Restructuring Tagged Corpora with Morpheme Adjustment Rules

Toshihisa Tashiro *Noriyoshi Uratani* *Tsuyoshi Morimoto*
ATR Interpreting Telecommunications Research Laboratories
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02, JAPAN
{tashiro,uratani,morimoto}@itl.atr.co.jp

Abstract

A part-of-speech tagged corpus is a very important knowledge source for natural language processing researchers. Today, several part-of-speech tagged corpora are readily available for research use. However, because there is wide diversity of morphological information systems (word-segmentation, part-of-speech system, etc.), it is difficult to use tagged corpora with an incompatible morphological information system. This paper proposes a method of converting tagged corpora from one morpheme system to another.

1 Introduction

Recently, many natural language processing researchers have concentrated on corpus-based approaches. Linguistic corpora can be classified as word-segmented corpora, part-of-speech tagged corpora, and parsed corpora. Because a part-of-speech tagged corpus is the most important corpus, much corpus-based natural language processing research has been performed using part-of-speech tagged corpora.

However, building a large part-of-speech tagged corpus is very difficult. It is even more difficult to build a corpus for languages without explicit word boundary characters, such as Japanese. Therefore, researchers always complain of the scarcity of data in the corpus.

To solve this data scarcity problem, previous works proposed methods of increasing the productivity of the labor required for building a part-of-speech tagged corpus. [1].

This paper proposes another method of acquiring large part-of-speech tagged corpora: restructuring tagged corpora by using morpheme adjustment rules. This method assures good use of the sharable part-of-speech tagged corpora that are already available such as the ATR Dialog Database (ADD) [2; 3].

Ideally, these corpora could be used by all researchers and research groups without any modifications. However, actual part-of-speech tagged corpora have the following problems:

- Diversity of orthography:
  A word can be spelled in various ways. In Japanese, there are three types of character sets: kanji (漢字), hiragana (ひらがな), and katakana (カタカナ). Also, people can use these character sets at their discretion.
- Diversity of word segmentation:
  Because the Japanese language has no word boundary characters(i.e. blank spaces), there are no standards of word segmentation. A single word in a certain corpus may be considered multiple words in other corpora, and vice versa.
- Diversity of part-of-speech systems:
  There are no standards for part-of-speech systems. It is true that a detailed part-of-speech system can help the application of part-of-speech information, but the labor required for building corpora will continue to increase. This problem is language-independent.

Diversities of word-segmentation and part-of-speech systems are fatal problems. The simplest way to solve these problems is to perform a morphological analysis on the raw text in the corpus, with no regard to the word-segmentation and part-of-speech information. However, making a high-quality morphological analyzer demands much time and care. Additionally, it is wasteful to ignore the word-segmentation and part-of-speech information that has been acquired with much effort.

In restructuring tagged corpora with morpheme adjustment rules, the word-segmentation and part-of-speech information of the original corpus is rewritten, making good use of the original corpus information. This method is characterized by reduced manual effort.

In the next section, the method of restructuring tagged corpora is described in detail. Section 3 reports the result of an experiment in rewriting the corpus using this method.

2 Restructuring Tagged Corpora

Restructuring tagged corpora involves the following three steps:

- preparation of training set
- extraction of morpheme adjustment rules
- rewriting of corpora

2.1 Preparation of Training Set

First, sentences for the training set are chosen from the corpus to be rewritten. New word-segmentation and part-of-speech information (morphological information) is given to the sentences by a morphological analyzer or by hand. Consequently, the training set has two sets of morphological information for the same raw text. Figure 1 shows an example of the training set.

A large number of training sentences is desirable, but preparing many sentences requires much time and effort. A vast number of sentences would be required to extend coverage to content words (such as nouns, verbs, etc), but functional words (such as particles, auxiliary verbs, etc) can be covered with a smaller number of sentences.

raw text	割引料金はあるんでしょうか
morphological information A	(割引料金 n-com)(は postp-topic) (あ vstem)(る vinfl)(ん fn) (でしょう auxv-polt-aux-nom) (か auxv-sfp-1)
morphological information B	(割引 普通名詞)(料金 普通名詞) (は 係助詞) (ある 本動詞) (ん 準体助詞)(でしょ 助動詞) (う 助動詞)(か 終助詞)

Figure 1: An Example of the Training Set

2.2 Extraction of Morpheme Adjustment Rules

The method of extracting morpheme adjustment rules from the training set involves finding correspondence between rewriting units and extracting rules for unknown words:

2.2.1 Correspondences of Rewriting Units

In languages without explicit word boundary characters, such as Japanese, a single word in a certain morphological information system may be divided into multiple words (one-to-many correspondence) in other morphological information systems, multiple words may be unified (many-to-one correspondence), or the segmentation of multiple words may be changed (many-to-many correspondence). Figure 2 shows these correspondences.

We developed an algorithm to find these correspondences (Appendix A). By using this algorithm, morpheme rewriting rules (Figure 3) can be extracted.

2.2.2 Rules for Unknown Words

Rewriting rules such as those shown in Figure 3 can rewrite only the words that appeared in the training set. If the training set is small, the coverage of the rules will be limited. However, because this morpheme adjustment is a method of rewriting part-of-speech tagged corpora, the treatment of unknown

[one-to-one]
"用紙 (NOUN)" ↔ "用紙 (NOUN)"
[one-to-many]
"送る (VERB)" ↔ "送 (V-STEM)" "る (INFL)"
[many-to-one]
"登録 (NOUN)" "用紙 (NOUN)" ↔ "登録用紙 (NOUN)"
[many-to-many]
"て (PARTICLE)" "いる (VERB)"
↔ "ていん (AUXV-STEM)" "る (INFL)"

Figure 2: Various correspondences

て (接続助詞) いる (補助動詞)
↔ ていん (auxvstem-aspc) る (vinfl)
を (格助詞) ↔ を (postp-oblg)
送っ (本動詞) ↔ 送 (vstem) っ (vinfl)
二十一 (数詞) ↔二 (n-digit) 十 (digit-suffix-zyuu) 一 (n-digit)
申込み (普通名詞) 用紙 (普通名詞) ↔ 申込み用紙 (n-com)

Figure 3: An Example of Extracted Rules

words is easier than with an ordinary morphological analyzer, because that our method can make good use of the part-of-speech information of the original corpus. Rules for unknown words without word-segmentation changes between two morphological information systems can be extracted automatically from one-to-one correspondence rules in the rewriting rules.

Rules for unknown words with word-segmentation changes can also be extracted automatically by using information concerning the length of the word's characters. For example, when a single verb with two characters in a certain morphological information system corresponds to two words (verb-stem with one character and verb-inflection with one character) in another morphological information system, the following rewriting rule is extracted.

2(verb) → 1(verb-stem) 1(verb-inflection)

Figure 4 shows sample rules for unknown words.

The heuristic knowledge of character sets that an ordinary Japanese morphological analyzer uses (such as "katakana words are usually proper nouns", "verb inflection words are spelled using hiragana", etc.) are also available in this morpheme adjustment technique.

2.3 Rewriting of Tagged Corpora

2.3.1 Application of Rewriting Rules

By applying the rewriting rules described in the last subsection to the tagged corpus, a lattice structure

格助詞 ↔ FADN
格助詞 ↔ POSTP-OBLG
2(本動詞) 1(補助動詞) ↔ 3(VSTEM) 1(VINFL)
2(副助詞) ↔ 1(POSTP-OPTN) 1(POSTP-CONTR)

Figure 4: An Example of Rules for Unknown Words

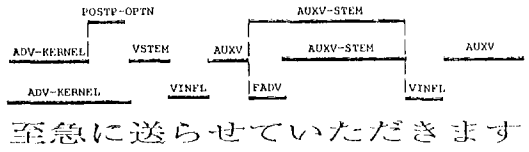

Figure 5: An Example of the Lattice Formed by the Morpheme Adjuster

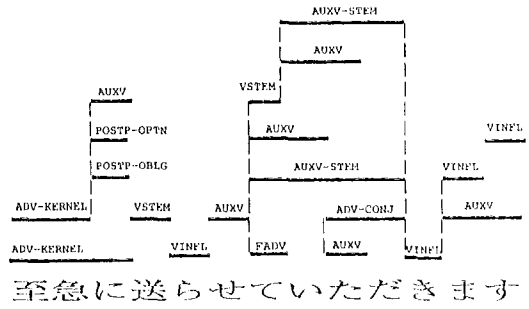

Figure 6: An Example of the Lattice Formed by an Ordinary Morphological Analyzer

(Figure 5) is formed because of the ambiguity in rewriting rules.[1]

However, this ambiguity is not as great as the ambiguity that occurs in ordinary morphological analysis because our method makes good use of the information of the original corpus. Figure 6 shows the lattice structure formed when using the ordinary morphological analysis on the same raw text. Note that the size of this lattice is greater than the size of the lattice made by our method.

2.3.2 Lattice Search

The last step in restructuring tagged corpora can be considered a lattice search problem. In this step, all of the following knowledge sources for ambiguity resolution used in ordinary morphological analysis is also available in our method:

- connection matrix
- heuristic preferences (longest word preference, minimum phrase preference, etc.)
- stochastic preferences (word n-gram, HMM, etc.)

By using these knowledge sources, the most plausible candidate is chosen. In effect, the original corpus is converted to a new corpus that uses a different morphological information system.

3 Experiment

3.1 Experimental Condition

The targets in our experiment are a morphological information system for the ATR Dialog Database[2; 3] and a morphological information system for the unification-based Japanese grammar used in ATR's spoken language parser[4]. These two morphological information systems have the following characteristics.

- The ATR Dialog Database was developed as material for analyzing the characteristics of spoken-style Japanese. Therefore, the part-of-speech granularity is coarse. Additionally, because the word-segmentation is based on a morphological and etymological criterion, compound nouns and compound words that function as a single auxiliary verb (e.g. "ている") are divided into several shorter word units. On the other hand, because this database gives little consideration to mechanical processing, stems and inflections of inflectional words are not segmented.
- The unification-based Japanese grammar has a medium-grained part-of-speech (pre-terminal) system to make it both efficient and easy to maintain[5]. Because the objective of the grammar is to extract the syntactic structures of Japanese sentences automatically and efficiently, compound words that function as a single word are usually recognized as a single word. On the other hand, stems and inflections of inflectional words are segmented for convenience of mechanical processing.

The above descriptions show that these morphological information systems differ. The objective of this experiment is to examine whether our method can adjust the differences between the two morphological information systems to a considerable extent.

First, we chose 1,000 sentences from the ATR Dialog Database as the training set and provided the morphological information (word-segmentation and part-of-speech) of the unification-based Japanese grammar. We prepared 350 sentences as the test set, separate from the training set. The test sentences were also given the morphological information.

We extracted 1,538 correspondences of rewriting units (i.e. rewriting rules) and 428 rules for unknown words. These rules can be used for the bi-directional rewriting experiment.

As the knowledge source in searching lattices, word bigrams and part-of-speech bigrams were trained with the training set. To perform the bi-directional rewriting experiment, these bigrams were trained in both morphological information systems.

To compare our method with ordinary morphological analysis, we developed a simple stochastic morphological analyzer that uses the same bigrams as the knowledge sources[2]. Because this morphological analyzer has been developed for the comparative experiment, it cannot manage unknown words. Therefore, the rewriting test was performed by using not only the

[1] This ambiguity mainly comes from the difference in part-of-speech granularity between the two morphological information systems.

[2] Of course, the ordinary morphological analyzer can rewrite the corpus much more accurately by using richer knowledge sources. However, it must be noted that our method also can use such knowledge sources.

Morphological Information	Unification-Based Japanese Grammar	ATR Dialog Database
Training Set sentences (words)	1,000 (10,510)	1,000 (10,723)
Test Set (Full) sentences (words)	350 (3,894)	350 (4,066)
Test Set (Sub) sentences (words)	148 (904)	148 (949)
Vocabulary	1,284	1,168
POS System	75	26
Word Bigram	4,325	4,292
POS Bigram	503	262

Table 1: Experimental Condition

test sentences, but also the training sentences (close experiment) and the sentences having no unknown words (a subset of the test set).

Table 1 shows the experimental conditions in detail.

3.2 Rewriting of Morphological Information

The experiment was performed bi-directionally between the morphological information system of the ATR Dialog Database (ADD) and the morphological information system of unification-based Japanese grammar.

3.2.1 From Unification-Based Grammar to ADD

This experiment rewrites from a medium-grained morphological information system to a coarse-grained morphological information system. Table 3.2.1 shows the result of this rewriting. The segmentation error rate and part-of-speech error rate were calculated using the same definition in [1]. Table 2 shows the result.

The error rates seem to be rather large, but it should be noted that only simple knowledge sources are used both in our method (the morpheme adjuster) and by the morphological analyzer. Also, it is significant that our targets are spoken-style Japanese sentences. Ordinary morphological analyzers can analyze written-style Japanese sentences with a less than 5% error rate, by using richer knowledge sources[1]. However, previous work reported that the error rate for automatic morphological analysis of the ADD text is more than 15%[6].

In comparing the two methods, the part-of-speech error rates of our method are clearly better than those of the morphological analyzer. This shows that our method can make good use of the original part-of-speech information.

3.2.2 From ADD to Unification-Based Japanese Grammar

This experiment is more difficult because this rewriting is from the coarse-grained morphological information system to the medium-grained morphological information system. Table 3 shows the result.

The part-of-speech error rates of our method are better in this rewriting experiment, too.

4 Conclusion

This paper proposed restructuring of tagged corpora by using morpheme adjustment rules. The eventual goal of this work is to make precious knowledge sources truly sharable among many researchers. The results of the experiment seem promising.

Our morpheme adjustment method has some resemblance to Brill's part-of-speech tagging method[7]. Brill's simple part-of-speech tagger can be considered a morpheme adjuster that adjusts differences between initial (default) tags and correct tags.

As Brill applied his part-of-speech tagging technique to the syntactic bracketing technique[8], we believe that our method can be applied to the adjustment of parsed corpora. In the work of Grishman et al.[9], tree rewriting rules to adjust differences between Tree Bank and their grammar were probably prepared manually. By applying our method to parsed corpora, such rewriting rules can be extracted automatically.

Acknowledgments

The authors would like to thank Dr. Yasuhiro Yamazaki, President of ATR Interpreting Telecommunications Laboratories, for his constant support and encouragement.

References

[1] Maruyama, H., Ogino, S., Hidano, M., "The Mega-Word Tagged-Corpus Project," TMI-93, pp.15-23, 1990

[2] Ehara, T., Ogura, K. and Morimoto, T. "ATR Dialogue Database," ICSLP-90, pp.1093-1096, 1990.

[3] Sagisaka, Y., Uratani, N., "ATR Spoken Language Database," The Journal of the Acoustical Society of Japan, Vol. 48, 12, pp. 878-882, 1992. (in Japanese)

[4] Nagata, M. and Morimoto, T.: "A Unification-Based Japanese Parser for Speech-to-Speech Translation," IEICE Trans. Inf. & Syst., Vol.E76-D, No.1, pp.51-61, 1993.

[5] Nagata, M. "An Empirical Study on Rule Granularity and Unification Interleaving. - Toward an Efficient Unification-Based Parsing System," in Proc. of COLING-92, 1992.

[6] Kita, K., Ogura, K., Morimoto, T., Yano, Y., "Automatically Extracting Frozen Patterns from Corpora Using Cost Criteria.", IPSJ Trans. Vol.34, No.9,pp.1937-1943, 1993.(in Japanese)

[7] Brill, E.,: "A Simple Rule-Based Part of Speech Tagger," Proceedings of the Third Conference on Applied Natural Language Processing, 1992.

[8] Brill, E., "Automatic Grammar Induction and Parsing Free Text: Transformation-Based Error-Driven Parsing," ACL93, 1993.

[9] Ralph Grishman, Catherine Macleod and John Sterling "Evaluating Parsing Strategies Using Standardized Parse Files," Proceedings of the Third Conference on Applied Natural Language Processing, pp.156-161,1992.

Method	segmentation error rate	part-of-speech error rate	Total
Test Set (Full)	7.8%	2.8%	10.6%
Test Set (Sub)			
Morpheme Adjuster	5.1%	1.5%	6.6%
(Morphological Analyzer)	(6.3%)	(3.4%)	(9.7%)
Training Set(close test)			
Morpheme Adjuster	0.2%	1.5%	1.7%
(Morphological Analyzer)	(1.3%)	(3.7%)	(5.0%)

Table 2: From Unification-Based Grammar to ADD

Method	segmentation error rate	part-of-speech error rate	Total
Test Set (Full)	8.2%	6.9%	15.1%
Test Set (Sub)			
Morpheme Adjuster	4.2%	3.1%	7.3%
(Morphological Analyzer)	(8.5%)	(6.8%)	(15.3%)
Training Set (close test)			
Morpheme Adjuster	0.5%	3.3%	3.8%
(Morphological Analyzer)	(0.5%)	(6.3%)	(6.8%)

Table 3: From ADD to Unification-Based Grammar

Appendix
A.The Rule Extraction Algorithm

```
type
    word = record
        symbol          :string;  {ex. "会議"}
        part-of-speech :string;  {ex. "NOUN"}
    end
    wordlist = record
        elem : array[1..MAXLENGTH] of word
        last : integer
    end
procedure FIND_CORRESPONDENCES(A,B:wordlist);
   {The arguments of this  procedure are
    two kinds of morphological information
    of the same sentence .
    For example:
       A = (わか vstem)(り vinfl)
           (ま auxv-stem)(し auxv-infl)
           (た auxv-tense)
       B = (わかり 本動詞)(まし 助動詞)
           (た 助動詞)
    The OUTPUT subroutine outputs the correspon-
    dences such as:
      (わか vstem)(り vinfl)<--> (わかり 本動詞)
    Because the total "LENGTHs" of two arguments
    are the same, this algorithm is guaranteed to
    be completed normally.}
   var
      lhs,rhs: wordlist;
      cur_a, cur_b: integer; {cursors}
   begin
      cur_a := 1;  cur_b := 1;
      lhs.last := 1;  rhs.last := 1; {Initialize}
      lhs.elem[lhs.last] := A.elem[cur_a];
      lhs.last := lhs.last+1;
      rhs.elem[rhs.last] := B.elem[cur_b];
      rhs.last := rhs.last+1;
      while ( A.last > cur_a )  do begin
         if LENGTH(lhs) = LENGTH(rhs) then begin
            OUTPUT(lhs, rhs);
            cur_a := cur_a+1; cur_b := cur_b+1;
            Initialize(lhs,rhs);
            lhs.elem[lhs.last] := A.elem[cur_a];
            lhs.last := lhs.last+1;
            rhs.elem[rhs.last] := B.elem[cur_b];
            rhs.last := rhs.last+1;
         end
         else if LENGTH(lhs) > LENGTH(rhs) then beg
            cur_b := cur_b+1;
            rhs.elem[rhs.last] := B.elem[cur_b];
            rhs.last := rhs.last+1;
         end
         else begin
            cur_a := cur_a+1;
            lhs.elem[lhs.last] := A.elem[cur_a];
            lhs.last := lhs.last+1;
         end
      end;
function LENGTH(A: wordlist);
   {This function returns the total length of
    wordlist. When the arg is "((わ
か vstem)(り vinfl))",
    this function returns 3.}
   var
      length,count : integer;
   begin
      length := 0;
      for count := 1 to A.last do
         length := length+|A.elem[count].symbol|;
      return length;
   end
```