

표현기술언어를 이용한 한국어 생성에 관한 연구

○
권일재, 송만석

연세대학교 전산학과, 한국어정보처리 연구실

A Study on the Generation of Korean using Description Directed Control

Iljae Kwon, Mansuk Song

Korean Information Processing Lab., Dept. of Computer Science, Yonsei Univ.

요약

자연스러운 문장을 생성하는 것은 자연어 생성에서 중요하다. 자연스러운 문장은 개념과 화자의 의도에 의해서 이루어진다. 따라서, 화자의 의도를 반영할 수 있는 한국어 생성 시스템의 설계가 필요하다. 본 논문에서는 언어 현상을 바탕으로 얻은 사람의 발화 모델에 대하여 살펴보고, 설정한 한국어의 기본격에 대한 무표어순을 고찰한 후, 이를 바탕으로 화자의 의도를 반영할 수 있는 생성 시스템을 설계한다. 그리고, 이 시스템에서 몇 가지 사람의 언어 행위가 재현되는 과정을 보인다.

1. 서론

자연어 생성이란 추상적인 생각을 구조 및 형태를 갖는 말로 옮겨 내는 작업이다. 생성되는 문장들이 문맥이나 화자의 의도에 의해 영향을 받고 있음은 심리언어학적 관찰을 통해 지지를 받고 있다[1]. 자연스러운 문장을 얻기 위해서는 이러한 문맥과 화자의 의도를 적용할 수 있는 생성 시스템의 설계가 필요하다.

생성 시스템은 문장 구성 전략을 세우는 단계, 대상 언어의 특성에 맞도록 구조를 만들어 내는 단계, 단어의 형태소적인 특성 등에 따라 변형, 합성하는 단계의 3단계로 이루어진다[2]. 첫번째 단계에서는 문장 구성 전략으로써 문맥에 대한 분석과 화자의 의도가 결정되어야 하고, 두번째 단계에서는 앞에서 정해진 문맥의 분석과 화자의 의도를 반영하여 문장의 구조를 구성하여야 한다. 본 논문에서는 두번째 단계를 위한 설계를 논의의 대상으로 한다.

한편, 사람은 항상 문법적인 말만을 유창하게 하지는 않는다. 이는 사람이 머리 속에서 문장 전체를 완성한 후에 발화하는 것이 아니라 부분적인 개념들에 대해 점

진적으로 발화를 해 나가는 증거가 된다. 따라서 점진적인 언어 생성을 하도록 생성 시스템을 설계해야 한다.

기존의 한국어 생성에 관한 연구는 주로 문장의 올바른 구조 및 조사의 설정에 관한 것이었고[3,4,5], 문맥 구조를 반영한 생성의 연구는 드물었다[6].

본 논문에서는 D. McDonald의 표현기술언어에 의한 생성 제어 모델[7]을 한국어 생성 시스템 설계에 도입한다.

본 논문의 구성은 다음과 같다. 2장에서 표현기술언어에 의한 생성 제어에 대하여 소개하고, 3장에서는 한국어 격문법 연구에 기반한 의미 구조의 격관계에 대한 고찰을 하며, 4장에서는 생성 시스템의 구성 및 작동에 대해 설명을 하고, 5장에서는 결론을 맺는다.

2. 표현기술언어에 의한 생성 제어 (Description Directed Control)

D. McDonald의 표현기술언어에 의한 생성 제어 모델은 <그림 1>과 같은 실현 명세(realization specification)를 통해 생성 과정을 제어한다. <그림 1>에서 'house3', 'red' 등은 개념이고, 'color-of', 'part-of' 등은 개념들의 관계를 나

타낸다. 4 번째 실현 명세의 'condense-on-property'는 화자의 의도를 반영하는 수사학적 표지로, 전체 문형 구조를 정하는 단서가 된다.

```
(r-spec1
  (r-spec2 color-of (door3 red))
  (r-spec3 part-of (door3 house1))
  (r-spec4 condense-on-property (r-spec2 r-spec5 red))
  (r-spec5 color-of (gate4 red))
  (r-spec6 part-of (gate4 fence2)))
```

그림 1: 실현 명세(realization specification)의 예

이 모델은 자료 중심 제어(Data Directed Control)개념을 확대한 것이다. 자료 중심 제어란 고정된 하나의 프로그램이 존재하는 것이 아니고, 상대적으로 작은 모듈로 나뉘어 그곳에 지식이 분산되어 있는 제어 구조이다. 이 모델에서는 관계 표지 및 화자의 의도 표지마다 의사 결정 모듈(decision procedure)이 정의되어, 실현 명세에 기술된 순서에 따라 깊이우선방식(depth-first-realization)으로 생성을 제어한다.

이 모델의 구조 및 제어는 다음과 같은 사람의 발화 행위와의 유사성을 띤다.

첫째, 사람은 점차적으로 말을 만들어 나간다. 이는 간투사 '에...' 등을 삽입하는 사람의 언어 행위를 그 근거로 삼을 수 있다[1].

둘째, 한번 발화가 된 후에 다시 발화 내용을 고칠 수 없다. 즉, 발화된 이후에 발생하는 제약 조건에 대해서 이전에 발화된 내용은 수정될 수 없다.

셋째, 사람은 한번에 전체 개념의 일부만을 발화 대상으로 삼는다. 즉, 자신이 문장을 어떻게 끝낼 지 미리 결정하지 않은 상태에서 발화를 시작한다.

이 모델의 깊이우선방식 제어로 문장의 왼쪽 부분(이미 생성된 부분)에서의 생성이 오른쪽 부분의 생성(앞으로 생성될 부분)에 영향을 끼치게 된다. 대명사가 왼편에 의존하거나, 주어가 동사에 앞서 나타나면서 동사의 형태 변화에 영향을 주는 등의 실제 언어 현상은 이 모델의 타당성을 뒷받침해 준다.

또한, 이 모델에 의하면 왼쪽 부분에서 이루어져야 할 제약 조건이 명확히 설정되지 않아 의사결정이 이루어지는 시점에서 충분한 판단 근거가 부족하거나, 왼쪽 부분에서 이루어진 제약 조건들 사이에 위배되는 조건들이 발생했을 경우에 사람이 말실수를 하는 것으로 설명할 수 있다.

이 모델에서는 실현 명세에 쓰이는 개념들 간의 관계 표지의 종류에 대해서는 언급하고 있지 않으나, 시스템의 구현을 위해서는 관계의 설정이 있어야 한다. 따라서 한국어 격문법 연구를 토대로 기본격을 설정하고, 한국어 어순에 관한 연구들을 토대로 화자의 의도가 첨가되지 않았을 경우의 어순, 즉 무표어순을 설정한다.

3. 기본격과 무표어순

3.1 기본격

자연스럽고 보다 정확한 자연어 생성을 위해서는 개념 간의 의미 관계가 명확히 규정되어야 한다. Fillmore에 의해 격문법이 제안된 이후, 국어학자들에 의해 격분류가 활발히 진행되었다[8,9]. 그러나, 학자들에 따라 약간씩의 차이를 보이고 있다. 본 논문에서는 기존의 연구를 기본으로 하여 본 생성 시스템에서 사용할 기본격들을 다음과 같이 설정하였다.

- 행위자격(AGNT), 경험격(EXPR), 도구격(INST), 수여격(DATV), 대상격(OBJ), 공간_상태시원격(SSRC), 시간시원격(TSRC), 공간_상태도달격(SGOL), 시간도달격(TGOL), 처소격(LOC), 시간격(TIME), 거리격(DIST), 기간격(DUL), 경로격(PATH), 주격자격격(SQL), 목적격자격격(OQL)

3.2 무표어순

같은 의미 표현 구조라도 생성되면서 화자의 의도가 첨가되면 조사에서뿐 아니라 어순에서도 변화가 일어난다. 다음 예에서 어순이 다름으로 인해 강조하는 바가 다름을 우리는 쉽게 찾아 볼 수 있다.

- (1) 그들은 단오날에 광한루에서 만났다.
- (2) 단오날에 그들은 광한루에서 만났다.

한국어의 어순에 관해서 [10]은 사격성(obliqueness)위계를 바탕으로 하위범주화 항목들에 대한 선형 순서가 있음을 주장하고 있으며, [9]는 기본 문장에서 문장 성분들의 순서에 대해 고찰하고 있다. 본 논문에서는 이 결과들을 토대로, 강조 표지가 들어가지 않은 문장에서의 개념관계들의 선후 관계를 얻어냈다<그림 2>.

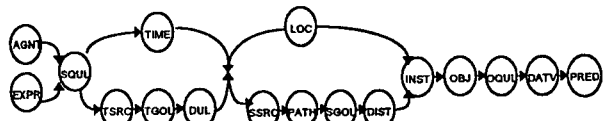


그림 2: 기본격관계들의 선후관계

4. 생성 시스템 설계

4.1 시스템의 구성

본 논문에서는 개념 그래프[11]에 의한 개념 구조의 표현과 화자의 의도를 나타내는 의도 표지를 입력으로 받는다.

시스템의 내부 구성도는 <그림 3>과 같다.

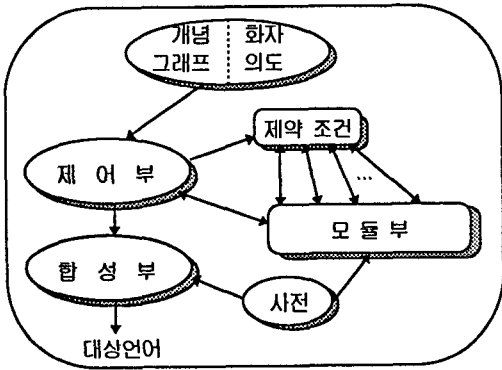


그림 3: 시스템의 내부 구성도

제약 조건이란 이후에 생성될 부분에 대한 제약으로, 생성 도중에 발생한다. 그 쓰임은 4.3에서 자세히 보겠다.

4.2 무표어순에서의 작동

3.2에서 논의한 무표어순에 대한 고찰 결과를 적용하여 문장을 생성하게 되는데, 그 기본적인 알고리즘은 <그림 4>와 같다.

```

main() {
    개념 그래프 입력;
    의도 표지 입력;
    개념 생성( PRED, 시작 중심 개념);
}
개념 생성(판제, 개념) {
    if (지배 개념 존재)
        for(지배 개념[i])
            개념 생성( DECO, 지배 개념[i]);
    if (의존 개념 존재) {
        무표 어순화 정렬;
        판제 의도 표지 반영;
        /* 어순 변화, 생성의 제약 조건 설정 */
        for (의존 개념[i])
            개념 생성(의존 개념[i]에 대한 판제, 의존 개념[i]);
    }
    "판제"와 관련된 의사 결정 모듈(제약 조건, 개념);
}
    
```

그림 4: 생성 알고리즘

위의 알고리즘에 의하여 <그림 5>의 입력에 대해 <그림 6>과 같은 과정을 거쳐 "철수가 어제 큰 방을 청소하

는 동생을 도왔다."를 생성하게 된다.

```

[ 돕다 ]-(AGNT)>> [ 철수 ]
-(OBJ)>> [ 동생 ]<-(AGNT)-[ 청소하다 ]
[ 어제 ]<-(TIME)-
[ 크다 ]<-(ATTR)-[ 방 ]<-(OBJ)-
()
    
```

그림 5: 입력 예 1

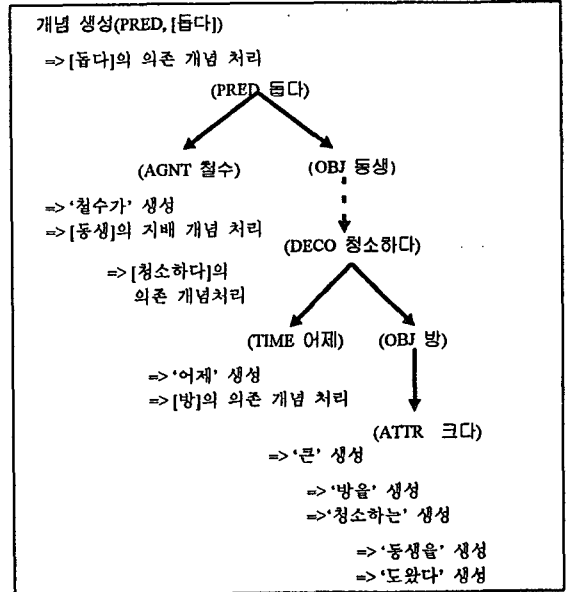


그림 6: 생성 과정

4.3 의도 표지

4.3.1 주제화

어순에 관한 연구에 따르면 주제화되는 문장 성분은 다른 보어들에 선행한다[10]. <그림 5>와 같은 입력에서 개념 [동생]을 주제화하기 위해서는, 입력에 (주제화 [동생])이라는 의도 표지를 첨가하면 된다. 이 경우, 무표어순에 의한 개념의 배열에 변형이 가해져서, 개념 [동생]의 생성 순서가 가장 우선하게 된다. 그 결과, "어제 큰 방을 청소하는 동생을 철수가 도왔다."라는 문장을 생성한다.

4.3.2 수동태

<그림 7>과 같은 개념 그래프는 다음과 같이 생성된다.

(3)영희가 주먹으로 철수를 때렸다.

여기에 (수동태)라는 의도 표지를 첨가하면, 수동태 의사 결정 모듈은 제약 조건에 '수동태' 표지를 등록하고, 대상격의 위치를 가장 앞으로 옮긴다. 이후 순서대로 각

[때리다]-(AGNT)-> [영희]
 -(INST)-> [주먹]
 -(OBJ)-> [철수]

그림 7: 문장(3)의 개념 그래프

각의 의사 결정 모듈들이 이 표지를 참조하여 상황에 맞는 조사를 써서 문장 (4)를 생성한다. PRED 의사 결정 모듈에서는 '때리다'의 반대말 '맞다'를 사용한다.

(4) 철수가 영희에게 주먹으로 맞았다.

4.3.3 높임말

다음과 같은 두 문장을 생각해 보자.

(5) 할머니가 밥을 먹는다.

(6) 할머니께서 진지를 잡수신다.

여기서 일률적으로 문장 (5)가 잘못 났다고 말할 수는 없다. 높임말은 말하는 이와 듣는 이와 관계에 따라 쓰이기 때문이다[12]. 따라서, 높임말에 관한 설정은 화자의 의도로써 입력되어야 한다. <그림 8>과 같은 입력에 대해 <그림 9>와 같은 과정을 통해 문장 (6)을 생성한다.

[먹다]-(AGNT)-> [할머니]
 -(OBJ)-> [밥]
 ((주체높임 [할머니]))

그림 8: 입력 예 2

주체높임 의사 결정 모듈 :
 개념 [할머니]에 '높임' 표지 부착.
 AGNT 의사 결정 모듈 :
 주격조사의 높임말을 써서 '할머니께서'를 생성.
 제약 조건에 '주격 높임'을 등록.
 OBJ 의사 결정 모듈 :
 제약 조건 중의 '주격 높임'을 참조.
 '밥'의 높임말을 골라 '진지를'을 생성.
 PRED 의사 결정 모듈 :
 제약 조건 중의 '주격 높임'을 참조.
 '먹다'의 높임말인 '잡수시다'에
 시제를 고려하여 '잡수신다'를 생성.

그림 9: 화자 의도 반영 과정

5. 결론

본 논문에서는 화자의 의도를 반영할 수 있는 생성 시스템의 설계에 대하여 기술하였다. 사람의 발화 모델을 반영할 수 있는 표현기술언어에 의한 생성 관리 모델에

대하여 알아보았고, 이 모델을 한국어의 생성에 적용하기 위하여 기존의 격문법 연구를 기반으로 기본격을 설정하였으며, 화자의 의도가 들어가지 않았을 때의 어순인 무표어순에 대하여 고찰했다. 또한 본 생성 시스템에서 몇가지 사람의 언어 행위가 제한되는 과정을 보였다.

사람의 언어 행위에는 더욱 많은 의도 및 수사학적 기법이 존재한다. 이러한 의도 표지들의 설정에 관한 연구와 의사 결정 모듈 설계에 관한 연구가 앞으로 필요하다.

참고문헌

- [1] I.Taylor, M.Taylor, *Psycholinguistics*, Prentice Hall, 1990
- [2] D.McDonald, M.Vaughan, J.Pustejovsky, "Factors Contributing to Efficiency in Natural Language Generation," *Natural Language Generation* edited by G. Kempen, Martinus Nijhoff Publishers, 1987
- [3] 김재훈, "중간언어방식을 이용한 기계번역에서의 한국어 격조사 생성을 위한 한국어 격률 설정", 한국과학기술원 전산학과, 석사학위논문, 1988
- [4] 김효준, 최기선, 김길창, "의미구조로부터 한국어 내포문의 생성에 관한 연구", *한국정보과학회 봄 학술발표 논문집*, pp.241-244, 1989
- [5] 이상호, 서정연, "의미 중심어 주도 방식을 이용한 한국어 생성 시스템의 구현", *제 6 회 한글 및 한국어 정보처리 학술발표 논문집*, pp.434-438, 1994
- [6] 조영환, 김길창, "문맥 친밀도를 이용한 대화체 문장의 생성", *제 1 회 지능기술 공동학술회의 논문 및 작품요약집*, pp.106-115, 1995
- [7] D.McDonald, "Description Directed Control: Its Implications for Natural Language Generation", *Readings In Natural Language Processing* edited by Grosz et al., Morgan Kaufmann Publishers Inc., 1986
- [8] 김영희, "한국어의 격문법 연구", 연세대학교 국어국문학과, 석사학위 논문, 1973
- [9] 김승곤, *국어토씨연구*, 서광학술자료사, 1992
- [10] 신효필, "HPSG를 기초로 한 한국어 동사의 하위법 주화", *언어학연구 제 7 호*, 서울대학교 대학원 언어학과, 1990
- [11] J. Sowa, *Conceptual Structures: Information Processing in Mind and Machine*, Addison-Wesley Publishing Company, 1984
- [12] 남기심, 고영근, *표준 국어문법론*, 탑출판사, 1993