Journal of The Korea Society of Computer and Information
Vol. 22 No. 9, pp. 17-24, September 2017
www.ksci.re.kr
https://doi.org/10.9708/jksci.2017.22.09.017

A Term Importance-based Approach to Identifying Core Citations in Computational Linguistics Articles

In-Su Kang*

Abstract

Core citation recognition is to identify influential ones among the prior articles that a scholarly article cite. Previous approaches have employed citing-text occurrence information, textual similarities between citing and cited article, etc. This study proposes a term-based approach to core citation recognition, which exploits the importance of individual terms appearing in in-text citation to calculate influence-strength for each cited article. Term importance is computed using various frequency information such as term frequency(tf) in in-text citation, tf in the citing article, inverse sentence frequency in the citing article, inverse document frequency in a collection of articles. Experiments using a previous test set consisting of computational linguistics articles show that the term-based approach performs comparably with the previous approaches. The proposed technique could be easily extended by employing other term units such as n-grams and phrases, or by using new term-importance formulae.

▸Keyword: Core Citation Recognition, Term Frequency, Citation, Citing Text

I. Introduction

핵심인용인식(core citation recognition)은 한 편의 논문이 인용하는 선행 논문들 중에서 핵심 선행 논문(들)을 구별해 내는 작업으로, 서로 다른 피인용논문(cited paper)들은 인용논문(citing paper)에 미치는 영향의 정도가 동일하지 않다는 가정에 기초한다[1, 2]. 예를 들어 논문 P가 인용하는 선행 논문들 중에서 어떤 논문은 P에서 사용한 데이터 처리 툴킷의 출처를 밝히기 위해 인용되었을 수 있으며(아래 인용텍스트 예 1 참조), 다른 논문은 P에서 제안하는 방법론의 기반이 된 최초 방법론을 기술하기 위해 인용되었을 수 있다(아래 인용텍스트 예 2 참조).

- 인용텍스트 예 1: *"실험에 사용된 논문 텍스트 파일로부터 참고문헌 목록을 추출하기 위해 ParsCit 툴킷(Councill et al., 2008)을 사용하였다."*

- 인용텍스트 예 2: *"이 논문에서는 Smith 등(Smith et al., 2011)이 고안한 부트스트래핑 방법의 재현율 향상을 위해 대용량 동질 코퍼스 기반의 연계 자질 추출을 시도한다."*

핵심인용인식은 저널 및 연구자 영향력 지수 계산을 단순 피인용 횟수 대신 핵심 인용 강도에 기반한 방식으로 개선하거나, 인용 기반 논문 요약[3]에서 사용되는 인용텍스트들을 핵심 인용 논문에 대한 인용텍스트인지 여부에 따라 차별화하여 요약문 생성에 활용하는 등 다양한 응용에 이용될 수 있다.

핵심인용인식의 기존 연구에서는 인용논문과 피인용논문 간의 인용 관계가 핵심인용인지 여부를 결정하기 위해 SVM 분류기를 적용하거나[2, 4], 인용논문에서의 피인용논문의 핵심인용 강도를 추정하기 위해 Support Vector Regression을 시도하였다[1]. Chakraborty와 Narayanam은 그래프 상의 인접성

• First Author: In-Su Kang, Corresponding Author: In-Su Kang
*In-Su Kang (dbaisk@ks.ac.kr), Dept. of Computer Science & Engineering, Kyungsung University
• Received: 2017. 05. 31, Revised: 2017. 06. 18, Accepted: 2017. 09. 04.
• This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (NRF-2015R1D1A1A01060489).

에 기반하여 레이블을 전파하는 방식의 그래프 기반 반교사 방법을 인용강도 추정에 적용하였다[5].

핵심인용인식의 학습 자질 표현을 위해 기존 연구에서는 대표적으로 (1) 인용논문 내에서 피인용논문에 대해 기술된 인용텍스트의 출현 정보(출현 횟수, 논문 내 출현 위치 등), (2) 인용논문과 피인용논문 간에 계산되는 내용 유사도, 출판년도 차이, 저자공유 여부 등의 정보, (3) 피인용논문 자체에 대한 피인용 횟수나 페이지랭크 값 등의 정보를 사용하였다. 인용논문 내에서 하나의 피인용논문에 대해 기술된 인용텍스트는 인용논문 내에서 1회 이상 출현할 수 있다. 기존 연구에서는 인용텍스트 단위의 출현 정보를 핵심인용결정의 주요 자질로 활용하였다.

본 연구에서는 인용텍스트 내에 출현한 개별 용어의 중요도에 기반하여 계산되는 인용텍스트의 인용강도를 기계학습 기반 핵심인용인식에 활용하는 용어 기반 방법(term-based method)을 제안한다. 인용텍스트에 출현한 개별 용어 t의 중요도 계산을 위해 t의 인용텍스트 내 빈도, t의 인용논문 내 빈도, t의 인용논문 내 역문장빈도, 외부 문서 집합에서의 t의 역문서빈도 등을 활용한다. 실험에서는 전산언어학 분야 논문들로 구성된 기존 평가집합을 사용하여 용어 기반 방법과 기존 핵심인용인식 방법들의 SVM 분류 성능을 비교 제시한다.

논문의 구성은 다음과 같다. 2장에서는 관련 연구를 기술한다. 3장에서는 용어 중요도에 기반한 핵심인용인식 방법에 대해 기술한다. 4장에서는 제안된 방법의 성능 평가에 대해 기술하고 5장에서 결론을 맺는다.

II. Related Works

Wan과 Liu는 인용논문과 피인용논문 사이에 1~5 사이의 인용강도(citation strength)를 추정하기 위해 Support Vector Regression을 적용하였으며, 인용논문 내에서 인용텍스트의 출현 횟수, 위치, 평균 문장 길이, 인접 인용문 간의 평균 최소 거리와, 인용논문-피인용논문 간 저자공유 여부, 연도차이의 6개 자질들을 사용하였다[1]. 이들의 연구에서는 수작업 구축된 정답 데이터셋을 통해 학습된 인용강도 추정 모델이 논문 영향도 및 저자 영향도 계산에서 긍정적 효과가 있는 것으로 보고하였다.

Zhu 등은 영향력 있는 인용(influential citation)인지 여부를 결정하기 위해, 인용텍스트 출현 정보, 인용논문-피인용논문 간 유사도, 통제 어휘에 기반한 인용텍스트 내 특정 의미 유형 발현 정도 등을 포함하는 총 38개 자질을 SVM 분류 성능에 기반하여 평가한 후 최종적으로 4개 자질을 선택 제시하였다[2]. 그 자질들은 인용텍스트가 출현한 총 횟수 및 섹션 개수, 피인용논문 제목과 인용논문 핵심 섹션 간 내용 유사도, 저자공유 여부이다.

Valenzuela 등은 중요 인용(important citation)을 분별하기 위해 SVM과 Random Forest 분류기를 적용하였으며, 인용텍스트 출현 정보, 인용논문-피인용논문 간 초록유사도, 인용이표나 그림의 캡션에 출현하는지 여부, 저자공유 여부, 페이지랭크 등 총 12개 자질을 사용하였다[4]. 특히 인용문을 명시적 인용표시(예: (Smith et al., 2007) 혹은 [3,4] 등)의 출현 여부에 따라 직접 인용문과 간접 인용문으로 구분하여 인용문 출현 자질을 추출하였다. 간접 인용문은 명시적 인용 표시가 사용되지 않고 피인용논문의 저자명이나 피인용논문에서 다루는 알고리즘 명칭 등이 사용되어 간접적으로 피인용논문에 대해 기술된 문장을 의미한다. Valenzuela 등은 간접인용문 출현 정보가 핵심인용인식의 전체 성능에 기여하는 바가 크지 않았다고 보고하였으나, 본 논문에서는 Valenzuela 등의 시도로부터 영향을 받아 개별 용어가 피인용논문에 대한 인용강도 표현에 활용될 수 있을 것이라는 생각을 하게 되었다.

Chakraborty와 Narayanam은 피인용논문의 인용강도를 결정하기 위해 그래프 기반 반교사(semi-supervised) 방법을 사용하였으며, 시도된 다양한 자질 중 인용텍스트 단위의 출현 횟수 자질 유형이 정답 레이블과의 상관계수가 가장 높았다고 보고하였다[5].

Akram도 중요 인용 분류를 위해 인용텍스트의 출현 정보를 사용하였다. 특히 피인용논문을 기술할 때 흔히 사용되는 단서어구들(예: following, according to, extending 등 135개)을 미리 준비해 두고, 단서 어구 목록과 인용텍스트 사이에 계산되는 코사인 유사도를 기계학습 자질 중 일부로 사용하였다[6].

전술한 기존 연구들은 인용텍스트 단위의 출현 정보를 핵심인용인식의 유용한 자질로 사용하고 있다. 기존 연구와 달리 본 논문에서는 인용텍스트 단위의 출현 정보라기보다는 인용텍스트 내에 출현한 개별 용어의 출현 정보에 기반하여 계산된 인용강도를 활용하여 핵심인용논문 결정을 시도한다.

본 논문의 이러한 시도와 관련하여 인용텍스트에 출현한 개별 용어를 다룬 대표적 연구로는 전술한 Zhu의 연구가 있는데[2], 이 연구에서는 특정 의미 유형의 어휘 목록을 미리 구축해 두고, 목록 내 어휘 중 인용텍스트에 출현한 어휘의 수에 기반하여 해당 의미 유형이 인용텍스트에서 발현된 정도를 표현하고자 하였다. 이들의 연구에서는 다양한 의미 유형들(relevant, recent, extreme, comparative, positive, strong, active, emotional, sentimental) 각각에 대해 전술한 방법을 시도하였다.

III. Term-based Method

본 연구에서는 인용논문 d에서 인용한 특정 피인용논문 p가 d의 핵심인용인지 여부를 결정하기 위해 d 내에서 p에 대해 기술된 인용텍스트의 인용강도를 기계학습 자질로 사용한다. 이를 위해 인용텍스트에 출현한 용어의 빈도 정보에 기반하여 p에 대한 인용 강도를 계산하는 수식들을 식 1~6과 같이 정의한다. 식 1~6에서 p는 인용논문 d에서 인용된 특정 피인용논문

www.kci.go.kr

을 의미하며 핵심인용인지 여부가 결정되어야 할 대상 논문이다. c는 d 내에서 p에 대해 기술된 하나의 인용텍스트이다. 일반적으로 피인용논문에 대한 인용텍스트는 인용논문 내의 여러 섹션 위치(예: 서론, 기존연구, 실험 등)에 출현할 수 있으며 하나의 섹션 내에 다수 출현할 수도 있다. $C(p)$는 d 내에서 발견되는 p에 대한 모든 인용텍스트들의 집합으로 정의하고, $T(c)$는 하나의 인용텍스트 c에 출현한 용어의 집합으로 정의한다.

$$s_{qtf}(p) = \sum_{c \in C(p)} \sum_{t \in T(c)} log(1 + tf(t,c)) \quad \text{식 1}$$

$$s_{isf}(p) = \sum_{c \in C(p)} \sum_{t \in T(c)} log\left(1 + \frac{M}{sf(t,d)}\right) \quad \text{식 2}$$

$$s_{qtfisf}(p) = \sum_{c \in C(p)} \sum_{t \in T(c)} log(1 + tf(t,c)) \times log\left(1 + \frac{M}{sf(t,d)}\right) \quad \text{식 3}$$

$$s_{tf}(p) = \sum_{c \in C(p)} \sum_{t \in T(c)} log(1 + tf(t,d)) \quad \text{식 4}$$

$$s_{idf}(p) = \sum_{c \in C(p)} \sum_{t \in T(c)} log\left(1 + \frac{N}{df(t)}\right) \quad \text{식 5}$$

$$s_{tfidf}(p) = \sum_{c \in C(p)} \sum_{t \in T(c)} log(1 + tf(t,d)) \times log\left(1 + \frac{N}{df(t)}\right) \quad \text{식 6}$$

식 1~6은 피인용논문 p에 대한 인용텍스트가 다수 출현하는 경우, 각 인용텍스트의 인용강도들의 합을 p의 인용강도로 고려한 것이다. 개별 인용텍스트에 대한 인용강도는 인용텍스트 내 출현 용어들의 *qtf, isf, qtfisf, tf, idf, tfidf* 값들에 기반하여 계산된다. t는 하나의 인용텍스트 c 내에 출현한 특정 용어이고, $tf(t,c)$는 c 내에 출현한 용어 t의 빈도수(term frequency), $sf(t,d)$는 인용논문 d 전체에서 용어 t가 출현한 문장의 개수(sentence frequency), M은 인용논문 d 내의 문장의 총 개수, $tf(t,d)$는 인용논문 d 전체에서 용어 t가 출현한 빈도수, $df(t)$는 외부 문서 집합에서 용어 t가 출현한 문서의 개수(document frequency), N은 외부 문서 집합 내 총 문서의 개수이다.

식 1~3은 Allan 등[7]이 시도한 문장적합도 수식의 변형들로 적합문장검색(relevant sentence retrieval) 관점에서, 피인용논문에 대한 인용텍스트 c를 질의로 고려하여 인용논문 d 내에서 질의와 일치하는 적합 문장이 검색되는 상황을 가정하고 계산되는 질의-문장 적합도로 해석할 수 있다. 식 1~3에서, 인용강도는 인용텍스트 내에 출현한 용어들의 중요도로부터 결정되며 용어 t의 중요도는 t의 인용텍스트 내 빈도(query term frequency)나 t의 역문장빈도(inverse sentence frequency), 혹은 그들의 결합에 비례하여 결정된다고 가정한 것이다.

식 4~6은 추출식 문서 요약[8, 9, 10]에서 사용되는 문장 중요도 수식들을 피인용논문의 인용강도 표현을 위해 도입한 수식들이다. 이는 문서 요약 관점에서 핵심인용결정 대상이 되는 각 피인용논문의 인용텍스트를 문서 요약의 추출 대상 문장으로 해석한 것에 해당한다. 식 4~6에서도 식 1~3에서처럼 인용강도는 인용텍스트 내 출현 용어들의 중요도로부터 계산된다고 가정한다. 그러나 식 4~6의 경우 용어 t의 중요도는 t의 인용논문 내 빈도(term frequency)나 t의 역문헌빈도(inverse document frequency), 혹은 그들의 결합에 비례하여 결정된다고 가정하고 있다.

이후 설명에서는 인용논문 d 내의 7번째 피인용논문을 p_7이라고 가정할 때 p_7에 대해 식 1, 식 2, 식 4, 식 5의 인용강도들을 계산하는 예를 보인다. 아래 예에서는 p_7에 대한 인용문이 d 내에서 1회만 출현한 것으로 가정하였고, 그 인용문을 7번 피인용논문에 대한 1번째 인용문이라는 의미로 c_{71}로 표기하였다. 또한 인용강도 계산을 위해 인용문 내의 명사 용어들만 사용하는 것으로 가정하였다.

$C(p_7) = \{c_{71}\}$
c_{71} = "데이터로는 CMU 데이터[7]를 사용하였다."
$T(c_{71}) = \{CMU, 데이터, 사용\}$

식 1에 해당하는 $s_{qtf}(p_7)$의 계산을 위해서는 $T(c_{71})$ 내 각 용어의 인용텍스트 c_{71} 내에서의 tf 값이 필요하다. 그러한 tf 값들은 다음과 같다.

$tf(CMU,c_{71})=1,\ tf(데이터,c_{71})=2,\ tf(사용,c_{71})=1$

그러면, $s_{qtf}(p_7)$은 아래와 같이 계산된다.

$s_{qtf}(p_7) = log(1+1)+log(1+2)+log(1+1)$

식 2에 해당하는 $s_{isf}(p_7)$의 계산을 위해서는 $T(c_{71})$ 내 각 용어 t에 대해 인용논문 d 내에서의 문장 빈도 값 $sf(t,d)$가 필요하다. 그러한 sf 값들이 다음과 같다고 가정하자.

$sf(CMU,d)=5,\ sf(데이터,d)=11,\ sf(사용,d)=19$

인용논문 d 내의 전체 문장수 M=135라면, $s_{isf}(p_7)$은 아래와 같이 계산된다.

$s_{isf}(p_7) = log(1+135/5)+log(1+135/11)+log(1+135/19)$

식 4에 해당하는 $s_{tf}(p_7)$의 계산을 위해서는 $T(c_{71})$ 내 각 용어의 인용논문 d 내에서의 tf 값이 필요하다. 그러한 tf 값들이 다음과 같다고 가정하자.

www.kci.go.kr

$tf(CMU,d)=5$, $tf(데이터,d)=12$, $tf(사용,d)=19$

그러면, $s_{tf}(p_7)$은 아래와 같이 계산된다.

$s_{tf}(p_7) = log(1+5)+log(1+12)+log(1+19)$

식 5에 해당하는 $s_{idf}(p_7)$의 계산을 위해서는 $T(c_{71})$ 내 각 용어 t에 대해 외부 문서 집합에서의 문헌 빈도 값인 $df(t)$가 필요하다. 그러한 df 값들이 다음과 같다고 가정하자.

$df(CMU)=10$, $df(데이터)=10^2$, $df(사용)=10^3$

외부 문서 집합 내 총 문서의 수 $N=10^6$이라면, $s_{idf}(p_7)$은 아래와 같이 계산된다.

$s_{idf}(p_7)$
$= log(1+10^6/10)+log(1+10^6/10^2)+log(1+10^6/10^3)$

IV. Experiments

핵심인용인식 방법의 성능 평가를 위해 기존 연구의 평가집합[1]을 사용하였다. [1]에 따르면 이 평가집합은 문서요약, 감성분석 등의 자연어처리 분야를 다루는 40편 논문 집합을 대상으로 각 논문이 인용하는 피인용논문에 대해 인용강도(1,2,3,4,5 중 하나의 값)를 두 명의 작업자가 각각 수작업 부여하여 구축된 것이다. 인용강도 1,2,3은 인용논문에 미친 영향이 크지 않은 피인용논문에 부여되어 있고, 인용강도 4,5는 인용논문에 중요한 영향을 미친 피인용논문에 부여되어 있다. 본 연구에서는 전술한 평가집합을 핵심인용 분류에 사용하기 위해, 평균 인용강도 4~5인 피인용논문들을 핵심인용 클래스(core citation class)로, 평균 인용강도 4 미만인 피인용논문들을 비핵심인용 클래스(non-core citation class)로 설정하여 사용하였다. 표 1은 실험에 사용된 평가집합의 클래스 분포를 보인 것이다.

Table 1. Statistics of the Test Set

Class	Number of Samples
Core citation class	81
Non-core citation class	739
Total	820

기계학습 방법으로 scikit-learn[11]에서 구현된 선형, 비선형 SVM 방법들을 사용하였다. 선형 SVM으로 Libsvm[12]과 Liblinear[13] 기반 구현들을 사용하였고 비선형 SVM으로 Polynomial, RBF 커널을 사용하였다. 실험 결과에서 Libsvm 기반 선형 SVM은 SMOlinear로 표기한다.

Table 2. Grid Search Ranges for SVM Parameters

Parameter	Range
C	2^{-5}, 2^{-3}, ..., 2^{15}
γ	2^{-15}, 2^{-13}, ..., 2^{3}
r	0, 1
d	1, 2, 3

기계학습 방법의 평가를 위해 5-fold 교차검증을 사용하였고, 각 fold 내 학습 집합에 대한 최적 SVM 파라미터 결정을 위해서는 3-fold 교차검증을 통한 격자탐색(grid search)을 적용하였다. SVM 커널 파라미터의 격자탐색 범위는 표 2와 같으며, C, γ의 경우 Hsu 등[14]이 권고한 탐색 범위를 따른 것이다. 표 2에서 r, d는 식 7에 보인 polynomial 커널과 관련된 파라미터들이다. 식 7은 [14]의 polynomial 커널 수식에서 γ=1로 설정한 것이다.

$$K(\mathbf{x}_i,\mathbf{x}_j)=(\mathbf{x}_i^T\mathbf{x}_j+r)^d \qquad \text{식 7}$$

성능 평가 지표로 정확률(precision), 재현율(recall), F1을 사용하였다. 이들 지표는 핵심 및 비핵심 인용클래스에 대해 각각 계산될 수 있다. 핵심인용클래스에 대한 정확률, 재현율, F1을 P_C, R_C, $F1_C$라고 하고 비핵심인용클래스에 대한 정확률, 재현율, F1을 P_N, R_N, $F1_N$이라고 하자. P_C는 시스템이 핵심인용으로 예측한 샘플 중에서 실제 핵심인용클래스에 속하는 것들의 비율로 정의된다. R_C는 실제 핵심인용클래스에 속하는 샘플 중에서 시스템이 핵심인용클래스로 올바르게 예측한 것들의 비율로 정의된다. $F1_C$는 P_C와 R_C의 조화평균이다. P_N은 시스템이 비핵심인용으로 예측한 샘플 중에서 실제 비핵심인용클래스에 속하는 것들의 비율로 정의된다. R_N은 실제 비핵심인용클래스에 속하는 샘플 중에서 시스템이 비핵심인용클래스로 올바르게 예측한 것들의 비율로 정의된다. $F1_N$은 P_N과 R_N의 조화평균이다.

그러나 평가집합 내 핵심인용클래스에 속하는 샘플의 비율은 9.88%로 핵심인용과 비핵심인용 클래스의 분포가 불균형이므로, Zhu[2]의 평가 방식을 따라 특별한 명시가 없는 한 성능 평가 결과는 핵심인용클래스에 대한 정확률, 재현율, F1에 해당하는 P_C, R_C, $F1_C$으로 제시하였다.

제안된 용어 기반 방법의 성능 수준을 비교하기 위한 기존 핵심인용인식 방법들로 2장에 기술한 Wan의 방법[1]과 Zhu의 방법[2]을 직접 구현하여 사용하였다. 다음은 이후 성능 평가 결과에서 사용될 세 가지 핵심인용방법들의 레이블들이다.

- Term-based method(본 논문의 용어 기반 방법)
- Wan's method
- Zhu's method

실험에 사용된 데이터셋 내에는 40편 각 논문에 대해 ParsCit 툴킷[15]을 적용한 분석 결과 파일이 포함되어 있다. ParsCit 툴킷은 텍스트 형식의 논문을 입력으로 받아 논문의

www.kci.go.kr

논리적 구조를 분석하는데 실험에서는 데이터셋 내에 포함된 다음의 ParsCit 분석 결과들을 활용하였다.

- 논문에서 인용한 각 피인용논문에 대한 참고문헌 레코드
- 각 피인용논문에 대한 인용텍스트 및 그 위치 정보
- 논문 내 섹션 제목, 섹션 유형 및 위치 정보

비교 시스템인 Wan과 Zhu의 방법들은 인용텍스트가 출현한 섹션 위치 정보를 사용하는데 이를 위해 논문을 구성하는 섹션들을 ParsCit 툴킷을 통해 인식되는 *introduction, related work, method, evaluation, conclusions*의 5개 섹션 유형들로 구분하여 사용하였다. 특히 섹션 유형 결정을 위해 ParsCit 툴킷이 인식한 섹션 제목에 기반한 다음의 처리를 추가하였다.

- 섹션 제목에 단어 introduction이 포함된 경우 섹션 유형을 *introduction*으로 수정한다. 섹션 제목에 related, relevant, prior, previous 중 하나의 단어가 포함된 경우 섹션 유형을 *related work*로 수정한다. 섹션 제목에 evaluation, result, experiment 중 하나의 단어가 포함된 경우 섹션 유형을 *evaluation*으로 수정한다.

식 2~3의 *sf(t,d)* 계산을 위해서는 간단한 정규표현식을 사용하여 논문 텍스트의 문장 분할을 수행하였다. 식 5~6의 *df(t)* 계산을 위한 외부 문서 집합으로 ACL Anthology 참조 코퍼스[16]를 사용하였고(2008년 3월 25일 버전), 데이터셋 내 논문이 포함된 경우 사전 제거 후 문헌빈도를 계산하였다.

용어 기반 방법에서는 인용텍스트에 출현한 개별 용어에 대한 중요도를 계산할 필요가 있다. 이를 위해 인용텍스트로부터 대상 용어들을 추출해야 하는데 그 추출 절차로 불용어 제거와 스테밍 적용 여부를 비교하는 실험을 먼저 진행하였다. 불용어 제거(stopword removal)는 핵심인용인식에서 의미가 없는 용어들을 제거하기 위해 적용되며 실험에서는 정보검색 분야의 불용어 목록[17]을 사용하였다. 스테밍(stemming)은 굴절 및 파생으로 인한 단어의 변형들을 동일 표현으로 정규화할 목적으로 단어의 어미부를 절단하는 것으로 실험에서는 Porter 스테밍[18]을 적용하였다. 이 실험에서는 불용어 혹은 스테밍 적용 이후 얻어진 세 문자 이상의 용어들을 대상으로 용어 중요도 자질들을 계산하여 Liblinear 커널로 학습 및 분류한 성능을 평가하였고 그 결과는 표 3과 같다. 표 3의 결과로부터 불용어는 제거하고 스테밍은 적용하지 않는 것이 다른 방식들에 비해 성능이 우수하여 이후 실험의 용어 기반 방법에서는 이 방식의 용어 추출 전처리를 적용하였다. 핵심인용인식 문제에서 불용어 제거와 스테밍 적용 유무에 따른 성능 효과 분석과 관련하여서는 향후 추가 연구가 진행될 필요가 있다.

Table 3. Comparing Term Extraction Methods For the Term-based Method Over Stopword Removal and Stemming Using Liblinear SVM

Term Extraction Method	F1
None	0.5546
Stopword removal	0.5593
Stemming	0.5333
Stopword removal + Stemming	0.5254

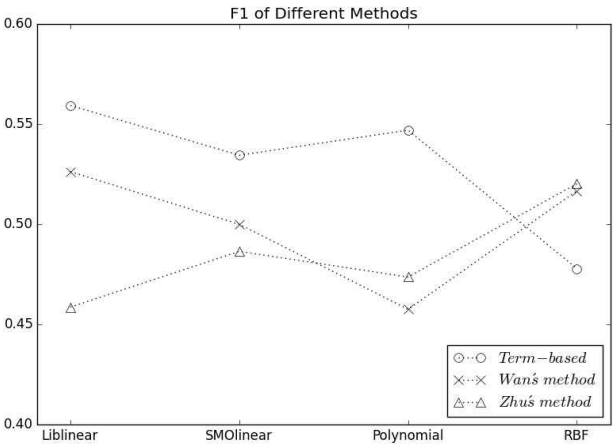

Fig. 1. Performance of Core Citation Recognition Methods Using Different SVM Kernels (Y-axis indicates F1)

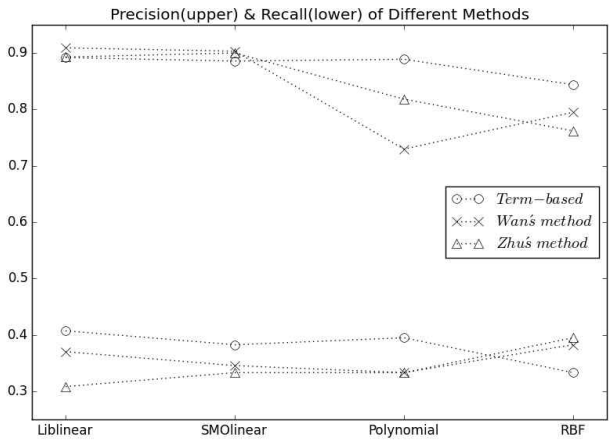

Fig. 2. Precisions and Recalls of Core Citation Recognition Methods Using Different SVM Kernels (Upper and lower graphs correspond to precision and recall figures, respectively)

그림 1은 핵심인용인식 방법들의 F1 성능을 Liblinear, SMOlinear, Polynomial, RBF 커널 각각에 대해 비교 제시한 것이다. RBF 커널의 경우 제안된 용어 기반 방법이 기존 방법들의 성능에 미치지 못하였으나 선형 및 다항식 커널에서는 기존 방법들보다 높은 성능을 보였다. 이러한 결과는 본 논문에서 제안한 식 1~6의 용어 중요도에 기반한 인용강도 표현이 핵심인용과 비핵심인용을 구별하는데 효과가 있음을 의미한다. 즉 피인용논문에 대한 인용텍스트를 적합문장검색에서의 질의나 문서요약에서의 추출 후보 문장으로 고려한 본 논문의 접근법

www.kci.go.kr

이 핵심인용인식 문제에 적용 가능함을 의미한다.

그림 2는 핵심인용인식 방법들의 성능 차이 분석을 위해 각 방법의 정확률과 재현율을 서로 다른 커널들에 대해 비교 제시한 것으로 그림 상단부는 정확률에 해당하고 하단부는 재현율에 해당한다. RBF 커널의 경우를 제외하면 제안된 방법은 Wan 및 Zhu의 방법들에 비해 상대적으로 높은 재현율을 보였으며 선형 커널들에서 정확률 저하가 있었으나 그 폭이 크지 않아, RBF 이외의 SVM 커널들에서 기존 방법들보다 높은 성능을 보인 것으로 판단된다.

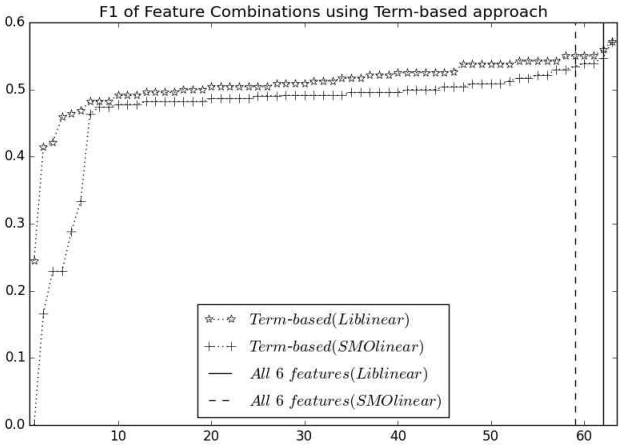

Fig. 3. Comparing Performances of All Feature Combinations of Term-based Approach Using Linear SVMs (Y-axis indicates F1)

그림 3은 용어 기반 방법에서 6개 자질(식 1~6)로부터 생성 가능한 63개 각 자질 조합의 Liblinear 및 SMOlinear 기반 F1 성능을 오름차순으로 비교 제시한 것으로 이를 통해 서로 다른 자질 선택에 따른 핵심인용인식 성능의 변화 추이를 확인할 수 있다. 그림에서 일부 자질 조합들은 현저한 성능 저하를 보였다. 대표적으로 두 선형 SVM 커널에서 공통적으로 최저 성능을 보인 하위 두 개 자질 조합들은 *idf* 혹은 *tfidf* 기반 자질(식 5 혹은 식 6)을 단독 사용한 경우에 해당한다. 그러나 이는 *idf* 계산에 사용된 외부 논문 집합의 규모와 무관하지 않을 수 있으므로 이 부분에 대해서는 향후 대규모 논문 집합을 통한 재검증이 필요할 것으로 판단된다.

그림 3에서 수직 실선과 수직 점선은 6개 모든 자질들을 사용한 자질 조합의 위치를 Liblinear와 SMOlinear 커널에 대해 표시한 것이다. 그림에서 전체 자질을 사용한 경우보다 높은 성능을 보인 자질 조합들은 각 수직선의 오른편에 위치하며 Liblinear의 경우 1개, SMOlinear의 경우 4개임을 알 수 있다. 특히 두 선형 커널에서 최고 성능을 보인 자질 조합은 동일하였는데 그 자질 조합은 식 2를 제외한 나머지 5개 식들을 자질로 사용한 경우였다.

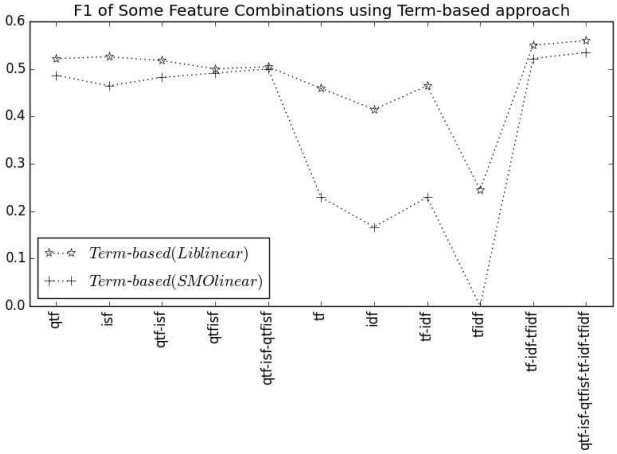

Fig. 4. Comparing Performances of Some Feature Combinations of Term-based Approach Using Linear SVMs (Y-axis indicates F1)

그림 4는 식 1~3과 식 4~6으로 대비되는 자질 조합들에 대해 용어 기반 방법의 F1 성능을 두 선형 커널들에 대해 비교 제시한 것이다. 그림에서 x축의 첫 5개 자질 조합들은 적합문장검색 관점에서 도입된 용어 중요도 수식들(식 1~3)을 개별 및 결합 사용한 경우이며, 다음 5개 자질 조합들은 추출식 문서요약 관점에서 도입된 용어 중요도 수식들(식 4~6)을 사용한 경우에 해당한다. 예를 들어 그림 4의 *qtf*는 식 1의 자질만을 사용한 것을 의미하며, *qtf-isf*는 식 1과 식2의 두 개 자질을 사용한 경우에 해당한다. 본 논문의 실험에서는 적합문장검색 관점의 자질들은 핵심인용인식의 성능 수준에 큰 차이가 없었으나 문서요약 관점의 자질 조합들의 경우 서로 다른 자질 선택에 따른 인식 성능의 변화 폭이 컸다.

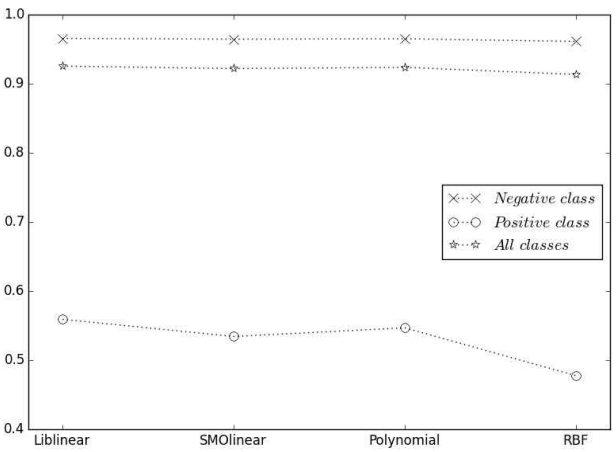

Fig. 5. Performance (F1) of Core Citation Recognition Using Term-based Method (All 6 features were used. Positive and negative classes indicate core citation and non-core citation class, respectively)

그림 5는 6개 모든 자질을 사용한 용어 기반 방법에서 핵심인용클래스(Positive class)와 비핵심인용클래스(Negative

www.kci.go.kr

class) 각각에 대한 성능($F1_C$ 및 $F1_N$)과 두 클래스 전체에 대한 F1 성능(클래스 분포에 기반한 $F1_C$와 $F1_N$의 가중 평균)을 비교 제시한 것이다. 소수 집단에 해당하는 핵심인용클래스(positive class)의 경우 분류 성능이 60%에 미치지 못하는 수준이나 다수 집단에 해당하는 비핵심인용클래스(negative class)는 높은 분류 성능을 보여 전체적으로는 90% 이상의 성능 수준을 보였다.

V. Conclusions

이 연구에서는 핵심인용인식을 위한 용어 기반 방법을 제안하고 실험 결과를 제시하였다. 이 방법은 인용텍스트에 출현한 각 용어의 가중치로부터 인용텍스트에 대응하는 피인용논문의 인용강도를 계산한다. 이를 위해 *qtf, isf, qtfisf, tf, idf, tfidf* 기반 용어 가중치 수식을 활용하였다. 제안된 방법은 전산언어학 분야 논문에 대해 구축된 기존 데이터셋을 사용한 SVM 기반 분류 성능 평가에서 기존 방법들과 대등한 핵심인용분류 능력을 보였다. 현재의 방법은 n-gram이나 구(phrase) 단위의 용어들을 추가 사용하거나 다양한 용어 가중치 수식들을 결합함으로써 그 확장이 용이한 장점을 갖는다.

제안된 방법에서는 피인용논문에 대한 인용텍스트가 인용논문 내에 다수 출현하는 경우 개별 인용텍스트에 대한 인용강도들의 단순 합으로 피인용논문의 인용강도를 계산하였다. 그러나 개별 인용텍스트의 인용강도는 해당 인용텍스트가 인용논문 내에 출현한 위치 혹은 인접 문맥에 따라 그 중요도가 다를 수 있다. 향후 연구에서는 이 부분을 반영하여 용어 기반 방법을 발전시킬 예정이다. 또한 현재 방법은 전산언어학 분야 논문들을 대상으로 SVM 알고리즘을 사용하여 평가된 것으로 다른 분야 논문들 혹은 다른 학습 알고리즘에서의 성능 수준 및 기존 연구와의 비교에 대해서는 추가 연구가 필요할 것이다.

REFERENCES

[1] X. Wan, and F. Liu, "Are All Literature Citations Equally Important? Automatic Citation Strength Estimation and its Applications," Journal of the Association for Information Science and Technology, Vol. 65, No. 9, pp. 1929-1938, 2014.

[2] X. Zhu, P. D. Turney, D. Lemire, and A. Vellino, "Measuring Academic Influence: Not All CitationsAre Equal," Journal of the Association for Information Science and Technology, Vol. 66, No. 2, pp. 408-427, 2015.

[3] A. Abu Jbara, and D. R. Radev, "Coherent Citation-Based Summarization of Scientific Papers," Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies(ACL), pp. 500-509, 2011.

[4] M. Valenzuela, V. Ha, and O. Etzioni, "Identifying Meaningful Citations," AAAI Workshop: Scholarly Big Data, 2015.

[5] T. Chakraborty, and R. Narayanam, "All Fingers are not Equal: Intensity of References in Scientific Articles," Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing(EMNLP), pp. 1348-1358, 2016.

[6] A. Akram, "Distinguishing Important Citations Using Contextual Information in Scholarly Big Data," Master's Thesis, Information Technology University, 2017.

[7] J. Allan, C. Wade, and A. Bolivar, "Retrieval and Novelty Detection at the Sentence Level," Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval(SIGIR), pp. 314-321, 2003.

[8] M. Galley, "A Skip-Chain Conditional Random Field for Ranking Meeting Utterances by Importance," Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing(EMNLP), pp. 364-372, 2006.

[9] C. Y. Lin, and E. H. Hovy, "The Automated Acquisition of Topic Signatures for Text Summarization," Proceedings of the 18th International Conference on Computational Linguistics(COLING), pp. 495-501, 2000.

[10] S. Xie, and Y. Liu, "Improving Supervised Learning for Meeting Summarization using Sampling and Regression," Computer Speech & Language, Vol. 24, No. 3, pp. 495-514, 2010.

[11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. VanderPlas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, 12, pp. 2825-2830, 2011.

[12] C. C. Chang, and C. J. Lin, "LIBSVM: A Library for Support Vector Machines," ACM Transactions on Intelligent Systems and Technology, Vol. 2, No. 3, pp. 27:1-27:27, 2011.

[13] R. E. Fan, K. W. Chang, C. J. Hsieh, X. R. Wang, and C. J. Lin, "LIBLINEAR: A Library for Large Linear Classification," Journal of Machine Learning Research, Vol. 9, pp. 1871-1874, 2008.

[14] C. W. Hsu, C. C. Chang, and C. J. Lin, "A Practical Guide to Support Vector Classification," *https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf*,

www.kci.go.kr

2003.
[15] I. G. Councill, C. L. Giles, and M. Y. Kan, "ParsCit: an Open-source CRF Reference String Parsing Package," Proceedings of the International Conference on Language Resources and Evaluation(LREC), 2008.
[16] S. Bird, R. Dale, B. J. Dorr, B. R. Gibson, M. T. Joseph, M. Y. Kan, D. Lee, B. Powley, D. R. Radev, and Y. F. Tan, "The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics," Proceedings of the International Conference on Language Resources and Evaluation(LREC), 2008.
[17] C. J. Fox, "A Stop List for General Text," SIGIR Forum, 24(1-2), pp. 19-35, 1990.
[18] M. F. Porter, "An Algorithm for Suffix Stripping," Program, Vol. 14, No. 3, pp. 130-137, 1980.

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 currently an associate professor in the Department of Computer Science & Engineering, Kyungsung University. He is interested in natural language processing and information retrieval.

www.kci.go.kr