다국어 말뭉치 분석기의 한국어 처리 구현에 관한 연구

허현규*, 정혜명**
*IGM
**김포대학 멀티미디어과
e-mail : minpere@nate.com

A Study on Implementation of treatment of Korean in multi-Language Corpus Analyzer

Hyun-Gue,Huh*, Hye-Myoung,Chung**
* Laboratoire d'informatique Gaspard-Monge
**Dept. of Multimedia, KimPo College

요 약

말뭉치 분석기는 언어 연구에 필요한 도구로써 망뭉치 분석을 통한 언어 정보의 추출, 적용 및 확인용으로 사용할 수 있다. 본 논문에서는 언어 기술을 국부 문법에 의한 그래픽적인 기술방법으로 처리하는 말뭉치 분석기를 이용하여 한국어 텍스트를 연구하기 위하여 기존의 굴절어 중심으로 구현되어진 다국어 말뭉치 분석기에 한국어와 같은 교착어들의 텍스트 처리를 위한 기능을 구현한다.

1. 서론

말뭉치 분석기는 언어 분석에 따른 결과를 일반적인 텍스트에 적용하여 언어 분석 결과를 확인하는 도구로 사용된다. 기존의다국적 말뭉치 분석기는 유럽어를 이용한 언어학적인 분석 도구로 사용되어저 왔으나 한국어와 같은 굴절어에 대해서는 동일 한 방법으로 처리를 하지 못 한다. 본 논문에서는 한국어 텍스트를 처리 분석할 수 있도록 기능 구현을 위하여 말뭉치 분석기에 사용될 언어학적인 분석 지식을 기술 할 수 있는 방법을 제안하고 이를 이용한 수집된 한국어 어휘 정보 처리 및 한국어 텍스트의 처리를 할 수 있는 기능 구현에 관한 것이다.

본 논문에서는 언어 정보의 기술 방법으로 국부 문법을 이용하고 그래픽 사용자 인터페이스를 이용하여 언어 정보를 표현하며 한국어 말뭉치에 대한 처리 결과를 문장단위의 오토마타로 표현하였다. 유럽어를 처리하는 다국어 말뭉치 분석기에 단어 단위의 처리가 아닌 형태소 단위의 처리 방법에 의한 텍스트 처리를 하였다. 언어학적인 분석 및 기술이 도구인 다국어 말뭉치 시스템에 한국어 및 교착어 처리를 구현을 방법을 제시한다.

2. 기존 연구

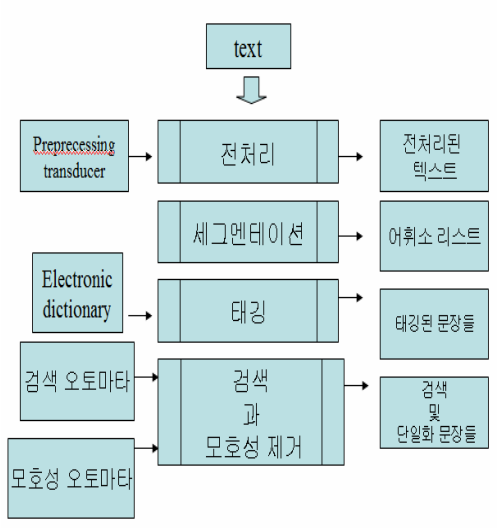
말뭉치 분석 시스템의 구축하는 데 있어서 언어 정보의 추출을 위한 전산적인 데이터 처리 뿐 아니라 언어학적 분석에 위한 언어 정보 데이터의 수집과 유지 및 정보의 기술이 중요시 된다. 언어 정보요소에 어휘, 문법 등의 정보를 기술하기 위하여 어휘 문법을 근간으로하는 국부 문법[1]을 이용한 방법을 이용하였다. 어휘간의 연결 정보에 대한 언어학적인 기술을 FST(Finite State Transducer)[2] 근간의 RTN(Recusive Transition Network)을 이용하여 각 어휘의 어휘 정보의 표기 방법을 언어 정보 기술 방법으로 사용 하였다. RTN 을 이용한 다국어 말뭉치 분석기인 INTEX[3]는 그래픽 인터페이스를 이용한 국부 문법을 표현하여 사용자 즉 언어 학자등이 쉽게 어휘소의 정보를 기술할 수 있는 방법을 지원한다. 기존 상용화된 INTEX 에 비교해서 유니코드를 이용한 다국어 텍스트 처리를 가능하게 하고 프로그램의 운영체제에 따른 호환성의 문제를 최소화 하기 위한 JAVA 와 C 로 구현하면서 언어 정보 기술 방법을 그래프에 의한 FST 기법 지원하고 순수한 학문적인 언어 도구로써 말뭉치 프로그램의 연구 목적의 개방형 프로그램을 전제로 개발된 UNITEX 가 개발 되게 되었다. 현재 UNITEX 는 WINDOW, UNIX, MAC 운영체제상에 JAVA 기반으로 수행되며 쉽게 설치하여 사용이 가능하다.

UNITEX[1] 말뭉치 분석기는 크게 4 개의 부분으로 기능을 나눌 수 있다. 하나는 사전 구성부로써 어휘들의 정보를 기술하여 사전을 구축하는 부분이다 두 번째는 텍스트 처리 단계로 흐름을 제어 하는 부분이

[1] http://www-igm.univ-mlv.fr/~unitex/

며 세번째는 언어 규칙을 기술하는 데 이용하는 그래픽 사용자 인터페이스 부분이다. 그리고 용례 추출을 위한 부분으로 나누어져 있다.

UNITEX 말뭉치 분석 시스템에서 유럽어의 처리는 (그림 1)과 같다. 시작단계에서는 입력 텍스트에 대해서 텍스트의 전처리 과정을 이용하여 축약된 어휘들에 대한 원형을 복구 및 문장의 단위로의 텍스트의 분리를 수행한다.

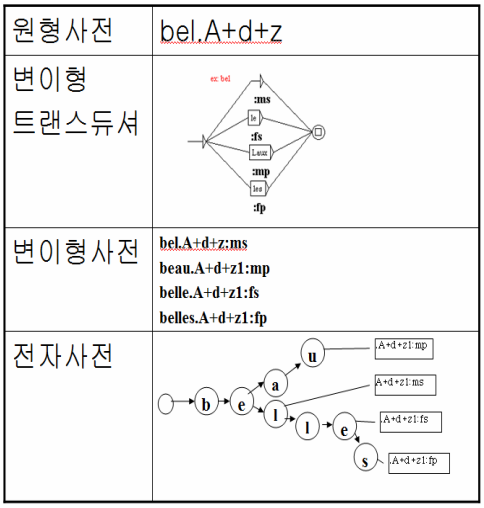


(그림 1) 텍스트 처리 흐름

입력 텍스트의 전처리후 공백, 심볼들의 구분자에 의한 텍스트의 세그멘테이션을 수행하여 텍스트 상의 단어와 특수문자를 분리 추출하여 입력 텍스트를 구성하는 어휘 리스트를 추출한다. 이 추출 어휘 리스트는 텍스트에 분포하는 어휘의 빈도를 자동적으로 보여준다. 어휘 리스트에 기 분석되어진 언어 정보를 가진 전자 사전을 이용하여 텍스트상의 어휘 리스트에 언어 정보를 테깅한다. 각 문장을 이루는 어휘소에 대해서 태그된 정보와 함께 각 어휘의 활용 모습에 대해 어휘 정보를 하나의 노드로 하여 어휘간의 연결을 전이로 표시하는 오토마타 형태로 하나의 문장을 나타내는 그래프를 보여준다. 이때 어휘마다 모호성이 존재하는 단어에 대해서는 여러 개의 경로로 각각의 의미를 보여 주게 된다. 이러한 모호성이 존재하는 경로에 대해서는 모호성 제거를 위한 오토마다를 이용하여 모호성이 제거된 문장그래프를 얻을 수 있다. 또한 검색을 위해서 검색 조건을 정규규칙으로 표현한 방법을 그래프 오토마타로 표현하여 원하는 형태의 문장이나 구 또는 단어를 찾아낸다.

유럽어의 전자 사전의 생성의 흐름은 (그림 2)와 같다. 형용사 “bel"의 언어 정보 “A+d+z"에 대한 어휘의 리스트 형태로 기술된 원형사전의 어휘에 대해서 변이형의 파생 방법을 기술한 트랜스듀서를 이용하여 리스트 형태의 변이형 사전을 얻게 된다. 그림에서 트랜스듀서는 단수/복수(s/p)와 남/여성(f/m)에 따른 정보 및 변이에 따른 치환하는 명령을 보여준다. 트랜스듀셔 처리과정에서 얻어진 변이형 사전 리스트를 입력으로 FST 형태로 압축된 전자사전을 얻을 수 있다. 이때 획득된 전자사전은 텍스트 상의 형태소의 표면형태(surface form)의 형태의 사전을 이룬다.

유럽어 말뭉치 처리기의 언어 처리 방식은 한국어에 대해서 언어학적으로 어휘적 구성 형태의 교착어와 굴적어의 차이와 전산적으로는 알파펫과 음절 문자 처리, 한국어 텍스트상의 한문, 한자, 알파펫의 혼용에 따른 처리가 달라야 한다. 한국어 말뭉치 처리에 있어서 UNITEX 가 기존 굴적어인 유럽어 중심의 전산 처리 구조를 가지고 있어서 단어 사전 기반의 언어학적인 기술 방법과 단어 기반의 사전 분석의 방법을 이용하므로 한국어와 같은 교착어에 대하여 굴적어 처리 방식으로 사전 형성 및 텍스트 처리 흐름으로 처리할 수 없다. 본 논문은 한국어 처리를 위하여 전자사전의 구축과 관련된 한국어의 어휘의 기술 방법 및 한글 텍스트 처리를 위한 한글의 음절 처리, 한글 코드의 처리를 위하여 단어 중심 사전이 아닌 형태소 단위의 형태소 열에 의한 어절 사전을 구축하였다.



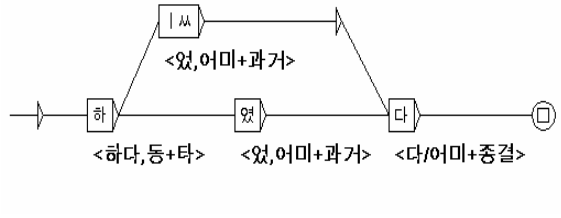
(그림 2) 사전의 구축 순서

III. 본론

한국어와 유럽언어의 텍스트 처리에서 기본적으로 기술 되는 문자의 차이와 언어적인 구문의 차이와 어절 단위의 처리에 의해서 차이가 있을 수 있다. 구문의 차이로 인한 문장의 술어와 주어의 위치의는 차이를 가진다. 유럽어의 텍스트의 단어는 즉 형태소의 표면형태는 단일 형태소에 성이나 수에 표현하는 한두개의 형태소가 붙어 있는 모습이지만 한국어와 같은 굴적어는 여러개의 형태소가 연결 되어 있는 형태소열로 보이며 각 형태소가 음성학적 규칙에 의해서 변이형을 가질 수 있다. 하나의 문자에 대해서 유럽어는 하나의 알파벳 문자가 대응 되지만 한국어는 음절을 이루는 음소의 집합임을 나타내며 또한 형태소간의 융합이 하나의 음절 안에 나타나는 모습을 볼 수 있다(예:나는(나:Nom+는:Postpostion subjective =>난).

문자 처리에 대한 문자 코드에 있어서 단일 바이트 코드와 다중 바이트 코드의 차이와 동일 코드 방식이어도 한국어에서는 조합형과 완성형 있다. 이러한 코드의 다양성에 따른 문제는 언어에 상관없이 단일한 코드 시스템인 유니코드의 사용으로 입력 텍스트의 통일성을 보장한다.

다국어 말뭉치 분석기에 한국어의 처리를 위해서 기존 유럽언어의 텍스트 처리를 위한 흐름과 동일한 흐름을 유지하면서 한국어와의 다른 점을 구현하기 위해서 먼저 사전의 구축 및 텍스트의 처리의 단위를 알파벳 즉 한국어에서는 음소 단위로 처리를 하였다. 텍스트상의 단어 즉 띄어쓰기를 위한 공백과 특수 문자에 의한 분리 되는 텍스트의 세그멘트 단위에 대한 처리에서 원형(lexical form, canonical form)에 대한 표면형태(surface form) 에 대한 텍스트사의 어절 사전으로 한국어 전자 사전을 구축하였다. 일반적으로 한국어의 형태소 분석이라는 것은 단어의 표면형태에 대해서 단어를 이루는 어휘소의 열을 발견하는 것이다. 이를 위해서 형태소 분석에서는 어휘 사전을 이용한 가능한 형태소 열들을 모든 추출하고 추출 된 형태소 열에 대해서 연결 규칙을 통계학적인 방법으로 올바른 어휘열을 찾아 나가는 방법을 쓰고 있다. 이러한 방법은 어휘 사전과 형태소 열결 규칙을 분리하여 저장 처리 및 관리하게 되어있다. 본 논문에서는 하나의 어절 사전을 형태소 열로 표현한 어절 사전을 구축하였다. 형태소 열로 구성된 어절 사전을 구축하기 위해서 하나의 형태소의 표현은 그래픽적으로 (그림 3)과 같이 기술하도록 제안 하다.

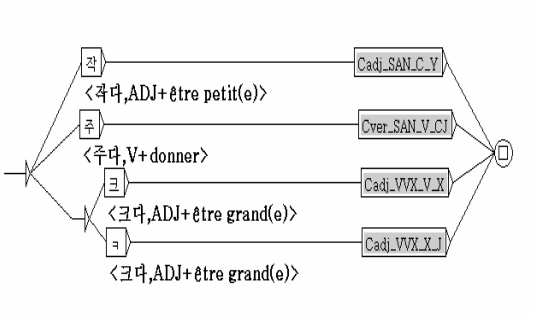


(그림 3) 어절의 기술

즉 기존의 어휘 언어 정보와 연결 규칙을 따로 구분하여 기술하는 방식에서 어휘소와 이에 연결될 수 있는 규칙을 동시에 기술하는 방식을 이용하여 기술한다면 이를 이용한 형태소 열로 구성된 어절 사전은 어휘소 열 추출후 어절내 올바른 형태소열을 찾는 규칙을 따로 가지고 있지 않아도 되며 한번의 사전 검색으로 모호성을 가지는 모든 형태소 열이 추출될 수 있다. 어절 사전을 이루는 언어 정보 데이터는 모든 어절에 대해서 그래프로 표현하는 것은 불가능하며 어휘의 어근에 대한 사전은 리스트 형태로 존재한다

우리는 (그림 4)처럼 오토마타를 이용한 그래프 표현으로 3 개의 어휘를 표현해 줄 수 있다.

오토마타의 하나의 노드는 어휘의 표면 형태로 표현하고 이에 대한 언어 정보를 표시하여 주며 또한 다른 형태의 노드는 서브 그래프를 표시하여(회색) 각 서브 그래프에 각 어휘에 따른 어미들의 열을 표시하여 줄 수 있다.



(그림 4) 어절의 그래프 표현

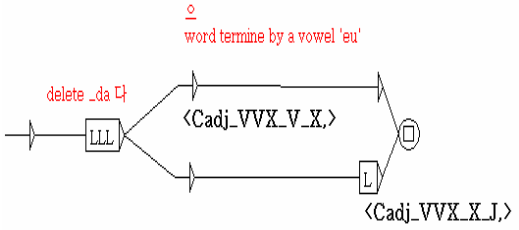
예로서 "크다"는 형용사는 변이형을 두가지 가지며 각각은 음운의 변화에 따라 두가지 형태의 어미 (Cadj_VVX_V_X, Cadj_VVX_X_J)를 가짐을 나타내 준다. 두개의 어미를 표현하는 것은 다른 서브 그래프로 가질 수 있는 어미열들 나타낸다. (그림 4)처럼 어휘들에 대한 그래프적인 기술 방식은 아래와 같이 리스트 형태로 나타내 줄 수 있으며 이는 (그림 2)에서의 변이형 형태소 사전과 같은 형태로 보여준다.

```
;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
```

리스트 형태의 원형의 어휘에 대해서 어휘 원형의 형태와 언어 정보 및 언어의 어절내의 연결 정보를 아래와 같이 표현할 수 있다.

```
크다,V,Cadj_VVX
```

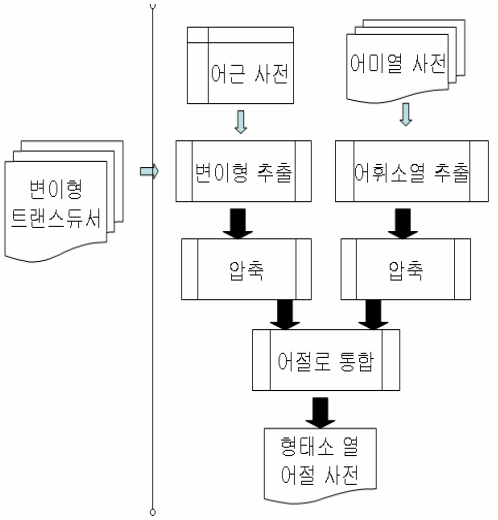
어휘소의 원형에 대한 정보의 기술은 원형이 가지는 언어 정보 및 형태소의 변이형 정보를 기술한다. 이러한 원형 리스트를 이용하여 그림 2 와 같이 (그림 5)의 트랜스듀서를 적용한다면 우리는 위의 형태의 변이형 리스트를 얻을 수 있다. 변이형 트랜스듀서는 원형에 대한 변이형의 구축 방법으로 'L'은 하나의 음소의 삭제를 의미하며 변이형이 가지는 연결 정보를 기술하게 된다. 이때 변이형 트랜스듀서를 가지는 파일의 명은 원형의 트랜스듀서 정보명과 동일하게 하다. 즉 변이형 트랜스듀서는 각 원형에 대한 변이와 연결정보를 같이 기술하도록 하였다. 여기서 연결정보란 각 변이형에 연결될 수 있는 형태소 집단을 의미한다.



(그림 5) 변이형 추출기

본 논문에서는 사전을 구성하기 위한 데이터의 형태인 어절의 어근은 리스트 형태의 사전에서 변형 트랜스 듀서를 이용하여 어근의 변이형 사전을 구축하였다. 각 어근의 어휘소에 언어학적 연결 관계를 가지는 어미들의 어휘소 열을 그래프로 기술 하였다.

두 개의 형태의 변이형에 대한 데이터에 대해서 각개의 압축 사전을 만들었다. 이때 어근들이이루는 오토마타는 최종 노드에는 자신들의 언어 정보와 가능한 연결 정보를 가진다. 어미들로 이루어진 압축 사전은 하나의 시작 노드가 아닌 다중의 시작노드로 구성하여 다중의 최종 노드가 존재한다. 어미열에 대한 오토마다의 최종 노드는 해당 어휘소의 정보와 연결정보의 종단임을 나타내는 정보를 가지게 된다. 어근을 이룬 오토마타의 최종 노드에 연결정보를 어미열들을 이룬 오토마타의 초기 노드의 위치로 바꾸어 주면 두 개의 오토마타는 연결이 된다. 이렇게 연결된 오토마타는 하나의 어휘소 열의 오토마타로 구성된 어절 사전으로 된다. (그림 6)은 어휘소 형태소 열로 구성된 어절 전자 사전의 형성 흐름을 보여 준다



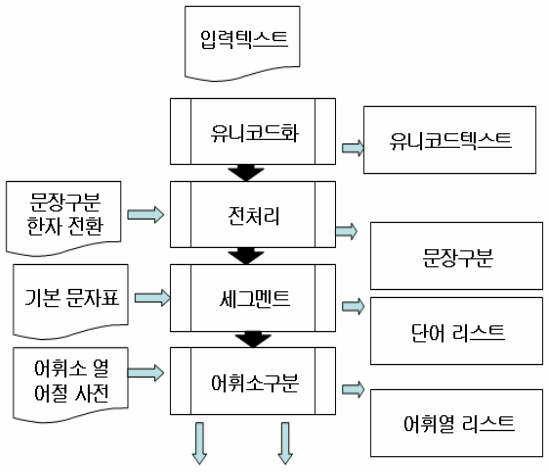
(그림 6) 어절 사전 구축

.

한국어 텍스트 파일의 처리를 위하여 그림 7 과 같은 흐름을 갖는다. 처음 입력파일를 유티코드로 전환하고 문장 구분 심볼을 삽입한다. 영문의 경우 문장의 첫글자는 대문자로 되었 있는데 이러한 단어들에 대한 균등한 전산 처리를 위하여 대소문자 및 특수문자의 일반화한다. 이를 응용하여 한국어 처리에서는 한자에 대응 한글로 전환하였다. 즉 한자와 한글의 전환 테이블을 구성하여 처리하였다.

유니코드는 각 나라의 텍스트를 하나의 코드로 통일 시키며 각 나라마다 기본 문자 집합, 자소(알파펫)를 가지게 된다. 언어별 텍스트의 처리에서 언어별 기본문자를 구분하여 주어야 한다. 세그멘트의 입력으로 쓰이는 “기본 문자표”는 처리하고자 하는 언어를 구분하는 데 사용된다. 한국어 텍스트의 경우는 한글, 한자, 영어가 혼합되어 사용되어 진다. 기본 자소 이외로 쓰여진 단어에 대해서는 심볼열로 처리하여 언어적 분석에서 제외 시킬 수 있다.

본 논문에서는 3 만여개의 어근 어휘를 사용하였고 어미의 형태소 열을 표현하기 위해서 64 개의 음운학적 언어학적으로 어미군을 분리하여 표현하기 위해서 280 개의 그래프로 표현하였다.



(그림 7) 한글 텍스트 처리

표 1

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

V. 결론

언어학적으로 국부 문법을 이용한 다국어 말뭉치 분석기에 형태소 단위의 처리를 이용한 한국어 처리를 가능하게 하는 형태소 열 어절 사전을 구축하고 한국어의 구조를 개방적이고 명시적으로 표현할 수 있도록 하여 주었다. 현재까지는 어절내의 모호성이 존재하는 형태의 형태소 열의 표현을 하고 있지만 모호성 구문단계에서 표현하여 제거하여 주는 개방적인 표현 방법과 처리 방법이 형태소 단위에서 가능하도록 하여 주어야 하고 검색에서도 어절단위 검색만 가능한 것을 형태소 단위 검색이 가능하도록 프로그램의 추후 개발이 진행되어야 한다.

참고문헌

[1] Maurice Gross.. The Construction of Local Grammars, in E.Roche et Y.Schabes (eds.), Finite-State Language Processing, Cambridge, Mass./London, The MIT Press, pp. 329-352. 1997

[2] Woods W.A."Transition networ 사 grammars for natural languge analysis, Comm. of the ACM, Vol 13, Issue 10, 1970

[3] Max D. Silberztein. INTEX: a corpus processing system, in COLING 94 Proceedings, Kyoto, Japan.1994.