The Selection of the Most Probable Dependency Structure in Japanese Using Mutual Information

Eduardo de Paiva Alves
University of Electro-Communications
1-5-1 Chofugaoka Chofushi Tokyo Japan
ealves@phaeton.cs.uec.ac.jp

Abstract

We use a statistical method to select the most probable structure or parse for a given sentence. It takes as input the dependency structures generated for the sentence by a dependency grammar, finds all triple of modifier, particle and modificant relations, calculates mutual information of each relation and chooses the structure for which the product of the mutual information of its relations is the highest.

1 Introduction

Computer Aided Instruction (CAI) systems are important and effective tools, especially for teaching foreign languages. Many students of Japanese as a foreign language are aware of the Computer Assisted TEchnical Reading System (CATERS) that provides helpful information for reading texts in science and technology fields (Kano and Yamamoto, 1995).

One of the difficulties in learning Japanese lies in recognizing dependency relations in Japanese sentences. This is because the language allows relatively free word orders. Take an example from a leading newspaper:

全国的な実態調査にもとづいた原因の速やかな究明を期待したい
We would like to expect a prompt study of the causes, based
on a national investigation

To understand this sentence it is necessary to know that 実態調査に (investigation) modifies もとづいた (based) but not 期待したい (expect); もとづいた (based) modifies 究明 (study) but not 原因 (cause). CATERS is useful because it provides such information through several user-friendly functions.

As effective as it is for foreign students, however, the texts in CATERS are fixed and the dependency structure of every sentence in them is all hand-coded. This inability to handle new text poses a serious problem in its general applicability and extensibility.

This paper describes a method for selecting the right or most probable structure for a Japanese sentence among multiple probable structures generated by Restricted Dependency Grammar (RDG) (Fukumoto, 1992). If this method works, then its results will be quite valuable for facilitating the development of new texts for CAI systems like CATERS.

2 Background

As pointed out earlier, the dependency relation of elements in Japanese sentences are fairly complicated due to relatively free word orders. RDG is designed to determine dependency relations among words and phrases in sentences. To do so, it classifies the phrases according to grammatical categories and syntactic attributes. However, it fails to reject semantically unacceptable dependency structures. The inevitable consequence is that RDG often produces multiple parses even for a simple sentence.

Kurohashi and Nagao (1993) try to determine the dependency relations of a sentence by means of using sample sentences. When the sentence is structurally ambiguous, they determine its structure by comparing it to structurally similar patterns taken from a manually generated set of examples and calculating similarity values.

Our method, on the contrary, uses a statistical approach to select the most probable structure or parse of a given sentence. It takes as input dependency structures generated by RDG for a sentence, finds all of modifier-particle-modificant relations, calculates their mutual information and chooses the structure for which the product of the mutual information of its relations is the highest.

In order to calculate the mutual information for any modifier-particle-modificant pattern, we use the Conceptual Dictionary[1] (CD) to build a taxonomic hierarchy of the modifiers which occur

[1] The Co-occurrence Dictionary and Conceptual Dictionary used in the process are part of a set of machine readable Japanese dictionaries compiled by the Japan Electronic Dictionary Research Institute (EDR, 1993). The Conceptual Dictionary is a set of graphs consisting of 400,000 concepts and a number of taxonomic as well as functional relations between them. The Co-occurrence Dictionary consist of a list of 1,100,000 dependency relations (modifier, particle and modificant) taken from a corpus. Each entry includes syntactic information, concept identifiers (a numerical code) and the number of occurrences in the corpus.

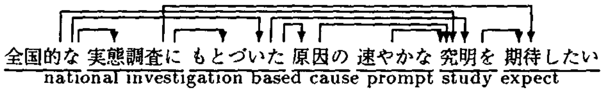
with the particle-modificant sub-pattern in the Co-occurrence Dictionary (COD). The mutual information for any pattern is the maximum mutual information between the sub-pattern and the concepts in the taxonomic hierarchy which generalize the modifier in the pattern.

Resnik and Hearst (1993) use a similar approach to calculate preferences for prepositional phrase attachment. While they use data on word groups, our method directly uses word co-occurrence data to estimate the preferences using the CD to identify the most adequate grouping for each relation.

While Kurohashi and Nagao compare the sentence with a single sample of patterns, we use all occurrences of the pattern in COD to calculate the mutual information. Our approach automatically extracts the occurrences from the dictionary as well as builds the taxonomic hierarchy. Unlike Kurohashi and Nagao (1993), which uses only verb and adjective patterns, we cover all dependency relations.

3 Selecting the Most Probable Structure

RDG identifies all possible dependency structures which consist of modifier-modificant relations between elements in a sentence. The arcs in the following example show modifier-modificant relations which can be combined into six different dependency structures.


全国的な 実態調査に もとづいた 原因の 速やかな 究明を 期待したい
national investigation based cause prompt study expect

Our objective is to develop a method to automatically select the correct dependency structures accurately or at least those which have the highest probability of being correct. We evaluate the various possible structures according to the mutual information between modifiers and particle-modificants. In some cases there is no particle and the modificant directly precedes the modifier (see example in section 3.2). To calculate the mutual information for each relation, we obtain form the COD the conceptual identifiers (a numerical code) for the modifiers that appear with the particle-modificant and the number of their occurrences in the corpus. If the pattern is not present, backing off, we search this information for the modificant only. For each of those concept identifiers we obtain from the CD all generalizers (concept identifiers that express a similar meaning in a more general way) and build a taxonomic hierarchy with them. Using the number of occurrences obtained, we calculate the mutual information for the concepts in the taxonomic hierarchy. We also build a taxonomic hierarchy for the modifier that appears with the particle-modificant in the sentence. Then comparing these two taxonomic hierarchies (one for the modifiers in the COD, one for the modifiers in the sentence), we look for the concept identifier common to both hierarchies that has the highest mutual information. This is the mutual information for the relation itself. For each dependency structure we calculate a score by multiplying the mutual information for all ambiguous relations (the non-ambiguous do not contribute to the evaluation). The dependency structure with highest probability of being correct is the one with the highest score. Since all structures have the same number of relations, this multiplication reflects the likelyhood of the structure.

3.1 The Algorithm

The process described above is written in an algorithmic form as follows:

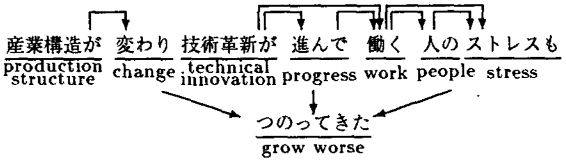
1. Select the ambiguous relations (those with more than one modificant) for each structure.
2. Search COD for the particle-modificant sub-pattern, in the corresponding positions. If there is no entry, search for the modificant only.
3. Obtain from the COD the concept identifiers for the modificant (there may be multiple meanings) and the concept identifiers with the number of their occurrences in the corpus for the modifiers which occur with the particle-modificant pattern.
4. For each modificant concept identifier, build a taxonomic hierarchy with its modifiers using CD to find the generalizer for each concept identifier.
5. Calculate the mutual information [2]
   for all the concept identifiers in the taxonomic hierarchies.
6. For the modifiers in the sentence, extract their concept identifiers from COD and build the taxonomic hierarchies using CD to find the generalizers for each concept identifier.
7. For each relation (modifier-particle-modificant pattern), search the concept identifier that generalizes the modifier word and has maximum mutual information. This value is the mutual information for the relation.
8. For each dependency structure, multiply the mutual information of its ambiguous dependency relations to obtain the score for that structure.
9. Arrange the structures according to their scores.

[2] The mutual information tells how much information one outcome gives about the other and is given by the formula:

$$I(w_1, w_2) = ln\left(\frac{P(w_1, w_2)}{P(w_1)P(w_2)}\right) \qquad (1)$$

3.2 Examples

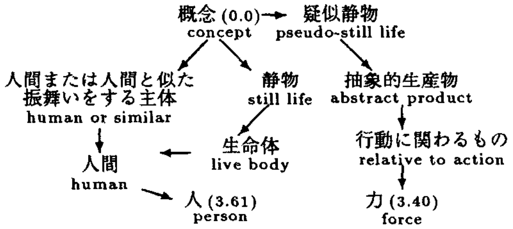
The following figure shows the output from RDG for a given sentence. The arrows in the figure indicate the dependency relations.



The ambiguous relations are 技術革新 が進んで, 技術革新 が働く, 働く 人, 働く ストレス, 進んで 働く and 進んでつのって. Accordingly the occurrences for the modificants in these relations (が進む, が働く, 働く, 進む, and つのる) are extracted from COD, obtaining a list of modifier concept identifiers with the number of their occurrences. Note that in the pattern 働く 人 and 働く ストレス the modificant precedes the modifier. The following figure shows some modifiers for 働く (work) with their number of occurrences.

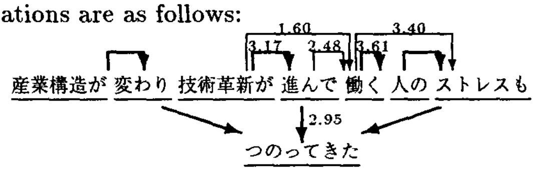
人	女性	母親	意欲	自分	者	工場	主婦	事情	作業員
person	woman	mother	drive	each	person	factory	wife	fact	worker
32	18	6	6	3	3	3	2	2	2

Next, the taxonomic hierarchy for each particle-modificant is built and the mutual information calculated for each concept identifier. An extract of the hierarchy for 働く is shown in the following figure.

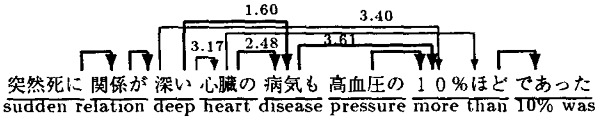


Next the generalizers for (革新, 人, and ストレス) are searched in the hierarchies for their modificants to obtain the mutual information for the relations. For 働く人 (working person) it happened to be the concept 人 (person) itself with mutual information of 3.61. For 働くストレス (working stress) the match occurred for 力 (force) giving a mutual information of 0.69.

Multiply the mutual information for all the dependency relations in each structure. For the example sentence the mutual information for the ambiguous relations are as follows:



From this the algorithm selects the parse with highest score which is drawn in thick lines. The next figure shows the result for the first example sentence.



4 Results and Evaluation

We have applied our method to 35 sentences taken from a leading newspaper and included with RDG software. The average number of dependency structures per sentence is 8.68. The method we used selected the correct structures for 25 sentences. The correct structures for 8 sentences were found as the second most probable structure by the method.

In another experiment, we parsed 70 sentences using a grammar similar to the one used in Kurohashi and Nagao (1993). Our method selected the most likely relation among the multiple generated in 95% of the cases.

Although the size of the test data is small, we say that our method provided a way to identify the most probable structure more efficiently than RDG. Since the sentences used are extracted from a newspaper, it's also general in its applicability. Therefore it can be used in preparing teaching materials such as the structures used by a CAI system such as CATERS, saving the instructor of hand-coding them. In future work we shall extract the co-occurrences directly from the corpora, and use other grouping techniques to replace the CD.

5 Acknowledgments

I am thankful to my thesis advisor Dr. T. Furugori and the anonymous referees for their suggestions and comments.

References

Fukumoto, F.; Sano H., Saitoh, Y.; and Fukumoto J. 1992. A Framework for Dependency Grammar based on the word's modifiability level - Restricted Dependency Grammar. In *Trans. IPS Japan*, 33(10), (in Japanese).

Resnik,P. and Hearst M. 1993. Structural Ambiguity and Conceptual Relations. In *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives*. Ohio State University.

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

Kano, C. and Yamamoto, H. 1995. A System for Reading Scientific and Technical Texts, Classroom, Instruction and Evaluation. In *Jinbunkagaku to computer* 27(1) (in Japanese).

Kurohashi, S., and Nagao, M. 1993. Structural Disambiguation in Japanese by Evaluating Case Structures based on Examples in Case Frame Dictionary. In *Proceedings of IWPT93*.