Sharing Syntactic Structures

Masahiro Ueki
Tokyo Institute of Technology

Takenobu Tokunaga
Tokyo Institute of Technology

Hozumi Tanaka
Tokyo Institute of Technology

Abstract

Bracketed corpora are a very useful resource for natural language processing, but hard to build efficiently, leading to quantitative insufficiency for practical use. Disparities in morphological information, such as word segmentation and part-of-speech tag sets, are also troublesome. An application specific to a particular corpus often cannot be applied to another corpus.

In this paper, we sketch out a method to build a corpus that has a fixed syntactic structure but varying morphological annotation based on the different tag set schemes utilized. Our system uses a two layered grammar, one layer of which is made up of replaceable tag-set-dependent rules while the other has no such tag set dependency.

The input sentences of our system are bracketed corresponding to structural information of corpus. The parser can work using any tag set and grammar, and using the same input bracketing, we obtain corpus that shares partial syntactic structure.

1 Introduction

The ready availability of large corpora, especially bracketed corpora, facilitates corpus-based research such as probabilitic parsing. However each corpus has its own part-of-speech tag sets and notation schemes. Corpus-based markup schemes can become customized to a specific corpus, and incompatible with other corpora with different tag sets or notation schemes.

A number of morphological information mapping methods have been proposed[5][6][1]. Mapping systems have rewrite rules that are derived automatically or manually, and map part-of-speech tags word by word. But mapping between part-of-speech tags, for example noun to pronoun, cannot be performed because of the large numbers of words with multiple parts-of-speech. Conventional rewrite rules consider both the word itself and its original part-of-speech tag in selecting all possible alternative tags. But, in practical use, this method has drawbacks.

We consider that dependencies between phrases coincide between tagging schemes, and should help to solve the above problem. Thus, given a structured corpus, we can use its structural information to relabel it according to a second tag set. In this paper, we sketch out a method to build a corpus that partially shares syntactic structures and has alternative part-of-speech tags for each word as stipulated by different part-of-speech tag sets. Sentences in the source corpus are bracketed to describe the basic phrase dependencies, and analyzed by a parser. The parser can work using any tag set and grammar, using the same input bracketing. As a result, we obtain an efficient corpus representation.

2 Our Method

2.1 Two Layered Grammar

First of all, let us look at examples of a parse tree.(Figs.1,2)

These two trees describe possible syntactic structures of the Japanese sentence *kare ga watashi ni atarashi-i jisho wo kure ta* "He gave me a new dictionary", where the auxiliary *ta* is the past tense marker. One difference between the two parse trees is the interpretation of this auxiliary.

Semantically, the structure in Fig.1 tells us that the event "He gives me a new dictionary" occurred in the past, while that in Fig.2 only shows the occurence of the action "gave". The overall meanings are not so very different, but structures like that in Fig.1 are difficult for parsing systems to produce correctly.

Now, let us compare these two structures from a grammatical point of view. The black circles in these figures denote nodes which have words as direct children (possibly with part-of-speech tags), with the subtrees subsumed by each such node corresponding to a grammar rule dependent on the part-of-speech tag set. In Fig.1, such nodes are scattered across the whole tree, and cannot be distinctly separated from nodes without a word as a direct child. In this case, differences in the part-of-speech tag set influence the whole

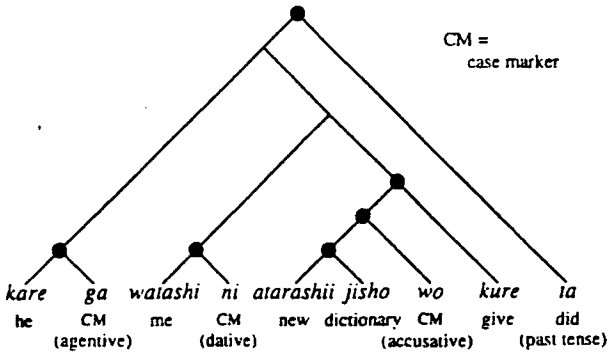


Figure 1: An example of a parse tree (1)

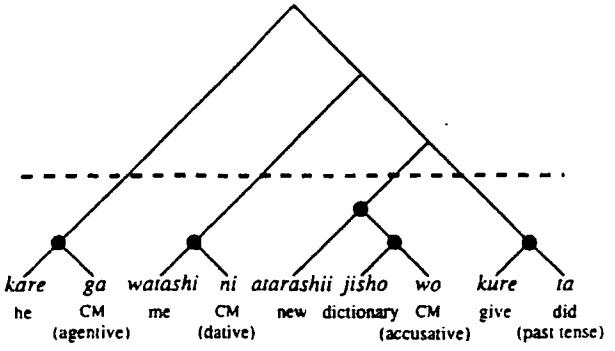


Figure 2: An example of a parse tree (2)

grammar. When we want to apply a second tag set, we have to prepare a separate grammar. On the other hand, markers in Fig.2 are clustered in the lower reaches of the tree, and the structure in the upper reaches gives us dependencies between phrases, and would seem to be tag set independent. We can easily distinguish between these two levels, and the same can be applied to grammars to produce two distinct parts.

We call these two levels the intra-phrase grammar and the inter-phrase grammar. The former is a set of rules which combines sequences of words into phrases, such as postpositional phrases and predicate phrases, and the latter describes dependencies between phrases. Each phrase holds information as to what kind of phrases it can modify and be modified by. This information can be considered as dependency restrictions that every phrase must satisfy.

We consider the inter-phrase grammar to be free of tag-set-dependencies, and the inter-phrase structure to be a common reusable structure across all tag sets.

2.2 MSLR Parsing System

The MSLR parsing system is based on the GLR parsing method, and integrates morphological and syntactic analysis.[2] In this system, morphological constraints are described as connection constraints between part-of-speech tags, and integrated into the LR parsing table. Connection constraints are described in the form of a connection matrix. This method has two advantages for our method.

1. Connection constraints greatly simplify intra-phrase grammar descriptions.
2. Given a tagged corpus, connection constraints can be produced automatically.

To describe a grammar for an unfamiliar tag set is difficult work, but the above factors make it much easier.

We extended the MSLR system to accept bracketed input sentences that specify dependency restrictions between phrases, and to output all possible parse trees that satisfy those restrictions.

3 Experiment

3.1 Corpus features and experiment methodology

We used structural information from the EDR corpus[3] as the shared syntactic structure. All trees are first transformed to structures separable into a tag-set-dependent and tag-set-independent component, as described in the previous section, and finally to bracketed sentences to form input strings for the parsing system.

The intra-phrase grammar is described by the tag set for the RWC corpus[4], because the RWC tag set is much more detailed than that of the EDR corpus[1]. However as no word dictionary is available for the RWC tag set, we extracted all word-tag pairs from the RWC corpus to use as a dictionary. Connection constraints between part-of-speech tags were also extracted. Almost all high-frequency words, particles, auxiliaries, etc., could be obtained, but nouns were still insufficient. We thus extracted all proper nouns from the EDR corpus and used them to supplement the original dictionary.

3.2 Result and evaluation

We applied the RWC tag set and dictionary to 13047 bracketed sentences from the EDR corpus, with the results shown in Table 3. The success rate of 74.8% does not seem to high. We randomly selected 500 sentences and examined them to determine the reason for this.

Of the 500 sentences, 362 (72.4%) were parsed successfully, but 17 sentences had incorrectly tagged words. In most cases, verbal words were tagged as nouns. The overall success rate can be estimated to be about 70%. On the other hand, 138 sentences could not be parsed for the following reasons.

[1]The number of parts-of-speech need to tag the RWC corpus is almost the same as for EDR, but they can be divided into detailed tags using grammatical information.

	EDR	RWC
# of morphemes	5×10^6	1×10^6
Morphological information	○	○
Structural information	○	×
# of part-of-speech tags	15	383

Table 1: Features of the two corpora

Inter-phrase grammar (# of rules)	371
Intra-phrase grammar (# of rules)	304
# of dictionary entries	323683

Table 2: Features of the grammar and dictionary

Success (sentences)	9760	(74.8%)
Failure (sentences)	3287	(25.2%)
Total (sentences)	13047	

Table 3: Parsing results using the RWC tag set

- Insufficiency of dictionary coverage
- Overconstraint in the connection matrix
- Incorrect bracketing of the source corpus

A tag-set-independent shared structure is shown in Fig.3(a), with the corresponding tag-set-dependent structures in Fig.3(b) and Fig.3(c). P1, P2, ···, P5 denote the phrase nodes and can be replaced by any intra-phrasal structure.

P4 is an adjectival phrase and modifies the proceeding noun in both tag-set-dependent structures, but the intra-phrase structures are somewhat different, in that the EDR tag set divides verbs and adjectives into a stem and inflecting suffix, while the RWC tag set does not. The three nouns in the source sentence are classified as either a pronoun or common noun.

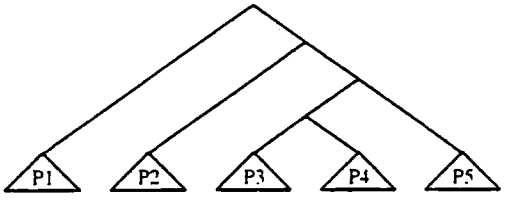
Given structural information, the parsing system can select the appropriate part-of-speech tagging. For example, *John to Mary ga kekkon shita* has two different interpretations.

1. "To John, Mary got married."
2. "John and Mary got married (to their respective partners)."

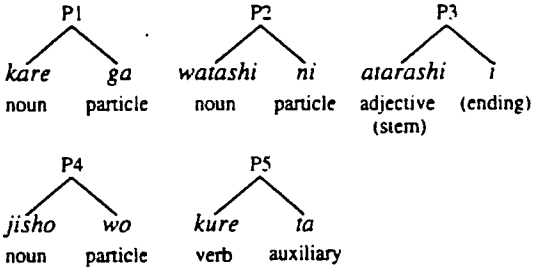
The word *to* functions both as an comitative case marker (interpretation 1) and as a coordinate conjunction (interpretation 2). Word-level n-grams cannot help to distinguish between these two, because there is no difference in lexical appearance. Given structural information like:

[[*John to*] [[*Mary ga*] [*kekkon shita*]]]

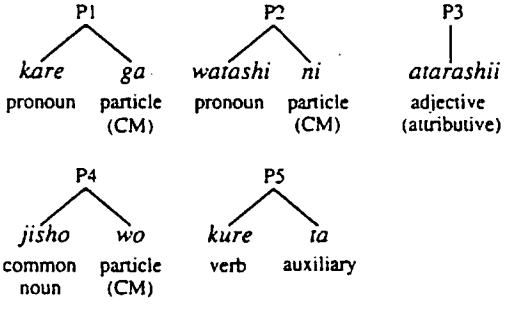
only interpretation 1 is correct.



(a) Shared structure



(b) The EDR tag-set-dependent structures



(c) The RWC tag-set-dependent structures

Figure 3: Comparison between EDR and RWC structures

4 Discussion

1. Word boundary ambiguities

In Japanese, ambiguous word boundaries are also problem. In the experiment, to avoid conflicts in word boundaries between the source corpus and target dictionary, all morphological information from the source corpus was discarded. Most particles and auxiliaries were segmentated as in the original corpus, but the segmentation of compound nouns could not be determined without semantic analysis, resulting in combinatorial ambiguity.

Nominals are very large in number, which causes problems such as in the analysis of unknown words and combinatorial segmentation ambiguity. As such, the morphological information of the source courpus

must be relied upon to some degree.

2. Phrase-boundary ambiguity
Different tag sets can sometimes lead to differences in phrase boundaries. In most cases, what is one phrase for some tag set can be divided into two or three phrases for another. Idiomatic expressions provide the most frequent occurence of this situation.
An adjustable layer boundary may be the solution to this problem, but automatic adjustment seems difficult.

Only item 2 above seems to pose a real problem for our approach and require further consideration.

5 Conclusion

In this paper, we proposed a method for building a corpus with shared syntactic structure for any tag set, by which detailed functional words can be disambiguated correctly.
As future work, it is important to evaluate hybrid methods that map part-of-speech tags between different noun subtypes and annotate other categorical words using structural information.

References

[1] John Hughes, Clive Souter, and Eric Atwell. Automatic extraction of tagset mappings from parallel-annotated corpora. In *EACL 95*, 1995.

[2] LI Hui and TANAKA Hozumi. A method for integrating the connection constraints into an LR table. In *Proceedings of Natural Language Processing Pacific Rim Symposium*, pages 703-708, 12 1995.

[3] Japan Electronic Dictionary Research Institute, Ltd. *EDR Electronic Dictionary Technical Guide*, 1995.

[4] Real World Computing Partnership. RWC text database. Technical report, (In Japanese), 1997.

[5] Toshihisa Tashiro, Noriyoshi Uratani, and Tsuyoshi Morimoto. Restructuring tagged corpora with morpheme adjustment rules. In *COLING 94*, 1994.

[6] Simone Teufel. A support tool for tagset mapping. In *Workshop SIGDAT (EACL 95)*, 1995.