한국어 미등록어 인식을 위한 단계별 접근방법[1)]

박소영[*]

[*]상명대학교

Step-by-step Approach for Effective Korean Unknown Word Recognition

So-Young Park[*]

[*]SangMyung University

E-mail : ssoya@smu.ac.kr

요 약

최근 웹 문서 뿐만 아니라 신문기사에서도 미드(미국드라마)나 안습(안구에 습기차다)와 같은 신조어를 사용하고 있다. 그러나, 사전에 등록되지 않은 이러한 단어는 한국어 분석기의 성능을 떨어뜨리는 주요인이 된다. 이러한 미등록어를 자동으로 인식하기 위해서, 본 논문에서는 전문분석 기반 미등록 명사 인식 단계, 웹 출현빈도 기반 미등록 용언 인식 단계, 웹 출현빈도 기반 미등록 명사 인식 단계로 구성된 단계별 접근방법을 제안한다. 제안하는 방법은 문서에서 여러 번 나타난 미등록어를 정확하게 인식할 수 있도록 전문분석 기반 단계를 포함한다. 한편, 문서에 한번 나타난 미등록어도 광범위하게 인식할 수 있도록 웹 출현 빈도 기반 단계도 포함한다. 그리고, 다양한 한국어 미등록어를 인식하기 위해서 미등록 명사 인식 단계와 미등록 용언 인식 단계를 구분한다. 실험결과 기존 접근방법에 비해 제안하는 접근방법은 정확률 1.01%와 재현율 8.50%를 개선하였다.

ABSTRACT

Recently, newspapers as well as web documents include many newly coined words such as “mid”(meaning “American drama” since “mi” means “America” in Korean and “d” refers to the “d” of drama) and “anseup“(meaning “pathetic“ since “an“ and “seup“ literally mean eyeballs and moist respectively). However, these words cause a Korean analyzing system’s performance to decrease. In order to recognize these unknown word automatically, this paper propose a step-by-step approach consisting of an unknown noun recognition phase based on full text analysis, an unknown verb recognition phase based on web document frequency, and an unknown noun recognition phase based on web document frequency. The proposed approach includes the phase based on full text analysis to recognize accurately the unknown words occurred once and again in a document. Also, the proposed approach includes two phases based on web document frequency to recognize broadly the unknown words occurred once in the document. Besides, the proposed model divides between an unknown noun recognition phase and an unknown verb recognition phase to recognize various unknown words. Experimental results shows that the proposed approach improves precision 1.01% and recall 8.50% as compared with a previous approach.

키워드

미등록어 인식, 한국어 처리, 웹 기반 접근방법, 전문분석 기반 접근방법

1) 이 논문은 2008년 정부(교육과학기술부)의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구임(*KRF-2008-531-D00036)*

Ⅰ. 서 론

지금 세계는 인터넷 기술을 바탕으로 산업사회에서 지식정보사회로 빠른 속도로 이동하고 있으며, 이에 따라 지식정보의 양은 급격히 증가하고 있다. 게다가, 지식정보는 항상 새로운 것을 추구하므로, “위키백과(Wikipedia)”와 같이 기존 사전에 포함되어 있지 않은 단어를 사용하여 새로운 지식정보를 표현하기도 한다. 그리고 웹 문서, 블로그, SMS 메시지 등에서는 “훈남(훈훈한 매력이 있는 남자)”, “지름신(충동구매 하게 만드는 신)”과 같이 흥미위주로 새로운 단어를 만들어 사용하기도 한다.

이러한 엄청난 양의 지식정보에서 누구나 쉽게 유용한 최신 정보를 찾을 수 있는 효과적인 지식정보 처리 기술에 관한 연구가 현재 활발히 진행되고 있다. 그러나 대부분의 지식정보 처리 기술은 시스템 개발 당시 확보한 자료를 바탕으로 자원을 구축하므로, 새로운 지식이나 단어에 대해 신속하게 대처할 수 없다는 한계가 있다. 따라서 사전에 등록되지 않은 새로운 단어를 효과적으로 인식할 수 있는 방법이 요구되고 있다.

따라서, 사전에 등록되지 않은 미등록어를 효과적으로 인식하기 위해서, 그동안 다양한 접근방법이 제안되었다. 이들은 크게 언어지식 기반 접근방법, 주변문맥 기반 접근방법, 명사추출 기반 접근방법, 전문분석 기반 접근방법, 웹문서 기반 접근방법이 있다.

첫째, 언어지식 기반 미등록어 인식방법은 형태소 패턴, 어절내 형태소 결합 정보와 같은 언어지식을 바탕으로 미등록어를 인식한다[1,2]. 그러나, 이러한 미등록어 인식방법은 구축한 언어지식에 따라 성능이 크게 좌우될 수 있고, 언어지식의 구축 자체가 쉽지 않다는 문제가 있다[3].

둘째, 주변문맥 기반 미등록어 인식방법은 미등록어 주변에 나타나는 어휘의 통계정보를 바탕으로 미등록어를 인식한다[4,5,6]. 예를 들어, 영어에서는 단어 첫 글자의 대문자 여부, 하이픈 존재 여부, 접두사, 접미사,어미 정보를 사용하여 미등록어를 인식할 수 있다[4,6]. 그러나, 교착어인 한국어에서 기본단위인 형태소는 서로 조합하여 매우 다양한 형태로 어절에서 나타나므로 자료 부족 문제가 심각하게 나타날 수 있다[7].

셋째, 명사추출 기반 미등록어 인식방법은 명사가 나타나는 특성을 고려하여 문서에서 명사를 추출한다[8]. 그러나, 이러한 접근방법은 형태소 기본형이 변하지 않고 어절에 그대로 나타나는 명사만을 추출하므로, “텔레반스럽다”, “텔레반스러운”, “텔레반스럽게”와 같이 형태소의 기본형이 변형하여 나타날 수 있는 용언은 미등록어로 인식할 수 없다.

넷째, 전문분석 기반 미등록어 인식방법은 문서에서 반복적으로 나타나는 문자열을 바탕으로 미등록어로 인식한다[3,9]. 전문분석 기반 접근방법은 비교적 정확한 인식결과를 보여준다. 그러나 미등록어가 문서에서 단 한번만 나타난 경우 미등록어를 제대로 인식할 수 없다[7].

다섯째, 웹문서 기반 미등록 명사 인식방법은 주어진 미등록어 후보와 조사를 조합하여 웹 문서에서 검색하고, 출현빈도가 임계값보다 높으면 미등록 명사로 인식한다[7]. 웹문서 기반 미등록어 인식방법은 대량의 웹문서를 이용하므로 자료 부족 문제를 완화할 수 있다. 그러나, 임계값과 단순히 비교하는 접근방법으로 다소 정교함이 떨어진다. 그리고, 명사추출 기반 미등록어 인식방법과 마찬가지로 미등록 용언을 인식할 수 없다.

본 논문에서 제안하는 접근방법은 언어지식의 구축이 쉽지 않고 관리가 어렵다는 기존 언어지식 기반 접근방법의 단점을 개선하기 위해서, 미등록어 인식을 위한 주변 문맥 규칙을 형태소 분석 말뭉치에서 자동으로 학습하여 사용한다. 그리고, 교착어인 한국어 특성을 고려하여 휴리스틱이 아니라 음소 또는 음절을 바탕으로 적절한 크기의 주변 문맥 규칙을 학습한다. 또한, 제안하는 접근방법은 기존 명사추출 기반 접근방법이나 웹문서 기반 미등록 명사 인식방법과 달리 기본형이 다양한 형태로 변형하여 어절에 나타날 수 있는 미등록 용언도 인식한다. 게다가, 기존 전문분석 기반 접근방법이 문서에 한번만 나타나는 미등록어를 인식하는 것은 불가능하다는 한계를 보완하기 위해서, 제안하는 접근 방법은 문서에 한번만 나타나는 미등록어에 대해서는 대량의 웹문서에서의 출현빈도를 바탕으로 검증한다.

Ⅱ. 한국어 미등록어 인식 방법

한국어 미등록어 인식 방법은 [그림1]과 같이 전문 분석 기반 미등록 명사 인식 단계, 웹 출현빈도 기반 미등록 용언 인식 단계, 웹 출현빈도 기반 미등록 명사 인식 단계를 통해 미등록어를 인식한다. 전문 분석 기반 미등록어 인식 단계에서는 문서내 출현 빈도 및 유형 을 분석하여 미등록어를 인식한다. 반면 웹 출현빈도 기반 미등록어 인식단계에서는 웹 문서에서의 출현빈도 및 유형을 분석하여 미등록어를 인식한다.

이때, 두 번째 단계의 용언 인식에 유용한 어미 리스트와 세 번째 단계의 명사 인식에 유용한 조사 리스트는 형태소 분석 말뭉치에서 학습하여 추출한다. 예를 들어, “-는”은 “들리는”처럼 어미로도 사용되지만 “미드는”처럼 조사로도 사용되므로 추출하지 않는다. 반면, “-ㄴ데”는 대부분 어미로 사용되므로 용언 인식에 유용한 어미로 추출하며, “-를”은 대부분 조사로 사용되므로 명사 인식에 유용한 조사로 추출한다.

첫 번째 단계인 전문 분석 기반 미등록 명사 인식 방법은 기존 전문 분석 기반 접근방법[4,7]과 같다. 먼저, 미등록어를 포함하는 어절들을 가나다순으로 정렬한다. 그리고 앞뒤어절의 음절을 비교하여 2음절이상 동일한 음절열이 있으면 이를 최장 공통 부분문자열로 추출한다[4]. 이렇게

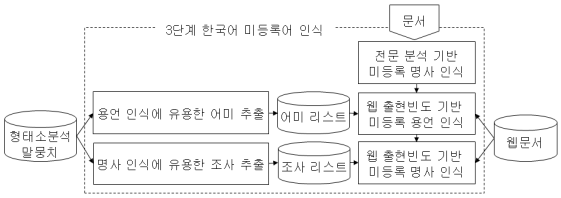

그림 1. 3단계 한국어 미등록어 인식 방법

추출된 최장 공통 부분문자열을 미등록 명사로 인식한다. 예를 들어, 주어진 문서에 나타난 "미드", "미드가", "미드는", "미드를"에서 최장 공통 부분문자열 "미드"를 추출할 수 있다. 따라서, "미드"를 미등록 명사로 인식한다.

두 번째 단계인 웹 출현빈도 기반 미등록 용언 인식 방법은 주어진 어절에 용언 인식에 유용한 어미가 있는지를 확인한다. 해당하는 어미가 있으면 그 어미를 제거한 후 기본형 어말어미 '-다'와 함께 문자열을 생성한다. 그리고 그 문자열을 웹 문서에서 검색하고, 그 출현빈도가 임계값보다 높으면 해당 음절열은 미등록 용언으로 인식한다. 예를 들어, "이뿐데"는 용언 인식에 유용한 어미 "-ㄴ데"를 포함하고 있으므로, "이뿌다"를 웹에서 검색하여 미등록 용언 "이뿌-"를 인식한다.

세 번째 단계인 웹 출현빈도 기반 미등록 명사 인식 방법은 주어진 각 어절에 대해 1음절씩 줄이면서 미등록 명사 후보를 생성한다. 각 후보를 명사 인식에 유용한 조사들과 조합하여 문자열을 구성한다. 그리고, 그 문자열을 모두 웹에서 검색한다. 먼저 모든 조합 문자열의 출현빈도가 임계값보다 높은 경우 해당 음절열은 미등록어로 인식한다. 주어진 미등록 명사 후보에 조사가 없으면서 주요 조사와 조합한 문자열의 웹 문서에서의 출현빈도가 임계값보다 높은 경우 미등록 명사 후보를 명사로 인식한다. 위 두 경우에 해당하지 않으면 1음절을 제거한 후 위과정을 반복한다.

예를 들어, "리오펠의"에 대해 "리오펠의가", "리오펠의를", "리오펠의도", "리오펠의에"를 생성하여 웹에서 검색하면 출현빈도가 0으로 나타난다. 따라서, 1음절을 제거한 "리오펠"에 대해 "리오펠이", "리오펠을", "리오펠도", "리오펠에"를 생성하여 웹에서 검색하면 주요 조사와 결합한 "리오펠이", "리오펠을"만 저빈도로 나타난다. 미등록 명사 후보 "리오펠"이 조사를 포함하지 않으므로,"리오펠"을 미등록 명사로 인식한다.

III. 실험 및 평가

제안하는 미등록어 인식 방법의 성능을 평가하기 위해서, 44개의 신문기사에 나타는 미등록어를 인식하고, [표1]과 같이 정확률, 재현율, F-measure를 사용하여 평가한다. 실험 문서는 총 13,429어절 중 247개의 미등록어를 포함하고 있다.

표 1. 단계별 미등록어 인식 결과

	정확률	재현율	F-measure
전문 분석 기반 미등록 명사 인식	97.01	52.63	68.24
웹출현빈도 기반 미등록 용언 인식	44.44	1.62	3.13
웹출현빈도 기반 미등록 명사 인식	93.27	39.27	55.27
통합	93.52	93.52	93.52

먼저, 전문 분석 기반 미등록 명사 인식 단계는 [표1]과 같이 정확률은 높지만 재현율은 상대적으로 낮다. 이는 미등록어가 주어진 문서에서 여러 형식형태소와 결합하여 다양하게 나타나는 경우 정확하게 인식한다는 것을 보여준다. 그러나, 미등록어의 45.75% 정도가 문서에서 한번만 나타나므로 이들을 제대로 인식하지 못하였다.

다음으로, 웹 출현빈도 기반 미등록 용언 인식 단계는 Google 검색엔진을 바탕으로 용언을 인식한다. 전체 미등록어중 4.86%에 해당하는 미등록 용언을 인식하기 위해 적용되므로 재현율을 특히 낮게 나타났다. 게다가, 용언은 "두터운"(두텁-)과 같이 활용활 때에 어간이나 어미의 기본 형태가 달라져서 정확률도 낮았다.

마지막으로 웹 출현빈도 기반 미등록 명사 인식 단계도 Google 검색엔진을 사용한다. 세 번째 단계의 자체 정확률은 다소 떨어지지만, 문서에서 한번 나타난 39.27%의 미등록 명사를 추가로 인식한다.

표 2. 기존 연구와 비교

	정확도	재현율	F-measure
(김선호2002)	97.01	52.63	68.24
(박소영2008)	92.51	85.02	88.61
제안하는 방법	93.52	93.52	93.52

제안하는 3단계 한국어 미등록어 인식 방법과 기존 접근 방법을 비교하면 [표2]와 같다. 전문 분석 기반 미등록어 인식인 (김선호2002)는 미등록어가 주어진 문서에서 한번만 나타난 경우 적용할 수 없었지만, 제안하는 방법은 웹 문서를 검색할 수 있으므로 재현율이 40.89%나 개선되었다. 한편, (박소영2008)은 미등록 명사만 인식할 수 있지만, 제안하는 방법은 미등록 명사와 함께 미등록 용언도 인식하여 재현율을 다소 개선하였다. 게다가, 제안하는 방법은 (박소영2008)와 달리 주어진 미등록 명사 후보가 조사를 포함하고 있는지 여부를 고려하여 좀더 정교하게 설계하여 정

확률을 1.01% 개선하였고, 이로 인해 재현율도 약 7%정도 올라갔다.

Ⅳ. 결 론

본 논문에서는 한국어 미등록어를 인식하기 위해서 전문분석기반 미등록 명사 인식 단계, 웹 출현빈도 기반 미등록 용언 인식 단계, 웹 출현빈도 기반 미등록 명사 인식 단계로 구성된 3단계 접근방법을 제안하였다. 먼저, 전문 분석 기반 미등록 명사 인식 단계는 한 문서에서 나타난 단어는 한 가지 의미로 사용된다고 가정하고, 문서에서 반복적으로 나타나는 문자열을 미등록어로 정확하게 인식한다. 그리고, 웹 출현빈도 기반 미등록 용언 인식 단계와 웹 출현빈도 기반 미등록 명사 인식 단계는 주어진 어절 뒤쪽의 어미나 조사를 다른 어미나 조사로 바꾼 후 대량의 웹 문서에서 검색하고, 그 출현빈도를 분석하여 미등록어를 인식한다. 실험결과 제안하는 접근 방법은 웹 문서에서의 출현빈도를 바탕으로 미등록 명사와 함께 미등록 용언을 인식하므로, 기존 접근방법에 비해서 8.5% 정도 재현율을 개선하였다.

참고문헌

[1] 이도길, 한국어 형태소 분석과 품사부착을 위한 확률 모형, 고려대학교 박사학위 논문, 2005.

[2] 박봉래, 전문분석에 기반한 한국어 미등록어의 인식, 고려대학교 박사학위 논문, 1999.

[3] David Yarowsky, “Unsupervised Word Sense Disambiguation Rivaling Supervised Methods”, Proceeding on 33rd Annual Meeting of the Association for Computational Linguistics, pp.189 -196, 1995.

[4] 김선호, 윤준태, 송만석, “한국어 문서 처리를 위한 동적 생성 로컬 사전 기반 미등록어 분석”, 정보과학회논문지:소프트웨어 및 응용, 제29권 제6호, 407쪽-416쪽, 2002.

[5] 양장모, 김민정, 권혁철, “언어정보를 이용한 한국어 미등록어 추정”, 한국정보과학회 봄 학술발표논문집, 제23권 제1호, 957쪽-960쪽, 1996.

[6] 차정원, 이원일, 이근배, 이종혁, “형태소 패턴 사전을 이용한 일반화된 미등록어 처리”, 정보과학회 인공지능연구회 춘계학술대회 논문집, 37쪽-42쪽, 1997.

[7] 박소영, “웹문서에서의 출현빈도를 이용한 한국어 미등록어 사전 자동 구축”, 한국컴퓨터정보학회 논문지, 제13권 제3호, 27쪽-33쪽, 2008.

[8] Ralph Weishedel, Marie Meteer, Richard Schwartz, Lance Ramshaw, and Jeff Palmulcci, "Coping with Ambiguity and Unknown Words through Probabilistic Models", Computational Linguistics, Vol.19, No.2, pp.359-382, 1993.

[9] Masaaki Nagata, “Automatic Extraction of New Words from Japanese Texts using Generalized Forward-Backward Search,” Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp.48-59, 1996.