IEICE TRANS. INF. & SYST., VOL.E93–D, NO.8 AUGUST 2010

LETTER

Minimizing Human Intervention for Constructing Korean Part-of-Speech Tagged Corpus

Do-Gil LEE[†*a)], **Gumwon HONG**[†], **Seok Kee LEE**[††], ***Nonmembers***, ***and*** **Hae-Chang RIM**[†*b)], ***Member***

SUMMARY The construction of annotated corpora requires considerable manual effort. This paper presents a pragmatic method to minimize human intervention for the construction of Korean part-of-speech (POS) tagged corpus. Instead of focusing on improving the performance of conventional automatic POS taggers, we devise a discriminative POS tagger which can selectively produce either a single analysis or multiple analyses based on the tagging reliability. The proposed approach uses two decision rules to judge the tagging reliability. Experimental results show that the proposed approach can effectively control the quality of corpus and the amount of manual annotation by the threshold value of the rule.

key words: *part-of-speech tagging, morphological analysis, part-of-speech tagged corpus*

1. Introduction

Annotated corpora are important and useful resources in natural language processing (NLP). Especially in statistical NLP, knowledge acquired from the annotated corpora plays an important role in resolving inherent ambiguities in natural languages. The construction of such corpora, however, requires considerable manual effort which is usually costly and time consuming. Therefore, it is sufficiently worthwhile to produce high-quality annotated corpora with less human effort.

This paper discusses the construction of a part-of-speech (POS) tagged corpus, an annotate corpus where a word is assigned with a corresponding syntactic category, i.e., a POS tag. Generally, the construction of such a POS tagged corpus is performed by human annotators with guidance of automatic POS taggers. In English, the reported precision of state-of-the-art POS taggers is 95–97% [1]–[4]. This performance, however, does not imply that we can readily use the result as it is; in order to make a 100% error-free POS tagged corpus, every POS tag should be examined by human annotators regardless of the POS tagger's performance.

Suppose a POS tagger with 97% precision produces a tagged result, and annotators consult this output to construct an annotated corpus. The annotators cannot identify where the 3% of erroneous words are located in the corpus. Thus, in order to construct an error-free corpus, they should examine all the words of the corpus. From the annotators' point of view, without a perfect POS tagger, the amount of manual effort would make no difference with regard to the performance of automatic taggers.

On the other hand, what if one can know which annotation is correct and which one is not? Then, one can construct a POS tagged corpus with much less cost. As long as a perfect POS tagger does not exist, a 'discriminative' tagger would be more desirable than a 'state-of-the-art' tagger in that it can tell correct words from erroneous ones. Thus, the annotators can intervene only for the erroneous words, and we expect this strategy can considerably reduce the manual effort.

2. POS Tagging Strategies for Minimizing Human Intervention

The construction of Korean POS tagged corpus generally requires the following two steps: morphological analysis and POS tagging. In the former, all possible interpretations for a given word are generated. In the latter, the best interpretation for each word is selected by referring to the neighboring words. Note that in a broad sense the POS tagging is regarded as comprising both steps. Figure 1 shows an example of morphological analysis and POS tagging. In Fig. 1, the path with a bold line shows the correct interpretations, and BOS and EOS indicate the beginning and end of a sentence, respectively.

In this study, we employ a morphological analyzer [5] based on probabilistic models considering three different linguistic units, and a probabilistic POS tagger [6] based on trigram Hidden Markov model considering surface forms. Unlike conventional morphological analyzers, the morphological analyzer [5] provides each interpretation with a probability generated by the probabilistic models. Therefore, the morphological analyzer can rank the interpretations by their probabilities.

An ideal morphological analyzer should produce all possible correct interpretations for each word, and an ideal POS tagger should choose the correct sequence of interpretations from the context throughout all words in a sentence. However, in the real world, the implementation of such a system is almost impossible. Therefore, we propose the following pragmatic and desirable strategy for constructing a POS tagged corpus. For each word in a sentence, if a tagger analyzes a word and decides the analysis is accurate (i.e.,

Manuscript received February 25, 2010.

[†]The authors are with Korea University, Korea.
[††]The author is with KAIST, Korea.
[*]Corresponding authors.
a) E-mail: motdg@korea.ac.kr
b) E-mail: rim@nlp.korea.ac.kr

DOI: 10.1587/transinf.E93.D.2336

Fig. 1 An example sentence "*na-neun hag-go-e gan-da.*" (I go to school.)

the analysis is not likely to contain errors), then the tagger generates a single analysis, otherwise, the tagger generates all possible analyses. In the latter case, a human intervention is required to select the most appropriate one among all the candidate analyses. The most important question here is how we decide if the analysis is correct.

In order to distinguish correctly tagged (analyzed) words from others, we propose two rules that judge the confidence (reliability) of a decision as follows:

- **Agreement-based rule**: If the best candidate (i.e., a sequence of morpheme-tag pairs) produced by morphological analysis is the same as the one produced by the tagger, then we regard the word is correctly tagged and generate single output. Otherwise, multiple candidates are generated.
- **Relative threshold-based rule**: If the probability of the best candidate differs from the probability of the second best candidate by greater than a certain threshold[†], then we regard only the best candidate as correctly tagged. Otherwise, all candidates within the threshold are generated.

Agreement-based rule: In morphological analysis, a word is broken down into grammatically allowed morpheme-tag pairs without referring to neighboring context words. On the other hand, POS tagging is performed with referring to neighboring context. Thus, if the best candidate of a morphological analyzer and the result of POS tagging are identical, then the resulting output is reliable.

Relative threshold-based rule: If a probability gap between the best candidate and the second best candidate of a morphological analysis is sufficiently large, then it is reasonable to use only the first candidate.

In this paper, we employ the following measures to evaluate the performance of a POS tagger that is used in constructing a POS tagged corpus.

- **Hand-validation rate**: the proportion of the words that must be validated by annotators. That is, it is the ratio of words with multiple analyses out of all words, or the ratio that the tagger fails to disambiguate.
- **Error rate**: the proportion of incorrect analyses that were erroneously reported as correct analysis by the system. No that this measure is identical to 'false positive'.

For example, suppose a tagger produces multiple results for $x\%$ of the words, and the words with one-best results have $y\%$ of errors. Then, manual annotation is required only for the $x\%$ of words and consequently the final tagged corpus can contain $y\%$ of error words. A desirable tagger may show a lower hand-validation rate and error rate. The goal of this study is to develop such a tagger to help in constructing high-quality tagged corpus with minimum human effort.

Table 1 Evaluation results of baseline tagger and agreement-based rule.

	Hand-validation rate	Error rate
Baseline	0	4.45
Agreement-based rule	4.55	4.15

3. Experiments

To evaluate our approach, we used the Sejong POS-tagged corpus[††], which contains about 10 million words. We randomly extracted 90% of them for training, and the rest for testing.

The evaluation is performed in the following manner. For each word in the test data, if it is confirmed as correct by the tagger, only one result is produced. Otherwise, two or more results are produced.

As for the morphological analyzer's option, "EMS", which utilizes Eojeol, morpheme, and syllable-unit models, is adopted.

Table 1 shows the result of the baseline tagger (without using any rules) and the result of applying the agreement-based rule. As can be seen in the table, the baseline tagger, which provides only one analysis per word, shows 95.55% of accuracy, i.e., an error rate of 4.45%. This figure implies that without performing any manual post-editing at all, the automatically tagged corpus contains 4.45% of errors in words. On the other hand, when the agreement-based rule is applied, the error rate decreases by 0.3% point from the baseline error rate, which implies 4.55% of words require the hand-validation process.

Figure 2 shows the result of applying both the agreement-based rule and relative threshold-based rules. As can be seen in Fig. 2, the error rate and the amount of annotation work are inversely correlated. In addition, the figure shows that every score converges when a relative threshold is over 10, indicating that the valid relative threshold ranges from 0 to 10. The minimum error rate is 0.88 when the hand-validation rate is 38.46. On the other hand, the minimum hand-validation rate is 9.50 where the error rate is 3.15. This implies that even a 99% accurate POS tagged

[†]More specifically, the threshold is compared with the difference of logs of probabilities between the best candidate and the second best candidate.

[††]http://www.sejong.or.kr/eindex.php

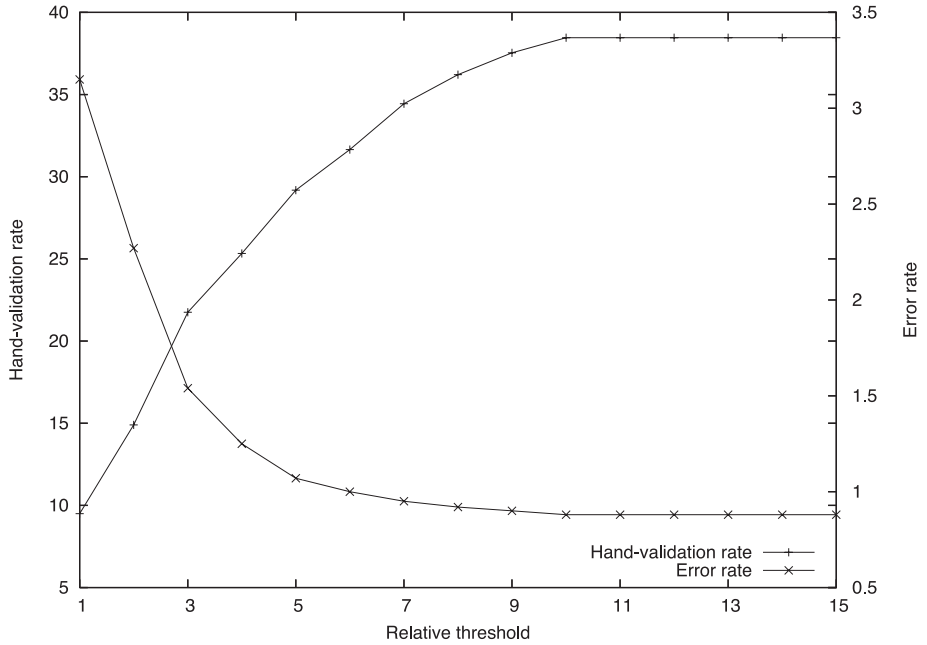

Fig. 2 Results of agreement-based and relative threshold-based rules.

corpus can be obtained with an examination of only 38% of the whole corpus, which greatly reduces the manual annotation work.

4. Conclusion

This paper has presented a method to minimize the manual effort in constructing a Korean POS tagged corpus. We propose a new tagging strategy that can selectively produce either a single analysis or multiple analyses based on the tagging reliability. In order to measure the reliability of the POS tagging, we proposed two decision rules using the morphological analyzer and the POS tagger.

The experimental results exhibited a 'trade-off' relationship between the error rate of an annotated corpus and the amount of annotation work. Even a small decrease in the amount of manual annotation work can achieve significant cost savings in constructing a large-scale POS tagged corpus. The proposed method can provide a new and convenient way to control the quality of the corpus and the amount of manual annotation with the relative threshold.

For future work, we plan to devise a new rule to measure the tagger's reliability. Voting a POS tagger from among multiple POS taggers can be an alternative method for the proposed tagging strategy. We also plan to apply the proposed approach to the construction of real POS tagged corpora.

Acknowledgment

This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MEST) (KRF-2007-361-AL0013).

References

[1] E. Brill, "Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging," Computational Linguistics, vol.21, no.4, pp.543–565, 1995.
[2] A. Ratnaparkhi, Maximum entropy models for natural language ambiguity resolution, Ph.D. thesis, University of Pennsylvania, Philadelphia, PA, 1998.
[3] T. Brants, "TnT – A statistical part-of-speech tagger," Proc. Sixth Applied Natural Language Processing (ANLP-2000), Seattle, WA, 2000.
[4] L. Màrquez, L. Padró, and H. Rodríguez, "A machine learning approach to pos tagging," Mach. Learn., vol.39, no.1, pp.59–91, 2000.
[5] D.G. Lee and H.C. Rim, "Probabilistic modeling of Korean morphology," IEEE Trans. Audio Speech Language Process., vol.17, no.5, pp.945–955, July 2009.
[6] D.G. Lee and H.C. Rim, "Part-of-speech tagging considering surface form for an agglutinative language," Proc. ACL 2004 on Interactive poster and demonstration sessions, p.10, Association for Computational Linguistics, Morristown, NJ, USA, 2004.