공개와 협업을 통한 세종 형태 분석 말뭉치 오류 개선 방법

한경은○, 백슬예, 임재수

(주)카카오

grace.han@kakaocorp.com, cecil.rosa@kakaocorp.com, jamie.lim@kakaocorp.com

Open Sourced and Collaborative Method to Fix Errors of Sejong Morphologically Annotated Corpora

Gyeong-Eun Han○, Seul-Ye Baek, Jae-Soo Lim

KAKAO Corp

요 약

본 논문에서는 21세기 세종계획 “현대문어 형태 분석 말뭉치"에서 나타나는 오류를 개선하는 방법으로 패치 시스템을 제안한다. 이 패치 시스템은 패치 파일과 패치 적용-생성 스크립트로 구성되며, 사용자들은 패치 파일을 사용하여 원래의 말뭉치에서 어떤 파일과 어절을 수정하였는지 확인할 수 있어 개발 목적에 맞는 학습 말뭉치를 생성할 수 있다. 또한 이 시스템을 이용해 서로의 수정 사항을 공유하고, 지속적으로 세종 말뭉치의 오류를 개선할 수 있다. 본 논문에서는 총 1,015만 어절을 대상으로 31만여 개의 오류를 수정하였다. 오류의 유형으로는 문장, 어절 분리 오류, 철자 오류, 불일치 오류, 분석 오류, 형식 오류가 있으며, 오류 수정 사항을 패치 파일에 반영하였다.

주제어: 세종 형태 분석 말뭉치, 오류 수정, 공개 및 협업

1. 서론

자연언어처리 분야에서 세종 형태 분석 말뭉치는 형태소 분석기나 품사 태거를 개발하는 데 활용된다. 그러나 세종 형태 분석 말뭉치 자체에는 철자 오류, 분석 오류, 형식 오류 등이 포함되어 있어 원래의 말뭉치 그대로를 학습 말뭉치로 사용하는 데 어려움이 있다. 따라서 대부분의 연구에서는 세종 형태 분석 말뭉치를 학습 말뭉치로 활용하기 위해 1차적으로 말뭉치의 오류를 수정하는 작업을 수행한다.

그러나 위 연구들의 결과물이 공개되어 있지 않아 수정된 말뭉치를 활용하는 것이 쉽지 않다. 또한 공개된 말뭉치라도 원래의 말뭉치에서 어떠한 부분이 수정되었는지 파악하기 어려워 개발 목적에 따라 반영하고 싶지 않은 수정 사항이 있을 경우, 2차적으로 그러한 부분을 일일이 확인하고 수정하는 데 오랜 시간이 걸린다.

본 논문에서는 위와 같은 어려움을 개선하기 위해 패치 시스템에 기반한 말뭉치의 오류 수정 방법을 제안한다. 사용자들은 이 시스템으로 누구나 오류가 수정된 세종 형태 분석 말뭉치를 생성할 수 있고, 말뭉치에서 어떤 파일과 어절을 수정하였는지 확인할 수 있어 개발 목적에 맞는 학습 말뭉치를 생성할 수 있다. 또한 서로의 수정 사항을 공유해 지속적으로 세종 말뭉치의 오류를 개선할 수 있다.

본 논문의 구성은 다음과 같다. 2장에서는 세종 말뭉치의 오류 수정과 관련된 연구와 오류를 수정한 대용량 말뭉치에 관해 살펴본다. 3장에서는 패치 시스템에 기반한 오류 수정 방법을 제안하고, 4장에서는 3장에서 제안한 방법을 적용한 결과를 제시한다. 5장에서는 본 논문의 결과와 앞으로의 연구 방향에 대해 설명하고 논문을 마친다.

2. 관련 연구

세종 말뭉치에서 나타나는 오류들을 검출하고 수정하기 위해서 다양한 연구들이 선행되었다. [1]에서는 원어절과 형태 분석 결과를 자모 단위로 분리하고 서로의 대응관계를 비교하여 오류를 수정하였다. [2]에서는 형태소 분석 결과에 포함된 형태소들을 결합하여 이를 원어절과 비교하는 방식으로 오류를 검출하였고, 이렇게 검출된 오류들을 수정할 수 있는 도구를 개발하였다. [3]에서는 세종 말뭉치로 학습한 품사 태거의 결과와 정답 결과를 비교해 오류 어절을 추출하고, 고빈도로 나타난 어절을 우선적으로 수정하여, 약 15만여 개의 오류를 수정하였다. [4]에서는 1500만 어절 규모의 세종 형태 의미 말뭉치를 어휘 분석 말뭉치로 재가공하는 과정에서 세종 지침을 수정하고, 오류 유형별로 검출 도구를 이용하여 오류를 수정하였다. [5]에서는 학습 데이터로 사용하는 말뭉치의 신뢰도를 검증하기 위해 오류 유형을 분류하고, 각 오류 유형별 검증 방법을 제안하였다.

위와 같이 세종 말뭉치의 오류 수정과 관련된 연구들은 세종 말뭉치의 오류 유형을 정의하고, 이러한 오류를 효율적으로 검출할 수 있는 오류 검출 도구나 오류 수정 도구 개발에 관한 것이다.

그러나 수정된 말뭉치의 활용 측면을 고려해 볼 때 이러한 연구들의 연구 결과물들은 학습 말뭉치로 사용하기에 어려움이 있다. 수정된 말뭉치가 공개되어 있지 않거나, 공개가 되어 있더라도 어떤 어절이 수정되었는지 수정된 내용을 파악하기 힘들기 때문이다.

또한, 대용량의 세종 형태 분석 말뭉치를 대상으로 오

류를 수정한 창원대 말뭉치, 고려대 말뭉치(SJ-RIKS), 울산대 말뭉치(UCorpus-HG)에서는 기존의 세종 말뭉치 지침을 수정하거나 원문 자체를 수정하였기 때문에 수정된 말뭉치 그대로 사용하기 어렵다.

구체적으로 창원대 말뭉치에서는 접두사는 인정하지 않고, 접두사와 후행 명사가 결합한 형태를 일반명사로 분석하였고, 명사와 명사로 이루어진 어절은 하나의 명사로 분석하였다[3]. 예를 들어서, '응시자격을' 이라는 어절이 있다면, 세종 말뭉치에서는 '응시/NNG + 자격/NNG + 을/JKO'로 명사와 명사의 결합을 분리하여 분석하지만, 창원대 말뭉치에서는 '응시자격/NNG + 을/JKO' 와 같이 명사와 명사의 결합을 단일 명사로 분석한다.

울산대 말뭉치에서는 원어절에서 철자 오류가 발생한 경우, 원본 데이터를 수정하거나 형태 중심이 아닌 표준국어대사전에 등재된 어휘 중심으로 분석하였다. 예를 들어서 '생산적' 이라는 어절을 형태 중심으로 분석한다면, '생산/NNG + 적/XSN'으로 분석해야 하지만, 울산대 말뭉치에서는 표준국어대사전 표제어를 기준으로 해당 어절을 단일어인 '생산적/NNG'로 분석하였다.

고려대 말뭉치에서도 어휘 중심의 정보를 추출할 수 있는 말뭉치로 재가공하였기 때문에 접두사를 인정하지 않았고, 접미사도 세종 말뭉치에서 지정한 54개의 목록과 달리 16개의 형태를 제외하고 선행하는 명사와 통합하여 분석하였다[4].

창원대, 고려대, 울산대 말뭉치와 같이 어휘 중심으로 분석한 말뭉치로 학습한 형태소 분석기나 품사 태거는 정보 검색에 사용할 경우, 재현율이 떨어질 가능성이 있다.

따라서 본 논문에서는 원래의 세종 지침에 따라 형태 중심의 분석을 따르면서 지침에 벗어나는 오류들을 수정하고, 원래의 코퍼스에서 어떤 부분이 수정되었는지 수정된 내용을 파악할 수 있는 패치 시스템을 제안한다.

3.제안 방법

본 논문에서 제안하는 패치 시스템은 패치 파일과 패치 적용 및 생성 스크립트로 구성된다.

그림1과 같이 패치 시스템은 동일한 세종 형태 분석 말뭉치 원본을 가지고 있는 사용자와 협업할 수 있으며, 협업은 커미터(committer)가 배포한 패치 적용 스크립트, 패치 생성 스크립트, 패치 파일을 이용해 이루어진다.

사용자1처럼 세종 말뭉치의 원본을 가지고 있는 경우, 커미터가 배포한 패치 적용 스크립트와 패치 파일을 이용해, 오류가 수정된 세종 말뭉치 수정본을 생성할 수 있다. 또한 수정된 말뭉치에서 더 수정하고 싶은 부분이 있다면, 패치 파일을 직접 수정하여 새로운 세종 말뭉치 수정본을 생성할 수 있다.

사용자2와 같이 세종 말뭉치의 원본과 세종 말뭉치 수정본을 가지고 있는 경우, 커미터가 배포한 패치 생성 스크립트를 이용해 패치 파일을 생성할 수 있다. 사용자2는 이러한 패치 파일과 커미터가 배포한 패치 파일을 비교하여, 자신이 반영하지 못한 오류 사항이 있다면 자신의 패치 파일을 보완할 수 있다. 반대로 커미터가 배포한 패치 파일에 반영되지 못한 수정 사항이 있다면 커미터에게 피드백을 줘 배포용 패치 파일을 보완할 수 있다.

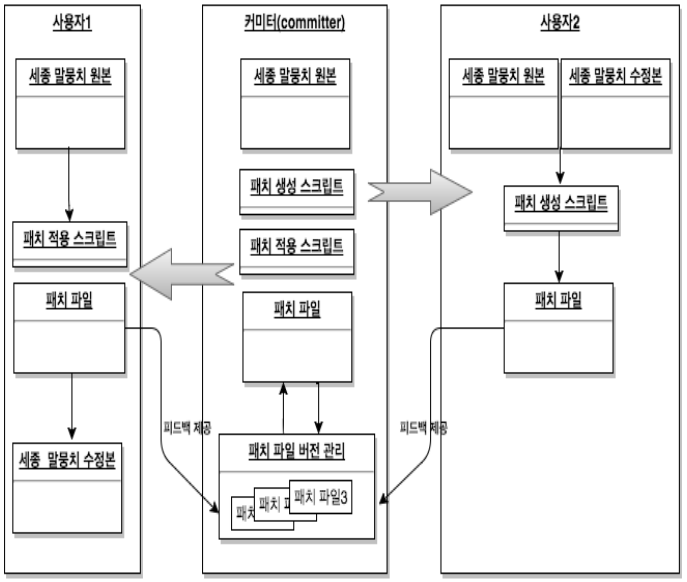

그림1. 패치 시스템을 이용한 협업 과정

이와 같이 패치 파일을 매개로, 패치 파일을 배포하는 커미터와 패치 파일을 사용하는 사용자가 서로 협업하여 지속적으로 세종 코퍼스의 오류를 개선할 수 있다.

패치 파일			
=	BTHO0397-00058767	소년들에서	소년/NNG + 들/XSN + 에서/JKB
+	BTAA0011-00012666-1	보리는	보리/NNG + 는/JX
-	BTBZ0074-00028385		
M	BTGO0348-00013286	BTGO0348-00013287	
S	BTAE0200-00042670	BTAE0200-00042671	

그림2. 패치 파일 예

패치 파일에는 세종 말뭉치에서 수정한 어절의 정보를 담고 있다. 패치 파일은 그림2와 같이 패치 종류, 어절 식별 번호, 원어절, 형태소 분석 결과로 구성되며 패치 종류는 아래와 같이 총 5개이다.

(1) 어절 변경 및 삭제/추가
- = : 해당 어절을 다음과 같이 변경
- + : 해당 어절 추가
- - : 해당 어절 삭제

(2) 문장 분리 오류
- M : 두 어절 사이 문장 분리 표지를 삭제하고 하나의 문장으로 합침(merge)
- S : 두 어절 사이에서 문장 분리 표지를 추가하여 두 문장으로 분리(split)

어절 관련 패치는 오류로 인해 어절을 변경할 경우, '=' 패치가 적용되며, 어절을 추가하거나 삭제할 때는 '+' 패치나 '-' 패치가 적용된다.

문장 분리 관련 패치는 문장 종결(예:)과 시작 마커(예:)로 잘못 분리된 어절들을 병합하는 'M, 문장 종결과 시작 마커로 분리되지 못한 두 어절을 분리하는 'S' 로 구성된다.

본 논문에서 구축한 패치 파일에는 이와 같은 패치의 종류뿐만 아니라, 세종 형태 분석 말뭉치와 동일한 어절 식별 번호를 함께 기술하기 때문에 사용자들은 어떠한 파일의 어절이 변경되었거나 삭제, 추가되었는지 확인할 수 있다.

4. 적용결과

4.1 오류 유형 정의

본 논문은 21세기 세종계획의 "현대문어 형태분석 말뭉치"를 대상으로 오류를 검출하고 수정을 진행하였다. 말뭉치의 오류 수정은 세종 지침에 따라 형태 중심의 분석을 따르면서 지침에 벗어나는 오류들을 수정하고 가급적 원문을 수정하지 않았다.

원문 자체를 수정한 경우는 의미를 파악할 수 없어 분석이 불가능하거나 오탈자 때문에 의미 없는 문장이 되었을 때이다.

오류의 검출과 수정은 1, 2차로 나누어 진행되었다. 형태 분석 말뭉치의 오류를 발견하고 수정하기 위해 모든 어절을 검증하는 것은 현실적으로 불가능하므로 오류 검출을 위해 1차적으로 세종 코퍼스와 창원대, 울산대 코퍼스를 각각 비교하여 고빈도로 나타나는 수정 사항을 참고하여 오류를 검출하였다. 그리고 2차적으로 원어절과 형태소 분석 어절의 음절 대응관계를 비교하여 오류를 검출하고 수정하였다.

현재까지 이루어진 형태 분석 말뭉치에서 수정한 오류를 유형별로 분류하면 다음과 같다.

(1) 문장 및 어절 분리 오류

하나의 문장임에도 여러 개의 문장으로 분리된 오류, 각각 다른 문장임에도 문장 분리가 이루어지지 않은 오류 유형이 해당한다.

예) 집회(?)를 집회/NNG + (/SS + ?/SF +)/SS +
를/JKO

이루고 이루/VV + 고/EC

(2) 철자 오류

원어절 및 형태 분석 결과에 나타나는 철자 오류이다. 원어절에 나타난 철자 오류의 경우, 사람들이 빈번하게 사용하는 오타(예: 요드, 횡경막)는 수정하지 않고, 코드 변환 과정에서 발생한 오류나, 분석이 불가능하거나 의미 없는 어절이 된 경우에 한해 수정하였다.

예) 밀?수입 밀/NNG + 수입/NNG

형태 분석 결과에 철자 오류가 나타난 경우, 오타는 원어절에 따라 분석 결과를 수정하였다.

예) 웬놈의 왠/MM + 놈/NNB + 의/JKG

(3) 불일치 오류

원어절과 형태 분석 결과의 음절 대응 관계를 비교하였을 때, 원어절에는 나타난 형태지만 형태 분석 결과에는 누락되었거나, 원어절에 나타나지 않은 형태가 형태 분석 결과에 추가되어 나타나는 오류이다.

예) 것"이라며 것/NNB + "/SS + 이/VCP + 며/EC
도쿄의 도쿄/NNP + (/SS + 동경/NNP +)/SS + 의/JKG

(4) 분석 오류

원어절을 엉뚱하게 잘못 분석한 경우로, 분석 표지를 잘못 부착하거나, 하나의 형태소로 분석해야 하는데 잘못 분할한 오류 유형이다.

예) 홍걸씨 홍/NNP + 것/NNB + 이/VCP + ㄹ/ETM + 씨/NNB
소년들에서 소년/NNG + 이/VCP + ㄴ들/EC + 에서/JKB

(5) 형식 오류

세종 형태 분석 말뭉치에서 제시하는 분석 형식을 준수하지 않은 오류로, 형태 분석 결과를 기술하는 과정에서 '+' 또는 '/' 를 누락하였거나, 태그를 이중으로 기술하는 오류 유형이 해당한다.

예) 국제자유도시로 국제/NNG + 자유/NNG + 도시/NNG 로/JKB
청계고가도로 청계/NNP+고가/NNG+도로/NNG/NNG

4.2 패치 적용 결과

본 논문에서는 총 1,015만여 개의 어절을 대상으로, 약 31만여 개의 오류를 수정하였고, 패치 적용 결과는 표1과 같다.

'M' 패치와 'S' 패치는 문장 및 어절 분리 오류에 적용된다. 'M' 패치는 하나의 문장으로 묶여야 할 어절이 다른 문장으로 분리된 오류에 적용된다. 표1의 예처럼 한 문장 내에서 목적어-술어 관계를 이루는 어절(예: 곤욕을 치렀다)들이 문장 시작 (예:) 및 종결 마커(예:)로 잘못 분리된 오류가 있다. 이러한 오류는 'M'

패치 적용 후, 하나의 문장으로 수정된다.

표1. 패치 적용 결과

패치 종류	세종 말뭉치	수정된 말뭉치
M	BTAB0170-00002897 곤욕(?)을 곤욕 /NNG + (/SS + ?/SF +)/SS + 을/JKO BTAB0170-00002898 치렀다. 치르/VV + 었 /EP + 다/EF + ./SF	BTAB0170-00002897 곤욕(?)을 곤욕 /NNG + (/SS + ?/SF +)/SS + 을/JKO BTAB0170-00002898 치렀다. 치르/VV + 었 /EP + 다/EF + ./SF
S	BTBD0236-00089764 오세요 오/VV + 시/EP + 어요/EC BTBD0236-00089765 !양경숙 !/SF + 양경숙 /NNP	BTBD0236-00089764 오세요! 오/VV + 시/EP + 어요/EC + !/SF BTBD0236-00089765 양경숙 양경숙/NNP
+		BTAA0155-00051766-1 지난주 지난주/NNG
-	BTBZ0074-00028385 는 늘/VV + ㄴ /ETM	
=	BTGO0345-00002470 가사로부터 가사 /NNG + 부터/JX	BTGO0345-00002470 가사로부터 가사 /NNG + 로부터/JKB

이와 반대로 ‘S’ 패치는 서로 다른 문장으로 분리되어야 하는 어절임에도 하나의 문장으로 분석된 오류에 적용된다. 적용 후에는 표1처럼 두 어절 사이에 문장 분리 표지가 추가되어 각각의 문장으로 분리된다.

‘+’ 패치와 ‘-’ 패치는 오류로 인해 해당 어절을 추가하거나 삭제한 경우에 적용된다.

‘+’ 패치는 세종 형태 분석 말뭉치에 없던 새로운 어절이 생성되었을 때 적용되는데 새로운 어절은 문장 및 어절 분리 오류를 수정하는 과정에서 생성된다. 예를 들어서, 문장 및 어절 분리 오류가 나타난 다어절은 각각의 어절로 분리되는 과정에서 새로 어절이 생성되기 때문에 ‘+’ 패치가 적용된다. 이러한 어절은 표 1의 예처럼 ‘-1’ 이라는 새로운 어절 번호를 할당받는다.

‘-’ 패치는 단독 어절로 쓰일 수 없는 형태가 단독 어절로 나타난 경우, 그 어절을 삭제할 때 적용된다. 표1의 예처럼, ‘-’ 패치가 적용된 ‘는’ 은 원문을 살펴보면, 어미 ‘라는'의 일부이다. 따라서 ‘-’ 패치 적용 결과, ‘는’ 이 단독 어절로 나타난 어절은 삭제되고, ‘는’ 은 선행 어절에 결합되어 분석된다.

‘=’ 패치는 철자 오류, 불일치 오류, 분석 오류, 형식 오류에 적용된다. 표1의 예는 원어절을 구성하는 형태소가 형태 분석 결과에 누락된 오류로, ‘=’ 패치 적용 결과 누락된 형태소가 형태소 분석 결과에 생성되었다.

5. 결론

본 논문에서는 가급적 원문을 훼손하지 않고, 원래의 세종 지침에 따라 형태 중심의 분석을 따르면서 지침을 벗어나는 오류들을 수정하였다. 수정된 오류 유형으로는 문장 및 어절 분리 오류, 철자 오류, 불일치 오류, 분석 오류, 형식 오류가 있었다.

수정된 내용이 반영된 패치 파일을 패치 적용 및 생성 스크립트와 함께 배포하여, 패치 파일을 배포하는 커미터와 사용자가 서로 협업하여 지속적으로 세종 말뭉치의 오류를 개선할 수 있는 방법을 제안하였다.

향후에는 동일한 어절을 일관성 없게 분석한 오류 유형을 수정할 것이며, 말뭉치 대상을 확대해 문어체 말뭉치뿐 아니라 구어체 말뭉치에 대해서도 지속적으로 오류를 검증하고 수정해 나갈 것이다.

참고문헌

[1] 김재훈, 서형원, 전길호, 최명길, “세종말뭉치의 오류 수정 방법", 한국마린엔지니어링학회 학술대

회 논문집, pp.435-436, 2010.
[2] 최명길, 서형원, 권홍석, 김재훈,“한국어 품사 부착 말뭉치의 오류 검출 및 수정", 한국마린엔지니어링학회지, 제 37 권, 제 2 호, pp.227-235, 2013.
[3] 홍진표, 차정원,“품사 태거와 빈도 정보를 활용한 세종 형태 분석 말뭉치 오류 수정”, 정보과학회논문지: 소프트웨어 및 응용, 제 40 권, 제 7 호. pp.417-428, 2013.
[4] 김일환, 이도길, 강범모, “SJ-RIKS Corpus : 세종 형태의미 분석 코퍼스를 넘어서”, 민족문화연구, 제 52 권, pp.373-403, 2010.
[5] 이미경, 정한민, 성원경, 박동인. “품사 표지 부착 말뭉치 검증", 한국정보과학회 언어공학연구회 학술발표 논문집, pp.145-150, 2005.