COM 기반의 다목적 형태소 분석기를 이용한 명사 추출기

이중영, 신병훈, 이공주, 김지은, 안상규
한국 마이크로소프트 개발부
서울특별시 강남구 대치동 892 POSCO 센터 서관 6층, 우: 135-777
{joonglee, bhshin, kjoolee, jeeeunk, sgahn}@microsoft.com

Noun Extractor based on a multi-purpose Korean morphological engine implemented with COM

Joong Young Lee, Byuoung Hoon Shin, Kong Joo Lee, Jee Eun Kim, Sahng Gyou Ahn
R&D, Microsoft Korea

요 약

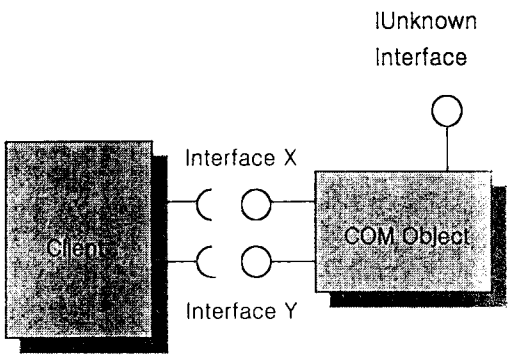
한국어 형태소 분석기는 한국어를 분석하여 여러 다른 응용프로그램에 적용할 수 있는 기본적인 도구이다. 형태소 분석기를 응용하여 맞춤법 검사기나 정보검색, 기계번역, 음성인식 등에 적용할 수 있다. 본 논문에서는 형태소 분석기를 이용하여 여러 응용프로그램에 다목적으로 적용할 수 있도록 COM(Component Object Model)으로 인터페이스를 설계하고, 일례로 명사를 추출하는 응용프로그램을 구현하였다.

1. 서론

형태소 분석기는 자연언어를 처리하는 기본적인 도구이다. 형태소 분석기를 통해서 사용자가 입력한 문장이나 음성 등의 데이터가 의미 있는 단위로 분석된다. 이 분석된 데이터는 다시 여러 응용프로그램에 입력되어 사용자가 원하는 형태로 출력된다. 여러 응용프로그램의 예를 들면, 사용자가 입력한 문장의 오류를 검사하는 맞춤법 검사기일 수도 있고, 정보 검색에 쓰이는 명사 추출기일 수도 있다. 현재 한국 마이크로소프트 내에서 사용하는 한글에 관련된 엔진들은 맞춤법 검사기, 한글/한자 변환기, 도움말 마법사, 인덱스 서버, 구문 분석기(parser) 등이 있다. 이 프로그램들은 모두 형태소 분석기를 기반으로 하고 있으며, 그 용도에 따라서 각각의 형태소 분석기는 조금씩 다르게 구현되어 있다. 또한 사용되는 사전도 다르기 때문에 각각의 프로그램들을 유지, 관리하는 과정에서 많은 시간과 인력이 낭비된다. 이러한 문제를 해결하기 위해서 형태소 분석기를 COM으로 설계하여 각각의 응용프로그램이 형태소 분석기를 공유하도록 하였으며, 명사 추출기를 구현하여 ETRI에서 주관한 MATEC99 (Morphological Analyzer and Tagger Evaluation Contest 99)에 참여하였다[1].

2. COM의 소개

COM(Component Object Model)이란 component와 client 간의 표준화된 인터페이스를 지칭한다. component를 형태소 분석기로, client를 명사 추출기로 가정하면 <그림 1>과 같은 기본 개념을 얻을 수 있다.



< 그림 1: COM Object와 Interfaces >

모든 COM Object는 IUnknown Interface를 가지고 있으며 client는 IUnknown Interface를 통해서 COM Object가 가지고 있는 Interface X, Interface Y를 얻은 다음에 각 Interface가 가지고 있는 method를 호출하여 사용한다.

2.1 COM의 특성

COM은 다음과 같은 특성을 가진다.

Language Independence(언어 독립적): COM은 프로그램 개발 언어가 아니라 개발 언어에 대한 규약이므로 어떤 프로그래밍 언어로도 개발할 수 있다.

Binary Standard(이진 표준): COM으로 개발된 component는 어떤 언어로 개발되었든지 DLL, EXE 등의 binary만 존재한다면 client가 사용할 수 있다.

Version Control(버전 관리 능력): COM으로 개발된 component는 여러 client들에 각각의 버전에 맞는 기능을 제공할 수 있다.

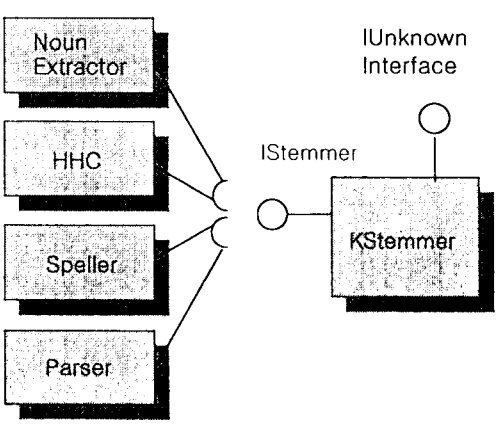
Location Transparency(장소의 투명성): component가 어디에 존재하든지 상관없이 client는 레지스트리에서 component의 GUID (Globally Unique IDentifier, 128bit 숫자)를 참조하여 사용한다.

Encapsulation(코드의 재활용): COM은 표준화된 인터페이스이기 때문에 component나 client의 설계가 변경된 경우에도 인터페이스만 수정하면 사용할 수 있으므로 코드를 재활용하기가 쉽다.

2.2 COM의 응용

여러 client들은 하나의 COM object를 공유하면서 사용할 수 있다. 이와 같은 원리에 의해 COM으로 구현된 형태소 분석기는, 맞춤법 검사기, 명사 추출기, 한글/한자 변환기 등의 client들에 공통된 인터페이스로서 뿐만 아니라 각각의 고유 인터페이스로서도 이용할 수 있다. <그림 2>는 형태소 분석기의 IStemmer 인터페이스를 여러 client들이 공유하는 것을 보여준다.

< 그림 2: 응용프로그램들의 IStemmer 공유 >



3. 다목적 형태소 분석기

본 형태소 분석기는 개발 초기부터 한글/한자 변환기, 한글 맞춤법 검사기, 한국어 구문 분석기 등에 응용할 목적으로 개발되었다. 그러므로, 저장하고 있는 정보의 양이 단일 목적의 형태소 분석기보다 많으며, 표준화된 인터페이스를 통해서 모든 응용프로그램들이 사용할 수 있도록 하였다.

3.1 태그세트, 사전 및 결합규칙

형태소 분석기가 일반적으로 사용하는 지식(knowledge)은 사전(dictionary) 정보와 단어들 간의 결합정보라고 할 수 있으며, 사전의 형태 및 특성에 따라서 각각의 형태소 분석기를 특징지을 수 있다.

태그세트: 본 시스템이 사용하는 태그세트는 명사(NOUN), 대명사(PRON), 부사(ADV), 동사(VERB), 형용사(ADJ), 감탄사(IJ), 어미(FUNCW), 조사(POSP)이다. 본 형태소 분석기에서는 단순 품사 외에도 세부정보를 사용함으로써 좀 더 세분화된 품사 정보를 나타낼 수 있도록 하였다.

< 표 1: 태그세트 및 세부정보 >

품사	세부정보	설명
명사 (76,406개)		보통명사
	C9	의존명사
	Mezhr	단위성 명사
	Num	수사
	VN	동작/상태성 명사
	PrprN	고유명사
	Suffix	접미사
대명사 (145개)		대명사
부사 (8,695개)		일반부사
	SentAdv	문장접속부사
동사 (25,169개)		동사
형용사 (5,914개)		형용사
	ADNOM	관형사
감탄사(315개)		감탄사
어미(6,848개)		어미
조사(2,552개)		조사

본 형태소 분석기는 <표 1>에서 보이는 세부정보 외에도 각각의 목적에 따라 40여 개의 세부정보를 더 사용하고 있다.

사전: 본 형태소 분석기가 기반으로 하고 있는 사전은 조재수님[5]의 국어사전이다. 이 사전은 약 13만 단어를 포함하고 있으며, 형태소 분석기는 이 중 고어, 비표준어 등을 제외한 약 11만 단어만을 선택적으로 사용하고 있다. 본 시스템이 사용하는 사전 표제어의 특징을 살펴보면, (1) '~하다', '~되다'가 결합된 형태를 모두 사전의 표제어로 등록하고 있으며, (2) '그게'와 같은 준말의 형태도 표제어로 사용하고 있다. (3) 또한, '비공개'나 '공개적'과 같은 한자형 접두사, 접미사들이 결합된 단어들을 표제어로 등록하였으며, (4) 어미와 조사의 경우, 복합어미, 복합조사의 형태에 대한 대표형을 모두 사전에 등록하였다.

<표 2>는 형태소 분석기가 사용하는 사전 정보를 보여주고 있다. 명사 '가공'은 세부정보로서 인성(Humn)과 동작성명사(VN)의 정보를 갖고 있으며, 동사 '가결하다'는 여-불규칙 동사로서 목적어를 가질 수 있다(T1)는 세부정보를 갖는다. 사전에 포함된 대부분의 세부정보는 구문분석 단계에서 사용된다.

< 표 2: 일반단어의 사전 정보 예 >

Entry	품사	불규칙정보	세부정보
가공	Noun	NOUN-REG	Humn, VN
가결하	Verb	VERB-여	T1

<표 3>은 어미와 조사에 대한 사전 정보의 예이다. 어미 '쓰음'은 '었음', '왔음', '했음', '았음' 등의 대표형으로서, 왼쪽(Left Category)에는 동

사, 형용사, 서술격 조사 등이 결합될 수 있으며, 오른쪽(Right Category)에는 서술격 조사가 결합될 수 있다. 활용정보는 어미 '쓰음'이 각각의 불규칙 용언에서 어떠한 형태로 변형되는지를 알려주는 정보이고, 세부정보는 '쓰음'이 명사형 어미(Gnd), 과거시제(Past), 그리고 '쓰'(FEss)과 '음'(FUm)의 결합으로 구성되어 있음을 알려준다. 이와 같은 어미나 조사의 세부정보는 대부분이 형태소 분석의 다음 단계인 구문분석 단계에서 사용된다.

< 표 3: 예 - 어미 '쓰음'의 사전 정보 >

Entry	쓰음
Category	FUNCW
Left Category	동사, 형용사, 서술격조사
Right Category	서술격조사
활용정보	InflectionPattern5
세부정보	Gnd, Past, FEss, FUm

결합규칙: 본 형태소 분석기에서 사용되는 결합규칙은 한 어절 내의 가능한 품사의 나열로서, < 표 4>에서 그 예를 볼 수 있다. 사용되는 응용분야에 따라 분석된 결과의 형태나 분석의 우선순위가 다를 수 있으므로 각각의 결합규칙은 그 규칙이 적용되어야 하는 응용분야를 명시하여 기술된다. 첫 번째 결합규칙은 HHC(한글/한자 변환기)와 명사추출에서 사용되며, 곡용하지 않은 명사(-Infld)와 '없다', '같다', '스럽다'와 같은 형용사가 결합한 어절을 처리할 수 있다. 두 번째 규칙은 일반적인 형태소 분석에서 사용되는 규칙으로 복합명사를 처리할 수 있다. 마지막 규칙은 맞춤법 검사기에서 "할수 있다"와 같은 띄어쓰기 오류가 포함되어 있는 어절을 처리할 수 있다. 각각의 결합규칙에는 우선순위가 주어지며, 작은 값을 가진 규칙이 먼저 적용된다. 본 시스템은 개발자에 의해 손으로 작성된 170여 개의 결합규칙을 갖고 있으며, 이 중 약 50여 개는 띄어쓰기 오류를 가진 어절에 대한 규칙이다.

< 표 4: 결합규칙의 예 >

응용분야	결합규칙	우선순위
HHC, 명사추출	noun{-Infld} + adj{ Lemma=[없 같 스럽]}	0
일반	noun{-Infld -Suffix -Num -C9} + noun{-Suffix -Num -C9}	5
맞춤법검사	noun{-Suffix} + verb	1000

3.2 형태소 분석기의 구현

본 형태소 분석기는 COM으로 구현된 DLL (Dynamic Link Library)이다. 따라서 단독으로 실행되지 못하고 다른 응용프로그램에서 필요로 하는 함수만을 COM으로 구현된 인터페이스를 통해 제공한다. 형태소 분석의 기본 알고리즘은 차트 기반의 CYK방식이다[6].

예를 들면, "감기는"이라는 어절은 <그림3>과 같이 동사, 형용사, 명사 4가지로 분석된다.

```
RESULT1:
감기는(1-8) VERB: Lemma 감기, (는/funcw)
        Bits Infld Adnom FNun Pres
RESULT2:
감기는(1-8) VERB: Lemma 감, (기는/funcw)
        Bits Infld FKi Gnd PUn Topick
RESULT3:
감기는(1-8) ADJ: Lemma 감, (기는/funcw)
        Bits Infld FKi Gnd PUn Topick
RESULT4:
감기는(1-8) NOUN: Lemma 감기, (ㄴ/posp)
        Bits Infld Erg PUn Topick SinoChar
```

< 그림 3: 어절 " 감기는" 의 형태소 분석 결과 >

3.3 명사 추출기의 구현

명사 추출기는 형태소 분석기의 결과를 받아

서 명사로 판정된 단어만을 결과로 출력한다. 명사가 포함된 어절의 유형은 여러 가지가 있고, 이 중 연속된 명사의 결합은 구문 분석기에서 과분석되는 것을 막기 위해서 최대 6개까지만 분석하도록 제한하였다. 이러한 어절의 유형은 결합규칙으로 기술되어 있으며, 명사 추출기에서 사용된 결합규칙은 모두 91개이다.

그 외에도 명사 추출기를 구현할 때 고려된 것은 미등록어에 대해서 명사를 추측해 주는 기능이다. 본 명사 추출기에서는 형태소 분석에서 사용되는 조사와 어미 사전을 이용해서 기능어 부분을 제거하고 간단한 접미사 처리를 통하여 미등록어에 대해서도 명사를 추출할 수 있도록 한다. 또한, 불용어사전(stop word list)을 이용하여 의존명사나 수사 등 불필요하게 추출된 명사를 제거한다.

<표 5>는 각종 명사들의 추출 결과를 보여준다. 미등록어에 대해 정확히 추정하며 한글과 한자 또는 영어와 한글이 결합된 입력에서도 각각의 명사를 추정해 낸다.

< 표 5: 명사 추출 예 >

원어절	명사 추출 결과
강조한다	강조
재옥아	재옥
데	(없음)
박회장님은	박회장
KBS에서	KBS
장두형(長頭形)	장두형, 長頭形

4. 결론

본 논문에서는 여러 응용프로그램들이 형태소 분석기를 공유하면서 수행할 수 있도록 COM을 기반으로 한 형태소 분석기를 제안하였다. 또한, 이를 응용한 명사 추출기를 구현하고 기술하였다. 본 논문에서 다루고 있는 형태소 분석기는 여러 응용프로그램이 사용할 수 있도록 고안되었기 때문에, 다소 비효율적이거나 지식의 중복적인 표현이 있을 수 있다. 그러나, 통합환경에서 여러 응용프로그램들이 형태소 분석기를 필요로 할 경우, COM 인터페이스를 통해서 이를 쉽게 공유하게 함으로써 형태소 분석기 및 사전 정보들에 대한 유지 및 보수 작업을 효율적으로 운용할 수 있으며, 전체 시스템의 효율성도 높일 수 있다.

본 명사 추출기는 형태소 분석기로부터 분석 중인 어절의 결과만을 받아서 처리하기 때문에 중의성을 해결하지 못한 채 명사를 추출한다. 또한 원어절에 오류가 있을 경우 명사 추출기에는 오류를 허용하는 규칙을 적용하지 않았기 때문에 미등록어로 잘못 인식한다.

향후 과제로는 동일한 형태소 분석기를 공유하는 한글/한자 변환기, 한글 맞춤법 검사기, 한글 구문 분석기의 구현이 있다.

참고 문헌

[1] http://aladin.etri.re.kr/~nlu/STANDARD /matec99.html

[2] Mastering COM Development, Microsoft Press, 1997.

[3] Dale Rogerson, Inside COM, Microsoft Press, 1997.

[4] Understanding ActiveX and OLE, Microsoft Press, 1996.

[5] 조재수, 한국어 사전, The Korean Monolingual Dictionary (c) 1996-1999, Microsoft Corporation.

[6] 김성용, 최기선, 김길창, "Tabular Parsing 방법과 접속정보를 이용한 한국어 형태소 분석기," 한국정보과학회 춘계 인공지능 발표 논문집, pp.133-147, 1987.