Yet Another Format of Universal Dependencies for Korean

Yige Chen[†*] **Eunkyul Leah Jo**[‡*] **Yundong Yao**[‡*]
KyungTae Lim[¶] **Miikka Silfverberg**[‡] **Francis M. Tyers**[§] **Jungyeul Park**[‡]

[†]The Chinese University of Hong Kong, Hong Kong   [‡]The University of British Columbia, Canada
[¶]Hanbat National University & TeddySum, South Korea   [§]Indiana University, USA

[†]`yigechen@link.cuhk.edu.hk` [‡]`{eunkyul,iameleve}@student.ubc.ca`
[¶]`ktlim@hanbat.ac.kr` [§]`ftyers@iu.edu` [‡]`{msilfver,jungyeul}@mail.ubc.ca`

Abstract

In this study, we propose a morpheme-based scheme for Korean dependency parsing and adopt the proposed scheme to Universal Dependencies. We present the linguistic rationale that illustrates the motivation and the necessity of adopting the morpheme-based format, and develop scripts that convert between the original format used by Universal Dependencies and the proposed morpheme-based format automatically. The effectiveness of the proposed format for Korean dependency parsing is then testified by both statistical and neural models, including UDPipe and Stanza, with our carefully constructed morpheme-based word embedding for Korean. MORPHUD outperforms parsing results for all Korean UD treebanks, and we also present detailed error analyses.

1 Introduction

Dependency parsing is one of the tasks in natural language processing that have been investigated extensively. Using the dependency grammar, it finds the relations between the words in a sentence, and forms an acyclic dependency graph that explains the grammatical structure of the sentence. With various machine learning techniques, the current dependency parsers are able to approach human performances given English corpora.[1] Previous studies trying to probe or improve Korean dependency parsers are lacking, especially for those considering the underlying reason and rationale based on the linguistic properties of Korean. The parsers do not perform as well as their English counterparts, partly due to the fact that Korean is a language that has more complicated linguistic features that make parsing on the word level difficult.

There have been previous studies trying to cope with the word-level representation issues of Korean (Choi and Palmer, 2011; Park et al., 2013; Kanayama et al., 2014). Given that Korean is an agglutinative language that heavily relies on morphemes, and the natural segmentation does not correctly reflect either the words or the morphemes of Korean texts, Park and Tyers (2019) suggested an annotation scheme that decomposes Korean texts into the morpheme level, and applied the morpheme-based format to POS tagging.

In this study, we propose a morpheme-based scheme for Korean dependency parsing that is developed based on Park and Tyers (2019), and adopt the proposed scheme to Universal Dependencies (Nivre et al., 2016, 2020), which contains two Korean dependency parsing treebanks, namely the GSD treebank (McDonald et al., 2013) and the Kaist treebank (Choi et al., 1994; Chun et al., 2018). While the two Korean treebanks meet the standards of Universal Dependencies and have been studied for dependency parsing tasks extensively (Kondratyuk and Straka, 2019; Qi et al., 2020), the treebanks are formatted in a way that the natural segmentations of Korean texts are preserved, and even with some morpheme-level information, only the language-specific part-of-speech tags on the morpheme level are included in the treebanks, and both treebanks do not have any morpheme-level parsing tags. Different from the traditional scheme based on natural segmentation, this scheme utilizes the inherent morphological and typological features of the Korean language, and the morpheme-level parsing tags can therefore be derived using a set of linguistically motivated rules, which are further used to produce the morpheme-level dependency parsing results and automatic conversions between the morpheme-based format and the traditional format.

The proposed morpheme-based representation is examined using several dependency parsing

---

[*]Yige Chen, Eunkyul Leah Jo, and Yundong Yao contributed equally. [¶]Corresponding author.

[1]`https://ai.googleblog.com/2016/05/announcing-syntaxnet-worlds-most.html`

5432

models, including UDPipe (Straka et al., 2016; Straka and Straková, 2017) and Stanza (Qi et al., 2020). Compared to the baseline models trained using the two treebanks without modification, our proposed format makes statistically significant improvements in the performances of the parsing models for the Korean language as reported in the error analysis.

2 Representation of MORPHUD

In this study, we adopt a morpheme-based format that captures the linguistic properties of the Korean language proposed by Park and Tyers (2019). The natural segmentation of Korean is based on eojeol, which does not necessarily reflect the actual word or morpheme boundaries of the language. For example, an eojeol of Korean may contain both a noun and its postposition, or both a verb and its particles marking tense, aspect, honorifics, etc. While this is typical for Korean as an agglutinative language, it creates difficulties and challenges for NLP tasks regarding the Korean language, including dependency parsing. It is not ideal that the tokens dependency relations are annotated on are sometimes words, and sometimes phrases as an eojeol may consist of more than a word. Furthermore, Korean as an agglutinative language has very regular conjugations, which makes it easy and natural to split those words and phrases into morphemes when analyzing the language since nearly every piece of an eojeol can be identified to be of a certain meaning or function.

The morpheme-based format aims at decomposing the Korean sentences further into morphemes, which means that dependency relations are no longer marked on the eojeol level. Instead, they are marked on morphemes such that within each eojeol that is not monomorphemic, a head of that eojeol will be found and all other morphemes will be attached directly to the head. As a result, the head of a non-monomorphemic eojeol carries the dependency relation this eojeol originally has, and all other morphemes will be attached to it. In order to find the head, we develop a script and apply some heuristics which include that the head of an eojeol is usually a noun, a proper noun, or a verb, and while there is no noun or verb in an eojeol, the script we implemented continues to find other morphemes such as pronouns, adjectives, adverbs, numerals, etc. The script also excludes the use of adpositions, conjunctions, and particles as

heads in most cases, unless these are the only part-of-speeches in an eojeol except for punctuations. While there are multiple morphemes that can be heads in an eojeol, the script will decide which one to take based on the part-of-speeches of the morphemes. For instance, when there are multiple nouns, the last one will carry the dependency relation of the eojeol, whereas when there are multiple verbs, the first verb will carry the dependency relation as Korean is a head-final language. Once the head of a non-monomorphemic eojeol is found, the other morphemes will be dependent on the head and be assigned with other dependency relations such as compound, case, auxiliary depending on their UPOS and XPOS.

3 Experiments and results

3.1 Data and systems

In this study, we deploy two parsers to evaluate our proposed format, namely UDPipe (Straka et al., 2016) as a baseline system and Stanza (Qi et al., 2020) as one of the state-of-the-art dependency parsers. UDPipe is a pipeline designed for processing CoNLL-U formatted files, which performs tokenization using Bi-LSTM, morphological analysis, part-of-speech tagging, lemmatization using MorphoDiTa (Straková et al., 2014), and dependency parsing using slightly modified Parsito (Straka et al., 2015). Since the whole pipeline needs no language-specific knowledge, which means that it can be trained using corpora in a different scheme, we choose UDPipe as our baseline. Stanza is another natural language processing toolkit that includes Dozat's biaffine attention dependency parser (Dozat and Manning, 2017). Dozat's dependency parser uses the minimum spanning tree algorithm that can deal with non-projectivity dependency relations, and more importantly it excelled all of dependency parsers during CoNLL 2017 and 2018 Shared Task (Zeman et al., 2017, 2018) In this study, the two dependency parsing pipelines take both the original word-based form[2] (the current scheme adopted by Universal Dependencies), which we denote as WORDUD, and the morpheme-based form, which we denote as MORPHUD, of the GSD and KAIST treebanks as the input.

We develop the script to convert between

---

[2]While words and eojeols are not the same in Korean based on their definitions, in this study, the terms "word-based" and "eojeol-based" are interchangeable.
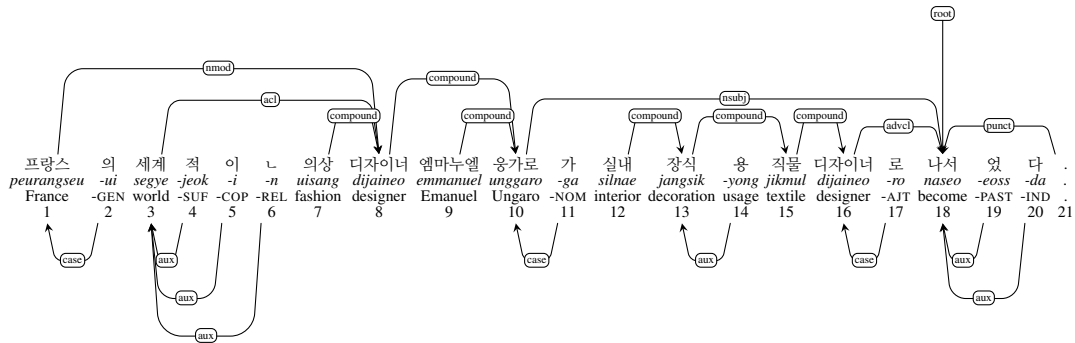
Figure 1: Example of morpheme-based universal dependencies for Korean: while dependencies in top-side are the original dependencies between words, dependencies in bottom-side are newly added dependencies for between morphemes.

the WORDUD format and our proposed MOR-PHUD format. The script consists of two major components, which are WORDUD to MOR-PHUD (`Word2Morph`) and MORPHUD to WORDUD (`Morph2Word`). The Word2Morph component splits the word tokens in the CoNLL-U treebank of Korean into morphemes using the lemmas already provided, and assigns dependency relations on the resegmented tokens based on the original dependency relations annotated on the word tokens. The Morph2Word component, on the other hand, firstly pairs the tokens in the WORDUD dataset and the MORPHUD dataset, and then assigns the dependency relations from morpheme tokens in MORPHUD to word tokens in WORDUD. Within both components, a root detector for the word is implemented in order to find the root (or stem) of a word when the word is multimorphemic (i.e., needs to be split into morphemes and attach dependency relations on it correspondingly). Evaluations of the conversion scripts are not conducted in this study, since the morphemes are inherited from the lemmas in the treebanks, and the part-of-speech tags, roots, and dependency relations are predicted and assigned to the morphemes based on the linguistic features and the grammar of Korean that are regular, as presented in Section 2.

3.2 Results

We report the labeled attachment score (LAS), which is a standard evaluation metric in dependency parsing, using the evaluation script (2018 version) provided by CoNLL 2018 Shared Task.[3] Table 1 shows results of udpipe as a baseline sys-

tem and stanza as one of the state-of-the-art systems. All results are reported in the WORDUD format. That is, all experiments are trained and predicted in the proposed MORPHUD format, and then the result is converted back to the WORDUD format for comparison purposes. We train udpipe once because it can produce the same parsing model if we train it on the same machine. For stanza, we provide average LAS and its standard deviation after five training and evaluation. Both systems use the finely crafted 300d embedding file by fastText (Bojanowski et al., 2017): WORDUD and MORPHUD use words and morphemes as their embedding entries, respectively to make sure that their input representation would be correctly matched. For embeddings, there are 9.6M sentences and 157M words (tokenized) based on WORDUD. The set of documents for embeddings includes all articles published in *The Hankyoreh* during 2016 (1.2M sentences), Sejong morphologically analyzed corpus (3M), and Korean Wikipedia articles (20201101) (5.3M). As expected, all results of MORPHUD outperform WORDUD in Table 1.

3.3 Error analysis and discussion

Figure 2 shows the confusion matrix between WORDUD and MORPHUD, in which the column and the row represent the arc direction of `gold` and `system`, respectively. MORPHUD outperforms WORDUD in predicting all directions except for right (`gold`) / left (`system`). The system predicts the left arc instead of the correct right arc (212 arc direction errors in WORDUD vs. 240 in MORPHUD). This is because we spuriously added left arcs for functional morphemes in MORPHUD where the system learned more left arc instances

	ko_gsd		ko_kaist	
	WORDUD	+MORPHUD	WORDUD	+MORPHUD
udpipe	70.90	77.01	77.01	81.80
stanza	84.63 ($\pm$0.18)	84.98 ($\pm$0.20)	86.67 ($\pm$0.17)	88.46 ($\pm$0.14)

Table 1: Dependency parsing results: for the comparison purpose all MORPHUD results are converted back to WORDUD after training and predicting with the format of MORPHUD
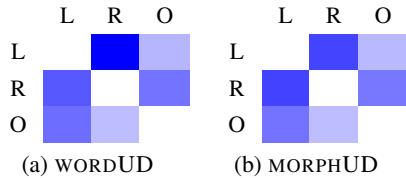

(a) WORDUD (b) MORPHUD

Figure 2: Confusion matrix for the direction of arcs where the column represents `gold`, and the row `system`: Left, Right, and O for TO ROOT.
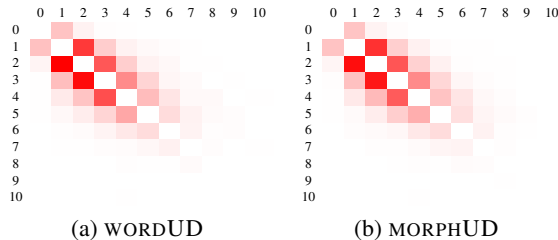

(a) WORDUD (b) MORPHUD

Figure 3: Confusion matrix for the depth of arcs where the column represents `gold`, and the row `system`.

during training.

Figure 3 presents the confusion matrix for the arc depth. The most frequent arc depth error is 2 (`gold`) / 1 (`system`) (901 arc depth errors in WORDUD vs. 855 in MORPHUD). Figure 4 shows an example of parsing errors generated by MaltEval (Nilsson and Nivre, 2008). The parsing error shows that whereas the gold's arc requires the depth 2, the system predicts the depth 1. This is mainly because the analysis of compound nouns for the NP modifier in the Korean treebank prefers a left skewed tree as shown in Figure 5 where some nouns are a verbal noun, and it plays a role as a predicate of the precedent NP modifier. This is a quite different from the English treebank where the right skewed tree dominates: [NP [PRPS *its*] [N [NN *Micronite*] [N [NN *cigarette*] [NNS *filters*]]]]. This is a well-known problem when parsing the Korean treebank because it requires the semantics of the noun to distinguish between the right and the left skewed trees. One possible remedy for this problem was to build a fully lexicalized parsing
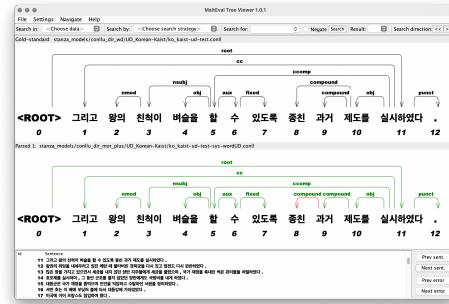

Figure 4: Example of the 2/1 error by MaltEval where the `gold`'s arc depth is 2 and the `system`'s depth is 1. Note that MORPHUD results are converted back to WORDUD: *geuligo wang-ui chinjog-i byeoseul-eul hal su issdolog jongchin gwageo jedo-leul silsihayeossda* 'And they introduced a clan system to make sure that the king's relatives can obtain the government position'
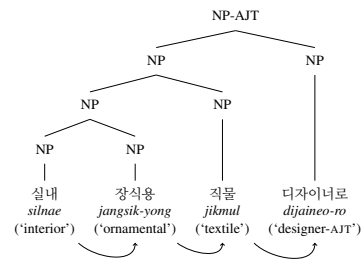

Figure 5: Compound noun with a left-skewed tree for NP modifiers in the Korean treebank

system (Park et al., 2013).

4   Conclusion

We proposed a new annotation scheme for Universal Dependencies for Korean. We have already worked on NER, in which we outperformed the word-level representation dataset by using the morphologically enhanced dataset, and we are planning to extend our idea to the semantic role labeling task. We are also trying to create a consortium to develop the morphologically enhanced Universal Dependencies for other morphologically rich languages such as Basque, Finnish, French, German, Hungarian, Polish, and Swedish. All conversion scripts (WORDUD to MORPHUD,

and vice versa), and MORPHUD datasets for `ko_gsd` and `ko_kaist` will be available through author's `github` at https://github.com/jungyeul/morphUD-korean.

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

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 Sharable 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).

Timothy Dozat and Christopher D. Manning. 2017. Deep Biaffine Attention for Neural Dependency Parsing. In *Proceedings of the 5th International Conference on Learning Representations (ICLR 2017)*, Toulon, France. The International Conference on Learning Representations (ICLR).

Hiroshi Kanayama, Youngja Park, Yuta Tsuboi, and Dongmook Yi. 2014. Learning from a Neighbor: Adapting a Japanese Parser for Korean Through Feature Transfer Learning. In *Proceedings of the EMNLP'2014 Workshop on Language Technology for Closely Related Languages and Language Variants*, pages 2–12, Doha, Qatar. Association for Computational Linguistics.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing universal dependencies universally. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2779–2795, Hong Kong, China. Association for Computational Linguistics.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, 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.

Jens Nilsson and Joakim Nivre. 2008. MaltEval: an Evaluation and Visualization Tool for Dependency Parsing. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, page 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Jungyeul Park, Daisuke Kawahara, Sadao Kurohashi, and Key-Sun Choi. 2013. Towards Fully Lexicalized Dependency Parsing for Korean. In *Proceedings of the 13th International Conference on Parsing Technologies (IWPT 2013)*, pages 120–126, Nara, Japan. Assocation for Computational Linguistics.

Jungyeul Park and Francis Tyers. 2019. A New Annotation Scheme for the Sejong Part-of-speech Tagged Corpus. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 195–202, Florence, Italy. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Milan Straka, Jan Hajic, and Jana Straková. 2016. UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4290–4297, Paris, France. European Language Resources Association (ELRA).

Milan Straka, Jan Hajič, Jana Straková, and Jan Hajič Jr. 2015. Parsing Universal Dependency Treebanks using Neural Networks and Search-Based Oracle. In *14th International Workshop on Treebanks and Linguistic Theories (TLT 2015)*, pages 208–220, Warszawa, Poland. IPIPAN.

Milan Straka and Jana Straková. 2017. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Jana Straková, Milan Straka, and Jan Hajič. 2014. Open-source tools for morphology, lemmatization, POS tagging and named entity recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 13–18, Baltimore, Maryland. Association for Computational Linguistics.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium. Association for Computational Linguistics.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria de Paiva, Kira Droganova, Héctor Martínez Alonso, Çagrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonça, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.