영한기계번역에서 계층적 한국어 어순 생성

°서진원†, 이신원††, 정성종†, 안동언†

† 전북대학교 컴퓨터공학과, ††정인대학교 컴퓨터정보과

{jin|swlee}@calhp1.chonbuk.ac.kr, {sjchung|duan}@moak.chonbuk.ac.kr

A Hierarchical Korean Word-order Generation in English-Korean Machine Translation

Jin-Won Seo*, Shin-Won Lee, Sung-Jong Chung, Dong-Un An

Chonbuk National University Chongin College

Dept of Computer Engineering, Dept. of Computer Imformation

요 약

본 논문에서는 영한기계번역 시스템에서 한국어 문장을 생성할 때 올바른 한국어 어순 규칙을 제안한다. 한국어 생성은 영한기계번역의 최종 단계로서 이전단계에서 얻어진 정보를 가지고 목적 언어인 한국어 문장을 만드는 곳이다. 본 논문에서 제안하는 계층적 어순 생성 규칙은 한국어 의존구조를 기본으로 하며 규칙 적용은 4가지 함수를 단계적으로 적용시킨다. 인터넷의 발달은 언어 장벽이라는 새로운 문제를 부각시켰으며 이를 위해서 기계번역은 활발히 연구가 진행되고 있는 분야이다. 한국어 문장에 대한 올바른 어순 생성 규칙은 번역 결과의 품질을 증가시키며, 기계 번역뿐만 아니라 한국어 생성을 필요로 하는 모든 시스템에 적용할 수 있다.

1. 서 론

컴퓨터로 인간의 언어를 처리하고자 하는 노력은 컴퓨터의 개발과 함께 진행되어 왔다. 자연언어처리 중에서도 기계번역은 연구된 기간에 비해서 만족할 만한 성과가 드러나지 않는 분야중의 하나이다. 최근들어 인터넷의 발달은 사용자들에게 다양하고도 방대한 정보를 제공하게 되었다. 그로 인해 사용자들은 자신에게 필요한 정보를 얻으려고 하면서 외국어라는 새로운 장벽을 만나게 되었다. 여기에 대한 대안으로 기계번역이 떠오르게 되면서 기계번역은 새로운 양상을 맞게 되었다.

하지만 국내의 상용제품들은 목적 언어라 할 수 있는 한국어의 생성에서 적지 않은 부족함을 보여왔다. 번역 대상이 되는 원시언어와 우리말 사이의 구조적인 차이 혹은 의미적인 차이를 고려하지 않은 한국어 생성 결과는 낮은 품질로 인해 일반 사용자들에게 외면을 받아왔다.

기계번역의 품질을 높이기 위해서는 모든 단계의 고른 성능 개선을 요구하지만 올바른 대역어 선정과 한국어 문법을 사용하여 문장을 생성하는 것으로도 많은 품질 향상을 얻을 수 있다. 이를 위해서 목적 언어인 한국어의 다양한 특성을 연구할 필요가 있다 또한 번역 대상 언어와 번역 목적어인 한국어의 차이점을 비교 분석해서 한국어 생성에 적용해야 한다

2. 기존의 연구

한국어 어순에 대한 연구는 많은 연구기 진행되지 않았다 반복 및 병렬을 중심으로 하는 연구가 있었으며[1], 문장의 핵심 성분과 여타 성분간의 상대적인 위치로 인하여 어순이 결정된다[2]는 연구가 있다. 하지만 이러한 연구들은 다수의 언어 사용자들이 사용하게 되면서 관용적으로 굳어진 부분만을 어순의 규칙으로 보고 있다. 더구나 국내에서 기계 번역에 초점을 두고 사용할 한국어 어순에 대한 연구는 찾기 어려운 실정이다

외국의 경우는 대부분 동양권 언어를 목적 언어로 번역하는 연구에서 어순규칙을 언급하고 있다 영어에서 터키어로 번역을 하면서 어순에 대한 논의를 한 연구[7]는 문장의 내용을 Topic과 Focus라는 두 부분으로 나누어서 어순 정렬을 한다. 새로운 문장의 Topic과 이전 문장의 Topic을 비교하여 새로운

문장의 Topic을 문장의 head로 이동하는 휴리스틱 기반으로 어순을 정렬하였다. 그러나 이 경우 휴리스틱 기반이라는 자체가 너무 모호하며, Topic을 결정해야 하는 경우 항상 이전 문장의 Topic 정보를 가지고 있어야 하는 부담이 따른다.

본 논문에서는 영한 기계번역 시스템 및 한국어 생성을 필요로 하는 모든 시스템에서 실적용이 가능한 계층적 한국어 어순 생성 규칙을 제안한다. 문장을 구조적으로 표현하는 방법은 의존구조를 사용하며, 한국어 생성 규칙에 사용되는 문법 및 용어는 남기심,고영근[3]을 사용하며, [그림 3]의 규칙이 [그림 2]에 적용됨을 보였다

3. 문장 의존 구조

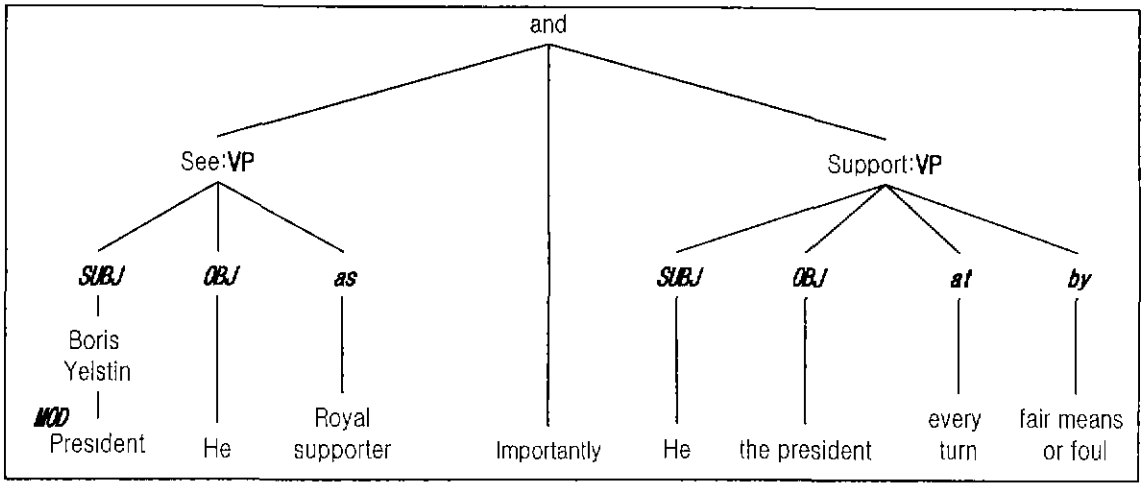
문장이란 하나 혹은 여러 개의 형태소(Morpheme)로 구성되는 단어들의 구조화된 집합이다[4]. 구조화 집합이라는 말은 일정한 규칙에 의하여 문장이 구성됨을 의미하며 의존 문법에서는 이러한 규칙을 문장 성분들 사이의 기본 관계로서 표현한다 의존문법 관련 용어는 의존문법개론[4]에서 사용하는 용어를 사용한다.

의존문법의 기본관계는 연결(Connect)이다. 모든 구성 성분들은 문장의 중심 성분을 중심으로 연결되는 구조를 가지며, 연결된 구성 성분들은 지배와 의존이라는 관계를 가지게 된다. 의존문법에서는 항상 동사가 최상위 지배 성분이 되도록 결정되어 있으므로 동사가 가지는 문장의 필수 성분을 이용하여 문장 구조를 파악한다.

본 논문에서는 다음의 예문을 통하여 한국어 어순 규칙을 계층적으로 적용하고자 한다

- importantly, Boris Yeltsin saw him as a loyal supporter, and he would support the president at every turn, by fair means or foul.

<그림 1>은 예문을 의존 구조로 표현한 것이다. 이 문장은 대등 접속사 'and'로 결합된 이어진 문장이다 의존구조에서 상위 노드와 하위 노드의 연결은 두 노드 사이의 관계를 의미한다. 이러한 관계는 연결선으로 표시하였으며, 문장의 주성분은 연결선 왼쪽에 대문자로, 부속 성분은 소문자 형태로 연결선 왼쪽에 표기하였다. 단어 옆에 있는 구/절 정보는 문장 구성 성분이다. 대문자로 표기되어 있는 주성분 다시 말해서, 동사의 필수격은 동사의 지배를 받고 있으며, 역으로 필수격들은 동사에 의존한다. 예문에서 전치사구로 구성된 부속성분들은 한국어 생성시에 부사구로 번역이 되서 용언을 수식, 한정하게 된다. 하지만 수식어의 부재는 문장의 형성에 아무런 제약을 주지 못하므로 지배 의존 관계라고 할 수는 없다. <그림 1>에서 'President of Ressia"가 'Boris Yelstin"의 지배를 받고 있지 않다. 본 논문은 수식과 피수식의 관계를 지배 의존 관계



<그림 1> 영어 의존 구조

가 아닌 지배 의존관계의 부분 집합으로 간주한다

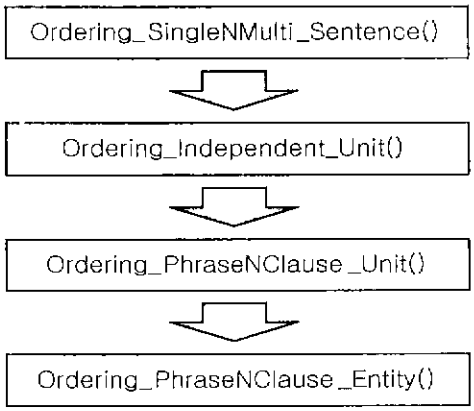
4. 계층적 한국어 어순 생성

의존 문법은 문장 구성 성분간의 모든 관계를 지배와 의존 관계로 표현한다고 하였다. 하지만 한국어 문장을 생성할 때 영어 문장에서 통사적인 지배 의존 관계를 명확하게 분석하는 것은 쉬운 일이 아니다. 또한 한국어에서 수식어와 피수식어의 관계는 지배와 의존의 관계가 아니므로 지배의존 관계만을 가지고 한국어 문장을 생성할 수 없다

이를 위해서 본 연구에서는 전체 기저 구조는 의존구조를 사용하며 생성에 사용되는 어순 규칙은 문맥자유문법의 생성규칙을 혼용하는 방법을 사용한다. 어순 생성 규칙은 문장부터 구성성분까지 단계적으로 적용된다. 한국어 문장의 기본어순은 SOV어순으로 알려져 있다[2]. 이 원리를 기본으로 하여 <그림 2>와 같은 계층적 한국어 어순 생성 규칙을 제안한다.

어순 생성 함수에서 사용하는 규칙은 생성 규칙을 그대로 사용하도록 한다 <그림 3>은 본 논문에서 사용하는 생성 규칙들이다 모든 규칙들은 좌변에서 우변으로 생성되지만 동사구(VP)는 생성 규칙이 아닌 지배관계를 의미한다. 그러므로 규칙 2, 3, 4는 의존 구조의 지배관계를 표현하는 것이다.

본 논문에서 계층적이라는 용어는 어순 규칙을 적용할 때 현 단계에 적합한 규칙을 적용하면서, 모든 규칙을 차례대로 적용한다는 의미로 사용한다. 어순 생성의 네종류의 함수들은 각 단계에서 <그림 3>의 규칙들을 적용시키면서 단어들을 정렬시키게 된다. SC는 종속문을, MC는 주문을 나타내며 I는 독립언을 나타낸다.



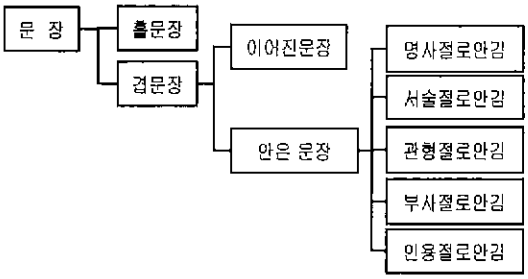
<그림 2> 한국어 계층적 어순 생성

1. S → SC MC	7. NP → NN NP
2. VP → [I] NP ADVP	8. NP → PRO NP
3. VP → [I] NP NP ADVP	9. NP → MODP NP
4. VP → [I] NP NP	10. ADJP → ADJ
5. NP → ADVP NP	11. ADVP → ADV
6. NP → ADJP NP	

<그림 3> 한국어 생성 규칙

4.1 Ordering_SingleNMulti_Sentence

이 함수는 문장이 홑문장인지 겹문장인지를 구별하여 문장간의 어순을 정렬한다. 한국어 문장을 구분하는 방법은 국어학자마다 상이하게 나뉜다. 본 논문의 한국어 생성기는 남기심,고영근[3]이 분류한 <그림 3> 방식에 의거 한국어 문장을 생성한다. 겹문장과 홑문장의 어순을 결정하는 문장 정렬은 번역 대상 문장이 대등 및 종속 접속사로 연결된 겹문장일 경우에만 정렬을 수행한다.



<그림 3> 한국어 문장 구분

```
Ordering_SingleNMulti_Sentece()
{
    if (Sentence=='홑문장')
       홑문장인 경우 다음 단계로 이동;
    elsief (Sentence == '겹문장')
       if (Sentence == '이어진 문장')
          원시언어 어순으로 정렬;
       elseif (Sentence == '안은 문장')
          안긴문장을 하나의 문장성분으로 간주;
          홑문장에서 정렬 시도
       elsief (Sentence == '주문/종속문')
          종속문+주문의 순서로 정렬;
       endif
    endif
}
```

<그림 4> 문장 정렬

Ordering_SingleNMulti_Sentence()는 생성규칙 1번을 사용한다. <그림 1>의 예문은 이어진 문장이므로 이 규칙이 적용되지 않는다.

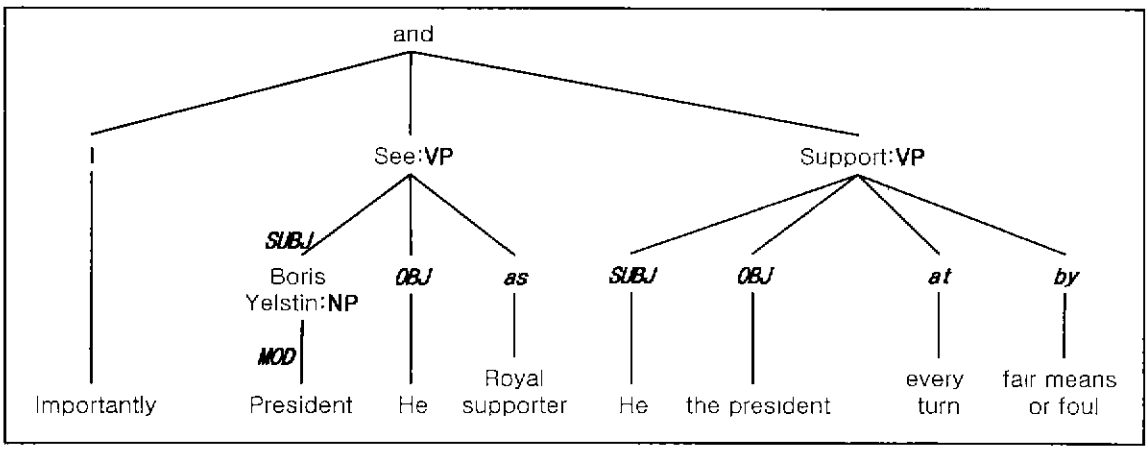
4.2 Ordering_Independent_Unit

독립어는 문장의 어떤 성분과도 직접적 관계가 없는 독립된 성분이다[3]. 주로 감탄사, 호격 등이 있다. 그리고 독립어는 아니지만 비슷한 속성의 문장부사가 있다. 문장부사는 문장의 특정 성분을 꾸며주는 것이 아닌 문장 전체를 꾸미는 성분으로서 감탄사 및 호격과 마찬가지로 문장의 처음에 나타나는 경향을 가진다. 본 논문에서는 어순 처리의 일관성을 위해 문장 부사를 독립언과 동일한 성분으로 간주하고 어순을 정렬한다. 문장 부사 외에 접속 부사 등도 동일한 방식으로 처리한다.

```
Ordering_Indepedent_Unit( )
{
   if (word=={'독립언'|| '문장부사'|'접속부사'})
    switch word
        case '독립언' .
        case '문장부사' .
        case '접속부사' .
            의존구조에서 노드의 최좌측으로 이동;
   endif
}
```

<그림 5> 독립언 정렬

독립언 정렬을 위한 Ordering_Independent_Unit()함수는 생성 규칙 2, 3, 4번을 사용한다



<그림 6> 독립언 정렬 후 의존구조

2, 3, 4번의 우변은 의존 구조에서 동사의 지배를 받는 성분들을 의미한다. <그림 1>의 예문에서 문장부사라 할 수 있는 성분은 ADVP의 'Importantly'가 있다. 이 함수의 적용 후 단어 'Importantly'는 의존 트리의 가장 왼쪽으로 이동이 되며 <그림 6>과 같은 결과를 얻게 된다.

4.3 Ordering_PhraseNClause_Unit

홑문장부터 이 규칙이 적용된다. 이 규칙은 한국어 어순의 기저구조인 SOV형식으로 문장의 어순을 정렬한다. 대등적, 이어진 문장인 경우에는 양쪽 문장에 본 규칙부터 적용한다.

이 단계에서는 문장 구성 성분을 품사별로 정렬하지 않는다. 우선 여러 구성 성분들을 구/절단위로 결합 시킨 후에 구/절 단위로 어순을 정렬하며 한국어 문장 어순을 SOV형식으로 정렬한다.

품사별 정렬은 다음 함수 Ordering_PhraseNClause_Entity 에서 수행한다. 어순 정렬에 앞서서 문장 성분을 구/절 형태로 결합시켜야 한다. 구/절의 결합은 문장 주성분을 중심으로 진행되며 생성 규칙 5, 6, 7, 8, 9을 이용하여 규칙의 역으로 결합한다.

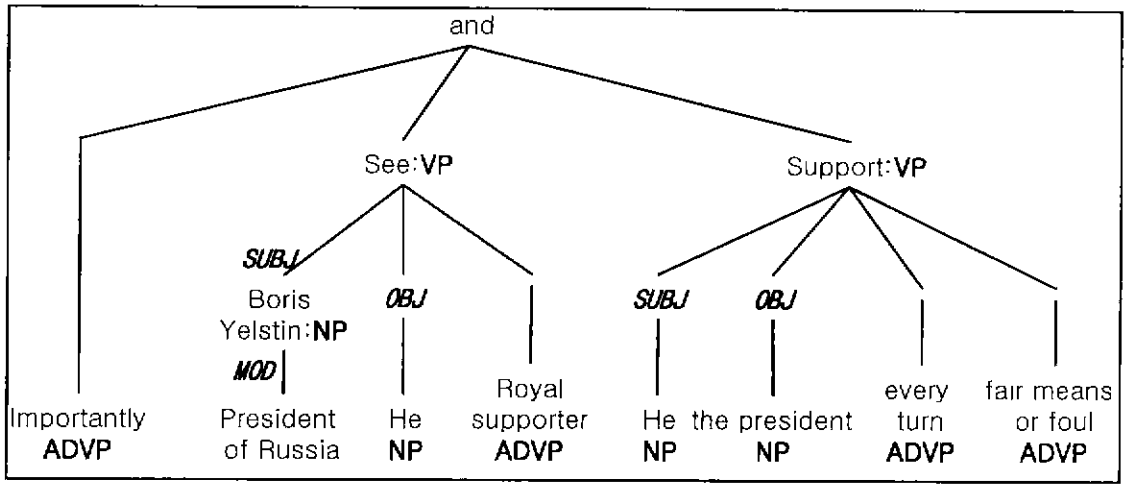
```
Ordering_PhraseNClause_Unit( )
{
    //구,절단위 결합 선행
Union_PhraseNClause_Unit( );
if (CASE=='SUBJ')
  Move_Node_Left();
elseif(CASE=='ADVP' && Not Sentence_ADVP)
  Move_Node_Right(),
elseif(CASE=='OBJ')
//목적격은 주어와 동사 사이에 위치
  Move_NODE();
else
// 연급하지 않은 격의 노드 이동
  Move_Node_Sub();
}
Union_PhraseNClause_Unit( )
{
  생성 규칙 3,4,5,6,7을 이용하여 결합;
}
```

<그림 7> 구/절 단위 어순 정렬

영어에서 전치사구는 한국어에서 부사구로 번역되므로 부속 성분으로 결합시킨다 Order_PhraseNClause_Unit 적용후에 <그림 1>의 구조는 <그림 8>로 변경된다.

4.4 Ordering_PhraseNClause_Entity()

구와 절을 이루는 성분별 정렬은 생성 규칙에서 이미 보여주고 있다. 한국어에서 수식어는 피수식어 앞에 오는게 일반적



<그림 8> 문장 구성 성분 결합

인 현상이다. 본 연구에서는 두 겹이상의 관형사구나 부사구는 고려하지 않는다.

이상과 같이 계층적 어순 정렬 함수가 모두 적용되면 영어 의존구조는 한국어 의존구조로 구조 변환되고 한글 대역어로 변환하면 <그림9>의 결과를 얻게 되며, 얻어진 한국어 의존구조를 DFS로 순회하여 한국어로 생성을 한다.

5. 결 론

본 논문에서는 영한 기계번역 시스템이나 한국어 문장 생성 시스템에서 적용 가능한 한국어 계층적 어순 생성 규칙을 제안하였다 여기에서 사용하는 계층적이란 규칙들이 단계적으로 적용됨을 의미한다.

한국어 생성은 기계번역 부분의 제일 마지막에 해당하는 단계이다. 그러므로 원시언어보다는 목적언어에 맞게 개발이 이루어져야 할 것이다. 또한 각 단계에 맞는 규칙을 제시 및 적용함으로서 시스템의 효율을 높일 수 있다.

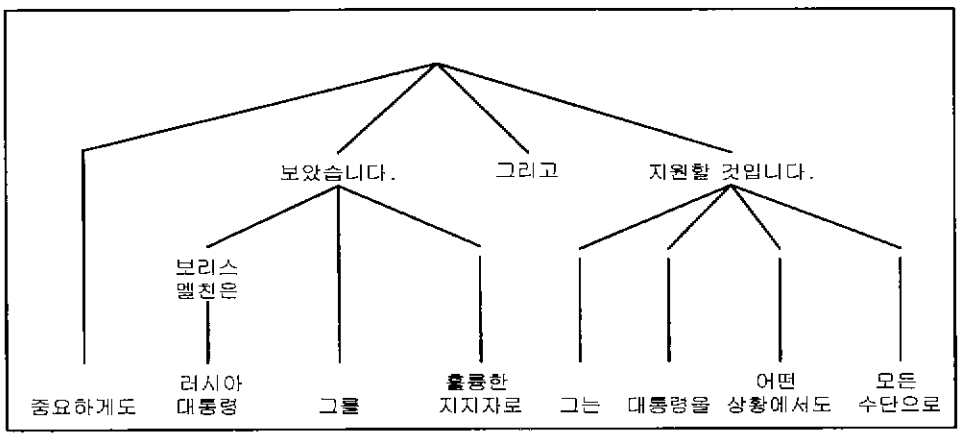
본 연구에서 제시한 계층적 한국어 어순 규칙의 가장 큰 장점은 원시언어에 의존적이지 않고 한국어 생성을 필요로 하는 모든 시스템에 적용이 가능하다는 점이다.

본 논문에서 실시하지 못한 평가는 고교 영어 문법서 및 영한 대역 문고 문장을 이용하여 평가를 시도하고자 한다.

본 논문에서 제안한 어순 생성 규칙은 문장 성분들을 정렬하고 있다. 앞으로의 연구 방향은 본 연구에서 다루지 못한 관형어나 부사어가 두 겹 이상 출현하는 경우나 용언의 보조용언 등의 결합 순서 등을 보완해야 하며 더 나아가, 격 프레임의 확장을 통한 문장 성분간의 의미적인 조응도 살펴보려 한다.

6. 참고문헌

1. 국어 어순의 연구: 반복 및 병렬을 중심으로, 채완, 1986, 탑출판사
2. 국어 어순 연구, 김승렬, 1990, 한신문화사
3. 표준 국어 문법론, 남기심-고영근,1998, 탑출판사
4. 의존문법개론, 이정출, 1993, 한신문화사
5. 자연언어처리, 황도삼 외, 1998, 홍릉과학출판사
6. 자연언어이해, 황도삼 외, 1999, 홍릉과학출판사
7. Translation into Free Word Order Language, Beryl Hoffman, 1996, University of Edinburgh
8. Separation Surface Order and Syntactic Relations In a Dependency Grammar, Norbert, Broeker, 1998, University of Stuttgart



<그림 9> 변환된 한국어 의존 구조