웹문서를 이용한 단계별 한국어 미등록어 인식 모델※

박소영*

Phase-based Model Using Web Documents for Korean Unknown Word Recognition

So-Young Park*

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

요 약

신문이나 블로그와 같은 실제 문서에서는 위키백과(Wikipedia)와 같은 기존에 없던 새로운 단어를 포함하고 있다. 그러나, 대부분의 정보 처리 기술은 시스템 개발 당시 확보한 자료를 바탕으로 사전을 구축하므로, 이러한 새로운 단어에 대해 신속하게 대처할 수 없다는 한계가 있다. 따라서 본 논문에서는 사전에 등록되어 있지 않은 한국어 미등록어를 자동으로 인식하는 모델을 제안한다. 제안하는 모델은 전문분석 기반 미등록명사 인식 단계, 웹 출현빈도 기반 미등록용언 인식 단계, 웹 출현빈도 기반 미등록명사 인식 단계로 구성된다. 제안하는 모델은 문서에서 여러 번 나타난 미등록어에 대해 전문분석을 통해 정확하게 인식할 수 있다. 그리고, 제안하는 모델은 문서에 한번 나타난 미등록어에 대해서도 웹문서를 바탕으로 광범위하게 인식할 수 있다. 또한, 제안하는 모델은 기본형이 어절에 그대로 나타나는 미등록명사뿐만 아니라 기본형이 변형하여 나타날 수 있는 미등록용언도 인식할 수 있다. 실험결과 기존 미등록어 인식방법에 비해 제안하는 접근방법은 정확률 1.01%와 재현율 8.50%를 개선하였다.

ABSTRACT

Recently, real documents such as newspapers as well as blogs include newly coined words such as "Wikipedia". However, most previous information processing technologies cannot deal with these newly coined words because they construct their dictionaries based on materials acquired during system development. In this paper, we propose a model to automatically recognize Korean unknown words excluded from the previously constructed dictionary. The proposed model consists 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 model can recognize accurately the unknown words occurred once and again in a document by the full text analysis. Also, the proposed model can recognize broadly the unknown words occurred once in the document by using web documents. Besides, the proposed model can recognize both a Korean unknown verb, which syllables can be changed from its base form by inflection, and a Korean unknown noun, which syllables are not changed in any eojeol. Experimental results shows that the proposed model improves precision 1.01% and recall 8.50% as compared with a previous model.

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

※ 이 논문은 2009년 한국해양정보통신학회 춘계종합학술대회에서 "한국어 미등록어 인식을 위한 단계별 접근방법"의 제목으로 발표된 논문을 확장한 것임

* 상명대학교 디지털미디어학부

접수일자 2009. 07. 02
심사완료일자 2009. 07. 31

Ⅰ. 서 론

현재 세계는 인터넷 기술을 바탕으로 산업사회에서 지식정보사회로 빠른 속도로 이동하고 있다. 이에 따라 지식정보의 양은 급격히 증가하고 있고, 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. 즉, "위키백과(Wikipedia)"와 같이 새로운 지식정보를 표현하는 단어나 "훈남(훈훈한 매력이 있는 남자)", "지름신(충동구매 하게 만드는 신)"과 같이 흥미 위주로 사용하는 새로운 단어를 효과적으로 인식할 수 있는 방법이 요구되고 있다.

따라서 사전에 등록되지 않은 이러한 미등록어를 효과적으로 인식하기 위해서, 그동안 다양한 접근방법이 제안되었다. 이들은 언어지식 기반 미등록어 인식방법, 주변문맥 기반 미등록어 인식방법, 명사추출 기반 미등록어 인식방법, 전문분석 기반 미등록어 인식방법, 웹문서 기반 미등록어 인식방법이 있다.

첫째, 언어지식 기반 미등록어 인식방법은 형태소 패턴, 어절내 형태소 결합 정보와 같은 언어지식을 바탕으로 미등록어를 인식한다[1,2]. 그러나, 이러한 미등록어 인식방법은 어떻게 언어지식을 구축하는가에 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]. 예를 들어, 영어에서는 단어 첫 글자의 대문자 여부, 하이픈 존재 여부, 접두사, 접미사, 어미 정보를 사용하여 미등록어를 인식할 수 있다[4,5]. 그러나, 교착어인 한국어에서 기본단위인 형태소는 서로 조합하여 매우 다양한 형태로 어절에서 나타나므로 자료 부족 문제가 심각하게 나타날 수 있다[6].

셋째, 명사추출 기반 미등록어 인식방법은 명사가 어절에서 나타나는 특성을 고려하여 문서에서 명사를 추출한다[7]. 그러나, 이러한 접근방법은 형태소 기본형이 변하지 않고 어절에 그대로 나타나는 명사만을 추출하므로, "이뿌게", "이뿐", "이뿐데"와 같이 형태소의 기본형이 변형하여 나타날 수 있는 용언은 미등록어로 인식할 수 없다.

넷째, 전문분석 기반 미등록어 인식방법은 문서에서 반복적으로 나타나는 문자열을 미등록어로 인식한다[3,8]. 전문분석 기반 접근방법은 비교적 정확한 인식결과를 보여준다. 그러나 미등록어가 문서에서 단 한번만 나타난 경우 미등록어를 제대로 인식할 수 없다[6].

다섯째, 웹문서 기반 미등록명사 인식방법은 주어진 미등록어 후보와 조사를 조합하여 웹 문서에서 검색하고, 출현빈도가 임계값보다 높으면 미등록명사로 인식한다[6]. 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 미등록용언을 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 미등록명사 인식방법과 달리 기본형이 다양한 형태로 변형하여 어절에 나타날 수 있는 미등록용언도 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. 단계별 한국어 미등록어 인식

한국어 미등록어 인식은 사전에 등록되어 있지 않지만 신문기사, 웹 문서 등의 실제 문서에서 나타나는 단어와 그 품사를 인식하여 추출하는 것이다. 예를 들어, [그림1]와 같은 신문기사가 주어지면, ①~⑨와 같이 기존 형태소 사전에 등록되지 않은 단어를 미등록명사로 인식하여 추출하는 것이다.

2007년 가을, ①동인문학상이 준비한 화려한 축배는 ②은희경의 손에 쥐어졌다. 종신 심사위원제 도입, 파격적 상금(5000만원)을 내걸고 지난 2000년 새롭게 출범한 ①동인문학상은 해마다 국내 문단에서 최고의 성취를 이룬 작품을 찾아내고 시상함으로써 한국 문학 발전에 기여해 왔다. ①동인문학상 개편 이후 역대 수상작가(③이문구·김훈·④성석제·⑤김연수·김영하·⑥권지예·이혜경)와 이름을 나란히 하게 된 ②은희경 작가는 "떨린다. 나를 진정시켜야 한다"는 말로 수상의 기쁨을 표현했다.

·②은희경은 누구 · 1959년 전북 고창 출생 · 숙명여대 국문과 졸업 · 연세대 대학원 국문학 석사 · 1995년 데뷔 첫해에 장편 '새의 선물'로 문학동네 소설상 · 1997년 소설집 '타인에게 말 걸기'로 ⑦동서문학상 · 1998년 단편 '아내의 상자'로 ⑧이상문학상 · 2000년 단편 '내가 살았던 집'으로 한국소설문학상 · 2006년 장편 '비밀과 거짓말'로 ⑨이산문학상 · 2007년 한국신문협회 선정 '올해의 신문 읽기 스타'

그림 1. 미등록어를 포함하는 신문기사의 일부
Fig. 1 A Piece of Newspaper with Unknown Words

이를 위해, 문서내 모든 어절에 대해 형태소 사전을 바탕으로 형태소 분석[9]을 수행하고, 모든 형태소 분석 후보의 생성 확률이 낮게 분석되는 어절은 미등록어를 포함한다고 가정하여 미등록어 포함 어절리스트에 등록할 수 있다[6]. 그리고, 제안하는 웹 문서를 이용한 단계별 미등록어 인식 모델은 [그림2]처럼 전문분석 기반 미등록명사 인식단계, 웹 출현빈도 기반 미등록용언 인식단계, 웹 출현빈도 기반 미등록명사 인식단계의 순서로 미등록어를 인식한다.

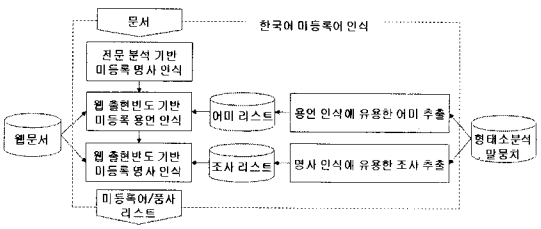

그림 2. 제안하는 한국어 미등록어 인식 모델
Fig. 2 Proposed Korean Unknown Word Recognition Model

첫째, 전문분석 기반 미등록명사 인식 단계는 기존 전문분석 기반 접근방법[6,8]과 같다. 즉, 한 문서에서 나타난 단어들은 한 가지 의미로 사용된다는 경향[3]을 바탕으로 문서에서 반복되는 문자열을 미등록어로 인식할 수 있다. 먼저, [그림3]과 같이 미등록어를 포함하는 어절들을 가나다순으로 정렬한다. 그리고 앞어절의 음절을 비교하여 2음절이상 동일한 음절열이 있으면 이를 최장 공통 부분문자열로 추출한다[6,8]. 이렇게 추출된 최장 공통 부분문자열을 미등록명사로 인식한다.

최장공통 부분문자열 (2음절이상)	미등록어포함 어절리스트 (가나다순 정렬)	
	일치부분	불일치
동인문학상	동인문학상	김연수
	동인문학상	동서문학상
	동인문학상	은
	동인문학상	이
		성석제

그림 3. 전문분석 기반 미등록명사 인식 단계
Fig. 3 Unknown Noun Recognition Phase based on Full Text Analysis

예를 들어, 미등록어 포함 어절리스트에 등록된 어절 "동인문학상", "동인문학상은", "동인문학상이"에 대해 최장 공통 부분문자열인 "동인문학상"을 추출한다. 이렇게 추출된 음절열을 명사로 추정하여 인식한다. 그리고, 해당 어절 "동인문학상", "동인문학상은", "동인문학상이"는 미등록어 포함 어절리스트에서 제거한다.

둘째, 웹 출현빈도 기반 미등록용언 인식 단계는 [그림4]와 같이 주어진 어절에 미등록용언 인식에 유용한 어미가 있는지를 먼저 확인한다. 해당하는 어미가 있으면 그 어미를 제거한 후 기본형 어말어미 '-다'와 함께 문자열을 생성한다. 그리고 그 문자열을 웹문서에서 검색하고, 검색된 출현빈도가 임계값보다 높으면 해당 음절열은 미등록용언으로 인식한다.

예를 들어, 어절 "오바마스럽구나"는 용언 인식에 유용한 어미 "-구나"를 포함하고 있으므로, 어미를 제거한 "오바마스럽-"에 기본형 어말어미 "-다"를 조합한 "오바마스럽다"를 웹에서 검색한다. 검색결과 출현빈도가 임계값보다 높으면 미등록용언 "오바마스럽-"를 인식한다.

명사는 "동인문학상", "동인문학상은", "동인문학상이"의 "동인문학상"처럼 기본형 음절열 그대로 어절에 나타난다. 반면, 용언은 "오바마스러워요", "오바마스러운", "오바마스럽고"의 "오바마스럽-"처럼 기본형 음절열이 변형하여 어절에 나타날 수 있다. 따라서, 웹 출현빈도 기반 미등록용언 인식 단계는 내부적으로 음절이

```
웹 출현빈도 기반 미등록용언 인식( 어절리스트, 어미리스트, 미등록어리스트 )
{
 // 어절 리스트에 미등록어를 추출해야하는 모든 어절에 대해서 while문 수행
 while ( 어절리스트에 분석할 어절 존재 )
 {
  // 어절에 어미가 포함되어 있는지 확인
  while( 어미리스트에 비교할 어미 존재 )
  {
   웹출현빈도 = 0; // 웹 출현빈도 초기화

   // 어절내 뒤부분과 어미가 음소단위로 같으면 어절에 미등록용언 후보가 포함되어 있다고 판단
   if ( 어절 뒤쪽 == 어미 )
   {
    // 미등록용언 후보에 기존 어미 대신 기본형 어미 "다"를 조합하여 웹에서 검색
    문자열 = 어절 - 어미 + "다";
    웹출현빈도 = search( 문자열, 웹검색엔진 Google );
   }

   // 문자열이 웹에 자주 나타나면 미등록어 리스트에 등록
   if ( 웹출현빈도 ≥ 최소허용빈도 )
   {
    insert( 어절 - 어미, "용언", 미등록어리스트 );
    break; // 다음 어절 처리
   }
  } // while( 어미 리스트 ) 종료
 } // while ( 어절 ) 종료
} // 웹 출현빈도 기반 미등록용언 인식 종료
```

그림 4. 웹 출현빈도 기반 미등록용언 인식 단계
Fig. 4 Unknown Verb Recognition Phase based on Web Document Frequency

아닌 음소를 기본단위로 사용한다.

셋째, 웹 출현빈도 기반 미등록명사 인식 방법은 [그림5]와 같이 각 어절에서 1음절씩 줄이면서 미등록명사 후보를 생성한다. 각 후보를 명사 인식에 유용한 조사들과 조합하여 문자열을 구성한다. 그리고, 그 문자열을 모두 웹에서 검색한다. 먼저 모든 조합 문자열의 웹 출현빈도가 임계값보다 높은 경우 해당 음절열은 미등록어로 인식한다[6]. 또한, 일부 문자열의 웹 출현빈도가 임계값보다 높은 경우도 주어진 미등록명사 후보에 조사가 없으면 미등록명사 후보를 명사로 인식한다. 이러한 두 가지 경우에 모두 해당하지 않으면 1음절을 제거한 후 위 과정을 반복한다.

예를 들어, "리오펠의"에 대해 "리오펠의가", "리오펠의를", "리오펠의도", "리오펠의에"를 생성하여 웹에서 검색하면, 출현빈도가 모두 0으로 나타난다. 따라서, 1음절을 제거한 "리오펠"에 대해 "리오펠이", "리오펠을", "리오펠도", "리오펠에"를 생성하여 웹에서 검색하면, "리오펠도"와 "리오펠에"만 출현빈도가 임계값보다 높게 나타난다. 이러한 경우, 미등록명사 후보 "리오펠"이 조사를 포함하지 않으므로, "리오펠"을 미등록명사로 인식한다.

위에서 살펴본 웹 출현빈도 기반 미등록용언 인식 단계에서 사용하는 어미 리스트와 웹 출현빈도 기반 미등록명사 인식 단계에서 사용하는 조사 리스트는 [그림2]와 같이 형태소 분석 말뭉치에서 학습하여 추출한다[6]. 예를 들어, "-는"은 "들리는"처럼 어미로도 사용되지만 "드라마는"처럼 조사로도 사용되므로 추출하지 않는다. 반면, "-ㄴ데"는 대부분 어미로 사용되므로 어미 리스트에 등록하며, "-를"은 대부분 조사로 사용되므로 조사 리스트에 등록한다.

```
웹 출현빈도 기반 미등록명사 인식( 어절리스트, 조사리스트, 미등록어리스트 )
{
 // 어절 리스트에 미등록어를 추출해야하는 모든 어절에 대해서 while문 수행
 while ( 어절리스트에 분석할 어절 존재 )
 {
  // 어절에서 우측 음절을 하나씩 줄여나가면서 for문 수행
  for( 음절열 = 어절; 음절열크기 > 1음절; 새 음절열 = 기존 음절열에서 우측1음절 제거 )
  {
   // 조사 리스트의 모든 조사에 대해서 음절열을 조합하여 웹검색엔진에서 검색
   for( i=0; 조사리스트[i] != 비어있음; i++ )
   {
    문자열 = 음절열 + 문맥규칙;
    웹출현빈도[i] = search( 문자열, 웹검색엔진 Google );
   }

   // 음절열과 모든 조사를 조합한 결과 항상 웹에 자주 나타나면 해당 음절열을 미등록어 리스트에 등록
   if ( ∀i (웹출현빈도[i] ≥ 최소허용빈도) )
   {
    insert( 음절열, "명사", 미등록어리스트 );
    break; // 다음 어절 처리
   }
   // 음절열내 조사가 없고, 일부조사를 조합한 결과가 웹에 자주 나타나면 해당 음절열을 미등록어 리스트에 등록
   else if ( (음절열에 조사가 포함되어 있는가? == false )
        && (웹출현빈도[0] ≥ 최소허용빈도) && (웹출현빈도[1] ≥ 최소허용빈도) )
   {
    insert( 음절열, "명사", 미등록어리스트 );
    break; // 다음 어절 처리
   }
  } // for ( 음절열 ) 종료
 } // while ( 어절 ) 종료
} // 웹 출현빈도 기반 미등록명사 인식 종료
```

그림 5. 웹 출현빈도 기반 미등록명사 인식 단계
Fig. 5 Unknown Noun Recognition Phase based on Web Document Frequency

Ⅲ. 실험 및 평가

제안하는 웹문서를 이용한 단계별 한국어 미등록어 인식 모델의 성능을 평가하기 위해서, 실험집합에 나타난 미등록어에 대해 [표1]과 같이 정확률, 재현율, F-measure를 사용하여 평가한다. 정확률은 모델에서 미등록어를 인식하려고 시도한 경우 중에서 올바르게 인식한 비율을 나타내고, 재현율은 전체 정답 미등록어중 모델에서 올바르게 인식한 비율을 나타낸다. F-measure는 정확률과 재현율을 조화평균을 나타낸다. 웹에서 무작위로 추출한 44개의 신문기사로 구성한 실험집합은 13,429개의 어절을 포함하고 이중 247개의 어절이 미등록어를 포함하고 있다.

실험집합으로 교과서나 블로그 대신 신문기사를 선택한 이유는 다음과 같다. 신문기사는 최신 정보를 전달하는 것을 목표로 하므로, 새로운 단어를 종종 포함한다. 블로그도 새로운 단어를 많이 포함하지만 이와 함께 철자오류도 자주 나타나므로, 모델에서 인식한 미등록어가 철자오류로 인한 것인지 순수한 미등록어인지 판단하는 작업이 부가적으로 필요하다. 반면, 신문기사는 철자오류가 거의 없으므로 기존 사전에 등록되어 있지 않은 단어를 모두 미등록어라고 가정하고 평가해도 크게

문제되지 않는다. 따라서, 철자오류를 고려하지 않고, 효과적으로 미등록어 인식 성능을 평가하기 위해서 신문기사를 실험집합으로 구성한다.

[표1]은 제안하는 한국어 미등록어 인식 모델을 구성하는 전분분석기반 미등록명사 인식 단계, 웹출현빈도 기반 미등록용언 인식 단계, 웹출현빈도 기반 미등록명사 인식 단계의 각 성능을 나타낸다. 그리고, 통합은 이러한 세 단계를 순서대로 모두 수행하였을 때의 성능을 나타낸다.

표 1. 단계별 미등록어 인식 결과
Table 1. Unknown Word Recognition Performance per Phase

	정확률	재현율	F-measure
전문분석 기반 미등록명사 인식	97.01	52.63	68.24
웹 출현빈도 기반 미등록용언 인식	44.44	1.62	3.13
웹 출현빈도 기반 미등록명사 인식	93.27	39.27	55.27
통합	93.52	93.52	93.52

첫째, 전문분석 기반 미등록명사 인식 단계에서는 [표1]과 같이 정확률은 높고 재현율은 상대적으로 낮게 나타났다. 즉, 247개의 어절 중에서 134개 어절에 포함된 미등록어는 주어진 문서에서 여러 형식형태소와 결합하여 두 번 이상 다양하게 나타나 미등록어 인식을 시도하였고, 이들 중 130개를 올바르게 인식하여 52.63%의 재현율과 97.01%의 정확률을 보였다. 반면, 113개 어절에 포함된 미등록어는 문서에서 한 번만 나타나서 이들에 대해서는 인식을 시도할 수 없다는 한계가 있었다.

둘째, 웹 출현빈도 기반 미등록용언 인식 단계는 Google 검색엔진을 이용하여 용언을 인식한다. 첫 번째 단계에서 처리하지 못한 113개 어절 중 9개 어절에 대해 미등록용언 인식을 시도하였으며, 그 중 4개 어절에서 용언을 올바르게 인식하여 44.44%의 정확율과 1.62%의 재현율을 보였다. 이처럼 재현율이 특히 낮게 나타나는데, 이는 247개 어절 중에서 미등록용언을 포함하는 정답 어절 자체가 12개뿐이기 때문이다. 그리고, 미등록용언을 포함하는 3개의 어절에 대해서는 인식 시도조차 하지 않았다. 또한, 정확률도 상대적으로 낮게 나타나는데, 이는 용언 "오바마스럽-"이 어미 "-ㄴ"을 만나 "오바마스러운"으로 활용되듯이, 용언은 어미가 붙어 음절 자체가 달라질 수 있으므로, 인식이 쉽지 않다는 것을 나타낸다.

셋째, 웹 출현빈도 기반 미등록명사 인식 단계도 Google 검색엔진을 사용하여 문서에서 한번 나타난 미등록명사를 인식한다. 위 두 단계에서 처리하지 못한 104개의 어절에 대해 미등록명사 인식을 시도하였으며, 그 중 97개의 어절을 올바르게 인식하여 39.27%의 재현율과 93.27%의 정확률을 보였다. 명사는 조사와 결합해도 기본형이 달라지지 않으므로 정확률이 높게 나타났다.

넷째, 이러한 세 단계를 순서대로 통합하여 수행하면, 각 단계에서 134개, 9개, 104개의 미등록어에 대해 인식을 시도하고, 이중 130개, 4개, 97개의 미등록어를 올바르게 인식한다. 결국, 247개 중 231개의 미등록어를 올바르게 인식하여 93.52%의 재현율과 정확률을 보였다.

표 2. 기존 연구와 비교
Table 2. Comparison with Previous Models

	정확률	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 한국어 미등록어 인식 방법과 기존 접근 방법을 동일한 실험환경에서 비교한 결과는 [표2]와 같다. 전문분석 기반 미등록어 인식인 (김선호2002)는 미등록어가 주어진 문서에서 한번만 나타난 경우 적용할 수 없어서 재현율이 낮게 나타났다. 반면, 제안하는 방법은 웹문서를 검색하여 이용하므로 재현율을 40.89%정도 개선하였다. 한편, (박소영2008)은 미등록명사만 인식하지만, 제안하는 방법은 미등록명사와 함께 미등록용언도 인식하여 재현율을 1.68% 개선하였다. 그리고, 제안하는 방법은 (박소영2008)와 달리 주어진 미등록명사 후보가 조사를 포함하는지를 고려하여 좀 더 정교하게 미등록어를 인식하도록 설계해서 재현율이 6.82% 올라갔다. 즉, 제안하는 모델은 미등록용언 인식단계에서 1.68%, 미등록명사 인식단계에서 6.82%정도 재현율을 개선하여, (박소영2008)에 비해 총 8.5%정도 재현율을 개선하였다.

Ⅳ. 결 론

본 논문에서는 효과적으로 한국어 미등록어를 인식하기 위해서 전문분석 기반 미등록명사 인식 단계, 웹 출현빈도 기반 미등록용언 인식 단계, 웹 출현빈도 기반 미등록명사 인식 단계를 포함하는 단계별 접근방법을 제안하였다.

먼저, 전문분석 기반 미등록명사 인식 단계는 한 문서에서 나타난 단어는 한 가지 의미로 사용된다고 가정하고, 문서에서 반복적으로 나타나는 문자열을 미등록어로 인식한다. 실험결과 미등록어를 97.01% 정도로 정확하게 인식한 반면, 문서에 한번만 나타난 미등록어는 인식하지 못해 재현율은 52.63%정도 낮게 나타났다.

그리고, 웹 출현빈도 기반 미등록용언 인식 단계와 미등록명사 인식 단계는 문서에 한번만 나타난 미등록어 후보에 어미나 조사를 덧붙여 웹 문서에서 검색하여 그 출현빈도를 분석하여 미등록어를 인식한다. 실험결과 웹 출현빈도 기반 미등록용언 인식 단계는 적용대상 미등록용언이 많지 않고 용언이 다양한 형태로 활용하여 나타나므로 정확률과 재현율이 낮게 나타났다. 반면, 웹 출현빈도 기반 미등록명사 인식 단계는 명사는 기본형 그대로 문서에 나타나므로 정확률과 재현율이 93.52%로 높게 나타났다.

이와 같이 제안하는 한국어 미등록어 인식 모델은 자료부족문제로 인해 전문분석 단계에서 인식하지 못한 미등록어에 대해 웹 출현빈도 기반 단계에서 대량의 웹 문서를 활용하여 인식할 수 있다는 특징이 있다. 실험결과 기존 접근방법에 비해서 8.5% 정도 재현율을 개선하였다.

참고문헌

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

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

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

[4] 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.

[5] 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.

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

[7] 이도길, 이상주, 임해창, "명사 출현 특성을 이용한 효율적인 한국어 명사 추출 방법", 정보과학회논문지: 소프트웨어 및 응용, 제30권 제2호, 173쪽-183쪽, 2003.

[8] 김선호, 윤준태, 송만석, "한국어 문서 처리를 위한 동적 생성 로컬 사전 기반 미등록어 분석", 정보과학회 논문지:소프트웨어 및 응용, 제29권 제6호, 407쪽-416쪽, 2002.

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

저자소개

박소영(So-Young Park)

1997년 2월: 상명대학교 전자계산학과(이학사)
1999년 8월: 고려대학교 컴퓨터학과 (이학석사)
2005년 2월: 고려대학교 컴퓨터학과(이학박사)
2007년 3월 ~ 현재: 상명대학교 디지털미디어학부 조교수
※관심분야 : 자연어처리, 기계학습, 텍스트마이닝