양상에 따른 자연스러운 주격 조사의 선정

이 강천 서 정연
서강대학교 전산학과

The Selection of a Subject Case Auxiliary Word According to Modality in Korean Generation

Kangchun Lee Jungyun Seo
Department of Computer Science
Sogang University

요 약

한국어 생성기의 성능은 여러 가지 요소로 평가될 수 있다. 속도, 생성 문장의 복잡성 등 여러 가지 요소가 평가 대상이 될 수 있다. 그 중에서 가장 중요한 요소로 평가될 수 있는 것은 생성되는 문장이 얼마나 자연스러운 것인가 하는 것이다. 자연스러움의 정도는 정확히 측정할 수 없지만 그 중에서 어절의 순서 배치, 대응되는 정확한 어휘의 선정, 조사, 어미 등의 적절한 선정을 들 수 있다. 본 논문에서는 특정한 양상을 술어가 가질 때 주격조사의 선정에 주안점을 두었다. 기존의 생성기[1,3,7,9]에서는 대표격 조사 '가(무종성)'나 '이(유종성)'를 사용하였는데 양상을 동반할 때에는 '는(무종성)'이나 '은(유종성)'을 사용하는 것이 더 자연스럽다는 것을 보이도록 하겠다.

1. 서론

한국어 생성기의 성능은 여러 가지 요소로 평가될 수 있다. 예로 속도가 빠르다던지, 아니면 더 복잡한 문장을 생성할 수 있다던지 하는 것들이다. 그 중에서도 가장 중요한 요소로 평가될 수 있는 것은 얼마나 자연스러운 문장을 생성하느냐이다. 생성은 분석과는 달리 모든 가능한 것들을 처리하는 것이 아니라 여러 개 중에서 가장 적당한 것을 생성하여야 한다. 여기서 가장 적당한 것이라고 할 수 있는 것은 의미에 맞는 적절한 단어(어휘)라든지 격에 어울리는 조사, 문맥의 흐름에 맞는 어미들이다. 그 중에서도 양상에 따른 주격조사의 선정에 본 연구의 초점을 맞추려 한다.

한국어에서는 대표 주격조사로 '이(유종성)'나 '가(무종성)'가 쓰인다[2,6,8]. 즉 주격의 생성시에는 격조사를 '가'나 '이'를 사용하게 된다. 그러나 특정 양상(modal)이 들어간 문장에서는 주격조사 '이'나 '가'보다는 '은'이나 '는'을 쓰는 경우가 더 자연스러운 때가 있다. 아래의 예문을 보자.

a. 철수가 과자를 먹는다.
b. 철수는 과자를 먹는다.
c. 철수가 과자를 먹을 수 있다.
d. 철수는 과자를 먹을 수 있다.

위의 예문에서 보듯 a와 b는 일반적인 문장이다. a와 b를 비교해 보면 a가 더 자연스러우며 b보다 a를 쓰는데 아무런 문제가 없는 것 같다. 그러나 '가능(possible)'이라는 양상[5,7]이 들어간 c와 d를 보면 c보다는 d가 더 자연스러움을 느낄 수 있다. 즉 양상이 첨가됨으로써 주격조사가 변화하는 것이 필요한 경우가 있다. 본 논문에서는 이와 같이 주격조사를 변화시키는 것이 더 자연스런 양상에 대하여 조사를 하였다.

2. 본론

2.1 학습 데이터의 구성.

격조사가 변화하는 것을 알아보기 위해 사용하는 데이터는 서강 대학교 말뭉치에서 추려낸 것을 사용하였다. 아래의 [표 1]은 학습 데이터의 구성을 보여준다.

	어절	문장	구성비
소 설	30446	269628	47.7%
교 과 서	17979	175963	28.1%
신문, 잡지	6881	82174	10.8%
설명문, 논설문	8568	115258	13.4%
합 계	63874	643023	100%

[표 1] 학습 데이터의 구성

2.2 격조사의 변화를 요구하는 양상.

본 연구에서는 31개의 양상[5] 중에서 격조사의

변화가 필요할 것 같은 11개의 양상을 뽑았다. 11개의 양상은 아래의 [표 2]에 나와 있다.

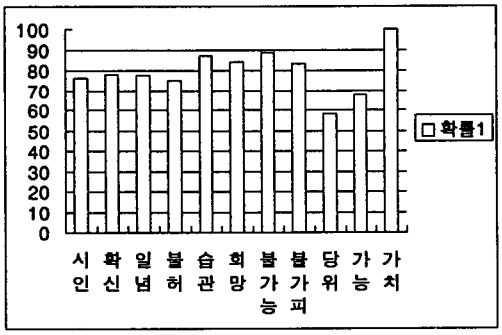
양상 명	특성 명	보조 용언
확신	certainty	기 마련이
불가피	inevitableness	ㄹ 수 밖에 없
가치	worthy	ㄹ 만하
희망	hope	고 싶
가능	possibility	ㄹ 수 있
습관	habit	곤 하
불허	disapproval	서는 안 되
불가능	impossiblity	ㄹ 수 없이
시인	approval	기는 하
당위	need	어야만 하
일념	concentration	ㄹ 뿐이

[표 2] 주격조사의 변화를 요구하는 양상

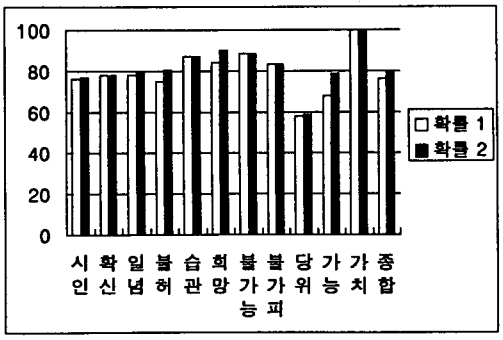
즉 위의 양상이 들어간 문장의 주격조사는 '이'나 '가'보다 '은'이나 '는'을 쓰는 것이 더 자연스러움을 알 수 있다. [그림 1]은 11개의 양상에서 주격조사로 '은', '는'이 사용된 확률을 나타낸다.

2.3 사용하는 휴리스틱

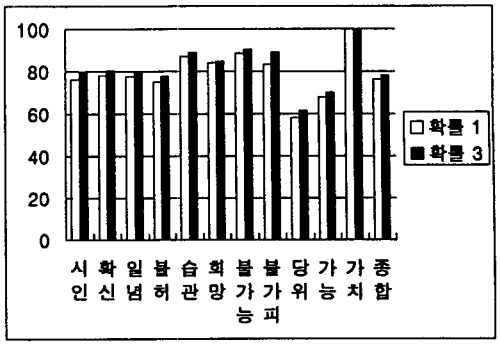
위의 2.2에서 알 수 있듯이 11개의 양상에서 주격조사로 '은', '는'이 사용되는 경우는 76.3%로 나타났다. 그러나 문장을 분석하여 보면 더욱 향상된 결론을 얻을 수 있는 것으로 나타났다.



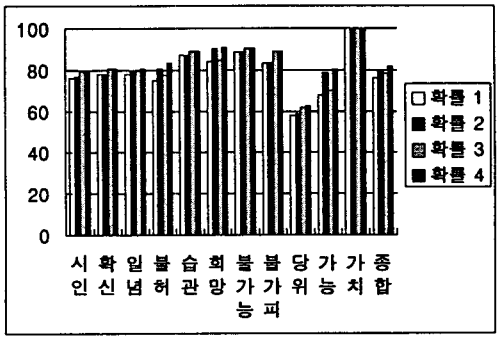
[그림 1] '은', '는'이 주격조사로 사용된 경우



[그림 2] 조사가 단어에 종속되는 경우



[그림 3] 주변에 보조사 '은', '는'이 사용된 경우



[그림 4] 휴리스틱 1, 2를 동시에 사용한 경우

2.3.1 조사가 단어에 종속되는 경우

다음의 문장을 보자.

e. 나는 단지 비둘기를 원할 뿐 다른 먹이를 원하지 않습니다.
f. 그러나 내가 듣고 싶은 것은 한 소녀의 소식이었습니다.

위의 예문 e, f에서 밑줄을 그은 부분이 양상을 나타내는 부분이고 이텔릭 글자체는 양상의 주격조사를 나타낸다. 그러나 위의 주격조사는 양상에 따라 변하는 것이 아니라 단어인 '나'와 '내'에 항상 종속이 된다. 즉 '나' 다음에는 항상 '는'이 쓰이며 '내' 다음에는 항상 '가'가 주격조사로 쓰인다[8]. 이것은 항상 최우선이 되며 양상이나 그 밖의 것에 우선한다.

[그림 2]는 이것을 보여주며 확률 1은 [그림 1]의 확률 1과 같고 확률 2는 조사가 단어에 종속될 때 자연스러운 조사를 선정할 확률이다. 이것을 사용하면 4.3%의 향상을 보인다.

2.3.2 주격조사의 좌우에 보조사 '은', '는'이 쓰일 때

보조사 '은', '는'이 주격조사의 좌우에서 사용이 되면 '은'이나 '는'은 연속적으로 쓰이지 않기 때문에 주격조사는 '이'나 '가'를 사용한다. 다음의 예문을 살펴 보자.

g. 북한에서는 개인이 자유로운 경제활동을 할 수 없다.
g'. 북한에서는 개인은 자유로운 경제활동을 할 수 없다.
h. 우리가 왜 뒤쪽은 청소할 수 없는지 의아했다.
h'. 우리는 왜 뒤쪽은 청소할 수 없는지 의아했다.

위의 예문에서 볼 수 있듯이 g, h가 g', h'보다 더 자연스러운 문장임을 알 수 있다. 이것은 앞에서 사용했던 조사가 단어에 종속되는 것보다는 낮은 우선 순위를 갖는다.

[그림 3]은 이것을 보여주며 확률 3은 이 휴리스틱을 사용했을 때 자연스러운 조사 선정 확률이다. 이것은 2.4% 정도의 성능향상을 가져온다. 또 [그림 4]는 휴리스틱 1과 2를 동시에 사용한 확률을 나타내며 7.3%의 성능 향상을 가져온다.

2.3.3 양상에 따른 휴리스틱.

위의 2.3.1과 2.3.2에서 사용한 휴리스틱은 모든 양상에 공통적으로 사용되는 것인데 비해 여기서 사용하는 양상은 특정 양상에 대하여만 쓰인다. 예를 들어 '가능'이라는 양상을 가진 다음의 예문을 보자.

i. 사람이 어찌 밥만 먹고 살 수 있겠는가?
i'. 사람은 어찌 밥만 먹고 살 수 있겠는가?
j. 사람은 밥만 먹고 살 수 있다.
j'. 사람이 밥만 먹고 살 수 있다.

위의 문장에서 i와 j가 i', j'보다 더 자연스러움을 알 수 있다. 실험결과 92.3%의 경우에서 의문문일 경우에 주격조사 '이'나 '가'를 사용한다.

하나의 예를 더 들어보자.

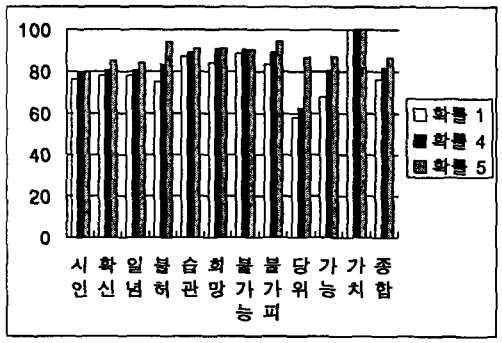
k. 사람다워지기 위해서는 끈기 있는 노력이 있어야 한다.
k'. 사람다워지기 위해서는 끈기 있는 노력은 있어야 한다.

위의 예문에서 볼 수 있듯이 양상 필요(~어야 하)는 술어 '있'과 결합할 때에는 87.4%의 경우에 있어서 주격 조사 '이', '가'가 사용되었다.

위의 양상 '가능', '필요'이외에도 '확신', '일념', '불허', 습관', '불가피', '당위'의 양상에서 휴리스틱을 사용하면 자연스러운 조사를 선정하는데 도움이 된다.

이 휴리스틱은 위의 휴리스틱 1, 2보다는 낮은 우선 순위를 가지며 가장 나중에 적용되어야 한다.

아래의 [그림 5]에서 확률 1은 학습 데이터에서 뽑은 '은', '는'은 확률이고 확률 4는 휴리스틱 1, 2를 동시에 적용한 경우이고 확률 5는 휴리스틱 1, 2, 3을 동시에 적용한 것을 보여준다. 이것을 사용함으로써 확률 4보다는 6.1% 확률 1보다는 13.5%의 성능 향상을 가져왔다.

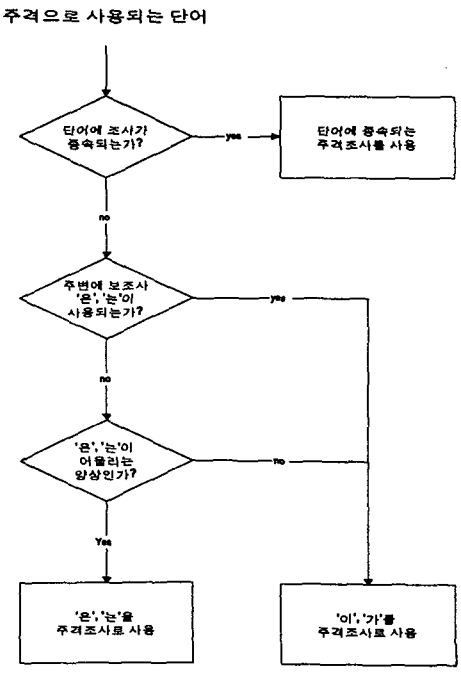


[그림 5] 각각의 양상에 대하여 휴리스틱 3을 적용했을 경우

2.3.4 휴리스틱 적용 순서.

위에서 사용한 휴리스틱에는 우선 순위가 존재한다. 즉 2.3.1이 가장 우선 순위가 높고 2.3.3이 우선 순위가 가장 낮다. 우선 순위에 의해 휴리스틱을 적용하여야 한다.

아래의 그림은 주격조사를 얻는 과정에 대한 흐름도(flow chart)이다.



[그림 6] 휴리스틱의 적용 흐름도

2.4 실험 데이터의 구성

실험은 휴리스틱을 뽑아내기 위한 학습데이타와 휴리스틱이 잘 적용되었는지를 테스트하는 실험데이타로 구성되어 있으며 학습데이타와 실험데이타는 서로 겹치지 않는다. 아래의 [표 3]은 실험 데이터의 구성을 보여준다.

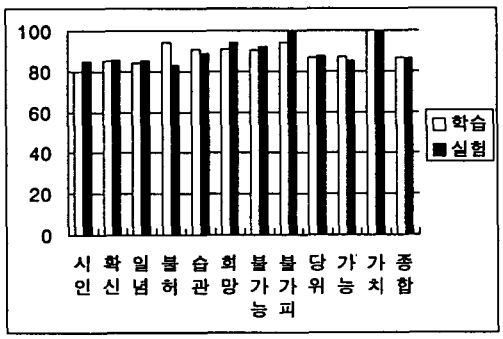
	어절	문장	구성비
소 설	9684	79522	35.0%
신문, 잡지	6527	69198	23.6%
수 필	11470	119667	41.4%
합 계	27681	268387	100%

[표 3] 실험 데이터의 구성

2.5 실험 결과 및 분석

위 2.4에서 구성된 실험 데이터로 2.3에서 사용한 방법을 이용하여 자연스러운 조사를 가진 경우를 통계를 내어 보았더니 아래의 [그림 7]과 같이 나왔다.

학습 결과와 실험 결과 모두 위에서 사용한 방법에 의해 도출한 결론으로 학습은 86.9% 정도의 성공률을 보였고 실험은 86.3%의 결과를 보였다. 학습과 실험에서 비슷하게 좋은 결과가 나왔음을 알 수 있다. 그리고 다른 종류의 데이터 보다 교과서에서 더 나은 결과가 나왔다. 이것은 다른 신문, 소설보다는 교과서가 더 정확하게 쓰여진 것에서 기인한다고 볼 수 있고 이것에 기인한다면 더 좋은 성공률을 보일 수 있을 것이다.



[그림 7] 학습 데이터와 실험 데이터의 비교

3. 결론

한국어 생성기의 가장 중요한 성능은 얼마나 자연스러운 문장을 생성하느냐에 달렸다. 그 중에서 본 논문에서는 양상에 따른 주격조사를 어떻게 선정하느냐에 초점을 맞추었다. 실험 결과에 의하면 약 87% 정도의 경우에서 좀 더 자연스러운 주격조사를 선정하는 것으로 나타났다.

생성은 여러 가지 중에서 가장 적절한 어휘, 조사, 어미를 선정하여야 하는 만큼 앞으로 연구되어야 할 것이 많은 것같다.

참고 문헌

[1] 권일재, 송만석. "표현기술 언어를 이용한 한국어 생성에 관한 연구", 제 7회 한글 및 한국어 정보처리 학술대회, 1995

[2] 김영희, "한국어의 격문법 연구", 석사학위논문, 연세대학교, 1973

[3] 김우영, 이호석, 김영택. "기계 번역에서 언어 생성기 구현에 관한 연구. 제 1회 기계번역 WORKSHOP 발표 논문집, 1989.

[4] 김재훈, "중간언어방식을 이용한 기계번역에서의 한국어 격조사 생성을 위한 한국어 격틀 설정", 석사학위논문, 한국과학기술원, 1988.

[5] 안동언. "Corpus를 기반으로 하는 한국어 술어의 양상 생성". 박사학위논문, 한국과학기술원, 1995

[6] 안동언, 최기선. "한국어 격이동 패턴". 인지과학 추계학술발표 논문집, 1990

[7] 이강천, 이상호, 서정연. "의미 중심어 주도 방식에 기반한 한국어 생성시스템", 제 23회 한국정보과학회 학술발표 논문집, 1996

[8] 조규빈. "하이라이트 고교문법", 지학사. 1986

[9] 한국과학기술원. "영한 기계번역 시스템 개발(2)", 1990.