A KOREAN MORPHOLOGICAL ANALYZER FOR SPEECH TRANSLATION SYSTEM

Youngkuk Hong *Myoung-Wan Koo* *Gijoo Yang*[1]

Multimedia Technical Laboratory, Korea Telecom Research Laboratories, Korea Telecom
[1]Dept. of Computer Science and Statistics, Dongguk Univ., Korea

ABSTRACT

This paper describes a Korean morphological analyzer which can be used as a part of language processor for a speech translation system. We have modified the CYK algorithm so that we are able to analyze many phenomena occurring in spontaneous speech such as ellipsis, shorter words, poor and mispronounced words and so on. And we also have constructed a rule set with 112 connection rules and seven kinds of dictionaries, in which there are totally about 81,000 keywords. Currently, we have achieved the success rate of 93.0% with a text corpus of dialogue for hotel reservation domain.

1. INTRODUCTION

At present, many researches for Natural Language Processing have been carried out actively at home and abroad. Domestically, language processing for written texts has been major topic for a long time. However, the research for spoken language is still in the beginning phase. In the technically advanced nations, many researches for spoken language have been done, especially as a part of automatic speech-to-speech translation systems for limited-domain applications [1, 2, 3]. Those researches are based on foreign languages, which are known to be quite different from Korean language in terms of linguistic and cultural characteristics, so that the results from these researches are inadequate to apply to Korean language.

In view of morphological typology, Korean language is an agglutinative language in which functional words such as a *josa*[josa][1] and an *eomi*[əmi][2] are well defined and frequently used [4]. In addition, many characteristics in spoken language such as ellipsis, shorter words, poor and mispronounced words depend on language and culture. So, for handling such problems in Korean, it is necessary to study Korean language and culture.

In this paper, we present a Korean morphological analyzer for speech translation system. In section 2, the overview of characteristics in Korean spontaneous speech is described. In section 3, we explain the methods for handling the characteristics described in section 2. Then, the structures of the system and its sub-modules are roughly presented in section 4. In section 5, we show the results of experiments done for our system. Finally, a conclusion is made in section 6.

[1] *josa* is a Korean noun-ending.
[2] *eomi* is a Korean verb-ending.

2. CHARACTERISTICS IN KOREAN SPONTANEOUS SPEECH

An *eojeol*[əjəl] is a unit for Korean morphological analyzer. An *eojeol* is composed of one or more morphemes and is separated from another *eojeol* with a space and/or a punctuational mark. Of cause, in dialogues, sometimes two or more *eojeols* may be merged into an *eojeol*. Thus, in this section, we discuss the characteristics of the *eojeol* occurring in Korean spontaneous speech. We have collected and analyzed large speech corpus to capture the characteristics [5, 6].

2.1. Ellipsis of Josa

As compared to written texts, ellipsis of *josa* occurs more often in spoken language and mainly used to be case-*josa* is omitted.

"예약할 방 있어요?" (1)
("Do you have room available?")

"서울가는 기차표를 예매할 수 있나요?" (2)
("May I book a train ticket bound for Seoul?")

In the above two example sentences, there are two types of *eojeols* where their *josa*'s were omitted. In the first type, a noun '방'(room) is used as an isolated *eojeol* without a *josa*, and in the second type, a noun '서울'(Seoul) is directly connected to the next *eojeol* '가는' without pause. For the above phenomena, there is no rule available. Only the speaker's habits or communication intention may cause the ellipsis of *josa*, and may merge two or more *eojeols* into an *eojeol*.

2.2. Shorter Words

Another characteristic of spoken language is shorter words.

"10월 1일로 예약돼 있습니다." (3)
("Your room was reserved from October 1.")

"이십일에 출발하는 걸로 할께요." (4)
("I want what is starting on 20th of this month.")

"서울행 기찬데요." (5)
("It's a train bound for Seoul.")

The sentence (3) shows that a verb "예약되" and an *eomi* "어" are shortened into an *eojeol* "예약돼". Seen in the sentence, the inflection of verb is one of the primary factors which make Korean morphological analysis more difficult.

In the sentence (4), a bound noun '것'(thing) and a *josa* '으로' are shortened into an *eojeol* '걸로'. As in the sentence, when a *josa* appears after noun with a final consonant, it is found that a shortened form is used in most cases.

The next sentence shows that a noun and an *eomi* can be merged into an *eojeol* like a verb, when a stem-supplement[3] '이' is omitted. The original form of the *eojeol* '기찬데요' is '기차인데요'[4](it's a train). But, after ellipsis of the stem-supplement '이', the noun '기차' was connected into the *eomi* '느데요' to make a new *eojeol* '기찬데요'. Such a case is possible only when the noun has no final consonant.

2.3. Combination of Bound Noun and Its Modifier

In spontaneous speech, a bound noun is often to be connected to its preceding modifier without space like the following sentence.

"예약할수 있습니까?" ("May I reserve it?") (6)

There is a bound noun '수' which is connected to the preceding modifier '예약할' in the sentence. Generally, a bound noun must be separated from the preceding modifier with a pause, but in spontaneous speech, this rule is often ignored.

2.4. Separation of Eojeol

Another characteristic is that an *eojeol* can be separated to be two *eojeols* by speaker's intention.

"구월 이십오일에 예약 되셨습니다." (7)
("Your room was reserved from September 25.")

In the above sentence, the two *eojeols* "예약" and "되셨습니다" are separated from the original *eojeol* "예약되셨습니다" for the purpose of speaker's stressing a word '예약'. Such separation of an *eojeol* mainly occurs in the case of the *eojeol* formed as [predicative-noun[5] + stem-supplement].

2.5. Interjections

The last characteristic described in this section is Korean interjections such as "아, 어, 그, 에, 음, 저, 뭐, ···" As other languages, many interjections are often inserted between *eojeols* in Korean spontaneous speech.

[3]stem-supplement changes a noun into a verb.

[4]'기차'(noun) + '이'(stem-supplement) + '느데요'(*eomi*).

[5]it is a noun which can become a verb combining with a stem-supplement.

3. ANALYSIS METHODS

In this section, we present handling methods for analyzing *eojeols* in spontaneous speech based on characteristics described in the previous section.

3.1. Method for Handling the Ellipsis of Josa

As mentioned in section 2.1, there are two types of *eojeols* whose *josa* is omitted. The first type is so trivial as to be simply analyzed by dictionaries. However, the second type is hard to be analyzed, because the *eojeol* should be recovered to two original *eojeols*, and then each *eojeol* should be analyzed respectively. In this paper, we use pseudo-code algorithm described in Algorithm 1 as a method for handling the connected *eojeol*.

Algorithm 1 A method for handling the connected *eojeol*

E_N : input eojeol with N-jaso's[jasoz]
E_i : estimated eojeol of which josa is omitted
E_j : estimated eojeol followed by E_i

$E_N \Leftarrow J_1 + \cdots + J_N$
for $i = 1$ to N **do**
  $E_i \Leftarrow J_1 + \cdots + J_i$
  if the eojeol E_i is in the Noun-Dictionary **then**
    $E_j \Leftarrow J_{i+1} + \cdots + J_N$
    if the eojeol E_j can be analyzed **then**
      $E_N \Leftarrow E_i + E_j$
    end if
  end if
end for

3.2. Methods for Handling the Shorter Words

As described in section 2.2, a verb can be shortened with *eomi*, and a noun can be shortened with *josa* or *eomi*.

Since the inflection of verbs may occur at vowel in Korean, we analyze the inflected verbs with regards to a vowel. We will give a full detail of the method for handling the inflected verbs in section 4.2.

There are two types of shorter nouns: One is '걸' in sentence (4) and the other is '기찬데요' in sentence (5). The problem case is the former. Generally, a word with a final consonant cannot become a shorter word with a *josa*, but shorter words of bound-noun '것' are frequently found in the corpus of spontaneous speech. However, the shortened form is not so various that we can construct a dictionary for shorter words easily to handle the shorter words of bound-noun '것'.

The shorter words of noun with *eomi* often occur in the case that a stem-supplement '이' between the noun and the *eomi* has been omitted. Our connection rule says that *eomi* can be

connected to stem-supplement '이'. So, if a new rule saying that the *eomi* can be connected to stem-supplement '이' is added to rule-set, the shorter words can be easily analyzed by this rule.

3.3. Method for Handling the Connection of Bound Noun and Its Modifier

For the connection of bound-noun and its modifier, we have constructed a new rule. This rule is based on the syntactic property of bound noun which can restrict its preceding modifier. For example, the *eojeol* '예약할수' is divided into the preceding modifier '예약할' and the bound-noun '수' by the rule, and they are subsequently analyzed by other rules.

3.4. Method for Handling the Separation of Eojeol

For handling the separation of an *eojeol*, we first analyze each *eojeol* in a sentence. Then, we find out two neighbored *eojeols* which must have been separated from an *eojeol* and try to connect them into an *eojeol*. To connect them into an *eojeol*, we apply the proposed algorithm shown in Algorithm 2 to those two neighbored *eojeols* when both of them were successfully analyzed. There is only one precondition in this algorithm: the first morpheme of the second *eojeol* must be one of the functional words such as *josa*/stem-supplement/*eomi*/*eomi*-preceding[6].

Algorithm 2 A method for handling the separated *eojeols*

E_i : successfully analyzed eojeol
S_N : input sentence with N-it eojeols
$M_{i,j}$: jth-morpheme of ith-eojeol

$S_N \Leftarrow E_1 + \cdots + E_N$
for $E_i = E_1$ to E_{N-1} **do**
  if E_i has been successfully analyzed **then**
    if E_{i+1} has been successfully analyzed **then**
      $E_i \Leftarrow M_{i,1} + \cdots + M_{i,l}$
      $E_{i+1} \Leftarrow M_{i+1,1} + \cdots + M_{i+1,m}$
      if $M_{i+1,1}$ is a functional word **then**
        if $M_{i+1,1}$ can be connected with $M_{i,l}$ **then**
          $E_new \Leftarrow E_i + E_{i+1}$
        end if
      end if
    end if
  end if
end for

3.5. Method for Handling Interjections

Although the insertion of interjections is one of the characteristics in spontaneous, interjections can be removed in the stage of speech recognition processing. Thus, there is no need for us to handle interjections in the stage of morphological analysis processing.

[6]it always precedes an *eomi*, so it is called '*eomi*-preceding' in this paper.

4. SYSTEM STRUCTURES

Seen in the Figure 1, the Korean morphological analyzer for spontaneous speech is composed of four modules: Preprocessing module, Analyzing module, Separating/Combining Module, and Postprocessing module.

4.1. Preprocessing Module

In this module, input *eojeol* is divided into a sequence of Korean characters and a sequence of non-Korean characters. But, when the analyzer is pipelined with speech recognition system, this module will be ignored.

4.2. Analyzing Module

A sequence of Korean characters is analyzed by CYK algorithm [7]. The CYK algorithm has been modified by taking into account the characteristics of Korean spontaneous speech. The proposed algorithm find out all sequences of morphemes which are grammatically possible. And a right-to-left search technique is used to improve efficiency of the algorithm. Since the right-to-left search technique can easily separate *josa* or *eomi* from an *eojeol*, we can get the only information required for further analysis from the dictionaries on the basis of the separated *josa* or *eomi*. As a result, we can save the dictionary access time which has heavy influence on the analyzing speed.

In addition, we analyze inflected verbs with regards to a vowel as described in the previous section. In this way, we define all possible inflection types for each vowel, and then we analyze all kinds of inflected verbs with regards to the inflected vowel.

4.3. Separating/Combining Module

In this module, the connected *eojeol* may be separated into two or more sub-*eojeol*, and/or the separated *eojeols* may be connected into an *eojeol*. This module will interact with the analyzing module if necessary, as showed in the Figure 1.

4.4. Postprocessing Module

Over-generated results and unexpected results are removed in this module. Also the recovery of inflected verbs is handled. In addition, the filtering for compound nouns is also processed. All processing in this module are based on the heuristic knowledge.

4.5. Dictionaries and Connection Rules

We have constructed seven kinds of dictionaries which contain totally about 81,000 keywords. We also have constructed a rule set with 112 connection rules for controlling connection between two morphemes. In order to obtain correct analyses, we have developed a new classification of Korean part of speech, in which all morphemes are classified

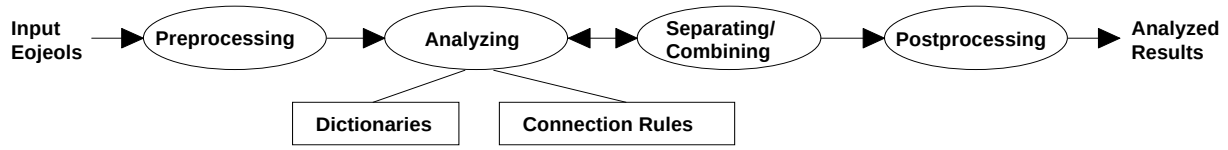

Figure 1: Korean morphological analyzer for spontaneous speech

# of input eojeols	# of analyzed eojeols	# of not-analyzed eojeols
30,240	24,948(82.5%)	5,292(17.5%)

Table 1: Experimental results by the analyzer for written text

# of input eojeols	# of analyzed eojeols	# of not-analyzed eojeols
30,240	27,904(92.3%)	2,336(7.7%)

Table 2: Experimental results by the analyzer for spoken language

into 17 categories.

4.6. Experimental Results and Analysis

For the performance evaluation of the analyzer described in this paper, we have tested two kinds of analyzers using dialogues for hotel reservation. One is designed for written texts and the other is for spoken language. The experimental results are shown in Table 1 and Table 2. The former analyzer showed very poor performance and the latter showed better performance than the former.

We tested our system with another larger dialogues extracted from newspapers and magazines. Table 3 shows that the our analyzer are working well.

5. CONCLUSION

In this paper, we have described a morphological analyzer which has been developed to be used as a part of language processor for a speech translation system.

Our system has been designed to handle idiosyncrasies of spontaneous speech such as ellipsis of *josa*, shorter words, connection of bound-noun and its modifier, separation of *eojeol* and so on. To handle such problematic cases, we have classified Korean part-of-speech into 17 categories and modified the CYK algorithm so that we can analyze shorter words easily. In addition, we used the right-to-left search technique which can easily separate *josa* or *eomi* from an *eojeol*. Furthermore, we have constructed a rule set with 112 connection rules and seven kinds of dictionaries which contain about 81,000 morphemes. Using these tools, our proposed algorithm is capable of finding all sequences of morphemes which are grammatically possible.

# of input eojeols	# of analyzed eojeols	# of not-analyzed eojeols
72,617	67,082(92.4%)	5,535(7.6%)

Table 3: Experimental results by the analyzer for spoken language with interviews extracted from newspapers and magazines

We have made two kinds of experiments. In the first experiment, we compared the performance of the proposed morphological analyzer with that of the conventional analyzer with regard to the same dialogues for hotel reservation domain. The conventional one shows the success rate of 82.5% and the new one yields the success rate of 92.3%. In the second experiment, we have tested the new analyzer with the dialogues extracted from other domain. The analyzer has shown that it is quite stable for other kinds of dialogues and texts.

Currently, we are concentrating on tuning and improving our morphological analyzer to achieve the better success rate.

6. REFERENCES

1. Dinesh Tummala, Stephanie Seneff, et al, "CCLINC: System Architecture and Concept Demonstration of Speech-to-Speech Translation for Limited-Domain Multilingual Applications," *Proc. of the Spoken Language Systems Technology Workshop*, pp.227-232, 1995.
2. B. Suhm, P. Geutner, et al, "JANUS: Towards Multilingual Spoken Language Translation," *Proc. of the Spoken Language Systems Technology Workshop*, pp.221-226, 1995.
3. L. Mayfield, M. Gavalda, W. Ward, and A. Waibel, "Concept-based Speech Translation," *ICASSP-95*, pp.97-100, 1995.
4. Ju-Haeng Lee, *Contemporay Korean Grammar*, Korea Textbook Inc., Korea, 1992.
5. Ki-Yong Lee, Seung-Ku Kang, Seung-Won Rho, *Text Corpus for Hotel Reservation Domain*, Dept. of Linguistics, Korea Univ., Korea, 1995.
6. Yong-Ju Lee, et al, *Spntaneous Speech and Text Corpus for Train-Ticket Reservation Task*, Human Interface Lab., Dept. of Computer Engineering, Won-Kwang Univ., Korea, 1995.
7. Eun-Chul Lee, *An Improved Method on Korea Morphological Analysis Based on CYK Algorithm*, A Master Thesis in Dept. of Computer Science, POSTECH, Korea, 1993.