웹기반 말뭉치 정보 검색 시스템

이정호*, 임희석**
*한신대학교 소프트웨어학과
**고려대학교 컴퓨터교육과
e-mail: ezis@hanmail.net

Web-based Corpus Information Retrieval System

Jeong-Ho Lee*, Heui-Seok Lim**
*Hanshin University, **Korea University

요 약

본 논문은 대용량의 한글어 말뭉치를 이용하여 언어학적 통계정보를 자동으로 검색할 수 있는 웹기반 언어정보 검색 시스템을 제안하고 구현하였다. 구현한 시스템을 통해 형태소, 품사, 어절 정보를 자동으로 획득할 수 있었다. 본 시스템은 언어학적 지식이 부족한 비전문가도 말뭉치 검색을 효율적으로 수행할 수 있으며, 웹기반으로 구현되었기 때문에 시스템 접근의 용의성에 의의가 있다.

1. 서론

표본선정이 잘 된 말뭉치는 언어를 연구하는데 매우 중요하다. 말뭉치는 그 크기가 유한하고 자연언어의 모든 예를 나타내기 어렵다는 한계에도 불구하고 단어 의미나 문법 구조 연구에 믿을 만한 정보를 제공한다.

이러한 말뭉치는 언어학, 언어 심리학, 자연어처리 등의 연구에서 통계 정보를 추출해서 실험데이터 구축에 많이 사용된다. 실험에서 믿을만한 데이터를 얻기 위해서는 가능한 큰 크기의 말뭉치에서 통계정보를 추출해야 오차를 줄이고 신뢰성이 높아진다. 하지만 말뭉치의 크기가 커질수록 실험 데이터를 위한 통계 정보를 추출하기 위한 노력과 시간이 배가 된다.

그리고 오늘날까지는 말뭉치에서 통계정보를 얻기 위해서는 사람이 직접 분류하거나 Excel과 같은 툴을 사용하는 수동적인 방법으로 추출했는데 이러한 방법은 인력과 시간이 많이 걸리고 Excel또한 효과적으로 사용하기 위해선 말뭉치를 구축하는 비용 외로 Excel 파일을 구축하는 비용이 들고 노력에 비해 효과가 미비해서 불편했다. 그래서 컴퓨터 비전문가도 언제 어디서나 한글 말뭉치에서 효율적으로 통계 정보를 검색 할 수 있게 웹을 기반으로 한 한글 말뭉치 검색 시스템을 개발하고자 한다.

■ 형태소 검색

형태소는 '의미의 기능을 부여하는, 언어의 형태론적 수준에서의 최소단위'로 한국어 처리, 분석에 있어서 가장 중요한 부분이다. 형태소의 통계정보를 통해서 고빈도 단어를 추출할 수 있다.

■ 품사 검색

품사란 '단어를 기능, 형태, 의미에 따라 나눈 갈래'로 본 시스템에서는 '[부록] 태깅 정보 분류표'와 같이 나눈 품사를 검색할 수 있다.

■ 어절 검색

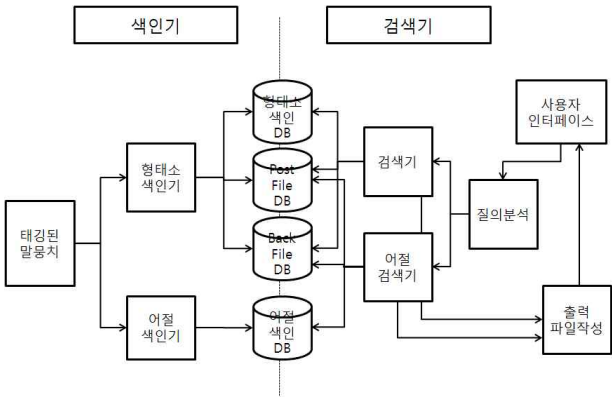
어절이란 '문장을 구성하고 있는 각각의 마디'이다. 문장 성분의 최소 단위로서 띄어쓰기의 단위가 된다.

2. 웹기반 말뭉치 정보 검색 시스템

웹기반 말뭉치 정보 검색 시스템이란 심리학자나 언어학자와 같은 컴퓨터를 잘 활용하지 못하는 비전

문가도 웹을 기반으로 제작된 말뭉치 (통계) 정보 검색 시스템이다.

본 논문에서 제안하는 시스템의 구조는 <그림 1>과 같다. 색인기의 구조는 태깅된 말뭉치를 형태소 색인기를 통해 형태소 색인 DB, Post File DB, Back File DB를 구성하고 어절 색인기를 통해 어절 DB를 구성한다.



<그림 1> 시스템 구조

그리고 검색기는 HTML과 구성된 사용자 인터페이스에서 사용자의 요구를 받아들여 질의를 분석해서 검색 파일이 있는 경우 파일에서 검색어를 추출하고, 검색 질의에 따른 각각의 형태소, 품사, 어절, 용례, 자소, 이웃단어 검색기로 보낸다. 그리고 각각의 검색기에서는 색인 DB에서 검색해서 결과를 출력하는데, 출력하기 전에 품사 오류 제거를 하고 상세 검색의 경우 중복되는 문장을 제거하고 사용자가 파일 저장을 원한 경우에는 사용자가 다운로드 할 수 있는 파일을 작성한다.

2.1. 색인 시스템

비단 말뭉치 검색에 제한하지 않더라도 검색 작업의 소요 시간을 향상시키기 위해서는 색인 파일의 구축이 필요하다. 색인 파일은 구축하는데 자원과 시간이 많이 소모 되지만 한번 구축된 색인 파일을 이용하면 검색 속도의 비약적인 향상을 가져 올 수 있다.

2.1.1. 자소 분리

자소 검색을 하기 위해서는 한글을 자소 단위로 분리해야 한다. 본 논문에서 자소 분리에 사용한 방법은 문자열을 읽어서 한글일 경우 2 Byte 씩 끊어 완성형 한글(KSC5601 또는 KS X 1001:2004)을 한글 상용 조합형(KSSM)으로 변환한 뒤, 초성, 중성, 종성을 분리해서 다시 완성형으로 바꾸어준다. 이 과정에서 종성이 없는 경우에는 다음 음절의 초성과 구분할 수 있고 검색의 편의를 위해서 '-'기호를 넣어주었다. 따라서 '시스템'이라는 단어를 자소 분리하면 'ㅅㅣ-ㅅㅡㅡㅌㅖㅁ'이 된다.

2.1.2. 역화일

코퍼스를 색인 하는 방법에 있어서 정보 검색에서 많이 사용되는 역화일(Inverted file)을 사용했다. 역화일(또는 역색인) 기법은 대표적인 사전 인덱스 방법의 하나로 파일이나 데이터베이스에서 레코드를 빨리 검색하기 위해 별도의 색인 파일을 만드는 것을 말한다. 이때 색인 파일에는 검색의 기준이 되는 키 필드의 값과 그 키 값을 가지는 레코드에 대한 포인터들이 저장된다.

본 시스템에서 첫 번째 역화일(post file database)에서는 품사, 형태소의 출연 빈도 값, 두 번째 역화일의 포인터 값, 다음 포인터(next pointer)를 저장하고 두 번째 역화일(back file database)에서는 원본 코퍼스 파일의 포인터 값과 다음 포인터 값을 저장한다. 이렇게 역화일을 두개로 나눈 이유는 같은 형태소에 품사가 다른 경우(예를 들어 형태소 '배'의 경우 품사가 NA, NNB, VA, NNP, VV, NNG 6개이다.)에도 각각의 품사별 결과(예제)를 빠르게 출력하기 위해서이다.

2.2. 검색 시스템

검색 시스템은 리눅스 환경에서 구성하였다. 검색 시스템은 색인된 코퍼스, 색인 DB, 어절 DB, Post File DB, Back File DB와 검색시스템(CGI), 웹페이지(HTML, PHP)로 구성된다.

웹페이지에서 사용자가 질의를 입력하면 검색시스템에서 질의를 분석해서 형태소, 품사, 어절, 용례, 자소, 이웃단어 검색기로 보낸다. 각각의 검색기는 색인 DB에서 질의어를 검색해서 출력기에서 자동색인 과정에서 발생할 수 있는 품사의 오류 제거, 복수 문장 제거, 사용자가 결과를 Excel 파일로 원하는 경우 파일 작성 등을 한 후 출력한다.

3. 실험 및 결과

본 논문에서는 문화관광부가 국립국어원 및 관련 학계와 더불어 지난 1998년부터 2007년까지 추진해 온 국어 정보화 사업인 21세기 세종 계획의 부산물인 1,000만어절의 세종코퍼스를 태깅하여 만든 코퍼스를 가지고 색인하여 검색 가능하게 했다.

1,000만 어절을 형태소 색인한 결과 193,667개의 유일한 형태소가 나와서 색인어로 색인 하였고 그 중 출현 빈도수가 10,000회 이상 되는 고빈도 형태소는 '한국' 외 186개였다. 그리고 형태소 중 일반명사가 85,536개로 가장 많은 수를 차지했다.

3.1. 형태소 검색 결과

본 시스템에서는 형태소를 검색해서 그림과 같이 품사와 형태소의 사용 빈도수, 사용 예, 같이 사용되는 형태소 정보를 알 수 있다. 동음이의어인 '배'를 검색 해본 결과 검색 결과는 아래 <그림 2>과 같다.

검색 단어	배			
입력빈도범위	0~1000000			
1 번째 단어	배			
구성 자소	ㅂㅐ-			
	1번째 품사	NA: 분석불능범주	빈도수	1
	2번째 품사	NNB: 의존명사	빈도수	6
	3번째 품사	VA: 형용사	빈도수	9
	4번째 품사	NNP: 고유명사	빈도수	72
	5번째 품사	VV: 동사	빈도수	420
	6번째 품사	NNG: 일반명사	빈도수	5859
	총 출현 빈도수 6367			

<그림 2> 형태소 검색 결과

3.2. 품사 검색 결과

품사 중 가장 많은 수를 차지하는 NNG(일반명사)를 검색 한 결과 <그림 3>와 같이 형태소, 사용 빈도수등이 가나다 순으로 정렬되는 것을 볼 수 있다. 상세 검색 시에는 각 형태소의 사용 예를 알 수 있다.

검색 품사	NNG			
입력빈도범위	1000~1000000			
	1 번째 단어	가격	빈도수	1769
	2 번째 단어	가난	빈도수	1215
	3 번째 단어	가능	빈도수	1054
	4 번째 단어	가슴	빈도수	4044
	5 번째 단어	가운데	빈도수	5655
	6 번째 단어	가을	빈도수	1448
	7 번째 단어	가정	빈도수	2498

<그림 3> 품사 검색 결과

3.3. 어절 검색 결과

어절 검색을 통해서 아래의 <그림 5>와 같이 어절을 구성하는 형태소, 태거, 사용 빈도를 알 수 있다.

검색 단어	한국
입력빈도범위	0~1000000
1 번째 어절	"…한국은
빈도수	1
	"/SS+한국/NNP+은/JX
2 번째 어절	"대다수의(한국)
빈도수	2
	"/SS+대다수/NNG+의/JKG+(/SS+한국/NNP+)/SS
3 번째 어절	"신한국
빈도수	3
	"/SS+신/XPN+한국/NNP
4 번째 어절	"이야기-한국체육사"시리즈7번째인
빈도수	2
	"/SS

<그림 4> 어절 검색 결과

3.4. 자소 검색 결과

자소 검색은 형태소의 구성 자소를 검색해서 검색 자소와 일치하는 형태소, 품사, 사용 빈도, 사용 예를 알 수 있다.

자소 검색은 다시 '자소 검색<그림 5)>'과 '자소 포함 검색<그림 6>' 두 가지 기능으로 나눴다. 그 이유는 '자소 검색'은 질의어와 일치하는 자소 또는 규칙(기호 '?'은 한 음절, '-'는 한 자소를 대신한다.)에 맞는 자소를 출력하고 자소 포함 검색은 질의어를 포함하고 있는 자소를 전부 출력한다.

검색 자소	ㅇ--?ㄷㅗㅇ?			
입력빈도범위	0~1000000			
1 번째 단어	아랫동네			
구성 자소	ㅇㅏ-ㄹㅐㅅㄷㅗㅇㄴㅔ-			
품사	NNG: 일반명사		빈도수	8
2 번째 단어	아랫동서			
구성 자소	ㅇㅏ-ㄹㅐㅅㄷㅗㅇㅅㅓ-			
품사	NNG: 일반명사		빈도수	1
3 번째 단어	아랫동이			
구성 자소	ㅇㅏ-ㄹㅐㅅㄷㅗㅇㅇㅣ-			
품사	NNG: 일반명사		빈도수	1
4 번째 단어	아침동자			
구성 자소	ㅇㅏ-ㅊㅣㅁㄷㅗㅇㅈㅏ-			
품사	NNG: 일반명사		빈도수	3
5 번째 단어	안서동위			
구성 자소	ㅇㅏㄴㅅㅓ-ㄷㅗㅇㅇㅟ-			
품사	NNG: 일반명사		빈도수	1

<그림 5> 자소 검색 결과

검색 자소	ㅈㅣ-ㄱㅏㅂ			
입력빈도범위	0~1000000			
1 번째 단어	담뱃지갑			
구성 자소	ㄷㅏㅁㅂㅐㅅㅈㅣ-ㄱㅏㅂ			
품사	NNG: 일반명사		빈도수	1
2 번째 단어	돈지갑			
구성 자소	ㄷㅗㄴㅈㅣ-ㄱㅏㅂ			
품사	NNG: 일반명사		빈도수	12
3 번째 단어	반지갑			
구성 자소	ㅂㅏㄴㅈㅣ-ㄱㅏㅂ			
품사	NNG: 일반명사		빈도수	1
4 번째 단어	손지갑			
구성 자소	ㅅㅗㄴㅈㅣ-ㄱㅏㅂ			
품사	NNG: 일반명사		빈도수	9
5 번째 단어	지갑			
구성 자소	ㅈㅣ-ㄱㅏㅂ			
품사	NNG: 일반명사		빈도수	207

<그림 6> 자소 포함 검색 결과

4. 결론 및 향후 연구 과제

정보 검색에서 이미 널리 사용되고 그 효과가 검증된 역화일을 이용해서 색인하고 검색하는 말뭉치 검색 시스템은 색인 된 말뭉치 1,000만개의 어절을 통한 실험 결과 만족할 만한 결과를 보여주었다. 하지만 웹으로 서비스하는 시스템을 개발하다보니 향후 연구되어야 할 부분들이 있었다.

먼저, 색인 과정의 복잡으로 인해서 색인 시간이 오래 걸리고 그로 인해 이미 색인되어 있는 말뭉치가 아닌 새로운 말뭉치에서의 검색을 어렵게 한다. 따라서 앞으로 색인 속도를 높여서 이미 기존에 색인되어 있는 데이터베이스뿐만 아니라 사용자가 자신의 개인 말뭉치를 색인해서 검색 할 수 있게 하는 방법에 대한 연구가 필요하다.

그리고, 말뭉치의 특성상 상세 검색을 할 경우 많은 예제가 출력되는데 웹에서는 출력 결과의 양이 많은 경우 한 화면에 출력하기 위해서 속도가 눈에 띄게 저하되는 현상을 볼 수 있었다. 본 논문에서는 이를 해결하기 위해서 Table태그를 이용해서 출력하던 것을 Div태그로 구성한 결과 속도의 향상을 볼 수 있었다. 그리고 형태소를 결과가 많은 품사 검색 등을 출력할 때 빈도 범위를 상향 조정(무제한->1000번 이상)했고 문장 검색 시 출력되는 문장의 개수를 100개로 제한했다. 하지만 향후 출력 결과가 많으면 페이지를 나누는 방법 등의 출력 속도의 향상을 위한 기술적인 연구가 필요하다.

참고문헌

[1] Colin J. Davis, "N-Watch: A program for deriving neighborhood", Behavior Research Methods, 65-70, 2005

[2] 임희석, 박기남, 남기춘, “계산주의적 시각단어재인 모델에서의 시각이웃과 음운이웃 효과”, 한국산학기술학회논문지, 제8권 4호, pp.803~809, 2007.

[3] 박기남, 임희석, 남기춘, “어휘판단 과제 시 보이는 언어현상의 계산주의적 모델 설계 및 구현”, 한국컴퓨터교육학회 논문지, 제 9권 2호, pp.89~100, 2006.