확률통계적 구문태깅의 자동화

김행곤○, 정성영, 서광준, 최기선

한국과학기술원 전산학과

Toward Automatic Probabilistic Syntactic Tagging

Hiongun Kim○, Sungyoung Jung, Kwangjun Seo, Key-Sun Choi
Department of Computer Science, KAIST

요약

언어처리에 통계 확률적인 방법이 도입되면서 현실적으로 상당한 진전이 있었지만 한국어의 경우에는 대부분 형태소 해석과 품사 태깅에 그치고 있다. 본 논문에서는 구문분석 수준에서의 통계적인 한국어 분석에 쓰일 자료 구축으로서의 구문 태깅의 방법론과 그 자동화에 대해 보고한다.

1. 서론

규칙기반의 언어처리에서부터 통계및 확률 기반으로 옮겨가면서 자연언어 처리는 상당한 진전이 있었다고 할 수 있다. 왜냐하면 만족할 만한 정밀성은 없지만 많은 경우에 성공적으로 언어를 분석해 낼 수 있는 방법들이 제공되기 때문이다. 이에 대해 [Foster91] 에서는 다음과 같은 세가지 중요한 장점들을 제시한다. 첫째, 통계적인 모델들은 더 단순하면서도 계산적으로 부담이 더 적다. 둘째, 모델의 중요한 변수들이 자동적으로 언어지므로 대상이 바뀌어도 쉽게 적응할 수 있다. 이는 점차로 자연어처리 시스템들이 많은 분야에 응용되고 있는 실정이기 때문에 꼭 필요한 성질이다. 셋째, 견고성을 부여할 수 있다. 자연언어가 갖는 개방성때문에 이 역시 꼭 필요한 기능이다.

모든 언어처리에 통계확률적인 방법론이 적용될 수 있지만 그중에서도 특히 형태소 분석 및 품사태깅에서 상당한 진전이 있었는데 [아운재93, 임철수94, 이상호94a, 이상호94b], 이는 품사태깅의 경우 문제가 비교적 단순하며 통계적인 자료인 품사태그가 풍부 코퍼스의 구축이 용이하기 때문이다.

한국어의 구문분석의 경우 어순의 자유로움 때문에 단일화에 기반한 HPSG나 LFG를 사용한 구문분석과 [장석진92, 김윤호92, 이상국93, 권혁철92] 의존문법에 기반한 구문분석들이 주류를 이루고 있다 [서광준93a, 김창현93]. HPSG에 기반한 구문분석은 기본적으로 자질구조라는 정보구조를 단일화라는 연산에 의해 점차 큰 구조로 완성해 나가는 방법으로서 구문규칙은 매우 적지만 매 단이마다 커다란 자질 구조를 갖추고 있어야 하므로 사전 정보가 무척 세밀하게 주어져야 한다. 자유로운 어순에 대해서는 효과적인 해결책이 제시되었지만 빈번한 생략현상의 경우는 여전히 해결하기 어려운 문제로 남아 있다. 한편, 의존문법에 기반한 구문 분석의 경우는 단어들간의 의존관계를 기본 단위로 보기 때문에 어순의 변화나 빈번한 생략현상등에 대해 매우 강한 견고성을 보이기 때문에 최근에 상당이 많은 주목을 받고 있는 방법론이다.

영어에서는 이미 구문분석 단계까지 나아간 통계확률적인 방법론들이 개발되었으며 다양한 실험이 진행되고 있다. 특히 대량의 텍스트로부터 자동으로 구문규칙을 추출하는 단계까지 시도되고 있다 [Charniak93]. 한국어에서는 세밀하게 기술된 구문분석용 지식을 사용해서 애매성을 해소하려는 노력은 다수 제시되었지만, 통계확률을 중심적인 도구로 사용하여 구문분석을 시도한 예가 아직 없다.

본 논문에서는 통계를 바탕으로 하는 구문분석시 나타나는 문제점들을 지적하고 이에 대한 해결책과 통계적인 정보의 축적 방법을 제안한다.

2. 통계적인 구문분석상의 문제

구문분석에 통계확률을 도입하려고 시도할 때 직면하는 두 가지 문제가 있는데, 첫째는 구문분석 자체에 대한 견해가 일치하지 않는다는 것과, 수학 적인 분석모델이 부족하다는 것이다. 각각에 대해서 간단하게 살펴보면 다음과 같다.

첫째, 위에서 소개한 한국이 구문분석의 두 흐름을 살펴보면 구문분석에서 해야 할 분석의 수준에 상당한 차이가 있음을 알 수 있다. 의존문법의 경우에는 문장에서 단어들의 지배-의존의 관계를 밝히고 그 관계가 무엇인지를 밝히는 통사적인 수준에 머무는 데 반해 [서광준93a], HPSG에 기반한 방법들에서는 상당한 수준의 의미분석까지를 포함하고 있는 [장석진92], 구문분석의 범위에 명확한 불일치가 존재한다. 따라서 구문 분석의 수준을 명확히 하는 것이 필요하다.

둘째로, 통계적인 방법들은 완전히 수치적인 해결책이기 때문에 기본적으로 잘 정리된 수학적인 분석모델이 갖추어져야 한다. 확률적인 구문분석이 활발히 연구되고 있는 영어의 경우, 형식언어 오토마타 이론에서 개발된 형식적인 분석모델이 있다. 일단 기반이 되는 문법체계에 대한 수학적 모델이 마련되면 확률을 적용할 수 있는 나양한 기술이 있는데 대표적인 것으로는 은닉 마르코프모델 (HMM)이 있다.

본 논문에서는 통계적인 자료의 축적을 위해서 의존문법의 경우만을 고려하는데, 이는 의존문법이 HPSG에 비교하여 다음과 같은 장점이 있기 때문이다.

첫째, 요구하는 정보가 더 단순하다. HPSG의 경우는 어휘정보에 많은 자질정보가 들어 있는 사전이 필수적으로 요구되지만 의존문법은 각 어절의 품사만 가지고 규칙을 기술한다. 의존관계에 대한 규칙은 어절과 어절 사이의 품사환경만으로 기술하는데, 대략 100개 이하의 관계규칙이 필요하다 [서광준93a]. 현재 95%정도의 정확도를 가지는 확률적 품사태깅이 한국어에도 이루어지고 있는 실정이므로 [이상호94a, 이상호94b]의 결과를 바로 연계하여 이용할 수 있다는 장점이 있다.

둘째로 의존관계를 밝히는 것이 구문분석의 전부라고 여기기 때문에 자유로운 어순이나 생략현상에 대해 강한 견고성을 보인다. 특히 대량의 텍스트를 다루려면 견고성은 대단히 중요 요구되는 성질이다. [서광준93b] 에서는 의존구조의 분석에 있어서, 탈문법적인 문장에 대해서도 견고하게 파싱할 수 있는 방법을 제시하고 있다.

셋째로 HPSG이론에서는 단일화연산을 통해서 통사와 의미, 화용정보를 자질구조 체계로 통합하려는 경향이 강하기 때문에 통계적으로 추출하기나 처리하기 어려운 대상들이 포함되어있다.

결국 한 마디로 표현하자면 단순하기 때문에 의존문법 체계가 더 적합하다고 할 수 있다. 다음은 본 논문에서 제시하는 통계자료 구축용 의존문법의 정의이다.

통계적인 구문분석과 의존문법

정의 : 의존문법은 단어집합과 단어에 부여되는 품사집합, 품사와 품사간의 방향성 의존관계를 기술한 규칙집합으로 나타낼 수 있다. 두 단어가 어떤 의존관계로 묶일 때 그 관계에서 지배를 하는 단어를 지배소, 지배를 당하는 단어를 의존소라고 한다. 하나의 의존소에는 지배소가 둘 이상이 될 수 없으며 한 문장에는 그 문장 전체의 지배소가 반드시 하나 있어야 하고 그 문장지배소는 다른 단어의 의존소가 될 수 없다. 문장내에서 의존관계들을 연결했을 때 엇갈린 관계(cross link)가 존재하지 않으면 투영적(projective)이라고 하는데, 어떤 문장이 의존문법에서 주어지는 의존관계를 가지고 투영적인 의존관계만으로 모든 단어를 연결하는 트리구조를 구성할 수 있으면 그 문장은 문법적이라고 말한다.

실제로 적용되는 의존문법은 이것 외에도 다양한 제약조건들이 첨가되지만 [서광준93a] 여기서는 통계적인 정보의 구축이라는 목표때문에 더 구체적인 정보들은 무시하기로 한다. 이 문법체계는 완전히 구구조문법의 일종인 X-bar문법으로 변환될 수 있다.

이에 대한 구문분석 알고리즘은 다음과 같이 간단한 다이내믹 프로그래밍 수식으로 표현가능하다.

$$V_{i,j}=\bigcup_{i\le k<j}\{h \mid (x\in V_{i,k} \text{ and } h\in V_{k+1,j} \text{ and } x\leftarrow h) \text{ or } (h\in V_{i,k} \text{ and } x\in V_{k+1,j} \text{ and } h\rightarrow x)\} \quad (2)$$

위 수식에서 $V_{i,j}$는 문장이 주어졌을 때, i번째 단어로부터 j번째 단어까지에서의 지배소가 될 수 있는 단어의 집합을 나타낸다. 의존관계 $h\rightarrow x$ 는 지배소 h가 의존소 x를 지배한다는 것을 나타낸다. 이때 초기값인 $V_{i,i}$는 i번째 단어 자체이다. 이 알고리즘에 따라 분석했을 경우 n개의 단어로 된 문장에 대해서 $V_{1,n}$에 해당되는 원소를 찾을 수 있으면 문법적인 문장이 된다. 이 알고리즘은 CYK 알고리즘과 동일한 성질을 가지고 있으므로 문장길이 n에 대해 $O(n^3)$에 동작하며, 분석을 한 후에 트리구조를 추출하려면 지배소를 어떤 $V_{i,j}$에 넣을 때, 어떤 두 원소로부터 그 지배소가 나오게 되었는지를 포인터로 연결해두면 된다. 특히 한국어의 경우에는 지배소가 항상 뒤에 온다는 지배소후위의 원칙(headword final)이 있으므로 위의 알고리즘에서 (2)번 조건은 빼버려도 상관없다.

이 알고리즘은 통계적인 분석에는 상당히 중요한 의미가 있는데, 그 이유는 [서광준93a], [김창현93], [Covington94] 에 의해 제시된 의존문법 알고리즘들은 모두 백트래킹을 동반하기 때문에 기하급수적인 복잡도를 가지는데 반해, 이 알고리즘은 $O(n^3)$에 동작한다. 기하급수적인 복잡도를 가지는 알고리즘에서는 일단 계산적으로 최적해 선정이라는 문제를 생각할 수 없다. 이런 문제 때문에 [서광준93b] 에서는 최적해를 포기하고 근사해를 구 한다. 무작위로 만들어진 구조들을 공유하기 때문에 통계적인 방법에 의해서 얻어낸 가중치를 부여한다면 [Viterbi73]에 의해 제시된 최적해 선정 알고리즘을 적용할 수 있는 기반이 마련되는 셈이다.

3. 구문태깅의 표현

구문태깅이란 구문분석에 필요한 통계적 정보를 추출하기 위해 미리 텍스트를 분석해 두는 작업을 일컫는 것으로 품사태깅과 구분하여 트리태깅이라고도 한다. 여기서는 구문태깅의 표현과 그 표현의 정당성에 대해서 고찰해보고자 한다.

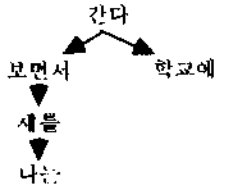
형태소분석 및 품사태깅의 경우 이미 95%이상의 높은 정확도를 얻고 있으므로 태깅작업의 거의 대부분을 자동화하고 수작업으로 수정하는 양은 상당히 축소시킬 수 있는 단계에 이르렀다. 그럼에도 불구하고 결국에 가서는 사람이 개입하여 수작업하는 것이 이쩔 수 없는 것으로 여겨진다. 구문태깅의 경우는 일단 구문분석의 예제성의 수가 너무 크며 그중에서 올바른 것을 선택하는 것이 실제 실용적으로 쓰일 수 있을 만큼 정확하지 않으므로 수작업을 고려하지 않는 것은 무의미하다.

이미 대량의 트리코퍼스를 만든 경험이 있는 [Marcus93]에 의하면 속도와 일관성, 정확도 측면에서 수작업만으로 코퍼스를 구축하는 것에 비해서 자동화된 태거와 수작업을 결합하는 것이 더 낫다고 한다. [Marcus93]은 또한 태깅작업을 두가지 방향으로 단순화시키는데, 첫째는 품사태그집합의 단순화이다. 원래 브라운코퍼스 (Brown Corpus)에서 사용했던 품사태그 집합에서 어휘 정보로 복구할 수 있는(lexically recoverable) 품사구분은 아예 하지 않고, 일관성을 유지하도록 하며, 구문적 기능도 고려하여 태그집합을 재정의한다. 본 연구의 기반이 되는 품사태그 집합 [김계후93] 역시 이와 같은 목표하에 재정리된 것으로서 이미 형태소 분석기및 태거에 적용되고 있다 [이상호94a, 이상호94b].

|Marcus93|에서는 또한 태깅하는 수작업의 속도를 심각하게 떨어뜨리는 애매성은 사람이라도 판단하기 어려우며, 판단한다 하더라도 교정자에 따라 의견이 다른 경우가 많으므로 통계적인 정보를 추출하기에는 의미가 없어지는 것을 발견하고는 이런 종류의 애매성은 처리하지 않고 그대로 놔두도록 하고 있다. 다음은 이러한 관점에서 구문태깅의 양식을 정리한 것이다

1) 가장 경제적인 표현이 요구된다.

의존문법으로 중국어의 구문태깅을 한 |Ming94|의 예를 보면 다음과 같은 문장에 대해서 의존문법에서 기본단위로 삼고 있는 (지배소, 의존소, 의존관계)의 튜플의 리스트로 표현한다.



[표현 1]{(가/pv+ㄴ다/ef,보/pv+면서/ecs,종속절)"
(보/pv+면서/ecs, 새/nc+를/jc, 목적어)
(새/nc+를/jc, 날/pv+는/exm, 관형어)
(가/pv+ㄴ다/ef, 학교/nc+에/jca, 부사어) }

|표현 1|은 의존구조가 가지는 표현의 경제성에 상당히 위배된다. 왜냐하면 똑같은 단어가 여러번 반복적으로 나타나는 것을 피할 수 없기 때문이다. 또한 문장의 어순과 태깅된 결과의 어순이 달라서 처리하기도 쉽지가 않다.

2) 문장의 어순을 유지시킬 필요가 있다.

[표현 2] (가/pv+ㄴ다/ef
(종속절 보/pv+면서/ecs
(목적어 새/nc+를/jc
(관형어 날/pv+는/exm)))
(부사어 학교/nc+에/jca))

반복을 피해 |표현 2|와 같이 기술한다고 해도 여전히 어순이 뒤바뀌는 문제가 남아 있다.

[표현 3] (S (NP 나/npp+는/jc)
(VP(SX(NP 새/nc+를/jc) 보/pv+면서/ecs)
(VP 학교/nc+에/jca 가/pv+ㄴ다/ef)))

* 여기서사용한 태그는 |김재훈94|에서 제시된 태그이다.|이상호93a,93b|에 의해 출력된 결과이다. pv:동사, ef:어말어미, ecs:종속적 연결어미, nc:보통명사, jc:격조사, jca:부사격조사, exm:관형형어미.

이에 대해 |표현 3|과 같이 구구조 문법으로 문장을 분석할 경우에는 문장에서의 단어들의 순서가 문장구조에서 그대로 유지되기 때문에 표현의 자연스러움이라는 측면에서는 더 선호된다. 그러나 우리말의 경우 의존 구조와 구구조는 별개일 수 없으며, 항상 지배소가 어떤 구조의 맨 마지막 단어라는 원칙이 지켜진다면 구구조로 표현한 것이나 의존구조로 표현한 것이나 상관없이 완전히 1-1 대응관계를 구성할 수 있다.

3) 각 어구마다 기능을 밝혀줄 필요가 없다.

우리말의 경우에는 매 어절마다 기능어인 조사나 어미가 붙어있게 마련이므로 굳이 매 구문 단위마다 그 단위의 기능에 해당하는 코드를 부여할 필요가 없다. 따라서 위의 문장을 다시 더욱 단순한 표현으로 바꾸면 다음과 같다.

[표현 4](나/npp+는/jc
(((새/nc+를/jc) 보/pv+면서/ecs)
(학교/nc+에/jca 가/pv+ㄴ다/ef)))

|서광준93a|에 의하면 임의의 두어절의 품사정보를 가지면 그 두 품사들 사이의 의존관계의 종류은 유일하게 결정되므로 이 표현에는 정보의 손실이 없다고 할 수 있다. 이는 복구가능한(recoverable)한 정보에 대해서는 코딩할 필요가 없다는 |Marcus93|의 관점과 같은 것이다.

4) 최소한의 구문분석단위는 어절이다.

위의 |표현 4|는 한가지 중요한 성질이 더 들어 있는데, 그것은 기능어인 조사/어미의 구분을 괄호에 의해서 표시하지 않았으며 최소한의 분석 단위로 띄어쓰기 단위인 어절을 썼다는 것이다. 이에 대해 두가지 이견이 있는데, 첫째는 한 어절내의 구조까지를 구문분석 단계에서 고려해야 한다는 것과, 여러 어절을 묶어서 하나의 구문 분석 단위를 만들어 주어야 한다는 구문요소화의 견해가 그것이다. 그러나 이 두가지 모두 단순화의 원리에 어긋나며 정보의 복구가능성 때문에 통계용 자료구축에서는 배제할 수 있는 사항이다.

[표현 5]
((나/npp) 는/jc
(((새/nc) 를/jc (보/pv) 면서/ecs)
((학교/nc) 에/jca (가/pv) ㄴ다/ef)))

|표현 5|는 한 어절내부까지 쪼개서 분석하는 경우인데, 한 어절 내에서의 형식형태소와 실질형태소 사이의 관계는 형태소 해석단계에서 다 분석되며 구문분석 단계에서는 형식형태소가 구조에 미치는 영향만을 따지면 대부분 지배-의존관계가 밝혀지기 때문에 굳이 세분해서 분석할 필요가 없다. |서광준93a|에서는 조용사 "~이다"의 경우는 특별히 따로 두 개의 서로 다른 의존관계가 성립가능하므로 체언과 "~이다"를 분리하여 분석하도록 제안하는데, 이 경우는 구조적인 표현력이 커지기는 하지만 경쟁하는 두 구조사이의 구분이 기계적으로 이루어지기 어렵고 설사 사람이 구분한다 할지라도 의견의 불일치가 일어날 가능성이 높기 때문에 그 구분이 통계적인 정보획득에 도움이 되지 않는다. 따라서 조용사 "~이다"의 경우에도 띄어쓰기 단위인 어절 을 구문태깅의 최소단위로 사용하는 원칙을 따르는 것으로 한다.

구문태깅의 최소 단위에 대한 또 다른 해석은 [안동언87]에 제안된 구문 요소화이다. 이는 의미적인 수준에서 보았을 때 하나의 단위가 되거나 관용적으로 자주 연달아 나타나는 경우에 형태소 분석의 후처리 단계에서 구문요소화를 통해 한 단위로 묶어주고 그 결과를 구문 분석에 넘겨줌으로써 구문 분석의 모호성을 줄인다는 것인데, 이는 통계적인 정보축적의 결과를 가지고 할 수 있는 것이지 통계적인 정보 축적과정에 쓸 수 있는 것은 아니다.

6) 지배소 후위원칙을 지킨다.

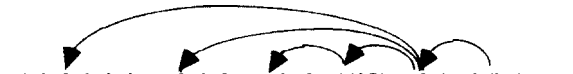
우리말의 경우 모든 어구, 어절에서 일반적으로 지배소가 뒤로 온다는 원칙이 지켜지는데 단 병렬어구만 그 원칙이 깨지는 경향이 있다. 예를 들어 다음과 같은 구문이 있으면 여러가지 문제가 발생한다.

"매체기법의 설계와 효율적 구현에 관해서"

이때 "매체기법의"가 "설계와 효율적 구현"을 수식할 경우, 가능한 표현들은 다음과 같지만,

[표현 6]

[표현 7]



[표현 8]

이 모두다 문제가 있다. [표현 6]은 의존트리에 가상의 노드가 나타나지 않는다는 특성이 깨진 것이고 [표현 7]이나 [표현 8]은 조사 "와"와 "에"의 상호관계가 불분명하다. [표현 6]과 [표현 7]은 중간의 연결구조가 하부구조를 지배하는 것으로 보는 관점이기 때문에 지배소후위 원칙에도 위배되는 경우이다. 보통 꾸밈말이 병렬어구 전체를 꾸미는 경우는 병렬어구의 마지막 요소를 꾸미는 것과 같은 것으로 보면 표현도 간단하며 정보의 손실을 없앨 수 있다. 왜냐하면 투영성 (projective, no crossing) 때문에 꾸밈 말이 병렬어구의 마지막 요소만을 꾸미는 일은 없기 때문이다. 다음의 [표현 9]는 이를 나타낸 것이다.

[표현 9]

마찬가지로 보조용언과 본용언 사이에도 지배소 후위원칙을 고수하여 다음과 같이 분석한다.

[표현 10]

요약하면 우리말의 구문태깅에 있어서는 어절 단위를 기본으로 하고, 지배소후위 원칙을 고수하며, 괄호만 치고 따로 구문기능을 부여하지 않아도 된다. 따라서 구문태깅이라 함은 품사태깅된 결과에 괄호를 메기는 작업이라고 정의할 수 있다.

4. 구문태깅의 자동화

끝으로 구문태깅을 자동화는 완전 자동화가 어렵기 때문에, 기본적으로 구문분석기와 교정하는 사람의 수작업이 협동하여 이루어지는 것이다. 자동화율을 높인다는 것은 교정자의 수작업 정도를 낮춘다는 것을 의미한다. [Marcus93]의 경우는 결정적인 영어 구문 분석기인 Fidditch파서를 쓰는데, Fidditch파서의 경우는 애매성이 있을 경우 결정을 유보하고 덜 완성된 트리를 출력한다. 이 결과를 수작업을 통해서 부분해석 결과를 "붙여주는"작업만으로 교정이 이루어진다. [Ming94]의 경우에는 구문분석기가 결정하지 못할 경우에는 교정자로하여금 결정을 내리도록 기다리는 방법을 쓴다.

이에대해서는 중간적인 입장을 취할 수 있는데, 먼저 어떤 방식으로든 최상의 구문 분석 트리를 만들어 낸 다음, 그 결과를 이용해서 교정자가 구문태깅을 하는 동안에 내려야할 결정의 횟수를 최소화 하는 방향으로 유도할 수 있다.

최상의 구문분석트리는 앞에서 제시한 알고리즘에 의해서 형성되는 중간구조에 가중치를 부여하면 이미 알려진 [Viterbi73] 알고리즘을 사용하면 최적의 구문트리를 찾을 수 있다. 이 가중치는 품사태거로부터 얻은 것도 있고 각각의 의존규칙에 부여된 가중치도 있을 수 있다.

5. 결론

이 논문에서는 한국어의 통계적인 구문분석을 하려고 할때 나타나는 문제점들을 지적하고 통계적인 구문 분석을 위한 자료구축인 통계적인 구문태깅의 방법을 제시했다. 이를 위해서 한가지 기본적인 원칙이 제시되었는데, 정보의 복구가능성이 있는 한 최대한 단순화된 형식으로 자료를 구축한다는 것이다. 또한 하부구조를 공유하도록 파싱하는 알고리즘을 제시하고 그것이 최적의 구문트리를 내는데 쓰일 수 있으며 구문태깅에 적절적인 응용이 있음을 밝혔다.

최적해를 선정하는 작업에 있어서 가중치를 어떤 식으로 부여할 지에 대한 연구가 앞으로 더 필요하고, 구문태깅 교정작업을 모델링하여 최소한의 노력으로 구문태깅 작업을 할 수 있도록 하는 것이 필요하다.

참고

[Charniak93] Eugene Charniak, "Statistical Language Learning," The MIT Press, 1993.

|Covington94| Michael A. Covington, "Discontinuous Dependency Parsing of Free and Fixed Word Order," Research Report AI-1994-02 of Univ. of Georgia, 1994.

|Foster91| George F. Foster, "Statistical Lexical Disambiguation," MS-Thesis of School of Computer Science, McGill University, Montreal, 1991.

|Marcus93| Mitchell P. Marcus, Beatrice Santorini, "Building a Large Annotated Corpus of English: The Penn Treebank," Computational Linguistics, Vol.19, No.2, 1993.

|Ming94| Ming Zhou, Changning Huang, "An Efficient Syntactic Tagging Tool for Corpora," Proceedings of COLING'94, Aug. 1994, Kyoto, Japan.

|Viterbi73| G. David Forney Jr., "The Viterbi Algorithm," Proceedings of the IEEE, Vol.61, No.3, March 1973.

|권혁철92| 권혁철, 최준영, "단일화 기반 의존문법을 이용한 한국어 분석기," 정보과학회논문지, 19권 5호, 1992.

|김윤호92| 김윤호, 이상조, "HPSG에 기반한 국어 분석기의 구현," 정보과학회 '92 가을 학술발표논문집, 1992.

|김재훈94| 김재훈, 서정연, "자연언어 처리를 위한 한국어 품사태그," 한국과학기술원 인공지능연구센터, CAIR-TR--94-55, 1994.

|김창현93| 김창현, 김재훈, 서정연, "자배가능경로를 이용한 오른쪽 우선 구문분석," 제5회 한글 및 한국어 정보처리 학술발표 논문집, 1993.

|서광준93a| 서광준, "어절사이의 의존관계를 이용한 한국어 구문분석기," 한국과학기술원 석사학위논문, 1993.

|서광준93b| 서광준, 최기선, "어절사이의 의존관계를 이용한 한국어파서의 구현," 정보과학회 '93 가을 학술발표논문집, 1993.

|안동언87| 안동언, "기계번역을 위한 한국어 해석에서 형태소로부터 구문요소의 형성에 관한 연구," 한국과학기술원 석사학위논문, 1987.

|이상국93| 이상국, 김윤호, 김재문, 이상조, "용언의 하위범주화 정보를 이용한 특수 문형의 처리 방안," 정보과학회 '93 가을 학술발표논문집, 1993.

|이상호94a| 이상호, 김재훈, 조정미, 서정연, "부분 문석결과를 공유하는 한국어 형태소 분석," 동 학술대회지, 1994.

|이상호94b| 이상호, 김재훈, 조정미, 서정연, "한국어 품사 애매성 해소를 위한 통계적 모델," 동 학술대회지, 1994.

|이운재93| 이운재, "한국어 문서 태깅시스템의 설계및 구성," 한국과학기술원 석사학위논문, 1993.

|임철수94| 임철수, "HMM을 이용한 한국어 품사 태깅시스템의 구현," 한국과학 기술원 석사 학위논문, 1994.

|장석진92| 장석진, "한국어 문법 - NLP를 위한 HPSG/K" 한국과학기술원 인공지능연구센터 기술보고서, CAIR-TR-92-33, 1992.