개념간 상호 정보를 이용한 효율적인 개념기반 한국어 대화체 파싱

노 서 영
충북대학교 컴퓨터공학과
충북 청주시 개신동 산48
우 : 361-763
rsyoung@dcenlp.chungbuk.ac.kr

정 천 영
구미 1대학 전자계산학과
경북 구미시 부곡동 407
우 : 730-170
cyjung@mail.kumi.ac.kr

서 영 훈
충북대학교 컴퓨터공학과
충북 청주시 개신동 산48
우 : 361-763
yhseo@cbucc.chungbuk.ac.kr

An Efficient Concept-based Spoken Language Parsing for Korean using Mutual Information between Concepts

Seo-Young Noh
Dept. of Computer Engineering,
Chungbuk National University

Chun-Young Jung
Dept. of Computer Science,
Kumi 1st College

Young-Hoon Seo
Dept. of Computer Engineering,
Chungbuk National University

요약

개념기반 한국어 대화체 분석 시스템에서 어려운 점으로 대두될 수 있는 것 중의 하나가 대화체 파싱에서 과다한 탐색공간의 생성이다. 과다한 탐색공간의 생성은 대화체 발화문으로부터 불필요한 탐색공간을 제거하는 메커니즘의 결여 때문이다. 따라서 본 논문에서는 이러한 문제점을 해결하고자 개념에 기반 되어서 작성된 문법을 통해서 얻어진 동사정보를 구성하여 단일 최상위 레벨 개념들로 분리하고 이를 가장 최소 개수의 최상위 레벨 개념으로 제한해서 제한된 개념으로 대화체 토큰열을 전사시키는 방법을 제시하였다. 그 결과 기존 탐색공간의 40%정도의 탐색공간을 제약할 수 있었다.

1 서론

인간이 사용하는 언어를 크게 두 가지 유형으로 나누면, 그것은 대화체와 문어체로 나누어 볼 수가 있다. 자연언어 처리 분야에서 문어체에 대한 연구는 활발하게 진행되어 온 반면 대화체에 대한 연구는 국내에서 아직까지 연구가 활성화되지 않고 있다.[1,2] 최종 사용자 인터페이스의 측면에서 인간이 사용하는 대화체 분석에 대한 연구는 절실하며 향후의 시대적 요구 또한 이 방면에 많은 중점을 둘 것임이 틀림없다. 대화체 문장을 분석함에 있어서 여러 가지 문제점이 발견될 수 있지만, 본 논문에서는 두 가지 난점에 대해 초점을 맞추었다. 첫째는 개념기반 문법을 통한 한국어 대화체 분석 시스템에서 파싱을 해 나갈 때 과다한 탐색공간의 생성이다. 과다한 탐색공간의 생성은 대화체 발화문으로부터 불필요한 탐색공간을 제거할 수 있는 메카니즘의 결여 때문이다. 과다한 탐색공간의 생성은 시스템의 전체 성능 저하로 연결된다. 문법을 따라 파싱을 해 나가다가 문법과 매칭이 일어나지 않았을 때 파싱 수행과정의 모든 정보가 불필요한 정보가 되고 만다. 둘째는 대화체 문장을 단일 문장단위로 분리하기가 어렵다는 것이다. 문어체 문장은 구두점에 의해서 문장과 문장사이의 경계가 명확히 들어 나지만 대화체 문장은 문장사이의 구별이 없이 단일 문장으로 형성이 된다. 여러 개의 문장이 단일 문장으로 형성된 입력에 대해서 개념기반 문법을 적용했을 때 개념들을 분리해 내는데 많은 과부하가 발생하게 된다.

2 개념기반 문법 구성

본 논문에서 이용하는 개념기반 문법은 '여행안내' 영역 1575개의 발화문을 기반으로 하여 작성된 문법이다. 여기에서 사용된 최상위 레벨 개념들은 [표 2-1]과 같다.

최상위개념	개념의 의미
[give_info]	청자에게 발화자가 정보를 주는 토큰
[i_want]	발화자의 희망을 나타내는 문장에 대한 토큰
[i_will]	발화자의 의지를 나타내는 문장에 대한 토큰
[nicety]	인사말이나 헤어질 때의 인사를 위한 토큰
[query]	발화자의 질의를 나타내는 문장에 대한 토큰
[request]	발화자의 요구를 나타내는 문장에 대한 토큰
[respond]	짧은 응답을 나타내는 문장에 대한 토큰
[othertop]	7개의 최상위 개념외의 문장에 대한 토큰

[표 2-1] '여행 안내' 영역의 최상위 개념

8개의 최상위 개념과 137개의 하위 개념들로 개념기반 문법이 작성되어 있다.[4,5] 하나의 개념은 토큰열과 다른 개념의 조합으로 구성되어 있으며 개념들간의 포함관계에서 하나의 개념으로 생성되기 위해서는 반드시 최하위 개념은 토큰으로부터 시작되어야 한다. 개념을 나타내는 문법은 크게 세 가지 Type으로 구분될 수 있는데 이를 일반화하면 다음 Type1, 2, 3의 형태를 가지게 된다.

Type 1:
$\langle t_1\ a_1\rangle\langle t_2\ a_2\rangle\cdots\langle t_n\ a_n\rangle$

Type 2:
$\langle t_1\ a_1\rangle\langle t_1\ a_2\rangle\cdots\langle t_k\ a_k\rangle[C_r]$

Type 3:
$\langle t_1\ a_1\rangle\langle t_2\ a_2\rangle\cdots\langle t_i\ a_i\rangle[C_j]\langle t_k\ a_k\rangle\cdots\langle t_n\ a_n\rangle$

Type 1, 2, 3은 자체가 하나의 개념을 나타내는데 상위 개념이나 하위 개념들은 Type 1, 2, 3의 조합으로 구성되어 있다. 최상위 레벨 개념은 다시 여러 개의 하위 개념들을 포함하게 되는데 최상위 레벨 개념으로부터 포함될 수 있는 하위 레벨 개념들을 [표 2-2]에 일부를 나타내었다.

최상위 레벨 개념	하위 레벨 개념
[give_info]	[airline][busy][go][guide]...
[i_want]	[about][depart][end_point]...
[i_will]	[because][by][package]...
[nicety]	[calling][confirm][return]...
[query]	[depart][differ][pay][plan]...
[request]	[about][busy][call][know]...
[respond]	[about][cost][detailed][go]...
[*othertop*]	[alphabet][range][time]...

[표 2-2] 최상위개념이 포함하는 하위 개념

3 동사 비트패턴 구성과 적용

기존 한국어 대화체 파싱에서 첫 번째 문제점의 원인은 탐색공간을 제약할 수 있는 방법의 부재를 들 수 있는데 탐색공간을 제약하기 위해서는 입력 문장으로부터 탐색공간 제약 가능한 단서를 찾아내어야 한다. 개념에 기반한 문법의 구성이 동사를 중심으로 한 개념들의 조합으로 구성되어 있다는 점을 착안할 때 탐색공간의 제한 방법 또한 개념들에 포함된 동사의 정보들을 체계화하고 이를 적용할 수 있는 메카니즘을 고안해 냄으로써 해결할 수 있다. 두번째 한국어 대화체 파싱의 문제점의 원인은 개념 단위로 대화체를 분리하기가 어렵다는 점인데 개념단위로 대화체를 분리하기 위해서는 개념을 분리할 수 있는 문어체에서의 구두점과 같은 정보를 이용해야 한다. 개념기반 대화체 분석 시스템은 문장의 구분을 하나의 단위 개념들로 분리함으로 개념들의 분리가 문장의 분리를 의미하게 될 것이다. 따라서 개념을 분리해 낼 수 있는 방법을 통해서 문장의 경계를 구분할 수가 있고 문장 경계 분리는 여러 개의 문장으로 구성된 단일 문장을 다시 여러 개의 개념으로 분리하여 파싱을 수행함으로 해서 과도한 탐색공간 생성을 막을 수 있는 메커니즘이 될 것이다. 이를 위해서 최상위 개념과 동사토큰간의 포함관계를 구성해야 하는데 [표 2-2]에 구성된 정보를 기반으로 해서 포함된 개념들에 속한 동사토큰을 분리해 낸다. 사용된 토큰들 중 토큰명이 verb인 것을 추출해 내고 $\langle$ verb attr$_i\rangle$ 가 최상위 레벨 개념에 포함되는지를 총 167개의 동사 토큰에 대해서 분석하였다. [표 3-1]은 167개의 동사토큰 중에서 일부분만을 나타내었다.

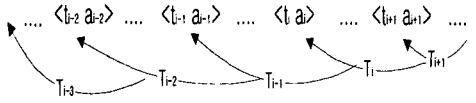
동사토큰	give_info	i_want	i_will	nicety	query	request	respond	othertop
<verb 가르치>	1	1	1		1	1		
<verb 감사하>						1	1	
<verb 같>	1							
<verb 계시>				1				
<verb 계획하>	1				1			
<verb 고맙>						1	1	
<verb 곤란하>	1				1			
<verb 그러하>			1					
<verb 받>	1	1	1		1	1	1	
<verb 보내>		1	1		1	1		
<verb 자세하>	1	1	1	1	1	1	1	
<verb 전화주>	1	1	1	1	1	1	1	

[표 3-1] 최상위개념과 동사토큰의 포함관계

발화문이 한국어 분석 시스템의 입력문장으로 입력이 되면 형태소 분석을 거치고 토큰열들로 생성되게 된다. 형태소 분석을 거친 토큰열의 형태는 다음과 같이 일반화된다.

$\langle \text{token}_1\ \text{attr}_1\rangle\langle \text{token}_2\ \text{attr}_2\rangle\cdots\langle \text{token}_n\ \text{attr}_n\rangle$

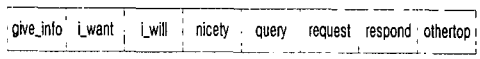
중심어인 동사토큰을 찾는데 한국어의 특성인 중심어 후치 특성을 이용해서 오른쪽에서 왼쪽으로 동사 토큰을 찾아간다.[6] 처음으로 만나는 동사토큰의 비트 패턴을 개념기반을 통해서 얻은 정보로부터 구하고 다음 동사토큰을 찾아가게 된다. 이렇게 해서 찾은 두번째 동사토큰에 대한 비트패턴을 구하게 된다. [그림 3-1]은 동사토큰의 비트패턴을 통해서 문장의 개념 패턴을 구하는 방법을 설명한 것이다.



$T = \bigcap_{i=1}^{n} T_i$ (단, T_i 는 동사의 *bit pattern*
T는 문장의 *concept pattern*)

[그림 3-1] 문장의 개념 패턴 결정

찾아진 두개의 동사토큰의 비트패턴을 AND연산을 하고 이를 통해서 얻어진 비트패턴의 의미는 두개의 동사가 비트패턴에 나타난 최상위 레벨에 포함될 수 있음을 의미하게 된다. 만일 두 동사 토큰의 비트패턴의 AND연산이 0으로 나타났다면 두개의 동사는 하나의 최상위 개념으로 결합될 수 없음을 의미하는 것이고 이는 문장이 두개의 개념으로 분리될 수 있음을 나타낸다. 비트패턴의 구성은 [그림 3-2]와 같다.



[그림 3-2] 비트 패턴의 구성

비트패턴 10000000의 의미는 동사토큰이 [give_info]에만 나타남을 의미한다.

$\langle token_i\ attri_i \rangle$: 00000110
$\langle token_j\ attr_j \rangle$: 11111110

i > j이면 i번째 token은 j번째 나오는 $token_j$ 다음의 동사토큰이다. AND연산을 통한 비트패턴 00000110은 최상위 레벨개념 [request]와 [respond]에 동시에 포함될 수 있다. 여기에서 동사토큰 $token_i$와 $token_j$의 $\langle token_{i+1}\ attr_{i+1} \rangle$와 $\langle token_{j+1}\ attr_{j+1} \rangle$은 어미를 나타내는데 일반적인 형태는

$\langle token_{i+1}\ attr_{i+1} \rangle \Leftrightarrow \langle eomi\ attr_{i+1} \rangle$
$\langle token_{j+1}\ attr_{j+1} \rangle \Leftrightarrow \langle eomi\ attr_{j+1} \rangle$

와 같은 형태이다. 탐색 공간을 줄이는데 있어서 동사토큰 비트패턴과 더불어 중요한 역할을 한다. 형태소 분석을 거친 어미들의 attribute값($attr_i$) 들은 [표 3-2]와 같이 분류되는데 이 정보와 동사토큰 비트와 AND연산을 통해서 가장 작게는 하나의 탐색 공간를 줄일 수 있다.

어 미	비트 패턴
<eomi decl>	11111111
<eomi will>	00100110
<eomi want>	01001100
<eomi request>	00000100
<eomi quest>	00001000
<eomi adj>	11111111
<eomi is>	11110111

[표 3-2] 어미 토큰의 비트패턴

만일 $\langle token_{i+1}\ attr_{i+1} \rangle$의 어미 정보 비트패턴이 00000100인 청유형일 때 동사토큰 $\langle token_i\ attr_i \rangle$ 와의 AND연산에 의한 결과는 00000100으로 두개의 최상위 레벨 개념 [request], [respond]가 다시 하나의 최상위 레벨 개념 [request]로 어미 정보를 이용함으로서 탐색공간이 줄어들 수 있음을 알 수가 있다. 다시 $\langle token_{i+1}\ attr_{i+1} \rangle$와 $\langle token_{j+1}\ attr_{j+1} \rangle$의 비트 패턴정보가 아래와 같이

$\langle token_i\ attr_i \rangle$: 00000110
$\langle token_j\ attr_j \rangle$: 11111000

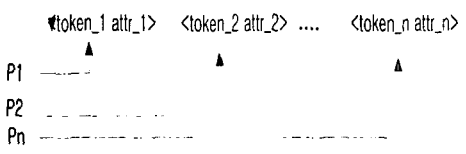
라 할때 두 동사토큰의 비트패턴의 AND연산의 결과는 다음과 같다.

$\langle token_i\ attr_i \rangle$: 00000110
AND $\langle token_j\ attr_j \rangle$: 11111000
00000000

두 동사토큰의 비트패턴 AND연산의 결과는 00000000이 되는데 이 결과 값이 의미하는 바는 $\langle token_i\ attr_i \rangle$ 와 $\langle token_j\ attr_j \rangle$ 의 동사는 하나의 최상위 레벨 개념으로 이루어질 수 없음을 의미하고 토큰 $\langle token_i\ attr_i \rangle$ 와 토큰 $\langle token_j\ attr_j \rangle$ 사이에는 단일 문장 단위의 개념으로 분리된다는 것을 의미하게 된다. 이러한 정보를 기반으로 하여 탐색 공간을 제한할 수 있는데 전체 알고리즘은 [알고리즘 3-1]과 같이 표현된다.

Set V : 동사 토큰들의 비트패턴 집합
Set E : 어미 토큰들의 비트패턴 집합

P_i : *TokenPosition*

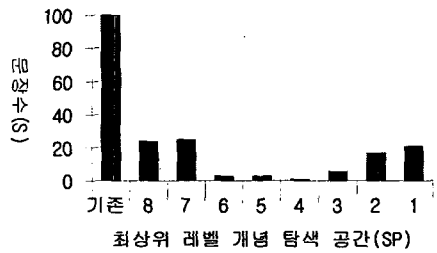



```
/* 오른쪽에서 왼쪽으로 검사하기위한 초기값*/
i ← n

/* 개념 분리가능 여부 비트패턴 */
T1 ← 11111111

while ( Pi ≠ P0 )
{ /* token position n에서 0까지 position number
        를 감소시키면서 알고리즘을 적용한다*/

  /* token 명이 동사인지 여부를 체크 */
    if ( tokeni = verb) then
                        {
        b[i] ← attri의 비트패턴 ∈V
        e[i] ← attri+1의 비트패턴 ∈E

/* 동사정보 비트패턴과 어미정보 비트패턴을 AND */
        Si ← attri and attri+1
/* 이전 동사토큰의 비트패턴과 AND */
        Ti ← Ti and Si

        /* 개념이 분리됨을 의미 */
        if ( Ti = 0 ) then
        {
            SP[i] ← Pi+2
            Ti ← 11111111
        }
      }
      i ← i-1
}
SP[i] ← P1

/* 파싱을 해나가는 루틴 */
for i = 1 to n
{
    if( SP[i] ≠ NULL) then
        Parsing( SP[i])
}
```

[알고리즘 3-1] 탐색공간을 제약하는 알고리즘

4 실험 및 분석

본 논문의 실험은 전화상의 대화를 전사한 '여행안내' 영역의 말뭉치[1]를 대상으로 동사를 포함하고 최상위 레벨 개념으로 분리될 수 있는 개념단위 문장 100개를 선정하여 최상위 레벨 개념으로의 탐색공간을 조사해 보았다. [그림 4-1]은 실험 결과를 나타내 주고 있다.



[그림 4-1: 실험결과]

기존의 대화체 파싱에서 탐색공간을 찾아가는 방식은 모든 8개의 최상위 개념을 찾아갔으나, 제안된 방법으로 수행했을 경우에

$$\frac{\sum_{i=1}^{8} SP_i \times S_i}{(\text{기존 } SP) \times S} = 0.61$$

로서 약 40%정도의 탐색공간을 줄일 수 있었다. 최상위 개념에서 하위 개념으로 파싱을 해 나간다는 점을 고려할 때 전체적인 탐색공간의 제약은 40%이상이 나올 것으로 기대된다.

분석 예)

입력문장 :

예 그럼 제가 그 팩스를 받구요 바로 자세한 내용을 보내 드리겠습니다 예 전화 주셔서 감사합니다 .

<yes 예><noun 팩스><verb 받><eomi after> <adv 바로>
SP[1] → SP[2]

<verb 자세하><eomi adj> <noun 내용><verb 보내><eomi wil>

<yes 예> <verb 전화주> <eomi because> <verb 감사하><eomi decl>
SP[11] → SP[12]

1) 여행안내 영역의 ETRI-Corpus 1575개의 발화문

〈verb 감사하〉	: 00000110
〈eomi decl〉	: 11110011
〈verb 전화주〉	: 11111110
AND 〈eomi because〉	: 11111111
SP[12]	00000010

SP[12] = [respond]

〈verb 보내〉	: 01101100
〈eomi will〉	: 00100100
〈verb 자세하〉	: 11111110
AND 〈eomi adj〉	: 11111111
SP[2]	00100100

SP[2] = [i_will], [request]

분석결과 :

[respond](<yes 예>)

[i_will]([after]([receive](<noun 팩스><verb 받>)<eomi after>)[send](<directly 바로>[detailed](<verb 자세하><eomi adj>)<noun 내용><verb 보내>)<eomi will>)

[respond](<yes 예>)

[respond]([reason]([calling](<verb 전화주>)<eomi because>)[thankyou](<thanks 감사하>)<eomi decl>)

5 결론

본 논문에서는 개념 기반 문법에 기술된 동사정보를 비트패턴 형태로 표현하고 이를 한국어 대화체 분석 시스템이 파싱을 할 때 탐색공간을 제약시키는 방법을 제시하였다. 동사가 포함된 최상위 레벨 개념의 수와 동사의 수에 대한 것은 [표 5-1]에 나와 있다.

구 분	8개	7개	6개	5개	4개	3개	2개	1개
동사의수	5	77	28	10	12	3	11	21

[표 5-1] 비트패턴에서 비트의 수와 동사의 수

[표 5-1]에서 볼 수 있듯이 8개의 최상위 레벨 개념에 모두 포함되는 전체 동사토큰의 3%를 차지한다. 탐색공간의 3%만이 8개의 모든 최상위 레벨 개념에 포함된다는 것이다. 하나의 개념으로 묶일 수 있는 동사토큰들이 n개일 때 n-1개가 전체의 68%을 차지하는 최상위 레벨 개념 7~6의 비트 패턴을 갖더라도, 하나의 토큰이 최상위 레벨 개념을 하나만 포함한다면 탐색공간은 유일한 하나로 제한된다. 그러나 대화체 문장의 특성인 간투어 사용과 반복 발화문 등이 토큰열을 형성할 때 탐색공간 제약을 가하는 비트 패턴 형태가 나타난다면 기대하는 결과를 얻어낼 수 없다. 따라서 자연 발화문의 특성을 고려하여 설계된 탐색공간 제약에 대한 연구가 향후 연구과제로 이루어져야 할 것이며, 문법을 구성할 때 동사정보의 독립성을 유지하는데 초점을 맞추어 구성이 되어야 할 것이다.

참고문헌

[1] 서영훈, '음성 언어 번역을 위한 개념 기반의 한국어 분석 및 생성', 정보과학회 논문지, 1996.11

[2] 서영훈, '대화체 및 문어체 기계 번역을 위한 한국어 구문/의미 해석시스템 개발', 연구보고서, 한국전자통신연구소, 1995

[3] 최재웅, '대화분석에 있어서의 몇가지 문제: 호텔 예약 전화 대화를 중심으로', 한글 및 한국어정보처리, 1996.10

[4] 왕지현, 서영훈, '개념 및 구문 정보를 이용한 한국어 대화체 분석 시스템', 한글 및 한국어정보처리, 1997.10

[5] 왕지현, '구문 정보를 이용한 개념기반의 한국어 대화체 분석기', 충북대학교 석사학위 논문, 1998. 2

[6] 김영택, '자연언어 처리', 교학사, 1994. 5

[7] 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', Proceeding of TMI95, 1995

[8] B.Suhm, P.Geutner, A.Lavie, L.Mayfield, 'JANUS:TOWARDS MULTILINGUAL SPOKEN LANGUAGE TRANSLATION', Interactive System Laboratories, Carnegie Mellon University