방향성을 이용한 한국어 비재귀 명사구 인식 모델

이신목[0] 강인호 김길창
한국과학기술원 전자전산학과
{smlee, ihkang, gckim}@csone.kaist.ac.kr

Korean BaseNP Identification Model using Forward and Backward Processing Characteristics

Sheen-Mok Lee[0] In-Ho Kang Gil-Chang Kim
Department of Electrical Engineering & Computer Science,
Division of Computer Science, KAIST

요 약

비재귀 명사구(baseNP)는 단순한 단어 패턴과 품사 패턴에 의하여 쉽게 인식되므로, 자연어처리의 다양한 분야에서 활용한다. 교착어의 지배 성분 후위 원칙에 의하여 한국어 비재귀 명사구 인식은 보다 많은 광역 정보를 필요로 하므로, 본 논문에서는 광역 정보의 활용이 쉬운 상태 기반 모델을 사용한다.

본 논문은 상태 기반의 한국어 비재귀 명사구 인식에서 방향성을 고려한다. 교착어의 특성상 한국어 비재귀 명사구는 처음 위치가 끝 위치에 비하여 인식이 어려운 특징을 가지므로 방향성을 고려하여 오른쪽 우선의 방법을 활용할 경우, 모델의 특성 및 성능이 변화한다. 본 논문에서는 기존의 왼쪽 우선 방법과 새로이 제안하는 오른쪽 우선 방법을 각각 적용하고, 양 방법을 통합하는 방법들을 제안한다. 통합 결과 92.55%의 정확률과 90.90%의 재현률을 얻었다.

1. 서론

비재귀 명사구는 내부에 다른 명사구를 포함하지 않는 명사구이다.

[Elco/NNP] earned/VBD [$/$ 7.8/CD million/CD] ,/, or/CC [$/$ 1.65/CD] [a/DT share/NN] ./.

위 문장에서 *$1.65 a share*도 명사구이지만, 내부에 다른 명사구를 포함하므로 비재귀 명사구는 아니다. 반면, *$1.65*와 *a share*는 각각 비재귀 명사구이다.

비재귀 명사구는 문장의 전체 구조를 파악하지 않고서도 비교적 쉽게 인식이 가능하므로, 구문 분석 전처리기, 정보 검색 시스템, 문서 요약 시스템 등 다양한 분야에서 사용한다.

비재귀 명사구 인식에 관한 연구는 영어권에서 시작하였으며, 현재 한국어에서도 시도하고 있다. 한국어는 영어와는 달리, 지배 성분 후위 원칙을 가지므로, 비재귀 명사구의 길이가 길어지는 경향이 있다. 따라서, 영어의 경우에 비하여 광역적인 정보의 유용성이 크다. 본 논문에서는 이러한 성질을 반영하여, 광역 정보의 활용이 비교적 쉬운 상태 기반 모델을 사용한다.

본 논문은 교착어적 특성상 명사구 끝 위치 인식이 명사구 시작 위치 인식에 비하여 쉬운 한국어의 특성을 이용하여, 상태 기반 모델에 상태 전이의 방향성을 새로이 고려한다. 즉, 기존에 사용하던 왼쪽 우선의 상태 기반 모델 뿐만 아니라 오른쪽 우선의 상태 기반 모델도 함께 적용한다. 두 모델의 특성 및 성능에 대한 비교 관찰 결과를 이용하여 두 방법을 통합함으로써 성능을 향상시키는 방법을 모색한다.

본 논문의 구성은 다음과 같다. 기존의 비재귀 명사구 인식 관련 연구들을 살펴보고, 한국어 비재귀 명사구의 특징을 보인 뒤, 본 논문에서 사용한 상태 기반 모델 및 상태 전이의 방향성을 설명한다. 또한, 방향성을 이용한 두 가지 통합 모델을 보인 뒤, 실험 분석 결과를 살펴 본다.

2. 관련연구

2.1 영어권에서의 연구

현재까지의 비재귀 명사구 연구는 크게 수동 규칙에 의한 방법과 학습에 의한 방법으로 구분된다.

2.1.1 수동 규칙에 의한 방법

수동으로 작성된 규칙들을 이용하는 방법이다. Abney(1996) 등은 유한 상태 오토마타를 이용하였고, Voutilainen(1993)은 제약 규칙 문법을, Bourigault(1992)는 휴리스틱 규칙을 사용하였다. 이와 같은 방법들은, 규칙 작성자가 모든 언어적 현상을 규칙을 이용하여 표현할 수 없으므로, 학습에 의한 방법에 비하여 성능 향상의 가능성이 작다.

2.1.2 학습에 의한 방법

실제 명사구 인식 결과가 표시된 언어 코퍼스를 분석하여 언어 모델을 만든 후, 이 모델을 실제 데이터에 적용한다. 학습에 의한 방법에는 바이그램 모델, 은닉 마르코프 모델, 변형 기반 모델 및 메모리 기반 학습 모델 등이 있다.

Church(1988)는 바이그램 확률을 이용하여 단어와 단어 사이에 명사구 시작 표식과 끝 표식이 들어갈 확률을 구하고, 가능한 경우들 가운데 가장 가능성이 높은 순서열을 선택한다.

품사 정보에 대한 확률을 이용하는 바이그램 모델과는 달리, 은닉 마르코프 모델에서는 명사구 인식 결과들을 이용하여 확률을 참조한다. 즉, 바로 이전 상태의 명사구 인식 결과를 참조하여 다음 상태로의 전이 확률을 구한다. Xun et al.(2000)은 은닉 마르코프 모델을 이용하여 품사 태깅과 명사구 인식을 동시에 시행하는 방법을 제시한다. 이 방법은 92.3%의 정확률과 93.2%의 재현률을 보이며, 이는 단일 방법을 사용한 비재귀 명사구 인식 결과 가운데 가장 좋은 성능으로 생각된다. 은닉 마르코프 모델과 같은 상태 기반 모델들은 광역 정보를 참조한다는 장점이 있다. 그러나, 현재까지 본 논문에서 제안하는 "상태 전이의 방향성"을 고려한 경우는 아직 없다.

Ramshaw and Marcus(1995)는 변형 기반 학습 모델(Brill, 1993)을 명사구 인식에 사용한다. 학습 말뭉치를 초기 상태에서 목표 상태까지 변화시키면서 생성되는 순차적 변형 규칙을 얻는다. 이 규칙은 변형 조건과 변형 방법으로 구성되며, 변형 조건에 맞는 경우에 변형 방법을 적용한다. 인식 결과는 B(begin), I(in), O(out)와 같은 태그로 나타내며, 91.8%의 정확률과 92.3%의 재현률을 보인다. 이 방법은 학습 시간 및 공간을 많이 필요로 한다는 단점이 있다.

메모리 기반 명사구 인식 모델로는 Argamon et al.(1998)이 있다. Argamon et al.(1998)에서는 메모리에 저장된 품사 패턴이 해당 패턴의 명사구 인식에 대한 긍정적 증거 혹은 부정적 증거의 역할을 한다. 이러한 증거들을 바탕으로 명사구 인식 결과를 산출한 결과, 92.4%의 정확률과 재현률을 얻었다. 이 방법은 실행 속도가 느리고 공간을 많이 필요로 한다는 단점이 있다.

2.2 한국어에서의 연구

Yoon(1999)은 한국어에서의 세 가지 유형의 청크 구조를 규칙을 이용하여 단계적으로 묶은 후, 이를 이용하여 구문 구조를 분석하는 방법을 제시한다. 이 가운데 두번째 청크 구조가 명사구이다. 이 방법은 모델의 결과에 조사가 포함되므로, 다양한 용도로 활용하기 어렵다는 단점이 있다.

강인호(2000)는 코퍼스에서 명사구로 해석된 모든 것을 인식의 대상으로 한다. 조사가 명사구의 뒤에 나올 확률이 높다는 성질을 이용하여, 명사구의 끝 위치를 확정지은 상태에서 최대 엔트로피 모델을 이용하여 시작 위치를 찾는다. 추출 가능한 모든 명사구를 대상으로 하므로, 만족할 성능을 거두지 못하며, 결과의 적용 범위도 축소된다.

양재형(2000)은 한국어에서는 최초로 규칙 학습 기법에 의한 비재귀 명사구 인식 방법을 제안한다. 국어정보베이스의 15만 단어 규모의 트리태그 부착 말뭉치를 이용한 실험 결과, 91.8%의 정확률과 90.7%의 재현률을 보인다. 이 방법은 최초로 한국어에 적용된 비재귀 명사구 인식 방법으로서의 의미를 가지나, 지역적 규칙만을 사용한다는 단점이 있다.

3. 한국어 비재귀 명사구의 특징

영어 비재귀 명사구의 대상은 한정사를 포함하여 명사구의 지배 성분으로 끝나는 명사구의 앞부분으로써, 후치 수식 구문은 제외한다. 다음 문장은 Penn Treebank 코퍼스로부터 자동으로 추출한 영어 비재귀 명사구의 예이다(Ramshaw and Marcus, 1995).

Even [Mao Tse-tung] [' s China] began in [1949] with [a partnership] between [the communists] and [a number] of [smaller, non-communist parties].

영어 비재귀 명사구는, 위 예문에서 보듯이 관사와 소유격을 이용하여 그 시작 위치를 알 수 있는 경우가 많다. 따라서, 명사구 시작 위치 인식이 끝 위치 인식에 비하여 쉽다고 볼 수 있으나,

인식 난이도의 차이는 그리 크지 않다.

다음 문장은 영어권에서의 비재귀 명사구 정의를 한국어에 그대로 적용한 비재귀 명사구의 예이다(양재형,2000).

[이 동등한 권리]+가 [불평등한 노동]+에 대해서는 [불평등한 권리]+인 것이다.

한국어의 지배 성분 후위 원칙에 의하여, 한국어 명사구에는 후치 수식 구문이 존재하지 않는다. 즉, 동일한 의미를 가지는 문장에서, 영어에서의 후치 수식 구문은 한국어에서 전치 수식 구문으로 사용한다. 후치 수식 구문은 일반적으로 비재귀 명사구에서 제외한다. 따라서, 후치 수식 구문이 존재하지 않는 한국어에서의 비재귀 명사구는 영어에 비하여 그 길이가 길 것으로 예상 가능하다. 일반적으로 종속 성분과 지배 성분 간의 관계 인식이 비재귀 명사구 범위 결정에 중요한 역할을 함을 미루어 본다면, 한국어 비재귀 명사구 인식에서 광역적인 정보의 참조가 유용함을 알 수 있다.

한국어에서는 지배 성분 후위의 원칙에 의하여 용언의 관형형이 비재귀 명사구에 포함되는지 여부를 판단하기 쉽지 않는 등의 애매성이 자주 발생한다. 따라서, 명사구의 처음 위치의 판단이 쉽지 않다. 위의 문장에서 예를 들자면, '불평등한 노동' 에서 '불평등한' 이 '권리가' 의 지배 성분인지, '노동' 의 종속 성분인지 판단하여야 한다. 반면, 대부분의 명사구의 뒤에는 조사가 붙으므로, 끝 위치는 알기 쉬운 특징이 있다.

즉, 한국어 비재귀 명사구는 광역적인 정보를 더 유용하게 활용하며, 명사구의 시작 위치와 끝 위치 판별의 난이도 차이가 크다는 특성을 지닌다.

4. 상태 기반의 한국어 비재귀 명사구 인식

한국어의 상태 기반 명사구 인식에 관하여 논하고, 본 논문에서 제안하는 방향성에 대하여 설명하며, 실제 시스템에 적용한 모델을 소개한다.

4.1 상태 기반 명사구 인식

현재까지의 상태 전이 과정의 일부와 전이할 다음 상태에 대하여 미리 학습된 전이 확률을 이용하여 명사구 인식 결과를 선택하는 모델을 상태 기반 모델이라고 한다.

상태 기반 모델은 은닉 마르코프 모델을 포함하며, 은닉 마르코프 모델과 다른 점은 바로 이전의 상태들의 결과에 의하여 전이 확률을 정하는 마르코프 가정을 따를 필요가 없다는 점이다.

그림 1은 상태 기반 모델의 일부를 나타낸다.

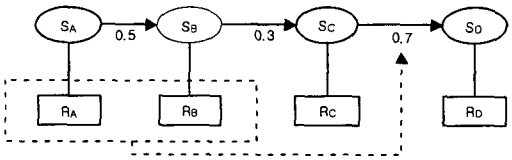


그림 1 상태 기반 모델의 일부

그림 1에서 타원은 각각의 상태를, 사각형은 각 상태에 따른 산출 결과를 나타낸다. S_C에서 S_D로의 전이 확률은 0.7로 주어지는데, 이것은 R_A,R_B와 R_D에 대하여 미리 학습된 전이 확률이다.

상태 기반 모델을 명사구 인식에 적용한 경우, 상태 기반 명사구 인식이라고 한다. 현재까지의 상태 기반 명사구 인식 연구들은 영어에서의 은닉 마르코프 모델을 중심으로 진행되었다. 현재까지 연구된 은닉 마르코프 모델은 왼쪽 우선의 분석 모델이었다. 그러나, 지배 성분 후위 원칙을 가지는 한국어에 대하여, 오른쪽 우선으로 분석하는 경우, 왼쪽 우선 모델과는 다른 장점이 있으리라 여겨진다.

4.2 상태 전이의 방향성

상태 전이의 방향성이란 입력 문장의 입력 순서를 역순으로 치환하여 상태 기반 모델에 적용한 경우에 생성되는 모델은 원래의 모델과 다르다는 성질이다. 즉, 두 상태 기반 모델은 단순히 상태 전이의 방향만 바뀌는 것이 아니라, 상태들 사이의 전이 확률 및 상태 전이 여부 또한 바뀜으로써, 문제의 난이도가 달라진다.

상태 기반 명사구 인식 모델에서도 상태 전이의 방향성은 존재한다. 일반적으로 왼쪽 우선의 모델과 오른쪽 우선의 모델은 문제의 정의가 다르며, 산출 결과 역시 달라질 수 있다.

다음 예문을 통하여 왼쪽 우선 및 오른쪽 우선 모델의 차이점을 알아본다.

영수는 이런 학생입니다. \(1)

문장(1)의 바른 명사구 인식 결과는 "[영수]는 [이런 학생]입니다." 이다. 다음에 문장(1)에 대하여 두 가지 상태 기반 모델을 적용한다. 참조하는 정보는 현재 상태를 포함하여 가장 최근 출현한 네 개의 상태로 정한다.

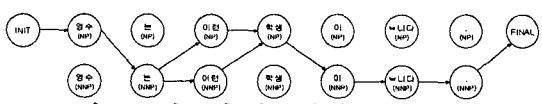


그림 2 문장(1)의 왼쪽 우선 모델 적용

그림 2는 문장(1)을 왼쪽 우선 모델에 적용한 결과이다. 그림 2에서의 각 상태는 해당 형태소와 명사구 인식 결과를 이용하여 표현한다. 명사구 인식 결과는 NP와 NNP로 표현하는데, NP는 명사구 내에 있음을, NNP는 명사구 밖에 있음을 나타낸다. 상태들 간의 화살표를 이용하여 상태 간의 가능한 전이를 나타낸다.

그림 2의 모델에서 <는(NNP)> 상태로부터 전이 가능한 다음 상태를 결정하려 한다. 이 경우, "[이런 좋은 결과]"와 같이 "이런"이 명사구에 속하는 경우와 "이런 [직접적]인 [결과] "와 같이 그렇지 않은 경우가 모두 존재할 수 있다. 즉, 이 모델에서 얻어지는 명사구 인식 결과 후보는 "[영수]는 [이런 학생]입니다."와 "[영수]는 이런 [학생]입니다." 이다.

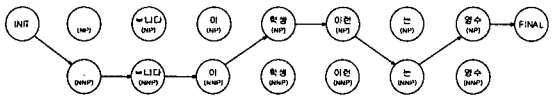


그림 3 문장(1)의 오른쪽 우선 모델 적용

문장(1)을 오른쪽 우선 모델을 사용하여 분석한 결과는 그림 3과 같다. 이 모델에서는 <학생(NP)> 로부터 전이 가능한 다음 상태를 결정하기 위하여, < .(NNP)> 상태와 <ㅂ니다(NNP)> 상태, 그리고 <이(NNP)> 상태를 함께 참조한다. 참조 정보를 통하여 "학생"의 뒷부분은 문장의 끝부분임을 알 수 있다. 만약, 관형사 "이런"이 비재귀 명사구에 포함되지 않는다면, "학생"을 제외한 다른 명사를 수식하여야 하는데, 수식 가능한 다른 명사가 존재하지 않음을 참조 정보를 통해 알 수 있다. 따라서, <학생(NP)>로부터 전이 가능한 상태는 <이런(NP)> 상태뿐이다. 이와 같이, 이 모델은 오직 한 개의 인식 결과만을 산출한다.

즉, 문장(1)을 왼쪽 우선 모델에 적용할 경우, 잘못된 인식 결과를 낼 수 있지만, 오른쪽 우선 모델에 적용할 경우, 항상 올바른 결과를 낸다.

이상으로부터, 왼쪽 우선 모델과 오른쪽 우선 모델은 일반적으로 서로 다른 형태의 모델을 산출하며, 이로 인하여 결과값 또한 달라질 수 있음을 알 수 있다.

4.3 두 모델 분석 결과의 통합

4.3.1 왼쪽 우선 및 오른쪽 우선 모델의 특징

왼쪽 우선 모델의 가장 큰 특징은 명사구의 시작 위치를 끝 위치에 비하여 먼저 인식한다는 점이다. 따라서, 명사구의 시작 위치 탐색에 있어서 유리하다. 시작 위치 탐색은 한국어의 특성상 정확률이 낮으므로, 왼쪽 우선 모델은 정확률에 있어서 불리한 방법이다. 반면, 모든 가능한 명사구 시작 위치를 고려하는데 유리하므로, 재현률에 있어서는 유리하다.

오른쪽 우선 모델은 반대로 명사구의 끝 위치 탐색에 유리하다. 끝 위치 탐색은 정확률이 높은 방법이므로, 정확률에 있어서 유리한 반면, 가능한 명사구 시작 위치의 탐색에는 불리하므로 재현률에 있어서는 불리하다.

두 모델 간의 이와 같은 차이점을 이용하기 위한 두 가지의 통합 방법을 소개한다.

4.3.2 문맥 정보를 이용한 결과 선택 방법

두 모델의 결과에서 인식 결과가 서로 다른 부분에 대하여, 주변 품사 문맥에 따라서 미리 학습된 가중치를 이용하여, 인식 결과를 선택하는 방법이다. 이 방법을 그림으로 표현하면 그림 4와 같다.

그림 4는 X,Y 라는 서로 다른 결과를 내며, A,B가 좌우 품사이고, A,B에 대하여는 서로 같은 결과를 내는 경우의 예이다.

가중치 산출 방법은 다음과 같다. 그림과 같은 경우의 학습 코퍼스 부분에 대하여, X가 정답인 경우, 두 모델 각각의 확률 정보에 가중치를 첨가하여 계산하면 X에 대한 결과가 Y에 대한 결과보다 커야 한다.

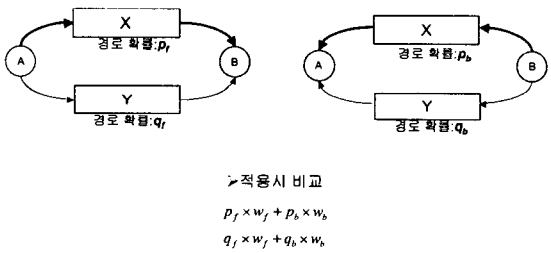


그림 4 문맥 정보를 이용한 결과 선택 방법

즉, 다음과 같은 부등식을 만족하여야 한다.

$$p_f \times w_f + p_b \times w_b > q_f \times w_f + q_b \times w_b \quad \ldots(1)$$

미리 정의된 상수 c를 이용하여 (1)을 등식으로 바꾸면 다음과 같다.

$$p_f \times w_f + p_b \times w_b = q_f \times w_f + q_b \times w_b + c \ldots(2)$$

(2)를 이용하여 가중치 w_f, w_b의 비율을 구한다. 학습 코퍼스의 모든 동일한 경우에 대하여 비율을 구한 뒤, 비율의 평균값을 가중치로 선정한다.

적용 문장에서는 다음과 같은 두 확률값을 비

교하여 결과를 선택한다.

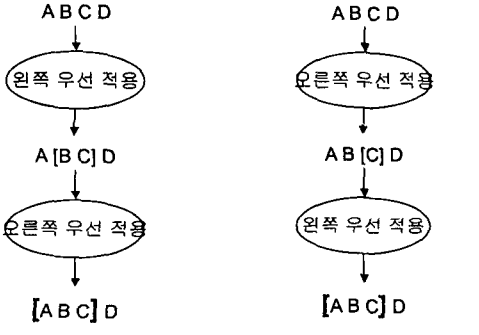
$$p_f \times w_f + p_b \times w_b,\ q_f \times w_f + q_b \times w_b$$

4.3.3 순차적 통합 방법

왼쪽 우선 모델과 오른쪽 우선 모델을 각각 순차적으로 적용하는 방법이다. 왼쪽 우선 모델을 우선 적용하는 방법과 오른쪽 우선 모델을 우선 적용하는 방법으로 구분한다.

두번째로 적용하는 모델은 첫번째로 적용한 모델의 결과값을 입력으로 받는다. 따라서, 두번째 모델의 전이 확률은 첫번째 모델의 명사구 인식 결과에 대하여 학습한다. 이 방법을 그림으로 표현하면 그림 5와 같다.

입력 문장은 품사 태깅 정보만을 가진 형태이고, 중간 결과는 입력 문장에 첫번째 모델을 적용한 결과이고, 최종 결과는 중간 결과를 입력으로 두번째 모델을 적용한 결과이다.



(a) 왼쪽 우선 모델→ 오른쪽 우선 모델

(b) 오른쪽 우선 모델→ 왼쪽 우선 모델

그림 5 순차적 결과 통합 방법

5. 실험 및 분석

5.1 실험 환경

본 실험에서는 양재형(2000)에서 사용한 국어 정보베이스의 322,057 형태소의 말뭉치를 이용한다. 이 가운데, 학습 말뭉치는 240,510 형태소(8903문장)로 구성되며, 시험 말뭉치는 81,547 형태소(3180 문장)로 구성된다. 이 말뭉치는 200,000 단어 규모의 학습 말뭉치와 50,000 단어 규모의 시험 말뭉치로 이루어진 영어 비재귀 명사구 인식 표준 말뭉치(Ramshaw and Marcus, 1995)에 비하여 뒤지지 않는다.

말뭉치는 54개의 태그셋을 사용하며, 비재귀 명사구의 정의는 양재형(2000)을 따른다.

5.2 왼쪽 우선 모델과 오른쪽 우선 모델의 적용

왼쪽 우선 모델과 오른쪽 우선 모델을 적용한 결과는 표 1과 같다. 4.3.1에서 논한 바와 같이, 왼쪽 우선 모델은 재현률에서, 오른쪽 우선 모델은 정확률에서 높은 성능을 보이며, F-value는 왼쪽 우선 모델이 약간 높다.

	정확률	재현률	F-value (α=0.5)
왼쪽 우선 모델	92.05	**90.64**	**91.34**
오른쪽 우선 모델	**92.33**	90.27	91.29
(양재형, 2000)	91.8	90.7	91.25

표 1 왼쪽 우선 및 오른쪽 우선 모델의 성능 비교

두 모델의 각 경계 표식에 대한 성능은 표 2와 같다. 시작 표식의 인식은 왼쪽 우선 모델이, 끝 표식의 인식은 오른쪽 우선 모델이 높은 성능을 보인다.

		정확률	재현률	F-value
시작 표식	**왼쪽 우선**	95.21	97.12	**96.16**
	오른쪽 우선	95.13	96.99	96.05
끝 표식	**왼쪽 우선**	98.54	96.92	97.72
	오른쪽 우선	98.90	96.72	**97.80**

표 2 각 경계 표식에 대한 두 모델의 성능 비교

5.3 두 모델의 통합 실험

두 모델을 지역적 품사 문맥 정보를 이용한 결과 선택 방법과 순차적 통합 방법을 이용하여 통합한 결과는 표 3과 같다. 표 3에서 보듯이, 본 실험의 통합 방법들 가운데 가장 높은 성능을 보인 것

	정확률	재현률	F-value
순차적 통합 (왼쪽→오른쪽)	92.46	**90.90**	91.67
순차적 통합 (오른쪽→왼쪽)	**92.50**	90.88	91.68
문맥 정보 이용 선택방법	**92.55**	**90.90**	**91.71**
양재형(2000)의 최종결과	91.8	90.7	91.25
양재형(2000)의 초기결과	81.1	78.2	79.62
Xun et al.(2000)의 결과	92.3	93.2	92.75

표 3 두 모델의 통합 결과 및 다른 연구와의 비교

은 문맥 정보를 이용한 결과 선택 방법이다.

또한, 순차적 적용 방법에서 오른쪽 우선 모델을 먼저 적용한 경우 정확률에서, 왼쪽 우선 모델을 적용한 경우 재현률에서 더 높은 성능을 보인다. 이 결과를 바탕으로 순차적 통합 방법에서 첫번째로 적용하는 모델의 특성이 전체 결과에 미치는 영향이 더 큼을 알 수 있다.

모든 통합 결과는 양재형(2000)에 비하여 높은 수치를 기록하였다. 영어에서 단일 방법으로 최고의 성능을 보인 Xun et al.(2000)에 비하여서는 정확률 면에서는 더 높은 수치를 기록하였으나, 재현률 면에서는 낮은 수치를 기록하였다. 이 같은 차이는 한국어 비재귀 명사구 인식에 있어서 시작 위치를 찾는 작업의 난이성 때문으로 해석된다.

6. 결론 및 향후 과제

본 논문에서는 한국어 비재귀 명사구 인식을 위하여,상태 기반 모델에 한국어의 특성을 반영한 상태 전이 방향성을 고려하였다. 기존의 왼쪽 우선 모델 뿐 아니라 분석 방향에 따른 차이점을 이용하기 위하여 오른쪽 우선 모델 또한 적용하였다. 그 결과, 왼쪽 우선 모델은 명사구의 시작 표식을 잘 인식하는 반면, 오른쪽 우선 모델은 명사구의 끝 표식을 잘 인식하는 결과를 보였다. 또한, 실험 결과, 왼쪽 우선 모델은 정확률에서, 오른쪽 우선 모델은 재현률에서 높은 성능을 보였다.

두 가지 모델의 성질이 서로 다름을 이용하여, 이들을 통합하였다. 통합 방법은 두 가지 방법을 사용하였다. 첫번째 방법은 두 가지 방향 모델의 결과를 선택하는 방법이다. 왼쪽 우선 모델과 오른쪽 우선 모델을 각각 적용하여 결과를 추출한 후, 서로 결과가 다른 부분에 대하여 주변의 문맥 정보에 따른 가중치를 이용하여, 결과를 선택하는 방법이다. 두번째 방법은 순차적 통합 방법이다. 한 가지 모델을 이용하여 하나의 결과를 추출한 후, 그 결과에 다른 모델을 적용하는 방법이다. 이러한 두 가지 통합 방법을 이용한 결과 왼쪽 우선 모델을 통한 결과나 오른쪽 우선 모델을 통한 결과보다 더 나은 성능을 얻었다.

본 연구의 결과를 통하여, 한국어 상태 기반 명사구 인식에 있어서 방향성이 존재함을 알 수 있었고, 두 모델을 통합함으로써 기존의 단방향 방법에 비하여 높은 성능의 결과를 얻을 수 있었다.

상태 전이의 방향성은 참조 정보의 선택에 따라서 더욱 크게 나타날 수 있다. 참조 정보를 어절 단위로 늘린다면 더욱 좋은 결과를 얻을 수 있으리라 생각한다.

왼쪽 우선 방법과 오른쪽 우선 방법을 통합하는 방법의 개선 또한 가능하다. 품사 문맥 가중치를 이용한 방법에서 가중치 산출 방식의 변화 등을 통하여 성능을 향상시킬 수 있으리라 예상된다.

7. 참고 문헌

(Church,1988) K. W. Church. 1988. A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text. Proceedings of the 1st Conference on Applied Natural Language Processing, pages 136-143.

(Ramshaw and Marcus, 1995) L. Ramshaw and M. Marcus. 1995. Text Chunking Using Transformation-Based Learning. In Natural Language Processing Using Very Large Corpora., pages 82-94

(Voutilainen, 1993) Atro Voutilainen. 1993. NPTool, a detector of English noun phrase. Proceedings of the Workshop on Very Large Corpora, page 48-57

(Bourigault, 1992) D. Bourigault. 1992. Surface Grammatical Analysis for the Extraction of Terminological Noun Phrase. Proceedings of the Fifteenth International Conference on Computational Linguistic

(Brill, 1993) E. Brill. 1993. A Corpus-Based Approach to Language Learning. Ph.D. thesis, University of Pennsylvania.

(Xun et al., 2000) E. Xun, C. Huang, M. Zhou. 2000. A Unified Statistical Model for the Identification of English BaseNP. Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics.

(Yoon et al., 1999) Juntae Yoon, Key-Sun Choi, Mansuk Song. 1999. Three Types of Chunking in Korean and Dependency Analysis Based on Lexical Association. the Eighteenth International Conference on Computer Processing of Oriental Languages, pages 59-65.

(**양재형**,2000) **양재형**. 2000. **규칙 기반 학습에 의한 한국어의 기반 명사구 인식. 정보과학회논문지 제** 27 **권 제** 10 **호**, pages 1062-1071.

(**강인호**, 2000) **강인호, 전수영, 김길창**. 2000. **최대 엔트로피 모델을 이용한 한국어 명사구 추출. 제** 12 **회 한글 및 한국어 정보처리 학술대회**, pages 127-132