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

E-mail: young_ng@kmou.ac.kr

국립 한국해양대학교 KOREA MARITIME AND OCEAN UNIVERSITY

Ⅰ. 초 록

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 **정규표현식**으로 표현한다. 이렇게 작성된 정규규칙을 이용해서 구묶음을 수행한다. 본 논문에서는 주로 문장의 핵심 구성요소인 술부의 **보조용언구**와 **조사상당어구**을 인식하는 정규표현식을 작성하였다. 제안된 방법은 학습말뭉치가 없는 환경에서 매우 유용한 방법이며, 이 방법을 토대로 학습말뭉치를 구축하여 기계 학습 방법으로 구묶음을 수행할 계획이다.

Ⅱ. 연구배경 및 목적

❖ 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 된다.

- 보조용언구 및 조사상당어구의 구묶음(chunking)을 위한 규칙 도출
  ⇒ 품사 태그(POS tag) 이용
- NLTK를 이용하여 정규표현식으로 구현
- 실제 구묶음에 적용

Ⅲ. NLTK를 활용한 구묶음 인식

- 구묶음(Chungking)

  예문) 선생님에 대한 이야기를 할 수 있었다.

(1) 형태소분석

```
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
```

(2) 구묶음 대상

```
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
```

조사상당어구 (에/JKB, 대하/VV, ㄴ/ETM)

보조용언구 (ㄹ/ETM, 수/NNB, 있/VA)

세종말뭉치에서 보조용언구와 조사상당어구에 관한 규칙 찾음

⇓

NLTK의 정규표현식으로 작성

⇓

- Input: 형태소가 부착된 문장
- Output: [Figure 1]과 같은 트리 구조의 구묶음 결과

- NLTK(Natural Language Toolkit)[5]

  NLTK의 정규표현식을 이용하여 구묶음 규칙 작성 ⇒ 쉽고 간단하게 구묶음 수행 및 트리 형식으로 시각화 가능.

Figure . 형태소 분석 (1)의 구묶음 결과

'에/JKB, 대하/VV, ㄴ/ETM': 조사상당어구로서 관형격조사(JKG)로 인식

'ㄹ/ETM, 수/NNB, 있/VA': 보조용언구로서 보조용언(VX)로 인식

Ⅳ. 토 의

- 많은 구묶음 규칙 작성할수록 과적합(overfitting) 현상 나타남
- 규칙을 적용하는 순서에 따라 구묶음 결과가 다소 달라지는 현상 발견
- 품사 부착 단계에서 기인한 오류의 영향

⇓

- **의미적**인 부분도 함께 고려함으로써 언어학적으로 수정 보완 계획
- 모호성을 줄이고 구문분석을 위한 토대 마련
- **성능분석(evaluation)**을 통해 정밀도, 재현율 산출할 계획

Ⅴ. 제 언

이 논문에서는 형태소에 부착된 품사 태그를 이용하여 보조용언구와 조사상당어구에 대한 규칙을 찾아내고 이를 이용하여 구묶음을 해 보았다.

정규표현식을 활용한 규칙만을 이용하여 구문 구조 분석에 필요한 모든 어구들을 구묶음 하는 것은 어려움이 따른다고 보여진다. 따라서, 한국어 문장에서 형태소를 구묶음 할 때에는 규칙을 이용할 뿐만 아니라 통계 기반, 기계 학습 등 다른 대안적인 방법도 함께 고려해서 적용해야 할 것이다.

참 고 문 헌

[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] 성재형, 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