문장부호를 사용한 효과적인 중국어 최장명사구 식별기법

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

An Effective Approach Using Sentence Symbols to Identify Maximal-Length Noun Phrase in Chinese

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

요 약

일반적으로 중국어의 명사구는 최단명사구, 기본명사구, 최장명사구로 분류된다. 최장명사구에 대한 정확한 식별은 문장의 전체적인 구조를 파악하고 문장의 정확한 지배용언을 찾아내는데 중요한 역할을 한다. 본 논문에서는 특성에 따라 5개의 클래스로 세분화된 문장부호를 학습자질로 사용하여 최장명사구 자동식별을 진행한다. 제안된 기법은 평균길이가 4인 최장명사구의 식별실험에서 기본모델(baseline)보다 4.5% 향상된 평균 85.1%의 우수한 F-measure 성능을 보인다.

1. 서론

Abney[1]가 처음으로 구문분석을 두 단계로 나누어 처리하는 방법, 즉 구묶음(chunking)을 구문분석의 전처리 단계로 적용한 방법을 제시한 이래 중국어 구문분석도 이와 비슷한 접근방법을 이용하고 있다.

여러 가지 구 단위에서도 특히 명사구는 문장에서 가장 중요한 구성성분이자 의사전달에 있어서 없어서는 안될 기본단위이다. 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, 기본명사구, 최장명사구 3가지[2]로 나눌 수 있는데, 일반적으로 최장명사구(MNP, maximal-length noun phrase)는 "다른 명사구에 포함되지 않은 명사구"로 정의된다. 하지만 이런 정의에는 논쟁이 뒤따른다. 중국어는 주제어 부각형(topic prominent) 언어이다. 따라서 문장에 주제어와 주어가 동시에 나타나는 경우가 많다[3]. 예를 들면

예) MNP[那/DT 只/M 狗/NN] MNP[我/PN] 已经/AD 看/VV 过/AS 了/AS 。/PU[1]
(해석: 저 개는 내가 본 적이 있어.)

이 문장은 주제어와 주어가 동시에 한 문장에 출현한 예제이다. 주제어와 주어 모두 명사구이기 때문에 같이 묶어서 하나의 최장명사구를 구성하게 되면 "那只狗我(저 개는 내가)"와 같이 부자연스러운 최장명사구가 생성된다. 따라서 이런 문장에서는 주제어와 주어를 나누어 최장명사구를 구성해야만 문장이 표현하고자 하는 정보를 정확하게 전달하게 된다. 기존의 연구들에서는 이 문제에 대하여 명확한 정의를 내리지 않았기 때문에 시스템마다 다르게 구현되었다. 그러므로 최장명사구 식별연구라 해도 서로 구의 경계를 다르게 보기 때문에 명확한 성능의 비교가 힘들다. 따라서 본 논문은 중국어 최장명사구를 "주제어와 주어를 동시에 포함하지 않으면서 다른 명사구에 포함되지 않은 명사구"로 좀 더 상세한 정의를 내린다. 이렇게 최장명사구를 구분해주어야만 구문분석과 변환, 나아가서 기계번역에 보다 좋은 정보를 제공할 수 있게 된다. 본 논문은 이런 최장명사구에 대하여 식별한다.

본 논문은 문장부호가 문맥에 주는 영향을 고려하여 문장부호를 기능에 따라 분류하였는데 이 정보를 학습자질로 사용하여 최장명사구 식별성능을 향상시킨다.

2. 관련 연구

중국어 최장명사구 식별에 대한 기존연구를 살펴보면 아래와 같다.

1) Li[4]는 단어의 품사 태그 정보를 이용하여 통계적 방법으로 최장명사구의 자동 식별을 시도하였는데 open test의 정확률과 재현율이 각각 71.3 %, 69.1% 이었다.

2) Tse[5]는 통계와 규칙의 혼합 방법을 사용하였다. 이 논문에서는 "的" 를 포함한 명사구만을 식별하였는데

[1] Penn Tree Bank 의 품사 태깅을 사용하였다.
DT: 한정사, M: 양사, NN: 일반명사, PN: 대명사, AD: 부사, VV: 일반동사, AS: 어기사, PU: 문장부호

정확률과 재현율이 각각 75%, 90% 이었다.

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 식별하고 계속하여 왼쪽으로 확장해 나가면서 왼쪽 경계를 식별하였다. 이 방법은 5단어 이상의 최장명사구에 대하여 정확률 85.4%, 재현율 82.3%의 성능을 보여주었다.

4) Yin[6]은 두 단계 학습모델을 이용하여 최장명사구 자동식별을 진행하였는데 평균길이가 7인 최장명사구의 식별에서 4개 태그[2]의 평균 F-measure가 92.5%의 성능을 보여주었다.

위의 기존 연구들에서의 문장부호의 품사 태그는 모두 단일 태그였다. 다만 Zhou[2]가 전처리 단계에서 문장부호를 단어의 좌우경계를 식별하는데 사용했을 뿐, 문장부호가 기존의 연구들에서 최장명사구 식별의 중요한 자질로 사용된 적은 없었다.

3. 문장부호의 분류

Chinese Penn Tree Bank에는 총 32종류[3]의 문장부호가 출현하지만 이들은 모두 단일 품사태그인 PU로 태깅되었다. Penn Tree Bank에서 각 문장부호의 사용상황이 아래의 표1에 제시되어 있다.

[표 1] 문장부호 사용상황[4]

Punctuation	Total	Inside	Outside
Comma (,)	15,262	754	14,508
Period(。)	7,380	16	7,364
Slight-pause mark (、)	5,590	5,258	332
Bracket (「)	1,178	835	343
Bracket ())	1,178	830	348
Quotation mark (")	830	689	141
Quotation mark (")	830	689	141
Quotation mark (《)	417	409	8
Quotation mark ())	417	409	8
Semicolon (;)	409	26	383
기타	1,665	1,040	625
합계	35,156	10,955	24,201

위의 표에서 알 수 있다시피 문장부호는 중국어 최장명사구를 식별함에 있어서 중요한 역할을 하게 된다. 본 연구에서는 문장부호의 문법적 기능[7]과 Penn Tree Bank에서의 사용상황에 근거하여 문장부호를 아래와 같은 5개 그룹으로 분류하였다.

[표 2] 문장부호 분류

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

4. 실험 및 결과 분석

실험은 Chinese Penn Tree Bank 3.0 을 코퍼스로 사용하였으며, 최장명사구가 태깅된 학습코퍼스는 Penn Treebank 의 정의에 따라 자동으로 구축하였다. 실험 코퍼스는 8,000 문장, 약 25 만 단어이고 문장의 평균 길이는 문장부호를 포함하여 31 단어이다. 최장명사구는 37,285 개이고 평균길이는 4 이다. 기계학습 알고리즘은 Naïve Bayes 를 사용하였고 테스트는 10 fold cross validation 으로 진행하였다.

Baseline 은 Penn Treebank 의 단어 품사정보(문장부호가 PU 로만 태깅됨) 만을 학습자질로 사용하여 최장명사구 식별을 한 것이다. Baseline 의 실험결과는 윈도우 크기 9 에서 제일 좋은 성능을 보여줬는데 4 개 태그의 평균 F-measure 는 80.6%였다.

4.1 최장 명사구 식별 모델 구축

[표 3] 최장 명사구 태그 형식

태그	설명
MNP-S	최장 명사구 시작 태그
MNP-E	최장 명사구 종결 태그
MNP-I	최장 명사구 시작 태그와 종결 태그 사이의 태그
MNP-O	위의 태그가 아닌 기타 태그

최장 명사구 식별은 위의 표 3 에 기술된 4 가지 클래스의 식별문제로 변환된다. 학습자질로는 단어 레벨에서 단어의 품사정보(문장부호가 단일 태그로만 태깅됨)와 분류를 나눈 문장부호의 품사정보이다. 그리고 윈도우 크기를 바꾸면서 성능이 제일 좋은 윈도우 크기를 찾았다.

4.2 문장부호 자질 추가에 의한 성능 향상

Baseline 이 윈도우 크기 9 에서 제일 좋은 성능을 보였으므로 일관된 성능비교를 위해 윈도우 크기를 9 로 고정시킨 다음, 표 2 에서 분류한 문장부호의 부류를 다양하게 조합시키면서 성능을 살펴보았다. 아래는 문장부호 태그를 각각의 다양한 조합으로 세분화시켰을 때 시스템 성능의 변화결과이다.

---

[2] MNP-O: 최장명사구 시작, MNP-C: 최장명사구 종결, MNP-I: 최장명사구 시작과 종결 사이의 태그 MNP-NUL: 기타 태그

[3] 전체 Penn Tree Bank 에서는 총 36종류의 문장부호가 사용되었지만 본 논문에서 사용한 코퍼스에서는 32종류가 사용되었다

[4] inside: 문장부호가 MNP의 경계 안에 있는 상황
outside: 문장부호가 MNP의 경계 밖에 있는 상황

[표 4] 각 문장부호 태그 사용시 성능

문장부호 태그의 설정	F-measure 평균
그룹1 과 기타로 분류	82.2% (+1.6%)
그룹2 와 기타로 분류	82.3% (+1.7%)
그룹3 과 기타로 분류	81.2% (+0.6%)
그룹4 와 기타로 분류	82.2% (+1.6%)
그룹5 와 기타로 분류	80.7% (+0.1%)
그룹1,2 와 기타로 분류	83.2% (+2.6%)
그룹1,2,3 과 기타로 분류	84.9% (+4.3%)
그룹1,2,3,4,5로 분류	85.1% (+4.5%)

위의 실험결과에서 보다시피 그룹 2(comma) 가 가장 높은 성능향상을 보였고 그 다음은 그룹 1(slight-pause mark)과 그룹 4(bracket)의 순서였다. 하지만 그 성능차이는 그다지 크지 않았다. 비록 그룹 5 가 가장 낮은 성능향상을 보이고 있지만, 그룹 5 에 속한 문장부호의 낮은 출현빈도 때문으로 판단되므로 이 그룹을 유지시킨다. 그 다음은 각 그룹을 하나하나씩 추가하면서 실험을 진행한다. 위의 실험결과에서 보다시피 각 그룹을 추가하였을 조금씩 성능향상을 보였으며 문장부호를 표 2 에서와 같이 5 개의 그룹으로 분류하였을 때 최고의 성능을 보였다.

4.3 최장명사구 식별결과

아래는 Penn Tree Bank 의 기본 품사정보(PU 를 제외)와 표 2 에서 분류한 문장부호를 학습자질로 사용하여 각기 다른 윈도우 크기에서 실험한 결과이다.

[표 5] 각 태그 별 정확률

문맥 길이	MNP_S	MNP_E	MNP_I	MNP_O
3	77.6%	80.6%	87.5%	85.8%
5	78.5%	83.5%	88.4%	87.4%
7	78.8%	84.5%	88.4%	87.9%
9	79.0%	84.6%	88.4%	88.1%
11	79.0%	84.6%	88.3%	88.1%
13	79.0%	84.6%	88.4%	88.1%

[표 6] 각 태그 별 재현율

문맥 길이	MNP_S	MNP_E	MNP_I	MNP_O
3	69.9%	87.5%	82.2%	94.7%
5	74.6%	87.8%	84.7%	93.1%
7	74.9%	87.3%	85.7%	92.9%
9	75.0%	87.1%	86.0%	92.7%
11	74.9%	87.2%	86.1%	92.7%
13	74.9%	87.1%	86.1%	92.7%

[표 7] 각 태그 별 F-measure

문맥 길이	MNP_S	MNP_E	MNP_I	MNP_O	평균
3	73.5%	83.9%	84.8%	90.0%	83.0%
5	76.5%	85.6%	86.5%	90.1%	84.7%
7	76.8%	85.9%	87.0%	90.3%	85.0%
9	76.9%	85.8%	87.2%	90.4%	85.1%
11	76.9%	85.9%	87.2%	90.4%	85.1%
13	76.9%	85.8%	87.2%	90.4%	85.1%

위의 실험에서 보시다시피 윈도우 크기 9 이상부터는 별다른 성능 차이를 보이지 않았다.

5. 결론 및 향후 연구

본 논문에서는 문장부호를 특성에 따라 5 가지 분류로 나누었다. 이 5 가지 그룹을 하나하나씩 추가할 때 모두 성능향상을 보였으며 본 논문에서 제시한 방법대로 분류하였을 때 baseline 보다 4.5%가 향상되어 가장 높은 성능을 보였다. 본 논문은 실험을 통하여 문장부호가 중국어 최장명사구 식별에 중요한 자질로 작용한다는 것을 입증하였다.

향후 과제로는 후처리 작업의 추가 및 오류 분석을 통해 발생빈도가 높은 오류를 식별할 계획이다. 그리고 보다 최적의 문장부호의 분류를 찾기 위한 실험을 수행한다. 자질 선택에서도 valency 정보 등을 더 추가하여 성능을 높일 수 있다. 또 다양한 기계 학습 방법들을 적용함으로써 본 최장명사구 식별 문제 해결에 가장 효과적인 기계 학습 방법을 찾는다.

6. 참고문헌

[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*", 1998

[3] Charles N.Li, Sandra A. Thompson Mandarin Chinese (A Functional Reference Grammar) 2001

[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, (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, 1995. (252-257)

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

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

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