JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 25 No. 4, pp. 141-148, April 2020
https://doi.org/10.9708/jksci.2020.25.04.141

Evaluation of English Term Extraction based on Inner/Outer Term Statistics

In-Su Kang*

*Associate Professor, Dept. of Computer Science, Kyungsung University, Busan, Korea

[Abstract]

Automatic term extraction is to recognize domain-specific terms given a collection of domain-specific text. Previous term extraction methods operate effectively in unsupervised manners which include extracting candidate terms, and assigning importance scores to candidate terms. Regarding the calculation of term importance scores, the study focuses on utilizing sets of inner and outer terms of a candidate term. For a candidate term, its inner terms are shorter terms which belong to the candidate term as components, and its outer terms are longer terms which include the candidate term as their component. This work presents various functions that compute, for a candidate term, term strength from either set of its inner or outer terms. In addition, a scoring method of a term importance is devised based on C-value score and the term strength values obtained from the sets of inner and outer terms. Experimental evaluations using GENIA and ACL RD-TEC 2.0 datasets compare and analyze the effectiveness of the proposed term extraction methods for English. The proposed method performed better than the baseline method by up to 1% and 3% respectively for GENIA and ACL datasets.

▸ **Key words**: Term extraction, Inner term set, Outer term set, Term importance score, Domain term

[요 약]

용어추출은 도메인 텍스트 모음으로부터 도메인 용어 목록을 인식하는 작업이다. 용어추출의 기존 효과적인 방법들은 비교사 방식으로 동작하며, 후보 용어 집합을 추출하는 작업과 후보 용어에 용어중요도를 할당하는 작업을 주요 단계로 포함한다. 후보 용어의 용어중요도 계산과 관련하여 본 논문에서는 후보 용어의 내부 및 외부용어집합을 활용한다. 내부용어집합은 후보 용어에 포함된 다른 짧은 용어들의 집합이며, 외부용어집합은 후보 용어가 포함된 다른 긴 용어들의 집합이다. 본 논문에서는 후보 용어의 내부 혹은 외부용어집합으로부터 후보 용어의 용어 강도를 계산하는 다양한 강도 함수들을 제시하고, 이들 용어 강도 값들과 C-value 점수를 결합하는 용어중요도 계산 방법을 소개한다. 생물학 및 전산언어학 분야 영어 데이터셋을 사용한 성능 평가에서는 제안된 방법의 용어추출 성능을 비교하고 분석한다. 제안된 방법은 생물학 및 전산언어학 분야 데이터셋에 대해 각각 최대 1%와 3% 차이의 성능 향상을 보였다.

▸ **주제어**: 용어추출, 내부용어집합, 외부용어집합, 용어중요도, 도메인 용어

• First Author: In-Su Kang, Corresponding Author: In-Su Kang
*In-Su Kang (dbaisk@ks.ac.kr), Dept. of Computer Science, Kyungsung University
• Received: 2020. 03. 09, Revised: 2020. 04. 20, Accepted: 2020. 04. 20.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

용어추출(Term Extraction)은 입력 도메인 코퍼스로부터 해당 분야를 대표하는 도메인 용어들을 추출하는 작업이다[1, 2]. 용어추출은 도서 후반부 색인 목록의 자동 생성[3, 4], 특정 도메인 텍스트로부터 해당 분야 용어의 자동 추출[5], 온톨로지 학습[6, 7] 등 다양한 자연어처리 응용에 활용 가능하다.

용어추출의 효과적인 선행 접근법들은 비교사(Unsupervised) 방식으로 제안되었다[2]. 비교사 기반 용어추출 절차는 용어추출대상 입력 코퍼스로부터 후보 용어 집합을 추출하고, 후보 용어 집합 내 각 용어에 대해 용어중요도를 계산한 다음, 상위 용어 중요도를 갖는 용어들을 추출하는 방식으로 진행된다.

기존 비교사 용어추출 연구에서는 후보 용어가 도메인 코퍼스 내에서 빈번히 출현할수록 용어중요도를 높이는 방법들이 시도되었는데, 후보 용어의 빈도수와 함께 후보 용어와 관련된 용어들의 출현 정보가 추가 활용되었다. C-value[8], Basic[9], RAKE[10]에서는 후보 용어(예: '정보 검색')를 그 부분으로 포함하는 다른 긴 용어들(예: '특허 정보 검색')의 출현 정보를 추가 활용하여 후보 용어의 중요도를 계산하였고, GM[11], ComboBasic[12]에서는 후보 용어(예: '정보 검색')가 그 내부에 포함하는 다른 용어들(예: '검색')의 출현 정보를 추가 활용하여 용어중요도를 결정하였다. 이러한 기존 연구들은 후보 용어가 포함되는 다른 긴 용어들(외부용어)이나 후보 용어에 포함되는 다른 짧은 용어들(내부용어)의 출현 정보를 용어중요도 계산에 활용한 시도들이다.

본 논문에서는 전술한 용어 간 부분-전체 포함 관계에 기반한 용어집합들(즉, 내부용어들 혹은 외부용어들의 집합)이 용어추출에 미치는 영향을 탐구한다. 이를 위해 후보 용어의 특정 포함관계용어집합으로부터 용어의 용어성 강도를 수치화하는 포함관계용어집합 강도 함수들을 정의하고, 이러한 포함관계용어집합 강도 함수들을 활용하는 용어중요도 계산 방법을 제안한다. 실험에서는 전산언어학 및 생물학 분야 영어 용어추출 데이터셋을 사용하여 후보 용어의 내부 및 외부용어집합들로부터 계산된 용어 강도들이 용어추출 성능에 미치는 영향을 평균정확률 관점에서 분석 및 비교 제시한다.

논문의 구성은 다음과 같다. 2장에서는 기존 연구에 대해 기술한다. 3장에서는 본 논문에서 제안하는 포함관계용어집합 기반 용어 추출 방법에 대해 기술한다. 4장에서는 제안된 방법의 성능 평가 결과를 제시하고 5장에서 결론을 맺는다.

II. Related Works

기존 비교사 용어추출 연구에서는 후보 용어의 도메인 코퍼스 내 출현 정보를 중심으로 용어중요도를 계산하는 TF-IDF[13], C-value[8], Basic[9], ComboBasic[12], RAKE[10], GM[11] 등의 방법과 후보 용어의 지역 문맥 내 공기 용어들의 출현 정보를 고려하는 NC-value[8] 방법 등이 시도되었다. 또한 후보 용어의 도메인 코퍼스 및 참조 코퍼스에서의 출현 정보의 차이를 이용하는 방법으로는 도메인 및 참조 코퍼스에서의 후보 용어 빈도수의 비율 정보를 활용하는 DomainPertinence 방법[14]이나 코퍼스 크기로 정규화된 빈도수의 비율을 사용하는 Weirdness 방법[15] 등이 시도되었다. 본 논문에서는 후보 용어의 문맥 정보나 일반 코퍼스 등 추가적인 참조 코퍼스는 활용하지 않는다.

기존 비교사 용어추출의 대표적 방법은 Frantzi 등[8]이 제안한 C-value 방법으로, 후보 용어 t에 대해 t를 구성하는 총 단어의 개수, t의 코퍼스 출현 빈도수, t가 내포된 다른 긴 용어들의 빈도수, t가 내포된 다른 긴 용어들의 개수의 네 가지 통계치들을 결합하여 용어중요도를 계산하였다. C-value에서는 여러 단어로 구성된 긴 복합 용어이고 도메인 코퍼스에서 빈번히 출현하면서 다른 긴 용어들에 내포되어 출현하지 않을수록 높은 용어중요도 값이 부여된다. 본 논문에서는 후보 용어의 C-value 점수를 최종 용어중요도 계산의 한 부분으로 활용하며 C-value의 자세한 수식은 3장에 제시된다.

Basic[9]은 후보 용어가 코퍼스 출현 빈도가 높으면서 많은 단어들로 구성되어 있고 해당 용어를 내포하는 다른 긴 용어들의 개수가 많을수록 높은 중요도를 할당하는 방식이다. Basic의 변형인 ComboBasic[12]는 후보 용어에 내포된 다른 짧은 용어들의 개수를 추가 고려하였다.

Rose 등[10]은 후보 용어를 구성하는 각 단어 점수의 합으로 용어중요도를 결정하는 RAKE 방법을 제안하였는데, 각 단어의 점수로는 코퍼스 출현 빈도가 높고 다양한 다른 후보 용어들의 부분으로 출현할수록 높은 값을 부여하였다.

Nakagawa와 Mori[11]는 복합명사와 내부 단일 명사 간 통계정보에 기반한 용어추출 방법을 제시하였다. 그들은 하나의 단일명사가 도메인 용어라면 그 용어는 빈번히 출현할 뿐 아니라 다른 많은 복합명사의 부분으로 사용될 것이라는 점에 착안하였으며, 후보 용어 t에 대해 t를 구성하는 각 명사들이 코퍼스 전반에 걸쳐 그 좌측 혹은 우측 위치에 다른 명사들과의 결합 다양성이 평균적으로 클수록 높은 용어중요도를 부여하는 GM 용어중요도를 제안하

였다. 또한 후보 용어가 코퍼스 내 명사구로 개별 출현하는 빈도수를 전술한 GM 중요도와 곱하여 얻어지는 FGM 중요도를 제안하였다.

기존 용어추출 연구 중 외부용어집합은 [8, 9, 10, 12]의 연구에서, 내부용어집합은 [11, 12]의 연구 등에서 활용되었다. 그러나 [12]의 경우를 제외하면 기존 연구들에서 후보 용어의 내부 및 외부용어집합을 동시 고려하는 용어중요도 방법에 대한 연구를 찾기 힘들다.

III. The Proposed Method

1. Set of Inner Terms and Set of Outer Terms

이 절에서는 용어 간 포함 관계에 기반한 용어집합의 유형으로 내부용어집합과 외부용어집합을 정의한다. 용어 t의 내부용어집합 B(t)는 용어 t가 그 내부에 포함하고 있는 모든 부분 용어(Subterm or Inner Term)들의 집합을 의미하며, t를 구성하는 단어열로부터 생성되는 t보다 작은 길이의 연속된 단어 n-gram들의 집합으로 정의된다. 예를 들어, 용어 t가 '정보 검색 시스템 개발'인 경우, B(t)는 {정보, 검색, 시스템, 개발, 정보 검색, 검색 시스템, 시스템 개발, 정보 검색 시스템, 검색 시스템 개발}이 된다.

용어 t의 외부용어집합 P(t)는 용어 t가 그 부분으로 포함되는 t보다 긴 용어(Superterm or Outer Term)들의 집합을 의미한다. 예를 들어, 용어 t가 '정보 검색'인 경우, P(t)는 '정보 검색 시스템', '정보 검색 연구', '도서 정보 검색', '인터넷 정보 검색 시스템' 등을 그 원소로 포함할 수 있다.

전술한 내부 및 외부용어집합들은 그 원소들의 특정 코퍼스 출현 제약이 추가되면 용어가 사용되는 분야에 따라 그 용어집합의 원소 구성에 차이가 발생할 수 있다. 코퍼스 C에 대해 코퍼스 C 내에 k회 이상 출현한 모든 출현 용어들의 집합을 term(C)라고 할 때, B(t,C)는 B(t) ∩ term(C)로 정의되며, 이는 코퍼스 C에서 k회 이상 발견되는 t의 부분용어들의 집합에 해당한다. 마찬가지로 P(t,C)는 P(t) ∩ term(C)로 정의되며, t를 포함하는 용어들 중 코퍼스 C에서 k회 이상 발견되는 용어들의 집합에 해당한다.

예를 들어, k=3인 경우 코퍼스 C 내에 '정보 검색 시스템', '검색 시스템 개발'이 3회 이상 출현하지 않았다면, B('정보 검색 시스템 개발',C)={정보, 검색, 시스템, 개발, 정보 검색, 검색 시스템, 시스템 개발}이 될 것이다. 또한 코퍼스 C 내에 '정보 검색'을 포함하는 3회 이상 출현하는 큰 용어가 '정보 검색 연구'뿐이라면 P('정보 검색',C)={정보 검색 연구}가 될 것이다.

2. Term Importance Score

본 논문에서 제안하는 용어중요도는 입력 코퍼스 C와 용어 t에 대해 식 1과 같이 $w_{div}(t,f,g)$, $w_{sub}(t,f,g)$, $w_{add}(t,f,g)$, $w_{mul}(t,f,g)$의 네 유형으로 정의된다. CV(t)는 용어 t의 C-value 값에 해당한다. f(t) 및 g(t)는 용어 t의 포함관계용어집합에 적용되어 용어 t의 용어집합 강도를 반환하는 함수들로, f 혹은 g에 대체 가능한 함수의 구체적인 형태는 식 3, 4에 제시하였다. 용어 t의 포함관계용어집합은, 3.1절에서 정의한 내부용어집합 B(t,C) 혹은 외부용어집합 P(t,C) 중 하나에 해당한다. 식 1의 $w_{div}(t,f,g)$, $w_{sub}(t,f,g)$, $w_{add}(t,f,g)$, $w_{mul}(t,f,g)$는 용어 t에 대한 용어집합 강도 함수값들 f(t) 및 g(t)를 각각 f/g, f-g, f+g, f*g의 방식으로 C-value 값과 결합하여 정의한 용어중요도 수식들이다. 식 1에 사용된 CV(t) 수식은 식 2에 제시하였는데, 이는 Ventura 등[16]의 연구를 참조하여 [1]에서 수정 제시된 C-value 수식으로, #(t) 및 #(v)는 용어 t 혹은 v의 C 내에서의 출현 빈도수이고 |t|는 t를 구성하는 총 단어 수이다.

$$
\begin{aligned}
w_{div}(t,f,g) &= CV(t) \times \left(1 + \log\left(1 + \frac{f(t)}{g(t)}\right)\right) \\
w_{sub}(t,f,g) &= CV(t) \times (1 + \log(1 + f(t) - g(t))) \\
w_{add}(t,f,g) &= CV(t) \times (1 + \log(1 + f(t) + g(t))) \\
w_{mul}(t,f,g) &= CV(t) \times (1 + \log(1 + f(t) \times g(t)))
\end{aligned}
\tag{1}
$$

$$
CV(t) = \begin{cases} \log(0.1 + |t|) \times \#(t) & , if\ |P(t,C)| = 0 \\ \log(0.1 + |t|) \times \left(\#(t) - \dfrac{\sum_{v \in P(t,C)} \#(v)}{|P(t,C)|}\right) & , else \end{cases} \tag{2}
$$

식 3은 용어 t의 내부용어집합에 적용되어 t의 용어집합 강도 값을 계산하는 13개 서로 다른 함수들의 목록을 보인 것이다. 식 3에서 $len_B(t)$는 t의 내부용어집합의 크기를 의미하며, $max_B(t)$, $min_B(t)$, $sum_B(t)$, $ame_B(t)$, $gme_B(t)$, $hme_B(t)$는 각각 t의 내부용어집합 내 용어 빈도수들의 최대값, 최소값, 총합, 산술평균, 기하평균, 조화평균에 해당한다. 식 3의 하단 6개 함수들은 용어 t의 내부용어집합 내 용어 빈도수들의 총합, 최대값, 산술평균, 기하평균, 조화평균, 최소값 간의 차이를 내부용어집합 강도값으로 정의한 것이다. 이는 최소값 혹은 최대값을 이상치(outlier)로 고려하여 총합, 최대값, 평균에서 제거하는 방식으로 총합, 최대값 및 다양한 평균값들의 조정을 시도한 것이다.

$$
\begin{aligned}
&len_B(t) = |B(t,C)| \\
&max_B(t) = max_{v \in B(t,C)} \#(v) \\
&min_B(t) = min_{v \in B(t,C)} \#(v) \\
&sum_B(t) = \sum_{v \in B(t,C)} \#(v) \\
&ame_B(t) = \frac{\sum_{v \in B(t,C)} \#(v)}{|B(t,C)|} \\
&gme_B(t) = \left(\prod_{v \in B(t,C)} \#(v) \right)^{1/|B(t,C)|} \\
&hme_B(t) = \frac{|B(t,C)|}{\sum_{v \in B(t,C)} \frac{1}{\#(v)}} \\
&dsx_B(t) = sum_B(t) - max_B(t) \\
&dsn_B(t) = sum_B(t) - min_B(t) \\
&dxn_B(t) = max_B(t) - min_B(t) \\
&dan_B(t) = ame_B(t) - min_B(t) \\
&dgn_B(t) = gme_B(t) - min_B(t) \\
&dhn_B(t) = hme_B(t) - min_B(t)
\end{aligned}
\tag{3}
$$

$$
\begin{aligned}
&len_P(t) = |P(t,C)| \\
&max_P(t) = max_{v \in P(t,C)} \#(v) \\
&min_P(t) = min_{v \in P(t,C)} \#(v) \\
&sum_P(t) = \sum_{v \in P(t,C)} \#(v) \\
&ame_P(t) = \frac{\sum_{v \in P(t,C)} \#(v)}{|P(t,C)|} \\
&gme_P(t) = \left(\prod_{v \in P(t,C)} \#(v) \right)^{1/|P(t,C)|} \\
&hme_P(t) = \frac{|P(t,C)|}{\sum_{v \in P(t,C)} \frac{1}{\#(v)}} \\
&dsx_P(t) = sum_P(t) - max_P(t) \\
&dsn_P(t) = sum_P(t) - min_P(t) \\
&dxn_P(t) = max_P(t) - min_P(t) \\
&dan_P(t) = ame_P(t) - min_P(t) \\
&dgn_P(t) = gme_P(t) - min_P(t) \\
&dhn_P(t) = hme_P(t) - min_P(t)
\end{aligned}
\tag{4}
$$

식 4는 식 3의 각 함수를 내부용어집합 B(t,C) 대신 외부용어집합 P(t,C)에 대해 재정의한 것이다. 기존 연구 중 $sum_P(t)$는 [8]에서 활용되었으며, $len_P(t)$는 [8, 9, 12]에서, $len_B(t)$는 [12]에서 활용되었다. $gme_B(t)$ 및 $gme_P(t)$는 빈도수 기반 기하 평균을 계산한다는 점에서 [11]에서 제안된 GM 수식과 관련이 있으나 내부용어집합 구성 방식 및 활용되는 빈도수 기반 정보 유형에서 차이가 있다.

IV. Experiments

1. Experimental Setup

본 논문에서 제안된 용어 추출 기법의 성능 평가를 위해 [2]에서 사용된 데이터셋들 중 수백 건 이상 문서를 대상으로 구축된 ACL RD-TEC 2.0 데이터셋(이후 'ACL 데이터셋'으로 표기)과 GENIA 데이터셋을 사용하였다. ACL 데이터셋[17]은 전산언어학 분야 논문 모음인 ACL Anthology Reference Corpus로부터 추출된 300개 초록 텍스트에서 발견되는 용어들을 수작업 인식해 둔 것이다. GENIA 데이터셋[18]는 MEDLINE 데이터베이스로부터의 2000개 초록 텍스트에서 발견되는 생물학 관련 용어들을 수작업 인식해 둔 것이다.

성능 평가 지표로 기존 연구[1, 2]를 따라 평균정확률 AvP을 사용한다. 식 5에 보인 것처럼 평균정확률 AvP는 R 순위 지점까지의 순위화된 용어 목록에 대해 정답 용어 검색 지점에서의 정확률들의 합을 R로 나눈 값이다. 본 논문에서는 기존 연구[19]를 따라 정답용어집합의 크기를 ACL 및 GENIA 데이터셋의 초록에서 발견되는 전체 정답 용어의 수로 각각 설정하였다. 식 5에서 Pre(i)와 Rec(i)는 각각 i번째 순위 지점에서의 정확률과 재현율이다.

$$
AvP = \sum_{i=1}^{R} Pre(i) \times (Rec(i) - Rec(i-1)) \tag{5}
$$

성능 비교를 위해 기존 대표적 용어 추출 기법인 C-value 방법을 베이스라인 방법으로 사용하였다. 제안된 방법의 내부용어집합 B(t,C) 및 외부용어집합 P(t,C)는 GENIA 및 ACL 데이터셋의 원시 문장들의 집합으로부터 3.1절의 k=1로 설정하여 생성하였다.

2. Extraction of Term Candidates

용어 추출의 대상이 되는 ACL 및 GENIA 데이터셋들로부터 후보 용어 집합을 추출하기 위해 기존 연구[1]에서 제시된 후보 용어 추출 절차를 따랐다. 후보 용어 집합은 입력 코퍼스 내 각 문장에 대해 후보용어제약조건들을 만족하는 모든 단어 n-gram들을 추출함으로써 생성되었는데, 본 논문에서는 기존 연구[1]를 참조하여 다음과 같은 후보용어제약조건들을 적용하였다.

- 조건 1: n-gram 내 총 단어 수는 1이상 4이하일 것
- 조건 2: n-gram 내 각 단어는 영어 대소문자, 숫자 문자 및 대쉬(-) 문자로만 구성될 것
- 조건 3: n-gram을 구성하는 단어 중 불용어가 없을 것
- 조건 4: n-gram을 구성하는 단어열에 대응하는 품사열이 명사 및 명사구일 것

위 조건 3의 불용어 판단을 위한 불용어 사전으로는 SMART IR 시스템의 불용어 목록을 사용하였다. 위 조건 4의 검사를 위해 입력 코퍼스 내 각 문장에 대해 spaCy 패키지 내 품사태거[20]를 사용하여 품사 태깅(part-of-speech tagging)을 수행한 다음, n-gram의 명사 및 명사구 여부 검사를 위해 다음의 후보 용어 품사열 정규표현식 패턴을 사용하였다.

후보 용어 품사열 정규표현식:
(NN(S|P|PS)? |JJ)*(NN(S|P|PS)?)

위 정규표현식은 기존 연구[1]에서 사용된 후보 용어 품사 패턴을 일부 수정한 것이다. 위 식에서 NN, NNS, NNP, NNPS, JJ는 Penn Treebank Project에서 사용된 품사태그들[21]로 각각 단수명사, 복수명사, 단수고유명사, 복수고유명사, 형용사에 대응하는 품사 태그들이다.

3. Experimental Results

그림 1은 ACL 데이터셋을 사용하여 용어중요도 수식(식 1)의 f와 g에 용어집합 강도 함수들(식 3, 4)의 서로 다른 결합을 적용한 용어추출 성능을 비교 제시한 것이다. 그림 1의 w_{div}, w_{sub}, w_{add}, w_{mul}은 차례대로 식 1에 제시한 네 유형의 용어중요도 방법들 $w_{div}(t,f,g)$, $w_{sub}(t,f,g)$, $w_{add}(t,f,g)$, $w_{mul}(t,f,g)$에 해당한다. 예를 들어 그림 1의 좌측 상단은 $w_{div}(t,f,g)$의 f와 g에 식 3, 4의 26개 함수들 중 임의의 두 함수를 대체하여 생성될 수 있는 총 650개 서로 다른 용어중요도 방법들(x축에 해당)의 성능을 오름차순으로 표시한 그래프이며, 그림 1의 좌측 하단은 $w_{add}(t,f,g)$로부터 생성 가능한 총 325개 용어중요도 방법들(x축에 해당)의 성능을 표시한 그래프에 해당한다. 우측 두 그래프들도 유사하게 해석될 수 있다. 그림 1에서 점선은 C-value 방법의 성능에 해당한다. 그림 2는 그림 1에 적용된 것과 같은 성능 평가를 GENIA 데이터셋에 대해 수행하여 얻은 결과이다.

그림 1의 ACL 데이터셋에 있어 베이스라인보다 높은 성능을 보인 용어집합 함수 결합 경우의 비율은 w_{div}, w_{sub}, w_{add}, w_{mul}의 네 유형에 대해 각각 19%, 20%, 55%, 40%였고, 그림 2의 GENIA 데이터셋에 대해서는 30%, 31%, 35%, 31%였다. 이는 내부 및 외부용어집합 강도 함수의 결합은 많은 경우에 있어 C-value의 성능을 저하시킬 수 있으나, 내부 및 외부용어집합 강도 함수들을 적절히 선택하여 결합하는 경우, C-value 방법보다 높은 성능을 보이며, 식 1에 제시된 네 유형의 각 용어중요도 수식들은 그 정도의 차이는 있으나 용어추출 성능 향상에 긍정적 효과를 만들어 낼 수 있음을 의미한다.

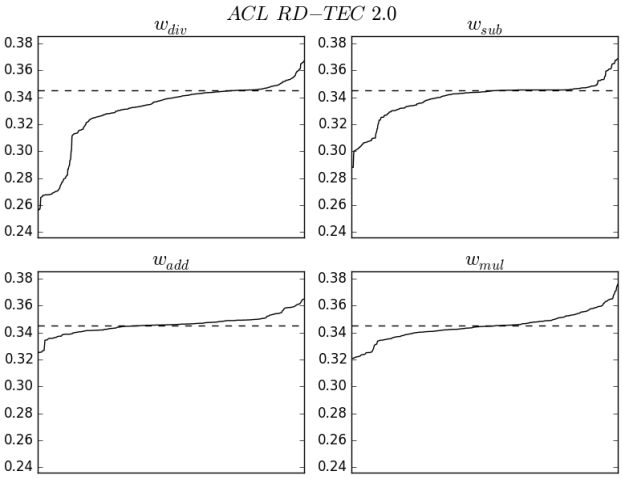

Fig. 1. Performance in AvP of four types of term extraction methods against ACL dataset. w_{div}, w_{sub}, w_{add}, w_{mul} indicate the corresponding methods in Eq. 1. Dotted lines indicate C-value method.

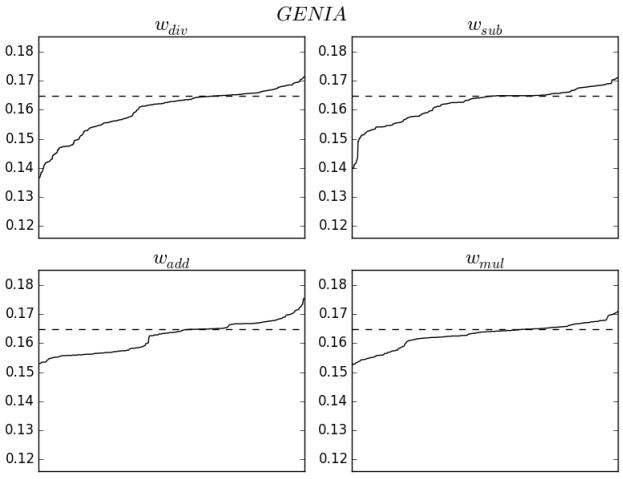

Fig. 2. Performance in AvP of four types of term extraction methods against GENIA dataset. w_{div}, w_{sub}, w_{add}, w_{mul} indicate the corresponding methods in Eq. 1. Dotted lines indicate C-value method.

제안된 용어중요도 방법에 있어 내부용어집합 및 외부용어집합 강도 함수 적용의 개별 효과를 비교하기 위해 식 1의 $w_{div}(t,f,g)$에서 함수 g를 1로 설정하고 함수 f에 식 3 혹은 식 4의 각 함수를 적용한 용어추출 성능을 그림 3에 제시하였다. 그림에서 가로축은 식 3, 4의 각 함수에 해당하며 마커 'x'와 'o'는 각 함수의 용어집합으로 각각 내부용어집합에 해당하는 B(t,C)와 외부용어집합에 해당하는 P(t,C)를 적용한 것들에 해당한다. ACL 데이터셋의 경우 내부용어집합 강도 함수를 통한 용어추출이 대부분의 경우 외부용어집합 강도 함수를 사용한 용어추출에 비해 높은 성능을 보였으나 GENIA 데이터셋의 경우 그 반대 결과를 보여 내부 및 외부용어집합 강도 함수의 단일 적용에 따른 명확한 차이를 발견하기 어려웠다. 그러나 두 데이터셋에 공통적으로, 내부용어집합 강도 함수 적용에 있어 총합, 최대값, 평균 등의 빈도수 대표값을 그대로 사용하는 대신 최소값을 차감하여 사용하는 것이 성능 상승 효과가 있음을 알 수 있다. 또한, 식 3, 4의 26개 강도 함수 중 내부용어집합에 기반한 dgn_B 및 dhn_B 함수들이 ACL 및 GENIA 데이터셋에 공통적으로 가장 높은 용어추출 성능을 보였다.

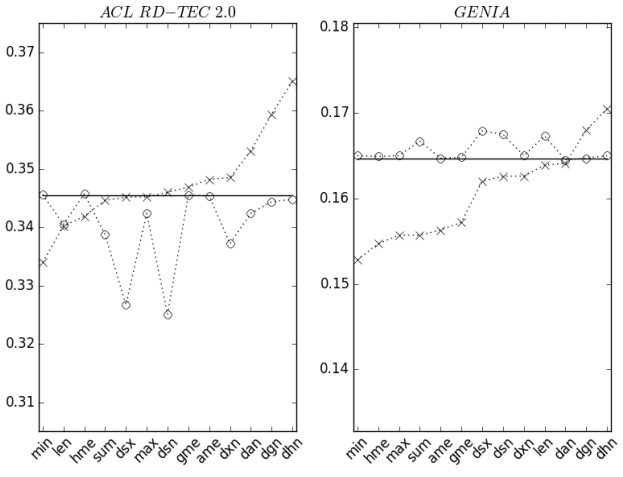

Fig. 3. Performance comparison in AvP of term extraction methods between using inner-term set and using outer-term set. Marker 'x' indicates the methods that applied B(t,C) to f with g set to 1 in $w_{div}(t,f,g)$ of Eq. 1 over various term-set functions. Marker 'o' means the same methods using P(t,C) instead of B(t,C). Solid lines indicate C-value method.

그림 4는 그림 3에 제시한 내부 및 용어집합의 단독 사용을 통한 용어추출 방법들이 내부 및 외부용어집합 강도 함수를 추가 결합함으로써 그 용어 추출 성능에서 향상이 가능한지를 비교 제시한 것이다. 그림에서 마커 'x'는 내부 및 외부용어집합을 단독 적용한 용어추출 성능 중 높은 AvP 수치에 해당한다. ACL 데이터셋의 경우 마커 'x'는 min, len, hme를 제외하면 모두 내부용어집합 단독 적용에 해당하고 GENIA 데이터셋의 경우 dgn, dhn을 제외하면 모두 외부용어집합 단독 적용을 통한 용어추출 성능에 해당한다. 마커 'o'는 마커 'x'에 해당하는 용어집합 강도 함수값에 식 3, 4의 강도 함수를 식 1의 네 유형 중 하나로 추가 결합한 경우의 최고 성능을 표시한 것이다. 그림을 통해 용어집합 강도 함수를 단독 사용한 용어추출 방법들은 추가 강도 함수를 결합함으로써 그 용어추출 성능 향상이 가능함을 알 수 있다.

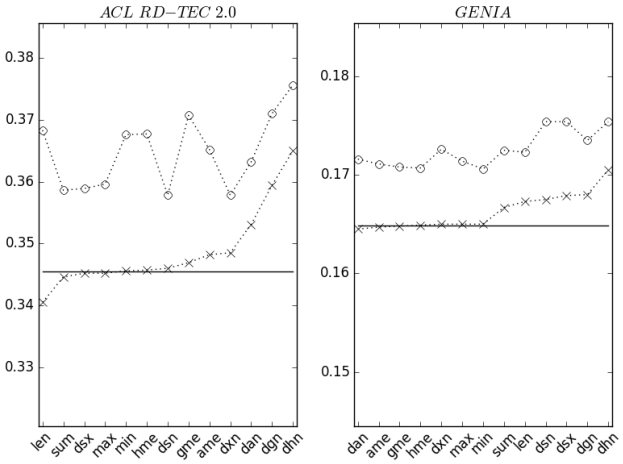

Fig. 4. Performance comparison in AvP of term extraction methods between using single term-set function and combining two term-set functions. Marker 'x' indicates the better among the two methods that applied B(t,C) or P(t,C) to f and g set to 1 in $w_{div}(t,f,g)$ of Eq. 1. Marker 'o' means the best methods that combined one of functions in Eq. 3 and 4 to the corresponding method indicated by 'x'. Solid lines indicate C-value method.

그림 5는 제안된 방법에서 상위 AvP 성능을 보인 100개 용어중요도 수식들(그림 5의 x축에 해당)의 성능을 전체 용어(마커 'o')와 합성 용어(마커 'x') 집합에 대해 구분하여 표시한 것으로 합성 용어는 하나의 단어로 구성된 단일 용어를 제외한 길이 2이상의 용어를 의미한다. C-value의 경우 전체 용어와 합성 용어 집합의 성능은 각각 실선과 점선에 해당한다. 그림을 통해 상위 100개 방법들은 ACL 데이터셋과 GENIA 데이터셋 모두 C-value 대비 성능 향상을 보였으며 합성 용어 집합으로 한정할 경우 성능 향상 폭이 더 컸다.

그림 5에 제시한 상위 100개 방법들은 ACL 및 GENIA 데이터셋에 대해 각각 다른 수식들일 수 있다. 표 1은 ACL 및 GENIA 데이터셋에 대해 상위 성능을 보인 그림 5의 각 100개 방법들의 교집합을 추출하여 해당 방법들 중 내부 및 외부용어집합 활용 빈도 및 용어집합 강도값들의

결합 함수 유형 빈도를 표시한 것이다. 두 데이터셋에 대해 공통적으로 상위 100 순위에 포함된 방법들은 총 61개였고, 이 중 60개 방법들은 dhn_B 혹은 dgn_B를 f 혹은 g에 적용한 방법들이었으며, $w_{div}(t,f=dhn_B,g=max_P)$이 두 데이터셋에서의 순위 기준으로 가장 좋은 성능을 보였다.

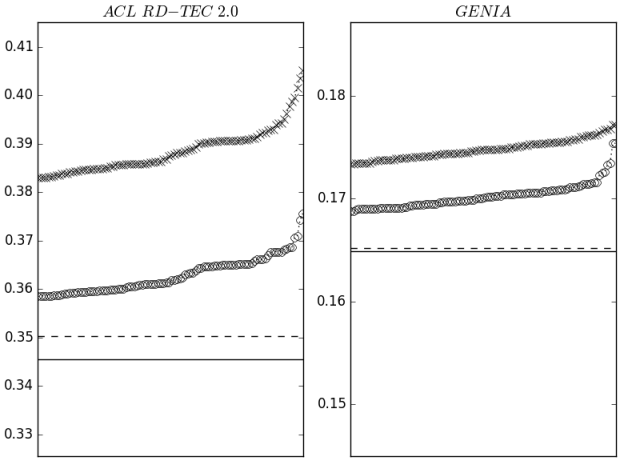

Fig. 5. Performance in AvP of highly-ranked 100 term extraction methods. Solid and dotted lines indicate C-value method. Maker 'o' and solid lines are for all terms, marker 'x' and dotted lines for complex terms.

표에서 BP는 식 1의 함수 f와 g에 각각 내부용어집합 강도 함수와 외부용어집합 강도 함수를 적용한 용어중요도 방법을 의미하며 BB는 f와 g에 모두 내부용어집합 강도 함수를 적용한 방법에 해당한다. PB 및 PP도 유사하게 해석될 수 있다. $w_{div}(t,f,1)$은 식 1의 $w_{div}(t,f,g)$에서 함수 g를 1로 설정한 용어중요도 방법에 해당한다. B와 P는 $w_{div}(t,f,1)$에서 함수 f에 각각 내부용어집합 강도 함수 및 외부용어집합 강도 함수를 적용한 방법에 해당한다. 예를 들어 표에서 BP 행의 w_{div} 열에 해당하는 값 7은 $w_{div}(t,f,g)$ 용어중요도 수식에서 함수 f에 내부용어집합 강도 함수를, 함수 g에 외부용어집합 기반 강도 함수를 적용한 방법들의 개수이다.

Table 1. Analysis of top term extraction methods.

	$w_{div}(t,f,1)$	w_{div}	w_{sub}	w_{add}	w_{mul}
P	0	N/A	N/A	N/A	N/A
PP	N/A	0	0	0	0
B	1	N/A	N/A	N/A	N/A
BB	N/A	2	3	0	6
BP	N/A	7	15	20	7
PB	N/A	0	0		

표를 통해 알 수 있는 것처럼 두 데이터셋에 대해 상위 100 순위 내에 공통적으로 포함된 61개 방법들 중 외부용어집합 강도 함수(들)만을 단독 혹은 결합 활용하는 방법(표에서 P 혹은 PP)은 발견되지 않았으며 내부용어집합 강도 함수(들)을 단독(B) 혹은 결합(BB) 활용하는 방법들은 일부 포함되었다. 특히 61개 방법 중 36%를 차지하는 22(=7+15)개 방법들은 f와 g에 각각 내부용어집합과 외부용어집합을 각각 적용하여 w_{div} 혹은 w_{sub}의 함수 결합을 사용한 수식들이었고, w_{add} 혹은 w_{mul}의 수식을 통해 내부용어집합과 외부용어집합을 결합한 방법들은 27(=20+7)개로 44%를 차지하였다. 이는 내부 및 외부용어집합으로부터의 강도값들을 결합 활용하는 방법이 용어중요도 계산에서 도움이 됨을 시사한다.

V. Conclusions

본 논문에서는 용어 간 포함 관계에 기반한 내부용어집합과 외부용어집합이 용어추출에 미치는 영향을 다루었으며 C-value와 결합하여 새로운 용어중요도 방법을 제시하였다. ACL 및 GENIA 데이터셋을 사용한 실험을 통해 내부 및 외부용어집합 강도 함수 결합의 네 가지 유형에 대한 용어추출 성능 차이를 분석 제시하였고, 내부 및 외부용어집합 강도 함수들의 적절한 결합 사용이 내부 혹은 외부용어집합 강도 함수를 단독 적용하는 방법의 용어추출 성능을 향상시킬 수 있음을 보였다.

본 논문에서는 용어추출 대상이 되는 도메인 코퍼스만을 사용하여 내부 및 외부용어집합을 추출하고 용어집합 내 각 용어의 빈도수를 수집 활용하였으나, 향후 일반 코퍼스를 추가 활용한다면 일반 코퍼스 기반 내부 및 외부용어집합 정보의 결합 사용을 통해 보다 향상된 용어추출 시도가 가능할 것이다.

REFERENCES

[1] N. Astrakhantsev, "ATR4S: Toolkit with State-of-the-art Automatic Terms Recognition Methods in Scala," CoRR abs/1611.07804, 2016.

[2] Z. Zhang, J. Gao, and F. Ciravegna, "SemRe-Rank: Incorporating Semantic Relatedness to Improve Automatic Term Extraction Using Personalized PageRank," CoRR abs/1711.03373, 2017.

[3] T. Koutropoulou, and E. Gallopoulos, "TMG-BoBI: Generating

Back-of-the-Book Indexes with the Text-to-Matrix-Generator," Proceedings of 10th International Conference on Information, Intelligence, Systems and Applications, 2019.

[4] Z. Wu, Z. Li, P. Mitra, and C. Giles, "Can back-of-the-book indexes be automatically created?," Proceedings of the 22nd ACM International Conference on Information and Knowledge Management, pp. 1745-1750, 2013.

[5] N. Simon, and V. Keselj, "Automatic Term Extraction in Technical Domain using Part-of-Speech and Common-Word Features," Proceedings of the ACM Symposium on Document Engineering, 2018.

[6] G. Petasis, V. Karkaletsis, G. Paliouras, A. Krithara, and E. Zavitsanos, "Ontology Population and Enrichment: State of the Art," Knowledge-Driven Multimedia Information Extraction and Ontology Evolution, pp. 134-166, 2011.

[7] M. Asim, M. Wasim, M. Khan, W. Mahmood, and H. Abbasi, "A survey of ontology learning techniques and applications," Database, Vol. 2018, 2018.

[8] K. Frantzi, S. Ananiadou, and H. Mima, "Automatic recognition of multi-word terms:. the c-value/nc-value method," International Journal on Digital Libraries, Vol 3, No. 2, pp. 115–130, 2000.

[9] G. Bordea, P. Buitelaar, and T. Polajnar, "Domain-independent term extraction through domain modelling," Proceedings of the 10th International Conference on Terminology and Artificial Intelligence, 2013.

[10] S. Rose, D. Engel, N. Cramer, and W. Cowley, "Automatic keyword extraction from individual documents," Text Mining: Applications and Theory, John Wiley & Sons Ltd, 2010.

[11] H. Nakagawa, and T. Mori, "A Simple but Powerful Automatic Term Extraction Method," COLING-02: COMPUTERM 2002: Second International Workshop on Computational Terminology, 2002.

[12] N. Astrakhantsev, "Methods and software for terminology extraction from domain specific text collection," Ph.D. thesis, Institute for System Programming of Russian Academy of Sciences, 2015.

[13] Z. Zhang, J. Gao, and F. Ciravegna, "JATE 2.0: Java Automatic Term Extraction with Apache Solr," Proceedings of the Tenth International Conference on Language Resources and Evaluation, 2016.

[14] K. Meijer, F. Frasincar, and F. Hogenboom, "A semantic approach for extracting domain taxonomies from text," Decision Support Systems, Vol. 62, pp. 78-93, 2014.

[15] K. Ahmad, L. Gillam, and L. Tostevin, "University of Surrey Participation in TREC8: Weirdness Indexing for Logical Document Extrapolation and Retrieval (WILDER)," Proceedings of The Eighth Text REtrieval Conference, 1999.

[16] J. Ventura, C. Jonquet, M. Roche, and M. Teisseire, "Combining c-value and keyword extraction methods for biomedical terms extraction," International Symposium on Languages in Biology and Medicine, pp. 45–49, 2013.

[17] B. QasemiZadeh, and A. Schumann, "The ACL RD-TEC 2.0: A Language Resource for Evaluating Term Extraction and Entity Recognition Methods," Proceedings of the Tenth International Conference on Language Resources and Evaluation, 2016.

[18] J. Kim, T. Ohta, Y. Tateisi, and J. Tsujii, "GENIA corpus - a semantically annotated corpus for bio-textmining," ISMB (Supplement of Bioinformatics), pp. 180-182, 2003.

[19] A. Sajatovic, M. Buljan, J. Snajder, and B. Dalbelo, "Basic: Evaluating Automatic Term Extraction Methods on Individual Documents," Proceedings of the Joint Workshop on Multiword Expressions and WordNet, pp. 149-154, 2019.

[20] SpaCy. *https://spacy.io/*

[21] M. Marcus, B. Santorini, and M. Marcinkiewicz, "Building a Large Annotated Corpus of English: The Penn Treebank," Computational Linguistics, Vol. 19, No. 2, pp. 313-330, 1993.

Authors

In-Su Kang received his bachelor's degree from Kyungpook National University in 1995, and master's and doctoral degrees from POSTECH, in 1999, and 2006, respectively. He is an associate professor at the Department of Computer Science, Kyungsung University. He is interested in natural language processing and information retrieval.