세종 문어체 말뭉치를 위한 말뭉치 데이터 추출 도구

박일남, 장우석, 강승식
국민대학교 전자정보통신대학 컴퓨터공학부
e-mail : pin0156@naver.com, rufia00@gmail.com, sskang@kookmin.ac.kr

Corpus Data Extracting Tool for Sejong Text Corpus

Il-Nam Park, Wu-Seok Jang, Seung-Shik Kang
School of Computer Science, College of EECS, Kookmin University

요 약

본 논문에서는 세종 말뭉치 데이터를 활용할 때 한글코드의 변환 및 말뭉치에서 필요한 정보 추출 등 한국어 말뭉치에서 통계 정보를 추출하는데 사용되는 여러 가지 기능들을 한데 묶어, 말뭉치 작업의 사용자 편의성을 개선시키기 위한 도구를 설계, 구현하였다. 이 말뭉치 활용 도구는 세종 말뭉치의 원시, 형태, 형태의미, 구문 말뭉치들을 다양한 옵션에 따라 사용자가 원하는 데이터를 추출할 있을 뿐만 아니라 일반적인 한글 텍스트 파일에 공통적으로 사용되는 코드 변환, 파일 합병, 빈도 계산 등을 제공하기 때문에 말뭉치 작업을 하는 사용자들이 편리하게 사용할 수 있게 하였다.

1. 서론

세종 문어체 말뭉치는 원시, 형태, 형태의미, 구문으로 이루어진 말뭉치들의 집합이며, 현재 원시 말뭉치 6363 만 어절, 형태 분석 말뭉치 1529 만 어절, 형태의미 분석 말뭉치 1269 만 어절, 구문 분석 말뭉치 82 만 어절의 말뭉치가 존재한다[1]. 각 말뭉치들은 세종 계획에 정의되어 있는 메타정보 태그들로 잘 정리되어 있어 언어처리를 하는 연구자들이 사용하기 좋은 코퍼스이다. 이러한 세종 말뭉치를 이용하면 각종 통계 자료를 구할 수 있고, 이를 토대로 몇 년치 기사들에 대해 구문적 특징을 구할 수 있다[2,5]. 그러나 세종 말뭉치를 활용하는데 있어 세종 말뭉치 규칙에 대하여 쉽게 자료를 추출, 재가공할 수 있는 다양한 도구들이 부족하여 세종 말뭉치에서 사용자가 필요한 형태로 가공하여 활용하는데 불편한 점이 있다. 이러한 세종말뭉치의 활용도를 높이고자 보다 손쉽게 데이터 추출, 가공을 위한 말뭉치 추출 도구를 구현하였다.

본 논문에서 소개하는 말뭉치 활용 도구 CMT (Corpus Management Tool) version 1.0 은 현재까지 구축되어 있는 세종 말뭉치를 이용하여 사용자가 원하는 정보를 추출하고 손쉽게 정보를 재가공할 수 있게 해준다. 세종 말뭉치가 제공하는 원시, 형태, 형태의미, 구문을 기준으로 4 가지의 기능을 구성하였으며, 각 기능마다 여러 가지 옵션을 두어 사용자가 원하는 형태의 데이터 정보를 추출할 수 있도록 하였다.

또한, 말뭉치를 활용할 때 공통적으로 사용되는 기능에 대해서는 세종 말뭉치 이외의 말뭉치에서도 사용할 수 있도록 일반 기능을 구현하여 범용성을 높였다. 일반 기능에서는 한글 코드 변환, 파일 합병, 빈도 계산 기능을 제공한다.

2. 말뭉치 활용 도구

본 절에서는 CMT version 1.0 의 구조와 각 기능들을 상세히 설명한다.

2.1 말뭉치 활용 도구의 구조

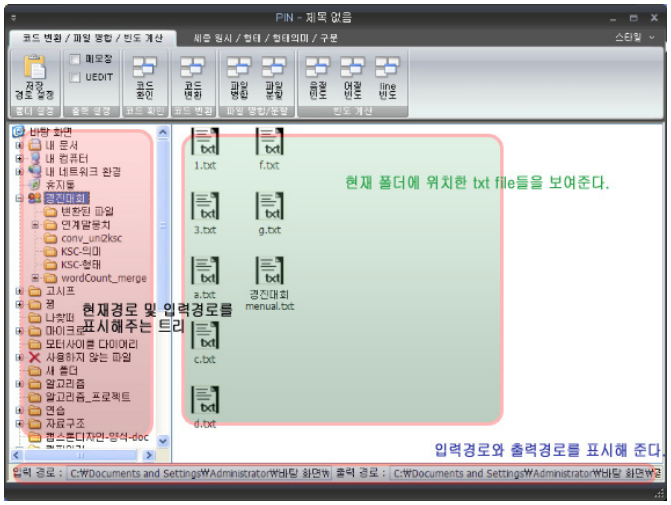
CMT version 1.0 은 한글 텍스트 파일로 저장된 일반 말뭉치에 대해 공통적으로 사용할 수 있는 기본 탭과 세종 말뭉치의 원시, 형태, 형태의미, 구문 말뭉치에만 적용되는 기능 탭으로 구성된다. 그림 1 은 일반 텍스트 파일에 적용되는 기본 탭의 기능이며, 그림 2 는 세종 말뭉치에만 적용되는 탭이다.

(그림 1) 일반 기능 탭 화면

(그림 2) 세종 말뭉치 탭 화면

기능 탭들은 도구의 위쪽에 배치해 놓았으며, 가운데를 기준으로 왼쪽에는 탐색기, 오른쪽에는 현재 탐색기가 가리키고 있는 폴더에 존재하는 텍스트 파일

들을 보여주고, 최하위 부분에는 입력 경로와 출력 경로를 표시해 준다. 그림 3 은 말뭉치 활용 도구의 전체 화면이다.



(그림 3) 말뭉치 활용 도구 전체 화면

2.2 일반 기능

말뭉치 활용 도구는 세종 말뭉치에만 적용되는 기능과, 모든 말뭉치(한글 텍스트 파일)에서 적용이 가능한 일반 기능 두 가지 인터페이스를 제공한다. 그 중에서 일반적인 한글 텍스트 파일에 공통적으로 사용되는 기능은 다음과 같다.

- 한글 코드 변환
  완성형, UTF-8, 유니코드 LE, 유니코드 BE
- 말뭉치 파일 관리
  파일의 병합 및 분할
- 빈도 계산
  음절 빈도, 어절 빈도, 라인 빈도

(1) 저장 경로 설정

도구를 실행시키면 그림 3 과 같은 화면이 나타난다. 사용자는 먼저 작업을 수행할 디렉터리를 탐색기를 통하여 이동한다. 탐색기 오른쪽에 있는 화면은 탐색기가 가리키고 있는 현재 디렉터리에 있는 모든 텍스트 들을 검색하여 보여준다.

그림 1 을 보면 저장 경로를 설정하는 버튼이 있다. 이 버튼을 클릭하면 현재 디렉터리의 경로가 저장되게 되는데 이는 앞으로 수행할 모든 결과물 파일들이 이 디렉터리에 생성되도록 설정하기 위한 것이다. 저장 경로를 설정하지 않으면 default 경로인 현재 작업 디렉터리에 결과물 파일들이 생성된다. default 경로를 사용하면 데이터 추출 시 원본 파일과 출력 파일이 혼동될 수 있으므로 저장 경로를 설정한 후에 작업하는 것이 바람직하다.

(2) 한글 코드 변환

한국어 텍스트 파일을 작성할 때 사용되는 한글코드는 KS 완성형 또는 유니코드이다. 유니코드는 국제 표준을 따르는 장점이 있는 반면에 파일로 저장할 때 UTF8 과 UTF16(LE,BE)이라는 3 가지 인코딩 방식이 사용된다. 즉, KS 완성형 한글코드를 포함하여 한글 문서는 4 가지 형태의 한글코드 파일이 사용되고 있다. 이 중에서 대용량 말뭉치에서 필요한 정보를 추출하거나 빈도를 계산하는 등 말뭉치 파일을 활용하는 도구는 KS 완성형 파일을 기준으로 작업하는 것이 편리하다.

그런데 세종 말뭉치는 UTF-16 인코딩 방식으로 저장되어 있으며, 특히 국외에서 생성된 대부분의 파일은 UTF-8 인코딩 방식으로 저장된다. 이와 같이 한글 코드의 다양성으로 인해 발생하는 문제점을 해결하기 위해 말뭉치 파일의 한글코드 혹은 유니코드의 인코딩 방식을 변경하여 저장할 필요가 있다. 말뭉치 활용 도구는 코드 확인 기능을 통해 현재 사용자가 다루는 파일의 코드 유형을 자동으로 검사하고 원하는 형태로 변환할 수 있도록 코드 변환 기능을 제공한다. 인코딩 변환 방식[3] 에는 UTF-8, Unicode, KS 완성형이 서로 호환 변환이 가능하도록 하였다.

특히, 유니코드에서 KS 완성형으로 변환할 때 독일어 움라우트, 중국어 한자 등 KS 완성형으로 변환이 될 수 없어서 정보가 손실되는 유니코드 문자들에 대해서는 해당 유니코드 값을 16 진수값 U+xxxxxx 형태로 변환하여 보존하게 하였다. 역으로 KS 완성형 파일을 유니코드로 변환할 때는 U+xxxxxx 부분이 원래 유니코드 문자로 원상 복구되도록 하였다. 유니코드 값을 16 진수 4 자가 아니라 6 자로 저장한 이유는 유니코드 확장 영역의 경우 4 자로 저장할 수가 없기 때문이다.

(3) 파일 병합 및 분할

파일의 병합 및 분할 기능은 말뭉치를 구축할 때 여러 개의 파일로 나누어 저장한 경우에 필요한 기능이다. 세종 말뭉치의 경우 원시, 형태분석, 형태의미 말뭉치 등이 수백 개의 파일로 분할되어 구축되어 있다. 수백 개 파일로 구성되어 있는 말뭉치에서 빈도를 계산하려면 각 파일들에 대해 빈도를 계산하여 합산해야 한다. 또한, 여러 파일에서 특정 단어가 포함된 문장을 추출하거나 특정 유형의 에러 수정 등의 작업을 할 때 불편한 점이 있다. 파일의 병합-분할 기능은 이러한 불편한 점을 해결하기 위한 것이다. 여러 말뭉치 파일을 병합한 후 작업을 하고 말뭉치가 수정된 경우 다시 분할 기능을 이용하여 원래의 말뭉치 파일들로 나누어 저장한다. 원래 파일로 분할하기 위하여 파일을 병합할 때 해당 파일명을 각 파일의 첫 부분에 저장한다.

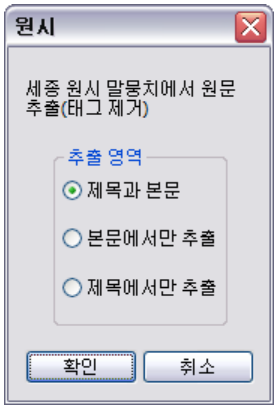
(4) 빈도 계산

말뭉치 데이터에 대한 빈도 계산 작업은 자연어 처리 연구를 수행하는데 가장 기본적인 기능이다. CMT version 1.0 은 말뭉치로부터 다양한 통계 자료를 구축할 수 있도록 음절, 어절, 줄(라인) 단위의 3 가지 형태의 빈도 계산 방법을 제공함으로써 사용자가 필요로 하는 형태의 빈도를 계산할 수 있도록 한다.

3. 세종 말뭉치 정보 추출 기능

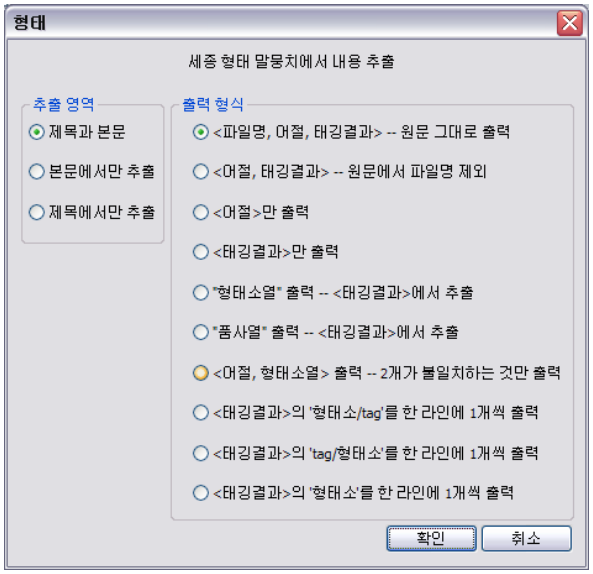
CMT version 1.0 의 메뉴에서 세종 말뭉치 탭에는 원시, 형태, 형태의미, 구문 버튼이 있다. 이 기능들은 세종 말뭉치에만 사용되는 것이므로 다른 일반적인 텍스트 파일 말뭉치에는 사용할 수 없다. 다만, 세종 말뭉치 메타정보 태그 규칙에 따라 말뭉치를 구축하였다면 이 기능을 사용할 수 있을 것이다.

'원시' 기능은 원시 말뭉치에서 태그를 제거하고 원문만 추출하기 위한 것이다. 원문을 추출할 때 제목, 본문의 추출 영역에 따라 그림 4 와 같이 세 가지 형태로 원문을 추출할 수 있다.



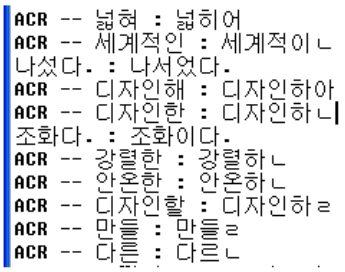
(그림 4) 원시 말뭉치 기능

원시 말뭉치에서 형태소, 태그 등 빈도를 계산하려면 말뭉치에 포함된 메타데이터를 제외한 빈도를 계산해야 한다. 즉, 작업일자와 원문 출처 등 메타데이터는 빈도 계산에서 제외되어야 한다. 이러한 필요성에 의해 원시 말뭉치는 그림 4 의 옵션과 같이 세종 원시 말뭉치에서 제목 영역, 본문 영역, 제목과 본문 모두에서 정보를 추출할 때 등에 대해 세종 원시 말뭉치로부터 메타데이터 태그들이 제거된 결과물을 얻을 수 있다.



(그림 5) 형태 말뭉치 기능

'형태' 기능은 형태 말뭉치에서 그림 5 와 같이 다양한 옵션들을 제공하며 사용자의 선택에 따라 출력물을 결정짓는다. 형태 말뭉치 구축 작업 과정에서 형태소열 부분에 오류가 포함된 경우가 있다. 이러한 오류 유형을 발견하기 위하여 그림 5 의 7 번째 옵션으로 <어절, 형태소열>을 출력하는 기능을 추가하였다. 이 기능은 입력어절과 형태소열을 결합한 스트링을 비교하여 두 개가 서로 다른 것들만 출력해주는 것으로 그 예는 그림 6 과 같다.



(그림 6) 형태소열 불일치 추출 예

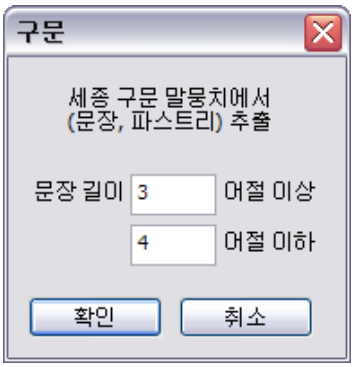
세종 형태 말뭉치에서 각 형태소별 태그 형태[4] 에 대해서 빈도 계산이나 각 형태소 구성 요소들을 구하고자 한다면 그림 5 의 기능에서 '형태소/tag' 출력 기능을 사용하면 된다. 추출된 결과에 대해서 일반 기능의 빈도계산을 사용하면 형태소/태그별 빈도를 구할 수 있을 것이다. 반대로 각 태그별 형태소 결과에 대하여 빈도 계산을 하고 싶다면 'tag/형태소' 출력 기능을 이용한다. 그림 7 은 형태소/tag, tag/형태소 결과를 보여준다.

(그림 7) 형태소와 태그쌍 추출 예

'형태 의미' 기능은 형태 기능의 요소들을 모두 가지고 있으며, 어깨번호 제거 기능을 추가하였다. 세종 말뭉치의 내용들을 보면 같은 이름을 가진 말뭉치들이 원시, 형태, 형태의미 각각에 모두 존재한다. 이는 원시 말뭉치로부터 형태, 형태의미에 대하여 데이터를 추출한 것이라고 볼 수 있다. 형태, 형태의미 추출 과정 중에 실수로 그 결과가 미 반영되거나 잘 못된 결과가 나타날 수도 있는데 이러한 오류에 대해서 검사를 원한다면 어깨번호 제거 기능을 사용하면 된다.

이는 형태의미 말뭉치의 어깨번호를 제거하게 되면 형태 말뭉치와 구조가 같아짐을 의미한다. 그렇게 되면 어깨번호를 제거한 형태의미 말뭉치와 형태 말뭉치 두 말뭉치를 비교하여 말뭉치 오류를 찾아낼 수 있다.

'구문' 기능은 세종 구문 말뭉치로부터 문장의 어절수 범위에 따라 구문 말뭉치를 추출하는 기능이며 그림 8 과 같이 제공 된다. 구문 말뭉치에는 문장의 크기 작아서 어절수가 3 이하인 문장들이 다수 포함되어 있으며, 또한 어절수가 90 이상인 문장도 있다. 구문 말뭉치에서 의미 있는 정보를 추출하기 위해서는 문장의 길이가 어느 정도 이상(예: 5 어절 이상의 문장)이어야 하며, 또한 필요에 따라 50 어절 이상과 같이 문장 길이가 너무 긴 것을 제외할 필요성이 발생할 수 있다.



(그림 8) 구문 말뭉치 추출 기능

그림 8 의 예는 3 어절 이상 4 어절 이하인 문장에 대한 파스 트리들만 추출하는 것으로 구문 말뭉치에서 모든 문장 중에서 3 어절과 4 어절 길이인 것만 추출하겠다는 의미이다. 추출과 동시에 해당 말뭉치의 문장 전체의 개수와 선택된 어절 구간의 문장 개수를 출력해 주기 때문에 전체 문장 비율 대 선택된 어절 비율을 구할 수 있도록 하였다. 이를 이용하면 전체 말뭉치에서 특정 어절수로 이루어진 문장들의 개수를 계산할 수 있다.

4. 기대 효과

CMT version 1.0 은 말뭉치를 활용하는데 필요한 기능들을 제공하는 도구로서 일반 말뭉치 파일과 세종 말뭉치에 대한 두 가지의 형태에 대해 적용되는 기능으로 구성되었다. 이 도구는 말뭉치 활용 도구로 사용될 수 있다. 말뭉치 활용 작업은 빈도 계산이 필수적이다. 여러 파일들로 구성되어 있는 말뭉치들에 대하여 각각 빈도를 계산하고, 그 계산된 빈도들을 더한다는 것은 매우 비효율적일 뿐만 아니라 빈도결과를 합산하는 과정에 있어서 오류가 발생할 가능성도 있다. 이러한 문제점을 해결하기 위하여 파일 병합-분할 기능을 제공하며, 하나로 병합된 파일에 대해 바로 빈도를 계산할 수 있도록 하였다. 이는 파일들을 한 개로 병합하는 프로그램을 실행시키고, 그 결과에 대하여 빈도를 계산해 주는 프로그램을 실행하는 두 번의 작업을 한 개의 툴 안에서 해결할 수 있게 하여 쉽고 편하게 작업할 수 있게 하였다.

이 도구는 세종 말뭉치에서 정보를 추출하거나 오류를 수정하는 도구로 사용된다. 세종 말뭉치는 자동 및 수작업 과정으로 구축되었다. 이 과정에서 오류가 포함되기도 한다. 말뭉치 활용 도구는 세종 말뭉치의 메타 정보인 태그들을 기준으로 해당 문서에서 누락되거나 혹은 원시로부터 얻어진 형태, 형태의미 말뭉치에 대하여 두 말뭉치를 비교함으로써 오류를 검출할 수 있다. 또한 여러 가지의 정보 추출 방법들을 제공함으로써 사용자가 세종 말뭉치를 이용하여 원하는 형태로 정보를 가공할 수 있다. 이러한 가공된 데이터를 이용하여 일반 말뭉치에 대한 보조 활용 기능을 추가적으로 사용함으로써 가공된 데이터의 빈도 계산을 수행할 수 있다.

5. 결론

본 연구에서는 세종 말뭉치들을 처리하는데 있어서 반복 작업, 조건에 따른 데이터 추출 및 가공 등 여러 기능들을 하나의 도구에서 사용할 수 있도록 하는 말뭉치 활용 도구를 개발하였다. 이 도구를 사용하면 세종 말뭉치 처리에 있어서의 빈도 계산, 오류 검출 및 수정이 용이하며, 특히 각종 기능들의 조합으로 여러 가지 형태로 필요한 데이터를 추출할 수 있어서 세종 말뭉치를 활용하는 사용자들이 편하게 사용할 수 있다.

CMT version 1.0 은 세종 문어체 말뭉치에서 오류 수정 작업을 수행해야 하는 필요성에 의해 개발되었으며, 국립국어원이 주관한 "2009 국어 정보처리 시스템 경진대회"에서 장려상을 수상하였다[6]. 이 도구의 파일 병합-분할 기능은 세종 말뭉치 파일에서 특정 유형의 오류가 발견되었을 때 수백 개의 파일에 대해 해당 오류가 있는지를 검사해야 하는 비효율성 문제를 해결해 주었다. 코드변환 기능의 경우 유니코드로 구축된 말뭉치를 KS 완성형으로 변환하여 작업을 수행하고 다시 유니코드로 변환하는데 활용하였다.

참고문헌

[1] 문화관광부, 21 세기 세종계획 국어 기초자료 구축, 2000.

[2] 국립국어원, 21 세기 세종계획 성과물 관리 및 배포 지원 최종보고서, 2007.

[3] 안대혁, 박영배 "유니코드의 한글 인코딩 표준안", 정보과학회논문지: 소프트웨어 및 응용, 34 권 12 호 , pp.1083-1092, 2007.

[4] 김형준, 임동희, 강승식 "세종 계획 말뭉치를 이용한 품사 태거의 성능 개선" , 한국정보과학회 종합학술대회논문집(C) , pp.117-180, 2007.

[5] 박만규, 이선웅, 나윤희 "21 세기 세종계획 관용표현 전자사전 구축에 대하여", 제 13 회 한글 및 한국어 정보처리 학술대회, pp.334-340, 2001.

[6] 국립국어원, 2009 국어 정보처리 시스템 경진대회 발표자료집, 2009.