Automatic Generation of Japanese Dependency Grammar Entries

4B−5

Eduardo de Paiva Alves and Teiji Furugori *

1 Introduction

This paper describes part of a system to help students of Japanese as a foreign language read technical papers. To achieve this goal, the system identifies structures of the sentences that the native speakers intuitively understand.

The system uses Restricted Dependency Grammar (RDG), which consists of a parser and a set of word entries, the last of these being hand-coded. We propose a method for generating these entries automatically using information from the Word Dictionary, Concept Dictionary and Co-occurrence Dictionary provided with EDR.

2 Approach

Obtaining the lexical information to build word entries for a dependency grammar is a hard task since a restricted classification of the words is not sufficient to rule out undesirable parsings. Working with a comprehensive classification is very costly. It involves both syntactic and semantic information. This paper proposes a way, using information from a machine readable dictionary in the form of a detailed hierarchy of concepts as well as the necessary information on syntactic features, to build the word entries used in a dependency grammar.

2.1 Dependency Grammar

Restricted Dependency Grammar (RDG, Fukumoto et all,1992) is a Japanese dependency grammar which builds a dependency structure from the phrases of a sentence. The restrictions are either linguistic or structural. Linguistic information consists of grammatical category and syntactic attributes. Structural information consists of restrictions on the possible dependency relations.

This information is stored in RDG in the form of word entries and rules. Linguistic information is included in the word entries and is used to classify the phrases of a sentence in ranks, which are used in the rules to decide whether a relation can be built. Figure 2 shows a word entry from RDG.

surface	[ジョギング, 中, の]
modified	体言
modifier	連体
rank	a1
case	の
sem	[act]

Fig 1: Example of a word entry

2.2 Dictionary

EDR (Japan Electronic Dictionary Research Institute, 1993) is a machine readable set of japanese dictionaries : the Word Dictionary includes syntactic information, definition, examples and associates each entry with an entry in the Concept Dictionary (CD); the CD is a set of graphs consisting of concepts and a number of relations. These include both taxonomic (kind-of) as well as functional (agent, object, etc) relations. In fig. 2 is an extract of CD which shows all the generalizers for the concept 人間 (human).

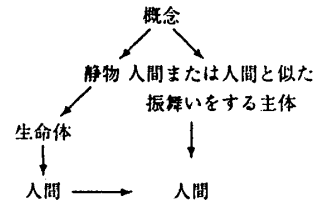

Fig. 2: Extract from the **Concept Dictionary**

The Co-occurrence Dictionary is a list of word pairs (co-occurrences), the relation which holds between the word pair, and a probability (0/1) of this relation. It includes also patterns for verbs. In fig. 3 is an extract from the Co-occurrence Dictionary which shows one possible pattern for the verb 働く.

働く	agent	が	act
	人間		生計を立てるために仕事をする

Fig. 3: Co-occurrence Dictionary Pattern

*Department of Computer Science, University of Electro-Communications, 1-5-1 Chofugaoka Chofushi Tokyo Japan {ealves,furugori}@phaeton.cs.uec.ac.jp

2.3 Method

The process which generates the word entries for a given sentence in Japanese is outlined in fig. 4. The input is first morphologically analized using a version of JUMAN (Matsumoto et all, 1994) which uses EDR Word Dictionary. Then for each phrase, an entry to RDG is produced using information from dictionaries in EDR set. This entries are added to RDG which performs the parse and the results can be used in the system to display the structures to the users. The method which generates the word entries is described in this section.

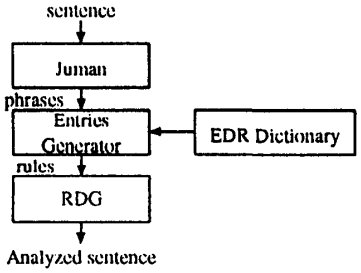

Fig. 4: The outline of the system

The input is an unanalized sentence in Japanese:

産業構造が変わり技術革新が進んで働く人のストレスも募ってきた。

The sentence is first input to JUMAN, which performs the morphological analysis and provides grammatical and inflectional information. When there is more than one possible morphological analysis, each one generates a separate sub-process.

2.4 Generating Entries

RDG is basically a parser and its dictionary is suitable only to parse 35 example sentences from Asahi Shinbun included with the system. In order to use it with arbitrary sentences it's necessary to build entries for all new words.

The morphological analysis using EDR provides the necessary syntactic information to build the word entry to RDG. The possible word entries are identified and classified into ranks according to the correspondences in EDR. Fig. 5 shows these correspondences for the example sentence.

Word Entry	Cat		Rank
産業構造が	体言	連用	a1
変わり	用言	連用	b
技術革新が	体言	連用	a1
進んで	用言	連用	a3
働く	用言	連体	a1
人の	体言	連体	a4
ストレスも	体言	連用	a1
募ってきた	用言	nil	

Fig. 5:Identifying and classifying word entries

The verb frames are built from the patterns in the Co-occurrence dictionary. Fig. 6 shows the frame generated for 働く.

働く　が　人間
　　が　言葉; 文字; 文法用語
　　が　動物の部分; 具体
　　が　器具; 具体物の空間的属性名; 法則; 構造

fig. 6:Example of a verb frame

For nouns, the word stems are identified and the possible concepts are extracted from the Word Dictionary. All their generalizers in the CD are included to allow matching with the restrictions in the verb frames. For instance one of the concepts corresponding to 人 is 人間 and the generalizers are those shown in fig. 2.

It's possible to build the word entries like that shown in fig. 1 combining the syntactic information provided by JUMAN morphological analysis, the semantic information for nouns in the form of the hierarchy of concepts and the frames built from the patterns in the Co-occurrence dictionary

3 Discussion and Future Work

The method was tested for the sentences from Asahi Shinbun provided with RDG so that we could compare the results using our automatically generated rules with those of hand-coded rules included in RDG.

The method is proven efficient as a tool for generating automatically the word entries to be used in parsing structures directly from the unanalized sentence. This spares the user the trouble of both identifying the phrases or adding entries in the dictionary.

The restrictions included in RDG allow a great number of parses, which could not be reduced using a more refined classification such as the CD from EDR. In future work we attempt to solve this problem using mutual information between modifier/modified pairs to identify the most probable parse for a given sentence.

References

[1] Fukumoto, F.; Sano H., Saitoh, Y.; and Fukumoto J. (1992). " Restricted Dependency Grammar" *Trans. IPS Japan*, 33(10), (in Japanese).

[2] Matsumoto, Y; Kurohashi, S; Utsurou, T; Nyoki, Y; Shinho, H; Nagao, M (1994). *User's Guide for the Juman System. Version 2.0* (in japanese)

[3] Japan Electronic Dictionary Research Institute, Ltd. (1993). *EDR Electronic Dictionary Specifications Guide* (in Japanese)