Adding Co-occurence Information to Dependency Grammar

Eduardo de Paiva Alves **Teiji Furugori**
Department of Computer Science **University of Electro-Communications**

2R－6

1 Introduction

This paper proposes a method to reduce the number of structures generated by a Dependency Grammar. The method consists of adding information extracted from a conceptual graph and corpora in the form of co-occurence pairs.

2 Dependency Grammar

In the Dependency Grammar framework only the relations between two arbitrary constituents of a sentence is considered in the analysis of a sentence. Tho build these relations, and the resulting dependency structures, linguistic and structural constraints are used. Nevertheless useless solutions are generated. The following example shows a dependency structure.

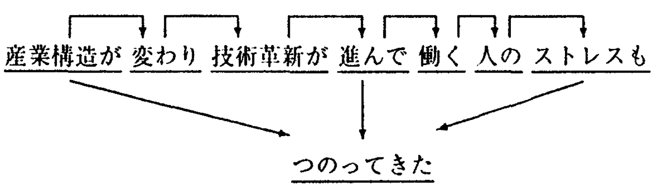

Fig. 1: Example of a dependency structure

Restricted Dependency Grammar (RDG, 福本 1992) is a Japanese dependency grammar which builds a dependency structure from the phrases of a sentence. It uses global information to reduce the number of structures for a given sentence. For this purpose, a classification of phrases and arcs is used. Using constraints among arcs it's possible to treat adequately modifying, coordinated, and complex modification sentences.

RDG uses information extracted from **IPAL** (IPAL 1987, 1990) on verbs and adjectives, as well as a reduced classification of concepts. This classsification and restrictions allow a great number of structures which could be further reduced using a more elaborate dictionary. Contrary to verbs and adjectives, no restriction on nouns to noun dependencies is included, which leads to some impossible word combinations.

3 Electronic Dictionary

As part of the research in natural language, a machine readable dictionary for Japanese was built by **EDR Institute**, which consists to of a set of dictionaries.

The **Word Dictionary** includes syntactic information, definition, examples and asssociates each entry with an entry in the **Conceptual 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 (object) relations. A third dictionary is a list of word pairs (co-occurences), the relation which holds between the word pair, and a probability (0/1) of this relation. The following picture shows an extract from the (**CD**).

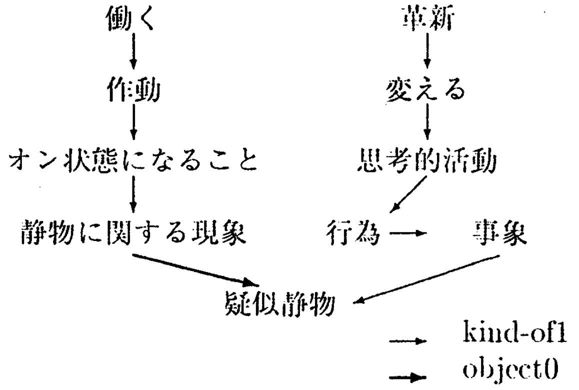

Fig. 2: Extract from the **Conceptual Dictionary**

4 Improving Efficiency

RDG is availiable from **ICOT**, and includes a set of 30 example sentences. According to the type of dependency (modification, coordination), it's possible to devise different approaches to reduce the number of solutions. For coordinate sentences, similarity between components may be used to choose among different structures (Kurohashi 1992). For modification dependencies coocurrence information may improve efficiency.

For the case of verbs and adjectives, the selectional restrictions used in **RDG** are not sufficient to avoid building some dependency that are semantically impossible. This is basically due to the fact that the classification of concepts used is insuficient. Refining these classification using the **CD** would improve drastically the performance of the system.

In case of verbs modifying nouns only syntactic restrictions are availiable. Ideally it's necessary to use word co-occurence information to cover all the cases because of the specificality of some constructions. Similarly, in noun-noun modifications, **RDG** includes no restriction, which can also be obtained from **Co-occurence Dictionary**.

5 Examples

In the analysis of the sentence

そのため酒やたばこが増え眠れなくなったり病気になったりする人も多い。

RDG provided 18 dependency structures in which the phrase 病気に may modify either なったり、する or 多い. Nevertheless, the pattern 病気にする doesn't appear in the coocurence dictionary. Also, 病気に人が多い can be avoided refining the selectional restrictions. These aditions can reduce the number of parses to 10.

Similarly, in the sentence

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つのってきた。

RDG provides 6 dependency structures in which the phrase 働く is modified by 技術革新が or 進んで and may modify 人の or ストレスも. Analysing the trees in the **CD** shown in fig 2, it's possible to deduce that 技術革新が cannot modify 働く. The figure shows only the concepts which link the two phrases, in which 働く has the sense of 作動. In the case of 進んで, an approach similar to that used to handle coordinate structures can be used since if this dependency is built 技術革新が would also modify 働く. Together these restrictions can reduce the number of dependency structures to two.

6 Future Research

From these few examples, a direction of research is shown, which points to the automatic selection from the dictionary of the necessary information to parse a given sentence. Since this information requires traversing the networks in conceptual dictionary and is of probabilistic nature, a probabilistic network seems to be the indicated framework to build such a system.

References

[1] 福本 文代, 佐野 洋, 斉藤 葉子, 福本 淳一 · 「係り受けの強度に基づく依存文法 − 制限依存文法 − 」 情報処理学会論文誌 33 巻 10 号 (1992)

[2] Kurohashi, S.,Nagao, M. , Dynamic Programming Method for Analysing Conjunctive Structures in Japanese, in *Proceedings of COLING-92. Nantes, August* (1992)

[3] Japan Electronic Dictionary Research Institute, Ltd., EDR 電子化辞書仕様説明書 (1993)

[4] 計算機用日本語基本動詞辞書 IPAL、情報処理振興事業協会技術センター (1987).

[5] 計算機用日本語基本形容詞辞書 IPAL、情報処理振興事業協会技術センター (1990).