한국어 대용량 코퍼스의 오류 어휘 탐지 방안

최민주○, 박지훈, 손성환, 강승식

국민대학교, 컴퓨터공학과
mjchoi0831@kookmin.ac.kr, hoonzinope@kookmin.ac.kr, ssh121@kookmin.ac.kr , sskang@kookmin.ac.kr

Error Word Detection in Korean Corpus

Min-Joo Choi○, Ji-Hoon Park, Sung-Hwan Son, Seung-Shik Kang
Kookmin University, Dept. of Computer Science

요 약

대용량의 언어 코퍼스를 이용할 때, 오류 어휘가 코퍼스에 포함되어 있는 경우 해당 코퍼스를 이용한 실험의 성능이 저하될 수 있다. 이 때문에 정확한 문장들로 이루어진 코퍼스를 구축하기 위해 다량의 문장 중에서 정확하게 오류 어휘를 탐지할 필요가 있다. 본 논문에서는 대용량 데이터에서 빈도수가 낮은 음절을 이용해 오류 어휘를 탐지하는 방법을 제안하고, 제안 방법을 이용하여 오류 어휘 탐지 시 고려하여야 할 점에 대해 서술한다.

주제어 : 철자 오류 탐지, 오류 어휘, Unigram, 자연어처리

1. 서론

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.

2. 관련 연구

철자 오류에 관한 기존 연구로는 형태소 분석결과를 이용한 방법[2], 자소 단위 철자 오류 교정 방법[3], 교정 사전을 사용한 방법[4], N-gram 모델을 이용한 방법[5] 등이 있다. 이 중 교정 사전 및 N-gram 을 이용하는 방법은 사전에 구축된 데이터가 필요하다.

교정 사전의 경우 오류 문자열과 교정 문자열을 한 쌍으로 하여 사전에 포함된 오류 문자열이 데이터에 존재하는 경우 교정 문자열로 변환하는 방법이다. 이 방법은 사전에 포함된 오류 문자열을 정확하게 탐지하고 변환할 수 있으나 미리 교정 사전을 구축해야 하고, 교정 사전에 존재하지 않는 어휘는 탐지할 수 없는 단점이 있다.

N-gram 모델은 문맥에 의존하여 오류 어휘를 탐지 및 교정하는 방법으로, 미리 N-gram 모델을 이용하여 어절 또는 음절 데이터 세트를 생성해야 한다. 이 경우 N의 값이 커질 때마다 데이터 세트를 생성하기 위한 계산 시간이 늘어나며, 데이터 세트에 존재하지 않는 어휘는 탐지할 수 없다.

본 논문에서는 음절 빈도수에 기반하여 미리 데이터 세트를 구축하지 않고 오류 어휘를 탐지할 수 있는 방법을 제안하고, 탐지 결과에 대해 설명한다.

3. 대용량 코퍼스의 오류 어휘 탐지

3.1 음절 빈도수를 이용한 오류 어휘 탐지 방법

대용량 코퍼스 데이터로 'KCC150'에서 추출한 문장들을 이용하였다. KCC150은 한국어 원시 말뭉치로 총 11,961,347문장(1억5천만 어절)으로 이루어져 있다.

KCC150에 포함된 대부분의 문장에는 오류 어휘가 존재하지 않으나, 일부 문장에서 오류 어휘가 발견되었다. 수많은 어휘 중 오류 어휘는 일부분을 차지하므로 오류 어휘는 코퍼스 내에서 그 등장 빈도가 적은 어휘로 추정할 수 있다.

코퍼스 내의 어휘들을 음절 단위로 확인하기 위해 Unigram 을 이용하여 KCC150에 존재하는 모든 음절과 그 빈도수를 추출한다. 이어서 추출된 음절 중 코퍼스 내에 가장 많이 나타난 음절 10개, 가장 적게 나타난 음절 중 10개를 확인하여 비교하였다.

표 1 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),	('궂', 1), ('볅', 1), ('킻', 1), ('귭', 1), ('뿤', 1), ('쎞', 1), ('윝', 1), ('멶', 1), ('뉸', 1), ('끍', 1),

문장의 종결어미로 주로 쓰이는 음절 '다' 는 KCC150 말뭉치에 총 14,882,952 개로 가장 많이 포함되어 있다. 이어서 한국어 조사 '는', '이/가' 또한 빈도수 상위 10 개 이내에 존재하므로 자주 쓰이는 어휘일수록 코퍼스 내에서 빈도수가 높다는 것을 알 수 있다.

반면 빈도수가 낮은 음절을 살펴본 결과 '귲',

‘볅’, ‘킲’ 과 같이 빈도수가 적고 한국어 텍스트에서 찾아보기 어려운 음절들은 철자 오류로 확인되었다.

표 2 빈도수가 낮은 음절이 철자 오류 음절인 예

삼성증**귖** 희망자들의 폭을 넓힐 예정이라고 **볅**혔다. 자본 **킲** 자본재수출과 정부는 2015년까지 원조**귬**를 30억 달러로 검은 연기를 내**뿈**고 있는

표 3 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

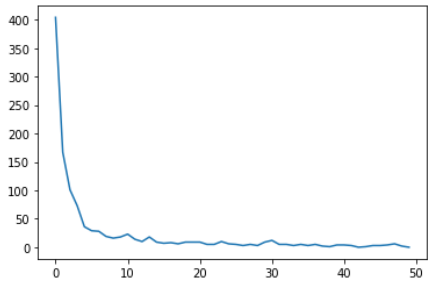

그림 1 음절 빈도수에 따른 음절 종류의 수

3.2 오류 어휘 탐지 결과

‘귖’, ‘볅’ 과 같이 KCC150 에 10회 이하로 등장하는 음절은 총 891 개 존재하며, KSX1001 에서 자주 쓰이지 않는 705개 음절을 포함한다. 따라서 KCC150으로부터 오류 어휘일 가능성이 높은 빈도수 10 이하 음절 891개를 포함하는 문장을 추출한다.

추출 결과 총 2259개 문장이 추출되었으며, 1개의 문장에 최대 4개의 오류 음절이 포함되어 있다. 실제 오류 어휘가 포함되어 있는지 수작업으로 검증한 결과 2259개의 문장 중 1277개 문장에 실제 오류 어휘가 존재하여 약 56% 의 정확도를 보인다.

표 4 실제 오류 어휘가 포함된 문장 예

짧은(짧은) 머리에 초록색 상의를 입고 있었다. 계약이 성사되**먄**(성사되면) 적응을 얼마나 빨리 하느냐에 따라서 달라진다고 생각한다. 그 길로 그녀와 헤어져 나는 장인을 **탖**아갔지(찾아갔지). 호쾌한 타격**웅**(타격을) 선보이고 있다. **멇**은(많은) 분들이 나에게 기대를 하고 있는데 부족했다. 한화는 이날 **툿**(투수) 이동걸, 포수 엄태용, 외야수 최진행을 1군으로 올렸다. 2층의 선사 시대실로 들어**갂**다(들어갔다).

3.3 결과 분석

탐지 및 검증 결과 2259개의 문장 중 982개 문장에 빈도수는 낮지만 실제 오류 어휘로 간주할 수 없는 음절이 존재하며, 전체의 약 43%를 차지한다.

표 5 실제 오류 어휘가 아닌 예

유형	예시
외래어	호텔 커피**숖** 구석자리에 미리 와서 기다리고 있던 명수는 내가 들어서자마자 컬러사진 한 장을 내밀었다. 대신 버섯, 호두, 오이 등으로 **케춥**을 만들었다.
의성어(의태어)	나는 올곧지 않게 대답했고 늙은이는 **츳츳** 혀를 찼다. **슉 슉** 하는 김빠지는 소리만 날 뿐 총성도 들리지 않았다.
사투리	다긔차기가 마른 건천이 돌팍 같은 애가 왜 그렇게 총긔가 **옰**어. 내 손으로 늴 쥑여**쁘**릴껴!
준말(줄임말)	**걜** 시중 들어주는 언니도 있어. 그렇다면 경북고 몇 **횝**니까?
고유명사	카를 **뵘**, 레너드 번스타인, 폰 카라얀 대학생 봉사단원, 청춘**삘**딩 센터장 등 다양한 분야에서 선발했다.
두루 쓰지 않는 어휘	감독인 김기호는 돌아가는 일의 **켯**속을 꿰고 있는 듯 보였다. 나는 그 다섯 마을 중의 웃**똠**에서 났다.

예시 중 ‘켯속’은 ‘일이 되어 가는 속사정’을 뜻하는 말로 표준국어대사전에 등재되어 있다. 하지만 그 사용빈도가 낮아 빈도수에 기반한 오류 어휘를 탐지한 실험에서 부득이하게 오류로 탐지되었다.

또한 실제 오류가 발생하였음에도 탐지되지 않은 어휘가 존재하는데, ‘깍’은 KSX1001 에 포함된 자주 사용되는 음절이므로 빈도수를 이용하는 방법으로 탐지할 수

없다.

표 6 실제 오류 어휘가 탐지되지 않은 문장 예

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

4. 결론

본 연구에서는 음절 빈도수를 이용하여 대용량 한국어 코퍼스로부터 오류 어휘를 탐지하는 방법을 제안하였다. KCC150과 같이 대용량의 원시 말뭉치로부터 Unigram 을 이용하여 음절 단위로 코퍼스 내 등장 빈도를 추출한 후, 빈도수에 기반하여 오류 어휘를 탐지하는 방법에 대해 설명하였다.

빈도수가 낮은 음절이 포함된 문장을 수작업으로 검증한 결과 총 2259개의 문장 중 1277 개의 문장에 실제 오류 어휘가 존재하였다. 정상적인 어휘임에도 오류 어휘로 탐지된 경우 대부분 한국어의 특성상 자주 쓰이지 않는 어휘임을 알 수 있었다.

향후 본 연구는 오타 탐지 및 교정 기술이 필요한 모든 응용분야, 키보드 오타 교정 및 오탈자 탐지 등에 응용될 수 있다.

감사의 글

이 논문은 2017 년 정부 (과학기술정보통신부) 의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2017M3C4A7068186)

참고문헌

[1] 이세희, 김학수, “교정 사전과 신문기사 말뭉치를 이용한 한국어 철자 오류 교정 모델”, 정보처리학회논문지B 16권5호, pp. 424-434, 2009

[2] Eric Brill, Robert C. Moore, “An Improved Error Model for Noisy Channel Spelling Correction", In Proc. of the 38소 Annual Meeting of the ACL, pp.286-293, 2000.

[3] 윤근수. 권혁철, “교정률 최적화를 위한 한국어 철자교정기의 모듈 배열”, 정보과학회논문지: 소프트웨어 및 응용, 제32권 제 5호, pp.366-377, 2005.

[4] 강승식, 장두성, “SMS 변형된 문자열의 자동 오류 교정 시스템", 정보과학회논문지: 소프트웨어 및 응용, 제35권 제6호. pp.386-391, 2008.

[5] 김민호, 권혁철, 최성기, “어절 N-gram 을 이용한 문맥의존 철자오류 교정”, 정보과학회논문지, 제 41권 제 12호, pp. 1081-1089, 2014

[6] 국민대학교 한국어 원시 말뭉치 KCC150, http://nlp.kookmin.ac.kr/kcc/