HWP 문서의 구조적 특징 분석을 통한 스타일 자동 인식 및 마크다운 변환

정현수*, 권유진*, 김민혁*, 오민준*, 최수지*, 김현석*
*동아대학교 컴퓨터공학과
e-mail : 1801145@donga.ac.kr

Automatic Style Recognition and Markdown Conversion through Structural Feature Analysis of HWP Documents

Hyeon-Su Jeong*, Yu-Jin Kwon*, Min-Hyeok Kim*, Min-Jun Oh*,
Su-Ji Choi*, Hyeon-Seok Kim*
*Department of Computer Engineering, Dong-A University

요 약

최근 생성형 AI 를 활용한 문서 초안 작성이 활발하지만, 최종 결과물은 특정 서식이 적용된 HWP 파일로 제출해야 하는 경우가 많아 수동 변환의 비효율이 발생한다. 본 논문에서는 HWP 의 개방형 포맷인 HWPML 의 구조적 특징을 분석하여 문서의 스타일(제목, 목록 등)을 자동으로 추론하는 휴리스틱 기반 시스템을 제안한다. 제안하는 시스템은 HWP 템플릿의 header.xml 과 section0.xml 을 파싱하여 글자 크기, 글머리 기호 등의 특징을 추출하고, 이를 기반으로 제목과 목록 스타일 후보를 선정한다. 구현된 시스템은 마크다운으로 작성된 문서를 사용자가 등록한 HWP 템플릿의 서식에 맞춰 자동으로 변환함으로써, 문서 작업의 생산성을 향상시키고 서식의 일관성을 확보할 수 있음을 보였다

1. 서론

ChatGPT 와 같은 생성형 AI 의 발전으로 마크다운(Markdown) 형식의 문서 초안 생성이 보편화되고 있다. 마크다운은 간결하고 직관적인 문법으로 빠르게 콘텐츠를 작성할 수 있는 장점이 있지만, 다수의 공공기관, 기업, 학계에서는 여전히 특정 서식이 적용된 한글(HWP) 파일을 공식 문서로 요구한다. 이로 인해 사용자는 AI 가 생성한 마크다운 문서를 HWP 에 옮긴 후, 제목, 목록, 본문 스타일을 일일이 수동으로 지정하는 반복적이고 비효율적인 작업을 수행해야 한다.

기존 문서 변환 도구들은 단순 텍스트 변환에 그쳐, 사용자가 원하는 HWP 템플릿의 고유한 스타일, 즉 글자 크기, 글꼴, 문단 모양 등을 정확히 인식하고 적용하는 데 명확한 한계를 가진다. 이러한 문제를 해결하기 위해, 본 연구에서는 HWP 문서의 내부 구조를 직접 분석하여 문서 내 스타일의 용도를 자동으로 추론하고, 이를 마크다운-HWP 변환에 활용하는 지능형 시스템을 제안하고 구현한다.

2. 제안하는 시스템

본 연구에서 제안하는 시스템은 사용자가 등록한 HWP 템플릿의 스타일 구조를 자동으로 분석하고, 이를 기반으로 마크다운 문서를 변환하는 것을 목표로 한다. 전체 시스템의 흐름은 [그림 1]과 같으며, 핵심인 스타일 분석은 특징 추출과 후보 선정의 2 단계로 구성된다.

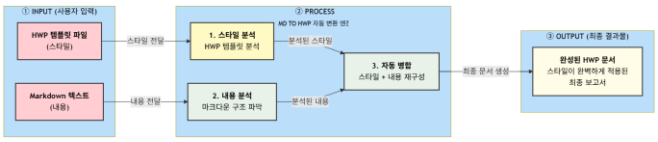

그림 1. 제안하는 시스템의 전체 구성도

2.1. 특징 추출 (Feature Extraction)

첫 단계로, HWP 의 개방형 포맷인 HWPX 파일 내 header.xml 과 section0.xml 을 파싱하여 각 문단의 구조적 특징을 추출한다.

글자 크기 정보 수집: header.xml 에는 문서에 사용된 모든 글자 스타일(charPr)의 속성이 정의되어 있다. 이 파일을 분석하여 각 스타일 ID 와 이에 해당하는 실제 글자 크기(height) 값을 추출하여 딕셔너리 형태로 저장한다.

문단 특징 분석: section0.xml 의 모든 문단(hp:p)을 순회하며 각 문단을 대표하는 글자 스타일 ID 와 글머리 기호(bullet)의 존재 여부 및 종류 등의 정보를 추출한다.

2.2. 후보 선정 (Candidate Selection)
두 번째 단계로, 추출된 특징들을 바탕으로 '제목'과 '목록' 스타일 후보를 추론하는 휴리스틱 알고리즘을 적용한다.
제목 후보 선정: 글머리 기호가 없는 문단 중에서, 1 단계에서 수집한 글자 크기 정보를 바탕으로 텍스트 크기가 큰 순서대로 상위 N 개의 스타일을 '제목' 후보로 선정한다. 이때, 동일한 글자 스타일 ID 를 가진 문단은 한 번만 후보로 고려하여 중복을 방지한다.
목록 후보 선정: 글머리 기호가 있는 문단 중에서, 서로 다른 종류의 기호(-, · 등)를 사용하는 스타일들을 중복 없이 추출한다. 이후 이들을 글자 크기가 큰 순서대로 정렬하여 상위 M 개의 스타일을 '목록' 후보로 최종 선정한다.
이러한 과정을 통해 별도의 사용자 설정 없이 HWP 템플릿의 주요 스타일을 자동으로 분류하고, 이를 변환에 활용할 기반을 마련한다.

3. 구현 및 결과

3.1. 구현 환경
제안하는 시스템은 사용자가 쉽게 접근할 수 있도록 데스크톱 애플리케이션으로 구현하였다. 개발 환경은 다음과 같다.
· 언어: Python 3.10
· GUI 라이브러리: PySide6
· HWP 구조 분석: xml.etree.ElementTree
· HWP 자동화: win32com

3.2. 실행 결과
시스템의 유효성을 검증하기 위해, 간단한 마크다운 문서를 사전 등록된 HWP 템플릿으로 변환하는 실험을 진행하였다. [그림 2]는 실제 변환 전후를 비교한 것이다.

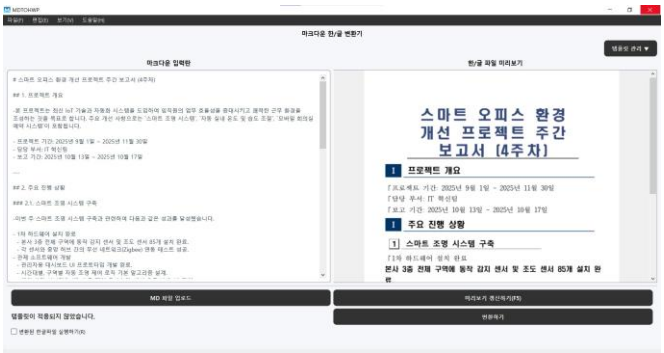

그림 2. 마크다운 입력(좌) 및 HWP 변환 결과(우)

실험 결과, 마크다운의 제목 문법(#, ##)은 템플릿에서 가장 큰 글자 크기를 가진 스타일(제목 1)과 두 번째로 큰 스타일(제목 2)에 각각 정확히 매핑되었다. 또한, 목록 문법(-)은 템플릿의 글머리 기호 스타일로 성공적으로 변환되었다. 이를 통해 제안하는 시스템이 HWP 템플릿의 스타일 구조를 효과적으로 분석하고, 마크다운 문서를 해당 서식에 맞춰 일관성 있게 변환할 수 있음을 확인하였다.

4. 결론

본 논문에서는 HWP 문서의 내부 XML 구조와 명시적 특징에 기반한 휴리스틱 알고리즘을 통해, 문서 스타일을 자동으로 추론하는 시스템을 구현하고 그 유효성을 검증했다. 제안된 시스템은 마크다운으로 작성된 문서의 HWP 변환 자동화 기반을 마련하여, 반복적인 서식 지정 작업으로 인한 비효율을 크게 개선할 수 있음을 보였다.
다만, 현재 시스템은 글자 크기와 글머리 기호 등 제한된 특징에 의존하므로 복잡한 문서 구조에서는 추론 정확도에 한계가 있을 수 있다. 향후 연구로는 들여쓰기, 줄 간격, 굵기(bold) 등 더 다양한 스타일 속성을 특징으로 추가하고, 이를 가중치 모델과 결합하여 다단계 문서 구조에서의 분류 정확도를 높이는 연구를 진행할 예정이다.

참고문헌

[1]J. Gruber, "Markdown: Syntax," 2004.
[2]Haansoft, "Hangul Word Processor Markup Language (HWPML) Specification," 2011.

감사의 글

이 논문은 2023 년도부터 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 결과임(No.2023-0-00076, SW 중심대학(동아대학교))