A New Annotation Scheme for the Sejong Part-of-speech Tagged Corpus

Jungyeul Park  
Department of Linguistics  
University at Buffalo  
jungyeul@buffalo.edu

Francis Tyers  
Department of Linguistics  
Indiana University  
ftyers@indiana.edu

Abstract

In this paper we present a new annotation scheme for the Sejong part-of-speech tagged corpus based on Universal Dependencies style annotation. By using a new annotation scheme, we can produce Sejong-style morphological analysis and part-of-speech tagging results which have been the de facto standard for Korean language processing. We also explore the possibility of doing named-entity recognition and semantic-role labelling for Korean using the new annotation scheme.

1 Introduction

In 1998 the Ministry of Culture and Tourism of Korea launched the 21st Century Sejong Project to promote Korean language information processing. The project is named after Sejong the Great who conceived and led the invention of hangul, the Korean alphabet. The corpus was released in 2003 and was continually updated until 2011, producing the largest corpus of Korean to date. It includes the several types of texts: historical, contemporary, and parallel texts. The section of contemporary corpora contains both oral and written texts. In this paper we focus on the contemporary written text which is annotated for morphology. This is referred to as the Sejong part-of-speech tagged corpus.

The contents of the Sejong POS-tagged corpus represent a variety of sources: newswire text, magazine articles on various subjects and topics, several book excerpts, and crawled texts from the internet. The current version of the morphologically annotated POS-tagged corpus consists of 279 files with over 802K sentences and 9.2M eojols.¹ The current annotation scheme in the Sejong corpus is exclusively based on the eojol concept. The corpus uses the Sejong tagset that contains 44

¹An eojol is a word separated by blank spaces.
Park et al. (2016, 2017), etc. This is done in order to avoid data sparsity, because longer segmentation granularity can combine words in an exponential way.

We propose a new approach to annotation using a morphologically separated word based on the approach for annotating multiword tokens (MWT) in the CoNLL-U format.\(^2\) Using the new annotation scheme, we can also explore tasks beyond POS tagging such as named-entity recognition (NER) and semantic role labelling (SRL). While there are a number of papers looking at NER for Korean (Chung et al., 2003; Yun, 2007), and SRL (Kim et al., 2014)\(^3\), these tasks have hardly been discussed in previous literature on Korean language processing. It has been considered to be difficult to deal with using the current annotation scheme of the Sejong POS corpus because of the limitations of the current eojel-based annotation and the agglutinative characteristics of the language. For example, for NER, having postpositions attached to the last word in the phrase they modify can make it more difficult to identify the named entity. The annotation scheme we propose (see Figure 3) is also different from the current annotation scheme in Universal Dependencies for Korean morphology, which represents combined morphemes for eojels (see Figure 4).

2 CoNLL-U Format for Korean

We use CoNLL-U style Universal Dependency (UD) annotation for Korean morphology. We first review the current approaches to annotating Korean in UD and their potential limitations. The CoNLL-U format is a revised version of the previous CoNLL-X format, which contains ten fields from word index to dependency relation to the head. This paper concerns only the morphological annotation: word form, lemma, universal POS tag and language-specific POS tag (Sejong POS tag). The other fields will be annotated either by an underscore which represents not being available or dummy information so that it is well-formed for input into applications that process the CoNLL-U format such as UDPipe (Straka and Straková, 2017).

\(^2\)\url{http://universaldependencies.org/format.html}

\(^3\)There is also Penn Korean PropBank (\url{https://catalog.ldc.upenn.edu/LDC2006T03})

| Sejong POS (S) | description | Universal POS (U) |
|---------------|-------------|------------------|
| NNG, NNP, NNB, NR, XR | noun related | NOUN  |
| NP, PROPN | proper noun | PROPN |
| PRON | pronoun | PRON |
| ADV | adverb | ADV |
| CONJMM | conjunctive adverb | CONJ |
| DET | determiner | DET |
| VBR | verb-related | VBR |
| ADJ | adjective | ADJ |
| POST | postpositions (case markers) | POST |
| VERB | verb | VERB |
| CONJ | conjunctive | CONJ |
| VAR | variable | VAR |
| AUX | auxiliary | AUX |
| NOUN | noun | NOUN |
| PROPN | proper noun | PROPN |
| PRON | pronoun | PRON |
| ADV | adverb | ADV |
| COVID | COVID-19 related | COVID |
| X | unknown | X |

Table 1: POS tags in the Sejong corpus and their 1-to-1 mapping to Universal POS tags

2.1 Universal POS tags and their mapping

To facilitate future research and to standardize best practices, (Petrov et al., 2012) proposed a tagset of Universal POS categories. The current Universal POS tag mapping for Sejong POS tags is based on a handful of POS patterns of eojels. However, combinations of words in Korean are very productive and exponential. Therefore, the number of POS patterns of the word does not converge even though the number of words increases. For example, the Sejong treebank contains about 450K words and almost 5K POS patterns. We also test with the Sejong morphologically analysed corpus which contains 9.2M eojels. The number of POS patterns does not converge and it increases up to over 50K. The wide range of POS patterns is mainly due to the fine-grained morphological analysis, which shows all possible segmentations divided into lexical and functional morphemes. These various POS patterns might indicate useful morpho-syntactic information for Korean. To benefit from the detailed annotation scheme in the Sejong treebank, (Oh et al., 2011) predicted function labels (phrase-level tags) using POS patterns that improve dependency parsing results. Table 1 shows the summary of the Sejong POS tagset and its detailed mapping to the Universal POS tags. Note that we convert the XR (non-autonomous lexical root) into the NOUN because they are mostly considered nouns or a part of a noun: e.g., minju/XR (‘democracy’).

2.2 MWTs in UD

Multiword token (MWT) annotation has been accommodated in the CoNLL-U format, in which MWTs are indexed with ranges from the first token in the word to the last token in the word, e.g., 1-2. These have a value in the word form field, but have an underscore in all the remaining fields. This
multiword token is then followed by a sequence of words (or morphemes). For example, a Spanish MWT vamotos (‘let’s go’) from the sentence vamotos al mar (‘let’s go to the sea’) is represented in the CoNLL-U format as in Figure 2a. Vamotos which is the first-person plural present imperative of ir (‘go’) consists of vamos and nos in MWT-style annotation. In this way, we annotate the Korean eojoel as MWTs. Figure 2b shows that naseossda (‘became’) in Korean can also be represented as MWTs, and all morphemes including a verb stem and inflectional-modal suffixes are separated. Sag et al. (2002) defined the various kinds of MWTs, and Salehi et al. (2016) presented an approach to determine MWT types even with no explicit prior knowledge of MWT patterns in a given language. (Çoltekin, 2016) describes a set of heuristics for determining when to annotate individual morphemes as features or separate syntactic words in Turkish. The two main criteria are (1) does the word enter into a labelled syntactic relation with another word in the sentence (e.g. obviating the need for a special relation for derivation); and (2) does the addition of the morpheme entail possible feature class (e.g. two different values for the Number feature in the same syntactic word).

3 A New Annotation Scheme

This section describes a new annotation scheme for Korean. We propose a conversion method for the existing UD-style annotation of the Sejong POS tagged corpus to the new scheme.

3.1 Conversion scheme

The conversion is straightforward. For one-morpheme words, we convert them into word index, word form, lemma, universal POS tag and Sejong POS tag. For multiple-morpheme words, we convert them as described in §2.2: word index ranges and word form followed by lines of morpheme form, lemma, universal POS tag and Sejong POS tag. For the lemma of suffixes, we use the Penn Korean treebank-style (Han et al., 2002) suffix normalisation as described in Table 2. The whole conversion table is provided in Appendix A. Figure 3 shows an example of the proposed CoNLL-U format for the Sejong POS tagged corpus. As previously proposed for Korean Universal Dependencies, we separate punctuation marks from the word in order to tokenize them, which is the only difference from the original Sejong corpus which is exclusively based on the eojeol (that is, punctuation is attached to the word that precedes it). One of the main problems in the Sejong POS tagged corpus is ambiguous annotation of symbols usually tagged with SF, SP, SE, SO, SS, SW. For example, the full stop in naseo/VV + eoss/EP + da/EF + ./SF (‘became’) and the decimal point in 3/SN + ./SF + 14/SN (‘3.14’) are not distinguished from each other. We identify symbols whether they are punctuation marks using heuristic rules, and tokenize them. Appendix B details and discusses the tokenisation problem, and how we can further process other symbols.

3.2 Experiments and Results

For our experiments, we automatically convert the Sejong POS-tagged corpus into CoNLL-U style annotation with MWE annotation for eojeols. We evaluate tokenisation, morphological analysis, and POS tagging results using UDPipe (Straka and Straková, 2017). We use the proposed corpus division of the Sejong POS tagged corpus for experiments as described in Appendix C. We obtain 99.88% f1 score for segmentation and 94.75% accuracy for POS tagging for language specific POS tags (Sejong tag sets). Previously, Na (2015) obtained 97.90% and 94.57% for segmentation and POS tagging respectively using the same Sejong corpus. While we outperform the previous results.
including Na (2015), it would not be the fair to make a direct comparison because the previous results used a different size of the Sejong corpus and a different division of the corpus.\(^5\) (Jung et al., 2018) showed 97.08% \(f_1\) score for their results (instead of accuracy). They are measured by the entire sequence of morphemes because of their seq2seq model. Our accuracy is based on a word level measurement.

3.3 Comparison with the current UD annotation

There are currently two Korean treebanks available in UD v2.2: the Google Korean Universal Dependency Treebank (McDonald et al., 2013) and the KAIST Korean Universal Dependency Treebank (Chun et al., 2018). For the lemma and language-specific POS tag fields, they use annotation concatenation using the plus sign as shown in Figure 4. We note that Sejong and KAIST tag sets are used as language-specific POS tags, respectively. However, while the current CoNLL-U style UD annotation for Korean can simulate and yield POS tagging annotation of the Sejong corpus, they cannot deal with NER or SRL tasks as we propose in §4. For example, a word like `peurangseu` (‘of France’) is segmented and analysed into `peurangseu/NNP` and `ui/GEN`. The current UD annotation for Korean makes the lemma `peurangseu+ui` and makes NNP+JKG language-specific POS tag, from which we can produce Sejong style POS tagging annotation: `peurangseu/NNP+ui/JKG`. While a named entity `peurangseu` (‘France’) should be recognised independently, UD annotation for Korean does not have any way to identify entities by themselves without case markers. In addition, as we described in §2.1 the number of POS patterns of the word which is used in the language-specific POS tag field does not converge. Recall that the language-specific POS tag is the sequence of concatenated POS tags such as NNP+JKG or NNG+XSN+VCP+ETM. The number of these POS patterns is exponential because of the agglutinative nature of words in Korean. However, it can be a serious problem for system implementation if we want to deal with the entire Sejong corpus.

\(^5\)Previous work often used cross validation or a corpus split without specific corpus-splitting guidelines. This makes it difficult to correctly compare the POS tagging results. For future reference and to be able to reproduce the results, we propose an explicit-split method for the Sejong POS tagged corpus in Appendix C.
which contains over 50K tags and tag combinations.  

4 Discussion on Moving Beyond POS Tagging

Named entity recognition and semantic-role labelling for Korean have hardly been explored compared to other NLP tasks mainly because they are difficult to deal with using the current annotation scheme of the Sejong corpus or other Korean language related corpora such as the KAIST treebank (Choi et al., 1994) and the Penn Korean treebank (Han et al., 2002). It is an eojeol-based annotation problem of agglutinative language characteristics without the sequence level morpheme’s boundary. For example, a named entity emmanuel unggaro without a nominative case marker instead of emmanuel unggaro-ga (‘Emanuel Ungaro-NOM’) should be dealt with for NER. Using the proposed annotation scheme, we can deal with these problems directly using sequence labelling algorithms. This section describes possible annotation for NER and SRL using the new annotation scheme for Korean.

Because of the characteristics of agglutinative languages previous work on NER (Chung et al., 2003; Yun, 2007) or SLR (Kim et al., 2014) used the sequence of morphemes which can be viewed as being similar to our approach for morpheme-wise aspects. However, our approach uses CoNLL-U style annotation which can be used for upstream tasks such as dependency parsing, semantic parsing, etc. These tasks usually share the same CoNLL-like format. Figure 5 shows an example of NER annotation for Korean. It contains following labels:

- B-Entity: beginning of the entity
- I-Entity: inside of the entity
- B-case: beginning of the functional morpheme (e.g. case marker) of the argument  
- I-case: inside of the functional morpheme of the argument  
- Frame: predicate

5 Conclusion

In this paper we have explored the Sejong corpus in order to determine best practices for Korean natural-language processing. We have defined a standard corpus division for training and testing and have tested POS tagging and syntactic parsing. In addition we have proposed a new tokenisation scheme and applied it to the corpus.

One of the other advantages of our approach is that it is compatible with universal morphological lattices (More et al., 2018), which can be easily converted. Language resources including the scripts and POS tagging models presented in this paper will be freely available (Appendix §D).
References

Anne Abeillé, Lionel Clément, and François Toussenel. 2003. Building a Treebank for French. In Anne Abeillé, editor, Treebanks: Building and Using Parsed Corpora, pages 165–188. Kluwer.

DongHyun Choi, Jungyeul Park, and Key-Sun Choi. 2012. Korean Treebank Transformation for Parser Training. In Proceedings of the ACL 2012 Joint Workshop on Statistical Parsing and Semantic Processing of Morphologically Rich Languages, pages 78–88, Jeju, Republic of Korea. Association for Computational Linguistics.

Jinho D. Choi and Martha Palmer. 2011. Statistical Dependency Parsing in Korean: From Corpus Generation To Automatic Parsing. In Proceedings of the Second Workshop on Statistical Parsing of Morphologically Rich Languages, pages 1–11, Dublin, Ireland. Association for Computational Linguistics.

Key-Sun Choi, Young S Han, Young G Han, and Oh W Kwon. 1994. KAIST Tree Bank Project for Korean: Present and Future Development. In Proceedings of the International Workshop on Shareable Natural Language Resources, pages 7–14, Nara Institute of Science and Technology. Nara Institute of Science and Technology.

Jayeol Chun, Na-Rae Han, Jena D. Hwang, and Jinho D. Choi. 2018. Building Universal Dependency Treebanks in Korean. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

Euisok Chung, Yi-Gyu Hwang, and Myung-Gil Jang. 2003. Korean Named Entity Recognition using HMM and CoTraining Model. In Proceedings of the Sixth International Workshop on Information Retrieval with Asian Languages, pages 161–167, Sapporo, Japan. Association for Computational Linguistics.

 Çağrı Çöltekin. 2016. (When) do we need inflectional groups? In Proceedings of The First International Conference on Turkic Computational Linguistics (TurCLing 2016), pages 38–43, Konya, Turkey.

Chung-Hye Han, Na-Rae Han, Eon-Suk Ko, Martha Palmer, and Heejong Yi. 2002. Penn Korean Treebank: Development and Evaluation. In Proceedings of the 16th Pacific Asia Conference on Language, Information and Computation, pages 69–78, Jeju, Korea. Pacific Asia Conference on Language, Information and Computation.

Jeen-Pyo Hong. 2009. Korean Part-Of-Speech Tagger using Eojeol Patterns.

Sangkeun Jung, Changki Lee, and Hyunsun Hwang. 2018. End-to-End Korean Part-of-Speech Tagging Using Copying Mechanism. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 17(3):19:1–19:8.

Young-Bum Kim, Heemoon Chae, Benjamin Snyder, and Yu-Seop Kim. 2014. Training a Korean SRL System with Rich Morphological Features. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 637–642, Baltimore, Maryland. Association for Computational Linguistics.

Do-Gil Lee and Hae-Chang Rim. 2009. Probabilistic Modeling of Korean Morphology. IEEE Transactions on Audio Speech and Language Processing, 17(5):945–955.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. Computational linguistics, 19(2):313–330.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Tackström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal Dependency Annotation for Multilingual Parsing. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 92–97, Sofia, Bulgaria. Association for Computational Linguistics.

Amir More, Özlem Çetinoğlu, Çağrı Çöltekin, Nizar Habash, Benoît Sagot, Djamel Seddah, Dima Taji, and Reut Tsarfaty. 2018. CoNLL-UL: Universal Morphological Lattices for Universal Dependency Parsing. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Paris, France. European Language Resources Association (ELRA).

Seung-Hoon Na. 2015. Conditional Random Fields for Korean Morpheme Segmentation and POS Tagging. ACM Transactions on Asian and Low-Resource Language Information Processing, 14(3):1–10.

Jin-Young Oh. 2009. Robust Korean Dependency Parsing Using Cascaded Chunking. Ph.D. thesis, Changwon National University.

Jin-Young Oh and Jeong-Won Cha. 2010. High Speed Korean Dependency Analysis Using Cascaded Chunking. Korean Simulation Journal, 19(1):103–111.

Jin-Young Oh, Yo-Sub Han, Jungyeul Park, and Jeong-Won Cha. 2011. Predicting Phrase-Level Tags Using Entropy Inspired Discriminative Models. In International Conference on Information Science and Applications (ICISA) 2011, pages 1–5, Jeju, Korea. Information Science and Applications (ICISA).

Jungyeul Park, Loic Dugast, Jeen-Pyo Hong, Chang-Uk Shin, and Jeong-Won Cha. 2017. Building a Better Bitext for Structurally Different Languages through Self-training. In Proceedings of the First Workshop on Curation and Applications of Parallel and Comparable Corpora, pages 1–10, Taipei,
Taiwan. Asian Federation of Natural Language Processing.

Jungyeul Park, Jeen-Pyo Hong, and Jeong-Won Cha. 2016. Korean Language Resources for Everyone. In Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation: Oral Papers (PACLIC 30), pages 49–58, Seoul, Korea. Pacific Asia Conference on Language, Information and Computation.

Jungyeul Park, Daisuke Kawahara, Sadao Kurohashi, Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann A. Copestake, and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. In Proceedings of COLING ‘02, pages 1–15, London, UK, UK. Springer-Verlag.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann A. Copestake, and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. In Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing, CICLing ’02, pages 1–15, London, UK. Springer-Verlag.

Bo-Hyun Yun. 2007. HMM-Based Korean Named Entity Recognition for Information Extraction. In Knowledge Science, Engineering and Management, pages 526–531, Berlin, Heidelberg. Springer Berlin Heidelberg.

A A Full List of Suffix Conversions

This appendix provides a full list of Penn Korean treebank (KTB)-style suffix conversions. Note that in the the Sejong-style the surface form of the morpheme is used, while in the KTB-style annotation a generic form is used (like a lemma) which is normalised with respect to allomorphy.

| Sejong-style ‘word form’ | KTB-style normalised ‘lemma’ | Sejong-style ‘word form’ | KTB-style normalised ‘lemma’ |
|-------------------------|----------------------------|-------------------------|----------------------------|
| 아/EC                   | 아/EC                      | 아/EC                   | 아/EC                      |
| 음/EP                   | 음/EP                      | 음/EP                   | 음/EP                      |
| ../ETM                  | ../ETM                    | ../ETM                  | ../ETM                    |
| 껆/ETM                  | 껆/ETM                    | 껆/ETM                  | 껆/ETM                    |
| 껇/ETM                  | 껇/ETM                    | 껇/ETM                  | 껇/ETM                    |
| 껇/JKO                  | 껇/JKO                    | 껇/JKO                  | 껇/JKO                    |
| 껇/JKS                  | 껇/JKS                    | 껇/JKS                  | 껇/JKS                    |
| 껇/JKB                  | 껇/JKB                    | 껇/JKB                  | 껇/JKB                    |
| 껇/JX                   | 껇/JX                     | 껇/JX                   | 껇/JX                     |
| 껆/JKO                  | 껆/JKO                    | 껆/JKO                  | 껆/JKO                    |
| 껆/JKS                  | 껆/JKS                    | 껆/JKS                  | 껆/JKS                    |
| 껆/JKB                  | 껆/JKB                    | 껆/JKB                  | 껆/JKB                    |
| 껆/JX                   | 껆/JX                     | 껆/JX                   | 껆/JX                     |

B Tokenisation and Rough Entity Detection

Since the annotation scheme in the Sejong corpus is exclusively based on the eojel, most Korean NLP systems have been developed based on eojels as their segmentation scheme. Therefore, the problem of tokenisation of Korean has often been ignored in the literature. However, there are also other word segmentation schemes for Korean as described in the Korean Penn treebank (Han et al., 2002). Korean dependency parsing (Choi and Palmer, 2011), Korean FrameNet (Park et al., 2014) and Korean UDs (Chun et al., 2018) have used the Penn treebank-style tokenisation scheme, in which punctuation marks are separated from the word.

For Korean tokenisation, we separate all punctuation marks in the eojel by identifying whether symbols are punctuation marks or not. Therefore, entities such as numbers with the decimal point (3.14), email addresses (name@email.com), web address (http://www.web.info), dates (23/9/2017), etc. can be presented as a single token while punctuation marks are separated from the eojel. This idea was originally proposed by (Choi et al., 2012).
to improve constituent parsing results by grouping possible entities. The punctuation mark is separated from the word and the corresponding word is annotated with `SpaceAfter=No`. The tokenisation script from the Sejong corpus will be provided through the DOI system.

### C Where to Train and Evaluate?

Other languages such as English and French have standard training/development/test divisions, especially for the purposes of parsing. For example, the English Penn treebank (Marcus et al., 1993) uses Sections 02-21 for the training set, Section 22 for the development set, and Section 23 for the test set. The French treebank (Abeilé et al., 2003) also defines its own treebank splits for training and evaluation (Seddah et al., 2013). For POS tagging using the Sejong corpus, (Hong, 2009; Lee and Rim, 2009) used 10-fold cross-validation, and (Na, 2015) used 80-20 training/test data sets. We propose to use common treebank 15 files as a test data set and their nearest files can be used as a development data set for the Korean POS tagging task. Since BGAA001 is in the treebank, BTAA0001 in the POS tagging corpus would be a part of the test data, and its nearest file BTAA0002 is a part of the development data. Table 4 provides the entire list of test and development files. In this way, we have a standard evaluation data set for POS tagging, and a similar type of the development data set for system tuning regardless of a variety of sources in the Sejong corpus. The remaining 249 files can be used as a training data set. Table 3 shows the brief statistics of the split corpus.

### D Conversion Tools

We provide scripts to convert the original POS tagged Sejong corpus in XML into the CoNLL-U format (without syntactic annotation) for Korean. We verify the POS tagging format, and remove sentences which contain words with tagging format errors. Note that the script checks only annotation format errors, not analysis errors.