradigm for the Chinese-to-English speech-to-speech (S2S) translation, which is interactive under the guidance of dialogue management. In this approach, the input utterance is first pre-processed and then serially translated by the template-based translator and the inter-lingua based translator. The dialogue management mechanism (DMM) is employed to supervise the interactive analysis for disambiguation of the input. The interaction is led by the system, so the system always acts on its own initiative in the interactive procedure. In this approach, the complicated semantic analysis is not involved.

1 Introduction

Over the past decade, many approaches to S2S translation have been proposed. Unfortunately, the S2S translation systems still suffer from the poor performance, even though the application domains are restricted. The common questions are: what translation strategies are necessary? What do the problems exist in the current S2S systems? And what performance of a system is acceptable?

Based on the questions, we have analyzed the current approaches to machine translation (MT) and investigated some experimental systems and the user's requirements. A novel paradigm for the Chinese-to-English S2S translation has been proposed, which is interactive under the guidance of DMM. In this approach, the input utterance is first pre-processed and serially translated by the template-based translator and the inter-lingua based translator. If the two translators are failed to translate the input, the dialogue management mechanism is brought into play to supervise the interactive analysis for disambiguation of the input. The interaction is led by the system, so the system always acts on its own initiative in the interactive procedure. In this approach, the complicated semantic analysis is not involved.

Remainder of the paper presents our motivations and the proposal scheme in detail. Section 2 gives analysis on the current MT approaches and the user's requirements. Section 3 describes in detail our approach to Chinese-to-English S2S translation. Section 4 draws conclusions and presents the future work.

2 Analysis on MT approaches and S2S translation systems

2.1 Analysis on MT approaches

In the past decades, many MT approaches have been proposed. We roughly divided the current approaches into two types, which are respectively named as the mainstream approaches and the non-mainstream approaches. The mainstream approaches include four basic methods: the analysis-based method, the example-based method, the template-based method and also the statistical method as well. The analysis-based method here includes the rule-based method, the inter-lingual method, or even the knowledge-based method. In the recent years, the approach based on multi-engine has been practiced in many systems (Lavie, 1999; Wahlster, 2000; Zong, 2000a). However, the engines employed in these experimental systems are mainly based on the four mainstream methods. The strong points and the weak points of the four methods have been analyzed in many works (Zong, 1999; Ren, 1999; Zhao, 2000).

The non-mainstream approach here refers to any other methods exclusive of the four methods mentioned above. To improve the performance of MT systems, especially to cope with the specific problems in S2S translation, many schemes have been proposed. Ren (1999) proposed a super-function based MT method, which tries to address the MT users' requests and translates the input without thorough syntactic and semantic analysis. The super-function based MT system is fast, inexpensive, easy to control and easy to update. However, the fluency and the correctness of the translation results are usual not high. Moreover, to extract the practical super-functions from the corpus is also a hard work. Yamamoto et al. (2001) proposed a paradigm named Sandglass. In the sandglass system, the input utterances from a speech recognizer are paraphrased firstly, and the paraphrased text is passed to the transfer controller. The task of the paraphrasing module for the source language is to deal with noisy inputs from the speech recognizer and provides different expressions of the input. An obvious question about the Sandglass is why the system would rather rewrite the input than to translate it directly? Zong et al. (2000b) proposed an MT method based on the simple expression. In the method the keywords in an input utterances are spotted out firstly and the dependence relation among the keywords are analyzed. Then, the translation module searches the examples in the knowledge base according to the keywords and their dependence relation. If an example is matched with the conditions, the target language expression of the example is sent out as the translation result of the input. When the input is not very long, and the domain and the type of the input are restricted, the method is very practical. However, to develop the knowledge base with dependence relation of keywords and to match an input with all examples in the knowledge base are sometimes difficult. Wakita et al. (1997) proposed a robust translation method which locally extracts only reliable parts, i.e., those within the semantic distance threshold and over some word length. This technique, however, does not split input into units globally, or sometimes does not output any translation result (Furuse et al, 1998). In addition, the method closely lies on the semantic computation, and sometimes it is hard to compute the semantic distance for the spoken utterances.

In summary, both mainstream MT methods and non-mainstream methods have been practiced in many experimental S2S translation systems. However, all methods mentioned above are unilateral and based on user's own wishful thinking. The system is passive and blind in some extent. The task that machine translates is imposed by human, and some problems are also brought by the speaker, e.g., the topics are changed casually, or the ill-formed expressions are uttered. In these cases, it is unreasonable to expect the system to get the correct translation results, but not to give the system any rights to ask the speaker about his or her intention or some ambiguous words. In fact, if we examine the procedures that human interpreters use, we can see that the translation is usually interactive. When an interpreter is unable to directly translate an utterance due to an ill-formed expression or something even worse, the interpreter may have to ask the speaker to repeat or explain his / her words. Based on the ideas, the interactive paradigms for S2S translation have been proposed (Blanchon, 1996; Waibel, 1996; Seligman, 1997; Seligman, 2000; Ren, 2000). Seligman (2000) proposed a ' quick and dirty' or 'low road' scheme, in which he suggested that, by stressing interactive disambiguation, practically usable speech translation systems may be constructable in the near term. In addition, two interactive MT demos were shown respectively in 1997 and 1998 (Seligman, 2000). However, all the proposed interactive schemes and the demos put the emphasis on the interface between speech recognition (SR) and analysis. The interface can be supplied entirely by the user, who can correct SR results before passing them to translation components. That means the translation system is still passive. Actually, as we know that the parsing results and the translation results are not certainly correct even though the input is completely correct, but some noisy words usually have not any influence whether they are correct or not. In this sense, the user should know what the system needs? And what brought the system ambiguity? This means, the system has rights and obligations to tell the user what the system want to know. In another words, the system necessitates a DMM to guide the interaction between the system and user, and sometimes the system should play the leading role.

2.2 Analysis on user's requirements

Although much progress in SR and spoken language parsing has been made, there is still a long way to reach the final and ideal goal that the translation results are complete correct. In this situation, let's think does a user always need the complete correct translation results? Please see the following three examples:

(1) *Input*: 喔，那个…… 这样吧，就给我预订一个单人间吧，对，单人间。(*Oh, that ... well, please reserve a single room for me, sure, a single room.*)

In the input, there are many redundant words, such as, 喔(Oh)，那个(that), 这样吧(well) and so on. If all words in the input are translated, the translation result is verbose and wordy. In fact, in the input only three keywords are useful, which are: 预订(reserve), 一个(one), and 单人间(single room) as well. The preposition phrase '给我(for me)' is not obligatory. Even the word '预订' is also not obligatory.

(2) *Input*: 是 向 个 里 拉 饭店 吗 ? (*Is this ... Xiang Ge Li La... Hotel?*)

In the example, the four characters with underline are originally a hotel name '香格里拉'(Shangri-la), but they are wrong transliterated and separated due to the absence of the word in the SR dictionary. In this case, it is impossible to correctly parse the input without user's help.

(3) *Input*: 有没有 去 黄山 的 问 有 路线? (*Is there any ... ask ... have... route to Huangshan mountain?*)

The input is a result of the SR component. Obviously, in the input two characters with stressing dots are wrong recognized from the original word '旅游 (tour)'. In this case, if all words are translated, the results will be inconceivable. On the contrary, the result is quite understandable if the two characters with stressing dots are omitted or ignored.

The example (1) shows that if the input is recognized completely correct, the parsing result is still probably wrong due to the ill-formed expression of the input. The example (2) means that it is impossible to correctly parse the input due to the unknown word and its incorrect recognition. The example (3) shows that even though the expression is formal and there is not any unknown word in the input, the result of SR is still probably wrong. The parser is impossible to correctly analyze the wrong SR result.

From the three examples we can easily get the following standpoints: a) the user expects his or her intentions to be translated rather than his (her) all words. The keywords and their dependence relations are the main objects to hold the user's intentions. b) For the translation component, it is not indispensable to correct all mistakes in the input from the SR component. c) If the parser is failed to parse the input, and the system only translates the keywords, the translation results may be still understandable and acceptable.

3 Interactive translation based on dialogue management

3.1 Overview of the paradigm

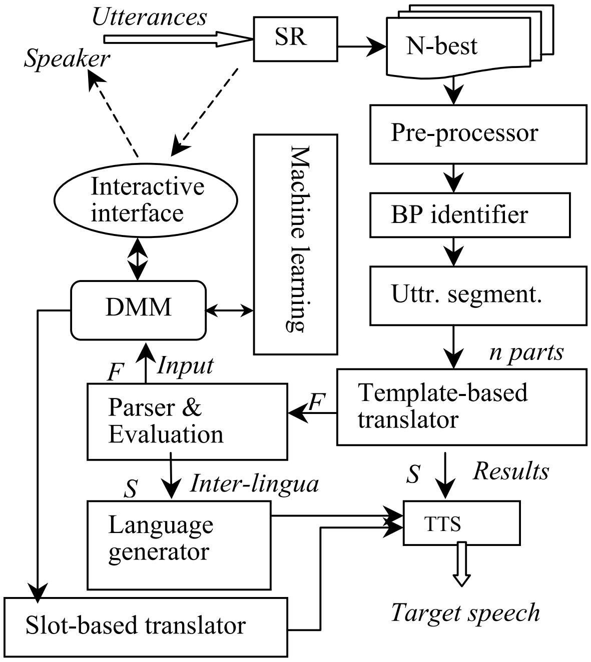

Figure 1. The paradigm of interactive translation

Based on the analysis on MT approaches and the user's requirements, we propose an interactive paradigm for the S2S translation, which is based on the template-based translation, inter-lingual translation and the DMM based translation as well. The paradigm is shown as Figure 1.

Where, the letter *S* beside the line with arrow means that the results of the former module are successful, and the letter *F* means the results are failure.

According to the paradigm, an input from the SR component is probably processed and translated by the following four steps. First, the input is pre-processed. Some noisy words are recognized, some repeated words are deleted, and the numbers are processed (Zong, 2000a). Then the base phrases (BP) in the input are identified, which include noun phrase (NP) and verb phrase (VP) mainly. And also, if the input is a long utterance containing several simple sentences or some fixed expressions, the input is possibly segmented into *n* parts. *n* is an integer, and $n \geq 1$. Second, each part of the input is passed to the template-based translator. If the input part is matched with a translation template, the translation result is sent to the text-to-speech (TTS) synthesizer directly. Otherwise, the input part will be passed to the inter-lingual translator. Third, in the inter-lingual translator, the input is parsed and the parsing results are evaluated. If the evaluation score is bigger than the given threshold value, the parsing results will be mapped into the inter-lingua, and the translation result will be generated by the inter-lingua based target language generator. Otherwise, the system performs the fourth step. Fourth, DMM works to supervise the interaction for disambiguation of the input. In the interaction, the user is asked to answer some questions regarding to the input part. The system will fill the slots according to the question-answers. The slots are designed to express the user's intentions in the input. The system directly generates the translation result according to the slots. So, the translation in the fourth step is named as slot-based translation.

Where, the template-based translator employs the forward maximum match algorithm (Zong, 2000c). The inter-lingua uses the interchangeable format (IF) developed by C-STAR (Consortium for Speech Translation Advanced Research). The parser oriented to IF is realized on the basis of HMM spoken language understanding model. In the experimental system we use the tri-gram to compute the probability of the sequence of semantic units (Xie, 2002). The IF-based language generator employs a task-oriented micro-planner and a general surface realizer. The target language is generated by the combination of template method and generation technology (Wu, 2000). The generic DMM has been proposed by (Xu, 2001), which combines both interaction patterns and task structure. The machine learning module is taking charge of recording the dialogue patterns, topics and modifying the dialogue history, and so on. This module is still under construction.

3.2 Utterance segmentation

In an S2S translation system, how to split the long input utterances is one of the key problems, because an input is often uttered by the spontaneous speech, and there is not any special mark to indicate which word is the beginning or the end of each simple sentence inside the utterance. In our system an input Chinese utterance is first split by the SR component according to the acoustic features, including the prosodic cues and pause etc. Suppose an input utterance has been transcribed by SR and separated into *k* parts P_1, P_2, … P_k (*k* is an integer, and $k \geq 1$.). Each part P_i ($i \in [1 .. k]$) is possibly further segmented into *m* (*m* is an integer and $m \geqslant 1$) units U_1, U_2, …, U_m by the segmentation module based on the linguistic analysis (SBLA). Where, all P_i ($i \in [1 .. k]$) and U_j ($j \in [1 .. m]$) are called as the split units in our system. A split unit is one of the following expressions:

- A single word.
- A fixed expression, such as a greeting phrase in Chinese.
- A simple sentence.
- A clause indicated by some special conjunctions. For example, an input similar with the pattern "因为(because) … , 所以(therefore) … " will be separated into two parts "因为(because)…" and "所以(therefore) … ".

Each P_i ($i \in [1 .. n]$) is analyzed and segmented by SBLA through the following three steps: splitting on the shallow level, splitting on the middle level, and splitting on the deep level. This means if a string *S* is separated into *n* parts by using the method on the shallow level, each part will possibly be further segmented by the method on the middle level, and so on.

3.3 Slot-based translation with DMM

The slot-based translation with DMM is built on the following viewpoints and hypothesis: 1) there are some noisy words or ambiguous words in the results from SR component, but the keywords are recognized correctly; 2) the user's intentions lie on the keywords and their dependence relations; and 3) the translation results based on the keywords are understandable and reflect the main intentions of the user. The slot-based translation under the guidance of DMM is performed as the following steps:

i) *Re-analyze the original input string, spot out the keywords, and also do the analysis on the dependence relation of the keywords.*

ii) *Interact with the user, make decision about the keywords and their dependence relation, and fill the slots for the translation.*

iii) *Generate the translation results according to the slots.*

iv) *DMM writes down the keywords and their dependence relations and modifies the dialogue history.*

3.3.1 Keywords spotting and dependence analysis

According to the evaluation score, if the parsing result of an input part is too worse, the parsing is treated as failure, and all analysis results, including base phrases, are ignored. The system will spot out the keywords from the original input and analyze the dependence relation among the keywords. Please note that the dependence relation of the keywords in this component is used for seizing the user's intentions and generating the translation results. It is different with the function in the simple expression based translation (Zong, 2000b).

In a specific domain, it is easy to define some keywords according to the statistical results of the collected corpus. In our system, a word is treated as the keyword if the following two conditions are met:

✧ The part-of-speech (POS) of the word is one of the following three POSs: noun (N), verb (V), and adjective (A), and the word occurs with high probability in the specific domain.

✧ The word is a number or a time word.

In our method, the verb keyword is always treated as the center when the dependence relations are analyzed. The dependence relations between the verb keyword and the noun keywords are defined as four types: (1) agent, (2) direct object, (3) indirect object, and (4) the pivot word as well. The agent is usually located at the left of the verb keyword. In general, the direct object, indirect object, and the pivot word all occur at the right of the verb keyword. The pronoun is treated as the noun. Other content words are treated as the modification words of the keywords. The search direction and the position relation may be shown as the following Figure 2. Where, W_i means a common word, and KW_i means a keyword..

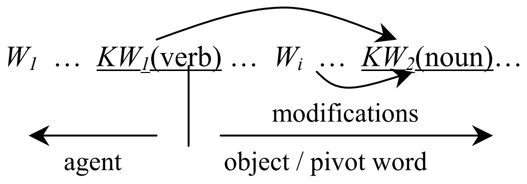

Figure 2. Keywords and their relations

According to the characteristics of the Chinese verbs, there are five cases respectively:

✧ There is no object after the verb;

✧ There is one object only;

✧ There are two objects. One is the direct object and another one is the indirect object.

✧ The object is a clause.

✧ After the verb keyword, the first noun is the object (pivot word) and acts as the role of agent of another followed verb.

In the keyword dictionary, each verb is tagged with its all possible POSs and relative features. The DMM asks the user questions according to the features of a specific verb, its context, and the slots.

3.3.2 Interaction and slot filling

In the DMM module, a frame is designed to express the user's intentions, which consists of a series of slots as follows.

Frame: ACTION: Keywords (verb)
TENSE: {Present/Past/…}
EXP. TYPE: {Interrogative/…}
AGENT: noun;
OBJECT1: noun;
OBJECT2: noun;
QUANTITY1: numeral;
UNIT1: classifier;
QUANTITY2: numeral;
UNIT2: classifier;
TIME: numeral & classifier;
HOW: adjective;

Figure 3. Frame of slots

Where, QUANTITY1 and UNIT1 modify the agent, QUANTITY2 and UNIT2 modify the OBJECT1 or OBJECT2. Because the keywords have been spotted out and their dependence relations have been analyzed, the DMM asks the user according to the analysis results and the concrete context. Please see the following example.

Input: 是向个里拉饭店吗？(Is this … Xiang Ge Li La … Hotel?)

Two keywords, '是(be)' and '饭店(hotel)', are spotted, and the word '饭店' is recognized as the object of the verb '是'. i.e., ACTION=是; TENSE=Present; EXP. TYPE= Interrogative; OBJECT1=饭店; and other slots are empty. However, there are four noisy characters between the two keywords. The DMM will ask the user by using the question pattern: '什么(what) X ?'. The variable X is just replaced with the keyword '饭店'. The user needs to answer the hotel name. Because the SR module still does not recognize the speech of the word '香格里拉(Shangri-la)', the DMM is unable to parse the user's answer. The following dialogue will be done:

System: '香格里拉'一词是形容词还是名词？(Is the word '香格里拉' an adjective or a noun?)

User: (是)名词(It is a noun.)

System: 是饭店的名字吗？(Is it the hotel name?)

User: 是的(yes).

System: 请输入这个词的英语名字(Please input the English name of the word '香格里拉').

The DMM will append the word '香格里拉' both into the SR dictionary and translation dictionary and treat it as the attribute of the keyword '饭店'. The input is finally translated by using the template 'Is this the X ?'.

3.3.3 Generation based on slots

After the interaction, the translation result will be generated based on the templates that are consisted of the slots. For example, if AGENT and ACTION are filled, the EXP. TYPE = Statement, and other slots are empty. The generation template is: !AGENT !ACTION. Where, !AGENT means the English word corresponding to the Chinese word in the AGENT slot. !ACTION is the English word corresponding to the Chinese word in the ACTION slot. However, the morphology of the verb will be changed according to the agent.

From the frame of slots we can see that the frame can only express the analysis results of simple sentence. So, the translation result is always expressed by the simple sentence. If the subject or the object of a Chinese input is a clause, the input will be translated into two or more simple English sentences. For instance,

Input: 我预订两个单人间需要多少钱？(How much does it cost if I reserve two single rooms?)

The input will be mapped into two frames. In the first frame, AGENT=我; ACTION=预订; EXP. TYPE=Statement; QUANTITY2=两; UNIT2=个; OBJECT1=单人间. In the second frame, ACTION=需要; EXP. TYPE= Interrogative; QUANTITY1=多少; OBJECT1=钱. Therefore, the input is separately translated into two simple English sentences: 'I reserve two single rooms.', and 'How much does it cost?'. Obviously, in the specific context, the results are completely understandable and acceptable.

4 Conclusion

This paper describes a new paradigm for S2S translation system, which is based on DMM. According to the description we can see that the paradigm is of the following features:

(1) The S2S translation is realized in the combination of direct translation engines and the interaction led by DMM. The interaction is not always brought

into the role, and it only works when the former translation engines work failed.

(2) The interaction is impersonative, target-oriented, and led by the system, not blind. The user does not need to correct all of the errors in the results of SR. He or she only needs to concern what the system asks.

(3) The system can always give the results for an input speech despite of the ill-formed expressions and the worse recognition results.

Although the whole experimental system is under construction, some preliminary results have been gained. Zong (2000c) reported the performance of the template-based translator; Xie (2002) reported the results of the robust parser for the Chinese spoken language; Xu (2001) presented the results of dialogue model; and so on. The results have made us confident to develop the practical S2S translation system based on the dialogue management. However, we are facing much hard work that involve the following aspects at least:

- Develop the reasonable strategies and standards to evaluate the parsing results;
- Design the effective templates to ask the user questions according the keywords and the concrete context;
- Define the practical templates to generate the translation results;
- Build the machine learning mechanism to enrich the knowledge base of the system.

References

Blanchon, H. 1996. A Customizable Interactive Disambiguation Methodology and Two Implementations to Disambiguate French and English Input. In *Proceedings of MIDDIM-96 (International Seminar on Multimodal Interactive Disambiguation)*, Col de Porte, Fance.

Furuse, O., Satsuo Yamada and Kazuhide Yamamoto. 1998. Splitting Long or Ill-formed Input for Robust Spoken-language Translation. In *Proceeding of COLING-ACL,* Canada. Vol. I, pp. 421-427.

Lavie, A., Lori Levin et al. 1999. The JANUS-III Translation System: Speech-to- Speech Translation in Multiple Domains. In *Proceedings of C-STAR II Workshop*, Schwetzingen of Germany, 24 Sept., 1999.

Ren, F., Shigang Li. 2000. Dialogue Machine Translation Based upon Parallel Translation Engines and Face Image Processing. In *Journal of INFORMATION*, Vol.3, No.4, pp.521-531.

Ren, F. 1999. Super-function Based Machine Translation, in *Communications of COLIPS*, 9(1): 83-100.

Seligman, M. 1997. Interactive Real-time Translation via the Internet. In *Working Notes, Natural Language Processing for the World Wide Web*. AAAI-97 Spring Symposium, Stanford University. March 24-26, 1997.

Seligman, M. 2000. Nine Issues in Speech Translation. In *Machine Translation.* 15: 149-185.

Waibel, A. 1996. Interactive Translation of Conversational Speech. In *Proceedings of ATR International Workshop on Speech Translation.* pp. 1~17.

Wahlster, W. 2000. Mobile Speech-to-Speech Translation of Spontaneous Dialogs: An Overview of the Final Verbmobil System. In *Verbmobil: Foundations of Speech-to-Speech Translation*. Springer Press. pp. 3-21.

Wakita, Y., Jun Kawai, Hitoshi Iida. 1997. Correct Parts Extraction from Speech Recognition Results Using Semantic Distance Calculation, and Its Application to Speech Translation. In *Proceedings of a Workshop Sponsored by the ACL and by the European Network in Language and Speech (ELSNET)*. pp. 24-29.

Wu, H., Taiyi Huang, Chengqing Zong, and Bo Xu. 2000. Chinese Generation in a Spoken Dialogue Translation System. In *Proceedings of COLING*. pp. 1141-1145.

Xie, G., Chengqing Zong, and Bo, Xu. 2002. Chinese Spoken Language Analyzing Based on Combination of Statistical and Rule Methods. Submitted to the *International Conference on Spoken Language Processing (ICSLP-2002).*

Xu, W., Taiyi Huang, and Bo Xu. Towards a Generic Dialogue Model for Information-seeking Dialogues. In *Proceedings of the National Conference on Man-Machine Speech Communications (NCMMSC6).* Shenzhen, China. pp. 125-130.

Yamamoto, K., Satoshi Shirai, Masashi Sakamoto, and Yujie Zhang. 2001. Sandglass: Twin Paraphrasing Spoken Language Translation. In *Proceedings of the 19th International Conference on Computer*

Processing of Oriental Languages (ICCPOL- 2001). pp. 154-159.

Zhao, T. et al. 2000. The Principle of Machine Translation (in Chinese). *Press of Harbin Institute of Technology*.

Zong, C., Taiyi Huang and Bo XU. 1999. Technical Analysis on Automatic Spoken Language Translation Systems (in Chinese). In *Journal of Chinese Information Processing*, 13(2):55-65.

Zong, C., Taiyi Huang and Bo Xu. 2000a. Design and Implementation of a Chinese-to-English Spoken Language Translation System. In *Proceedings of the International Symposium of Chinese Spoken Language Processing (ISCSLP-2000)*, Beijing, China. pp. 367-370.

Zong, C., Yumi Wakita, Bo Xu, Kenji Matsui and Zhenbiao Chen. 2000b. Japanese-to-Chinese Spoken Language Translation Based on the Simple Expression. In *Proceedings of International Conference on Spoken Language Processing (ICSLP-2000)*. Beijing, China. pp. 418-421.

Zong, C., Taiyi Huang and Bo Xu. 2000c. An Improved Template-based Approach to Spoken Language Translation. In *Proceedings of International Conference on Spoken Language Processing (ICSLP-2000)*. Beijing, China. pp. 440-443.