An Integrated Error Detection System For POS-tags In Korean

Duhyeon Jin
Department of Linguistics, Chungnam National University, 99 Deahakro, Yuseong-gu, Daejeon, Rep. of Korea
E-mail: dhjin@cnu.ac.kr

Abstract. A POS-tagged corpus is valuable data for NLP tasks. However, a large amount of POS-tagged corpus contains errors and errors debase the quality of corpus. To deal with this problem, we adopt Loftsson(2009)’s approach for POS-tagging error detection with the consideration of linguistic features of Korean. We introduce an integration of two methods to detect errors. One is variation n-gram method and another is multiple taggers’ disagreement method. Then, we reconsider ‘Sejong tagset’ that is used in Sejong corpus which is a representative Korean corpus. By measuring error detection accuracy in POS-tagging result per each tagset candidates with our system, we find that combination of two methods can enhance the accuracy of error detection.

1. Introduction
Annotated corpora are crucial resources for NLP. A POS(part-of-speech)-tagged corpus is a kind of annotated corpus and it is a base of other resources for further NLP tasks such as syntactic parsing and named entity recognition. No matter what kind of annotated corpus has been built, errors appear in general and the errors in a corpus necessarily lead to a lower performance of applications that depend on the corpus. POS-tagging error is fundamental because it causes problems for further steps of NLP. It becomes the origin of all problems.

Automated ways of detecting errors and correcting it in corpora were proposed by Dickinson and Meurers (2003). It is called ‘variation n-gram’ method. Loftsson(2009) proposes a hybrid way of detecting POS-tag errors with the combination of variation n-gram, multiple taggers' disagreement and shallow parsing. Loftsson(2009) discusses the way of detecting errors on an Icelandic corpus. In contrast with other Indo-european languages, Icelandic has complex morphological components that are similar to that of Korean and Japanese. The rich morphology in the language can bring many tags in a POS tagset and a fine-grained tagset makes it hard for taggers to tag correctly. Given that these difficulties are common to both two languages, Icelandic and Korean, we adopt two complementary methods from Loftsson(2009) to detect errors in a Korean corpus. In our discussion, Sejong POS-tagged corpus is our target corpus to do this task.

We integrate two methods into one system and optimize the system by adjusting a weight of parameters of variation n-gram method. After this work, we reconsider a tagset of Sejong corpus and suggest an alternative tagset. Then we evaluate accuracies of error detection with our re-organized corpus in comparison with an original version of Sejong corpus.
The remainder of this paper is as follows. Section 2 reviews previous works in detecting POS-tag errors. Section 3 describes main features of Korean morphology briefly and shows how Sejong corpus is organized and how we derive an alternative tagset for Sejong corpus. Section 4 proposes our method of detecting errors in our target corpus. Section 5 shows experiments and we demonstrate why our approach is better than others for language specific reason. Section 6 concludes our discussion.

2. Previous works on POS-tagging error detection
Dickinson and Meurers(2003) propose variation n-gram method for detecting errors to this end. This method finds different tags for the same strings in a corpus. In the experiment with Wall Street Journal corpus of about 1.3 million words, it is proven that this method is quite efficient in finding errors. With the use of variation n-gram method, they report 2,436 errors out of 2,495 error candidates were found as true errors.

Loftsson(2009) notes that this method is not suitable for corpora tagged with a large fine-grained tagset. It is because a large ratio of variation n-grams tends to reflect true ambiguity rather than inconsistent tagging. In a corpus with a morphologically rich language like Icelandic which has 700 possible tags in the set, a complementary approach is needed to deal with this problem. He uses five taggers and gives POS-tagging results with identical inputs per each tagger, and catches disagreements between taggers to detect true errors. Another complementary method to detect errors is a use of shallow parsing. With the use of IceParser, the result of shallow parsing of IFD corpus provides agreement errors between constituents referring its terminal’s POS tags. By adding last two methods to variation n-gram method, the detected true errors increased from 254 to 1,334 in the research.

In our work, we adopt two methods from previous researches to detect errors in Sejong POS-tagged corpus. One is the variation n-gram method and the other is multiple taggers’ disagreement method. Contrary to Icelandic, agreement phenomena are rare in Korean syntax in spite of its morphological richness. Accordingly, we do not adopt shallow parsing method. In the process of finding best accuracy in detection of POS-tagging errors, we first optimize a parameter to select best candidates from variation n-gram methods, and then we merge them with error candidates from multiple taggers’ disagreement method. Meanwhile, we apply an alternative tagset to our target corpus and observe an accuracy of error detection in the same process. This step may give an intuition to further researches to reconsider Korean POS tagset system.

3. Target language and corpus specification

3.1. Target language
Korean is an agglutinate and head-final language. A remarkable characteristic of Korean is an agglutination of morphemes. In Korean morphology, morphemes for verbal features are attached behind a verb stem systematically. The features are presented below and they are classified into 5 classes.

- **Clause form**: AdverbiaClauseForm, IndirectQuot, NominalClauseForm, RootClauseForm, NounComplementClauseForm, GovernedForm, RelativeClauseForm,
- **Sentence type**: Declarative, Imperative, Interrogative, Propositive
- **Honorificity**: SubjectHonorific, Non-honorific
- **Tense**: Past, RemotePast, Nonpast
- **Modality**: Realis, Irrealis, Retrospective, Non-retrospective

The classes and features above are adopted from No(2007). One could notice how tags correspond to each verbal feature in grammar by its name. These features are realized as inflectional suffixes in Korean. A verb which contains many features is presented below.
As we see at the example above, a word with three characters is analyzed to one verb lexeme and four verbal features. A difficult issue in morphological processing for Korean is how to treat these features. If we choose to use these features with fine-grained tags, the complexity of tagging model will increase so that tagging accuracy will go down accordingly. Sejong corpus suggests a compromise for this problem. Various verbal features were simplified and independent tags stand for verbal features.

3.2. Corpus and tagset
Sejong tagged corpus consists of about 12 million words. A part of Sejong tagset is presented in table 1.

| Class         | Main Category | Tags          | Description                           |
|---------------|---------------|---------------|---------------------------------------|
| Nominals      | NN            | NNG, NNP, NNB | general nouns                         |
|               |               | NP, NR        | proper nouns                          |
|               |               | VV, VA, VX, VCP, VCN | general verbs, adjectival verbs, auxiliary verbs, copulas |
| Verbals       | VV            | JKS, JKC, JKG | nominative suffixes, complement suffixes, genitive suffixes |
|               |               | JKO, JKB, JKV | accusative suffixes, conjunction suffixes, vocative suffixes |
|               |               | JQ           | indirect quotation                    |
| Phrasal suffixes | JK        | JKS, JKO, JKB, JKV | nominative suffixes, accusative suffixes, conjunction suffixes, vocative suffixes |
| Delimiter     | JX            | JX           | delimiters                            |
| Inflectional suffixes | E       | EC, ETN, ETM | temporal or honorific suffixes, clausal suffixes, nominalization suffixes, modification suffixes |

Table 1: A part of Sejong tagset.
POS tags in Sejong corpus are coarse-grained and it is hard to extract specific verbal features which are presented in previous subsection. We suggest a way to use grammatical features maximally in our proposed tagset.

### 3.3. Proposed tagset

We propose an alternative tagset for Sejong corpus. In this tagset, all tags for inflectional suffixes (‘E’ category in the ‘table 1’) are replaced by fine-grained sub-tags which represents a grammatical feature of an inflectional morphemes. All sub-tags for inflectional suffixes are merged into verbal tags and they mark a grammatical feature in the tagset. The feature tags of inflectional suffixes are organized by the mode of No(2007) that was previously described.

In our experiment, Sejong corpus is re-organized with this tagset. For instance, original tags of a verb are replaced by a combination of a verbal tag and feature tags as we see below. Thus, we use another version of Sejong POS-tagged corpus.

- Sentence: “그는 달렸으나 결국 경찰에게 잡혔다.” (He ran but was eventually caught by police)
  
  He run but eventually by police was caught

- Sejong analysis: 달렸으나 → 달리/VV+었/EP+으나/EC

- Extended analysis: 달렸으나 → 달리었으나/VV.PST.ACF
  
  (verb: “run”, past, adverbial clause form)

The size of proposed tagset is bigger than original tagset as we see in table 2. We will observe how this difference affects error rate and accuracies of error detection in our system.

### Table 2: # of POS-tags in Original and Proposed tagsets.

| Method          | # of tags |
|-----------------|-----------|
| Original tagset | 45        |
| Extended tagset | 344       |

To see the effect of the larger tagset, we observed performances of taggers per each tagset. We used identical data set(50,000 sentences to train, 1,000 sentences to test) in both original corpus and re-organized corpus for taggers. The result in table 3 shows that accuracies of taggers with extended tagset decreased.

### Table 3: Accuracies of taggers with two corpora (identical training data, different tagsets).

| Method    | Original tagset | Extended tagset |
|-----------|-----------------|-----------------|
| TnT tagger| 0.971           | 0.961           |
| CRF tagger| 0.978           | 0.967           |
| DT tagger | 0.905           | 0.895           |
4. Proposed method
We propose an integrated POS-tagging error detection system that utilizes two error detection methods. One is variation n-gram algorithm which is implemented in DECCA package(http://decca.osu.edu) in python. Another is multiple taggers’ disagreement method that is implemented as a comparison algorithm which uses each tagger’s tagging results. For this system, we use three taggers, TriT tagger(Brants, 2000), CRF tagger(sklearn-crfsuite package) and Decision Tree tagger(Schmid, 2013). To train each POS tagger, we use 50,000 sentences from Sejong corpus. In initial step, we collect error candidates per each method.

In the integrating step, we first gather duplicate error candidates between two methods. Since we suppose these candidates are certain, these are regarded as true errors by the system. Secondly, we obtain most probable error candidates from variation n-gram method and merge them to error candidates from tagger’s disagreement method. An optimized parameter for filtering error candidates from variation n-gram methods is used in this step.

4.1. Variant ratio parameter
We derive an optimal variant ratio by measuring f1-score of error candidates that filtered by the ratio of n-gram variants. As we see a chart in figure 1, by filtering n-gram error candidates with variant ratio, we can observe maximum f1-score of error detection accuracy. In the development stage, we apply hand-annotated data set which consists of 100 sentences for the optimization. An optimal variant ratio parameter is applied to variation n-gram in the experiment step, and it function as a filter so that only variants that has ratio lower than optimal ratio can be merged into 3 taggers’ disagreement method’s error candidates. In the development stage, we could be convinced ourselves that filtering error candidates of variation n-gram method with optimized parameter and merging them into other candidates can enhance a performance of our system.

Figure 1: We suppose that we have word sequence ‘a b c’ and tag sequence ‘x v y’ and ‘x u y’ for the word sequence in a corpus. Generally, a lower percentage of variant indicates it is likely to be an error. For instance, if ‘x v y’ occupies 1% of all variations and ‘x u y’ occupies 99% of them, ‘x v y’ is more probable error candidate. We filter error candidates that account for more than optimized variant ratio. So we get the best error candidates.

5. Experiments
We sampled 300 sentences from Sejong corpus and we organized a gold standard test set manually. We evaluated an accuracy of two methods with the precision, recall and F1-score measure. In this step, we set the variant ratio parameter as 66% to filter error candidates from variation n-gram method. Then, the integrated error detection system is evaluated. Experiments proceeded for two version of corpus with test sets in same text. The results are presented in table 4 and table 5.
### Table 4: Experiment I: Error Detection Accuracy in Corpus with an Original tagset.

| Method                  | Precision | Recall | Fi-score |
|-------------------------|-----------|--------|----------|
| Variation n-gram        | 0.53      | 0.12   | 0.20     |
| 3 taggers’ disagreement | 0.05      | 0.08   | 0.06     |
| Integrated Method       | 0.12      | **0.20** | 0.15     |

### Table 5: Experiment II: Error Detection Accuracy in Corpus with an Extended tagset.

| Method                  | Precision | Recall | Fi-score |
|-------------------------|-----------|--------|----------|
| Variation n-gram        | 0.88      | 0.10   | 0.19     |
| 3 taggers’ disagreement | 0.06      | 0.18   | 0.09     |
| Integrated Method       | 0.09      | **0.25** | 0.13     |

In the first experiment, we showed that our integrated error detection system can enhance a recall of detection. An increased recall demonstrates that two methods are complementary in detecting errors. However, it turns out that F1-score of integrated method is lower than that of variation n-gram method. In the process of second experiment, length of sentences in the test set has to be shortened, because all tokens of verbs and its inflectional suffixes were merged into one token in the corpus. In order to make test sets of a second experiment, we had to adjust index of true errors in original test set. The result of second experiment is similar to first experiment’s result, but it shows that our corpus with extended tagset can enhance a detection rate of true errors in the variation n-gram method in comparison with original version. However, an F1-score of integrated method decreased. In spite of that, the recall of the integrated method increased dominantly. This also shows the complementary power of our system. It also suggests that that our system can find more ‘true errors’ in a corpus with a fine-grained tagset.

### 6. Conclusion

Detecting errors in a POS-tagged corpus is challenging for morphologically rich languages like Korean. We adopted two methods for detecting errors. One is variation n-gram method and the other is multiple taggers’ disagreement method. In this research, we use two methods in an integrated fashion for error detection. The results of experiment on a Sejong corpus showed that merging inflectional morpheme tags to one verbal tag helps to increase recall of error detection. In conclusion, we showed that the integrated approach can be effective way to detect errors in POS-tagged corpus. Our future work is to increase precision of error detection in our system.
References

[1] Brants T 2000 TnT: a statistical part-of-speech tagger In Proc. of the sixth conference on Applied natural language processing pp 224-231

[2] Chae H R, No Y 1998 A survey of morphological issues in Korean: Focusing on syntactically relevant phenomena (Korean Linguistics vol 9) pp 65-109

[3] Dickinson M, Meurers W D 2003 Detecting errors in part-of-speech annotation In Proc. of the tenth conference on European chapter of the Association for Computational Linguistics vol 1 pp 107-114

[4] Petrov S, Das D, and McDonald R 2011 A universal part-of-speech tagset Preprint arXiv/1104.2086

[5] Loftsson H 2009 Correcting a POS-tagged corpus using three complementary methods In Proc. of the 12th Conf of the European Chapter of the Association for Computational Linguistics pp 523-531

[6] No Y 2007 KWGInterpreter: a lemmatizing POS tagger for the Korean language In Proc. of the 2007 Joint Conf. of LAK, MLSK, and KSLI pp 86-94

[7] Schmid H 2013 Probabilistic part-of-speech tagging using decision trees In New methods in language processing p 154