Chinese Utterance Segmentation in Spoken Language Translation

Chengqing Zong[1] and Fuji Ren[2]

[1] National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences
Beijing 100080, China
cqzong@nlpr.ia.ac.cn
[2] Department of Information Science and Intelligent Systems,
Faculty of Engineering, The University of Tokushima
2-1, Minamijosanjima, Tokushima, 770-8506, Japan
ren@is.tokushima-u.ac.jp

Abstract. This paper presents an approach to segmenting Chinese utterances for a spoken language translation (SLT) system in which Chinese speech is the source input. We propose this approach as a supplement to the function of sentence boundary detection in speech recognition, in order to identify the boundaries of simple sentences and fixed expressions within the speech recognition results of a Chinese input utterance. In this approach, the plausible boundaries of split units are determined using several methods, including keyword detection, pattern matching, and syntactic analysis. Preliminary experimental results have shown that this approach is helpful in improving the performance of SLT systems.

1 Introduction

In spoken language translation (SLT) systems, an input utterance often includes several simple sentences or relatively independent fixed expressions. However, unlike in written language, there are no special marks to indicate which word is the beginning or the end of a simple sentence. Although some boundaries may be detected by the system's speech recognition module through the analysis of acoustic features, some boundaries may still remain hidden in an utterance. For example, a Chinese speaker may pronounce an utterance as follows: 我来确认一下，您是要带浴缸的单人间，预算在一晚一百美元左右，最好是在闹市区，是吗？("Let me confirm. You would like to reserve a single room with a bath. The budget is about one hundred dollars per night. You prefer a downtown location. Is that right?"). In this utterance, there are four simple sentences and one fixed expression, the confirmation question. The potential difficulty of understanding the utterance without punctuation will be easily imagined. To make matters worse, the speech recognition result often contains incorrectly recognized words and noise words. Thus it is clearly quite

important in SLT systems to split input utterances into simple units in order to facilitate the job of the translation engine.

To cope with the problem of boundary detection, many approaches have been proposed over the last decade. Some of these approaches detect boundaries by analyzing the acoustic features of the input utterance, such as its energy contour, the speaking rate, and the fundamental frequency F_0 (Swerts 1997, Wightman 1994). It is true that some of the approaches take into account the linguistic content of the input utterance (Batliner 1996, Stolcke 1996) to some degree. For instance, automatic detection of semantic boundaries based on lexical knowledge as well as acoustic processing has been proposed (Cettolo 1998). However, we believe that none of these approaches have applied sufficient linguistic analysis for reliable sentence boundary detection in speech recognition. We would argue that linguistic analysis on multiple levels, including lexical analysis, syntactic analysis, and semantic analysis, is indispensable. Therefore, we propose a new approach based on linguistic analysis, in order to supplement or enhance this function.

The remainder of this paper will give emphasis on our methods of linguistic analysis in approaching Chinese utterance segmentation. In Section 2, some related work on utterance segmentation is briefly reviewed, and our motivations are presented. Section 3 describes in detail our methods based on multi-level linguistic analysis. Experimental results are presented in Section 4. Finally, Section 5 gives our conclusion.

2 Related Work and Our Motivations

2.1 Related Work

Stolcke et. al. (1998, 1996) proposed an approach to detection of sentence boundaries and disfluency locations in speech transcribed by an automatic recognizer, based on a combination of prosodic cues, modelled by decision trees, and word-based event N-gram language models. In Stolcke's approach, syntactic and semantic analysis were not involved. Ramasway (1998) introduced a trainable system that can automatically identify command boundaries in a conversational natural user interface. Ramasway's system employs the maximum entropy identification model, trained using data in which all of the correct command boundaries have been marked. The linguistic features employed in this method include only words and phrases and their positions relative to the potential command boundaries. However, this method is impractical for segmenting input utterances for an SLT system, since sentences in such systems are generally considerably longer than the commands used in dialogue systems. Cettolo et. al. (1998) used lexical knowledge in his approach to automatic detection of semantic boundaries, but his approach still treats acoustic knowledge as the main basis for detecting semantic boundaries. Kawahara (1996) proposed a novel framework for robust speech understanding, based on a strategy of detection and verification. In this method (Kawahara 1996), the anti-subword model is used, and a

key-phrase network is used as the detection unit. Linguistic analysis is performed on a shallow level.

Batliner (1996) proposed a syntactic-prosodic labeling scheme in which two main types of boundaries and certain other special boundaries are labeled for a large VERBMOBIL spontaneous speech corpus. The method only aims at segmentation of these special boundaries. Furuse (1998) proposed an input-splitting method for translating spoken language which includes many long or ill-formed expressions. The proposed method splits input into well-balanced translation units based on a semantic distance calculation. However, the method relies heavily on a computational semantic dictionary. Wakita (1997) proposed a robust translation method which locally extracts only reliable utterance segments, but the method does not split input into units globally, and sometimes fails to output any translation result. Zhou (2001) proposed a method of splitting Chinese utterances by using decision trees and pattern matching techniques, but the method lacks robustness when the input utterance is long and ill-formed or when the results from the speech recognizer contain many incorrectly recognized words. Reynar (1997) introduced a maximum entropy approach for identifying sentence boundaries. However, Reynar's approach focused on the boundary identification of English sentences in written language: potential sentence boundaries are identified by scanning the text for sequences of characters separated by white space (tokens) containing one of the symbols !, . or ?. Of course, in spoken language, there are no such specific symbols. Palmer (1994) and Riley (1989) also described methods of identifying sentence boundaries in written text.

Unfortunately, before beginning our work, we found few papers specifically addressing Chinese utterance segmentation.

2.2 Our Motivations

As outlined in Section 1, an utterance segmentation module operates between the speech recognition module and the translation component of a spoken language translation system (Figure 1).

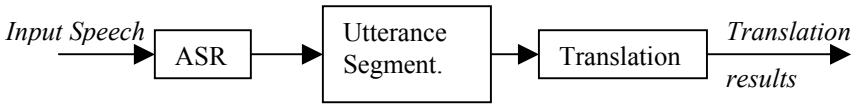

Figure-1. Location of the utterance segmentation module

In Figure 1, ASR signifies automatic speech recognition. A Chinese input utterance is first recognized by ASR; then the speech recognition result is analysed and possibly split by the utterance segmentation module (USM) before being passed to the translation module. In fact, the input utterance may already have been segmented by the speech recognizer using acoustic feature analysis. Thus in our experimental system an utterance can be split at both the acoustic and the linguistic levels. And so the input to the translation module is usually a simple sentence or a fixed expression, at least in theory. In this SLT design, some analysis work is separated out of the

translation module and moved to the segmentation module. Thus the translation module may employ simple direct translation methods, for example using template-based or pattern-based translation engines.

Suppose an input utterance has been transcribed by ASR, and a part of the recognition result is $P = W_1\ W_2\ \ldots\ W_n$ (where W_i is a Chinese word and $n \geq 1$.). P is possibly separated into k units $U_1, U_2, \ldots U_k$ ($1 \leqslant k \leqslant n$) by USM. A split unit is one of the following expressions:

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

Each part P is analysed and segmented by USM through the following three steps: 1) splitting using keyword detection; 2) splitting using pattern matching; and 3) splitting using syntactic analysis.

In this approach, a long utterance, especially an utterance containing more than two verbs, is usually split into small units, even if the original utterance is a complete simple sentence. As shown in the following examples,

Ex1. 我预订两个单人间需要多少钱？(How much does it cost if I reserve two single rooms?)

⇒ 我预订两个单人间 (I reserve two single rooms) ||

需要多少钱 (How much does it cost ?)

Ex2. 晚上9点以后办理入住手续可以吗？(May I check in after 9 o'clock in the evening?)

⇒ 晚上9点以后办理入住手续 (Register after 9 o'clock in the evening) ||

可以吗? (Is it OK?)

The examples show that it is no problem to understand the user's intension even if the utterance is split. This technique relies on the fact that the listener and the speaker both know what they're talking about. That is, they understand the discourse context. By taking advantage of such mutual knowledge, this splitting technique greatly reduces the work of the SLT system's translation component.

3 Segmentation Based on Multi-level Linguistic Analysis

In our methodology, if a string S from ASR is separated into n parts using the method of keyword detection, each part will be further segmented using, in succession, pattern matching methods and methods based on syntactic analysis.

3.1 Splitting by Keyword Detection

In the Chinese language, certain special words always indicate the beginning or the end of a simple sentence. For instance, the Chinese characters '呢(ne)', '吗(ma)' and '吧(ba)' always indicate the end of a question sentence. Two words '如果(if)' and '的话(a mood word)' often imply that the utterance is a conditional clause. Based on these facts, several special rules have been designed to split an input string. The rules are expressed by the two types of expressions as follows:

$$\#KW_1, KW_2, ..., KW_n \quad \ldots (1)$$

$$\%KW_{11}, KW_{12}, ..., KW_{1m}$$
$$\$KW_{21}, KW_{22}, \ldots, KW_{2k} \quad \ldots (2)$$

where KW is a keyword, and n, m and k are all integers greater than zero. In formula (1), $KW_1, KW_2,\ldots, KW_n$ are synonyms, and perform the same role in the utterance. Formula (1) means that if the keyword KW_i ($i \in [1 .. n]$) is present in the analysis input, the input will be split into two parts after the keyword KW_i. In formula (2), $KW_{11}, KW_{12}, \ldots, KW_{1m}$ and $KW_{21}, KW_{22}, \ldots, KW_{2k}$ are two sets of synonyms. Any KW_{1i} ($i \in [1 .. m]$) and KW_{2j} ($j \in [1 .. k]$) compose a pair of keywords that collectively determine the boundary of a split unit. KW_{1i} is treated as the starting word and KW_{2j} is treated as the last word of the split unit.

Since the splitting procedure is based only on character comparison and does not involve any lexical analysis, syntactic analysis, or semantic analysis, we say that the splitting is performed at a shallow level. The algorithm is as follows.

Input: a string S_{in} from ASR;
Output: a string S_{out} with boundary marks of split units.

Suppose all keywords given in formula (1) are denoted as a set KS_{single}, and all pairs of keywords given in formula (2) are denoted as a set KS_{pair}.

```
for ∀K∈KS_single
  {set the boundary mark after the keyword K; }
if S_in is separated into n parts: P_i (i = 1..n){
  for ∀P_i{
     for ∀K_p∈KS_pair {
       set the boundary mark after the second word
       of the keyword pair K_p;
     }
  }
}
Output S_out and return;
```

Algorithm 1. Segmentation based on keyword detection

3.2 Splitting by Pattern Matching

Once an input has been separated into *n* parts after splitting at the shallow level, each part $P = W_1, W_2, \ldots, W_m$ (where $m \geq 1$, $i \in [1..m]$, and W_i is a Chinese word) will be parsed and tagged with its phrase structure. Each part may be split using the pattern matching method. All patterns are expressed by Chinese words and part-of-speech or phrase symbols. For example,

#太 *AP* 了 … (3)
#*IJ* … (4)

where *AP* indicates an adjective phrase, and *IJ* is the symbol for fixed greeting expressions in Chinese. Pattern (3) signifies that all strings matching the pattern will be treated as a split unit, e.g. 太贵了(It is too expensive.), 太高了(It is too high.). Pattern (4) means that all fixed greeting expressions are treated as split units, e.g. 你好(Hello), 不客气(You are welcome), etc.

For phrase recognition, a partial parser is employed, based on the chart parsing algorithm using a PCFG (Probabilistic Context-Free Grammar). In our system, the goal of the parser is to recognize phrases rather than whole sentences. Although there are large differences between spoken and written Chinese, we think these differences are mainly reflected at the sentence level, e.g., by various orderings of constituents containing redundant words in spoken Chinese expressions. By contrast, spoken and written Chinese follow the same phrase construction patterns. Accordingly, the PCFG rules employed in our system are directly extracted from the Penn Chinese Treebank[1]. All of the rules comply with the condition of $\sum_i P(\text{LHS} \rightarrow \alpha_i)=1$. For example:

NN NN → NP, 1.00
MSP VP → VP, 0.94
MSP VP → NP, 0.06

3.3 Splitting by Syntactic Analysis

Splitting on the syntactic level is carried out by recognizing syntactic components and their dependency relations.

Suppose *S* is a string to be split on the syntactic level. After phrase recognition, $S = H_1 H_2 \ldots H_n$, where H_i ($i \in [1..n]$) is a phrase, and *n* is an integer ≥ 1. As motivated in Section 2.2, when analyzing dependency relations, we treat the verb phrase as the centre of the segment to be analyzed. Notice that we do not treat the *predicate* as the centre of the sentence, as is commonly done. There are two reasons: 1) in SLT systems, an input is often not a complete sentence, and it is frequently difficult to recognize the predicate; and 2) analysis of the dependency relation between a verb phrase and other phrases is relatively simple, as compared with analysis involving predicates, so analysis accuracy is increased.

[1] Refer to Fei Xia, "The Part-of-Speech Tagging Guidelines for the Penn Chinese Treebank (3.0)", http://www.ldc.upenn.edu/ctb/

In our approach, six dependency relations are defined between the verb phrase and other components: agent, quantifier, complement, direct object, indirect object, and adverbial adjunct. There are also six types of verb phrases:

1) The verb phrase does not take any object, denoted as V_0.
2) The verb phrase takes only one object at most, denoted as V_1.
3) The verb phrase usually takes two objects, denoted as V_2. One object is the direct object, and the other one is the indirect object.
4) The verb phrase probably takes a clause as its object, denoted as V_c.
5) The verb phrase takes a noun as its object, but the noun acts as the agent of another following verb or verb phrase, denoted as V_j. In this case, the noun is called a pivot word. The pivot word's action verb is denoted here as V_p.
6) The verb is a copula, such as 是(be), denoted as V_{be}.

In the dictionary, each verb is tagged with one of the six types. From V_0 to V_{be}, the level is considered to increase. A higher level type may override a lower-level type. For example, if a verb probably acts as a V_1, but also as a V_2, it will be tagged as V_2 in the dictionary. The type of a verb phrase and its context in an utterance can then be used to identify boundaries in an utterance according to the following algorithm:

Input: a part of an input utterance tagged with phrase symbols;
Output: split units of the input.

```
for each phrase XP {
   if XP = V0
      {the boundary mark (BM) is set after the XP;}
   if XP = V1
      {the BM is set after XP's object;}
   if XP = V2 {
      if there is indirect object {
         the BM is set after XP's indirect object;
      }
      else{
         the BM is set after XP's direct object;
      }
   if XP = Vc || XP = Vbe {
      if there is only a noun after the XP {
         the BM is set after the noun;
      }
      else{
         the BM is set after the XP;
      }
   if XP = Vj {
      if there is only a noun after the XP {
         the BM is set after the noun;
      }
      else{
         the BM is set after the Vp's object;
      }
}
```

Algorithm 2. Segmentation based on syntactic analysis

Figure 2 shows a sample application of the splitting algorithm based on syntactic analysis.

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?)

Analysis procedure:

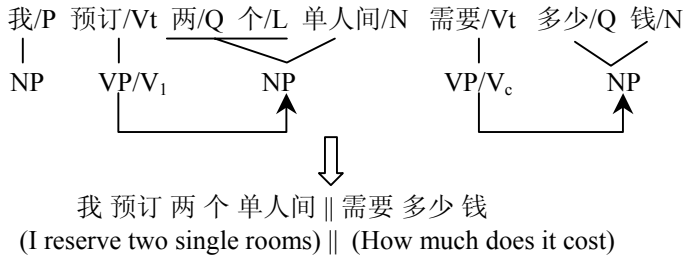

Figure-2. Sample application of the splitting algorithm based on syntactic analysis

4 Experimental Results

An experimental USM has been developed for a Chinese-to-English SLT system. The Chinese USM is built on 64800 collected utterances in the travel domain. From this corpus, we extracted 18 rules for splitting input on the shallow level, 32 patterns for splitting on the middle level, and 224 PCFG rules for partial parsing. Another 300 long utterances not included in the 64800 utterances are used as the test corpus, which contain 560 simple sentences or clauses, and 210 fixed greeting expressions. Thus each utterance consists of 2.57 split units on the average. The experimental results are shown in Table 1.

RESULTS	FIXED EXPRESSIONS	SIMPLE SENTENCES OR CLAUSES
Output	203	523
Correct	203	411
Correct Rate (%)	100.	78.6
Recall (%)	96.7	73.4

Table 1. Experimental Results

The table shows that the correct rate for the total output can be calculated by the formula: ((203 + 411) / (203 + 523)) ×100% = 84.6%. The recall rate is ((203 + 411) / (560 + 210))×100% = 79.7%. For the 560 simple sentences and clauses contained in the 300 input utterances, 37 simple sentences or clauses are not successfully separated out, and 112 utterances are split incorrectly. There were three main reasons for erroneous segmentation: (A) incorrect phrase parsing results, (B) incorrect

dependency analysis, and (C) lack of semantic consistency checking. Table 2 gives the distribution of the three error types.

RESULT	INCORRECT PARSING RESULTS	INCORRECT DEPENDENCY ANALYSIS	LACK OF SEMANTIC CONSISTENCY CHECKING
Number	71	24	17
Ratio (%)	63.4	21.4	15.2

Table 2. Error Distribution

Clearly, incorrect phrase parsing is the main cause of incorrect utterance segmentation.

5 Conclusion

This paper introduces a new approach to Chinese utterance segmentation for Chinese-to-English SLT systems, based on linguistic analysis. The preliminary results have given us confidence to improve the performance of our SLT system. However, much hard work remains for further research, including the development of robust approaches to phrase boundary recognition, to identification of the field that a verb phrase dominates, to verification of semantic consistency, etc. In the next step, we will focus mainly on the following two points:

- ◇ Research on approaches to identifying the semantic boundaries of sentences;
- ◇ Combining segmentation methods based on linguistic analysis with statistical methods, including the maximum entropy method, hidden Markov models (HMM), and decision-tree methods.

6 Acknowledgements

This work is sponsored by the Natural Sciences Foundation of China under grant No.60175012, as well as partly supported by the Education Ministry of Japan under Grant-in-Aid for Scientific Research (14380166, 14022237) and a grant funded by the University of Tokushima, Japan.

The authors are very grateful to Dr. Mark Seligman for his very useful suggestions and his very careful proofreading. The authors also thank the anonymous reviewers for their helpful comments.

References

1. Batliner, A. and R. Kompe *et. al*. (1996) Syntactic-Prosodic Labelling of Large Spontaneous Speech Data-Bases. In *Proceedings of ICSLP*. USA.
2. Cettolo, Mauro and Daniele Falavigna. (1998) Automatic Detection of Semantic Boundaries Based on Acoustic and Lexical Knowledge. In *Proceedings of ICSLP*. pp. 1551-1554.
3. Furuse, Osamu, Setsuo Yamada and Kazuhide Yamamoto. (1998) Splitting Long Ill-formed Input for Robust Spoken-language Translation. In *Proceedings of COLING*, vol. I, pp. 421-427.
4. Kawahara, Tatsuya, Chin-Hui Lee and Biing-Hwang Juang. (1996) Key-Phrase Detection and Verification for Flexible Speech Understanding. In *Proceedings of ICSLP*, USA.
5. Nakano, Mikio, Noboru Miyazaki and Jun-ichi Hirasawa *et. al*. (1999) Understanding Unsegmented User Utterances in Real-time Spoken Dialogue Systems. In *Proceedings of ACL*.
6. Palmer, David D. and Marti A. Hearst. (1994) Adaptive Sentence Boundary Disambiguation. In *Proceedings of the 1994 Conference on Applied Natural Language Processing (ANLP)*. Stuttgart, Germany, October.
7. Ramasway, Ganesh N. and Jan Kleindienst. (1998) Automatic Identification of Command Boundaries in a Conversational Natural Language User Interface. In *Proceedings of ICSLP*, pp. 401-404.
8. Reynar, Jeffrey C. and Adwait Ratnaparkhi. (1997) A Maximum Entropy Approach to Identifying Sentence Boundaries. In *Proceedings of the Fifth Conference on Applied Natural Language Processin*g. USA. pp.16-19.
9. Riley, Michael D. (1989) Some applications of tree-based modelling to speech and language. In D*ARPA Speech and Language Technology Workshop*. Cape Cod, Massachusetts. pp. 339-352.
10. Seligman, M. (2000) Nine Issues in Speech Translation. In *Machine Translation.* 15: 149-185.
11. Swerts, M. (1997) Prosodic Features at Discourse Boundaries of Different Strength. *JASA*, 101(1): 514-521.
12. Stolcke, Andreas and Elizabeth Shriberg (1996) Automatic Linguistic Segmentation of Conversational Speech. In *Proceedings of ICSLP*, vol. 2, pp. 1005-1008.
13. Stolcke, Andreas and Elizabeth Shriberg *et. al.* (1998) Automatic Detection of Sentence Boundaries and Disfluencies Based on Recognized Words. In *Proceedings of ICSLP*, pp. 2247-2250.
14. Wakita, Yumi, Jun Kawai *et. al*. (1997) Correct Parts Extraction from Speech Recognition Results Using Semantic Distance Calculation, and Its Application to Speech Translation. In *Proceedings of Spoken Language Translation*. Spain. pp. 24-31.

15. Wightman, C. W. and M. Ostendorf. (1994) Automatic Labelling of Prosodic Patterns. *IEEE Transactions on Speech and Audio Processing*, 2(4): 469-481.
16. Zechner, Klaus and Alex Waibel. (1998) Using Chunk Based Partial Parsing of Spontaneous Speech in Unrestricted Domains for Reducing Word Error Rate in Speech Recognition. In *Proceedings of COLING-ACL'98*, pp. 1453-1459.
17. Zhou, Yun. (2001) Analysis on Spoken Chinese Corpus and Segmentation of Chinese Utterances (*in Chinese*). *Thesis for Master Degree*. Institute of Automation, Chinese Academy of Sciences.