□신기술해설□

부분 구문분석 방법론

김 재 훈†

◈ 목 차 ◈

1. 서 론

강건하고 정확한 자연언어처리 시스템은 음성 합성과 문체 검사에서부터 메시지 이해와 자동번역에 이르기까지 아주 다양한 분야에 응용될 수 있다. 이를 위해서 많은 연구자들이 부단한 노력에도 불구하고, 영역에 무관하고 실용적인 구문분석기는 개발되지 않았다[15]. 구문분석에 고질적인 문제는 다음 세 가지로 요약될 수 있다. 첫째, **분석단위의 분리 문제**이다. 일반적으로 자연언어 문장은 적절한 단위로 분리하고, 분리된 단위의 합성에 의해서 문장의 구조나 의미를 분석한다. 그러나, 자연언어 문장에서는 이들의 분리를 명확하게 정의하기 어렵다. 예를 들면, 다중단어에 속하는 "on account of, because of"는 여러 개의 단어로 구성되었지만, 구문적으로는 하나의 역할을 수행한다. 한국어 처리에서 이 문제는 더욱 심각하다. 예를 들면 각 연구자들에 의해서 정의된 형태소, 즉, 형태소분석기의 결과가 서로 다르다. 둘째, **중의성 문제**이다. 중의성 문제는 자연언어처리에서 가장 고질적인 문제 중 하나이다. 대부분의 자연언어처리 시스템은 언어적인 현상에 따라 여러 단계(형태소분석, 구문분석, 의미분석, 담화분석)로 나뉘어서 처리된다. 각 단계에서는 이용 가능한 언어정보를 이용해서 다양한 형태의 중의성을 해결하고 있다. 셋째, **미등록어 문제**이다. 미등록어 문제는 시스템에서 알 수 없는 단어나 문법 등이 입력되었을 경우에 합리적인 방법을 이용해서 추정해야 한다. 최근 인터넷 문서를 대상으로 하는 여러 응용 시스템(인터넷 검색 엔진)이 등장하면서 이 분야에 대한 연구가 더욱 활발히 진행되고 있다.

부분 구문분석은 이러한 문제를 다소 완화시키기 위해서 구문분석의 복잡한 문제를 단계적으로 해결하고자 하는데 그 목적이 있다. 즉, 가장 분명한 구문구조를 먼저 분석하고, 분석된 구조를 이용해서 좀더 복잡한 구조를 분석하는 방법이다. 응용분야에 따라 부분 구문분석의 정의는 조금씩 다르지만 모든 구들 중에서 연속적이고, 비재귀적인 구성성분에 해당하는 단순한 구를 인식하는 것으로 정의할 수 있다.[3][7].

부분 구문분석의 필요성은 아래와 같이 요약될 수 있다. 첫째, **견고한 분석이 가능하다**. 구문분석은 완전한 문법을 가진 상황에서 주어진 문장 구조를 밝히는 것이므로 입력 문장이 주어진 문법에 벗어난다면 분석에 많은 어려움이 있다. 더구나, 많은 실생활 문장에서 오류를 다소 포함하고 있기 때문에, 문법이나 사전은 항상 불완전하다고 가정되어야 한다. 일반적인 문장에 대해 완

† 정회원 : 한국해양대학교 컴퓨터공학과 조교수

전한 분석은 조금 희생되더라도 효율적이고 확실한 구문구조를 찾기 위해서 부분 구문분석을 사용한다. 둘째, **응용 분야가 많다**. 부분 구문분석만으로도 이용할 수 있는 응용 분야들이 많이 있다. 예를 들면, 정보검색[37], 정보추출[25], 정보요약[26], 언어정보 추출 도구[24] 등이 그 예이다. 셋째, **부분 구문분석으로 완전한 구문분석이 가능하다**. 부분 구문분석은 이론적이라기보다는 아주 실용적 측면에서 다루어지며, 부분 구문분석 기법을 여러 번 적용함으로써 완전한 구문분석이 가능하다[6][14]. 예를 들면, 부분 구문분석 기법 중 하나인 유한상태 오토마타를 이용해서 단어들로부터 말덩이를 구성하고, 말덩이들로부터 구를 구성하고, 구들을 이용해서 문장의 구조를 분석할 수 있다.

본 논문은 다음과 같이 구성된다. 제2장에서는 기존 연구들에서 정의된 부분 구문분석의 단위에 대해서 기술한다. 제3장에서는 여러 종류의 부분 구문분석 방법에 대해서 소개하고, 제4장에서는 부분 구문분석 시스템의 응용 사례에 대해서 기술한다. 끝으로 제5장에서 부분 구문분석에 대한 앞으로의 연구 방향과 한국어에서 부분 구문분석의 역할에 대해서 간단히 기술하고자 한다.

2. 부분 구문분석의 단위

부분 구문분석 시스템에서 인식 대상이 되는 부분 구문분석의 기본 단위는 여러 연구자들에 의해서 매우 다양하게 정의되고 있다. 예를 들면, 말덩이(chunk)[6], 기저명사구(baseNP)[33], 최장명사구(maximal length NP)[11][14], 구문요소[4] 등이 부분 구문분석의 단위들이다. 앞에서도 언급했듯이 부분 구문분석은 응용분야와 밀접한 관계를 가지고 있기 때문에 모든 분야에 잘 적용되는 부분 구문분석의 단위를 결정하는 것은 대단히 어려운 일이라고 생각한다. 본 장에서는 기존 연구에서 부분 구문분석의 단위에 대해서 고찰하고자 한다.

2.1 말덩이

말덩이[6]는 중의성을 가지는 문의 구성성분에 대해서 중심어에 부착하지 않은 상태의 구문분석 조각이다. 즉, 구문적인 문장 구성성분이다. 영어의 경우, 명사구만이 말덩이에 포함되는 것이 아니라, 전치사구, 수식어구, 접속어구, 동사의 필수성분 등이 모두 여기에 속한다. 말덩이는 아래와 같은 성질을 가지고 있다.

1. 절(clause)은 말덩이의 연속으로 구성된다(no discontiguous).
2. 말덩이 내에 말덩이가 속할 수 없다(no recursive).
3. 말덩이는 반드시 연속적이다(no centerembedded).

아래는 말덩이의 예이고[6],[x]로 표시된 단어열은 하나의 말덩이를 나타내며, NX는 명사구, VX는 동사구, INF는 부정사구를 의미한다[1].

(1) [NX we] [VX lack] [NX the ways] and [NX means] [INF to ...]

(2) [NX the Ways and Means Committee]

2.2 기저명사구

영어에서 명사구는 중심어를 기준으로 해서 앞이나 뒤에서 수식어구(형용사구, 또 다른 명사구, 전치사구 등)로부터 수식을 받을 수 있다. 어떤 명사구의 범위는 정확하게 결정하는 문제는 완전 구문분석 문제와 거의 같다. 예를 들면 복합명사구를 결정하는 문제에 대해서도 많은 의미론적인 지식이 요구된다[39]. 또한 명사구의 내부 구조를 결정할 경우에는 중심어들의 수식관계를 결정하기 위해서 많은 어휘 지식이 요구된다[29]. 그러나, 정보검색이나 정보추출을 위해서는 복잡한 언어적인 지식을 이용할 경우, 실시간 처리 뿐만 아

1) 말덩이에 대한 상세한 설명은 [7]를 참고하기 바란다.

니라 강인한 시스템을 구현하기가 매우 어렵다. [33]에서는 비교적 간단한 구문지식을 이용해서 명사구를 인식할 수 있도록 기저명사구를 정의하였으며, 다음과 같은 제약조건을 가지고 있다.

1. 중심어에 해당하는 명사를 포함한다.
2. 관사를 포함한 한정사들이 비재귀적으로 포함된다.
3. 뒤에서 수식하는 구나 절(전치사구, 관계절)은 포함되지 않는다.

구문트리 부착 말뭉치에서 내포 명사구를 포함하지 않는 명사구를 추출함으로써 기저명사구 부착 말뭉치를 쉽게 구축할 수 있다. 즉, 구문트리 부착 말뭉치를 가지고 있을 경우에 기저명사구 부착 말뭉치를 구축하는 일은 어려운 일이 아니다. 다음은 기저명사구에 대한 몇 가지의 예이다.

(3) During [N the third quarter N], [N Compaq N] purchased [N a former Wang Laboratories manufacturing facility N] in [N Scotland N], which will be used for [N international service and repair operations N].

(4) [N The government N] has [N other agencies and instruments N] for pursuing [N these other objectives N].

2.3 최장명사구

최장명사구는 기저명사구와는 달리 앞에서 수식하는 수식어 뿐 아니라 뒤에서 수식하는 수식어를 포함한다. 최장명사구가 정확하게 인식되면, 완전한 구문분석은 인식된 명사구들의 부착만 고려하면 된다. 아래에는 불어에 대한 최장명사구의 예이다[12].

(5) mesure du debit du ventilateur dextraction avec trappe en position fermee

품사열 : noun prep det noun prep det noun prep noun prep noun prep noun adj

2.4 구문요소

한국어에서 구문요소는 **문장의 구성성분**을 말하며, 이는 구문분석의 기본 단위가 된다[4]. 구문

〈표 1〉 부분 구문분석 시스템의 개요

출처 및 시스템	방법론	분석단위	재현률(%)	정확률(%)
[11]	단순규칙	명사구		
[23] -- Fidditch	수정된 구구조 문법			
[40] -- NPTool	수정된 구구조 문법 유한상태 오토마타	명사구	98.5	95.0
[33]	변환기반 학습 변환규칙	기저명사구 명사구/동사구	92.0 88.0	92.0 88.0
[18] -- PARTS	HMM	비재귀적명사구	98.0	
[27] -- Supertag	LTAG HMM(Viterbi 탐색)	말덩이		77.3
[36] -- Chunk Tagger	HMM	말덩이		81.3
[14]	CMM	말덩이	84.8	91.4
[35]	최대 엔트로피 모델	말덩이		84.2
[19]	메모리기반 학습 유사도기반 인식	명사구 동사구	94.0	93.7
[16]	메모리기반 학습 최장일치 규칙	기저명사구	94.0	94.0
[17]	HMM 유한상태 오토마타	말덩이 최장명사구	96.0	95.0

요소는 문의 의미를 알리는 구문정보를 지니고 있어야 하며, 구문요소는 주어, 서술어, 목적어, 부어, 관형어, 부사어, 독립어로 분류된다. 구문요소는 하나의 단어로 구성되는 경우도 있지만, 경우에 따라서는 구나 절이 될 수도 있다. [4]에서는 후자와 같이 광범위한 구문요소를 다루는 것이 아니라, **하나의 의미를 가지는 최소단위**를 구문요소로 정의하여 처리하였다. 아래는 구문요소의 예이다.

(6) *[과학이내] [공학의] [여러] [**분야 중에서**] [컴퓨터는] [가장] [급속히] [발달되는] [분양의] [하내[이대.*

(7) *[1946년] [ENIAC으로] [시작된] [컴퓨터는] [현재는] [**헤아릴 수 없을**] [만큼] [많은] [종류개] [**나와 있으며**] [그] [기능도] [다양하다.*

3. 부분 구문분석에 대한 방법론

부분 구문분석 방법은 크게 규칙기반 방법[33], 통계기반 방법[18], 기계학습 방법[19], 혼합 방법[17]으로 나눌 수 있으며, <표 1>은 부분 구문분석 시스템의 특성을 요약한 것이다.

3.1 규칙기반 부분 구문분석 방법

3.1.1 말덩이와 경계어

말덩이의 경계로서 기능어 혹은 불용어를 이용하는 방법이며, 가장 간단한 방법이다. 여기서, 경계가 되는 말들(주로 기능어)을 경계어(chicks)라고 하였고, 경계와 경계 사이의 말들(주로 내용어)을 말덩이라고 하였다. [11]에서는 이 방법을 불어에서 명사구를 찾는 데 사용하였다. 경계어는 명사구에 속하지 않는 모든 단어들이다. 예를 들면, 동사, 대명사, 접속사, 전치사, 한정사 등이다. 경계와 경계 사이에 있는 단어열은 말덩이, 즉, 명사구이다. 이 방법은 말덩이에 해당하는 기술용어를 정확하게 추출하기 위해서 대량의 품사 패턴이 필요할 것이다.

3.1.2 Fidditch

Fidditch[23]는 원래 부분 구문분석을 목적으로 개발한 것은 아니고, 비제한적인 문장을 분석하기 위해서 개발되었으며, Marcus 구문분석기[32]에 펀트(punt)라고 하는 행위를 첨가하였다. 펀트는 구의 역할을 정확하지 않을 경우에 해당하는 구를 입력에서 제거하고(skip and fit), 즉, 부착을 하지 않은 상태로 두고 파싱을 계속 진행하는 방법을 말한다. 파서는 절의 핵심요소(보통 절의 경계표지, *that, which* 등), 주어, 술어, 펀트된 절에 부착된 요소들을 인식한다. Fidditch는 대량은 문장 즉 말뭉치에 대한 구문분석에 매우 유용하다(매우 빠르기 때문에). 또한 파서가 결정적으로 수행되기 때문에 유한상태 오토마타를 이용해서 구현할 수 있다.

3.1.3 수정된 구구조 문법

NPTool[40]은 ENGCG(English Constraint Grammar)를 이용해서 부분 구문분석을 수행하며, 형태소분석에서 각 단어에 품사 정보 이외에 구문 정보를 할당하여 부분 구문분석을 수행한다. 규칙은 정규표현으로 기술되며 항상 일관된 분석이 가능하도록 하였다. (8)은 ENGCG를 이용해서 명사구를 인식한 예이다. 명사를 추출하기 위해서 (8)의 결과를 바탕으로 하는 정규표현을 이용해서 (9)과 같은 명사를 추출한다.

(8) *the/@>N inlet/@>N and/@CC exhaust/@>N manifolds@/NH are/@V mount/@V on/@AH opposite/@>N sides/@NH of/@>N the/@>N cylinder/@>N head/@NH*

(9) *the*
np: inlet and exhaust manifold
are mounted on
np: opposite sides of the cylinder head

여기서, @>N, @CC, @NH 등은 명사구를 추출하기 위한 구문표지이다.

3.1.4 유한상태 오토마타

유한상태 오토마타를 이용하는 방법은 여러 연구자들에 의해서 수행되었다[8][9][20][22][27][30][34]. 이들의 공통점은 계층형 유한상태 오토마타를 이용한다는 것이다. 계층형 유한상태 오토마타는 유한상태 오토마타가 여러 층을 이루고 있는 경우를 말하며, 각 층에서는 정규표현에 의해서 특별한 구를 인식한다. 아래에 그 구체적인 예를 보이고 있다[7].

1: NP → D? A* N+ | Pron
VP → Md Vb | Vz | Hz Vbn | Bz Vbn | Bz | Vbg
2: PP → P NP
3: SV → NP VP
4: S → (Adv | PP)? SV NP? (Adv | PP)*

여기서 번호는 층을 의미하며, 층 0는 품사 태깅의 결과이다. 층 s에서는 층 s-1의 결과에 대해서 정규표현을 적용한다. 주어진 입력은 여러 개의 정규표현에 일치될 수 있다. 이 경우에는 일반적인 정규표현에서 중의성을 해결하는 방법과 같이 최장일치된 정규표현을 선호하도록 한다. 어떤 정규표현에도 일치되지 않는 입력은 무시되어 그대로 출력된다. 중의성 해결하는 다른 방법을 HMM을 이용하는 방법을 사용할 수도 있다[17]. [20]의 결과에 따르면 HMM을 이용한 방법이 더 좋은 성능을 보였다[2]. 그러나 최장일치 방법을 전체 문장을 분석하지 않고도 매우 신뢰성 높게 구를 찾을 수 있으며, 계층형 파서는 신뢰성이 가장 높은 구에 대해서 먼저 수행하여 점점 더 큰 구를 형성해 갈 수 있다. 또한 이 방법은 추가적인 정보를 쉽게 첨가할 수 있기 때문에 유연성이 매우 좋다.

3.1.5 변환기반 방법

이 방법은 변환기반 학습 방법을 부분 구문분석에 적용한다[33].

기저명사구를 인식하는 문제에서는 말덩이 표지는 {I, O, B}[3]이고, 명사군과 동사군을 인식하는 문제(NV 말덩이)에서는 {BN, N, BV, V, P}[4]이다. (10)은 기저 명사구의 예이고, (11)은 NV 말덩이의 예이다.

(10) Even/O Moa/I Tse-tung/I 's/B China/I begin/O in/O 1949/I with/O a/I partnership/I between/O the/I communist/I and/O a/I number/I of/O smaller/I ,/O non-communist /I parties/I ,/O

(11) Indexing/BN for/BN the/N most/N part/N has/BV involved/V simply/V buying/V and/BV then/V holding/V stocks/BN in/BN the/N correct/N mix/N to/BV mirror/V a/BN stock/N market/N barometer/N ./P

먼저, 초기 시스템은 학습말뭉치의 각 단어나 품사에 대해서 가장 높은 빈도수를 갖는 말덩이 표지를 할당한다. 그리고 나서 변환기반 학습 방법에 의해서 추출된 변환규칙을 이용해서 초기 시스템의 결과를 수정한다. Penn Treebank를 대상으로 변환기반 학습을 통해 추출된 변환규칙은 (그림 1)와 (그림 2)과 같다. (그림 1)은 기저명사구 말덩이 규칙의 일부이고, (그림 2)는 NV 말덩이 규칙의 일부이다. 그림에서 T_i는 현재 단어를 중심으로 i위치에 있는 말덩이 품지를 의미한다.

2) 명사구에 대해서 유한상태 오토마타 모델은 97.8%의 정확률을 보였고, 통계적인 방법은 98.6%의 정확률을 보였다. 또한 모든 비재귀적 구에 대해서 유한상태 오토마타 모델은 87.0%의 정확률을 보였고, 통계적인 방법은 93.5%의 정확률을 보였다.

3) I : 기저명사구에 속하는 단어, O : 기저명사구에 속하지 않는 단어, B : 연속적인 기저명사구가 있을 때, 시작하는 단어.

4) BN : 명사군의 시작 단어, N :명사군 내의 단어, BV : 동사군의 시작 단어, V : 동사군 내의 단어, P: 문장 부호. P는 명사군이나 동사군 밖에 있으나, 경우에 따라서는 명사군이나 동사군 내에 어디에나 속할 수 있다.

단계	잘못된 태그	문 맥	올바른 태그
1.	I	T_1 = O, P_0 = JJ	O
2.	-	T_{-2} = I, T_{-1} = I, P_0 = DT	B
3.	-	T_{-2} =O, T_{-1} = I, P_{-1} = DT	I
4.	I	T_{-1} =I, P_0 = WDT	B
5.	I	T_{-1} =I, P_0 = PRT	B
6.	I	T_{-1} =I, W_0 = who	B
7.	O	T_{-1} =I, P_0 = CC, P_1 = NN	I
8.	O	T_1 =I, W_0 = &	I
9.	O	T_{-1} =I, P_0 = CC, P_1 = NNS	I
10.	O	T_{-1} =O, W_0 = about	I

(그림 1) 기저명사구 추출을 위한 변환규칙

단계	잘못된 태그	문 맥	올바른 태그
1.	BN	T_{-1} = BN, P_0 = DT	N
2.	N	T_{-1} =Z, W_{-1} = ZZZ	BN
3.	N	T_{-1} =P, P_{-1} =','	BN
4.	BN	T_1 =V, P_1 = VB	BV
5.	N	T_{-1} =BV, $P_{-1,-2,-3}$ = VBD	BN
6.	N	P_{-1} =VB	BN
7.	BV	T_{-1} =V, $P_{-1,-2,-3}$ = RB	V
8.	V	T_{-1} =N, $P_{-1,-2,-3}$ = NN	BV
9.	BV	T_{-1} =BV, $P_{1,2,3}$ = VB	V
10.	BN	T_{-1} =BN, P_0 = PRP$	N

(그림 2) NV말덩이 추출을 위한 변환규칙

예를 들면, T_{-1}은 현재 단어 바로 뒤에 있는 단어의 말덩이 표지이다. 이와 비슷하게 P_i와 W_i도 각각 품사 태그와 단어를 의미한다.

3.2 통계기반 부분 구문분석

3.2.1 PARTS

PARTS[18]는 통계적인 방법의 시초이며, 품사 태깅 결과를 입력으로 하여 비재귀적 명사구(기저명사구)를 찾는 것으로 목적으로 하였다. 품사 태깅 결과를 HMM에 적용하기 위해서 단어(품사)에 대응하는 표지 [,],][, λ를 사용하여 명사구를 인식하였으며, (그림 3)는 명사구 인식을 위한 HMM이고, 이를 이용한 시스템의 최종 결과는 (12)과 같다[5].

상태 (출력)		[	λ	]	λ	[	]
품사	$	DT	NN	VBD	IN	NN	CS
단어		the	prosecuter	said	in	closing	that

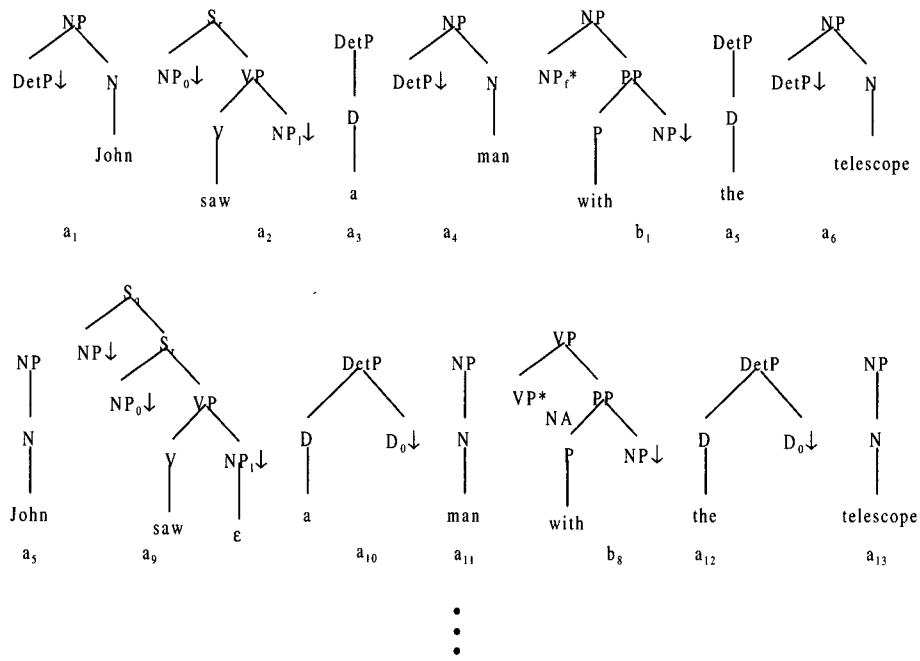
(그림 3) PARTS에서 명사구 추출을 위한 HMM

(12) [The/AT company/NN] is/BEZ cooperating/VBG in/IN [the/AT investigation/NN] ,/, [Tucker/NP] said /VBD ./.

3.2.2 Supertag

Supertag[27]는 LTAG(lexicalized tree-adjoining grammar)를 사용한다. LTAG는 각 기본 트리에 어휘정보가 포함된다. LTAG의 대치과 부가는 의존그래프에서 부착과 같은 연산이다. 각 단어는 여러 개의 기본 트리를 가질 수 있으며, 이들 트리들은 다른 문법구조와 의존소의 다른 집합을

5) (11)에서 사용된 품사표지(part-of-speech tag)는 을 참조하기 바란다.

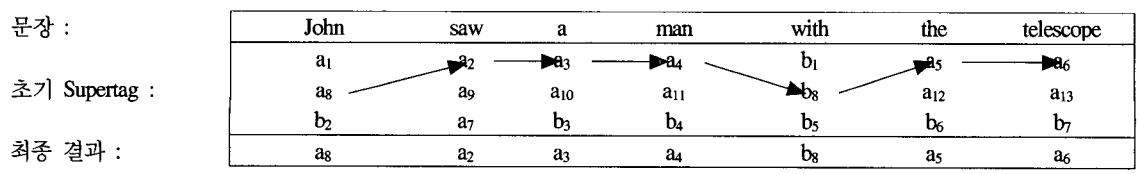


(그림 4) LTAG의 기본 트리

가지고 있다(그림 4). 기본 트리를 supertag라고 하며, 이것은 [40]의 구문정보와 매우 유사하다. 부분 구문분석은 이를 여러 개의 트리 중에서 주어진 문장에 적합한 하나의 트리를 선택하는 과정이며, supertag를 검사하고, Viterbi 탐색을 수행함으로써 파싱이 수행된다(그림 5). 이 시스템은 부분 구문분석을 위해서 개발된 것은 아니라, 품사 태깅을 위한 것이다. 그러나 품사 태깅은 결과는 LTAG의 개념에 의해서 조합하면 완전 구문분석 혹은 부분 구문분석과 같은 결과를 가져온다.

3.2.3 Chunk Tagger

Chunk Tagger[36]는 말뭉치가 작을 경우에 실세계의 문장분석을 목적으로 하고 있으며, 언어에 독립적인 모델이다. [18]에서는 비재귀적 명사구를 인식하나, 이 방법은 복잡한 명사구, 즉, 재귀적 명사구도 인식할 수 있도록 설계되었다. 인접한 단어들 사이에 괄호를 생성하는 것이 아니라, 구조적 관계를 생성하는 문제로 모델링하였다. 즉, 주어진 단어열 $<w_0, w_1, ..., w_n>$에 대해서 구조적 관계열 $<r_1, r_2, ..., r_n>$을 찾는 문제로 모델링하였다. 여기서 확률 변수 r_i는 {0, +, ++, -, --, =, 1}의 값을 가질 수 있다. (그림 6)는 부호화된 각 구조적 관계에 대응하는 실제 구문 구조를 보이고 있다. 이 구조는 인접한 단어 사이에서 하나 이상의 값을 가질 수 있다. 즉, 중의성을 가지고

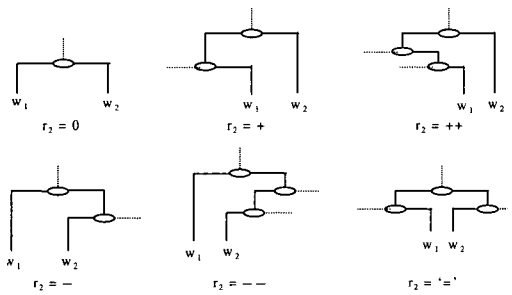


(그림 5) Supertag의 할당

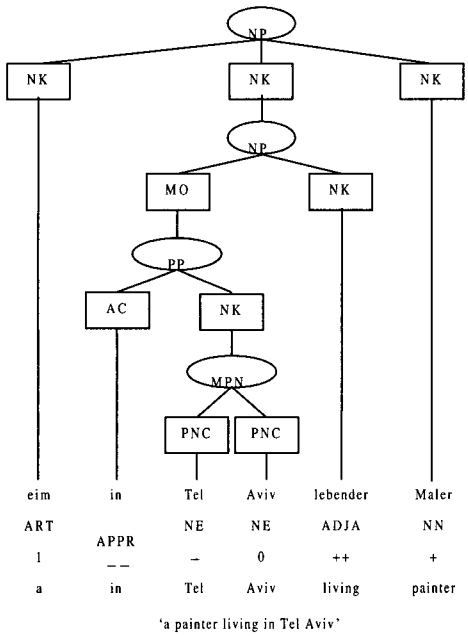
있다. 이를 해결하는 방법으로 HMM을 사용하는데, 주어진 품사열 T에 대해서 구조적 관계열 R을 생성하는 HMM은 아래와 같다.

$$\arg\max_R P(R \mid T) = \arg\max_R \prod_{i=1}^{n} P(r_i \mid r_{i-2}r_{i-1})P(t_i \mid r_i)$$

프랑스 문장 "ein in Tel Aviv lebender Maler"에 대해서 "1, --, -, 0, ++, +"를 생성한다. 이를 구문트리로 표현하면 (그림 7)과 같다. 실험 결과에 따름면, R의 원소수가 너무 작아서 정확하게 분별할 수 있는 힘이 부족하여 r_i 대신에 <r_i, t_i>를



(그림 6) 단어 w_2에 해당하는 구문구조 표지



(그림 7) 구문구조와 말덩이 표지

생성하게 하여 분별력을 크게 향상할 수 있었다. 위의 결과는 구문구조의 범주를 말할 수 없다. 이를 해결하기 위해서 위의 모델에서 r_i 대신에 <r_i, t_i, c_i>를 생성하게 한다. 여기서 c_i는 NP, ADJ와 같은 구의 범주를 나타낸다. 또한 정확률을 향상시키기 위해서 구문트리의 조부모 노드에 대한 정보를 이용한다. 따라서 위의 식에서 r_i대신에 <r_i, t_i, c_i, g_i>를 생성하도록 모델을 수정하였다. 따라서 최종적인 모델은 아래와 같다.

$$\arg\max_S P(S \mid T) = \arg\max_s \prod_{i=1}^{n} P(\langle r_i, t_i, c_i, g_i \rangle \mid \langle r_{i-2}, t_{i-2}, c_{i-2}, g_{i-2} \rangle \langle r_{i-1}, t_{i-1}, c_{i-1}, g_{i-1} \rangle) P(t_i \mid \langle r_i, t_i, c_i, g_i \rangle)$$

3.2.4 최대 엔트로피 모델

이 방법은 최대 엔트로피 모델을 부분 구문분석에 적용한다[35]. 최대 엔트로피 모델의 가장 큰 장점은 자질을 더 세분할 수 있고 중복된 자질을 제거할 수 있다는 큰 장점을 가지고 있다. 이 방법의 기본 골격은 통계적인 방법 중 하나인 [36]를 그대로 이용하고 있으며, 문맥정보에 해당하는 자질은 최대 엔트로피 이론을 적용하여 최적화하였다. 또한 각 자질의 확률을 최대 엔트로피 이론을 근거로 계산되었다.

3.3 기계학습 기반 부분 구문분석

3.3.1 메모리기반 학습

이 방법은 메모리기반 학습 방법을 이용해서 부분 구문분석에 필요한 규칙을 추출하고 유사도 기반 방법에 의해서 부분 구문분석 방법을 수행하는 방법이다[10][19][38]. 메모리기반 학습은 지도학습 방법이고 예제로부터 귀납적 추론에 의해서 학습된다. 또한 이 방법은 지연 학습(lazy learning)이다. 즉, 보여지지 않은 자료들에 대한 추정을 위해서 모든 학습 자료를 그대로 지니고 있다는 점이다. 반면에 조기학습(eager learning)은 학습 자료로부터 확률분포나 추상적인 지식을 추정

한 후에 학습 자료에 대한 개별적인 정보는 더 이상 기억하지 않는다. 또한 메모리기반 학습은 자질에 가중치를 자동적으로 부여할 수 있기 때문에 많은 자질 집합을 가진 영역에 잘 적용될 수 있다. 부분 구문분석에서 메모리기반 학습을 위한 하나의 예제는 자질벡터와 그 자질벡터가 속한 클래스로 구성된다. 새로운 자질벡터가 주어지면 메모리에 있는 모든 자질벡터 중에 가장 비슷한 벡터가 가지는 클래스로 인식한다.

이 방법에서도 품사 태깅 방법과 비슷하고 말덩이 인식을 위해 각 단어에 말덩이 표지를 할당해야 하는데, 대부분의 방법이 [33]의 말덩이 표지와 비슷한 방법을 채택한다. [19]에서는 중복되지 않고 비재귀적인 기저명사구와 기저동사구(baseVP)를 인식하기 위해서 5개의 말덩이 표지 I_NP (기저명사구 내), O(기저명사구나 동사구 밖), B_NP(인접한 기저명사구의 시작), I_VP(동사구 내), B_VP (인접한 동사구의 시작)를 사용하였다.

3.3.2 TreeBank 방법

[16]에서는 학습말뭉치로부터 기저명사구 문법(base NP grammar)를 추출하고 추출된 문법 중에서 유용한 문법만을 선택하기 위해서 각 문법에 대해 적절한 점수를 계산한다. 그리고 나서 어떤 기준치(threshold) 이상의 점수를 얻은 문법을 선택한다. 규칙을 추출하는 방법은 메모리기반 학습 방법을 사용하는데, 이 방법은 개념적으로 [13]에서 제안된 변환기반 학습 방법과 비슷하다. 즉 각 규칙 r의 점수 B_r은 C_r - E_r로 계산된다. 여기서 C_r은 규칙 r이 정확하게 인식한 기저명사구의 수이고, E_r은 잘못 인식한 기저명사구의 수이다. 기저명사구를 인식하는 과정에서 문법이 중복에서 적용될 경우, 즉 중의성이 발견된 경우에는 최장일치법을 적용하여 해결한다.

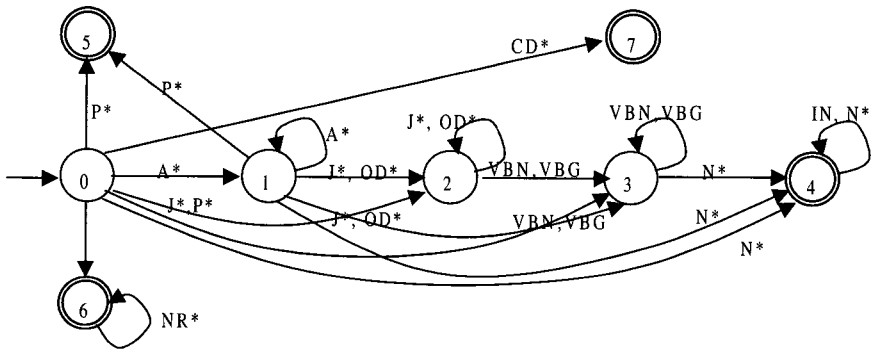
3.4 혼합 방법

이 방법은 먼저 통계적인 방법과 언어적인 정보를 유한상태 오토마타로 표현하여 최장명사구를 추출하는 방법이다[17]. 먼저 품사 태깅 시스템에 의해서 주어진 문장을 품사 태깅하고, 이를 말덩이 단위로 분리한다. 분리된 말덩이는 적절한 품사열로 표현되는데 이것이 말덩이의 표지가 된다. 이 말덩이 표지는 중의성이 발생된다. 이 중의성은 Viterbi 알고리즘을 이용해서 해결하며 Viterbi 알고리즘에서 사용되는 점수는 아래와 같은 식에 의해서 계산된다.

$$C^* = \arg\max_{c_1 \ldots c_n} \prod_{k=1}^{n} P(c_k \mid c_{k-1}) \times P(c_k)$$

여기서 c_i는 i번째 말덩이를 나타내고, P(.)는 확률을 의미한다.

이와 같은 방법에 의해서 주어진 문장의 말덩



(그림 8) 명사구 추출을 위한 유한상태 오토마타

이가 결정되며, 언어정보를 이용해서 구문적 중심어와 의미적 중심어를 결정하고 (그림 8)과 같은 유한상태 오토마타를 이용해서 명사구를 추출한다.

(그림 8)에서 사용된 기호을 LOB 말뭉치의 품사 표지이다. 예를 들면, N*는 명사, P*는 대명사, J*는 형용사, A*는 한정사, IN는 전치사, CD*는 양수사, OD*는 서수사, NR*는 부서적 명사를 나타낸다. 그리고 *는 정규표현에서 사용되는 Kleene closure를 나타낸다.

4. 응용 분야

부분 구문분석 시스템은 완전한 구문분석 시스템의 전처리기, 정보검색을 위해서 다중단어나 전문용어 추출, 정보추출 시스템에서 정보의 기본 단위의 인식 등에 이용될 수 있다. 또한 자연언어처리 시스템에서 강인한 시스템을 구축하기 위해서 많은 시스템들이 부분 구문분석의 개념을 도입한다. 이 밖에도 부분 구분분석은 여러 응용 분야에 두루 사용될 수 있는데 본 절에서는 몇 가지의 응용 사례를 소개하고자 한다.

4.1 구문분석을 위한 bootstraping 정보의 획득

Bootstraping의 주된 목적은 완전한 구문분석을 위해 필요한 어휘정보를 획득하는 데 있으며, 여기서 획득된 정보는 공기, 하위범주화, 선택제약이다. [24]에서는 Fidditch를 주어와 동사, 동사와 목적어 쌍을 구하는데 이용하였다. 이들 쌍의 연관성을 측정하여 불완전하기는 하지만 선택제약 모델을 얻을 수 있었으며, 연관성은 상호정보과 t 점수를 이용한다. 또한 [24]에서는 VP-NP-PP 쌍을 추출하여 전치사 부착 문제를 해결하기도 하였다. [31]에서는 부분 구문분석기를 이용해서 동사의 하위범주화 정보를 획득하기 위한 전처리기로 사용하였다.

4.2 정보추출(information extraction)

부분 구문분석은 MCU(message understanding conference)에 출품되는 대부분의 시스템들이 사용하고 있다. 주로 정보 추출 용도로 사용되는데 MCU에서 정보추출의 대상은 날짜, 사람이름, 회사이름, 지명, 작품명 등이다. 정보추출 시스템에서는 부분 구문분석에 의해서 인식된 정보의 조각을 적절히 결합하는 원하는 정보를 찾는데, 말덩이-부착 파싱이 매우 효과적이다. 일반적으로 정보추출 과정을 여러 과정을 거친다. 첫째, 정보추출과 무관한 문서를 제거한다. 둘째, 문서를 적절한 단위의 토큰(단어 혹은 형태소)으로 분리한다. 셋째, 키워드 주변의 말들을 구문분석한다. 넷째, 의미틀에 적절한 정보를 채운다. 다섯째, 의미틀을 결합하여 원하는 정보를 구축한다. 이 과정에서 키워드와 정보가 담긴 구성성분을 찾는데 부분 구문분석이 이용된다. 이와 같은 정보의 조각을 찾아서 영역에 특별한 의미틀을 이용하여 결합하는 방법은 부분 구문분석에서 말덩이를 찾고 말덩이의 부착으로 구문분석을 하려고 하는 기본 개념과 일치하게 된다.

4.3 음성인식

대부분의 음성인식에서 언어모델로 n그램 모델을 사용한다. n그램 모델은 지역적인 정보만 이용하기 때문에 정확률을 높이기 위해서 구문적인 정보 및 담화정보를 요구한다. [42]에서 제안한 음성인식 시스템의 개략적인 절차를 살펴보면 아래와 같이 요약된다.

1. 음성인식 시스템에서 N개의 가능한 문장을 출력한다.
   참조발화 : you werent born just to soak up sun
   1위 : you werent born justices so cups on
   190위 : you werent born just to soak up sun
2. 각 문장에 대해서 품사 태깅을 수행한다.
   (1) : you/prp werent/vb born/vbn justices/nns so/rb cups/nns on/in[6]

(190) : you/prp werent/vb born/vbn just/rb to/to soak/vb up/rp sun/nn

3. 품사 태깅 결과를 이용해서 말덩이를 분석한다.
   (1) : [np you] [vc werent born] [np justices] [advp so] [np cups] [advp on]
   (190) : [np you] [vc werent born] [advp just] [vc to soak up] [np sun]
4. 분석된 말덩이에 대해서 신경망 모델을 이용해서 점수를 계산한다.
   (1) : np-vc-np-advp-np-advp // worse score
   (190) : np-vc-advp-vc-np // more natural
5. 계산된 점수를 이용해서 N개의 문장에 대해서 순위를 다시 조정한다.
   1/(1) : you werent born justices so cups on
   4/(190) : you werent born just to soak up sun

4.4 정보검색 및 다국어검색

정보검색에서 색인어의 대부분은 명사이다. 그러나 단일어 즉, 하나의 명사를 색인어로 할 경우, 정확한 결과를 얻기 어려울 때가 많다. 이런 이유로 많은 연구자들은 복합어를 색인어로 간주한다[1][2][5]. 많은 부분 구문분석 방법은 명사구를 추출하기 위해서 제안되었다[11][16][17][38][40]. 복합 색인어를 구하기 위해서는 단순히 연속적인 명사들의 결합 관계만 이용하는 것은 아니다. 때로는 명사들의 의존관계, 특별한 구문 패턴 등 간단한 구문관계를 이용하는데, 이와 같은 것이 부분 구문 분석을 이용하는 대표적인 예이다. 특히 가장 일반적인 다국어검색은 먼저 원시언어의 색인어 및 구색인어를 추출하고, 추출된 색인어를 여러 대상언어로 번역하여 대상언어의 정보검색 시스템을 이용해서 문서를 검색하여 질의자에게 보내진다[21]. 다국어 검색에서는 색인어를 정확하게 번역하는데, 일반적으로 단어 대 단어 번역에서는 많은 중의성을 가지고 있다. 이들 중의성은 여러 개의 단어, 즉, 전문용어나 복합 명사구를 번역함으로써 중의성을 줄일 수 있다.

4.5 정보요약

정보요약 시스템은 크게 자연언어처리 기술을 이용해서 문서의 내용을 이해하고 요약하는 정보요약시스템과 통계적인 기술을 이용해서 문서의 내용을 대표할 만한 문장들을 추출하는 대표문장 추출시스템으로 분류된다. 후자의 경우에는 정보검색과 비슷한 방법으로 문서에서 명사를 추출하여 문서에 포함된 문장을 벡터로 표현한다. 각 문장의 벡터들 사이의 유사도를 측정하여 가장 유사도가 높은 문장을 요약문에 포함되는 것을 가정한다. 문장 벡터는 일반적으로 명사구에 의해서 표현되는데, 이들 명사구는 부분 구문분석에 의해서 구해진다. 또한 정보요약에서는 실마리 단어들이 중요한 자질이 된다. 일반적으로 이들은 명사구는 아니지만, 유한상태 오토마타와 같은 부분 구문분석 기법을 이용해서 구하기도 한다.

5. 결 론

본 논문은 최근에 활발히 연구되고 있는 부분 구문분석 연구 현황을 조사하였다. 인터넷 문서들은 일반 문서와는 달리 철자오류나 문법 오류들이 많이 포함되어 있다. 이와 같은 오류에 능동적으로 대처하지 않는다면, 자연언어처리를 필요로 하는 인터넷 서비스를 원활히 제공할 수 없을 것이다. 부분 구문분석은 강인한 자연언어처리 기술 중 하나이며, 구문분석의 복잡도를 크게 줄일 수 있다. 한국어 처리에서 부분 분석에 대한 연구는 그다지 활발히 연구되고 있지 않으며, 부분 구문분석의 단위 조차도 명확히 정의되어 있지 않다.

본 논문은 한국어 부분 구문분석 연구를 체계적으로 수행하기 위해서 국내외 부분 구문분석 방법론을 조사분석하였다. 앞으로는 이 연구를 기

6) 여기서 사용된 품사 표지는 에서는 언급하지 않았으나 편의상 Penn Treebank에서 사용한 것이다.

반으로 한국어 부분 구문분석의 단위를 정의하고 한국어 부분 구문분석 시스템을 설계하고자 한다.

감사의 글

본 연구는 첨단정보기술 연구센터를 통하여 과학재단과 지원을 받았으며, 또한 과학기술부 STEP 2000 프로젝트에 의해 지원되고, 전문용어언어공학연구센터에 의해 수행중인 "대용량 국어정보 심층처리 및 품질관리 기술개발" 연구과제의 일환으로 수행되었습니다.

참고문헌

[1] 강병주, 최기선, 윤준태, "한국어 정보검색에서 복합명사 색인 실험," *제10회 한글 및 한국어 정보처리 학술대회 발표논문집*, 고려대학교, pp. 130-136, 1998.

[2] 김미진, 박미성, 장혁창, 최재혁, 이상조, "고빈도어를 이용한 복합명사 색인어 추출 방안," *제10회 한글 및 한국어 정보처리 학술대회 발표논문집*, 고려대학교, pp. 121-129, 1998.

[3] 김재훈, 한국어 부분 구문분석 단위와 그 표지, 한국해양대학교, 컴퓨터공학과, 기술문서 KMU-NLP-TR-2000-006, 2000.

[4] 안동언, *기계번역을 위한 한국어 해석에서 형태소로부터 구문요소의 형성에 관한 연구*, 한국과학기술원, 전산학과, 석사학위 논문, 1987.

[5] 원형석, 박미화, 이근배, "복합명사 분할과 명사구 합성을 이용한 통합 색인 기법," *정보과학회논문지: 소프트웨어 및 응용*, 제27권, 제1호, pp. 84-95, 2000.

[6] Abney, S., "Chunk and dependencies : Bringing processing evidence to bear on syntax," *Computational Linguistics and the Foundations of Linguistic Theory*, CLSI, 1995.

[7] Abney, S., "Part-of-speech and partial parsing," *Corpus-Based methods in language and Speech Processing*, eds. Young, S and Bloothooft, G., Kluwer Academic Publishers. pp. 118-173, 1996.

[8] Abney, S., "Partial parting via finite-state-cascades," *Proceedings of the ESSLLI-96 Robust Parsing Workshop*, 1996.

[9] Aït-Mohtar, S. and Chanod, J.-P., "Incremental Finite-State Parsing," *Proceedings of ANLP'97*, Washington, pp. 72-79, 1997.

[10] Argamon-Engleson, S., Dagan, I. and Krymolowski, Y., "A memory-based approach to learning shallow natural language pattern," *Journal of Experimental and Theoretical Artificial Intelligence*, vol. 11, no. 3. 1999.

[11] Bourigault, D., "Surface grammatical analysis for the extraction of terminological noun phrases," *Proceedings of COLING-92*, pp. 977-981, 1992.

[12] Bourigault, D., "An endogeneous corpus-based method for structure noun phrase disambiguation," *Proceedings of EACL-93*, pp.81-86, 1993.

[13] Brill, E., *A Corpus-based Approach to Language Learning*, Ph.D. Thesis, Dept. of Computer and Information Science, University of Pennsylvania, 1993.

[14] Brants, T., "Cascaded Markov Models," *Proceedings of EACL-99*, Bergen, Norway, 1999.

[15] Briscoe, T., Carroll, J., Carroll, G., Federici, S. Grefenstette, G. Pirrelli, V. Prodanof, I., and Rooth, M., *SAPAKLE Package 3 Phrasal Parser Software Deliverable 3.1.* 1997

[16] Cardie, C. and Pierce, D., "Error-driven pruning of treebank grammars for base noun phrase identification," *Proceedings of COLING-ACL-98.*, 1998.

[17] Chen K.-H. and Chen H.H., "Extracting noun phrase phrases from large scale texts: Hybrid approach and its automatic evaluation," *Proceedings of ACL-94*, pp. 234-241, 1994.

[18] Church, K., "A stochastic PARTS program and noun phrase parser for unrestricted texts," *Proceedings of ANLP-88*, Austin, Texas, 1988.

[19] Daelemans, W., Buchholz, S. and Veenstra, J., "Memory-Based Shallow Parsing," *Proceedings of CoNLL-99*, Bergen, Norway, 1999.

[20] Ejerhed, E and Church, K., "Finding clauses in unrestricted text by finitary and stochastic methods," *Proceedings of ANLP-88*, 1988.

[21] Grefenstatte, G. *Cross-Language Information Retrieval*, Kluwer Academic Publishers, 1998.

[22] Grefenstatte, G., "Light parsing as finite state filtering," *Extended Finite State Models of Language*, Kornai, A. eds, Cambridge University Press, pp. 86-94, 1999.

[23] Hindle, D., *User manual for Fidditch, Technical Memorandum*, #7590-142, Naval Research Laboratory, 1983.

[24] Hindle, D., and Rooth, M., "Structural Ambiguity and Lexical Relations," *Computational Linguistics*, vol. 19, no. 1, pp. 103-120, 1994.

[25] Hobbs, J., Appelt, D., Bear, J., Israel, D., Kameyama, M., Stickel, M. and Tyson, M., "FASTUS: a cascaded finite-state transducer for extracting information from natural-language text," *Finite State Devices for Natural Language Processing*, E. Roche and Y. Schabes, eds., Cambridge MA: MIT Press, 1996.

[26] Hovy, E. and Lin, C.-Y., "Automated text summarization in SUMMARTRIST," *Advances in Autmatic Summarization*, The MIT Press, pp. 81-94, 1999.

[27] Joshi, A. and Hopely, P., "A parser from antiquity: an early application of finite state transducers to natural language parsing," *Extended Finite State Models of Language*, Kornai, A. eds, Cambridge University Press, pp. 6-15, 1999.

[28] Koskenniemi, K., Tapanainen, P. and Voutilainen, A., "Compiling ans using finite-state syntactic rules," *Proceedings of COLING-92*, pp. 156-162, 1992.

[29] Lauer, M., "Conceptual association for compound noun analysis," *Proceedings of ACL-94*, pp. 337-339, 1994.

[30] Light, M., "CHUMP: Partial Parsing and Underspecified Representations," *Proceedings of the 12th European Conference on Artificial Intelligence Workshop: Corpus-Oriented Semantic Analysis*, 1996.

[31] Manning, C. D., "Automatic acquisition of a large subcategorization," *Proceedings of ACL-93*, pp. 235-242, 1993.

[32] Marcus, M. P., A *Theory of Syntactic Recognition for Natural Language*, The MIT Press, 1980.

[33] Rawshaw, L. A. and Marcus, M. P., "Text chunking using transformation-based learning," *Proceedings of the 3rd Workshop .on Very Large Corpora*, MIT, pp. 82-94, 1995.

[34] Schiller, A., "Multilingual finite-state noun phrase extraction," *Proceedings of ECAI-96 Workshop on Extended Finite State Models for Language*, pp. 65-69, 1996.

[35] Skut, W. and Brants, T., "A maximum-entropy partial parser for unrestricted text," *Proceedings of the Sixth Workshop on Very Large Corpora.* Montreal, Canada, 1998.

[36] Skut, W. and Brants, T., "Chunk tagger -

statistical recognition of noun phrases," *Proceedings of the ESSLLI Workshop on Automated Acquisition of Syntax and Parsing*. Saarbrücken, Germany, 1998.

[37] Strzalkowski, T., "Robust text processing in automatic information retrieval," *Proceedings of ANLP-94*, Stuttgart, Germany, pp. 168-173, 1994.

[38] Tjong Kim Sang, "Noun phrase representation by system combination," *Proceedings of ANLP-NAACL 2000*, Seattle, Washington, USA, 2000.

[39] Vanderwender, L., "SENS: the system for evaluating noun sequence," *Natural Language Processing : the PLNLP approach*, eds Jensen, K. Heidorn, G. E., and Richardson, S. D., Kluwer Academic Publishers, 1993.

[40] Voutilainen, A., "NPtool, .a detector of English noun phrases," *The Computation and Language E-Print Archive* (http://arXiv.org/), cmp-lg/ 9502010, 1995.

[41] Voutilainen, A. and Padro, L., "Developing a hybrid NP parser," *Proceedings of ANLP-97*, 1997.

[42] Zechner, K. and Waibel, A., "Using Chunk Based Partial Parsing of Spontaneous Speech in Unrestricted Domains for Reducing Word Error Rate in Speech Recognition," *Proceedings of COLING/ACL 98*, Montreal, Canada, pp. 1453-1459, 1998.

김 재 훈

1986년 계명대학교 전자계산학과 (이학사)
1988년 한국과학기술원 전산학과 (공학석사)
1996년 한국과학기술원 전산학과 (공학박사)

1988년-1997년 한국전자통신연구원, 선임연구원
1997년-현재 한국해양대학교, 컴퓨터공학과, 조교수
관심분야 : 자연언어처리, 정보검색, 코퍼스 중심 언어처리, 음성언어처리