협성대학교 소프트웨어공학과

A Python Library for Measuring the Passage of Time in Fiction with a Labeled Dataset of Contemporary Korean Novels

소설 속 시간 흐름 측정 파이썬 라이브러리와
국문 현대소설 라벨링 데이터셋 개발 연구과제신청서

이시헌, 황태검

2025-9-24

A Python Library for Measuring the Passage of Time in Fiction with a Labeled Dataset of Contemporary Korean Novels

1. 서론

본 연구는 세 부분으로 구성된다. fic-time 이라 명명한 라이브러리를 구축하는 것, 국문 현대소설 시간흐름 데이터셋 개발, 국문 시간흐름 계측 기능을 fic-time에 추가하는 것.

이때 첫 번째 목표만을 달성하여 학술지에 게재하는 경우에는, 10/1 마감인 『디지털인문학』(Korean Journal of Digital Humanities, KJDH)에 투고한다. (한국디지털인문학협의회, 25년도 2호)

- 학술지 종류: 미등재
- 게재예정학회지명: 디지털인문학

만약 연구의 전 과정을 수행하는 경우에는, 26년도 2/28 마감인 『언어와 정보』에 투고한다. (한국언어정보학회, 26년도 1호)

- 학술지 종류: KCI 등재
- 게재예정학회지명: 언어와 정보

2. 연구목적 및 배경

문학 텍스트를 자동적으로 분석(automate content analysis)하는 기존의 방법론들이 있다. 대표적인 것은 'Named entity extraction'과 'Topic modeling'이 있다. 그러나 이들 방법론은 기본적으로 '단어를 세어' 통계량을 얻는 것이고, 명백한 한계를 갖는다. 예를 들면 "소설 한 페이지에 평균적으로 얼마나 긴 시간이 흐르는가?"와 같은 질문이 있다.

"wow, it's been thirty years but feels like yesterday" (Underwood, 2023)

Underwood (2023) 는 위와 같은 문장을 제시하며, 이 장면에서 30년의 시간이 흘렀다고 계산해서는 안된다는 점을 지적하며, GPT-4를 사용한 시간 흐름 추정이 유의미하다는 연구를 제시한다. 그러나 Underwood의 연구는 그 결과물이 디지털 문학 연구에 유효함을 입증했음에도, 여전히 개념 증명 단계에 머무르고 있다.

기존의 자연어처리 생태계에 시간 계측 라이브러리가 없는 것은 아니다. 그러나 이들 라이브러리는 특정한 어휘를 찾는 것에 의존한다. 'Spark NLP'의 'dateMatcher'와 같은 도구는

명시적인 날짜 표현을 찾는 것으로 동작[1]하며, 행동 묘사로부터 시간 흐름을 추론하는 기능은 없다(Yauney, 2019). 'pyTLEX'와 같은 라이브러리는 연구자의 주석에 의존[2]한다(CognacLab, n. d.). 이들 방법을 Gregory Yauney가 더욱 발전시켜 시간흐름 측정 정확도를 향상시켰지만, 여전히 인간에 비하면 현저히 낮은 성능을 보인다.[3]

이러한 상황에서, 인간 연구자에 준하는 시간 흐름 추정이 가능한 Underwood의 방법론이 문학 연구 전반에 도움이 됨에도 불구하고 적절한 후속 연구가 진행되지 않고 있다. 그러한 이유 중 하나로, 인문 연구자들의 소스코드 작성 능력 부재를 꼽을 수 있다. Underwood의 연구 결과는 GitHub에 게시되어 있지만, 어디까지나 연구 결과의 재현을 위한 소스코드에 불과하다. 인문 연구자들이 실제 연구에 즉각적으로 사용할 수 있는 상태가 아니라는 뜻이다. 이에 따라 본 연구는, Underwood의 연구 결과를 토대로, **'사용하기 편한 연구 도구로서의, 언어모델을 사용해 인간 연구자와 유사한 시간 흐름 추정 결과를 산출하는 파이썬 라이브러리'를 제작하여 배포하는 것을 목표로 한다.**

3. 주요 연구내용 및 방법

연구는 다음 세 단계로 진행된다.

- 파이썬 라이브러리(fic-time)로 Ted Underwood의 연구 재현 및 검증.
- 국문 소설 시간흐름 데이터셋 개발.
- Underwood의 연구를 국문으로 전환 및 검증.

먼저 Underwood가 작성한 소스코드를 pip 라이브러리 형태로 가공한다. 가공 과정에서 포함되어야 할 모듈의 목록은 다음과 같다.

모듈 이름	기능
Preprocessor	사용자 입력 텍스트 정제. 불필요한 공백 제거, 비표준 문자 정규화. 텍스트 분할.
Prompter	사용자 정의 프롬프트 템플릿 저장. 시스템 메시지, 프롬프트 등의 입력 인터페이스 제공.
Gateway	외부 LLM API 통신 추상화 제공. API 키 관리, HTTP 요청 전송, 오류처리 제공.
Parser	LLM 응답에 대한 안전한 파싱 수행. 예외처리 제공. 오류 검증 수행.
Schemas	사용자 정의 출력 구성 기능 제공. 다루기 쉽게 잘 구조화 된, 깨끗한 캡슐로 데이터 출력. (Pandas DataFrame 등의 데이터 분석 도구에 즉시 통합될 수 있는 출력 형식 지정 가능.)

[1] "DateMatcher and MultiDateMatcher extract *exact* & *normalized dates* from relative date-time phrases and convert these dates to a *provided date format*. DateMatcher can only extract one date per input document while MultiDateMatcher can multiple dates." [2]

[2] "**pyTLEX** is the Python library to perform temporal analysis of TimeML annotated texts."

[3] 문학사적 맥락에서는 유효하나, 정성연구의 정밀도에는 미치지 못한다.

그리하여 Underwood가 제시한 데이터 흐름의 파이프라인을 높은 수준으로 추상화한 단일 시스템을 구축한다. 연구자는 본 시스템이 외부에 노출한 몇 개의 함수와 변수를 조작하는 것으로 간편하게 문학 텍스트의 시간 흐름을 추정할 수 있게 되어야 하며, 각 모듈의 작동 방식을 얼마든지 수정할 수 있어야 한다. 연구자의 역량에 따라 얼마든지 다른 시스템이나 모듈과 연동할 수 있도록, 확장성 있는 구조를 갖추어야 한다는 뜻이다.

이때 fic-time의 성능을 향상시킬 단서로서 참고할 만한 연구로서, 서사학의 이론적 틀 속에서 LLM을 사용한 시간흐름 계측을 수행한 연구(Piper & Bagga, 2024)가 있다. 그러나 본 연구의 목표는 시간흐름 계측의 정확도 향상이 아니므로, 성능에 관한 추가적인 연구를 진행하지 않는다.

둘째로, 해당 연구를 국문으로 전환하기 위한 선행 작업으로, 국문 소설을 대상으로 하는 '시간 흐름'이 라벨링 된 데이터셋을 개발한다. 해당 작업은 협성대학교 소속의 4학년 이상 현대소설 전공자를 동원하여 진행하며, 저작권이 소멸한 채만식, 이상, 현진건 등의 소설을 대상으로 한다.

셋째로, Underwood의 방법론을 적용한 국문 전용 프롬프트를 개발하여, 국문 소설을 대상으로 한 시간흐름 추정의 유효함을 검증할 예정이다. Underwood가 계측했던 것과 유사한 상관계수(r=.74)를 달성하는 것을 목표로 하며, 프롬프트만으로 목표 성능에 도달할 수 없는 경우, 모델에 대한 파인튜닝을 수행할 수 있다.

4. 연구진행 계획

9월 30일까지 fic-time 라이브러리가 Underwood의 연구를 정확히 재현했는지 검증한다. 다수의 소형 언어모델로 Underwood가 사용한 123문장에 대한 r 값을 산출하는 것을 포함한다.

10월 31일까지 국문 현대소설 데이터셋의 개발을 완료한다. 다른 국문학 연구에 응용될 수 있는 충분한 수량의 데이터 라벨링을 수행한다. 이때 어느 정도가 충분한지를 정하는 것 또한 연구에 포함된다.

11월 31일까지, 앞서 개발한 데이터셋에 기반하여 Underwood의 프롬프트에 준하는 성능을 보이는 국문 전용 프롬프트와 알고리즘을 구성한다. 개발이 완료되면 fic-time에 해당 기능을 포함시키고, pip 저장소에 라이브러리를 등록한다.

12월 전체 기간동안 추가 기능 구현에 돌입한다. 인물별 시간흐름 계측, 인물별 공간변화 집계, 계산 결과의 시각화 기능 등을 포함한다. 모든 기능의 구현과 동작 검증이 완료되면, 연구 논문을 작성하고 학술지에 게재한다.

5. 기대효과

- 질적연구와 양적연구가 혼합된 방식의 국문학 연구에 응용할 수 있는 데이터셋을 산출하여, 향후 새로운 인문 연구를 촉발한다.
- 인문 연구자들을 위한 편리한 연구 도구를 제공하는 것으로, 멀리서 읽기에 기반한 새로운 국문학 양적연구들이 촉발될 것이다.

참고문헌

- Underwood, T. (2023, March 19). Using GPT-4 to measure the passage of time in fiction. *Ted Underwood's Blog.* https://tedunderwood.com/2023/03/19/using-gpt-4-to-measure-the-passage-of-time-in-fiction/
- Yauney, G. (2019). Computational prediction of elapsed narrative time. *New Literary History, 85*(2), 351 – 365.
- CognacLab. (n.d.). *Pytlex.* Florida International University. https://cognac.cs.fiu.edu/pytlex/
- Piper, A., & Bagga, S. (2024). Using large language models for understanding narrative discourse. In *Proceedings of the 6th Workshop on Narrative Understanding (WNU 2024).*