2023-2 언어데이터과학 기말프로젝트 최종보고서 샘플

한국어 부사 '완전'의 통사적·의미적 확장에 관한 연구를 위한 '모두의 말뭉치' 활용

2020-20202 박수민

(이 샘플은 하나의 예시일 뿐, 실제 제출하는 보고서는 아래의 구조와 일치하지 않아도 무방하다.)

1 서론

1.1 연구 목적

(연구 목적을 간명하게 쓴다. 연구계획서에 썼던 내용을 다 옮겨 적을 필요는 없다.)

[예시]

이 프로젝트의 목적은 현대 한국어에서 '완전'의 용법이 발화자의 연령과 성에 따라 차이를 보이는지를 알아보는 것이다.

1.2 연구 계획 대비 목표 달성 결과

(연구계획서에 작성한 내용이 계획대로 실행되었는지 쓴다.)

[예시]

	계획	**실행**	**결과**
데이터 수집	[[모두의 말뭉치]] 중 [일상 대화 말뭉치] 및 [온라인 대화 말뭉치] 다운로드.	[[모두의 말뭉치]] 중 [온라인 대화 말뭉치] 버전 1.1을 다운로드함.	일부 달성
데이터 가공	JSON 파일 --> 데이터프레임 변환.	pandas를 사용하여 실행함. [해당 코드]	전체 달성
데이터 활용	데이터프레임이 구축되면 pandas 라이브러리와 정규표현식을 사용하여 발화 형태로부터 '완전'의 출현 문맥을 추출할 수 있다.	[코드] 참조.	전체 달성
...	...	...	...

1.2.1 미달성 사유

(계획 중 달성하지 못한 부분이 있으면 그 이유를 쓴다.)

[예시]

- [[모두의 말뭉치]] [일상 대화 말뭉치]는 파일이 너무 커서 로컬 및 코드스페이스의 저장 가능한 용량을 초과했기 때문에 이 연구에서 사용할 수 없었다.

2 연구 방법

2.1 코퍼스

(코퍼스 크기, 포맷 등을 기술한다.)

[예시]

[[모두의 말뭉치]] [온라인 대화 말뭉치]는 총 74,665건의 온라인 대화가 저장된 47,421개의 JSON 파일로 이루어져 있으며, 전체 크기는 835MB이다.

2.2 텍스트 전처리

(문장 단위 분리, 불필요한 문장 부호 제거 등 해당하는 것이 있는 경우 방법을 설명한다.)

[예시]

[온라인 대화 말뭉치]는 국립국어원에서 공개한 버전에 이미 특수 메시지 코딩({emoji:✨ }, {share:url} 등)이 완료되어 있어서 전처리가 많이 필요하지 않다. 단, 발화가 특수 메시지만으로 이루어진 경우 언어적 표현이 없으므로 해당 발화는 제외해야 한다.

2.3 데이터 가공

(전처리를 마친 데이터의 가공 방법을 간략하게 설명하고 최종 통계량을 제시한다.)

[예시]

우선 47,421개의 JSON 파일을 발화 단위로 분리하고 특수 메시지만으로 이루어진 경우를 제외하면 모두 2,977,840건의 데이터를 얻었다. 코퍼스에서 제공하는 데이터의 속성(attribute) 중 id, form, speaker_id 세 가지만을 남기고 데이터프레임으로 저장했다.

다음으로 각 파일의 발화자 목록을 읽어서 age, sex, speaker_id 세 개의 변수를 가진 새로운 데이터프레임으로 저장했다.

최종적으로는 발화 데이터프레임과 발화자 데이터프레임을 공통 변수인 speaker_id로 통합하여 각 발화에 발화자의 연령과 성 정보를 추가했다.

지금까지 기술한 과정의 구현 방법은 아래의 코드에서 차례로 확인할 수 있다.

- [관련 코드 노트북 1]
- [관련 코드 노트북 2]

이 과정에서 얻은 데이터 파일은 아래와 같다.

- [데이터 파일(발화)]
- [데이터 파일(발화자)]
- [데이터 파일(통합)]

3 연구 결과

3.1 현상 기술

(데이터에서 발견한 사실을 기술한다. 시각화 내용이 있는 경우 이 절에 포함시킨다.)

[예시]

발화 정보와 발화자 정보를 통합한 데이터에서 연령별로 '완전'과 '아주'의 사용 비율을 구한 결과는 아래 표와 같다.

	완전	아주
10대	90.755008	9.244992
20대	74.070822	25.929178
30대	67.589049	32.410951
40대+	59.757155	40.242845

연령이 높아짐에 따라 '완전'의 사용이 상대적으로 증가하는 현상을 볼 수 있다.

(후략: '완전'과 '아주'의 성별 비율, 연령별*성별 비율, '완전'과 '아주'가 수식하는 단어 목록 등을 같은 방식으로 서술한다.)

3.2 분석

(새로 발견하거나 확인한 사실의 원인이나 함의를 분석한다.)

[예시]

발화자의 연령이 낮을수록 '완전'의 사용 비율이 '아주'에 비해 높아지므로, '완전'의 쓰임이 확장되는 현상이 현재에도 진행 중임을 알 수 있다.

(후략)

4 결론

(연구 결과를 요약한다. 추후 계획이 있다면 쓴다.)

[예시]

이 연구에서는 '완전'이 정도부사로서 그 쓰임이 확장되고 있는지를 검토하기 위해 [모두의 말뭉치] [[온라인 대화 말뭉치]]를 활용하여 실제로 발화된 데이터에서 '완전'의 사용량이 연령별·성별 요인에 따라 어떻게 달라지는지를 조사했다. 조사 결과 '완전'의 사용량이 연령이 감소함에 따라 높아지는 경향이 나타났고, 이를 통해 현대 한국어에서 '완전'의 쓰임이 확장되고 있다는 결론을 내릴 수 있다.

참고 문헌

- 국립국어원(2022). 국립국어원 온라인 대화 말뭉치(버전 1.1). URL: https://corpus.korean.go.kr
- 김다미(2021). 〈부사적 쓰임을 보이는 '완전'에 대한 통시적 고찰〉. 《국어학》 97, 439-475. URL: https://kiss.kstudy.com/thesis/thesis-view.asp?key=3878140