확장청크와 세분화된 문장부호에 기반한 중국어 최장명사구 식별

(Identification of Maximal-Length Noun Phrases Based on Expanded Chunks and Classified Punctuations in Chinese)

백 설 매 † (Xue-Mei Bai) 이 금 희 †† (Jin-Ji Li) 김 동 일 ††† (Dong-Il Kim) 이 종 혁 †††† (Jong-Hyeok Lee)

요 약 일반적으로 명사구는 기본명사구와 최장명사구로 분류되는데 최장명사구에 대한 정확한 식별은 문장의 전체적인 구문구조를 파악하고 정확한 지배용언을 찾아내는데 중요한 역할을 하게 된다. 본 논문에서는 확장된 청크(chunk) 개념과 다섯 개의 클래스로 세분화된 문장부호 정보를 자질로 사용한 두 단계 최장명사구 식별 기법을 제안한다. 제안한 기법은 기본모델보다 2.65% 향상된 평균 89.66%(F_1-measure)의 우수한 성능을 보인다.

키워드 : 최장명사구, 확장청크, 세분화된 문장부호

Abstract In general, there are two types of noun phrases(NP): Base Noun Phrase(BNP), and Maximal-Length Noun Phrase(MNP). MNP identification can largely reduce the complexity of full parsing, help analyze the general structure of complex sentences, and provide important clues for detecting main predicates in Chinese sentences. In this paper, we propose a 2-phase hybrid approach for MNP identification which adopts salient features such as expanded chunks and classified punctuations to improve performance. Experimental result shows a high quality performance of 89.66% in F_1-measure.

Key words : Maximal-Length Noun Phrase(MNP), Expanded Chunk, Classified Punctuation

1. 서 론

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 구묶음(chunking)의 여러 가지 구 단위에서도 특히 명사구에 대한 정확한 식별은 구문분석뿐만 아니라 기계번역이나 정보검색, 정보추출과 같은 다양한 자연언어처리 분야에서 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 3 가지로 분류할 수 있는데, 기본명사구와 최단명사구를 동일시하는 관점도 있다. 중국어의 기본명사구에 대한 정의는 연구자에 따라 약간씩 차이가 있지만, 최장명사구는 일반적으로 "다른 명사구에 포함되지 않는 명사구"로 정의된다. 최장명사구는 명사구내에 포함될 수 있는 단어들이 매우 다양할 뿐만 아니라 장거리 의존성(long dependency) 문제도 포함되기 때문에 최단명사구나 기본명사구에 비해 정확한 식별 작업이 어렵다.

하지만 최장명사구 식별은 구문분석의 복잡도를 크게 낮출 수 있을 뿐만 아니라 문장의 전체적인 구조를 이해하는데 도움이 되며, 특히 중국어에서는 정확한 지배

† 비 회 원 : (주)4U Applications
xuemei@4uapplications.com
†† 비 회 원 : 포항공과대학교 컴퓨터공학과
ljj@postech.ac.kr
††† 비 회 원 : 연변과학기술대학교 컴퓨터전자통신학부 교수
dongil@ybust.edu.cn
†††† 종신회원 : 포항공과대학교 컴퓨터공학과 교수
jhlee@postech.ac.kr
논문접수 : 2008년 1월 31일
심사완료 : 2009년 2월 6일

용언을 찾아내는데 많은 정보를 제공하기에 꼭 필요한 작업이다.

본 논문에서는 보다 효과적인 최장명사구 식별을 위해 기존의 구묶음 단계의 묶음단위(chunking unit)를 보다 확대시킨 확장청크(expanded chunk)의 개념을 도입한다. 또한 중국어에서 사용되고 있는 다양한 문장부호들의 특성에 주목하여 이들의 세분화된 정보를 최장명사구의 식별에 필요한 자질로 제안한다. 전체 시스템은 먼저 확장청크를 식별해내고 그 다음 최장명사구를 식별하는 두 단계 기법을 사용한다.

본 논문의 구성은 다음과 같다. 우선 2장에서는 구묶음 및 최장명사구 식별에 대한 관련 연구들을 살펴보고 3장에서는 왜 확장청크와 세분화된 문장부호 정보를 자질로 제안했는지에 대하여 설명한다. 4, 5장에서는 각각 시스템 구조와 실험 및 분석에 대해서 기술한다. 6장에서는 최장명사구 식별에 필요한 후처리에 대해서 설명하고, 에러분석은 7장에서, 그리고 최종 결론과 향후 과제는 8장에서 기술한다.

2. 관련 연구

최장명사구 식별은 영어, 불어, 중국어 등 언어권에서 연구가 진행되고 있다. 먼저 영어와 불어에서의 최장명사구 식별에 대한 기존 연구들을 살펴보면 간단한 규칙이나 규칙과 통계의 혼합형 방법을 사용하였고 현재는 거의 연구를 하지 않는 상태이다[1-3].

중국어에서 최장명사구 식별에 대한 연구는 초반에는 규칙과 통계기반 등 방법을 사용하였으나, 점차 먼저 기본구를 식별하고 나중에 최장명사구를 식별하는 두 단계 모델을 이용한 방법으로 전환하였다.

1) Li[4]는 단어의 품사 태그 정보를 이용하여 통계적 방법으로 최장명사구의 자동식별을 시도하였는데 정확률과 재현율이 각각 71.3%, 69.1%로 보고되었다.
2) Tse[5]는 통계와 규칙의 혼합 기법을 사용하였는데 이 논문에서는 "的(de)"를 포함한 명사구만을 식별하였고 정확률과 재현율을 각각 75%와 90%로 보고하고 있다.
3) Zhou[6]는 최장명사구 식별을 두 단계로 나누어 진행하였다. 먼저 코퍼스의 통계정보를 이용하여 문장부호, 공기정보, 등위접속구조의 좌우경계를 식별하고, 규칙기반 방법으로 최장명사구의 오른쪽 경계를 확정한 후 계속하여 왼쪽으로 처리범위를 확장해 나가면서 최장명사구의 왼쪽 경계를 식별하였다. 이 방법은 평균길이가 3~4 단어인 최장명사구 식별에 대하여 정확률 85.4%, 재현율 82.3%의 성능을 보여주었다.
4) Yin[7]은 두 단계 학습모델을 사용하였는데 먼저 기본구를 식별한 다음 기본구의 중심어만을 추출하여 최종 최장명사구 식별에 사용하였다. 그러나 기본구에 대한 정의와 식별 방법으로 인한 문제점으로, 최종 최장명사구 식별 성능 향상도가 그리 높지 않았다. 정확률과 재현율은 각각 77.1%, 75.5%로 보고하였다.

두 단계 방법은 효율성이 있다고 주장하지만 첫 번째 단계에서 기본구의 식별 오류가 바로 두 번째 단계로 이어지기 때문에 기본구의 식별 정확률이 높아야 되는 전제조건을 가지고 있다. 그리하여 본 논문에서는 기본구를 확장한 확장청크의 개념을 제시하고자 한다.

영어권에서는 현재 구묶음에 기계학습 방법을 많이 사용하고 있는데 이는 기존의 규칙이나 통계기반 방법보다 훨씬 좋은 성능을 낸다고 알려지고 있다. 본 논문 역시 기계학습 방법을 제안하며 또 기계학습 방법의 성능을 높이는데 꼭 필요한 유용한 자질을 개발하고자 문장부호가 최장명사구 식별에서 보이는 쓰임새에 따라 세분화된 문장부호를 자질로 제안한다.

3. 확장청크와 세분화된 문장부호

본 논문에서는 확장청크와 세분화된 문장부호 정보를 최장명사구 식별에 유용한 자질로 개발하였다. 먼저 최장명사구 식별에 확장청크와 세분화된 문장부호 정보가 왜 필요한지 그 원인 및 타당성에 대하여 설명하려고 한다.

3.1 확장청크

기존 청크는 품사와 구문관계에 따라 분류하고 정의하였기 때문에 문장에 흔히 등장하는 공기(co-occurrence) 패턴이나 따옴표("") 등 문장부호에 의해 묶인 덩어리를 표현할 수 있는 방법이 없다. 본 논문에서 사용하고자 하는 확장청크는 최장명사구 식별 시 미리 하나의 단위로 묶어서 처리해 줄 수 있는 단어 집단들까지 포괄함으로써 기존의 청크 개념을 확장시킨 것이다. 우리가 정의한 확장청크에는 기본구(base phrase), 공기패턴, 인용부호 및 slight pause mark[1]에 의해 묶인 단어들이 모두 포함된다.

확장청크 개념의 사용은 최장명사구에 포함되거나 포함되지 않는 단어들을 미리 묶어줌으로써 이후 단계에서 소요되는 불필요한 과정을 줄이고 효율성을 높이는데 목적이 있다. 이것은 보다 많은 주변 문맥 정보를 참조하여 최장명사구를 식별할 수 있게 한다. 그러나 일단 하나의 청크로 묶이게 되면 이들은 하나의 단어처럼 간주되기 때문에 잘못된 청크의 설정은 최장명사구 식별에 오히려 악영향을 미치게 된다. 그 대안으로 높은 정확률을 위해 확장청크의 식별은 규칙 기반으로 수행되도록 구성하였으며, 애매하거나 틀릴 가능성이 있는 규칙은 모두 제외하

1) slight pause mark (、)는 중국어에서 대등접속관계를 나타내는 문장부호로서 단순한 단어들의 나열에 사용된다.

도록 하였다. 현재 확장청크를 묶어주는 규칙 기반 시스템의 정확률과 재현율은 각각 97%와 65%이다.

하나의 확장청크에 포함된 단어들은 하나의 단위로 간주되어 처리되기 때문에 각 단어들이 확장청크 내에서 갖는 특성을 놓칠 가능성이 있다. 그리하여 확장청크 내부 단어들의 시작과 종결, 그리고 중간 태그 정보를 유지하여 그 정보를 활용함으로써 보다 정확한 최장명사구의 식별이 이루어지도록 하였다.

3.1.1 기본구

구묶음 연구는 많이 진행되었지만 시스템마다 다른 묶음 단위를 정의하여 사용하는 문제점이 존재한다. 이런 문제점을 감안하여 CoNLL-2000 shared task에서는 11가지 영어 기본구를 정의하였는데 정의한 기본구로는 명사구(NP), 동사구(VP), 전치사구(PP), 부사구(ADVP), 종속어구(SBAR), 형용사구(ADJP), particle 구(PRT), 접속사구(CONJP), 감탄사구(INTJ), 리스트구(LST), unlike coordinated phrase(UCP) 등이다. 이런 기본구는 하나의 중심어(head)와 여러 개의 pre-modifier로 구성될 수 있지만 post-modifier나 argument는 가질 수 없다는 제약을 두었다[8].

중국어에서도 구묶음에 관한 연구가 활발히 진행되었으나 아직도 기본구에 대한 정확하고 통일된 정의가 없는 현실이다. 본 논문에서는 CoNLL-2000 shared task에서 제시한 영어의 기본구 정의에 기반하여 중국어 기본구 정의를 하였고 그 제약조건도 그대로 따르고자 한다. 다만 영어에만 존재하고 중국어에는 존재하지 않는 particle 구는 제외하고, 또 중국어에만 존재하는 방향사구를 추가하여 중국어와 영어의 차이점을 반영하였다. 그리고 최장명사구 식별에 도움이 안되고 영어에서도 거의 쓰이지 않는 UCP도 제외하였다. 우리가 최종 정의한 중국어 기본 명사구 리스트는 표 1과 같다.

이러한 기본구 중에서 종속어구, 전치사구, 접속사구, 감탄사구, 리스트구, 방향사구 등은 기본구의 제약조건에 의하면 그 단어 자체로서 하나의 구를 이루기 때문에 최장명사구 식별에 도움을 주지 못한다. 따라서 본 논문에서는 이들의 사용을 배제하고 나머지 기본구만을 최장명사구 식별에 사용하였다. 그리고 유용한 정보를 최대한 사용하기 위하여 중국어 언어학적 특성을 충분히 고려하여 명사구는 고유명사구와 일반명사구, 시간사구(temporal NP), 양사구(quantifier NP)로 세분화 하였고 동사구는 일반 동사구와 계사(copula)구, "有(you)"동사구, 그리고 형용사가 지배용언이 되는 동사구로 좀 더 세분화시켜서 사용하였다.

표 1 중국어 기본 명사구

기본구 유형	설명	한국어 설명
NP	Noun Phrase	명사구
VP	Verb Phrase	동사구
ADVP	Adverb Phrase	부사구
ADJP	Adjective Phrase	형용사구
SBAR	Subordinated Phrase	종속어구
PP	Prepositional Phrase	전치사구
CONJP	Conjuction Phrase	접속사구
INTJ	Interjection Phrase	감탄사구
LST	List Mark Phrase	리스트구
LP	Location Phrase	방향사구

3.1.2 " "로 묶인 인용문

Ex1. (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

한국어: 상해 푸둥(지명)은 단지 "일정한 시기 일을 하여 경험을 축적한 이후 다시 법규와 규정을 제정하는" 방법을 채택하는 것이 아니다.

위의 예문에서 [ ]로 묶인 부분이 최장명사구이며 괄호 ( )로 묶인 부분이 확장청크다. 예문에서 볼 수 있다시피 " "로 묶인 부분에서 밑줄 친 부분들은 품사가 명사가 아니기 때문에 일반적으로 명사구에 포함되는 경향이 낮으며, 또한 따옴표 안의 단어들은 그 자체로서 하나의 문장을 이루고 있기 때문에 부가적인 정보가 없다면 이 문장에서 최장명사구 식별은 매우 힘들다. 만약 따옴표 안의 문장을 하나의 청크로 간주하여 미리 묶어주면 최장명사구 인식 시 발생할 수 있는 오분석의 가능성을 감소시킬 수 있을 뿐만 아니라, 이후 따옴표 내부의 단어들에 대한 분석도 안전하게 수행될 수 있다.

또한 인용문 내부의 단어들이 하나로 묶였기 때문에 인용문의 앞뒤에 존재하는 '简单(jiandan)', '采取(caiqu)', '做法(zuofa)'와 같은 단어들이 참조할 수 있는 문맥의 범위를 손쉽게 확대시킬 수 있어 최장명사구 식별에 보다 많은 정보를 활용할 수 있게 된다. 그러나 따옴표로 묶인 부분을 무조건 하나의 청크로 간주하는 것은 위험하며 먼저 앞뒤 문맥을 살펴서 따옴표 안의 내용이 인용문인지 아닌지를 판단하는 과정이 필요하다.

3.1.3 공기 패턴

Ex 2. [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]

한국어: 가공무역이 광동 대외경제무역 발전에서의 위치

위 예문에서는 괄호로 묶인 부분을 하나의 확장청크로 간주한다. 중국어에는 '在(zai)……中(zhong)'과 같이 쌍으로 나타나는 공기 패턴들이 존재한다. 이런 공기 패턴들은 대부분 전치사와 동사, 방향사로 구성되었으며 최장명사구에 포함되는 경향이 낮고 최장명사구를 내포하는 경우가 다수이다.

하지만 이런 공기 패턴 뒤에 '的(de)' 또는 '동사구 + 的(de)'가 오는 경우에는 최장명사구에 속하게 되는데 이런 경우에는 이 공기 패턴과 그 내부에 포함된 단어들을 하나의 확장청크로 묶어주어 최장명사구 식별에 도움을 주도록 하였다.

3.1.4 Slight pause mark로 나열된 단어들

Ex 3. (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

한국어: 상해 푸둥은 최근에 와서 경제, 건설, 계획, 과학기술, 문화교육 등 영역에 관련된 71가지 법규 문서를 반포하고 실행했다.

위 예문에서 slight pause mark로 나열된 단어들을 먼저 확장청크로 묶어서 일괄 처리함으로써 slight pause mark의 앞뒤에 존재하는 단어들이 참조할 수 있는 문맥의 범위를 확대시켜 정확한 식별을 하는데 도움을 주도록 하였다.

지금까지 위 예문들에서 살펴본 바와 같이 최장명사구 식별에 도움을 줄 수 있도록 특정 단어나 문장부호에 의해 단어들을 미리 묶어줌으로써 구묶음의 범위를 확장시켜 사용하였다. 다만 하나의 청크로 묶이게 되면 하나의 단어처럼 취급되기 때문에 기존 정보가 손실될 위험이 높으므로, 확장청크의 내부에 동사구의 상(aspect)정보와 같은 유용한 기능성 단어가 포함되지 않도록 하여 그 기능성 단어들이 제공하는 정보를 그대로 유지할 수 있게 하였다.

3.2 세분화된 문장부호

중국어에서는 문장부호들이 서로 다른 용도로 문장에 선 쓰이지만 이들을 특별히 세분화하지 않고 품사처리 단계에서 단일 태그로만 태깅한다. 본 논문에서 사용한 코퍼스에도 총 38 종류의 문장부호가 출현하는데 이들도 모두 단일 품사태그인 'PU'로 태깅되었다.

문장부호가 최장명사구에 미치는 영향을 Penn Chinese Treebank 코퍼스에 출현한 빈도에 따른 상위 10개 문장부호의 사용상황이 표 2에 제시되어 있다.

위의 표에서 빈도수는 문장부호가 출현한 총 빈도수를 말하고 Inside MNP는 문장부호가 MNP에 포함된 상황을 말하며 Outside MNP는 문장부호가 MNP에 포함되지 않는 상황을 표시한다. 그리고 outside MNP 비율은 문장부호가 최장명사구에 포함되지 않는 비율을 표시한다. 표 2에서 알 수 있다시피 문장부호에 따라서 최장명사구에 포함되거나 포함되지 않는 뚜렷한 경향을 띠고 있어 중국어 최장명사구를 식별함에 있어서 중요한 역할을 할 것이라고 생각한다.

본 논문에서는 문장부호의 문법적 기능과 Penn Chinese Treebank에서의 사용상황에 근거하여, 또 학습코퍼스에서 문장부호의 데이터 희귀성(data sparseness) 문제를 고려하여 문장부호를 5개 그룹으로 나누었다. 예를 들면 코퍼스에서 95% 이상이 최장명사구에 포함되지 않은 문장부호 중에서 comma(,)와 period(。)는 문법적 기능이 다르기에 서로 다른 그룹으로 나누었지만, 문법적 기능이 같은 question mark(?)와 period(。)는 같은 그룹으로 분류하였다. 또 그룹 5는 코퍼스에서 아주 적게 출현하는 문장부호들이지만 다른 그룹들과 기능면에서 구별이 되기 때문에 하나의 그룹으로 간주했다. 문장부호의 상세한 분류는 표 3에 제시되어 있다.

표 2 빈도수에 따른 상위 10개 문장부호의 사용상황

문장부호	빈도수	Inside MNP	Outside MNP	Outside MNP 비율 (%)
Comma (,)	12, 695	404	12,291	96.82%
Period (。)	4,698	9	4,689	99.81%
Slight-pause mark (、)	2,725	2,306	419	15.38%
Brackets (「」)	1,666	1,217	449	26.95%
Question Mark (?)	308	12	296	96.10%
Semicolon (;)	302	10	292	96.69%
Colon (:)	191	9	182	95.29%
Quotation Mark (" ")	163	144	19	11.66%
Exclamation Mark (!)	131	0	131	100%
Brackets(())	114	86	28	24.56%

표 3 세분화된 중국어 문장 부호

그룹	문장부호
그룹1	Slight-pause Mark (、)
그룹2	Comma (,)
그룹3	Period (。), Question Mark (?), Exclamation Mark (!), Semicolon (;), Colon (：), ...
그룹4	Quotation Marks (“”, ‘’, 《》, <>, 「」), Brackets ({}, ()), ·····
그룹5	Hyphen (-), dash (--), apostrophe (’), Slash Mark (/), dot (.), ·····

4. 시스템 구조

본 논문은 두 단계에 거친 최장명사구 식별 모델을 제안한다. 첫 번째 단계에서는 규칙 기반의 방법으로 확장청크 식별을 수행한다. 이 단계에서 입력문장은 품사정보가 태깅된 문장이다. 본 논문에서 최종적으로 식별하고자 하는 것은 최장명사구이다. 확장청크 식별 단계에서 최종 최장명사구 식별에 악영향을 미치는 에러를 포함하면 이후 단계에서는 그 에러를 수정할 수 없다. 이런 에러 전달(error propagation) 문제를 줄이고자 첫 번째 단계에서는 재현율보다는 높은 정확률을 필요로 하는 규칙 기반 시스템을 제안한다.

두 번째 단계에서는 기계학습방법을 이용하는데 이 단계에서는 첫 번째 단계에서 확장청크가 묶인 문장이 입력된다. 기본적으로 품사정보와 그리고 제안한 세분화된 문장부호 정보, 첫 번째 단계에서 식별한 확장청크의 정보를 학습자질로 사용하여 최장명사구 식별을 진행하였다. 그리고 주변 문맥정보가 주는 영향을 살펴보기 위하여 윈도우 사이즈를 변화시키면서 학습을 수행하여 가장 적합한 윈도우 사이즈를 선택하도록 하였다.

본 논문에서는 Kudo[9]가 제안한 IOBES 태그를 채택하였고 따라서 최장명사구 식별은 아래 표에 기술된 5가지 클래스의 식별문제로 변환된다.

표 4 최장명사구 식별에서 IOBES 태그별 의미

MNP 태그	설명
MNP_B	최장명사구 시작 태그
MNP_E	최장명사구 종결 태그
MNP_I	최장명사구 시작 태그와 종결 태그 사이의 태그
MNP_S	최장명사구에 속하지 않는 태그
MNP_O	한 단어로 구성된 최장명사구 태그

5. 실험 및 분석

본 논문에서는 Penn Chinese Treebank 4.0을 실험 코퍼스로 사용하였다. Penn Chinese Treebank 4.0에서는 괄호를 이용하여 구 구조(phrase structure)를 표시하고 있는데, 이 구조체에서 제일 상위 노드 중 명사구, “的(de)”가 포함된 명사구, 숫자 리스트구(list mark phrase), 양사구(quantifier phrase)를 자동 추출하여 학습을 위한 최장명사구로 사용하였다[10].

실험 코퍼스는 5,000 문장을 사용하였고 문장부호를 포함하여 약 16 만 단어이며 문장의 평균 길이는 약 32 단어이다. 최장명사구는 약 28,000개이고 평균길이는 3~4 단어이다. 기계학습 알고리즘은 Support Vector Machines(SVM)[2]을 사용하였고 모든 실험은 10 fold cross validation을 진행하였다. 성능평가는 5개 태그의 F_1-measure 평균 값으로 하였다. 본 논문에서 사용한 정확률, 재현율과 F_1-measure 는 각각 아래의 방법으로 계산된다.

$$\text{정확률} = \frac{\text{정확하게 식별된 단어수}}{\text{시스템의 식별 결과}}$$

$$\text{재현율} = \frac{\text{정확하게 식별된 단어수}}{\text{정답에 있는 단어수}}$$

$$F_1 - measure = \frac{2 * \text{정확률} * \text{재현율}}{\text{정확률} + \text{재현율}}$$

본 논문에서는 최장명사구 식별을 위하여 두 단계로 나눈 접근방법과 제안한 학습자질들의 효율성에 대하여 각각 실험을 수행하였다. 그리고 윈도우 사이즈를 변화시키면서 최적의 윈도우 사이즈를 찾는 실험도 병행하였다. 윈도우 사이즈는 중국어 단어 분리기에 의해 분리된 토큰의 개수를 기반으로 설정되었다.

기본 모델은 품사정보만 유일한 자질로 사용했을 때를 가리킨다. 다시 말하면 모든 문장부호가 학습 코퍼스에서 ‘PU’라는 단일태그로 태깅되었을 때의 경우이다.

5.1 기본구와 확장청크 자질

이 실험에서 우리는 최장명사구 식별을 두 단계로 나눠서 진행하는 것 및 제안하는 확장청크의 자질이 기본구보다 얼마나 효율이 높은 지에 대해서 알아보고자 한다. 섹션 4 시스템 구조에서 설명하였다시피 첫 단계인 기본구 및 확장청크의 식별은 섹션 3.1에서 기술된 다양한 규칙들에 의해 이루어지게 되며, 에러 누적 문제를 효과적으로 해결하기 위하여 높은 정확률(97%)을 보여주는 규칙기반으로 진행하였다.

단어들이 기본구 및 확장청크로 묶이어서 청크에 포

2) LIBSVM(http://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html)에서 제공되는 기본모델과 파라미터 값 및 다중 분류 시스템을 사용하였다. LIBSVM이 제공하는 기본 커널은 RBF이다.

함된 단어들이 하나의 단위로 간주되어 처리되면 각 단어들이 확장청크 내에서 갖는 특성을 놓칠 가능성이 있다. 때문에 우리는 기본구 및 확장청크의 클래스 정보와 내부 단어들의 시작(B)과 종결(E), 그리고 중간(I)[3] 태그 등 위치정보를 유지하여 그 정보를 활용함으로써 보다 정확한 최장명사구의 식별이 이루어지도록 한다. 즉, 두 번째 단계인 기계학습방법에서는 확장청크로 식별된 것은 단어들의 품사정보 대신 확장청크를 나타내는 클래스정보와 확장청크의 시작과 종결 및 중간 태그를 사용하게 되고, 확장청크 이외의 단어들은 종전대로 품사정보를 자질로 사용하게 된다.

표 5에서 보다시피 두 단계로 나눠서 최장명사구 식별하는 것이 더 좋은 성능을 보였고, 특히 확장청크를 먼저 식별해 내는 것이 기본구보다 더 도움이 됨을 알 수 있다. 다만 성능제고의 폭이 우리가 기대하는 것보다 다소 낮게 나왔는데 이는 규칙 기반의 확장청크 식별이 아직은 낮은 재현율(65%)에 그치고 있음에 있다고 짐작할 수 있을 것 같다.

표 5 기본구와 확장청크를 사용했을 때 성능비교

윈도우 사이즈	기본모델 (F_1-Measure)	기본모델+ 기본구 (F_1-Measure)	기본모델+ 확장청크 (F_1-Measure)
3	84.02%	84.64%	86.01%
5	85.36%	86.02%	87.93%
7	86.65%	87.22%	88.36%
9	87.01%	87.52%(+0.51%)	88.42%(+1.41%)
11	87.01%	87.48%	88.36%

표 6 세분화된 문장부호 자질을 사용했을 때 성능

윈도우 사이즈	기본모델 (F_1-Measure)	기본모델 + 세분화된 문장부호 (F_1-Measure)
3	84.02%	85.90%
5	85.36%	88.33%
7	86.65%	89.14%
9	87.01%	89.58% (+2.57%)
11	87.01%	89.58%

5.2 세분화된 문장부호 자질

표 6이 보여주다시피 세분화된 문장부호 자질은 윈도우 사이즈가 9일 때 기본모델보다 무려 2.57% 향상되었다. 이는 코퍼스 분석을 통해서 문장부호가 최장명사구 식별에서 유용하게 쓰일 것이라는 우리의 추측을 확실히 검증하였다고 볼 수 있다.

3) 예를 들면 NP로 묶인 단어들이 있다면 그 단어들은 기본모델에서 사용한 품사정보 대신 NP_B, NP_E, 혹은 NP_I 등 클래스와 확장청크 내 위치정보를 자질로 사용하게 된다.

5.3 확장청크와 세분화된 문장부호

위 실험 5.1과 5.2에서 보다시피 윈도우 사이즈가 9일 때 기본모델을 포함한 기타 모든 시스템 성능이 수렴함을 알 수가 있었다. 그리고 제안한 확장청크와 세분화된 문장부호 정보를 모두 자질로 채택했을 때 기본모델보다 2.65% 향상됨을 표 7에서 보여줬다. IOBES 각 태그별 성능은 표 8에 제시하였다.

표 8에서는 MNP_B와 MNP_S 태그의 성능이 다른 태그들보다 약간 낮은 성능을 보이는데 이 부분에 대해서는 더 효율적인 자질을 필요로 하고 있음을 알 수 있다.

표 7 윈도우 사이즈가 9일 때 확장청크와 세분화된 문장부호 정보를 사용했을 때 성능

윈도우 사이즈	F_1-Measure
Baseline	87.01%
+ Classified Punctuations	89.58%
+ Expanded Chunks	89.66% (+2.65%)

표 8 윈도우 사이즈가 9일 때 IOBES 태그 별 성능

MNP 태그	정확률	재현율	F_1-Measure
MNP_B	84.48%	81.22%	82.99%
MNP_E	90.72%	93.52%	92.10%
MNP_S	86.35%	88.33%	87.33%
MNP_I	93.41%	87.66%	90.44%
MNP_O	93.19%	97.77%	95.43%

5.4 비교 분석

표 9는 기타 시스템과의 성능비교이다. Yin[7]과는 같은 코퍼스에서 실험을 했는데 제안한 방법이 10% 이상의 성능 향상을 보였다. 하지만 Zhou[6]의 방법과는 같은 코퍼스에서 실험을 진행한 것이 아니기에 객관적인 비교는 안되지만 Zhou[6]가 사용한 코퍼스의 최장명사구의 평균길이가 모두 3~4 단어임을 감안할 때, 우리가 제안한 중국어 최장명사구 식별 방법이 비교적 높은 성능을 보이고 있음을 알 수 있다.

표 9 다른 시스템과의 성능 비교

방법	F_1-Measure
Yin [7]	76.30%
Zhou [6]	83.82%
Proposed Method	89.66%

6. 후처리 작업

후처리 작업은 최장명사구 태그 수정 작업과 좌우 경계 매칭 작업 두 가지를 포함한다.

구묶음 태그 에러들을 수정하여 시스템의 정확률을 향상시키기 위해 태그 수정 규칙을 적용한다. 수정 규칙

은 자주 등장하는 오류를 자동으로 추출하고 그들의 좌우 문맥정보를 사용하여 정확한 최장명사구 태그로 수정하여 주는 규칙 10개를 사용하였다.

최장명사구 식별은 왼쪽과 오른쪽 경계를 정확하게 식별하고 이들을 매칭시켜야만 후속으로 이루어지는 구문분석이나 기계번역 등에 사용될 수 있다. 본 논문에서는 최장명사구 식별이 끝난 후 설정된 좌우 경계의 짝이 일치하지 않는 경우, 가장 가까운 좌우 경계를 매칭시킴으로써 최장명사구의 경계를 확정하였다.

표 10이 보여주다시피 수정 규칙을 적용한 후 시스템 성능은 다소 향상되었으나 최종 경계 매칭 작업을 수행한 후 시스템 성능이 저하되는 양상을 보였는데 이것은 우리가 아주 간단한 좌우 경계 매칭 알고리즘을 사용하였기 때문이다. 향후 보다 정교한 문법 정보나 코퍼스에서의 통계 정보 등을 사용하여 성능 향상을 꾀하는 것이 과제로 남아있다.

표 10 후처리 작업 후 IOBES 태그 별 성능

MNP 태그	제안된 모델	수정규칙 적용 후	경계 매칭 적용 후
MNP_B	82.99%	83.50%	81.95%
MNP_E	92.10%	92.50%	92.90%
MNP_S	87.33%	88.20%	88.83%
MNP_I	90.44%	90.60%	89.74%
MNP_O	95.43%	95.40%	94.28%
평균	89.66%	90.04%	89.54%

7. 에러 분석

후처리 작업 후 실험에서 발생한 에러를 품사 별로 분류해보면 표 11과 같다. 에러 유형 중 명사가 가장 큰 비율을 차지하고 있는데, 이는 모든 명사는 기본적으로 최장명사구에 속하지만 그 단어가 최장명사구의 시작이나 내부, 종결 또는 한 단어로 된 명사구 등 4개의 태그 중 하나로 태깅될 수 있기에 정확하게 식별하기 어렵다는 점이 작용한 것이다. 품사가 명사인 단어들을 더 잘 처리할 수 있는지 여부가 향후 최장명사구 식별의 성능을 크게 좌우할 것이다.

그리고 중국어에서는 주제어(topic)와 주어가 한 문장에 동시에 출현하며 또 단순 명사들의 나열로만 표현되는 경우가 있는데, 이때 주제어와 주어의 경계 즉 최장명사구의 경계를 정확히 식별해 주지 못하는 경우가 있다.

표 11 품사 별 에러율

품사	빈도수	비율
Noun	5479	37.53%
Verb	2897	19.85%
Adverb	1263	8.65%
Preposition	993	6.80%
Others	3966	27.17%

그리고 동사, 전치사 및 부사는 일반적으로 최장명사구에 속하는 경향이 낮지만 특정한 경우에는 최장명사구에 속하게 되는데 이러한 경우 역시 오류를 유발하게 된다.

Ex 4: [中国(중국)/NR 边境(변경)/NN 开放(개방)/NN 城市(도시)/NN] [经济(경제)/NN 发展(발전)/NN 成就(성과)/NN] 显著/VA 。/PU

Ex 5: [最(가장)/AD 常(자주)/AD 挂(사용)/VV 在/P 嘴/NN 上/LC 的/DEG 一/CD 句/M 话/NN] 是/VC……

위 예문에서 [ ]으로 묶인 부분이 최장명사구이다. 예문 4에서 '中国边境开放城市'는 주제어이고, '经济发展成就'는 주어이다. 한국어에서는 조사가 사용되고 또 띄어쓰기가 있기 때문에 주제어와 주어를 쉽게 판단할 수 있지만 중국어에서는 보다시피 단순한 명사의 나열이다. 때문에 이 문장에서 '중국 변경 개방 도시'라는 주제어와 '경제 발전 성과'라는 주어를 따로 최장명사구로 정확하게 식별해주지 못했는데 이런 종류의 에러는 명사에서 자주 발생하는 에러 중의 하나이다. 이 경우는 앞으로 결합가 정보(valency infomation)를 이용하여 명사와 동사 사이의 유사성을 계산하는 방법으로 문제 해결을 시도 해볼 수 있을 것이다.

예문 5에서 '가장', '자주', '사용' 등 단어들의 품사는 부사 및 동사로서 일반적으로 최장명사구에 속하지 않지만 가끔 예문과 같이 관형구를 형성하여 최장명사구에 속하게 된다.

이러한 문제를 해결하기 위하여서는 보다 정교한 언어학적 지식이 필요할 것이며 우리가 앞으로 해결해야 할 과제이다.

8. 결론 및 향후 연구

본 논문에서는 두 단계로 나누어 접근하는 방법 즉 먼저 확장청크를 식별하고 후에 최장명사구를 식별하는 방법이 문제 해결에 더 많은 도움이 된다는 것과, 또 문장부호의 그룹별 사용은 최장명사구를 식별하는데 있어서 중요한 역할을 한다는 것을 실험을 통해 보여줬다. 즉, 확장청크와 세분화된 문장부호의 자질은 최장명사구 식별에 아주 유용한 자질임을 확인하였다.

그러나 확장청크의 인식은 이후에 수행될 처리 과정에 큰 영향을 미치기 때문에 확장청크 인식을 위한 규칙의 설정에는 세심한 검토가 필요할 것을 지적하였다. 현재 확장청크의 식별 성능에서 정확률은 97%로 높으나 재현율은 65%로 아주 낮기에 정확률을 훼손하지 않으면서

재현율을 높이는 방법으로 확장청크의 식별에서의 성능 향상이 필요하다. 에러분석에서 알 수 있다시피 많은 에러가 명사에서 발생하는데 명사에서 나타나는 에러를 줄이는 방법도 모색되어야 할 것이다. 후처리 작업에서도 단순히 가장 가까운 좌우 경계를 매칭시키는 방법을 사용하여 감소된 성능을 보다 정교한 규칙을 제안하여 후처리 작업의 성능향상을 꾀할 수 있을 것 같다.

최장명사구의 식별 결과를 구문분석이나 통계기반 기계번역 시스템에 적용시켜 그 간접성능을 구해 보는 것 역시 향후에 해야 할 과제로 남아있다.

참 고 문 헌

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

[2] 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.

[3] 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 38 Computational Linguistics, New York: Academic Press, pages 234-241, 1994.

[4] 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.

[5] 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.

[6] Qiang Zhou, Maosong Sun and Changning Huang, "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.

[7] Changhao Yin, "Identification of Maximal Noun Phrase in Chinese: Using the Head of Base Phrases," Master Dissertation, 2005.

[8] 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.

[9] 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.

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

[11] 김영택 외 공저, "자연언어처리", 생능출판사, 2001.

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

[13] Ming Zhou, "A Block-Based Robust Dependency Parser for Unrestricted Chinese Text," In: Proceedings of the Second Chinese Language Processing Workshop, Hongkong, pages 78-84, 2000.

[14] 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.

[15] Yuqi Zhang, Qiang Zhou, "Chinese Base-Phrases Chunking," In: COLING-02: The First SIGHAN Workshop on Chinese Language Processing, 39 pages 131-135, 2002.

[16] 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.

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

[18] Shui-fang Lin, "study and application of punctuation," (標点符号的學習与應用). People's Publisher, P.R.China (in Chinese), 2000.

[19] 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.

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

[21] LIBSVM: A Library for Support Vector Machines http://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html

백 설 매

2003년 중국 연변과학기술대학교 학사 2006년 포항공과대학교 정보통신대학원 석사. 2006년~2008년 (주)JUPICON 2008년~현재 (주)4U Applications. 관심분야는 자연언어처리, 중국어분석 등

이 금 희

1999년 중국 길림공업대학교 학사. 1999년~2001년 연변과학기술대학교 강사 2003년 포항공과대학교 컴퓨터공학과 석사. 2003년~2005년 포항공과대학교 정보통신연구소 연구원. 2005년~현재 포항공과대학교 컴퓨터공학과 박사과정. 관심분야는 자연언어처리, 중한 기계번역, 중국어 분석 등

김 동 일

1983년 부산대학교 기계공학과 학사. 1989년 Florida Institute of Technology 전산학 석사. 1995년 University of Florida 전산학 박사 수료. 2003년 포항공과대학교 컴퓨터공학과 박사. 1995년~현재 중국 연변과학기술대학교 컴퓨터전자통신학부 교수. 1999년~현재 중국 연변과학기술대학교 언어공학연구소 소장. 관심분야는 자연언어처리, 중한 기계번역, 중국어 분석 등

이 종 혁

1980년 서울대학교 수학교육학과 학사 1982년 한국과학기술원 전산학과 석사 1988년 한국과학기술원 전산학과 박사 1989년~1991년 일본전기(NEC) 중앙연구소 초청연구원. 1991년~현재 포항공과대학교 컴퓨터공학과 교수. 1998년~1999년 미국 CRL/NMSU(뉴멕시코주립대학) 방문교수. 2007년~2008년 캐나다 RALI/University of Montreal 방문연구원. 관심분야는 자연언어처리, 기계번역, 정보검색 등