(제 17 회 한글 및 한국어 정보처리 학술대회)

품사 표지 부착 말뭉치 검증

이미경, 정한민, 성원경, 박동인
한국과학기술정보연구원
{jerryis, jhm, wksung, dipark}@kisti.re.kr

Verification of POS tagged Corpus

Mikyoung Lee, Hanmin Jung, Won-Kyung Sung, Dong-In Park
Korea Institute of Science and Technology Information

요 약

본 논문에서는 자연어 처리 연구에서 이용되는 품사 표지 부착 말뭉치의 오류 검증 방안에 대해 제안한다. 현재까지의 품사 표지 부착 말뭉치들은 정제보다는 구축에 중점을 두고 있으며, 기존의 오류 검출과 정정 방안에 관련된 연구들은 기 구축된 말뭉치를 대상으로 한 것이 아니라, 품사 표지 부착 시스템의 후 처리에 집중하고 있다. 형태소 분석기나 품사 표지 부착 시스템의 학습에 이용되는 품사 표지 부착 말뭉치가 오류 검증 단계를 거친다면 이 시스템들은 좀 더 높은 신뢰성을 가지게 될 것이다. 본 논문에서는 품사 표지 부착 말뭉치 검증을 위한 어절 분할 오류, 철자 오류, 표지 부착 오류, 형식 오류, 일관성 오류의 5가지 오류 유형과 검증 방안을 제안한다. 또한 제안한 방법에 따라 세종 계획의 형태소 분석 말뭉치의 오류를 검증해 보았으며, 그 결과 말뭉치 오류 정제가 말뭉치의 신뢰도를 향상시킬 수 있음을 보인다.

1. 서 론

자연어 처리 연구에서 어절을 분석하여 각 형태소에 품사 표지를 부착한 품사 표지 부착 말뭉치는 형태소 분석기와 태거의 학습에 주로 이용되고 있다. 형태소 분석기의 결과물은 높은 정확성을 요구하는 자동 번역에 있어서 번역 시스템의 전체적인 성능을 좌우하기 때문에 형태소 분석기의 성능 향상은 자연어처리 연구에서 중요한 이슈로 대두되고 있다. 형태소 분석기의 오류는 구문 분석 시스템, 정보 검색 시스템, 기계 번역 시스템 등에 그대로 반영되기 때문에 또 다른 오류를 야기시킨다. 따라서 오류 문제를 다소 완화시키기 위해서 많은 연구자들은 여러 가지 방법을 이용하여 품사 표지 부착 시스템의 성능을 향상시키고 있다[1].

기존의 품사 표지 부착 말뭉치는 대부분 수작업을 통해 구축되었고 여러 작업자들의 협업에 의해 구축하였다. 그리고 다년간의 작업에 따른 지침의 변경이나 작업 지침의 미비 등으로 인해 상당수 오류가 존재할 수 있다. 그리고 각기 다른 품사 태그셋간의 변환으로 인한 오류 또한 다수 존재한다. 따라서 말뭉치의 오류 수정 작업이 계속적으로 수행되어 정확성 높은 품사 표지 부착 말뭉치를 구축해야 한다. 하지만, 아직까지 국내에서는 이미 구축된 말뭉치에 대한 검증이나 보완에 대한 연구가 미비한 실정이다.

따라서, 본 논문에서는 신뢰도가 높은 품사 표지 부착 말뭉치를 구축하기 위해서 이미 구축된 말뭉치에서 발견될 수 있는 오류의 유형들을 정의하고 유형에 따른 검증 방안을 제안하여 협업으로 말뭉치를 구축할 때 구축 지침이 보완될 수 있도록 피드백을 제공하고자 한다. 본 논문의 구성은 다음과 같다. 2장에서 기존의 오류 처리 방안에 대해서 살펴보고 3장에서 품사 표지 부착 말뭉치에서 나타날 수 있는 오류들의 유형을 정의한다. 4장에서는 각 오류 별 검증 방안에 대해 제안하고 5장에서는 오류 검증 방안에 따른 실험 및 평가를 수행한다.

2. 기존 연구

기존의 오류 처리 연구는 자동 표지 부착 시스템에서의 후 처리를 위한 것이 대부분이었다. [2]의 논문에서는 형태소 분석 대상 어절의 좌우 어절 내의 대표 형태소 어휘 문맥 정보에 기반한 형태소 오류 정정 방안을 제안하였다. [3]에서는 품사 표지

부착 시스템에서 신뢰도 측정 방법에 대해 기술하였다. 이 논문에서는 품사 표지 부착 결과의 오류 정도로 신뢰도 측정 방법을 제시하였다. 품사 표지 부착에서 오류 확률을 구하는 방법으로는 품사 표지 부착 시스템을 이용하여 학습 말뭉치를 표지 부착하고 그 결과로부터 오류 확률을 구하게 된다.

기존의 연구들은 품사 표지 부착 시스템을 이용하여 말뭉치를 표지 부착할 때 발생하는 오류들의 처리에 대해 주로 다루었다. 그러나, 본 논문에서는 품사 표지 부착 시스템이 학습할 품사 표지 부착 말뭉치의 오류를 검증하여 말뭉치의 신뢰도를 높이고자 한다. 신뢰도 높은 말뭉치를 학습 데이터로 이용한 형태소 분석기나 품사 표지 부착 시스템은 기존 오류가 포함된 말뭉치를 학습 했을 때보다 높은 성능을 나타낸다. 현재까지 기 구축된 말뭉치의 오류 검증과 정정에 대한 연구는 알려진 바가 없으므로 논문에서 제안한 오류 검증 방식을 이용하여 기 구축된 품사 표지 부착 말뭉치들을 정제한다면, 정제된 말뭉치를 학습하는 여러 시스템들의 정확성을 높일 수 있을 것이다.

3. 오류 유형 정의

본 장에서는 품사 표지 부착 말뭉치에서 나타나는 오류의 유형들에 대해 정의해본다. 철자, 어절 분할, 형식, 표지 부착, 일관성 등의 오류에 대한 검증을 처리 하기 위한 유형에 대해 정의한다. 그리고 일관성 있는 전산 처리가 가능할 수 있도록 형식적으로 통일되지 않는 부분도 오류로 분리한다. 오류 유형 목록은 표 1과 같다.

표 1. 오류 유형 목록

유형	종류
어절 분할 오류	다어절 오류 형태소 분할 오류
철자 오류	어절 철자 오류 형태소 철자 오류 품사 태그 사용 오류
표지 부착 오류	어절 분할 특정 형태소 누락 특정 형태소 추가
형식 오류	형태소 품사 쌍 오류 형태소 분석 결과
일관성 오류	문맥 의존 오류 문맥 비의존 오류

3.1 어절 분할 오류

어절 분할 오류의 유형으로는 여러 어절이 하나의 어절로 묶여 표지 부착된 오류인 다어절 오류와 하나의 어절이 과분할되거나 잘못된 위치에서 분할되어 생기는 형태소 분할 오류를 들 수 있다.

예) 다어절 오류 : "있고,대반도가", "것의중심이"
형태소 분할 오류 : "부터"

위의 예를 살펴보면 "있고,대반도가"라는 어절에는 ','가 포함되어있는데, 이 기호는 어절의 끝을 나타내는 구분자이다. 그리고 "것의중심의"의 경우에 '의'는 조사이기 때문에 한 어절 내에서는 뒷부분에 명사가 위치할 수 없으므로 다어절 오류에 해당된다.

형태소 분할 오류인 "부터"의 경우 '부터'라는 것은 명사 뒤에 나오는 조사로 단독으로 어절을 구성할 수 없다. 따라서 잘못된 형태소 분할 오류가 된다.

3.2 철자 오류

철자의 오류 유형으로는 어절이나 형태소 내에서 철자 오류가 발생하여 의미 없는 형태를 가지게 되는 문맥 비의존 오류와 어절이나 형태소 내에서 철자 오류가 발생했지만, 의미 있는 형태로 우연히 변경된 문맥 의존 오류로 나눌 수 있다.

예) 문맥 비의존 오류 : "교욱인", "돌어가면"
문맥 의존 오류 : "새명의", "새대의"

"교욱인"의 경우는 문맥에 상관없이 형태소 분석 시 미등록어나 오류로 처리되게 된다. 앞뒤 문맥을 고려했을 경우 "교육인"이 맞는 어절임을 알 수 있다. "새명의"의 경우에는 형태소 분석이 가능하기 때문에 문맥을 보지 않고는 철자 오류로 인식하기 힘들다. 이런 경우에는 앞뒤 문맥을 파악하여 "세명의"가 철자 오류라는 것을 파악해야 한다.

철자 오류의 유형 중의 하나인 품사 태그 사용 오류는 구축 지침에서 정의되지 않은 품사 태그를 사용한 경우 오류로 처리한다.

3.3 표지 부착 오류

표지 부착 오류는 크게 어절 분할, 형태소 누락, 형태소 추가의 3가지 오류로 나눌 수 있다.
어절 분할 오류는 어절에 대한 잘못된 해석으로 인해 발생하는 분할 및 표지 부착 오류이다.

예) 같은해 → 같/VA+은/ETM+하/XSV+아/EC

위의 예의 "같은해"라는 어절에서 "같은"은 하나의 형태소로 해석되어야 하므로 잘못된 분할 오류이다.

표지 부착 오류의 또 다른 유형으로 형태소 분석 결과 특정 형태소가 누락되거나 추가된 경우에 나타나는 오류가 있다. 아래의 예는 '라'가 누락된 경우를 보여준다.

예) 태산이라던 태산/NNG + 이/VCP + 던/ETM

3.4 형식 오류

형식 오류는 품사 표지 부착 시 형식적으로 잘못된 오류를 지적한다. 아래의 예와 같이 형태소와 품사 쌍에서 형태소나 품사가 누락된 오류, 어절과 이에 대응하는 "형태소-품사" 쌍이 순서적으로 뒤섞이거나 생략된 경우, 분리자 사용의 오류로 인해 자동적으로 처리될 수 없는 오류 유형이 여기에 속한다.

예) 대학생들의 대학+ 생/NNG+ 들/XSN+ 의/JKG
대학생들의 대학/NNG+ 생+ 들/XSN+ 의/JKC
대학/NNG+ 생/NNG+ 들/XSN+ 의/JKC

형식 오류를 파악하기 위해서는 말뭉치에서 어절과 형태소-품사 쌍을 구분하여 형태소와 품사의 쌍이 맞게 표현되었는지를 검증한다.

3.5 일관성 오류

품사 표지 부착 말뭉치 내에서 일정한 출현 빈도를 보이는 어절들에 대해서 형태소 분석 결과가 상이하거나 동일한 경우를 비교하여 일관성 오류를 찾아낸다. 어절들의 문맥을 검토하여 문맥에 의존적인 형태소의 경우에는 문맥을 고려한 일관성 오류를 파악하고, 문맥 비의존적인 형태소인 경우에는 다른 형태로 분석된 경우를 찾아내어 일관성 오류를 파악한다.

4. 오류별 검증 방안

오류별 검증 방안 절차는 그림 1과 같다. 품사 표지 부착 말뭉치들은 형식 오류를 검사하게 된다. 형식 오류 검사를 거친 어절들을 이용하여 어절 분할 오류 검사와 철자 오류 검사, 표지 부착 오류 검사, 일관성 오류 검사의 과정을 거치게 된다.

어절 분할 오류를 검사하기 위해서 한글 코드 변환 및 코드 범위를 검사하고, 품사열 규칙을 적용한다. 형태소 분석 및 미등록어 검사, 형태소 결합 후 어절 비교, 품사 집합을 검증하여 철자 오류 검사를 하고 표지 부착 오류를 검사한다. 일관성 오류 검사를 위해서는 연도별 출현 어절 목록을 추출한 후, 이를 이용하여 연도별 일관성 정보와 문맥 정보를 파악하여 일관성을 검사하게 된다.

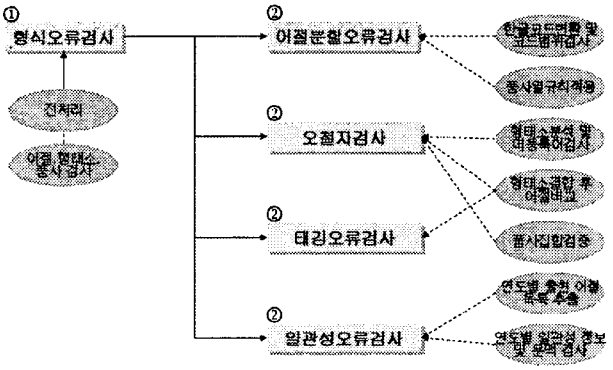

그림 1. 오류 검증 방안 절차

그림 1의 절차에 따라 오류 검사를 수행하면, 표 1에서 나타난 오류별 유형들이 검증된다.

4.1 형식 오류 검사

(1) 전처리

말뭉치들은 오류 검증에 앞서 전처리 과정을 거친다. 형태소 분석 말뭉치의 경우 구분자로 스페이스, 탭, "+", "[]", XML태그 등의 다양한 형식을 사용하고 있다. 따라서 어절 및 분석 결과의 순서를 유지하는 범위 내에서 일관된 형식으로 정규화 시키고, 품사 표지 부착 말뭉치에서 어절과 형태소 목록을 제외한 부가적인 데이터는 삭제시킨다.

(2) 어절, 형태소, 품사 검사

어절과 형태소, 품사를 분리하여 어절의 위치에 어절이 존재하고 형태소-품사 위치에 형태소-품사가 존재하는지를 체크하여 어절이나 형태소-품사 쌍이 생략되거나 순서적으로 뒤섞인 경우의 오류, 분리자 사용으로 인한 오류를 검사한다. 그리고 형태소-품사 쌍을 분리하여 형태소나 품사가 생략된 경우가 없는지를 파악한다.

4.2 어절 분할 오류 검사

(1) 한글 코드 변환 및 코드 범위 검사

한글에서 만든 말뭉치의 경우 한글 전용 코드 값을 이용하여 작성한 곳이 있는지를 검사해야 한다. 특정 코드의 범위를 지정하여 그 범위에 해당되지

않는 한글 전용의 코드 값을 사용했다면 이를 오류로 처리한다.

(2) 품사열 규칙 적용

한 어절 내에서 존재할 수 없는 품사 간 결합을 검사하기 위해 품사열을 규칙화하여 적용한다. 품사열과 매칭되는 어절 분석 결과는 어절 분할 오류로 처리된다.

4.3 철자 오류 검사

(1) 형태소 분석 및 미등록어 검사

품사 표지 부착 말뭉치의 어절에 대해 다른 형태소 분석기를 이용하여 재분석한다. 어절 자체에서 철자 오류가 발생하는 경우에는 미등록어로 분석될 가능성이 크므로 이들에 대해 수작업으로 검증한다.

(2) 형태소 결합 후 어절 비교

분리된 형태소를 결합하여 원래의 어절이 되는지를 확인한다. 같은 어절이 나오지 않는 경우는 형태소 분석 시 오류가 발생한 것으로 처리한다.

(3) 품사 집합 검증

구축 지침에 정의되지 않은 품사를 사용한 경우를 파악한다. 형태소-품사 쌍에서 품사의 위치에 오는 데이터를 검사하여 정의된 품사가 아닐 경우에 품사 태그 사용 오류로 처리한다.

4.4 표지 부착 오류 검사

표지 부착 오류 검사 시, 철자 오류 검사의 형태소 결합 후 어절 비교 방법을 이용한다.

4.5 일관성 오류 검사

(1) 연도별 출현 어절 목록 추출

말뭉치 내에 출현하는 모든 어절들에서 연도별 한번 이상 출현하는 어절들에 대한 목록을 추출한다. 이는 같은 어절이 연도에 따라 다른 결과를 나타날 경우의 일관성 정보를 파악하여 일관성 오류를 판단하기 위한 사전 작업이다.
표 2는 말뭉치에서 추출한 출현 어절 목록의 일부를 보여준다.

표 2. 말뭉치 출현 어절 목록 예시

출현 어절	출현 빈도수
북한이	23
아니에요	15
소개하는	13
않나?	28
북한에	21
얘긴데,	17
작용을	12
으흠.	29
여보세요	7
부분은,	13

(2) 연도별 일관성 정보 및 문맥 검사

일관성 정보를 추출하기 위해서 (1)에서 추출한 어절 목록을 이용하여 전체 말뭉치에 포함된 어절의 형태소 분석 정보를 추출한다. 형태소 분석 정보는 각 어절이 말뭉치에서 일관성 있게 사용되었는지를 비교하여 일관성 오류를 검사 위해 사용된다. 일관성 오류를 파악하기 위해서 어절의 문맥을 참고 한다.
일관성 오류의 후보로 파악된 어절이 문맥에 의존된 경우인지 의존되지 않은 경우인지를 파악하기 위해 해당 어절이 포함된 문장의 문맥 정보가 필요하다. 따라서 말뭉치에서 일관성 오류의 후보 목록의 어절과 함께 그 어절 뒤에 나오는 다섯 어절들을 추출하여 문맥 정보로 이용하고, 이 정보를 참고하여 문맥 일관성 오류를 검사한다.

5. 실 험

본 논문에서 제안한 방법을 이용하여 세종계획 형태소 분석 말뭉치에 나타난 오류들을 검증해보았다. 그 결과 유형별로 아래와 같은 오류들이 나타났다. 유형별 오류 목록으로 추출된 데이터들은 말뭉치에서 수정 과정을 거쳐 정제될 것이다.

(1-1) 다어절 오류

만났다.이날 만나/VV + 았/EP + 다/EF + ./SF + 이날/NNG
걱정이다.50%에 걱정/NNG + 이/VCP + 다/EF + ./SF + 50/SN + %/SW + 에/JKB

(1-2) 형태소 분할 오류

의 의/ukall
으로 으/uknc + 로/j

(2-1) 어절 철자 오류

붙는다든다 붙/pv + 는/ef + 다/ma + 들/pv + ㄴ다/ef
내로라하는 내/pv + 로라/ef + 하/ma + 는/j

일찌기 일찌기/MAG
교육인 교육/NNG+ 이/VCP+ ㄴ/ETM

(2-2) 형태소 철자 오류

구경하러 구경/NNG + 하/XSV + 어/EC

(2-3) 품사태그 사용 오류

당신은 당신은/NNP1
막. 막/MAG.+/SF

(3-1) 어절 분할 오류

같은해 같/VA+ 은/ETM+ 하/XSV+ 아/EC

(3-2) 특정 형태소 누락

태산이라던 태산/NNG + 이/VCP + 던/ETM

(4-1) 형태소-품사쌍 누락 오류

대학생들의 대학+ 생/NNG+ 들/XSN+ 의/JKG
그거여 그거/NP+ 이 ㅕ/EC

(4-2) 형태소 분석 결과 오류

대략+ 대략/NNG
경우/NNG+ 는/JX+ ,/SP
뭔가

(5-1) 일관성 오류

가능성을 가능/XR 성/XSN 을/JKO
가능성을 가능/NNG 성/XSN 을/JKO
가능성을 가능/NNG 성/XSN 을/JKO
가능성을 가능성/NNG 을/JKO
가능성을 가능성/NNG 을/JKO
가능성을 가능/NNG 성/XSN 을/JKO
가능성을 가능성/NNG 을/JKO
가능성을 가능/NNG 성/XSN 을/JKO

가능하게 가능/XR 하/XSA 게/EC
가능하게 가능/NNG 하/XSA 게/EC
가능하게 가능하/VV 게/EC
가능하게 가능/NNG 하/XSV 게/EC
가능하게 가능/NNG 하/XSA 게/EC

아래의 표 3은 세종계획 구어 전사 형태소 분석 말뭉치에 출현한 형태소들의 빈도들을 나타낸 것이다.

표 3. 구어 전사 형태소 출현 빈도 분석

출현빈도	Lexical coverage
1번	40.3%
2번	15.1%
3번	8.1%
4번	5.2%
5번	3.7%
6번	2.6%
7번	2.1%
8번	1.8%
9번	1.4%
10번	1.25%
10번 이상	18.45%

표 3에서 나타난 빈도에서 보듯이 총 125만 형태소들 중에서 출현 빈도가 1번인 형태소들이 전체의 40.3%를 구성한다. 단 한번만 출현하는 형태소들은 오류 후보로 의심해야 한다.

출현 빈도 별로 100개의 형태소를 샘플 조사하여 오류 검사를 해보았다. 출현빈도가 1번만 나오는 형태소는 24%의 오류 확률을 보인다. 5번 출현하는 형태소들의 경우에는 오류 확률이 11%이고, 10번 이상 출현하는 형태소들의 경우에는 4%의 오류 확률을 보인다.

따라서, 자원이 허락하는 범위 내에서 출현 빈도가 낮은 형태소들에 대해 우선적으로 오류를 검사하는 것이 바람직하다.

6. 결 론

본 논문에서는 품사 표지 부착 말뭉치를 검증하기 위한 방법으로 말뭉치 내에서 발견되는 오류들의 유형을 파악하고, 오류 유형별 검증 방안을 제안하였다. 기존의 오류 정정에 관련된 연구들은 품사 표지 부착 시스템에서나 형태소 분석기를 이용한 결과로 나타나는 오류들에 대한 연구이다. 그러나, 본 논문에서 제안하는 것은 수작업에 의해 구축된 품사 표지 부착 말뭉치의 신뢰도를 높이기 위한 방안이다. 이 방법을 이용하여 말뭉치들을 검증, 정제하면 말뭉치를 이용하여 학습하는 시스템들의 형태소 분석의 정확성을 높이게 되어 기존의 말뭉

치를 이용한 시스템보다 높은 신뢰성을 나타낸다.

본 논문에서 제안한 오류 검증 방법은 어절 분할 오류 검사, 철자 오류 검사, 표지 부착 오류 검사, 형식 오류 검사, 일관성 오류 검사의 5가지 유형으로 나누며, 이런 검사들을 통해 추출된 오류 목록들은 말뭉치 오류 정제의 데이터로 제공된다. 따라서 오류 정제를 거친 말뭉치는 기존의 말뭉치에 비해 신뢰성이 향상된다.

향후, 본 논문에서 제안한 오류 검증 방법을 이용하여 세종 계획의 기초, 특수 분과의 모든 형태소 분석 말뭉치의 오류를 검증하고 정제해 나갈 것이다.

참고 문헌

[1] 김덕봉, 한국어 어절의 철자변환 현상 분류와 인식 방법, 정보과학회논문지, 제30권, 제5호, 2003.

[2] 김영길, 양성일, 홍문표, 박상규, 형태소 어휘 문맥에 기반한 태깅 오류 정정, 제15회 한글 및 한국어정보처리 학술대회 논문집, 2003.

[3] 김재훈, 품사 태깅 시스템의 신뢰도 측정, 정보처리학회논문지B, 제8-B권, 제4호, 2001.