Machine Aided Error-Correction Environment for Korean Morphological Analysis and Part-of-Speech Tagging

Junsik Park, Jung-Goo Kang, Wook Hur and **Key-Sun Choi**
Center for Artificial Intelligence Research
Korea Advanced Institute of Science and Technology
Taejon 305-701, Korea
{jspark,jgkang,hook,kschoi}@world.kaist.ac.kr

Abstract

Statistical methods require very large corpus with high quality. But building large and faultless annotated corpus is a very difficult job. This paper proposes an efficient method to construct part-of-speech tagged corpus. A rule-based error correction method is proposed to find and correct errors semi-automatically by user-defined rules. We also make use of user's correction log to reflect feedback. Experiments were carried out to show the efficiency of error correction process of this workbench. The result shows that about 63.2 % of tagging errors can be corrected.

1 Introduction

Natural language processing system using corpus needs the large amount of corpus (Choi et al., 1994), but it also requires the high quality.

The process of making the general annotated corpus can be viewed as Figure 1. There are some difficulties in processing the annotated corpus. First, the number of items in a dictionary is not so large. The second problem is in the difficulty of modifying the errors produced by automatic tagging. Manual error correction would require large amount of costs, and there may still remain errors after correcting process. There were also researches about automatic correction, but they had problems about the side-effects after automatic error correction (Lee and Lee, 1996; Lim et al., 1996).

In this paper, we will integrate the morphological analysis and tagging, and provide interactive user interface. User gives the feedback to resolve the ambiguities of analysis. To reduce the cost and improve the correctness, we have developed an environment which is enable to find errors and modify them.

In the following section, related works are described. In section 3, we propose our model. Then, implementation and experiment results are explained. Finally, discussion is followed.

2 Related Works

An automatic tagging is prone to errors that cannot be avoidable due to the lack of overall linguistic information. To model the automatic error-detection process, the statistical approach of detecting tagging error has been developed (Foster, 1991). In this section, we will describe some approaches about rule-based error correction method for Korean part-of-speech(hereafter, "POS") tagging system.

2.1 Transformation-Based Part-of-Speech Tagging System

(Lim et al., 1996) proposed tagging system that uses word-tag transformation rules dealing with agglutinative characteristics of Korean, and also extends the tagger by using specific transformation rule considering the lexical information of mistagged word.

General training algorithm of the transformation rule (Brill, 1993) is as follows:

1. Train initial tagger on initial training corpus C_0.
2. Make Confusion matrix with the result of comparing the current training corpus C_i (initially, $i = 0$) and C'_0, the output of a manual annotation on C_0.
3. Extract rules correcting the errors of Confusion matrix best.
4. Apply the extracted tagging rules to the training corpus C_i and generate improved version C_{i+1}.
5. Save the rule and increase i.

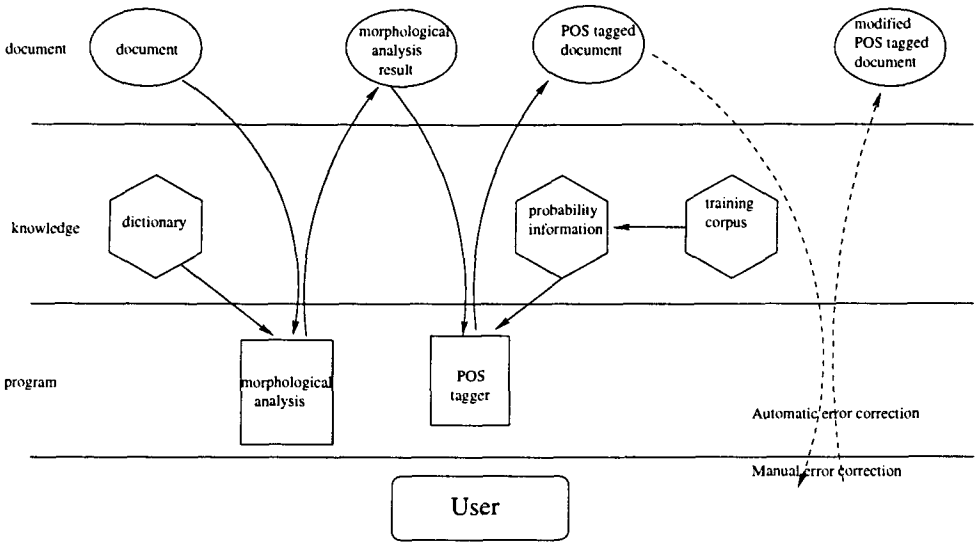

Figure 1: Process of making part-of-speech tag annotated corpus

6. Repeat steps 2 to 5 until frequency of error correction, which is done by rules found in the previous step, is less than threshold.

2.2 Rule-based Error Correction

This method (Lee and Lee, 1996) is based on Eric Brill's tagging model (Brill, 1993). This tagging system is a hybrid system using both statistical training and rule-based training. Rule-based training is performed only on the statistical tagging errors. The rules are learned by comparing the correctly tagged corpus with the output of tagger. The training is leveraged to learn the error-correction rules.

3 Proposed Model

3.1 The Causes of Part-of-Speech Tagging Error

We will mention important causes to make POS tagging errors. The first cause comes from the low accuracy at tagging unknown words, since assigning the most likely tag for unknown words cannot be expected to give a good result. Second, the linguistic information reflects only the morpheme concatenation, as mentioned in the previous section. Especially, errors occur because of the complex morphological characteristics of Korean. Third, the ambiguities of meanings cannot be resolved, since tagger would not distinguish them in the morphological level.

3.2 Processing Unknown Words

Some of the tagging errors come from the unknown word - absence of the word entry in the dictionary. If at least one sequence of morphological analysis can produce sequence of morphemes registered in the dictionary, the unknown word identification routine does not work even if other sequence contains unknown word. If no sequence is successful, then the system suggests the possible POS-tagged unknown words. In our system, if the morphological analyzer cannot find that all morphemes are in the dictionary, unknown words are supposed to be included in the word. Then, the user adds the unknown words into the dictionary with dictionary manager, if any. After adding the words, morphological analyzer is called once again. Because the user adds the identified unknown words into the dictionary, morphological over-analysis can be avoided.

3.3 Correction of Errors

The result produced by any tagger will contain errors, and correcting these errors would cost very much. Hence, it would be helpful to correct tagging errors using a system which finds errors and correct them. To correct errors in this proposed model is defined first to suggest candidate tags to the user and then to find words which is likely to be wrong tagged. *Correction rule*

and *manual correction log* are necessary for automatic error detection and candidate suggestion. Rule-based method is a way of finding the wrong tags with exact match using the predescribed rule and suggestion pair. The correction rules are in the form of:

(<current morpheme>
<current tag>)/position of wrong morpheme or tag/corrected morpheme or tag*

where * means the repetition. Four kinds of operators can be used in current morpheme or tag.

- **Don't Care(*)** indicates that matching with all morpheme or tag is permitted. If we replace all the tag α after noun word with tag β, the rule '$* < noun > * < \alpha > /4/ < \beta >$' is used.
- **Or(|)** allows to match any one of the expressions. If we replace all the tag α after common or proper noun word with tag β, the rule '$* < noun > | < propernoun > * < \alpha > /4/ < \beta >$' is used.
- **Closure(+)** matches only the content before "+". If we replace all the tag α after common noun(tagged as 'ncn', 'ncpa', 'ncps'), with tag β, the rule, '$*nc + * < \alpha > /4/ < \beta >$' is sufficient.
- **Not(!)** matches except expressions following "!" If we replace all the tag except α after noun word with tag α, the rule '$* < noun > *! < \alpha > /4/ < \alpha >$' is used.

For example, the following rule can replace all the tag 'jcs' before the word "되다(doeda)" with 'jcc'.

'* jcs 되(doe) pvg / 2 / jcc'

Another is the method of using manual correction log. Errors which are not detected by correction rules should be corrected by human tagger. The result of correction is compiled for the next time. Manual log is composed of part of error and part of suggestion. For example, when we change "다운(da'un)/ncpa" to "답(dab)/xsm+ㄴ(n)/etm", the entry will be 'da'un/ncpa, dab/xsm+n/etm'. We can adapt the entry to the augmented case, such as '사람(saram)/ncn+da'un/ncpa', '학교(hag'gyo)/ncn+da'un/ncpa'.

Correction rule can apply to the many kinds of word phrase; while manual log is concerned about only one instance of word phrase. With the manual correction logs, many repetitive errors in a document can be remedied.

4 Implementation

We have implemented error-correction environment to provide the human tagger with the interactive and efficient tagging environment. The overall structure of our environment is shown in Figure 2.

The process of making POS-tagged documents in this environment is as follows:

1. Identify unknown words through morphological analysis.
2. Add unknown word to the dictionary.
3. Repeat morphological analysis using updated dictionary until no more unknown word is found.
4. Run automatic POS tagging.
5. Detect unknown word error and suggest a correct candidate word.
6. Act according to reaction of human tagger - approving modificaton or not, receiving direct input from the human tagger.
7. Repeat steps 5 and 6 with automatic error correction using rules and correction logs so that incremental improvement of tagging accurarcy can be achieved.
8. Correct manually, if there is any error, which is not detected.
9. Save what the human tagger corrected at step 8, and start detecting errors and give suggestion on the POS-tagged document, with manual log.
10. If unknown word exists in the result from step 9, save the result in the dictionary; otherwise, add it to the manual log.
11. Repeat steps 8 and 10 until the human tagger finds no error in the POS-tagged document.

Figure 3 shows the Tagging Workbench.

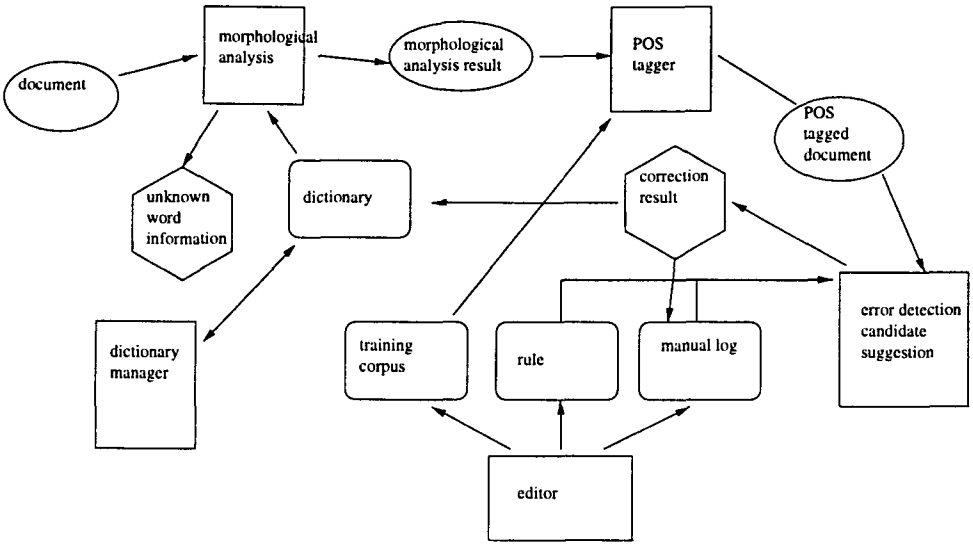

Figure 2: The Structure of Proposed Environment

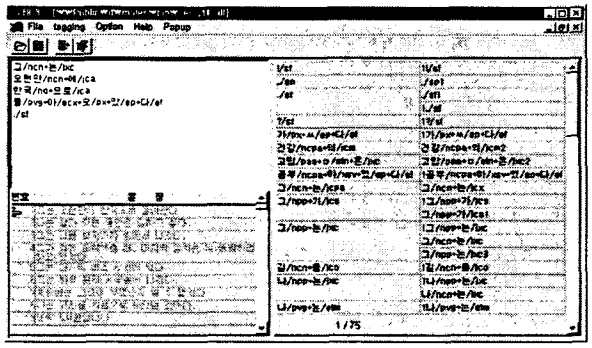

Figure 3: Tagging Workbench

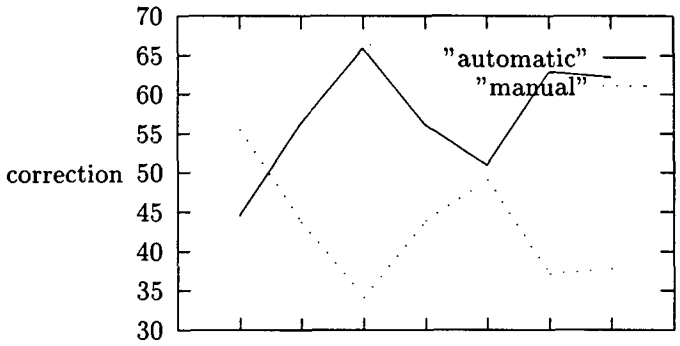

Figure 4: Comparison between automatic and manual correction

5 Experiments and Results

We have experimented on the documents, using morphological analyzer and tagger (Shin et al., 1995). The correction log of one document affects the tagging knowledge base. Then, the next tagging process is automatically improved. In the experimental result, error elimination rates are evaluated.

The result of experiment is in Figure 4. In Figure 4, automatic correction means the right correction made by error detection using rule and manual correction log. Manual correction means the correction made directly by user. We can see that the rate of automatic correction increased, while that of manual correction decreased.

We can correct about 7% of total errors by resolving unknown words. With the increasing number of entries, the probability of unknown word occurrence will decrease.

6 Conclusion

As the researches on the basis of corpus have become more important, constructing large annotated corpus is a more important task than ever before. In general, constructing process of POS-tagged corpus consists of morphological

analysis, automatic tagging and manual correction. But, manual error correction step requires a large amount of costs.

This paper proposed an environment to reduce the cost of correcting errors. In the morphological analysis process, we have eliminated the errors of unknown words, and find errors with error correction rules and manual correction log, suggesting the candidate words. Users can describe error correction rule easily by simplifying the format of error rule. As a result of experiment, about 63.2% of tagging errors were corrected.

Our environment needs further enhancements. One is the need of observation on the pattern of errors to make rules so that accuracy may be improved, and the other is the efficient use of manual logs; currently we use pattern matching. More general rules could be found by expressing the manual logs in other ways.

References

E. Brill. 1993. "A Corpus-Based Approach to Language Learning". *Ph. D. Thesis, Dept. of Computer and Information Science, University of Pennsylvania.*

K. Choi, Y. Han, Y. Han, and O. Kwon. 1994. "KAIST Tree Bank Project for Korean: Present and Future Development". *SNLP, Proceedings of International Workshop on Sharable Natural Language Resources*, pages 7–14.

G.F. Foster. 1991. "Statistical Lexical Disambiguation". *M.S. Thesis, McGill University, School of Computer Science.*

G. Lee and J. Lee. 1996. "Rule-based error correction for statistical part-of-speech tagging". *Korea-China Joint Symposium on Oriental Language Computing*, pages 125–131.

H. Lim, J. Kim, and H. Rim. 1996. "A Korean Transformation-based Part-of-Speech Tagger with Lexical information of mistagged Eojeol". *Korea-China Joint Symposium on Oriental Language Computing*, pages 119–124.

J. Shin, Y. Han, Y. Park, and K. Choi. 1995. "A HMM Part-of-Speech Tagger for Korean with wordphrasal Relations". *In Proceedings of Recent Advances in Natural Language Processing.*