PPeditor: 한국어 의존구조 말뭉치 구축 도구

박은진*, 김재훈*, 김강민*, 김창현**
*한국해양대학교 컴퓨터공학과, **한국전자통신연구원
*e-mail : jhoon@hhu.ac.kr, bakeunjin@yahoo.co.kr, shoutkkm@hanmail.net
**e-mail : chkim@etri.re.kr

PPeditor: A Corpus Annotation Tool for Korean Dependency Structures

Jae-Hoon Kim*, Eun-Jin Park*, Kang-Min Kim*, Chang-Hyun Kim**
*Department. of Computer Engineering, Korea Maritime University
**Electronics and Telecommunications Research Institute

요 약

효과적인 언어처리 시스템을 개발하기 위해서는 언어정보가 부착된 대량의 말뭉치가 필요하다. 그러나, 대량의 말뭉치를 구축하기 위해서는 많은 시간과 노력이 필요하다. 이와 같은 시간과 노력을 절약하기 위해서 일반적으로 말뭉치 구축 도구를 사용한다. 본 논문에서는 한국어 의존구조 말뭉치를 구축하기 위한 도구를 설계하고 구현하였다. 본 논문에서 개발된 구축 도구는 여러 가지 특징을 가지고 있다. 1) 특정 응용분야에 관계없이 두루 사용할 수 있다. 2) 분석 단계와 분석 오류를 연계하여 작업의 집중도를 높였다. 3) 가능한 한 오류는 축적되지 않도록 하여 구축된 말뭉치의 질을 크게 개선할 수 있었다. 4) 구축된 정보는 서로 공유할 수 있도록 하여 작업의 일관성을 극대화하였다. 5) 초보자로 사용자가 쉽게 도구를 사용할 수 있도록 인터페이스를 설계하였다. 본 논문에서 개발된 구축 도구를 이용하여 8 명의 연구원이 약 2 개월 (하루에 평균 4 시간)에 걸쳐서 10,000 문장의 의존구조 말뭉치를 구축할 수 있었다. 구축된 말뭉치에는 형태소 정보, 구묶음 정보, 의존구조 정보가 부착되어 있다.

1. 서론

최근 자연언어처리 연구 분야에서는 대량의 말뭉치를 많이 사용하고 있다. 이는 대량의 말뭉치로부터 자연언어처리 시스템에 필요로 하는 언어지식을 자동으로 추출하고자 하는 데 가장 큰 목적이 있다. 그러나 대량의 말뭉치를 구축하기 위해서는 많은 시간과 노력이 필요하다. 이와 같은 시간과 노력을 절약하기 위해서 일반적으로 말뭉치 구축 도구를 사용한다[1-3]. 본 논문에서는 한국어 의존구조 말뭉치를 구축하기 위한 도구를 설계하고 구현하였다.

일반적으로 구문구조는 구구조나 의존구조로 표현할 수 있으며, 서로 변환이 가능하다[4-5]. 본 논문에서는 구문구조를 의존구조로 표현한다. 구문정보에는 여러 가지 다양한 정보가 포함되어 있다. 즉, 형태소 분석 정보, 품사 정보, 구 경계 정보, 의존구조 정보, 의존 관계 정보가 포함되어 있으며, 이들 정보는 서로서로 밀접하게 연관되어 있다. 따라서 어떤 한 정보로 쉽게 무시할 수 없다. 본 논문에서 개발된 말뭉치 구축 도구는 이와 같은 다양한 정보를 한 도구에서 수정하고 편집될 수 있도록 설계되었다. 또한 각 정보를 분석 단계와 연계하여 낮은 단계(형태소 분석)의 오류가 높은 단계(구문분석)으로 전달되지 않도록 설계하였다. 그 외에도 개발된 말뭉치 구축 도구는 아래와 같은 특징을 가지고 있다. 1) 특정 응용분야에 관계없이 두루 사용할 수 있다. 2) 분석 단계와 분석 오류를 연계하여 작업의 집중도를 높였다. 3) 가능한 한 오류는 축적되지 않도록 하여 구축된 말뭉치의 질을 크게 개선할 수 있었다. 4) 구축된 정보는 서로 공유할 수 있도록 하여 작업의 일관성을 극대화하였다. 5) 초보자로 사용자가 쉽게 도구를 사용할 수 있도록 인터페이스를 설계하였다.

개발된 도구는 의존구조 말뭉치를 구축하는 데 사용되었다. 20 어절 이상으로 구성된 10,000 문장을 구축하였으며, 약 2 개월 동안, 8 명의 연구원 하루 평균 4 시간 정도 작업으로 구축할 수 있었다. 이 결과는 객관적으로 비교할 수는 없지만 많은 시간과 노력이 절약됨을 알 수 있었다.

본 논문의 구성은 다음과 같다. 2 장에서는 말뭉치 구축 도구에 대해서 간략하게 소개하고, 3 장에서는 한국어 의존구조 말뭉치 구축 도구에 대해서 자세한 기술한다. 4 장에서는 말뭉치 구축의 효율을 개선하기 위한 보조 기능에 대해 설명한다. 5 장에서는 실제 한국어 의존구조 말뭉치를 구축하는 과정에 대해 설명한다. 마지막으로 6 장은 토의 및 앞으로의 연구에 대하여 설명한다.

2. 관련 연구

본 절에서는 말뭉치 구축 도구로서 Alembic workbench[1], WordFreak[3], 세종계획에서 개발된 반자

동 구문분석 말뭉치 구축 도구[6]에 대해서 그 기능과 특징을 간략히 살펴보고자 한다.

가. Alembic Workbench

Alembic Workbench[1]는 정보추출 시스템의 학습을 위해서 필요한 개체명(named-entity)을 부착하기 위해서 MITRE[1]에서 개발된 말뭉치 도구이다. 주요 기능으로는 다국어 지원 가능하고, 구축된 말뭉치는 SGML 형식으로 출력된다. 또한 말뭉치를 구축하는 과정에서 축적된 패턴에 따라서 어느 정도 반자동 구축이 가능하다. 그래픽 인터페이스를 통해서 사용자가 쉽게 개체명을 부착할 수 있도록 설계되었고, 사용자의 행동을 학습하여 부착작업에 도움을 줄 수 있다. 이 시스템은 2004 년에 다국어 지원이 가능하도록 크게 개선하여 Callisto[2]라는 이름으로 공개하였다. 이 시스템을 한국어에 그대로 적용하기 위해서는 형태소 분석과 같은 전처리 작업이 필요하다.

나. WordFreak

WordFreak[3]도 Alembic workbench 와 같이 인터넷[3]을 통해서 공개된 말뭉치 구축 도구이다. 자바로 구현되어 다국어와 여러 종류의 운영체제에 쉽게 적용될 수 있다. 또한 구문정보, 개체명, 대용어 정보 등과 같은 다양한 언어정보를 쉽게 부착할 수 있도록 설계되었다. 시스템의 많은 구성 요소들은 쉽게 확장할 수 있고 재사용할 수 있도록 설계되었다. 주요 기능으로는 새로운 문서를 쉽게 그리고 빨리 언어정보를 부착할 수 있도록 자동 언어정보 부착 기능이 포함되어 있다.

다. 구문분석 말뭉치 종합 관리 도구

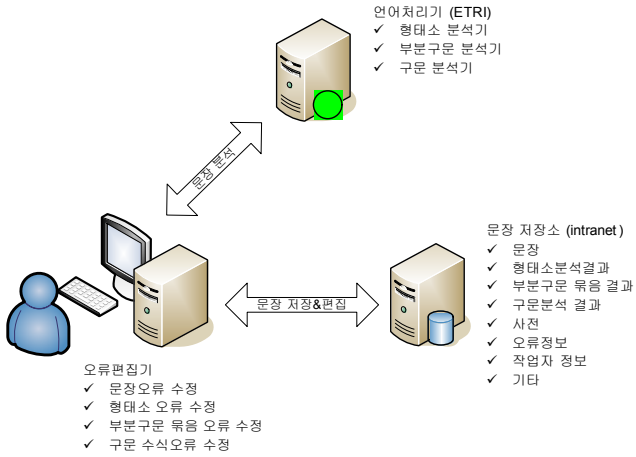
세종계획에서 구축하는 이진 구구조 구문분석 말뭉치를 변환/검색/수정의 편의성을 제공하기 위하여 개발된 종합 관리 도구이다[6]. 기존의 구문분석 말뭉치를 관리하는 작업이 독립적으로 수행되던데 비해서 구문분석 말뭉치를 관리하는 통합 환경을 제공한다. 그러나 다중 작업자 처리와 각 단계별로 문장이 파일로 존재하는 관계로 대용량 말뭉치를 구축하기엔 부적절한 도구이다.

3. PPeditor: 한국어 의존구조 말뭉치 구축 도구의 구조

본 절에서는 본 논문에서 개발된 한국어 의존구조 말뭉치 구축 도구의 구조에 대해서 자세히 기술하고자 한다. 본 논문에서 개발된 말뭉치 구축 도구는 형태소 분석기, 부분구문분석기, 구문분석기 등과 같은 언어처리기의 결과를 수정하도록 설계되었다. 반자동으로 말뭉치가 구축된다. 이런 이유로 본 논문에서는 말뭉치 구축 도구를 구문정보 편집기라고 하고 이하에서는 주로 구문정보 편집기와 말뭉치 구축 도구를 서로 혼용해서 사용하게 될 것이다. 구문정보 편집기는 언어처리기에 의해서 자동으로 분석된 언어정보의 오류를 빠르고 쉽게 수정할 수 있도록 설계되었다. 또한 본 시스템은 대량의 말뭉치 분석 작업을 다수의 작업자가 동시에 작업할 수 있도록 중앙에 데이터베이스를 구성해서 말뭉치 구축 작업이 분산되어 수행할 수 있도록 설계되었다.

가. 시스템 구조

구문정보 편집기는 언어처리기, 오류편집기, 문장 저장소로 구성되며, [그림 1]과 같다.



[그림 1] **PPeditor** 의 구조

언어처리기는 형태소분석기, 부분구문분석기, 구문분석기로 구성되어 있다. 그 밖에도 간단한 사전 검색 기능을 제공한다.

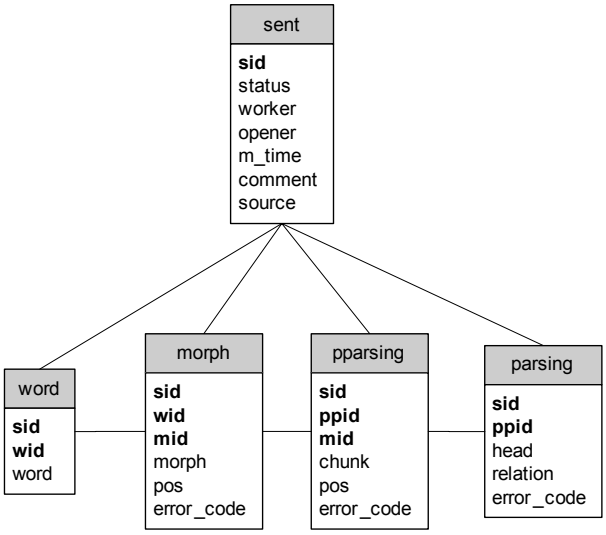
오류편집기는 자동으로 문장을 분석하며, 분석된 결과는 GUI 를 통해서 작업자(annotator)가 편리하게 관찰할 있도록 한다. 작업자는 문장 오류, 형태소 오류, 구묶음 오류, 구문구조 및 의존관계 오류를 각 분석 단계별로 관찰할 수 있다. 또한 오류 검사 기능을 통해서 이와 같은 오류가 말뭉치에 최대한 저장되지 않도록 하며, 자주 일어나는 오류에 대해서는 자동으로 수정되는 기능을 가지고 있다. 또한 난해한 언어현상에 대해서는 이전 작업자의 작업 결과를 관찰할 수도 있다.

말뭉치 저장소는 구축된 말뭉치가 저장되는 데이터베이스이며, 말뭉치를 통해서 다수의 작업자들이 동시에 작업할 수 있으며 일관성을 유지하기 위해서 필수적이다. 또한 말뭉치 저장소에 기록되어 완료 문장은 모두 다른 형식으로 쉽게 변환할 수 있다.

나. 말뭉치 저장소(데이터베이스)구조

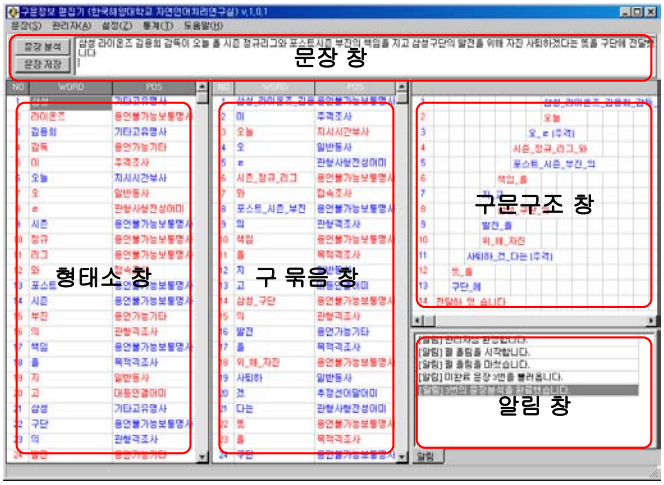
본 시스템에서 사용하는 데이터베이스 구조는 [그림 2]와 같다. 하나의 문장은 어절단위로 word 테이블에 저장되고, 형태소 분석정보는 morph 테이블에 저장되고, 구묶음 정보는 pparsing 테이블에 저장되며, 구문구조 정보는 parsing 테이블에 저장된다. 여기서 sent 테이블은 문장의 상태, 작업자, 소유자, 최종 수정 시간, 문장의 출처 등이 기록된다. 이 외 작업자 정보, 각종 사전, 오류, 문장상태를 관리하는 테이블을 유지함으로써 기계학습의 기초 자료로 사용할 수도 있다.

[1] http://www.mitre.org/tech/alembic-workbench/
[2] http://callisto.mitre.org/
[3] http://wordfreak.sourceforge.net/



[그림 2] 데이터베이스 구조

다. 오류편집기의 구성



[그림 3] 오류편집기 구성

구문정보 편집기는 크게 문장 창, 형태소 창, 구묶음창, 구문구조 창, 알림 창으로 구성된다. 각 창에서는 문장 정보, 형태소 정보, 구묶음 정보, 구문구조 정보를 시각화하여 표시하고 알림 창은 문장의 상태, 오류정보, 사전검색 결과 등을 나타낸다. 문장 창에서는 띄어쓰기 오류나 철자 오류가 수정되며 텍스트 에디터를 통해서 구현되었다. 형태소 창은 형태소 분리 오류와 품사 오류가 수정된다. 품사 수정이 입력 오류를 최소화하기 위해 콤보박스를 통해서 구현되었다. 구묶음 창에서는 구묶음 인식 오류가 수정되며 주로 잘못된 구묶음을 해제하거나 새로운 구묶음을 통해서 오류가 수정된다. 이 때 구묶음의 품사도 함께 수정된다. 품사 수정을 최소화하도록 구묶음을 해제할 경우에는 자동으로 형태소 품사 정보가 부착된다. 구문분석 창에서는 의존구조 오류 즉 중심어를 잘못 지정한 오류와 의존관계 오류가 수정된다. 주로 긴 문장에 대해 많은 오류가 있기 때문에 중심어의 위치를 쉽게 파악할 수 있도록 한 구성소(constituent)를 지정하면 그 구성소의 중심을 강조하여 쉽게 발견할 수 있도록 하였다. 이로 인해 문장 분석 작업자는 최단시간에 많은 문장을 분석하도록 구현하였다. 또한 문장분석 시간을 단축하기 위하여 각 기능별로 단축키를 제공하여 문장 분석 작업의 속도를 높이였다.

4. 말뭉치 구축의 효율을 개선하기 위한 보조 기능

앞에서도 언급했듯이 말뭉치 구축은 많은 시간과 노력이 요구된다. 이를 개선하기 위해서는 말뭉치 구축 도구가 어느 정도는 지능적이어야 하며 반복되는 작업을 가능한 한 배제할 수 있어야 한다. 본 논문에서는 이를 위해서 오류 자동 수정 기능과 오류 저장 방지 기능을 추가하였다.

가. 오류 자동 수정

오류 자동 수정 기능은 반복적인 오류 수정을 빠른 시간 내에 수정할 수 있도록 보강된 기능이다. 작업이 수정 전후의 분석 결과를 비교하여 수정 이력을 관리하고 언어처리기의 분석 결과를 수정 이력과 비교하여 문맥이 같고 빈도수가 기준값(환경 설정으로 변경할 수 있음) 이상일 경우에 자동으로 수정할 수 있도록 되어 있다. 문맥은 앞뒤 어절의 품사를 고려하여 현재의 형태소와 품사를 결정한다.

나. 오류 검사 및 저장 방지

오류 검사 및 저장 방지 기능은 구축된 말뭉치의 신뢰도를 보장하기 위한 기능이다. 작업자가 형태소 분석, 구묶음 분석, 구문구조 분석과 같이 문장분석을 할 때, 혹은 문장분석을 완료 할 때, 오류를 검사하여 오류가 있다면 문장분석이 완료되지 않도록 한다. 일반적인 오류에는 입력 오류(예: 단어와 품사에 공백이 삽입됨), 문법 오류(예: 명사 다음에 어미가 옴), 단위 오류(예: 어절 수와 형태소분석 어절수의 불일치) 등이 있다.

다. 관리 기능

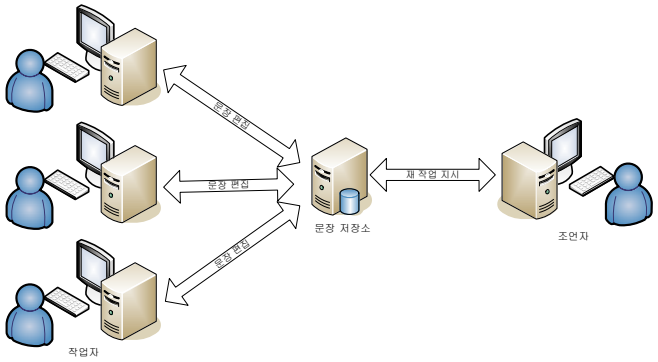
관리 기능에는 작업자 관리, 환경 설정 관리, 데이터베이스 관리, 사전 관리 등의 기능이 있다. 작업자 관리 기능은 대량의 말뭉치 분석작업으로 다수의 작업자가 문장을 분석함으로 인해 작업자를 관리하는 기능이다. 환경 설정 관리 기능은 작업 위치, 언어처리기의 주소, 데이터베이스 서버 주소 등을 관리한다. 데이터베이스 관리 기능은 말뭉치 저장소의 문장을 텍스트로 변환하는 기능인 가져오기/내보내기 기능을 제공한다. 사전 관리 기능은 각 단어의 뜻이나 품사정보를 검색할 수 있는 사전 검색 기능과 구축된 말뭉치로부터 새로운 단어를 추가하는 기능 등이 있다.

그 밖에도 작업자별 작업량과 같은 각종 통계치를 계산하는 기능을 제공한다. 작업자들을 관리하는 사용자 관리 도구와 실시간 작업 진행 상태를 파악할 수 있는 통계보기를 제공하고, 일관성 있는 분석작업이 되기 위하여 문장 작업 지침서를 실시간으로 제공하며, 이미 작업한 완료문장에서 검색하는 기능을 제공한다.

5. 한국어 의존구조 말뭉치 구축

본 논문에서 개발된 말뭉치 구축 도구로 한국어 의존구조 말뭉치를 구축하였다. 말뭉치의 구축 대상 문장은 20 어절 이상으로 구성된 10,000 문장을 구축하였다. 구축 작업에 투입된 인원은 1 개월 가량을 말뭉치 구축 훈련을 거친 후, 약 2 개월 동안 8 명의 연구원이 투입되었다. 본 절에서는 본 논문에서 개발된 말뭉치 구축 도구를 이용하여 말뭉치가 구축되는 과정과 구축 과정에서 발생되는 오류의 유형을 살펴보고자 한다.

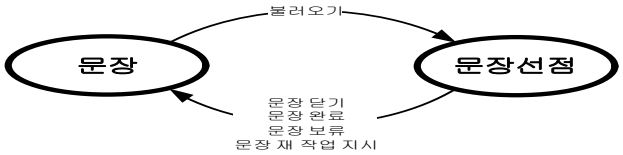
가. 말뭉치 구축 과정



[그림 4] 말뭉치 구축 과정

말뭉치 구축을 위한 작업자는 일반 작업자와 조언자로 구성된다. 일반작업자는 미완료 문장이나 재작업 지시문장을 불러와서 분석 결과의 오류를 수정한다. 조언자는 작업자 중에서 가장 선임 작업자를 말하며, 난해한 문장 구조의 해석이나 작업의 진행을 관리한다. 작업자는 난해한 문장에 대해서는 작업보류를 신청할 수 있으며 보류된 문장은 분명한 이유를 기록하여야 한다. 조언자는 문장저장소에 저장된 보류 문장이나 완료 문장 중에 적절치 않는 문장을 선별하여 그 문장에 별도의 주석을 부가하여, 원래 작업자에게 재작업을 지시함으로써 문장분석의 정확도를 높였다.

각 문장은 다른 작업자가 같은 문장을 불러올 수 없도록 문장을 선점함으로써 문장의 무결성을 유지하였다. 그리고 각 문장은 문장의 작업상태에 따라 분석되지 않은 미완료 문장, 작업자가 분석을 끝낸 문장인 완료 문장, 작업 중 난해한 문장이나 모호한 문장을 보류한 보류 문장, 완료 문장이나 보류문장을 조언자가 주석을 부가하여 재작업을 지시한 재작업 지시 문장으로 나뉜다. 각 문장은 [그림 5]과 같은 과정을 통해 작업이 완료된다.



[그림 5] 문장 상태 변화

나. 말뭉치 구축 과정에서 발생한 오류의 유형

자동 분석결과 일반적으로 나타나는 오류는 문장, 형태소, 구묶음, 구문구조 등으로 나누어 보면 다음과 같다. 문장 분석결과에 나타나는 오류로는 " 저희들 한테" 와 같은 띄어쓰기 오류, " 등안시" 와 같은 철자오류가 많이 나타나고, 형태소 분석결과에는 " 분리+하+는" 과 같은 분리 오류, " 집을 지+어" 와 같은 형태소 복원오류, " 김정일/기타고유명사" 와 같은 품사 부착 오류 등이 많이 나타난다. 그리고 구묶음 분석 결과에 나타나는 오류에는 " 하+게//되+ㄴ" 과 같은 인식오류, " [다음_,_자기_집_옆]+에" 와 같은 묶음 오류 등이 많이 나타난다. 이와 같은 오류를 빠르게 수정/편집하기 위하여 어절 토클, 형태소 삽입, 삭제, 구문 분리, 사전검색과 같은 기능을 제공함으로써 문장 분석 속도를 증가하였다. 또한 구문구조 분석결과에서 구문 수식관계를 시각화함으로써 수식구조를 빠르게 파악하도록 하였다. 구문구조 분석결과에는 구문 관계가 별도로 부착되어 있지 않음으로 인해 작업자가 구문의 의미를 파악하여 구문관계를 부착하는 기능을 제공한다.

6. 토의 및 앞으로의 연구

본 연구에서 개발된 구문구조 부착 도구는 기존의 많은 도구와 다르게 모든 분석 단계별 오류를 검사하고 수정할 수 있도록 하여 구문구조 분석 오류를 보다 쉽게 수정할 수 있도록 하였다. 이 방법은 오류를 수정하는 데 있어서도 각 오류를 분리해서 수정함으로써 오류 수정의 능률과 집중도를 크게 향상시킬 수 있었다.

앞으로 오류 자동 수정 기능을 더욱 개선하여 가능한 한 수정된 결과를 다시 수정하는 일을 최소화해야 할 것이다. 이를 위해서 변환 기반 기계학습 방법[7]을 적용하여 오류 수정 규칙을 더욱 다양화할 계획이다.

참고문헌

[1] D. Day, J. Aberdeen, L. Hirschman, R. Kozierok, P. Robinson, and M. Vilain, "Mixed-Initiative Development of Language Processing Systems", *Proc. of the ANLP*, pp. 348~355, 1997.

[2] 임준호, 박소영, 곽용재, 임해창, 김의수, 강범모, "구문패턴을 이용한 반자동 구문분석 말뭉치 구축 도구", 제 14 회 한글 및 한국어정보처리 학술발표 논문집, pp.343-350, 2002.

[3] T. Morton and J. LaCivita, "WordFreak: An Open Tool for Linguistic Annotation", *Proc. of the NAACL,* pp. 17-18, 2003.

[4] H., Gaifman "Dependency systems and phrase-structure systems", *Information and Contro*l, vol. 8, pp. 304-337, 1965.

[5] S. Höfler, *Link2Tree: A Dependency-Constituency Converter*, Ph.D. Dissertation, Institute of Computational Linguistics University of Zurich, 2002.

[6] 문화관광부, 21 세기 세종계획 국어 기초자료 구축, 연구보고서, 2003.

[7] G. Ngai , R. Florian. "Transformation-Based Learning in the Fast Lane", *Proc. of the NAACL*, pp. 40-47, 2001.