문장의 종결정보와 예문을 이용한 핵심개념 기반의 한국어 대화체 파싱

김홍국, 서영훈
충북대학교 컴퓨터공학과
hkdeer@dcenlp.chungbuk.ac.kr, yhseo@cbucc.chungbuk.ac.kr

Core Concept-based Korean Spoken Language Parsing Using Ending Information and Example Sentences

Hong-Kuk Kim, Young-Hoon Seo
Department of Computer Engineering, Chungbuk National University

요 약

핵심개념 기반의 분석 시스템은 기존의 CFG형태로 기술된 문법의 양을 현저히 줄이고 간투어, 중복발화등과 같은 파싱 불필요 성분을 처리하는 루틴을 개선해 파서의 부담을 줄인 분석 방법이다. 핵심개념 기반 분석 시스템은 동사를 기준으로 문법을 기술한다. 따라서, 발화자의 사투리 등에 의해서 동사 정보를 상실한 문장은 분석이 되지 않는 문제점을 가지고 있으며 또한, 문장 분석시 분석 문법을 구성할 수 없는 짧은 발화문같은 경우에도 분석을 하지 못하는 문제점을 가지고 있다. 이러한 문제점들을 해결하기 위해서 본 논문에서는 발화문의 예를 작성해 놓은 예문사전과 발화문이 가지고 있는 종결형 정보를 이용해서 그러한 문제를 해결하고 분석의 정확성을 높였다.

1. 서 론

현재까지 자연언어의 분석은 문어체를 대상으로 연구가 많이 이루어져 왔고 대화체에 대한 연구는 문어체의 연구에 비해서 연구가 많이 이루어 지고 있지 않은 상황이다. 이러한 이유는 문어체를 기반으로한 연구가 문법적으로 잘 다듬어진 문장들을 다루어 분석의 예외성이 적은 반면에 대화체 문장은 대화체에서 갖는 단어의 축약, 조사의 생략, 수정 또는 반복 발화, 간투어 등의 특성으로 인해 자연언어를 분석하는데 많은 문제점들을 포함하고 있기 때문이다[1,2]

이러한 대화체 문장을 분석하기 위한 가장 대표적인 방법이 개념기반 분석기법이다. 개념기반 분석기법은 강건성을 가장 큰 장점으로 가지며 비문법적인 요소를 많이 포함하고 있는 자연발화 처리에 유리한 기법 중 하나로 평가되고 있다[5,6]. 그러나, 개념기반 분석기법은 분석에 불필요한 성분을 제거하기 위한 분석의 오버헤드와 한국어 부분 자유어순 특성을 CFG(Context Free Grammar)형태의 문법으로 기술함으로써 문법이 방대해지는 문제점을 가지고 있다[1,2]. 이러한 문제점들을 해결하기 위한 분석기법으로 핵심개념기반 분석기법[1]이 있다. 핵심개념기반 분석기법은 대화체 문장에서 중요한 의미를 가지는 요소만을 핵심개념으로 정의하고 문법을 기술하여 문법이 방대해지는 것

을 막고 간투어, 중복발화 등과 같은 파싱 불필요 성분 처리를 위한 메커니즘을 따로 구성하지 않아 파서의 부담을 덜은 분석기법이다. 그러나, 핵심개념 기반 분석기법의 문법 구성이 동사를 중심으로 이루어져 발생하는 문제점이 몇 가지 있다. 첫째, 발화할 때 사투리를 사용하거나 발화 상태의 원인으로 동사가 출현하지 않는 발화문의 경우로서 입력문장이 동사 '이다'를 포함한 ~'입니다'로 종결될 때 발생한다. 둘째, 동사가 출현하더라도 그 동사에 포함될 핵심개념이 없는 짧은 발화문의 경우로서 이러한 예는 '연락하겠습니다'와 같은 짧은 발화문이 입력될 때 핵심개념 기반 분석 시스템으로 분석시 동사 '연락하다'에 포함될 핵심개념이 없기 때문에 정확한 결과를 생성하지 못한다는 것이다.

이에 본 논문에서는 이러한 문제점들을 해결하기 위해서 예문과 발화문의 종결형 정보를 이용하여 분석을 수행하고자 한다.

2. 핵심개념 기반 문법

2.1 핵심개념의 구성

본 논문에서 이용하는 핵심개념 기반 문법의 구성을 살펴보면 다음과 같다. 문법은 여행안내(travel arrangement) 영역 ETRI Corpus, 1575개의 발화문을 기반으로 하여 작성되었으며 개념은 3단계로 구성이 된다. 개념은 단어를 중심으로 하는 단위개념, 하나 이상의 단위개념이 형성하는 상위개념 그리고 단위개념이나 상위개념이 이루는 최상위개념으로 구분된다. 단위개념은 단어나 연속된 단어들에 부여하는 개념으로 형태소 분석과 전처리를 거친 토큰이나 토큰열로 구성된다. 예를 들어 장소개념 [local]은 단위개념 nation(국가), city(도시), island(섬) 등으로 구성되는 상위개념이 되며 표 1은 일부 상위개념들을 나타낸다.

표 1. 핵심개념 기반 문법 상위개념

상위개념	상위개념의 의미
[go]	'가다'의 의미를 내포한 개념
[book]	'예약하다'의 의미를 내포한 개념
[offer]	'제공하다'의 의미를 내포한 개념
[temporal]	시간의 의미를 내포한 개념
[local]	장소의 의미를 내포한 개념

[go], [book], [offer] 등은 동사를 중심으로 구성된 토큰들로서 발화자나 어떠한 객체의 행위를 나타내는 개념이며, [temporal], [local]등은 명사나 구를 중심으로 구성된 토큰들로서 동사를 중심으로 구성된 개념들에 의해 사용되는 것이다.

최상위개념은 상위개념과 단위개념들의 조합으로 구성이 되며 현재 7개의 최상위개념으로 구성이 되어 있다. 이는 실험에 사용된 여행안내 영역 Corpus에서 7개의 최상위개념만으로도 발화자가 의도하는 바를 표현할 수 있기 때문이다. 최상위개념은 표 2와 같다.

표 2. 핵심개념 기반 문법의 최상위개념

최상위개념	최상위개념의 의미
[give_info]	청자에게 발화자가 정보를 주는 최상위개념
[i_want]	발화자의 희망을 나타내는 최상위개념
[i_will]	발화자의 의지를 나타내는 최상위개념
[query]	발화자의 질의를 나타내는 최상위 개념
[request]	발화자의 요구를 나타내는 최상위 개념
[respond]	짧은 응답을 나타내는 최상위 개념
[nicety]	인사말을 나타내는 최상위 개념

2.2 핵심개념 기반 문법 구성

대화체 분석을 위한 기존의 문법은 매칭이 될 수 있는 개념이나 요소들을 나열하여 분석을 하지만 핵심개념 기반 문법은 한국어의 부분 자유어순의 특성을 고려하여 문법을 구성한다. 부분 자유어순 특성을 가진 한국어를 CFG형태의 문법으로 구성할 경우 문법이 방대해지므로 문법의 작성 및 관리가 어려워지게 된다.

이와 같이 한국어의 부분 자유어순 특성이 문법에 적용될 경우 문법으로 모든 순서를 처리해야 하므로 문법의 양이 방대해 지는 것을 피할 수 없으며 문법을 관리하는데 또한 어려움이 따르게 된다. 이에 반해서 핵심개념 기반 문법에서는 하나의 개념으로 묶일 수 있는 요소들은 개념 안에서 자유롭게 위치이동이 가능한데 이는 개념을 집합으로 간주하고 개념으로 묶일 수 있는 토큰들을 집합의 요소로서 간주함으로써 쉽게 문법을 구성할 수 있게 된다.

핵심개념 기반 문법에서는 문법을 구성할 때 대상 Corpus 문장을 동사 단위로 분리하고 i 번째

동사 v_i와 i+1번째 동사 v_{i+1} 사이에 있는 토큰열을 개념으로 생성한다. 토큰열에서 하나의 토큰이 개념을 형성할 수 있고 하나 이상의 토큰이 개념을 형성할 수 있다. 형성된 개념들이 동사 v_{i+1}에 필요한 성분인지를 검사하여 만일 필요한 성분이라면, 동사 v_{i+1}을 나타내는 개념에 대한 문법 G_{i+1}의 핵심개념으로 취해지게 된다. 이러한 절차로 모든 Corpus의 문장을 대상으로 동사 v_{i+1}이 가질 수 있는 개념들을 문법 G_{i+1}에 기술함으로써 동사에 대한 문법이 구성된다. 예를 들어 '~로 가는 비행기'라는 문장이 있을 경우 개념 [go]는 '비행기'를 수식하여 '비행기'에 대한 개념 [aircraft]에 포함되게 된다.

이를 적용하여 동사 '출발하다'에 대한 '여행안내'영역 Corpus에 대해서 핵심개념 기반 문법을 구성하면 그림 1과 같다.

```
[depart]
        [aircraft]
        [local]
        [temporal]
        [noun]
        directly
        when
```

그림 1. 동사 '출발하다'에 대한 핵심개념 기반 문법

3. 발화문의 종결형 정보와 예문을 이용한 대화체 분석

3.1 예문사전을 이용한 분석

대화체는 문어체와 달리 대부분의 대화에서 인사말이 출현하고 간단한 질의/응답이 출현하게 된다. 이러한 발화문은 대부분 문장이 짧아 개념을 포함하지 않게 되므로 동사를 중심으로 분석을 수행하는 핵심개념 기반 문법에서는 분석이 되지 않는다. 본 논문에서는 그러한 짧은 발화문의 분석을 위해서 예문기반 분석기법을 도입하여 이러한 문제를 해결한다. 이를 위해서 Corpus를 분석해서 짧은 발화문을 분석하고 그것을 바탕으로 예문사전을 구성하였다.

분석시스템의 전체적인 구성을 보면 다음 그림2와 같다.

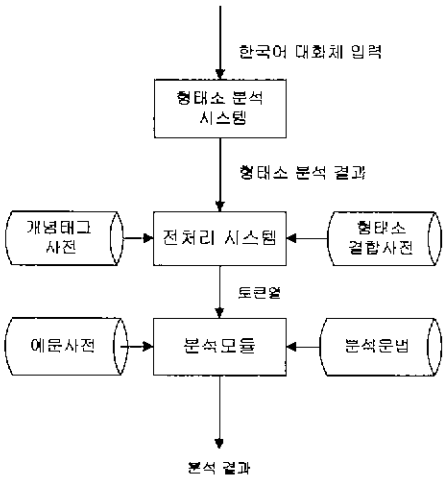

그림 2. 분석시스템의 처리과정

분석시스템은 형태소 분석 시스템, 전처리 시스템, 한국어 분석모듈로 구성되며 형태소 분석 시스템은 한국어 대화체 입력에 대한 형태소 분석 결과를 출력한다. 전처리 시스템은 개념태그 사전과 형태소 결합사전을 이용해서 토큰열을 출력하게 된다. 토큰열의 구성은 <@1 @2> 형태로 구성되는데 @1 필드는 개념 필드로서 토큰명(token name)을 나타내고 @2 필드는 속성(attribute)을 나타낸다. 입력된 단어가 '대화'라면 토큰명 'noun', 속성은 '대화'가 되어 <noun 대화>와 같이 출력된다. 따라서, 전처리 시스템을 거친 발화문의 출력은 다음과 같이 구성된다.

<token1 attr1><token2 attr2>···<token n attr n>

전처리 시스템을 거친 토큰열들은 분석모듈의 입력이 되고 분석모듈에서는 분석문법을 이용해서 입력된 토큰열들을 분석하고 번역을 위한 분석결과를 출력한다. 그러나, 분석과정에서 개념을 포함하지 않는 짧은 발화문이 출현할 경우에는 예문사전을 이용하여 입력된 문장과 가장 가까운 예문을 찾고 그에 해당하는 영문을 출력해 주게 된다. 발화문이 '전화주셔서 고맙습니다'일 경우의 예를 살펴보면 다음과 같다.

```
입력문장 : 전화주셔서 고맙습니다

전처리 결과 :
    <verb 전화주><eomi because><verb 고맙>
<eomi decl>

분석 결과 :
[give_info]([thanks]([because]([telephone]())))
```

그림 3. 핵심개념기반 분석 결과

입력문장 '전화주셔서 고맙습니다'의 핵심개념 기반 분석 결과는 그림 3과 같다. 핵심개념 기반 문법은 동사를 중심으로 문법을 구성한다. 즉, Corpus를 분석하여 발화문이 포함하고 있는 동사와 직접적으로 관련이 있는 문장들로 문법을 구성한다. 따라서, 위와 같은 문장이 입력될 경우 동사 '전화주'에 대한 문법을 구성할 수 있는 개념이 '<verb 전화주>' 앞에 출현하지 않기 때문에 정확한 분석을 하지 못하게 된다. 즉, 그림 3의 분석 결과처럼 토큰 '<verb 전화주>'에 대해서 상위개념 [telephone]만을 형성하고 [telephone]에 대한 하위개념을 형성하지 못하게 된다. 따라서, 예문사전에서 '전화주셔서 고맙습니다'와 같은 문장을 찾아 그에 해당하는 영문을 결과로 생성하게 된다. 해당하는 문장이 없을 경우에는 가장 유사한 문장을 결과로 생성한다. 예문사전의 구성 예와 분석결과는 그림 4,5와 같다.

```
감사합니다/thank you/
예 감사합니다/yes thank you/
네 감사합니다/yes thank you/
안녕하세요/Hello/
안녕하십니까/Hello/
알았습니다/I see/
예 알았습니다/yes I see/
연락하겠습니다/I will contact/
전화주셔서 고맙습니다/thanks for calling/
```

그림 4. 예문사전의 구성 예

```
입력문장 :
        알았습니다 그럼 구월 십육일 오전
열시 비행기로 예약해 주십시오

전처리결과 :
<verb 알><eomi decl><month 구><day
십육><t_o_d 오전><hour 열><aircraft
비행기><verb 예약하><eomi please>

분석결과 :
I see
[request]([book]([temporal](<month 구>
<day십육><t_o_d오전><hour열>)[airplane](
<aircraft 비행기>)))
```

그림 5. 예문사전을 이용한 분석결과

그림 5에서 전처리 결과 동사 '<verb 예약하>'와 종결어미 '<eomi decl>' 사이의 토큰들이 동사 '예약하다'의 핵심개념으로 취해지고 분석결과 발화자의 요구를 나타내는 최상위개념인 [request]를 형성한다. 그러나 동사 '<verb 알>'은 앞에 출현하는 개념이 없으므로 예문사전에서 입력문장 '알았습니다'에 해당하는 영문 'I see'를 분석결과로 출력하게 된다.

3.2 발화문의 종결형 정보를 이용한 분석

대화체에서 발화문이 동사 '이다'를 포함한 ~'입니다'의 형식으로 종결될 때 발화문의 입력에 따라 전처리 결과가 다르게 출력이 된다. 또한 ~'입니다'가 대화체에서 발화의 상태에 따라서 다르게 변형이 될 수 있는데 이때 발화문을 분석하게 되면 동사 정보를 상실한 결과를 출력하게 된다. 그림 6에 위와 같은 발화문의 예를 나타내었다. 그림 6에서 입력문장 1이 정확하게 분석이 된 문장이고 입력문장 2는 '동안입니다'가 띄워 쓰기 없이 입력이 되어 동사 정보를 상실한 경우이고 입력문장 3은 동사 ~'이다'를 포함한 문장 ~'입니다'가 대화상에서 변형되어 정확한 결과를 출력하지 못한 예이다.

입력문장 1 :
'경주 패키지는 삼박 사일 동안 입니다'
전처리 결과 :
<city 경주><noun 패키지><dur_sleep 삼><day 사><noun 동안><verb 이><eomi decl>
입력문장 2 :
'경주 패키지는 삼박 사일 동안입니다'
전처리 결과 :
<city 경주><noun 패키지><dur_sleep 삼><day 사><noun 동안><eomi is>
입력문장 3 :
'경주 패키지는 삼박 사일 동안 이구요'
전처리 결과 :
<city 경주><noun 패키지><dur_sleep 삼><day 사><noun 동안><eomi after>

그림 6. 입력문장의 형태에 따른 전처리 결과

이러한 결과는 동사 '이다'가 그림 6의 문장에서 보이는 것처럼 어떤 명사와 결합할 때 동사로 분석이 되기 보다는 어미로서 분석이 되기 때문이다. 본 논문에서는 위와 같은 문제를 해결하기 위해 문장의 종결형 정보를 이용한다.

문장의 종결형 정보, 즉 종결어미는 대화체 분석에서 화자의 의도를 분석할 수 있는 중요한 것으로서 평서, 의문, 청유, 가능형 등으로 분류할 수 있으며 본 시스템에서는 총 15가지로 종결형 어미를 분류하고 있다. 표 3은 종결형 어미의 일부를 보여주고 있다.

표 3. 종결형 어미의 종류와 예

종결형 어미	종결어미 예
decl	습니다, 해요, 니다,...
is	ㅂ 니다, 라고@합니다,...
after	이구요, 가지구요,...
quest	인가요, 있습니까,...
will	하겠습니다, 주겠습니다,...
can	@수@있습니다, @수@있어요,...
guess	을@것@같습니다, 을@겁니다,...
want	ㄹ 려고@합니다, 시기@바랍니다,...
please	ㄹ 켜@주십시오, 십시오,...
can&quest	ㄹ 수@있을까요, 실@수@있나요,...

본 논문에서는 편의상 전처리과정을 통해 분석된 종결형 어미 'decl', 'is' 등을 '종결형 어미'라 하고 입력문장의 종결형 어미 '습니다' '이구요' 등을 '종결어미'로 구분한다. 동사 정보를 상실한 문장을 분석하기 위해서 ETRI Corpus 여행안내 영역의 1575개 발화문을 분석한 결과 표 3에서 동사 '이다'가 생성할 수 있는 종결형은 decl, is, after, quest 이다. 그림 6의 입력문장 2와 같이 종결형 어미가 'is'일 경우에는 단지 띄워 쓰기 문제로 인하여 동사 정보를 상실한 경우이므로 동사 '이다'로 분석을 수행한다. 그러나, 그림 6의 입력문장 3과 같이 문장이 변형되어 입력이 된 경우에는 문장의 종결어미를 비교한다. 이를 위해서 종결형 어미 각각에 대해서 종결어미의 예를 표 3과 같이 작성한다. 입력문장의 종결어미가 표 3의 종결어미 내에 있을 경우에 동사 '이다'로 분석을 수행하고 그렇지 않을 경우에는 입력문장이 문장을 구성할 수 있는 형태의 문장이 아니고 단지 단어가 나열되어 있는 의미 없는 문장임을 나타낸다. 분석을 수행하기 위해서는 동사 '이다'에 대한 개념 '[is]'를 생성하고 분석을 수행한다. 분석결과는 그림 7과 같다.

입력문장 :
'경주 패키지는 삼박 사일 동안 이구요'
전처리 결과 :
<city 경주><noun 패키지><dur_sleep 삼><day 사><noun 동안><eomi after>
분석 결과 :
[give_info]([is]([local](<city경주>)
[pnoun](<noun패키지>)[temporal](<dur_sleep 삼><day 사>)[pnoun](<noun 동안>)))

그림 7. 종결형 정보를 이용한 분석 결과

그림 7의 분석결과에서 동사 '이다'에 대한 개념 '[is] 가 형성되고 동사 '이다'에 기술된 문법을 바탕으로 결과를 생성하게 된다.

4. 실험 및 결과

본 논문은 핵심개념에 기반한 분석기법에 예문과

종결형 정보를 이용한 분석기법을 적용한 한국어 분석 시스템이다. 본 논문의 실험을 위해서 전화상의 대화를 전사한 ETRI Corpus의 여행안내 영역 1575개의 발화문을 대상으로 이루어졌다. 먼저, 1575개의 발화문에 나타나는 짧은 발화문을 분석하고 그것을 기반으로 예문사전을 구축하였다. 대화체의 특성상 거의 모든 발화문에서 인사말이 출현하였고 또한 짧은 질의 및 응답도 출현 빈도가 높게 나타나고 있다.

코퍼스를 분석한 결과 1575개의 발화문중 인사말이나 짧은 질의/응답이 총 677번 출현하였다. 이러한 결과는 전체 발화문의 약 43%에 해당되는 것으로 대화체에서 짧은 발화문이 차지하고 있는 중요성을 수치로서 보여주고 있다. 또한, 중복되는 발화를 제외한 문장은 135번 출현하였으며 예문사전은 중복된 발화를 제외한 문장으로 구성이 되었다. 이렇게 예문사전을 구성하여 분석을 수행함으로써 기존의 핵심개념 기반 분석시스템이 분석하지 못했던 짧은 발화문을 정확하게 분석하고 분석 결과의 정확도를 향상시켰다.

또한, 대화체 분석시에 동사 정보가 출현하지 않는 문장은 376개로서 전체 Corpus의 약 23.8%를 차지했다. 이러한 문장들은 동사 정보를 상실한 문장들이므로 기존의 핵심개념 기반 분석기법에서는 분석이 되지 않았던 것들이다. 이러한 발화문을 분석함으로써 대화체 분석의 성능을 크게 향상시킬 수 있었다.

5. 결 론

본 논문에서는 예문과 문장의 종결형 정보를 이용해서 핵심개념 기반 대화체 분석 시스템이 가진 문제점을 해결하였다.

대화체는 문어체와 달리 대부분의 대화에서 인사말이 출현하고 간단한 질의/응답으로 구성된 발화문이 많이 출현하게 된다. 이러한 발화문은 대부분 문장이 짧아 개념을 포함하지 않게 되므로 정확한 분석이 되지 않는다. 이러한 대화체의 문제점을 해결하기 위해서 ETRI Corpus '여행안내' 영역의 1575개의 발화문을 분석하고 간단한 인사말, 질의문 그리고 응답문을 찾아 예문사전을 구축하고 여기에 발화문과 이에 상응하는 영문표현을 기술해 놓았다. 따라서, 분석 시스템이 개념을 포함하지 않는 짧은 발화문을 분석할 때는 예문사전을 찾아보아 일치하는 발화문이 있을 때 그에 상응하는 영문표현을 결과로 생성하고 그렇지 않을 경우에는 가장 유사한 발화문을 찾아 결과를 출력하여 한국어 대화체 분석의 정확성을 향상시킬 수 있었다. 또한 동사 정보를 상실한 발화문의 유형을 분석하고 이를 문장의 종결형 정보를 이용해서 분석함으로써 기존의 분석 기법이 분석하지 못했던 많은 문장들을 정확하게 분석하여 대화체 분석의 성능을 크게 개선시켰다.

참고문헌

[1] 노서영, 정천영, 서영훈 "핵심개념 기반의 강건한 한국어 대화체 파싱", 한국정보처리학회 논문지, 제6권, 제8호, pp. 2113-2123,1999.

[2] 정천영, 임희동, 서영훈 "대화체 기계번역을 위한 중심어 기반 한국어 분석", 충북대학교 산업과학기술연구소 논문지, 제13권 제1호, pp. 47-56, 1999

[3] 왕지현, "구문 정보를 이용한 개념기반의 한국어 대화체 분석기", 충북대학교 석사학위논문, 1998

[4] 김영훈, "개념 구조를 이용한 한국어 대화체 구문 분석", 충북대학교 석사학위논문, 1997

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

[6] Mayfield, L., M.Cavalda, Y-H S대, N. Suhm, W. Ward, A. Waibel, "Parsing Real Input in JANUS: A Concept-based Approach to Spoken Language Translation", Proceeding of TMI95, 1995

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

[8] 이현정, 서정연, "문장의 화행을 반영한 한-영 대화체 기계번역", 한글 및 한국어 정보처리, pp.271-276, 1997

[9] Bonnie J. Dorr, Pamela W. Jordan, John W. Benoit, " A Survey of Current Paradigms in Machine Translation", 2000