

접속사가 없는 경우의 중국어 대등접속문 인식

김미훈[○] 김미영 김동일 이종혁

포항공대 정보통신대학원[○], 포항공대 컴퓨터공학과, 첨단정보기술 연구센터
{meixun[○], colorful, dongil, jhlee}@postech.ac.kr

Identification of Chinese Coordinate Structure in case of no conjuncts

Meixun Jin[○] Mi-Young Kim Dongil Kim Jong-Hyeok Lee
Dept. of Graduate School for Information Technology, POSTECH[○]
Div. of Electrical and computer Engineering, POSTECH
Advanced Information Technology Research Center(AITrc)

요 약

일반적인 언어에서 대등접속문은 접속사와 함께 나타나지만, 중국어 대등접속문은 접속사 없이도 이루어 질 수 있다. 이러한 경우 조사정보가 대등접속문을 판단하는데 중요한 정보가 된다. 본 논문에서는 중국어 조사(助词)정보를 고려하여 접속사가 없는 중국어 대등접속문을 인식하는 방법을 제안한다.

1. 서 론

일반적으로 대등접속문은 접속사(영어에서 and... 등; 한국어에서는 -고, -와 ... 등)가 필수적인 요소로 쓰여, 접속사를 중심으로 대등접속문임을 판단할 수 있지만, 중국어에서는 접속사 없이도 대등접속문을 이룰 수 있다.

기존연구들은[1,2,6] 접속사가 이미 파악된 상황에서 접속사 좌우의 병렬성을 이용하여 대등접속문을 판단하는 방법을 제안했다. [1]에서는 임의의 문장에서 접속사 좌우의 단어(bunsetsu)간 유사도를 계산하여 선접속구(preconjunct)와 후접속구(postconjunct)를 판단한다. [2]에서는 선머리어(선접속구의 머리)와 후머리어(후접속구의 머리) 및 선접속구와 후접속구의 구조적 병렬성을 이용하여 판단한다. [2]는 [1]과 달리, 선접속구와 후접속구의 유사성 판단시 각 접속구 내의 모든 단어를 이용한 것이 아니라, 명사구의 헤드만을 이용하였다.

또한 기존의 대등접속문은 구성 단어의 품사에 따라 NP 대등접속문과 VP 대등접속문으로 나누고 있다. 이러한 대등접속문 판정 및 분류법은 중국어 대등접속문을 판단하고 분석하기에는 부족하다.

중국어에서는, 접속사 없이도 대등접속문을 만들 수 있으므로, 중국어 문장 속에 대등접속문이 존재하는지 여부를 먼저 판단할 수 있어야 한다. 접속사가 있을 경우에는 접속사에 의해 자연스럽게 대등접속문이 있음이 판단 된다. 하지만, 접속사가 없는 대등접속문은 조사와 함께 나타나는 경우가 비교적 많이 있어, 조사 정보로 대등접속문이 있는지 여부에 대한 판단을 간접적으로 할 수 있다.

본 논문에서는 먼저 접속사가 있을 경우, 대등접속의 병렬성을 중국어에 맞게 계산하는 방법을 제안하고, 접속사가 없을 경우 조사정보를 이용하여 대등접속문을 분별하는 방법을 제안한다.

2. 중국어 대등접속분석

2.1 중국어 대등접속문의 분류

중국어 대등접속문을 접속사가 있는 대등문과 접속사가 없

접속사가 있을 경우	예제 1: <u>学生和老师们</u> 都来到学校礼堂. <u>학생들과 선생님들</u> 은 학교강당에 모였다.
	예제 2: <u>他吃光</u> 了 <u>苹果和梨</u> . 그는 <u>사과와 배</u> 를 다 먹었다.
접속사가 없을 경우	예제 3: ... <u>宽敞明亮</u> 的教室... <u>넓고 밝은 교실</u> ...
	예제 4: ... <u>全面科学地</u> 揭示了市场经济规律... <u>전반적으로 과학적으로</u> 시장경제의 규칙을 제시했다.
	예제 5: ... <u>参与组织指挥</u> 了这一工程. ...이 프로젝트에 <u>참여했고 조직했고 지휘했다</u> .

[표1] 대등접속문의 구문적 분류 예제

는 대등문으로 분류하고, [표1]은 그 예들을 보여준다. 표 안의 밑줄 친 부분은 대등접속구를 나타내고 있다. 예제 1에서는 두 명사 学生() 老师()가 접속사 和(와)와 함께 대등접속문이 구성되어 이 문장에서 주어역할을 한다.

예제 2에서는 명사 대등접속문이 문장에서 목적어 역할을 하고 있다.

예제 3에서 두 형용사 宽敞() 明亮() 접속사 없이 대등접속문이 구성되어 뒤에 오는 명사 教室()을 수식하고 있다. 예제 4, 예제 5에도 접속사가 없이 대등접속문을 이루어서 목적어와 서술어기능을 하고 있다.

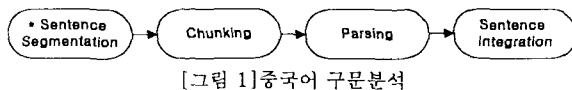
2.2 접속사가 있는 대등접속문의 판단

예제 1과 예제 2에서 보여 준 것과 같이 주어역할과 목적어 역할을 하는 대등접속문은 명사나 명사구로 이루고 있어 이는 기존연구에서 NP 대등접속문에 해당한다. 중국어에서 NP 대등접속문인 경우에는 대등접속연사(连词)로 연결되어 있다.

중국어에서 NP 대등접속문 접속사들은 아래와 같다. “和”, “与” () “及” “以及” () “或” “或者” “抑或” () “, ” “ ”는 명사나 명사구의 대등접속을 나타내는 부호이다.

중국어 NP 대등접속문 또한 병렬성 및 대등성이 두드러지게 나타나고 있다. 본 논문에서 중국어의 병렬성에 대한 검사는 휴리스틱에 의해 판단하고, 휴리스틱 값들은 중국어 긴 문장 약 50개에서 테스트하고 여러가지 영향요소를 조절함으로써 얻은 결과이다.

먼저 본 연구실의 Total- C/K의 중국어구문분석을 살펴보면 아래 그림 1과 같다.



[그림 1] 중국어 구문분석

(*부분은 중국어 문장을 쉼표로 나눈 분할이다.)

그림 1에서 문장분할은 중국어 쉼표로 나눈 분할을 말하고 있다. 중국어에서 쉼표는 아래와 같은 상황에서 사용한다. [7 참조] (1) 한 문장의 각 절 사이에; (2) 주어에 수식성분이 많이 붙어서 길어졌을 경우 주어 뒤에; (3) 목적어에 수식성분이 많이 붙어서 길어졌을 경우 목적어 뒤에; (4) 부사어 뒤에; (5) 대등접속문의 연결 ;등의 상황이 있다.

예를 들면, 凋谢的时候() (1), 花瓣也不会枯萎或各自纷纷飘落()

() (2), 只是在某一瞬间() (3), 整朵花像一团火或咯血一样同时掉落()

() (4), (1)과 (3)은 부사어 뒤에 사용된 경우이고, 쉼표(2)는 대등접속 연결을 나타내고 있다.

문장분할 다음 단계로 각 분할 안에서 구둑음이 이루어지고, 각각에 대하여 파싱을 하게 된다. 마지막으로 파싱을 한 각 분할들을 합치고 의존관계를 설정한다. NP 대등접속문 판단은 구둑음에 포함된 단계로서 다른 구둑음이 이루어진 뒤에 마지막 구둑음 과정에 진행된다.

두 단어의 유사도 계산에 사용한 리소스는 [표2] 과 같다. 이 중에서 북경대와 동북대에서 사용하는 하위품사는 분류의 기준에 차이가 있다. 그러므로 분류한 단어에 따라서 어떤 단어는 북경대 하위품사를 적용하는 것이 바람직하고, 또 어떤 단어는 동북대 하위 품사를 적용하는 것이 더 바람직하여, 표2와 같이 계산하는 방법을 제안했다. 중국어의 품사정보만으로 중국어 구문분석에 정보를 주기에 너무 부족하기에, 한국어 대역어의 품사의 정보도 사용하여, 부족한 부분을 보완하고자 한다.

문자열 정보	같은 단어	+15->끝
	같은 글자수	+ 같은 글자수
	단어의 길이	+ (strlen- 5)
품사 정보	품사	n, a, v 일 경우 같은 품사 +3; 다른 품사일 경우 +2
	북경대하위품사 동북대하위품사	북경대, 동북대 다 같은 경우 +3; 하나만 같은 경우 +1
의미 정보	[3]	First Level: +1 Second Level: +2 Third Level: +4 Fourth Level: +5
대역어 정보		같은 경우 +1

[표 2] 유사도 계산요소

2.3 접속사가 없는 대등접속문의 판단

[표1]의 예제에서 보여주는 것처럼 중국어에서 대등접속문은 접속사가 없이도 가능하며 해당 문장에서 관형사, 부사 또는 서술어가 될 수 있다.

예제(1)에서 조사(助词) “的”는 한국어 “-은, -의”에 해당하고 앞에 宽敞() 明亮() 두 형용사로 구성된 대등접속문이 뒤에 오는 명사 教室(교실)을 수식하도록 연결하고 앞에 오는 구가 뒤에 오는 명사의 관형어가 된다는 것을 나타낸다. 예제(2)에서 조사 “地”는 한국어 “-으로”에 해당하며 앞에 全面(전면적)과 科学(과학적) 두 형용사로 구성된 대등접속문과 함께 부사어가 되어 뒤에 오는 동사 揭示()를 수식하여 앞에 오는 대등접속문이 뒤에 오는 동사의 부사라는 것을 말해주고 있다. 예제(3)에서는 조사 “了”는 한국어로 “-었다”이고, 본 예제에서는 동사 参与() 组织() 指挥() 세 개로 된 대등접속문 뒤에 연결되어서 서술어가 되고 있다.

위의 예제에서도 본 것처럼 조사 “的” “地” “了”는 대등접속문과는 아무런 관계가 없다. 하지만, 이러한 조사들이 연결어가 없는 대등접속문과 많이 나타나고 있기에 이러한 조사들은 대등접속문이 있을 가능성을 보여주고 있으므로, 조사 앞에 같은 품사나 패턴이 반복적으로 여러 개가 나타날 경우, 아래와 같은 휴리스틱을 이용하여 대등접속문을 판단할 수 있다.

(1) 조사와의 결합정보: 조사 앞에 같은 품사가 연속적으

로 반복될 경우, $A_n A_{n-1} A_{n-2} \dots A_1$ (의(地, 了)와 같은 패턴은 A_1 에서 A_n 까지 어법정보사전[5]으로부터 해당 조사와의 결합정보를 참조하게 된다. 결합이 가능하면 대등접속 후보로 선정이 되고 (2)단계로 연결된다. 결합이 불가능하면 대등접속을 이루지 못한다.

(2) 조사뒤의 명사구나 동사구와의 결합정보: 조사앞에 A_1 에서 A_n 까지 조사가 " 的" 일경우 조사 뒤의 명사구, 조사 " 地" 일경우 조사 뒤의 동사구, 조사 " 了" 일경우 조사 뒤의 명사구와의 결합정보를 보게 된다. 결합이 모두 가능하면 대등접속문을 판단할 수 있다.

예제 5에서 "...参与组织指挥了这一工程. 이 문장은 参与了这一工程() 组织了这一工程()

이렇게 세 부분을 묶어 놓은 것으로 볼 수 있다. (1)단계 판단에서는 각 동사가 뒤에 오는 조사와의 결합이 가능하여 (2)단계에서는 뒤에 오는 목적어와의 결합정보를 보게 된다. 모두 가능하기에 대등접속문으로 판단된다. 예제 5에서 보여준 것과 같이 접속사가 없는 대등접속문에는 병렬성 또한 존재한다.

이 방법은 대등접속문 분별뿐만 아니라, 뒤에 오는 명사구나 동사구와의 관계 설정 정보도 조사로부터 바로 알 수 있다.

4. 실험 및 평가

실험에 사용한 데이터는 중국 인민일보 말뭉치에서 출현한 길이가 70단어 이상인 문장 200개 이다. 실험결과를 각각 다음과 같다.

(1) 접속사가 있는 경우

문장수	문장내의 모든 대등접속구가 바르게 판단된 문장수	문장내의 대등접속구 중, 하나라도 잘못 판단된 문장수	정확률
200	158	42	79.0

[표3] 실험결과

대등접속문수	정확하게 판단된 대등접속문수	잘못 판단된 대등접속문수	정확률
583	471	112	80.7

[표4] 실험결과

(2) 접속사가 없는 경우

대등접속문수	정확하게 판단된 대등접속문수	잘못 판단된 대등접속문수	정확률
61	52	9	85.2

[표5] 실험결과

$V_n V_{n-1} V_{n-2} \dots V_1$ (NP)와 같은 구조에서 대등접속이 아닌 경우에는 V_1 가 조동사일 경우였다. 예, ...建起了高楼.()에서 조사 " 了" 앞에 동사 " 建" (짓-) 과 " 起" 로

구성되어 있고, 두 번째 동사 起는 방향성을 가리키는 조동사이다.

$V_n V_{n-1} V_{n-2} \dots V_1$ (NP)와 같은 패턴에서 항상 이러한 상황인지에 대해서는 더 많은 연구가 있어야 할 것 같다.

5. 결론 및 향후 작업

본 논문에서는 중국어 대등접속 접속사가 있는 대등접속문과 접속사가 없는 대등접속문의 판단방법을 제안했다.

실험결과에서 보여준 것과 같이 비교적 좋은 성능을 보여주고 있다.

향후 작업으로는 아래와 같다.

(1) 조사가 나타나지 않고 접속사 없이 대등접속문이 이루어 지는 경우: 아직 이러한 경우를 찾지 못했지만, 코퍼스에서 더욱 세밀한 확인작업이 필요하다고 생각한다. 만약 이러한 경우가 있다면 어떻게 판단할 것인지에 대한 연구가 필요하다.

(2) 점표에 대한 연구: 중국어 점표는 앞에서도 설명했던 것처럼 다양한 기능을 가지고 있다. 특히 구조가 복잡한 대등접속문을 연결할 경우가 비교적 많다. 그러므로 중국어 문장내의 점표에 대한 기능적 분별이 필요하다. 이러한 분별은 중국어 긴 문장 구문분석 및 점표로 연결된 대등접속 구문분석에 많은 도움을 줄 것으로 기대된다.

6. 감사의 글

본 연구는 첨단정보기술 연구센터를 통하여 과학재단의 지원을 받았습니다.

7. 참고문헌

- [1] S.Kurohashi and M.Nagao, " A syntactic analysis method of long Japanese sentences based on the detection of conjunctive structures", computational Linguistics, V.20, No.4, pp.507-534, 1994
- [2] 윤준태, " 한국어의 대등접속구문 분석", 정보과학회논문지(B) 제24권 제3호, P326-335, 1997
- [3] 梅家驹, 竺一鸣, 高蕴琦, 殷鸿翔, 《同义词词林》, 上海辞书出版社, 上海, 1983
- [4] Ming Zhou, " A Block-based Robust Dependency Parser for Unrestricted Chinese Text", Proceedings of the Second Chinese Language processing Workshop, HONGKONG, 2000
- [5] 王惠, 詹卫东, 刘群, " 《现代汉语语义词典》的概要及设计" in Proceedings of 1998 International Conference on Chinese Computing, Beijing, 1998
- [6] Rajeev Agarwal, Lois Boggess, " A Simple But Useful Approach to Conjunct Identification.", Proceedings of the 30th annual meeting of ACL, 1992
- [7] 胡明亮, " 逗号, 汉语教学的一个难点", 《国际汉语教学学术研讨会论文集》, pp.189-192, 2001