확률모형과 수식정보를 이용한 와/과 병렬명사구 범위결정

(Range Detection of Wa/Kwa Parallel Noun Phrase using a Probabilistic Model and Modification Information)

최 용 석 † 신 지 애 †† 최 기 선 †††
(Yong-Seok Choi) (Ji-Ae Shin) (Key-Sun Choi)

요 약 한국어 구문 분석의 초기 단계로서 병렬구조의 해석은 파싱의 효율을 높일 수 있다. 본 논문은 병렬구조 해석을 위한 비지도식 언어에 독립적인 확률 모델을 제안한다. 이 모델은 병렬구조의 대칭성과 상호교환성에 근거한다. 대칭성은 같은 구조가 반복된다는 것이고, 교환성은 좌우 구성요소를 교환해도 같은 의미를 지닌다는 것이다. 병렬구조는 일반적으로 대칭성을 따르지만, 수식어의 성질에 따라서 한쪽만을 수식하는 비대칭적인 구조가 출현하기도 한다. 비대칭 병렬구조 해석을 위해서 추가적으로 수식관계 통계정보를 사용한다. 제안된 모델을 본 논문에서는 "와/과" 조사로 이루어진 한국어의 명사구 병렬구조를 해석하는데 사용되는 것[1]을 중점으로 보여준다. 지도적 방식에 의한 모델을 포함한 다른 모델들에 비해 효율적임을 실험적으로 보여준다.

키워드 : 한국어 구문분석, 자연어 처리, 언어 독립적 모델, 확률모델, 병렬구조

Abstract *Recognition of parallel structure at early stage of sentence parsing can reduce the complexity of parsing. In this paper, we propose an unsupervised language-independent probabilistic model for recongition of parallel noun structures. The proposed model is based on the idea of swapping constituents, which replies the properties of symmetry (two or more identical constituents are repeated) and of reversibility (the order of constituents is inter-changeable) in parellel structures. The non-symmetric patterns that cannot be captured by the general symmety rule are resolved additionally by the modifier information. In particular this paper shows how the proposed model is applied to recognize Korean parallel noun phrases connected by "wa/kwa" particle. Our model is compared with other models including supervised models and performs better on recongition of parallel noun phrases.*

Key words : *Korean parsing, Natural Language Processing, Unsupervised Probabilistic Model, Language-Independent Model, Parallel Structure*

1. 서 론

의존구조 구문분석은 그림 1[1]과 같은 방식으로 이루어질 수 있다. 명사구 분석기 부분에서 기본구를 인식한 후에 기본구를 바탕으로 구문분석을 시도하여 구문분석의 복잡도를 줄인다. 명사 기본구를 인식한 후에 기본구간의 병렬적인 관계에 대한 분석을 수행할 수 있다.

기본구(base phrase 혹은 chunk)의 개념은 1991년 Abney[2]에 의해 처음 도입되었다. Abney는 문장에서 끊어 읽는 운율적 휴지의 단위와 구문구조에 있어서의 단위가 상응한다고 보고 "하나의 중심어를 포함한 겹쳐지지 않는(non-overlapping) 단어들의 묶음"을 기본구로 정의하였다. 이 논문에서는 병렬구를 한 덩이로서 구문요소화하는 논제를 연구한다.

† 학생회원 : 한국과학기술원 전산학과
angelove@kriss.re.kr
†† 정 회 원 : 정보통신대학교 공학부 교수
jiae@icu.ac.kr
††† 종신회원 : 한국과학기술원 전산학과 학과장
kschoi@world.kaist.ac.kr
논문접수 : 2007년 11월 14일
심사완료 : 2008년 1월 19일

정보과학회논문지 : 소프트웨어 및 응용 제35권 제2호(2008.2)

병렬구조는 같은 구조가 두 개 이상 반복되어 나타나는 대칭적 구조에서 비롯된다. 즉, "a는 X와 Y가 V이다"와 같은 구조에서 X, Y는 같은 수준의 지시 대상을 가진다. 예를 들어, "수돗물에는 염소와 석회 성분이 들어있다"를 보자. '염소'와 '석회 성분'이 각기 X, Y라 하면, 대칭성에 문제가 생긴다. '염소성분'과 '석회성분'이어야 대칭성(symmetry)을 만족한다. 따라서 병렬구조는 '염소와 석회'까지라고 할 수 있다. 또한 병렬구조는 교환성(reversibility)를 만족해야 한다[3]. 즉, 'X와 Y' 구조에서 'Y와 X' 구조로 바뀌어도 문장에 영향을 주지 않는 구조이다. 앞의 예에서 두 성분을 바꾸어서 '석회 성분과 염소'하면 원 문장의 의미에 변화가 있으므로 교환성을 만족시키기 위해서는 병렬구조의 범위를 '염소와 석회'라고 보는 것이 옳다.

본 논문에서는 대칭성과 교환성을 이용하는 언어 비의존적인 확률 모델을 제안한다. 확률 모델을 적용한 이후에 수식 구조를 이용해서 병렬 명사구 범위를 확정한다. 수식구조를 이용하는 모델에서는 관용적 어휘 정보와 시소러스 정보를 이용한다.

본 논문의 구성은 다음과 같다. 먼저, 2장에서 본 논문에서 다루고자 하는 문제의 범위에 대하여 살펴본다. 3장에서는 기존연구를 알아보고, 4장에서는 병렬명사구 범위 탐지를 위한 모델을 설계하여 제안하고, 5장에서는 수식구조를 이용하여 병렬구조 범위를 확정하는 방법을 제안한다. 6장에서는 실험에 대해 다루고, 마지막으로 8장에서 결론을 내리고, 향후 연구 과제를 제시한다.

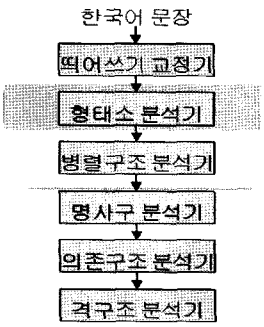


그림 1 구문분석 과정

2. 병렬구와 구문적 병렬범위 추정

2.1 병렬구에 관련된 정의들

이항병렬구조라 함은 'X op Y'와 같이 병렬어휘 op의 양 옆에 비슷한 대칭적 구조를 가지는 X와 Y로 구성된다. 일반적으로 N항의 병렬구조가 가능한데, 이 때 'X1 op1 X2 op2 … Xn (opn) '에서 각 Xi는 비슷한 구조를 갖고, 각 opj는 쉼표나 대등접속사 등으로 표시된다. 단, 마지막 연산자 opn은 그 외에 여러 개의 비슷한 병렬 대상이 있다는 뜻에서 '등'으로 나타난다.

병렬구조는 '비슷한 대칭적 병렬구조'와 '완전한 대칭적 병렬구조'로 구분될 수 있다. 예를 들면, '훌륭한 부모와 성실한 자녀'는 완전한 대칭적 병렬구조를 이룬다. 구문적으로 '훌륭한'과 '성실한'은 각각 형용사이며, '부모'와 '자녀'는 보통명사이다. 또 의미적으로도 같은 부류임을 알 수가 있다. '비슷한 대칭적 구조'라 함은 구문과 의미적으로 완전한 대칭을 이루지 않았으나, X, Y의 문맥에 따른 제약으로 X, Y가 의도적인 대칭화가 이루어 질 수 있는 경우를 의미한다. 이 의도적인 대칭화의 과정에서 병렬구조의 경계 인식은 어렵다. 예를 들어,

수돗물 안에는 염소와 석회 성분이 들어있다.(2)

여기서 op는 '와'이다. X = '염소'인 반면, Y = '석회'인가 혹은 '석회 성분'인지 혼동이 된다. 구문적 대칭관계에서 보면 분명히 Y = '석회'가 되어야 한다. 그러나 다음의 예를 보자.

강원도에는 염소와 석회 동굴이 많다.(2')

X = '염소'는 틀림이 없지만, X = '석회 동굴'이 더 의미적으로 옳다고 본다. '염소 동굴' (즉, 염소가 많이 사는 동굴)도 있을 수 있겠지만 의미적이 아닌 지극히 상황의존적일 수 밖에 없다.

위의 예 (2)에서 보는 바와 같이 완벽한 명사구는 '염소와 석회 성분'이 되어야 한다. 이와 같이 병렬구조를 포함하는 명사구를 이 논문에서는 '병렬내포명사구'라고 하고, 병렬내포명사구의 내부에서 완전한 대칭적 구조를 이루는 요소를 '병렬핵' (혹은 '병렬덩이')라고 한다. 예를 들면, '염소와 석회 성분'은 병렬내포명사구이고 '염소와 석회'는 병렬핵이 된다. 병렬내포명사구는 예 (2)에서 "수돗물 안에는 Z가 들어 있다"의 Z에 대입할 수가 있는 것과 같이, 완전한 구문요소이다. 병렬내포명사구는 대칭적 병렬구조인 병렬핵을 포함한다. (2)과 (2')에서 병렬핵 '염소와 석회'의 예에서 '성분' 혹은 '동굴'이 어떻게 수식되어야 하는지는 비슷한 대칭적 구조를 인식하는 문제가 된다.

2.2 병렬내포명사구 분류

병렬내포 명사구의 구조는 그 안의 대칭적 병렬구조 분석의 복잡도에 따라 4단계 정도로 구분할 수 있다. 그 복잡도 단계는 아래 표 1과 같다.

표 1 병렬내포 명사구의 복잡도

복잡도	완전한 대칭적 구조	비슷한 대칭적 구조
1	(X op Y)	
2	(X op Y) Z	(X op Y Z)
3	MOD (X op Y)	(MOD X op Y)
4	MOD (X op Y) Z (MOD X op Y Z)	MOD (X op Y Z) (MOD X op Y) Z

제1복잡도는 단순하게 2개의 구성 요소를 묶어 주면 되는 간단한 경우이다. "[사회정학과 사회동학]은 각각 사회질서와 사회진보를 이룩하는데 그 목적이 있었다." 가 이에 해당한다.

제2복잡도는 Y 뒷 부분에 수식이 가능한 요소 Z가 왔을 경우이다. 앞에서 논한 '[염소와 석회] 성분'과 같이 'X op Y Z'에서 op의 범위를 결정하기 위해서 X와 Z의 의미를 알고 그 유사도를 측정할 수 있다면 유용하게 사용할 수 있다.

제3복잡도는 병렬 명사구 앞에 수식하는 어휘가 출현할 경우이다. '커다란 슬픔과 고통'를 'MOD X op Y'로 보자. 수식 어휘 MOD = '커다란'이 X = '슬픔' 만을 수식한다면 비슷한 대칭적 구조를 가진 병렬 명사구 '(커다란 슬픔)과 고통'이 된다. 이것을 판단하기 위해서는 수식하는 어휘와 X, Y의 의미를 알고 있어야 한다. 따라서, 병렬구조의 의존구조 인식단계 중 제2단계의 구문적 완전대칭구조는 '커다란 (슬픔과 고통)'를 구문적 완전대칭구조로서 출력을 하며, 제3단계의 의미적 완전대칭구조는 '(커다란 슬픔)과 (커다란 고통)'이 된다. 그러나, 제2단계에서 의미해석 결과 '커다란 고통'이 성립을 안하면 제3단계의 결과는 '(커다란 슬픔)과 (고통)'이 될 것이다.

제4복잡도는 제2, 3복잡도가 복합적으로 나타난 경우로, 복합적인 판단과 우선순위를 결정할 수 있는 과정이 정의되어 있어야 한다. 제4복잡도의 경우 병렬내포 명사구의 구성 요소 수가 늘어남에 따라 문제가 더 복잡해짐을 알 수 있다. 실제 병렬내포 명사구에서는 구성요소를 여러 개 묶어서 X나 Y같이 하나의 단위로 표현할 수도 있다. [커다란 슬품과 경제적인 고통] 을 (MOD X op Y) 구조로 분류하여, Y = '경제적인 고통'으로 둔다.

3. 관련 연구

3.1 패턴 분석

박준식[4]은 병렬구 분석을 위해 품사열의 패턴을 분석했다. 형태/품사적인 특징을 이용하여 패턴을 만들어 병렬구의 범위를 결정했다. 패턴은 두 가지 계층으로 교차 적용하여 자료 부족 문제를 회피하려 했다. 이런 패턴 모델은 모든 병렬형태의 문제에 적용할 수 있으나, 같은 패턴이라면 항상 같은 결과를 내준다. 예를 들어 '[문학과 자연 과학]'과 '[자유와 평등] 관계'처럼, 같은 패턴이면서도 다른 분석결과를 가지는 문장에 대해서는 약점을 가질 수밖에 없다.

3.2 완전한 대칭성 분석[5]

구로하시[5]는 병렬명사구나 접속절 등의 접속구조를 인식하는 모델을 제안했다. 접속사 좌우에 있는 단어열 사이의 유사성을 계산하여, 비슷한 형태를 가진 단어열을 접속구조를 이루는 구성성분으로 결정한다.

예를 들어 '커다란 슬픔과 경제적 고통'이라는 병렬구에서 접속사 '과' 좌우에 출현하는 문절 간의 유사도를 측정한다. 두 문절 간의 유사도는 실험을 통해 얻은 가중치값 부여를 통해 구해진다. 유사도를 구하는 방법은 아래와 같다. (일본어에서 문절은 독립적인 기능을 갖는 자립어와 기능어로 이루어진 가장 작은 유의미한 문자열을 말한다.)

구로하시[5]는 앞서 설명한 방법과 같이 완전한 대칭적 구조에 기반한 분석을 했다. 대칭성에 대해서 실험결과 얻어진 규칙을 만들어 성능을 높일 수 있는 방안을 제시했다. 이 방법의 경우, 비슷한 대칭적 구조를 가지는 병렬 구문에 대해서도 점수를 통한 해결방식을 제안해 놓았으나, 완전한 대칭적 구조에 기반하였기 때문에 비슷한 대칭적 구조에 약점을 가진다. 제2절에서 정의한 제2복잡도 수준의 문제는 대칭성만을 이용해 어느 정도 해결이 가능하나, 제3복잡도는 수식하는 어휘의 의미를 활용하지 않기 때문에 문제 해결이 어려운 약점을 가진다. 즉, 병렬구 내부의 일부 구성요소 생략에 의해 완전한 대칭적 구조가 아닐 때, 약점을 가지게 된다. 예를 들어 '[커다란 슬픔과 경제적인 고통]'으로 분석은 잘 하지만 '커다란 [슬픔과 경제적인 고통]'와 같이 비슷한 대칭적 구조로 분석해야 할 경우에 완전한 대칭적 구조로 분석하는 것을 선호한다.

예는 표 2와 같은 행렬로 표현될 수 있다. '[커다란 슬픔과 경제적인 고통]'은 합이 20점이 되고, '커다란 [슬픔과 경제적인 고통]'과 같은 분석은 합이 12점이 된다. 따라서 전자의 분석을 선호하게 되고, 이런 점수체계에서는 완전한 대칭적 구조로 분석하는 것을 선호하게 된다.

표 2 대칭성 분석 점수계산 방법의 예

커다란	10	
슬픔	2	10
	경제적	고통

또한, 규칙들과 점수 부여 방법이 실험적으로 정해졌기 때문에 환경이 바뀌거나, 언어가 바뀔 경우 성능을 보장할 수 없다. 위의 표에서와 같이 언어가 달라지면 점수를 10점을 줄것인가 아니면 다른 점수를 부여할 것인가에 대한 판단도 달라질 수 밖에 없다. '커다란 [슬픔과 경제적인 고통]'과 같은 분석에서는 '슬픔' 1어절에 '경제적인 고통' 2어절이 대응되고 있는데, 여기에 얼마나 큰 벌점을 부여하냐에 따라 경로탐색의 결과가 달라지게 된다.

3.3 비교를 위한 지도식 학습 모델

기존의 연구결과는 병렬기호 op를 중심으로 하여, 좌우구성요소의 유사성을 계산하여 병렬의 범위를 결정짓는다. 구로하시[5]의 경우, Viterbi 산법에 의한 최대값 경로를 구하고, 박준식[4]의 방법은 품사열의 유사성을 이용한 것이다.

결정나무(decision tree) 모델은 규칙을 자동으로 구하는 모델이다. 결정나무(decision tree) 방법은 ID3, C4.5[6] 등의 알고리즘으로 귀납 추론한 학습의 결과가 결정나무(decision tree)로 표현되는 방법으로 여러 응용에 성공적으로 적용됐다.

확률기반 모델로 최대 엔트로피 모델을 사용하여 다른 모델들과 비교해 본다. 최대 엔트로피 모델은 서로 모양이 다른 형태의 정보를 합치는데 유용한 확률 모델이다. 패턴은 속성과 그에 대한 값으로 표현된다. 각 패턴들이 여러 개의 특징들로 표현되고, 이러한 특징들은 분류를 위한 제약조건으로 사용한다.

지지벡터 기계(Support Vector Machine) 모델[7]은 이진분류 문제를 해결하기 위한 학습 알고리즘으로 Vladimir Vapnik에 의해 1995년에 제안되었다. 지지벡터기계는 학습데이타를 이진 분류할 수 있는 최적의 초평면(hyper plane)을 찾는다.

지도식 학습 방법에는 모두 어절 표지 정보와 어절의 위치 정보를 학습을 위한 자질값으로 사용한다. 좌우측 어절 표지에 따라서 좌측의 병렬구 범위를 결정하고, 결정된 좌측의 범위값과 어절 표지를 특징값으로 사용해서 우측 병렬구 범위를 결정한다.

4. 제안 모델

4.1 병렬명사구의 대칭성

병렬내포 명사구는 대칭성[8]이 있어서 서로 대응하는 부분이 있다. 그림 2[9]과 같이 문장을 "와/과"조사를 중심으로 좌우로 나눈 후, 병렬구 내부의 서로 대응하는 요소끼리 연결할 수 있다.

구로하시[5]도 이런 대칭성을 이용해서 대응하는 부분간의 점수를 정해서 유사도를 이용했다. 본 논문에서는 자동으로 확률을 학습하는 방법을 제시한다. 이를 통해 경험에 의한 점수부여 방법을 배제하고, 최대확률 대응을 구하려 한다.

병렬구조로 이루어진 문장에서는 문장을 좌우로 분리한 다음에 좌측의 어느 부분이 우측과 대응하는 지를 살펴보고 병렬구조의 범위를 정하는 모델이다. 모델의 수식은 아래와 같다. 확률값이 최대가 되는 쪽으로 문장 단위들을 서로 대응시키는 것이다. 실제 언어상황에서 일어나는 확률을 정확히 쓴다면, 결과가 더욱 정확해 질 것이다. 실제로 정확한 확률을 쓸 수 없기 때문에 수식에서 확률값을 무엇으로 정의하느냐에 따라 각각의 모델들의 성질이 결정된다. 수식에서 l_1^J는 병렬기호 op 왼쪽의 어절들로 1부터 J까지의 어절을 가진다. r_1^I는 병렬기호 op 오른쪽의 어절들로 1부터 I까지의 어절들을 가진다. a_1^J는 왼쪽의 어절들의 오른쪽에 대한 대응 정보를 의미한다.

$$\hat{a}_1^J = \arg\max_{a_1^J} P(l_1^J, a_1^J \mid r_1^I)$$

결과로 나온 대응을 보고 병렬구조의 범위를 결정한다. 우측의 시작 단어와 대응하는 좌측의 단어가 병렬구문의 시작이고, 좌측의 마지막 단어와 대응하는 우측의 단어가 병렬구문의 끝이 된다. 그림 2에서 우측의 시작 단어는 '인쇄술의'이고 이 단어와 대응된 단위는 '종이의'이다. 병렬구의 시작은 '종이의'로 결정된다. 같은 방법으로 좌측의 마지막 단어 '발명과' 대응된 단어는 '발달로'이고 '발달로'가 병렬구의 마지막 범위로 결정된다.

접속조사 왼쪽에 출현한 단위와 오른쪽에 출현한 단위의 확률을 이용하여 첫번째 모델을 정의한다. 확률값을 사용하기 위해서는 사용할 단위를 결정해야 한다. 본 논문에서는 실험을 통해 가장 우수한 결과를 나타내는 어절단위를 사용한다. 어절 단위는 어절표지로 표현되며, 어절표지는 품사정보를 바탕으로 결정된다.

제안한 모델은 가장 확률적으로 대응될 가능성이 높은 단위를 연결하게 된다. 이를 위해 처음으로 필요한 확률은 왼쪽에 출현한 단위와 오른쪽에 출현한 단위간의 대응확률이다. 이는 가장 기본적인 확률로 그 확률식은 아래와 같다.

$$p(l \mid r)$$

두 번째로 사용할 확률에는 위치정보도 사용한다. 명사가 명사와 대응될 확률이 높다고 하더라도, 문장의 첫 번째 위치에 온 명사와 병렬구를 나타내는 조사 '와/과' 옆에 나타난 명사의 대응 확률은 다를 수 밖에 없다. 위의 예 문장에서 접속사 좌측에 나타나는 '시대를'과 '종

그 시대+를 지나+면 종이+의 발명+과

인쇄술+의 발달+로 책+의 문화+는 놀랍+ㄴ 발전+을 거듭+하+ㄴ다+.

그림 2 문장에서 병렬명사구의 대응

이의'가 같은 명사로 분류될 수 있으나, 이 두 명사가 우측에 나오는 '인쇄술의'와 대응할 확률은 달라야 한다. 이러한 위치 정보를 사용한 확률식은 아래와 같다.

$$p(a_j \mid j, I, J)$$

이 확률식의 의미는 왼쪽 문장단위의 길이가 J이고 오른쪽 문장 단위의 길이가 I인 상황에서 현재 왼쪽의 j번째 위치가 오른쪽 aj 번째 위치와 대응될 확률이다. 여기서 i를 사용하지 않고 aj를 기호로 사용한 것은 i의 위치가 j에 의존하기 때문이다. 많은 문장으로 학습하면 '현재 위치와 멀리 떨어진 위치와 대응할 확률은 작은 값이다.'와 같은 자료들을 얻어낼 수 있다.

기본확률과 위치확률을 모두 곱한 것이 문장 단위 정렬의 기본확률이 된다. 이 정렬확률이 높은 것을 선택하면 높은 확률로 대응된 결과를 얻을 수 있다.

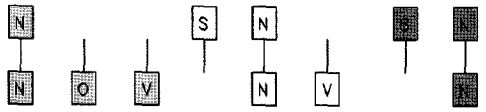
병렬구조로 이루어진 문장에서는 문장을 좌우로 분리한 다음에 좌측의 어느 부분이 우측과 대응하는 지를 살펴보고 병렬구조의 범위를 정할 수 있다. 우측의 시작 단어와 대응하는 좌측의 단어가 병렬구문의 시작이고, 좌측의 마지막 단어와 대응하는 우측의 단어를 병렬구문의 끝으로 가정하였다.

4.2 교환정렬 모델 제안

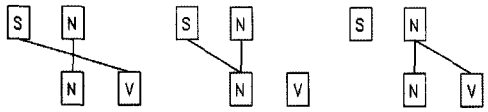
병렬구에서는 내부 요소간의 교환이 가능하다. 내부 요소의 위치를 바꿔어도 문장이 성립한다. (X op Y)의 형태의 병렬구를 (Y op X)로 바꿔도 전체 문장에는 큰 변화가 없다. 예를 들어 [커다란 슬픔과 경제적인 고통]은 [경제적인 고통과 커다란 슬픔]으로 바꿔도 전체문장에 영향을 끼치지 않는다. 제안하는 모델은 이러한 교환성(reversibility)을 사용한 교환정렬 모델이다.

모델을 사용하면 많은 수의 빈 정렬이 생길 것으로 예상하였다. 그림 3의 (a)와 같이 병렬구가 문장내에서 여러 역할을 수행할 때, 빈정렬이 많을 것으로 보았다. 가령 병렬구가 목적어이면 병렬구는 병렬구끼리 대응되고 주어는 빈정렬, 술어도 빈정렬이 될 것으로 예상하였다. 실제로 정렬을 해 본 결과는 빈정렬 이 많이 나타나지 않았다.

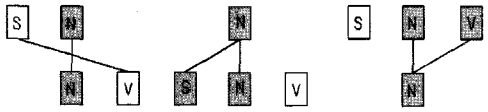
그림 3의 (b)와 같이 좌측에 주어가 나타나면 우측에 술어와 대응되는 경우가 많았다. 하지만, 좌측에 나타난 주어가 우측의 목적어와 대응되는 경우도 있었다. 이는 실제로 주어가 좌측에서 출현하면 우측에 목적어가 출현할 확률이 어느 정도 있기 때문이다. 대응확률의 의미는 이렇게 좌측과 우측의 대응할 확률을 자동으로 학습하는 것으로, 사람이 체계화 시킨 문법과 다른 의미를 가진다. 이러한 정렬은 범위를 결정하는 데에 문제를 발생시킨다. 가령 그림 2에서 '발달'이 '시대'와 대응되면, 범위를 결정하는데 오류를 일으키기 때문에 '대응오류' 문제라고 칭하겠다.



(a) 주어, 목적어, 보어 병렬구



(b) 목적어 병렬구의 정렬예



(c) 병렬구의 교환

그림 3 교환 정렬의 예

'대응오류' 문제를 해결하기 위해서 병렬구 요소의 교환을 제안한다. 그림 3의 (c)와 같이 병렬구의 범위로 나온 부분을 교환하는 것이다. 교환을 했을 때, 올바른 병렬구의 범위였다면 문장의 정렬 확률은 높게 나타날 것이고, 올바른 정렬이 아니면 교환된 문장의 정렬확률은 낮게 나타날 것이기 때문이다. 이 성질을 이용하여 제안하는 교환모델은 기존 정렬확률에 교환했을 때의 확률을 더한 식으로 아래와 같다. 수식에서 l_1^J는 병렬기호 op 왼쪽의 어절들로 1부터 J까지의 어절을 가진다. r_1^I는 병렬기호 op 오른쪽의 어절들로 1부터 I까지의 어절들을 가진다. a_1^J는 왼쪽의 어절들의 오른쪽에 대한 대응 정보를 의미한다.

$$\hat{a}_1^J = \arg\max_{a_1^J}[P(l_1^J, a_1^J \mid r_1^I) + \max_{b_1^I} P(l_1^{s-1} r_1^e, a_a^{s-1} b_1^e \mid l_s^J r_{t+1}^I)]$$

교환 정렬의 실제 예를 보면 그림 4와 같다. 그림 4의 (a)에서 2개의 정렬확률을 볼 수 있다. 우리가 원하는 정렬 모습은 왼쪽의 정렬 모습이지만, 약간의 확률차이로 인해서 기존의 정렬모델은 오른쪽의 정렬을 정답으로 선택하게 된다. 여기에 병렬 명사구 요소 교환을 적용한 것이 그림 4의 (b)의 모습이다. 병렬 명사구 범위로 선택된 부분을 교환한 후에 다시 정렬확률을 계산한다. 기존의 서로 다른 언어간의 정렬에서는 교환을 생각할 수 없었지만, 병렬구의 경우 좌와 우가 같은 언어로 이루어져 있기 때문에 교환이 가능하다. 교환 후에 좌측은 병렬구를 교환한 형태이기 때문에 정렬확률값은 변함이 없다. 하지만, 우측은 최소격 부분이 병렬명사구 중간에 들어간 꼴이 되므로 정렬확률값은 많이 떨어지게 된다. 그림 4의 (c)는 초기 정렬확률과 교환 후 정렬

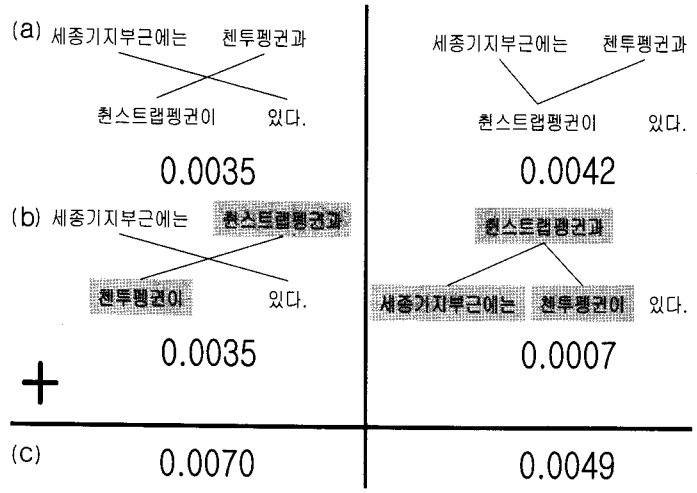


그림 4 교환 정렬의 실제 예

확률을 더한 부분으로 이를 이용해서 교환모델은 좌측의 정렬을 결과로 내놓게 된다. 초기 정렬확률이 높았다 해도, 교환 후 확률이 낮으면 결과로 선택될 수 없다.

5. 수식관계정보 이용

5.1 수식관계로 변하는 범위

품사표지의 구조는 같은데 구문분석의 결과가 달라지는 경우에 품사표지만으로 규칙을 만들려 하면 애매성이 발생한다. 이런 애매성을 해소하기 위해서는 어휘의 출현빈도라든지, 격틀 등을 사용해서 해결해야 한다.

병렬구에서는 수식구조의 영향으로 같은 표지인데 다른 구조인 결과가 나타난다. 수식 구조가 있는 경우 같은 수식구조라도 다른 형태의 분석결과가 나올 수 있다. 그림 5와 같이 병렬구 내부에 있는 수식어는 내부 경계선인 접속사를 넘어가는 수식을 하지 않지만, 외부에 있는 수식어는 경계선을 넘어 양쪽의 지배소를 수식한다.

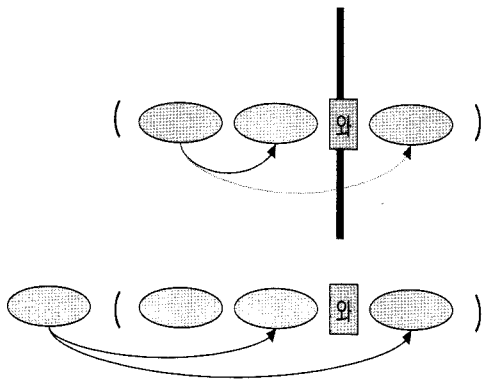


그림 5 내부 수식어와 외부 수식어

5.2 수식관계정보를 이용한 범위 조절

앞에서 정렬모델로 구한 범위를 수식관계를 이용해서 조절한다. 내부 수식관계인지 외부수식 관계인지를 결정하면 범위의 변경이 필요한지 아닌지를 알 수 있다.

이를 위해 단어 사이에 어떤 관계가 있는지 관계정도를 구한다. 그 식은 다음과 같다.

$$rel(a,b) = P(a,b)$$
$$= \max_c P(a,c) * sim(b,c)$$

두 단어 사이의 관계확률을 사용한다. 데이타 부족 문제로 인하여 두단어 사이의 관계가 항상 나타나지 않을 수도 있다. 이럴 경우 관계를 가지고 있는 다른 단어를 이용한다. 즉, a와 b사이의 관계값이 존재하지 않는 경우 a와 관계값이 존재하는 c중에서 최대의 관계값을 사용하되, b와 c의 유사도를 구해서 곱해준다. 유사도로는 시소러스의 정보량을 사용한다.

5.3 시소러스의 정보량

기존에 사용되었던 방법들을 살펴보면, 일단 모두 2단어간의 공통 부모점을 구한다. 공통점이 두 단어 사이의 관계를 시소러스에서 표현해 주기 때문에, 이를 통해 유사도를 구한다[10].

구로하시는 단순하게 이 공통점의 깊이를 유사도로 사용했다. 그럴 경우, 비교하려는 두 단어가 어떤 위치에 있던, 공통점만 일치한다면 모두 같은 유사도를 가지게 된다. 이를 해결하기 위해 비교 단어의 깊이로 공통점의 깊이를 나눠주는 정규화 방법을 사용했다.

이런 공통점의 깊이를 기반으로 하는 방법의 문제점은 같은 깊이를 가지는 단어라고 해도, 실제로 시소러스상의 최초 점에서 그 깊이까지 같은 깊이를 가진다고 가정할 수 없다는 것이다. 시소러스를 정교하게 만들지 않는 한 같은 깊이에 있는 형제점들이라도 상위에 대한

거리는 같다고 가정할 수 없다. 이 문제를 해결하기 위해 Resnik[10]의 공통점에 대한 정보량을 사용하였고, 본 논문에서도 이 방법을 적용하였다.

5.4 관용적 어휘 정보 사용

관용적으로 사용되는 어휘에는 범위 결정의 규칙성이 있다. 명사 파생 접미사 "간"이 출현 했을 때 명사구의 범위를 그 접미사 앞까지로 한다. "사이"도 의미에 의해 비슷한 역할을 하기 때문에 범위결정 요소로 사용했다. 쉼표도 같은 방법으로 사용했다. 예를 들어 '[현실과 꿈 사이]'의 잘못된 병렬내포 명사구 범위를 범위결정 요소를 사용하여 '[현실과 꿈] 사이'로 바로잡을 수 있다.

왼편의 중심어(head)의 어휘 정보를 이용해서 다음과 같은 알고리즘도 사용했다.

(1) 오른쪽에 중심어와 같은 어휘가 출현하면 오른쪽 범위를 그 어휘까지 확장한다. 단, 쉼표와 같이 분리자가 그 사이에 있을 경우 확장하지 않는다.
(2) 오른쪽 범위의 어절 패턴과 같은 방식으로 왼쪽의 어절 패턴이 전개된다면 왼쪽의 범위를 그 패턴까지 확장한다.
(3) 왼쪽의 범위 밖에 수식어나 복합명사를 이루는 명사가 있다면 범위를 거기까지 확장한다.

6. 실 험

6.1 실험 환경

실험에는 KIBS과제로 1996~1997년 구축된 과기원 코퍼스[9] 중에서 나무부착(tree-tagged) 코퍼스 31,086문장을 사용하였다. 4,176(13.4%) 문장에서 조사 와/과가 출현하고 있었다. 이 중에 평가가능한 3,383개를 정답집합으로 사용했으며, 이를 가지고 수식관계정보를 위한 단어간의 관계확률을 구하였다.

와/과로 묶이는 자료를 살펴보면 좌우 한 형태소씩 묶이는 자료가 48.98%로 대부분을 차지한다. 이것이 병렬구 범위결정 모델 성능평가의 최저 기준선이 된다.

결정나무(decision tree) 구현을 위해서는 C4.5[6]를 사용했다. 최대 엔트로피 모델[11]을 위해서는 MEMT[12]을 사용했다. 지지벡터기계 모델을 사용하기 위해서는 SVM light[7]를 사용하였다. 제안한 모델은 브라운[13]이 통계적 기법을 사용해서 기계번역을 하는 방법으로 제안한 정렬 모델과 유사성을 가진다. 정렬 모델 적용을 위해서는 GIZA++[14]을 사용했다.

6.2 정확도 비교 실험

과기원 코퍼스[9]에서 형태소 분석결과를 가지고 있는 문장을 학습집합으로 썼다. 와/과 조사가 들어간 학습집합의 문장 수는 나무 부착 코퍼스의 4,176문장을 합하여 총 43,575문장이다. 제안한 정렬에 바탕을 둔 모델에서는 이 모든 문장이 학습집합으로 사용할 수 있었다. 하지만, 결정나무(decision tree) 기반 모델이나 최대 엔트로피 기반 모델에서는 정답이 확실한 3,383문장만을 학습 문장으로 사용할 수 있었다. 실험은 어절단위를 기반[15]으로 이루어졌다.

수식구조 관계를 파악하기 위해서 사용한 시소러스는 코어넷[16]이고, 코어넷의 각 단어에 정보량을 부여하기 위해서 사용한 자료는 앞에서 기술한 43,575문장이다.

대칭성 분석 모델[5]을 한국어에 적용할 때, 일본어와의 언어적 차이가 있는 부분을 고려하여 적용하였다. 어절을 기본 단위로 했으며, 두 어절간의 일치 품사 개수에 2점을 준 후에, 품사열의 길이로 나누어 주어 정규화시킨 유사도를 구했다.

각 모델별 정확도는 아래 표와 같다.

6.3 교환정렬 모델의 정확도

교환정렬 모델에서 병렬구 종류별 정확도는 다음 표와 같다. 병렬구 종류는 1장에서 예로 든 병렬구들을 바탕으로 4종류로 나눴다. 수식구조 정보를 사용해서 전체 정확도가 70.47에서 73.13으로 증가하였다. 수식관계정보는 제일 결과가 좋은 교환정렬 모델에만 사용하였다. 복합명사 범위에 대해서는 다른 방법의 처리[17]가 필요함으로 복합명사에 대한 결과는 전체 정확도에 포함시키지 않았다. 수식관계정보 처리 과정에서 관용적 어휘 정보를 사용함으로 인해 통사수준의 병렬구에서는 오히려 정확도가 떨어졌으나, 다른 경우에서 정확도가 올라가서 전체적인 정확도가 높아졌다.

표 4 수식관계정보처리 이전/이후 정확도 비교

병렬구 종류	수식관계정보처리 이전 정확도	수식관계정보처리 이후 정확도
(1) 통사수준	96.54 (978/1013)	95.26 (965/1013)
(2) 복합명사	31.31 (103/329)	39.82 (131/329)
(3) 수식구조	62.48 (1114/1783)	66.91 (1193/1783)
(4) 단어의미 필요	49.74 (292/587)	53.83 (316/587)

7. 결 론

그림 6은 병렬내포 명사구 범위 탐지를 위해 사용 가능한 모델들의 개념을 비교해서 표현한 것이다. 구로하시[5]는 (b)와 같은 완전한 대칭적 구조에 기반한 모델

표 3 다양한 모델과의 정확도 비교

모델	대칭성 분석	결정 나무	최대 엔트로피	지지벡터기계	교환정렬
정확도	65.36	67.34	50.01	60.28	73.13

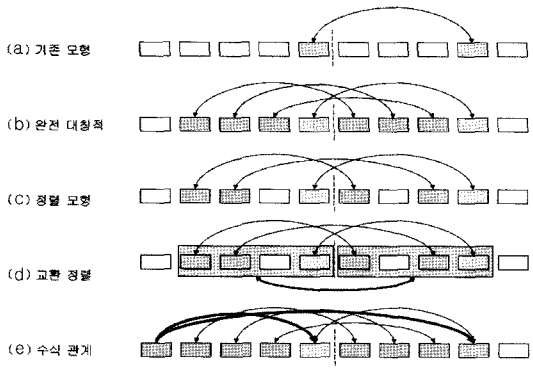


그림 6 다양한 모델들의 개념 비교

을 제시하면서 기존 모델은 (a)와 같이 지배소만을 고려했기 때문에 병렬구조 파악에 문제가 있음을 지적하였다. (b)와 같이 모든 요소들의 대칭성을 비교하는 것이 (a)와 같은 모델보다 발전한 형태지만, 비슷한 대칭적 구조에 약점을 가지는데 (c)와 같은 정렬모델을 통해서 비슷한 대칭적 구조도 확률적으로 포괄할 수 있는 방안을 제안하였다. 병렬구는 교환이 가능하다는 교환성을 활용하여 (d)와 같은 병렬명사구의 좌측과 우측을 비교할 수 있는 교환정렬 모델도 제안하였다. 마지막으로 (e)와 같이 수식관계를 활용하는 방법도 제안하였다.

본 논문에서는 와/과 주변의 병렬구조 해석을 위해 정렬과 수식구조 정보를 사용하는 방법을 제안하고, 다른 모델들을 사용하여 비교 평가를 하였다. 제안한 모델은 병렬구의 대칭성을 이용한 확률식을 사용한 정렬 모델을 제안하고, 병렬구조만의 특징인 교환성을 표현할 수 있는 교환모델로 변형하는 방법을 제안했다. 이 방법은 비지도식 언어에 비의존적인 확률에 기반한 결과에, 수식구조 정보와 어휘 정보 등을 사용해서 병렬명사구의 범위를 확정하는 방법을 제안했다. 제안한 방법은 다른 모델들에 비해서 우수한 성능을 보였으며, 확률식 자체에 지향점이 표현되는 비지도식 학습 모델로 정답 자료가 없을 경우에도 학습이 가능한 강점을 가지고 있다. 이 강점은 다른 모델들보다 더 많은 양의 학습을 가능하게 함으로서, 정확도가 더 높을 수 있었다.

현재, 격틀 등을 이용한 개념체계 간의 관계정보를 사용해서 더 높은 성능을 올릴 수 있는 방법에 대한 연구와 어휘정보를 효율적으로 사용할 수 있는 방법에 대한 연구와 더불어 코퍼스에서 병렬내포 명사구 조사로 제일 많이 출현하는 '와/과'뿐만 아니라 다른 병렬조사에 의한 접속구문에 대해서도 연구도 진행 중이다.

참 고 문 헌

[1] Kurohashi, Sadao and Makoto Nagao, 1994a. KN Parser: Japanese dependency/case structure analyzer. In Proceedings of Workshop on Sharable Natural Language Resources, pages 4855.

[2] Abney, S., "Parsing by Chunks," In R.C. Berwick, S.P. Abney and C. Tenny, editors, Principle-Based Parsing: Computation and Psycholinguistics, Kluwer, pp. 257-278, 1991.

[3] 이관규, "국어 대등구성 연구", 서광학술 자료사, 1992

[4] 박준식, "품사 패턴을 이용한 한국어 병렬 구문의 해석", 한국과학기술원 석사학위 논문, 1998.

[5] Kurohashi, S. and Nagao, M., "A Syntactic analysis method of long Japanese sentences based on detection of conjunctive structures," Computational Linguistics, Vol.20, No.4, pp. 507-534, 1994.

[6] Quinlan, J. Ross, "C4.5:Programs for Machine Learning", Morgan Kaufmann Publishers, 1993.

[7] Joachims, Thorsten, Learning to Classify Text Using Support Vector Machines. Dissertation, Kluwer, 2002.

[8] Corbett, Edward P. J. Classical Rhetoric for the Modern Student. 3rd ed. NY: Oxford University Press, p. 428. 1990.

[9] The KAIST corpus 1996-1997, Korea Advanced Institute of Science and Technology, http://korterm.org/, 1997.

[10] Resnik, Philip, "Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language," Journal of Artificial Intelligence Research, Vol.11, pp. 95-130, 1999.

[11] Jaynes, E.T., "Information theory and statistical mechanics," Physics Reviews106, pp. 620-630, 1957.

[12] Eric Sven Ristad. 1998. Maximum entropy modeling toolkit, release 1.6 beta. http://www.mnemonic.com/software/memt.

[13] Brown, P. F., S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. "The mathematics of statistical machine translation: Parameter estimation. Computational linguistics, Vol.19, pp. 263-312, 1993.

[14] Och, Franz Josef, Hermann Ney, "A Systematic Comparison of Various Statistical Alignment Models," Computational Linguistics, 29(1):19-51, 2003.

[15] Choi, Yong-Seok, Ji-Ae Shin, Key-Sun Choi (2006), Identification of Boundaries in Parallel Noun Phrases: A Probabilistic Swapping Model, International Journal of Computer Processing of Oriental Languages, 19(2&3), 109-132.

[16] Choi, Key-Sun, Hee-Sook Bae, Procedures and Problems in Korean-Chinese-Japanese Wordnet with Shared Semantic Hierarchy, WordNet Conference, pp. 320-325, 2004.1, Brno, Czech.

[17] Yoon, Juntae, Key-Sun Choi, Mansuk Song "Corpus-Based Approach for Nominal Compound Analysis for Korean Based on Linguistic and Statistical Information," Natural Language Engineering vol 7/No 3, 251-270, 2001.

최 용 석

1995년 한국과학기술원 전산학과 학사 1997년 한국과학기술원 전산학과 석사 1997년~현재 한국과학기술원 전산학과 박사과정. 2005년~현재 한국표준과학연구원 지식정보팀 선임. 관심분야는 자연언어처리, 기계번역, 정보검색

신 지 애

1984년 부산대학교 계산통계학과 학사 1986년 한국과학기술원 전산학과 석사 2004년 New York University, Computer Science, Ph.D. 2005년~현재 한국정보통신대학 공학부 교수. 관심분야는 AI Planning & Scheduling, Semantic Technology

최 기 선

1978년 서울대학교 자연과학대학 수학과 졸업(학사). 1980년 한국과학기술원 전산학과 졸업(석사). 1986년 한국과학기술원 전산학과 졸업(박사). 1987년~1988년 일본 NEC C&C 정보연구소 연구원. 1988년~현재 한국과학기술원 전산학과 교수 1997년~1998년 미국 스탠포드대학 CSLI 객원교수. 2002년~2003년 일본 NHK 방송기술연구소 초빙연구원. 2006년 한국인지과학회 회장. 2003년~현재 국가지정 언어자원특수소재은행장 http://bola.kaist.ac.kr. 1998년~현재 전문용어 언어공학연구센터 http://korterm.or.kr/. 2006년~현재 시맨틱웹첨단연구센터 센터장 http://swrc.kaist.ac.kr/ 관심분야는 온톨로지, 텍스트마이닝, 인공두뇌, 지식획득, 창의계산론, 언어공학, 시맨틱웹