중국어 V+NP1+的+NP2 구문 패턴의 애매성 해소

최정[0] 김미영 김동일* 이종혁
포항공대 정보통신 대학원[0], 포항공대 컴퓨터 공학과
(cuizheng[0], colorful, dongil, jhlee)@postech.ac.kr

Resolving structural ambiguity of Chinese V+NP$_1$+的+NP$_2$ syntactic pattern

Cui Zheng[0] Mi-Young Kim Dongil Kim* Jong-Hyeok Lee
Dept. of Graduate School for Information Technology, POSTECH[0]
Div. of Electrical and Computer Engineering, POSTECH
and Advanced Information Technology Research Center (AITrc)

요 약

중국어 V+NP$_1$+的+NP$_2$ 형 패턴은 동사구와 명사구로 분석이 가능하여 중국어 구문분석의 결과에 중요한 영향을 미친다. 본 논문은 중국어 V+NP1+的+NP2 형 패턴의 구조적 중의성 문제를 해결하기 위한 방법을 제안한다. 제안하는 방법은 통계정보로 보완된 동사의 결합가 정보, 두 명사간의 결합도 정보 및 휴리스틱으로 구조적 애매성을 해소하고자 한다.

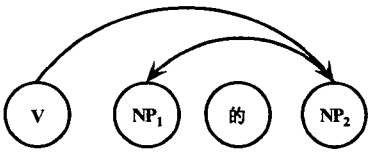
1. 서 론

명사구의 분석은 자연언어를 처리함에 있어서 아주 중요한 역할을 한다. 따라서 명사구 분석은 정보검색, 기계번역 등 광범위한 응용분야에서 필수적인 부분으로 되고 있다. 중국어 명사구 분석에서 관계절 명사구 분석의 오류가 전체 명사구 분석 실패의 30% 정도를 차지한다[2,3]. 본 연구실의 중-한 기계 번역 시스템 Total-C/K[1]에서도 관계절 명사구 분석의 문제가 존재하여 번역성능에 큰 영향을 미치고 있다.

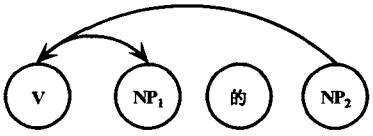
중국어 관계절 명사구는 K.F. Wong이 관계절의 유형에 따라 verb-object, subject predicate, adverbial verb, verb-complement coordinate verb등 4가지로 분류하였다. K.F. Wong의 명사구 추출 NPext[3] 시스템에서 verb-object형 관계절 명사구의 분석실패가 전체 명사구 분석 실패의 44%를 차지한다고 밝혔다. Verb-object형 관계절 명사구에서 일반형태는 V+NP$_1$+的+NP$_2$으로 그림 1과 같이 동사구와 명사구로 분석될 수 있다.

2. 기존 연구

Verb-object 패턴의 구조적 애매성을 해소하는 방법 중의 하나는 어휘정보를 이용하는 것이다. 그림 1의 a에서 명사구 NP$_1$는 NP$_2$의 수식어고 V는 NP$_2$와 관련되고 그림 1의 b 에서는 동사



(a) 동사구로 분석



(b) 명사구로 분석

<그림 1> 중국어 V+NP1+的+NP2형 명사구의 중의성

V는 명사구 NP$_1$와 관련을 가지는데 NP$_1$는 NP$_2$ 와 직접적인 관계를 가지지 않는다. 이러한 애매성을 해소하기 위하여 논문[3]에서 V와 NP1 및 V와 NP2의 어휘 결합도(lexical association)를 이용하는 방법을 제시하였다.

통상적으로, V NP1 的 NP2의 구조에서 V와 NP1의 어휘 관련도가 V와 NP2 의 어휘 관련도보다 크면 V와 NP1는 동사구로 결합하여 NP2 의 수식 성분으로 된다.

* 중국 길림성 연길시 연변과학기술대학 부교수

2. 1 어휘 결합도 (Lexical association)

어휘 결합도는 두 단어간의 의미적 관계를 수량으로 표현하는 방법이다. 어휘 결합도를 이용하여 구조적 애매성을 해소하는 방법은 이미 영어 구문분석에 많은 응용이 이루어져 있다. 어휘 결합도의 값을 구하는데 두 단어가 학습 말뭉치(trainging corpus)에서 같은 단어 혹은 의미 분류(semantic class)로 공기하는 빈도수를 이용하였다. 단어 레벨(level)에서 공기정보를 얻으려면 대 규모의 학습 말뭉치가 필요한데 학습 말뭉치가 적으면 단어를 의미 분류한 시소러스를 사용하여 의미 분류에서의 공기정보를 얻을 수 있다.

$$LA(v,n) = P(n \mid v)MI(v,n) \quad (1)$$

$$= P(n \mid v)\log_x \frac{P(v,n)}{P(v)P(n)} \quad (2)$$

V+NP_1+的+ NP_2의 구조적 애매성을 해소하는데 V와 NP_1의 공기정보가 필요하다. 수식 1와 2는 단어 v와 n의 어휘 결합도를 계산하는 공식이다. MI는 단어 v와 n의 상호 정보량으로 동사와 명사간의 통계적 결합도를 나타낸다. 실제 사용할 때는 $P(n \mid v)$를 곱하여 동사와 더 자주 나타나는 명사를 선호하게 한다. (1),(2) 식은 단어간에 결합도를 나타내는데 데이터 부족문제가 생길 수 있으므로 실제 사용할 때는 의미 분류로 결합도를 계산한다.

$$LA(Cv,Cn) = P(Cn \mid Cv)MI(Cv,Cn) \quad (3)$$

$$= P(Cn \mid Cv)\log_x \frac{P(Cv,Cn)}{P(Cv)P(Cn)} \quad (4)$$

수식 (3), (4)에서 Cv와 Cn는 각각 동사와 명사의 의미 분류를 표시한다. 여기에서 <<同义词词林>>[5]의 개념 분류 체계를 사용하고 각 부분의 확률 값은 MLE방법으로 계산한다.

2. 2 알고리즘

중국어 V+ NP_1+ 的+ NP_2패턴의 애매성을 해소하는데 통계데이터 추출 및 애매성 해소 두 개의 단계로 진행된다. 통계데이터 추출 단계에서 2천만자의 1991년 <<인민일보>> 코퍼스를 사용하여 동사-명사 쌍을 추출하고 수작업으로 verb-object패턴을 추출한다. 그리고 추출된 verb-object들에 대하여 의미 태깅을 하고 단어간과 개념간의 공기 정보를 얻는다.

If $\dfrac{LA(Cv,Cn_1) - LA(Cv,Cn_2)}{LA(Cv,Cn_1)} > \varepsilon$

Then NP_1는 동사 v의 목적어로 전체 구조는 명사구가 된다

Elseif $\dfrac{LA(Cv,Cn_2) - LA(Cv,Cn_1)}{LA(Cv,Cn_2)} > \varepsilon$

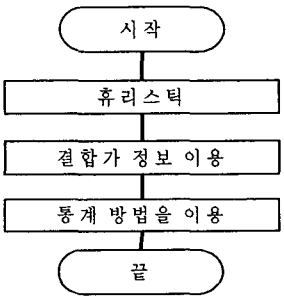
Then NP_2는 동사 v의 목적어로 전체 구조는 동사가 된다

<그림 2> 중국어 V+NP1+的+NP2형 명사구 중의성 해소 알고리즘

다음 단계는 앞 단계에서 얻은 단어간 및 개념간 공기정보로 그림 2와 같이 동사와 NP_1, NP_2사이의 결합도에 따라 구조적 애매성을 해소한다.

3. 애매성 해서 모델

기존의 방법에서 2천만자의 코퍼스에서 Verb-object정보를 얻는 것이 인력이 많이 필요 되는 작업으로 현재 상황에서 대규모 작업이 어렵다. 그러나, 현재 북경대학의 의미정보사전[6]에 결합가 정보가 구축되어 있어 대용으로 쓸 수 있다. 결합가 정보를 사용하면 높은 적용범위가 있어 애매성 해소에 많은 도움이 된다. 따라서 트리 뱅크(tree bank)에서 얻은 동사와 명사간의 정보와 두 명사간 결합도 정보를 이용하여 결합가 정보가 부족한 부분을 보완한다. 이 과정에 휴리스틱도 적용하여 정확도를 더 높이는 시도를 하였다. 전체 흐름을 보면 그림 3과 같다.



<그림 3> Total-CK RCLDis 프로그램의 흐름

3.1 결합가 정보 및 휴리스틱 이용방법

결합가 정보의 높은 적용범위를 충분히 이용하기 위하여 결합가 정보를 통계 방법이전에 적용한다. V와 NP_1만 결합 가능하면 명사구로, V와 NP_2만 결합 가능하면 동사구로 결정된다.

북경대학 의미정보사전[6]에는 단어들의 결합가 정보가 기술되어 있다. 결합가 정보는 단어와 단어 사이 선택제약 조건을 가리킨다. 여기에서는 동사와 명사간의 선택제약 정보를 이용한다.

V와 NP_1 및 NP_2 가 동시에 결합 가능하면 다음 단계인 통계 방법을 이용하는 단계로 넘어 간다. 결합가 정보를 이용한 방법으로 V+ NP_1+ 的+ NP_2형 패턴의 애매성을 해소한 결과의 오류를 관찰한 결과 아래와 같은 휴리스틱(heuristic)을 발견하였다. 즉 NP_2이 동사성 명사인 경우의 오류가 27%차지 하였다. 예를 들면, 维护(v) 秩序(n1) 的(u) 稳定(n2)[1]. 휴리스틱1은 NP2이 동사

[1]원문의 뜻은 "질서(n)의 온정(vn)을 유지(v)하다" 이다.

성 명사일 경우 V+ NP_1+的+NP_2 패턴을 동사구로 분석한다.

그리고 V+ NP_1+的+NP_2 패턴을 분석한 결과 NP_2이 동사의 주어로 사용되는 경우에 명사구로 분석 될수 있는 것을 발견하였다. 예를 들면, 符合(v) 实际(n1) 的(u) 方案(n2)[2]에서 n2(방안)은 v(부합되다)주어로 쓰이므로 명사구로 분석된다.

3.2 통계정보를 이용한 방법

현재 북경대학 사전에 기술된 결합가 정보는 동사는 10,747개, 명사는 27,792개로 V+ NP_1+的+NP_2형 명사구의 애매성을 해소하는데 부족하다. 따라서 동사V가 NP_1,NP_2 와 동시에 결합 가능하거나 동시에 결합불가능할 때에 또 다른 정보를 사용하여 애매성을 해소할 필요성이 있다.

본 논문에서는 K.F. Wong[4]가 제시한 통계적 방법을 기초로 하고 NP_1이 NP_2의 수식어로 나타나는 확률과 결합하는 방법을 제안 한다. n_1, n_2 는 NP_1와 NP_2 의 의미 중심인 헤드를 표시한다.

$$\frac{LA(Cv,Cn_1)-LA(Cv,Cn_2)}{LA(Cv,Cn_1)}-k_1\cdot LA(Cn_1,Cn_2)>\varepsilon \quad (5)$$

$$\frac{LA(Cv,Cn_2)-LA(Cv,Cn_1)}{LA(Cv,Cn_2)}+k_2\cdot LA(Cn_1,Cn_2)>\varepsilon \quad (6)$$

수식 5,6의 $LA(Cn_1,Cn_2)$는 Cn_1,Cn_2이 트레닝 코퍼스에서 수식 관계로 나타나는 강도를 표시하고 k는 두 명사간의 수식 가능성에 대한 가중치이다.

4. 실험 및 평가

실험을 시작하기 전에 단어 결합도를 계산하기 위하여 통계데이터를 준비해야 한다. 중국어 **Penn Treebank**에서 모든 verb-object로 나타나는 동사와 명사 정보를 추출하는 것으로 verb-object패턴 추출단계의 수작업을 대체 한다. 그리고 추출된 패턴들에 대해 의미 태깅을 한다. 북경대학의 44개의 명사 의미분류 체계는 구분도가 떨어 지므로 4,000여 개의 부류로 구성된 동북대학의 분류체계를 사용하였다. 동북대학의 의미 분류는 7개 레벨로 구성되었는데 여기에서 상위 4 레벨을 사용한다.

실험에서 테스트한 데이터는 **Penn Treebank**에서 추출한 960개의 V+ N_1+的+N_2패턴이다. 각 패턴에 동사구나 명사구로 표시하고 전반부에서 제시한 방법으로 각 패턴에 대해 애매성 해소를 하였다.

다음과 같이 세가지 실험을 통하여 우리의 방법의 유효성을 검증하였다. 첫 번째 실험은 결합가 정보만 사용하였고, 두 번째 실험에서 결합가 정보와 휴리스틱을, 마지막에는 결합가 정보, 휴리스틱 및 통계정보를 사용한 방법으로 실험을 하였다. 실험 결과는 도표 1과 같다.

[2]원문의 뜻은 "실제상황(n1)에 부합(v)되는 방법(n2)" 이다.

도표에서 결합가 정보를 이용한 방법의 재현율이 현저하게 낮음을 발견 할 수 있다. 이는 현재 사용하고 있는 북경대학 의미정보사전의 엔트리수가 부족하기 때문이다. 이는 향후 보완 작업으로 개선될 여지가 있다. 실험에서 看望(v) 孩子(n) 的(u) 家长(n) 처럼 의미 분류로 해결 불가능한 예들도 발견하였다. 이는 부가적인 구조적 정보와 주변 문맥정보로 해결할 수 있다.

	Correct	Error	정확도
Model 1	670	260	72%
Model 2	780	189	80.5%
Model 3	789	161	83.0%

<도표 1> 세 가지 실험의 성능 비교

5. 결론 및 향후 작업

본 논문에서 개선된 중국어 V+ NP_1+的+NP_1 패턴의 구조적 애매성을 해소하는 휴리스틱과 통계정보를 이용한 방법론을 제시 하였다. 실험에서 보여주는 바와 같이 제안한 방법이 구조적 애매성을 해소하는데 효과적이었으나 북경대학 의미 정보사전에 기술된 정보의 빈약으로 결과가 기대이하 이다. 향후 작업에서 북경대학 의미 정보사전의 정보를 보완하고 더 넓은 문맥에서의 구조적 애매성 해소 방법에 대한 연구가 필요하며 다양한 관계절 분석의 연구가 진행될 것이다.

6. 감사의 글

본 연구는 첨단정보기술 연구센터를 통하여 과학재단의 지원을 받았습니다.

7. 참고문헌

[1] Dong-Il Kim, Zheng-Cui, Jinji-Li, Jong-Hyeok Lee, "Resolving Structural Transfer Ambiguity in Chinese-to Korean Machine Translation," in ICCLC '2002

[2] Guo Zhili, Yuan Chunfa and Huang Changning, "A statistical Approach to study the structure and boundary of 的-phrases," in Proceedings of 1996 International Conference on Chinese Computing (ICCC '96), Singapore, 1996

[3] WenJie.Li, K.F. Wong, B.T.Low, A.S.Y. Tse, and V.Y. Lum, "Chinese noun phrase extraction based on the statistical distribution information," in Proceedings of 1996 International Conference on Chinese Computing (ICCC '96), Singapore, 1996,

[4] Kam-fai, Wong, W.J.Li, V.Y.Lum, and B.T.Low, "Resolving Relative Clause Structural Ambiguity in Chinese Noun Phrase Extraction," Computer Processing of Oriental Languages, Vol.14, No.2, 1997

[5] 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

[6] 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 (ICCC '98), Beijing, 1998