새로 생겨난 단어의 의미를 기술하는 프로그램의 설계 및 구현

*김 웅 희, 이 상 곤
*전주대학교 컴퓨터공학과 언어과학실
e-mail : kuh7770@naver.com, samuel@jj.ac.kr

Design and Implementation of Meaning Collecting Tool for New Words

Unghee Kim, and Samuel Sangkon Lee
Dept of Computer Science & Engineering, Jeonju University

요 약

웹에 실시간으로 등록되는 언론 기사를 수집하는 웹 에이전트를 개발하여 텍스트를 추출하고, 어휘 분석을 통하여 어미/조사를 자동으로 제거하고, 국어사전에 등록된 표제어를 제외하여 새롭게 생성된 신조어의 추출 작업을 지원하는 조사 도구를 본 논문에서 제작하였다. 본 프로그램은 웹 에이전트를 개발하여 신어의 의미를 기술하고 그 결과물을 검색엔진 시스템의 내부에 준비해 두고 있다가 고객의 검색 요구에 따라 새로 생성된 신어의 의미를 국민들에게 대민 서비스하는데 본 논문의 목적이 있다.

1. 서론

현대는 문화의 다양화가 급격하게 이루어짐에 따라 인터넷에 새로운 개념이나 문물이 계속하여 유입되고 새로운 제도도 생기고 있다[1]. 또한 토착 지식의 관리 미비로 인해 표준어 규범의 외연에 있는 생활 어휘의 수집과 관리 강화가 필요하다[2]. 이에 따라 전에 없던 개념이나 사물을 표현하기 위해 새로운 말(새말)도 생겨나게 된다. 외국으로부터 들어오는 전문(학술) 지식의 증대로 인해 국민들이 방송, 통신, 금융, ICT(정보 통신 기술), BT (생명공학), CT(문화 기술), NT(나노 기술), ST(우주 항공 기술) 등 관련 용어, 디지털 신기술 용어에 대한 이해도가 낮아져 국민들의 의사소통이 불편해지고 있는 실정이다. 이를 위해 실용 국어의 시대를 지향할 적극적인 사전(辭典) 정보의 국가적인 종합 지원이 필요하다.

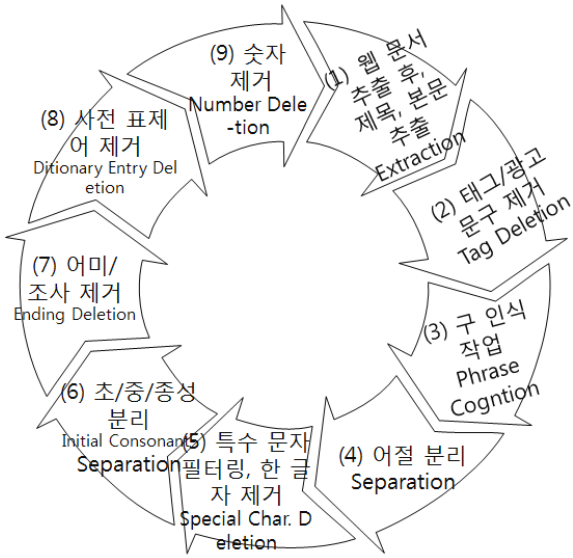
본 논문에서는 다음과 같은 방법으로 연구를 진행한다. 웹 에이전트를 통해 언론 기사를 사이트 별로 자동으로 수집한다. 국내의 유명 인터넷 포털 사이트인 네이버의 언론사별 뉴스 사이트와 방송사의 웹 사이트에서 URL 패턴을 확인하여 원하는 년/월/일자 별로 원문 기사를 수집하고, 신어 작업자의 컴퓨터에 체계적으로 저장한다(원문 기사는 저작권 정보이므로 조사 후에는 삭제한다). URL과 HTML 소스 분석을 통하여 정규화 되지 못한 HTML 소스에서도 특정 정보인 기사의 제목, 게시 날짜, 기사 내용 등을 수집한다. 이와 동시에 기사의 해당 분야를 구분하여 원하는 카테고리별 폴더에 저장한 후에 간단한 어휘 분석을 하고, 사전 표제어와 기존에 조사한 신어(이 두 가지 작업을 '확장형 비교'라 지칭하자)를 비교하여 제거한다. 왜냐하면 사전 표제어와 기존에 조사한 신어는 반드시 새로운 신어가 될 수 없기 때문이다. 기사를 구성하는 문장을 배열로 저장하여 불필요한 어미와 조사를 제거한다. 한국어의 특성을 잘 나타내 주는 조사와 어미는 일반 언어학의 측면에서는 중요한 정보를 제공하지만, 신어 발견을 위한 정보 처리 관점에서는 필요 없는 정보이다. 따라서 이 정보를 제거하고 추출된 어절에서 (중복된 신어 조사가 되지 않도록) 기존 신어와 이미 등록된 국어사전의 표제어를 자동으로 제거하고 새롭게 생성된 단어로 신어 후보 단어를 조사한다.

2. 연구 방법 및 내용

신어를 식별하기 위해서는 말뭉치에서 수집한 용례들을 기준으로 이 용례에서 특정 형태소를 결합하거나, 삭제하거나 혹은 대체하는 등 변형이 일어난 문자열을 프로그램이 찾아 주고 인간이 그 문법성 여부를 판단한다. 새로운 신어를 조사하기 위한 연구의 조사 과정을 다음과 같이 아홉 단계로 나누어 살펴본다. ① 자료 수집 : 조사 대상이 되는 사이트에서 언론 자료를 <표 1>과 같이 수집하고, 이를 체계적으로 저장하여 말뭉치화 한다. 이 기능을 하는 프로그램은 반자동화[1]된 자료 수집기라 할 수 있다. 이 수작업은 인간이 하면 매우 힘든 작업이다. 반자동 수집기를 작성하는 것만으로도 인간의 작업을 크게 덜어 줄 수 있다. ② 자료의 구조화 : 본 연구의 결과물 형태를 세

1) 컴퓨터에 의한 자동화 된 작업과 사람의 개입이 필요한 작업, 이 것을 '반자동화라 표현'하였다.

종 기초 말뭉치의 기본 형식과 유사하게 유지하고, 표제항 추출 : 각 신어의 용례별 탐색기를 개발하여 표제어를 추출하고, 그 사용 용례를 검색하여 보여준다. 국어 전문가에게 사용 용례를 실시간으로 보여 주어야 신어를 판단하는데 크게 도움이 된다. ③ 어미/조사 복합체의 제거, ④ 표제항 비교 : 표준국어대사전, 기 조사된 신어 표제항과



(그림 1) 자료의 구조화 과정의 세부 작업(자동화)

비교 분석, 기존에 조사된 신어(이하 기존 신어라 통칭)와 사전에 이미 등재된 표제어(국립국어원에서 편찬한 표준국어대사전의 표제어 수 420,957개(중복 제거), 이하 사전 표제어라 통칭)를 비교하여 추출하고, 그 결과를 이용하여 신어 1/2차 후보 추출 : 1차(단어 분석 과정에 의한 컴퓨터의 추출)/2차(국어 전공자의 선별 작업에 의한 추출)/3차(국어 전문가의 최종 판단 작업) 신어 후보어의 목록을 각각 작성한다. ⑤ 2차 신어 후보어의 선택 및 추출 ⑥ 신어 표제항 확정 : 국어 고위 전문가가 신어의 확정 작업을 하는데 기초 자료(3차 신어 후보어 선정)로서 제공하고, ⑦ 신어 용례 사전을 구축 : 신어가 사용된 전체 용례를 제시한다. 신어 자료집의 출판을 위해 조사된 신어 중에서 전부 혹은 일부를 선택하여 출간량을 조절할 수 있는 기능을 추가한다. ⑧ 신어의 어원, 분야, 의미 등을 기술, ⑨ 신어 구축에 이용된 전체 자료와 통계 정보의 제시 : 최종적으로 출간하기로 결정된 신어의 원어 정보와 뜻풀이 등을 기술할 때 참고 자료를 제공하고, 실제 언론 기사에서 사용되는 적절한 용례를 탐색하여 최초의 출현 시기와 시간의 흐름에 따라 계속적인 사용 여부를 모니터링 하는 기능을 추가한다. 이와 같은 아홉 가지의 자동 혹은 반자동 과정을 (그림 1)에 제시하였다. 이 그림에서 (2) 자료의 구조화 과정을 완전 자동화가 되도록 하였다.

앞의 ①에서 제시한 일간지의 정치, 경제, 사회, 생활/문화, 스포츠, 연예, 국제, 정보통신 등 7개 분야의 기사와 뉴스 혹은 방송용 대본을 모아 모두 11개 분야가 하나의 말뭉치로 구성한다. 말뭉치는 일간지 및 방송사 홈페이지의 기사 및 뉴스 원문을 매일 스캐닝/트래핑(trapping) 할 수 있는 프로그램을 이용하여 구성하였다. 이렇게 말뭉치화 된 자료를 바탕으로 용례를 탐색하는 탐색기를 이용하여 잘 정돈된 표제어를 추출한다. 추출된 표제어는 사전 표제어와 기존 신어 목록을 비교하여, 표제어로 등록되지 않은 것을 최종 신어 후보어로 추출한다. 신어 후보어의 추출에는 표제어 비교 프로그램을 제작하여 이용한다. 추출된 후보어는 1차/2차/3차로 나누어 시간적/질적 검토를 통해 점진적으로 정제해 나가야 한다.

<표 1> 수집 자료의 1년분 분량

수집 기간	2012년 1월 1일 ~ 12월 31일
용량	263 GByte 각 처리 단계별 중복 허용
문자 수	1,415,000,000,000 문자
어절 수	361,336,577
파일 수	16,602,867
폴더 수	1,440,059
분야	11

<표 2> 신어 정보의 기술 내용

순서	기술 내용	태그
1	표제어	NW
2	단어 분석(형태소 분석)	COMP
3	품사 정보	POS
4	분야 정보	F
5	뜻풀이	SEM
6	용례	ILL
7	출전	FAM
8	보도 연월일	DATE
9	참고 및 기타 특이 사항	REF

발견된 신어에 대한 원어 정보는 위의 <표 2>와 같이 아홉 가지 정보로 나누어 기술한다. 이들 중에서 형태소 분석, 품사 정보, 사용 영역, 뜻풀이, 용례, 출전, 보도 연월일 등이 연구자에게 의미 있는 정보가 될 것이다. 신어가 한자어나 외래어인 경우에는 모두 원어를 밝혀 줌을 원칙으로 한다. 또한 원어가 어느 나라 언어인지 함께 제시한다. 품사 정보, 전문어 영역, 뜻풀이의 기본 원칙은 국어사전의 편찬 지침에 따라 기재한다. 용례는 가능한 한 그 의미를 정확히 보여줄 수 있는 것으로 제시하고, 그렇지 않을 경우에는 다양한 용례를 골라 제시하여 용례를 통해 표제어의 여러 쓰임새를 살펴볼 수 있는 연구를 하도록 한다. 용례나 인용문은 국립국어원의 어문 규범에 어긋나는 부분은 손질하여 제시하고, 해당 신어가 조사연도 이전에 발견되면, 인터넷 검색(네이버의 옛날 신문 검색)을 통해 가장 이른 시기에 사용된 것으로 보이는 인용문을 제시하여 그 최초 생성 시기를 사용자가 참고할 수 있도록 한다[2]. 용례 다음에는 반드시 출전을 표기한다. 필요에 따라 관련 어휘나 참고할 만한 어휘가 있는 경우에

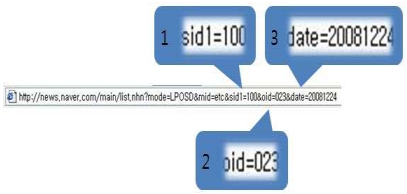
는 뜻풀이 다음에 표제어의 구분, 기타 특기사항 등을 기록한다.

본 논문에서 설계한 프로그램은 다섯 가지 기능을 하며, 각 기능에 해당하는 개별 폼(form)을 제공한다. 이 기능을 열거하면 기사의 URL을 얻기 위한 기능(Collection for URL)과 얻어진 기사의 URL을 가져오는 기능, 기사에서 불필요한 노이즈를 제거하는 필터링(Filtering), HTML 태그의 분리 기능(Separation), 수집된 기사 본문을 여러 종류의 파일로 변환하는 기능(File Converting) 등이다. 국어 전공자의 신어 조사 작업을 용이하게 하기 위해서는 다음의 두 가지 방법으로 신어 후보어를 추출한다. 첫째는 어휘 분석 방법(Word Analysis)이며, 둘째는 신어 조사에 특화된 형태소 분석 방법(Morpheme Analysis)을 이용하였다.

본 프로그램의 실행은 언론 기사 URL의 수집에서 출발한다. 주소 수집의 패널에서 수집 정보(수집을 요하는 시작 날짜, 끝나는 날짜, 언론사의 사이트 선택 등)를 입력하면 처음에는 기사의 URL만 수집되며, 수집된 URL 목록은 그림의 오른쪽과 같이 목록 화면(리스트 뷰)을 통해 사용자에게 제공되고, 실제 목록은 파일 형태로 저장 장치에 저장된다. 2013년도의 102개 사이트의 어절 수가 687,033,694 건 이었다.

신어 조사기는 신문, 방송 등의 언론 기사 수집 로봇 및 신어 후보어의 추출을 목적으로 하는 프로그램이다. 통합적이고 일괄적인 제어가 가능하도록 구현하였다. 탐색기는 도구의 옵션에 설정된 저장 경로의 하위 디렉토리의 내용을 보여 주며, 수집 URL(인터넷의 유일한 주소)은 수집 기사에 대한 수집 정보를 입력하고, 수집 URL에 수집되는 URL의 목록을 보여준다. 분류기[3]는 뉴스의 URL을 입력 받아 HTML 소스, 제목, 내용, 날짜 정보를 추출한다. 수집 파일 뷰어(Viewer)는 수집된 URL을 가져와 URL 마다 각 기사 파일을 생성하며, 그 목록을 수집 파일 뷰어에서 보여 준다. 가공 파일 뷰어는 수집 파일 뷰어에 저장된 각 파일의 위치를 이용하여 /collected, /textized, /sentenced, /segmented, /word-analyzed-compared, /morphologicalized, /word-analyzed, /word-analyzed-and-compared, /compared 등의 파일로 각각 나누어 생성하며, 각종 목록을 가공 파일 뷰어에서 제공한다. 메뉴 “단어 분석기 + 확장형 비교기”는 텍스트 파일 내에 출현하는 어절에서 어미(3,221개의 어미 연결체)와 조사(5,443개의 조사 복합체)를 분리하고, 표제어, 기존 신어(구 형태도 포함)와 매칭하여 제거한다. 이들을 제거하는 이유는 이들 단어들은 결코 신어가 될 수 없기 때문이다. 어미와 조사의 정리 작업은 한국어 어미와 조사, 어미와 조사의 이형태, 계사의 활용형과 조사, 조사의 일부와 조사, 계사(繫辭[2]), 용언의 활용형과 조사 등의 어미 연결체와 조사 구조체[3]를 대상으로

2) 연결 동사(be, become 처럼 주어와 주격 보어를 이어주는 동사), = linking verb(copula)



(그림 2) 인터넷 포털 사이트 네이버에서 기사 URL의 자동 수집

삭제하였다. 한국어 어미는 이형태를 포함하여 1,087개이며, 조사는 163개로 조사되었다. 따라서 이들 두 가지로 조합 가능한 어미 연결체는 3,221개, 조사 구조체(혹은 조사 복합체)는 5,443개 이었다. 이들 두 가지를 합한 문자열 8,664개가 신어 조사 작업의 중요한 지식베이스가 되었다.

일반적으로 어휘 분석은 형태소 분석 방법이 많이 사용되는데, 자연 언어 이해(Natural Language Understanding)의 입장에서 형태소 분석을 수행하면 한국어 의미 단위(가장 작은 단위)인 형태소로 분리된다. 일반적인 형태소 분석 작업은 어휘 분석 작업이 매우 강력하게 이루어져 자칫 후보어의 조사 작업을 어렵게 만들 수 있다. 다시 말하면 분석 프로그램이 신어의 구성 형태소를 너무 잘게 분해하여 사전 표제어에 이미 존재하는 구성요소들로 분리되어 버려 뒤이어 진행되는 사전 매칭 작업에서 기본 단어들로 인식되어 모두 빠져 나갈 수 있다. 따라서 본 논문의 연구 방법에서는 체언과 용언(어절)에서 어미와 조사만을 분리하는 단어 분석 방법[4]을 이용하였다. 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 목록을 제외시켜 신어 조사자가 선택한 단어가 표제어 목록과 기존 신어 목록에 없는 단어만을 대상으로 작업하도록 조사하는 작업이다.

신어 조사용 프로그램 개발을 위해 준비한 지식베이스[4]에 대해 설명하면 다음과 같다. 첫째, 본 연구실에서 지금까지 조사한 1차 신어 후보어를 DB화 할 수 있도록 통일하여 정리하였다. 둘째, 국립국어원에서 제공하는 표

3) 향후에는 명사구와 조사, 파생 접미사와 조사, 의존 명사와 조사, 부사와 조사 등도 추가하여야 한다.

4) 지식베이스로 사용한 목록은 다음과 같이 12가지이다. 열거하면 (1) 전각문자(불필요한 태그) 3, 952개, (2) 표준국어대사전의 표제어 목록 420,957개, (3a) 고유명사 27,309개, (3b) 고유명사의 기술 목록(기초) 167,793개, (4a) 인명 12,693개, (4b) 외래어로 된 인명 8,963개, (5a) 지명 7,373개, (5b) 외래어로 된 지명 9,671개, (6) 개인적으로 조사한 단어 278개, (7) 조사 복합체 9,453개, (8a) 지금까지 추출한 2차 신어 후보어 37,607개, (8b) 올해 조사한 2차 신어 후보어 63,988개, 기존에 조사하여 확정된 기존 신어-2009까지 29,005개, (10) 어미로 구성한 구조체 11,340개, (11) 표준국어대사전의 동사/형용사 목록 58,998 개(“-다”는 는 삭제), 마지막으로, (12) 우리말 샘 연구팀[9]의 신어 분과에서 제안한 배제 단어 목록 30,783 개 등이다.

준국어대사전의 표제어에서 동음이의어를 제외하고 420,957개를 준비하였다. 셋째, 한국어 어미와 조사의 이형태, 계사의 활용형, 계사/어미/조사 등이 결합 가능한 복합체를 조사하였다. 여기서 어미는 1,087개, 조사는 163개, 어미 구조체는 3,221개, 조사 구조체는 5,443개[5]이었다. 본 프로그램은 MS-Windows 7 운영체제에서 Microsoft Visual Studio 2010의 개발 도구를 이용하여 C# 언어로 개발하였다.

3. 결론

본 논문에서는 언론 자료에 나타나는 신어 및 미등재어를 조사/정리하여 체계적으로 관리할 수 있는 프로그램을 개발하였다. 본 논문에서 개발한 프로그램은 언론 자료를 통해 수집된 신어 및 미등재어에 대해 원어, 전문 영역, 뜻풀이, 용례 등을 기술하는 작업자에게 편리하게 작업하도록 설계하였으며, 전체 신어의 유형과 특징을 분석하여 단일어, 파생어, 합성어 등으로 나누어 연구하고, 그 내적 구성 방식의 연구도 가능하게 하는 도구이다. 이 조사 도구를 이용하여 신어의 의미 기술과 용례, 최초 출현일 등을 지식서비스로 구축할 수 있다. 본 논문에서 제안하는 신어 조사 도구는 신문, 방송 등의 언론 기사 전용 수집 웹 로봇의 개발, 신어 후보어 자동 추출 프로그램의 개발, 신어 정리 및 통합 관리 프로그램의 개발 등 세 가지 기능을 갖고 있다. 이 도구를 이용하여 최종적으로 472개의 신어와 2,022개의 미등재어를 조사하였다.

본 연구 결과를 통해 다음과 같이 인터넷 포털 사이트의 지식 서비스 개선 방안들을 생각해 볼 수 있다. 첫째, 기존의 자연언어처리 연구에서 사용하던 KWIC; Key Word In Context의 개념을 기반으로 신어 조사에 적용하여 NWIC; New Word In Context의 개발 방안을 마련할 수 있다. 둘째, 수집된 12만 건의 기사 목록을 조사하여 중복성 문제를 대부분 해결하였으나 여전히 중복 기사가 있을 것으로 추정된다. 따라서 이에 대한 획기적인 연구 방안의 마련이 필요하다. 셋째, 신어를 모니터링 할 수 있는 메뉴의 필요성에 대한 조사 연구를 시작하여야 한다. 넷째, 배제 정보(결코 후보 신어가 될 수 없는 부분적인 문자열)에 추가되어야 할 사항에 대해 조사 및 토의가 필요하다. 다섯째, 기존에 조사되었던 신어 혹은 미등재어의 결과물과 당해연도에 조사된 결과물 간의 표준화 된 저장 방식을 마련할 수 있다. 여섯째, 2013년도 이전의 언론 기사 수집을 통해 신어의 생성, 발전(계속적인 사용), 소멸의 과정 즉, 신어의 추적 조사 작업이 가능함을 입증할 수 있다. 일곱째, 신어의 선정 기준의 정밀한 분석 기준을 필요, 신어/미등재어의 선정을 위한 표준안을 마련할 수 있다. 마지막으로, 신어가 새로 발생하면 신어의 생성 여부를 알려주는 신어 경보기에 대한 연구 개발이 가능함을 보여주었다.

참고문헌

[1] 송 인 성, 정 희 석, 이 상 곤, 이 래 호, "언론 기사에 나타난 신(조)어 조사 도구의 설계 및 구현", 제31회 한국 정보처리학회 춘계 학술대회 논문집, 제16권, 제1호, pp. 114-117, 2009.

[2] 김 동 의, 이 상 곤, "신어를 찾아내고 의미를 기술하여 관리하는 신어 조사용 프로그램의 설계 및 구현", 정보과학회논문지: 소프트웨어 및 응용, 제40권, 제12호, pp. 882-894, 2013.

[3] 이 상 곤, "확장된 벡터 공간 모델을 이용한 한국어 문서 분류 방안", 정보처리학회 논문지(B), 제18권, 제2호, pp. 93-108, 2011.

[4] 안 동 언, 김 재 훈, 남 영 준, 박 혁 로, 이 상 곤 공역, "최신 정보검색론", pp. 001-514, 교보문고, 2010.

[5] 이 상 곤, "신조어 자동 추출 방법론과 신어 조사 도구의 개발", 제21회 한글 및 한국어 정보처리 학술대회 논문집, pp. 271-276, 2009.

[6] 이 래 호, "한국어 교육에서의 신어 교육 방안에 대한 연구", 한국중원언어학회 언어학 연구 20, pp. 155-178, 2011.

[7] 소 강 춘, 이 래 호, 주 경 미, 개방형 한국어 지식 대사전, 신어 분과의 표제어 선정과 그 실제, 한국사전학회 한국사전학 20, pp. 52-85, 2012.