문장구조분석을 위한 서술성 명사 복원

임수종[0] 이창기 장명길
한국전자통신연구원 음성/언어정보연구부 지식마이닝연구팀
{isj, leeck, mgjang}@etri.re.kr

Restoring a Predicative Noun to Verb for Parsing

Soojong Lim[0], Changi Lee, Myun-Gil Jang
Knowledge Mining Research Team, ETRI

요 약

본 연구는 신문기사나 백과사전 등의 문서에서 빈번히 발생하는 동사 파생 접미사와 어미가 생략된 형태의 서술성 명사를 동사로 복원하는 방법에 대한 것으로 이러한 복원은 문장구조 분석에 영향을 미친다. 기존 연구는 간단한 규칙만을 사용하지만 규칙을 사용하는 방법은 재현률에서 성능 저하를 보이기 때문에 본 연구에서는 이러한 생략 형태를 구분하여 규칙과 통계 방법을 사용하여 각각 적합한 형태에 적용하였다. 본 연구의 접근 방법은 규칙 기반에 비해 약 30%, 통계 기반에 비해 약 8%의 성능 향상을 보여서 문장 구조 분석에서는 3.6%의 성능 향상을 보였다.

1. 서론

한국어 문장을 분석하기 위한 과정 중에서 문장의 구조를 분석하기 위해서는 주로 용언을 중심으로 하여 분석을 한다. 그러나, 한국어 문장의 구조 분석에 있어서 주어 생략에 대한 연구는 많이 되고 있지만 실질적으로 문장에서 중요한 역할을 하는 용언의 복원에 대한 연구는 미미하다.

특히 자유어순인 한국어 문장 구조 분석에서 가장 적합한 의존 문법의 경우 명사-명사 의존 관계 분석과 명사-동사 분석은 특징이 다르기 때문에 형태소 분석 결과가 명사이더라도 구문적으로 용언의 역할을 하는 경우는 이러한 사실을 인지하여 문장구조 분석에 반영하는 것이 중요하다.

그러나 문장 구조 분석에 있어서 인터넷 문서 중에서도 정확한 문서를 담고 있는 백과사전이나 신문에서는 서술성 명사에 대해 용언화 접사와 어미가 생략된 형태의 표현을 사용하여 한 문장의 구조 분석을 어렵게 한다.

다음은 용언화가 가능한 명사에 용언화 접사와 어미가 생략된 예이다.

W.처칠이 이끄는 보수당에 압승, 노동당 단독 내각의 총리로 임명되었다.

'압승'의 경우 실제로는 '압승하여'라는 표현이 정확한 표현이다. 이러한 서술성 명사에 대해 처리를 하지 않을 경우에는 그림 1과 같이 '보수당에'이 '압승하다'의 논항이 되어야 하지만 용언으로 인식이 되지 않을 경우에는 '임명되었다'의 논항 후보가 되어 문장의 구조를 분석하는데 오류 가능성을 갖는다.

이러한 서술성 명사의 경우 한글로 된 신문기사와 백과사전 문서에서 발생하는데 연합뉴스와 백과사전에서 무작위로 추출한 각 50문서씩 100문서에서 출현한 용언 중에서 이러한 형태의 서술성 명사가 차지하는 비율은 0.2%에 불과하지만 의존관계에 기반한 문장구조 분석기[2]에서 서술성 명사를 복원하여 문장을 분석한 경우와 비교하여 약 3.6%정도의 성능 하락을 보였다.

서술성 명사의 복원은 다음과 같은 문장 때문에 사전만을 기반으로 하였을 경우에는 모호성이 발생하여 서술성 명사로 분류해야 하는지를 판단해야하는 발생한다.

아버지는 병용(炳鎔), 어머니는 경주 김씨이다.
병용하다[1]: 아울러 같이 씀.

핵분열 생성물, 우주선 사워, 질서·무질서의 문제,
샤워하다[1]: 소나기처럼 뿜어내리는 물로 몸을 씻는 일

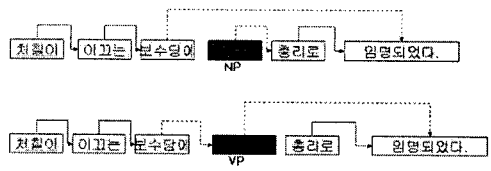

그림 1 서술성 명사 복원이 문장 구조 분석에 미치는 영향

[1] 표준국어 대사전

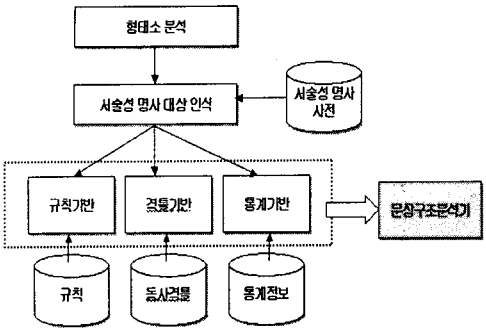

그림 2 서술성 명사 처리 흐름

이러한 문제점을 극복하기 위해서 본 연구에서는 기존 전자사전에서 용언화 접사와 부착하여 용언으로 사용이 가능한 명사 리스트와 백과사전의 용례를 이용하여 형태소 분석 결과 일반 명사로 판명된 명사 중에서 실제적으로 용언화 접사와 어미의 생략 형태를 인식하기 위한 규칙을 구축한다. 그러나 일반적으로 규칙 기반의 접근 방법의 경우에 정확한 규칙을 만들기에는 많은 노력이 필요하고 만들어진 규칙도 정확율 면에서는 효과적이지만 재현율 면에서는 성능이 떨어지는 단점이 있다. 이러한 문제점을 극복하기 위해서 본 연구에서는 규칙을 우선 적용하고 규칙으로 처리하기 힘든 유형에 대해서는 ME를 이용한 통계 기법을 사용하여 보완하고자 한다.
의존관계를 분석한 후에 술어-논항 구조를 생성하는 문장구조 분석기에서 본 연구가 차지하는 역할은 그림2와 같다.

2. 관련 연구

앞에서도 언급했듯이 주어 복원이나 기타 다른 성분 혹은 바꿔쓰기에 대한 연구는 상대적으로 활발하지만 이러한 서술성 명사를 동사로 복원하는 연구는 활발하지 못하였다.
관련 연구로는 한국어 구문 분석 정보를 이용한 정보추출 시스템[1, 3]에서 논항 정보를 이용하여 정보를 추출할 때 서술성 명사의 서술어화 처리 규칙을 사용하였다. 규칙은 단순하게 서술성 명사와 기호 ','를 사용하고 배제 규칙으로 문장의 첫 어절인 경우는 배제하였다.

3. 특성에 맞는 서술성 명사 인식

서술성 명사 인식의 경우에는 규칙기반의 방법, 격틀 기반 방법, 통계 기반 방법 등 3가지 방법을 서술성 명사로 인식된 명사의 앞 어절에서 사용된 문맥 정보를 이용하여 각자 유형에 맞는 방법을 사용하도록 하였다.

3.1 유형별 분류

신문기사와 백과사전에서 후보가 되는 용례를 추출하여 가장 좋은 방법을 적용하기 위해서 앞 어절 출현 형태별로 분류하였을 때 유형과 각각에 대한 접근 방법은 표1과 같다.

표 1 서술성 명사 유형 및 접근 방법

유형	접근 방법	예
격조사 부착 명사	규칙	이후 서울시립교향악단을 [창립,] 지휘하는 한편,
보조사 부착 명사	통계	1889년 1월 회사는 [파산,]
속격조사 부착 명사	규칙	리골레토의 [공작,]
공동격조사 부착 명사	규칙, 격틀, 통계	참주 테론과 [결탁,]
조사부착없는 명사	통계	고대오리엔트박물관 [관장,]
명사 이외	통계	로마탈취를 시도하였다가 [실패,]

보조사, 속격 조사, 조사 부착없는 명사의 경우에서 알 수 있듯이 비록 서술성 명사로 동사로 사용이 가능하더라도 실제 문장에서는 동사가 아닌 명사로 쓰인 경우가 많다.

3.2 규칙 기반 방법

용언화 명사를 복원하기 위한 규칙은 품사 정보, 정답유형(개체명) 정보, 부분 구문 분석의 기본 정보와 어절의 위치, 기호 사용 여부등의 부가 정보로 구성되며 용언화 명사 후보를 중심으로 앞 2어절을 사용하여 구축된다. 수작업으로 규칙을 추출하였으며 규칙은 [1]을 참조하여 간단하게 만들었다. 사용된 규칙은 다음과 같다.

채택 규칙
{명사+격조사} {서술성명사+, or .}

배제규칙
{명사+속격조사} {서술성명사+, or .}
문장의 첫 어절.

3.1의 유형별 분류 중에서 앞 어절이 격조사인 경우에는 실제로 동사의 역할을 하는 경우가 99% 이상이기 때문에 이러한 점을 감안하여 규칙으로 채택하였고 반대로 속격조사가 부착된 경우에는 예외없이 복원의 대상이 아니다. 그리고, 맨 앞 어절에 출현하는 경우에도 예외없이 명사 나열형의 일부이다.

3.3 격틀 기반 방법

격틀기반 방법은 규칙 기반 방법의 일종으로 격조사, 속격조사 이외에 접속 조사를 처리하기 위한 방법이다. 실제 용례를 조사해 봤을 때 격조사, 정답 유형 명사, 속격조사와는 다르게 접속 조사 다음에 출현하는 명사는 서술성 명사로 인식이 되어야 하는 경우와 인식되지 말아야 하는 경우 공존하게 되어 확실한 규칙을 만들 수는

없다. 그러나, 규칙을 만들고 규칙을 검증하는 수단으로 격틀을 채택하여 특정 서술성 명사에 대해서 해당하는 용언의 격틀을 참조하여 접속 조사를 격으로 갖을 수 있는 경우만을 인식하면 된다. 물론, 이 경우에는 용언의 격틀의 불완전성은 고려되어야 하기 때문에 격틀에 공동격 조사를 취하는 경우에만 적용하였다.

3.4 통계 기반 방법

규칙과 격틀 기반 방법으로 결정하기 힘든 문제를 통계 정보를 사용하고자 해결하고자 한다. 최대 엔트포리(Maximum Entropy, ME) 모델[5]은 주어진 제약 조건을 만족하는 여러 확률 분포 중에서 가장 균일한 분포 상태를 가지는 모델이다. 바꾸어 말하면, ME 모델은 주어진 제약 조건 하에서 최대 엔트로피를 가지는 확률 분포를 가지고 있다. 이를 수식으로 나타내면 다음과 같다.

$P = \{ \text{models consistent with constraints} \}$
$H(p) = \text{Entropy of } p,\ p \in P$
$P_{ME} = \text{argmax}_{p \in P} H(p)$

여기서 P_{ME} 가 최대 엔트로피 확률 분포를 가지는 모델이다.

ME 모델의 가장 두드러진 특징은 모델의 특성을 완전히 드러내는 후보 자질들을 선택해 주기만 하면 되는데 본 논문에서는 후보 자질로 어휘자질, 서술성 명사 사전 자질, 형태소 품사 자질, 거리 자질을 사용하였다. 특히, 서술성 명사의 복원과 같은 이진 결정을 내리는 경우에 ME 모델이 적합한 것으로 알려져 있다.

ME 모델의 매개변수 추정에 사용되는 알고리즘에는 Generalized Iterative Scaling(GIS)[6], Improved Iterative Scaling(IIS)[4], 그리고 Limited Memory BFGS(L-BFGS)[7]등 잘 알려진 것이 몇가지 있다. 본 연구에서는 GIS 알고리즘을 사용하였다.

4. 실험 및 분석

학습 데이터를 구축하기 위해서 백과사전 문서 중에서 어미가 부착되지 않은 문장을 추출하여 규칙을 만들고 규칙에 해당되지 않는 문장을 사용하여 ME 모델의 학습 데이터로 사용하였다. ME 모델의 학습 데이터는 수작업으로 265개의 긍정적인 데이터와 143개의 부정적인 데이터로 분류하였고 통계 방법의 경우 실험의 base line이 64.9%이다.

규칙과 격틀 기반 방법만을 사용한 실험, 통계 정보만을 사용한 실험, 규칙, 격틀 그리고 통계 정보를 모두 사용한 실험으로 분류하여 진행하였고 실험 데이터로는 백과사전 문서에서 무작위로 추출된 332문장을 사용하였다.

표 2 실험결과

	정확률	재현률	F-measure
규칙+격틀	86.89	46.70	60.75
통계(ME)	83.25	83.25	83.25
규칙+격틀+통계	86.38	97.80	91.74

실험의 평가 척도로는 일반적으로 잘 알려진 재현률, 정확률, 그리고 F-measure를 이용하였다.

격틀을 포함한 규칙을 사용한 실험에서는 정확률에 비해 재현률이 현저하게 낮은 성능을 보여서 규칙의 한계를 보여줬고 통계만을 사용한 경우에는 작은 규모의 학습 데이터 만으로도 base line에 비해 약 19% 정도의 성능 향상을 보였다.

규칙을 먼저 적용하고 규칙으로 판단을 내릴 수 없는 경우에 한해서 통계를 적용한 경우에는 정확률은 일정한 수준을 유지하고 재현률이 현저하게 향상되는 결과를 보였다. 규칙과 통계를 함께 사용한 경우에 통계 방법을 적용한 부분의 성능은 정확률 87.22%, 재현률 95.87%, F-measure 91.34의 성능을 보여서 규칙과 통계를 함께 사용하는 것이 서술성 명사를 동사로 복원하는 문제에서 효과가 있음을 볼 수 있다.

5. 결론 및 향후 연구방향

본 연구에서는 형태소 분석 결과는 명사이지만 문장구조에서는 동사 역할을 하여 문장 구조 분석 시에 오류를 일으키는 서술성 명사를 동사로 복원하는 문제를 해결하고자 규칙과 작은 규모의 학습 데이터를 사용하는 통계 방법을 함께 적용하였다. 실험 결과 규칙과 통계만을 사용하는 실험에 비해서 각각 30%, 8% 정도의 성능이 향상되는 것을 볼 수 있다.

이번 연구에서는 서술성 명사를 동사로 복원 여부만을 연구대상으로 하였지만 실제로 문장 구조 분석의 최종 목표결과인 정확한 술어-논항 구조를 파악하기 위해서는 복원된 동사가 실제로 -하다, -되다, -시키다 등의 다양한 용언 파생 접미사 중에서 어떤 접미사가 생략된 것인지를 파악하는 연구를 진행할 계획이다.

6. 참고 문헌

[1] 유혜원, "한국어 구문 분석 정보를 이용한 정보추출 시스템 -신문 기사문을 중심으로-", 제 29 차 한국어학회 전국학술대회, 2003 년 8 월..

[2] 임수종, 정의석, 장명길, "백과사전 질의응답을 위한 격틀 기반 의존 관계 분석", 제 16 회 한글 언어 인지 학술대회, 2004 년 10 월, pp.167-172.

[3] 차준경, "서술성 명사의 의미부류 설정", 제 24 차 한국어학회 전국학술대회, 2002 년 2 월.

[4] Berger, A., "The Improved Iterative Scaling Algorithm: A Gentle Introduction", School of Computer Science Camegie Mellon University, December, 1997.

[5] Berger, A., Della Pietra, S. and Della Pietra, V., "A maximum entropy approach to natural language processing", Computational Linguistics, 22(1):39-71, 1996.

[6] Darroch, J. and Ratchli, D., "Generalized Iterative Scaling for Log-Liner Models", The Annal of Mathematical Statistics, Vol. 43, No.5, pp.1470-1480, 1972..

[7] Liu, D.C. and Nocedal J., "On the Limited Memory BFGS Method for Large Scale Optimization", Math. Programming, 1989.