제25회 한국정보처리학회 춘계학술발표대회 논문집 제13권 제1호 (2006. 5)

형태소분석에 의한 어절사전의 구축

허현규,
e-mail : hhuh@univ-mlv.fr

Construction of Korean word dictionary by morphological analysis

Hyun-Gue HUH
Laboratoire d'infomatique de l'Institut Gaspard-Monge
Universite de Marne-la-vallee /CNRS(UMR8049)
5, bd Descartes-F77454 Marne-la-vallee CEDEX 2-France

요 약

언어 기반의 형태소 분석에 따라 언어 지식을 국부 문법을 이용하여 그래픽적으로 기술하여 이를 기반으로 한 어휘 정보와 연결 정보를 함께 가지는 어절 사전을 구축하였다.

1. 서론

본 논문에서는 굴절어에 대한 어절 사전을 구성 함에 있어 일반적인 어휘사전의 나무 구조의 오토마타의 방식이 아닌 다중 트랜스두서의 연속적인 연결에 의한 어절 사전을 소개한다. 본 논문은 한국어의 어절 사전을 구축함에 있어서 단어간의 구문을 기술하는 국부 문법[1]에 의한 기술을 한국어 어절내의 형태소간의 구문 기술 기법으로 적용하여 어절을 이루는 각 형태소마다 언어 정보를 지니는 출력을 가지는 오토마타인 트랜스두서를 연속적인 연결에 의한 어절 사전을 구성한다.

한국어 텍스트의 전산적 처리에서 사용되는 형태소 분석이란 텍스트에서 추출된 어절에 대해서 2 단계로 일반적으로 첫번 단계는 어휘소 사전을 적용하여 가능 형태소를 찾고 두 번째 단계는 형태소간 연결 규칙을 통계학적 HMM등의 확률 규칙을 이용하여 올바른 어절을 추출하는 방법이다[2 3 4 5]. 또 다른 방법은 단계 순서를 바꾸어 각 단어에 에티켓의 열을 태그하여 이를 이용하여 형태소 어휘들과 함께 형태적인 세그먼트를 행하는 방법이 있다[6]. 또한 두개 레벨에 의한 언어 기술인 Klex system[i]이 있다. 이러한 방법들에서는 어휘의 추가 시 사전과 형태소 규칙에 대한 각기 자료를 추가 수정하여야 하며 또한 규칙을 정의하기 위한 전산 기술 언어들을 습득하여야 한다.

본 논문에서는 어휘소간의 연결을 가지는 어절 사전을 구축함에 있어서 유한 오토마타에 의한 국부 문법의 그래픽적 기술을 이용한 형태소의 언어학적 분석 결과를 기술한 자료를 이용하여 어절 사전을 구축 하였다. 어절내의 형태소 분석 결과를 정확하고 풍부한 언어정보로 기술함으로 인하여 어절간의 더욱 정확한 어절간의 분석에 도움이 될 수 있으며 언어 정보 기술을 쉽게 하여 언어 정보의 재활용을 높이는데 있다.

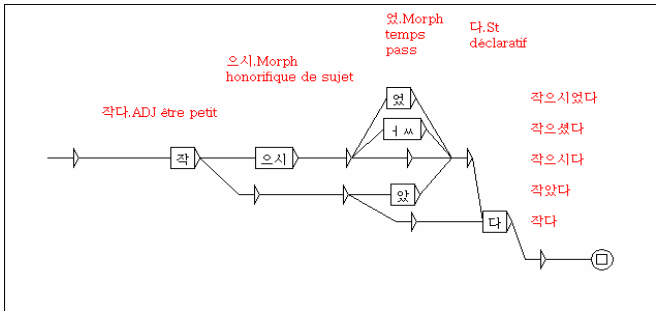
2. 어절내의 형태소들의 기술

한국어 어절을 형태소의 열로 정의한다면 각 형태소는 연결되어지는 다른 형태소와의 관계를 표현하는데 있어서 단어간의 구문의 기술로 사용한 그래프에 의한 국부 문법의 기술[7]을 굴절 언어인 한국어 어절내의 형태소간의 연결을 기술하는 방법으로 사용하였다. 국부 문법에 의한 형태소간의 연결은 오토마타를 이루며 오토마타를 이루는 각 상태 상자에 각 형태소별 텍스트상의 활용형을 표기하고 각 형태소의 언어 정보를 기술한다. 오토마타의 두 개의 상태를 연결하는 즉 각 형태소간의 연결정보는 상태를 연결하는 전이 화살표에 의해서 자동적으로 표현되어 질 수 있다. 그래픽적으로 오토마타로 표현된 형태소 언어 정보를 기술하는 방법을 통하여 구축되어진 언어 정보를 이용하여 한국어의 형태소 열에 의한 어절 사전을 구현하였다. 각 형태소간의 정보 표현도구로써 UNITEX[ii], [8]

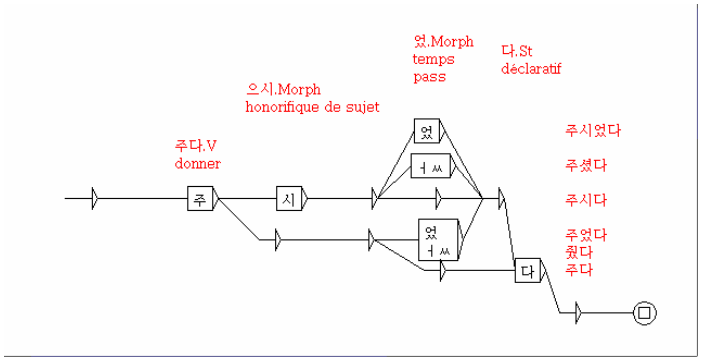
i http://www-cis.upen.edu/~nrh/klex.html

[ii] 코퍼스 언어 분석기http://www-igm.univ-mlv.fr/~unitex

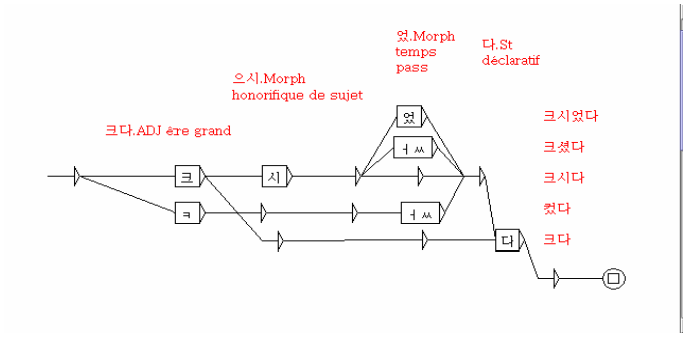
시스템을 이용하였다. UNITEX 시스템은 RTN[iii] 을 지원한다[9]. 어휘소 간의 연결을 오토마타로 나타내어 줄 수 있다. <그림 1,2,3>은 작다, 주다, 크다의 어근에 대한 어미들의 연결을 보여주고 있다.



<그림 1> 어근 '작다'와 어미들의 열 표현

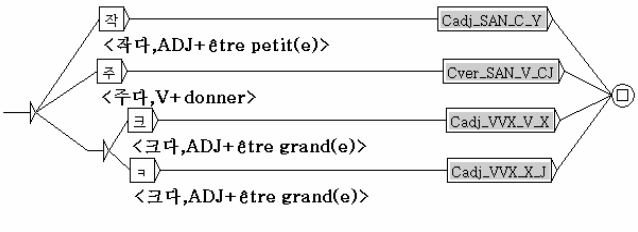


<그림 2> 어근 '주다' 와 어미들의 열 표현



<그림 3> 어근 '크다'와 어미열의 표현

어근에 대한 어휘 기술의 작업은 대부분 리스트적 언어 정보 표현 방법 이루어져 왔다. <그림 4>는 부 그래프를 이용하여 어근과 각 어근이 가지는 어미열에 대한 부분으로 다시 표현할 수 있다



<그림 4> 어근의 변화형과 호환 어미열로 표현

[iii] Recursive Transition Network

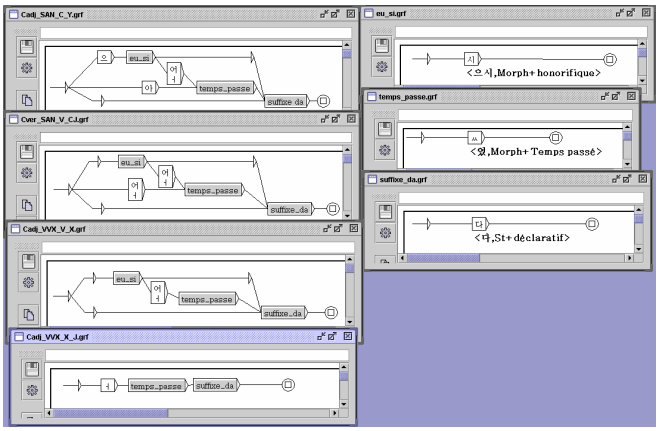
각 어근의 음성학적 형태소적인 차이에 따라 각기 다른 어미열에 대한 호환성을 가질 수 있다. <그림 4>에서 각 어근의 활용형에 따른 어미열들은 부 그래프를 나타내는 회색 박스로 표현하였다. 본 논문에서는 하나의 형태소에 대한 언어 정보 표현을 텍스트상의 활용형에 대한 기본형, 형태소 언어 정보 및 뒤에 따르는 호환성 정보로 표현하였다.

위의 그래픽으로 표현된 어근과 어근에 상응되는 어미열에 대한 표현은 리스트적인 기술 표현으로 나타낸다면 다음과 같이 나타내 줄 수 있다. 즉 한 어절 내의 어근의 정보는 언어 정보 및 그 어근에 연결되는 어미들의 호환성 정보로써 표현되어 질 수 있다.

```
;작,,작다,ADJ,Cadj_SAN_C_Y
;주,,주다,V,Cver_SAN_V_CJ
;크,,크다,ADJ,Cadj_VVX_V_X
;ㅋ,,크다,ADJ,Cadj_VVX_X_J
```

본 논문에서는 모든 어근을 어절 내에서 하나의 형태소로 정의 하며 어간에 대한 사전 리스트를 구성시 이를 기술함에 있어서 어간의 언어적인 정보 및 호환 어미 열에 대한 정보를 같이 기술하여 주었다.

<그림 5>는 각 어근에 대해서 호환되어 지는 어미열에 대해서는 여러 부 그래프로 표현되는 오토마타로 구축함을 보여준다.



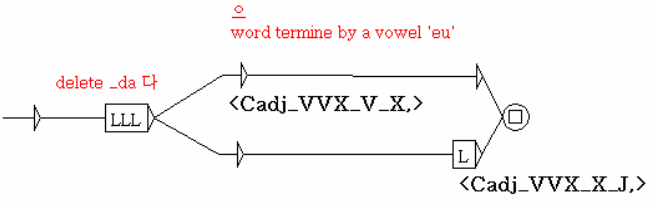
<그림 5> 어미열에 대한 부 그래프를 이용한 기술

어근의 사전은 일반적으로 아래와 같이 어근의 변화를 가지지 않는 일반 사전상의 형태인 기본형으로 기술하며 어근이 변화를 가지는 불규칙 어근에 대해서는 이에 대한 변화를 기술하는 트랜스두서를 그래픽으로 기술하여 사용하였다. 이를 이용하여 자동적으로 각 기본형에서부터 변이형 및 호환성을 추출한다.

```
작다,,ADJ,Cadj_SAN_C_Y
주다,,V,Cver_SAN_V_CJ
크다,,V,Cadj_VVX
```

어근 '크다'에 대해서 변이형과 이에 호환되는 어미 열에 대한 추출을 위한 기술 방법으로 트랜스두서

(Cadj_VVX)는 <그림 6>과 같이 표현하였다. 'L'은 하나의 음소를 지우라는 명령이다. 두개의 경로는 각기 다른 형태의 변이형을 만들어 낸다.



<그림 6> 어근의 변위형을 추출하는 트랜스두서

이러한 트랜스두서를 이용한 변이형의 추출에 대한 기술은 한국어를 이해하려는 이들에게 언어의 이해에 도움이 될 수 있다. 본 논문에서는 모든 정보 기술을 쉽게 이해 시킬 수 있도록 노력하였다.

3. 다중 트랜스두서를 이용한 어절사전의 구현

일반적 어휘 사전은 전자 사전을 위한 압축에 의해서 하나의 초기 상태와 각 어휘의 정보를 가지는 다중의 최종 상태의 나무구조의 오토마타로 표현되어지는 압축 형태를 갖는다. 유럽어는 사전상의 하나의 기본형이 수와 시간에 따라 몇 개의 변화형을 가지게 된다. 하지만 한국어와 같은 굴적어에서는 하나의 기본 어근이 수백 개에서 수천개의 굴절부를 가질 수 있다. 하나의 단어를 나타내는 어절의 사전을 위해서는 너무 거대한 크기의 어절 사전을 이룰 수 있다. 본 논문에서는 리스트 형태로 구성된 어근과 그래프로 구성된 어미열에 대한 개별적인 압축을 하여 이를 통합하는 방식으로 한국어 형태소 열에 의한 어절 사전을 구축하였다.

유한 문자 집합 B 와 유한 상태의 모임 Q 와 초기 상태 I 와 종료 상태 T 의 오토마타의 정의 5 개 요소로 표현한다면 하나의 오토마타는 A=(B, Q, I, T, F)로 나타낼수 있다. 여기서 F 는 전의를 나타내며 F(p, a, q) : 하나의 상태 'p' 에서 하나의 문자 요소 'a' 는 다음 상태 'q'로 가기 위한 전이 조건이 된다.

전이에 출력이 나오는 트랜스두서는 6 개의 요소로 표현되어 질 수 있다. 이때 오토마타는 A=(B1, B2, Q, I, T, F)로 나타내며 입력 문자 B1 과 출력 문자 B2 및 이때의 전이는 F(p, m1, q, m2)로 나타내면 상태 'p' 에서 하나의 요소 'm1'에 의해서 상태 'q'로 전의시 하나의 요소 'm2'를 출력하는 전의를 표현한다.

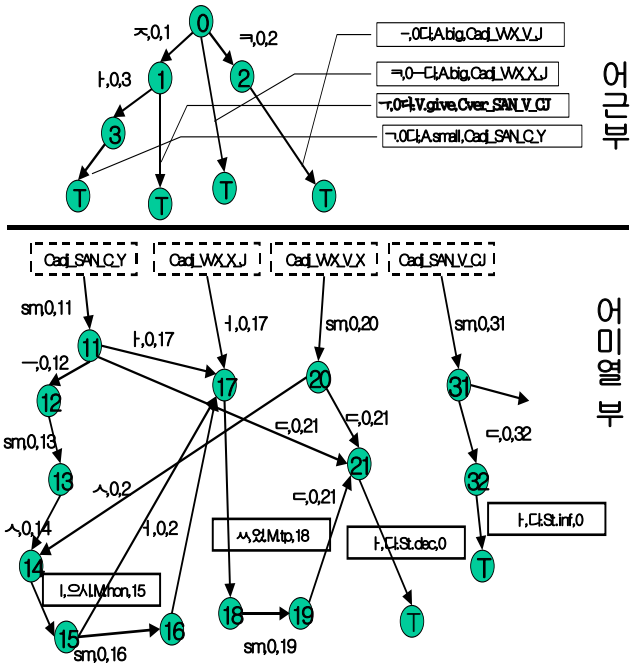
일반적인 전자 어휘사전은 하나의 초기 상태에서 각기 어휘에 따른 정보를 저장하는 복수개의 최종 상태를 가지는 나무형의 구조를 가지는 오토마타(I(n)=1, T(n)>=1)를 가진다. 본 논문에서는 다중 초기 상태와 다중 최종 상태의 트랜스두서(I(n)>=1, T(n)>=1)를 다단으로 연결하여 하나의 전자 어절 사전을 구축한다.

먼저 각 어근부와 어미열부의 트랜스두서를 구축한다. 그리고 어근부에서 어미열부를 어근의 호환성에 따라 연결하였다.

리스트형태의 어근 부분은 나무형의 하나의 초기 상태 노드와 다중 최종 노드의 트랜스두서를 이루나 그래프로 기술한 어미 열 부분은 다중 초기 상태와 다중 최종 상태 트랜스두서를 이룬다. 각각 부분에 대한 트랜스두서의 압축 기법[10]을 사용하여 압축하였다. 어휘들에 대한 오토마타에 의한 구축은 자동적인 압축을 효과를 가져 다 준다. 또한 최종 노드에서부터 동일 입력과 동일 출력을 가지는 전이들에 대해서 각기 향하는 목적지 노드의 값을 일괄적으로 통일 시켜 준다면 노드들에 대한 축소를 가져 다 줄 수 있다.

즉 초기노드에서부터 같은 깊이의 두개의 노드 Pi와 Pj에서 각기 전이 F_i(Pi, Ii, Qi, Oi), F_j (Pj, Ij, Qj, Oj)에 대하여 만일 (Ii==Ij) 이고 (Oi == Oj) 라면 Qj를 Qi로 대치할 수 있다.

<그림 7>은 각 어근부와 어미열에 대한 압축의 예를 보여주고 있다.

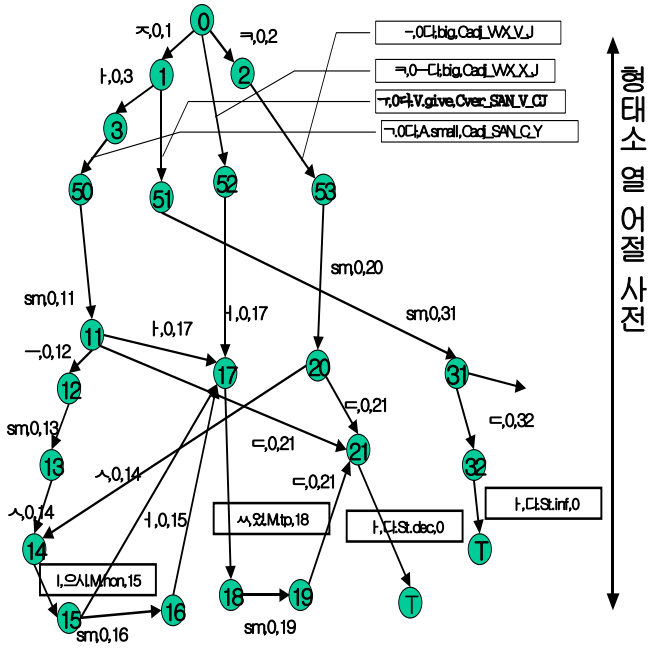


<그림 7> 어근과 어미의 압축예

어근과 어미열의 오토마타의 각 상태간의 전이는 음소, 정보, 전위되는 상태 노드들을 표시한다.

각기 어근과 어미열들의 트랜스두서로 구축된 후 어근부의 마지막 노드로의 전이에 있어서 목적지 노드의 값은 각 어근의 호환 어미열을 표시하며 실제적인 메모리의 주소를 가지지 않는다. 이는 어근의 최종 전이의 목적지 노드의 값은 프로그램의 컴파일 시, 대상 파일내에 주소값을 갖지 않는 변수와 같이 다음 노드의 주소를 갖는 것이 아니라 어미열부에 각 초기 노드의 이름을 가지고 있게 된다.

어간부와 어미열부를 통합하면서, 컴파일의 링커처럼, 실제적인 주소를 가지지 않는 어근들의 마지막 노드들에 대해서 실제적인 노드의 주소값을 어미열의 초기 상태 노드의 주소로 대치하여 준다. 'sm'은 음절간을 표시한다.



<그림 8> 어근과 어미열을 결합

본 논문은 <그림 8>의 두개의 부분에서 어근부의 미확정된 노드 값을 가지는 전이에 대해서 이에 대응되는 어미열부의 초기 상태 노드의 주소로 대치시켜 주어서 각기 압축한 어근부과 어미열부를 하나로 통합하여 하나의 형태소 열 어절 사전을 구축하였다.

본논문에서 사용한 리스트형 어근의 어휘는 <표 1>과 같다.

<표 1> 어근의 입력 어휘수

	명사	동사	형용사	관형사
어휘	14123	3603	4593	463
파생어휘	7823	5309	628	0
소계	21946	8912	5221	463
합				36542

어미열에 대해서는 64 개의 명사, 동사 및 형용사의 어근의 변화형에 따른 음성학적 형태론적 호환성을 구현하기 위해서 약 280 개의 그래프를 구성하여 어미열을 표현하여 다시 리스트 열로 총 127,461 개를 어미열을 이용하여 930 Kb 의 어근과 어미열의 트랜스두서에 의한 단일 사전을 얻었다. 이 사전을 이용하여 1 초에 41,222 단어를 처리 한다.

본 어절 사전으로 처리된 텍스트의 단어는 가능한 형태소 열을 각 형태소의 변화형 및 기본형 및 언어정보를 보여주는 문장단위의 그래프로 결과를 보여준다.

4. 결론 및 해야 할 일

본 논문에서는 언어 지식을 근간으로 하는 어절내의 형태소의 연결을 기술한 정보를 이용한 형태소 열 어절 사전을 구축하였다. 본 논문에서는 그래픽을 이용한 어휘간의 기술을 이용하여 언어기술에 있어서 언어를 기술하는 언어에 대한 지식이 없이 단순한 그래픽 도구에 의해서 언어의 지식에 의해 기술된 정보를 이용한 한국어 어절 사전을 구축하였다.

본 논문의 어절 사전을 텍스트에 적용한 결과의 정밀성은 67%를 나타내며 78%의 오류는 언어 자원의 미비이며 22%는 계속시스템을 정비하여야 하는 오류이다. 장기적으로는 언어 자원 측면에서는 어휘의 첨가 및 복합명사에 대한 기술 및 한 어절안에 발생하는 불확실성 제거를 위한 언어 정보 기술[11] 및 어절간의 정보를 이용한 문장 기술을 통한 문장 단위 구문 연구가 수행되어져야 하며 전산적인 측면에서는 한국어 텍스트상의 단어의 형태에 검색 및 언어정보에 의한 단어 검색을 처리하는 단계를 개발해 나가야 한다.

핀란드언어와 같은 굴절어에도 적용할 수 있다.

본 논문에서 사용한 UNITEX 시스템 및 한국어의 언어 자원 및 프로그램은 모두 공개 자원으로 누구나 사용할 수 있다.

참고문헌

[1] Maurice Gross, 1997, MIT Press "The Construction of Local Grammars, Finite-State Language", pp 329-354

[2] Cha J.W, Lee G.B, LEE J.H 1998,"Generalized Unknown Morpheme Guessing for Hybrid POS Tagging in Korean",Proc. Of Workshop on Very Large Corpora, Monreal, pp85-93.

[3] KIM D.B, LEE S.I, CHOI K.S, KIM G.GH, 1994, "A Two-Level Morphological Analysis of Korean", COLING, Vol.1

[4] CHOI,S.W,1999,"Implantation de dictionnaires electroniaues du coreen par autoamates finis",these de doctrant, Univ. Marne-la-Vallee,IGM

[5] Lee, G.B, LEE J.H et Yoo J.H, 1997, "Multi-level Post-Processing for Korean Character Recognition Using Morphological Analysis and Linguistic Evaluation

[6] HAN Ch.H, Palmer,,2005,"A Morphological Tagger for Korean: Statistical Tagging Combined With Corpus-based Morphological Rule Application. MT Journal

[7] SILBERZTEIN Max, 1997,"The Lexical Analysis of Natural Language, Finite-State Language Processing, MIT Press, pp 175-203

[8] Sebastien PAUMIER, 2003,"De la reconnaissance de formes linguistiques a analyse syntaxique", these de doctora, univ. de Marne-la-vallee.

[9] WOODS W.A, 1970, "Transition network grammars for natural language analysis"

[10] Dominique REVUZ,1991, "Dictionnaires et Lexiques, Methodes et algorithme", these de doctorat,Univ. Paris 7

[11] Laporte, Eric, 2001,"Reduction of lexical ambiguity", Linguisticae Investiigaiones 24-1 pp67-103