두 단계 학습을 통한 중국어 최장명사구 자동식별

윤창호° 이용훈 김미훈 김동일+ 이종혁
포항공대 정보통신대학원 정보처리학과 포항공대 컴퓨터공학과° 중국연변과학기술대학 언어공학연구소+
첨단정보기술 연구센터
{yustian, yhlee95,meixunj,dongil, jhlee}@postech.ac.kr

Two-Level Machine Learning Approach to Identify Maximal Noun Phrase in Chinese

Changhao Yin° Yong-Hun Lee Meixun Jin Dong-il Kim+ Jong-hyeok Lee
Dept. of Graduate School for Information Technology, POSTECH, Dept. of Computer Science &Engineering, POSTECH2, Language Engineering Institute, YUST+ China, and Advanced Information Technology Research Center(AITrc)

요 약

일반적으로 중국어의 명사구는 기본명사구(base noun phrase), 최장명사구(maximal noun phrase) 등으로 분류된다. 최장명사구에 대한 정확한 식별은 문장의 전체적인 구조를 파악하고 정확한 구문 트리(parse tree)를 찾아내는데 중요한 역할을 한다. 본 논문은 두 단계 학습모델을 이용하여 최장명사구 자동식별을 진행한다. 먼저 기본명사구, 기본동사구, 기본형용사구, 기본부사구, 기본수량사구, 기본단문구, 기본전치사구, 기본방향사구 등 8가지 기본구를 식별한다. 다음 기본구의 중심어(head)를 추출해 내고 이 정보를 이용하여 최장명사구의 식별을 진행한다. 본 논문에서 제안하는 방법은 기존의 단어레벨의 접근방법과는 달리구레벨에서 학습을 진행하기 때문에 주변문맥의 정보를 많이 고려해야 하는 최장명사구 식별에 있어서 아주 효과적인 접근방법이다. 후처리 작업을 하지 않고 기본구의 식별에서 25개 기본구 태그의 평균 F-measure가 96%, 평균길이가 7인 최장명사구의 식별에서 4개 태그의 평균 F-measure가 92.5%로 좋은 성능을 보여주었다.

1. 서 론

자연언어처리에 있어서 명사구에 대한 식별은 정확한 구문 트리(parse tree)를 분석함에 있어서 아주 중요한 역할을 한다. 특히 중국어와 같이 조사가 발달하지 않은 언어는 명사구 식별에 있어서 많은

중의성(ambiguity)이 존재한다. 통계[12]에 따르면 중국어에서 중의성이 많이 일어나는 37개 구 패턴 중 29개의 패턴이 명사구 때문에 중의성이 생기게 된다. 따라서 구문분석을 위해서는 구묶음(chunking) 단계에서 명사구에 대한 정확한 식별이 꼭 필요하다. 중국어 문장 중의명사구는 구성성분의 결합관계와

개수에 따라 크게 2가지로 볼 수 있다.

첫째, 기본명사구 : 명사구 안에 다른 어떤 구도 포함하지 않은 명사구이다.

둘째, 최장명사구 : 기본명사구가 아니면서 문법적으로 명사구의 역할을 하는, 다른 어떤 구에도 포함되지 않는 명사구이다.

다음은 2가지 명사구에 대한 예문이다.

예 1) 기본명사구

半 . [世] 之后 [上海 浦东] 成为 [外资 金融 机构] 入驻
반 개 세기 후 상해 푸뚱 성장하다 외국 금융 기구 입주하다
[上海] 的 一 个 无可争论 的 [地区]
상해 적(의) 일 개 영순위 적(의) 지역
한국어 대응문 : 반 세기 후 상해푸뚱은 외국금융기구가 상해에 입주하는 영순위의 지역으로 성장하였다.

예 2) 최장명사구

[半 . 世] 之后 上海 浦东 成为 [外资 金融 机构 入驻
반 개 세기 후 상해 푸뚱 성장하다 외국 금융 조직 입주하다
上海 的 一 个 无可争论 的 地区]
상해 적(의) 일 개 영순위 적(의) 지역

예문 1, 2에서 보다시피 기본명사구로부터 최장명사구까지 명사구의 구성성분의 개수는 점점 증가하며 따라서 자동식별의 난이도도 증가한다. Abney[1]가 처음으로 구묶음(chunking)을 구문분석의 전 처리 단계로, 구문분석을 두 단계로 나누어 해결하는 방법을 제시하면서 일반적으로 중국어 구문분석 역시 비슷한 접근방

법을 이용하고 있다. 전처리 단계 즉 구묶음(chunking) 단계에서 기본명사구만을 식별한다면 많은 중의성을 해소하지 못하고 그대로 다음 단계에 넘기기 때문에 구문분석의 복잡도를 크게 낮추지 못한다. 따라서 구문분석의 복잡도를 낮추고 성능을 높이기 위해서는 구묶음(chunking) 단계에서 최장명사구까지 정확히 식별함으로써 문장의 전체적인 구조를 파악하는데 도움이되도록 해야 한다. 최장명사구의 식별이 어려운 것은 바로 명사구의 구성성분 사이에 많은 중의성이 존재하기 때문이다. 특히 중국어의 언어적 특성 상 대부분의 품사가 형태적인 변형이나 전치사의 도움이 없이도 최장명사구의 구성성분이 될 수 있기 때문에 중국어
최장명사구의 식별은 더욱 어려울 수 밖에 없다.

예 4) 명사구와 동사구에서 생기는 중의성

명사구 : [打算/v 逃跑/v] 的 囚犯/n
plan to escape DE criminal
한국어 대응문 : 탈옥을 시도하는 죄수

동사구 : 射击/v [跑动/v 的 猎物/n]
shoot running DE animal
한국어 대응문 : 달리는 동물을 사격하다

예문4에서 보다시피 **'동사+동사+ 돨(의)+명사'** 패턴은 명사구와 동사구 사이에서 중의성이 존재한다. 영어 같은 경우에는 위에서처럼 to 부정사가 두 동사를
연결해 주거나 동사가 ing형으로 변형되어명사를 수식한다. 이와 달리 중국어는 형태적인 변형이나 다른 품사의 도움이 없이 직접 동사들 사이에 수식관계가 일어나기 때문에 중의성의 해소에어려움이 크다.

2. 관련 연구

기본명사구의 자동식별에 대한 연구는 많이 진행되었지만 최장명사구에 대한 연구는 많지 않다. 특히 중국어의 최장명사구에 대한식별은 아직 시작 단계이다. 구묶음(chunking) 식별에서 대부분의 연구는 Ramshaw[5]가 제시한 IOB[1] 태그의 변형기반의 방법을 이용한다.

영어 기본명사구에 대한 기존연구를 살펴보면 :

1) Erik[3]는 먼저 IOB IOB[1] : Inside, Outside, Beginning of a consecutive chunk 등 5가지 변형된 구 표기형식으로 5종류의 학습모델을 만들고 다음 그 결과에 가중치를 부여하는 방법으로 내부 시스템 합성(internal-system combination)을 진행하여 단일 classifier로는 최고성능의 학습모델을 구축하였다. 같은 방법으로[2] 7종류의 classifier로 학습모델을 구축하고 외부 시스템 합성(external-system combination)을 진행하는 방법으로 영어 기본 명사구 식별에서 최고 성능을 보여 주었다.
2) Taku[4]는 SVMs으로 4가지 구 표기형식[3]과 양방향 학습에 기반한 8종류의 학습모델을 구축하고 내부 시스템 합성하는 방법으로 영어 기본명사구 식별에서 좋은 성능을 보여 주었다.
시스템 합성방법에는 여러가지voting 방법이 존재하지만 그 차이는 미미하며, 일반적으로 majority voting 방법을 사용한다. 위의 두 논문은 시스템 합성방법이 단일 시스템보다 좋은 성능을 보임을 증명하였다.

중국어 기본 명사구에 대한 기존 연구를 살펴보면 :

3) Huang[13]은 중국어 기본명사구 자동식별에서 규칙기반의 접근방법을 사용하였다. 먼저 명사구의 왼쪽경계와 오른쪽경계에 대한20가지 규칙 탬플릿을 구축한 다음 코퍼스의 통계정보를 이용하여 결과를 분석하고 임계치를 넘는 새로운 규칙을 생성하여 탬플릿에 추가하고 새로운 규칙이 더 생기지 않을 때까지 같은 알고리즘을 반복하는 방법을 제안하였다. 문제점은 사람이 복잡한 언어현상을 모두 감안하면서 정교한 규칙 템플릿을 만든다는 것은 현실적으로 한계가 있다는 것이다.
4) Zhang[14]은 Memory-based learning 으로태그기반의 기본구자동식별을 진행하였다. 주변문맥의 크기를 늘리면서 실험을 진행하였는데 window size 5에서 결과가 수렴함을 보여 주었다.

중국어 최장명사구에 대한 기존연구를 살펴보면 :

5) Zhou[2]는 중국어 최장명사구 식별을 두 단계로 나누어 진행하였다. 먼저 코퍼스의 통계정보를 이용하여 문장부호, 공기정보(co-occurrence), 등위접속(co-ordinate)구조의 좌우경계를 식별해 낸 다음 규칙기반의 방법으로 최장명사구의 오른쪽 경계를 식별하고 계속해서 왼쪽으로 확장해 나가는 방법으로 최장명사구를 식별하였다. 이 논문에서 5단어 이상의 구성성분을 가지고 있는 최장명사구에 대해서 정확률 70.8%, 재현율 61.7%를 보여주

1) IOB1 : Initialized with B, others same with IOB
IOE : Inside, Outside, End of chunk followed by another chunk
IOE1 : Initialized with E, others same with IOE1
O+C : Open and Close word of chunk

2) MaxE-nt MaxEnt, ALLis, IGTree, SNoW, MBSL, C5.0

3) IOB, IOB1, IOE, IOE1

었다.위 결과는 현재까지의 중국어 최장명사구 식별에서 최고의 성능이지만 재현율이아주 낮음을 알 수 있다. 기존의 영어 기본명사구 연구에서는 학습자질로 품사정보만을 사용하였지만 정확률과 재현율이 낮은 중국어 구 식별에 있어서 품사정보만 사용해서는 중의성을 효과적으로 해소 할 수 없다.

예 5) 의미정보를 이용한 명사구와 동사구 중의성 해소

명사구 : [打算/v/Gb05 逃跑/v/Fb01] 的 囚犯/n/An02
plan to escape DE criminal
한국어 대응문 : 탈옥을 시도하는 죄수

동사구 : 射击/v/Hb06 [跑动/v/Fb01 的 猎物/n/Bi01]
shoot running DE animal
한국어 대응문 : 달리는 동물을 사격하다

학습자질로 품사정보만을 사용한다면 중의성을 해소할 수 없지만 의미정보를 추가한다면 중의성해소가 가능할 것이다. Gb05(계획하다)와 An02(사람)는 결합강도가 아주 약하지만 Hb06(사격하다)와 Bi02(동물)은 결합강도가 아주 강하기 때문에 이런 의미 제약 관계는 학습자질로서 유용하게 쓰일 수 있다.

[표 2.1] 의미 코드

단어	코드	의미정보	단어	코드	의미정보
打算	Gb05	심리활동	囚犯	An02	사람
射擊	Hb06	군사활동	獵物	Bi01	동물

3. 두 단계 학습을 통한 중국어 최장명사구 식별

본 논문은 중국어 최장명사구 식별에서 변형기반의 접근방법을 이용하였다. 일반적으로 최장명사구 안에포함된 구성성분이 일반 명사구보다 훨씬 많기 때문에 학습할 때 주변의 문맥정보를 보다 많이 고려해야 한다는 점을 감안하여 두 가지 접근방법을 시도하였다. 첫 번째 접근방법[15]은 직접 최장명사구 식별을 진행하는 것으로서 주변의 문맥정보에 초점을 맞추는 것이다. 두 번째 접근방법은 먼저 기본구를 식별하고 그것의 중심어(head) 정보를 이용해서 다시 최장명사구의 식별을 진행하는 두 단계 학습접근방법이다.

3.1 접근방법1 : 다양한 윈도우 사이즈에 기반한 SVM 학습모델[15]

관계절이 포함되어 있는 비교적 긴 명사구를 식별하기 위해서는 참조해야 할 윈도우 사이즈가 상대적으로 커야 하고 반면에 기본명사구와 같은 작은 명사구는 작은 윈도우 사이즈에서도 식별이 가능할 것이다.

그러나 일괄적으로 주변 윈도우 사이즈를 크게 잡으면 오히려 작은 명사구에는 정보의 과적용(over fitting)이 일어나 부작용이 클 것이다. 즉 명사구의 길이에 맞는 학습에 적당한 윈도우 사이즈가 있을 것이다. 따라서 윈도우 사이즈에 기반한 서로 다른 학습 모델을 만든다면 이들 각각의 처리 능력이 다를 것이며 따라서 시스템합성을 한다면 보다 좋은 성능을 보일 것이다. 이런 생각에 기인하여 window size 3, 5, 7, 9, 11으로 5가지 다른 SVM 학습모델을 구축하고 시스템 합성을 통하여 성능을 향상하고자 하였다.

표 3.1.1은 O,C,S,N[1] 태그에 대한 정확률과 재현율이다.

[표3.1.1] 각 태그별 정확률과 재현율

#	O		C		N		S	
	P	R	P	R	P	R	P	R
3	**79.2**	65.3	79.3	56.3	62.4	75.6	85.7	93.7
5	77.2	**69.1**	80.8	61.3	64.0	**78.9**	86.2	94.2
7	77.9	68.0	80.6	**60.3**	**64.1**	78.9	86.0	94.2
9	77.6	66.8	**80.9**	59.7	63.8	77.5	**86.0**	**94.3**
11	77.5	66.1	80.6	59.1	63.8	77.5	85.9	94.2

표 3.1.2는 시스템합성을 한 후의 시스템 성능이다.

[표3.1.2] 시스템합성을 한 후의 정확률

전체 태그	O	C	S	N
85.2% (+2.5%)	79.6% (+0.4%)	84.2% (+3.3%)	70.6% (+6.5%)	88.1% (+2.1%)

시스템합성 후 전체 태그 정확률은 단일시스템에서의 최고성능인 82.6%에서 85.2%로 2.6% 향상되었고, 각 태그별 정확률도 따라서 향상되었음을 볼 수 있다. 그러나 시작태그와 종결태그의 재현율은 아직 낮다. 이는 단순히 주변 문맥정보를 늘리는 것으로는 한계가 있음을 보여준다. 때문에 명사구 내부의 구성성분 사이의 제약정보는 반드시 고려되어야 한다.

3.2 접근방법2 : 두 단계 학습을 통한 최장명사구 식별

접근방법1[15]의 결과를 분석해 보면 최장명사구 식별은 단순한 접근방법보다는 여러가지 측면을 종합적으로 고려해야 함을 알 수 있다. 다음은 최장명사구 식별에서 반드시 고려해야 할 3가지 중요한 특성이다.

3.2.1 최장명사구 식별에서 고려해야 할 3가지 특성

첫째, 경계적인 특성(boundary characteristic)

4) O : 명사구의 시작 C : 종결 S : 구성성분이 하나인 명사구 N : 그 외 태그

예 6) 최장 명사구의 왼쪽경계 정보[1]

[VP 赚/v 钱/n] 的 机会/n
벌다 돈 적(의) 기회
한국어 대응문: 돈 버는 기회

[PP 在/p 天津/n] 投资/v 的 外国/n 企业/n
재(에) 천진 투자하다 적(의) 외국 기업
한국어 대응문: 천진에 투자한 외국기업

[MP 一/cd 个/m] 城市/n 的 兴衰/n
하나 개 도시 적(의) 흥망
한국어 대응문 : 한 도시의 흥망성쇄

예 6에서 알 수 있다시피 동사구, 전치사구, 수량사구 등 기본구들은 최장명사구의 왼쪽 경계를 차지하는 경향이 있다. 단어 레벨에서 단순히 문맥 정보를 늘린다는 것은 이러한 특성을 반영하기는 하지만 동시에 단어들 사이에 존재하는 복잡한 중의성으로 인해 그만큼 많은 부작용을 감안해야 한다는 것을 의미한다.

둘째, 최장 명사구 내부의 구성패턴[2]

[표3.2.1.1] 최장명사구의 구성패턴

	패턴	분포
1	NN+NN	25.6%
2	VA+NN	10.0%
3	PN+NN	9.5%
4	M+NN	7.9%
5	VV+NN	5.9%
6	M+CD+NN	5.3%
7	VA+DE+NN	5.2%
8	NN+NN+NN	4.7%
9	NR+NN	4.5%
10	PN+CD+NN	4.5%
11	NN+DE+NN	4.4%
12	PN+DE+NN	3.3%
13	JJ+NN	3.1%
14	AD+VA+DE+NN	1.8%
15	NN+NN+DE+NN	1.2%
16	V+DE+NN	1.0%
17	VA+DE+NN+NN	0.8%
18	NR+DE+NN	0.6%
19	VV+NR+DE+NN	0.4%
20	… …	…

표 3.2.1.1은 최장명사구 내부 구성패턴에 대한 통계정보이다. 품사정보만을 고려한다면 당연히 내부 구성성분 사이에는 중의성이 존재하겠지만 그 구성패턴을 분석해보면 문법적으로 일정한 그룹으로 묶여지는 경향이 있음을 알 수 있다. 단어레벨에서 구성성분 사이의 패턴을 분석할 때의 문제점은 최장명사구의 길이가 아주 길 때에 참조할 수 있는 문맥정보의 크기에 한계가 있기 때문에 생긴다. 반면에 구레벨에서 접근하면 참조하는 문맥의 크기가 상대적으로 작아도 단어레벨과 비교할 때 참조되는 범위가 크기 때문에 이런 문제에 대한 보다 높은 처리능력을 가질 수 있을 것이다.

셋째, 공기 정보(co occurrence information)

예 7) 공기 정보

① 保. /v [NP 我. 伟. 的 祖. /n]
수호하다 우리 위대하다 적(의) 조국
한국어 대응문: 우리의 위대한 조국을 수호하다

② 在/p [NP 明媚 的 .光 下/l]
재(에) 눈부시다 적(의) 햇살 아래
한국어 대응문: 찬란한 햇살아래에서

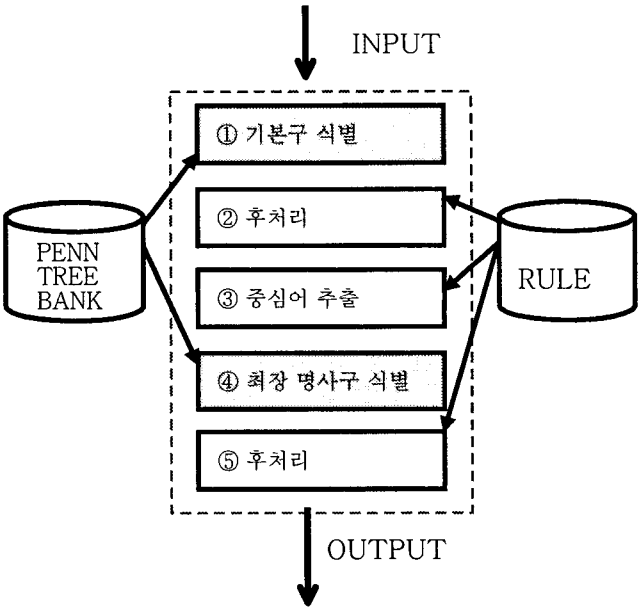
예 7에서 동사 '保卫(수호하다)'는 명사구의 왼쪽 경계로서 명사구 내의 '祖国(조국)'과 아주 강한 결합관계를 가진다. 마찬가지로 전치사 '在(에)'는 명사구 내의 오른쪽 경계인 '下(아래)'와 아주 강한 결합관계를 가진다. 이러한 특성은 명사구의 왼쪽경계를 결정하는데 아주 유용한 자질로 사용될 수 있다. 왼쪽경계 단어와 결합관계를 가지는 단어는 구 내의 오른쪽경계 단어 즉 중심어이다. 따라서 명사구 내의 중심어를 수식하는 형용사나 기타 수식 절은 이러한 결합관계를 고려할 때에는 필요 없는 정보이며 따라서 구레벨 접근방법을 사용한다면 중복된 정보를 피하고 학습에 유용한 자질을 추출해 낼 수 있다. 중국어 최장명사구 식별에서 위의 3가지 측면을 염두에 둘 때 단어 레벨에서의 학습보다는 상대적으로 적은 윈도우 사이즈에서도 좋은 성능을 낼 수 있는 구레벨 접근방법이 보다 효과적인 방법임을 알 수 있다.

3.2.2 시스템 설계

먼저, 기본구와 최장명사구가 태깅된 코퍼스에 대한 학습을 통해 기본구 식별모델과 최장명사구 식별모델을 구축한다. 최장명사구 모델의 구축은 구레벨에서 진행된다. 품사정보와 의미정보가 태깅된 문장이 들어오면 먼저 기본구 식별모델을 통하여 기본구가 태깅된다. 규

5) NP : 명사구 VP : 동사구 SP : 단문구 PP : 전치사구 LP : 방향사구
6) NN : 일반명사 VA : 형용사 AD : 부사 PN : 대명사 M : 양사 VV : 동사 CD : 수사 DE : NR : 고유명사 JJ : 구별사(술어가 될 수 없는 형용사), L : location **자세한 내용은 참고논문[16] Penn TreeBank 4.0 참조

칙에 기반한 후처리 작업을 통하여 기본구의 좌우 경계를 조정한다. 다음 1단계에서 식별한 기본구에서 중심어를 추출해 내고 기본구 정보와 중심어 정보를 이용하여 최장명사구 식별을 진행한다. 최장 명사구 모델을 통해 얻은 결과 역시 2단계와 비슷한 방법으로 좌우경계를 조정한다.



[그림 3.2.2.1] 시스템 개요

3.2.2.1 기본구 식별모델 구축

영어에서의 기본명사구는 다른 명사구를 내포하지 않으면서 문법적으로 연관되는 비 중첩 비 내포의 명사구를 뜻한다[1]. 영어와는 달리 중국어는 기본명사구에 대한 표준화된 정의가 없으며 시스템마다 조금씩 다르다.[13,14] 본 실험에서는 기본구는 다른 기본구를 포함하지 않는다는 원칙을 지키고 있으며 시스템의 수요에 따라서 기본구의 종류와 정의를 확장하였다. 본 실험에서 식별할 기본구로는 명사구, 동사구, 형용사구, 부사구, 수량사구, 단문구, 전치사구, 방향사구 등 8가지이다. 그 중 단문구와 전치사구는 기존의 다른 시스템과 달리 새로 추가한 기본구이다.

그중 단문구는 영어나 한국어 같은 경우에는 보통 관계절로 인식되지만 형식형태소가 발달하지 않은 중국어에서는 구레벨로 접근하는 것이 타당하다고 생각한다.

예8) 최장 명사구 내부의 단문구

[MNP [SP 国家/n 利用/v] [NP 外国 政府 贷款]] 来自
국가 이용하다 외국 정부 차관 왔다
한국어 대응문: 국가가 이용한 외국정부의 차관은 …에서 왔다.

예 8에서 [단문구 國家(국가) 利用(이용하다)] 는 명사구 안에서 貸款(차관)을 수식하는 단문이다. '국가'와 '이용하다'를 각각 기본명사구와 기본동사구로 태깅하기보다는 전체를하나로 묶고 거기서 중심어 '이용하다'를 추출하는 것이 중의성을 해소하는데 바람직하다. 같은 맥락에서 기본전치사구를 구 식별에 추가하였다.

예 9)구 내부의 구성성분 확장

[VP 均/d 获得/v 过/as]
모두 얻다 었
한국어 대응문: 모두 얻었었다

예9도 마찬가지로 **'부사+동사+조사'** 조합을 부사구와 동사구로 세분하는 기존의 다른 시스템과는 달리 위의 패턴에서 중심어는 동사이고 앞의 부사와 뒤의 조사는 명사구 식별에서 상대적으로 필요 없는 정보라고 판단하여 전체를 동사구로 묶었다.

[표3.2.2.1.1] 각 기본구의 태그 표기형식

태그	설명
BASE-O	기본구의 시작
BASE-C	기본구의 종결
BASE-I	시작과 종결 사이의 태그
NUL	기타

접근방법1[15]과는 달리 하나의 단어로 구성된 기본구는 식별에 의미가 없고 구문분석에 도움이 안되기때문에 학습에서 배제하였고 시작태그와 종결태그 사이의 단어들을 특별히 I 태그로 구별하였다.8가지의 기본구와 각 종류마다 3개 타입을 갖고 있기에 전체 25(8*3+1)가지의 클래스 식별문제로 변환하였다.

학습자질로는 왼쪽2개 단어와 오른쪽 3개 단어의 품사정보,의미정보를 이용하였다. 그리고 전 단어의 기본구 태그가 현재 단어의 타입을 결정하는데 중요한 역할을 할 수 있다[3]는 점을 감안하여 바로 전 단어의 기본구 태그를 자질로 사용하였다. 예를 들어 전 단어가 NP-O로 태깅되었다면 현재단어는 NP-I 혹은 NP-C 로 태깅될 확률이 상대적으로 높을 것이다.

3.2.2.2 중심어 추출

[표 3.2.2.2.1] 중심어 추출 규칙

기본구	중심어	예문
명사구	***Rightmost word*** NN+NN, JJ+NN VA+DE+NN,…	銀行(은행) 帳戶(계좌) 管理(관리)
동사구	**1.V+N(CD)→ V**	進行(진행하다) 合作(합작),
	2.V+V → **leftmost V**	開發(개발하다) 生產(생산하다)
	3.(AD)+V+(AS)→ V	剛剛(방금) 好(좋아졌다)
수량사구	***Rightmost word*** CD+M, DT+M,…	一(한) 輛(대)
부사구	***Rightmost word*** AD+ AD,…	几乎(거의) 成倍(배로)
형용사구	**Rightmost word** JJ+JJ,CD+JJ,…	大型(대형의) 國有(국가의)
전치사구	**Rightmost word** P+NN,P+PN,…	依(의하여) 法(법)
방향사구	**Leftmost word** NN+L,…	國際(국제) 上(상)
단문구	**Rightmost word** NN+VV,NN+VA,..	國家(국가) 利用(이용하다)

표3.2.2.2.1은 기본구의 정의와 코퍼스에 근거하여 만든 중심어 추출 규칙이다. 명사구나 동사구의 대등접속에서 구성성분들의 의미정보나 품사정보가 비슷하기 때문에 가장 오른쪽단어를 중심어로 결정한다. '동사+명사' 유형의 동사구와 단문구에서 중심어는 동사이고 목적어나 주어는 필요 없는 정보이다. 전치사구의 전치사와 방향사구의 방향사 정보는 3.2.1에서 알 수 있다시피 명사구의 왼쪽 경계와 오른쪽 경계를 결정하는데 중요한 자질로 사용된다. 명사구 학습에서 구 정보를 이미 자질로 사용하기 때문에 구 정보에서 전치사나 방향사 정보가 충분히 반영되었다고 보고 이 두 유형의 구에서는 구 안의 명사를 중심어로 결정한다. 중심어 추출을 거친 후 최장명사구의 구성성분 개수는 절반 이상으로 줄어들며 대부분 명사구들이 참조 가능한 문맥의 범위에 들어오게 된다.

3.2.2.3 최장 명사구 식별 모델 구축

중심어가 추출된 기본구의 정보를 이용하여 최장명사구식별모델을 구축한다. 식별모델은 기본구 식별모델 구축과 비슷하며 다른 점은 전자는 단어 레벨에서 학습하지만 후자는 구 레벨에서 학습을 진행한다는 것이다.

[표3.2.2.3.1] 최장명사구 태그 표기형식

태그	설명
MNP-O	최장명사구 시작
MNP-C	최장명사구 종결
MNP-I	최장명사구 시작과 종결사이의 태그
MNP-NUL	기타 태그

최장명사구 식별은 5가지의 클래스의 식별문제로 변환된다. 학습자질로는 왼쪽의 2개 기본구와 오른쪽의 3개 기본구의 중심어의 품사정보와 의미정보, 그리고 기본구의 구 타입, 바로 전 단어의 최장명사구 태그 등을 사용하였다. 구레벨로 접근하면서 단어레벨보다 문맥정보의 참조범위가 커졌기 때문에 기본구식별에서와 마찬가지로 참조할 윈도우 사이즈는 왼쪽 2개, 오른쪽 3개로 고정하였다. 윈도우 사이즈를 늘리는 것은 시스템의 성능 대 수행속도를 비교해볼 때 별로 바람직한 방법이 아니라고 생각된다.

3.2.2.4 기본명사구와 최장명사구의 좌우경계 조정을 위한 후처리 작업

① 시작태그 수정규칙

IF $Y\text{-}C_{-1}$,(NUL_0 | $X\text{-}I_0$),$X\text{-}C_{+1}$
THEN (NUL_0|$X\text{-}I_0$) → X-O

만일 바로 전 태그가 Y타입의 종결태그이고 태그가 NUL[1] 타입이거나 X타입의 가운데 태그이고 다음 태그가 X타입의 종결 태그이면 현재 태그를 X타입의 시작 태그로 수정한다.

IF $Y\text{-}C_{-2}$,NUL_{-1},NUL_0,$X\text{-}C_{+1}$
THEN NUL_0 → X-O

만일 왼쪽 두번째 단어가 Y타입의 종결태그이고 바로 전 단어가 NUL 태그이고 현재 단어가 NUL태그이고 바로 다음 단어가 X타입의 종결태그이면 현재태그를 X타입의 시작태그로 수정한다.

② 종결 태그 수정규칙

IF $Y\text{-}O_{-1}$,(NUL_0 | $Y\text{-}I_0$),$X\text{-}O_{+1}$
THEN (NUL_0|$Y\text{-}I_0$) → Y-C

만일 바로 전 태그가 Y타입의 종결태그이고 현재 태그가 NUL 태그이거나 Y타입의 중간태그이고 바로 다

7) NUL 태그는 기본구 식별에서는 어떤 구에도 포함되지 않는 단어 즉 BP NUL 태그이며 최장명사구 식별에서는 시작도 끝도 가운데도 아닌 기타 태그 즉 MNP-NUL태그를 의미한다.

음 태그가 X타입의 시작태그이면 재태그를Y타입의 종결태그로 수정한다.

IF $Y\text{-}O_{-2}$,$Y\text{-}I_{-1}$,$Y\text{-}I_{0}$,$X\text{-}O_{+1}$
THEN $Y\text{-}I_{0}$ → Y-C

만일 왼쪽 두번째 단어가 Y타입의 시작태그이고 바로 전 단어가 Y타입의 중간태그이고 현재 단어가 Y타입의 중간태그이고 바로 다음 단어가 X타입의 시작태그이면 현재태그를 Y타입의 종결태그로 수정한다.

③ 중간태그 수정규칙

IF $X\text{-}O_{-1}$,$X\text{-}C_{0}$,$X\text{-}C_{+1}$
THEN $X\text{-}C_{0}$ → X-I

만일 바로 전 태그가 X타입의 시작태그이고 현재태그가 X타입의 종결태그이고 바로 다음 단어가 X타입의 종결태그이면 현재태그를 X타입의 중간태그로 수정한다.

위의 수정규칙은 아직 완성되지 않았으며 코퍼스의 통계정보를 이용하여 후처리 작업 규칙을 추가하려고 한다.

4. 실험 결과

실험은 기본구 식별과 최장명사구 식별 두가지로 나누었다. 두 실험 모두 Chinese Penn Tree Bank 4.0[16]을 학습코퍼스로 사용하였다. 기본구와 최장명사구가 태깅된 학습코퍼스는 Penn Tree Bank를 위의 기본구 정의에 따라 자동으로 구축하였다. 구성성분이 하나인 구는 기본구 식별에서 배제하였다. 실험 코퍼스는 8000문장,22만 단어이다 최장명사구는 13087개이고 평균길이는 7이다. 기본구와 최장명사구의 후처리 실험은 아직 포함되지 않았으며 최장명사구 식별은 기본구가 이미 정확하게 태깅된 코퍼스를 학습 코퍼스로 사용하였다. 기본구 식별의 정확도가 비교적 높고 중심어 추출 방식의 최장명사구 식별방법의 타당성을 검증하는 차원에서 실험을 진행하였다. 기계학습 classifier는 Naïve Bayes[12], Decision Tree[12] 등 2가지를 사용하였다. 의미정보와 전 단어 태깅정보가 학습자질로 사용되었을 때의 성능 개선여부를 판단하기 위하여 기본구 학습에서는 품사정보만 사용했을 때, 의미정보를 추가하였을 때, 전 단어 태깅정보를 추가하였을 때,3가지로 실험을 진행하였다. 최장 명사구 학습에서는 중심어의 품사정보와 의미정보를 사용하였을 때, 기본구의 타입정보를 추가하였을 때, 전 노드의 최장명사구 태그정보를 추가하였을 때 3가지로 나누어 실험을 진행하였다. 테스트는 5 cross validation으로 진행하였다.

4.1 기본구 식별결과

[표 4.1.1] 기본구의 각 classifier 별 정확률

	Naïve Bayes	Decision Tree
POS[1)]	76.9%	88.6%
+SEM	78.6%(+1.7)	90.5%(+1.9)
+prv.Tag[2)]	88.6%(+11.7)	96.3%(+7.7)

표 4.1.1에서 알 수 있다시피 의미정보를 추가하였을 때가 품사정보만 사용했을 때보다 성능이 다소 향상되었으며 전 단어의 태그를 자질로 추가했을 때는 품사정보만 사용했을 때보다 Naïve Bayes에서는 11.7%, Decision Tree에서는 7.7%로 현저하게 향상되었음을 볼 수 있다. 이로부터 전 단어의 태그정보가 기본구 식별에 있어서 아주 중요한 자질임을 알 수 있다. 그리고 Decision Tree가 Naïve Bayes보다 좋은 성능을 보여주고 있다.

[표 4.1.2] Decision Tree에서의 기본구 각 태그별 실험결과

	정확률	재현율	F-measure
NP-O[3)]	93.9%	96.5%	95.2%
NP-C	98.1%	97.0%	97.5%
NP-I	94.9%	96.3%	95.6%
VP-O	93.3%	92.7%	93.0%
VP-C	97.3%	98.3%	97.8%
VP-I	92.6%	90.5%	91.5%
AP-O	94.2%	91.5%	92.8%
AP-C	98.5%	99.2%	98.8%
AP-I	97.3%	95.5%	96.4%
DP-O	96.2%	94.8%	95.5%
DP-C	99.9%	99.5%	99.7%
DP-I	96.5%	97.9%	97.2%
MP-O	99.1%	99.2%	99.1%
MP-C	99.4%	99.9%	99.6%
MP-I	98.7%	93.4%	96.0%
SP-O	85.3%	84.1%	84.7%
SP-C	98.3%	100%	99.1%
SP-I			
PP-O	88.2%	88.2%	88.2%
PP-C	100%	100%	100%
PP-I			
LP-O	96.6%	98.3%	97.4%
LP-C	100%	100%	100%
LP-I			
BP-NUL	96.9%	96.6%	96.7%

8) POS : 품사정보이다. 본 실험에서는 펜 트리뱅크의 품사체계를 그대로 사용하였다. SEM : 의미정보이다.

9) prv.Tag 는 바로 전 단어의 기본구 태그를 의미한다

표 4.1.2는 Decision Tree에서 모든 기본구 태그의 정확률과 재현율, F measure에 대한 실험결과이다. 단문구(SP)의 시작태그와 전치사구(PP) 시작태그를 제외한 모든 태그는 F measuer가 90% 이상 이었으며 대부분 태그가 95%이상 이었다. 단문구의 시작태그의 정확률과 재현률이 낮은 원인은 단문구와 동사구, 명사구 사이에 중의성이 존재하기 때문이다. 대부분 기본구의 시작태그의 성능이 종결태그보다 낮으며 반면에 종결태그와 중간태그의 성능은 아주 높다. 이는 후처리 작업을 통하여 시작태그의 성능을 개선할 수 있음을 보여주며 후처리 작업은 시작태그의 수정에 초점을 맞춰야 함을 알려준다. 기본구 식별의 성능이 후처리 작업 없이 96% 이상에 도달하였으므로 다음 단계인 최장명사구 식별모델의 입력으로서 이용가치가 충분히 있다.

4.2 최장명사구 식별 결과

[표 4.1.3] 최장명사구의 각 classifier별 전체태그의 정확률

	Naïve Bayes	Decision Tree
POS+SEM	81.9%	93.1%
+baseTag[1)]	83.7%(+1.8)	93.2%(+0.1)
+prv.Mtag[2)]	91.2%(+9.3)	96.5%(+3.4)

선택된 자질은 기본구 모델과 비슷한 양상을 보인다. 기본구 타입정보를 학습자질로 이용했을 때가 중심어의 품사와 의미정보만 이용했을 때보다 성능이 다소 향상되었으며 전 노드의 최장명사구 태그정보를 학습자질로 추가했을 때에 성능이 Naïve Bayes에서는 9.3%, Decision Tree에서는4% 향상되었음을 볼 수 있다.

4.2.1 실험결과 비교

[표4.1.4] Decision Tree에서의 최장명사구 각 태그별 결과

	정확률	재현율	F-measure
M-O	94.4%	91.6%	93.0%
M-C	95.8%	96.8%	96.3%
M-I	98.6%	97.3%	97.9%
MNP-NUL	97.1%	99.3%	98.2%

표 4.1.4는 Decision Tree에서의 최장명사구 각 태그별 정확률과 재현율을 보여준다. 역시 시작태그의 정확률이 종결태그보다 낮음을 알 수 있다. 접근 방법1[15]과 비교했을 때 정확률과 재현율이 11.3% 와 20.6%로 현저하게 향상되었다. 아직 좌우경계 매칭 실험을 하지 않았기 때문에 직접적인 비교는 불가능하지만 Zhou[2]의 논문에서 재현율이 61.7%것을 감안할 때 최장명사구 태그의 평균 F measure 96.4%는 아주 고무적인 결과이다.

10) NP:명사구 VP:동사구 AP:형용사구 DP:부사구 MP:수량사구 SP:단문구 PP:전치사구 LP: 방향사구 BP NUL: 기본구에 속하지 않은 기타 태그 O: 시작 C: 종결 I:중간

11) baseag는 기본구의 타입 정보를 의미한다.

12) prv.Mtag는 바로 전 노드의 최장명사구 태그를 의미한다.

5.결론 및 향후 작업

식별의 두 단계는 후처리 작업을 거치지 않은 상태에서 각각 평균96.0%와 96.4%의 F measure 성능을 보여 주었다. 본 실험은 기본구가 식별이 100% 정확하다고 가정하고 진행하였기 때문에 기본구 식별 성능이 96%인 것을 감안할 때 시스템의 성능은 대략 92.5%(96.% * 96.4%) 이다. 각 단계에서 후처리 작업을 진행한다면 성능향상이 가능하며 전체 시스템의 성능을 95% 대로 더 끌어올릴 수 있다. 이 실험은 두 단계 학습을 통한 중국어 최장명사구 자동식별방법이 한 단계로 직접 식별하는 방법보다 현저하게 좋은 성능을 보여주는 접근방법임을 보여 주었다. 향후 작업으로는 코퍼스의 통계정보와 문법정보를 이용한 규칙기반의 후처리 작업과 좌우경계 매칭(matching) 작업을 완성하는 것이다. 본 실험에서는 Naïve Bayes 와 Decision Tree를 사용하였는데 Neural Network이나 SVM과 같은 다른 학습 툴들을 사용하여 시스템 합성을 통한 성능향상을 기대해 볼 수 있다.

감사의 글

본 연구는 첨단정보기술 연구센터(AITrc)를 통하여 과학재단의 지원을 받았습니다.

6. 참고문헌

[1] Steven P. Abney, "*Parsing by Chunks*", In Principle Based Parsing, pages 257 278. Kluwer Academic Publishers, Dordrecht, 1991

[2] Zhou Qiang, Sun Maosong and Huang Changning "*Automatically Identify Chinese Maximal Noun Phrase*", 1998

[3] Erik F.Tjong Kim Sang &al "*Applying System Combination to Base Noun Phrase Identification*" In : Proceedings of COLING 2000, 857 863,2000

[4] Taku Kudo and Yuji Matsumoto,"*Chunking with Support Vector Machines*", In NAACL, 2001

[5] Lance A. Ramshaw and Mitchell P. Marcus, "*Text*

Chunking using Transformaion Based Learning". In Proceedings of the Third ACL Workshop on Very Large Corpora.Cambridge, MA, USA,1995

[6] 梅家駒, 竺一鳴 & al 《同義詞詞林》, 上海辭書出版社, 上海, 1983

[7] Georage Forman, "*An Extensive Empirical Study of Feature Selection Metrics for Text Classification*" In : The Journal of Machine Learning Research Volume 3 , March 2003

[8] Urich. H. GKreBel. "*Pairwise Classification and Support Vector Machines*" In Advance in Kernel Mathods.MIT Press

[9] Chin Chung Chang and Chih Jen Lin. "*a Library for Support Vector Machines*" , http : //www.csie.ntu.edu.tw/~cjlin/libsvm/ November 12, 2003

[10] Academia Sinica Balanced Corpus of Modern Chinese, http : //www.sinica.edu.tw/

[11] Zhan Weidong "*A Study of Constructing Rules of Phrases in Contemporary Chinese for Chinese Information Processing*" 2000

[12] WEKA machine learning toolkit http : //www.cs.waikato.ac.nz/~ml/

[13] Huang Changning "*Chinese baseNP detection*" 1999

[14] Zhang Yuqi and ZhouQiang "*Chinese Base Phrase Chunking*" 2000

[15] Changhao Yin "*Identification of Chinese Maximal Noun Phrase on Different Context Size Settings Using SVMs*" 2004

[16] Penn TreeBank 4.0 http : //www.cis.upenn.edu/~chinese/