확장 정의된 유사어절의 분석에 근거한 실시간 미등록어 인식*

박봉래, 황영숙, 임해창
고려대학교 전산과학과 자연어처리연구실

Real Time Recognition of Unknown Words based on the Analysis of Similar Words with an Extended Definition

Bong-Rae Park, Young-Sook Hwang, Hae-Chang Rim
NLP Lab., Department of Computer Science, Korea University

요약

기존의 미등록어 추정 방법은 대부분 단일 어절 접근 방법으로 단일 어절에서 추출할 수 있는 추정 정보가 부족하여 과분석과 오분석의 가능성이 높았다. 그래서 동일 미등록어를 가진 어절들을 동시에 분석하는 유사 어절 접근 방법이 제시되었다. 그러나 이 방법도 유사 어절의 범위를 조사나 어미만 다른 어절로 정의함으로써 수집될 수 있는 유사 어절의 수가 제한되어 대량의 텍스트에서만 적용이 가능하였다. 이에 본 논문은 유사어절을 동일 음절열을 공유하는 어절들로 확장 정의하여 작은 크기 N의 텍스트 윈도우에서 유사 어절의 발견 가능성을 높임으로써 실시간으로 미등록어를 추정할 수 있게 하는 방법을 제시한다. N을 100으로 한 실험결과는 미등록어 추정 정확도가 99.3%였고 재현율은 약 32%였다.

1. 서론

기존의 미등록어 추정 방법은 대부분 단일 어절 접근 방법이다. 이 방법은 한국어의 어절이 실질 형태소 자체이거나 실질 형태소와 형식 형태소의 결합이라는 특성을 이용하여, 형태소 분석에 실패한 어절에 대해 미지의 실질 형태소를 가정하고 1차로 가능한 모든 분석 후보를 제시한다. 그리고 나서 응용 분야에 따라 1차 분석 후보를 모두 적용하거나[6] 최장 조사 분리 방법으로 하나의 후보를 선정한다[2]. 그런데 <표 1>에 나타난 바와 같이 1차 분석 후보가 여러 개 존재하고 최장 조사 분리 방법만 이용하여 하나의 후보를 선정할 때는 잘못 선정되는 경우가 발생한다. 또한 미등록어를 명사에 국한할 경우에도 다른 품사의 미등록어는 1차 후보로조차 제시되지 않는다.

이러한 과분석 및 오분석의 주요인은 단일 어절에서 미등록어 인식을 위해 추출할 수 있는 정보가 모호하다는 점이다. 형식 형태소 중에 미등록 명사는 단독으로 어절을 구성하기도 하고, 형식 형태소인 조사나 어미의 각 음절들은 기타 다른 단어의 끝 음절과 동일한 경우가 많아 형식형태소를 명확하게 분별할 수 없다.

이러한 문제를 해결하는 방법으로 유사 어절 접근 방법이 제시되었다[1]. 이때의 유사 어절이란 동일한 실질 형태소를 공유하고 서로 다른 형식 형태소를 가진 어절들로서 이를 이용하면 구체적으로 실질형태소를 분별할 수 있기 때문이다. <표 2>에 제시된 예는 <표 1>에 제시된 형태소 실패 어절들에 대한 각각의 유사 어절(괄호로 둘러싸인 부분)을 추가하여 함께 분석한 결과이다. 이번 분석에서는 미등록 용언도 추정하였는데 '출장가'와

* 본 연구는 96년 교육부 학술연구조성비 자유공모과제 "한국어 정보처리 시스템의 전처리를 위한 미등록어 추정 및 철자오류의 자동 교정"의 지원을 받은 것입니다.

<표 1> 단일 어절 접근 방법에 기반한 미등록어 추정의 예

형태소 분석 실패 어절	단일 어절 분석에 근거한 추정		올바른 분석
	1차 분석 후보	선정된 단일 후보	
물깊이	물깊(명사)+이(조사) 물깊이(단독명사)	물깊(명사)+이(조사)	물깊이(단독 명사)
벨기에로만	벨기(명사)+에로만(조사) 벨기에(명사)+로만(조사) 벨기에로(명사)+만(조사) 벨기에로만(단독명사)	벨기(명사)+에로만(조사)	벨기에(명사)+로만(조사)
출장가서	출장가(명사)+서(조사) 출장가서(단독명사)	출장가(명사)+서(조사)	출장가(용언)+서(어미)
밥먹고	밥먹(명사)+고(조사) 밥먹고(단독명사)	밥먹(명사)+고(조사)	밥먹(용언)+고(어미)

<표 2> 유사 어절 접근 방법에 기반한 미등록어 추정의 예

형태소 분석 실패 어절	유사 어절 분석에 근거한 추정	올바른 분석
물깊이 (물깊이가)	물깊이(단독명사) (물깊이(명사) + 가(조사))	물깊이(단독 명사)
벨기에로만 (벨기에까지)	벨기(명사) + 에로만(조사) 벨기에(명사) + 로만(조사) (벨기(명사) + 에까지(조사)) (벨기에(명사) + 까지(조사))	벨기에(명사) + 로만(조사)
출장가서 (출장가면)	출장가(용언) + 서(어미) (출장가(용언) + 면(어미))	출장가(용언) + 서(어미)
밥먹고 (밥먹을)	밥먹(용언) + 고(어미) (밥먹(용언) + 을(어미))	밥먹(용언) + 고(어미)

'밥먹'이 용언의 어간으로 분석된 이유는 어미 '서'와 '면' 그리고 어미 '고'와 '을'이 동시에 붙을 수 있는 실질 형태소는 용언뿐이기 때문이다[5].

이와 같이 유사 어절 접근 방법은 실질 형태소를 명확하게 분리해 내는 장점을 가지고 있어 유사 어절이 존재하기만 하면 과분석과 오분석 문제를 해결하는 데에 크게 도움이 된다. 그러나 기존의 유사 어절 접근 방법은 대량의 텍스트에서 일괄처리를 통해 적용할 때 만족할 만한 효과가 있었고 작은 양의 텍스트에서는 유사어절의 획득이 어렵기 때문에 유사 어절 접근 방법을 시도할 수 없었다.

작은 양의 텍스트에서 유사 어절의 획득이 어려웠던 이유는 유사 어절의 범위를 조사나 어미만 다른 어절로 제한하였기 때문이다. 실질 형태소를 명확하게 분리하는 데에서 조사나 어미만이 구별 근거가 되지는 않는다. 실질 형태소가 동일하지 않아도 문제의 미등록어 부분을 제외하고 주변의 다른 형태소들이 명확하게 구별될 경우에 이 미등록어를 추출하여 각각의 실질 형태소를 분리해낼 수 있다. 예를 들어 <표 3>에 있는 신문기사 중의 어절 '한보주택'과 어절 '한보문제는'은 미등록 단어 '한보'로 인해서 형태소 분석에 실패한다. 그러나 '한보'와 결합한 '주택'과 '문제는'이 각각 단독 명사와 명사와 조사의 결합 형태임이 분명함으로 복합명사 생성 원리에 근거해서 '한보'를 '한보주택'과 '한보문제'의 공통되는 구성 명사로 추정할 수 있다. 이와 같이 유사 어절의 범위를 확장하여 정의하였을 때, 100여개의 어절들로 구성된 신문기사 <표 3>에서 우리는 32개의 유사 어절들을 발견할 수 있다. 반면 확장 정의되지 않은 유사

<표 3> 신문 기사의 일부

재판부는 한보주택 혹은 주거래은행인 조흥은행등 관계인이나 관계법인등에 대한 서면조사등을 통해 본안에 해당하는 회사정리개시여부를 심리할 필요성이 있다고 판단할 경우 보전처분결정을 내려 회사재산의 처분이나 채무의 변제를 금지시키게 된다.

또 회사재산의 강제경매나 임의경매 혹은 조세체납처분의 중지명령도 내려지게 돼 한보주택은 일단 도산의 위험에서 벗어날 수 있게 된다. 통상적인 경우 재산보전 처분결정은 신청후 빠르면 2-3일 늦을 경우 2개월여까지 지체되기도 한다.

재판부가 "한보문제는 사회적 관심도와 파장등을 고려, 여타의 사건에 우선, 집중적인 심리를 하고 있다"고 밝혀 보전처분 결정은 금주안으로 결정될 전망이다.

재판부는 이 결정 이후에는 대표이사나 관계인을 심문하거나 조사위원을 선임하여 정리절차개시여부에 대한 조사를 마친 뒤 법정관리 여부를 최종 판단하게 되며 기간은 보통 5-6개월이 걸리게 된다.

어절만을 고려하였을 때는 8개의 유사 어절만을 발견할 수 있다.

본 논문은 작은 양의 텍스트에서도 많은 유사 어절들을 획득할 수 있도록 유사 어절 범위를 확장 정의한다. 그리고 작은 크기의 텍스트 윈도우내 유사 어절들을 동일 음절열에 의해 색인하는 방법을 다루고 이를 실시간 미등록어 추정에 적용한 실험 결과를 제시한다.

2. 유사 어절 추출

확장 정의된 유사 어절을 실시간으로 추출하기 위한 첫단계는 길이 2이상 L이하의 음절열을 이용하여 일정 크기의 텍스트 윈도우내 각 어절을 색인하는 것이다. 즉, 동일한 색인용 음절열을 가진 어절들을 발견하여 유사 어절을 추출하는 과정이다. L값이 3인 경우 다음 (예 1)에 나타난 어절들은 동일 음절열 '**노래방**'에 의해 똑같이 색인되는 유사 어절들이다.

(예 1) **노래방**에서, **노래방**을, **노래방**시설은, 패꼬리**노래방**, ...

위 방법을 텍스트 윈도우에 적용하여 실시간에 유사 어절을 추출하기 위해서는 두 가지 변수값을 적합하게 설정하여야 한다. 그중 하나는 음절열의 최대 길이(L)이며, 또 하나는 대상 텍스트 윈도우의 크기(N)이다.

동일 음절열의 최대 길이 L의 값은 가장 적합한 유사 어절을 추출하는 데에 중요한 변수이다. 다음 (예 2)에서 L값이 2일 때 세 어절 모두 유사 어절이 된다. 하지만 어절 '체육고교'는 나머지 두 어절의 미등록어 '**체육인**' 또는 '**체육인협회**'를 인식하는 데에 별로 도움이 안된다. 반면에 L이 3이상일 경우 유사 어절은 어절 '**한국체육인협회는**'과 어절 '**체육인협회가**'가 우선 추출되어 효율적으로 미등록어 '**체육인협회**'와 '**한국체육인협회**'를 추정할 수 있다. 물론 이때 모호성이 발견되어 미등록어 분별이 어려울 때 더 작은 길이의 음절열을 고려한 유사 어절들을 분석할 수도 있다. 따라서 L값이 클수록 좋지만 시스템의 조건상 일정 값이하로 제한할 수밖에 없다.

(예 2) 한국**체육인협회**는, **체육인협회**가, 체육고교로

유사 어절의 양을 결정하는 텍스트 윈도우 크기(N)도 중요한 변수이다. N값이 작을수록 유사 어절이 나타날 가능성은 낮아지고 클수록 높아지는데, 너무 클 경우에는 실시간 처리의 장애 요인이 될 것이므로 적당한 N의 설정이 필요하다[1].

<표 4>는 크기 N의 윈도우를 스캔하면서 실시간으로 윈도우내에 위치한 유사 어절들을 추출하는 방법을 기술한 알고리즘이고, <표

<표 4> 유사 어절 추출 알고리즘

```
While each eojeol(어절) Enew in text
 begin
  IF text window(텍스트 윈도우) is not full Then
   begin
    Input eojeol(어절) Enew in text window(텍스트 윈도우)
    Index eojeol(어절) Enew by substrings(음절열) of eojeol(어절) Enew
   end
  Else
   begin
    Output the first inputed eojeol(어절) Eold from text window(텍스트 윈도우)
    Input eojeol(어절) Enew in text window(텍스트 윈도우)
    Index eojeol(어절) Enew by substrings(음절열) of eojeol(어절) Enew itself
    Delete the substrings(음절열) of the eojeol(어절) Eold
    Extract the similar words(유사어절) of the outputed eojeol(어절) by refering
     to the index list(색인리스트)
   end
 end
```

<표 5> 유사 어절 추출 과정의 텍스트 윈도우와 색인 리스트 상태 변화 예

어절 '리콜된'의 유사어절 추출 전				어절 '리콜된'의 유사어절 추출 후			
텍스트 윈도우		색인 리스트		텍스트 윈도우		색인 리스트	
번호	어절열	음절열	색인된 어절	번호	어절열	음절열	색인된 어절
①	리콜된	동차	② ④	②	자동차를	동차	② ④
②	자동차를	리콜	① ④	③	회사는	리콜	④
③	회사는	사는	③	④	자동차리콜	사는	③
④	자동차리콜	원칙	⑤	⑤	원칙에	원칙	⑤
⑤	원칙에	자동	② ④	①	준하여	자동	② ④
		차를	②			준하	①
		차리	④			차를	②
		칙에	⑤			차리	④
		콜된	①			칙에	⑤
		회사	③			하여	①
						회사	③

5>는 이 알고리즘을 가상의 텍스트 **"이렇게 리콜된 자동차를 회사는 자동차리콜 원칙에 준하여 ..."**에 실제로 적용한 한 과정이다. 텍스트 윈도우 크기(N)를 5개의 어절로 제한하고 음절열의 길이도 2음절만을 고려한 간단한 경우로서, 어절 **'준하여'**가 입력되고 어절 **'리콜된'**이 출력되는 전후 과정의 텍스트 윈도우와 색인 리스트의 변화를 보여주고 있다. 어절 **'리콜된'**이 출력되면서 색인 리스트에서 동일 음절열을 가지고 있는 어절 **'자동차리콜'**이 유사 어절로서 함께 출력된다.

3. 실시간 미등록어 인식

형태소 분석에 실패한 어절에 대해 유사 어절들이 추출되었으면, 이들을 서로 비교하여 미등록어를 인식한다. 미등록어가 명사일 경우에 가능한 유사 어절들을 다음과 같이 16가지로 분류해 볼 수 있다. 동일 음절열을 X라

1. 본 논문에서는 N값을 100으로 실험하였다. S를 100으로 선택한 이유는 100어절은 보통 초록의 분량으로서 일정한 의미를 전달하는 데에 충분할 것이라는 가정과 실험 환경인 PC의 메모리 사정을 고려하여 결정하였다.

<표 6> 미등록어 인식을 위한 유사 어절 비교 규칙

출력 어절 유형	유사 어절 유형	규칙(추정 조건)	추정 명사 형태
X	X	-	-
	H′X	-	-
	XT′	T′가 최장 조사 또는 명사 시작 어절	X
	H′XT′		X와 HX
HX	X	-	-
	H′X	-	-
	XT′	T′가 최장 조사 또는 명사 시작 어절	HX와 X
	H′XT′		HX와 H′X
XT	X	T가 최장 조사 또는 명사 시작 어절	X
	H′X		X와 H′X
	XT′	T와 T′가 최장 조사, 또는 T는 최장 조사 그리고 T′는 명사 시작 어절, 또는 T가 명사 시작 어절 그리고 T′가 최장 조사, 또는 T와 T′가 명사 시작 어절	X
	H′XT′		X와 H′X
HXT	X	T가 최장 조사 또는 명사 시작 어절	HX와 X
	H′X		HX와 H′X
	XT′	T와 T′가 최장 조사, 또는 T는 최장 조사 그리고 T′는 명사 시작 어절, 또는 T가 명사 시작 어절 그리고 T′가 최장 조사, 또는 T와 T′가 명사 시작 어절	HX와 X
	H′XT′		HX와 H′X

<표 7> 미등록어 인식의 유형별 예

출력 어절 유형별 예	유사 어절 유형별 예	규칙(추정 조건)	추정 명사
변호인	변호인	-	-
	국선변호인	-	-
	변호인으로	'으로'가 최장 조사	변호인
	국선변호인제도	'제도'가 명사	변호인 및 국선변호인
법조항	조항	-	-
	헌법조항	-	-
	조항만	'만'이 최장 조사	법조항 및 조항
	현조항을	'을'이 최장 조사	법조항 및 현조항
기무사가	기무사	'가'가 최장 조사	기무사
	국군기무사		기무사 및 국군기무사
	기무사를	'가'와 '를'이 최장 조사,	기무사
	국군기무사사령관이	'가'가 최장 조사이고 '사령관' 명사	기무사 및 국군기무사
신골프장 시설	골프장	'시설'이 최장 조사	신골프장 및 골프장
	컨츄리골프장		신골프장 및 컨츄리골프장
	골프장공사	'시설'과 '공사'가 각각 명사	신골프장 및 골프장
	동아골프장에서	'시설'이 명사이고 '에서'가 최장 조사	신골프장 및 동아골프장

하고, X 앞에 붙는 명사나 접두사를 H라 하며, X 뒤에 붙는 명사, 접미사, 조사 또는 이들의 복합형태를 T라 할 때, X가 미등록어일 경우 X에 의해 생성되는 가능한 어절의 유형은 X, HX, XT, HXT의 네 가지이다. 따라서 각각의 경우에 대한 유사 어절의 유형 네 가

지를 고려할 때 가능한 조합수는 모두 16가지이며, 이들에 대한 처리 규칙은 <표 6>에 제시되어 있다. 유사 어절 수가 하나 이상인 경우, 가장 긴 공통 음절열을 가진 어절부터 적용하고 명확하게 분석되지 않으면 다음으로 작은 음절열을 공유한 어절을 적용한다. 한편 인식된 미등록어들은 임시사전에 저장하여 두고 이들을 포함하는 어절이 발견되면 유사 어절이 없더라도 인식할 수 있도록 한다. <표 6>에서 최장 조사의 여부를 확인하는 이유는 동일 음절열의 끝부분이 최장 복합조사의 앞부분일 가능성 때문이다. <표 7>은 <표 6>에 나타난 유사 어절 유형 각각에 대해 예를 든 것으로 적용되는 규칙과 그 결과로 추정되는 미등록어들이 제시되어 있다.

4. 실험 및 평가

87,130개의 어절로 구성된 코퍼스의 각 텍스트에 대해서 텍스트 윈도우 크기 N을 100으로 하고 최대 동일 음절 길이 L을 6으로 하여 실험한 결과가 <표 8>에 제시되어 있고 이것에 대한 평가가 <표 9>에 제시되어 있다. <표 9>의 미등록어 인식 재현율은 정확한 미등록어의 수를 모르기 때문에 형태소 재분석 성공 어절 수의 비율을 이용하여 추정하였다. 또한 실험 평가는 형태소 분석 실패 어절이 모두 미등록어에 기인한다는 가정하에 실시되었다. 그러나 실제로는 형태소 분석기의 오류나 텍스트 자체내에 존재하는 철자 및 띄어쓰기 오류로 인해 형태소 분석에 실패하는 경우도 있음을 감안해야 하다.

N의 값을 100으로만 설정하여도 유사 어절 획득율이 52.6%로 높은 편이다. 그러나 실제에 있어서 유사 어절 활용도는 59.6%에 지나지 않으므로 아직도 개선해야 할 점이 많이 남아 있다.

<표 8> 실험 결과

형태소분석 실패 어절		인식 미등록어		형태소 재분석 성공 어절
유사 어절 존재	유사 어절 미존재	정인식	오인식	
4238		405		1,329
2,231	2007	402	3	

<표 9> 실험 평가

유사 어절	미등록어 인식		유사 어절
획득율	정확도	재현율	활용도
52.6%	99.3%	32%	59.6%

※ 유사 어절 획득율 = $\frac{\text{유사 어절 존재}}{\text{형태소 분석 실패 어절}} * 100$

※ 미등록어 인식 정확도 = $\frac{\text{정인식}}{\text{인식 미등록어}} * 100$

※ 미등록어 인식 재현율 ≒ $\frac{\text{형태소 재분석 성공 어절}}{\text{형태소 분석 실패 어절}} * 100$

※ 유사 어절 활용도 = $\frac{\text{형태소 재분석 성공 어절}}{\text{유사 어절 존재}} * 100$

5. 결론 및 향후 연구

본 논문에서는 미등록어 인식을 위한 유사 어절 방법을 개선하여 유사 어절 범위를 확장 정의하고 이를 이용하여 실시간에 미등록어를 인식하는 방법을 제시하였다. 텍스트 윈도우 크기 N을 100으로 하여 실험한 결과 형태소 분석 실패 어절들 중 절반에 대해 유사 어절을 발견하였고 이들을 분석하여 약 99%의 정확도로 미등록어들을 인식하였다. 인식된 미등록어들로 형태소 분석 실패 어절들을 재분석한 결과 전체의 32%가 분석에 성공하였고 이것으로 재현율을 대신하였다. 물론 윈도우 크기를 키우면 재현율도 높아질 것이다.

본 논문의 실시간 유사 어절 추출 방법은 미등록어 인식뿐만 아니라 형태소 분석 결과의 중의성 해결에도 이용할 수 있다. 형태소 분석 결과의 중의성에는 4가지의 유형이 있는데[4], 이들 중 결합 방식에 따른 중의성의 해결에 도움이 된다. 예를 들어 '**소라도**'라는 어절은 명사 '**소**'와 조사 '**라도**'의 결합 형태와 명사 '**소라**'와 조사 '**도**'의 결합 형태로 두 가지 분석이 가능하다. 하지만 근처에 이 어절의 유사 어절로 '**소라를**'이 존재한다면 어절 '소라도'는 명사 '소라'와 조사 '도'로 단일하게 분석해도 될 것이다.

또한 이 방법은 복합 명사 중의성 해결의 경우에도 적용할 수 있다. 예를 들어 복합 명사 '**대학생선교회**'가 명사 '**대학생**'과 명사 '**선교회**'의 결합 형태와 명사 '**대학**', '**생선**',

'교회'의 결합 형태로 두 가지 분석이 가능하지만[3], "이번 대학생선교회의 선교활동은 어느 때보다 ..."와 같은 문장처럼 주변에 '선교활동은'과 같은 유사 어절이 존재한다면 적어도 '선교'가 분리되지 않는 쪽으로 단일하게 분석할 수 있을 것이다.

앞으로 현재의 높은 정확도를 유지하면서 더 높은 재현율을 얻기 위해 1음절 명사나 서술격 조사 '이'와 동사화 접미사 '하', '되', '시키' 등을 포함하는 고빈도 접미사가 미등록어 뒤에 오는 경우도 연구할 계획이다.

참고문헌

[1] 박봉래, 황영숙, 임해창, "유사 어절의 TAIL 패턴 분석에 기반한 미등록 명사 추정," *정보과학회 봄 학술발표논문지*, pp907-910, 1996.

[2] 양장모, 김민정, 권혁철, "언어 정보를 이용한 한국어 미등록어 추정," *정보과학회 봄 학술발표논문지*, pp957-960, 1996.

[3] 윤보현, 임희석, 임해창, "통계정보를 이용한 한국어 복합명사의 분석방법," *정보과학회 봄 학술발표논문집*, pp925-928, 1995.

[4] 임희석, "어절의 중의성 유형 분류에 근거한 한국어 형태소 분석기," *고려대학교 전산과학과 석사학위 논문*, pp18-19, 1994.

[5] 조규빈, *고교문법*, 지학사, 1993.

[6] Ralph Weischedel, Marie Meteer, Richard Schwartz, Lance Ramshaw, "Coming with Ambiguity and Unknown Words through Probabilistic Models," *Association for Computational Linguistics*, 1993.