JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 20, No. 4, pp. 843-850, Apr. 2019

접미 형태소열 기반 미등록어 인식 확률 모델

한 경 수* · 홍 혁 준 · 지 한 솔
성결대학교 컴퓨터공학과

Probabilistic Model for Unknown Word Identification Based on Suffix Morpheme Sequences

Kyoung-Soo Han* · **Hyuck-Jun Hong** · **Hansol Ji**
Department of Computer Engineering, Sungkyul University, Anyang-si 14097, Korea

[요 약]

인간의 언어는 생산성이 높아 고유명사, 외래어, 전문용어 등 다양한 새로운 단어들이 생성되고 있다. 기존의 말뭉치, 사전 등이 이런 모든 단어를 미리 포함하고 있을 수 없으므로 효과적인 자연어처리를 위해서는 미등록어 인식이 필수적이다. 본 연구에서는 미등록어 인식과 관련된 기존 연구들에서 활용되어 왔던 다양한 정보들을 포괄할 수 있는 확률 모델을 제안한다. 제안하는 모델은 명사 뒤에 등장하는 형식 형태소열의 특징, 이들과 미등록어 후보 간의 결합 관계, 미등록어 후보의 웹 출현 빈도, 미등록어 후보의 처리 대상 문서에서의 출현 빈도 등을 활용하여 후보 단어가 미등록어일 확률을 추정한다. 일반 뉴스와 경제 뉴스에서 실험한 결과 제안한 모델이 우수한 성능을 보였다.

[Abstract]

Human languages are highly productive, and new words such as proper nouns, foreign words, and technical terms are being generated. Since existing corpus and dictionaries cannot contain all these words in advance, it is essential to recognize unknown words for efficient processing of natural language. In this paper, we propose a probabilistic model that can cover various information that has been used in previous researches related to unknown word identification. The proposed model estimates the probability that candidate words are unknown words by using the features of the functional morpheme sequences appearing after the noun, the linkage between these and the unknown word candidates, the frequency of unknown word candidates on the web, and the occurrence of unknown word candidates in the document to be processed. Experiments on general news and economic news show that the proposed model has good performance.

색인어 : 미등록어, 접미 형태소열, 지표 형태소, 공기 빈도, 확률 모델

Key word : Unknown word, Suffix morpheme sequence, Barometer morpheme, Co-occurrence, Probabilistic model

http://dx.doi.org/10.9728/dcs.2019.20.4.843

Received 25 February 2019 ; **Revised** 13 April 2019
Accepted 26 April 2019

***Corresponding Author; Kyoung-Soo Han**

Tel: +82-31-467-8189
E-mail: kshan@sungkyul.ac.kr

Ⅰ. 서 론

인간은 언어를 사용하여 각종 정보를 표현하고 이들을 서로 공유하며 지식을 넓혀 나간다. 인공지능이 인간 수준의 지능을 확보하고 지속적으로 업데이트해가기 위해서는 인간의 언어로 표현된 정보들을 습득할 수 있어야 한다. 이를 위해 자연언어처리(natural language processing) 분야의 연구가 활발히 진행되고 있다. 자연언어처리 기술을 통해 인간 언어를 처리하기 위해서는 인간들이 기존에 어떤 종류의 단어를 어떤 방식으로 사용해 오고 있는지에 대한 경험적인 언어 지식이 필요하며, 이 지식들은 각종 사전, 말뭉치, 규칙 등으로 표현된다.

그런데 인간이 사용하는 언어는 생산성이 높다. 즉, 신조어, 고유명사, 외래어, 전문용어, 복합명사 등 다양한 새로운 단어들이 생성되고 있다. 이런 상황에서 사전, 말뭉치, 규칙 등이 새로이 생겨나는 단어들을 모두 포함할 수는 없게 된다. 결국 기존의 언어 지식에 포함되지 않는 단어가 존재할 수밖에 없으며, 이런 단어들을 미등록어라고 한다. 사회가 전문화되고 다양화될수록 미등록어 문제는 더욱 심각해지고 있으며, 미등록어가 포함된 정보를 강건하게 처리할 수 있는 기술이 중요하게 되었다.

미등록어의 대부분은 명사이다[1]. 미등록 명사를 인식하기 위하여 대부분의 기존 연구들은 한 어절 내에서 명사 뒤에 출현하는 형태소 정보를 활용하였다. 이 형태소들을 명사 접미 형태소로 칭하기로 한다. [2]에서는 조사 사전을 바탕으로 형태소 분석에 실패한 어절에서 최장 조사를 떼어낸 부분을 미등록어로 인식하였다.

[3]은 어절이라는 분석 범위를 탈피하여 문서나 문서집합에서 반복적으로 등장하는 유사 어절들을 비교함으로써 미등록어를 인식하였다. 사전에 등록되지 않은 형태소라도 문서에서 반복적으로 등장한 것들을 미등록어로 인식하는 방법이다. [4]는 형태소 분석 말뭉치로부터 미리 구축해둔 명사 접미 형태소열 사전을 이용하여 규칙에 따라 어절을 미등록어 후보 명사와 명사 접미 형태소열로 구분한 후, 명사 사전에 등록되어 있지 않은 후보 명사를 미등록어로 인식하였다. [5]는 어절 내의 형태소 결합 관계나 문서 내에서 중복 출현 정보이외에 좌우 어절과의 연관관계를 이용하였다.

[6], [7]은 웹 문서에서의 출현빈도를 활용함으로써 문서에서 반복해서 등장하는 미등록어뿐만 아니라 단 한번 등장한 미등록어까지 인식하는 방법을 제안하였다. 형태소 분석 말뭉치로부터 각각 추출된 명사 및 용언 인식에 유용한 조사 및 어미 리스트와 미등록어 후보를 함께 웹 검색을 수행하여 그 결과로 후보의 적절성을 검증하였다.

대부분의 기존 연구들은 다양한 휴리스틱에 따라 미등록어를 인식하는 방법을 제안하였으나, 확률에 기반한 미등록어 인식 모델에 관한 연구들이 있었다. [8]은 말뭉치로부터 동일한 명사와 결합하는 명사 접미 형태소 패턴을 학습하고, 미등록어 후보가 명사일 확률을 계산하여 임계값 이상의 후보를 미등록어로 인식하였다. 이 연구에서는 단지 접미 형태소열 S의 명사 추정 확률 $P(N=1|S)$만을 사용하였다. [9]에서는 어절 간의 최장 공통 앞 문자열(LCP, longest common prefix)에 대해 LCP와 명사 접미 형태소열 사이의 분할 확률을 이용하여 미등록어를 추출하였다. 분할 확률은 명사 접미 형태소열 S가 주어졌을 때 직전 음절이 미등록어 LCP U의 마지막 음절 C_U일 확률 $P(C_U|N=1, S)$로 계산되었다. [1]은 확률기반의 미등록어 분리 및 태깅 모델을 제안하였다. 명사 접미 형태소들을 형태소 분석 말뭉치에서 추출하여, 미등록어 U의 확률 $P(U)$와 미등록어로부터 접미 형태소로의 전이확률 $P(S|U)$를 사용하여 미등록어를 인식하였다. 그러나 사전이나 말뭉치에 등장하지 않은 미등록어의 사전 확률 $P(U)$를 추정하거나, 알려지지 않은 미등록어에 대한 조건부 확률 $P(S|U)$를 적절히 추정하는 것은 쉽지 않아 보인다.

표 1. 어절 내 명사열의 개수
Table 1. Number of NS(noun sequences) in a Eojeol

Number of NS	Number of Eojeol	Ratio
1	4614519	99.9%
≥ 2	4958	0.1%
Total(≥ 1)	4619477	100.0%

본 연구는 기존 미등록어 인식 연구들에서 활용했던 미등록어 관련 다양한 정보들을 하나로 포괄할 수 있도록 미등록어 인식을 위한 확률 모델을 제안하고자 한다.

Ⅱ. 연구 방법

2-1 명사 포함 어절의 특징

[1]에 따르면 미등록어의 93.2% 가 명사로서, 대부분의 기존 연구들은 미등록어 명사를 인식하는데 초점을 두고 있다. 본 연구도 미등록어를 미등록 명사로 한정하기로 한다. 미등록어 인식 모델을 유도하기 전에 명사를 포함하는 어절은 어떤 특징이 있는지 살펴본다. 이 특징은 세종계획 품사 부착 말뭉치[10]를 사용하여 분석하였다.

표 1은 한 어절 내에서 명사열이 몇 개나 등장하는지를 보인다. 명사열이란 하나 이상의 연속된 명사(보통명사, 고유명사, 의존명사)를 뜻한다. 단, 의존명사 단독으로 등장한 경우는 제외하였다. 명사열은 단일명사나 명사가 연속된 복합명사를 의미한다. 명사열이 한번 이상 등장한 어절의 99.9%는 어절 내에 단 하나의 명사열만 존재하였다. 2개 이상의 명사열이 존재하는 어절은 “가정.학교.사회가”, “관계자,회계사,호텔” 등의 명사가 나열된 형태이거나, “10승2패가”, “10인10색이요” 등처럼 숫자와 명사가 반복적으로 결합된 경우이거나, “TV중계권료를”, “개혁신당측과의” 등처럼 명사들 중간에 접미사(‘권’)나 접두사(‘신’)가 결합되어 있는 경우들이었다. 일부는 띄어쓰기

표 2. 명사열의 등장 위치
Table 2. Position of NS(noun sequences) in an Eojeol

Position of NS	Number of NS	Ratio
1	4548242	98.3%
2	65110	1.4%
≥ 3	11470	0.3%
Total(≥ 1)	4624822	100.0%

오류로 인한 것도 있었다.

표 2는 각 명사열이 한 어절 내에서 몇 번째 형태소로 등장하는지를 보인다. 한 어절 내에 여러 명사열이 존재할 경우 각 명사열을 각각 카운트하였다. 명사열의 98.3%가 어절의 첫 번째 위치에 등장하였다. 두 번째 형태소로 등장한 명사열은 "초개인주의는", "준국가", "첫공판이" 등 접두사('초', '준')나 관형사('첫')와 결합한 경우였다. 세 번째 이후로 등장한 경우는 "한국신문학사를", "평준화제도" 등처럼 명사열 앞쪽에 접두사('신')나 접미사('화')가 놓여 명사열이 끊긴 경우, "9년제대학", "Y2K문제는" 등과 같이 숫자나 영문자가 앞쪽에 결합된 경우, "가정.학교.사회가", "관계자,회계사,호텔"과 같이 명사들이 나열된 경우였다.

2-2 미등록어 인식 환경 정의

앞 절에서 파악한 명사 포함 어절의 특징에 따라 미등록어 인식 문제에 대해 다음과 같이 미등록어 인식 환경을 정의한다.

첫째, 본 연구에서는 독립적인 미등록어 인식 환경을 정의한다. 미등록어 인식기는 형태소 분석 및 품사 부착 단계에서 정상적인 품사 부착에 실패했거나 형태소 분석 후보의 생성 확률이 낮게 분석되는 어절을 대상으로 후처리처럼 적용할 수 있을 것이고, 미등록어 사전 구축 등을 목적으로 문서나 문서 집합을 입력으로 하여 독립적으로 적용할 수도 있을 것이다. 형태소 분석과 연계되어 동작하는 환경이더라도 처리 대상 어절의 형태소 분석 및 품사 부착 정보가 잘못되어 있을 가능성이 높다. 따라서 본 연구에서는 입력 어절 표층 형태만으로 미등록어 인식을 시도한다. 한편, 사전에 얼마나 많은 단어가 등록되어 있는지에 따라 미등록어 정의 자체가 달라질 수 있다. 다른 요소의 영향을 최대한 배제하기 위해 관련 연구들에서 널리 사용하는 세종계획 품사 부착 말뭉치에 등장하는 형태소만으로 사전을 구축하고 이 사전에 등장하지 않은 명사를 미등록어로 정의하기로 한다. 말뭉치에 등장한 명사열은 복합명사 사전을 별도로 구축한다.

둘째, 사전에 존재하지 않는 명사열과 그 명사에 결합된 접두사 및 관형사까지 포함하여 미등록어로 인식한다. 예를 들어, "신한동해오픈의", "초시대를"이 입력 어절이라면 각각 '신한동해오픈', '초시대'를 미등록어로 추출한다. 복합명사와 같은 명사열 또는 접두사나 관형사가 결합된 명사열 자체가 미등록어를 형성하는 것이므로 분해하지 않은 형태로 미등록어를 인식한다. 복합명사 분해는 본 연구의 범위가 아니다.

셋째, 하나의 어절에 미등록어는 하나 존재한다고 가정한다. 표 1에서 살펴본 바와 같이 99.9%의 명사 포함 어절은 이 가정을 만족한다. 또한 명사가 나열된 어절의 경우 나열 기호를 기준으로 어절을 분리하여 처리하면 이 어절들 또한 가정을 만족한다.

넷째, 미등록어는 어절의 첫 위치에 등장한다고 가정한다. 접두사나 관형사가 결합된 형태로 미등록어를 인식하고 나열 기호로 분리하여 처리하면 이 가정도 대부분의 미등록어에 대해 만족한다.

2-3 미등록어 인식을 위한 확률 모델

하나의 미등록어 포함 어절 e는 미등록어 u와 명사 접미 형태소열 s의 결합인 $<u, s>$로 정의해볼 수 있다. 명사 접미 형태소열이란 한 어절에서 명사 뒷부분에 등장할 수 있는 한 개 이상의 형식 형태소 나열을 뜻한다. 어절 e에 포함되어 있는 미등록어 u를 인식해내기 위해, 어절 e를 $\{<u_1, s_1>, <u_2, s_2>, ...\}$ 형태로 미등록어 후보 u_i와 그에 따른 접미 형태소열 s_i로 분할해 볼 수 있다. 이 후보들 중 미등록어 포함 어절 e를 구성할 가능성이 가장 높은 후보를 미등록어로 인식한다.

이를 모델링하기 위해 미등록어 후보를 U, 미등록어 후보의 품사가 명사인지의 여부를 N, 접미 형태소열을 S로 표기하기로 하자. 미등록어 후보가 명사일 경우 N은 1, 아닐 경우 0이다. 미등록어는 다음과 같이 확률 $P(U, N=1, S)$를 최대화시키는 후보로 정의할 수 있다.

$$\begin{aligned} u &= argmax_{u_i} P(U, N=1, S) \\ &= argmax_{u_i} P(S)P(N=1|S)P(U|N=1, S) \end{aligned} \quad (1)$$

예를 들어, "유로존에서는"이라는 어절에 대해, $\{<u_1$=유로존, s_1=에서는>, $<u_2$=유로존에서, s_2=는>, $<u3$=유로존에서는, $s_3=\phi>\}$ 등의 미등록어 후보를 추출할 수 있으며, 이 후보들 중 확률 $P(U=u_i, N=1, S=s_i)$를 최대로 만드는 후보 u_i를 미등록어로 인식한다.

수식 (1)의 각 확률을 추정하는 방법에 대해 차례로 살펴본다. $P(S)$는 접미 형태소열의 확률로서 다음과 같이 추정한다. $C(<*, *>)$는 학습 말뭉치에서 등장한 모든 어절의 개수이고, $C(<*, S>)$는 어절의 뒷부분이 접미 형태소열 S로 구성된 어절의 개수이다.

$$P(S) \approx \frac{C(<*, S>)}{C(<*, *>)} \quad (2)$$

$P(N=1|S)$는 접미 형태소열이 주어졌을 때 선행하는 형태소가 명사일 조건부 확률로서, 접미 형태소열 S로 끝나는 어절의 개수 $C(<*, S>)$와 S가 명사 뒤에 사용된 어절 개수 $C(<*^N, S>)$를 사용하여 다음과 같이 추정한다.

$$P(N=1|S) \approx \frac{C(<*^N, S>)}{C(<*, S>)} \quad (3)$$

그러나 우리의 관심 대상인 미등록어들은 말뭉치에서 등장하지 않은 단어들이므로 수식 (1)의 $P(U|N=1,S)$를 학습 말뭉치로부터 적절히 추정하기가 쉽지 않다. 본 연구에서는 이 확률을 추정하기 위해 숨겨진 명사 접미 형태소열 S^B를 설정한다. S^B는 특정 접미 형태소열 s_i에 어울리는 명사 u_i의 결합 가능성을 파악하기 위한 지표 역할을 하는 명사 접미 형태소열들이다. 이를 지표 형태소열이라고 부르기로 한다. 예를 들어, <유로존, 에서는>이라는 후보에 대해 '에서는'이라는 접미 형태소열로부터 '유로존'이 미등록어인지를 추정하기 어렵다면, '에서는'과 관련성 높은 '에서도', '에는' 등의 접미 형태소열과 '유로존'이 결합하여 등장할 가능성에 따라 미등록어를 추정한다. 한 어절의 접미 형태소열이 s_i이고 선행 형태소의 품사가 명사라는 것을 알고 있을 때, 그 명사가 미등록어 후보 u_i일 확률은 학습 말뭉치로부터 u_i의 관련 증거를 찾아볼 수 없으므로 증거를 찾아볼 수 있는 지표 형태소열 S^B를 이용한다. 지표 형태소들을 이용하여 다음과 같이 $P(U|N=1,S)$를 추정한다.

$$P(U|N=1,S) = \sum_{S^B} P(S^B|N=1,S) P(U|S^B,N=1,S) \quad (4)$$

$P(S^B|N=1,S)$는 특정 접미 형태소열과 선행 형태소가 명사라는 사실이 주어졌을 때 지표로서 S^B가 사용될 확률로서, S^B가 명사 접미 형태소열 S를 대신하여 지표 역할을 얼마나 잘 수행할 수 있는지를 표현하는 값이다. 이 확률은 다음과 같이 추정된다.

$$P(S^B|N=1,S) \approx \frac{Q(<*^N, S \wedge S^B>)}{Q(<*^N, S>)} \quad (5)$$

$Q(<*^N,S>)$는 학습 말뭉치에서 $<u_i,s_i>$ 형태로 접미 형태소열 s_i와 결합하여 등장한 명사 u_i의 중복을 제거한(unique) 개수를 뜻한다. $Q(<*^N,S \wedge S^B>)$는 명사 u_i에 대해 $<u_i,s_i>$ 형태의 어절과 $<u_i,s_j^B>$ 형태의 어절이 모두 존재하는 명사 u_i의 중복을 제거한 개수를 의미한다. 이 값을 접미 형태소열 쌍의 공기 빈도라고 부르기로 한다. 확률 값 계산을 위해 학습 말뭉치에 등장하는 명사 접미 형태소열의 각 쌍에 대해 공기 빈도를 계산해 두어야 한다. 예를 들어, <기차, 를>, <기차, 에서>, <기차, 에서는>, <기차, 로>, <안양, 을>, <안양, 에서>, <안양, 에서는>, <안양, 으로> 등의 어절이 학습 말뭉치에서 등장했다면, (에서는, 에서)의 공기 빈도는 두 형태소열이 '기차'와 '안양'에 모두 결합된 형태로 등장했으므로 2가 된다. (를, 에서)는 1, (을, 에서)는 1, (를, 을)은 0, (을, 로)는 0의 값을 갖는다. 이 공기 빈도는 자연스럽게 선행 명사의 종성 유무에 따라 교환 가능한 접미 형태소열인지도 판별할 수 있는 정보를 제공한다.

미등록어 u_i의 생성 확률은 지표 형태소열 S^B가 주어진 상황에서는 S와 독립적이라고 가정하면 수식 (4)의 $P(U|S^B,N=1,S)$는 다음과 같이 풀어 쓸 수 있다.

$$\begin{aligned} P(U|S^B,N=1,S) &\approx P(U|S^B,N=1) \\ &\approx \lambda P(U|S^B) + (1-\lambda) P(U|N=1) \end{aligned} \quad (6)$$

λ는 $P(U|S^B)$와 $P(U|N=1)$의 반영 정도를 조절하는 파라미터이다. 미등록어는 학습 말뭉치에 등장하지 않은 단어이므로 위 확률 추정을 위해서는 다른 방법이 필요하다. 확률 $P(U|S^B)$의 추정에는 웹 문서를 활용한다.

$$P(U|S^B) \approx \frac{C_W(<U,S^B>)}{C_W(<*,S^B>)} \quad (7)$$

$C_W(<U,S^B>)$는 웹 문서에서 $<u_i,s_j^B>$ 형태로 등장한 사례의 개수로서 편의상 문서 개수로 계산한다. $C_W(<*,S^B>)$는 지표 형태소열 S^B가 등장한 문서 개수이다. 형태소 분석 등의 처리가 이루어지지 않은 웹 문서에서 $C_W(<*,S^B>)$ 값을 산출하는 것은 어려움이 있으므로, 학습 말뭉치에서 발생한 비율에 따라 그 값을 다음과 같이 추정한다.

$$C_W(<*,s_j>) \approx \frac{\gamma \cdot \sum_{n_j \in FN_j, s_j} C_{W_{approx}}(<*,s_j>,n_j)}{|FN_j|} \quad (8)$$

$$C_{W_{approx}}(<*,s_j>,n_j) = \frac{C(<*,s_j>) C_W(<n_j,s_j>)}{C(<n_j,s_j>)} \quad (9)$$

FN_j는 접미 형태소열 s_j에 대해 $C(<n_j,s_j>)$가 큰 명사 집합을 뜻한다. 즉, 학습 말뭉치에서 s_j와 자주 결합하여 등장한 고빈도 명사 집합이다. $|FN_j|$는 집합 FN_j의 원소 개수를 뜻한다. 고빈도 명사를 이용하여 학습 말뭉치에서 등장하는 비율 $C(<*,s_j>)/C(<n_j,s_j>)$ 에 따라 웹 문서에서의 접미 형태소열 빈도를 추정하고 그 평균값을 $C_W(<*,S^B=s_j>)$로 사용한다. 이때 어절 빈도 비율을 문서 개수로 변환하기 위해 1이하의 양수인 변환 상수 γ를 곱한다. 효과적인 추정을 위해 FN_j는 $C(<n_j,s_j>)$가 가장 큰 2음절 이상으로 구성된 m개의 명사로 구성한다.

수식 (6)의 $P(U|N=1)$도 또한 학습 말뭉치에서는 알아낼 수 없는 값이다. 웹 문서가 아무리 방대하더라도 처리하려는 문서에 처음 등장하는 신조어라면 수식 (7)의 $P(U|S^B)$ 값도 추정이 어려울 것이다. 이런 경우에 대응하기 위해 $P(U|N=1)$은 처리 대상 문서들에서 등장하는 단어 통계를 활용하여 다음과 같이 추정한다.

$$P(U|N=1) \approx \frac{C_L(<U^{\hat{N}},*>)}{C_L(<*^{\hat{N}},*>)} \quad (10)$$

$C_L(<*^{\hat{N}},*>)$는 처리 대상 문서들에 등장한 어절 중 명사가 포함된 것으로 판정되었거나 어절의 접미 형태소열을 바탕으로 미등록어(명사)가 포함되어 있을 것으로 추정되는 총 어절의 개수, $C_L(<U^{\hat{N}},*>)$는 처리 대상 문서들에서 임의의 명사 접미 형태소열 s_i에 대해 $<u_i, s_i>$ 형태로 분리가 가능한 어절의 개수를 의미한다.

앞서 서술한 요소들을 결합하면 수식 (1)의 $P(U,N=1,S)$는 결국 다음과 같이 추정된다.

$$P(U,N=1,S) \approx \frac{C(<*,S>)}{C(<*,*>)} \cdot \frac{C(<*^N,S>)}{C(<*,S>)} \cdot \sum_{S^B} \frac{Q(<*^N,S \wedge S^B>)}{Q(<*^N,S>)} \left(\lambda \frac{C_W(<U,S^B>)}{C_W(<*,S^B>)} + (1-\lambda) \frac{C_L(<U^{\hat{N}},*>)}{C_L(<*^{\hat{N}},*>)} \right) \quad (11)$$

$$P(U,N=1,S) \approx \frac{C(<*^N,S>)}{C(<*,*>)} \cdot \sum_{S^B} \frac{Q(<*^N,S \wedge S^B>)}{Q(<*^N,S>)} \left(\lambda \frac{C_W(<U,S^B>)}{C_W(<*,S^B>)} + (1-\lambda) \frac{C_L(<U^{\hat{N}},*>)}{C_L(<*^{\hat{N}},*>)} \right) \quad (12)$$

본 연구에서는 모든 u_i, s_i에서 동일한 값을 갖는 $C(<*,*>)$를 제거하는 등 위 확률을 다음과 같이 간략화한 점수 함수 $Score(u_i,s_i)$를 사용하여 미등록어를 인식한다.

$$u = argmax_{u_i} Score(u_i,s_i) \quad (13)$$

$$Score(u_i,s_i) = \frac{C(<*^N,s_i>)}{Q(<*^N,s_i>)} \cdot \sum_{s_j \in S_i^{Top}} Q(<*^N,s_i \wedge s_j>) \left(\lambda \frac{C_W(<u_i,s_j>)}{C_W(<*,s_j>)} + (1-\lambda) \frac{C_L(<u_i^{\hat{N}},*>)}{C_L(<*^{\hat{N}},*>)} \right) \quad (14)$$

수식 (14)에서 $C_W(<*,s_j>)$는 수식 (8), (9)로 계산된다. 효율적인 계산을 위하여 모든 지표 형태소열에 대해 값을 계산하지 않고, 미등록어 후보 u_i에 결합된 명사 접미 형태소열 s_i에 대해 공기 빈도 $Q(<*^N,s_i \wedge s_j>)$ 값이 큰 상위 k개의 지표 형태소열 집합 S_i^{Top}에 대해서만 계산한다. 수식 (13)에 의해 선택된 최고 후보의 점수가 임계값 δ를 초과하는 경우에만 최종 미등록어로 인식한다.

본 모델은 기존 연구들에서 활용했던 미등록어 인식 관련 다양한 정보들을 하나의 확률 모델로서 포괄하고 있다. 기본적으로 명사 접미 형태소열의 통계를 기반으로 접미 형태소열 간의 공기 빈도를 통하여 미등록어 후보와 접미 형태소열 간의 결합 정보를 간접적으로 활용한다. 또한 웹 문서에서의 등장 빈도나 처리 대상 문서에서의 중복 출현 정보를 모두 고려하여 가장 적합한 미등록어 후보를 선택한다.

표 3. 테스트 문서의 통계
Table 3. Statistics of test documents

Test set	Domain	Docs	Eojeols	Unknown words	Unknown words (unique)
Gen	General	50	21,329	1,509 (7.1%)	770
Eco	Economy	50	15,021	1,257 (8.4%)	513

Ⅲ. 실험 결과 및 분석

3-1 실험 환경

미등록어 인식 성능 테스트를 위해 온라인 뉴스 기사를 직접 수집하여 미등록어 정보를 부착하였다. 표 3은 테스트 문서의 통계를 보인다. 테스트 문서로는 일반뉴스(Gen)와 경제뉴스(Eco) 등 두 개의 세트를 구축하였다. 일반뉴스는 특정 섹션 제한 없이 네이버에서 제공하는 2018년 11월 1일 ~ 2019년 1월 31일까지의 주요뉴스[11]를 수집하여 무작위로 선별하되 너무 짧지 않으면서도 일자나 내용의 중복도가 최소화되도록 50개의 뉴스로 구성하였다. 경제뉴스는 네이버 뉴스 기사 '글로벌 경제' 섹션의 2019년 3월 28일 뉴스[12] 총 112개 중 너무 짧은 기사를 제외하여 무작위로 선택된 50개의 뉴스로 구성하였다. 일반뉴스 세트는 어절의 약 7.1%가 미등록어를 포함하고 있었으며 경제뉴스 세트는 약 8.4%가 미등록어를 포함하여, 경제뉴스가 일반뉴스에 비해 더 많은 비율의 미등록어를 포함하고 있었다. 일반뉴스 세트는 모델의 적절한 파라미터를 선정하는 용도로 사용하였으며, 경제뉴스 세트는 최종 시스템의 성능을 평가하는데 사용하였다.

수식 (14)의 $C()$, $Q()$, $C_W(<*,s_j>)$ 등의 계산을 위해서는 품사 부착 말뭉치가 필요하다. 본 연구에서는 세종계획 품사 부착 말뭉치[10]를 사용하였다. 또한 이 말뭉치에 등장한 각 형태소와 명사열들을 사전으로 구축하여 이 사전에 등장하는 형태소는 등록어로 처리한다. 또한 한 어절 내에서 명사열 뒷부분에 등장하는 형식 형태소열들을 추출하여 명사 접미 형태소 사전으로 사용하였다. 이 사전에 따라 입력 문서의 각 어절을 $\{<u_1, s_1>, <u_2, s_2>, ...\}$로 분할한다. 이 과정에서 u_i가 사전에 등록된 형태소라면 등록어로 간주하여 처리 대상에서 제외하였다. 사전에 등록되지 않은 각 후보에 대해 수식 (14)를 계산하였다.

수식 (14)의 $C_W(<u_i,s_j>)$를 계산하려면 특정 단어가 등장한 웹 문서의 개수를 알아야 한다. 이를 위해 웹 검색을 이용한다. 노이즈를 최소화하고 처리 대상 문서와 유사한 통계값을 구하기 위해 본 연구에서는 네이버 뉴스 검색[13]을 이용한다. 즉, u_i와 s_j가 결합한 형태의 질의를 구성하여 네이버 뉴스 검색을 수행하여 검색 결과 건수를 $C_W(<u_i,s_j>)$로 사용한다. 수식

표 4. 지표 형태소열 크기에 따른 성능
Table 4. Performance according to barometer morpheme sequence(BMS) size

BMS Size(k)	Gen		
	Precision	Recall	F1
3	0.7431	0.8891	0.8096
5	0.7444	0.8919	0.8115
10	0.7420	0.8890	0.8088
20	0.7396	0.8872	0.8067
30	0.7389	0.8874	0.8063

표 5. 파라미터 λ에 따른 성능
Table 5. Performance according to the parameter λ

Parameter λ	Gen		
	Precision	Recall	F1
0.0	0.7362	0.8826	0.8028
0.1	0.7418	0.8902	0.8092
0.3	0.7448	0.8916	0.8116
0.5	0.7444	0.8919	0.8115
0.7	0.7471	0.8942	0.8141
0.9	0.7508	0.8984	0.8180
1.0	0.7951	0.8615	0.8269

(8)의 γ로는 학습 말뭉치 문서당 평균 어절 수의 역수를 사용하였고, 사전 실험에 따라 m은 20으로 고정하여 실험하였다.

평가 척도는 다음과 같이 계산되는 정확률(Precision), 재현율(Recall), F1값을 사용한다.

$$Precision = \frac{\text{시스템이 올바르게 인식한 미등록어 개수}}{\text{시스템이 인식한 미등록어 개수}} \quad (15)$$

$$Recall = \frac{\text{시스템이 올바르게 인식한 미등록어 개수}}{\text{미등록어 총 개수}} \quad (16)$$

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad (17)$$

3-2 미등록어 인식 결과

표 4는 수식 (14)에서 지표 형태소열 S_i^{Top}의 크기 k에 따른 성능 변화를 보인다. 본 실험에 사용된 파라미터는 $\lambda = 0.5$, $\delta = 0$이다. 점수 임계값 δ를 0으로 설정하여 점수가 존재하는 모든 후보를 출력한 결과이다. 성능 척도 중 재현율에 집중해 볼 필요가 있다. 표 4의 재현율은 정답 미등록어가 어절의 최고 점수 후보에 얼마나 많이 포함되어 있는지를 보여주는 것이다. 실험 결과 5개 정도의 소규모 지표 형태소열만 사용하더라도 좋은 성능을 보였다. 사용된 지표 형태소열을 예를 들면 다음과 같다.

표 6. 시스템 성능 비교
Table 6. Performance comparison with other system

System	Gen			Eco		
	Precision	Recall	F1	Precision	Recall	F1
Sim_Web	0.9056	0.5287	0.6676	0.9499	0.5413	0.6896
Prop_P	0.9079	0.5207	0.6618	0.9565	0.5458	0.6951
Prop_R	0.7508	0.8984	0.8180	0.8967	0.9526	0.9238
Prop_F	0.8028	0.8652	0.8328	0.9147	0.9112	0.9129

- 에서는: {에, 에서, 의, 에는, 을}
- 가: {를, 는, 의, 로, 와}
- 마저: {도, 만, 에, 의, 를}
- 하며: {하고, 하는, 할, 한, 하여}
- 까지: {도, 에, 의, 과, 는}

표 5는 수식 (14)에서 웹 문서에서의 빈도와 처리 대상 문서에서의 빈도 사이의 반영 비율을 조절하는 λ에 따른 성능 변화이다. 파라미터는 $k = 5$, $\delta = 0$이 사용되었다. 전반적으로 λ 값이 높을수록, 즉 웹 문서에서의 빈도를 더 많이 고려할수록 더 좋은 성능을 보이고 있다. 그러나 $\lambda = 1$로 설정하여 처리 대상 문서에서의 중복도를 전혀 고려하지 않고 웹 빈도만 사용하는 경우에는 재현율이 낮아졌다. 즉, 정답 미등록어 중 최고 점수 후보에 포함되지 않는 비율이 증가하게 된다는 것이다. λ가 0.9일 때 최고의 재현율을 보였다. 반대로 $\lambda = 0$으로 설정하여 웹 빈도를 전혀 사용하지 않은 경우에도 성능이 크게 하락하지는 않았다. 웹에서의 빈도와 처리 대상 문서에서의 중복도가 자체적으로 의미 있는 정보이며 본 모델 내에서 서로 상호 보완하는 역할을 하고 있는 것이다.

일반뉴스에 대해 수행한 이전 실험들을 통하여 파라미터들을 $k = 5$, $\lambda = 0.9$로 결정하고, 이 값들로 제안 시스템 설정을 완료하였다. 표 6은 일반뉴스 및 경제뉴스에서 기존 연구와 유사하게 자체 구현한 시스템(Sim_Web)과 제안 시스템(Prop)의 성능을 비교한 결과이다. Sim_Web은 유사 어절을 비교하는 기존 연구 [3]-[5]와 웹 문서에서의 출현 빈도를 활용한 [6], [7]의 연구를 참고하여 구현된 것이다. 어절의 시작 부분이 동일한 다른 어절이 존재하는 경우, 이들을 모아 (공통 문자열) + (명사 접미 형태소열)로 구분 가능하면 최장 길이 공통 문자열을 후보로 추출한다. 시작 부분이 동일한 다른 어절이 없는 경우에는 어절의 뒷부분이 명사 접미 형태소와 매칭 가능하면 앞부분을 미등록어 후보로 추출한다. 미리 구축해둔 명사 추정에 도움이 될 만한 조사와 후보를 결합하여 웹 검색을 수행하고 그 결과가 임계값 이상일 경우 최종 미등록어로 인식한다.

목표 성능에 따라 제안 시스템은 3가지 조합을 제시한다.

Prop_P는 정확률이 중요한 상황에 사용할 시스템으로서 $\delta = 800$으로 설정되었다. Prop_R은 $\delta = 0$으로 설정하여 재현율이 중요한 상황의 시스템을, Prop_F는 $\delta = 10$으로 설정하여 정확률과 재현율이 동시에 중요한 상황에 어울리는 시스템을 뜻한다.

Sim_Web은 정확률은 높지만 재현율이 50% 수준에 머무르는 문제점을 보인다. 이에 비해 제안 시스템은 정확률과 재현율 모두 높은 수준을 보이고 있다. 이런 경향은 일반뉴스와 경제뉴스에서 공통적으로 보이고 있다. 특히 경제뉴스에서 더 높은 성능을 보이는데, 이는 뉴스 세트의 특징에 기인하는 것으로 보인다. 시스템 처리 과정에서 미등록어 후보가 하나만 생성된 경우가 일반뉴스는 전체 어절의 약 2.6%, 경제뉴스는 약 3.8%에 해당했다.

Sim_Web은 웹 출현 빈도를 기반으로 검증하기 때문에 웹 검색 결과 건수가 많지 않을 경우 미등록어로 추출하지 못한다. 이에 비해 제안한 모델은 웹 출현 빈도, 문서 내 출현 빈도, 결합된 접미 형태소열의 특징 등을 동시에 고려하므로 웹 출현 빈도가 낮더라도 미등록어를 정확히 인식해낼 수 있었다. 예를 들어, "...이로써 석유를 생산하는 업스트림부터 석유화학제품을 생산하는 다운스트림까지 보유하게 됐다..."와 같은 내용이 포함된 문서에서 미등록어 '업스트림', '다운스트림'을 Sim_Web은 인식해내지 못했다. 미리 정해 둔 조사와 결합된 표현이 웹에서 출현한 빈도가 높지 않았기 때문이다. 그러나 제안한 모델은 '업스트림'에 대해서는 '부터'의 지표 형태소열로 {에, 의, 도, 에서, 을}이 선택되었고, 각 지표 형태소열과 결합된 웹 검색 결과는 공기 빈도에 따라 서로 다르게 점수 산정에 기여하게 된다. '다운스트림'에 대해서는 '까지'의 지표 형태소열로 {도, 에, 의, 과, 는}이 선택되어 웹 검색이 수행되었다. 제안 모델은 웹 검색 결과뿐만 아니라, '부터', '까지'가 갖는 명사 접미 형태소로서의 특징과 후보의 문서 내에서의 빈도까지 함께 고려하여 해당 단어들을 모두 미등록어로 인식할 수 있었다.

반면, 본 모델은 입력 어절에서 후보를 추출하는 과정에서 (사전에 등록된 형태소) + (명사 접미 형태소열)로 분할할 수 있는 경우 단순히 등록어로 간주하여 처리 대상에서 제외함으로써, '김성은', '손재권'과 같이 명사 접미 형태소열로 간주 가능한 표현으로 어절이 끝나는 경우 미등록어 인식에 실패했다. 특히 인명의 경우에 흔히 발생되는 문제였다. 이 문제는 인명사전을 추가 활용함으로써 보완할 수 있으리라 예상한다.

Ⅳ. 결 론

본 연구는 미등록어 인식을 위한 확률 모델을 제안하였다. 미등록어 포함 어절이 미등록 명사 후보와 명사 접미 형태소열로 구성될 확률을 계산하여 이 확률이 가장 높은 후보를 미등록어로 인식한다. 이 확률 모델은 명사 접미 형태소열의 등장 확률과 그 명사 접미 형태소열이 미등록 명사 후보와 결합하여 함께 등장할 확률로 구분하여 추정하였다. 명사 접미 형태소열의 등장 확률은 품사 부착 말뭉치에서의 출현 빈도를 바탕으로 추정하였다. 명사 접미 형태소열에 대해 미등록어 후보가 결합하여 등장할 확률은 해당 미등록어 후보에 대한 증거가 충분하지 않으므로, 해당 증거의 수집이 용이한 지표 형태소열을 활용하여 확률을 추정하였다. 이를 위해 접미 형태소열 간의 공기 빈도를 사용하였다. 이와 함께 미등록어 후보가 지표 형태소열과 결합하여 웹에서 출현한 빈도와 미등록어 후보가 처리 대상 문서에서 출현한 빈도를 모두 고려하여 확률을 추정하였다. 이렇게 함으로써 미등록어 인식 관련 연구들에서 활용했던 명사 접미 형태소열, 후보와 명사 접미 형태소열 간의 결합 관계, 웹에서의 출현 빈도, 처리 대상 문서에서의 출현 빈도 등을 하나의 확률 모델로 포괄할 수 있었다. 단편적으로 활용되던 다양한 휴리스틱을 하나의 모델로 결합하여 정립하였다는 점에서 본 논문의 의의가 있다. 실험 결과 제안한 모델은 F1 기준으로 경제뉴스에서 0.9238의 성능을 달성할 수 있었다. 본 모델은 미등록어의 대부분을 차지하는 미등록 명사를 집중하여 다루었으나 다른 품사로도 충분히 확장할 수 있을 것이다. 향후 품사 제한 없이 미등록어를 인식하는 모델로 확장하기 위한 연구와 인식된 미등록어를 각 단위 형태소로 분리하는 연구가 필요하겠다.

감사의 글

이 논문은 2018년도 정부(과학기술정보통신부)의 재원으로 한국연구재단-차세대정보ㆍ컴퓨팅기술개발사업의 지원을 받아 수행된 연구임(No. 2017M3C4A7068186)

참고문헌

[1] B. Kim and J. S. Lee, "Probabilistic Segmentation and Tagging of Unknown Words," *Journal of KIISE*, Vol. 43, No. 4, pp. 430-436, April 2016.

[2] S. S. Kang, *Korean Morphological Analysis Using Syllable Information and Multiple-Word Units*, Dept. of Computer Engineering, Seoul National University, Ph. D. Thesis, 1993.

[3] B. R. Park, *Korean Unknown Word Recognition Based on Fulltext Analysis*, Dept. of Computer Science and Engineering, Korea University, Ph. D. Thesis, 2000.

[4] K. T. Park and Y. H. Seo, "Korean Unknown-noun Recognition using Strings Following Nouns in Words," *Journal of the Korea Contents Association*, Vol. 17, No. 4, pp. 576-584, April 2017.

[5] J. M. Yang, M. J. Kim, and H. C. Kwon, "Extraction Method of the Unknown-Words with Linguistic Knowledge in Korean," *Proceedings of KIISE Conference*, Vol. 23, No.

1A, pp. 957-960, April 1996.
[6] S. Y. Park, “Automatic Construction of Korean Unknown Word Dictionary using Occurrence Frequency in Web Documents,” *Journal of the Korea Society of Computer and Information*, Vol. 13, No. 3, pp. 27-33, May 2008.
[7] S. Y. Park, “Phase-based Model Using Web Documents for Korean Unknown Word Recognition,” *Journal of the Korea Institute of Information and Communication Engineering*, Vol. 13, No. 9, pp. 1898-1904, September 2009.
[8] B. R. Park, Y. S. Hwang, and H. C. Rim, “Estimation of an Unknown Noun Based on a Tail Pattern Analysis of Morphologically Similar Words,” *Proceedings of KIISE Conference*, Vol. 23, No. 1A, pp. 907-910, April 1996.
[9] S. H. Kim, J. T. Yoon, and M. S. Song, “Analysis of Unknown Words for Korean Document Processing based on Dynamically-generated Local Dictionary,” *Journal of KISS: Software and Applications*, Vol. 29, No.5·6, pp. 407-416, June 2002.
[10] The National Institute of the Korean Language, *21st Century Sejong Project Final Result*, Revised Edition, 2011.
[11] Naver. Main News [Internet]. Available: https://news.naver.com/main/history/mainnews/index.nhn.
[12] Naver. Global Economy [Internet]. Available: https://news.naver.com/main/list.nhn?mode=LS2D&sid2=262&mid=shm&sid1=101&date=20190328.
[13] Naver. News Search [Internet]. Available: https://search.naver.com/search.naver?where=news.

한경수(Kyoung-Soo Han)

1998년 : 고려대학교 컴퓨터학과 (학사)
2000년 : 고려대학교 대학원 (이학석사)
2006년 : 고려대학교 대학원
(이학박사-전산학)

2006년～2009년: SK텔레콤
2009년～현 재: 성결대학교 컴퓨터공학과 교수
※관심분야 : 정보검색, 질의응답시스템, 텍스트 마이닝 등

홍혁준(Hyuck-Jun Hong)

2014년～현재: 성결대학교 컴퓨터학과 재학

※관심분야 : 자연어처리, 텍스트 마이닝 등

지한솔(Hansol Ji)

2017년～현재 : 성결대학교 컴퓨터학과 재학

※관심분야 : 인공지능, 자연어처리 등