NLTK를 활용한 보조용언구 및 조사상당어구 인식

남궁영[1] · 천민아[2] · 박호민[3] · 윤호[4] · 최민석[5] · 김재훈+

Chunking Auxiliary Verb and Particle Equivalent Phrases through NLTK

Young Namgoong[1] · Min-Ah Cheon[2] · Ho-Min Park[3] · Ho Yoon[4] · Minseok Choi[5] · Jae-Hoon Kim+

Abstract: 구묶음은 구문분석의 전처리 단계로서 구문의 모호성을 미리 줄여줌으로써 구문분석의 효율성을 크게 높일 수 있다. 한국어에서는 구묶음 부착 말뭉치가 많이 부족한 실정임으로 본 논문에서는 규칙 기반의 구묶음 방법을 제안한다. 본 논문은 한국어 품사 부착 말뭉치(세종말뭉치)로부터 가능한 구묶음(chunking)을 찾아내고 이들을 일반화하여 정규표현식으로 표현한다. 이렇게 작성된 정규규칙을 이용해서 구묶음을 수행한다. 본 논문에서는 주로 문장의 핵심 구성요소인 술부의 보조용언구와 조사상당어구을 인식하는 정규표현식을 작성하였다. 제안된 방법은 학습말뭉치가 없는 환경에서 매우 유용한 방법이며 이 방법으로 토대로 학습말뭉치를 구축하여 기계 학습 방법으로 구묶음을 수행할 계획이다.

1. 연구 배경

한국어 처리를 연구함에 있어 그 근간이 되는 형태소 분석과 품사 부착 연구는 다년간 다양한 방법으로 행해져 왔다.[1][2] 이를 토대로 구문 분석을 해 나가기 위해 먼저 긴밀하게 연결되어진 문장 구성 성분을 하나의 단위인 말덩이(chunk)[3][4]로 묶게 되는데, 이때 품사 태그(POS tag)를 이용하게 된다. 본 논문은 이러한 품사 태그들을 통해 보조용언구 및 조사상당어구의 구묶음(chunking)을 위한 규칙을 도출하고, 이를 NLTK를 이용하여 정규표현식으로 구현한 뒤, 실제 구묶음에 적용해 보고자 한다. 예를 들면, '선생님에 대한 이야기를 할 수 있었다.'라는 문장의 형태소 분석은 다음과 같다.

(1) 선생님/NNG + 에/JKB
대하/VV + ㄴ/ETM
이야기/NNG + 를/JKS
하/VV + ㄹ/ETM
수/NNB
있/VA + 었/EP + 다/EF + ./SF

여기서 각 품사는 세종품사태그이고, '에/JKB, 대하/VV, ㄴ/ETM'은 조사상당어구와 'ㄹ/ETM, 수/NNB, 있/VA'는 보조용언구가 구묶음 대상이 될 수 있다.

2. NLTK를 활용한 구묶음 인식

이러한 구묶음을 인식하기 위해 자연언어 처리에 유용한 도구인 NLTK(Natural Language Toolkit)[5]를 이용한다. NLTK의 정규표현식을 이용하여 구묶음의 규칙을 작성함으로써 쉽고 간단하게 구묶음을 수행하고, 이를 트리 형식으로 시각화할 수 있다. 세종말뭉치에서 보조용언구와 조사상당어구에 관한 규칙을 찾고, 이를 NLTK의 정규표현식으로 작성하고 형태소가 부착된 문장을 입력으로 받아서 Figure 1과 같은 트리 구조의 구묶음 결과를 얻을 수 있다. Figure 1에서 '에/JKB, 대하/VV, ㄴ/ETM'은 조사상당어구로서 관형격조사(JKG)로 인식하고 'ㄹ/ETM, 수/NNB, 있/VA'는 보조용언구로서 보조용언(VX)로 인식한다.

+ 김재훈(한국해양대학교 컴퓨터공학과), E-mail: jhoon@kmou.ac.kr, Tel: 051)410-4574
1 남궁영 (한국해양대학교 대학원, 컴퓨터공학과)
2 천민아 (한국해양대학교 대학원, 컴퓨터공학과)
3 박호민 (한국해양대학교 대학원, 컴퓨터공학과)
4 윤호 (한국해양대학교 대학원, 컴퓨터공학과)
5 최민석 (한국해양대학교 대학원, 컴퓨터공학과)

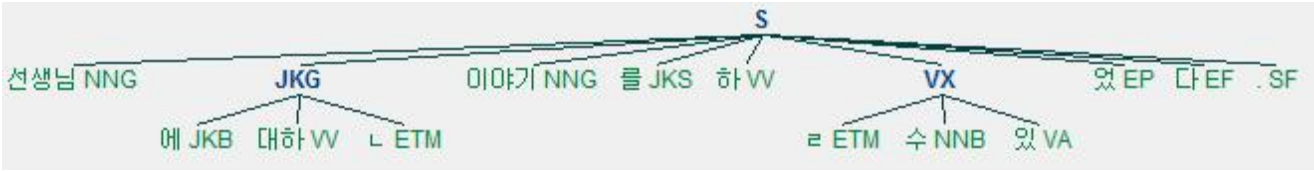

Figure 1. 형태소 분석 (1)의 구묶음 결과

3. 토의

본 논문에서는 보조용언구와 조사상당어구를 인식하는 정규표현식을 기술하였다. 불행히도 세종말뭉치에 구묶음 부착 말뭉치가 없어서 제안된 시스템의 성능은 개관적으로 평가할 수 없었으나 비교적 좋은 성능을 보였다. 모든 규칙 기반 시스템에서 가지고 있는 문제이지만 많은 구묶음 규칙을 작성할수록 참고자료에 과적합(overfitting)되는 현상이 나타난다. 또한, 규칙을 적용하는 순서에 따라 구묶음의 결과가 다소 달라지는 현상을 발견하였다. 이는 규칙을 작성할 때 단순히 형태소의 문법적인 면을 고려할 뿐만 아니라 의미적인 부분도 함께 고려함으로써 언어학적으로 수정·보완할 계획이다. 이외에도 품사 부착에서 발생한 오류가 구묶음 단계까지 영향을 미치는 경우도 있었다. 이러한 오류들을 반영하여 규칙을 보완하고 발전시켜 구묶음을 함으로써 문장의 모호성을 줄이고, 향후 구문 분석을 위한 토대를 마련해 나갈 수 있을 것이다. 오류를 수정해나가는 과정으로 변환기반학습(Brill's Transformation-based learning)[6] 등 다양한 방법을 고려해 볼 수 있다. 또한, 생성한 규칙을 통해 어느 정도 구묶음이 완료되면, 이를 다른 코퍼스에도 적용해 봄으로써 성능분석(evaluation)을 통해 정밀도(precision)와 재현율(recall) 등을 산출해 나갈 것이다.

4. 결론

이 논문에서는 형태소에 부착된 품사 태그를 이용하여 보조용언구와 조사상당어구에 대한 규칙을 찾아내고 이를 이용하여 구묶음을 해 보았다. 찾아낸 규칙만을 문장에 재적용하여 구묶음을 한 경우, 의미적으로 구묶음이 될 수 없는 형태소도 해당 규칙에 영향을 받아 주변의 다른 형태소들과 함께 규칙을 부여한 품사 태그로 묶여지는 예외가 발생하였다. 이는 형태소의 의미를 고려하지 않은 채 규칙만을 적용하여 나타나기도 하였으며, 때로는 구묶음 이전 단계인 코퍼스의 품사 부착에 발생한 오류에서 기인하기도 하였다. 일부는 규칙의 적용 순서에 따라 구묶음되지 않거나, 반대로 의도치 않은 부분이 과도하게 구묶음되는 경우도 발생하였다. 이를 통해, 규칙을 찾아내어 구묶음 할 때 단순히 품사의 종류뿐만 아니라 형태소의 의미적인 부분 등 언어학적인 측면을 함께 고려하여야 함을 알 수 있었다. 정규 표현식을 활용한 규칙만을 이용하여 구문 구조 분석에 필요한 모든 어구들을 구묶음 하는 것은 어려움이 따른다고 보여 진다. 따라서, 한국어 문장에서 형태소를 구묶음할 때에는 규칙을 이용할 뿐만 아니라 통계 기반, 기계 학습 등 다른 대안적인 방법도 함께 고려해서 적용해야 할 것이다.

감사의 글

이 논문은 2017년 정보(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2017M3C4A7068187)

참고문헌

[1] 김재훈, 서정연, 자연언어 처리를 위한 한국어 품사 태그, 한국과학기술원, 인공지능연구센터, CAIR-TR-94-55, 1994년.

[2] 신준철, 옥철영, 기분석 부분 어절 사전을 활용한 한국어 형태소 분석기, 정보과학회논문지:소프트웨어 및 응용, 39권 5호, pp. 415-424, 2012

[3] 김재훈, 한국어 부분 구문분석의 단위와 그 표지, 한국해양대학교 컴퓨터공학과 기술문서 006, 2000

[4] 양재형, 규칙 기반 학습에 의한 한국어의 기반 명사구 인식, 정보과학회논문지:소프트웨어 및 응용, 제27권, 제10호, pp. 1062-1071, 2000

[5] Bird, Steven, Edward Loper, and Ewan Klein, Natural Language Processing with Python. O'Reilly Media Inc., 2009

[6] E. Brill, Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging, Proceedings of the ACL, Vol.21, No.4, pp.543-565 1995.

[7] 김태웅, 조희영, 서형원, 김재훈, 의존명사를 포함하는 보조용언의 구묶음, 제18회 한글 및 한국어 정보처리 학술대회 발표논문집, pp. 279-284, 2006년