미리내 검색시스템의 명사추출 시스템

김영관, *권혁철
부산대학교 전자계산학과, *부산대학교 정보·컴퓨터 공학부

Noun Extraction System in Information retrieval System of "Mirine"

Young-kwan Kim, Hyuk-Chul Kwon
Department of Computer Science, Pusan National University
*Division of Information and Computer Engineering, Pusan National University

요 약

이 논문은 한국어 정보검색 시스템 "미리내"의 내부 모듈인 색인어 추출 시스템의 성능 평가에 관한 내용이다. 성능 평가를 위해서 99년 ETRI에서 실시한 "형태소분석기 및 태거 비교 분석대회(MATEC99)"의 시험어절을 사용하였다. 정보검색 시스템 "미리내"는 한국어 정보검색을 위해 부산대학교에서 개발한 시스템이다. 한국어 형태소분석 및 태거 대회(MATEC99)를 위해 미리내 검색엔진의 색인어 추출 모듈을 일부 수정하여 명사를 추출하였다.

명사추출기이든 형태소분석기이든 응용프로그램의 특성에 맞춰져서 동작한다. 정보검색의 하위 모듈인 색인어 추출 시스템은 정보검색을 위해 변형된 결과를 출력하므로 성능 비교를 위해 일부 모듈의 수정이 불가피하였다.

ETRI에서 실시한 MATEC99는 지금까지 객관적인 평가 기준이 없었던 한국어 형태소분석기, 태거, 명사추출기의 표준화에 중요한 역할을 하였다.

1. 서 론

개인 홈페이지의 수가 기하급수적으로 증가하면서 인터넷의 정보가 급증하고 있다. 인터넷의 정보의 증가로 정보검색엔진에 대한 의존도도 높아지고 있다.

사용자의 질의어에 정보검색시스템이 많은 문서를 결과로 찾아 주면서 결과의 양보다는 질을 우선하게 되었다.

정보검색시스템의 질은 사용자가 원하는 문서를 보다 높은 순위에 보여주는 것이다. 질을 높인다는 것은 검색 시스템의 recall보다 precision을 더 가중치를 두는 것이다. Precision을 높이려면 무엇보다 정확한 색인어의 추출이 되어야한다. 색인어를 정확하게 추출하기 위해서는 형태소분석기, 미등록어 추정 시스템, 중의성 제거시스템 모두의 성능향상이 필요하다.

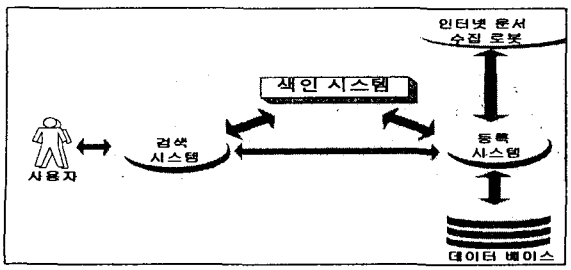
형태소 분석 시스템, 중의성 제거시스템(태거) 등의 한국어 정보처리 분야는 아직 표준화가 미흡한 단계이다. 형태소 분석을 위한 태그셋까지 개발자마다 다르게 사용하는 실정이다. 또한 분석의 가장 기본이 되는 사전의 표제어에 대한 표준안도 현재 만들고 있는 단계이어서 개발자마다 서로 다른 사전 표제어를 사용하여 각자의 시스템을 개발하고 있다. 표준화 작업이 미흡하여 각자의 시스템 성능은 자체 평가만 가능할 뿐 객관적인 기준에 맞춘 비교 분석이 어려운 실정이다. 따라서 사전 표제어, 테그셋 등의 표준화를 위한 노력이 절실히 요구된다.

이러한 측면에서 ETRI에서 실시한 MATEC99는 큰 의미를 가진다.

2. 시스템 구성

1. 정보 검색시스템 '미리내'의 시스템 구성

색인시스템은 독립된 서버이지만, 정보 검색시스템의 내부 모듈이다 정보 검색시스템 "미리내"의 구성은 [그림1]과 같다.



[그림 1] 정보 검색시스템 "미리내"의 전체 구조

정보 검색시스템 "미리내"는 4개의 큰 모듈로 구성된다. 인터넷의 문서를 수집하는 로봇, 수집한 문서를 분석하는 색인시스템, 색인된 결과를 저장하는 등록시스템, 사용자의 질의어를 검색하고 순위를 부여하는 검색시스템으로 구성된다.

인터넷 로봇 모듈은 인터넷의 방대한 문서를 수집하는 일을 한다. 인터넷의 문서는 날마다 변화한다. 새롭게 생성되는 것은 물론이고 내용이 갱신되기도 하며 없어지기도 한다. 로봇 모듈은 문서를 수집한 후 이러한 변화를 데이터 베이스에 반영하는 기술까지를 포함한다.

색인 시스템은 로봇이 수집한 문서를 분석하여 색인어를 추출하고 한 문서 안에서 색인어의 출현 빈도를 반영한 분석결과를 등록시스템에게 넘겨주는 역할을 한다[1]. 또한 검색 시스템에서 받아들인 사용자의 자연언어 질의어를 분석하여 명사를 추출하고 그 결과를 검색 시스템에게 되돌려 주는 역할도 한다.

등록시스템은 색인 시스템의 분석결과를 가지고 데이터베이스를 구축하는 역할을 한다. 빠른 검색과 적은 메모리를 사용하기 위해서 역파일구조로 압축하여 저장한다[3][5].

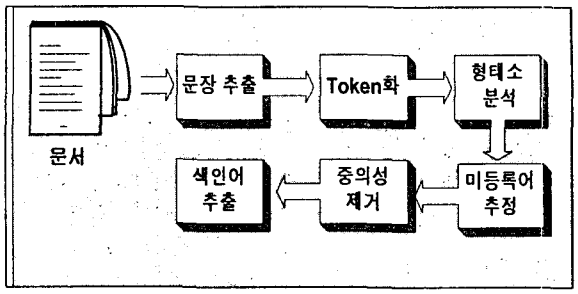
검색시스템은 사용자의 질의어에 대한 데이터베이스 검색결과를 가지고 순위를 부여한다. 즉, 많은 검색결과를 유사도가 높은 순서로 정렬하여 사용자가 보기 편하게 하는 일을 한다[4].

2."미리내"의 색인시스템 구성

a. 색인시스템의 구조

색인시스템의 기본 기능은 문서에서 색인어로서 가치가 있는 단어를 추출하는 것이다. 색인시스템에서 추출하는 색인어를 사용하여 정보검색 시스템에서 검색을 수행하므로 정보 검색 시스템의 성능과 밀접한 관계가 있다. 즉 색인어로 추출되지 않은 단어는 검색할 수 없으며, 잘못된 색인어는 관련없는 문서를 검색하는 결과를 낳는다. 따라서 정확한 색인어의 추출은 검색엔진과는 독립적이면서도 성능에는 중요한 역할을 한다.

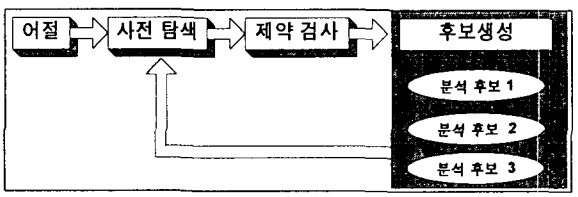
검색시스템 "미리내"에서 사용하는 색인시스템의 전체 구조는 [그림2]과 같다.



[그림 2] 색인시스템의 구성

b. 형태소 분석 시스템

형태소 분석 시스템에서는 어절을 앞에서부터 분석을 한다[1]. 우리말은 용언의 활용이 다양해서 분석이 까다롭다. 형태소 분석기에서는 기본적으로 사전의 단어와 제약조건 검사로 분석을 수행한다.



[그림 3] 형태소 분석 과정

미리내 검색 시스템의 형태소 분석기에서 사용하는 사전은 [표 1]과 같다.

사전 종류	사전의 단어 수	사용하는 메모리
기본사전	88,255개 (명사,동사,형용사 등)	1MB
조사사전	1,926개	19,442 bytes
어미사전	3,315개	34,046 bytes
수사사전	784개	7,910 bytes
보조용언	50개	750 bytes

[표 1] 형태소 분석 모듈에서 사용하는 사전

c. 미등록어 추정 시스템

형태소 분석기에서는 사전에 들어 있는 단어만 가지고 분석을 수행하기 때문에 사전에 들어 있지 않은 지역 이름, 신조어, 사람 이름 등이 포함되어 있는 어절은 분석을 하지 못한다. 이러한 단점으로 보완해 주는 역할을 미등록어 추정 시스템에서 한다[2]. 미등록어 추정시스템은 형태소 분석기의 분석결과가

없거나 오류를 포함한 분석결과일 때만 수행된다.

형태소 분석 시스템에서는 한 어절을 앞쪽에서 뒤쪽으로 분석하지만 미등록어 추정시스템은 뒤에서 앞으로 분석을 한다. 어절의 뒤쪽에서부터 조사, 어미 등을 제거하고 미등록어를 찾아낸다. 미등록어 추정과정에서도 사전을 사용한다. 사용되는 사전에는 역명사사전, 역어미사전, 역조사사전 등이다.

미등록어 추정에서는 사용하는 역사전은 음절의 순서를 거꾸로 해놓은 사전이다. 이것은 미등록어 추정과정이 어절의 뒤쪽 음절에서 앞쪽으로 분석하기 때문이다.

d. 중의성 제거 시스템 (태깅 시스템)

형태소 분석시스템과 미등록어 추정시스템에서 생성한 분석결과가 두 개 이상일 때 좌우 어절(3개)의 분석결과를 참조하여 하나의 분석 후보를 선택한다[1]. 좌우 어절의 분석결과에서 품사정보,

e. 색인어 추출 시스템

색인어 추출 시스템은 문장의 분석결과를 바탕으로 색인어로서 가치가 있는 단어를 추출하는 모듈이다[1]. 분석과정에서는 명사와 명사가 결합된 복합명사 어절은 분리하여 분석한다. 따라서 색인어를 추출할 때는 이것을 감안하여 단일 명사와 복합명사 모두를 추출한다. 즉, "학교 + 생활"이라는 분석결과에서 "학교", "생활", "학교생활"을 색인어로 추출한다. 또한 지나치게 사용 빈도가 높은 단어(불용어)는 색인어로서 가치가 없으므로 제거한다.

한 문서에서 사용되는 단어의 빈도도 문서를 대표하는 색인어의 가치를 판단하는 요소이므로 색인어와 단어의 빈도를 같이 추출한다.

3. 자체 성능 평가

시험용 말뭉치와 "미리내"시스템의 색인어 추출 시스템의 입력 형식의 차이로 많은 오류가 발생했다.

MATEC99의 시험용 말뭉치에는 어절의 구분이 new line으로 되어 있었다. 이 형태의 입력은 색인기 시스템의 태깅과정에서 많은 오류를 유발 시켰다. 태깅과정에서는 문장 부호, 공백, new line 등의 delimiter들이 중요한 역할을 한다. 미리내 시스템에서는 어절사이의 delimiter로 공백을 사용한다. 공백이 아닌 new line 기호는 올바른 분석 후보를 제거하는 오류를 유발 시켰다.

오류의 예를 보면, new line으로 시작하는 의존 명사는 분석 후보에서 제거하게 되어있다. 이것은 모든 의존명사로 시작하는 분석 후보를 제거하였으며, 따라서 의존명사들이 분석 결과로 나왔다.

new line은 동사의 분석 후보 태깅과정에도 영향을 주었다.

이러한 오류를 수정한 후 자체 평가한 결과는 recall과 precision 모두 96% 이상이었다.

4. MATEC99와 표준안

MATEC99는 한국어 형태소 분석기, 태거, 명사추출기의 개발을 위한 표준화 작업에 선구적 역할을 하였다. 하지만 아쉬운 점은 MATEC99을 위한 사전 표제어 선정기준이 명확하지 않아 각각의 팀이 서로 다른 기준으로 결과를 제출한 것이다. 또한 선정기준을 알더라도 이미 가지고 있는 사전을 선정기준에 맞추어 다시 만드는 작업이 쉬운 일이 아니다. 표제어 선정 기준에 맞는 사전이 함께 배포하였으면 하는 아쉬움이 남는다.

[참고 문헌]

[1] 김민정, "규칙과 말뭉치를 이용한 한국어 형태소 분석과 중의성 제거", 박사학위 논문 , 1997

[2] 양장모, "언어 정보를 이용한 한국어 미등록어 추정시스템의 구현", 석사학위 논문 , 1997

[3] 박 승, "실시간 정보검색 시스템 환경의 구현", 석사학위 논문 , 1998

[4] 강상배, "한국어 문서의 통계적 정보를 이용한 문서 요약 시스템 구현", 석사학위 논문 , 1998

[5] 이준영, "다중색인과 압축저자에 의한 정보검색 시스템 개발에 관한 연구", 석사학위 논문 , 1997