한국어의 이중주어 분석

언어처리연구팀, 음성/언어정보연구센터, 한국전자통신연구원
김 창 현·박 상 규

Double Subject Analysis in Korean

Chang-Hyun Kim, Sang-Kyu Park
NLP Team, Speech/Language Technology Research Center, ETRI, Daejeon, Korea

요 약

이중주어 문장이란 하나의 용언이 두개의 주격조사를 취하는 경우를 말한다. 이러한 이중주어 현상은 영어권에서는 없는 현상으로, 한국어 분석 측면에서 어려움을 야기할 뿐 아니라, 자동 번역 측면에서도 반드시 고려되어야 하는 현상이다. 그러나 이중주어의 분석에 대한 기존 연구는 국어학 분야에서만 진행되었을 뿐 자연어처리 분야에서는 다루어 진 적이 없다. 본 논문에서는 이중주어 현상에 대한 분석을 통해, 이중주어 현상이 한국어 문장에서 빈번히 발생하는 현상이며, 기존의 '명사-격조사-용언'으로 구성되는 공기정보만으로는 이중주어 문장을 올바로 분석할 수 없음을 보인다. 이를 해결하기 위해 본 논문에서는 이중주어의 특성을 파악하고, 이들 특성에 맞게 '명사-격조사-용언' 공기정보뿐 아니라 명사들 간의 공기정보 및 관형형 용언과 명사 공기정보, 그리고 주격조사의 교체를 통한 이중주어 분석 방법을 제안한다.

서 론

이중주어 문장이란 아래의 예와 같이 하나의 용언이 두개의 주격 조사를 취하는 경우를 말한다.

a. 그 차가 속도가 빠르다
b. 사장님이 결심이 섰다

이러한 이중주어 현상은 주어가 발달된(subject-prominent) 인도-유럽 어족에서는 없는 현상으로, 한국어 분석 측면에서 어려움을 야기한다. 한국어 분석 시의 기본적인 규칙으로 일문 일격의 원칙을 적용하기 때문이다.

c. 목이 염증이 생겨 물을 마실 수 없다
d. 목소리가 부드러운 음악과 잘 어울린다.
e. 목소리가 부드러운 사람과 어울리고 싶다.

문장 c의 경우 '마시다'의 주어가 생략된 문장으로서, 이중주어를 고려하여 분석하지 않는다면 '목-이'는 주어가 없는 '마시다'의 주어로 분석된다. 문장 d, e는 모두 형용사 '부드럽다'를 포함하고 있으나, d는 이중주어 문장이 아닌 경우이고 e는 이중주어 문장인 경우이다. 이러한 이중주어의 분석은 구조분석에서의 가장 정확한 정보라 할 수 있는 '어휘-조사-용언' 공기정보를 이용하더라도 제대로 분석하기가 어렵다. 문장 d, e에 '명사-조사-용언' 공기정보를 적용할 경우 아래와 같은 3가지 종류의 공기정보를 비교하게 된다.

목소리-가-부드럽다
음악-가-부드럽다
사람-가-부드럽다

그러나, 위 3가지 공기정보는 모두 높은 빈도수로 발생되며, 이들을 서로 비교하더라도 올바른 구조를 파악하기 힘들다.

자동 번역 측면에서도 이중주어의 분석은 고려되어야 하는 현상이다.

f. 산이 좋은 사람⇒the person who likes mountain

E-mail : chkim@etri.re.kr
E-mail : parksk@etri.re.kr

g. 산이 좋은 이유⇒the reason why mountain is good

예문 f, g는 모두 문장 '산이 좋다'를 포함하고 있으나, 이중주어인지의 여부에 따라 각각에 대한 번역이 달라지게 된다.

그간 이중주어 분석을 위한 연구는 국어학 분야에서만 진행되었을 뿐 자연어 처리 분야에서는 다루어진 적이 없다. 이중주어 현상이 일반적이지 않으며 발생 빈도수가 낮다는 것이 그 주된 이유이다. 문장 c와 같이 연결 어미를 취하는 용언 앞에 이중주어가 나타나는 경우는 상대적으로 발생 빈도가 낮을 수 있으나, 문장 d, e, f, g와 같이 관형사형 전성어미를 취하는 용언과 함께 나타나는 이중주어는 높은 빈도로 국어 문장에서 발생하는 현상이며, 이중주어에 대한 처리는 자연어 처리에서 반드시 다루어져야 한다. 이중주어 처리를 위해서는 기존의 '명사-격조사-용언'으로 구성되는 공기정보만으로는 충분하지 않다. 따라서 이를 해결하기 위해 본 논문에서는 명사들 간의 공기정보 및 관형형 용언과 명사 정보를 이용하는 방법, 그리고 주격조사의 교체를 이용한 방법을 제시하고, 이들 정보를 이용함으로써 보다 정확한 이중주어 문장의 분석이 가능함을 보인다.

이중주어 문장의 유형

국어학에서 이중주어를 설명하는 입장은 다양하다. 주제설, 변형설, 동격주어설, 서술절설, 구동사절 등이 그것이다.[3] 본 논문에서는 국어학에서의 이중주어 설명에 대한 입장에 서지 않고 자연어 처리의 관점에서 이중주어의 유형을 크게 3가지로 구분하였다. 이때 NP1과 NP2는 이중주어문의 첫번째 주어 및 두번째 주어를 지칭한다.

유형 1 : NP2가 보어인 경우
  물이 얼음이 되었다
  그는 천재가 아니다

유형 2 : NP2가 주어이고 NP1은 '의/에/에서/에게'의 변형으로 볼 수 있는 경우
  '의' 변형 :
  그녀가 마음이 곱다(그녀의 마음이 곱다)
  코끼리가 코가 길다(코끼리의 코가 길다)
  '에' 변형 :
  그 일이 돈이 든다(그 일에 돈이 든다)
  이 집안이 자손이 귀하다(이 집안에 자손이 귀하다)
  '에게' 변형 :
  내가 뱀이 무섭다(내게 뱀이 무섭다)
  내가 산책이 즐겁다(내게 산책이 즐겁다)
  '에서' 변형 :
  보석이 빛이 난다(보석에서 빛이 난다)
  이 밭이 채소가 잘 자란다(이 밭에서 채소가 잘 자란다)

유형 3 : NP2가 수량사 혹은 재귀대명사인 경우
  학생이 열명이 모였다
  농경지가 2천ha가 침수되었다
  그는 자신이 직접 차를 몰고 여행을 했다.

유형 1은 주격조사를 2개 취한다는 점에서 이중주어라고 분류하였으나, 기존의 학교 문법을 따르는 관점에서 이들 용언들은 보어를 취하는 용언으로 분류된다. 이 부류에 속하는 용언은 '되다'와 '아니다'이다. 유형 3은 형태적으로 파악이 용이하다. 따라서, 본 논문에서는 유형 2에 대해서만 다루고자 한다.

이중주어 발생 유형

이중주어 문장은 형용사 및 자동사인 경우에 발생한다. 본 논문에서는 이중주어 문장에 대한 처리를 아래와 같이 두 단계로 구분한다.

- 이중주어 처리 여부의 판단
- 이중주어 분석 단계

모든 용언 및 주어에 대해 이중주어 처리를 수행하는 것은 심각한 부하를 야기하며, 따라서 이중주어 처리 여부의 판단 단계에서는 자동사 및 형용사에 대해 다음과 같은 경우에 이중주어 처리를 수행한다.

- 자동사 혹은 형용사 바로 앞에 두 개의 주어가 나타나고, 해당 용언이 대등절이나 종속절을 이끄는 경우, 예) 그녀가 마음이 고와서…
- 자동사 혹은 형용사 바로 앞에 한 개의 주어가 발생하고 관형절을 이끄는 경우, 예) 마음이 고운 그녀가…

이때, 자동사나 형용사와 주어 사이에 부사가 있는 경우에는 바로 인접한 것으로 간주한다. 이중주어 처리가 필요하다고 판단된 문장 형태를 간략화하면 다음과 같이 나타낼 수 있다.

- NP1-가 NP2-가 p_{s|c}[1]
- NP2-가 p_m[2] NP1

1 P_{s|c} : {종속|대등} 연결어미 취하는 용언
2 P_m : 관형형 연결어미를 취하는 용언

이중주어의 발생 빈도를 보면 대등절이나 종속절을 이끄는 경우보다는 관형절을 이끄는 경우가 더 자주 발생한다. 실제 이중주어 문장인지의 여부와 상관 없이 이중주어 처리 여부의 판단이 필요한 경우만을 조사해 본 결과 위의 두 가지 형태가 발생하는 전체 빈도수와 대비하여 관형절 형태의 발생 빈도수가 전체의 95% 이상이었다. 그러나, 이중주어 여부를 결정하기 위해 두 가지 형태에 적용하는 방법론은 기본적으로 동일하다.

자료획득

이중주어 여부를 판단하기 위한 가장 기본적인 정보로는 어휘공기패턴인 <명사, 조사, 용언> 정보를 이용한다. 어휘공기패턴 정보는 형태소 분석기와 태거를 원시말뭉치에 적용하여 얻은 결과에 대해 휴리스틱 규칙을 적용하여 추출한다. 이때 발생하는 오류를 최소한으로 하기 위해 최대한 정확한 쌍들만을 추출하는 휴리스틱을 이용한다. 기본 휴리스틱은 다음과 같다.

규칙 1 : 문장의 가장 마지막 용언인 p(n)과 바로 이전 용언인 p(n-1) 사이의 어절 'N(1)…N(k)'에 대해 <N(1), 조사, p(n)>, ..., <N(k), 조사, p(n)>을 추출한다. 이때 <N(i), 가, p(n)> 추출 시 <N(j), 가, p(n)>(i<j)가 존재하면 <N(i), 가, p(n)>을 추출하지 않는다.[3]

그러나, 위의 규칙만으로 추출되는 자료는 그 양이 적을 수밖에 없다. 따라서, 조금은 불확실하더라도 더 많은 자료 확보를 위해 다음과 같은 규칙을 첨가한다.

규칙 2 : 용언 p가 타동사일 때

- 관형절이 아닌 경우, 부사를 제외하고 바로 인접한 어절이 '명사-격조사'이면 <명사, 격조사, p>를 추출한다.
- 관형절인 경우, 부사를 제외하고 바로 인접한 어절이 '명사-목적격조사'이면 <명사, '를', p>를 추출한다.[4]

규칙 3 : 용언 p가 자동사일 때

- 관형절이 아닌 경우, 부사를 제외하고 바로 인접한 어절이 명사-격조사이면, <명사, ... 한다. 이때, p가 대등적 연결어미를 취하... 이 관형형을 취하는 동사이면 해당 공기정보를 추출하지 않는다.

규칙 4 : 용언 p가 형용사이고, 부사형 전성어미와 관형형 전성어미를 취하지 않는 경우, 바로 인접한 어절이 '명사-{주격조사, 부사격조사}'이면 <명사, 격조사, p>를 추출한다. 이때, p가 대등적 연결어미를 취하고 바로 다음 용언이 관형사형 전성어미를 취하는 형용사이면 추출하지 않는다.

규칙 3, 4의 경우 관형절과 대등적으로 연결된 경우에 대해서는 오류 가능성이 많으므로 추출하지 않는다. 예를 들어, '그에게 착하고 예쁜 여자를 소개시켜 주었다'에서 <그, 에게, 착하>를 추출하지 않는다. '하다' 동사의 경우 서술성 명사와 결합하여 '-를 하다'와 같은 형태로 많이 발생한다. 이 경우에는 '-를 하다'를 '-하다'와 같은 형태의 하나의 동사로 취급해서 위의 어휘공기패턴 추출 알고리듬을 적용한다.

그러나, 이렇게 추출된 어휘공기정보도 여전히 많은 오류를 포함하고 있다. 말뭉치 자체의 오류 및 품사 태깅 오류 때문이다. 동일한 어휘를 갖는 동사와 형용사 간의 구분 및 자타동사 간의 구분은 품사 태깅 단계에서 정확히 할 수 없다. 따라서, <명사, 조사, 용언> 정보 이외에 본 논문에서는 <p, N> 정보를 추출한다. <p, N> 정보는 이중주어 구문 여부를 파악할 때에 사용된다. 즉, 이중주어 문장 'NP1 NP2 p'에 대해 'NP2 p_m NP1' 형태는 가능하지만 'NP1 p_m NP2'는 불가능하다는 것을 이용한다. 예를 들어, 이중주어문 '그녀가 마음이 곱다'에 대해 '마음이 고운 그녀'는 가능하지만, '그녀가 고운 마음'은 가능하지 않다. 따라서, <p, N> 정보는 이중주어를 판단하는 중요한 정보가 된다.

규칙 5 : 용언 p가 관형형이고, p와 바로 인접하여 명사 N이 있는 경우

- N이 관형격 조사 '의' 이외의 조사를 취하고, N이 의존명사가 아닌 경우 <p, N> 추출

<p, N> 정보와 함께 명사 쌍 정보도 이중주어 여부를 판단하는 중요한 정보로 사용된다. 이중주어가 '의' 변형인 경우 NP1과 NP2 간에는 일반적으로 대상과 속성, 전체와 부분 등의 관계가 존재하며,[3] 이러한 관계는 복합명사 형태로도 자주 사용된다. 예를들어, '코끼리가 코가 길다'와 같은 경우 '코끼리코'와 같은 형태로도 사용된다.

3 이중주어의 특성을 고려한 것이다. '그 ... 은 경우, <돈, 이, 들다>만 추출하고 <일, 이, ...
4 관형절인 경우에는 '집으로 잡은 물고기 ... 이 오류가 많이 발생한다.

규칙 6 : 하나의 어절이 'N(1)…N(k-1)N(k)' 로 구성되어 있을 때, 해당 어절을 구성하는 명사들 중 가장 마지막 두 개의 명사에 대해 <N(k-1), N(k)>를 추출한다.

<N(k-1), N(k)>만을 추출하는 것은 추출되는 자료의 정확도를 높이기 위해서이다. N(k-2)의 경우에는 <N(k-2), N(k-1)>과 <N(k-2), N(k)> 가운데 어느 것이 정확한지 결정을 해야 하기 때문이다. 명사 N(i), N(i+1)은 하나의 어절 내에서 복합명사로 공기할 뿐 아니라 어절 단위로도 공기한다. 즉, '코끼리코' 와 같은 하나의 어절 형태로도 발생하지만 '코끼리 코' 와 같이 두개의 어절로도 공기한다. 이뿐 아니라 관형격 조사 '의' 를 이용하여 '코끼리의 코' 로도 공기한다.

규칙 7 : 연속된 두 어절이 'N(1) N(2)_{조사}' 형태로 공기할 때, <N(1), N(2)>를 추출한다. 이 때, N(1) 바로 앞 어절이 관형어 상당어구일 경우에는 추출하지 않는다.

규칙 8 : 연속된 두 어절이 'N(1)_의 N(2)_{조사}' 형태로 공기할 때, <N(1), N(2)>를 추출한다. 이 때, N(2)가 관형격 조사 '의' 를 취하는 경우 및 N1 바로 앞 어절이 관형어 상당어구일 경우에는 추출하지 않는다.

N(1) 바로 앞에 관형어 상당어구가 나타나는 경우에는 부정확한 데이터가 추출될 가능성이 높다. 예를 들어, '이 경우 건물이 불안정하다' 에서 <경우, 건물>을 추출하거나, '아이들이 뛰노는 모습의 그림' 에서 <모습, 그림>을 추출하는 것을 적절하지 않다.

분석 알고리듬 및 실험

이중주어의 두 가지 발생 형태인 'NP1-가 NP2-가 p_{s|c}' 와 'NP2-가 p_m NP1' 에 대한 분석 방법은 동일하다. 여기서는 'NP2-가 p_m NP1' 형태를 기준으로 처리 알고리듬을 기술한다. 이 때 사용되는 기호는 다음과 같다.

문장 형태 : NP2-가 p_m NP1

Fp1 : <p, NP1>의 빈도수

Fp2 : <NP2, 가, p>의 빈도수

Fn12 : <NP1, NP2>의 빈도수

Fn21 : <NP2,NP1>의 빈도수

MI1 : 조사 '에/에게/에서' 에 대해 <NP1, 조사, p>의 상호정보 중 가장 높은 값

MI2 : <NP2, 주격조사, p>의 상호정보

명사들 간의 공기 정보는 빈도수를 사용하고,[5] <명사, 조사, 용언> 간의 공기정보는 상호정보(mutual information)을 이용한다. 상호정보 값은 다음과 같이 각각의 확률값 pr을 이용하여 구한다.

MI(명사, 조사, 용언)=log(pr(명사, 조사, 용언)/(pr(명사, 조사)*pr(용언)))

전체 분석 알고리듬은 다음과 같다.

if *Fn12*>*T1*[6] && *Fn21*==0 : 1)
then 이중주어문장
elif *Fp2*>T2 && *Fp1*>T3 : 2)
then 이중주어문장
elif *MI1*>T4 && *MI2*>T5 : 3)
then 이중주어문장
else 이중주어문장이 아님

1)에서 명사들 간의 공기정보를 이용하는 이유는 NP1이 '의' 변형인 경우 NP1과 NP2 사이에 전체와 부분, 대상과 속성의 관계가 존재하여 NP1-NP2가 자연스러운 복합명사를 구성하는 반면 NP2-NP1은 그렇지 않다는 사실에 기반한다. 예를 들어, 아래의 두 문장에서

h. 소녀가 아름다운 미소가 돋보인다.
i. 미소가 아름다운 소녀가 돋보인다

'소녀-{의}-미소' 는 자연스러우나, '미소-{의}-소녀' 는 부자연스럽다. 따라서, '미소가 아름다운 소녀' 는 이중주어로 분석이 되지만 '소녀가 아름다운 미소' 는 이중주어로 분석되지 않는다. 문장 h, i를 <명사, 조사, 용언> 공기정보만을 이용해서 분석하는 경우, 위 두 문장은 모두 동일한 구조를 갖게 된다. <소녀, 가, 아름답>과 <미소, 가, 아름답>, <소녀, 가, 돋보이>와 <미소, 가, 돋보이> 모두 높은 연관도를 가지며, 따라서 위 두 문장은 <명사, 조사, 용언> 공기정보를 이용할 경우 동일한 구조를 가지게 된다. 그러나, 본 논문에서 제안하는 명사들 간의 관계를 이용할 경우에는 '소녀-미소' 와 '미소-소녀' 를 비교하게 되므로 두 문장 모두 올바르게 분석한다.

5 [1]에 의하면, 빈도수 정보를 그대로 이용하는 것도 충분히 좋은 결과를 보인다고 보고하고 있다.

6 threshold value

j. 우승이 가능한 상황이다

문장 j는, 명사들 간의 관계만으로는 이중주어 분석이 쉽지 않다. 따라서, 이 경우에는 2)를 적용하여, <우승, 이, 가능하>와 <가능하, 상황> 정보를 이용한다. Fp1의 빈도, 즉 <가능하, 상황>의 빈도가 높다는 것은 '상황'이 NP1일 가능성이 높다는 것이고, Fp2의 빈도, 즉 <우승, 가, 가능하>의 빈도가 높다는 것은 '우승'이 NP2일 가능성이 높다는 것이다. 따라서, 이 두 정보를 결합하여 이중주어 분석을 수행한다. 2)을 적용해서 이중주어문장으로 판정이 나지 않으면, 상호정보를 적용하는 3)을 수행하게 된다. 관형형 이중주어의 경우, 이중주어가 '에/에서/에게' 변형인 경우는 기존의 어휘공기정보를 이용하는 방법론과 동일하다. 관형형 이중주어의 경우 1)과 2)는 새로운 방법론이지만 3)은 기존의 방법론과 차이가 없다. 그러나, 대등/종속절 이중주어 유형의 경우에는 1)과 2) 뿐 아니라 3)도 새로운 적용 방법론이다. 즉, 첫번째 주어에 대해 주격조사를 '에/에서/에게'로 대치한 후 분석을 수행하게 되는 것이다.

실험 대상 문장은 발생 빈도가 높은 명사들만을 포함하도록 하였으며, 이중주어인 관형형 문장 10개와 이중주어가 아닌 관형형 문장 10개, 그리고 이중주어인 대등/종속절 문장 10개, 이중주어가 아닌 대등/종속절 문장 10개로 구성하였다. 관형형 문장은, 아래와 같이 용언의 개수는 2개이고, 주격조사는 한 번 혹은 두 번 발생하도록 선정하였다.

전망이 좋은 공원에 갔다
과일이 싼 가격에 팔린다

대등/종속절의 경우, 용언의 개수가 2개이고 주격의 개수도 2개인 문장을 선정하였다.

주가가 폭락이 지속되자 우울해졌다.
사람들이 물가인상이 지속되자 우울해졌다.

발생 빈도가 높은 명사들을 대상으로 수행하여 40문장 모두에 대해 정확한 구조를 파악할 수 있었다.

Table 1.

	관 형 형		대등/종속	
	10문장(이중)	10문장(이중X)	10문장(이중)	10문장(이중X)
정확도	100%	100%	100%	100%

대등/종속 이중주어 문장 선정 과정에서 NP1이 앞뒤 용언에 모두 걸리는 문장들도 종종 발생하였다.

아이들이 배가 아파 병원에 갔다

위와 같은 경우, '아이들'은 '아프다'와 '가다'의 주어에 해당한다. 실험 문장에서는 위와 같은 문장들은 모두 배제하였으나, 실제 구조분석 과정에서는 이러한 문제에 대한 처리를 고려해야 할 것이다.

현재의 실험 문장은 고빈도 명사만을 대상으로 하였으며, 어휘에 기반하고 있기 때문에 자료부족 문제에 대한 대처 방안이 필요하다.

결 론

본 논문에서는 이중주어문의 분석을 위해 새로운 분석 방법을 제안하였다. 첫째는 이중주어문을 구성하는 명사들 간의 명사공기정보를 이용하는 방법을 제안하였으며, 둘째는 관형형 용언과 피수식어인 명사를 이용하는 방법을 제안하였으며, 셋째는 이중주어문의 주격조사 대신에 '에/에서/에게'를 이용하여 이중주어문의 여부를 파악하는 방법을 제안하였다. 그러나, 현재의 방법론은 어휘 데이터만을 이용하고 있으므로 자료부족문제에 직면할 수 밖에 없다. 따라서, 자료 부족 현상을 완화하기 위하여 명사의 의미정보를 이용하는 연구를 현재 진행중에 있으며, 이를 통해 처리 범위를 확대할 수 있으리라 기대한다.

REFERENCES

1) Brigitte Krenn(2000) : *"Empirical Implications on Lexical Association Measures", Proceedings of the 9th EURALEX International Con-gress. Stuttgart, Germany. Brigitte Krenn*
2) 서정수(1971) : "국어의 이중주어 문제-변형생성 문법적 분석-". 국어국문학 52
3) 안명철(2001) : "이중주어구문과 구동사". 국어학 38, pp181-208