한국어 품사 부착 말뭉치의 오류 검출 및 수정

요약: 품사 부착 말뭉치의 품질은 품사 부착기를 개발하는데 있어서 매우 중요한 역할을 수행한다. 그러나 세종 말뭉치를 비롯하여 한국에서 구축된 많은 품사 부착 말뭉치들은 여전히 다양한 형태의 오류를 포함하고 있다. 이런 오류들을 살펴보면 품사 부착 오류는 물론이고 철자 오류, 문자의 삽입 및 삭제 등 매우 다양하다. 본 보고서에서는 오류 패턴을 이용하여 품사 부착 오류를 검출하는 방법을 제시하고자 한다.

1.서론

-말뭉치

자연언어처리 분야에서는 대량의 학습 자료를 사 용해서 보다 쉽고, 지능적이며, 빠르게 시스템을 개발하고 있다. 자연어 처리의 가장 밑단, 즉 기단부 역할을 하는 게 한글 말뭉치다. 이를 개발용어로는 코퍼스 (Corpus)라고 부른다.

한글 말뭉치는 쉽게 말하면 한글 어휘와 어휘 특성의 저장소다. 디지털 사전에 비유될 수 있지만 그보단 복잡하다. 일반적으로 문서를 디지털화한 뒤 해당 문서에 사용된 모든 어휘를 문장, 어절, 형태소별로 추려내고 각각에 특성을 부여한다. 뿐만 아니라 동사의 경우 다양한 활용형태, 예를 들어 감사합니다, 감사하다, 감사해요? 등과 결과치들을 데이터베이스화 한다. 서울대학교 컴퓨터공학과 교수등 은 “말뭉치는 분석 결과가 태그로 부착된 파일들의 집합” 이라고 했다.

즉, 자연언어처리 분야에서 대량의 학습 자료를 일반적으로 언어정보 부착 말뭉치라고 한다. 한국어 정보처리를 위해도 다양한 말뭉치가 구축되었으며, 이 중에 한국어 정보처리 연구자가 쉽게 그리 고 널리 이용할 수 있는 말뭉치가 세종 말뭉치이다. 세종 말뭉치는 원시 말뭉치, 형태분석 말뭉치, 구문 분석 말뭉치 등을 포함하고 있다.

-말뭉치(코퍼스)가 개발되어야 하는 이유

한국 말뭉치가 없다면 자연어처리 기술로 인공지능 시장에 진입하고자 하는 스타트업이 부담이 된다. 당장 IT 대기업과 이 분야에서 간극을 좁히기가 어렵다. 네이버와 같은 IT 대기업들은 이미 십수년 간 수집한 뉴스 문서, 기타 블로그와 전문 자료들을 광범위하게 확보해 정교한 말뭉치를 자체적으로 구축해왔다. 오랜 공을 들여 구축한 말뭉치이기에 공개하기도 어렵다. 반면 스타트업이 이에 준하는 수준으로 말뭉치를 구축하기는 비용 문제로 사실상 불가능하다. 말뭉치라는 디지털 커먼스가 공공에서 제공돼야 할 이유이기도 하다.

따라서 국립국어원은 1998 년부터 2007 년까지 150 억원의 예산을 투입해 이 프로젝트를 완료했다. 여기엔 세종 말뭉치라는 한글 말뭉치 사업이 포함돼있었다. 그것도 무려 2 억 어절에 해당하는 방대한 구축 작업이었다. 이는 영국이나 미국, 일본에 뒤지지 않는 규모로 평가됐다.

그렇게 탄생한 세종 말뭉치는 자연어 처리 분야에서 다양하게 활용되고 있다. 대표적으로는 서울대에서 개발한 꼬꼬마 세종 말뭉치 활용 시스템이다. 세종 말뭉치를 데이터베이스에 저장해 웹 프로그램 형태로 구현한 사례다. 말뭉치 통계 정보 조회, 말뭉치 검색, 한국어 쓰기 학습의 세 가지의 기능을 갖추고 있다.

일본어 형태소 분석기 엔진에 세종 말뭉치를 학습해 한국형 형태소 분석기로 탄생한 사례도 있다. 은전한닢 프로젝트에 따라 개발된 'mecab-ko' 형태소 분석기다. 'mecab-ko'는 세종 말뭉치에서 2 만여 문장을 가져와 분석 엔진을 학습시켰다.

-세종 코퍼스의 현재 상황

자연어 처리에 광범위하게 활용되던 세종 말뭉치는 2007 년 기점으로 업데이트가 사실상 중단됐다. 더 이상 제 2 의 21 세기 세종 프로젝트가 진행되지 않고 있는 탓이다. 10 년이 지나는 동안 수많은 인터넷 신조어가 탄생하고 있고 광범위하게 일상에 활용되고 있지만 세종 말뭉치에는 반영이 되지 않는다. 비교적 변화가 적은 언어 유형은 상관이 없지만 구어와 같은 메신저 용어 등은 세종 말뭉치에서 품사 구별조차 하지 못하는 실정이다.

세종 코퍼스는 많은 학습량 만큼 오류도 많았다. 기계가 인간의 언어를 인식하기 위해서는 인간이 수천년 간 사용한 문자 언어를 통달해야 한다. 특히나 한국어를 이해하려면 한글이라는 언어의 역사적 굴곡까지도 파악해낼 수 있어야 한다. 이를테면, 현대 한글 뿐만 아니라 고어, 심지어 사자성어와 같은 자주 쓰는 한자어도 인식할 수 있어야 한다.

방언도 예외일 순 없다. 로봇의 모양을 한 인공지능 로봇이 제주에 거주하는 고령층의 대화를 인식하기 위해서는 제주어에 대한 데이터베이스가 갖춰져야 한다. 언어의 역사와 사회성에 대한 방대한 학습이 전제돼야 한다는 얘기다.

뿐만 아니라, 문법에서의 오류가 가장 크다.

특히 세종 형태 분석 말뭉치에는 형태소에 품사가 잘못 부착되었거나, 문장 내에서 단어가 잘못 분리된 경우, 그리고 불필요 한 단어가 삽입된 경우나 단어가 삭제되는 경우 등의 오류를 포함하고 있다. 이러한 오류들이 포함된 말뭉치를 학습 자료로 사용할 경우 품사 부착기 등과 같은 자연언어처리 시스템의 좋은 성능을 기대할 수 없다.

언어는 끊임없이 변화한다. 신조어 탄생하는가 하면 일부 어휘는 사멸하기도 한다. 예전 단어에 새로운 의미가 부여되는가 하면 단어와 단어의 새로운 결합이 나타나 전혀 다른 용례로 쓰이기도 한다. 말뭉치는 이처럼 역동적인 언어의 변동에 대응할 수 있어야 한다. 자연어 처리를 기반으로 하는 인공지능 기술은 역동성에 대응하는 말뭉치의 힘에 의존할 수밖에 없다.

문제는 막대한 비용이 투입되는 말뭉치를 누가 구축할 것 인가이다. 말뭉치 구축은 따지면 기초연구에 해당한다. 비용은 많이 들지만 곧장 수익이 나지 않는 사업이다. 그것의 공적 가치는 무한하지만 그 자체의 상업적 가치는 제한적일 수밖에 없다.

연구주제

이와 같은 문제점을 해결하기 위해서 본 보고서에서 는 품사 부착 말뭉치로부터 오류 유형을 분석한다. 보고서에서 오류 검출 방법으로 형태소 생성에 기반한 오류 패턴을 이용한다.

2.한국어 품사 부착 말뭉치에서 오류 검출

-한국어 품사 부착 말뭉치의 특징

한국어 품사 부착 말뭉치는 영어 품사 부착 말뭉치 와 다르게 어절과 형태소 분석 결과를 함께 저장해야 한다. 어절은 형태소 분석 결과의 형태소 생성 결 과로 볼 수 있다. 그러나 일반적으로 어절에 대한 형태소 분석 결과가 모호하므로 품사 부착 말뭉치에서는 정확한 형태소 분석 결과를 저장하고 있어야 한다. 따라서 대부분의 한국어 품사 부착 말뭉치들은 Figure 1 과 같이 어절과 형태소 분석 결과를 함께 저장하고 있다. Figure 1 에서는 문장을 표시하는 태그이다. 첫 번째 열은 어절 번호이며 말뭉치 내에서 구별되는 번호를 가지고 있다. 두 번째 열은 어절 자체이고 세 번째 줄은 그 어절의 형태소 분석 결과이다. 세 번째 열에서 형태소는 '+'로 구분되며 각 형태소는 형태소 자신과 품사로 구성되어 있다.

-세종 형태분석 말뭉치의 오류 분석(출처: 카카오 정책산업 연구)

세종 형태분석 말뭉치의 3%에 해당하는 450,000 어절에 대하여 오류를 분석하였으며, 그 결과 29,253 개의 오류가 발견되어 대략 6.5%의 오류를 포함하고 있음을 알 수 있었다. Table 1 에서 그 일부를 보여 주고 있다. Table 1 에서 대부분의 오류는 부가적인 설명 없이도 충분히 이해할 수 있다. '한자변환 오류'는 말뭉치를 구축하는 과정에서 한자 정보가 한글로 변환된 형태로 품사 정보가 부착되었다. 또 '다중 품사 부착 오류'는 말뭉치 구축 과정에서 품사 부착기의 출력이 수정되지 않은 상태로 존재하는 오류들이고, '영어 대소문자 오류'는 어절의 영어 철자 와 형태소 분석 결과의 영어 철자가 다를 경우 오류이다.

Figure 1

```
<p>
BTAA0001-00000164 빙수기,        빙수기/NNG + ,/SP
BTAA0001-00000165 사베트기,      사베트기/NNG + ,/SP
BTAA0001-00000166 얼음           얼음/NNG
BTAA0001-00000167 몰통           몰통/NNG
BTAA0001-00000168 등             등/NNB
BTAA0001-00000169 여름           여름/NNG
BTAA0001-00000170 주방           주방/NNG
BTAA0001-00000171 생활           생활/NNG
BTAA0001-00000172 용품이         용품/NNG + 이/JKS
BTAA0001-00000173 인기다.        인기/NNG + 이/VCP + 다/EF + ./SF
</p>
<p>
BTAA0001-00000174 특히           특히/MAG
BTAA0001-00000175 여름철         여름철/NNG
BTAA0001-00000176 더위를         더위/NNG + 를/JKO
BTAA0001-00000177 식히는데       식히/VV + 는데/EC
BTAA0001-00000178 최고인         최고/NNG + 이/VCP난/ETM
BTAA0001-00000179 빙수를         빙수/NNG + 를/JKO
BTAA0001-00000180 만드는         만들/VV + 는/ETM
BTAA0001-00000181 기계는         기계/NNG + 는/JX
BTAA0001-00000182 소비자들이     소비자/NNG + 들/XSN + 이/JKS
BTAA0001-00000183 가장           가장/MAG
BTAA0001-00000184 많이           많이/MAG
BTAA0001-00000185 찾고           찾/VV + 고/EC
BTAA0001-00000186 있는           있/VX + 는/ETM
BTAA0001-00000187 품목.          품목/NNG + ./SF
</p>
```

Table 1: Examples of annotation errors in the Sejong POS tagged corpus

오류 유형	어절의 예	오류 수정의 예시	
		오류	수정
분석	끔찍한	끔직/XR+ 하/XSA+ ㄴ/ETM	끔찍/XR+ 하/XSA+ ㄴ/ETM
한자 변환	다소비(다소비)	다소비(다소비)	다소비(多消費)
특수 문자	㈜	(/SS+ 주/NNG+)/SS	㈜/SS
다중품사 부착	여권에서	여/NNG+ 권/XSN/NNG + 에서/JKB	여/NNG+ 권/NNG + 에서/JKB
어절 철자	혜택을	혜택을	혜택을
띄어쓰기	디자인세계	디자인세계	디자인 세계
형태소 분리	국내의	국/NNG+ 내의/NNG	국내/NNG+ 의/NNG
영어 대소문자	content	Content/SL	content/SL

-형태소 생성에 의한 오류 검출 제안점

한국어 품사 부착 말뭉치에는 어절과 형태소 분석 결과를 모두 포함하고 있다. 먼저 형태소 분석 결과에 포함된 형태소들을 결합하여 어절을 생성하면 여러 개의 어절이 생성된다. 그 중에 하나가 원래의 어절과 다르다면 오류일 가능성 이 매우 높다. 형태소 생성이 목적이 아니므로 직접 형태소를 생성하는 것이 아니라 원래의 어절과 문자열의 차이를 구하고, 그 차이와 주변 문맥이 정당한 형태소 생성이라면 오류로 추정하지 않는다.

Table 2: An example of morphological generation patterns detected through identification of string difference

왼쪽 문맥	불일치	오른쪽 문맥
ㅇㅏㄹㅡㅁㄷㅏ	ㅇㅜ	ㄴ
ㅇㅏㄹㅡㅁㄷㅏ	ㅂ	ㄴ

Table 2 에서 그 예를 보여주고 있다. Table 2 에서 어절 ‘아름다운’과 이에 대한 형태소 분석 결과에 속한 형태소들 ‘아름답+ㄴ’를 자소 단위 로 비교하여 그 차이를 보이고 있다.

여기서 형태소 생성 패턴은 좌우의 한 자소를 형태소 생성 패턴으로 저장한다. 이렇게 정당한 형태소 생성 패턴을 저장하여 품사 부착 말뭉치의 오류를 검출할 수 있다고 생각한다.