단위 명사간 보-술 관계를 이용한 한국어 복합 명사의 문장 복원

양성일*, 김영길*, 서영애*, 박은진*, 나동렬**
*한국전자통신연구원 음성/언어연구센터 언어처리연구팀
**연세대학교 정보기술학부
e-mail : {siyang, kimyk, yaseo, ejpark}@etri.re.kr*, dyra@dragon.yonsei.ac.kr**

Restoring Functional Word and Noun-Verb Syntactic Relations for Korean Compound Noun Analysis

Seong-Il Yang*, Young-Kil Kim*, Seo Young-Ae*, Eun-Jin Park*, Dong-Yul Ra**
*NLP Team, Speech/Language Technology Research Center, ETRI
**Div. of I.T., Yonsei University

요 약

한국어 문장의 구성은 명사, 동사와 같은 내용어와 조사, 어미와 같은 기능어로 크게 나눌 수 있다. 문장의 핵심적인 의미 전달은 내용어에 의해 이루어지며, 한국어 명사구의 경우 잦은 기능어의 생략으로 명사 나열에 의한 복합 명사가 발생된다. 이렇게 발생되는 복합 명사를 구성하는 단위 명사들은 일부 문장 성분을 생략시켜 발생된 것으로, 생략 성분의 복원에 의해 본래의 문장 형태를 추정할 수 있다. 한국어 복합 명사의 경우, 생략되는 문장 성분은 대부분 접사, 조사와 같은 기능어로 국한되며, 기능어의 복원은 단위 명사 간의 격 관계와 의미 관계를 분석하여 이루어질 수 있다. 본 논문에서는 단위 명사간의 보-술 관계를 이용하여 복합 명사를 구성하는 단위 명사 간의 의존 관계를 추정하고, 추정된 의존 관계에 의해 생략된 격조사와 용언화 접사를 복원하는 방법을 제안한다. 구조 분석에서 사용되는 의미 격틀에 의해 결정되는 격 관계는 격조사와 용언화 접사의 복원을 결정하며, 올바른 본래의 문장 표현 복원을 위해 관형격 조사와 관형격 어미를 비롯한 특별한 형태의 복원은 통계 정보와 휴리스틱 규칙으로 결정한다.

1. 서론

한국어는 굴절어의 특성을 갖고 있어, 기능어가 발달하였으며, 이로 인해 문장 성분의 대부분을 조사, 접사와 같은 기능어와 명사가 결합된 형태가 차지한다. 이렇게 발생하는 명사의 나열은 기능어나 연결 구문의 생략으로 문장 내 복합 명사의 사용이 빈번하도록 만든다. 아울러, 부분 자유 어순과 자유로운 띄어쓰기의 사용은 한국어 분석의 어려움을 가중 시킨다. 따라서, 한국어 복합 명사의 처리는 매우 중요하며, 나열되는 명사의 처리를 위해 기반 명사구 분석, 복합 명사 분해, 명사구 묶음등 명사 처리를 위한 많은 연구가 진행되고 있다.

자연어 처리에서 명사가 나열된 복합 명사의 분석은 단위 명사간 관계를 명사-명사간의 관계로 간주하고 두 명사간의 어휘, 혹은 의미적 공기 관계를 대상으로 분석하여 왔다. 그러나, 단위 명사를 서술성과 비서술성 명사로 구분하여 명사-명사의 공기 관계를 명사-동사 공기 관계로 변환하여 분석하고자 하는 방법이 시도되고 있다.

한국어 복합 명사의 경우, 내포된 의미는 생략된 기능어를 복원하여 더 명확히 나타낼 수 있다. 따라서, 복합 명사 분석에 의해 생략된 문장 성분 정보를 복원하여 문장으로 표현 되었을 때, 동일한 의미를 갖는 경우를 복합 명사의 분석이 성공한 것으로 간주할 경우, 분석 결과는 좀더 분명해질 수 있다.

다음은 이렇게 복원된 복합 명사로 생성되는 문장의 예이다.

공무원 입시 지망생 나이 제한 제도 철폐
→ 공무원(의) 입시(를) 지망(하는) (지망)생(의) 나이(를) 제한(하는) 제도(를) 철폐(하다)

위 예는 두개의 문장이 합쳐져 하나의 복합 명사가 생성된 경우이다. 각 단문은 관형격 어미 “는”으로 연결되어 있으며, 단문 표시는 밑줄로 구분되어 있다.

본 논문에서는 구조 분석에 사용되는 의미 격틀 정보를 사용하여, 복합 명사를 구성하는 단위 명사간의 관계를 서술성 명사와 비서술성 명사간의 보-술 관계에 의한 명사-동사 격 결정 문제로 치환하여 복합 명사 분석을 시도한다. 격 결정이 완료되는 경우, 서술성 명사를 중심어로 하는 단문으로 간주하고, 각 격 정보를 채우는 명사에 해당 격조사를 복원하고, 서술성 명사의 용언화 접사를 복원하여 문장 복원을 시도

한다. 이렇게 생성된 단문들과 격 결정에 실패한 명사들은 휴리스틱 규칙에 의해 관형격을 나타내는 기능어가 복원된다.

2. 의미 제약 조건

명사-동사간의 의미 제약 조건으로 사용되는 구문 구조 분석을 위한 의미 격틀은 서술성 명사와 비서술성 명사간의 격관계 결정을 위해 사용될 수 있으며, 이러게 결정된 격 관계는 올바른 격조사 생성과 단문 분할의 판단을 제공한다. 본 연구에서는 약 300 개로 분류된 의미 코드를 사용하는 한국어 동사구 의미 패턴을 사용한다. 다음은 의미 제약 조건으로 사용되는 동사구 의미 패턴의 예이다.

```
{조직}!가 {정치활동}!에 참여!하다
{사람}!가 {경제활동}!에 참여!하다
{사람}!가 {조직}!에 참여!하다
```

그림 1. 한국어 동사구 의미 패턴

대괄호 "{", "}"로 묶인 부분은 명사의 의미 코드를 나타내며, 구분자 "!"로 구분되는 격조사를 함께 나타낸다. 패턴에 기술되는 격조사는 주격은 "가", 목적격은 "를"과 같이 표기되도록 대표형을 지정하여 사용한다. 서술성 명사는 용언의 형태로 패턴의 오른쪽 마지막에 기술된다.

복합 명사를 구성하는 단위명사는 어휘 사전에 서술성과 비서술성으로 구분되는 품사 정보와 동사구 의미 패턴의 격 정보를 비교하기 위한 의미 코드를 등록하여 사용한다.

3. 보-술 관계 격 정보 결정

단위 명사 중 서술성 명사를 기준으로 비서술성 명사와의 격관계 결정을 위해 의존 관계 구문 분석기를 사용한다. 구문 구조 분석은 동사구 의미 패턴의 격 정보를 채운 단위 명사를 제외한 나머지 명사들을 상대로 의미 제약 조건을 적용하게 된다. 격 정보 결정을 위한 대상 단위 명사는, 서술성 명사가 관형절을 이끌 수 있으므로, 서술성 명사의 앞에 나타나는 비서술성 명사들과 바로 뒤에 나타나는 1 개의 단위명사를 비교하여 격 정보를 결정한다. 이때 격 정보 결정을 위한 의미 제약 조건의 가중치는 아래와 같은 수식에 의해 결정된다.

$$CRD_{vp}(w_{1,n}, v) = \arg\max_{s_{1,n}} P(s_{1,n}, w_{1,n}, v)$$

$$= \arg\max_{s_{1,n}} \sum_{i=1}^{n} P(s_i \mid w_i) * P(s_i, v \mid w_i)$$

$$\approx \arg\max_{s_{1,n}} \sum_{i=1}^{n} P(s_i \mid w_i) * P(v \mid s_i)$$

서술성 명사는 v, 격 정보 대상이 되는 비서술성 명사와 해당 의미코드는 $w_{1,n}$과 $s_{1,n}$으로 나타낸다.

4. 기능어 복원 휴리스틱

보-술 관계 격 정보 결정에 따른 격조사와 용언화 접사의 복원에 대해, 일부 복원이 불가능한 비서술성 명사의 나열은 기본적으로 관형격 조사 "의"로 연결되는 것을 가정할 수 있다. 아울러, 뒤에 비서술성 명사를 격 관계로 갖는 서술성 명사는 용언화 접사와 함께 관형격 어미를 부여할 수 있다. 이러한 휴리스틱 규칙은 동작 규칙과 배제 규칙으로 나뉘어 기술되며, 대용량의 말뭉치에서 수집된 어휘 공기 정보를 사용하여 확률 가중치를 지정할 수 있다. 아래는 일부 휴리스틱 규칙을 보인다.

<표 1> 휴리스틱 복원을 위한 동작 규칙의 예

동작 규칙	동작 예
{서술성명사 + 서술성명사} → {관형격어미 + 의존명사} 복원	국회 연설 불허 → 국회(에서) 연설(하는 것을) 불허(하다)
{서술성명사 + 비서술성명사} → {관형격어미} 복원	해충 방제 구역 → 해충(을) 방제(하는) 구역

접사를 취하는 서술성 명사의 경우 "입시 지망생"에서 알 수 있듯이 인명, 지명등 의미 속성을 얻어 격 정보 결정에 사용할 수 있다. 이러한 경우, "지망(하는) (지망)생"과 같이 복원할 수 있다.

5. 실험 및 결론

신문 기사 제목에서 추출한 평균 길이 8.4 음절, 평균 단위명사의 개수 4.1 개의 일반 명사로 이루어진 복합 명사 250 개에 대한 기능어 복원에 대해, 올바른 의미의 기능어 복원에 성공한 단위 명사는 총 850 개로 약 83%의 정확률을 보였다.

이러한 복합 명사의 문장 복원은 동일한 의미의 다른 표현에 대한 패러프레이징이나 기계번역, 문장 정규화를 통한 데이터 부족 해소등에 이용될 수 있다.

향후, 더 나은 성능 향상을 위해, 수동-능동형에 따른 격 정보 변환, 접사 처리 보완과 추가 휴리스틱 규칙의 연구를 해 나갈 예정이다.

참고문헌

[1] 강승식, "한국어 복합명사 분해 알고리즘", 정보과학회논문지(B), 25(1), pp172-182, 1998

[2] 김영길, 양성일, 박상규외 6 명, "국소 구문 관계 및 의미 공기 정보에 기반한 명사 의미 모호성 해소", 제 14 회 한글 및 한국어 정보처리 학술대회, 2002

[3] 윤준태, 정의석, 송만석, "명사간 어휘 정보를 이용한 한국어 복합 명사 분석", 정보과학회논문지(B) 제 25 권 제 11 호, pp1716-1725, 1998

[4] 임수종, 이창기, 장명길, "문장구조분석을 위한 서술성 명사 복원", 한국컴퓨터종합학술대회 2005 논문집 Vol. 32, No.1(B), pp475-477.