CRF를 이용한 한국어 문장의 복합명사 상당어구 묶음

박별○, 선충녕, 서정연
서강대학교 컴퓨터공학과, 서강대학교 컴퓨터공학과/바이오융합기술협동과정
{pstar0830, wilowisp, seojy}@sogang.ac.kr

Korean Composed Noun Phrase Chunking Using CRF

Byul Park○, Choong-Nyoung Seon, Jung-Yun Seo
Department of Computer Science and Engineering, Sogang University
Department of Computer Science and Interdisciplinary Program of Integrated Biotechnology, Sogang Univeristy

요 약

구분분석은 문장을 분석하여 문장의 구문 구조를 밝히는 작업으로, 문장이 길어질수록 문장의 중의성이 높아져 구문분석 복잡도를 증사시키고 성능이 떨어진다. 구문분석의 복잡도를 감소시키기 위한 방법 중 하나로 구묶음을 하는데 본 논문에서는 하나의 명사처럼 쓰일 수 있는 둘 이상의 연속된 명사, 대명사, 수사, 숫자와 이를 수식하는 관형사, 접두사 및 접미사를 묶어서 복합명사 상당어구라고 정의하고 복합명사 상당어구 인식 시스템을 제안한다. 본 논문은 복합명사 상당어구 인식을 기계학습을 이용한 태그 부착 문제로 간주하였다. 문장 내 띄어쓰기, 어절의 어휘 정보, 어절 내 형태소들의 품사 정보와 품사-어휘 정보를 함께 자질로 사용하였다. 실험을 위하여 세종 구문분석 말뭉치 7만여 문장을 학습과 평가에 사용했으며, 실험결과는 95.97%의 정확률과 95.11%의 재현율, 95.54%의 F_1-평가치를 보였고, 구문분석의 전처리로써 사용하였을 때 구문분석의 성능과 속도가 향상됨을 보였다.

주제어: 기계학습, 복합명사 상당어구 구묶음

1. 서론

구문분석은 문장을 분석하여 문장의 구문 구조를 밝히는 작업으로 기계번역과 정보검색 등의 분야에서의 활용을 위해 국내외에서 많은 연구가 진행되어왔다. 하지만 현재의 구문분석 기술은 실용 분야에 적용시키기에는 성능상의 한계를 가지고 있다.

특히 영어와 비교했을 때, 한국어 문장은 문장 성분들 간의 어순이 자유롭고 필수 성분의 생략이 빈번하며 문장 성분들 간 복잡한 수식 관계를 가지고 있다. 한국어의 이러한 특징들은 문장의 구조적 중의성을 높여 구문분석 복잡도를 증가시키고 더 나아가 성능을 떨어뜨리는 요인이 된다.

구문분석의 계산 복잡도는 문장 내의 어절 수가 증가할수록 증가하는데 CKY 모델을 이용하였을 때의 계산 복잡도는 $O(n^3)$이다. 예를 들어, "영희는 여름 생활 용품을 구입하였다."라는 문장은 $O(5^3 = 125)$ 의 구문분석 계산 복잡도를 갖지만 <여름 생활 용품>을 하나의 복합명사 상당어구로 구묶음 하였을 때 문장 내의 어절 수가 줄어드는 것과 같은 효과로 위의 문장은 $O(3^3 = 27)$ 의 복잡도를 가지게 된다. 이렇듯 하나의 대표 명사로 나타내어 질 수 있는 연속된 명사들의 집합을 명사구로써 묶는다면 구문 분석 시 고려해야 할 후보 단어의 수가 줄어들어 복잡도를 현저하게 낮추고 중의성을 감소시킬 수 있다. 이러한 이유에서 구문분석의 선행 작업으로 명사구, 동사구, 등의 구묶음을 수행하는 많은 연구들이 진행되어 왔다[1-3].

구묶음에 관한 연구는 크게 규칙 기반 구묶음[1-4]과 통계 기반 구묶음[5-6]으로 나눌 수 있는데, 규칙 기반 구묶음에 대한 연구들은 규칙에 위배되는 구묶음의 처리가 힘들다는 단점이 있다.

따라서 본 논문에서는 구문분석의 선행 작업으로 기계학습 방법인 CRF를 이용한 복합명사 상당어구 묶음 방법을 제안한다. 이 방법은 구묶음을 분류 문제처럼 다루는 방법으로 어휘 정보와 띄어쓰기 정보, 앞뒤 어절의 품사 정보 등을 사용하여 학습 말뭉치에 태그를 부착하고 학습시켜서 모델을 생성 한다. 그리고 이 모델을 이용해 복합명사 상당어구 구묶음을 하고자 하는 새로운 말뭉치에 태그를 자동으로 부착하는데 실험을 위해 세종 구문분석 말뭉치 중 7만여 문장을 사용하였다.

본 논문의 구성은 다음과 같다. 2장에서는 구묶음 관련 연구들에 대해서 기술하며, 3장에서는 복합명사 상당어구를 정의하고 인식 모델에 대해 기술한다. 4장에서는 실험을 위해 말뭉치에 태그를 부착한 후 실험을 통하여 본 논문의 방법에 대한 성능을 평가해보며, 5장에서는 향후의 연구 과제를 살펴보고 결론을 짓는다.

2. 관련연구

구묶음(chunking)은 Abney에 의해 처음 제안되었다[7]. Abney는 구묶음을 "문장에서 겹쳐지지 않는 부분"이라고 정의하였다. 그 이후로 구묶음이라는 개념을 기반으로 국내외에서 많은 연구가 진행되어 왔다. 구묶음에 관한 연구는 크게 규칙 기반 구묶음과 통계 기반 구묶음으로 나눌 수 있다.

국내에 있었던 규칙 기반 구묶음에 대한 연구로 일정한 규칙을 기반으로 명사구 단위화를 한 후 동사구 장벽 알고리즘을 이용하여 동사구 묶음을 수행한 연구[1]가 있

었다. 또한 구문분석 말뭉치에서 추출한 명사쌍과 이들의 의미부류정보를 이용한 복합명사구의 구묶음 방법[2], 규칙과 어휘 정보를 함께 이용한 방법[3]이 있었다. 이외에도 규칙 기반 구묶음에 대한 많은 연구가 있었지만 규칙 기반 구묶음은 규칙에 위배되는 구묶음의 처리가 어렵고, 수동으로 규칙을 구축해야 한다는 단점이 있다.

통계 기반 구묶음에 대한 연구도 꾸준히 진행되어 왔는데 국내의 연구로 [5]의 경우는 한국어 기본구(base phrase)인식의 성능을 향상시키고자 할 때 최적의 자질 집합이 무엇인지 논하였다. 또한, "점증적 유용성"에 따라 현재 형태소의 품사, 어휘 및 띄어쓰기 정보와 주변 형태소의 품사, 어휘 및 띄어쓰기 정보 등을 기본 자질집합으로 선택하였고 실험을 위해 결정트리 학습과 메모리기반 학습방법을 사용하였다. 또 다른 통계 기반 연구로 한국어의 기반 명사구 인식을 위해 학습의 자질로 어절의 중심어를 사용하는 tri-gram HMM모델을 제안한 것이 있었다[6].

본 논문은 통계 기반의 기계학습 방법을 이용하여 구묶음을 분류 문제로써 다루었다. 기계학습 방법으로는 HMM의 한계를 보완한 모델인 CRF를 사용하여 연속 데이터, 즉 주어진 문장에 복합명사 상당어구 태그를 부착하는 복합명사 상당어구 인식 모델을 제안한다.

3. CRF를 이용한 복합명사 상당어구 인식 모델

본 논문은 CRF(Conditional Random Fields)를 이용한 복합명사 상당어구 구묶음을 제안한다. CRF는 연속된 데이터를 분할하고 라벨을 붙이는 확률 모델 구축을 위한 프레임웍으로 전이함수와 상태함수의 곱만을 고려하는 HMM(Hidden Marcov Model)의 한계를 보완한 모델이다[8]. 본 논문에서는 앞, 뒤 여러 어절을 고려해야 할 뿐 아니라 다양한 자질의 사용이 필요하여 연속 데이터 처리에 유리한 CRF를 사용하였다.

3.1. 복합명사 상당어구 정의

명사구라고 하면 명사의 구실을 하는 구로, 관형사 및 형용사가 수식하는 명사열을 말한다. 본 논문에서는 구문분석의 전처리로서의 명사구를 다룬다. 추출된 명사구는 의존 관계가 명확해야 하므로 다양한 형태의 확장이 가능한 형용사와 부사의 경우는 대상에서 제외하였다. 새롭게 정의된 추출 대상을 복합명사 상당어구(CNP: Composed Noun Phrase)로 명명하고 다음과 같이 정의한다.

복합명사 상당어구 : 하나의 명사처럼 쓰여질 수 있는 둘 이상의 연속된 명사, 대명사, 수사, 숫자와 이를 수식하는 관형사, 접두사 및 접미사

복합명사 상당어구의 예로 '수입품 4 개', '탈식민주의 이론', '이 회의'를 들 수 있다. 첫 번째 예인 '수입품 4 개'는 명사와 숫자가 결합한 구이고, '탈식민주의 이론'은 명사'탈식민주의'와 '이론'이 결합한 형태의 구라고 볼 수 있다. '이 회의'의 경우 '이'는 '회의'를 수식하는 관형사로 관형사를 포함하는 명사구도 복합명사 상당어구에 포함한다.

3.2. 복합명사 상당어구 인식의 정의

복합명사 상당어구의 인식은 주어진 문장의 어절들에 대해 적절한 복합명사 상당어구 태그를 할당하는 것으로 정의한다. 복합명사 상당어구 태그는 문장 내의 복합명사 상당어구의 위치를 나타낼 수 있도록 B, I, O 태그 법을 사용하며, B는 경계(boundary), I는 내부(inside), O는 그 밖(outside)를 의미한다. 여기에 명사(noun)을 나타내는 N을 추가하여 태그셋은 NB, NI, OB, OI 의 네 가지 종류의 태그로 이루어지는데, 각 태그의 의미는 [표 1]과 같다. "NB"와 "NI"는 각각 CNP의 시작, 내부 어절 태그로, "서구 세력"의 경우 "NB NI"로, "경찰 관료 체제"의 경우 "NB NI NI"로 태그가 부착된다.

[표 1] 복합명사 상당어구의 태그 표기법

NB	복합명사 상당어구의 시작 어절
NI	복합명사 상당어구의 내부/끝 어절
OB	복합명사 상당어구를 제외한 다른 구의 시작 어절
OI	복합명사 상당어구를 제외한 다른 구의 내부/끝 어절

[표 1]의 태그 표기법을 이용하여 문장, "출근용 정장 스타일의 임부복이 많이 나오고 있다"에 태그를 부착한 결과는 [그림 1]과 같다. "출근용 정장 스타일"이 하나의 복합명사 상당어구로 취급되어 "NB NI NI"태그가 부착되었다.

[그림 1] 복합명사 상당어구의 태그 부착의 예

3.3. 복합명사 상당어구 인식을 위한 자질 선택

기계학습에 의한 복합명사 상당어구 인식은 인식 결과 태그가 부착된 학습 말뭉치로 학습을 시킨 뒤 실제 데이터에 적용하는 방법이다. 이 때 성능에 가장 큰 영향을 미치는 것이 학습에 쓰이는 자질 집합이다. 본 논문에서는 형태소가 아닌 어절 단위로 CRF를 학습시켰는데, 이 때 사용한 자질들은 [표 2]와 같다.

[표 2]에서 W는 각 어절의 어휘 정보, F는 각 어절의 첫 번째 형태소의 품사, L은 마지막 형태소의 품사를 나타낸다. 그리고 작은 첨자로 표시된 i는 현재 어절, i-1은 이전 어절, i+1는 다음 어절을 나타내며 i-2, i+2는 두 어절 전, 후를 나타낸다. "출근용 정장 스타일의 임부복이 많이 나오고 있다."라는 문장을 예로 들면, i가 3일 때

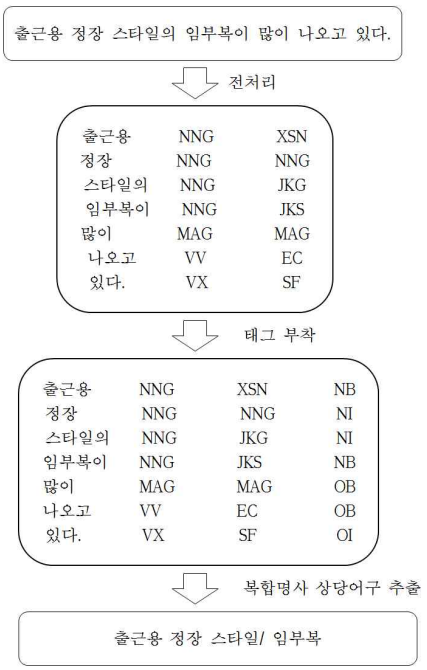
W_i은 “스타일의”라는 어절이고, F_i는“스타일”의 품사인 명사(NNG), L_i는 “의”의 품사인 조사(JKG)가 된다. 어휘와 품사 정보 뿐 아니라, ‘어휘+품사’를 혼합한 자질로 ‘이전 어절의 마지막 품사+현재 어절의 어휘 정보’를 함께 사용해서 성능의 향상을 보였다. 이는 어휘와 품사가 각각 쓰이는 것 외에도 현재 어휘의 바로 이전 품사가 모델 학습에 영향을 미치기 때문이다.

[표 2] 복합명사 상당어구 인식 모델의 자질 집합

종류	자질
어휘	W_{i-2}, W_{i-1}, W_i, W_{i+1}, W_{i+2}, W_{i-1}/W_i, W_i/W_{i-1},
품사	F_{i-2}, L_{i-2}, F_{i-1}, L_{i-1}, F_i, L_i, F_{i+1}, L_{i+1}, F_{i+2}, L_{i+2}, L_{i-1}/F_i, F_i/L_i, L_i/F_{i+1}, $F_i/L_{i+1}/F_{i+2}$
혼합	L_{i-1}/W_i

3.4. 복합명사 상당어구 인식 모델

문장이 입력되면 전처리 과정으로써 어절별로 자른 뒤 각 어절의 첫 번째 형태소와 마지막 형태소를 자질로써 어절의 오늘쪽 열에 차례로 기입한다. 여기에 태그 부착까지 끝난 문서로 학습을 시키면 학습된 모델이 생기고, 새로운 문장에 이 모델을 이용해서 태그를 자동으로 부착시킬 수 있다. [그림 2]는 본 논문에서 제안하는 CRF를 이용한 복합명사 상당어구 인식 모델의 예이다. “출근용 정장 스타일의 임부복이 많이 나오고 있다.”라는 문장이 입력되면 전처리로써 문장을 어절 별로 자른 뒤 어절의 첫 번째 형태소와 마지막 형태소가 오른쪽 열에 기입되고 이를 인식 모델에 입력하면 태그가 부착된 결과가 출력되고 이로부터 복합명사 상당어구를 출력하는 모델이다.



[그림 2] 복합명사 상당어구 인식 모델의 예

4. 실험 및 결과 분석

4.1. 실험 환경

복합명사 상당 어구는 최소 단위의 구 묶음에 해당한다. 따라서 구문 분석 말뭉치로부터 얻어질 수 있다. 본 논문에서는 실험을 위해 세종 구문분석 말뭉치의 구구조 구문트리의 단말 노드들에 아래의 [그림 3]과 같은 파싱 규칙을 이용하여 생성하였다.

```
NP <- NP_terminal
NP <- NP + NP
NP <- DP(관형사) + NP
NP <- NP + NP_SBJ
…
```

[그림 3] 파싱 규칙의 예

변환된 말뭉치는 모두 72801 개의 문장인데, 학습을 위해 58242개의 문장을, 평가를 위해 14559개의 문장을 사용하였다.

4.2. 실험 결과 및 분석

복합명사 상당어구 인식은 전체 태그열이 일치하는 경우에 정답으로 간주한다. 즉, 정답 태그열은 “NB NI NI OB”인데, 실험 결과가 “NB NI OB OB”라고 나왔을 경우 앞 쪽 “NB NI”는 일치하지만 전체 태그열은 일치하지 않기 때문에 틀린 것으로 간주한다.

[표 3]은 제안한 시스템에 대한 성능 평가 결과이다. 본 논문에서 제안한 시스템은 정확률 95.97%, 재현율 95.11%, F_1-평가치 95.54%로 95%를 웃도는 높은 성능을 보였다.

[표 3] 복합명사 상당어구 인식 모델의 성능

	정확률	재현율	F_1-평가치
제안 시스템	95.97%	95.11%	95.54%

$$\text{정확률} = \frac{\text{제안한 복합명사 상당어구 중 올바른 것의 수}}{\text{제안 시스템이 제안한 복합명사 상당어구의 수}}$$

$$\text{재현율} = \frac{\text{제안한 복합명사 상당어구 중 올바른 것의 수}}{\text{정답문서가 제안한 복합명사 상당어구의 수}}$$

$$F_1\text{-평가치} = \frac{2*(\text{정확률}*\text{재현율})}{(\text{정확률}+\text{재현율})}$$

오류 중 약 60%는 명사구 시작(NB)와 명사구 내부(NI)의 구별 실패에서 왔다. 예를 들어, “그날 오후 엑쌍프로방스를 거쳐...”의 정답 태그는 “NB NI NB OB...”인데, “엑쌍프로방스를”의 태그가 “NI”로 잘못 부착되어 총 태그가 “NB NI NI OB...”로 부착된 경우가 있었다. 이는 “오후 엑쌍프로방스를”이 명사구의 일부로 취급된 경우로, 학습 데이터에 있던 “오후 <명사>를”의 유일한 형태인 “오후 전투를”의 정답 태그인 “NB

NI"가 오류에 영향을 미친 것으로 보인다. 오류의 약 22%는 명사가 아닌데 명사구의 일부(NB 또는 NI)로 태그가 부착 된 것이었으며, 약 17%는 명사인데 OB나 OI 태그가 부착 된 것이었다.

본 논문이 제안하는 복합명사 상당어구 묶음이 구문분석의 전처리 과정으로써 유용함을 보이기 위하여, 전처리를 하지 않은 구문분석[9] 결과와 구묶음 후 구문분석 결과를 [표 4]에서 비교하였다. 구묶음 적용 여부에 따라 구문분석의 어절 정확률은 83.78%에서 83.91%로 구묶음을 한 경우 0.13%의 성능이 향상되었고, 문장 당 처리 속도 역시 5.81msec에서 4.06msec으로, 약 30%의 속도 향상이 있었다.

[표 4] 구묶음 적용에 따른 구문분석 성능 비교

	어절 정확률 (%)	문장 처리 속도 (msec)
구문분석	83.78	5.81
구묶음+구문분석	83.91	4.06

5. 결론

이 논문에서는 기계학습을 이용한 복합명사 상당어구 인식 모델을 제안하였다. 제안된 모델은 그 자질로써 띄어쓰기, 어휘 및 품사, 그리고 어휘와 품사를 혼합한 자질들을 사용하였으며, 연속적인 데이터의 태그 부착 성능이 우수한 CRF를 이용해서 95% 이상의 우수한 성능을 얻을 수 있었고, 구문분석의 전처리로써 사용하였을 때 구문분석의 성능 향상이 있었다.

향후 연구로 오류의 60%를 차지했던 명사구의 시작(NB)과 명사구의 내부(NI) 구별이 실패하는 문제를 해결하기 위해 명사의 의미 변별을 할 수 있도록 의미 정보를 고려하는 모델에 대한 연구가 필요하다. 추가적으로 동사구를 포함한 다른 기본구 인식도 한다면 구문분석의 중의성을 줄여 성능향상에 도움이 될 선행 시스템이 될 뿐만 아니라 정보 검색 등의 실용 분야에도 이용될 수 있을 것이라고 생각한다.

* 이 논문은 한국연구재단의 중견연구자 프로그램의 (No. 2009-0086194) 지원으로 이루어진 결과의 일부입니다.

참고문헌

[1] 신효필, "최소자원 최대효과의 구문분석," 제11회 한글 및 한국어 정보 처리 학술대회 발표자료집, 242-248, 1999.

[2] 안광모, 서영훈, "명사 의미 부류를 이용한 연속된 명사열의 구묶음,"한국콘텐츠학회논문지, 제10권, 제3호, pp. 10-20, 2010.

[3] 김미영, 강신재, 이종혁, "규칙과 어휘정보를 이용한 한국어 문장의 구묶음,"제12회 한글 및 한국어 정보처리 학술대회 발표논문집, 103-109, 2000.

[4] 임지희, 최호섭, 이정철, 옥철영, "자동 구축된 구문 패턴사전과 규칙을 이용한 구묶음,"제16회 한글 및 한국어 정보처리 학술대회 발표논문집, 35-39, 2004.

[5] 황영숙, 정후중, 박소영, 곽용재, 임해창, "자질집합 선택 기반의 기계학습을 통한 한국어 기본구 인식의 성능향상,"정보과학회논문지, 제29권, 제9호, pp. 654-668, 2002.

[6] 서충원, 오종훈, 최기선, "어절의 중심어 정보를 이용한 한국어 기반 명사구 인식," 제15회 한글 및 한국어정보처리 학술대회 발표논문집, 145-151, 2003.

[7] S. Abney, "Parsing by Chunks," In R. C. Berwick, S. P. Abney and C. Tenny, editors, Principle-Based Parsing: Computation and Psycholinguistics, Kluwer, pp. 257-278, 1991.

[8] J. Lafferty, A. McCallum, and F. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data,"In Proceedings of International Conference on Machine Learning, pp. 282-289, 2001.

[9] 박영민, "최대신장트리를 이용한 한국어 의존구문분석," 제22회 한글 및 한국어정보처리 학술대회 발표논문집, 68-72, 2010.