말뭉치에 근거한 한국어 사전 표제어 구성

박영환, 윤준태, 송만석
연세대학교 전산과학과

요 약

사전은 자연어를 처리하는 핵심 부분을 이루고 있다. 그러나 기존의 한국어 사전은 기계적인 처리에 직접 이용하기에는 크게 미흡하다. 특히, 사전의 기본을 이루는 표제어 수록에 관한 연구는 더욱 취약한 형편이다. 본 연구는 새로운 한국어 사전의 표제어를 구성하기 위하여 대형 말뭉치를 수집하였다. 이 말뭉치를 이용하여 기존 사전에서 빠져있는 미등록어들을 찾아내어 수록하고, 말뭉치에 나타난 각 단어의 출현 빈도를 조사하였다. 이 연구를 수행하기 위하여 형태소 분석기, 용례 분석기 등의 필수적인 텍스트 처리 도구들을 개발하였다. 또한, 말뭉치에 나타난 어절 단위의 오류 분포를 조사하여 밝히었다.

I. 서 론

한국어를 기계적으로 처리하려는 대부분의 연구는 사전을 바탕으로 하여 이루어지고 있다. 사전에 기반한 시스템에서는 사전의 효율이 시스템의 성능으로 직접 이어지는 특징을 가지고 있다. 그러나 지금까지 연구들의 대부분은 시스템에 알맞은 이상적인 사전을 가정하면서, 실제로 사용되는 사전은 소규모로 구성하고 적은 데이타로 성능 실험을 하고 있다. 이로 인하여 자연어 처리 시스템에서 중요시 되는 실생활 언어 처리와 동떨어지는 연구를 하게 되고, 스스로 연구의 한계를 규정짓는 잘못을 범하고 있다.

연세대학교 부설 한국어 사전 편찬실(이하 사전 편찬실이라 약칭한다.)은 새로운 한국어 사전의 편찬을 위하여 국내에서 출판된 대형 사전인 "새우리말 큰사전(신기철·신용철)"을 전산 입력하였다. 이 전산 입력 사전을 분석한 연구에 의하면 이 사전은 다음과 같은 몇가지 문제점을 쉽게 발견하게 된다[4].

① 실생활에서 사용되는 단어들의 누락
② 실생활에서 사용되지 않는 단어(특히 한자어)들의 과다 수록
③ 의존형태소 수록 기준의 비일관성

이 문제점은 시중에 출판된 기존의 다른 사전에도 공통적으로 나타나고 있다. 기존 사전의 표제어를 강화하지 않고 시스템 사전으로 직접 사용할 경우 위의 문제점들이 그대로 시스템에 이어져 많은 오류를 남겨주게 된다.

본 연구는 이러한 문제점들을 해결하기 위하여 대형의 말뭉치를 분석하여 그 결과를 사전의 표제어 수록에 반영하는 방법을 사용하였다. 말뭉치를 분석하기 위한 기본 도구로 한국어 형태소 분석기와 용례 분석기를 개발하였다. 말뭉치는 실생활에 기반을

두고 모아진 것이므로 ①, ②번 문제 해결의 근거를 제시한다. 말뭉치를 조사하여 사전에서 빠진 단어는 등록하고(①번 해결), 사용되지 않는 단어는 삭제를 고려할 수 있다(②번 해결). ③번 문제가 가장 두드러지게 나타나고 있는 접미사가 결합된 단어를 말뭉치에서 모두 추출하여 그 해결 방안을 제시하였다.

Ⅱ. 말뭉치 언어학 (Corpus Linguistics)과 형태소 분석기

2.1 말뭉치 언어학의 개요

최근 유럽에서는 발전된 전산기술을 이용하여 대량의 자연 언어를 처리하는 말뭉치 언어학이 발전하고 있다[10]. 이는 전산기를 이용하여 확률적이고 통계적인 방법으로 실생활에서 사용되는 언어를 분석·처리하려는 노력이다.

일반적으로 말뭉치는 한 저자의 저작 전부, 또는 한 특정 분야의 저작 전부를 뜻한다. 즉, 언어학적 의미의 한 말뭉치는 어떤 기준으로든 한 덩어리로 볼 수 있는 말의 뭉치를 가리킨다. 예를 들어 이 논문도 하나의 말뭉치가 될 수 있고, 교과서 한 권도 말뭉치가 될 수 있다.

문법책이나 교과서의 예문들은 그 저자가 알고 있는 문장 규칙을 준수하여 만들어낸 조작적이고 인위적으로 꾸며낸 문장이다. 이 문장들은 잘된(well-formed) 문장은 될 수 있어도 실생활에서 쓰이고 있다는 것을 뜻하는 것은 아니다. 말뭉치 언어학의 목표는 이런 조작성, 인위성을 피하여 언어의 자연 상태를 포착하려는 것이다. 그래서 뭉치 언어학은 실생활과 괴리된 언어 연구를 지양하고 실제로 사용되는 말을 대량으로 수집하여 이를 통계적으로 다루려는 것이다. 말뭉치를 이용한 통계 자료가 신뢰를 받기 위해서는 그 자료가 광범위하면서도 어떤 동질성을 띄어야 한다. 한국어 전체는 확실한 동질성을 가진 굉장히 큰 말뭉치이다. 이를 모두 다룰 통계적인 방법은 없으나, 한국어를 대표할 수 있는 말뭉치를 작성하여 다루는 방법이 유용할 것이다[6].

2.2 연세 말뭉치

사전 편찬실은 말뭉치에 근거한 사전을 편찬하기 위하여 연세 말뭉치 I (300만 마디)과 연세 말뭉치 II (100만 마디)를 구성하여 놓았다[6]. 연세 말뭉치 I을 구성한 자료의 구성 비율은 표 1에 나타내었다. 우리는 연세 말뭉치 I을 대상으로 사전을 강화하기 위한 연구를 수행하였다.

항목	비 율	항목	비 율	항목	비 율
신문	46.28	소설	11.24	수기	8.13
잡지	21.69	취미	6.80	교과서	5.86

표 1. 연세 말뭉치 I의 구성 비율(%)

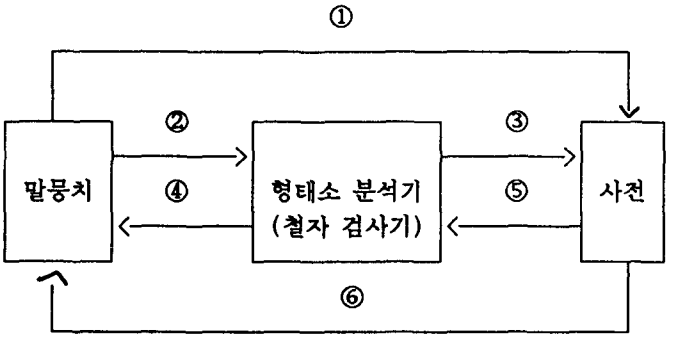
2.3 형태소 분석기

형태소 분석기는 하나의 한국어 어절을 입력으로 받아 각 형태소 단위로 분리하여

출력하는 기능을 수행한다[1]. 전산기를 이용한 형태소 분석기는 사전에 수록된 표제어를 기본 형태소로 가정한다. 즉, "분석기"라는 어절을 형태소 분석기로 해석한다면 "분석(명사)+기(접미사)+는(조사)"의 형태가 아닌 "분석기(명사)+는(조사)"의 출력을 하게 된다. 이는 "분석기"라는 단어가 사전에 수록되어 있기 때문이다.

형태소 분석기를 이용하여 철자 오류의 기능을 수행할 수 있다. 말뭉치 어절 중 형태소 분석이 되지 않는 어절은 철자(띄어쓰기 오류 포함) 오류이거나 사전에 수록되어 있지 않은 형태소를 포함하는 어절이다. 이중 철자 오류 어절을 사용하여 철자 오류의 유형 분포를 구할 수 있다. 또한, 철자 오류 어절을 뺀 나머지 어절을 분석하여 사전에 등록되지 않은 단어들을 찾아낼 수 있다.

또한 형태소 분석기를 이용하여 말뭉치에서 사용된 단어의 빈도도 추출할 수 있다. 즉, 분석기에서 출력되는 형태소를 셈함으로써 각 형태소의 출현 빈도를 조사할 수 있다. 이러한 자료는 많은 사전 표제어 보다는 실제 사용되는 단어로만 사전을 구성하는 것이 필요할 때 적용될 수 있다.



① : 사전 강화 정보 제공　　② : 알고리즘 개선 제시
③ : 사전 구성 정보 제공　　④ : 철자 오류 검색
⑤ : 시스템 기본 사전 제공　　⑥ : 빈도수 셈 기본 사전 제공

그림 1. 말뭉치, 형태소 분석기, 사전의 상호 관계도

Ⅲ. 접미사 결합 단어 정리

3.1 접미사의 분류

기존 한국어 사전을 기계적인 자연어 처리에 이용할때 가장 큰 문제점으로 대두되는 것은 접미사 부분이다.

접미사는 접미사가 결합된 단어의 품사적 성격에 따라 크게 두가지 종류로 나눌 수 있다. 결합된 단어가 체언의 성격을 가지는 체언화 접미사와 용언의 성격을 가지는 용언화 접미사로 나누어진다. 우리는 사전을 한 장씩 확인하여 체언화 접미사 211개와 용언화 접미사 21개를 추출하였다(표 2,3 참조).

표 2. 체언화 접미사 (사:사전 접미사, 규:규칙 접미사)

순서	접미사	종류	용 법	말뭉치에 나타난 예
1	-가	사	①전문인 ②성 ③노래 ④거리 ⑤값 ⑥원자가	①분석가, 소장가 ②김가 ③개국가 ④개울가, 고서점가 ⑤공급가 ⑥
2	-간	규 사	①사이 ②장소	①가부간, 가족간, 개인간 ②방앗간
⋮	⋮	⋮	⋮	⋮
211	-회	사	①그러한 모임	①경제학회, 공동회, 공예가회

표 3. 용언화 접미사 (사:사전 접미사, 규:규칙 접미사)

순서	접미사	종류	용 법	말뭉치에 나타난 예
1	-거리	사	의성·의태어와 결합	가물거리다, 갤갤거리다, 껑중거리다
⋮	⋮	⋮	⋮	⋮
2	-답	규	같다, 그런 가치가 있다	가을답다, 가장답다, 결승답다
⋮	⋮	⋮	⋮	⋮
21	-하	규	하다 동사	공부하다

또한, 접미사를 처리하는 방법에 따라 사전 접미사와 규칙 접미사로 두가지로 크게 분류하였다. 규칙 접미사는 그 앞에 결합하는 형태소가 자유로우며 기본적으로 규칙 처리가 가능하다. 반면 사전 접미사는 앞에 결합할 수 있는 형태소들의 제약 조건이 까다로우며 접미사가 결합된 단어를 모두 사전에 수록하여야 한다.

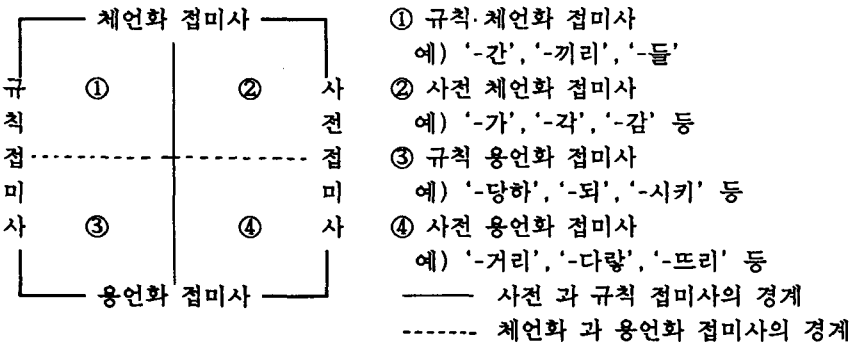


그림 2. 접미사의 종류와 예

기존 사전은 규칙 접미사를 처리하기에는 부족치 않으나, 사전 접미사의 처리에 있어서는 접미사 결합 단어 대부분을 사전에 수록하지 못하고 있다. 이는 사전 제작자들

이 접미사 결합 단어들을 찾을 수 있는 방법이나 대상 말뭉치가 없었기 때문이다. 사전 접미사 단어 미수록 문제는 실제 말뭉치의 분석 결과 커다란 문제로 등장한다 (3.2 절 참조).

사전 접미사와 규칙 접미사의 분류는 말뭉치에 나타난 각 접미사의 실제 예를 대상으로 조사하여 결정하였다.

3.2 접미사 처리 방법

규칙 접미사는 기계적 처리에서 규칙으로 처리하는 것을 원칙으로 하였다. 이 방법을 말뭉치에 적용하여 분석하여 보면 그 타당성을 확인할 수 있다.

사전 접미사는 결합 단어를 찾기 위해 말뭉치에서 아래와 같은 방법을 이용하여 처리를 하였다.

① 패턴 매치로 접미사 결합 가능성이 있는 어절을 추출.
② ①에서 접미사 결합 어절을 추출.
③ 접미사 결합 단어를 생성하여 사전에 수록

사전 체언화 접미사 '-가'와 사전 용언화 접미사 '-다랗'의 ①번 단계의 결과가 예 1,2의 (가)에 나타나 있다. 접미사 결합 가능 어절을 모두 추출하기 위해서는 패턴 매칭 방법을 사용하여야 하였다.

①의 결과를 연구원이 직접 확인하여 접미사가 결합된 어절을 추려내었다. 이 작업은 문맥에서 떨어져 있는 하나의 어절만을 보고 접미사 결합 유무를 확인하기는 불가능하다. 어절이 포함돼 있는 문맥을 보고 결정하기 위해서, 말뭉치를 대상으로 용례 분석 프로그램을 이용하여 문제 어절의 용례를 추출 확인하였다. 예 1,2의 (나)에 용례가 나타나 있다.

②의 작업 과정중 사전의 접미사 용법 중 나타나지 않은 용법도 다수 발견되었다. 예를 들어 사전 체언화 접미사 '-대'의 경우 사전에 나타난 어떤 용법보다도 "대학교"의 약자로 쓰인 경우가 다수였다. 총 2,318 번의 접미사 결합 단어 사용 중 1,758 번이 대학의 약자로 사용되었다.

②의 결과를 바탕으로 접미사가 결합된 형태의 단어를 생성하여 사전에 수록하였다. 이러한 방법으로 체언화 접미사 12,528 개, 용언화 접미사 440 개를 찾아내었다. 이 중 기존 사전에서 누락되었던 체언화 접미사(12,467 개), 용언화 접미사(107 개)의 결합 단어를 사전에 수록하였다. 예 1,2의 (다)는 사전에 수록되는 단어의 예이다. 연세 말뭉치 I에서 한 번 이상 출현했던 기존 사전 표제어 수(체언 37,656 개, 용언 3,840 개)를 감안하면 이는 기존 사전에 얼마나 많은 표제어가 누락되어 있는가를 나타내 주고 있다.

예 1) 사전 체언화 접미사 '-가'의 결합 단어 추출 과정

가도가도 2	각가 2	감은가를 1	개국가 1	개국가의 1
개울가로 5	개울가를 2	개울가에 2	개울가에서 2	거래가의 1
건너가고 1	건너가나 1	건너가는 1	건너가더라도 1	계신가 1
계획가도 1	고서점가를 1	고서점가에서 1	고평가가 1	골동가에 1
골동가에서 1	공급가 1	공시가를 1	공시가와 1	공유가를 1
구리가는 1	구리가의 1	구매가의 1	구연가 1	구입가 1

(가) 접미사 '-가'가 포함된 말뭉치 어절들의 일부분(어절 뒤 숫자는 빈도수).

하나 그 후의 개헌 정국은 가도가도 끝이 없는 험난한 도정이었다
다만, 가도가도 끝간 데가 없을 듯한 광막한
올림픽개최는 1 개국가 1 개도시가 주관하도록 국제 올
조금도 다를 바가 없는 게 개국가의 모습이다
리에서부터 저수지로 올라오는 개울가를 더듬었다
한 번은 토미천 개울가에서 삼 일 염불을 밤낮으로
삼선교에서 안암동으로 흐르는 개울가를 거닐면서 서로 얼굴만 확인하
(1) 소나기황 순원 소년은 개울가에서 소녀를 보자 곧 윤 초시네
다음 날은 좀 늦게 개울가로 나왔다
다음 날부터 좀 더 늦게 개울가로 나왔다
매일같이 개울가로 달려와 봐도 뵈지 않았다
흰 조약돌만 만지작거리며 개울가로 나왔다
낫거들랑 이사 가기 전에 한 번 개울가로 나와 달라는 말을 못해 둔 것

(나) "가도가", "개국가", "개울가"의 용례들

개울가, 거래가, 계획가, 고서점가, 골동가,
공급가, 공시가, 공유가, 구리가, 구매가,
구연가

(다) 접미사 '-가'에 의해 사전에 수록될 단어들

예 2) 사전 용언화 접미사 '-다랗'의 결합 단어 추출 과정

가느다랗게 2 곱다랗게 1 길다랗게 1 높다랗게 3 커다랗게 6
커다랗고 2 커다랗던 1
가느다란 25 것이다란 1 굵다란 2 기다란 5 길다란 2
넓다란 1 높다란 4 다란 1 바다란 3 베다란 1
야마다란 1 얇다란 1 없다란 2 잗다란 1 좁다란 4
짧다란 1 카다란 1 커다란 202

(가) 접미사 '-다랗'이 포함된 말뭉치 어절들의 일부분(어절 뒤 숫자는 빈도수).

남바위 위에다가 중산모를 쓰시고 기다란 지팡이를 끌고 가시는 것이 흡
땅에 잦아들어 뿌리로 변하고 몸은 기다란 줄기로 자라났으며 얼굴은
위인 남천 선생 김익주가 일어서서 기다란 두루마리를 읽어 내려갔다
그는 기다란 노끈에 초를 먹여서 반들반들ㅎ
e 재단을 한다 (사이끄) 5) 헤리 즉 기다란 나무 토막을 붙인다
문 작성 연습을 주로 해왔는데 막상 길다란 지문과 함께 자신의 주장을 솔
말까한데 뒤퉁수에 초승달 모양으로 길다란 흉터가 있었다
졸업반인데도 책임 있는 직책 없이 잗다란 심부름이나 등교 공작에 혹사

(나) "기다란", "길다란", "잗다란"의 용례들

가느다랗다, 곱다랗다, 굵다랗다, 기다랗다, 길다랗다,
넓다랗다, 높다랗다, 잗다랗다, 좁다랗다, 짧다랗다,
커다랗다,

(다) 접미사 '-다랗'에 의해 사전에 수록될 단어들

Ⅳ. 말뭉치 분석 결과

4.1 단어 빈도 조사

개발 된 형태소 분석기를 이용하여 말뭉치를 대상으로 체언류 및 용언류의 단어 빈도 조사를 실시하였다. 이는 연세 말뭉치 I 중 형태소 분석이 실패한 어절과 형태적 중의성을 갖는 어절을 뺀 나머지 1,668,718 개의 어절을 대상으로 한 결과이다[7].

것 37531	이 20700	그 16631	수 12926
등 9510	백 8255	때 6615	김 6613
사람 6394	우리 5898	천 5463	한국 4859
때문 4821	월 4659	또 4453	경우 3901
서울 3810	위해 3768	문제 3732	원 3626
해 3538	씨 3360	중 3277	명 2909
억 2907	개 2865	속 2693	및 2628
년 2589	여 2480	모두 2458	미국 2446
대해 2424	정부 2417	대회 2353	자신 2345
두 2322	사회 2316	내 2262	동안 2261

예 3) 연세 말뭉치 I에 나타난 다빈도 체언류

있 30625	되 12241	없 11222	않 10248
하 7959	같 5841	크 4111	많 3762
못하 3692	말하 2771	다르 2677	좋 2673
받 2490	보이 2465	아니 2414	생각하 2279
오 2126	보 1995	놓 1969	늘 1955
만들 1743	찾 1641	알 1626	따르 1616
밝히 1557	갖 1557	티 1532	이러하 1451
시작하 1437	싶 1388	어렵 1382	나오 1367
버리 1235	새롭 1204	나타나 1187	느끼 1142
맞 1096	주 1069	나가 1062	필요하 1040

예 4) 연세 말뭉치 I에 나타난 다빈도 용언류

4.2 오류 어절 분석 결과

표 4. 형태소 분석 실패 어절의 원인 분석

원 인	갯수	비율(%)
사전 수록	396	14.7
보조 용언	252	9.4
철자 오류	2028	75.7

형태소 분석이 실패한 75,907 개의 어절 종류 중 일부분인 2,676 개의 어절을 분석한 결과 다음과 같은 오류 유형과 형태소 분석 실패 원인을 찾을 수 있었다.

V. 결론

본 연구 결과에 의하면 한국어 기존 사전은 표제어 수록에 있어서 많은 문제점을 지니고 있다. 한국어 사용의 일부분인 연세 말뭉치 I에 기존 사전을 적용한 결과, 사전 표제어의 신뢰성을 의심받게 하고 있다. 사전에 수록되어 있지 않은 접미사 결합 단어의 수치가 실제 사전에서 사용된 표제어 수치의 20 %에 가깝다는 사실이 이를 증명해 주고 있다. 또한, 말뭉치에서 나타난 복합 명사 및 외래어, 신조어 등을 감안한다면 11,000 여개 정도의 새로운 표제어가 실려야 할 것으로 예상된다.

이러한 기존 사전의 문제점이 표제어 수록 부분에서만 나타나는 것이 아님은 최근의 연구 결과 나타나고 있다[4]. 조속한 한국어의 기계적 처리를 위해서는 근본적인 사전 재편찬이 이루어져야 할 것이다. 이는 한국인의 문화 사업일 뿐만 아니라 자연 언어를 기계로 처리하려는 많은 시도에 필수적인 요소가 된다.

기존 사전의 문제점을 발견, 개선할 수 있는 연구는 실생활 언어를 수집한 말뭉치에 의해서 가능하다고 본다. 이 연구는 사전 편찬실에서 새 한국어 편찬을 위해 수집되는 말뭉치의 일부분만으로 이루어졌다. 사전 편찬실은 연세 말뭉치를 3,000만 마디까지 확장할 예정이며, 확장되는 말뭉치를 대상으로 계속적인 연구를 진행해 나갈 것이다.

VI. 참 고

[1] 김대식, "Lexical Database 구축을 위한 어절 분석 도구에 관한 연구", 연세대학교 석사 학위 논문, 1990

[2] 김동찬, "단어 조성론", 북한 어학 자료 총서 314, 고등교육도서출판사, 1987

[3] 남기심, "표준 국어 문법론", 탑출판사, 1990

[4] 남기심외 4인, "한국어 사전의 어휘론적 분석 연구", 국어정보학회, '91 우리말 정보화 잔치, pp.327-335, 1991

[5] 신기철·신용철, "새우리말 큰사전", 삼성출판사, 1978

[6] 정찬섭외 공저, "새 한국어 사전 편찬을 위한 사전 편찬학 연구", 연세대학교 한국어 사전 편찬실, 1990

[7] 최현배, "우리말본", 정음문화사, 1989

[8] 한국어 사전 편찬실, "연세대학교 학술 연구비에 의한 연구 보고서", 한국어 사전 편찬실, 1991

[9] Gerard Salton, "Automatic Text Processing", Addison Wesley, 1989

[10] Using Corpora, "Procedings of the 7'th Conference of the UW Centre for the New OED and Text Research, 1991