최대 엔트로피 모델을 이용한 한국어 명사구 추출[1]

강인호° 전수영 김길창
한국과학기술원 전자전산학과
{ihkang, syjeon}@csone.kaist.ac.kr, gckim@cs.kaist.ac.kr

Korean Noun Phrase Identification using Maximum Entropy Method

In-Ho Kang° Su Young Jeon Gil Chang Kim
Dept. of Electrical Engineering & Computer Science, KAIST

요 약

본 논문에서는 격조사의 구문적인 특성을 이용하여, 수식어까지 포함한 명사구 추출 방법을 연구한다. 명사구 판정을 위해 연속적인 형태소열을 문맥정보로 사용하던 기존의 방법과 달리, 명사구의 처음과 끝 그리고 명사구 주변의 형태소를 이용하여 명사구의 수식 부분과 중심 명사를 문맥정보로 사용한다. 다양한 형태의 문맥정보들은 최대 엔트로피 원리(Maximum Entropy Principle)에 의해 하나의 확률 분포로 결합된다. 본 논문에서 제안하는 명사구 추출 방법은 먼저 구문 트리 태깅된 코퍼스에서 품사열로 표현되는 명사구 문법 규칙을 얻어낸다. 이렇게 얻어낸 명사구 규칙을 이용하여 격조사와 인접한 명사구 후보들을 추출한다. 추출된 각 명사구 후보는 학습 코퍼스에서 얻어낸 확률 분포에 기반하여 명사구로 해석될 확률값을 부여받는다. 이 중 제일 확률값이 높은 것을 선택하는 형태로 각 격조사와 관계있는 명사구를 추출한다. 본 연구에서 제시하는 모델로 실험을 한 결과 평균 4.5개의 구를 포함하는 명사구를 추출할 수 있었다.

1 서론

명사구 추출 문제는 자연언어처리의 기본 작업으로 볼 수 있다. 명사구 추출은 구문 분석 전처리 작업으로 사용되어, 문장 해석의 애매성을 줄여서 과도한 문장 해석을 줄일 수 있다. 또한 추출한 명사구는 정보 검색, 정보 추출 및 자연언어처리의 기본 단위로 사용할 수 있다. 구문 분석 전처리 작업으로 사용될 경우, 명사구의 범위를 지정해줌으로 해서 명사구 범위와 교차하는 범위를 가지는 구문 해석을 피할 수 있다. 그리고 입력 문장을 단순화 시켜서 용언과 체언간의 관계를 지어주는 형태로 구문 분석의 복잡성을 줄일 수 있다. 정보 추출에 사용될 경우에는 미리 작성된 격정보 기반의 지식 사전을 적용시킬 때 템플릿에 해당하는 부분을 찾아내어 대응 시킬 수 있다.

본 연구에서 추출하고자 하는 명사구는 단순한 명사의 나열뿐만 아니라 구문 트리 태깅된 코퍼스에서 명사구로 해석된 모든 것을 대상으로 한다[2]. 그림 1에서는 '달 탐색선', '달 탐색선에 실어 보낼 실험 기기', 그리고 '달 탐색선에 실어 보낼 실험 기기를 개발한 연구원'을 명사구로 본다. 즉 기존의 정보 검색에서 사용하던 명사의 나열이나 복합 명사뿐만 아니라 수식어구를 포함한 명사구를 대상으로 한다.

영어에서는 관사와 소유격을 시작으로 하여 명사구의 시작을 알 수 있다. 반면 한국어 문장에서는 격조사를 기반으로 하여 명사구의 마지막을 추정할 수 있다. 명사구는 홀로 문장 성분으로 기능할 수도 있다. 그러나 대개는 조사와 결합하여 일정한 문장 성분으로 기능한다. 명사에 결합해서 일정한 문장성분으로 기능하게 하는 조사가 격조사이다. '가', '을', '에' 등과 같은 주격, 목적격, 보격, 관형격, 호격, 부사격, 접속격, 공동격, 인용격 조사가 있다. 격이란 문장성분들이 문장안에서 차지하는 자리이다. 결국 격조사는 이러한 기능을 명사구에 부여하는 조사이다. 달리 얘기하면 격조사 앞에는 명사구가 있다. 따라서 특정 격조사를 이용하면 명사구의 마지막 위치를 알 수 있다. *KAIST* 구문 트리 태깅 코퍼스에서 복합 명사와 같은 연속된 명사구의 결합을 하나의 명사구로 간주할 경우 보조사와 격조사가 명사구를 바로 따라오는 경우가 94.38%였다. 표 1은 명사구를 뒤따르는 품사 중에서 분포가 높은 일부 조사를 나타낸 것이다. 격조사를 기반으로 할 경우에, 명사구 추출 문제의

[1]본 연구는 한국과학 재단의 연구과제 "통계적 한국어 구문 분석"에 의해 지원되었음.
[2]본 연구에서는 KAIST 구문 트리 태깅 코퍼스(이공주, 장병규, 김길창, *1997*)를 기반으로 한다

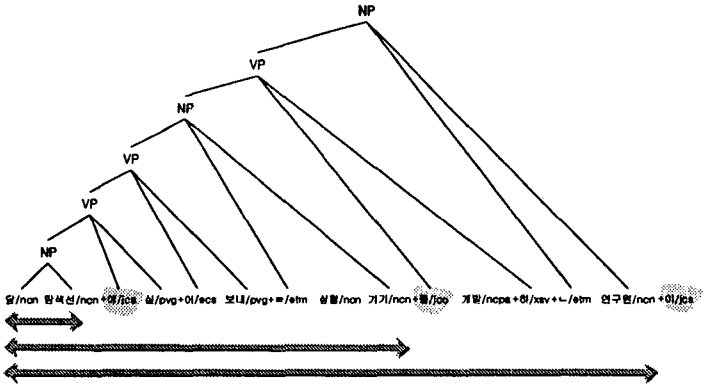

그림 1: 명사구 예제1

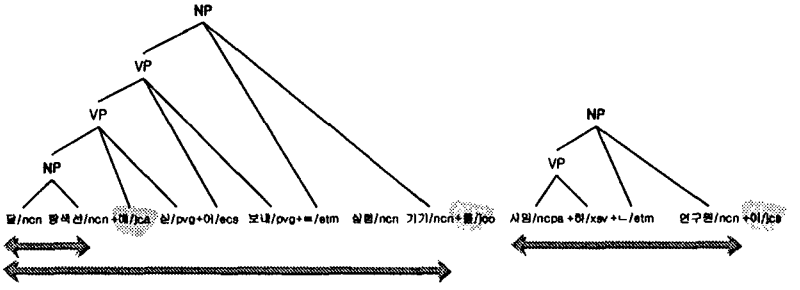

그림 2: 명사구 예제2

90% 이상이 시작 위치를 찾아내는 문제로 바뀐다.

표 1: 명사구를 뒤따르는 품사의 분포

품사	부사격	목적격	관형격	보조사	주격
분포	22.2%	17.2%	13.4%	12.8%	12.5%

명사구의 시작을 찾기 위해서는 단순한 품사열 만으로는 정보가 부족하다. 그림 1의 품사열과 동일한 형태의 그림 2에서 수식 어구의 범위가 달라져 서로 다른 범위의 명사구를 추출해야 함을 알 수 있다. 이를 해결하기 위해서는 '기기'와 '개발하다' 사이에는 새로운 명사구의 시작이 잘 나타나지 않는다는 정보가 필요하다.

명사구 시작 범위를 추출하는 것에 있어서는 단순한 주변 품사열 뿐만 아니라 수식어 간의 연관성도 고려해야 한다. 본 연구에서는 격조사가 앞에 있는 명사구와 결합한다는 성질을 이용하여 명사구의 마지막 위치를 추정하고, 수식어간의 상호 연관성을 고려하는 형태로 수식어의 정확한 범위를 알아낸다.

본 논문의 구성을 보면 다음과 같다. 먼저 명사구 추출에 대해서 설명하고, 본 연구에서 대상으로 하는 KAIST 트리태깅 코퍼스를 소개한 뒤, 명사구 추출을 위한 최대 엔트로피 모델 설명과 정보의 정형화 방법을 보인다. 그리고 학습 코퍼스에서 얻어낸 단순한 확률 정보를 이용하여 어느 정도 수준으로 명사구를 추정할 수 있는지를 보인다.

2 관련 연구

명사구 추출에 대한 연구는 크게 규칙 기반과 확률 기반의 방법으로 구분할 수 있다.

2.1 규칙 기반 명사구 추출

규칙 기반은 시간, 장소, 인물 같은 특정 의미 정보와 품사 정보를 이용하여 명사구에 대한 규칙을 기술한다. 이렇게 기술한 규칙을 이용하여 오토마타를 만들고 명사구를 추출한다(Abney, 1996).

```
## Noun Chunk
  nx -> such? DET? NUM? (ADJ | PTC)* (ADJ | N)* h=COMMON cd?
      | DET? NUM? (ADJ | PTC)* h=PROPER
      | DET h=(jjr | jjs | such)
      | cdql? h=dtp-q
      | h=( prp | cd | dtp | cd | dtp | qq | ex       # prp=Personal Pronoun
          | name | person | doll | ci-st | rbr | rbs
          )
      ;
```

그림 3: 명사구 규칙

예를 들어 그림 3의 제일 위의 규칙에 의해서 다음은 명사구로 인식된다.

- such a beautiful girl

(Abney, 1996)는 그림 3에서와 같이 인명과 지명을 포함하는 '*PROPER*'를 정의하기 위해서 인명 성씨 리스트와 미국의 주요 지명을 리스트로 가진다. 입력으로 사용하는 펜트리 뱅크 품사 체계로 부족한 정보는 리스트를 이용해서 세분화시켜 얻어낸다. 규칙 기반의 방법은 명사구 규칙을 작성하는 것이 쉽지 않다는 단점이 있으나 추출 과정에 대한 수정과 해석에 있어서 확률 기반 방법보다 용이하다는 장점을 가진다.

2.2 확률 기반 명사구 추출

확률 기반의 방법은 (Church, 1988)에서부터 시작한다. 단어 사이에 '['과 ']'을 삽입하여 임의의 단어 w_i에서 w_j가 명사구로 묶일 확률을 계산하여 문장 전체적으로 가장 확률값이 높은 경우를 찾는다. 예를 들어 "NN VB"인 문장이 주어질 경우, 5가지의 경우가 가능하다. 5가지 각각의 경우에 대해서 '[', ']', 및 아무것도 없을 확률값을 계속해서 곱하는 형태로 문장에 대한 전체 확률값을 구한다.

- NN VB
- [NN] VB
- [NN VB]
- [NN] [VB]
- NN [VB]

확률 기반의 다른 방법으로는 주어진 단어 사이에 명사구의 시작, 연결, 마지막, 그리고 명사구가 아님을 태깅하는 형태로 명사구 추출 문제를 정의한다(Skut and Brants, 1998a). 즉 주변 문맥정보와 앞선 명사구 추정 정보를 이용하여 고려하고 있는 위치가 4종류 중 무엇으로 태깅이 될지를 결정하는 것이다. 그러나 이러한 방법은 연속적인 문맥정보만 고려할 수 있다는 단점이 있다. 중간에 기호 등이 삽입되어 거리가 멀어질 경우에는 중심어와 직접적인 수식 성분을 문맥정보로 고려하지 못한다.

3 KAIST 구문 트리 태깅 코퍼스

본 연구에서 사용하는 KAIST 구문 트리 태깅 코퍼스는 구구조 문법에 기반하여 만들어졌다. 사용된 구구조 규칙은 기본적으로 다음과 같은 형태를 취하고 있다(이공주, 1998).

$$A \rightarrow B + \gamma \quad C$$

여기서 A, B, C는 구절을 의미하며, γ는 형식 형태소가 나타내는 문법적 관계를 의미한다. 이 규칙이 나타내는 정보는 B라는 구절이 C라는 구절을 γ라는 문법적 관계에 의해서 수식하고 있으면서, 두 구절이 결합되어 A라는 구절을 형성함을 의미한다. 이에 해당하는 규칙의 예를 보이면 다음과 같다.

$$VP \rightarrow NP + jcs \quad VP \tag{1}$$

$$NP \rightarrow VP + etm \quad NP \tag{2}$$

위의 수식 관계 표현에 의하면 수식 관계는 구절과 구절 간의 관계에서만 존재한다. 그런데 규칙의 변별력을 높이기 위해서 제한 사항을 수정한다. 수식받는 구절이 수식하는 구절과 바로 연접해 있는 어절의 관계에 있을 때는 규칙에서 수식받는 구절을 단어 수준(즉, 품사 수준)으로 기술한다. "밥을 먹는다."의 경우에는 '$VP \rightarrow NP + jco \ \ pvg$'로 규칙을 기술한다. 그러나 "밥을 학교에서 먹는다."의 경우에는 '학교에서 먹다'가 동사구절 '$VP \rightarrow NP + jca \ \ pvg$'를 형성하고, 다시 어절 '밥을'과 '$VP \rightarrow NP + jco \ \ VP$'의 규칙으로 결합되도록 한다. 따라서 그림 1에서 '실험 기기'를 명사구로 별도로 묶지 않는다. 이와 같이 KAIST 구문 트리 태깅 코퍼스를 기반으로 할 경우 명사구의 수식어가 존재할 경우 수식어구를 포함한 것만을 올바른 명사구로 간주한다. 또한 명사구와 명사구의 결합으로 더 큰 명사구로 해석되는 경우는 큰 명사구만 고려한다.

4 명사구 추출을 위한 최대엔트로피 모델

본 논문에서는 주어진 형태소열이 명사구로 해석 될 가능성을 최대엔트로피 모델에 기반하여 정의한다. 즉 주어진 형태소열과 그 주변의 정보를 $h(history)$로 두고, 명사구인지 아닌지를 $d(decision) \in \{0, 1\}$로 했을때, $p(d|h)$를 계산한다.

최대 엔트로피 모델에서는 주어진 여러 정보를 자질 함수(feature function)라는 형태로 정의한다. 자질 함수(f_i)는 trigger 형태로써, 정해놓은 제약 조건을 만족하였는지 그렇지 않은지를 구분해주는 0 또는 1의 값을 가지는 함수이다(Berger, Pietra, and Pietra, 1996). 명사구 추출에서는 고려되고 있는 형태소열에 대해서 특정 위치의 형태소 어휘와 품사의 조합으로 표현된다.

$$p(d|h) = \frac{\prod_{i=0}^{k} e^{\lambda_i f_i(h_i, d)}}{\prod_{i=0}^{k} e^{\lambda_i f_i(h_i, 0)} + \prod_{i=0}^{k} e^{\lambda_i f_i(h_i, 1)}} \tag{3}$$

최대 엔트로피 모델은 최대 엔트로피 원리(Maximum Entropy Principle)에 기반하여 만들어진다. 최대 엔트로피 원리란 랜덤 변수 $x_i (i = 1, 2, \ldots, n)$에 대한 확률 분포를 p_i라고 할 때, 자질 함수(feature function) f에 대해서 다음과 같은 제약

조건을 가한다.

$$E[f_j] = \tilde{E}[f_j],\ 1 \le j \le k \quad (4)$$

$$E[f_j] = \sum_{h \in H, d \in D} p(h,d) f_j(h,d) \quad (5)$$

$$\tilde{E}[f_j] = \sum_{i=1}^{n} \tilde{p}(h_i,d_i) f_j(h_i,d_i) \quad (6)$$

H는 있을 수 있는 모든 문맥의 집합을 나타내고, D는 원하는 출력값의 집합을 나타낸다. 그리고 n은 학습 데이터에서 발견된 문맥 h와 d의 곱집합으로 얻을 수 있는 총 가지수로써, 모델에서 고려하는 경우의 수를 의미한다. 식 5은 학습 데이터에서 발견된 경우만 고려하는 근사화 된 수식을 이용하여 계산한다.

$$E[f_j] = \sum_{i=1}^{n} \tilde{p}(h_i) p(d_i|h_i) f_j(h_i,d_i) \quad (7)$$

이렇게 두 개의 제약조건을 만족하는 확률 분포들 중에서 엔트로피가 최대가 되도록 모델을 구성하는 것이 최대 엔트로피 원리이다.

$$H(p) = -\sum_{h \in H, d \in D} p(h,d) \log p(h,d) \quad (8)$$

즉 최대 엔트로피 원리는 알려진 또는 사용하고자 하는 정보에 대해서는 확실히 지켜주고, 고려하지 않은 경우나 모르는 경우에 대해서는 어느 하나가 낫거나 못하다는 근거가 없기 때문에 동등하게 가중치를 줌으로 해서 특정 부분에 치우치지 않는 분포를 구한다는 뜻이다(Berger, Pietra, and Pietra, 1996).

최대 엔트로피 원리에 의한 파라미터 추정법은 (Jaynes, 1957)에 의해 제시되었고, 이를 수치적으로 측정하는 Generalized Iterative Scaling 방법이 (Darroch and Ratcliff, 1972)에 의해 고안되었다. GIS 방법을 발전시킨 Improved Iterative Scaling 방법(Berger, Pietra, and Pietra, 1996)도 많이 사용된다.

5 명사구 추출

5.1 명사구 구문 규칙 추출

격조사를 기반으로해서 명사구의 마지막 위치를 파악한 뒤에 명사구로 해석이 가능한 후보를 제시한다. 후보를 제시하기 위해서 전문가가 명사구에 관련한 규칙을 기술할 수도 있지만, 규칙 작성의 어려움으로 본 연구에서는 자동으로 명사구 규칙을 추출하는 방법을 사용한다. 즉 학습 코퍼스에서 명사구로 해석이 된 경우에 대해서 품사열 정보를 얻어낸다. 그림 1에서는 다음과 같은 명사구 규칙을 얻을 수 있다.

- ncn ncn[3]
- ncn ncn+jca pvg+ecs pvg+etm ncn ncn
- ncn ncn+jca pvg+ecs pvg+etm ncn ncn+jco ncpa+xsv+etm ncn

이렇게 얻어낸 품사열 정보를 이용하여 명사구 규칙을 대체한다.

5.2 명사구 추출 학습

자동으로 얻어낸 명사구 규칙을 학습 코퍼스에 적용하여 가능한 모든 명사구를 추정한다. 이렇게 추정한 명사구들에 대해서 제대로 추정한 경우와 그렇지 못한 것들을 이용하여 학습 데이터를 얻는다. 학습 데이터는 명사구의 바로 앞 두 형태소(m_{i-2}, m_{i-1})와 명사구의 시작 형태소(m_i), 마지막 형태소(m_{i+1}) 그리고 이때 이루어진 명사구 추정의 옳고 그름(d)을 정보로 가진다. 그림 1에서 표 2와 같은 정보를 얻을 수 있다.

표 2: 학습 데이터의 추출

m_{i-2}	m_{i-1}	m_i	m_{i+1}	d
		달/ncn	탐색선/ncn	1
보내/pvg	ㄴ/etm	실험/ncn	기기/ncn	0
		달/ncn	기기/ncn	1
기기/ncn	를/jco	개발/ncpa	연구원/ncn	0
하/xsv	는/etm	연구원/ncn	연구원/ncn	1
		달/ncn	연구원/ncn	1

이렇게 얻어낸 학습 데이터에 대해서 네가지 자질 함수를에 맞추어 실제 자질 함수를 얻어 낸다. 네가지 자질 함수 형태는 다음과 같다. 자질 함수 f_1은 수식 어구의 연결 여부를 알려주는 정보로 구의 시작인지 연결인지를 알 수 있다. 그림 1에서는 '기기/ncn'와 '개발/ncpa'사이가 새로운 구의 시작이 아니고 계속적인 연결인 것을 자질 함수 f_1을 이용하여 나타낸다.

$$f_1(h) = \begin{cases} 1 & \text{if } m_{i-2} = M_{i-2} \ \&\ m_{i-1} = M_{i-1} \\ & \quad \&\ m_i = M_i \ \&\ d = D \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

자질 함수 f_2, f_3, f_4는 모두 명사구의 중심어와 수식어구의 호응을 알 수 있게 한다. 이는 한국어에서 명사구의 중심어는 제일 마지막에 위치한다는 가정에 기반한다.

$$f_2(h) = \begin{cases} 1 & \text{if } m_{i-1} = M_{i-1} \ \&\ m_i = M_i \\ & \quad \&\ m_{i+1} = M_{i+1} \ \&\ d = D \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

[3]품사 체계는 KAIST 태깅 표준안을 따른다(강인호, 1999).

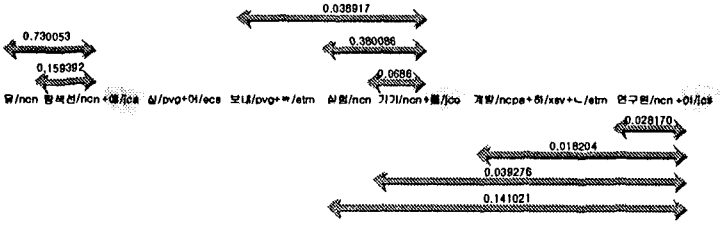

그림 4: 명사구 추정에(확률값이 클수록 하나로 묶일 가능성이 높다는 것을 의미한다.)

$$f_3(h) = \begin{cases} 1 & \text{if } m_{i-1} = M_{i-1} \ \& \ m_i = M_i \ \& \ d = D \\ 0 & \text{otherwise} \end{cases} \quad (11)$$

$$f_4(h) = \begin{cases} 1 & \text{if } m_i = M_i \ \& \ m_{i+1} = M_{i+1} \ \& \ d = D \\ 0 & \text{otherwise} \end{cases} \quad (12)$$

자료 희귀성 문제를 줄이기 위해 학습 코퍼스에서 3번 이상 출현한 단어에 대해서만 어휘까지 고려하고, 그렇지 않은 경우에는 품사만 고려한다. 그리고 f_1, f_2, f_3, f_4 각각의 경우에 대해서 품사만 고려하는 자질 함수도 함께 사용한다.

표 2의 네번째 예제인 "기기/ncn, 를/jco, 개발/ncpa, 연구원/ncn, 0"에서는 다음과 같은 자질 함수를 얻을 수 있다.

- m_{i-2} = 기기/ncn, m_{i-1} = 를/jco, m_i = 개발/ncpa, $d = 0$
- $m_{i-2} = ncn$, $m_{i-1} = jco$, $m_i = ncpa$, $d = 0$
- m_{i-1} = 를/jco, m_i = 개발/ncpa, m_{i+1} = 연구원/ncn, $d = 0$
- $m_{i-1} = jco$, $m_i = ncpa$, $m_{i+1} = ncn$, $d = 0$
- m_{i-1} = 를/jco, m_i = 개발/ncpa, $d = 0$
- $m_{i-1} = jco$, $m_i = ncpa$, $d = 0$
- m_i = 개발/ncpa, m_{i+1} = 연구원/ncn, $d = 0$
- $m_i = ncpa$, $m_{i+1} = ncn$, $d = 0$

만약 '연구원'이 학습 데이터에서 3번 이상 나타나지 않은 경우에는 위의 자질 함수에서 '연구원'을 제외한 형태로 자질 함수를 추출한다. 학습 코퍼스에서 얻어낸 자질 함수들은 최대 엔트로피 원리에 의해서 하나의 확률 모델로 합쳐진다.

5.3 명사구 추정

명사구 추정은 품사 태깅이 된 문장을 입력으로 사용한다. 격조사를 발견한 경우 보조사와 다른 조사를 제외한 명사구의 마지막 가능 품사를 찾아낸다. 그리고 미리 저장한 명사구의 시작 가능한 부분을 이용하여 명사구 후보를 추출하고 학습 코퍼스에서 얻어낸 확률값을 할당한다. 명사구로 판정 될 확률값이 경계값 0.5 이상인 것 중에서 가장 높은 후보를 명사구로 추정한다. 0.5 이상의 확률값을 가지는 후보 명사구가 존재하지 않는 경우에는 명사구 추정을 포기한다. 그리고 명사구 후보를 추정하는 경우에는 앞에서 명사구로 판정된 범위를 기억하여 범위가 교차하는 명사구는 후보에서 제외한다. 그림 4에서 '개발하는 연구원'이 하나의 명사구로 나타날 확률이 '실험 기기를 개발하는 연구원'보다 낮음을 알 수 있다. 그러나 0.5 보다 작으므로 명사구로 추정하지 않는다.

6 실험

구문 분석 정보가 부착된 KAIST 코퍼스 31,086 문장에 대해서 90%를 학습 데이터로 사용하고 나머지 10%를 실험 데이터로 사용한다. 학습 데이터에서 4만여개의 명사구 규칙을 얻어낼 수 있었다. 이는 단순히 품사열의 나열로 인한 결과이다. 명사구 추정에 있어서 동일한 기능을 하는 품사를 세분화한 결과와 동일한 패턴의 반복적인 사용을 달리 본 것에 기인한다. 본 논문에서 제시하는 모델과 명사구 후보 중 제일 긴 후보를 선택하는 최장 일치 방법의 결과는 표 3과 같다.

표 3: 명사구 추출 실험 결과

	학습 데이터		실험 데이터	
모델	정확률	재현율	정확률	재현율
최장 일치	77.5%	66.5%	60.5%	51.6%
최대 엔트로피	99.3%	81.0%	88.3%	60.0 %

표 3의 결과는 명사구 후보 중에서 확률값이 0.5 이상인 것을 대상으로 고려하였다. 그림 5는 경계값(alpha)에 따른 정확률과 재현율을 나타낸다.

7 토의

본 연구의 결과물로 나온 명사구를 분석하여 본 결과 명사구 내부에 평균 4.5개의 구가 존재했다. 이를 통해 단순 명사 나열을

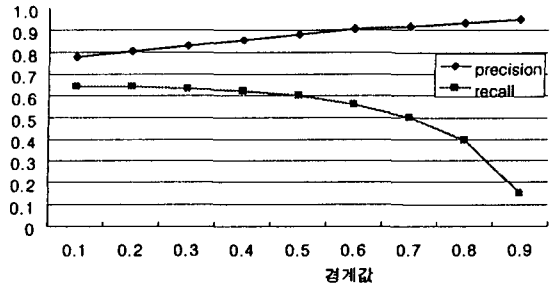

그림 5: 경계값에 따른 정확률과 재현율

찾아주는 방법보다는 구문 분석기 애매성 해소에 많은 도움을 줄 수 있을 것이라고 본다.

보다 나은 정확률과 재현율을 얻기 위해서는 세가지를 고려해야 한다. 첫번째는 격조사를 기반으로 마지막 위치가 동일한 명사구가 복수개가 존재할 수 있다는 것이다. 이를 해결하기 위해서는 명사구 후보 중에서 확률값이 최대인 것만 선택하는 것이 아니라 2순위와 1순위의 차이에 따른 복수개 추출도 필요하다.

두번째는 추출 대상으로 보는 명사구에 제약이 없다는 점이다. 수식어구의 복잡도 차이가 심하다. 보다 정형화되고 기본적인 형태의 명사구 선정 및 코퍼스 정형화 작업이 필요하다.

세번째는 재현율을 높이기 위해 격조사 외의 명사구 범위 추정자를 사용해야 한다. 이러한 예로 ', ", [,]와 같은 문장 기호를 들 수 있다. 위와 같은 문장 기호를 추가함으로 해서 보다 나은 결과를 얻을 수 있다.

8 결론 및 향후 연구

본 연구에서는 최대 엔트로피 모델을 이용한 명사구 추출 방법을 보였다. 격조사가 앞에 있는 명사구와 결합한다는 성질을 이용하여 명사구의 마지막 위치를 추정하고, 수식어간의 상호 연관성을 고려하는 형태로 명사구를 추출한다. 이는 구문 분석 전처리기의 입력으로 사용되어 입력문을 단순화 시켜 애매성을 줄여 구문 분석에 드는 비용을 줄인다. 또한 명사구만으로 용언의 격정보 추출이나 정보 추출의 사전 구성에 도움을 줄 수 있다.

본 연구에서는 복합 명사나 명사의 나열을 하나로 합쳐서 생각했다. 단순한 명사구 범위 추정뿐만 아니라 명사구안의 나열된 명사간의 수식 관계나 복합 명사의 분석에 대한 연구가 필요하다. 또한 명사구 내부의 연결 정도와 길이에 대한 정보를 고려하는 연구가 필요하다. 본 연구에서는 명사구만 고려하였지만 확장시켜 모든 구에 대해서 부분 구문 분석을 수행한 뒤 결과들을 합치는 형태로 구문 분석기를 만들 수 있다.

참고문헌

이공주, 김재훈, 장병규, 최기선, 김길창. 1996. 한국어 구문 트리 태깅 코퍼스 작성을 위한 한국어 구문 태그. Technical Report CS-TR-96-102, 한국과학기술원.

이공주, 장병규, 김길창. 1997. 한국어 구문 트리 태깅 코퍼스 작성 요령. Technical Report CS-TR-97-112, 한국과학기술원.

이공주. 1998. 언어 특성에 기반한 한국어의 확률적 구문 분석. Ph.D. thesis, 한국과학기술원.

강인호. 1999. 최대엔트로피 모델을 이용한 한국어 품사 태깅. Master's thesis, 한국과학기술원.

Abney, Steven. 1996. Partial parsing via finite-state cascades. In *the ESSLLI'96 Robust Parsing Workshop.*

Berger, A., V. Della Pietra, and S. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

Church, Kenneth Ward. 1988. A stochastic parts program and noun phrase parser for unrestricted text. In *Second Conference on Applied Natural Language Processing*, pages 136–143.

Darroch, J.N. and D Ratcliff. 1972. Generalized iterative scaling for log-linear models. In *The Annals of Mathematical Statistics*, volume 43, pages 1470–1480.

Jaynes, E.T. 1957. Information theory and statistical mechanics. *Physics Reviews106*, pages 620–630.

Skut, Wojciech and Thorsten Brants. 1998a. Chunk tagger - statistical recognition of noun phrases. In *ESLLI-98 Workshop on Automated Acquisition of Syntax and Parsing.*

Skut, Wojciech and Thorsten Brants. 1998b. A maximum-entropy parital parser for unrestricted text. In *the Sixth Workshop on Very Large Corpora.*

Yoon, Juntae, Key-Sun Choi, and Mansuk Song. 1999. Three types of chunking in korean and dependency analysis based on lexical association. In *the Eighteenth International Conference on Computer Processing of Oriental Languages*, pages 59–65.