KSII The 16th International Conference on Internet (ICONI) 2024.

Better Sentence Learning through Converting Simple Sentences based on C3Tree Model

Janghwan Kim, Woo Sung Jang, and R. Young Chul Kim*
Software Engineering Laboratory, Hongik University
Sejong, South Korea
Nextsol, Seoul, South Korea
[e-mail: lentoconstante@hongik.ac.kr, jwshurray@gmail.com, *bob@hongik.ac.kr]
*Corresponding author: R. Young Chul Kim

Abstract

With the rapid growth of large language model (LLM) applications, the complex grammar structures and diverse morpheme combinations in Korean have become crucial factors affecting the performance and interpretation of natural language processing (NLP) models. Korean is known for its unique and complicated grammatical structures, which makes it challenging for machines to understand and process accurately. To solve this issue, we propose a mechanism that generates training data through a rule-based transformation using the C3Tree model to simplify complex sentences. By converting complex sentences into simpler forms, our approach aims to reduce sentence redundancy and enhance the quality of training sentences. We expect to enhance the accuracy of Korean NLP models and improve data quality, making them more effective for use in a variety of AI applications.

Keywords: Korean Natural Language Textural Analysis, C3Tree, Natural Language Processing

1. Introduction

Natural Language Processing (NLP) is a field that enables computers to understand and process human language, playing a key role in various AI applications. However, Korean is a language that has complex grammar structures and diverse morpheme combinations, which presents huge challenges for machines to effectively understand and process it. These complex sentences can create difficulties in morphological and syntactic analysis, potentially negatively impacting the performance of NLP models [1]. Therefore, a process called 'simplification' is necessary to convert complex sentences into simpler forms.

Simplifying complex sentences helps machines improve the comprehension of text, enhances the performance of NLP models, and contributes to maintaining data consistency [2].

Simple sentences are easier to understand not only for machines but also for humans, making it easier to convey the meaning of information clearly [3]. Additionally, when the structure is simplified, the accuracy of analysis and prediction can be increased, and reducing structural differences within datasets leads to improved quality of training data [4].

We propose a rule-based training data generation mechanism to convert complex Korean sentences into simpler sentences, aiming to improve the performance of NLP models. We hope that this will enhance the accuracy of Korean NLP models and improve the quality of training data.

Chapter 2 mentions related studies that introduce existing methods for analyzing Korean sentences using machine learning and other techniques. Chapter 3 represents a C3Tree-based Korean sentence refinement mechanism to improve the quality of Korean data. Chapter 4 provides the conclusion.

2. Related Works

2.1 C3Tree Model-based Natural Language Requirements Analysis Methodology

The C3Tree stands for Conditional, Conjunction, and Clause Tree, that is, Korean sentence analysis approach designed to simplify complex sentences [5, 6].

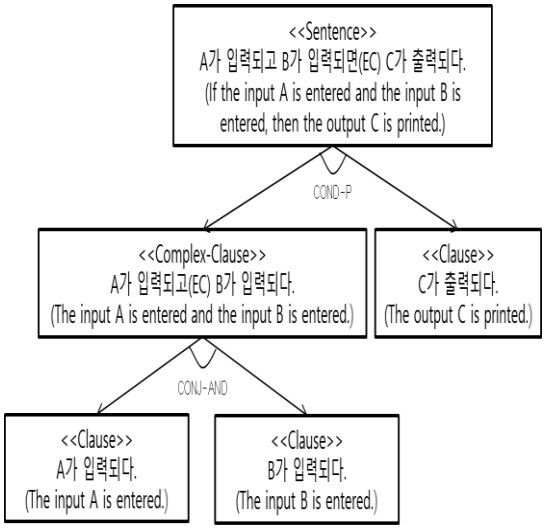

Fig. 1. C3Tree Model Example

Fig. 1 shows a diagram for the requirements analysis process using the C3Tree model.

The C3Tree model analyzes complex sentences by breaking them down into clauses using a tree structure and restores any omitted subjects in the process. It also converts passive sentences into active ones and merges simplified sentences with similar meanings into a single form. Through these steps, the model restructures complex sentences into a more concise and consistent form, which helps resolve ambiguities in requirements.

2.2 KLUE Benchmark Korean Language Learning Dataset

KLUE (Korean Language Understanding Evaluation) is a benchmark designed to evaluate the performance of Korean Natural Language Processing (NLP) models. It is used to test and assess various tasks related to understanding Korean. KLUE includes tasks like sentence classification, sentence similarity, sentiment analysis, question answering, and natural language inference. By evaluating these tasks, it helps measure how well Korean NLP models understand and process language [7]. This benchmark is an important tool for objectively comparing how models analyze complex Korean sentences, and it also emphasizes the need for high-quality training data in Korean.

Table 1. KLUE Benchmark Exam Categories

Name	Description
Topic Classification (TC)	This provides a classifier for predicting the topic of text snippets.
Semantic Textual Similarity (STS)	This measures the degree of semantic equivalence between two sentences.
Natural Language Inference (NLI)	This reads pairs of whole sentences and hypothesis sentences, and predicts whether the relationship is contradictory or neutral.
Named Entity Recognition (NER)	This detects the boundaries of named entities in unstructured text and classifies the types.
Relation Extraction (RE)	This identifies semantic relations between entity pairs in a text.
Dependency Parsing (DP)	This finds relational information among words.
Machine Reading Comprehension (MRC)	This evaluates the ability to understand questions and find answers.
Dialogue State Tracking (DST)	This predicts the dialogue states from a given dialogue context.

Table 1 shows the eight categories covered by KLUE. For semantic analysis, simplified sentences provide clear emotional signals, which can improve model performance. Simplified sentences express information more clearly, which can lead to more accurate and consistent answers.

3. Korean Sentence Simplification Mechanism for Proving Natural Language Analysis based on C3Tree

We use complex sentences as input and prompt them into the C3Tree model to generate several simplified sentences.

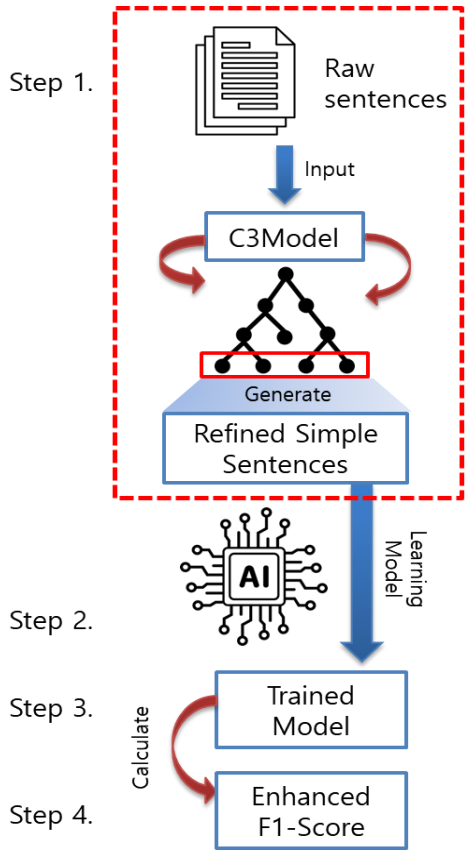

Fig. 2. Enhancing Model Learning Process based on C3Tree

Fig. 2 shows proposed process that enhances a trained model for better F1 score. We focus on step 1 in this paper.

When converting complex sentences into simpler ones, a common issue is that the subjects of the simplified sentences do not match properly. Jang's study addresses this problem by using a subject restoration method [6]. Compared to complex sentences, the simplified sentences produced through this conversion are presented in a more consistent structure, making them easier for models to classify because they enhance structural simplicity.

Fig. 3 shows a detail of Step 1 from the process of inputting a complex sentence into the C3Tree model and simplifying it through sentence refinement.

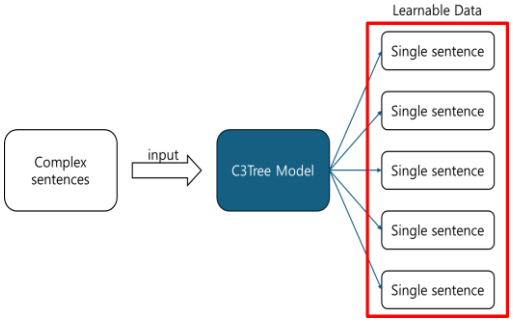

Fig. 3. The Process of Simplifying Complex Sentences using C3Tree

Table 2 shows the sentences that are input into the Korean natural language analyzer equipped with the C3Tree model.

Table 2. Complex Sentences Input

Input Sentences
사용자가 텍스트를 입력하고, save 버튼을 누르면, 입력된 내용을 저장한다. 입력된 내용을 저장하면, 프로그램이 종료된다. (If user enters text and presses the save button, the entered content is saved. If the entered content is saved, The program ends.)

The input sentences in **Table 2** consist of two sentences connected by three conjunctions. When this sample sentence is input into the analyzer, it produces results like those shown in **Fig. 4**.

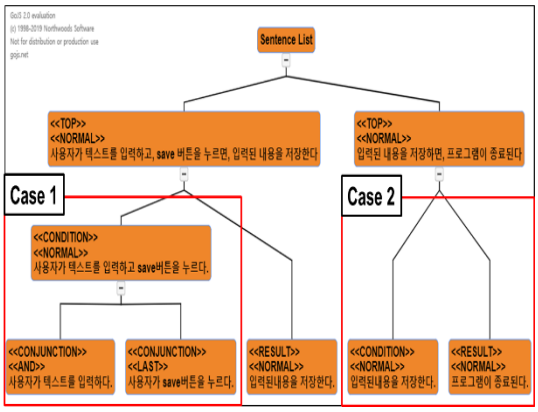

Fig. 4. Simplified Sentence Result in Tree form

Fig. 4 shows the results of entering the sentences from **Table 2** into the tool using the C3Tree model. **Table 3** displays the C3Tree results from **Fig. 4**. In Case 1, the sentence is a compound sentence. Therefore, this sentence

which is connected by "and" is simplified into two separate sentences. " In Case 2, the sentence is a complex sentence that includes a conditional clause. The word "if" is removed, and the sentence is split into two simple sentences. The simplified sentences are then stored in the training dataset and used to train the model.

Table 3. Refined Sentences by C3Tree Model

	Input	Output
Korean (English) Case 1	사용자가 텍스트를 입력하고, 입력된 내용을 저장한다. (User enters text and the entered content is saved.)	사용자가 텍스트를 입력하다 (User enters text.)
		입력된 내용을 저장한다. (The Enterted content is saved.)
Korean (English) Case 2	입력된 내용을 저장하면, 프로그램이 종료된다. (If the entered content is saved, The program ends.)	입력된 내용을 저장한다. (The Enterted content is saved.)
		프로그램이 종료된다. (Program ends.) Program ends.

4. Conclusions

We propose a rule-based mechanism to convert complex Korean sentences into simpler ones, improving the accuracy and consistency of training data for Korean natural language processing models. By using a C3Tree-based mechanism, we enhance data quality and model performance, reducing difficulties in morphological and syntactic analysis.

This approach aims to improve Korean NLP and has potential applications in various AI fields. We plan to verify its effectiveness across different domains and datasets, anticipating further improvements through AI integration. We expect to verify the effectiveness of the proposed mechanism by applying it to different domains and datasets and expect that integrating artificial intelligence with algorithms will lead to further improvements in Korean natural language processing.

Acknowledgement

This research was supported by Korea Creative Content Agency (KOCCA) grant funded by the Ministry of Culture, Sports and Tourism (MCST) in 2024 (Project Name: Artificial Intelligence-based User Interactive Storytelling 3D Scene Authoring Technology Development, Project Number: RS-2023-00227917307820870502 01) and National Research Foundation (NRF), Korea, under project BK21 Four.

References

[1] D. Khurana, A. Koli, K. Khatter, "Natural language processing: State of the art, current trends, and challenges," *Multimedia Tools and Applications*, vol.*82*, pp.3713-3744. 2023. DOI: https://doi.org/10.1007/s11042-022-13428-4

[2] A. Siddharthan, "A survey of research on text simplification," *ITL-International Journal of Applied Linguistics*, vol.165, no.2, pp.259-298, 2014.

[3] G. Song, S. Lee, "A study on the social relationship between humans and machines due to technological evolution," in *Proc of the 2017 Fall Conference of the Korean Society for Technology Innovation*, vol.2017. no,11, pp.425-438, Korean Society for Technology Innovation, 2017.

[4] Z. Kozareva, Q. Li, K. Zhai, and W. Guo, "Recognizing Salient Entities in Shopping Queries," in *Proc. of the 54th Annual Meeting of the Association for Computational Linguistics*, vol.2, pp.107-111, Berlin, Germany, 2016.

[5] W. Jang, R. Y. Kim, "Automatic cause–effect graph tool with informal Korean requirement specifications," *Applied Sciences*, vol.*12*, no.18, 2022.

[6] W. Jang, R. Y. Kim, "Automatic Generation Mechanism of Cause-Effect Graph with Informal Requirement Specification Based on the Korean Language," *Applied Sciences*, vol.11, no.24, 2021.

[7] S. Park, et al., "KLUE: Korean Language Understanding Evaluation," *arXiv preprint arXiv:2105.09680*. 2021.