언론 기사에 나타난 신(조)어 조사 도구의 설계 및 구현

송인성*, 정희석*, 이상곤*, 이래호**
*전주대학교 컴퓨터공학과 언어과학실
**전북대학교 인문한국 쌀・삶・문명연구원(HK사업단)
{fusionsong*, raeholee**}@hanmail.net, mosaick@naver.com*, samuel@jj.ac.kr*

Design and Implementation of Detecting Tool for New Word in Korean Journal Articles

In-sung Song*, Hee-seok Jeong*, Samuel Sangkon Lee*, and Raeho Lee**
*Dept of Computer Science and Engineering, Jeonju University,
**Dept of Korean Language, Chonbuk National University

요 약

신조어 조사용 프로그램은 웹에 실시간으로 등록되는 언론 기사를 수집하는 웹 에이전트를 개발하여 텍스트를 추출하고, 간단한 어휘 분석을 통하여 국어사전에 등록된 표제어와 이미 연구자가 발견한 기존의 신조어를 제외하고 새롭게 생성된 신조어를 추출하는 작업을 하는 도구이다. 인터넷의 언론 사이트에서 규칙적인 URL 패턴을 발견하고 뉴스 기사를 수집한다. HTML 소스 분석을 통하여 언론 기사만을 추출하고 이 기사에서 사전의 표제어와 기존에 조사된 신어를 제외하여 국어 전공자가 신어를 찾아내는 작업을 하는데 사용하는 시스템을 설계하고 구현하였다.

1. 서론

하루가 다르게 사회가 변하고, 문화의 다양화가 급격히 이루어짐에 따라 새로운 개념이나 문물이 유입되고 새로운 제도가 생기고 있다. 이에 따라 전에 없던 개념이나 사물을 표현하기 위해 새로운 말도 생겨나게 된다. 또한 새로운 직종과 첨단 학문이 도입되면서, 이를 표현하고자 하는 전문적인 성격을 띤 새로운 말들, 즉 신어(新語)[1] 혹은 신조어(新造語), 전문용어들이 대량으로 생겨나고 있다. 언어는 사회의 변화를 가장 민감하고 빠르게 드러내 보여준다. 오늘날처럼 신조어가 기하급수적으로 늘어가는 것은 우리 사회가 그만큼 급속히 변화하고 있음을 뜻한다[1]. 그러나 모든 신조어가 그대로 굳어 오랫동안 통용되는 것은 아니다. 어떤 말들은 일시적으로 사용되다가 사라지기도 하고 어떤 말들은 오래 살아남아 사전에 실리기도 한다. 본 논문에서 제안하는 프로그램의 개발 목적은 이상과 같은 신어를 효율적이고 통합적으로 조사하는 도구의 개발에 있다.

신어의 생성 유래에 대해서는 외국의 것을 차용해 오는 경우와 국내에서 자생적으로 만든 경우 등 크게 두 가지로 나눌 수 있다. 외래어의 차용은 역사적인 배경에 따르지만, 대부분 높은 문화권의 언어가 낮은 문화권의 언어로 차용되는 일방적인 양상을 띠는 경우가 많다. 한 문화권에 없던 문물이 다른 문화권에서 들어오면 새로운 명칭이 필요하게 되어 차용이 필연적이 된다. 이와 같이 새로 생겨난 신어들은 기존의 국어사전에는 등재되지 못한 상태이다. 따라서 신어들의 정확한 의미와 함께, 그 신어가 사용되는 전문적인 영역, 용례, 신어의 유형, 사용 양상, 시간적인 변화 등이 체계적으로 정리되어야 할 필요성이 있다. 이러한 작업은 국어의 어휘를 풍부하게 만드는 것이며, 동시에 새로운 어휘들이 어떤 유형으로 결합되어 단어나 구의 형태와 의미를 형성하는 지에 대한 객관적 기술은 국어 어휘론[5, 6, 7] 연구에 있어 중요한 기초로 이용될 수 있다. 또한, 언어 변화의 방향을 가늠하여 국어 정책[2]을 세우는 데 기초 자료로 활용될 수 있다.

본 논문은 신문, 방송 등 언론 자료에 나타나는 신어를 조사하고 이를 정리하기 위한 목적으로 신(조)어 조사용 프로그램을 만드는 것을 목표로 한다. 국어를 전공한 사람들이 언론 자료를 통해 새롭게 수집된 신어를 효율적으로 정리하도록 원어, 전문영역, 뜻풀이, 용례 등을 손쉽게 기술할 수 있도록 관리 도구를 제공하고, 신어 유형 및 특징을 분석하는 기초 자료를 제공하고자 한다. 본 논문의 결과물을 통해 크게 두 가지의 목표를 실현하고자 한다. 하나는 기존에 조사된 신어를 토대로 각 연도별로

1) 신어(신조어)란 (1) 새로운 개념이나 사물을 표현하기 위해 생긴 말, (2) 개념이나 사물은 존재하는데 명칭이 없는 경우에도 어휘 체계의 빈자리를 채우기 위해서 생긴 말, (3) 이미 있던 개념이나 사물일지라도 그것을 표현하던 말들의 표현력이 감소됐을 때 그것을 보강하거나 신선한 새맛을 가진 말로 바꾸기 위한 대중적 욕구에 의해서 생긴 말, (4) 국어 순화 운동의 일환으로 생긴 말이다.

신문, 방송 등 언론 자료에 나타나는 신어를 조사하는 것이고, 다른 하나는 새롭게 조사된 신어의 신어들의 유형과 특징을 분석하는 것이다.

본 논문의 주요 구성은 다음과 같다. 웹 에이전트를 통해 언론 기사를 자동으로 수집한다. 인터넷 포털 사이트 네이버(Naver)의 언론사별 뉴스 사이트와 방송사의 웹 사이트에서 URL 패턴을 확인하여 원하는 월/일자 별로 기사를 수집하고, 신어 작업자의 컴퓨터에 체계적으로 저장한다. URL과 HTML 소스 분석을 통하여 정규화 되지 못한 HTML 소스에서도 특정 정보인 기사의 제목, 게시 날짜, 기사 내용 등을 수집한다. 동시에 원하는 카테고리 별로 구분하여 폴더에 저장하는 기능을 한다. 어휘 분석과 확장형 비교(사전 표제어, 기존 신어)를 한다. 기사를 구성하는 문장을 배열로 저장하여 불필요한 어미와 조사를 제거한다. (중복된 신어 조사가 되지 않도록) 기존 신어와 이미 등록된 국어사전의 표제어를 제거하고 새롭게 생성된 단어로 추정된 후보 단어를 추출한다.

2. 대상 자료

본 논문에서 조사하는 신어와 그 신어가 사용된 인용문장은 21개의 주요 신문사(스포츠 한국, 스포츠 투데이, 경향신문, 내일신문, 뉴시스, 문화일보, 세계일보, 조선일보, 쿠키 뉴스, 한국일보, YTN, 국민일보, 노컷 뉴스, 동아일보, 서울신문, 연합뉴스, 중앙일보, 한겨레, 일간 스포츠, 스포츠 서울, 스포츠 조선 등)와 3개의 방송사(KBS 9시 뉴스, MBC 뉴스 데스크, SBS 8시 뉴스 등), 2개의 인터넷 언론사(프레시안과 오마이 뉴스)에 나타난 기사를 대상으로 하였다. 이 기사들은 인터넷 포털 사이트인 '네이버(NAVER)'에서 제공하는 언론사별 뉴스 페이지에서 수집하였다.

3. 조사 방법 및 내용

본 논문의 연구를 위해 선정한 대상 자료에서 신어를 조사하는 방법은 다음과 같다.

3.1 연구 조사 방법

① 조사 대상이 되는 언론 자료를 수집하고 이를 체계적으로 저장하여 말뭉치화 하고, ② 각 신어의 용례별 탐색기를 개발하여 표제어를 추출하고 그 사용 용례를 자동으로 작성한다. ③ 기존에 조사된 신어(23,220개[2], 이하 기존 신어라 통칭)와 사전에 이미 등재된 표제어(참고문헌 [2]의 국립국어원에서 편찬한 표준국어대사전의 표제어 356,982개, 이하 사전 표제어라 통칭)를 비교하여 추출하고, 그 결과를 이용하여 1차(컴퓨터에 의한 추출)/2차(국어 전공자의 선별작업에 의한 추출)/3차(국어 전문가의 최종 판단에 의한 추출 작업) 신어 후보의 목록을 각각 작성한다. ④ 국어의 고위 전문가가 신어의 확정 작업을 하는데 기초 자료를 제공하고, ⑤ 최종적으로 결정된 신어의 원어 정보와 뜻풀이 등을 기술할 때 참고 자료를 제공하고, 실제 언론 기사에서 사용되는 적절한 용례를 탐색하여 최초의 출현 시기와 계속적인 사용 유무를 추적하는 기능을 추가한다.

(1)에서 제시한 일간지의 정치, 경제, 사회, 생활·문화, 스포츠, 연예, 국제, 정보 통신 등 7개 부문의 기사와 뉴스 방송 대본을 모아 말뭉치를 구성한다. 말뭉치의 구성은 일간지 및 방송사 홈페이지의 기사 및 뉴스 원문을 매일 스캐닝/트래핑 할 수 있는 프로그램을 개발하였다. 이렇게 말뭉치화 된 자료를 바탕으로 용례를 탐색하는 탐색기를 이용하여 표제어를 추출한다. 이와 같이 추출된 표제어는 사전 표제어와 기존 신어 목록을 비교하여, 표제어로 등록되지 않은 것을 신어 후보어로 추출한다. 신어 후보 추출에는 표제어 비교 프로그램을 제작하여 이용할 것이다.

이렇게 추출된 후보를 바탕으로 올해 만들어진 신어를 확정한다. 이때 신어와 유행어를 구분하는 기능도 추가한다. 신어는 그 사회에서 대치할 말이 없거나, 어떠한 개념을 표현해야 할 사건이 계기가 되어 새로이 만들어진 새로운 표현의 어휘로 판단한다. 한편, 유행어란 정보 매체가 발달함에 따라 어떤 사건이나 계기를 통해 만들어진 말이 사람들의 대화 등에서 한 시기에 한시적으로 널리 흔하게 쓰이는 것으로 판단한다.

<표 1> 신어 정보의 기술 내용

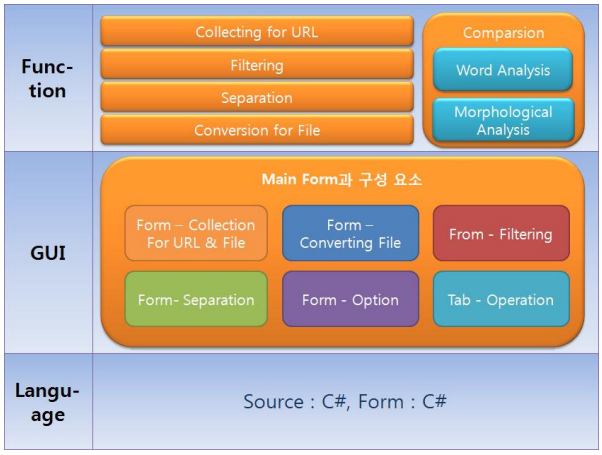
순서	기술 정보
1	표제어
2	단어 분석(형태소 분석)
3	품사 정보, 사용 영역, 의미
4	용례
5	출전
6	보도 연월일
7	구분
8	기타 특기 사항

신어 후보를 선정할 때 우선적으로는 단어가 표제어 후보가 될 것이지만, 표제어로 올릴 가치가 있다고 판단되는 것은 구도 표제어로 선정할 것이다. 특히 전문어는 구로 된 표제어가 많을 것으로 예상된다. 표제어가 선정되면, 신어의 원어 정보와 의미 등을 기술하고, 적절한 용례를 제시한다. 신어에 기술될 내용을 제시하면 다음과 같다.

3.2 신어 정보의 기술 내용

발견된 신어의 원어 정보는 <표 1>과 같이 형태소 분석, 품사 정보, 사용 영역, 뜻풀이 등을 기술한다. 신어가 한자어나 외래어인 경우에는 모두 원어를 밝혀 줌을 원칙으로 한다. 또한 원어가 어느 언어인지도 함께 밝혀 준다. 품사 정보, 전문어 영역, 뜻풀이의 기본 원칙은 국어사전의 지침에 따라 한다. 용례는 가능한 한 그 의미를 정확히 보여줄 수 있는 것으로 제시하고, 그렇지 않을 경우에는 다양한 용례를 제시하여 용례를 통해 표제어의 다양한 쓰임새를 살펴볼 수 있도록 한다. 용례나 인용문은 어문 규범에 어긋나는 부분을 손질하여 제시하고, 해당 신어가 조사연도 이전 시기에도 발견된다면, 인터넷 검색을 통해 가장 이른 시기에 사용된 것으로 보이는 인

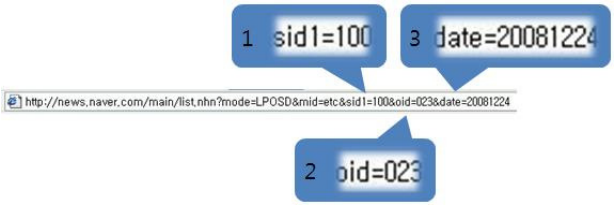
용문을 제시한다. 용례 다음에는 반드시 출전을 표기한다. 이와 함께 필요한 경우 관련 어휘와 참고 어휘가 있는 경우에는 뜻풀이 다음에 표제어의 구분, 기타 특기사항을 기록한다.



(그림 1) 시스템의 전체적인 구성

4. 시스템의 구성

(그림 1)에서 보는 바와 같이 프로그램은 다섯 가지 기능을 하며 각 기능에 해당하는 개별 폼(form)을 제공한다. (1) 기사의 URL[2]을 얻기 위한 기능(Collection for URL)과 (2) 얻어진 기사의 URL을 가져오는 기능, (3) 기사에서 불필요한 노이즈를 삭제하는 필터링(Filtering), (4) HTML 태그의 분리 기능(Separation), (5) 수집된 기사 본문을 여러 종류의 파일로 변환하는 기능(File Converting) 등이다. 국어 전공자의 신어 조사 작업을 용이하게 하기 위해 다음의 두 가지 방법으로 신어 후보어를 추출한다. 첫째는 어휘 분석 방법(Word Analysis)이며, 둘째는 형태소 분석 방법(Morpheme)을 이용하였다. 이 방법들을 이용하여 신어 1차/2차/3차의 신어 후보 추출 과정을 거치고 전문가가 최종적으로 국어사전에 등재될 신(조)어를 선정한다.



(그림 2) 인터넷 포털 사이트 네이버(NAVER)에서 URL 자동 수집의 예

인터넷에 언론사의 뉴스는 각 언론사별로 특정한 URL 형식이 존재한다. 언론 기사를 수집하기 위해서는 각 사이트의 특징을 파악하여야 한다. 예를 들어, 어떤 URL에 해당하는 기사를 수집하려면 위의 (그림 2)와 같은 URL 정보에서 아래와 같이 세 가지 인식 작업을 수행한다.

1. "sid = 번호": 번호는 특정한 분야를 나타낸다. 예를 들어, 100은 정치, 101은 경제 등 이와 같이 번호별로 구분된다.

2. "oid = 번호": 언론사의 이름을 나타내는데, 예를 들어, 숫자 023은 조선일보를, 024는 매경 이코노미를 나타낸다. 각 언론사가 특정 번호로 구분되어 있다.

3. "date = 날짜"는 웹 기사의 날짜를 나타내는 부분으로, 년/월/일이 연속되어 있다. 이러한 형태의 URL을 프로그램 내에서 자동 생성되며, 해당 URL에 있는 뉴스 기사를 컴퓨터로 수집하고 해당 기사를 파일로 저장하게 된다.

(그림 3) HTML 소스에서 분리 작업

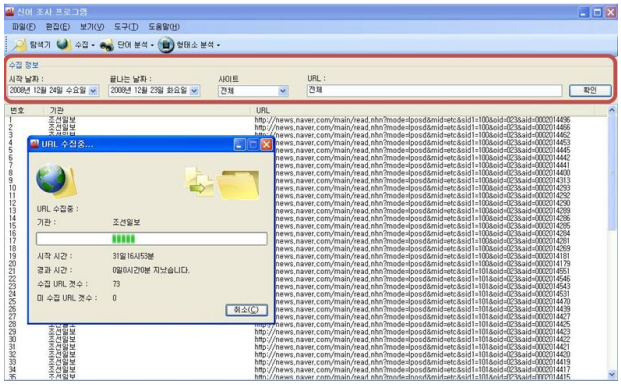
웹 사이트에서 기사를 가져올 때에는 한 사이트를 나타내는 HTML 소스가 들어가 있다. 이러한 HTML 소스를 보면 항상 기사에 포함되어야 되는 기사의 제목, 본문, 날짜가 포함되어 있으며, 각 언론사별로 고유한 형태로 분리할 수 있도록 되어 있다. 이렇게 언론사들을 조사하여 분류 지식을 정리해 놓으면 컴퓨터 프로그램이 소스 분석을 통하여 제목, 본문, 날짜 등을 분리할 수 있다.

(그림 4) 단어 분석과 지식베이스 비교 후의 후보 단어 추출

일반적으로 어휘 분석은 형태소 분석 방법이 많이 사용되는데, 자연언어처리 입장에서 형태소 분석을 수행하면 한국어 의미 단위(가장 작은 단위)인 형태소로 분리된다. 형태소 분석 작업은 어휘 분석 작업이 매우 강력하게 이루어져 자칫 신어 후보 발견을 어렵게 만들 수 있다. 따라서 본 논문의 연구방법에서는 체언과 용언(어절)에서 어미와 조사만을 분리하는 단어 분석 방법을 이용한다. 단어 분석 방법은 어미/조사 목록(6,725/5,442)에서 그 길이가 가장 긴 순서대로 정렬(최장일치법)하여 기사에서 나온 단어를 음소별로 매칭한 후 국어사전의 표제어 목록과 기존 신어 목록에서 차집합(差集合)을 계산하여 작

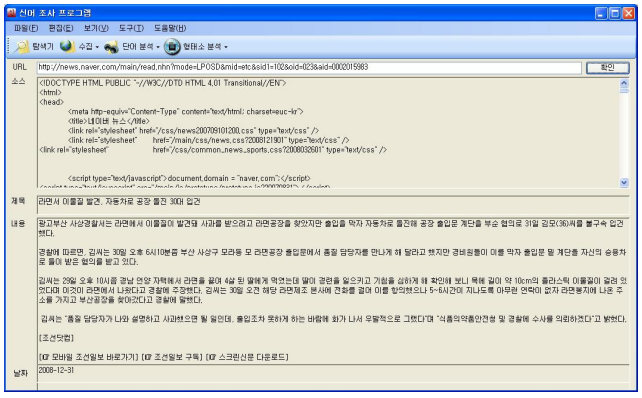
2) URL; Uniform Resource Locator(자원 위치 표시자)의 약자로 인터넷상의 파일 위치를 뜻하는 주소이다. 본 논문에서는 인터넷 상의 개별 기사의 웹페이지 요소나 웹 문서의 주소를 뜻한다.

업자가 선택한 단어가 표제어 목록과 기존 신어 목록에 없을 때 작업 화면에 신어 후보로 출력한다. 위의 (그림 4)에 단어 분석의 과정과 그 결과를 제시하였다.



(그림 5) URL 수집기의 예

본 프로그램의 구성은 언론 기사 URL의 수집부터 출발하였다. 주소 수집 패널에서 수집 정보(수집을 요하는 시작 날짜, 끝나는 날짜, 언론사 사이트 등)를 입력하면 기사의 URL만 수집되며, 수집 URL 목록은 목록 화면(리스트 뷰)을 통해 사용자에게 제공되며 목록은 파일 형태로 저장된다.

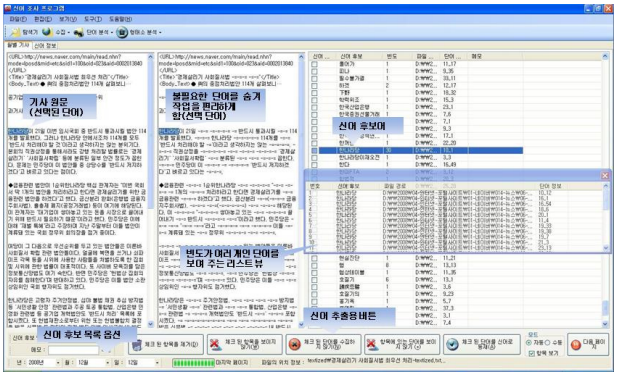


(그림 6) HTML 소스의 분리 과정

URL 입력 창에 기사 URL을 입력하면 링크들이 제목/본문 내용/날짜에 따라 자동으로 분리되어 텍스트 박스에 출력된다. 이 패널은 사용자에게 보여주기 위한 목적으로 제작하였으며, 아울러 모듈 태스팅(testing) 목적으로도 사용될 수 있다.

파일 변환을 시작하면 각 단계별 변환 작업(7가지 형태)이 시작된다. 다음은 단계별 변환 작업 중 주요 처리에 대한 설명이다. 문장 분리 작업(/sentenced)이 일어나고, 구 형태의 단어 처리(/phrasal-recognized), 어절별 분리 파일(/segmented)을 수행한다. 그리고 신어 추출 목록 파일을 만들어 신어 후보 목록에 보여 주도록 한다.

신어 후보 목록을 보여주는 패널은 추출 작업을 원하는 특정 날짜에 대한 신어 후보 목록을 보여줄 수 있는 “신어 후보 목록” 옵션이 있다. 이 신어 후보 목록을 계산하여 출력하면 작업자가 신어 후보를 선택 시, 기사의 원문과의 대조 작업이 용이하게 만들어 보여준다. 또한 빈도가 두 번 이상인 단어는 두번째 단어부터는 다른 항목에 보여 주도록 되어 있으며, 작업자의 작업이 불편하지 않도록 관리 기능을 추가하여 메시지 처리가 자동으로 수행되도록 하였다. 덧붙여 신어 추출 작업을 하면서 얻어진 필요 없는 단어(배제 정보)들을 다음부터는 수집되지 않도록 하여 작업자가 검토하여야 할 작업량을 점차 줄여갈 수 있고 작업 능률을 향상시켰다.



(그림 7) 신어 후보의 추출 과정

4. 결론

본 프로그램은 MS-Windows XP에서 Microsoft Visual Studio 2008 도구를 이용하여 C# 언어로 개발하였다. 본 조사 도구를 통해 얻을 수 있는 기대 효과는 다음과 같다. 먼저, 이번 조사 도구로 수집한 대용량의 대상 자료는 우리말 어휘의 생성과 관련한 연구 자료로 활용될 수 있다. 이를 위해 한국어 정보학을 전공한 전문가들과 협업이 필요하다. 둘째, 국어사전의 편찬과 보완 작업 시 표제어 선정을 위한 기초 자료로 활용할 수 있다. 이를 위해서는 신어사전 편찬 팀과의 꾸준한 연구가 진행되어야 할 것이다[8]. 셋째 국어 어문 정책의 기초 자료로 활용할 수 있으며, 마지막으로 우리말의 변천을 살피는 데 중요한 기초 자료로 활용할 수 있을 것으로 기대된다.

참고문헌

[1] 국립국어원, 2002년 이후 생겨난 새말, 사전에 없는 말 신조어, 태학사, 2007.

[2] 국립국어원 공식 사이트(http://www.korean.go.kr/), 자료마당/어문규정

[3] 김상형, 닷넷 프로그래밍 정복, 가메출판사, 2008.

[4] 최재규, Visual C# . NET 2005, 2nd Ed., 영진닷컴, 2004.

[5] 정성륜, 어휘조직론, 태학사, 1998.

[6] 시정론, 국어의 단어 형성 원리, 한국문화사, 1998.

[7] 이광호, 국어 어휘 의미론, 월인, 2004.

[8] 박형익, 신어 사전의 분석, 한국문화사, 2005.