

많이 사용되는 한국어 문형 패턴용 조사기의 설계 및 구현

이상곤^o, 소강춘
전주대학교 컴퓨터공학과, 전주대학교 국어교육과
{samuel, korean}@jj.ac.kr

Design and Implementation of Sentence Frame Analyzer for Korean

Samuel Sangkon Lee^o, Kang-chun So
Dept. of Computer Science & Engineering, Dept. of Korean Education,
Jeonju University

요 약

21C 세종계획 결과물의 태그된 말뭉치를 살펴보면 국어 정보화 측면에서 귀중한 자료가 많이 구축되어 있으나 말뭉치를 효율적으로 이용할 수 있는 저작 도구가 대단히 부족하다. 본 논문에서는 태그 정보의 조사자가 입력하는 태그의 여러 조합을 다양하게 검색하고 잘못 부착된 태그열 오류가 발견되면 즉시 수정하여 자료의 무결성을 보장하고, 한국어에서 많이 출현하는 문형 패턴을 검색할 수 있는 저작 도구를 설계하고 구현하였다.

주제어: 문형 패턴, 세종계획 결과물 활용, 태그 정보 추출

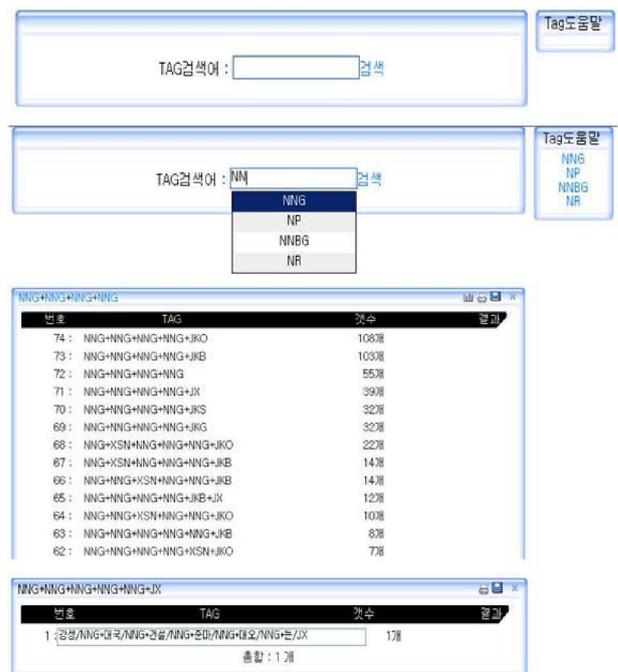
1. 서론

21C 세종 결과물과 같은 대용량의 국어 정보 말뭉치를 조사하여 한국어에서 많이 사용되는 문형 패턴을 발견해 내고, 이 중 많이 사용되는 문형을 추출하여 외국인들에게 우선적으로 가르쳐야 할 것을 선별해 내는 작업이 필요하다. 현재에도 말뭉치로부터 실용적인 격틀(case frame) 정보를 자동 구축하는 연구가 꾸준히 진행되고 있다[1].

본 논문의 시스템 개발을 위해 기존에 개발된 목표 시스템과 비슷한 프로그램에 대해 조사하는 것 보다는 연구자의 목적에 부응하는 시스템의 개발을 목적으로 한다. 여러 명의 국어정보처리학자들의 의견을 청취하고 조언을 얻어 제안된 의견 중 공통된 의견을 토대로 하여 본 논문에서 제안하는 프로그램을 제작하였다. 이 중 주요 의견은 국어 정보에 관한 방대한 자료가 있는데 이를 활용할 수가 없다는 문제점을 해결하고자 본 시스템을 개발하였다. 기존에 있던 파일 시스템을 데이터베이스화시켜 자료의 조회에 대해 융통성 있게 개발하고, 하나의 DB를 가지고 연구자의 다양한 목적에 부응하도록 웹 기반 소프트웨어로 구축하였다.

또한 세종 계획 결과물에는 잘못 부착된 태그열이 다수 존재하는데, 이것을 검출할 수 있는 자동화된 방법이 필요하다는 생각이 들었고, 국어 정보학자들과 토의해

본 결과 125 만 어절의 기준으로 8% 정도의 에러가 발견되어 이번 기회에 이를 수정하였다.



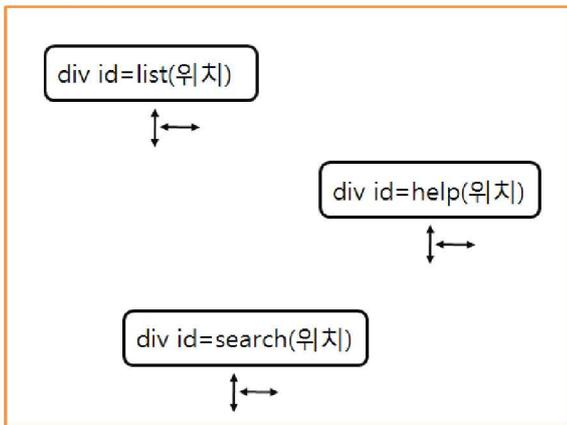
(그림 1) 문형 패턴 조사기의 구성

2. 인터페이스 및 데이터베이스의 설계

국어 정보 처리를 위한 작업자용 인터페이스의 요구사항을 열거하면 다음과 같다.

- 1) 태그열 조합 조회,
- 2) 에러의 발견과 오류 수정,
- 3) 작업 결과의 엑셀 파일 형태로 저장,
- 4) 작업 결과의 인쇄
- 5) 차트 작성 기능 등

메인 화면의 창이동 인터페이스
index.php



(그림 2) 메인 화면의 구성

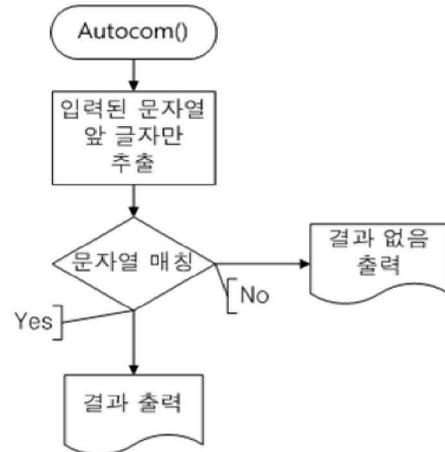
메뉴를 심플하게 배치하여 한 눈에 알아보기 쉽도록 (그림 1)과 같이 윈도우 방식으로 구성하고, 창과 창이 자유롭게 겹칠 수 있도록 하였다. 겹쳐진 창에서는 경우에 따라 맨 앞에 오는 창과 맨 뒤에 가는 창을 유동적으로 조절 가능하다. 새로운 검색 결과에 의해 새 창이 앞으로 나오는 방식으로 구성하였다. 검색은 (그림 1)에서는 태그열을 검색할 수 있고, 사용자가 쉽게 기입할 수 있는 자동 완성 기능과 태그열에 대한 설명을 넣어 본 시스템을 편리하게 이용하고 추천 단어는 키보드를 이용해서 이동하며 선택할 수 있다. 정렬은 그림에서와 같이 검색된 결과를 가지고 문자 또는 번호 순서 또는 개수에



(그림 3) 전체 시스템 구성

따라서 정렬되도록 인터페이스를 구성해야 한다. 저장/인쇄는 모든 검색된 결과가 엑셀로 저장하거나 바로 문서 출력이 가능하고 2차 저작물을 만들 때 편리하도록 지원하며, 검색 결과를 그대로 저장한다.

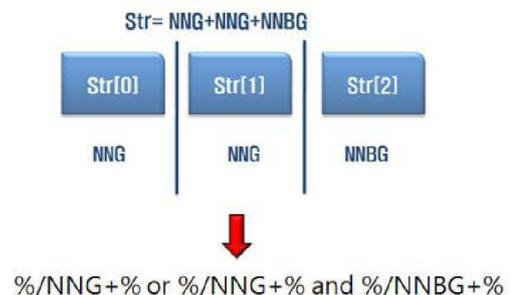
차트 기능을 제공하여 태그 조합열의 각종 통계 정보를 한 눈에 파악이 가능하도록 하였다. 잘못된 태그열이 조회되었을 경우 간편하게 수정이 가능한 인터페이스를 구성하고, 똑같은 오류가 여러 개가 존재하는 경우 한



(그림 4) 자동 완성 기능의 제어 흐름

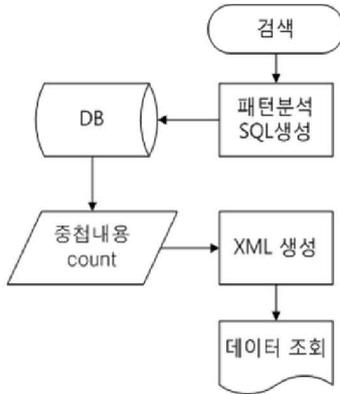
번에 모두 수정할 수 있다.

메인 화면의 구성은 다음과 같다. (그림 2)는 모든 창의 자유로운 움직임을 염두해 두고 설계하였다. 별도의 레이아웃은 없으며, 화면 전체를 div()로 분할하여 화면에 절대값으로 구성하고, div의 위치 정보에 따라서 화면 구성을 배치하였다. (그림 3)에 전체 시스템의 구성도를 나타내었다. (그림 4)에서는 작업자가 자동 완성 기능을 이용해서 태그열의 조합[2]을 입력하면 고유 문형의 패턴을 분석한 후, 쿼리문을 자동 생성하여 DB 검색을 하고, XML 문서를 생성하여 다시 사용자 인터페이스로 전송해 주는 방식이다.



자동 완성 기능(autocom)은 (그림 4)과 같이 입력된 문자열 중에서 앞 글자만 잘라내 미리 입력된 태그열 중

비슷한 것이 매칭 되면 곧바로 추천 단어로 출력되고, 매칭된 결과가 없다면 출력하지 않는다.



(그림 6) DB 검색의 과정도



(그림 7) DB 검색의 예-1

문형 패턴의 분석(Pattern)은 (그림 4)와 같이 쿼리문을 생성하는데 예를 들어 “Str = NNG + NNG + NNBG”가 검색어가 된다면 “+” 기준으로 각 문자열을 배열로 저장한다. 그 이유는 배열로 구분 하여 패턴 분석된 배열을 이용해 문자열을 검색하기 위해서 위의 그림과 같이 와일드 카드 기호를 삽입하여 SQL 문장을 생성한다.

DB의 검색은 (그림 6)과 같은 절차로 완성된 SQL 문을 통해 중첩 내용을 카운트 하는 방식으로 데이터를 검색해서 XML로 생성하고, 다음 클라이언트에 제공해서 데이터를 조회하는 방식으로 구성하였다.

<표 1> 데이터베이스 이름

데이터베이스 이름	사 용
corpus	말뭉치

데이터베이스의 이름은 다음의 <표 1>과 같으며, 데이



(그림 8) DB 검색의 예-2

블의 구성은 <표 2>와 같이 구성하였다.

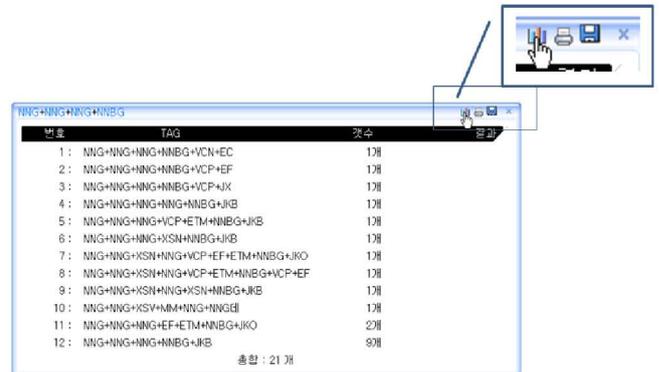
<표 2> 테이블의 구성과 이름

테이블	필드	종류	비고
hap2	number	int(11)	순서 번호 (index)
	spacing one	varchar (255)	원시 말뭉치
	spacing two	varchar (255)	태그 변환 말뭉치

3. 구현 내용 소개 및 기능 설명

(그림 7)에서는 태그 명령어를 입력하여 검색할 수 있으며, 검색 버튼 이용하여 검색이 가능하다. (그림 8)은 내용을 입력하면 글자별로 추천 단어 목록을 보여 주며, 도움말을 제공하여 초보자도 쉽게 내용을 알 수 있도록 낱말 자동 완성 기능을 제공한다.

(그림 8)은 검색창에 검색 태그를 입력하고, 기본적으로 글자를 기준으로 정렬해서 출력된다. (그림 9)는 내용 파악을 쉽게 하기 위해서 한 줄 단위로 줄의 색 변경이 된다. (그림 10)은 연구자가 특정 태그열을 상세하게 조

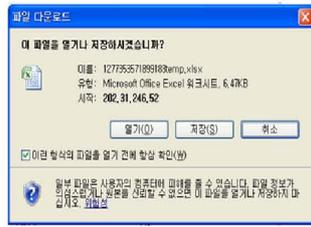


(그림 9) 다양한 출력 기능의 선택

사하고자 할 때 사용하는 메뉴로 특정 태그열을 선택하면 검색된 태그열의 원래 내용이 출력되고 총합과 동일 태그의 내용이 데이터베이스에서 중복되는 것이 몇 건이나 되는지 알 수 있다.

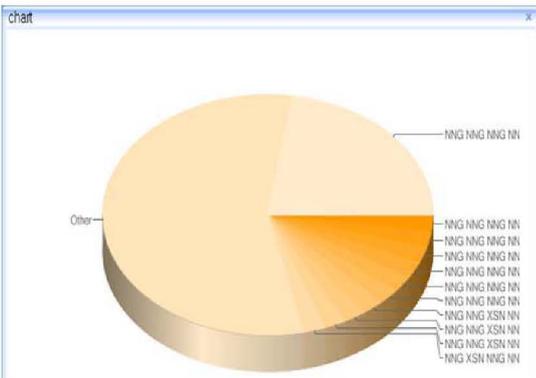


(그림 10)



(그림 11)

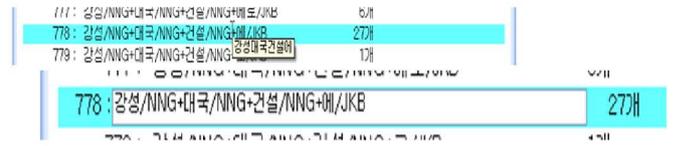
(그림 9)는 개수에 따라서 정렬되거나 태그의 첫 번째 글자에 따라서 정렬되거나 번호에 따라서 정렬되도록 구성하였으며, 한번 클릭할 경우 오름차순, 내림차순을 번갈아 가며 정렬한다. (그림 8)는 오른쪽 상단 디스켓 모양을 선택하면 엑셀 파일로 저장이 되도록 구성되어 연구자의 손쉬운 통계 정보를 만들 수 있다. (그림 10)는 오른쪽 상단 프린터 모양의 그림을 선택하여 인쇄하며 그림과 같이 인쇄 가능하다.



(그림 12) 그래프 출력 예

(그림 12)는 오른쪽 상단 막대그래프 모양을 선택하면 차트가 출력이 된다. (그림 12)는 원형 차트가 출력이 되도록 구현되어 있으며 색깔로 구분 가능하다. 내용을 확인을 하다가 보면 잘못된 내용을 발견하면 오류로 생각되는 부분을 밑에 그림과 같이 마우스의 포인터를 놓고 클릭하면 그림과 같이 변경할 수 있는 텍스트 박스가 나온다. 내용 수정하고 마우스를 아무 장소에다가 클릭을 하면 27 건의 동일한 오류를 한 번에 수정할 수 있다. 에러의 검출에 대해 서술한다. 검색어에 *를 입력할 경우, 전체 말뭉치에서 태그열만 추출하며 잘못된 태그를 발견할 수 있다. 잘못된 것의 개수를 세어 주는 기능도

있다. 에러 검출 기능으로 확인해 보니 전체 태그 중 2,619,084 개가 사용되었고, 그 중 에러로 판정된 것이 178,566개로 전체의 약 6.8%가 잘못된 태그열임을 알 수 있었다.



(그림 13) 원 클릭에 의한 손쉬운 수정

4. 결론

본 논문에서 개발한 시스템의 구현 환경은 Intel(R) Core(TM) 2 Duo CPU E 8400 3.00 GHz, Windows XP 이며, 사용된 프로그래밍 언어는 PHP, Java Script, XML, HTML, MYSQL, Photoshop 이다.

현재에는 125만 어절에서 3천만 어절로 DB를 늘려서 태그를 분석하여 그 결과를 토대로 태그 테이블을 생성하여 저장하고 있다. 결과물 확인을 위해 현재 운용 중인 사이트를 제시(<http://202.31.246.52/kor/index.php>)한다. 향후에는 여기에 일반적인 텍스트화된 말뭉치를 삽입하면 자동적으로 태그가 부착되는 프로그램으로 발전시키고자 한다.

참고 문헌

- [1] 양단희, 송만석, “말뭉치로부터 격틀 구축에 필요한 학습 데이터 추출“, 제10회 한글 및 한국어 정보처리, pp. 287-292, 1998.
- [2] 이윤진, 한국어 문형 표현 100, 건국대학교출판부, 2판, 2005.
- [3] 송유석, 이상근, 이인홍, “한국어 문형 패턴 조사기의 설계 및 구현“, 제33회 한국정보처리학회 춘계 학술 발표 논문집, 제17권, 제1호, pp. 409-412, 2010.

[부록 A] 태그의 분석에 사용한 표

번호	태그명	품사	번호	태그명	품사
01	EC	연결 어미	21	NP	대명사
02	EF	종결 어미	22	NR	수사
03	EP	선어말 어미	23	SE	출입표
04	ETM	관형형 전성어미	24	SF	마침표, 물음표, 느낌표
05	ETN	명사형 전성어미	25	SH	한자
06	IO	감탄사	26	SL	외국어
07	JO	접속 조사	27	SN	숫자
08	JKB	부사격 조사	28	SO	붙임표
09	JKC	보격 조사	29	SP	첨표, 가운데점, 물론, 빗금
10	JKG	관형격 조사	30	SS	따옴표, 괄호표, 줄표
11	JKO	목적격 조사	31	SW	기타 기호
12	JKQ	인용격 조사	32	VA	형용사
13	JKS	주격 조사	33	VCN	부정 지정사
14	JKV	호격 조사	34	VCP	긍정 지정사
15	JX	보조사	35	W	동사
16	MAG	일반 부사	36	XPN	체인 접두사
17	MM	관형사	37	XR	어기
18	NNBG	의존 명사	38	XSA	형용사 파생 접미사
19	NNG	일반 명사	39	XSN	명사 파생 접미사
20	NNP	고유 명사	40	XSV	동사 파생 접미사