국립국어원 2023-01-13

발간등록번호
11-1371028-000944-01

2022년 유사 문장 생성 말뭉치 연구 분석 사업

연구 책임자

안 희 돈

국립국어원

제 출 문

국립국어원장 귀하

국립국어원과 체결한 연구용역 계약에 따라 '2022년 유사 문장 생성 말뭉치 연구 분석'에 관한 연구 보고서를 작성하여 제출합니다.

■ 사업 기간: 2022년 8월 ~ 2023년 3월

2023년 3월 3일

연구책임자: 안희돈(건국대학교)

연구 기관: 건국대학교 산학협력단
㈜나라지식정보

연구책임자: 안희돈
공동연구원: 조용준, 위혜경, 박동근, 윤수원, 김성환, 박분선
보조연구원: 하지희, 이주연, 원유권, 홍정연, 이상순, 이정훈

<국문 요약>

2022년 유사 문장 생성 말뭉치 연구 분석

본 사업은 2021년 유사 문장 생성 말뭉치 연구 분석 사업의 후속 사업으로서 특히 표, (텍스트가 포함된) 그림, 그래프 자료 기반 유사 문장(문단) 생성 말뭉치를 수집 및 구축하고, 이를 위한 지침 개선과 수립을 목적으로 한다. 이를 위해, 표를 기반으로 한 말뭉치 수집 및 구축 지침을 개선하고, (텍스트가 포함된) 그림을 기반으로 한 말뭉치 수집 및 구축 지침을 수립하였다. 또한, 그래프를 기반으로 한 유사 문장(문단) 생성 말뭉치 수집 및 구축 지침을 수립하였다. 이러한 지침을 기반으로 표, 그림, 그래프 자료를 활용하여 유사 문장(문단) 생성 말뭉치를 구축하였으며, 납품 자료의 품질 관리 및 보안 유지 계획을 수립하여 실행하였다. 본 사업 결과는 다양한 유형의 데이터를 활용하여 유사 문장(문단)을 생성하는 자연어 처리 기술의 발전에 기여할 것으로 예상된다.

사업의 세부 내용을 요약하면 다음과 같다.

가. 표 기반 유사 문장 생성 말뭉치 수집 및 구축 지침 개선

본 사업의 첫 번째 세부 내용은 표 기반 유사 문장 생성 말뭉치 수집 및 구축 지침 개선이다. 이를 위해 다음과 같은 작업을 수행하였다. 우선, 표 기반 유사 문장 생성 말뭉치를 수집하기 위해서는 상업적 활용과 변형 등 2차적 저작물 작성이 가능한 자료를 수집해야 한다. 따라서, 이번 사업에서는 이러한 자료 수집을 위한 지침을 개선하였다. 또한, 2021년 유사 문장 생성 말뭉치 연구 분석 사업에서 마련한 '표 수집 및 선정 지침(2021년 유사 문장 생성 말뭉치 연구 분석 보고서, 국립국어원, 2021)'을 체계화 및 정밀화하였다. 또한, 표를 기반으로 유사 문장을 생성하기 위해서는 표 기반 문장 생성 지침을 정밀화하여야 한다. 이에 따라, 2021년 유사 문장 생성 말뭉치 연구 분석 사업에서 마련한 '표 유사 문장 생성 지침(2021년 유사 문장 생성 말뭉치 연구 분석 보고서, 국립국어원, 2021)'을 체계화 및 정밀화하였다. 이러한 작업을 통해 표 기반 유사 문장 생성 말뭉치 수집 및 구축 지침이 보다 체계적이고 정밀해졌으며, 향후 연구에서 더욱 유용하게 활용될 수 있도록 기반을 다졌다.

나. (텍스트가 포함된) 그림 기반 유사 문장 생성 말뭉치 수집 및 구축 지침 수립

본 사업의 두 번째 세부 내용은 그림 기반 유사 문장 생성 말뭉치 수집 및 구축 지침 수립이다. 이를 위해 다음과 같은 작업을 수행하였다. 우선, 텍스트가 포함된 그림 자료 수집 방법론을 제시하고 그 지침을 수립하였다. 그림 기반 유사 문장 생성 말뭉치를 수집하기 위해서는 상업적 활용과 변형 등 2차적 저작물 작성이 가능한 자료를 수집해야 한다. 이에 따라, 이번 사업에서는 이러한 자료 수집을 위한 방법론을 제시하였다. 또한, 그림에 한 단어 이상의 텍스트가 포함되어 있어야 하며, 이를 포함하는 그림 자료만을 수집하도록 지침을 수립하였다. 둘째로, 텍스트가 포함된 그림 기반 문장 생성 지침을 수립하였다. 그림을 기반으로 유사 문장을 생성하기 위해서는 그림 기반 문장 생성 지침을 수립하여야 한다. 이에 따라, 그림의 텍스트 중심 묘사 기술 방법론을 제시하였으며, 생성된 문장에는 그림에 드러난 텍스트의 구체적인 내용과 텍스트가 쓰여 있는 장소 및 물체에 대한 정보가 포함되어야 한다는 지침을 수립하였다. 이러한 작업을 통해 그림 기반 유사 문장 생성 말뭉치 수집 및 구축 지침이 보다 체계적이고 정밀해졌으며, 그림을 활용하여 유사 문장을 생성하는 기술 개발에 기반을 다졌다.

다. 그래프 기반 유사 문장(문단) 생성 말뭉치 수집 및 구축 지침 수립

본 사업의 세 번째 세부 내용은 그래프 기반 유사 문장(문단) 생성 말뭉치 수집 및 구축 지침 수립이다. 이를 위해 다음과 같은 작업을 수행하였다. 먼저, 그래프 자료 수집 방법론 제시 및 지침을 수립하였다. 그래프 기반 유사 문장(문단) 생성 말뭉치를 수집하기 위해서는 상업적 활용과 변형 등 2차적 저작물 작성이 가능한 자료를 수집해야 한다. 이에 따라, 이번 사업에서는 이러한 자료 수집을 위한 방법론을 제시하였다. 또한, 그래프와 관련된 문단을 함께 수집하도록 지침을 수립하였다. 그리고, 그래프 기반 문장(문단) 생성 지침을 수립하였다. 그래프를 기반으로 유사 문장(문단)을 생성하기 위해서는 그래프 기반 문장(문단) 생성 지침을 수립하여야 한다. 이에 따라, 그래프 자료 기반 문장(문단) 생성 방법론을 제시하였으며, 결과물에는 그래프와 함께 수집된 문단, 작업자가 직접 구축한 기준 문장(문단), 작업자 또는 기계가 생성한 생성 문장(문단)이 포함되어야 한다는 지침을 수립하였다. 단, 생성 문장(문단)의 경우 검수 과정을 반드시 거쳐야 하며, 그래프를 시각화할 수 있는 방법론을 제시하였다. 이러한 작업을 통해 그래프 기반 유사 문장(문단) 생성 말뭉치 수집 및 구축 지침이 보다 체계적이고 정밀해졌으며, 그래프를 활용하여 유사 문장

(문단)을 생성하는 기술 개발에 기반을 다졌다.

라. 표, 그림, 그래프 기반 유사 문장(문단) 생성 말뭉치 구축

본 사업의 세 번째 세부 내용은 위에 제시한 지침에 기반하여 표, 그림, 그래프 기반 유사 문장(문단) 생성 말뭉치를 수집 및 구축하는 것이었다. 이를 위해 다음과 같은 작업을 수행하였다. 우선 표, 그림, 그래프를 포함한 문서 수집 및 선별 작업이었다. 이번 사업에서는 표 10,000건 이상, 그림 9,000건 이상, 그래프 1,000건 이상의 자료를 수집 및 선별하는 것을 목표로 삼았다. 이때, 상업적 활용과 변형 등 2차적 저작물 작성이 가능하도록 저작권 처리를 수행하였으며, 그림 자료는 국내외에서 말뭉치(또는 데이터 세트)로 구축된 적이 있는 자료는 사용하지 않았다. 둘째로, 표, 그림, 그래프 기반 유사 문장(문단) 생성 작업이었다. 선별된 표, 그림, 그래프 자료를 바탕으로, 수립한 지침에 따라 자연스러운 국어로 기술한 유사 문장(문단)을 생성하였다. 표와 그림 자료는 건당 5건 이상의 유사 문장을 생성하였으며, 작업자 한 명은 표·그림별 하나의 문장을 생성하였다. 그래프 자료는 하나의 기준 문장(문단)과 하나 이상의 유사 문장(문단)을 생성하였다. 셋째로, 말뭉치로 정제 및 가공 작업을 수행하였다. 생성된 유사 문장(문단)들을 정의된 형식에 따라 말뭉치로 정제 및 가공하였다. 최종 산출물 말뭉치의 구조 및 세부 형식(파일명 부여 방식, 표지 부착 방식 등)은 주관기관과 협의하여 결정하였다.

사업의 수행 결과, 표 자료 10,065건, 텍스트가 포함된 그림 자료 9,212건, 그래프 자료 1,060건 등 총 20,337건을 수집하였고, 표 기반 유사 문장 50,325개, 텍스트가 포함된 그림 기반 유사 문장 46,060개, 그래프 기반 유사 문단 2,120개 등 총 98,505개의 문장(문단 포함)을 생성하였다.

주요어: 유사 문장, 표, 텍스트가 포함된 그림, 그래프, JSON, 지침, 인공 지능

<Abstract>

2022 Korean Paraphrase Generation and Corpus Development from Tables, Text-Embedded Images, and Graphs

This project, an extension of the 2021 initiative, focuses on developing a comprehensive corpus of paraphrases derived from tables, images embedded with text, and graphical data. The overarching aim is to refine and establish robust guidelines for this endeavor. To realize this goal, we have enhanced the guidelines for amassing and structuring a corpus from tables and formulated new guidelines for corpora based on text-embedded images. Additionally, we have set out guidelines for gathering and constructing a corpus from graph-based paraphrases. Utilizing these guidelines, we built a corpus of paraphrases using data from tables, images, and graphs while implementing stringent measures for quality control and security maintenance of the outputs. The outcome of this project is poised to advance natural language processing techniques for generating paraphrases using diverse data forms.

The project's specifics unfold as follows:

1. Enhancement of Guidelines for Table-Based Paraphrase Generation and Corpus Construction

The initial phase refined the guidelines for compiling a corpus of table-based paraphrases. This involved gathering materials suitable for commercial exploitation and derivative works. The 'Table Collection and Selection Guidelines (National Institute of Korean Language, 2021)' from the preceding project were meticulously systematized and enhanced, laying a more structured and precise foundation for future research endeavors.

2. Establishment of a Framework for Developing a Paraphrase Corpus from Text-Embedded Images

The second phase entailed the creation of guidelines for accumulating and building a corpus of paraphrases from text-incorporated images. This phase saw the development of a methodology for sourcing image data with embedded text, setting the stage for systematic collection. A critical aspect was establishing parameters for sourcing images that include at least one word of text. Moreover, the project delineated guidelines for formulating sentences from these images, ensuring that the resulting paraphrases accurately reflect the text's content, location, and contextual relevance within the image. This initiative has laid a foundation for pioneering image-based paraphrase generation technologies.

3. Formulation of Guidelines for Constructing Graph-Based Paraphrase Corpus

The third phase was dedicated to formulating guidelines for compiling and constructing a corpus of paraphrases rooted in graphical data. This entailed proposing a methodology for graph data collection and setting guidelines

for sourcing materials amenable to commercial use and further adaptation. The project established parameters for collating graph-related paragraphs and articulated guidelines for generating paragraphs from these graphs. This included methodologies for graph visualization and ensuring that the outputs comprised collected paragraphs, worker-constructed baseline paragraphs, and worker-generated paraphrases. This phase has significantly contributed to establishing more systematic and precise guidelines for future paraphrase generation using graphical data.

4. Constructing a Paraphrase Corpus from Tables, Images, and Graphs

The final phase of the project, grounded in the aforementioned guidelines, was dedicated to collecting and constructing a corpus of paraphrases based on tables, images, and graphs. The project set targets for accumulating a substantial number of documents containing tables (over 10,000), images with text (over 9,000), and graphs (over 1,000). Rigorous copyright processing was undertaken to facilitate commercial use and derivative work. The corpus development involved generating paraphrases based on the selected tables, images, and graphs, with a minimum quota set for each data type. The generated paraphrases were meticulously processed into a corpus in a predefined format, with its structure and detailed formatting (like naming conventions and cover attachment methods) being collaboratively determined with the overseeing organization.

As a result, the project culminated in the collection of 20,337 distinct items, including 10,065 table-based items, 9,212 image-based items (with text), and 1,060 graph-based items. The paraphrase generation resulted in 98,505 sentences (including paragraphs), encompassing 50,325 table-based paraphrases, 46,060 image-based paraphrases, and 2,120 graph-based paraphrases, marking a significant stride in natural language processing.

Key-words: paraphrase, table, text-embedded image, graph, JSON, guideline, artificial intelligence

차 례

표 차례

그림 차례

제 1 장

서 론

1.1. 사업의 목적과 배경

1.1.1. 사업의 목적

4차 산업혁명은 빠르게 진행되고 있으며, 이에 대비하기 위해서는 인공지능 기술의 개발과 활용이 필수적이다. 인공지능 기술의 핵심 구성 요소 중 하나인 자연어 처리는 인간의 언어를 이해하고 생성하는 것을 포함하고 있으며, 이를 위해서는 대규모 언어 데이터인 말뭉치가 필요하다. 이 사업은 다양한 언어 환경을 반영하는 문장 생성 말뭉치를 분석하여 국어 자원의 활용도와 가치를 제고하고, 인공지능 기술의 개발과 활용을 위한 연구를 수행하고자 한다.

특히 본 사업은 2021년 유사 문장 생성 말뭉치 연구 분석 사업의 후속 사업으로서 특히 표, (텍스트가 포함된) 그림, 그래프 자료 기반 유사 문장(문단) 생성 말뭉치를 수집 및 구축하고, 이를 위한 지침 개선과 수립을 목적으로 한다. 이를 위해, 표를 기반으로 한 말뭉치 수집 및 구축 지침을 개선하고, (텍스트가 포함된) 그림을 기반으로 한 말뭉치 수집 및 구축 지침을 수립하였다. 또한, 그래프를 기반으로 한 유사 문장(문단) 생성 말뭉치 수집 및 구축 지침을 수립하였다. 이러한 지침을 기반으로 표, 그림, 그래프 자료를 활용하여 유사 문장(문단) 생성 말뭉치를 구축하였으며, 납품 자료의 품질 관리 및 보안 유지 계획을 수립하여 실행하였다. 본 사업 결과는 다양한 유형의 데이터를 활용하여 유사 문장(문단)을 생성하는 자연어 처리 기술의 발전에 기여할 것으로 예상된다.

1.1.2. 사업의 세부 내용

가. 표 기반 유사 문장 생성 말뭉치 수집 및 구축 지침 개선

본 사업의 첫 번째 세부 내용은 표 기반 유사 문장 생성 말뭉치 수집 및 구축 지침 개선이다. 이를 위해 다음과 같은 작업을 수행하였다. 우선, 표 기반 유사 문장 생성 말뭉치를 수집하기 위해서는 상업적 활용과 변형 등 2차적 저작물 작성이 가능한 자료를 수집해야 한다. 따라서, 이번 사업에서는 이러한 자료 수집을 위한 지침을 개선하였다. 또한, 2021년 유사 문장 생성 말뭉치 연구 분석 사업에서 마련한 '표 수집 및 선정 지침(2021년 유사 문장 생성 말뭉치 연구 분석 보고서, 국립국어원, 2021)'을 체계화 및 정밀화하였다. 또한, 표를 기반으로 유사 문장을 생성하기 위해서

는 표 기반 문장 생성 지침을 정밀화하여야 한다. 이에 따라, 2021년 유사 문장 생성 말뭉치 연구 분석 사업에서 마련한 '표 유사 문장 생성 지침(2021년 유사 문장 생성 말뭉치 연구 분석 보고서, 국립국어원, 2021)'을 체계화 및 정밀화하였다. 이러한 작업을 통해 표 기반 유사 문장 생성 말뭉치 수집 및 구축 지침이 보다 체계적이고 정밀해졌으며, 향후 연구에서 더욱 유용하게 활용될 수 있도록 기반을 다졌다.

나. (텍스트가 포함된) 그림 기반 유사 문장 생성 말뭉치 수집 및 구축 지침 수립

본 사업의 두 번째 세부 내용은 그림 기반 유사 문장 생성 말뭉치 수집 및 구축 지침 수립이다. 이를 위해 다음과 같은 작업을 수행하였다. 우선, 텍스트가 포함된 그림 자료 수집 방법론을 제시하고 그 지침을 수립하였다. 그림 기반 유사 문장 생성 말뭉치를 수집하기 위해서는 상업적 활용과 변형 등 2차적 저작물 작성이 가능한 자료를 수집해야 한다. 이에 따라, 이번 사업에서는 이러한 자료 수집을 위한 방법론을 제시하였다. 또한, 그림에 한 단어 이상의 텍스트가 포함되어 있어야 하며, 이를 포함하는 그림 자료만을 수집하도록 지침을 수립하였다. 둘째로, 텍스트가 포함된 그림 기반 문장 생성 지침을 수립하였다. 그림을 기반으로 유사 문장을 생성하기 위해서는 그림 기반 문장 생성 지침을 수립하여야 한다. 이에 따라, 그림의 텍스트 중심 묘사 기술 방법론을 제시하였으며, 생성된 문장에는 그림에 드러난 텍스트의 구체적인 내용과 텍스트가 쓰여 있는 장소 및 물체에 대한 정보가 포함되어야 한다는 지침을 수립하였다. 이러한 작업을 통해 그림 기반 유사 문장 생성 말뭉치 수집 및 구축 지침이 보다 체계적이고 정밀해졌으며, 그림을 활용하여 유사 문장을 생성하는 기술 개발에 기반을 다졌다.

다. 그래프 기반 유사 문장(문단) 생성 말뭉치 수집 및 구축 지침 수립

본 사업의 세 번째 세부 내용은 그래프 기반 유사 문장(문단) 생성 말뭉치 수집 및 구축 지침 수립이다. 이를 위해 다음과 같은 작업을 수행하였다. 먼저, 그래프 자료 수집 방법론 제시 및 지침을 수립하였다. 그래프 기반 유사 문장(문단) 생성 말뭉치를 수집하기 위해서는 상업적 활용과 변형 등 2차적 저작물 작성이 가능한 자료를 수집해야 한다. 이에 따라, 이번 사업에서는 이러한 자료 수집을 위한 방법론을 제시하였다. 또한, 그래프와 관련된 문단을 함께 수집하도록 지침을 수립하였다. 그리고, 그래프 기반 문장(문단) 생성 지침을 수립하였다. 그래프를 기반으로 유사 문장(문단)을 생성하기 위해서는 그래프 기반 문장(문단) 생성 지침을 수립하여야 한다. 이에 따라, 그

래프 자료 기반 문장(문단) 생성 방법론을 제시하였으며, 결과물에는 그래프와 함께 수집된 문단, 작업자가 직접 구축한 기준 문장(문단), 작업자 또는 기계가 생성한 생성 문장(문단)이 포함되어야 한다는 지침을 수립하였다. 단, 생성 문장(문단)의 경우 검수 과정을 반드시 거쳐야 하며, 그래프를 시각화할 수 있는 방법론을 제시하였다. 이러한 작업을 통해 그래프 기반 유사 문장(문단) 생성 말뭉치 수집 및 구축 지침이 보다 체계적이고 정밀해졌으며, 그래프를 활용하여 유사 문장(문단)을 생성하는 기술 개발에 기반을 다졌다.

라. 표, 그림, 그래프 기반 유사 문장(문단) 생성 말뭉치 구축

본 사업의 세 번째 세부 내용은 위에 제시한 지침에 기반하여 표, 그림, 그래프 기반 유사 문장(문단) 생성 말뭉치를 수집 및 구축하는 것이었다. 이를 위해 다음과 같은 작업을 수행하였다. 우선 표, 그림, 그래프를 포함한 문서 수집 및 선별 작업이었다. 이번 사업에서는 표 10,000건 이상, 그림 9,000건 이상, 그래프 1,000건 이상의 자료를 수집 및 선별하였다. 이때, 상업적 활용과 변형 등 2차적 저작물 작성이 가능하도록 저작권 처리를 수행하였으며, 그림 자료는 국내외에서 말뭉치(또는 데이터세트)로 구축된 적이 있는 자료는 사용하지 않았다. 둘째로, 표, 그림, 그래프 기반 유사 문장(문단) 생성 작업이었다. 선별된 표, 그림, 그래프 자료를 바탕으로, 수립한 지침에 따라 자연스러운 국어로 기술한 유사 문장(문단)을 생성하였다. 표와 그림 자료는 건당 5건 이상의 유사 문장을 생성하였으며, 작업자 한 명은 표·그림별 하나의 문장을 생성하였다. 그래프 자료는 하나의 기준 문장(문단)과 하나 이상의 유사 문장(문단)을 생성하였다. 셋째로, 말뭉치로 정제 및 가공 작업이었다. 생성된 유사 문장(문단)들을 정의된 형식에 따라 말뭉치로 정제 및 가공하였다. 최종 산출물 말뭉치의 구조 및 세부 형식(파일명 부여 방식, 표지 부착 방식 등)은 주관기관과 협의하여 결정하였다.

1.1.3. 사업의 동향 분석

자연어 처리(natural language processing, NLP) 분야는 양상(modality)에 따라 단면적 자연어 처리(unimodal NLP)와 다면적 자연어 처리(multimodal NLP)로 나눌 수 있다. 여기서 양상이라 함은 인간의 내적 표상으로서 시각, 청각, 촉각, 후각, 미각 등의 감각과 내적 대화를 포함한다. 우선 단면적 자연어 처리는 텍스트 처리, 음성 처리, 이미지 혹은 비디오 처리 등 입력이나 출력의

양상이 한 모드로 고정되어 있는 것인 반면 다면적 자연적 처리는 이들 모드가 혼합되어 있는 것으로서 이미지에서 텍스트로(text-to-image), 그래프에서 텍스트로(chart-to-text), 비디오에서 텍스트로(video-to-text), 혹은 그 역방향으로의 자연어 처리에서처럼 입력과 출력의 양상이 각기 다른 경우를 말한다. 유사 문장(문단) 생성은 양상적 측면에서 보면 기계 번역, 요약(summarization), 텍스트 분류(text classification), 감성 분석(sentiment analysis)과 같은 텍스트에서 텍스트로의(text-to-text) 자연어 처리, 즉 단면적 자연어 처리의 한 분야이다.

유사 문장(문단)의 정의와 범위

유사 문장(paraphrase)[1]은 '같은 의미를 지녔으나 표현상 서로 다른 텍스트'를 말하는 것이며, 유사 문장 생성(paraphrasing)이란 그와 같은 텍스트를 생성하는 작업을 말한다. 이와 같은 텍스트는 어휘 층위, 구 층위, 문장 층위, 문단 층위에 걸쳐 다양하게 존재한다는 측면에서 유사 표현, 유사 문장, 유사 문단 등의 표현으로 다양하게 쓸 수 있다. 유사 문장이란 엄밀하게는 이 가운데 문장 층위에서의 것(sentential paraphrase)을 말한다고 볼 수 있으나, 여기에서는 편의상 유사 문단(paragraph paraphrase)까지 포괄하여 일컫는 것으로 하되, 엄밀하게 표현할 때에는 문장 층위일 경우 유사 문장으로, 문단 층위일 경우 유사 문단으로 표현할 것이다.

그러나 자연어 처리 분야에서 유사 문장의 정의와 그 범위에 관련해서는 크게 합의된 것이 없다(Vila et al., 2014; Zeng et al., 2019). 따라서 유사 문장 말뭉치와 유사 문장 생성 모형은 그에 대한 정의에 따라 매우 상이한 양상을 띠고 있다. 예를 들어, MS 연구 유사 문장 말뭉치(Microsoft Research Paraphrase (MSRP) dataset, 2005)의 경우, 의미적 등가(semantic equivalence) 혹은 동일한 의미(진리 조건)를 지닌 명제들 간의 관계, 즉 양방향 함의(bidirectional entailment)가 성립하는 경우를 유사 문장으로 보고 말뭉치를 구축하였다(Dolan and Brockett, 2005). 그러나 양방향 함의라는 것이 드문 현상이라는 점에서 말뭉치의 다양성, 나아가서 그 품질을 저하하게 되는 효과를 가질 수밖에 없으므로 실제로는 그 조건을 완화하여 적용하였다.

P4P 유사 문장 말뭉치(Paraphrase for Professionals paraphrase dataset, 2013)의 경우는 그 범

1) paraphrase는 '의역(意譯)', '다시 쓰기', '환언(이숙의, 2021)' 등으로 번역되거나, 외래어 표기로 '패러프레이즈'로 그대로 사용되기도 한다. 그러나 '의역'이 '원문의 단어나 구절에 지나치게 얽매이지 않고 전체의 뜻을 살리어 번역함. 또는 그런 번역(표준국어대사전)'이라는 점에서 정확한 대응어는 아니며 '패러프레이즈' 또한 외래어로서 일반화되어 있다고 보기는 어렵다는 점에서 적절한 번역어로 보기는 어렵다. 본 사업에서 사용하는 '유사 문장(문단)'이 이 용어가 의도하는 원래의 의미를 갖고 있는 것으로 판단되나, 이 용어가 어휘적 차원은 물론, 구, 절, 문장, 문단, 텍스트 단위까지 포괄한다는 측면에서 '유사 문장' 혹은 '유사 문단'은 그 현상의 일부만을 가리킨다는 측면에서 한계를 지니고 있다. 그러나 여기서는 본 사업의 대상이 sentential paraphrase 혹은 paragraph paraphrase에 국한되어 있다는 관점에서 '유사 문장' 혹은 '유사 문단'이라는 용어를 그대로 사용할 것이다.

위를 확대하여, 다음 표와 같이 그 유형을 3단계로 계층분류하여 형태어휘기반 변화(morpholexicon-based changes), 구조기반 변화(structure-based changes), 의미기반 변화(semantics-base changes), 기타 변화(miscellaneous changes) 등 4개 부류와 20개의 유형으로 하위분류하였다(Barrón-Cedeño et al., 2013).

부류	하위부류	유형
형태어휘기반 변화	형태기반 변환	• 굴절 변화(inflectional changes) • 양태동사 변화(modal verb changes) • 파생형 변화(derivational changes)
	어휘기반 변화	• 철자와 포맷 변화(spelling and format changes) • 동일극성 대체(same-polarity substitutions) • 통합적/분석적 대체(synthetic/analytic substitutions) • 반대-극성 대체(opposite-polarity substitutions) • 역대체(converse substitutions)
구조기반 변화	통사기반 변화	• 태 교체(diathesis alternations) • 부정 전환(negation switching) • 접속 변화(coordination changes) • 내포 및 중첩 변화(subordination and nesting changes)
	담화기반 변화	• 문장부호 및 포맷 변화(punctuation and format changes) • 직접/간접 화법 변화(direct/indirect style alternations) • 문장 양태 변화(sentence modality changes) • syntax/discourse structure changes(통사/담화 구조 변화)
의미기반 변화		• 의미기반 변화
기타 변화		• 어순 변화 • 추가/삭제

<표 1> Barrón-Cedeño et al(2013)의 유사 문장 유형 분류

Vila et al.(2014) 등 일련의 언어학적 접근에서는 이를 확장하여 유사 문장에 대한 포괄적 유형 분류와 그 예시를 제시하고자 노력하였다. 이는 유사 문장의 정의에 있어 '의미의 근사적 동일성(approximate sameness of meaning)'에 대해서는 어느 정도 의견이 일치하지만 '근사성'의 정도에 있어 모호할 뿐 아니라 여기서의 '의미'에 대해서도 의견이 일치하지 않으므로, 유사 문장의 정확한 경계란 존재하지 않으며 수행하는 과제의 유형과 수행 목적에 따라 상이할 수 있다고 보기 때문이다. 따라서 유사 문장이라는 것은 범주적 구분이 아니라 '유사 문장으로서의 가능성(paraphrasability)'이라는 정도성의 측면에서 바라봐야 한다고 보았다. Vila et al.(2014)에서 유사 문장으로서의 경계를 결정하는 것은 크게 세 가지로 첫째는 내용 손실(content loss)이다. 이는 어구 삭제나 추상화(혹은 일반화)에 의해 이뤄지는데 이 경우 '없어진 내용(missing content)'은 맥

락 내의 암묵적 어휘 지식에 의해 복구될 수 있기도 하다. 두 번째는 화용적 지식(pragmatic knowledge)으로서 의미적 유사성을 벗어나 있는 예로 Martin(1976)의 '화용적 유사 문장(pragmatic paraphrases)'으로 명명된 것에 해당한다. 화자의 의향에 있어 동일하거나 사실과 사건에 있어 동일한 경우가 포함된다. 세 번째는 문법적 특징(grammatical features)으로 이에는 인칭, 수, 시제의 변화 등이 포함된다. 그러나 이와 같은 세밀한 분류는 너무 복잡하여 신경망 기반 연구의 대규모 말뭉치의 주석 부착에는 적합하지 않다는 비판을 받는다.

이와 같은 이론적 정의를 시도하는 대신 자연어처리 관점에서의 절차적이거나 조작적인 정의에 기댄 말뭉치도 있다. PARANMT50M 말뭉치는 역번역(back-translation)[2]에 의해 생성된 문장은 유사 문장일 가능성이 높다는 유사 문장의 특징을 활용하여 대규모 유사 문장 말뭉치를 구축하였다(Wieting et al., 2017; Wieting and Gimpel, 2018).

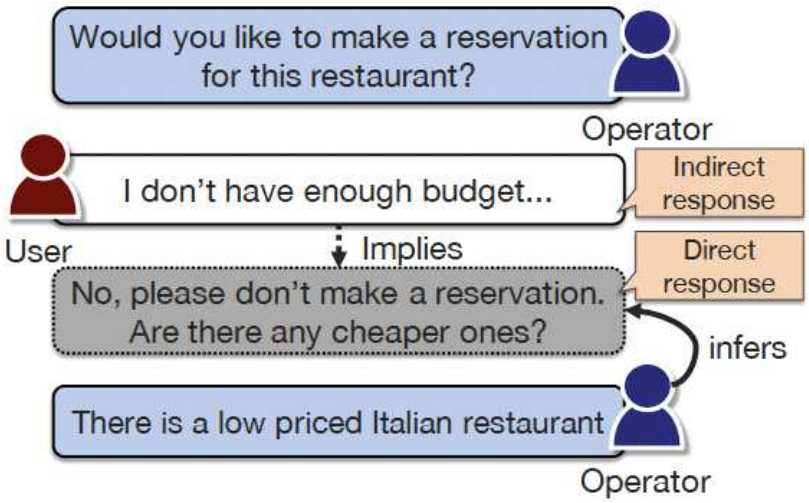
MSCOCO는 이미지 캡셔닝 말뭉치로서 동일한 이미지를 지시하는 두 문장은 유사 문장이라는 화용적 유사 문장의 관점에 서 있다. 약 12만 개의 이미지에 대해 각각 5명의 주석자가 만들어낸 5개의 이미지 캡션은 이미지에서 두드러진 대상 혹은 행위를 묘사하고 있다는 관점에서 유사 문장 말뭉치로서 기능할 수 있다고 보는 것이다. 최근 이들 두 말뭉치가 유사 문장 생성에 주로 쓰이고 있다(Fu et al., 2019; Gupta et al., 2018).

선행 연구를 살펴보았을 때, 유사 문장에 대한 고정불변의 정의는 원초적으로 불가능하다고 본다. 이런 점에서 본 사업의 성격과 목적을 돌아볼 수밖에 없는데, 본 사업은 기계 생성 말뭉치가 아니라 인간 생성 말뭉치를 목표로 하고 있으므로 절차적이거나 조작적 정의에 기초하기보다는 이론적 정의에 기대고 있다. 또한 유사 문장 말뭉치 생성은 기계번역 등 자연어 처리 분야에서 성능을 향상시키기 위한 데이터를 제공하는 것이 그 목표로서 이러한 말뭉치는 문장 간 의미적 유사성을 학습하고 이를 기반으로 자동 문장 생성, 요약, 번역 등의 과제를 수행하는 언어모형을 개발하는 데 도움을 주고, 자연어 이해 분야에서의 대화 시스템, 검색 엔진 등에서의 질문 응답 시스템의 성능을 향상시키는 데도 사용될 수 있다. 이러한 시각에서 의미적 유사성이 다소 떨어진다 하더라도, 화자의 발화 의향이 동일한 경우에 해당하는 '화용적 유사 문장'까지 포함하는 포괄적 관점에서 유사 문장에 대한 정의를 내려야 한다.

예를 들어, 일상 대화에서 비관습적 간접 화행의 발화는 주요 발화수반행위(primary illocutionary act)의 추론에 기대어 그에 대응하는 직접 화행의 발화와 유사 문장으로 간주하는 것

2) 역번역이란 하나의 언어에서 다른 언어로 번역한 다음, 다시 그 번역된 언어를 원래 언어로 번역하는 기술을 의미한다.

이 질문 응답 시스템 구축에 유리하다.



<그림 1> 대화에서의 간접화행과 유사 문장(Takeyama et al., 2021: 1980, Fig. 1)

위 대화에서 관리자(operator)는 사용자(user)의 발화인 "I don't have enough budget"을 그 주요 발화수반행위인 "No, please don't make a reservation. Are there any cheaper ones?"와 동일한 의미를 지닌 것으로 해석해야 적절한 대화가 이뤄질 수 있다. 효율적인 질의응답 시스템 구축을 목표로 한다면, 이와 같은 대화상의 특징을 반영할 수밖에 없다.

이런 점에서 유사 문장의 경계에는 다음과 같은 경우들이 들어가게 된다.

(1) a. Close the door please.
b. There is air flow.
(2) a. Trump sends new tweets.
b. The president sends new tweets. (Zeng et al., 2019: 80543, Fig. 1)

(1)은 앞서 소개한 '화용적 유사 문장(pragmatic paraphrase)'로 알려져 있는 예이며, (2)는 '지시적 유사 문장(referential paraphrase)'에 해당하는 예이다. 이와 같은 유사 문장의 생성은 문법적 변형에 관련될 뿐만 아니라 맥락 이해와 세상에 대한 지식을 요한다. 본 사업의 유사 문장의 범위에는 이들도 포함하여 말뭉치를 생성하도록 하였다.

유사 문장(문단) 말뭉치의 개관

여기에서는 기존의 유사 문장 말뭉치를 영어와 한국어를 중심으로 개관하도록 한다.

말뭉치 명칭	공개 연도	장르	생성 주체	건별 문장 수	생성 건수	말뭉치의 성격
DIRT	2001					
PPDB	2013					
MSRP	2005	뉴스	기계	2	5,801	유사도
TUC		트위터		2	56,787	
ParaNMT-50M		소설, 법규		2	51,409,585	
MSCOCO		이미지 캡셔닝		5	493,186	
Quora		질문		2	404,289	
PAWS	2019			2	108,463	유사도
ParaSCI-ACL		과학논문	기계	2	59,402	
ParaSCI-arXiv		과학논문	기계	2	479,526	
STS-B	2012~ 2017	이미지캡셔닝, 뉴스 헤드라인, 유저 포럼	기계/인간	2	8,628	

<표 2> 영어권 유사 문장 말뭉치 개관

유사 문장 말뭉치를 세분하면 말 그대로의 유사 문장 말뭉치와 쌍으로 주어진 문장 간의 유사도 측정값을 부착한 문장 유사도 말뭉치(sentence similarity corpus)로 구분할 수 있다. 후자의 경우 비유사 문장까지 포함하고 있는 특징이 있다.

test_id	question1	question2
0	How does the Surface Pro himself 4 compare with iPad Pro?	Why did Microsoft choose core m3 and not core i3 home Surface Pro 4?
1	Should I have a hair transplant at age 24? How much would it cost?	How much cost does hair transplant require?
2	What but is the best way to send money from China to the US?	What you send money to China?
3	Which food not emulsifiers?	What foods fibre?
4	How "aberystwyth" start reading?	How their can I start reading?
5	How are the two wheeler insurance from Bharti Axa insurance?	I admire I am considering of buying insurance from them

<표 3> 유사 문장 말뭉치의 예(Quora)

index	sentence1	sentence2	score
0	A plane is taking off.	An air plane is taking off.	5.00
1	A man is playing a large flute.	A man is playing a flute.	3.80
2	A man is spreading shredded cheese on a pizza.	A man is spreading shredded cheese on an uncooked pizza.	3.80
3	Three men are playing chess.	Two men are playing chess.	2.60
4	A man is playing the cello.	A man seated is playing the cello.	4.25
5	Some men are fighting.	Two men are fighting.	4.25
6	A man is smoking.	A man is skating.	0.50
141	A man is dancing.	A man is talking	0.00

<표 4> 문장 유사도 말뭉치의 예(STS-B)

<표 3>은 쿼라(Quora) 말뭉치의 예로서 좁은 의미의 유사 문장 말뭉치이다. 질문에 유사한 문장으로 구성되어 있다. <표 4>는 STS-B 말뭉치의 예로서 주어진 문장쌍의 유사도를 크라우드소싱으로 모집한 일반인 평가자들이 각 문장쌍별로 5명씩 6점 리커트 척도(0~5)로 평가한 후 이를 평균한 측정값(score)을 부착한 문장 유사도 말뭉치의 예이다.

문장 유사도 말뭉치 중에는 리커트 척도 대신 가부판단(yes/no)으로 측정한 말뭉치(MSRP)도 있다. MSRP(Microsoft Research Paraphrase Corpus) 말뭉치는 마이크로소프트에서 공개한 것으로 2년간 수집한 9,516,684개의 문장으로부터 추출한 13,127,938개의 문장쌍으로 구성되어 있다.

PAWS(Paraphrase Adversaries from Word Scrambling)은 구글에서 공개한 말뭉치로 위키피디아에서 추출한 문장과 쿼라(Quora) 말뭉치에서 추출한 문장을 어순 및 구조 변경, 그리고 역변역을 통해 기계적으로 생성한 유사 문장에 5명의 평가자의 6점 리커트 척도(0~5) 기반 유사도 측정치를 부착한 말뭉치이다. 특히 적대적 유사 문장(paraphrase adversary)을 포함한 말뭉치로서, 여기서 적대적 유사 문장이란 단지 1~2개의 구성 요소를 제외하고 나머지 모든 요소들이 동일한 문장을 말한다.

(3) a. Flights from New York to Florida.
b. Flights to Florida from NYC.
c. Flights from Florida to New York. (Zhang et al., 2019: 1, (1)-(3))

(4) a. 경찰청장은 아이유에게 홍보대사 임명장을 수여하였다.
b. 질병관리청장은 아이유에게 홍보대사 임명장을 수여하였다. (김민호 외, 2021: 450)

(3a-c)의 세 문장은 높은 BOW(Bag of Words) 중복값을 가짐에도 불구하고 (3b)는 (3a)의 유사 문장에 해당하지만 (3c)는 그렇지 않다. (3c)와 같은 예를 적대적 유사 문장이라고 한다. 한국어로 예를 들면 (4b)는 (4a)와 '경찰청장'과 '질병관리청장'으로 대체한 것 말고는 다른 구성 요소들이 모두 동일한 적대적 유사 문장의 예이다. 형식적 유사성이 매우 높기 때문에 언어 모형이 유사 문장으로 판단하기 쉽지만 의미상 유사 문장에 해당하지 않는다.

STS-B 데이터세트는 쌍으로 주어진 두 문장 간의 의미적 유사도를 6점 리커트 척도(0~5)로 평가한 측정값을 모은 것이다.

한국어 유사 문장 말뭉치는 다음과 같다.

말뭉치 명칭	연도	장르	건별 문장 수	생성 건수	말뭉치의 성격
KorSTS	2020	이미지캡셔닝, 뉴스 헤드라인, 유저 포럼	2	8,628	유사도
paraKQC	2020	질문, 명령	10	10,000	유사도/진성
Question-pair	2020	질문	2	7,576	유사도
styleKQC	2022	질문, 명령	2	15,000	유사도/진성
국립국어원 유사 문장 말뭉치 2019	2019	문어(신문), 문어(기타), 구어(공적대화), 구어(사적대화)	10	37,543	진성
엑소브레인 패러프레이즈 말뭉치	2018	법률 데이터, 특허 데이터, 뉴스 기사, 백과사전, 자연어 이해 언어 자원	2	2,000	유사도/진성
KLUE-STS	2021	airbnb 리뷰, policy-리뷰, paraKQC-스마트홈 쿼리	2	13,224	유사도/진성

<표 5> 한국어 유사 문장 말뭉치의 개관

KorSTS는 카카오브레인(kakaobrain)에서 구축한 의미 유사도 말뭉치로서 STS 데이터세트를 기계 번역한 것으로, 이 중 개발 세트(development set)와 테스트 세트(test set)의 경우 기계 번역한 것을 2명의 번역 전문가의 수정을 거친 자료이다. paraKQC는 10개의 비슷한 문장에 대해 1,000개의 집합으로 구성된 것으로 문장 유사도 데이터 494,500개와 유사 문장 데이터 45,000건으로 구성되어 있다. Question-pair는 두 개의 질문이 같은 질문인지 아닌지 레이블링한 데이터로서 학습 6,888건, 테스트 688건으로 구성되어 있다. 쌍으로 구성된 문장들 간 유사도에 대해 가부

판단(Yes/No task)을 한 문장 유사도 말뭉치라고 할 수 있다. styleKQC(style-variant paraphrase corpus for Korean questions and commands)는 화제, 화행, 격식성 등에 따라 유사 문장을 생성한 말뭉치이다.

<그림 2> styleKQC의 예(Cho et al., 2022: 3, Fig. 3)

국립국어원에서 제작한 유사 문장 말뭉치는 기계 생성한 유사 문장 5개와 인간이 생성한 5개를 세트로 구성한 유사 문장 말뭉치이다. 엑소브레인 패러프레이즈 말뭉치는 법률이나 특허와 같은 전문 분야의 자연어 처리 및 질의응답 기술을 개발하기 위한 말뭉치로, 기호적 접근과 딥러닝 기술을 융합하여 생성하였다. KLUE STS 데이터는 STS 과제를 해결하기 위해 만들어진 한국어 데이터 세트며, AIRBNB(구어체 리뷰), Policy(격식체 뉴스), ParaKQC(구어체 스마트홈 쿼리)의 세 가지 도메인으로 구성되어 있다. 전체 데이터 개수는 총 13,224개로, Train 데이터 11,668개, Dev 데이터 519개, Test 데이터 1,037개인 약 20:1:2의 비율로 구성되어 있다.

1.2. 사업의 수행 내용과 체계

<그림 3> 사업 수행 체계

자료 수집 및 선별 단계는 나라지식정보의 주도하에 다양한 문서를 수집하는 과정이다. 표와 그림, 그래프가 포함된 문서들을 대상으로 이뤄진 자료 수집은 다양한 정보원에서 충분한 데이터를 확보하는 데 중요한 역할을 한다. 수집된 문서들 중에서는 효율적인 자료 처리와 정확한 정보 활용을 위해 활용 가능한 문서들만이 선별되었다. 이 과정은 데이터의 질을 높이고, 불필요한 정보를 걸러내는 데 중요하다.

자료 구축 및 변환 단계는 선별된 자료를 활용 가능한 형태로 만드는 과정이다. 건국대학교의 다언어다문화연구소에서 이 단계를 주도하며, 자료를 체계적으로 정리하고, 필요에 따라 변환한다. 이 과정은 자료의 접근성과 사용성을 향상시키는 데 중요한 역할을 한다. 또한, 이 단계에서 구축된 자료는 머신러닝 기법에 활용될 수 있도록 말뭉치 형태로 전환된다. 이는 인공지능 분야의 연구와 발전에 필수적인 요소로, 머신러닝 알고리즘의 학습 및 개선에 큰 도움이 된다.

유사 문장 생성 과정에서는 구축 및 변환된 최종 자료를 바탕으로 유사 문장을 생성한다. 이 과정은 데이터의 다양성과 풍부함을 더하는 중요한 단계로, 다양한 응용 분야에서의 사용을 가능하게 한다. 유사 문장 생성 시에는 자체 회의 및 자문을 통한 의견 수렴과 국립국어원의 전문적인 판단을 고려한다. 이를 통해 생성된 문장의 품질과 적합성을 보장하며, 과정의 정확성과 신뢰성을 높이게 된다.

완료 점검 및 보완 단계는 프로젝트의 마지막 단계로, 전체 과정의 완성도를 높이고 최종 결과물의 품질을 보장하기 위해 필수적이다. 먼저, 유사 문장 검증 과정을 통해 생성된 문장들의 정확성과 적합성을 확인한다. 이 단계에서는 유사 문장 생성 지침에 의해 생성된 문장들이 원본 데이터의 의미를 정확히 반영하고 있는지, 그리고 언어적으로 적절한지를 평가한다. 원자료 재검토는 수집 및 변환 과정에서 사용된 원본 자료들을 다시 한번 확인하여 오류나 누락된 부분이 없는지 확인하는 과정이다. 이는 데이터의 정확성을 보장하고, 최종 결과물의 신뢰성을 높이는 데 중요하다. 결과 파일 변환 과정에서는 최종 데이터를 사용자가 접근하기 쉬운 형태로 변환한다. 이는 파일 형식의 변환, 데이터의 압축, 또는 다른 플랫폼에서의 호환성을 위한 조정을 포함할 수 있다. 최종 오류 수정 단계는 이전 단계에서 발견된 모든 오류를 수정하는 과정이다. 이는 프로젝트의 정확도를 최종적으로 보장하며, 품질 관리의 중요한 부분이다. 마지막으로, 보고서 작성 단계에서는 프로젝트의 전체 과정과 결과에 대한 설명을 문서화한다. 이러한 완료 점검 및 보완 단계를 통해 프로젝트의 완성도를 높이고, 최종 결과물의 품질을 보장한다.

이러한 체계적인 접근 방식은 고품질의 데이터를 제공하고, 보다 정확하고 신뢰할 수 있는 정보를 활용할 수 있도록 하게 한다. 이 과정들은 데이터 수집부터 처리, 그리고 활용에 이르기까지 모든 단계에서 철저한 관리와 정밀한 작업을 요구한다. 이는 끊임없이 변화하는 데이터 환경에서 중요한 역할을 하며, 지식 정보의 품질과 가치를 높이는 데 기여하게 된다.

제 2 장

자료의 수집과 선별

2.1. 표 자료 수집과 선별

2.1.1 개요

구분	표 수집 건수	비율
중앙행정기관 보도 자료	2,048	20.3%
공유저작물	5,517	54.7%
국회 예산정책처	2,500	24.8%
합계	10,065	100%

<표 6> 표 자료 수집 건수

표 자료는 상업적 활용과 변형 등 2차적 저작물 작성이 가능한 10,000건 이상의 표와 표를 설명하는 문장을 수집하는 것이다. 공공자료는 자유 이용이 가능한 자료에 공공누리 제1유형()을 부착하여 공개하고 있고 학술자료는 CC BY()로 저작물 이용에 제한을 두고 있어 수집 문서마다 이용 범위를 확인하여야 한다. 자유 이용으로 공개된 중앙 및 지방 행정기관의 보도 자료, 공유저작물 등에서 표와 문장을 추출하였으며, 국회예산정책처에서 작성된 표(국립국어원과 국회예산정책처 업무 협약을 통해 확보)는 원본 문서를 찾아 관련 문장을 추가하는 방법으로 표 자료를 수집하였다. 자료 출처에 따른 수집 비율은 다음 표와 같다.

2.1.2 수집

공공누리 제1유형 저작물은 공공기관 보도 자료 중 공공누리 제1유형으로 서비스하는 곳을 대상으로 하여 표를 포함하고 있는 보도 자료를 수집하였고 학술자료는 공유저작물을 서비스하는 지식공유사이트(http://share.nanet.go.kr/)에서 CC BY 저작물 표시를 확인한 후 표가 있는 문서를 수집하였다. 국회 자료 중 예산정책처 자료는 기관에서 제작한 표를 근거로 원본 문서의 관련 텍스트를 추출하였으며 입법조사처 자료는 서비스 사이트(https://www.nars.go.kr/)에서 표를 포함하고 있는 문서를 수집하였다.

2.1.3 선별 및 가공

수집된 문서 중 표와 표의 특정 셀의 내용을 설명하는 문장이 있는 자료를 선정하였으며 표 및

설명 문장 추출 가공 시 표의 크기가 1페이지를 넘어가거나 제목 행 또는 제목 열이 없는 표는 제외하였다. 설명 문장은 표 셀의 내용을 계산으로만 도출하는 문장이나 1개 이상의 표를 대상으로 하는 문장은 제외하고 표 셀의 내용이 변경되지 않고 명시적으로 들어있는 문장을 추출하였다. 또한 추출된 문장에서도 표의 내용과 직접 관련이 없는 구절은 삭선 표시를 하였다.

보도 자료 외 학술자료는 원본이 피디에프(pdf) 파일로 되어 있어 내용의 복사 추출이 불가능한 것은 입력하거나 추출 문장을 정리하는 작업을 별도로 하였다. 피디에프(pdf) 파일의 표는 변환과정(pdf-엑셀)을 거쳐 한글 파일(hwp)로 재작성하여 1표 1건 1파일로 메타정보 및 내용을 정리하였다.

(작성 예시)

①[파일 번호] TS70658-058

②[문서 제목] 2012 국정감사 정책자료 II: 농림수산식품위원회 : 국유림의 지속적 확대 필요

③[표 제목] 타 용도로의 산지전용 현황

④[보도 일자] 20120820

⑤[저작권] 국회입법조사처

⑥[출처 url]

https://www.nars.go.kr/report/view.do?categoryId=&cmsCode=CM0043&searchType=TITLE&searchKeyword=2012%20%EA%B5%AD%EC%A0%95%EA%B0%90%EC%82%AC%20%EC%A0%95%EC%B1%85%EC%9E%90%EB%A3%8C&brdSeq=240

⑦[수집 문장]

□ 한편 각종 국책사업, 산업용지 공급 등에 따른 산지개발 수요로 인하여 타 용도로의 산지전용이 증가하고 있는데, '10년도 산지전용 현황을 보면 총 전용면적 11,851ha 중 공장 2,240ha(18.9%), 택지 1,355ha(11.4%), 골프장 1,223ha(10.3%), 도로 1,115ha(9.4%), 농업용 450ha(3.8%) 등의 순으로 비농업용이 절반 이상을 차지하고 있음

⑧[표의 내용이 아닌 경우 삭제]

□ 한편 각종 국책사업, 산업용지 공급 등에 따른 산지개발 수요로 인하여 타 용도로의 산지전용이 증가하고 있는데, '10년도 산지전용 현황을 보면 총 전용면적 11,851ha 중 공장 2,240ha~~(18.9%)~~, 택지 1,355ha~~(11.4%)~~, 골프장 1,223ha~~(10.3%)~~, 도로 1,115ha~~(9.4%)~~, 농업용 450ha~~(3.8%)~~ 등의 순으로 비농업용이 절반 이상을 차지하고 있음

⑨[기준 문장]

⑩[표]

(단위: ha)

용도		2008년			2009년			2010년		
		계	보전	준보전	계	보전	준보전	계	보전	준보전
합계		13,739 (100)	4,142 (30)	9,597 (70)	15,877 (100)	5,368 (34)	10,509 (66)	11,851 (100)	4,446 (38)	7,405 (62)
농업용	계	571	147	424	535	169	366	450	175	275
	농지	516	124	392	494	159	335	393	147	246
	초지	55	23	32	41	10	31	57	28	29
비농업용	계	13,168	3,995	9,173	15,342	5,199	10,143	11,401	4,271	7,130
	택지	1,707	88	1,619	1,207	206	1,001	1,355	371	984
	공장	2,253	412	1,841	3,308	666	2,642	2,240	720	1,520
	광업	144	90	54	205	161	44	101	55	46
	도로	1,181	415	766	1,497	580	917	1,115	443	672
	골프장	2,130	1,218	912	2,181	897	1,284	1,223	744	479
	스키장	16	2	14	79	-	79	3	3	-
	묘지	78	28	50	117	77	40	61	30	31
	기타	5,659	1,742	3,917	6,748	2,612	4,136	5,303	1,905	3,398

⑪[출처 표기]]

(국회입법조사처)에 의해 창작된 (2012 국정감사 정책자료 II)은 크리에이티브 커먼즈(=[출처url])에 따라 이용할 수 있습니다.

위 작성 양식에 대해 자세히 설명하면 다음과 같다.

① [파일 번호]: '표(T)+저작물(P-보도 자료, B-국회예산정책처, S-공유저작물)+작업코드(1자리)+일련번호(4자리)+[문서 내 표 일련번호(3자리)]'로 생성하여 기재한다.
예) TB10001-001, TS10001-001

② [문서 제목]: 해당 보도 자료의 게시 제목, 정책자료, 연구보고서의 경우 문서 제목을 기재한다.

③ [표 제목]: 추출하는 표의 제목을 기재한다.

④ [보도 일자]: 해당 자료가 게시된 날짜나 출판일 및 발간자료의 경우 발간일을 기재한다.

⑤ [저작권]: 해당 자료의 저작권자를 기재(공공자료의 경우 기관명)한다.

⑥ [출처 URL]: 해당 자료를 다운로드할 수 있는 주소(URL)를 복사하여 넣는다.

⑦ [수집 문장]: 표의 내용을 구체적으로 기술하고 있는 문장으로 특정 셀의 단어나 수치가 포함되어 있는 문장을 폭넓게 추출한다.

⑧ [표의 내용이 아닌 경우 삭제]: 표와 관련이 없는 수치나 단어, 구문은 삭선으로 표시한다.

⑨ [기준 문장]: 유사 문장 작성팀에서 작성하도록 비워 둔다.

⑩ [표]: 원래의 음영을 제거하여 유사 문장 작성에 필요한 셀을 선택할 수 있도록 한다.

⑪ [출처 표기]: 자료 제공 기관에서 안내하고 있는 저작권 표시에 따른다.

2.1.4 산출물

총 10,065건의 표 및 문장을 수집 가공하였으며 제목, 저작권자, 가공 파일 번호가 기재된 엑셀 파일 1종과 1표 1건으로 된 표 가공 양식 1건을 수록한 한글파일 10,065개를 산출하였다.

2.2. 그림 자료 수집과 선별

2.2.1 개요

그림 자료는 상업적 활용과 변형 등 2차적 저작물 작성이 가능한 것으로 숫자나 한 단어 이상의 텍스트가 명확히 식별되는 그림 자료 9,000건 이상을 수집하는 것이다. 사업 초기 샘플 작업을 위한 사진은 참여 보조원의 기획 촬영으로 확보하였으며 이후 본격적인 수집은 다수 작가의 참여로 다양한 사진 수집이 가능한 공모전을 열어 확보하였다. 사업단 기획 촬영과 공모전의 수집 비율은 다음 표와 같다.

구분	수집방법	수집건수	수집비율
텍스트가 포함된 사진	샘플 수집	172	1.9%
	기획 촬영	544	5.9%
	공모전	8,496	92.2%
합계		9,212	100%

<표 7> 텍스트가 포함된 그림 자료 수집 건수

2.2.2 수집

사업 초기 작업의 빠른 착수를 위해서 기 촬영자료 172건을 수집함과 더불어 사업의 이해도가 충분한 보조원으로 하여금 촬영하게 하여 최종 선별 기준 544건을 확보하였다. 본 작업 자료 수집은 2022년 9월 1일부터 12월 31일까지 크라우드 수집 플랫폼에 4회에 걸쳐 공모전을 열었다. 공모전에는 총 522명이 참여하였으며 1차 2,444건, 2차 2,902건, 3차 2,690건, 4차 460건 총 8,496건을 선별 수집하였다. 각 개인의 저작권, 초상권은 공모절차(저작권양도승인-공모지원-사업자선정-결제완료)를 통해 양도되도록 하였으며 공모주최자와 사업수행자, 주관기관의 3자 저작권이용허락계약을 체결하여 최종적으로 주관기관이 저작권 이용권한을 가지도록 하였다.

2.2.3 선별

텍스트가 포함된 사진의 선별 조건은 다음과 같다.

○ 해상도: 장축 800픽셀 이상
○ 파일의 종류: jpg, jpeg
○ 사진에 한 단어 이상의 텍스트가 포함되어야 한다.
- 텍스트에 한글을 포함하고 있는 경우
- 단순히 한국어를 알파벳으로 표기한 경우 (예시: SEOUL, INCHEON, SAMSUNG, HYUNDAI)
○ 텍스트가 그림의 초점이 되어야 하지만 텍스트의 그림 내 위치는 상관없다.
○ 초점이 되는 텍스트가 명료하여야 한다.
○ 텍스트의 내용은 가능한 한 일반적이고 범용적으로 쓰이는 것으로 한다.
○ 고유명사 위주의 텍스트는 가능한 한 배제하도록 한다.
○ 배경은 텍스트의 맥락을 유추하기에 충분하여야 한다.

2.2.4 산출물

저작권자, 이미지파일명, 원본코드(공모코드)로 된 엑셀 파일 1종과 이미지파일 9,040건이 산출되었다. 공모사진의 파일명은 그림(p)+작업코드(1자리)+일련번호(4자리)로 되어 있으며 사업단에서 기획 촬영한 사진은 PK+일련번호(4자리)를 부여하였다.

2.3. 그래프 자료 수집과 선별

2.3.1 개요

그래프 자료는 상업적 활용과 변형 등 2차적 저작물 작성이 가능한 것으로 공공기관의 제1유형() 공개 또는 자유이용() 공유저작물에서 수집하였다. 샘플 제작 검토 과정에서 표로의 수치전환이 불가능하거나 변수 2개를 초과하는 자료를 배제하게 되어 1차 수집 자료 중 요건에 미달되는 자료를 삭제하고 2차 추가 수집을 진행하였다. 최종 수집 자료는 보도자료 1,017건, 공유저작물 43건 총 1,060건이다.

2.3.2 수집

그래프를 많이 사용하는 문서는 공공기관에서 배포하는 보도 자료였으며 최근 자료부터 열람하여 수집하였다. 기관별 수집 분포는 다음 표와 같다.

기관명	건수	기관명	건수
고용노동부	22	소방청	1
과학기술정보통신부	6	소비자보호원	82
교육부	117	여성가족부	231
국토교통부	29	중소벤처기업부	12
금융위원회	27	통계청	153
농림축산식품부	48	한국소비자원	20
문화체육관광부	86	행정안전부	10
보건복지부	129	국회입법조사처	43
산업통상자원부	44		
합 계			1,060

<표 8> 그래프 자료 수집 건수

2.3.3 선별 및 가공

그래프는 막대형(Bar chart), 선형(Line chart), 산점도(Scatter Plot) 3가지 유형의 그래프로 변수 2개를 넘지 않는 자료를 수집하기로 하였으며 세부 기준은 다음과 같다.

○ 문서 내에 그래프 제목이 반드시 있어야 한다.

○ 그래프에 대한 관련 설명 문장 또는 문단이 함께 있어야 한다.

○ 그래프의 유형

- 막대형(Bar chart), 선형(Line chart), 산점도(Scatter Plot) 3가지 유형 중 하나여야 한다.
- 그래프 이미지의 해상도는 장축 200픽셀 이상이어야 한다.
- 그래프가 포함하고 있는 텍스트(수치, 범례, 단위 포함)가 식별 가능해야 한다.
- 독립 변수는 한 개 또는 두 개여야 한다.

○ 그래프의 종속 변수의 값을 확인할 수 있어야 한다.

세부 기준에 의해 선별된 자료는 1개의 그래프를 1개의 한글 파일(hwp)에 동일한 양식을 적용하여 정리 가공하였다.

(작성 예시)

①[파일 번호] GP10001

②[문서 제목] 디지털 게임 국제거래 소비자 불만 전년 대비 11.3% 증가

③[그래프제목] 디지털 게임서비스 관련 국제거래 소비자상담 접수 현황

④[작성 일자] 2022-06-14

⑤[저작권] 한국소비자원

⑥[출처 url]

https://www.kca.go.kr/home/sub.do?menukey=4002&mode=view&no=1003323529

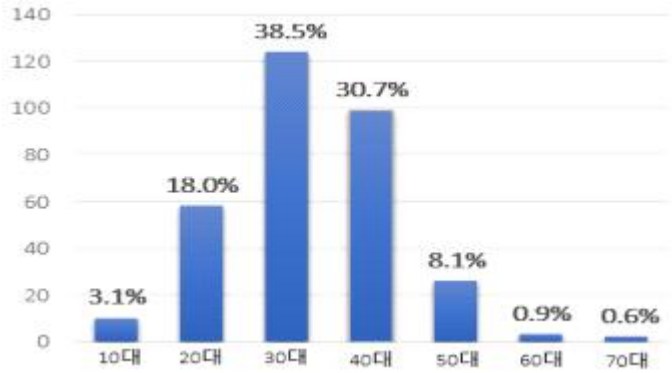
⑦[수집 문장]

(연령별, 성별) 연령이 확인된 322건을 분석한 결과, 30대가 38.5%(124건), 40대가 30.7%(99건)을 차지하고 있는 것으로 나타남.

⑧[기준 문장]

<정제 단계 작업>

⑨[그래프]



⑩ [출처 표기]

본 저작물은 '한국소비자원'에서 '2022-06-14' 작성하여 공공누리 제1유형으로 개방한 '디지털 게임 국제거래 소비자 불만 전년 대비 11.3% 증가'을 이용하였으며, 해당 저작물은 '한국소비자원 (https://www.kca.go.kr/)'에서 무료로 다운받으실 수 있습니다.

위 작성 양식에 대해 자세히 설명하면 다음과 같다.

① [파일 번호]: '그래프(G)+저작물(P-보도 자료, S-공유저작물)+작업코드(1자리)+일련번호(4자리)'로 생성하여 기재한다.

② [문서 제목]: 해당 보도 자료의 게시 제목이나 보고서의 문서 제목을 기재한다.

③ [그래프 제목]: 추출하는 그래프의 제목을 기재한다.

④ [보도 일자]: 해당 자료가 게시된 날짜나 출판일 및 발간자료의 경우 발간일을 기재한다.

⑤ [저작권]: 해당 자료의 저작권자를 기재(공공자료의 경우 기관명)한다.

⑥ [출처 URL]: 해당 자료를 다운로드할 수 있는 주소(url)를 복사한다.

⑦ [수집 문장]: 그래프의 내용을 구체적으로 기술하고 있는 문장이나 문단을 추출한다.

⑧ [기준 문장]: 유사 문장 작성팀에서 작성하도록 비워 둔다.

⑨ [그래프]: 단위와 범례, 수치 등이 누락되지 않도록 범위를 지정하여 이미지화 한다.

⑩ [출처 표기]: 자료 제공 기관에서 안내하고 있는 저작권 표시에 따른다.

2.3.4 산출물

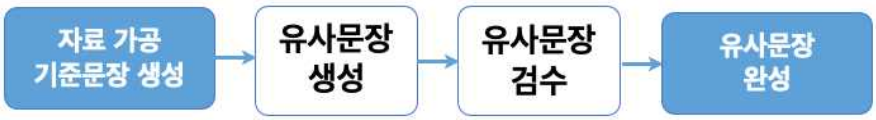
총 1,060건의 그래프 및 관련 문단을 수집하였으며 이에 대해 가공 파일명, 원본 출처 URL 정보, 원본 문서 제목, 저작권 정보를 수록한 엑셀 파일 1종과 1 그래프, 1 파일로 된 한글 파일을 산출하였다.

제 3 장

유사 문장의 생성

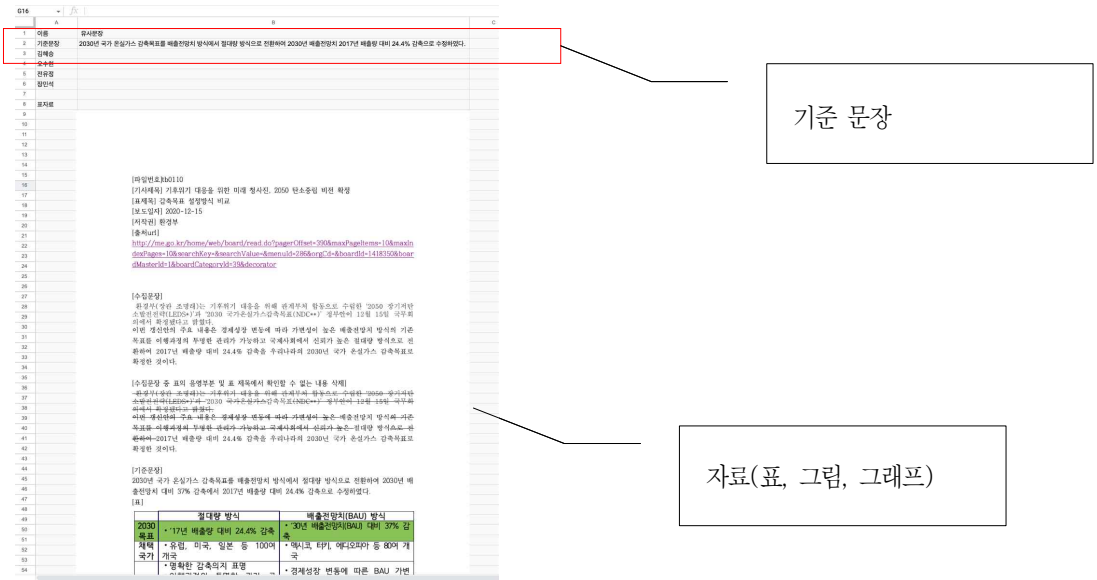
3.1. 유사 문장의 생성 도구 및 절차

유사 문장 생성 작업은 구글 워크스페이스를 사용하여 수행되었다. 구글 워크스페이스는 온라인 협업도구로서, 여러 애플리케이션 중 주로 구글 드라이브와 구글 스프레드시트로 작업이 이루어졌다. 또한 유사 문장 생성은 다음과 같은 절차를 밟았다.



<그림 4> 유사 문장 생성 절차

유사 문장 생성은 구글 드라이브에서 수행되었다. 표, 그림, 그래프(차트)로 나눠진 하위 폴더에서 날짜별로 구성된 스프레드시트에서 작업을 수행하였다. 각 스프레드시트는 자료별로 시트가 구성되어 있다. 한 시트 내에서는 표와 그래프의 경우 다음과 같이 기준 문장과 자료로 구성되었다.



<그림 5> 유사 문장 생성 작업 화면 예시

위에서처럼 본인의 이름 옆에 유사 문장을 작성하도록 하였다. 유사 문장 생성의 평가 기준은 크게 3가지로 요약할 수 있다.

- 문장의 유사성(similarity): 유사 문장의 기준 충족 여부
- 자연성(naturalness): 한국어의 고유한 특성을 충분히 반영한 자연스러운 한국어로 유사 문장이 구성되었는지 여부
- 다양성(variety): 어휘, 형태, 통사, 의미, 정보구조/화용의 각 층위를 모두 고려한 다채로운 문장 구성 여부

위의 기준에 의거 생성된 유사 문장을 종합적으로 검토하여 유사 문장으로서의 가부를 판단하여 부적합하다고 판단된 경우 문장을 수정하여 다시 검수를 받도록 하였다.

3.2. 유사 문장의 생성

3.2.1 표 기반 유사 문장의 생성

표 자료의 경우, 기준 문장은 수집 문장에 드러난 표현을 최대한 그대로 활용하도록 하였다. 표 음영과 관련된 정보만 포함하는 것을 원칙으로 하되, 유사 문장 작업을 고려하여 최대한 풍부한 정보를 포함하도록 하였다. 특히 문서 제목과 표 제목을 적극적으로 활용하여 기준 문장을 작성하였으며, 기준 문장에 이용할 수 있는 기본 정보에는 다음이 포함된다.

- 문서 제목
- 표 제목
- 보도 일자
- 저작권
- 표의 첫 행(heading)
- 단위 정보

예를 들어 수집된 표가 다음과 같을 경우, 잘못 작성된 기준 문장과 올바르게 작성된 기준 문장

은 이렇게 볼 수 있다.

[파일 번호] TS10014-001

[문서 제목] 미래인재 양성을 위한 지역 교육 사례 연구

[표 제목] 지역 규모별 마을공동체 제안

[작성일자] 20191200

[저작권] 국가균형발전위원회

[출처 URL]

http://share.nanet.go.kr/portal/work/workDetail.do?searchType=&queryText=&providing_org=%EA%B5%AD%EA%B0%80%EA%B7%A0%ED%98%95%EB%B0%9C&currPage=1&viewRowCnt=10&order=&type=&control_no=KSDB00000000154967&data_type=MONO&returnUrl=organ

[출처 표기]

(국가균형발전위원회, 한국생산성본부)에 의해 창작된 (미래인재 양성을 위한 지역 교육 사례 연구)은 크리에이티브 커먼즈(=[출처url])에 따라 이용할 수 있습니다.

[수집 문장]

□ 소규모 저개발 지역에서는 부모의 돌봄과 학생들 가정환경의 경제적 여건이 부족한 경우가 많기 때문에 돌봄 공동체, 경제 공동체를 추가해야 함.

□ 소규모 저개발 소멸 위험 지역은 총체적인 상황이 좋지 않기 때문에 통합형으로 마을 주민들과의 협의를 통해 필요한 부분을 총체적으로 관리할 필요가 있음.

[표]

지역 구분	중점 추진 유형
대규모 발전 지역	* 학습 공동체 * 주거 공동체
중간규모 중개발 지역	* 학습 공동체 * 주거 공동체 * 문화 공동체
소규모 저개발 지역	* 학습 공동체 * 주거 공동체 * 문화 공동체

	* 돌봄 공동체 * 경제 공동체
소규모 저개발 소멸 위험 지역	*** 통합형**

- 잘못 작성된 기준 문장: 소규모 저개발 소멸 위험 지역은 총체적인 상황이 좋지 않기 때문에 통합형으로 마을 주민들과의 협의를 통해 필요한 부분을 관리할 필요가 있다.
  잘못 작성된 기준 문장은 음영 처리된 내용을 활용하였으나, '총체적인 상황이 좋지 않기 때문에', '마을 주민들과의 협의를 통해'와 같이 기본 정보와 음영 처리된 표를 통하여 유추할 수 없는 내용임을 알 수 있다.

- 올바르게 작성된 기준 문장: 미래인재 양성을 위한 지역 교육 사례 연구에 따르면, 소규모 저개발 소멸 위험 지역에서는 통합형 마을공동체를 중심으로 관리해야 한다.
  올바르게 작성된 기준 문장은 문서 제목에 있는 '미래인재 양성을 위한 지역 교육 사례 연구'라는 정보와, 표 제목에 있는 '마을공동체 제안'이라는 정보, 그리고 표에 음영 처리된 내용을 활용하였음을 확인할 수 있다.

유사 문장 생성에 있어서는, 기본적으로 기준 문장과 의미가 유사한 문장을 생성하도록 하였다. 단, 표면적 유사성이 지나치게 높아지는 것을 막기 위해, 어순의 조정, 단어의 교체, 문장 구조의 변경 중 최소한 하나의 절차를 수행하도록 한다. 어미와 조사만 교체하는 최소한의 수정은 지양하도록 한다. 단순한 정보로 인해 유사 문장을 만들기 어려울 경우, 사칙연산 등 수치의 단순한 비교와 비교 술어의 사용은 가능하다. 또한, 기준 문장에서 사용한 표의 내용을 제거하는 것은 불가능하다.

예를 들어 수집된 표와 기준 문장이 다음과 같을 경우, 잘못 작성된 유사 문장과 올바르게 작성된 유사 문장은 이렇게 볼 수 있다.

[기준 문장]
정규직 근로자의 월임금총액은 2016년 약 3,283,000원으로 10년 비 증감률이 34.8%이다.

- 잘못 작성된 유사 문장: 2016년 정규직 근로자의 월임금총액은 10년 비 34.8%로 증가된 약

3,283,000원으로, 전체 증감률보다 높은 수치를 기록하고 있다.

- 올바르게 작성된 유사 문장: 2016년 정규직 근로자의 월임금총액은 10년 비 34.8%로 증가된 약 3,283,000원이다.
- 잘못 작성된 유사 문장은 기준 문장의 내용을 활용하였으나, 그에 추가적으로 표에 음영 처리되지 않은 부분인 전체의 월임금총액 10년 비 증감률을 사용하였다. 또한, 아래 작성된 다른 유사 문장과 비슷한 문장 표현으로 작성되었다.
- 올바르게 작성된 유사 문장은 기준 문장에 들어있는 정보만을 활용하여 동일한 내용을 가지고 있는 다른 표현의 문장임을 알 수 있다.

3.2.2 그림 기반 유사 문장의 생성

텍스트가 있는 그림 자료의 경우, 기준 문장에는 그림 내에 있는 구체적 텍스트가 반드시 포함되어야 한다. 또한, 텍스트가 쓰여 있는 장소 혹은 물체에 대한 정보를 포함하여야 한다. 또한, 그림 내에 있는 한글만 포함하도록 하되, 그램(g)이나 미터(m) 등의 도량형 단위나 이미 일반화되어 있는 축약형(B1, B2, AM, PM)은 포함할 수 있다.

예를 들어 수집된 그림이 다음과 같을 경우, 잘못 작성된 기준 문장과 올바르게 작성된 기준 문장은 이렇게 볼 수 있다.

<그림 6> 그림 기반 기준 문장 작성의 예시 1

- 잘못 작성된 기준 문장: 사리원 불고기 식당은 6월 25일 오픈하는 가게로서 3대에 걸쳐 운영되는 가게이다.

잘못 작성된 기준 문장은 그림에 포함된 텍스트 내용과 그림에서 확인할 수 있는 시각적인 내용을 모두 포함하였으나, 로마자로 표기되어 있는 'OPEN' 글자를 '오픈'으로 한글로 읽어 사용하였으므로 부적절하다.

- 올바르게 작성된 기준 문장: 사리원 불고기는 열두 가지의 과일과 야채로 만든 찍어 먹는 소스가 특징이다.

올바르게 작성된 기준 문장은 그림에 포함된 텍스트 내용과 그림에서 확인할 수 있는 시각적인 내용만을 활용하여 작성하였음을 확인할 수 있다.

<그림 7> 그림 기반 기준 문장 작성의 예시 2

- 잘못 작성된 기준 문장: 이 가게에서는 긴급재난지원금카드와 온누리상품권 등 각종 카드로 결제가 가능하다.

잘못 작성된 기준 문장은 그림에서 확인할 수 있는 시각적인 내용을 모두 포함하였으나, '긴급재난지원금카드'와 '각종 카드'라는 문구는 그림에서 잘려 있으므로 사용할 수 없다.

- 올바르게 작성된 기준 문장: 이 생선 가게에서 가자미와 동태는 5천 원, 그리고 오징어는 다섯 마리에 만 원이다.

올바르게 작성된 기준 문장은 그림에 포함된 텍스트 내용과 그림에서 확인할 수 있는 시각적인 내용만을 활용하여 작성하였음을 확인할 수 있다.

유사 문장의 경우, 표 기반 유사 문장과 마찬가지로, 기본적으로 기준 문장과 의미가 같은 문장을 만든다. 기준 문장에서 사용한 그림 내 텍스트(키워드)는 반드시 포함하되, 그 이외의 텍스트는 포함하지 않도록 한다. 어미와 조사만 교체하는 등의 '최소한의 수정'은 지양한다. 또한, 그림에서

확인할 수 있는 정보더라도 기준 문장과 키워드에 포함되어 있지 않은 정보는 사용할 수 없으며, 기준 문장에서 사용한 그림의 내용을 제거하면 안 된다.

예를 들어 수집된 그림과 기준 문장이 다음과 같을 경우, 잘못 작성된 유사 문장과 올바르게 작성된 유사 문장은 이렇게 볼 수 있다.

<그림 8> 그림 기반 기준 문장 작성의 예시 3

[기준 문장]
대한예수교 장로회 소속인 서촌교회 앞에는 빨간색 대형버스가 있다.

- 잘못 작성된 유사 문장: 빨간색 대형버스 뒤로 종탑과 함께 대한예수교 장로회 소속의 서촌교회가 있다.

잘못 작성된 유사 문장은 기준 문장에 포함되어 있는 정보와 키워드를 모두 활용하였으나, 추가적으로 '종탑과 함께'라는 시각적 정보를 포함하였다. 이는 기준 문장에만 적힌 내용 외에 그림에서 확인할 수 있는 정보를 추가적으로 사용하면 안 된다는 지침을 어겼으므로 부적절하다.

- 올바르게 작성된 유사 문장: 빨간색 대형버스 뒤로 대한예수교 장로회 소속의 서촌교회가 있다.

올바르게 작성된 유사 문장은 기준 문장에 포함되어 있는 정보와 키워드를 모두 활용하여 같은 내용을 담은 다른 문장 표현을 사용하고 있음을 알 수 있다.

3.2.3 그래프 기반 유사 문장의 생성

그래프 기반 유사 문장의 생성 지침 작성을 위해 수집된 1158건 중 50%(579건)을 샘플로 하여 그래프 분석 작업을 수행하였다. 아래는 각 그래프 유형별 분포를 보여 준다.

	단순형	복합형(2)	복합형(3)	합계
막대(세로)	216	191	4	411
막대(가로)	26	26	0	52
선그래프	49	59	0	108
산점도	2	0	0	2
기타	6	0	0	6
합계	299	276	4	579

<표 9> 독립변수의 수에 따른 각 그래프 유형별 분포

표에서 보듯이 세로 막대그래프가 전체 579건 중 411건에 해당해 대다수를 차지함(70.98%)을 알 수 있다. 다음으로 선그래프가 108건(18.65%)의 수가 많이 나타났으며, 가로 막대그래프의 경우 52건(8.98%)였고, 산점도 그래프의 경우 2건, 그밖에 6건이 발견되었다. 독립변수가 1개인 단순형의 경우 299건(51.64%), 독립변수가 2개인 복합형의 경우 279건(48.19%), 독립변수가 3개인 복합형의 경우 4건(0.69%)이 발견되었다. 막대그래프와 선그래프 중 독립변수가 1개인 경우와 2개인 경우가 전체 중 567건이 되어 대부분을 차지(97.93%)하고 있다는 점에서, 그래프 유형을 균형있게 생성하기는 어렵고, 수집된 그래프의 유형에 기반하여 말뭉치를 생성하는 방향으로 결정하였다.

독립변수가 2개인 경우 한 독립변수의 경우 범례에 들어가야 하므로 그 변수의 레벨 수, 즉 범례의 항목 수를 샘플에서 살펴본 바에 따르면 다음과 같다.

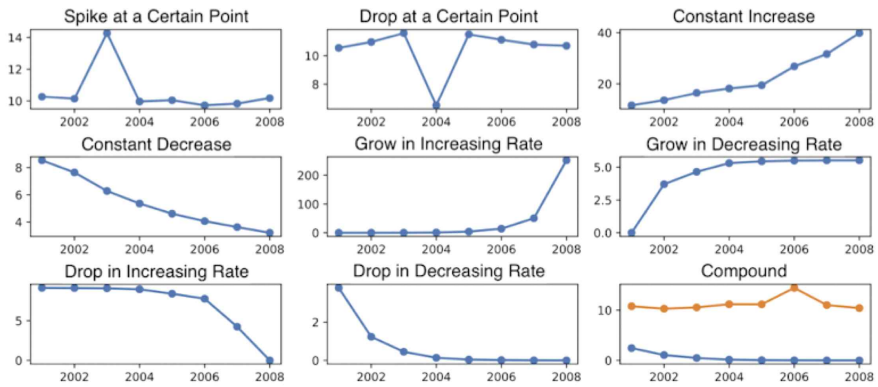
	범례(2)	범례(3)	범례(4)	범례(5)	범례(6)	범례(7)
막대(세로)	139	40	7	6	0	2
막대(가로)	24	0	1	1	0	0
선그래프	43	10	1	3	2	0
산점도	0	0	0	0	0	0
기타	0	0	0	0	0	0
합계	206	50	9	10	2	2

<표 10> 복합형(2)의 경우 범례의 항목 수에 따른 각 그래프 유형별 개수

위에서 보듯이 범례의 수가 2개인 경우 전체 276건 중 206건(74.64%)에 이르는 것을 확인할 수 있다. 유사 문장 생성의 대상이 되는 그래프의 범위를 단순화하는 방향으로 지침을 정하여, 범례의 수가 2개인 경우까지만을 대상으로 선정하기로 하였다.

본 사업에서는 효과적인 데이터 시각화 분석을 위한 기준 문장 작업 지침을 마련하였다. 이 지침은 수집된 문장 또는 문단을 분석하여 정보의 정확성과 문법적 정확성을 확보하고, 그래프 데이터와의 일관성을 유지하는 데 중점을 두고 있다. 수집된 문장 또는 문단을 근거로 하여 국어 문장을 자연스럽게 재구성하는 과정에서 그래프와 관련된 핵심 정보만을 추려내며, 불필요한 문구는 제거한다. 작성된 기준 문단은 문법적으로 정확해야 하며, 띄어쓰기와 맞춤법에 있어 통일성을 유지해야 하며, 기준 문단과 생성 문단 간의 수치 불일치 여부를 면밀히 확인한다.

수집된 문장 또는 문단이 추세곡선에 따른 기준 문장 생성 지침을 충족하지 않을 경우, 해당 지침에 따라 문장을 수정한다. 수집된 문장이 그래프 자료에 대한 충분한 정보를 제공하지 못하는 경우가 많으므로, 그래프의 주요 세 포인트를 기반으로 내용을 보충하고, 증감폭이나 추세 등의 정보를 반영하여 기준 문단을 작성한다. 추세 곡선에 따른 기준 문장 생성 지침은 다음과 같다. 우선 추세곡선은 다음과 같이 8가지이다. 마지막 그래프는 앞의 8가지 추세곡선의 복합형이다.



<그림 9> 그래프 추세 곡선의 유형(Zhu et al., 2021: 4, Figure 3)

그래프에 기반한 유사 문단 생성에 있어 기준 문단 생성 지침은 중요한 역할을 한다. 이 지침은 그래프의 추세를 해석하고, 그 결과를 명확하고 효과적으로 전달하기 위한 구체적인 방법을 제공한다. 그래프에 기반한 유사 문단 생성에 있어 기준 문단 생성의 첫 번째 지침은 그래프의 개요를 제공하는 것이다. 이는 그래프의 유형과 주요 내용을 명시하여 그래프의 맥락을 이해할 수 있게 한다. 예를 들어, “이 그래프는 1970년과 1990년 사이 한국에서의 패스트푸드 소비량을 보여준다.”와 같은 문장이 이에 해당한다.

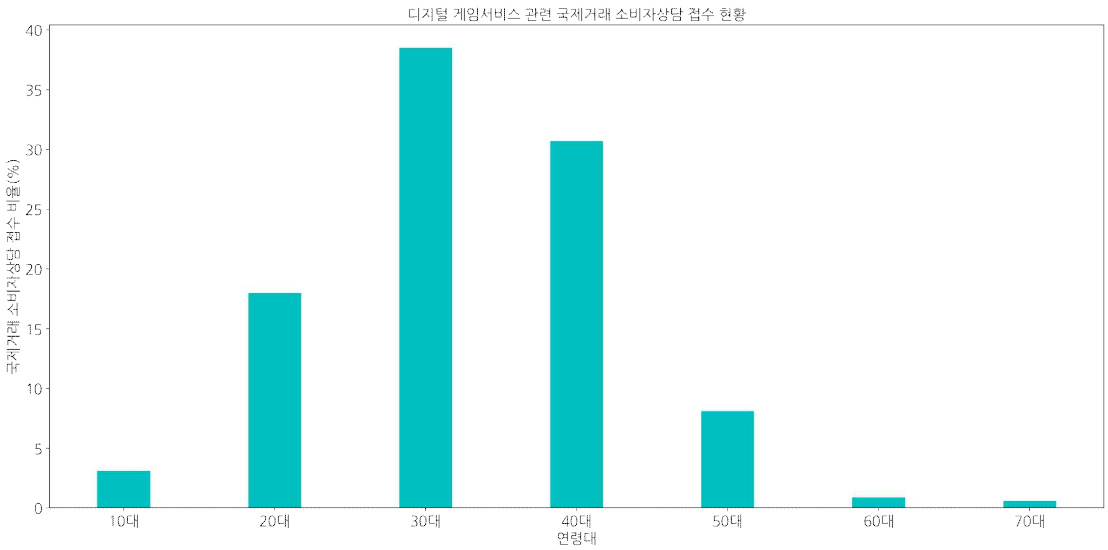
두 번째 지침은 그래프의 수의적 묘사에 초점을 맞춘다. 이는 그래프의 형상 및 요소 등 그래프를 구성하는 시각적 요소들에 대한 설명을 포함한다. 예를 들어, 유럽의 여러 국가를 표본으로 하는 그래프의 경우, “표본 추출한 국가는 핀란드, 프랑스, 조지아, 독일, 그리스, 헝가리 등 모두 유럽 국가이다”와 같은 설명이 이 지침을 충족시킨다. 세 번째 지침은 그래프 정보의 해석에 관한 것이다. 이 지침은 단순한 관찰을 넘어서, 그래프에서 보이는 추세나 변화를 구체적인 숫자를 통해 해석하고 설명하는 것을 요구한다. 예를 들어, “1970년에 소비량은 주당 300그램 정도였다가 1990년에 주당 220그램으로 하락했다.”는 구체적인 수치를 이용하여 변화를 설명하는 좋은 예이다. 네 번째 지침은 평가적 표현을 포함할 수 있는 옵션을 제공해야 한다는 것이다. 이는 특정 값이나 비교에 대한 해석이나 평가를 포함할 수 있음을 의미한다. 마지막으로, 다섯 번째 지침은 그래프를 바탕으로 한 결론, 요약, 예측, 또는 함의를 포함해야 한다고 명시한다. 이는 그래프에서 도출된 정보를 종합하여, 더 넓은 맥락에서의 의미나 잠재적인 영향을 독자에게 전달하는 것을 목표로 한다. 추세 분석에 대한 지침은 그래프에서 나타나는 추세를 정확하고 명확하게 분석하고 전달하는 데 중점을 둔다. 이는 그래프의 요소 간 차이를 검토하여 추세가 분명히 드러나는지 확인하고, 모호한 표현 대신 구체적인 수치를 제시하여 추세를 설명하는 것을 포함한다. 이러한 접근 방식은 그래프 해석의 정확성을 높이고, 정보를 보다 쉽게 이해할 수 있도록 돕는 역할을 한다.

종합적으로, 이 지침들은 그래프로부터 정보를 효과적으로 해석하고 전달하는 데 필수적인 요소들을 제시한다. 구체적인 예와 함께 명확한 해석, 정확한 수치 제시, 데이터 시각화의 전달력을 극대화하는 데 목표를 둔다. 특히, 분석 내용은 가능한 한 구체적으로 제시되어야 한다. 이를 위해 그래프에서 주어진 수치를 명확하게 서술하며, 최고 및 최저 수치를 포함하여 설명한다. 또한, 특정 값이나 영역에 치우치지 않고 그래프 내용을 전반적으로 포괄할 수 있도록 서술한다. 추세곡선을 기반으로 기준 문단을 생성하는 것도 중요하며, 이 연구에서는 총 8가지의 추세곡선을 식별하였으며, 마지막 그래프는 앞서 언급한 8가지 추세곡선을 종합한 복합형으로 구성된다. 이러한 지침

은 데이터 시각화의 해석과 분석에 있어서 일관성과 정확성을 확보하는 데 필수적이며, 연구의 신뢰성을 높이는 데 기여할 것으로 기대된다.

유사 문단 생성의 기본적인 목적은 기준 문단과 의미적으로 유사한 문단을 만드는 것이다. 이 과정은 원본 문단의 핵심 내용을 유지하면서 새로운 형식으로 표현하는 것을 포함한다. 중요한 점은 기존 문단과 의미적 유사성을 유지하되, 표면적으로는 다른 형태를 갖도록 하는 것이다. 이를 위해 어순 조정, 단어 교체, 문장 구조 변경 등의 방법이 적용된다. 문장 내에서 단어의 순서를 변경하거나, 유사한 의미의 다른 단어로 교체하거나, 문장의 구조 자체를 변경하는 것이 이에 해당할 수 있다. 이러한 변화 과정에서는 어미와 조사의 단순 교체에 의존하는 것을 피해야 한다. 이는 표면적인 변화에 그치고 근본적인 내용의 차이를 만들어내지 못하기 때문이다. 단순한 정보의 제한으로 인해 문단 생성이 어려운 경우가 있을 수 있다. 이러한 상황에서는 사칙연산, 수치 비교, 비교 술어의 사용 등을 통해 문단을 구성할 수 있다. 이는 복잡한 내용을 단순화하면서도 필요한 정보를 전달하는 효과적인 방법이다. 또한, 유사 문단 생성의 원칙은 기준 문단 생성과 동일하게 적용되어, 문서 전체의 일관성을 유지하고 콘텐츠의 질을 높이는 데 기여하게 된다. 이러한 접근은 정보 전달의 명확성을 보장하고, 전반적인 문서의 질을 향상시키는 데 중요한 역할을 한다.

❍ 그래프 기반 유사 문장의 작성 예시



<그림 10> 그래프 기반 유사 문장의 작성 예시

[그래프 정보]

· 문서 제목: 디지털 게임 국제거래 소비자 불만 전년 대비 11.3% 증가

· 문서 저자: 한국소비자원

· 보도 일자: 2022-06-14

· 표 제목: 디지털 게임서비스 관련 국제거래 소비자상담 접수 현황

[그래프와 함께 수집한 문단]

□ (연령별, 성별) 연령이 확인된 322건을 분석한 결과, 30대가 38.5%(124건), 40대가 30.7%(99건)을 차지하고 있는 것으로 나타남.

[기준 문단]

이 그래프는 10대에서 70대까지 세대별로 디지털게임 서비스 관련 국제 거래 소비자 상담 접수 현황을 분석한 것이다. [← 지침 1, 2에 따라 수정] 그래프에 따르면, 30대가 38.5%, 40대가 30.7%로 가장 많이 차지하는 것으로 조사되었다. [← 지침 3, 5에 따라 수정]

[유사 문단]

이 그래프는 디지털게임 서비스 관련 국제 거래 소비자 상담 접수 현황을 살펴본 것이다. 10대에서 70대까지 세대별로 분석한 이 그래프에 따르면, 30대와 40대가 각각 38.5% 그리고 30.7%로 가장 높게 나타났다.

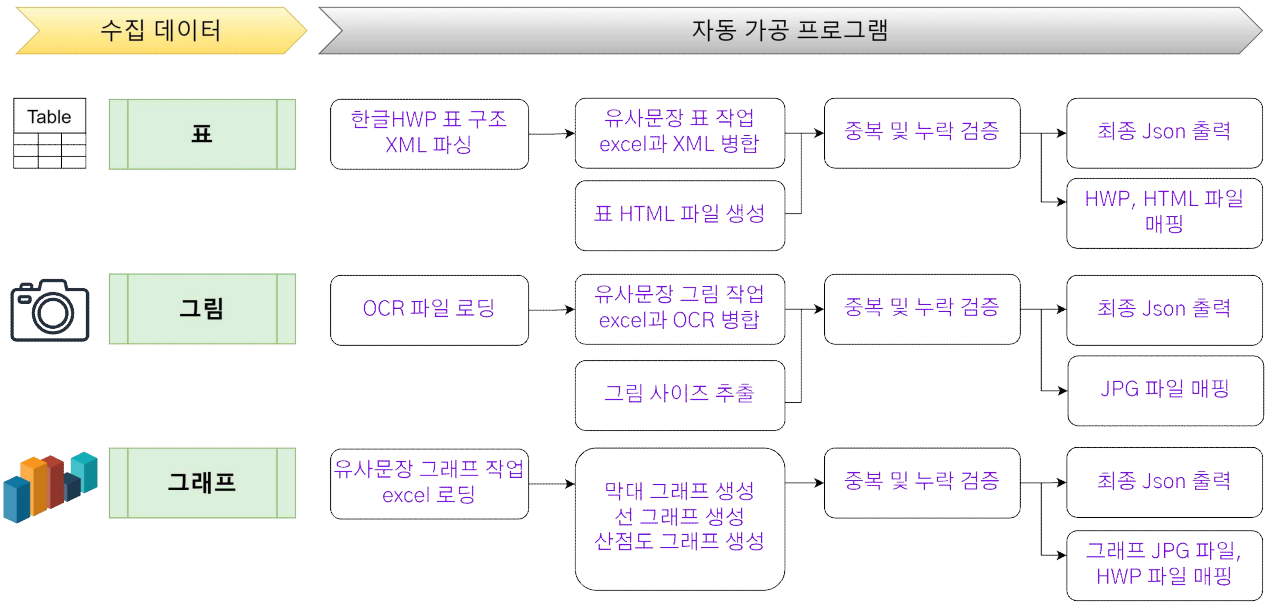
제 4 장

자료 가공 및 검증

4.1. 자료 가공

자료 선별 기준에 따라 수집한 자료 및 유사 문장 작업자가 작업한 결과물은 인공 지능 학습용으로 가공이 필요하다. 표, 그림, 그래프 각각의 데이터 양식에 따라 적절한 자동 가공 방식을 채택하여 최종적으로 작업자가 유사 문장을 작성할 수 있는 파일을 생성하였다.

4.1.1 개요

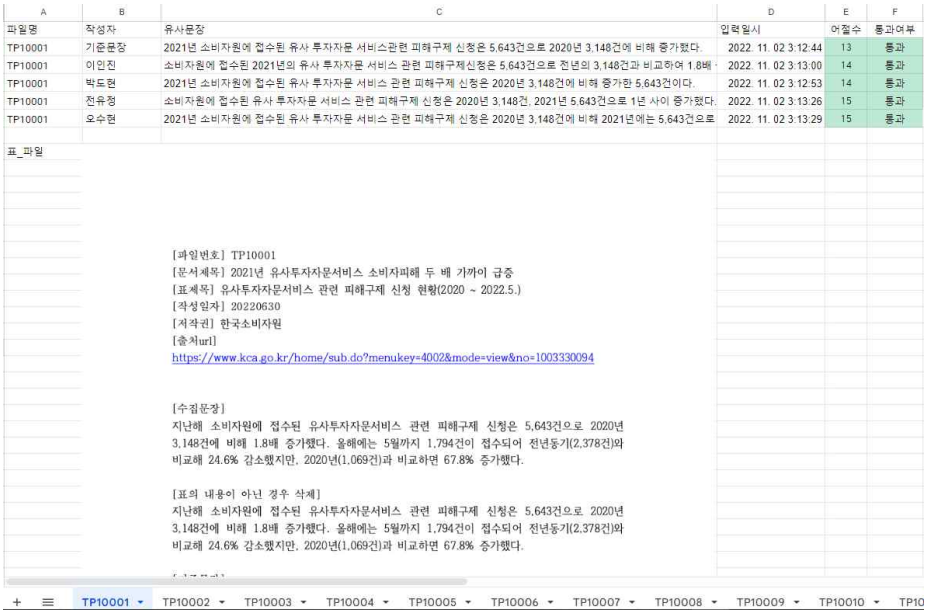


<그림 11> 데이터 가공 절차

<그림 11>은 작업자들이 작업한 엑셀 파일을 불러와서 최종 제이슨(json)을 산출하는 자동 가공 프로그램의 흐름을 나타낸다. 자동 가공 프로그램은 파이썬 스크립트로 실행되고, 크게 표, 그림, 그래프 세 가지를 처리하는 구조로 되어 있다. 앞선 표 자료와 그림 자료, 그래프 자료의 수집 및 가공된 데이터를 대상으로 병합 과정을 거친 후 중복 및 누락을 검증하고 최종 제이슨(json) 형식으로 산출한다. 제이슨(json) 파일로 선정된 아이디(ID)와 한글 파일(hwp), 표 HTML, 그래프 JPG 파일들을 매핑하여 별도의 폴더에 저장한다.

4.1.2 유사 문장 작업자용 엑셀 생성: 표, 그림

기준 문장만 있는 표의 한글 파일을 텍스트 파일로 변환하고, 텍스트 파일의 정보를 읽어서 엑셀로 자동으로 저장하는 파이썬 스크립트를 작성하였다. 그림의 경우에는 표의 엑셀 시트 양식을 기반으로 그림을 삽입해서 넣었다. 엑셀 시트의 수는 한글 파일의 아이디(ID) 수와 같다. 한 엑셀 파일의 시트의 수는 대략 100개 내외로 저장되었다.



파일명	작성자	유사문장	입력일시	어절수	통과여부
TP10001	기준문장	2021년 소비자원에 접수된 유사 투자자문 서비스관련 피해구제 신청은 5,643건으로 2020년 3,148건에 비해 증가했다.	2022. 11. 02 3:12:44	13	통과
TP10001	이인진	소비자원에 접수된 2021년의 유사 투자자문 서비스 관련 피해구제신청은 5,643건으로 전년의 3,148건과 비교하여 1.8배	2022. 11. 02 3:13:00	14	통과
TP10001	박도현	2021년 소비자원에 접수된 유사 투자자문 서비스 관련 피해구제 신청은 2020년 3,148건에 비해 증가한 5,643건이다.	2022. 11. 02 3:12:53	14	통과
TP10001	전유정	소비자원에 접수된 유사 투자자문 서비스 관련 피해구제 신청은 2020년 3,148건, 2021년 5,643건으로 1년 사이 증가했다.	2022. 11. 02 3:13:26	15	통과
TP10001	오수현	2021년 소비자원에 접수된 유사 투자자문 서비스 관련 피해구제 신청은 2020년 3,148건에 비해 2021년에는 5,643건으로	2022. 11. 02 3:13:29	15	통과

표_파일

[파일번호] TP10001
[문서제목] 2021년 유사투자자문서비스 소비자피해 두 배 가까이 급증
[표제목] 유사투자자문서비스 관련 피해구제 신청 현황(2020 ~ 2022.5.)
[작성일자] 20220630
[저작권] 한국소비자원
[출처url]
https://www.kca.go.kr/home/sub.do?menukey=4002&mode=view&no=1003330094

[수집문장]
지난해 소비자원에 접수된 유사투자자문서비스 관련 피해구제 신청은 5,643건으로 2020년 3,148건에 비해 1.8배 증가했다. 올해에는 5월까지 1,794건이 접수되어 전년동기(2,378건)와 비교해 24.6% 감소했지만, 2020년(1,069건)과 비교하면 67.8% 증가했다.

[표의 내용이 아닌 경우 삭제]
지난해 소비자원에 접수된 유사투자자문서비스 관련 피해구제 신청은 5,643건으로 2020년 3,148건에 비해 1.8배 증가했다. 올해에는 5월까지 1,794건이 접수되어 전년동기(2,378건)와 비교해 24.6% 감소했지만, 2020년(1,069건)과 비교하면 67.8% 증가했다.

TP10001 TP10002 TP10003 TP10004 TP10005 TP10006 TP10007 TP10008 TP10009 TP10010 TP10

<그림 12> 표 작업 엑셀 파일 예시

파일명	이름	유사문장	입력일시	어절수	통과여부
P10001	이주연	오른쪽으로 100미터를 가면 대왕릉이 나오고 왼쪽으로 80미터를 가면 소왕릉이 나온다.	2022. 09. 12 3:5	10	통과
P10001	오수현	대왕릉과 소왕릉으로 가는 길이 안내판 위에 새겨져 있다.	2022. 09. 12 3:5	8	통과
P10001	이인진	현재 위치에서 100미터 우측으로 가면 대왕릉이 나오고 80미터 좌측으로 가면 소왕릉이 나온다.	2022. 09. 12 3:5	12	통과
P10001	박도현	대왕릉으로 가려면 오른쪽으로 100미터를 가야 하고 소왕릉으로 가려면 왼쪽으로 80미터를 가야 한다.	2022. 09. 12 3:5	12	통과
P10001	전유정	현재 장소를 기점으로 우측으로 100미터 가면 대왕릉이, 좌측으로 80미터 가면 소왕릉이 나온다.	2022. 09. 12 3:5	12	통과

사진

P10001 P10002 P10003 P10004 P10005 P10012 P10006 P10007 P10008 P10009 P10010

<그림 13> 그림 작업 엑셀 예시

엑셀의 열은 “파일명(ID), 작성자, 유사 문장, 입력일시, 어절 수, 통과 여부”로 구성되어 있다. 파일명은 한글 파일의 아이디(ID)가 입력되고, 유사 문장은 한글 파일에 입력된 유사 문장을 불러온다. 입력 일시는 작업자들이 작업할 일시가 기록되고, 어절 수는 유사 문장 및 작업자들이 작업한 문장의 어절 수로 계산된다. 통과 여부는 작업자가 작업한 문장을 검수자가 검수한 후 이상이 없을 시 통과로 기록한다. 해당 엑셀 작업은 구글 드라이브로 작업하여 버전별로 형상관리되었다.

4.1.3 유사 문장 작업자용 엑셀 생성: 그래프

A	B	C	D	E	F	G
파일명	작성자	유사문장	입력일시	어절수	통과여부	
GP10003	기준문장	이 그래프는 10대부터 70대까지 연령별로 디지털 게임서비스 관련 국제거래 소비자상담 접수 현황을 나타낸 것이다. 연령대 중 30대가	2023. 02. 15 10:42:21	48	통과	
GP10003	정우석	이 그래프는 디지털 게임서비스 관련 국제거래 소비자상담 접수 현황을 보여준다. 10대에서 70대까지 세대별로 분석한 그래프	2023. 02. 15 10:41:32	56		
id	GP10003					
title	디지털 게임 국제거래 소비자 불만 전년 대비 11.3% 증가					
chart_title	디지털 게임서비스 관련 국제거래 소비자상담 접수 현황					
date	2022-06-14					
publisher	한국소비자원					
url	https://www.kca.go.kr/home/sub.do?menukey=4002&mode=view&no=1003323529					
chart_type	세로막대					
xlabel	세대별					
ylabel	소비자 상담 접수율(%)					
legend1						
legend2						
10대	20대	30대	40대	50대	60대	70대
3.1	18	38.5	30.7	8.1	0.9	0.(

<그림 14> 그래프 작업 엑셀 파일 예시

유사 문장 작업자용 그래프 엑셀 파일의 생성 결과는 <그림 14>와 같다. ‘파일명, 작성자, 유사 문장, 입력일시, 어절 수, 통과 여부’로 구성되어 있고, 하단에는 수집한 그래프에 대한 정보가 담긴 한글 파일에서 추출한 정보가 입력된다. ‘title’은 문서의 제목을 나타내고, ‘chart_title’은 그래프의 제목을 나타낸다. ‘xlabel, ylabel’은 각각 그래프 x, y축의 제목을 나타내고, ‘legend1, legend2’는 범례를 나타낸다. 범례 하단의 첫 번째 행은 x축의 자료 값을 나타내고, 두 번째 행은 y축의 자료 값을 나타낸다.

4.2. 자료 검증

구축된 데이터에 대한 유효성 검증은 '2021 언어 능력 평가 체계'의 기준 모델에 학습하여 평가 검증을 수행하였으며 해당 모델 개발 업체인 테디썸에서 학습 및 평가 검증을 담당하였다. 그러나 이미지 데이터의 경우 이미지 기반 문장 생성 데이터를 기반으로 하는 모델로 개발되었기 때문에 2022년의 이미지 내 텍스트를 기반으로 한 문장 생성 데이터를 평가하는 데에는 한계가 있었으며 검증 결과가 현저히 낮게 나오는 원인이 되었다. 또한 그래프 데이터는 비교 검증을 수행할 기존 모델이 없고 구축 데이터도 1,000여 건에 불과하여 인공지능학습 데이터로서의 유효성 평가를 수행하기에 적절하지 않아 유효성 평가에서 제외하였다.

4.2.1. 표 기반 텍스트 생성 데이터 검증

표 기반 텍스트 생성은 구조화된 형식의 표 데이터로부터 설명문을 만드는 태스크이다. 본 평가에는 BART 기반의 모델인 KoBART와 TeddyBART를 사용하여 평가가 이루어졌다. 해당 모델은 '2021 언어 능력 평가 체계'의 기준 모델인 korean_T2T_baseline 모델이 사용되었다. (https://github.com/teddysum/korean_T2T_baseline)

2021년 데이터세트의 경우 하나의 표 데이터로부터 5개의 정답 설명문이 존재하여, 각 데이터를 복수 정답으로 간주하여 학습 데이터에 포함시켰고, 평가 시 각 정답문에 대한 점수의 평균값을 최종 점수로 사용하였다.


```
  "output": [
      "협약 사업장의 감축량은 4,571톤, 비협약 사업장의 감축량은 539톤이다.",
      "협약 사업장과 비협약 사업장의 감축량은 각각 4,571톤, 539톤이다.",
      "협약 사업장의 감축량은 4,571톤인데 비해 비협약 사업장의 감축량은 539톤에 그쳤다.",
      "굴뚝원격감시체계 설치 사업장 중 협약 사업장의 감축량은 4,571톤, 비협약 사업장의 감축량은 539톤으로 나타났다.",
      "굴뚝원격감시체계 설치 사업장의 오염물질 감축량은 협약 사업장 4,571톤, 비협약 사업장 539톤으로 나타났다."
  ]
```


2022년 데이터세트의 경우 하나의 정제된 정답문이 있어 이를 학습데이터와 평가에 활용하여 진행하였다. 2022년 데이터세트의 예는 아래와 같다.

"sentence_annotation": {

"sentence_after_deletion": "소비자피해 유형 중 품질불량·하자에 관한 피해에 응답한 비율은 전체의 45.1%인 69명이고, A/S불량에 관한 피해에 응답한 비율은 전체의 43.1%인 66명이다.",

"worker1": "소비자피해 유형을 조사 결과에 따르면, 각각 69명(45.1%)와 66명(43.1%)이 품질불량·하자 유형과 A/S불량에 관한 피해 유형으로 응답했다.",

"worker2": "소비자피해 유형 중 품질불량·하자에 관한 피해에 응답한 비율과 A/S불량에 관한 피해에 응답한 비율은 각각 전체의 45.1%(69명), 43.1%(66명)이다.",

"worker3": "한국소비자원에서 조사한 소비자피해 유형 중 품질불량·하자에 응답한 비율은 전체 45.1%(69명)를, A/S불량에 응답한 비율은 전체 43.1%(66명)를 차지하여 두 항목은 비슷한 응답률을 보였다.",

"worker4": "소비자피해 유형을 조사한 결과, 전체 응답자 중 69명(45.1%)이 품질불량·하자에, 66명(43.1%)이 A/S 불량에 응답했다."

},

데이터세트는 학습 및 평가를 위해 아래와 같이 8:1:1 비율로 임의로 분류하였다. 사용한 데이터세트의 분포는 다음과 같다.

학습	개발	평가	총 합
8,000	1,000	1,112	10,112

<표 11> 표 기반 텍스트 생성 데이터세트의 분포

성능평가 결과는 아래와 같다.

데이터세트	모델	평가 결과	
		ROUGE-1	BLEU
2021 데이터	KoBART	41.47	42.46
	TeddyBART-small	44.46	43.98
2022 데이터	KoBART	37.15	28.48
	TeddyBART-small	36.12	28.32

<표 12> 표 기반 텍스트 생성 데이터세트의 성능 평가 결과

성능 평가에 대한 의견

2021년 데이터의 경우 5개의 복수 정답 요약문이 있는 것으로 가정하였기에 상대적으로 데이터 양이 많고 또한 평가에 있어서도 정답이 5개로서 대체로 성능이 잘 나올 수 있는 상황임을 감안할 때 크게 유효성 평가에서 문제가 있어 보이지는 않는다. 그러나 2021년 데이터에 비해 비교적 난도가 높고 설명문의 길이가 길어 복잡성이 늘어난 것은 명확해 보인다.

4.2.2. 이미지 기반 텍스트 생성 데이터 검증

이미지 기반 텍스트 생성은 이미지(그림) 데이터로부터 설명문을 만드는 태스크이다. 본 평가에는 이미지 인코딩을 위해 ViT 모델이 사용되었고 텍스트 생성을 위해 KoGPT-2 모델이 사용되었다. 해당 모델은 '2021 언어 능력 평가 체계'의 기준 모델인 korean_IC_baseline 모델이 사용되었다. (https://github.com/teddysum/korean_IC_baseline)

2021년 데이터세트의 경우 하나의 이미지 데이터로부터 5개의 정답 설명문이 존재하여, 각 데이터를 복수 정답으로 간주하여 학습 데이터에 포함시켰고, 평가 시 각 정답문에 대한 점수의 평균값을 최종 점수로 사용하였다.


```
{
    "id": "nikluge-2022-image-train-000001",
    "input": "K0A0002",
    "output": [
        "오락실에서 한 남성이 오락기 위에 가방을 둔 채 손을 빠르게 움직이며 게임에 열중해 있다.",
        "한 소년이 오락실에서 게임에 집중하고 있다.",
        "까만 옷과 안경을 착용한 남자가 오락실에서 게임기를 양손으로 조작하고 있다.",
        "안경을 쓴 남자아이가 오락실에서 게임기 화면을 바라보며 버튼을 누르고 있다.",
        "한 남자가 검은 옷을 입고 오락실에서 게임을 즐기고 있다."
    ]
}
```


2022년 데이터세트의 경우 하나의 정제된 정답문이 있어 이를 학습데이터와 평가에 활용하여 진행하였다. 2022년 데이터세트의 예는 아래와 같다. 2021년 데이터세트와 차이점은 설명문이 이미지에 있는 텍스트에 대한 정보를 포함하고 있다는 점으로, 오시알(OCR, Optical Character

Recognition) 기술을 요구하는 것으로 보이나 본 평가에서는 오시알(OCR)을 사용하지 않고 기존의 기준 모델을 그대로 사용하여 평가가 이루어졌다.


```
    "sentence_annotation": {
      "reference_sentence": "엘리베이터 버튼 위에 붙어 있는 안내문은 유모차와 장애인 우선 탑승을 안내하고 있다.",
      "worker1": "엘레베이터 오른쪽에 붙어 있는 안내문에는 유모차와 장애인이 우선 탑승할 수 있도록 안내하고 있다.",
      "worker2": "엘레베이터 오른쪽에 부착된 안내문은 유모차, 장애인 우선 탑승을 안내하고 있다.",
      "worker3": "2호기 표지판 아래에 부착된 안내문은 유모차, 장애인 우선 탑승을 지시하고 있다.",
      "worker4": "엘레베이터 층수 버튼 상단에 유모차 장애인 우선이라고 적힌 안내문이 부착되어 있다."
    },
    "image_width": 4032,
    "image_height": 3024,
    "ocr_info": [
      {
        "words": "유모차 장애인 우선",
        "type": "rect",
        "bbox": {
          "x": 3118,
          "y": 1663,
          "width": 335,
          "height": 162
        }
      }
```


데이터세트는 학습 및 평가를 위해 아래와 같이 8:1:1 비율로 임의로 분류하였다. 사용한 데이터세트의 분포는 다음과 같다.

학습	개발	평가	총 합
7,369	920	923	9,212

<표 13> 그림 기반 텍스트 생성 데이터세트의 분포

성능평가 결과는 아래와 같다.

데이터세트	모델	평가 결과	
		ROUGE-1	BLEU
2021 데이터	ViT+KoGPT-2	50.71	74.19
2022 데이터	ViT+KoGPT-2	26.34	21.85

학습	개발	평가	총 합
7,369	920	923	9,212

<표 14> 그림 기반 텍스트 생성 데이터세트의 성능 평가 결과

성능 평가에 대한 의견

2022년 데이터의 경우 2021년 데이터와 달리 OCR를 요구하는 태스크로서 기존의 모델로서는 한계가 있음이 명확해 보이며, 2021년 데이터에 비해 비교적 난도가 높은 과업임을 알 수 있다.

이미지	설명문
	<정답> 인조 잔디로 덮여 있는 원형 통이 땅에 매립되어 있고 그 주위로 울타리가 쳐져 있으며 그 옆에는 매립가스 포집 시설이므로 주의할 것을 알리는 경고판이 세워져 있다.
	<예측> 잔디밭 위에 세워진 안내문에 따르면 이 잔디밭은 잔디밭이 아니므로 잔디밭을 이용하지 말라는 안내문이 있다.

<그림 15> 그림 기반 유사 문장 평가 결과 예시 1

이미지	설명문
	<정답> 엄청난 양의 짚으로 만든 커다란 반달곰의 형상이 세워져 있고, 그 옆에는 반달가슴곰이라고 적힌 안내판이 놓여 있다.
	<예측> 길가에 세워진 표지판에는 길가에 있는 나무 여러 그루가 심겨 있고, 그 뒤로 나무 여러 그루가 심겨 있다.

<그림 16> 그림 기반 유사 문장 평가 결과 예시 2

4.3 말뭉치 구축

4.3.1 표 말뭉치 구축

표 기반 유사 문장 말뭉치의 제이슨(json) 형식은 다음 <표 15>와 같다.

1수준	2 수준	3 수준	4 수준	5 수준	타입	설명
id					str	
metadata					obj	*파일의 메타정보
	title				str	제목
	creator				str	*생성자: 국립국어원
	distributor				str	*배포자: 국립국어원
	year				str	*작업세트생성년도(2022)
	category				str	자료분류(표)
	annotation_level				str	유사 문장 생성
	sampling				str	본문 일부

document					arr	*문서정보
	id				str	*표 아이디(ID)
	metadata				obj	*표의 메타정보
		title			str	*문서 제목
		table_title			str	*표 제목
		date			str	*작성일시, 게시일시, 크롤링일시
		publisher			str	*공공기관, 신문사
		url			str	*URL 주소(표 출처)
		highlighted_cells			arr	*하이라이트 셀
	sentence_annotation				obj	*음영표시된 셀을 서술하는 원문과 수정한 문장 유사 문장 서술
		sentence_after_deletion			str	기준 문장
		worker1			str	*유사 문장1
		worker2			str	*유사 문장2
		worker3			str	*유사 문장3
		worker4			str	*유사 문장4
	table				arr	*표
		value			str	*셀의 값, 표의 처음 셀부터 시작
		is_header			bool	*셀이 header인지 표시
		col			int	열 번호
		colspan			int	확장된 열 번호
		row			int	행 번호
		rowspan			int	확장된 행 번호

<표 15> 표 기반 유사 문장 말뭉치의 제이슨(json) 형식

실제로 위의 형식이 적용된 표의 표본을 기준 문장과 음영이 표시된 것을 제시하면 다음과 같다.

[기준문장]
2019년 전국에 위치한 요양병원은 총 1,565개이며 그 중 강원도에는 총 34개의 요양병원이 존재한다.

[표]

(단위 : 개)

	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
전국	868	988	1,103	1,232	1,337	1,372	1,428	1,529	1,560	1,565
강원	20	21	24	26	27	31	31	30	33	34

<그림 17> 표 자료의 기준 문장과 음영이 표시된 예

자동 가공 프로그램은 엑셀의 파일들의 시트들을 순회하며 표의 아이디(ID)와 해당 시트 이름이 일치하는지 확인하고, 일치하는 항목이 있으면 하이라이트 셀(metadata['highligted_cells'])을 찾아 업데이트하고, 시트의 모든 행을 순회하며 5개의 유사 문장을 추출하고 제이슨(json)으로 산출한다.

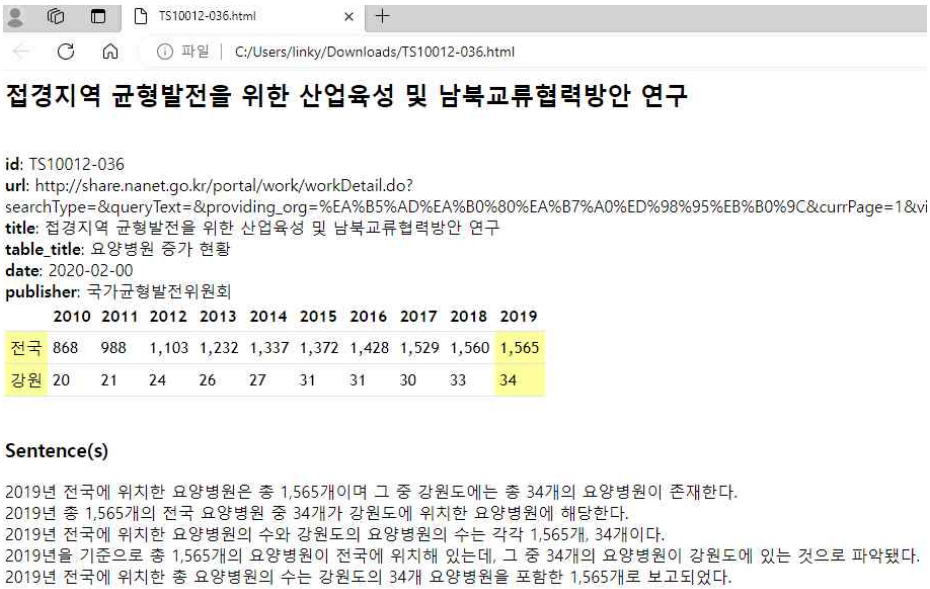
다음은 위 <그림 17>의 표를 제이슨(json) 형식에 맞게 산출한 결과를 보인 것이다.


```
{
    "id": "TS10012-036",
    "metadata": {
      "title": "접경지역 균형발전을 위한 산업육성 및 남북교류협력방안 연구",
      "table_title": "요양병원 증가 현황",
      "date": "2020-02-00",
      "publisher": "국가균형발전위원회",
      "url":
"http://share.nanet.go.kr/portal/work/workDetail.do?searchType=&queryText=&providing_org=%EA%B5%AD%EA%B0%80%EA%B7%A0%ED%98%95%EB%B0%9C&currPage=1&viewRowCnt=10&order=&type=&control_no=KSDB00000000154969&data_type=MONO&returnUrl=organ",
      "highlighted_cells": [
        [
          0,
          1
        ],
        [
          10,
          1
```

```
            ],
            [
              0,
              2
            ],
            [
              10,
              2
            ]
          ]
        },
        "sentence_annotation": {
          "sentence_after_deletion": "2019년 전국에 위치한 요양병원은 총 1,565개이며 그 중 강원도에는 총 34개의 요양병원이 존재한다.",
          "worker1": "2019년 총 1,565개의 전국 요양병원 중 34개가 강원도에 위치한 요양병원에 해당한다.",
          "worker2": "2019년 전국에 위치한 요양병원의 수와 강원도의 요양병원의 수는 각각 1,565개, 34개이다.",
          "worker3": "2019년을 기준으로 총 1,565개의 요양병원이 전국에 위치해 있는데, 그 중 34개의 요양병원이 강원도에 있는 것으로 파악됐다.",
          "worker4": "2019년 전국에 위치한 총 요양병원의 수는 강원도의 34개 요양병원을 포함한 1,565개로 보고되었다."
        },
        "table": [
          {
            "value": "",
            "is_header": true,
            "col": 0,
            "colspan": 1,
            "row": 0,
            "rowspan": 1
          },
          {
            "value": "2010",
            "is_header": true,
            "col": 1,
            "colspan": 1,
            "row": 0,
            "rowspan": 1
          },{(이하중략)},
      ]
  }
```


<표 16> 표 기반 제이슨(json) 산출물 파일 예시

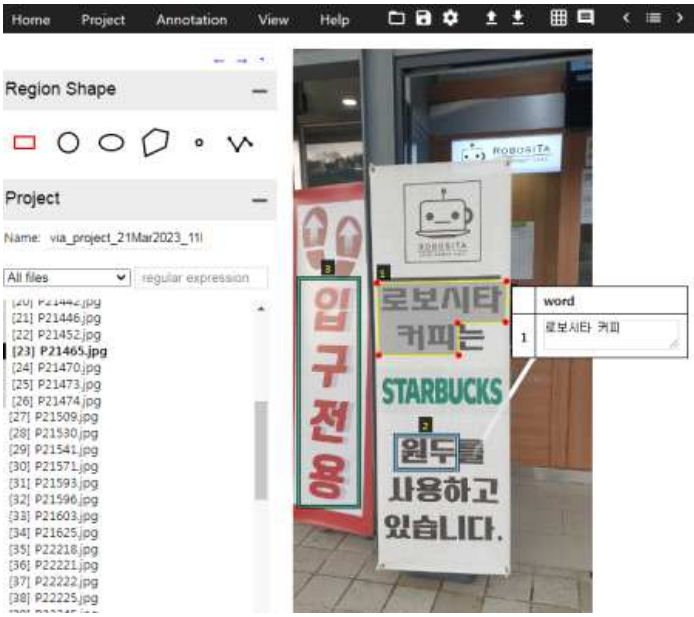
제이슨(json)을 저장한 후에 표에서 추출한 표의 열과 행의 값들이 잘 추출되었는지 확인하기 위해서 제이슨(json)을 html로 변환하여 표로 나타내어 교차 검수하는 과정을 거쳤다. 아래는 해당 제이슨(json)으로 원본의 표를 다시 그린 html을 나타낸다.



<그림 18> 제이슨(json) 표(table) 값을 html로 변환한 결과 예시

4.3.2 그림에서 텍스트와 텍스트의 위치정보(ocr-info) 작성

그림을 설명하는 문장 속에 포함된 단어나 문장이 위치한 위치정보는 공개 소프트웨어(BSD-2 license)인 VIA(VGG Image Annotator: via-2.0.12.버전, 옥스퍼드대학교 VGG그룹)를 사용하여 작성하였다. VIA는 HTML, Javascript 및 CSS를 기반으로 하여 별도의 설치 작업이 필요 없고 사용방법이 비교적 간단하여 작업 도구로 채택하였다.



<그림 19> VIA를 이용한 텍스트 위치 표시 작업 예시

VIA를 이용한 텍스트의 위치 표시는 직사각형, 원형, 다각형, 폴리선 등 다양한 방법이 있지만 본 사업에서는 직사각형과 다각형 두 가지로만 표시하였다. 위치정보를 표시할 텍스트는 유사 문장 생성팀으로부터 전달 받은 그림 설명 기준 문장과 키워드를 참고하여 선정하였으며 그림과 같이 텍스트 위치 표시의 속성값으로 입력하였다.

VIA를 이용한 텍스트 위치 표시 작업과 텍스트 입력을 모두 끝낸 후 제이슨(json) 파일로 저장하면 산출물 구성요소인 'ocr_info'에 필요한 정보가 저장되며 이를 추출하여 산출물을 구성하였다.

VIA에서 저장된 제이슨(json) 파일의 정보

"P21465.jpg3086039":
{"filename":"P21465.jpg","size":3086039,"regions":[
{"shape_attributes":{"name":"polygon","all_points_x":[550,1397,1393,1080,1083,559],"all_points_y":[1535,1517,1788,1794,2006,2006]},
"region_attributes":{"word":"로보시타 커피"}},
{"shape_attributes":{"name":"rect","x":668,"y":2527,"width":410,"height":223},
"region_attributes":{"word":"원두"}},
{"shape_attributes":{"name":"rect","x":37,"y":1501,"width":397,"height":1495},
"region_attributes":{"word":"입구전용"}}
],"file_attributes":{}},

산출물의 "ocr_info" 구성


```
"ocr_info": [          {
    "words": "로보시타 커피",
    "type": "polygon",
    "bbox": {
      "all_points_x": [550,1397,1393,1080,1083,559],
      "all_points_y": [1535,1517,1788,1794,2006,2006]}},
  {
    "words": "원두",
    "type": "rect",
    "bbox": {
      "x": 668,"y": 2527,"width": 410,"height": 223}},
  {
    "words": "입구전용",
    "type": "rect",
    "bbox": {"x": 37,"y": 1501,"width": 397,"height": 1495}}]
```


4.3.3 그림 말뭉치 구축

<그림 19>에서 나와 있는 것과 같이 4.3.2절에서 추출한 오시알(ocr) 정보를 불러온 뒤, 작업자들이 작업한 엑셀 파일과 병합하는 작업을 거치고, 아이디(ID) 중복 및 누락을 검증한 뒤에 최종 제이슨(json) 산출한다. 그림 기반 유사 문장 말뭉치의 제이슨(json) 형식은 다음의 표와 같다.

1수준	2 수준	3 수준	4 수준	5 수준	타입	설명
id					str	
metadata					obj	*파일의 메타정보
	title				str	제목
	creator				str	*생성자: 국립국어원
	distributor				str	*배포자: 국립국어원
	year				str	*작업세트생성년도(2022)
	category				str	자료분류(그림)
	annotation_level				str	유사 문장 생성

document					arr	*문서정보
	id				str	*그림 아이디(ID)
	metadata				obj	*그림의 메타정보
		date			str	*작성일시,게시일시, 크롤링일시
		publisher			str	*크라우드 소싱
	sentence_annotation				obj	*음영표시된 셀을 서술하는 원문과 수정한 문장 유사 문장 서술
		reference_sentence			str	기준 문장
		worker1			str	*유사 문장1
		worker2			str	*유사 문장2
		worker3			str	*유사 문장3
		worker4			str	*유사 문장4
	image_width				int	*이미지 너비
	image_height				int	*이미지 높이
	ocr_info				arr	*오시알(ocr)로 인식된 정보
		words			str	오시알(ocr)로 인식된 문자열
		type			str	인식된 객체의 유형(rect:사각형, polygon:다각형)
		bbox			int	인식된 경계상자
			x		int	경계 상자의 x 좌표
			y		int	경계 상자의 y좌표
			width		int	경계 상자의 너비
			height		int	경계 상자의 높이
			all_point_x		arr(int)	경계 상자 모든 점의 x 좌표
			all_point_y		arr(int)	경계 상자 모든 점의 y좌표

<표 17> 그림 기반 유사 말뭉치의 제이슨(json) 형식

자동 가공 프로그램은 이미지 정보와 작업자들이 작업한 엑셀 파일의 데이터를 매칭하여 최종 제이슨(json) 파일을 생성한다. 이미지 파일명과 제이슨(json) 그림 아이디(ID)가 일치하는 경우

원본의 그림을 특정 경로에 복사하게 된다.

실제로 위의 제이슨(json) 형식이 적용된 그림의 표본을 오시알(ocr) 정보와 제시하면 다음 그림과 같다.

<그림 20> P11603 아이디(ID)의 그림 예시


```
{
"id": "GIPS2202305020",
  "metadata": {
    "title": "국립국어원 그림 유사 문장 말뭉치 GIPS2202305020",
    "creator": "국립국어원",
    "distributer": "국립국어원",
    "year": "2022",
    "category": "그림",
    "annotation_level": "유사 문장 생성"
  },
"document": [
  {
      "id": "P11603",
      "metadata": {
        "publisher": "크라우드 소싱",
        "date": "2022"
      },
      "sentence_annotation": {
```

```
      "reference_sentence": "주차요금 정산기 왼쪽에 위치한 표지판을 보면 주차장 이용 방법을 알 수 있다.",
      "worker1": "우측 주차요금정산기의 옆에 놓인 표지판을 보면 주차장 이용 관련 안내가 적혀 있다.",
      "worker2": "주차장 이용에 관한 내용은 주차요금 정산기 왼쪽에 위치하고 있다.",
      "worker3": "주차장 이용 방법이 적힌 표지판 오른쪽에 주차요금 정산기가 있다.",
      "worker4": "주차요금 정산기 왼쪽에 주차장 이용을 설명하는 배너가 설치되어 있다."
    },
    "image_width": 4032,
    "image_height": 3024,
    "ocr_info": [
      {
        "words": "주차장 이용",
        "type": "rect",
        "bbox": {
          "x": 635,
          "y": 1076,
          "width": 737,
          "height": 177
        }
      },
      {
        "words": "주차요금 정산기",
        "type": "polygon",
        "bbox": {
          "all_points_x": [
            2779,
            3713,
            3705,
            2775
          ],
          "all_points_y": [
            252,
            169,
            288,
            370
          ]
        }
      }
    ]
  },{이하 생략}
}
```


<표 18> 그림 기반 제이슨(json) 산출물 파일 예시

<그림 20>은 4032x3024 픽셀로 구성되어 있다. 사각형(rect) 유형은 네 개의 모서리로 이루어진 도형이며, 일반적으로 '왼쪽 위'의 시작점('x', 'y' 좌표)과 '너비(width)' 및 '높이(height)'로 정의된다. x축은 왼쪽에서부터의 픽셀 거리를 표시하고, y축은 위에서부터의 픽셀 거리를 표시한다.

다각형(polygon)은 세 개 이상의 점으로 이루어진 도형이며, 각 점은 'x'와 'y'의 좌표쌍으로 표시된다. bbox(Bounding Box)의 키 'all_points_x': [2779, 3713, 3705, 2775], 'all_points_y': [252,169,288,370] 값은 시계 방향으로 제공된 좌표의 순서이다.

4.3.4 그래프 말뭉치 구축

그래프 기반 말뭉치는 한글 파일에서 필요한 정보를 불러와서 <그림 17>과 같은 엑셀 파일로 저장한다. 작업자들의 유사 문장 생성 작업이 끝난 후 이 엑셀 파일에서 필요한 정보를 불러와서 자동 가공 프로그램을 통해 중복 검사 및 누락을 검증한 뒤 그래프 생성과 제이슨(json)을 생성한다. 아래 표는 그래프 기반 제이슨(json) 구조를 나타낸다.

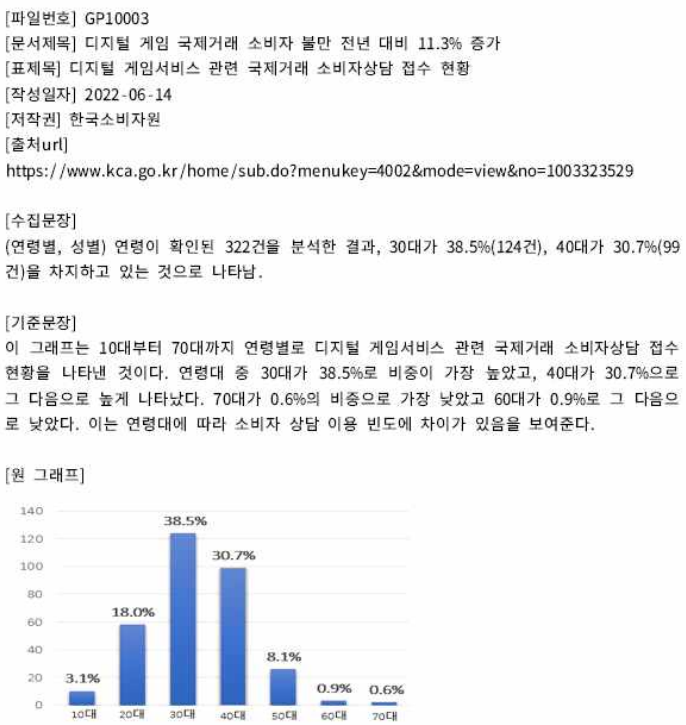
1수준	2 수준	3 수준	4 수준	5 수준	타입	설명
id					str	
metadata					obj	*파일의 메타정보
	title				str	제목
	creator				str	*생성자: 국립국어원
	distributor				str	*배포자: 국립국어원
	year				str	*작업세트생성년도(2022)
	category				str	자료분류(그래프)
	annotation_level				str	유사 문장 생성
document					arr	*문서정보
	id				str	*그림 아이디(ID)
	metadata				obj	*그래프의 메타정보
		title			str	문서 제목
		figure_title			str	표 제목
		date			str	*작성일시,게시일시,크롤링일시
		publisher			str	*공공기관
		url			str	*URL 주소(표 출처)
	str_annotation				obj	*음영표시된 셀을 서술하는 원문과 수정한 문장 유사 문장 서술
		reference_str			str	기준 문장
		paraphras			str	*유사 문장1

		e_str				
	figure				obj	그래프에 대한 정보
		type			str	그래프 유형
		x_label			str	x축 제목
		y_label			str	y축 제목
		x			str	x값
		y			float (int)	y값

<표 19> 그래프 기반 유사 말뭉치의 제이슨(json) 형식

자동 가공 프로그램은 엑셀 파일의 목록을 반복하면서 각 파일의 시트를 순회한다. 각 시트에서 필요한 데이터를 추출하여 제이슨(json) 파일로 저장된다. 모든 엑셀 파일과 시트의 처리가 완료되면 모든 제이슨(json) 파일을 읽어서 하나의 큰 제이슨(json) 파일로 병합하고, 한글 파일과 엑셀 시트 간의 불일치 항목을 확인하고, 결과를 저장한다.

그래프 기반 말뭉치에서 'figure' 오브젝트는 그래프에 대한 정보를 나타내며, 'type' 키는 가로 막대그래프, 세로 막대그래프, 선그래프의 유형으로 정의된다. 'x_label'과 'y_label'은 각각 x축과 y축의 제목을 나타내며, 'x, y'는 각 축의 자료값을 나타낸다.


[파일번호] GP10003
[문서제목] 디지털 게임 국제거래 소비자 불만 전년 대비 11.3% 증가
[표제목] 디지털 게임서비스 관련 국제거래 소비자상담 접수 현황
[작성일자] 2022-06-14
[저작권] 한국소비자원
[출처url]
https://www.kca.go.kr/home/sub.do?menukey=4002&mode=view&no=1003323529

[수집문장]
(연령별, 성별) 연령이 확인된 322건을 분석한 결과, 30대가 38.5%(124건), 40대가 30.7%(99건)을 차지하고 있는 것으로 나타남.

[기준문장]
이 그래프는 10대부터 70대까지 연령별로 디지털 게임서비스 관련 국제거래 소비자상담 접수 현황을 나타낸 것이다. 연령대 중 30대가 38.5%로 비중이 가장 높았고, 40대가 30.7%으로 그 다음으로 높게 나타났다. 70대가 0.6%의 비중으로 가장 낮았고 60대가 0.9%로 그 다음으로 낮았다. 이는 연령대에 따라 소비자 상담 이용 빈도에 차이가 있음을 보여준다.

[원 그래프]

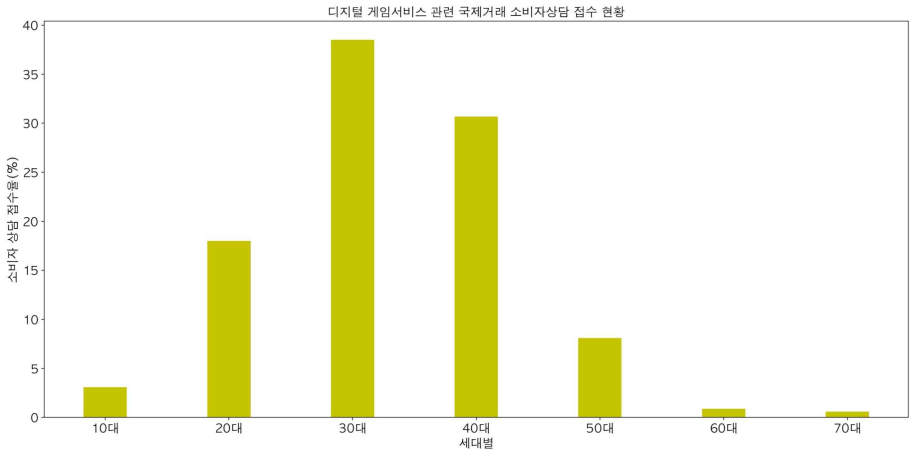
<그림 21> 그래프 한글 파일 예시

<그림 21>의 한글 파일을 엑셀 파일과 합치는 작업을 거친 뒤 <표 19>의 형식에 따라 제이슨(json)을 생성하고 그래프를 생성한 결과는 다음 <표 20>와 <그림 22>과 같다.


```
{
  "id": "GGPS2202305020",
  "metadata": {
    "title": "국립국어원 그래프 유사 문장 말뭉치 GGPS2202305020",
    "creator": "국립국어원",
    "distributer": "국립국어원",
    "year": "2022",
    "category": "그래프",
    "annotation_level": "유사 문장 생성"
  },
  "document": [
{
      "id": "GP10003",
      "metadata": {
        "title": "디지털 게임 국제거래 소비자 불만 전년 대비 11.3% 증가",
        "figure_title": "디지털 게임서비스 관련 국제거래 소비자상담 접수 현황",
        "date": "2022-06-14",
        "publisher": "한국소비자원",
        "url": "https://www.kca.go.kr/home/sub.do?menukey=4002&mode=view&no=1003323529"
      },
      "str_annotation": {
        "reference_str": "이 그래프는 10대부터 70대까지 연령별로 디지털 게임서비스 관련 국제거래 소비자상담 접수 현황을 나타낸 것이다. 연령대 중 30대가 38.5%로 비중이 가장 높았고, 40대가 30.7%으로 그 다음으로 높게 나타났다. 70대가 0.6%의 비중으로 가장 낮았고 60대가 0.9%로 그 다음으로 낮았다. 이는 연령대에 따라 소비자 상담 이용 빈도에 차이가 있음을 보여준다.",
        "paraphrase_str": "이 그래프는 디지털 게임서비스 관련 국제거래 소비자상담 접수 현황을 보여준다. 10대에서 70대까지 세대별로 분석한 그래프에 따르면 30대가 가장 높은 비율인 38.5%를 차지하고 있으며 40대가 30.7%로 그 뒤를 이었다. 반면 70대는 0.6%로 가장 낮은 비율을 차지하고 있으며 60대가 0.9%로 그 다음으로 낮은 모습을 보여주고 있다. 이를 통해 연령대 별로 소비자상담 이용 빈도에 차이가 있음을 알 수 있다."
      },
      "figure": {
        "type": "세로 막대 그래프",
        "x_label": "세대별",
        "y_label": "소비자 상담 접수율(%)",
        "x": [
          "10대",
          "20대",
          "30대",
          "40대",
          "50대",
          "60대",
          "70대"
        ],
```

```
            "y": [
              3.1,
              18,
              38.5,
              30.7,
              8.1,
              0.9,
              0.6
            ]
          }
        },{이하 생략}
}
```


<표 20> 그래프 기반 제이슨(json) 산출물 파일 예시



<그림 22> GP10003 아이디(ID)의 그래프 생성 예시

부 록

유사 문장 생성 지침

1. 표 유사 문장 생성 지침

○ 기준 문장 작업 지침

1) 수집 문장의 내용과 표현을 최대한 그대로 활용하도록 하되, 비맥락화하여 기준 문장 그 자체로 완전한 정보를 전달할 수 있도록 한다.
2) 기준 문장에 이용할 수 있는 기본 정보에는 기사 제목, 표 제목, 보도 일자, 저작권, 표의 첫 행(heading)이 포함된다. 그 이외의 정보는 삭제하도록 한다.
3) 명백한 오타, 오기로 보이는 표현의 경우 표에 있는 표기를 규범 표기로 변경하여 기준 문장을 작성하도록 한다.
4) 수집 문장을 기준 문장 생성 지침에 따라 자연스러운 국어 문장으로 변경하도록 한다.
5) 기준 문장 작성 시 문법적 오류가 없도록 하고, 맞춤법에 맞게 작성하도록 하도록 한다.
6) 기준 문장 작성 시 띄어쓰기의 통일성을 유지한다.
7) 음영 표시의 문제가 있을 경우에는 아래의 지침에 따른다.

※ 음영 표시에 문제가 있을 경우는 다음과 같다.

1) 음영이 추가되거나 삭제되어야 하는 경우
2) 음영 처리된 부분이 반영하고 있는 내용이 비대칭적으로 이루어져 하나의 문장을 만들기 어려운 경우
3) 음영 처리된 부분의 용어와 수집 문장의 용어가 일치하지 않을 경우

1)과 2)는 수집 문장에 맞추어 음영을 다른 색으로 수정한 후, 유사 문장 작업 준비팀에 인계한다.

3)의 경우, 용어의 차이를 기계가 인식하기 어렵다고 판단될 경우, 해당 표는 작업 대상에서 제외한다.

○ 기준 문장 작성 예시

[수집 문장]

ㅇ (소비자피해 대응 행동) ‘업체에 문의·항의·보상 요청’한 응답자가 56.2%(86명)로 가장 많았고, ‘이의제기하지 않았다’가 38.6%(59명), ‘관련 기관에 문의·피해구제 요청’ 3.9%(6명) 순으로 나타남.

[표의 내용이 아닌 경우 삭제]

ㅇ (소비자피해 대응 행동) '해당 안마의자 렌탈업체에 문의·항의·보상 요청'한 응답자가 56.2%(86명)로 가장 많았고, '이의제기하지 않았다'가 38.6%(59명), '~~관련 기관~~ 소비자보호기관에 문의·피해구제 요청' 3.9%(6명) 순으로 나타남.

[표]

(단위: 명, %)

구분	전체	바디프랜드	휴테크산업	LG전자	SK매직
해당 안마의자 렌탈업체에 문의·항의 혹은 보상요청	86(56.2)	15(45.5)	31(66.0)	11(37.9)	29(65.9)
이의제기하지 않음	59(38.6)	15(45.5)	15(31.9)	15(51.7)	14(31.8)
1372 소비자상담센터 등 소비자보호기관에 문의 혹은 피해구제 요청	6(3.9)	1(3.0)	1(2.1)	3(10.3)	1(2.3)
기타	2(1.3)	2(6.1)	-	-	-
총합	153(100.0)	33(100.0)	47(100.0)	29(100.0)	44(100.0)

[기준 문장]

피해를 얻은 소비자 중 '이의제기하지 않음'으로 응답한 소비자는 전체의 38.6%인 59건인 것으로 나타났다.

○ 유사 문장 작업 지침

1) 기본적으로 의미가 유사한 문장을 생성한다. 이때 의미는 진리조건적, 명제적 의미의 동일성을 말한다.
2) 단, 표면적 유사성이 지나치게 높아지는 것을 막기 위해, 어순의 조정, 단어의 교체, 문장 구조의 변경 중 최소한 하나의 절차를 수행하도록 한다. 어미와 조사만 교체하는 최소한의 수정은 지양하도록 한다.
3) 단순한 정보로 인해 유사 문장을 만들기 어려울 경우, 사칙연산 등 수치의 단순한 비교와 비

교 술어의 사용은 가능하다.

4) 유사 문장 작성 시 기준 문장에 오류가 있는 경우, 관리자에게 보고한 후 확인될 때까지 작업을 보류한다. 오류의 종류는 다음과 같다.

- 제시된 표의 정보와 기준 문장의 내용이 일치하지 않는 경우
- 음영 표시된 부분의 내용과 기준 문장의 표현이 일치하지 않는 경우
- 유사 문장을 작성하기 어려울 정도로 기준 문장의 내용이 충분하지 않은 경우
- 맞춤법 등 표기의 통일이 되어 있지 않은 경우

4) 유사 문장 작성 시 문법적 오류가 없도록 하고, 맞춤법에 맞게 작성하도록 하도록 한다.

5) 유사 문장 작성 시 띄어쓰기의 통일성을 유지한다.

○ 유사 문장 작성 예시

[수집 문장]

○ 국가유공자 본인 지원규모의 적정성(1년에 100만 원씩 총300만 원 한도 내)에 관한 평균은 4.00으로 나타나 확대되어야 한다는 의견이 다수인 것으로 조사됨

[표의 내용이 아닌 경우 삭제]

○ 국가유공자 본인 지원규모의 적정성(1년에 100만 원씩 총300만 원 한도 내)에 관한 평균은 4.00으로 ~~나타나 확대되어야 한다는 의견이 다수인 것으로~~ 조사됨

[기준 문장]

국가유공자 본인 지원 규모의 적정성(1년에 100만 원씩 총 300만 원 한도 내)은 평균 4.00의 수치를, 유가족 지원 규모의 적정성(1년에 50만 원씩 총 150만 원 한도 내)은 4.01을 차지했다.

[표]

구분	Mean	SD
국가유공자 등에게 취업능력개발을 지원하는 제도	3.90	.965
– 국가유공자 본인 지원 규모의 적정성	4.00	.807

(1년에 100만 원씩 총 300만 원 한도 내)		
- 국가유공자 유가족 지원 규모의 적정성 (1년에 50만 원씩 총 150만 원 한도 내)	4.01	.883

[유사 문장]

- 국가유공자 본인 지원 규모의 적정성(1년에 100만 원씩 총 300만 원 한도 내)은 평균 4.00의 수치를, 유가족 지원 규모의 적정성(1년에 50만 원씩 총 150만 원 한도 내)은 4.01을 차지했다.

- 국가유공자 본인 지원 규모의 적정성(1년에 100만 원씩 총 300만 원 한도 내)의 평균은 4.00, 유가족 지원 규모의 적정성(1년에 50만 원씩 총 150만 원 한도 내)의 평균은 4.01로 비슷한 수치를 보였다.

- 국가유공자 본인 지원 규모의 적정성(1년에 100만 원씩 총 300만 원 한도 내)와 국가유공자 유가족 지원 규모의 적정성(1년에 50만 원씩 총 150만 원 한도 내)의 평균은 각각 4.00, 4.01이다.

- 국가유공자 본인 지원 규모의 적정성을 조사했을 때, "1년에 100만 원씩 총 300만 원 한도 내" 항목은 평균치 4.00를, 유가족 지원 규모의 적정성을 조사했을 때, "1년에 50만 원씩 총 150만 원 한도 내" 항목은 4.01을 차지했다.

- 1년에 100만 원씩 총 300만 원 한도 내에서 국가유공자 본인을 지원하는 규모의 적정성은 4.00의 평균값을 가지며, 1년에 50만 원씩 총 150만 원 한도 내에서 유가족을 지원하는 규모의 적정성은 4.01의 평균값을 가진다.

2. (텍스트가 포함된) 그림 기반 유사 문장 생성 지침

○ 기준 문장 작업 지침

1) 텍스트가 있는 그림 자료의 경우에 기준 문장에는 그림 내에 있는 구체적 텍스트가 반드시 포함되어야 한다. 기준 문장에 포함한 그림 내의 텍스트('키워드'라 칭함)는 따로 분리하여 표시하도록 한다.

2) 기준 문장에 포함한 그림 내의 텍스트('키워드')는 결과물의 제이슨(json) 파일 내에 'ocr_info'의 형태로 저장된다.

3) 기준 문장에 사용된 텍스트가 동일 그림 내에서 두 개 이상 출현하는 경우, 오시알(ocr) 정보는 하나만 제시하도록 한다.

4) 오시알(ocr) 정보는 어절 단위로 잡고, 어절 바로 뒤에 종결어미가 나타나는 경우 종결어미까지 잡는다.

   예) '위험할 수가 있습니다.'에서 '위험할 수가'만 제공하지 않아야 함.

5) 민감 정보에 해당하는 개인 휴대전화 번호와 상점 등의 전화번호가 노출된 사진들은 사진에서 번호를 지우는 등의 후처리를 하거나 구축 목록에서 아예 제외한다.

6) 텍스트가 쓰여 있는 장소 혹은 물체에 대한 정보를 포함하여야 한다.

7) 그림 내에 있는 한글만 포함하도록 하되, 그램(g)이나 미터(m) 등의 도량형 단위나 이미 일반화되어 있는 축약형(B1, B2, AM, PM)은 포함할 수 있다.

8) 기준 문장 작성 시, 문법적 오류가 없도록 한다.

9) 기준 문장을 맞춤법에 맞게 작성하도록 하고, 띄어쓰기의 통일성을 유지한다.

10) 명백한 오타, 오기로 보이는 표현의 경우 사진에 있는 표기를 규범 표기로 변경하여 기준 문장을 작성하도록 한다.

○ 기준 문장 작성 예시

[기준 문장]

대한예수교 장로회 소속인 서촌교회 앞에는 빨간색 대형버스가 있다.

[키워드] 대한예수교 장로회, 서촌교회

○ 유사 문장 작업 지침

1) 기본적으로 의미가 유사한 문장을 생성한다. 이때 의미는 진리조건적, 명제적 의미의 동일성을 말한다.
2) 기준 문장에서 사용한 그림 내 텍스트(키워드)는 반드시 포함하되, 그 이외의 텍스트는 포함하지 않도록 한다.
3) 단, 표면적 유사성이 지나치게 높아지는 것을 막기 위해, 어순의 조정, 단어의 교체, 문장 구조의 변경 중 최소한 하나의 절차를 수행하도록 한다. 어미와 조사만 교체하는 최소한의 수정은 지양하도록 한다.
3) 단순한 정보로 인해 유사 문장을 만들기 어려울 경우, 사칙연산 등 수치의 단순한 비교와 비교 술어의 사용은 가능하다.

○ 유사 문장 작성 예시

[기준 문장]

- 대한예수교 장로회 소속인 서촌교회 앞에는 빨간색 대형버스가 있다.

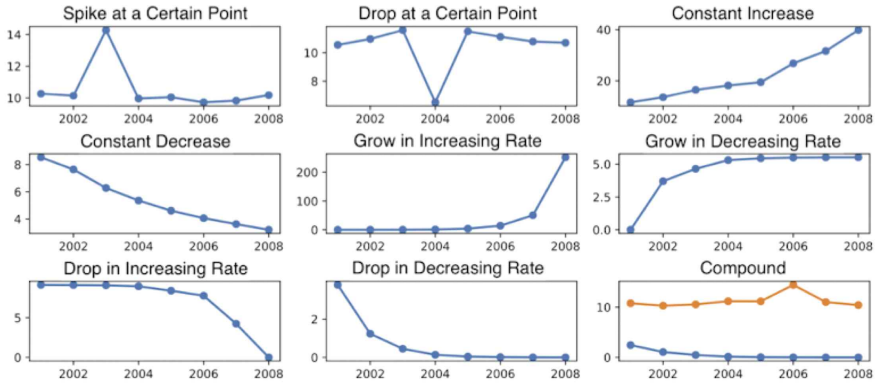
[유사 문장]

- 서촌교회는 대한예수교 장로회 소속이고, 그 앞에는 빨간색 대형버스가 있다.
- 서촌교회의 소속은 대한예수교 장로회이고, 교회 앞으로 서 있는 대형버스는 빨간색이다.
- 빨간색 대형버스가 있는 곳은 대한예수교 장로회 소속인 서촌교회의 앞이다.
- 서촌교회 앞에 있는 것은 빨간색 대형버스이며 교회는 대한예수교 장로회 소속이다.

3. 그래프 기반 유사 문장 생성 지침

○ 기준 문장 작업 지침

1) 수집 문장(문단)을 기준 문단 작성 지침에 따라 자연스러운 국어 문장으로 기준 문장(문단) 작성
   - 수집 문장(문단)에서 그래프와 관련된 내용만 남기고 불필요한 문구 삭제
   - 문법적 오류가 없는 기준 문단 작성
   - 띄어쓰기 통일성 유지
   - 맞춤법에 맞게 작성
   - 기준문단과 생성문단 간 수치 불일치 여부 확인
2) 수집 문장(문단)이 아래의 '추세곡선에 따른 기준 문장 생성 지침'에 따르지 않을 경우 문장을 그 지침에 따라 수정하여야 한다. 대체적으로 수집된 문장(문단)이 그래프 자료에 대한 충분한 정보를 제공하지 못하는 경우가 많으므로 작성 지침에 따라 수정하도록 한다.
   - 기준 문단 작성 시 그래프의 3개 포인트로 설명하도록 한다. 수집 문단에 2개 포인트만 표현되었더라도 그래프를 바탕으로 내용을 보충하도록 한다.
   - 수집 문단에서 제시한 증감폭, 추세 등의 정보를 반영하여 기준 문단 작성한다.
3) 분석 내용을 구체적으로 제시하도록 한다.
   - 그래프에서 주어진 수치 구체적으로 제시하도록 한다.
   - 최고, 최저 수치를 함께 설명할 것
   - 특정 값/영역에 편향되지 않고 그래프 내용을 전반적으로 아우를 수 있도록 서술할 것
4) 추세곡선에 따른 기준 문단을 생성하도록 한다. 추세곡선은 다음과 같이 8가지이다. 마지막 그래프는 앞의 8가지 추세곡선의 복합형이다.



5) 기준 문단 생성 지침은 아래와 같다.

- 지침 1(필수적): 그래프를 개관하여야 한다. 어떤 그래프인지, 그래프의 내용이 무엇인지 표현되어야 한다.

  예) "이 그래프는 1970년과 1990년 사이 한국에서의 패스트푸드 소비량을 보여준다."

- 지침 2(수의적): 그래프 자체의 묘사(description), 그래프의 형상이나 요소에 초점을 둔다.

  예) "표본추출한 국가는 핀란드, 프랑스, 조지아, 독일, 그리스, 헝가리 등 모두 유럽 국가이다."

- 지침 3(필수적): 그래프 정보의 해석, 그래프의 추세와 그래프 정보의 단순한 관찰을 포함하되 그래프의 숫자 정보를 보고하도록 한다. 그러나, 단순한 흐름 기술은 안 된다.

  적절한 예) "1970년에 소비량은 주당 300그램 정도였다가 1990년에 주당 220그램으로 하락했다."

  부적절한 예) "쌀 소비량은 약간 감소하였다."

- 지침 4(수의적): 특정 값이나 비교에 대한 평가적 표현을 포함할 수 있다.
- 지침 5(필수적): 그래프에 기초한 결론, 요약, 예측, 혹은 함의를 포함하여야 한다.

☞ 결론: 그래프에 주어졌거나 기술한 사실을 전제로 하여 최종적으로 이끌어낸 판단

☞ 요약: 그래프에 대한 앞선 기술을 짧게 정리한 것

☞ 예측: 그래프에 주어졌거나 기술한 사실을 기초로 하여 미래에 대해 예측하는 것

☞ 함의: 그래프에서 주어진 사실이나 그에 대한 기술에 숨겨져 있는 의미를 표현한 것

  예) "피자 소비량은 크게 증가하였지만, 피시앤칩스 소비량은 감소하였다."

- X축이 시간을 나타내는 시계열 그래프에서는 문장들이 자료의 흐름과 서로 다른 시점간의

비교에 초점을 두어야 한다. 흐름을 기술하는 지침 3과 4의 문장들은 추세곡선 유형 8개로 나눌 수 있다. 가시적인 흐름을 보여주지 않는 시계열 그래프의 경우, 자료 간 비교 혹은 특정 시점들에 초점을 두어야 한다.

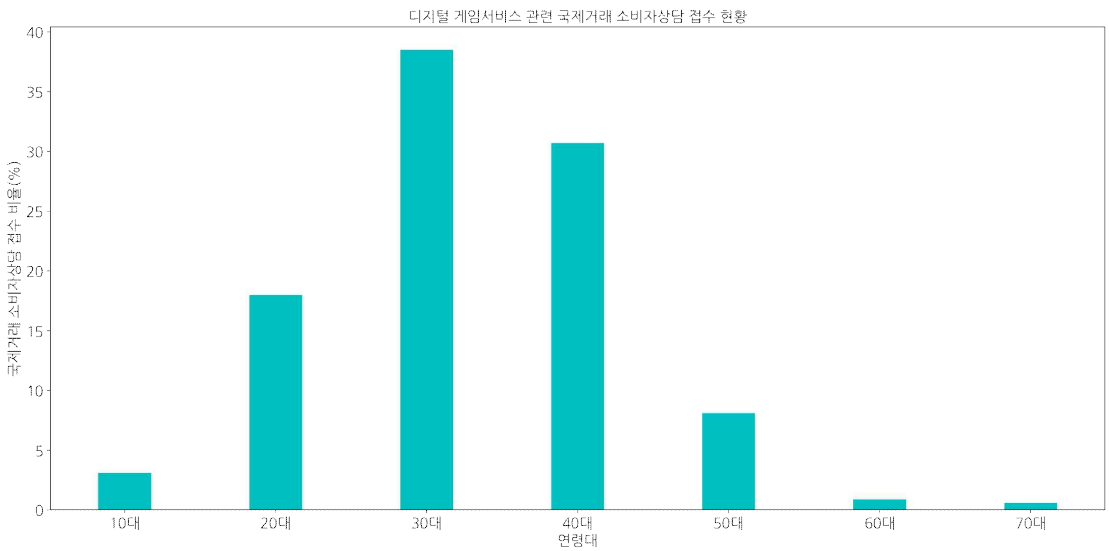
- X축이 도시, 음식, 등의 항목을 나타내는 범주형 그래프의 경우, 지침 3과 4의 문장들은 서로 다른 범주를 비교하거나 특정 항목들에 초점을 두도록 한다.

6) 추세 분석에 대한 기술은 다음과 같이 한다.

- 자료 적절성 검토: 그래프에서 각 요소들 간 차이가 분명하여 추세가 드러나는지 우선 검토하도록 한다. 변화 추이가 나타나는 그래프임에도 추세 분석하지 않았는지 여부 확인하여 이를 반영하도록 한다.
- '꾸준한 성장세', '지속적 증가', '전반적으로 많은'과 같이 모호한 표현을 피하되, 필요한 경우 구체적인 수치를 함께 제시한다.
- 지침 3에 따라 그래프에 나타난 수치 자료를 사용하여 추세를 구체적으로 설명하도록 한다.

  예) 2017년 15,684건에서 2018년 22,169건, 2019년 24,194건, 2020년 26,954건으로 꾸준히 증가하였으나, 2021년에는 전년 대비 47.7%가 감소한 14,086건으로 나타났다. 2017년 이후 점진적 증가세를 보이던 것과 달리 2021년에는 소비자상담 접수가 급감한 것이다.

❍ 기준 문단의 작성 예시



[그래프 정보]

· 문서 제목: 디지털 게임 국제거래 소비자 불만 전년 대비 11.3% 증가

· 문서 저자: 한국소비자원

· 보도 일자: 2022-06-14

· 표 제목: 디지털 게임서비스 관련 국제거래 소비자상담 접수 현황

[그래프와 함께 수집한 문단]

□ (연령별, 성별) 연령이 확인된 322건을 분석한 결과, 30대가 38.5%(124건), 40대가 30.7%(99건)을 차지하고 있는 것으로 나타남.

[기준 문단]

이 그래프는 10대에서 70대까지 세대별로 디지털게임 서비스 관련 국제 거래 소비자 상담 접수 현황을 분석한 것이다. [← 지침 1, 2에 따라 수정] 그래프에 따르면, 30대가 38.5%, 40대가 30.7%로 가장 많이 차지하는 것으로 조사되었다. [← 지침 3, 5에 따라 수정]

○ 유사 문장 작업 지침

1) 기본적으로 기준 문단과 의미가 유사한 문단(생성 문단)을 생성한다.
2) 단, 표면적 유사성이 지나치게 높아지는 것을 막기 위해, 어순의 조정, 단어의 교체, 문장 구조의 변경 중 최소한 하나의 절차를 수행하도록 한다. 어미와 조사만 교체하는 최소한의 수정은 지양하도록 한다.
3) 단순한 정보로 인해 생성 문단을 만들기 어려울 경우, 사칙연산 등 수치의 단순한 비교와 비교 술어의 사용은 가능하다.
4) 유사 문단 생성에 대한 원칙은 기준 문단 생성의 원칙과 동일하다.

○ 유사 문장의 작성 예시

[기준 문단]

이 그래프는 10대에서 70대까지 세대별로 디지털게임 서비스 관련 국제 거래 소비자 상담 접수 현황을 분석한 것이다. 그래프에 따르면, 30대가 38.5%, 40대가 30.7%로 가장 많이 차지하는 것으로 조사되었다.

[유사 문단]

이 그래프는 디지털게임 서비스 관련 국제 거래 소비자 상담 접수 현황을 살펴본 것이다. 10대에서 70대까지 세대별로 분석한 이 그래프에 따르면, 30대와 40대가 각각 38.5% 그리고 30.7%로 가장 높게 나타났다.

참고문헌

오교중·김현민·고보원·남제현·최호진(2019), 문장 유사성 분석을 위한 한국어 패러프레이즈 말뭉치 및 구축 가이드라인, 제31회 한글 및 한국어 정보처리 학술대회 논문집.

이숙의(2021), 유사 문장 말뭉치 분석을 통한 유사도 인식에 관한 연구, 어문연구 108, 63-89.

Barrón-Cedeño, A., M. Vila, M. A. Martí, and P. Rosso, “Plagiarism meets paraphrasing: Insights for the next generation in automatic plagiarism detection,” Comput. Linguistics, vol. 39, no. 4, pp. 917-947, Dec. 2013.

Dolan, W. and C. Brockett, “Automatically constructing a corpus of sentential paraphrases,” in Proc. 3rd Int. Workshop Paraphrasing (IWP), Jan. 2005, pp. 9-16.

Martin, B. 2004. Plagiarism: Policy against cheating or policy for learning?, Nexus (Newsletter of the Australian Sociological Association), 16(2):15-16.

Vila, M., M. A. Martí, and H. Rodríguez, “Is this a paraphrase? What kind? Paraphrase boundaries and typology,” Open J. Mod. Linguistics, vol. 4, no. 1, pp. 205-218, Mar. 2014.

Wieting, J. and K. Gimpel, “ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations,” in Proc. 56th Annu. Meeting Assoc. Comput. Linguistics, vol. 1, Jul. 2018, pp. 451-462.

Wieting, J., J. Mallinson, and K. Gimpel, “Learning paraphrastic sentence embeddings from back-translated bitext,” Jun. 2017, arXiv:1706.01847. [Online]. Available: https://arxiv.org/abs/1706.01847.

Zhang, Y., Baldridge, J., He, L. PAWS: Paraphrase Adversaries fromWord Scrambling, arXiv:1904.01130v1 [cs.CL] 1 Apr 2019.

Zhu, J., Ran, J., Lee, R. K., Choo, K., & Li, Z. (2021). AutoChart: A Dataset for Chart-to-Text Generation Task. ArXiv. /abs/2108.06897.

〈기획・연구〉
국립국어원 강미영 언어정보과장
국립국어원 유희정 학예연구사
국립국어원 이민주 연구원
국립국어원 박미은 연구원
국립국어원 정영은 연구원

〈연구 참여자〉
연구 책임자 안희돈(건국대학교)
공동 연구원 조용준(건국대학교)
위혜경(단국대학교)
박동근(대진대학교)
윤수원(서울시립대학교)
김성환(건국대학교)
박분선(나라지식정보)
연구 보조원 하지희(세종대왕기념사업회)
이주연(건국대학교)
원유권(건국대학교)
홍정연(건국대학교)
이상순(나라지식정보)
이정훈(나라지식정보)

발행인: 국립국어원장
발행처: 국립국어원
서울시 강서구 금낭화로 154
전화 02-2669-9775, 전송 02-2669-9727
인쇄일: 2023년 3월 4일
발행일: 2023년 3월 4일
인 쇄: 이호문화사

※ 이 보고서는 국립국어원의 용역비로 수행한 '2022년 유사 문장 생성 말뭉치 연구 분석' 사업의 결과물을 발간한 것입니다.