시간 표현에 대한 부분 문법 기술 및 FST를 이용한 시간 구문 분석

김윤관, 윤준태*, 송만석

연세대학교 컴퓨터과학과, 서울시 서대문구 신촌동 134, 우:120-749
{general, mssong}@december.yonsei.ac.kr, *jtyoon@linc.cis.upenn.edu

Representation of Local Grammar for Temporal Expression and Analysis of Temporal Phrase with FST

Youngwan Kim, Juntae Yoon, Mansuk Song
Dept. of Computer Science, Yonsei Univ.
*IRCS, University of Pennsylvania

요 약

시간표현은 문장에서 다른 명사와 결합하여 복합어를 이루는 경우가 있고, 용언과 결합하여 시간 부사의 역할을 하는 경우가 있는데, 이는 구문 분석에 있어서 중의적 해석이 두드러지며, 그 결과 구문 분석의 오류를 빈번히 야기하기도 한다. 본 논문에서는 이러한 시간 관계의 표현을 대량의 말뭉치로부터 획득하고 이들을 부분문법(local grammar)으로 표현한 후, 이것을 FST(Finite State Transducer)를 이용하여 부분 구문분석을 하고자 한다. 이를 위해 5천만 어절의 말뭉치에서 259개의 시간 단어를 추출하였고, 시간 단어들의 의미적 또는 기능적 사용에 의해서 26개의 어휘 범주로 분류하고 각 범주들의 결합관계를 일반화하였다. 실험을 통하여 인식을 위한 시간표현의 결합관계는 최고 97.2%의 정확률을 보였고, 품사태깅에 있어서는 평균 96.8%의 정확률을 보였다. 이는 시간 표현의 결합관계가 부분 구문분석에 있어서 유용한 정보임을 보여준다.

1 서 론

시간 표현은 정보 추출과 같은 응용분야에서 필히 인식을 필요로 하는 중요한 요소이다. 또, 구문 분석에 있어서는 구성 성분의 중의적 해석이 두드러지게 나타나는 부분이며, 이러한 결과로 구문 분석의 오류를 빈번히 야기하기도 한다. 다음의 예를 보자.

1a) *간밤* 꿈에 나타난 ...
1b) *간밤* 아내는 어디로 갔는지 ...
2a) *이번 여름* 휴가에 우리가 갔던 곳은...
2b) *이번 여름* 산사태가 일어난 곳은 ...
3a) *10월 9일 저녁 7시* 비행기표를 예약할 수 ...
3b) *10월 9일 저녁 7시* 김 대통령의 담화가 있다.

위 문장들을 시간명사와 관련하여 구문 분석해보면 크게 두 가지 경우로 나뉘는데, 첫 번째(a)는 시간 표현[1] 이 다음 명사를 수식하는 관형어 역할을 하여 하나의 명사구[TN[2]]를 이루고, 두 번째(b)는 시간표현이 용언을 수식하는 부사[TA[3]]로 사용된 경우이다. 따라서 이 둘이 구분되어야 한다. 그 결과는 다음과 같다.

1a) TN(T[4]간밤 NN꿈)에 ...
1b) TA(T간밤) ... (V갔는지)
2a) TN(T(이번 여름) NN휴가)에 ...
2b) TA(T(이번 여름)) ... (V일어난) ...
3a) TN(T(6월 18일 저녁 7시) NN비행기표) ...
3b) TA(T(6월 18일 저녁 7시)) ... (V있다)

1) 앞으로 한 어절을 포함하여 연속적으로 나열된 시간 관련 어휘들을 '시간표현'으로 지칭한다.
2) TN : Time Noun (시간명사)
3) TA : Time Adverb (시간부사)
4) T: 시간표현, NN: 명사, V: 동사

즉, 시간표현은 다른 명사와 결합하여 복합어를 이루는 경우가 있고, 시간 부사의 역할을 하는 경우가 있다. 이는 시간표현이 명사를 수식할지 아니면 동사를 수식할지를 결정하는 것이므로 실제 구문분석을 하기 전에 대단히 중요한 정보가 된다[15,17]. 이에 대한 하나의 해결책으로서 '오늘'과 같이 명사이나 부사로도 사용되는 단어들을 태깅 단계에서 구분하여 태깅하는 접근 방법이 있을 수 있다. 그러나, 실제로 이러한 시간 표현의 양상은 어휘 관계에 밀접하게 연관되어 있어 적은 양의 예문을 기반으로 해서는 정확히 인식하기 어렵다. 더욱이 여러 시간 관련 어휘들이 하나로 묶여서 시간 명사구와 시간 부사구를 구성하게 되므로 연속된 여러 어절들을 하나의 구로 인식할 수 있어야 한다. 그래서 이러한 시간 관계의 표현을 대량의 말뭉치로부터 획득하고 이들을 부분 문법으로 표현한 후, 이것을 FST를 이용하여 태깅하고자 한다. 본 연구에서는 먼저 KAIST 원시 말뭉치 5000만 어절로부터 용례 색인기[5]를 이용하여 시간 표현을 추출한 후, 대상 시간 표현 문법을 추출하고 이를 부분 문법으로 표현한다. 또, 표현된 어휘 관계의 문법은 FST로 표현되며, FST로부터 나오는 출력은 문법 기능을 가진 시간 관계의 표현으로서 구문 분석기에 직접적으로 입력되는 표현덩이(chunk)가 된다. 즉, FST를 이용해서 시간 문법에 대한 부분 문법을 기술함으로써 문장 내에서 시간표현의 시작과 끝을 찾아서 한 어절 이상으로 구성된 시간표현을 인식할 수 있고, final state에서 명사인지 부사인지를 판별하는 기능(output function)을 수행할 수 있다[10,12].

품사 태깅에 있어서 통계적 품사 태깅방법은 규칙을 만들 필요가 없고 높은 정확률을 가지는 반면에 언어 정보가 간접적으로 얻어지고 상당히 많은 양의 통계표가 필요하다는 단점이 있다. 최근 규칙기반의 한 방법으로 Brill[3]이 제안한 변형 규칙기반의 품사 태거는 직접적으로 언어 정보를 획득할 뿐 아니라 적은 양의 규칙으로도 통계적 방법의 정확률을 가진다. 또 Roche[10]는 Brill의 규칙기반의 품사태거를 결정적이고 최소 상태수를 사용하는 FST로 구현하여 태깅함으로써 수행속도를 linear time 으로 향상시켰다. 이러한 FSM(Finite State Machine)은 효율적인 수행속도와 저장공간, 그리고 표현 및 인지의 편리성으로 인해 전자사전[13,14], 음성인식[7], 패턴매칭[12], 품사태깅[9]등의 분야에서 사용되고 있다. 또한 최근에는 '가격표현[11]', '주식과 관련된 표현[19]', '특정 주제와 관련된 고유명사 표현[4]'등과 같은

표 1 어휘범주분류표

어휘범주분류			분류번호	단 어
수식어범주	수식어		1	꼬박,올,지난, ...
	수사		2	아라비아숫자, 몇,수, ...
시간명사범주	시간단위의존명사		3~10	세기,년,월,주,일,시, ...
	시대		11	고생대,후대,현대, ...
	년범주		12	금년,새해,올해, ...
	월범주		13	내달,선달,정월, ...
	주범주		14	격주,금주,내주, ...
	일범주	요 일	15	일요일,월요일, ...
		고유하루	16~17	하루,이틀,보름,추석, ...
		상대하루	18	오늘,어제,내일, ...
	시간및기간범주	하루시간	19	새벽,아침,점심, ...
		년중기간	20~21	선거철,명절,연말, ...
		계절	22	봄,여름철,한겨울, ...
		기간명사	23	건조기,성수기,환절기, ...
		보조기간	24~25	초기,중반,말엽, ...
기능어범주	기능어		26	경,무렵,동안,이내, ...
총 어휘 수			26범주 259단어	

세부적인 언어현상에 대해서 부분문법을 기술하고, 이를 FSM으로 표현하여 처리하고 있다.

2 시간표현의 결합관계 일반화

2.1 시간어휘범주 분류

시간표현의 결합관계를 얻기 위해서 본 논문에서는 먼저 시간표현 명사[18]에 대해서 학습말뭉치에서 용례를 추출하고, 이 용례에서 시간표현 명사와 결합하는 다른 시간관련 단어에 대해서 출현 빈도수를 고려하여 총 259개의 시간표현 단어를 획득하였다. 이러한 단어들을 시간표현 내에서의 기능을 고려해서 크게 3가지 수식어범주, 시간명사범주, 기능어범주로 분류하였고, 다시 각 범주에 대해서 결합관계의 유사성을 고려해서 표1과 같이 총 26개의 범주로 분류하였다.

2.2 시간표현 결합정보 획득

학습말뭉치에 기반을 두고 각 시간어휘범주에 속한 시간 단어들에 대해서 용례를 추출하고, 그 시간 단어와 좌.우 거리 1로 결합하는 시간 단어에 대해서 그림1과 같은 형태의 결합정보를 획득한다. 그리고 나서 어휘범주와 어휘범주 사이의 결합관계를 기술한다[15].

<결합정보> := <중심어><문맥정보>
<중심어> := TW TAG TC
<문맥정보> := <좌문맥><우문맥>
<좌문맥>:={앞어절}:= TW TAG TC
<우문맥>:={뒤어절}:= TW TAG TC
TW := 시간단어
TC := 시간어휘범주
TAG := 품사태깅정보

그림1 결합정보 형태

학습말뭉치에서 학습된 시간표현의 내부적인 결합관계는 기능적인 결합, 예를 들어 '수사+시간단위의존명사'와 같은 것을 제외하면, 의미적으로 대부분 수식 관계[5]에 있으며 큰 시간범주에서 작은 시간범주로 표현된다. 이러한 결합관계가 유지되면 4b)처럼 하나의 시간표현으로 결합될 수 있고, 그렇지 않으면 5b)처럼 시간표현은 분리된다.

4a) *이튿날 상오 9시 30분*까지...
5a) *어제 저녁 10월*의 행사에 대해 ...

4b) T_1(*이튿날 상오 9시 30분*)까지...
5b) T_1(*어제 저녁*) T_2(*10월*)의 행사에 대해 ...

2.3 시간표현의 부분구문 중의성 해소

시간표현이 명사를 수식하는지 동사를 수식하는지를 어떻게 하면 예측할 수 있을까? 이는 물론 전체 문장을 분석해야 알겠지만, 부분 문맥(local context)에서 단어들의 나열만 보고도 어느 정도 예측이 가능하다. 이러한 것을 텍스트 청킹(text chunking)이라고 하는데, 이것은 파싱(parsing) 이전에 유용한 예비 단계로서 결합 가능한 어절들의 표현덩이를 만드는 것이다[1]. 이러한 텍스트 청킹을 이용해서 '최대길이명사구(maximal length noun phrase)'를 찾아낼 경우 95%의 성능을 보인다[2].

본 논문에서는 시간표현의 부분 구문 중의성 해소를 위해서 한 문장의 전체를 보지 않고, 부분 문맥[6]만을 고려하여 시간명사와 시간부사의 구별이 가능하다고 가정한다. 그래서 그림2의 절차를 거쳐서 시간단어와 결합하는 어휘연관성이 높은 명사 리스트를 추출하여 시간표현의 결합사전으로 구축한다. 단, 사람이 판단하기에도 어려울 정도의 시간 구문에 대해서는 모든 문맥이 동원되어야 해결될 수 있는 부분이기 때문에 제외했다.

1. 학습말뭉치에서 각 시간단어들의 용례추출.
2. 4b)와 5b)와 같은 형태로 전체문장을 보고 시간표현에 대해서 TN과 TA로 품사태깅.
3. TN으로 태깅 된 시간 단어와 결합한 명사들에 대해서 그림3과 같은 방법으로 결합사전을 구축

그림2 시간표현 결합사전 구축 절차

<시간표현복합명사구> := <시간표현> NN
<시간표현> := Tw_1 Tw_2 ... Tw_n ($n \geq 1$)
if LA(Tw_n,NN) //어휘연관성(Lexical Assoc.)
add <시간표현 결합사전 엔트리>
:= Tw NN TC
Tw := 시간단어
NN := 시간단어와 결합하는 일반명사
TC := Tw의 시간어휘범주

그림3 시간표현 결합사전 엔트리 형태

그러면 예를 들어 1a)와 1b)에서처럼 '간밤'이라는 시간표현이 '꿈'과 같이 결합관계를 갖는 명사를 만나면 복합명사가 되고, 그렇지 않으면 시간 부사가 되는 것이다. 그리고 실제 문장에서 두 어절 이상으로 구성된 시간표현에 대해서도 그림 3의 방법을 적용하는데, 이때 그림4와 같은 몇 가지 사항에 대해서는 본 논문에서 예외 사항으로 간주하여 다루지 않는다.

1] a. 여름 바다에 갔다.
b. *지난 여름* 바다에 갔다.
2] a. *오늘 아침* ...
b. *오늘* 이른 *아침* ...
3] a. *금년* 매출액의 ...
b. *금년* 저조한 매출액의 ...

1]의 경우, a에서 '여름'은 '바다'와 어휘연관성이 높지만, b에서는 '지난'이 '여름'과 결합함으로써 '여름'은 '바다'에 대한 어휘연관성이 매우 낮아진다.
2]의 경우, a,b 모두 시간표현으로 인식되어야 하지만 어절 거리 2 이상의 시간 표현 결합에 대해서는 고려하지 않는다.
3]의 경우, a는 시간표현의 복합명사구로 인식하고 b는 결합명사에 수식어가 붙은 경우로 고려하지 않는다.

그림4 시간표현 결합관계의 예외사항

5) 뒤 명사가 앞 명사에 딸려 있으면서 보다 구체적인 시간 의미를 나타내는 관계

6) 한 문장을 구성하는 부분(part)으로서 본 논문에서는 어절단위로 Bigram 모델을 고려한 의미로 사용한다.

3 FST를 이용한 시간 결합 관계의 표현방법

이 장에서는 지금까지 기술한 시간표현 결합관계를 FST로 표현하고, 이 FST를 이용해서 문장 내에서 시간표현을 인식하고 그것이 시간명사인지 아니면 시간부사인지를 부분 구문 분석하는 방법을 다루고 있다.

Def) Finite State Transducer
6-튜플(tuple) $T = (\Sigma_1, \Sigma_2, Q, i, F, E)$
Σ_1 : 유한개의 입력 알파벳
Σ_2 : 유한개의 출력 알파벳
Q : 상태의 유한 집합
$i \in Q$: 초기 상태
$F \subseteq Q$: 최종 상태의 집합
$E : Q \times \Sigma_1^* \rightarrow \Sigma_2^* \times Q$: 전이출력함수

그림5 FST의 정의

3.1 FSM(Finite State Machine)의 소개[12]

FST는 그림5와 같이 정의되며, FST와 FSA(Finite State Automata)의 가장 큰 차이점은 FSA는 상태에서 결과를 출력하지만, FST는 전이(transition)할 때 결과를 출력한다. 그래서 모든 비결정적(Non deterministic) FSA는 결정적(Deterministic) FSA로 변환 가능하지만, FST는 그렇지 못한 경우가 있고 특별히 결정적 FST를 'Sequential FST'라고 한다. 그리고 최종상태에서 최종결과 출력함수(final output function)을 제공하는 것을 'Subsequential FST'라고 한다. 본 논문에서는 그림6과 같은 'Subsequential FST'를 이용해서 시간표현을 인식하고 바로 뒤의 단어를 더 봄으로써 마지막 상태에서 시간명사인지 시간부사인지를 판별해서 앞서 인식한 시간표현에 대해서 명사인지 부사인지를 태깅한다.

Def) 7-튜플, $T = (\Sigma_1, \Sigma_2, Q, i, F, E, \rho)$
$\Sigma_1, \Sigma_2, Q, i, F$ and E are same as the FST
$\rho : F \rightarrow \Sigma_2^*$: the final output function

그림6 Subsequential FST의 정의

3.2 시간 구문 분석을 위한 FST 구현

문장 내에서 시간표현을 인식하고 태깅하는 FST는 다음과 같이 정의된다.

Σ_1 = {시간어휘범주}, Σ_2 = {TN, TA ,NT[7]}
i = {S}, F ={S,F}

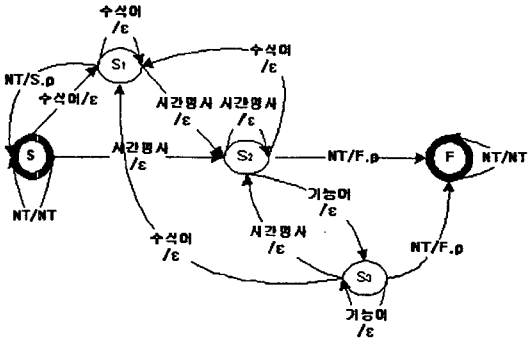

그림7 3가지 어휘범주에 대한 시간구문 FST

I_i : S_i에서 입력받을 수 있는 입력 알파벳의 집합
$= \{I_{i1}, I_{i2}, \ldots, I_{in}\} \subseteq \Sigma_1$
$N_{i,j}$: S_i에서 $I_{i,j}$가 입력되었을 때 전이되는 상태
$O_{i,j}$: S_i에서 $I_{i,j}$가 입력되었을 때의 출력

$$= \begin{cases} \varepsilon & \text{if } N_{i,j} \notin F \\ \text{NT} & \text{if } S_i \in F,\ I_{i,j} = \text{NT} \\ S.\rho & \text{if } N_{i,j} = \text{S},\ I_{i,j} = \text{NT} \\ F.\rho & \text{if } N_{i,j} = \text{F},\ I_{i,j} = \text{NT} \end{cases}$$

$S.\rho$: S에서 S_i까지 인식한 시간표현을 NT로 태깅
$F.\rho$: S에서 S_i까지 인식한 시간표현에 대해서 S_i 바로 이전 상태의 입력 알파벳과 $I_{i,j}$ 쌍이 시간표현 결합사전에 있으면 TN, 그렇지 않으면 TA로 태깅.
또한 $I_{i,j}$ 가 조사나 어미일 경우에는 TN으로 태깅.

시간어휘범주의 결합관계를 상태의 전이로 표현하기 위해 2차원 배열구조를 사용하고[12,14], 시간표현의 히스토리를 유지하기 위해 링크드 리스트구조를 사용한다. 그림7은 수식어범주, 시간명사범주, 기능어범주 이렇게 3개의 범주만을 고려한 시간표현 부분 구문 분석 FST이

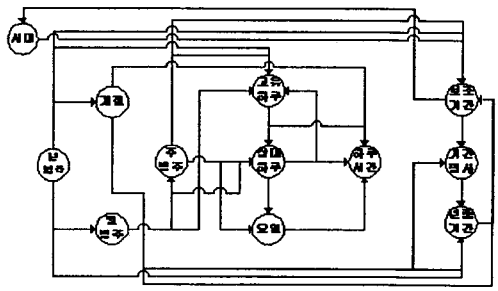

그림8 시간명사 어휘범주 결합도

7) NT(NonTimeword):시간어휘범주에 속하지 않은 단어

다. 그리고 시간명사범주에 속해 있는 각 범주들의 결합 관계를 다이어그램(diagram)으로 표현한 것이 그림8이다.

4 실험 및 평가

실험을 위해 학습말뭉치에서 학습에 이용되지 않은 100개의 문장과 '제 1회 형태소 분석기 및 품사 태거 평가 대회 MATEC 99'에서 배포한 학습데이터 중에서 시간표현을 포함하고 있는 문장 200개를 추출하여 총 300개의 문장에 대해서 표2와 같은 FST를 가지고 실험을 하였다.

표 2 실험한 FST의 종류

FST종류	상태수	FST의 내용
FST1	5	3개의 시간 어휘범주
FST2	28	26개의 시간 어휘범주
FST3	28	FST2에서 같은 시간범주의 단어에 대해 결합사전을 통합
FST4	32	FST2에 부족한 어휘들을 보강

본 실험에서는 먼저 학습말뭉치에서 획득한 시간어휘범주들의 결합관계가 얼마나 시간표현을 인식하여 하나로 태그 할 수 있는지를 살펴보았다. 표3는 그 결과를 보여준다.

표 3 시간표현 인식 결과

	FST1	FST2	FST3	FST4
정확률	92.59 %	94.48 %	96.35 %	97.15 %
재현률	84.45 %	81.85 %	88.58 %	90.56 %

표3에서 FST1과 FST2의 정확률이 비슷한 것은 실제 시간표현의 결합관계가 어느 정도 정형화되어 있어서 이를 예측하기가 용이함을 의미하고, FST2의 재현률이 더 낮은 것은 상태수가 증가함에 따라 고려되지 못한 전이관계가 있음을 의미한다. 그리고 FST3과 FST4처럼 시간단어의 어휘가 많을수록 그 인식률은 증가한다.

두 번째 실험으로는 위 FST가 앞서 인식한 시간표현에 대해서 태깅을 하였는데, 이 결과는 표4과 같다. 전반적으로 TN의 정확률이 TA보다 높은 것은 조사와 어미 앞에서는 무조건 TN이 된다는 결정적인 정보가 있기 때문이고, TN의 재현률이 현저하게 낮은 것은 시간단어의 어휘량이 적은 것을 의미한다. TA의 정확률이 낮은 것은 TN의 재현률이 낮은 것과 관계가 있는데, 이는 TN

표 4 시간표현 태깅 결과

	FST1	FST2	FST3	FST4
TN 정확률	98.52 %	99.40 %	99.20 %	98.75 %
TN 재현률	76.10 %	82.78 %	87.13 %	89.50 %
TA 정확률	83.11 %	86.47 %	93.30 %	94.80 %
TA 재현률	88.76 %	90.14 %	91.33 %	92.61 %

으로의 분석이 실패하면 TA로 태깅을 하기 때문에 실제로 TN인 시간표현을 시간표현의 결합사전에 미등록된 단어에 대해서 결합관계가 없는 것으로 판단을 하기 때문에 TA로 태깅을 하게 되고, 이것이 바로 TA의 정확률을 낮추는 요인이 된다.

FST3의 결과가 FST2보다 향상된 것은 같은 시간 어휘범주내의 단어들은 유사한 단어들과 결합 연관성이 높은 것을 의미한다.

FST4는 FST2의 실험에서 부족한 시간단어들을 보강한 것으로 시간결합관계만의 오류를 확인할 수 있다. 다음은 시간결합관계에서 나타나는 몇 가지 오류 유형들이다.

6) 1854년 3차례에 걸쳐..
[정 답] TN(1854년) NO(3차례)에 ...
[오분석] NO(1854년 3차례)에 ...
7) 82년 현재의 이곳으로 이주하여..
[정 답] TA1(82년) TN2(현재)의 ...
[오분석] TN1(82년 현재)의 ...
8)처음 몇 달 내 눈에 대한 기적은 ...
[정 답] TA(처음 몇 달) NO(내 눈)에 ...
[오분석] TA(처음 몇 달 내) NO(눈)에 ...

6)의 경우는 시간표현 뒤에 수식어 범주가 나와서 시간단어가 아닌 것과 결합하여 앞의 시간표현 마저도 태깅되지 못 하도록 하는 것인데, 이는 현재 상태에서 링크드 리스트에 저장된 시간표현에 대해 마지막 노드에서부터 수식어 범주들을 제거하고 시간부사로 태깅함으로써 해결될 수 있다. 7)과 8)의 예는 본 시스템이 부분 문맥만을 고려하여 태깅할 때의 한계로서 시간표현의 결합관계만으로는 해결하기 어렵다.

5 결론

본 논문에서는 말뭉치에서 시간표현에 대한 결합관계를 부분문맥을 고려하여 획득하였고, 이를 FST를 이용하여 문장 내에서 시간표현을 인식하고 품사 태깅하는

실험을 하였다. 실험을 통하여 인식을 위한 시간표현의 결합관계는 어느 정도 규칙 기반으로 일반화될 수 있음을 알 수 있었고, 특히 연접한 시간 표현의 인식에 있어서는 아주 높은 인식률을 가졌고 시간부사에 대한 태깅보다 시간명사에 대한 태깅이 더 좋은 결과를 보였다.

향후 연구에서는 품사태깅에서 중요한 정보가 되는 시간표현의 결합사전을 보강할 수 있는 방법에 대한 모색이 필요하다. 또 본 시스템은 규칙 기반의 한 방법으로 현재 응용 시스템을 고려해 하나의 해를 제안하고 있으나, 이는 향후 n개의 해를 제안하는 방법에 대해서도 연구되어야 할 것이다.

[참고문헌]

[1]Abney, S., "Parsing by Chunks", Kluwer Academic Publishers, 1991

[2]Bourigault, D., "Surface Grammatical Analysis for the Extraction of Terminological Noun phrases", In Proc. 15th International Conference on Computational Linguistics, p.977-981, 1992

[3]Brill, Eric., "A simple rule-based part of speech tagger", In Third Conference on Applied Natural Language Processing, p.152-155, 1992

[4]Jean Senellart, "Tools for Locating Noun Phrases with Finite State Transducers", In Proc. of COLING-ACL, 1998

[5]KAIST, KAIST 용례색인 프로그램, URL = http://csfive.kaist.ac.kr/kcp/

[6]Mohri, M., "Compact Representation By Finite-State Transducers, Proceeding of ACL, 1994

[7]Mohri, M., "Finite-state transducers in language and Speech Processing", Computational Linguistics, 1997

[8]Ramshaw, L. A. and Marcus, M., "Text Chunking using Transformation-Based Learning", ACL Third Workshop on Very Large Corpora, 1995

[9]Roche, E. and Schabes, Y., "Deterministic Part-of-Speech Tagging with Finite-State Transducers", Computational Linguistics, 1995

[10]Roche, E., Schabes, Y. "Finite-state language processing", The MIT Press, 1997

[11]강승식, "상거래용 형태소 분석기 및 구문분석기 개발", 연구보고서(한국전자통신연구원), 1999.

[12]김재훈, "Finite State Machine for Natural Language Processing", 자연언어처리 튜토리얼, 1999

[13]남지순, "검색엔진을 위한 '백과사전' 전자사전의 구축(I)", 한글 및 한국어정보처리 학술대회, 1998

[14]백대호, "Finite State Transducer를 이용한 한국어 전자 사전의 구조", 정보과학회 논문지, 22권 2호, 1995

[15]윤준태, "공기관계 기반 어휘연관도를 이용한 한국어 구문 분석", 연세대학교 컴퓨터과학과 박사학위논문, 1997

[16]이상주, "품사태깅을 위한 어휘규칙의 자동획득", 한글 및 한국어정보처리 학술대회, 1998

[17]임희석, "어절 태그 변형 규칙을 이용한 한국어 품사 태거", 정보과학회 논문지, 24권 6호, 1997

[18]윤준태, "한국어 시간 표현 명사의 분류", Memo, 1999

[19]LADL, a research laboratory of the University Paris 7, URL=http://www.ladl.jussieu.fr/INTEX/