한국어 대용량 코퍼스의 오류 어휘 탐지 방안

- 최민주, 박지훈, 손성환, 강승식 -

대용량 코퍼스의 오류 어휘 탐지

Abstract

텍스트 데이터를 분석에 이용하기 위해서는 우선 대용량 코퍼스에 존재하는 오류 어휘를 정확하게 탐지하여 정제하는 전처리 과정이 필요하다. 이러한 전처리 결과는 데이터 분석의 성능에 영향을 주기 때문에 텍스트 데이터로부터 오류 어휘를 효과적으로 탐지하는 방법이 필요하다.

본 논문에서는 대용량 텍스트 데이터에 존재하는 오류 어휘를 효과적으로 탐지하는 방법을 제시한다. 우선 대용량 코퍼스 내에서 자주 쓰이지 않는 음절을 오류 어휘로 가정하고, 해당 음절이 실제로 오류 어휘인지 검증한다. 이어서 검증 결과를 분석하여 음절 빈도 수를 이용하여 오류 어휘를 탐지할 때 고려하여야 할 점에 대해 기술한다.

대용량 코퍼스 데이터

- KCC150 (한국어 원시 문장 말뭉치)
  - 문장 11,961,347 개, 어절 150,705,457 개

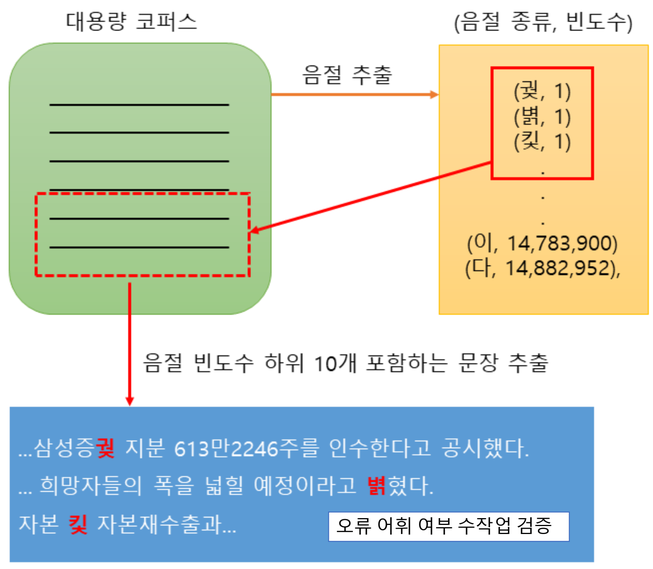
Main Idea

- 수많은 어휘 중 오류 어휘는 일부분에 불과
  -> 코퍼스 내에서 등장 빈도가 적은 어휘를 오류 어휘로 추정

장점

- 사전에 데이터 세트를 구축하지 않고 오류 어휘 탐지 가능

오류 어휘 탐지 방안 흐름도



음절 빈도수를 이용한 오류 어휘 탐지

KCC150 에 존재하는 음절 종류와 빈도수

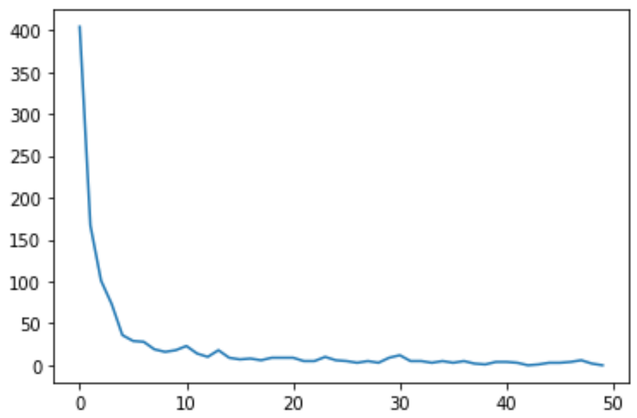
최상위 빈도수 음절 종류	최하위 빈도수 음절 종류
('다', 14882952), ('이', 14783900), ('는', 9485560), ('에', 9412842), ('을', 8908798), ('의', 7433544), ('지', 7069223), ('로', 6560180), ('가', 6524429), ('고', 6413492),	('귖', 1), ('볇', 1), ('킻', 1), ('귬', 1), ('뿕', 1), ('쒚', 1), ('윝', 1), ('멇', 1), ('뉸', 1), ('끍', 1),

빈도수가 낮은 음절이 철자 오류 음절인 예

- 삼성증**귖**
- 희망자들의 폭을 넓힐 예정이라고 **볇**했다.
- 자본 **킻** 자본재수출과
- 정부는 2015년까지 원조**귬**를 30억 달러로
- 검은 연기를 내**뿕**고 있는

음절 빈도수에 따른 음절 종류의 수

음절 빈도 수	음절 종류 수
1	404
2	167
3	101
4	73
5	36
6	29
7	28
8	19
9	16
10	18
합계	**891**



오류 어휘 탐지 결과

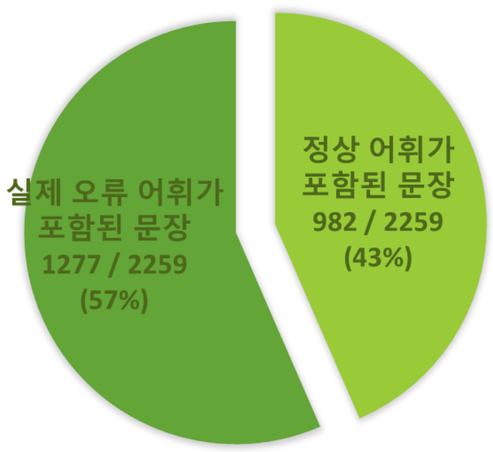
오류 어휘 포함 문장 추출

- 빈도수 10이하 음절 891개를 포함하는 문장 추출 →총 2259 문장

실제 오류 어휘 예시

- **짧**은(짧은) 머리에 초록색 상의를 입고 있었다.
- 그 길로 그녀와 헤어져 나는 장인을 **탖**아갔지(찾아갔지).
- 호쾌한 타격**읗**(타격을) 선보이고 있다.
- **멇**은(많은) 분들이 나에게 기대를 하고 있는데 부족했다.
- 2층의 선사 시대실로 들어**갂**다(들어갔다)

문장 분석 결과



정상 어휘 예시

유형	예시
외래어	커피**숖**, 케**촙**
의성어(의태어)	**츳츳, 슉 슉**
사투리	**욾**어(없어), 죅여**쁘**릴(죽여버릴)
준말(줄임말)	**걜** (걔를), 몇 **횝**니까?(몇 회입니까?)
고유명사	카를 **뵘**, 청춘**삘**딩
두루 쓰지 않는 어휘	**켯**속, 웃**똠**

결과 분석

예외

1) 오류 음절의 빈도수가 높은 오류 어휘 예

박달나무로 만들었는데, 한쪽을 **깍**아서, 몽둥이라기보단 창이었다.

- '깍' 은 KSX1001 에 포함된 자주 사용되는 음절
- 빈도수를 이용하는 방법으로는 탐지 불가능

- N-gram 모델을 이용하여 보완 가능
  - ex) unigram 어절 빈도 수 비교 ('깎아서' > '깍아서')

2) 어휘 호응이 맞지 않는 오류 어휘 예

어려운 **소**비를 넘고 (어려운 고비를 넘고)

- 각 음절 및 어절은 정상적이나 서로 어휘 호응이 맞지 않는 오류
- 빈도수를 이용하는 방법으로는 탐지 불가능

- N-gram 모델을 이용하여 보완 가능
  - ex) bigram 어절 빈도 수 비교 ('어려운 고비를' > '어려운 소비를')

결론

- 대용량 코퍼스에 존재하는 오류 어휘를 탐지하는 방법 제안
  - 음절의 빈도수가 낮은 어휘를 오류 어휘로 탐지

- 오류 어휘 탐지 시 고려할 사항
  - 실제 오류 어휘가 아닌 경우
    - 외래어, 의성어(의태어), 사투리, 준말(줄임말), 고유명사 등
  - 오류 어휘가 탐지되지 않는 경우
    - 오류 어휘의 빈도수가 높은 경우
    - 어휘 호응이 맞지 않는 경우