産業科學技術硏究所 論文集 第十三卷 第一號
(J. of Ind. Sci. and Tech. Institute)
Vol. 13, No. 1, 1999. 7.

대화체 기계번역을 위한 중심어 기반 한국어 분석

정천영* 임희동** 서영훈**
*구미1대학 전자계산과 **충북대학교 컴퓨터공학과

Korean Analysis Based on Core Words for Spoken Language Machine Translation

Jung, Cheon Young* Lim, Hee Dong** and Seo, Young Hoon**
*Kumi College **Chungbuk National Univ.

Abstract

This paper describes Korean analyzer based on core words for the machine translation of spoken language. Because the grammar based on core words can describe grammar regardless of word order, it can solve the partially-free word order feature of Korean and some problems occurring when parsing spoken language containing unnecessary syllables such as meaningless words, repetitions speech, etc. Also, because unnecessary syllables are removed by the grammar based on core words, we can analyze spoken language robustly without the mechanism such as skip of syllables. The data used for the experiment are scheduling domain of ETRI, and the parsing performance is 99.0% for trained utterances, 88.1% for untrained utterances.

1. 서 론

자연언어 처리 분야의 핵심은 자연언어의 분석 및 생성이다. 영어의 경우 수십년 동안 축적된 기초자료와 기반 기술을 바탕으로 다양한 응용소프트웨어와 시스템 소프트웨어에서 활용되고 있다. 그러나 한국어 정보처리는 불과 십여년 동안에 자료 부족 및 기초연구 미흡 등 어려운 속에서 연구가 진행되어 왔으며, 상대적으로 선진국에 비해 분석기술이 뒤떨어져 왔다. 특히 한국어의 교착어 특성과 비정형 언어라는 특성은 형태소 분석과 구문분석 기술의 발전 기술을 더디게 하는 가장 큰 요인이 되고 있다[1].

대화체 문장은 대화체에서 갖는 단어의 축약이나 탈락, 조사의 생략, 수정 또는 반복 발화, 간투어 등의 특성으로 형태소 분석이나 구문분석 등이 현재의 자연언어 처리를 위한 문법이나 파싱 기법으로 처리하기에는 상당한 문제점이 있다[2,3]. 대화체의 이러한 특성으로 이를 처리하기 위한 여러 가지 새로운 방법이 시도되고 있는데 단어간 확률정보를 이용하는 확장된 문맥 자유 문법을 이용하거나[4], 구문정보를 전혀 고려하지 않은 개념기반의 시스템을 구성하거나[5], 구문과 의미정보를 이용하되 기존의 자연언어 분

석 기법을 자연발화 처리에 적합하도록 변형하여 강건한 특성을 포함시킨 기법[6] 등이 대표적인 예이다. 대화체 처리의 어려움으로 문어체에 비해 대화체에 대한 연구가 미흡한 편이며, 특히 한국어 대화체는 연구가 활성화되고 있지 않은 상태이다.

부분 자유어순 특성을 가지는 한국어를 CFG(Context Free Grammar) 형태의 문법으로 기술할 경우 문법이 방대해지고 대화체의 특징으로 나타나는 불필요한 성분을 처리해야 하므로 파서의 부담이 커진다. 또한 기존의 개념기반 기법은 불필요한 개념으로 인한 파싱의 오버헤드와 한국어 부분자유어순 특성을 고려하지 않아 문법이 방대해지는 문제점이 있고, 핵심개념 기반 기법은 한국어의 조사를 고려하지 않고 문법을 기술하기 때문에 정확한 생성결과를 얻기 어려우며, 예제나 패턴을 이용한 기법은 패턴수가 방대해지고 패턴이 존재하지 않으면 실패하는 문제점이 있다.

따라서 본 논문에서는 말뭉치로부터 서술어가 나타나는 문장을 분석하여 서술어와 결합되는 문장에서 중요한 의미를 갖는 어절을 추출하여 중심어 문법을 기술하였다. 서술어에 따라 선택하는 중심어는 품사정보 뿐만 아니라 중심 어절에 따라 의미 Tag를 부여하고, 부분 자유어순 특성을 고려하여 문법을 작성하는데 문장의 형태, 양상 정보, 비종결어미 정보와 서술어에 따른 하위범주를 묶어 문법을 정리하였다.

의미 Tag는 응용분야나 의미해석이 요구되는 정밀도에 따라서 그 복잡도나 의미 Tag Set의 분류가 달라질 수 있다[7]. 본 연구에서는 말뭉치를 분석하여 많이 출현하는 명사의 의미속성을 추출하고 이들 속성에 명사의 의미 Tag를 설정하도록 하였다.

이렇게 구축된 중심어 기반 문법은 한국어의 부분 자유 어순 특성을 해결할 뿐만 아니라 간투어나 비문법적인 문장 등 대화체의 특성을 해결하는데 도움을 준다.

또한 대화체 분석은 강건한 분석이 요구되는데 대화체 분석에 불필요한 성분이 중심어 문법으로부터 제거되기 때문에 어절의 skip 등과 같은 별도의 메커니즘을 설계할 필요없이 강건한 분석이 가능하고, 불필요 성분을 처리하기 위한 파서의 부담을 줄였으며 비교적 단순한 문법만으로 대화체 분석이 가능하다.

2. 중심어 기반 문법

2.1 조사의 모호성

한국어는 격조사에 의해 구문적 역할이 결정되는 특징을 갖고 있으며, 특히 하나의 부사격 조사가 여러 가지 의미를 가지는 특징을 갖고 있다. 한국어 격조사는 크게 주격, 서술격, 목적격, 보격, 관형격, 부사격, 호격으로 나눌 수 있는데 한영 기계번역에서 문제가 되는 격조사는 부사격 조사이다.

한영 기계번역에서 조사와 어미의 번역 어휘를 선택하는 일은 그들의 의미를 결정하는 것과는 다른 문제이다. 주어진 문맥에 맞는 번역 어휘를 선택하여 목표 언어로 적절한 변환을 수행하는 것이 기계번역의 과제인 것이다. 원시문장에서 조사나 어미의 의미적 역할이 동일하다 하더라도 번역어가 갖는 용법과 어휘 특성에 따라 서로 다른 번역을 하여야 되는 경우가 있다[8].

영어는 문장 성분의 위치와 전치사구의 역할에 의해서 격이 결정되므로 한국어의 주격 조사나 목적격 조사는 영어로 번역할 때 큰 문제가 되지 않는다. 그러나 부사격 조사는 그 의미와 일치하는 전치사로 번역되어야만 정확한 의미 전달이 가능하다.

예를 들어,

(1) 연필로 쓰다.
(1′) write with a pencil.
(2) 산이 바다로 변하다.
(2′) Mountains turn to sea.

문장 (1)과 (2)에서 부사격 조사'로'는 동사가 '쓰다'인 경우 '수단'의 의미로 사용되었고, '변하다'인 경우에는 '변화'의 의미로 사용되었다. 이러한 의미 변화는 영어로 번역할 때 전치사 'with'와 'to'로 각각 번역된다. 따라서 한국어 조사는 동사에 의해서 영어의 전치사가 결정되므로 같은 문법에 의해서 번역을 할 수 없다.

(3) 나는 서울로 간다.
(3′) I go to Seoul.
(4) 나는 버스로 간다.
(4′) I go by bus.

문장 (3)과 (4)는 동사의 형태 및 의미가 같고 부사격 조사 '로'가 사용되었지만 영어로 번역될 때는 (3')와 (4')와 같이 각각 다른 전치사 to와 by로 다르게 번역되는데 이것은 조사 '로' 앞의 어휘의 의미로 발생된 것이다. 따라서 문장 (3)과 (4)를 같은 문법에 적용할 수 없으므로 조사 앞의 어휘의 의미를 밝히는 것이 필요하다.

또한 같은 구문 구조를 가지는 문장에서 명사의 의미에 따라 격이 결정되는 경우가 있으며 영어로 번역할 때 의미가 달라진다.

예를 들어
(5) 철수는 영희와 사과를 먹었다.
(5') Chulsu and Younghee ate the apple.
(6) 철수는 딸기와 사과를 먹었다.
(6') Chulsu ate the berry and the apple.

문장 (5)와 (6)는 같은 구문 구조를 보이는 형태로 문장 (5)의 영희를 문장 (6)에서 딸기로만 대체하였다. 그러나 영어로 번역될 때 문장 (5)에서 영희는 주어로, 문장 (6)에서 딸기는 목적어로 번역되기 때문에 문장 (5)와 (6)를 같은 문법에 적용할 수 없다.

이와 같이 어휘 모호성 및 구문 모호성은 문법을 구성할 때 품사, 구문정보, 개념정보만으로 문법을 구성 할 때 발생한다. 이러한 모호성을 해결하기 위하여 명사에 조사정보를 포함하는 의미자질을 도입하여 동사에 따라 문법을 구성한다.

문장 (1)과 (2)는 동사 문법에 의해 전치사를 결정할 수 있고, 문장 (3)과 (4)는 조사 '로' 앞의 명사 의미에 따라 전치사를 결정할 수 있으며, 문장 (5)와 (6)는 문법을 구성할 때 의미 자질이 같을 때에만 결합되도록 함으로써 모호성 해결이 가능하다. 따라서 한국어를 분석할 때 조사도 함께 고려되어야 정확한 생성 결과를 얻을 수 있다.

2.2 문법의 구성

문법을 구성할 때 문법의 수를 줄여 효율적으로 관리할 수 있는 방안이 요구되고, 문장을 분석할 때 구조적 모호성을 줄이고 번역할 때 정확한 의미전달이 가능하도록 구성하여야 하는데 명사의 의미자질을 도입하여 이러한 문제점의 해결이 가능하다. 의미자질은 실용적이면서 광범위하고 타당한 자료를 바탕으로 하여 객관적으로 설정되어야 한다. 명사의 의미를 분류하는 연구는 미국 Princeton 대학에서 개발한 언어 심리학적 원리에 기반을 둔 online 데이터베이스인 WordNet, 인간 생활에 전반적으로 통용되는 일반 상식을 대규모의 지식베이스로 구축한 CYC, 일영/영일 기계번역을 위해서 일본 정부 기관과 관련 업체가 참여하여 개발한 EDR 등이 대표적이다. 현재까지 한국어 처리에 대한 대부분의 연구는 주로 형태소 분석과 구문분석의 중의성 해소에 초점을 맞추어 진행되어 왔으며, 의미 해석시 중의성을 해소하려는 연구[[9,10]]가 진행되고 있으나 구문 분석시의 중의성 해소와 같은 특수 목적용으로 응용분야가 제한적이고, WordNet을 한국어 어휘에 적용하려는 시도가 연구되고 있고 의미격을 분류하거나 신경회로망을 이용한 중의성 해소 방법들이 연구되고 있다[7]. 그러나 이러한 연구는 연구실 수준의 실험적인 연구로 응용분야나 의미해석의 정밀도에 따라 분류가 달라질 수 있다.

명사의 의미 분류는 사전적 의미와 특정 영역에서 사용되는 대화체와 달라질 수 있으며 어떻게 응용하느냐에 따라 분류의 깊이도 달라질 수 있다. 따라서 본 연구에서 실험 대상으로 하는 말뭉치인 '여행안내' 영역 총 152개 대화인 1,607 문장을 조사하여 의미별로 많이 출현하는 단어를 추출하고, 이들 단어에 의미자질을 부여하였다. 즉, 여행안내 영역은 장소와 시간 등에 많은 의미가 있고 의미전달에 중요한 요소가 되기 때문에 이를 고려하여 도메인에서 나타나는 빈도가 많은 자질을 중심으로 <표 1>과 같이 14가지로 구분하여 선정하였다.

모든 명사에 의미표지를 부여하는 일은 구축비용이 높고 객관적으로 검증이 어려우나 실험 도메인에서 나타나는 명사 중 의미표지 부여가 가능한 명사를 대상으로 Tag 번호를 부여하여 만들었다. 의미자질은 도메인이 확장됨에 따라 확장될 수 있으며 다른 의미의 문장이 패턴에 일치되어 목적 문장이 잘못 생성될 경우 분류된 의미표지를 더 세분하여 모호성을 해결할 수 있다.

문장 성분상 서술어는 주어, 목적어와 같은 필수 성분으로 관형어나 부사어에 비해서는 기능상의 상위를 차지한다. 한편 목적어는 서술어의 특성에 따라 필요 유무가 정해지므로 서술어와 주어가 문장에서 가장

Tag 번호	의미표지	설 명
@1	animal	동물
@2	human	사람
@3	local	장소
@4	time	날짜,시간
@5	vehicle	교통수단
@6	number	수
@7	eating	음식,식사
@8	money	화폐
@9	proper	고유명사
@10	card	신용카드
@11	name	사람 이름
@12	bound	교통수단 방향
@13	alpha	알파벳
@14	travel	여행

<표 1> 의미자질

중요한 성분이다. 그러나 주어는 생략이 일반화되어 있고, 서술어는 특별한 이유 없이는 생략되지 않으므로 서술어가 문장 내에서 기능이 더 크다. 국어는 개별언어로서는 물론 보편언어로서도 주어 중심 언어라기 보다는 서술어 중심 언어라고 결론지을 수 있다[11]. 따라서 한국어의 문장을 서술어를 중심으로 분석 기반 문법을 구성하는 것이 효율적인 파싱이 가능하다.

대화체 분석을 위한 기존의 문법은 매칭이 될 수 있는 개념이나 요소들을 나열하여 분석을 하였다. 이렇게 문법을 구성할 경우 한국어의 부분 자유어순 특성상 많은 문제점들이 발생한다.

"사박 오일 동안 친구와 미국에 갑니다"에 대한 문장을 분석할 때 동사 '가다'에 대한 문법은 한국어의 부분 자유어순 특성으로 인하여 다음과 같이 표현할 수 있다.

{사박 오일 동안} {친구와} {미국에} 갑니다
{사박 오일 동안} {미국에} {친구와} 갑니다
{친구와} {사박 오일 동안} {미국에} 갑니다
{친구와} {미국에} {사박 오일 동안} 갑니다
{미국에} {사박 오일 동안} {친구와} 갑니다
{미국에} {친구와} {사박 오일 동안} 갑니다

위와 같이 한국어의 부분 자유 어순 특성에 의하여 6가지의 다른 형태로 표현될 수 있다. 동사 '가다'앞에 나타나는 어휘의 개수는 5개로 '사박 오일 동안'은 어순이 바뀌지 않고 순서대로 나타나야 하며, '사박 오일 동안', '미국에', '친구와'의 순서가 서로 바뀌어도 대화를 하는데 있어서 의미를 전달하는데는 큰 문제가 되지 않는다. n개의 어휘로 이루어진 하나의 문장이 순서가 바뀌어 나타날 수 있는 형태의 개수는 최대 n!이 되어 문법의 수가 방대해질 뿐만 아니라 문법을 관리하는데 어려움이 많다.

이러한 어려움을 극복하기 위하여 어휘의 위치 이동이 가능하도록 문법을 작성할 필요가 있으며, 이는 위치 이동이 가능한 어휘를 집합의 한 요소로 간주함으로써 해결 가능하다. 즉, 집합내의 원소는 문장을 구성하는 단위일 뿐이고 이들간의 순서 관계는 나타내지 않는다.

본 논문에서는 서술어에 따라 중심어 기반 문법을 구성한다. 중심어는 문장에서 서술어와 결합되는 중요한 의미를 갖는 어절로 정의한다. 중심어는 말뭉치로부터 서술어 단위로 추출하였다. 중심어 문법은 의미자질, 품사, 터미널 심볼, 옵션의 집합으로 서술어 단위로 구성하는데 문법에서 사용되는 품사는 명사, 동사, 부사, 형용사, 수사로 구분하였고, 옵션은 <표 2>와 같이 8가지 기호로 표현한다.

옵션 기호에서 '*'는 성분이 나타나지 않거나 한번 나타나는 것을 의미하고, '+'는 한번 또는 그 이상 나타나는 것을 의미하며 '*+'는 '*'과 '+'을 조합한 것으로 한번도 나타나지 않거나 한번 나타나거나 한번 이상 나타나는 것을 의미한다. '[ ]'는 열거된 여러 성분 중에서 오직 하나의 성분만 선택되는 것을 의미하고 '|'는 열거된 여러 성분을 구분하는 구분자이며 '/'는 설명문으로 파싱에는 영향을 미치지 않는다. 또한 '('와 ')'는 반복되는 성분의 범위를 나타내며, '('와 ')'는 한국어의 어순이 순서대로 나타나야 할 경우 그 범위

기 호	설 명
*	zero or one
+	one or more
*+	zero or one or more
[ ]	열거된 성분 중에서 오직 하나만 선택
|	열거된 성분을 구분
/	주석
{ }	반복되는 성분의 범위
( )	어순 순서 범위

<표 2> 옵션 기호

를 나타낸다.

문법은 분석문법과 생성문법으로 구성되어 있으며 분석문법에는 구(phrase) 문법과 동사 문법으로 구분하여 작성하였다. 구 문법은 입력 문장에서 나타나는 어휘로부터 결합이 가능한 어휘를 먼저 결합함으로써 동사에 따른 문법수를 줄이고 파싱의 오버헤드를 줄이기 위함이 목적이다. 즉, 구 문법은 동사 문법의 하위문법이라 할 수 있다. 예를 들면 "사박 오일 동안 친구와 미국에 갑니다"에 대한 문장을 분석할 때 동사 '가다'에 대한 분석을 하기 이전에 '사박 오일 동안'에 대한 어절을 분석한 후 동사 '가다'에 대한 문법을 분석한다. 구문법은 '@2 와,@2 가 : #1 and #2'의 형태로 표현되는데 어절과 어절은 콤마(,)로 구분하고 구분자 :를 기준으로 좌측은 한국어 구문법이고 우측은 구에 대한 대역어를 의미한다. 한국어 구문법에서 '@'+'숫자'는 의미 Tag 번호이고, 대역어에서 '#'+'숫자'는 구문법에서 의미 Tag가 나타나는 순서를 의미한다.

분석 문법은 말뭉치로부터 추출된 226개의 각 서술어에 대하여 중심어를 추출하여 작성하였다. 분석 문법은 순서에 관계없이 문법을 기술할 수 있으므로 중심어가 문장의 어느 위치에 나타나든 관계없이 분석이 가능하므로 한국어 부분 자유 어순 특성을 해결할 수 있다. 또한 서술어가 취할 수 있는 중심어는 중심어 문법에서 선택되기 때문에 대화체 특성을 효과적으로 해결할 수 있고, 분석에 불필요한 성분은 중심어 기반 문법으로부터 제거되기 때문에 어절의 skip과 같은 별도의 메커니즘이 필요없이 대화체에서 요구되는 강건한 분석이 가능하다.

동사 '가다'에 대한 분석 문법은 [그림 1]과 같이 구성된다. 분석 문법에서 어절과 어절은 콤마(,)로 구분하고 콤마 앞뒤에는 공백이 없으며, 어근과 어미는 하나의 공백(space)으로 구분하여 기술한다.

```
  ⋮
@12 [를|을]
@12 으로
@14 [를|을]
@2 [가|이]
@2 [는|은]
@2 [를|을]
@2 로
@2 [과|와]
@3
@3 [까지|까지는]
@3 [는|은|에서는|에서는요]
@3 [과|와]
@3 [로|으로]
@3 [를|을]
@3 에
@3 에서
@4
@4,동안 에
@4,[정도 로|정도 에]
@4 [에|에는]
@4 으로
@5 [가|이]
@5 으로
  ⋮
```

[그림 1] 동사 '가다'에 대한 분석 문법

3. 중심어 기반 분석

3.1 개요

한국어 대화체 문장 파싱은 대화체에서 갖는 단어의 축약이나 탈락, 조사의 생략, 수정 또는 반복 발화,

간투어 사용 등의 특성으로 인하여 기존의 분석 기법으로 처리하기에는 상당한 문제점이 있다. 따라서 한국어 대화체 문장의 분석은 의미전달이 가장 큰 목적이므로 불필요한 성분을 제거하고 문법적으로 옳지 않은 문장을 분석하는 강건한 분석이 요구되며 한국어의 특징 중의 하나인 부분 자유 어순 특성을 고려한 방법이 요구된다.

따라서 본 논문에서는 한국어 대화체 문장의 특성을 고려하고 부분자유 어순 특성으로 인하여 문법이 방대해지는 문제점을 해결할 수 있는 중심어 기반 분석 방법을 제안한다. 중심어 기반 분석은 입력 토큰에 대하여 중심어를 문법으로부터 선택함으로써 수행한다.

분석은 입력 문장에서 출현하는 모든 서술어에 대하여 서술어 단위로 분석을 하는데 입력 토큰의 앞부분부터 서술어가 나타나는 부분까지를 분석 단위로 분석한다. 분석 후 각 서술어는 문장형태, 양상, 비종결어미, 하위범주 및 대역어에 대한 정보를 출력한다.

한국어는 첨가어(agglutinative language)이므로 조사, 어미, 조동사 등의 기능이 대단히 발달되어 있다. 이러한 기능 중에서 술어 어간에 연결되어 술어를 구성하는 보조용언, 어미 등의 굴절접사(inflectional suffix)는 많은 문법적 기능을 담당한다. 또한 화자(speaker)의 양상을 표현하는 것으로 전달하고자 하는 문의 내용을 현실과 연결시켜 나타낸다. 모든 문에는 양상이 있으며 술어를 완성하므로 양상의 정확한 해석과 생성은 기계번역의 품질을 높이는데 중요한 역할을 한다. 본 논문에서 양상은 말뭉치를 분석하여 의지(will), 상태(state), 추측(guess), 당위(must), 원인(because), 희망(want), 가정(if), 간접의문(whe), 가능(can), 불가능(cant), 예정(be_to)의 11가지를 이용하고 있다.

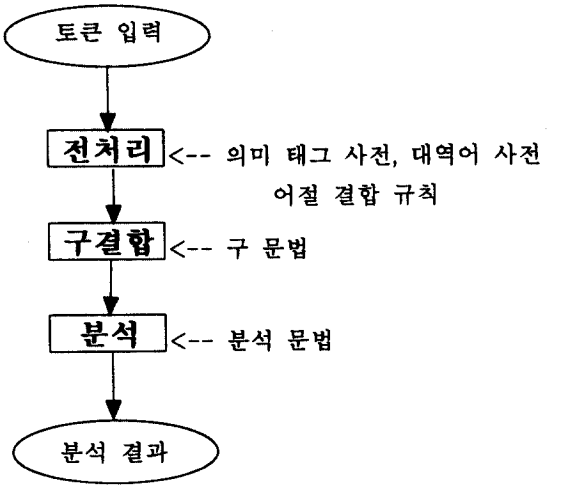
문장형태는 입력 문장의 종결 어미를 보고 판단하여 평서문, 의문문, 평서문으로 구분하는데 평서문인 경우는 decl, 의문문인 경우는 quest, 청유문인 경우는 please로 표현한다.

비종결어미는 입력 문장이 단문이 아닌 중문이나 복문인 경우 서술어 단위로 분석이 수행된 후 문장 단위로 결합을 하기 위한 정보이다. 비종결어미에 의하여 절이 문장 단위로 연결된다. 비종결 어미는 말뭉치를 분석하여 관형절, 부사절, 순접, 역접과 원인, 가정으로 분류하여 기술한다.

하위범주 정보는 입력 문장에 대하여 분석한 결과를 나타내는데 의미 자질, 품사, 터미널 심볼과 중문이나 복문의 경우 절 연결을 위한 비종결 어미의 집합으로 구성되는데 생성을 고려하여 입력어절에 대한 대역어까지 출력한다.

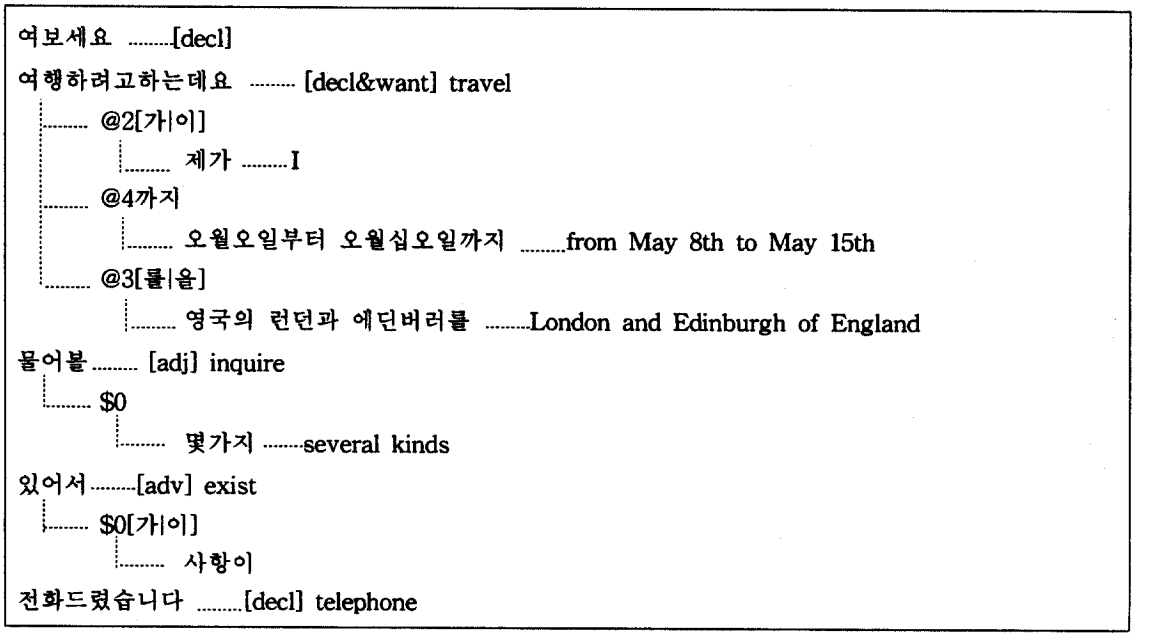
3.2 중심어 기반 분석

분석은 대화체 문장을 형태소 분석을 하여 얻은 토큰열 정보를 이용하여 수행한다. 대화체 문장은 문어체와는 달리 많은 특성을 가지고 있기 때문에 대화체 문장을 처리하기 위하여 문어체 형태소 결과를 그대로 사용할 수 없다. 따라서 문어체 형태소 결과를 대화체 특성을 고려하여 후처리한 후 그 결과를 파서의 입력으로 받는다. 현재 한글 형태소 분석기는 계속 연구 중이며
완벽한 형태소 분석을 위한 분석기는 존재하지 않기 때문에 형태소 분석 결과 중의성 및 모호성이 발생된다. 따라서 파서의 입력으로 들어오는 형태소 분석 결과는 중의성 및 모호성을 제거한 후 입력된다. 본 연구에서 사용하는 형태소 분석기는 충북대학교 자연언어처리연구실에서 보유하고 있는 문어체용 형태소 분석기를 사용하였다.



[그림 2] 분석 과정

형태소 분석 결과를 입력받아 분석을 하기 위하여



[그림 3] 분석 결과

전처리 과정을 거친다. 전처리 과정에서는 입력된 각 명사의 어절에 대하여 의미자질 사전을 이용하여 의미 태그를 부여한다. 명사의 의미 태그 부여는 의미자질 사전에 해당 명사가 있으면 의미 태그를 부여하여 분석을 수행하고, 태그 사전에 없으면 품사 태그로 대체하여 분석을 수행한다.

예를 들어 사람을 의미하는 명사(가이드, 개인, 고객 등)들은 '@2'로 태그되고, 장소를 나타내는 명사(대전, 영국, 설악산 등)들은 '@3'으로 태그된다. 태그된 어절은 생성을 고려하여 대역어 사전으로부터 대역어를 생성한다.

또한 파서의 부담을 줄이고 분석 결과의 모호성을 감소시키기 위하여 어절 결합 규칙을 이용하여 결합 가능한 어절을 결합한다. 어절 결합 규칙은 연속된 어절이 의미 태그가 같고 앞 어절이 조사를 갖지 않으면 두 어절을 하나로 묶어 새로운 하나의 어절로 결합한다. 앞 어절에 조사는 없으나 의미 태그가 다를 경우는 어절 결합 규칙을 이용하여 결합한다. 어절 결합 규칙은 의미 태그가 서로 다른 어절을 결합하기 위한 것으로 ':'로 구분하여 좌측은 결합되기 이전의 어절을 의미하고 우측은 결합된 어절로 의미 태그는 ':'우측의 의미 태그로 유지되며 두 번째 구분자 ':'의 우측은 결합된 어절을 생성하는 순서이다.

예를 들면 시간을 의미하는 의미 태그 '@4'로 태그된 두 어절 '오월'과 '십이일'은 의미 태그가 같고 앞 어절 '오월'에 조사가 없기 때문에 '오월 십이일'이라는 하나이 어절로 결합되고 의미 태그는 '@4'로 유지된다. 그러나 카드를 의미하는 의미 태그 '@10'으로 태그된 '신용카드'와 번호를 의미하는 의미 태그 '@6'으로 태그된 '번호'는 어절 결합 규칙'@10 @6 : @6 : #1 #2'를 이용하여 의미 태그가 '@6'인 새로운 어절로 생성되는데 생성 순서는 '#1 #2'로 '@10'이 앞에 생성되고 '@6'이 뒤에 생성된다.

전처리 과정을 수행한 후 구 문법을 이용하여 구 결합을 한다. 구 결합은 미리 정의된 구 문법 내에 존재하는 구를 찾아 매치되는 구를 생성하는 모든 어절들을 하나의 어절로 결합한다.

예를 들어 의미 태그가 '@3'이고 조사가 '에서'인 어절 '서울에서'와 의미 태그가 '@3'이고 조사가 '까지'인 어절 '부산까지'인 두 어절은 구 문법 '@3 에서,@3 까지 : from #1 to #2'를 이용하여 하나의 구를 생성한다. 새로운 하나의 구 결합시 의미 태그는 '@3'으로

훈련 여부	발화문 수	완전 실패	부분 실패	성 공	성공율(%)
훈련된 데이터	128		1	127	99.2
	275		2	273	99.3
	160		1	159	99.4
	149		3	146	98.0
	소계(712)		7	705	99.0
훈련되지 않은 데이터	211		26	185	88.1

<표 3> 실험 결과

유지되고 대역어는 'from 첫 번째 어절 대역어 to 두 번째 어절 대역어'인 'from Seoul to Pusan'이 된다. 구 결합을 함으로써 파서의 부담을 줄일 뿐만 아니라 입력된 어휘를 단순화함으로써 문법을 구축하는 비용을 많이 줄일 수 있는 효과가 있다.

구 결합이 끝난 후에 파서는 분석 문법을 이용하여 서술어 단위로 분석을 수행한다. 서술어 단위의 분석은 입력 토큰열의 앞부분부터 차례로 어절을 읽어들여 서술어가 나타나는 위치까지를 단위로 서술어가 취할 수 있는 어절을 취하는 것이다. 즉, 분석 문법에는 각각의 서술어가 취할 수 있는 어절들이 기술되어 있으며 이 정보를 이용하여 서술어 앞에 나타나는 어절들을 검사하여 분석을 한다. 한 문장에 여러개의 서술어가 존재할 경우 서술어 이후의 토큰부터 다음 서술어까지를 단위로 입력 토큰열이 끝날 때까지 계속하여 분석을 수행한다.

예를 들어 동사 '가다'가 취할 수 있는 어절들은 '$0[를|을]', '$0에', '@12[를|을]', '@12[로|으로]', '@2[가|이]' 등이다.

입력 발화문 "여보세요 제가 오월 팔일부터 오월 십오일까지 영국의 런던과 에딘버러를 여행하려고 하는데요 몇가지 물어볼 사항이 있어서 전화드렸습니다"에 대한 분석 과정은 다음과 같다.

파서는 형태소 분석 결과를 입력받아 분석을 수행하는데 전처리 과정에서 형태소 결합, 명사의 의미태그 부여, 대역어 변환을 한다. 전처리 과정을 거친 후 구결합을 수행하는데 구 문법을 이용하여 구 결합을 한다. 구 결합을 한 결과 대역어는 결합된 토큰열에 대한 대역어로 변환된다.

구가 결합된 후 분석을 수행한다. 분석은 발화문의 처음 입력 토큰인 '여보세요'부터 처음 나타나는 서술어인 '여보세요'까지, '여보세요' 다음 토큰인 '제가'부터 '여행하려고하는데요'까지, '몇가지'부터 '물어볼'까지, '사항이'부터 '있어서'까지, '전화드렸습니다'부터 '전화드렸습니다'까지가 각각 대상이 된다. 분석 결과는 [그림 3]과 같다.

4. 실험 및 평가

한국어는 부분 자유어순 특성을 갖는데 이러한 특성을 고려하여 CFG 등으로 문법을 기술하면 문법이 방대해 질뿐만 아니라 분석을 하는 과정에서 불필요한 성분을 만나면 분석이 실패하는 경우가 발생한다. 그러나 중심어 기반 분석은 분석 문법을 통하여 중심어를 입력 토큰으로부터 취하여 분석하기 때문에 수정발화나 부분자유어순에 대한 문제점을 효과적으로 해결할 수 있다.

또한 대화체는 분석에 불필요한 성분을 제거하여 강건한 분석이 요구되는데 분석 문법으로부터 불필요한 성분이 제거되기 때문에 입력어절의 skip과 같은 별도의 메커니즘을 설계할 필요가 없이 강건한 분석을 할 수 있다.

이와 같이 중심어 기반 한국어 분석은 한국어 부분 자유 어순 특성 및 대화체가 가지고 있는 여러 가지 특성을 효과적으로 분석할 수 있다.

본 논문에서 실험을 위하여 사용된 데이터는 ETRI corpus인 '여행 안내' 영역 중에서 712개의 훈련된 발화문과 211개의 훈련되지 않은 발화문을 대상으로 각

각 실험하였다. 훈련 방법은 일차적으로 128개의 발화문을 분석하여 문법을 만들고 실험을 위하여 712개의 발화문을 4부분으로 나누어 단계적으로 실험을 하면서 문법을 평가하고 확장하는 과정으로 훈련하였다.

실험 결과는 분석의 성공 여부에 따라 '완전 실패', '부분 실패', '성공'으로 구분하여 평가하였다. 완전 실패는 분석이 전혀 되지 않거나 잘못된 경우이고, 부분 실패는 입력 발화문 일부만이 분석이 된 경우이며, 성공은 완전 실패와 부분 실패를 제외한 결과로 올바른 분석 결과를 의미한다.

훈련된 데이터(trained data)는 발화문에 나타나는 어휘가 실험을 통하여 분석 문법에 반영되었다는 것을 의미하고, 훈련되지 않은 데이터(untrained data)는 발화문에 나타나는 어휘가 반영되지 않은 분석문법에 의하여 실험된 것을 의미한다.

실험 결과는 <표 3>과 같이 나타났는데 훈련된 데이터는 평균 99.0%의 분석 성공률을 보이는데 실패한 경우는 하나의 문장에 여러 개의 서술어가 존재할 경우, 서술어의 분석 대상이 되는 성분이 해당 서술어를 수식하지 않을 경우에 대부분 발생한다. 훈련되지 않은 데이터에서는 88.1%의 성공률을 보이는데 실패 원인은 분석 문법의 부족으로 입력 토큰의 일부가 분석되지 못하였다. 이러한 분석 실패는 분석 문법을 확장함으로써 분석 성공률을 높일 수 있다.

실험 결과에서 보듯이 중심어 기반으로 문법을 구성하였을 경우 대화체의 특성을 반영하여 분석이 가능하고, 한국어의 특성인 부분자유어순 특성을 해결할 수 있으며, 분석 문법이 비교적 단순하고, 대화체 분석을 위한 별도의 메커니즘이 필요없기 때문에 파서의 부담도 줄일수 있다. 또한 동일한 말뭉치로 실험한 결과, 개념 기반 기법(분석 성공률: 94.4%)이나 핵심개념 기반 기법(분석 성공률: 97.8%)보다 높은 분석 성공률을 얻을 수 있다.

<표 3>의 실험 결과는 형태소 분석 결과에서 중의성이 없다는 가정 하에서 수행하였다. 또한 입력 문장 자체의 의미가 애매하여 의미전달이 모호한 경우, 띄어쓰기나 인식이 잘못된 경우, 발화의 중복이 심하여 어해가 어려운 경우 등은 입력 문장을 일부 수정하여 형태소 분석을 하였다.

5. 결 론

한국어는 부분 자유 어순 특성이 있고, 대화체는 문어체와는 달리 비문법적이거나 단어의 축약이나 탈락, 조사의 생략 등의 특징이 있어 분석하는데 상당한 어려움이 있다.

부분 자유어순 특성을 가지는 한국어를 CFG(Context Free Grammar) 형태의 문법으로 기술할 경우 문법이 방대해지고 대화체의 특징으로 나타나는 불필요한 성분을 처리해야 하므로 파서의 부담이 커진다. 또한 기존의 개념기반 기법은 불필요한 개념으로 인한 파싱의 오버헤드와 한국어 부분자유어순 특성을 고려하지 않아 문법이 방대해지는 문제점이 있고, 핵심개념 기반 기법은 한국어의 조사를 고려하지 않고 문법을 기술하기 때문에 정확한 생성결과를 얻기 어려우며, 예제나 패턴을 이용한 기법은 패턴수가 방대해지고 패턴이 존재하지 않으면 실패하는 문제점이 있다.

이러한 문제점을 해결하기 위하여 본 논문에서는 문장에서 중요한 의미를 갖는 중심어를 이용하여 문법을 기술한다. 중심어 기반 문법은 순서에 관계없이 문법을 기술할 수 있으며 중심어가 문장의 어느 위치에 나타나든 관계없이 분석이 가능하므로 한국어의 부분자유 어순 특성과 대화체 특성을 중심어 기반 문법으로 효과적으로 해결하였다. 또한 중심어 기반 문법은 말뭉치를 분석하여 추출하였는데 대화체 분석에 불필요한 성분이 문법으로부터 제거되기 때문에 어절의 skip과 같은 별도의 메커니즘 없이 대화체에서 요구되는 강건한 분석을 하였다.

실험을 위하여 사용된 데이터는 여행안내 영역 중에서 712개의 훈련된 발화문과 211개의 훈련되지 않은 발화문으로, 실험 결과 훈련된 발화문은 평균 99.0%, 훈련되지 않은 발화문은 88.1%의 분석 성공율을 보였다.

앞으로 의미표지를 더욱 세분화하고 더욱 많은 말뭉치를 분석하여 중심어를 확장하면 더 높은 분석 성공률을 얻을 수 있고, 생성을 위하여 성(gender), 수(number), 일치(agreement), 대동사나 대명사 처리를 하기 위하여 문맥정보를 고려하는 방안도 생각할 수 있다.

참 고 문 헌

[1] 강승식, "한국어 정보처리의 문제점 및 방법론 고찰", 제10회 한글 및 한국어 정보처리 학술대회, 1998, pp329-334

[2] 서영훈, "음성언어 번역을 위한 개념기반의 한국어 분석 및 생성", 정보과학회 논문지 제23권 제11호, pp 1176-1184, 1996.11

[3] 최운천, 한남용, 김영성, "개념파서를 이용한 대화체 음성언어 번역", 정보처리학회 추계 학술발표논문집, 제2권 제2호, 1995

[4] Seneff, S., "TINA: A Natural Language System for Spoken Language Applications", Computational Linguistics, Vol.18, No.1, pp.61-86, 1992

[5] Levin, E., and R.Pieraccini, "Concept-based Spontaneous Speech Understanding System", Eurospeech'95, pp.555-558, 1995

[6] Levie, A., "GLR*: A Robust Grammar Focused Parser for Spontaneously Spoken Language", Doctoral Thesis, Carnegie-Mellon University, 1995

[7] 이수광 외5인, 의미속성에 기반한 한국어 명사 의미 TAG에 관한 연구, 제10회 한글 및 한국어 정보처리 학술대회, 1998. pp412-418

[8] 김나리, 김영택, "한국어 동사 패턴에 기반한 한국어 문장 분석과 한영 변환의 모호성 해결", 정보과학회 논문지 제23권 제7호, pp 766-775, 1996.7

[9] 김진현 외 3인, "용언의 구문 관계를 이용한 명사 분류", 한글 및 한국어 정보처리 학술 대회, 1997

[10] 류범모 외 4인, "구문구조부착 말뭉치를 이용한 술어의 하위범주화 정보 구축", 한글 및 한국어 정보처리 학술 대회, 1997

[11] 이관규, "국어 대등구성 연구", 서광학술자료사, 1992