Parsing Long English Sentences with Pattern Rules

Wei-Chuan Li, Tzusheng Pei, Bing-Huang Lee, and Chuei-Feng Chiou

System Software Department(X200)
Electronic Research and Service Organization(ERSO)
Industrial Technology Research Institute(ITRI)
Chutung, Hsinchu, Taiwan, R.O.C.
E-mail: X200PTS0@TWNITRI1.BITNET

Abstract

In machine translation, parsing of long English sentences still causes some problems, whereas for short sentences a good machine translation system usually can generate readable translations. In this paper a practical method is presented for parsing long English sentences of some patterns. The rules for the patterns are treated separately from the augmented context free grammar, where each context free grammar rule is augmented by some syntactic functions and semantic functions. The rules for patterns and augmented context free grammar are complimentary to each other. In this way long English sentences covered by the patterns can be parsed efficiently.

1. Introduction

A long English sentence, from the parsing point of view, is defined as a sentence which has complicated syntactic structure or has too many words in it. Some factors which may contribute to the syntactic complication are words with multiple part-of-speeches, conjunctions, prepositional phrases, and commas. Since the number of possible syntactic structures of a sentence grows with the factors, it is not easy for a machine translation system to pick a right syntactic structure, based on syntactic knowledge and a little semantic knowledge[1], among the large set of possible syntactic structures generated by the parser and the parsing time increases as well, due to the construction of many possible syntactic structures.

To put it in another way, sentence parsing is a searching problem. The parsing time increases exponentially as the branching factor, reflecting complicated syntactic structure, and searching depth, reflecting actual sentence length, increase. In some-path bottom-up parsing[2], reducing branching factor or using beam search method [3][4] to restrict the value of branching factor may decrease parsing time. However, basically, the parsing mechanism is still exponential.

As an example, for the English-Japanese machine translation system, ATLAS II [5], in translating the corpus of 220 sentences selected from software manuals and papers, among the English sentences with usable translation, the average number of words of the sentences is 33.5. For sentences with translation that needs some postediting and sentences with translation that can not be used, the average sentence lengths are 45.7 and 46.8 words respectively. In order to have better performance, a machine translation system should be able to translate sentences of reasonable length.

2. ERSO-ECMT

The English-Chinese machine translation system, ERSO-ECMT, has been developed in Electronic Research and Service Organization(ERSO), Industrial Technology Research Institute(ITRI), Hsinchu, Taiwan, since July 1986. The analysis/transfer/systhesis approach has been adopted. The system contains an augmented context free grammar, where each context free grammar rule is augmented by some syntactic functions, which reflect preference over some syntactic structures, and semantic functions. The status of parsing process should satisfy the syntactic and semantic conditions before the parser applies the grammar rule to derivation. The system translates one sentence at a time. There is no information from the context while translating the current input sentence. At present, the domains of translation for ERSO-ECMT are in computer science and environmental protection. The samples from environmental protection include 871 sentences, with 19539 words in total, excerpted from some abstracts of papers. The average sentence length is 18.3 words. There are 42 sentences, about 4.8 %, with number of words over 40.

The characteristics of translation speed of ERSO-ECMT, run on LAMBDA LISP machine, for the computer science corpus which is the first chapter of the UNIX manual are as follows: for sentences of length less than 34 words, the time for translation, in average, is within one minute and for sentences of length over 40 words and on, the speed increases drastically (exponentially). In order to have reasonable translation time, say one minute for ERSO-ECMT, the length of sentences should be limited.

3. Our Approach

In addition to some-path bottom-up parsing for reducing branching factors, the input sentence can be divided into several meaningful segments(i.e. reducing the searching depth), then each segment is parsed separately without exchanging information with the other segments, and finally the parsing results of all segments are combined. The Chinese translation will be based on the combined parsing result.

The parser of ERSO-ECMT first see if the input sentence matches the long sentence patterns. It will do parsing in accordance with the pattern having been matched, otherwise it will proceed parsing with the augmented context free grammar. In the case of failing in getting a complete parsing tree with the long sentence pattern, the parser will also do the same thing, trying to parse the sentence with the augmented context free grammar.

The procedure for parsing with the long English patterns is as follows:

a. Partitions the input long sentence into some meaningful segments:
   .Looks up the partition rules by pattern-matching with unification.
   .If the resultant segments are still with length greater than 40, does partitioning recursively on them, until no more pattern can be used.

   Note: in general, the resultant segments, such as Declarative Sentence(SDEC), Noun Phrase(NP), Infinitive Phrase(INF), and Verb Phrase(VP), are big structures with some key words or some special structures among them in the sentential form.

b. Parses or translates each segment separately.

c. Combines the results of all segments.

d. Generates the corresponding Chinese sentence.

   Note: The parser can either combine the syntactic parsing results of all segments and then generate corresponding Chinese sentence, or generate Chinese translations of all segments and then put them, by transformation rules, in a sequence with order not necessary that of the original English segments in the input sentence.

Before parsing a sentence, a sentence length threshold, say 40 for ERSO-ECMT, can be set to indicate how long a sentence will be parsed with the pattern rules.

The format of the rules for long English patterns in ERSO-ECMT is as follows:

```
SEGRULE := (LHS RHS)

LHS :=  (E ... E)
E   :=  E1 | (E1 test)
E1  :=  VAR | CAT | ving | ved | num
      | english | (closure LHS) | (plus LHS)
      | (opt E1)
CAT :=  a | z | x | art | b | c | h | m | n
      | p | r | v | u | w | wn | wu

RHS :=  (A ... A)
A   :=  (parse VAR node)
      | (parse_transfer_synthesis VAR)
      | chinese
VAR :=  %v1 | ... | %vn
```

where "(" and ")": all terminals,
SEGRULE : a rule for long English segmentation,
LHS: an augmented regular expression which is composed of a regular expression and test(s),
RHS: parsing action(s),
test : a LISP function which implements the designated test,
ving: a gerund, such as going, doing,
ved: a verb with ending "ed", where it indicates its past or past participle form,
num: a number,
english : an English word, or a symbol of punctuation,
closure and plus: the functions R^* and R^+ corresponding to R , where R is a regular expression, (The function are done by matching the shortest pattern, covered by the functions, in the input sentence.)
opt: an optional item,
Each symbol of the right-hand side of CAT: part-of-speech or category.
parse : the LISP function for doing parsing,
node : a grammar node, a nonterminal of the parsing tree,
parse_transfer_synthesis: the LISP function to do syntactic parsing, transformation, and generation,
chinese : Chinese character(s), and
%v1 ... %vn: each of them being a variable to which a segment of the input English sentence will be bound.

The reason for using the regular expression is that some repeated elements can be covered. Although the expressive power of the regular expression is less than that of the augmented context free grammar already in the system, they focus on two different things. The augmented context free grammar takes care of detailed phrase structures, though it can deal with long sentences, not quite well in general, whereas the long sentence pattern rules handle some of long sentences by breaking them down into segments of some big structures and then the augmented context free grammar takes care of all the segments.

4. Examples

Example 1. The following sentence matches the pattern (%v1 ving %v2, (closure (ving %v3,)) c ving %v4) where "c" is a conjunction.

The air-use plans may be used as the basic framework for achieving the desired air quality by such means as limiting the emissions from individual sources, limiting the emissions from sources in certain areas, or disallowing new pollutant sources in overburdened areas.

Example 2. The following sentence matches the pattern (In order INF, SDEC). The syntactic tree is the combination of the trees of INF and SDEC.

In order to evaluate or rank land use plans in terms of air quality, it is necessary for planners to be able to project emission density using only planning variables, because detailed source characteristics are not available at the time alternative plans are being developed and evaluated.

Example 3. The pattern (SDEC, because SDEC) can be used to parse the following sentence. The corresponding Chinese sentence can be obtained from the combination of the two Chinese segments for the two SDECs.

Epa's preferred decision is to approve and support funding for the proposed alternative, because this is the most cost-effective way of achieving federal and state water-quality goals, improving the quality of the rio grande, and protecting prime agricultural lands.

Example 4.

For some special sentence patterns, they have specific Chinese translation. The following is a pattern rule.

(Not that SDEC but that SDEC)

5. Result of parsing a sample of the environmental corpus

The results from parsing the environmental protection corpus mentioned in 1. are as follows.

a. Number of sentences with length > 40
   : 42 sentences
b. Correct partition
   : 36 sentences
c. Correct partition with correct translation
   : 28 sentences
d. The percentage of c. over b.
   : 66.7 %
e. Average parsing time for these long sentences
   : 2 min/sentence

For ERSO-ECMT, the rate of correct matching is about one third of the total long sentence. It can be improved by organizing more rules to cover a large range of sentences, but for some long sentences without apparent features, there is no pattern rules for them.

6. Discussion

In fact, the problems arising from parsing long English sentence are the combined effects of some problems, which could not be treated quite well, such as prepositional phrase attachment problems[6], compound noun phrases. They are all about the problems of the semantic relations among words or segments in sentences. At present the ways for encoding and using massive semantic information for a practical machine translation system of some domains, computer science and environmental protection, for example, are not clear. Basically the approach here is not to solve all the problems, but to break the sentence into some segments within manageable size and then to parse them separately.

In this way, since each segment is shorter than the original sentence, failing to construct a correct parsing result usually affects only a shorter range of words, but it wastes time in parsing some pattern which is found not appropriate for the input sentence eventually. In order not to do pattern matching so much, which is time-consuming, some patterns can be put under the word, which is the leading word of the pattern, of the dicionary, and some under the augmented context free grammar rules of appropriate nonterminals to guide the parser before the parsing mechanism is initiated. The parser will check to see if the sentence is a particular pattern by looking at the pattern rule encountered in the dictionary or in the grammar rules in course of parsing and then try to parse the other parts of the sentence with the pattern rule found if any.

7. Conclusion

For short sentence, a good MT system can usually generate readable translation, whereas for long sentence the translation is usually not satisfactory. A practical method is presented for parsing long English sentence. It bases on some patterns of long English sentences. The patterns can be inserted in the lexicon or the augmented context free grammar to guide the parser.

8. Acknowledgement

The paper is a partial result of the project No. 3131500 conducted by the ITRI under sponsorship of the Minister of Economic Affairs, R.O.C.

9. References

[1] Jane J. Robinson, "DIAGRAM: A Grammar for Dialogues", Communications of ACM, January 1982, pp. 27-47.

[2] W.S. Bennett and J. Slocum, "The LRC Machine Translation System", Computational Linguistics, Vol. 11, No. 2-3, April-September, 1985, pp. 112-121.

[3] Keh-Yih Su, et al., "A New Parsing Strategy in Natural Language Processing Based on the Truncation Algorithm", NCS 1987, Vol 2, pp. 580-586.

[4] P.H. Winston, "Artificial Intelligence", Addison-Wesley, Reading, MA, USA, 1984.

[5] Makoto Shiotsu, "Japanese Polish Support System for the Japanese-English Translation", The 35th National Conference of Information Processing Society of Japan, September 1987(in Japanese), pp. 1249-1250.

[6] Y. Wilks and X.M. Huang, "Syntax, Preference and Right Attachment", IJCAI-85, pp. 779-784.