대화체 번역을 위한 논항 구조에 기반한 한국어 분석

(A Korean Analysis based on Argument Structures for Spoken Language Translation)

정 천 영 † 서 영 훈 ††
(Cheon-Young Jung) (Young-Hoon Seo)

요 약 본 논문에서는 대화체 번역을 위한 논항 구조에 기반한 한국어 분석에 대하여 기술한다. 논항 구조 기반 문법은 순서에 관계없이 기술된다. 따라서 한국어의 부분 자유 어순 특성으로 문법이 방대해지는 문제점을 해결할 수 있다. 또한, 서술어가 지배하는 논항이 문법으로부터 선택됨으로서 대화체가 갖는 특성인 간투어나 중복 발화 현상 등을 효과적으로 해결할 수 있다.

실험을 위하여 사용된 데이타는 '여행 안내' 영역 중에서 1,335개의 훈련된 발화문과 420개의 훈련되지 않은 발화문이다. 실험 결과 훈련된 발화문에서는 99.7%, 훈련되지 않은 발화문에서는 93.3%의 분석 성공률을 보였다.

Abstract This paper describes a Korean analysis based on argument structures for spoken language translation. A rule in our grammar denotes only immediate dominance relation. That is, the order of components in a grammar rule is meaningless. Owing to characteristic of this grammar the our grammar size can be greatly reduced. Also, a problem of ungrammatical constituents is resolved because only arguments are described in a rule, and the parser ignores words in input utterance when they do not match in the rule. Experimental domain is travel arrangement. The total number of utterances in domain corpus is 1,755. We train a system by use of 1,335 utterances, and remain 420 utterances for test. Experimental result shows that 99.7% for trained utterances and 93.3% for untrained ones are analyzed successfully.

1. 서 론

대화체 문장은 문어체와는 달리 대부분 발화가 비문법적이고 간투어나 중복 발화 등 불필요한 성분을 많이 포함하고 있을 뿐만 아니라 단어의 축약이나 탈락, 조사의 생략 등의 특성이 있다. 또한 대화체 문장은 발화자의 의도를 모두 포함하고 있지 않을 경우도 있기 때문에 형태소 분석이나 구문 분석 등이 현재의 문어체 처리를 위한 문법이나 파싱 기법으로 처리하기에는 상당한 문제점이 있고, 발화자의 심리 상태나 발화 상황에 따라 의미의 해석이 달라 질 수 있다[1-6].

대화체의 이러한 특성으로 인하여 대화체를 처리하기 위한 여러 가지 새로운 방법이 시도되고 있는데 단어간 확률정보를 이용하는 확장된 문맥 자유 문법을 이용하거나[7], 구문 정보를 전혀 고려하지 않은 개념 기반의 시스템을 구성하거나[8], 구문과 의미 정보를 이용하되 기존의 자연 언어 분석 기법을 자연발화 처리에 적합하도록 변형하여 강건한 특성을 포함시킨 기법[9] 등이 대표적인 예이다.

기존의 개념 기반 기법[10,11]은 언어의 공통원리에 기반하여 강건성을 가장 큰 장점으로 가지기 때문에 비문법적인 요소를 많이 포함하고 있는 대화체 처리에 가장 유리한 기법중의 하나로 평가되고 있다. 1994년 ARPA ATIS 평가결과[8]는 AT&T, BBN, CMU, MIT, MITRE, SRI, UNISYS에서 개발한 음성 언어 처리 시스템 중 CMU와 AT&T에서 개발한 시스템을

† 정 회 원 : 혜천대학 컴퓨터통신계열 교수
cyjung@hcc.ac.kr
†† 종신회원 : 충북대학교 컴퓨터공학과 교수
yhseo@cbucc.chungbuk.ac.kr
논문접수 : 2000년 4월 18일
심사완료 : 2000년 12월 11일

가장 우수하면서 다른 시스템들과 상당한 격차가 있는 것으로 평가하고 있는데, 이들은 모두 개념 기반의 분석 기법을 이용하고 있다.

한편 국내에서 대화체 번역을 위한 자연언어 분석 기법에 관한 연구는 아직 미미한 실정이다. [1]에서는 개념 기반 기법으로 한국어 분석 및 생성을 하여 문법의 복잡도를 줄이고, 강건함을 강화시켰지만, 구문 정보를 전혀 이용하고 있지 않기 때문에 구문으로부터 쉽게 얻을 수 있는 최소한의 정보조차 이용하지 못하고, 불필요한 개념으로 인한 파싱의 오버헤드와 한국어 부분 자유 어순 특성을 고려하지 않아 문법이 복잡해지는 단점이 있다. 순수 개념 기반 기법으로 문법을 기술할 경우, n개의 어절로 구성된 한국어 문장은 어절의 위치 이동으로 최대 n!개의 문법을 기술해야 한다.

또한 한국어는 원시 문장에서 조사나 어미의 의미적 역할이 동일하다 하더라도 번역어가 갖는 용법과 어휘의 특성에 따라 서로 다른 번역을 하여야 되는 경우가 있다[12]. 그러나 개념 기반 기법을 적용한 한국어 대화체 분석[1,5]은 조사를 전혀 고려하지 않고 문법을 기술하기 때문에 목표언어로의 적절한 변환이 어렵다.

대화체 특징으로 나타나는 불필요한 성분이나 생략 등으로 인한 처리의 어려움과 한국어의 부분 자유 어순 특성으로 인하여 문맥 자유 문법, 패턴이나 예문 등으로 문법을 기술할 경우에 문법이 방대해진다. [3]에서는 의미 패턴에 기반하여 대화체를 분석하고 생성을 하여 비교적 높은 번역 성공률을 얻었지만 패턴이 존재하지 않으면 번역에 실패하고, 패턴수가 매우 많아진다는 단점이 있다.

따라서 본 논문에서는 한국어 대화체 번역의 성능 향상을 위하여 논항 구조에 기반한 한국어 분석 시스템을 제안한다. 논항 구조에 기반한 분석은 한국어의 부분 자유 어순 특성으로 인하여 문법의 수를 줄일 수 있고, 불필요한 성분을 처리하기 위하여 문법 규칙에 맞지 않는 성분을 건너 띄어 분석하기 위한 방식에 의한 처리방법[13,14]과 같은 별도의 메커니즘이 필요 없이 강건한 분석을 할 수 있다. 또한 문장을 분석할 때 구조적 모호성을 줄이고, 번역할 때 정확한 의미 전달을 위하여 문법을 구성할 때 조사를 포함하였다.

2. 논항 구조 기반 문법

2.1 논항 구조

문법을 구성할 때 문법의 수를 줄여 효율적으로 관리할 수 있는 방안이 요구되고, 문장을 분석할 때 구조적 모호성을 줄이고 번역할 때 정확한 의미 전달이 가능하도록 구성하여야 하는데, 논항 구조를 이용하여 이러한 문제점의 해결이 가능하다.

논항은 서술어가 지배하는 성분 중에서 중요한 의미를 갖는 어절로 정의하고 분석 문법은 논항에 의하여 작성한다. 논항 사전은 말뭉치에서 나타나는 서술어에 대한 논항을 추출하여 작성하였다. 논항은 체언과 조사로 구성되고, 체언은 의미 자질, 품사나 입력 어절 중의 하나이다.

명사의 의미 분류는 사전적 의미와 특정 영역에서 사용되는 대화체에 따라 달라질 수 있고, 어떻게 응용하느냐에 따라 분류의 깊이도 달라질 수 있다. 또한, 작성자의 주관이 많이 개입되기 때문에 신뢰성이나 객관성에 문제가 있을 수 있다. 본 연구에서 명사의 의미 자질은 실험 대상으로 하는 말뭉치인 '여행 안내' 영역 152개 대화인 1,607 문장을 조사하여 의미별로 많이 출현하는 단어를 조사하여 이들 명사의 의미 속성을 바탕으로 의미 자질을 정의하였다. 즉, '여행 안내' 영역은 장소와 시간 등에 많은 의미가 있고 의미 전달이 중요한 요소가 되기 때문에 이를 고려하여 도메인에서 나타나는 빈도가 높은 자질을 중심으로 동물, 사람, 장소, 날짜 시간, 교통수단, 숫자, 음식 식사, 화폐, 고유명사, 신용카드, 사람이름, 발착, 알파벳, 여행의 14가지로 구분하여 설정하였다.

표 1 의미 자질

의미 자질	의미 태그	예
동물	@1	원숭이, 강아지
사람	@2	가족, 아들
장소	@3	서울, 뉴욕
날짜,시간	@4	칠월, 다섯시
교통수단	@5	비행기, 기차
숫자	@6	사사삼, 이십칠
음식,식사	@7	아침식사, 부페
화폐	@8	금액, 숙박요금
고유명사	@9	고려관광, 설악호텔
신용카드	@10	비자카드, 크레디트카드
사람이름	@11	김경수, 송은영
발착	@12	부산행, 서울발
알파벳	@13	에이, 케이이
여행	@14	관광, 여행

논항의 구조는 다음과 같다.

308 [@3 를|@3 을] : #1

논항에서 왼쪽의 숫자는 논항에 대한 고유 번호이고, 구분자 :의 좌측은 한국어 구문 규칙이고, 우측은 목표언어에 대한 대역어 패턴을 의미한다. 한국어 구문 규칙

에서 '@'+'숫자'는 의미 태그 번호이고, 대역어에서 '#'+'숫자'는 한국어 구문 규칙에서 체언이 나타나는 순서를 의미한다. 말뭉치로부터 추출된 논항의 개수는 총 421개이다.

2.2 분석 문법

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 중요한 성분이다. 그러나 주어는 생략이 일반화되어 있고, 서술어는 특별한 이유 없이는 생략되지 않으므로 서술어가 문장 내에서 기능이 더 크다. 한국어는 개별언어로서는 물론 보편언어로서도 주어 중심 언어라기 보다는 서술어 중심 언어라고 결론지을 수 있다[13]. 따라서 한국어를 분석하기 위하여 문법을 구성할 때 서술어를 중심으로 분석 문법을 구성하는 것이 효율적으로 분석이 가능하다.

대화체 분석을 위한 기존의 문법은 매칭이 될 수 있는 개념이나 요소들을 나열하여 분석을 하였다. 이렇게 문법을 구성할 경우 한국어의 부분 자유 어순 특성상 문법이 방대해지므로 문법의 작성 및 관리가 어려워지는 문제점이 발생한다.

"사박 오일 동안 친구와 미국에 갑니다"에 대한 문장을 분석할 때 동사 '가다'에 대한 문법은 한국어의 부분 자유 어순 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 어절의 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. 서술어를 제외한 n개의 어절로 이루어진 하나의 문장이 순서가 바뀌어 나타날 수 있는 형태의 개수는 최대 n!이 되어 n!개의 문법을 작성해야 하므로 문법의 수가 방대해질 뿐만 아니라 문법을 관리하는데 어려움이 많다.

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

본 논문에서는 서술어에 따라 논항 구조에 기반한 문법을 구성한다. 분석 문법은 구(phrase) 문법과 서술어 문법으로 구분하여 작성한다. 구 문법은 입력 문장에서 나타나는 어절로부터 결합이 가능한 어절을 먼저 결합함으로써 동사에 따른 문법 수를 줄이고 파싱의 오버헤드를 줄이기 위함이 목적이다. 구 문법은 말뭉치로부터 결합이 가능한 구를 추출하여 작성하였다.

예를 들면 "사박 오일 동안 친구와 미국에 갑니다"에 대한 문장을 분석할 때 서술어 '가다'에 대한 분석을 하기 이전에 구 문법을 이용하여 '사박 오일 동안'에 대한 어절을 결합한 후 서술어 '가다'에 대한 문법을 분석한다. 구 문법의 구조는 다음과 같다.

@2 와,@2 가 : #1 and #2
@4, 동안 : during #1

구 문법에서 어절과 어절은 콤마(,)로 구분하고 구분자 ":"를 기준으로 좌측은 한국어 구 문법이고 우측은 한국어 구에 대한 대역어 패턴을 의미한다. 말뭉치로부터 추출된 구 문법은 총 189개이다.

서술어 문법은 말뭉치로부터 추출된 231개의 각 서술어에 대하여 논항을 추출하여 논항 번호에 의하여 작성하였다. 서술어 '가다'에 대한 서술어 문법의 일부를 표 2에 나타내었다.

표 2 서술어 '가다'에 대한 서술어 문법

논항	논항 번호
:	:
:	:
@2 [가|이]	202
@2 [과|와|하고]	204
@2 [는|은]	205
@2 [를|을]	208
@2 로	213
@3	302
@3 [과|와|하고]	304
@3 [는|은]	305
@3 [까지|까지는]	306
:	:
:	:

서술어 문법은 순서에 관계없이 문법을 기술할 수 있으므로 논항이 문장의 어느 위치에 나타나든 관계없이 논항을 취하여 분석이 가능하므로 한국어 부분 자유 어

순 특성으로 문법이 방대해지는 문제점을 해결할 수 있다. 또한 서술어가 취할 수 있는 논항은 논항 사전에서 선택됨으로써 분석에 불필요한 성분은 문법으로부터 제거되기 때문에 문법 규칙에 맞지 않는 성분을 건너띄어 분석하는 처리 방식과 같은 별도의 메커니즘이 필요 없이 대화체에서 요구되는 강건한 분석을 할 수 있다.

3. 논항구조 기반 분석

3.1 분석 시스템

한국어 대화체 문장의 분석은 의미 전달이 가장 큰 목적이므로 불필요한 성분을 제거하고 문법적으로 옳지 않은 문장을 분석하는 강건한 분석이 요구되며, 한국어의 특징 중의 하나인 부분 자유 어순 특성을 고려하여 분석하는 방법이 요구된다.

논항 구조에 기반한 분석은 입력 토큰에 대하여 서술어가 지배하는 논항을 문법으로부터 선택함으로써 수행한다. 분석은 입력 발화문에서 출현하는 모든 서술어에 대하여 서술어 단위로 분석을 하는데 입력 토큰의 앞부분부터 서술어가 나타나는 부분까지를 분석 단위로 구분하여 분석한다.

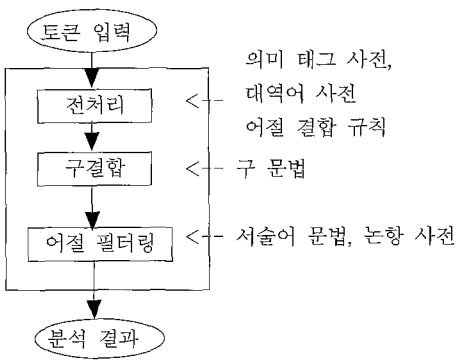

그림 1 시스템 개요

분석 시스템은 대화체 문장을 형태소 분석하여 얻은 토큰열 정보를 입력받아 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 특성[1,2,3]을 고려하여 후처리한 다음 그 결과를 분석 시스템의 입력으로 받는다.

형태소 분석 결과를 입력받아 분석을 하기 위하여 먼저 전처리 과정을 거친다. 전처리 과정에서는 입력된 각 명사의 어절에 대하여 의미 태그 사전을 이용하여 의미 태그를 부여한다. 명사의 의미 태그 부여는 의미 태그 사전에 해당 명사가 있으면 의미 태그를 부여하여 분석을 수행하고, 태그 사전에 없으면 품사 태그로 대체하여 분석을 수행한다.

또한 분석 시스템의 부담을 줄이고 분석 결과의 모호성을 감소시키기 위하여 어절 결합 규칙을 이용하여 결합 가능한 어절을 결합한다. 어절 결합은 이전의 어절에 조사가 없고, 의미 태그가 같으면 두 어절을 하나로 묶어 새로운 하나의 어절로 결합한다. 앞 어절에 조사는 없으나 의미 태그가 다를 경우는 어절 결합 규칙을 이용하여 결합한다.

예를 들어 입력 어절 '오월 팔일부터는'은 '오일'의 이전 어절 '오월'에 조사가 없고, '오월'과 '오일'의 의미 태그가 같으므로 '오월'과 '팔일부터'를 결합하여 하나의 어절로 만든다. 어절 결합 알고리즘을 그림 2에 제시하였다.

```
for 모든 어절 do
  begin
    if 이전 어절의 조사 <> NULL then continue
    if 이전 어절의 의미 태그 = NULL then continue
    if 이전 어절의 태그 = 품사 태그 then continue
    if 현재 어절의 의미 태그 = NULL then continue
    if 현재 어절의 태그 = 품사 태그 then continue
    if 이전 어절의 의미 태그 = 현재 어절의 의미 태그
    then 현재 어절과 이전 어절을 결합
    else 명사 결합 규칙을 이용하여 현재 어절과 이전 어절을 결합
  end
end
```

그림 2 어절결합 알고리즘

전처리 과정을 수행한 후 구 문법을 이용하여 구 결합을 한다. 구 결합은 미리 정의된 구 문법 내에 존재하는 구를 패턴 매칭하여 찾아내고 구를 생성하는 모든 어절들을 하나의 어절로 결합하고 구가 가지는 의미 태그는 그대로 이어 받는다.

예를 들어, 입력 어절 '서울에서 런던으로'는 구문법을 검사하여 구 패턴 '@3에서 @3으로'이 존재하면 '@3으로'라는 새로운 구를 생성한다.

구 결합을 함으로써 파서의 부담을 줄일 뿐만 아니라 입력된 어절을 단순화함으로써 문법을 구축하는 비용을 많이 줄일 수 있는 효과가 있다. 구 결합에 대한 알고리즘은 그림 3에 나타내었다.

구 결합이 끝난 후에 파서는 논항 사전과 서술어 문법을 이용하여 서술어 단위로 어절 필터링을 수행한다.

```
while true do begin
    find := false;
    for 남아 있는 어절 do begin
        if 어절의 품사 = 서술어
            then break;
            else 어절 리스트에 어절 추가
    end
    for 어절 리스트 집합의 모든 어절 do
        구 문법을 검사
        if 구 문법 <> NULL then begin
            새로운 구 생성
            find := true;
        end
    end
    // 구가 발견되지 않으면 탈출
    if find = false then break
    end
end
```

그림 3 구 결합 알고리즘

서술어 단위로의 어절 필터링은 입력 토큰열의 앞부분부터 차례로 어절을 읽어들여 서술어가 나타나는 위치까지를 단위로 분석 문법으로부터 서술어가 취할 수 있는 어절을 취하는 것이다. 즉, 서술어 문법에는 각각의 서술어가 취할 수 있는 논항들이 기술되어 있으며, 이 정보를 이용하여 서술어 앞에 나타나는 어절들을 검사하여 분석을 한다. 한 문장에 여러 개의 서술어가 존재할 경우 서술어 이후의 토큰부터 다음 서술어까지를 단위로 입력 토큰열이 끝날 때까지 계속하여 분석을 수행한다. 그림 4에 어절 필터링 알고리즘을 나타내었다. 어절 필터링을 하고 나면 입력 발화문에 대한 분석 결과는 그림 5와 같이 얻을 수 있다.

```
for 모든 어절 do begin
    if 어절의 품사 <> 서술어 then continue
    분석 문법을 검색
    if 문법 = NULL then continue
    for 서술어 이전의 모든 어절 do begin
        논항 사전 검색
        if 논항 = NULL then continue
        if 서술어에 의해 논항이 취해지면 then
            서술어 리스트에 어절 추가
    end
end
```

그림 4 어절 필터링 알고리즘

3.2 분석 결과

논항 구조에 기반한 한국어 대화체 분석은 전화상의 대화를 전사한 '여행안내' 영역의 말뭉치를 대상으로 하였다. 말뭉치에서 나타나는 발화문에 대한 분석 결과의 예를 그림 5에 나타냈다.

입력 발화문 : 오월 십사일날 서울에서 런던으로 가는 비행기편은요 대한항공이 열네시에 있는데요

```
있는데요    [decl] exist
  504 [@5는|@5은|@5은요] : #1
    비행기편은요    airplane
      가는    [adj] go
        401 [@4] : #1
          오월 십사일날    May 14th
        307 [@3로|@3으로] : #1
          서울에서 런던으로 ---- from Seoul
                                 to London
  504 [@9가|@9이] : #1
    대한항공이 ------ KAL
  412 [@4에] : #1
    열네시에 ------ 14 hours
```

그림 5 분석 결과

그림 5의 분석 결과는 서술어가 '가는'과 '있는데요' 두 개이기 때문에 입력 어절 '오월'부터 '가는'까지가 분석되고 난 후에 '비행기편은요'부터 '있는데요'까지가 분석되었다. 입력 어절 '오월 십사일날'은 형태소가 결합된 상태이고, '서울에서 런던으로'는 구가 0결합된 결과이다.

한국어는 부분 자유 어순 특성을 갖는데 이러한 특성을 고려하여 문맥 자유 문법 등으로 문법을 기술하면 문법이 방대해 질뿐만 아니라 분석을 하는 과정에서 불필요한 성분을 만나면 분석이 실패하는 경우가 발생한다. 그러나 논항 구조 기반 분석은 분석 문법을 통하여 논항을 입력 토큰으로부터 취하여 분석하기 때문에 수정 발화나 부분 자유 어순에 대한 문제점을 효과적으로 해결할 수 있다.

또한 대화체는 분석에 불필요한 성분을 제거하여 강건한 분석이 요구되는데 분석 문법으로부터 불필요한 성분이 제거되기 때문에 문법 규칙에 맞지 않는 성분을 건너띄어 분석하기 위한 처리 방식과 같은 별도의 메커니즘을 설계할 필요가 없이 강건한 분석을 할 수 있다. 그림 6에 강건한 분석 결과에 대한 예를 보여주고 있다.

그림 6의 입력 발화문 "육십달 육십불을 더 내셔야 됩니다"에서 어절 '육십달'은 수정 발화된 어절로 전체

분이다. 육십달은 서술어 문법에 의하여 취하지 않으므로써 분석되지 않고 '육십불을 더 내셔야 됩니다'만 분석됨으로써 강건한 분석을 할 수 있다.

입력 발화문 : 육십달 육십불을 더 내셔야 됩니다

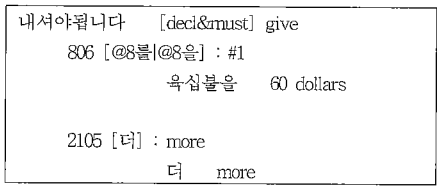

그림 6 강건한 분석 결과

분석은 서술어 앞에 나타나는 논항을 취하여 분석을 하는데 서술어 앞에 나타나는 논항이 해당 서술어의 지배 범위를 벗어나는 경우가 존재한다. 이러한 경우 서술어의 지배 범위를 벗어나 지배하는 서술어에 의해 논항이 분석되어야 한다. 그림 7에 논항이 다음 서술어의 지배 범위를 벗어난 경우의 분석 결과를 보여주고 있다.

그림 7의 입력 발화문 "요금은 출발 당일날 와서 지불하십시오"에서 논항 '요금은'이 서술어 '오다'에 의해 지배되도록 분석하면 분석에 실패한다. 그러나 서술어 문법에 의하여 논항 '요금은'은 서술어 '오다'에 의해 지배되지 않고 '지불하다'에 의해 지배되어 분석에 성공하였다.

이와 같이 논항 구조에 기반한 한국어 분석은 한국어 부분 자유 어순 특성 및 대화체가 가지고 있는 여러 가지 특성을 효과적으로 해결하고 강건한 분석할 수 있다.

입력 발화문 : 요금은 출발 당일날 와서 지불하십시오.

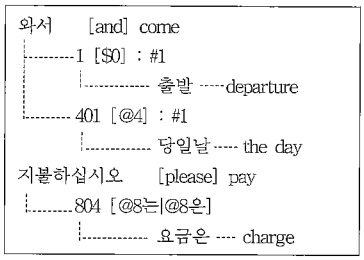

그림 7 서술어 지배 범위를 벗어난 분석 결과

4. 실험 및 평가

본 논문에서 실험을 위하여 사용된 데이타는 한국전자통신연구원(ETRI) corpus인 '여행 안내' 영역 중에서 1,335개의 훈련된 발화문과 420개의 훈련되지 않은 발화문을 대상으로 실험하였다.

표 3 실험 결과

훈련 여부	훈련된 발화문	훈련되지 않은 발화문
발화문 수	1,335	420
완전 실패	3	8
부분 실패	0	20
성 공	1,332	392
성공율(%)	99.7	93.3

훈련 방법은 일차적으로 403개의 발화문을 분석하여 논항 사전을 만들고 논항 사전을 바탕으로 분석 문법을 만들어 훈련된 발화문과 훈련되지 않은 발화문을 각각 실험을 하였다. 2차 실험에서는 712개의 발화문, 3차 실험에서는 1,020개의 발화문, 4차 실험에서는 1,335개의 발화문으로 훈련 발화문을 단계적으로 확대하면서 실험을 하였고, 실험 발화문을 확대하면서 논항 사전과 문법을 평가하고 확장하는 과정으로 훈련하였다. 훈련되지 않은 발화문은 훈련된 발화문에 의하여 단계적으로 확대된 논항과 문법을 적용하여 각각 실험하였다.

훈련된 발화문(trained utterance)은 발화문에 나타나는 어절이 실험을 통하여 문법에 반영되었다는 것을 의미하고, 훈련되지 않은 발화문(untrained utterance)은 발화문에 나타나는 어절이 반영되지 않은 문법에 의하여 실험된 것을 의미한다.

대화체 분석에 대한 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 발화문의 일부만이 분석이 된 경우이며, 성공은 완전 실패와 부분 실패를 제외한 결과로 올바른 분석 결과를 의미한다.

실험 결과는 표 3과 같이 나타났는데 1,335개의 훈련된 발화문은 99.7%의 성공률을 보이고, 420개의 훈련되지 않은 발화문에서는 93.3%의 성공률을 보이고 있다. 훈련된 발화문 수를 확대하여 문법을 더 확장할 경우 훈련되지 않은 발화문의 성공률은 계속 증가하여 훈련된 발화문의 성공률에 근접할 것으로 예상된다.

분석이 완전 실패하는 경우는 대화체의 특성상 발화시 서술어가 생략되어 발화됨으로써 서술어 중심으로 분석을 하는 본 논문에서는 분석을 할 수 없는 경우이거나 서술어 문법에 논항이 존재하지 않기 때문에 발생한다. 서술어가 생략되어 실패하는 경우는 입력 문장의 앞 뒤 문맥을 파악하여 생략된 어절을 복원함으로써 해결 가능하고, 서술어 문법에 논항이 존재하지 않아 실패

하는 경우는 훈련을 통하여 서술어에 논항을 추가함으로써 해결 가능하다. 부분 실패하는 경우는 훈련된 발화문에서는 훈련을 통하여 논항이 문법에 반영되었기 때문에 발생하지 않았고, 훈련되지 않은 발화문에서는 부분적으로 분석되지 않아 실패하는 경우가 대부분이었는데, 실패 원인은 서술어에 대한 논항의 부족으로 인하여 발화문의 논항 일부가 서술어 문법에 존재하지 않기 때문에 부분 실패하였다.

예를 들어 발화문 "엘리스 여행 경비는 신용카드로 지불되어 집니다"는 그림 8와 같이 분석되는데 분석 문법에서 서술어 '지불되다'에 대하여 '엘리스'에 대한 논항이 없기 때문에 입력 토큰 '엘리스'는 분석되지 않았다.

분석 결과는 형태소 분석 결과에서 중의성이 없다는 가정 하에서 수행하였다. 또한 입력 문장 자체의 의미가 애매하여 의미 전달이 모호한 경우, 띄어쓰기나 인식이 잘못된 경우, 발화의 중복이 심하여 이해가 어려운 경우 등은 입력 문장을 일부 수정하여 형태소 분석을 하였다

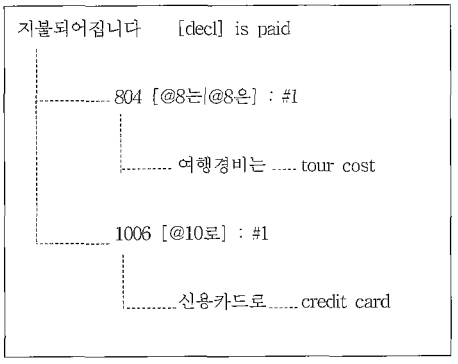

그림 8 분석이 실패한 예

5. 결 론

부분 자유 어순 특성을 가지는 한국어를 문맥 자유 문법 형태의 문법으로 기술할 경우 문법이 방대해지고 대화체의 특징으로 나타나는 불필요한 성분을 처리해야 하므로 파서의 부담이 커진다. 또한 기존의 개념 기반 기법은 불필요한 개념으로 인한 파싱의 오버헤드와 한국어 부분 자유 어순 특성을 고려하지 않아 문법이 방대해지는 문제점이 있고 한국어의 조사를 고려하지 않고 문법을 기술하기 때문에 정확한 생성 결과를 얻기 어려우며, 예제나 패턴을 이용한 기법은 문법수가 방대해지고 문법이 존재하지 않으면 실패하는 문제점이 있다.

이러한 문제점을 해결하기 위하여 본 논문에서는 서술어가 지배하는 성분 중에서 중요한 의미를 갖는 논항을 이용하여 문법을 기술하였다. 실험을 위하여 사용된 데이타는 여행안내 영역 중에서 1,335개의 훈련된 발화문과 420개의 훈련되지 않은 발화문으로, 실험 결과 훈련된 발화문은 99.7%, 훈련되지 않은 발화문은 93.3%의 분석 성공률을 보였다.

논항 구조 기반 문법은 순서에 관계없이 문법을 기술하고, 논항이 문장의 어느 위치에 나타나든 관계없이 분석이 가능하므로 한국어의 부분 자유 어순 특성과 대화체 특성을 효과적으로 해결하였다. 또한 대화체 분석에 불필요한 성분이 문법으로부터 제거되기 때문에 문법 규칙에 맞지 않는 성분을 건너띄어 분석하기 위한 방식과 같은 별도의 메커니즘 없이 대화체에서 요구되는 강건한 분석을 하였다.

대화체 특성으로 문장 성분이 생략되어 발화되는 경우가 있다. 특히 본 논문에서는 서술어 중심으로 분석을 하기 때문에 서술어가 생략되어 발화되는 경우는 분석에 실패한다. 따라서 발화문의 앞 뒤 문맥을 파악하여 생략된 성분을 복원할 필요가 있다. 또한 실험 도메인을 '여행 안내' 영역으로 제한하여 실험하였기 때문에 대상 도메인을 다른 영역으로 확대할 경우 의미 자질을 체계적으로 분류하고, 각 서술어마다 논항을 추출하여 확장할 필요가 있다. 그리고 번역의 성공률과 질을 높이기 위하여 문장 생성시 절의 수식관계를 정립하고, 대동사나 대명사 처리를 하기 위하여 문맥 정보를 고려하는 방안도 고려할 필요가 있다.

참 고 문 헌

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

[2] 최운천, 한남용, 김영섭, "개념파서를 이용한 대화체 음성언어 번역", 한국정보처리학회 추계 학술발표논문집, 제2권 제2호, 1995

[3] 정천영, 서영훈 "의미 패턴에 기반한 대화체 한영 기계번역", 한국정보처리학회 논문지, 제5권 제9호, pp. 2361-2368, 1998.9

[4] 서영훈 외, 대화체 및 문어체 기계번역을 위한 한국어 구문/의미 해석시스템 개발, 한국전자통신연구소, 1995

[5] 서영훈, "형태소 정보를 이용하는 개념 기반의 한국어 자연발화 분석기", 충북대 학교산업과학기술연구소 논문집, 제11권 2호, 1997.12

[6] 최재웅, "대화분석에 있어서의 몇 가지 문제: 호텔 예약 전화 대화를 중심으로", 한글 및 한국어 정보처리 학술 대회, 1996.10

[7] Seneff, S., "TINA : A Natural Language System for Spoken Language Applications," Computational

Linguistics, Vol.18, No.1, pp.61-86, 1992
[8] Levin, E., and R. Pieraccini, "Concept-based Spontaneous Speech Understanding System," Proceedings of Eurospeech '95, pp.555-558, 1995
[9] Levie, A., "GLR* : A Robust Grammar Focused Parser for Spontaneously Spoken Language," Doctoral Thesis, Carnegie-Mellon University, 1995
[10] Mayfield, L., M.Cavalda, Y.-H. Seo, N. Suhm, W. Ward, A. Waibel, "Parsing Real Input in JANUS: A Concept-based Approach to Spoken Language Translation," Proceedings of TMI95, 1995
[11] B. Suhm, P. Geutner, A. Lavie, L. Mayfield, "JANUS: Towards Multilingual Spoken Language Translation," Interactive System Laboratories, Carnegie Mellon University
[12] 김나리, 김영택, "한국어 동사 패턴에 기반한 한국어 문장 분석과 한영 변환의 모호성 해결", 한국정보과학회 논문지, 제23권 제7호, pp.766-775, 1996.7
[13] 이관규, 국어 대등 구성 연구, 서광학술자료사, 1992
[14] Levin, A and Tomita, M. "An Efficient Word-Skipping Parsing Algorithm for Context-Free Grammar," 3rd International Workshop on Parsing Technologies(IWPT93), Belgium, 1993
[15] Woszczyna, M. et al, "Recent Advances in JANUS : A Speech Translation System," Proceedings of Eurospeech '93, pp.1295-1298, 1993

정 천 영

1986년 충남대학교 계산통계학과 (학사) 1992년 충남대학교 전산학과 (석사). 2000년 충북대학교 컴퓨터공학과 (박사). 1986년 ~ 1997년 한국에너지기술연구소 연구원. 1997년 ~ 2001년 구미1대학 컴퓨터정보전공 조교수. 2001년 ~ 현재 혜천대학 컴퓨터통신계열 조교수.

서 영 훈

1983년 서울대학교 컴퓨터공학과 (학사). 1985년 서울대학교 컴퓨터공학과 (석사). 1991년 서울대학교 컴퓨터공학과 (박사). 1988년 ~ 현재 충북대학교 컴퓨터공학과 교수. 1994년 ~ 1995년 미국 Carnegie-Mellon 대학 기계번역센타 객원교수. 관심분야는 자연언어처리, 음성언어처리, 기계번역