정렬기법을 활용한 와/과 병렬명사구 범위 결정

Range Detection of Wa/Kwa Parallel Noun Phrase by Alignment method

최용석, 신지애*, 최기선**, 김기태, 이상태

한국표준과학연구원 지식정보팀
*정보통신대학교 공학부 교수
**한국과학기술원 전산학과 학과장

ABSTRACT

In natural language, it is common that repetitive constituents in an expression are to be left out and it is necessary to figure out the constituents omitted at analyzing the meaning of the sentence. This paper is on recognition of boundaries of parallel noun phrases by figuring out constituents omitted. Recognition of parallel noun phrases can greatly reduce complexity at the phase of sentence parsing. Moreover, in natural language information retrieval, recognition of noun with modifiers can play an important role in making indexes.

We propose an unsupervised probabilistic model that identifies parallel cores as well as boundaries of parallel noun phrases conjoined by a conjunctive particle. It is based on the idea of swapping constituents, utilizing symmetry (two or more identical constituents are repeated) and reversibility (the order of constituents is changeable) in parallel structure. Semantic features of the modifiers around parallel noun phrase, are also used the probabilistic swapping model. The model is language-independent and in this paper presented on parallel noun phrases in Korean language.

Experiment shows that our probabilistic model outperforms symmetry-based model and supervised machine learning based approaches.

Keyword:Korean parsing, natural language processing, parallel structure analysis

1. 서론

1.1. 연구의 배경 및 목적

정보검색에서 기존에는 명사 위주의 색인어가 많이 쓰였으나, 최근에는 수식어 정보가 중요한 역할을 하고 있다.

"What is the most popular classical music?"

위의 예에서와 같이 classical이 수식어로서 music을 꾸며주고 있으며, 이는 정보검색에서 중요한 정보로 이용된다. 정보검색에서 수식어 정보를 추출하기 위해서 다음과 같은 상황도 고려하게 된다.

"classical music and musician"의 경우에는 "classical music"과 "classical musician"이 색인어로 사용된다. 하지만, 같은 구조인 "modern music and computer"의 경우에는 "modern music"과 "computer"가 색인어로 사용된다.

병렬구조란 두 가지 이상의 문법단위가 동등한 자격으로 나타나는 것을 의미한다. 공통요소를 생략해서 표현의 효율성을 높이는 구조인다. 위의 예에서 "classical music and musician"의 경우 classical이 공통요소라 생략된 병렬 구조이고, "modern music and computer"는 생략된 요소가 없는 병렬구조이다.

병렬구조는 대칭성(symmetry)과 교호성(reversibility)을 갖는 특징이 있다. 대칭성은 동등한 문법요소가 대응되는 특징으로 특정요소가 생략될 경우 대칭성이 깨지게 된다. 교호성은 대응되는 단위를 맞바꿔도 의미에 변화가 없다는 특징으로 "classical music and musician"의 경우 "music"과 "musician"을 맞바꿔도 의미상 차이가 없다.

본 논문에서는 병렬구조의 두 가지 특징인 대칭성과 교호성을 이용해서 확률기법으로 병렬명사구 범위를 결정하는 방법을 제안한다. 실험을 통해 제안한 방법의 성과를 살펴본다.

2. 병렬구와 구문적 병렬범위 추정

2.1. 병렬구의 정의

이항병렬구조라 함은 대칭적 병렬구조로서, "X op Y"와 같이 병렬어휘 op의 양 옆에 X, Y라는 비슷한 대칭적 구조를 갖는다. 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, opn은 마지막 병렬연산자로서 그 외에 여러 개의 비슷한 병렬 대상이 있다는 뜻에서 "등" (영어에서는 "etc.")으로 나타난다.

"비슷한 대칭적 병렬구조"를 정의하기 위하여 "완전한 대칭적 병렬구조"를 먼저 정의하여 보자. 예를 들면, "훌륭한 부모와 성실한 자녀"와 같은 것이다. 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 강제로 대칭화됨을 의미한다. 이 강제 대칭화의 과정에서 병렬구조의 경계 인식이 문제가 된다.

2.2. 병렬구의 의존구조 인식 단계

병렬구의 의존구조 인식을 3 단계로 생각하여 보자. 제1단계는 상위구조의 하나의 문장성분으로서 병렬구의 범위 전체를 파악하여, 제2단계에서 완전한 병렬핵을 파악하므로서 완벽한 구문적 대칭구조인식을 종결한다. 이는 제3단계에서 의미 혹은 문맥해석에 따라 제1단계에서 인식한 병렬구의 내부구조를 의미적 완전대칭구조로 이끌기 위한 것이다.

제1단계: 문장성분으로서 병렬내포명사구의 경계를 인식한다. "α에는 β가 V"의 문장구조에서 β의 범위를 찾는다.

제2단계: 구문적 완전대칭구조인 병렬핵(X&Y)을 구한다. 병렬핵은 제1단계에서 구한

병렬내포명사구의 범위 안에 존재한다. 즉, β=β'(X&Y)β"이다. 따라서 제3단계에서 병렬핵(X&Y)과 병렬내포명사구 안의 병렬핵이 아닌 남은 구조인 β'과 β"간의 모호성을 해결하여, 구문-의미적 완전대칭구조를 구한다. 그 결과, 병렬명사구의 의존구조를 파악할 수 있다.

제3단계: 완전한 의존구조를 파악할 수 있다. 병렬내포명사구의 범위 내에서 의미적 완전대칭구조는 병렬범위모호성을 해결한 결과가 된다. (X&Y)β"와 (X&(Yβ"))로 됨을 알 수 있다. 여기서 (X&Y)β"는 의미적으로 (Xβ")&(Yβ")와 동치이다.

번역을 목적으로 한다면 제3단계까지 해석을 하여야 한다. 그러나, 제3단계는 의미 혹은 문맥해석이 필요하므로, 구문해석의 단계에서는 제2단계까지 진행하도록 한다. 즉, 구문해석의 결과는 수식모호성이 남아있는 구조로서 제2단계의 결과인 $\beta = \beta'(X\&Y)\beta''$이다.

4. 제안 모형

4.1. 병렬명사구의 대칭성

병렬구조로 이루어진 문장에서는 문장을 좌우로 분리한 다음에 좌측의 어느 부분이 우측과 대응하는 지를 살펴보고 병렬구조의 범위를 정하는 모형이다. 모형의 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 모형들의 성질이 결정된다. 수식에서 l_1^J 는 병렬기호 op 왼쪽의 어절들로 1부터 J까지의 어절을 가진다. r_1^I 는 병렬기호 op 오른쪽의 어절들로 1부터 I까지의 어절들을 가진다.

a_1^J 는 왼쪽의 어절들의 오른쪽에 대한 대응 정보를 의미한다.

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

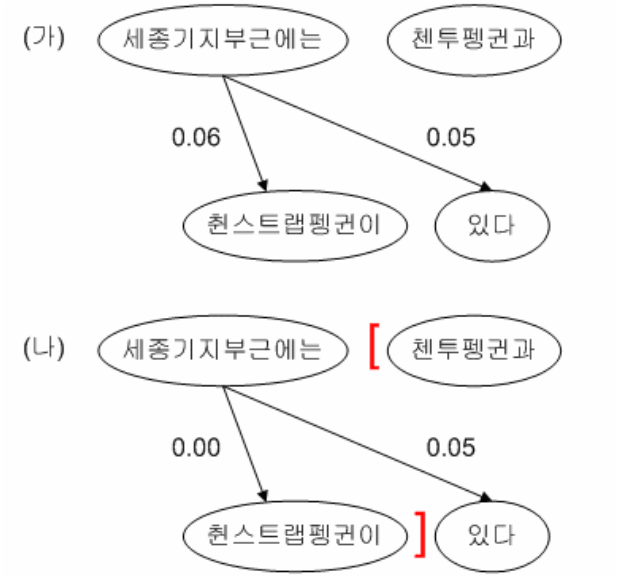
결과로 나온 대응을 보고 병렬구조의 범위를 결정한다. 우측의 시작 단어와 대응하는 좌측의 단어가 병렬구문의 시작이고, 좌측의 마지막 단어와 대응하는 우측의 단어가 병렬구문의 끝이 된다.

4.2. 조건부 대응확률 모형 제안

한 문장에서 병렬구가 결정되면 병렬구 외부에 있는 요소와 내부 요소간 대응 확률은 기본적으로 없어야 한다. 병렬구 내부 요소는 내부 요소끼리 대응되고, 외부요소는 외부요소끼리 대응되어야 한다. 대응 오류가 일어난 대응확률을 낮춰줘서 원하는 정렬이 일어날 수 있도록 한다. 정렬확률은 기본모형과 같은 확률을 쓴다. 조건부 대응확률 모형은 대응확률자체의 수식을 다음과 같이 정의한다.

$$P(l_1^J, a_1^J \mid r_1^I) = p(s \mid l_1^J, r_1^I)\, p(t \mid l_1^J, r_1^I) \prod_{j=1}^{J} p(l_j \mid r_{a_j}, s, t, l_1^J, r_1^I)$$

다음 그림과 같이 (가)에서는 0.06의 확률을 가지고 있었으나, 조건부로 범위 밖으로 나가게 되면 범위 밖의 요소와 범위 안의 요소가 대응될 확률은 0이 된다.



5. 결 론

5.1. 실험

과기원 말뭉치[KAIST 1997]에서 형태소 분석결과를 가지고 있는 문장을 학습집합으로 썼다. 와/과 조사가 들어간 학습집합의 문장 수는 나무 부착 말뭉치의 4176문장을 합하여 총 4만 3575문장이다. 제안한 정렬에 바탕을 둔 모형에서는 이 모든 문장이 학습집합으로 사용할 수 있었다. 하지만, 결정나무 기반 모형이나 최대 엔트로피 기반 모형에서는 정답이 확실한 3383문장만을 학습 문장으로 사용할 수 있었다. 실험은 어절단위를 기반으로 이루어졌으며 정확도는 2단계 수준의 평가 결과이다.

모형	대칭성 분석	결정 나무	최대 엔트로피	지지벡 터기계	조건부 확률
정확도	65.36	67.34	50.01	60.28	68.32

위의 표와 같이 조건부 확률 모형의 결과가 나왔다.

5.2. 결과 분석

다른 모형과 비교해서 조건부 확률 정렬 모형이 우수한 성능을 보여 주고 있다. 다른 학습 모형들은 정답 집합이 있어야만 학습이 가능하지만, 정렬 모형은 기본 자료 모두를 학습에 사용할 수 있다. 정렬 모형은 병렬구의 좌우 대칭성 정보만으로 학습하는 것으로 병렬구 구조 결정에 잘 어울리는 특성을 가지고 있는 모형이다.

참고문헌

[1] 최용석, 신지애, 최기선, (2008) “확률모형과 수식정보를 이용한 와/과 병렬명사구 범위결정”, 한국정보과학회논문지,

[2] 이관규, (1992) "국어 대등구성 연구", 서광학술자료사,

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

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

[5] 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.

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

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