문장부호 정보와 확장된 청크에 기반한 중국어 최장명사구 식별

백설매[0] 김미훈 이금희 정유진 이종혁
포항공과대학교 정보통신대학원 정보처리학과[0], 포항공과대학교 컴퓨터공학과
{xuemei[0], meixunj, ljj, prizer , jhlee}@postech.ac.kr

Maximal Length Noun Phrase Identification Based on Punctuations and Expanded Chunk

Xue-Mei Bai[0], Mei-Xun Jin, Jin-Ji Li, You-Jin Chung, Jong-Hyeok Lee
Dept. of Graduate school for information technology[0],
Dept. of Computer Science and Engineering, Pohang University of Science and Technology

요 약

명사구는 기본명사구와 최장명사구로 분류된다. 최장명사구에 대한 정확한 식별은 문장의 전체적인 구문구조를 파악하고 문장의 정확한 지배용언을 찾아내는데 중요한 역할을 수행한다. 본 논문에서는 확장된 청크(chunk) 개념과 다섯 개의 클래스로 세분화된 문장부호 정보를 사용한 최장명사구 식별 기법을 제안한다. 제안된 기법은 기본모델(baseline)보다 4.05% 향상된 평균 88.63%의 우수한 F-measure 성능을 보인다.

1. 서론

Abney[1]가 구 묶음(chunking) 작업에 의한 전처리 과정을 거친 뒤 구문분석을 진행하는 방법을 제시한 후 많은 구문분석 연구에서 이 방법을 사용하였다. 이는 구 묶음을 통하여 구문분석의 복잡도를 줄이고자 하는 목적에서이다. 중국어에서도 구 묶음을 구문분석의 전처리 단계로 진행하면 구문분석을 더 효율적으로 진행할 수 있다.

인지과학의 입장에서는 사람들이 문장을 정확하게 인식하려면 먼저 문장에 포함된 실체나 개념 등을 정확하게 인식해야 한다고 한다. 이런 실체나 개념을 구성하고 있는 것은 대부분 명사이다. 때문에 구 묶음의 여러 가지 구 단위에서도 특히 명사구에 대한 정확한 식별은 구문분석뿐만 아니라 기계번역이나 정보검색, 정보추출과 같은 다양한 자연언어처리 분야들에서 매우 중요한 문제로 인식되고 있다.

구조적인 측면에서 볼 때 명사구는 크게 최단명사구(Minimal length Noun Phrase), 기본명사구(Base Noun Phrase), 최장명사구(Maximal length Noun Phrase, 이하 MNP로 약칭)의 3가지[2]로 분류할 수 있는데, 기본명사구와 최단명사구를 동일시 하는 관점도 있다. 중국어의 기본명사구에 대한 정의는 연구자에 따라 약간씩 차이가 있지만(제3장 참조), 최장명사구는 일반적으로 "다른 명사구에 포함되지 않는 명사구"로 정의된다. 최장명사구는 그 안에 포함될 수 있는 단어들이 매우 다양할 뿐만 아니라 장거리 의존성(long dependency) 문제도 포함되기 때문에 최단명사구나 기본명사구에 비해 정확한 식별이 힘들다.

하지만 최장명사구 식별은 구문분석의 복잡도를 크게 낮출 수 있을 뿐만 아니라 특히 중국어에서는 정확한 지배용언을 찾아내는데 도움이 되기에 꼭 필요한 작업이다.

본 논문에서는 보다 효과적인 최장명사구 식별을 위해 기존의 구 묶음(chunking) 단계에서 수행하게 되는 일반적인 묶음 단위를 보다 확대시킨 확장된 청크(chunk)(제4장 1절 참조) 개념을 도입한다. 또한 중국어에서 사용되고 있는 다양한 문장부호들의 특성에 주목하여 이들의 세분화된 정보를 최장명사구의 식별에 활용한다.

2. 관련 연구

최장명사구의 식별은 언어 분석을 위한 중간 처리 과정으로서 그 중요성 때문에 많은 연구자들의 관심 연구대상이 되고 있다. 먼저 다른 언어권에서의 최장명사구 식별에 대한 기존연구들은 다음과 같다.

1) Bourigault[3]는 규칙기반 방법으로 불어에서의 최장명사구 식별을 하였는데 정확한 성능은 언급되지 않았다.
2) Voutilainen[4]은 최장명사구를 식별함에 있어서 'NP-hostile' 과 'NP-friendly' 의 두 유한상태 오토마타(finite state automata)를 사용하였는데 정확률과 재현율을 각각 95-98%, 98.5-100% 로 보고하였다. 하지만 Chen[5]은 이 논문의 부록에 수록된 예제들에 대한 성능 평가를 직접 수행한 결과 실제 재현율은 85%에 머물렀으며 또한 명사구의 정의에도 일관성이 없음을 지적하였다.
3) Chen[5]은 통계적 방법으로 문장을 청크(chunk) 단위로 나눈 다음 유한상태 오토마타로 명사구를 식별하는 혼합형(hybrid) 기법을 제안하였다.

한편 중국어에서 최장명사구 식별에 대한 기존연구들은 다음과 같다.

1) Li[6]는 단어의 품사 태그 정보를 이용하여 통계적 방법으로 최장명사구의 자동 식별을 시도하였는데 정확률과 재현율이 각각 71.3 %, 69.1% 였다.
2) Tse[7]는 통계와 규칙의 혼합 기법을 사용하였다. 이 논문에서는 "的" [1]를 포함한 명사구만을 식별하였는데 정확률과 재현율을 각각 75%와 90%로 보고하고 있다.
3) Zhou[2]는 최장명사구 식별을 두 단계로 나누어 진행하였다. 먼저 코퍼스의 통계정보를 이용하여 문장부호, 공기정보, 등위접속구조의 좌우경계를 식별해 내고, 규칙기반의 방법으로 최장명사구의 오른쪽 경계를 식별한 후 계속하여 왼쪽으로 처리범위를 확장해 나가면서 최장명사구의 왼쪽 경계를 식별하였다. 이 방법은 다섯 단어 이상의 최장명사구에 대하여 정확률 85.4%, 재현율 82.3%의 성능을 보여주었다.
4) Yin[8]은 두 단계 학습모델을 사용하였는데 먼저 기본구를 식별한 다음 기본구의 중심어를 추출하여 최장명사구 식별을 진행하였다. 그러나 기본구에 대한 정의를 명확하게 제시하지 않았으며, 기본구의 식별에 의한 최장명사구 식별 성능 향상도가 그리 높지 않았다.

위에 기술된 기존 연구들은 대부분 간단한 규칙기반이나 통계기반 방법론에 머물렀으며 기계학습 알고리즘을 사용한 방법론도 제시되었으나 주목할만한 성능을 얻지는 못하였다. 그리고 기존의 연구들에서는 Zhou[2]가 전처리 단계에서 몇 개의 문장부호를 단어의 좌우경계를 식별하는데 사용했을 뿐, 최장명사구의 식별을 위한 중요한 자질로서 문장부호를 사용한 기존연구는 제시되지 않았다. 또한 기본구 식별을 최장명사구 식별의 전 단계로 사용한 경우는 있지만 성능 향상에 도움을 주지는 못하였으며, 확장된 청크 개념을 도입한 적은 없었다.

3. 중국어 구에 대한 정의

영어에서 기본구는 하나의 중심어(head)와 여러 개의 pre-modifiers로 구성될 수 있지만 post-modifiers나 arguments는 가질 수 없다는 규정 하에 명사구, 동사구, 전치사구, 부사구, 종속어구, 형용사구, particle phrase, 접속사구, 감탄사구, list marker phrase, unlike coordinated phrase 등 11개 유형으로 설정하였다[9]. 하지만 중국어에서는 아직도 기본구에 대한 하나의 통일된 정의가 없으며 시스템마다 다르게 구현하고 있다. 아래는 중국어 기본구에 대한 여러 가지 정의이다.

1) Penn Chinese Treebank 에서는 15개의 구를 정의하여 사용하였다. Tan[10]은 중국어 기본구를 10가지로 정의하였는데 Penn Chinese Treebank에서 정의한 구 가운데서 동사구, 관형사구, 형용사구, 수량사구, FRAG[2], 명사구, 전치사구, 방향사구, 부사구, 양사구 10개를 사용하였다. 이 분류는 Penn Chinese Treebank 코퍼스에 지나치게 의존하고 있기 때문에 다른 코퍼스에서는 존재하지 않는 구도 포함하고 있다는 문제점을 갖고 있으며 기본구에 대한 정의도 명확하지 않다.
2) Zhang[11]은 명사구, 동사구, 형용사구, 구별사구, 부사구, 시간사구, 방향사구, 수량사구, 양사구 등 9가지의 기본구를 사용하였는

[1] '的' 는 중국어에서 수식어와 명사를 연결하는 조사이다.

[2] FRAG: FRAGMENT의 약자로 단장(段章)의 단어를 frag란 구로 따로 표기하였다.

데 아래와 같은 식으로 표현하고 있다.
{modifier}* + head + [complement]* or coordinate structure
다만 여기에서 "modifier + head" 혹은 "head + complement" 는 가능하지만 "modifier + head + complement" 는 허용되지 않는다는 부가적인 제약을 두고 있다.

3) Zhao[12]는 Zhao[13]의 기본명사구의 정의에 근거하여 기본구 정의를 내렸다. 즉 기본구는 단어나 다른 기본구로 구성될 수 있지만 같은 유형을 내포 할 수는 없다는 것이다. 이 논문에서는 형용사구, 부사구, 명사구, 시간사구, 방향사구, 동사구와 수량사구의 7가지 기본구를 정의하여 사용하였다.
4) Yin[8]은 명사구, 동사구, 형용사구, 부사구, 수량사구, 단문구, 전치사구, 방향사구의 8가지 기본구를 정의하였는데 주요하게는 Zhao[13]의 정의에 따랐고 시스템의 필요에 따라 수정하여 사용하였기에 기본구에 대한 명확한 정의가 없다.

위와 같은 기본구에 대한 다양한 설정들이 있지만 본 논문에서는 영어의 기본구 정의를 채택하였다. 기본구 유형으로는 명사구, 동사구, 형용사구, 시간사구, 전치사구, 수량사구, 부사구, 방향사구, 접속사구, list marker 등의 10가지로 설정하였다. 이 중에서 방향사구, 전치사구, 접속사구 list marker 등은 최장명사구 식별에 도움을 주지 못하기에 논문에서는 사용을 배제하고 나머지 6개의 기본구만을 최장명사구 식별에 사용하였다. 그리고 명사구는 고유명사구와 일반명사구로, 동사구는 일반동사구와 계사구, "有" 동사구로 좀 더 세분화시켜 사용하였다.

4. 확장된 청크(chunk)의 개념

기존의 chunk 는 품사와 구문관계에 따라 분류하고 정의하였기 때문에 문장 중에 흔히 등장하는 공기(co-occurrence) 패턴이나 따옴표 ("") 등의 문장부호에 의해 묶인 덩어리를 표현할 수 있는 방법이 없다. 본 논문에서 사용하고 있는 '확장된 청크(chunk)' 는 바로 이런 최장명사구 식별 시 하나의 단위로 처리해 줄 수 있는 단어 집단들까지 포괄함으로써 기존의 청크 개념을 확장시킨 것이다. 확장된 청크에는 기본구, 공기 패턴, 인용부호 및 slight pause mark 에 의해 묶인 단어들이 모두 포함된다.

확장된 청크 개념의 사용은 최장명사구에 포함되거나 포함되지 않는 단어들을 미리 묶어줌으로써 이후 처리단계에서 소요되는 불필요한 과정을 줄이고 처리의 효율성을 높이는데 목적이 있다. 또한 이렇게 함으로써 처리과정에서 보다 많은 주변 문맥 정보를 참조하여 처리를 수행할 수 있게 한다. 그러나 일단 하나의 청크로 묶이게 되면 이들은 하나의 단어처럼 처리되기 때문에 잘못된 청크의 설정은 최장명사구 식별에 오히려 악영향을 미치게 된다. 따라서 보다 높은 정확률 획득을 위해 최장명사구 식별을 위한 부가적인 청크의 확장은 규칙 기반으로 수행되도록 구성하였으며, 애매하거나 틀릴 가능성이 있는 규칙은 모두 제외하도록 하였다. 그리고 이때 동일한 청크에 포함된 단어들은 하나의 단위로 간주되어 처리되기 때문에 각 단어들이 갖는 특성 정보를 놓칠 가능성이 있다. 따라서 확장된 청크 내부 단어들의 시작과 종결, 그리고 중간 태그 정보를 유지하여 이들 정보를 활용함으로써 보다 정확한 최장명사구의 식별이 이루어지도록 구성하였다. 다음은 기본구를 제외한 최장명사구 식별을 위해 확장하게 되는 청크들의 예이다.

1) "" 로 묶인 인용문

(上海/NR 浦东/NR) 不/AD 是/VC 简单/VA 的/DEV 采取/VV ([**"/PU** 干/VV 一/CD 段/M 时间/NN , /PU 等/P 积累/VV 了/AS 经验/NN 以后/LC 再/AD 制定/VV 法规/NN 条例/NN **" /PU**] **的/DEC** 做法/NN) 。/PU[3]

위의 예문에서 괄호 ()로 묶인 부분이 최장명사구이며 꺾쇠 []로 묶인 부분이 확장된 청크이다. 이 예문에서 볼 수 있다시피 인용부호 "" 로 묶인 부분에서 밑줄 친 부분들은 품사가 명사가 아니기 때문에 일반적으로 명사구에 포함되는 경향이 낮으며, 또한 "" 안의 문장은 그 자체로서 하나의 문장을 이루고 있기 때문에 인용부호로 묶여있다는 사실을 인지하여 부가적인 처리를 수행해 주지 않는 한 하나의 최장명사구로 식별되기가 매우 힘들다. 따라서 인용부호로 묶인 "" 안의 문장을 하나의 청크로 간주하여 미리 이를 하나로 묶어주면 최장명사구 인식 시 발생할 수 있는 오분석의 가능성을 감소시킬 수 있을 뿐만 아니라, 이후의 처리과정에서 이루어지게 될 "" 내부의 단어들에 대한 분석도 안전하게 수행될 수 있게 된다. 또한 인용문 내부의 단어들이 하나로 묶였기 때문에 인용문의 앞뒤에 존재하는 '简单', '采取', '做法' 과 같은 단어들이 참조할 수 있는 문맥의 범위를 손쉽게

[3] 한국어 해석: 상해 푸뚱(지명)은 단지 "일정한 시기 동안 일을 하여 경험을 축적한 이후 다시 법규와 규정을 제정하는" 방법을 채택하는 것이 아니다.

확대시킬 수 있어 최장명사구 식별에 보다 많은 정보를 활용할 수 있게 된다. 그러나 인용부호로 묶인 부분을 무조건 하나의 청크로 간주하는 것은 위험하며 먼저 앞뒤 문맥을 살펴서 "" 안의 내용이 인용문인지 아닌지를 판단하는 과정이 필요하다.

2) 공기 패턴

(加工/NN 贸易NN) ([在/P 广东/NR 外经贸/NN 发展/NN 中/LC] 的/DEC 地位/NN)[4]

이 예제에서도 꺾쇠 []로 묶인 부분을 하나의 확장된 청크로 처리해 줌으로써 전치사와 방향사, 동사의 정확한 식별에 도움을 줄 수 있게 된다. 중국어에는 在……中의 경우와 같이 서로 쌍으로 나타나는 단어들이 많다. 이런 단어 쌍들은 대부분 전치사와 동사, 방향사로 구성되었으며 최장명사구에 포함되는 경향이 낮고 내부에 최장명사구를 포함하는 경우가 다수이다. 하지만 이들 단어 쌍 뒤에 的이 오거나 뒤에 "동사구 + 的"이 따라오는 경우에는 최장명사구에 속하게 되는데 이런 경우에는 이 단어 쌍들과 그 내부에 포함된 단어들을 하나의 청크로 묶어주어 최장명사구 식별에 도움을 주도록 하였다.

3) Slight pause mark[5]로 나열된 단어들

(上海/NR 浦东/NR) (近年/NT) 来/LC 颁布/VV 实行/VV 了/AS (涉及/VV [经济/NN 、/PU 建设/NN 、/PU 规划/NN 、/PU 科技/NN 、/PU 文教/NN] 等/ETC 领域/NN 的/DEC 七十一/CD 件/M 法规性/NN 文件/NN) 。/PU[6]

위의 예문에서도 slight pause mark 로 나열된 단어들을 먼저 청크로 묶어서 일괄 처리함으로써 slight pause mark 앞뒤에 존재하는 단어들이 참조할 수 있는 문맥의 범위를 확대시켜 정확한 식별을 하는데 도움을 주도록 하였다.

지금까지 위 예제들에서 살펴본 바와 같이 최장명사구 식별에 도움을 줄 수 있도록, 특정한 기능을 하는 단어나 문장부호, 품사정보에 기반하여 단어들을 미리 묶어줌으로써 구 묶음의 범위를 확장시켜 사용하였다. 다만 하나의 청크로 묶이게 되면 하나의 단어처럼 취급되기 때문에 기존 정보가 손실될 위험이 높으므로, 확장된 청크의 내부에 동사구의 상 정보와 같은 유용한 기능성 단어가 존재할 경우 청크로 묶지 않고 그 기능성 단어들이 제공하는 정보를 그대로 유지할 수 있게 하였다.

5. 학습자질

5.1 문장부호의 사용

중국어에서 다른 종류의 문장부호들이 서로 다른 용도로 문장 중에 쓰이고 있지만 이들을 특별히 세분화하지 않고 일반적인 품사처리 단계에서는 단일 태그로만 태깅한다. Penn Chinese Treebank에서는 총 38종류의 문장부호가 출현하는데 이들도 모두 단일 품사태그인 PU로 태깅되었다. Penn Chinese Treebank 코퍼스에 출현한 빈도에 따른 상위 10개 문장부호의 사용상황이 표 1 에 제시되어 있다.

[표 1] 문장부호 사용상황

Punctuation	Total	Inside	Outside
Comma (,)	12,695	404	12,291
Period(。)	4,698	9	4,689
Slight-pause mark (、)	2,725	2,306	419
Brackets (「」)	1,666	1,217	449
Question mark (?)	308	12	296
Semicolon (；)	302	10	292
colon(：)	191	9	182
Quotation marks (“”)	163	144	19
Exclamation mark (!)	131	0	131
Brackets (())	114	86	28

위의 표에서 Inside는 문장부호가 MNP에 포함된 상황을 말하는 것이고 Outside는 문장부호가 MNP에 포함되지 않는 상황을 표시하는데 위의 표에서 알 수 있다시피 문장부호에 따라서 최장명사구에 포함되거나 포함되지 않는 현상을 나타내기에 문장부호는 중국어 최장명사구를 식별함에 있어서 중요한 역할을 하게 될 것이라고 생각한다. 본 논문에서는 문장부호의 문법적 기능[14]과 Penn Chinese Treebank에서의 사용상황에 근거하여

---

[4] 한국어 해석: 가공무역이 광동의 대외경제무역 발전에서 (차지하는) 지위

[5] slight pause mark (、)는 중국어에서 대등접속관계를 나타내는 문장부호로서 단순한 단어들의 나열에 사용된다. [14]

[6] 한국어 해석: 상해 푸뚱은 최근에 와서 경제, 건설, 계획, 과학기술, 문화교육 등의 영역에 관련된 71가지 법규적 문건을 반포하고 실행하였다.

문장부호를 아래와 같은 5개 클래스로 분류하고 실험에 학습자질로 사용하였다.

[표 2] 문장부호 분류

분류	문장부호
그룹1	slight-pause mark (、)
그룹2	comma (,)
그룹3	period (。), question mark (?), exclamation mark (!), semicolon (;), colon (:), star (*), ellipses (。。。) …
그룹4	quotation marks (“”, ‘’, 《》, <>, 「」, 『』), brackets({}, ()) …
그룹5	hyphen (-), dash (--), , apostrophe (’), slash mark (/), dot(·) …

5.2 의미정보의 사용

의미정보는《同义词词林》[15]의 분류체계를 사용하였다. 《同义词词林》[15]에서는 단어들을 대분류 12 개, 중분류 94 개, 소분류 1428 개로 분류하고 있는데 본 논문에서는 중분류의 의미정보만을 학습자질로 사용하였다.

6. 시스템 개요

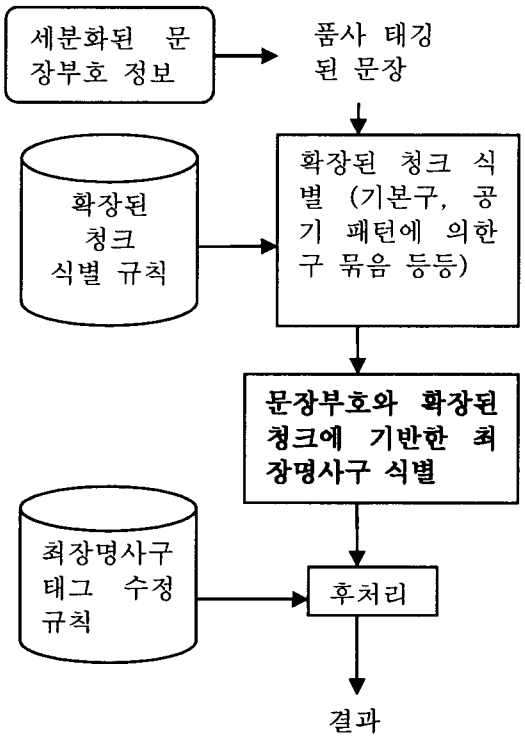

그림 1 시스템 개요

그림 1 은 전체 시스템의 구성도이다. 품사 태깅된 문장이 입력되면 확장된 청크 식별 규칙에 근거하여 먼저 청크들을 묶어준다. 그리고 하나의 단위로 된 청크와 단어의 품사정보, 문장부호 정보 등의 다양한 자질들을 활용하여 최장명사구 식별을 수행한다. 마지막으로 최장명사구 태그 수정규칙에 의한 후처리 작업을 통해 최종결과를 얻는다.

7. 최장명사구 식별 모델 구축

대부분의 구묶음(chunking) 연구는 Ramshaw[16]가 제시한 IOB[7] 태그를 사용하고 있다. Taku[17]에서 Ramshaw[16]의 IOB 태그를 좀 더 확장시킨 5가지의 IOBES 태그를 소개하고 있는데 본 논문에서는 Taku[17]에서 제안한 5가지 태그를 사용하였다. 그러므로 최장 명사구 식별은 아래 표에 기술된 5가지 클래스의 식별문제로 간주될 수 있다.

[표 3] 최장명사구 태그 형식

최장명사구 태그	설명
MNP_B	최장 명사구 시작 태그 (한 단어 이상으로 구성된 최장명사구)
MNP_E	최장 명사구 종결 태그
MNP_I	최장 명사구 시작 태그와 종결 태그 사이의 태그
MNP_O	최장명사구에 속하지 않는 태그
MNP_S	한 단어로 구성된 최장명사구

8. 실험 및 결과 분석

본 연구에서는 Penn Chinese Treebank 4.0을 실험 코퍼스로 사용하였다. Penn Chinese Treebank 4.0에서는 괄호를 이용하여 구 구조(Phrase Structure)를 기술하고 있는데, 이 구조체에서 제일 상위 노드의 명사구, 수량사구, “的”가 포함된 명사구, 숫자 리스트구, 양사구를 자동 추출하여 학습을 위한 최장명사구 정보로 사용하였다.

실험 코퍼스는 5,000 문장을 사용하였고 문장부호를 포함하여 약 16 만 단어이며 문장의 평균 길이는 32 단어이다. 최장명사구는 약 28,000 개이고 평균길이는 3 단어이다. 기계학습 알고리즘은 Naïve Bayes[18]와 Decision Tree[18],

[7] I: chunk 안에 포함되는 단어; O: chunk 안에 포함되지 않는 단어; B: chunk 의 시작; E: chunk의 끝; S: 하나의 단어로 된 chunk.

Support Vector Machine [8] [19]을 사용하였고 테스트는 10 fold cross validation, 성능평가는 5개 태그의 F-measure 평균 값으로 진행하였다.

기본모델은 Penn Chinese Treebank의 단어 품사정보(문장부호가 PU로만 태깅됨) 만을 학습자질로 사용하여 진행하였다. 기본모델의 실험결과는 Decision Tree의 윈도우 크기 9에서 수렴하였고 5개 태그의 F-measure 평균 값은 84.58%였다.

본 논문에서는 최장명사구 식별을 위하여 여러 학습자질들이 얼마간의 도움을 줄 수 있는지에 대하여 각각 실험을 하였다. 다음으로 좋은 성능을 보인 확장된 청크와 문장부호를 자질로 사용하여 다시 실험을 진행하였다.

8.1 문장부호 추가

[표 4] 문장부호 추가 시 성능

기계학습방 / 윈도우 사이즈	Naive Bayes	Decision Tree	SVM
3	82.16%	85.10%	83.48%
5	83.90%	87.00%	85.11%
7	84.16%	**87.16%**	86.25%
9	84.14%	**87.16%**	86.95%
11	84.14%	**87.16%**	87.01%

표 4는 분류를 나눈 문장부호의 품사정보를 학습자질로 추가하였을 때의 실험결과이다. 윈도우 사이즈 7에서 수렴하였고 Decision Tree 기계학습 방법이 제일 좋은 성능을 보였으며 기본모델에 비해 2.58%의 성능 향상이 있었다.

8.2 의미정보 추가

[표 5] 의미정보 추가 시 성능

기계학습방법 / 윈도우 사이즈	Naive Bayes	Decision Tree	SVM
3	76.92%	81.42%	76.32%
5	78.72%	84.08%	79.61%
7	79.32%	84.56%	82.24%
9	79.56%	**84.60%**	83.53%
11	79.70%	**84.60%**	83.93%

표 5에서 알 수 있다시피 중분류 94가지 종류의 의미정보를 추가하였을 때 성능에 거의 변화가 없었다. 그 원인으로는 시스템 사전에 모든 단어의 의미정보가 있는 것이 아니고 또한 한 단어에 의미정보가 여러 개일 경우 적절한 의미정보 부여에 어려움이 있었기 때문이라고 생각한다. 본 연구에 사용된 중국어 분석기 사전에 존재하는 단어 중 문장부호를 제외한 14만 단어에서 2만 단어는 의미정보가 없으며 8만 단어는 다중의미를 가지고 있다. 이런 단어들은 중국어 분석 시 하나의 정확한 의미정보를 제공해 주지 못하기 때문에 정보로서의 가치가 희석되어 시스템의 성능 향상에 큰 영향을 주지 못한 것으로 판단된다.

[8] Support Vector Machine 에서 커널은 linear로 하였고 C 값은 1.0으로 하였다.

8.3 기본구와 확장된 청크 사용

[표 6] 기본구 사용 시 성능

기계학습방법 / 윈도우 사이즈	Naive Bayes	Decision Tree	SVM
3	77.24%	81.90%	79.65%
5	78.84%	85.06%	81.89%
7	79.72%	85.46%	83.88%
9	79.82%	**85.56%**	84.92%
11	79.92%	85.52%	85.29%

위의 표는 규칙에 근거하여 기본구를 먼저 식별한 다음 기본구를 하나의 단위로 처리하여 다시 최장명사구를 식별하였을 때의 실험결과이다. 기본구를 한 개의 단위로 보았을 때 윈도우 사이즈 9에서 가장 좋은 성능을 보였으며 기본모델보다 0.98%의 성능 향상이 있었다. 이 실험결과는 기본구를 한 개의 단위로 보는 것이 최장명사구 식별에 도움이 된다는 사실을 보여준다. 이 실험결과에 근거하여 본 논문에서는 기본구를 확장된 청크 개념으로 확장하여 최장명사구 식별에 도움을 주고자 하였다.

[표 7] 확장된 청크 사용 시 성능

기계학습방법 / 윈도우 사이즈	Naive Bayes	Decision Tree	SVM
3	81.20%	85.22%	83.71%
5	82.12%	87.02%	85.68%
7	82.30%	87.30%	85.90%
9	82.18%	**87.32%**	86.30%
11	82.08%	**87.32%**	86.57%

위의 표에서 알 수 있다시피 확장된 청크 개념을 사용하였을 때 기본구를 먼저 식별했을 때보다 1.76%, 기본모델보다 2.74%의 성능 향상이 있었다. 때문에 확장된 청크를 먼저 식별한 다음 이를 하나의 단위로 처리하여 앞뒤에서 참조할 수 있는 문맥정보를 늘리는 것은 최장명사구 식별에 도움이 된다는 것을 알 수 있다.

이상의 실험결과에서 모든 성능은 윈도우 사이즈 7 이나 9 에서 최고 성능을 보였다. 때문에 본 논문에서는 윈도우 사이즈를 9 로 고정시키고 학습자질을 하나하나씩 추가하였을 때의 성능 향상 결과에 대하여 실험을 진행하였다. 표 8 에서 알 수 있다시피 기계학습 방법은 Decision Tree 가 가장 좋은 성능을 보였고 모든 학습자질을 다 추가하였을 때 기본모델보다 3.44% 향상된 최고의 성능을 보였다.

[표 8] 최장명사구 식별 성능

기계학습방법 / 윈도우 사이즈	Naive Bayes	Decision Tree	SVM
품사정보	79.90%	84.58%	82.53%
+ 문장부호	84.14%	87.16%	86.96%
+ 확장된 청크	82.80%	**88.02%**	87.81%

아래는 모든 학습자질들을 다 추가하여 윈도우 사이즈 9 에서 Decision Tree 로 진행한 실험의 5개 태그 각각의 성능이다. 이 표에서 알 수 있다시피 최장명사구 시작태그(MNP_B)에 대한 식별 작업이 가장 어려웠으며, 최장명사구에 속하지 않는 단어들(MNP_O)에 대한 식별은 상당히 좋은 성능을 보이고 있다.

[표 9] 각 태그별 성능

성능 / MNP 태그	정확률	재현율	F-measure
MNP_B	**82.30%**	**78.20%**	**80.20%**
MNP_E	89.00%	93.30%	91.10%
MNP_S	84.20%	85.80%	85.00%
MNP_I	91.80%	85.60%	88.60%
MNP_O	**92.80%**	**97.60%**	**95.20%**

9. 후처리 작업

최장명사구 식별은 왼쪽과 오른쪽 경계를 정확하게 식별하고 이들을 매칭시켜야만 후속으로 이루어지는 구문분석이나 기계번역 등의 응용에 사용될 수 있다. 본 논문에서는 최장명사구 인식이 끝난 후 설정된 좌우 경계의 짝이 제대로 일치하지 않는 경우, 가장 가까운 두 좌우 경계를 매칭시킴으로써 최장명사구의 경계를 확정하였다. 아래의 표에서 알 수 있다시피 후처리 작업을 한 후 F-measure 평균값이 다소 향상되었으며 최장명사구의 왼쪽과 오른쪽 경계 매칭에 직접적인 영향을 주는 세 태그, 시작, 종결, 하나의 단어로 구성된 최장명사구 태그 성능도 조금씩 향상되었다.

[표 10] 후처리 결과

후처리 / MNP 태그	후처리 전	후처리
MNP_B	80.20%	**80.23%**
MNP_E	91.10%	**92.48%**
MNP_S	85.00%	**88.29%**
MNP_I	88.60%	88.60%
MNP_O	95.20%	93.53%
평균	88.02%	**88.63%**

10. 에러 분석

위의 실험결과에 대하여 품사 별로 분석을 해보면 아래와 같은 에러현상을 나타낸다.

[표 11] 품사별 에러 상황

품사	에러 비율
명사	**35.76%**
동사	**20.23%**
부사	8.54%
전치사	6.92%
기타	28.52%

에러 유형 중에서는 명사가 가장 큰 비율을 차지하고 있다. 모든 명사는 기본적으로 최장명사구에 속하게 되지만 자신이 최장명사구의 시작이나 내부, 종결 또는 한 단어로 된 명사구 중 어느 쪽에 속하는지 경계를 정확하게 찾아주기 힘들기 때문에 에러가 많이 발생한다. 동사, 전치사, 부사는 일반적으로 최장명사구에 속하는 경향이 낮지만 특정한 경우에 최장명사구에 속하게 되는데 이러한 경우를 식별하기 힘들기 때문에 나타나는 문제이다.

11. 결론 및 향후연구

위의 실험을 통하여 문장부호의 그룹별 사용과 확장된 청크의 개념은 최장명사구를 식별하는데 있어서 중요한 역할을 한다는 것을 알 수 있다. 그러나 확장된 청크의 인식은 이후에 수행될 처리 과정에 큰 영향을 미치기 때문에 확장된 청크 인식을 위한 규칙의 설정에는 세심한 검토가 필요하다.

에러분석(제10장 참조)에서 알 수 있다시피 많은 에러가 명사에서 발생하는데 명사에서 나타나는 에러는 주로 최장명사구의 시작과 끝을 정확하게 식별해주지 못하기 때문이다. 만약 전 단어의 최장명사구 식별 태그를 학습자질로 준다면 현재 단어의 최장명사구 식별에 많은 도움이 될 수 있을 것이므로 전 단어의 최장명사구 식별 태

그를 학습자질로 사용하여 문제 해결을 시도해볼 수 있다. 그리고 명사에서 나타나는 또 하나의 문제점은 명사들이 단순하게 나열되어 있는 경우 주제어와 주어를 구분해주지 못하는 문제가 있다. 이 경우는 결합가 정보를 이용하여 명사와 동사 사이의 유사성을 계산하는 방법으로 문제 해결을 시도 해볼 수 있다.

후처리 작업에서도 단순히 가장 가까운 좌우 경계를 매칭시키는 방법을 보다 진보시킨 정교한 규칙을 사용하여 후처리 작업에서의 좌우 경계 매칭의 성능향상을 시도해볼 수 있다.

참고 문헌

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

[2]Zhou Qiang, Sun Maosong and Huang Changning "Automatically Identify Chinese Maximal Noun Phrase", Technical Report 99001, State Key Lab. of Intelligent Technology and Systems, Dept. of Computer Science and Technology, Tsinghua University. 1998

[3] Didier Bourigault. "Surface Grammatical Analysis for the Extraction of Terminological Noun Phrases," In: Christian Boitet ed. Proceedings of the 15th International Conference on Computational Linguistics (COLING 92), Nantes, France, 977-981, 1992

[4] Atro Voutilainen, "NPTool, a detector of English Noun Phrases" In: Ken Church ed. Proceedings of the workshop on Very Large Corpora: Academic and Industrial Perspectives Ohio State University, Columbus, Ohio, USA, pages 48-57, 1993

[5] Kuang-hua Chen, Hsin-His Chen, "Extracting Noun Phrases from Large-Scale Texts: A Hybrid Approach and Its Automatic Evaluation" In: Proceedings of 32nd Annual Meeting of Association of Computational Linguistics, New York: Academic Press. pages 234-241, 1994

[6]Wenjie Li, Haihua Pan, Ming Zhou, Kam-Fai Wong and Vincent Lum, " Corpus-based Maximal-length Chinese Noun Phrase Extraction" In: Key-Sun Choi ed. Proceedings of Natural Language Processing Pacific Rim Symposium(NLPRS' 95), Korea: Academic Press, pages 246-251 1995.

[7] Angel S. Y. Tse, Kam-Fai Wong, & al. "Effectiveness Analysis of Linguistics- and Corpus-based Noun Phrase Partial Parsers." In: Key-Sun Choi ed. Proceedings of Natural Language Processing Pacific Rim Symposium (NLPRS' 95), Korea: Academic Press, pages 252-257, 1995

[8] Changhao Yin, "Identification of Maximal Noun Phrase in Chinese: Using the Head of Base Phrases" Master Dissertation, 2004

[9] Erik F. Tjong Kim Sang, Sabine Buchholz, "Introduction to the CoNLL-2000 Shared Task: Chunking", In: Proceedings of CoNLL-2000 and LLL-2000, pages 127-132, 2000

[10] Yongmei Tan, Tianshun Yao, Qing Chen and Jongbo Zhu "Applying Conditional Random Fields to Chinese Shallow Parsing", In: The Sixth International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2005), LNCS, Vol.3406, Springer, pages 167-176, 2005

[11] Yuqi Zhang, Qiang Zhou "Chinese Base-Phrases Chunking", In: COLING-02: The First SIGHAN Workshop on Chinese Language Processing, pages 131-135, 2002

[12] Tie-jun Zhao, Mu-yun Yang, Fang Liu, Jian-min Yao, Hao Yu, "Statistics Based Hybrid Approach to Chinese Base Phrase Identification", In Proceedings of Second Chinese Language Processing Workshop , Hong Kong, China, pages 73-77. 2001

[13] Jun Zhao, Chang-ning Huang, "The Model for Chinese BaseNP Structure Analysis", In: Chinese J. Computer, 22(2) , pages 141-146 , 1999

[14] Shui-fang Lin , "study and application of punctuation" (标点符号的学习与应用). People's Publisher, P.R.China (in Chinese) , 2000

[15] 梅家驹, 竺一鸣 & al 《同义词词林》, 上海辞书出版社, 上海, 1983

[16] Lance A. Ramshaw and Mitchell P. Marcus. "Text Chunking using transformation-based Learning" In: Proceedings of the 3rd workshop on very large corpora, pages 88-94, 1995

[17] Taku Kudo and Yuji Matsumoto, "Chunking with Support Vector Machines" In: Proceedings of Second Meeting of North American Chapter of the Association for Computational Linguistics(NAACL), pages 192-199. 2001

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

[19] Multi-Class Support Vector Machine Learning Toolkit http://www.cs.cornell.edu/People/tj/svm_light/svm_multiclass.html