Journal of The Korea Society of Computer and Information
Vol. 21 No. 11, pp. 85-91, November 2016
www.ksci.re.kr
http://dx.doi.org/10.9708/jksci.2016.21.11.085

Using Collective Citing Sentences to Recognize Cited Text in Computational Linguistics Articles

In-Su Kang*

Abstract

This paper proposes a collective approach to cited text recognition by exploiting a set of citing text from different articles citing the same article. First, the proposed method gathers highly-ranked cited sentences from the cited article using a group of citing text to create a collective information of probable cited sentences. Then, such collective information is used to determine final cited sentences among highly-ranked sentences from similarity-based cited text recognition. Experiments have been conducted on the data set which consists of research articles from a computational linguistics domain. Evaluation results showed that the proposed method could improve the performance of similarity-based baseline approaches.

▸ Keyword : Cited Text Recognition, Sentential Similarity, Citation, Cited Text, Citing Text

I. Introduction

논문과 같은 학술 문헌들 간에 발생하는 인용 정보(예: 인용문, 피인용문)는, 인용 정보에 기반한 논문의 자동 요약[1,2,3]에서 중요하게 고려된다. 후행 논문 P1이 선행 논문 P0를 인용한다고 가정할 때, 다음의 S1은 논문 P1에 출현한 (논문 P0를 인용하는) 인용문의 예이며, 예문 S0는 S1과 관련된 (논문 P0에 출현한) 피인용문의 예이다.

- 인용문 S1: *Smith(2013)의 방법은 지식 구축 과정의 전문가 의존성으로 인해 다른 분야로의 확장 적용이 용이하지 않다.*
- 피인용문 S0: *제안된 방법은 초기 도메인 지식 구축에 분야 전문가의 개입이 요구된다.*

특정 논문 P0를 인용하는 논문들(예: P1)에 출현한 (논문 P0를 인용하는) 인용문들의 집합은 논문 P0에 대한 인용 기반 자동 요약 방법의 주요 입력 유형이다[1]. 이와 함께 인용문 집합 내의 각 인용문에 대응하는 피인용문들의 집합 또한 인용 기반 논문 요약의 유용한 입력이 된다. 이와 관련하여 최근 진행된 CL-SciSumm-2016 Shared Task[3]에서는 학술 논문의 자동 요약을 위한 기술로 (1) 피인용 텍스트 인식, (2) 피인용 텍스트의 패싯(facet) 결정, 그리고 (3) 인용 기반 논문 자동 요약의 세 가지 세부 작업들을 제시하고 참가팀들의 성능을 평가하였다. 이 중 피인용 텍스트 인식(CTR: Cited Text Recognition)은, 인용논문에 출현한 인용문(예: S1)에 대응하는 피인용 텍스트(예: S0)를 피인용논문 내에서 인식하는 것이다.

기존 CTR 접근법으로 비교사 방법과 학습 기반 방법이 시도되었다. 비교사 방법의 대표적 접근법인 문장유사도 기반 방법에서는 인용문과 피인용논문 내 각 문장 간 유사도를 계산하고 가장 유사한 문장(들)을 피인용텍스트에 포함시키는 방식이 사용되었다. 학습 기반 방법에서는 피인용논문 내 각 문장과 인용문의 쌍을 학습 샘플로 사용하여 자질을 추출하고 기계학습을 적용하는 방식이 사용되었다.

이 연구에서는 피인용논문을 인용하는 다수 인용논문들의 집합이 주어진다고 가정하고 서로 다른 인용논문에 출현한 인용문장들에 대응하는 피인용문장들의 집단 정보를 수집하여 피

• 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: 2016. 10. 25, Revised: 2016. 11. 02, Accepted: 2016. 11. 14.
• 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).

인용텍스트 인식에 활용하는 집단적 CTR 방법을 제안한다.

논문의 구성은 다음과 같다. 2장에서는 관련 연구를 기술한다. 3장에서는 피인용 텍스트 인식을 위해 피인용문장의 집단적 정보를 수집하고 활용하는 방법에 대해 기술한다. 4장에서는 제안된 방법의 성능 평가에 대해 기술하고 5장에서 결론을 맺는다.

II. Related Works

문장유사도 기반 비교사 CTR 방법들 중 [4]에서는 인용문과 피인용 후보문 간에 tf-idf 기반 코사인 유사도를 계산하고, 인용문 문맥 확장, 워드넷 신셋 확장, 문장 길이 제약 등을 결합 시도하였다. [5]에서는 용어 기반 Jaccard 및 idf sum 유사도, word2vec 기반 유사도 등 다양한 유사도들의 가중치 결합에 기반하여 피인용문을 결정하였다. [6]에서는 피인용 후보 문장과 인용문 간에 공유되는 최대 두 개 고빈도 단어의 빈도와 문장 내 거리 및 공유 구문 관계에 기반하여 피인용문 인식을 시도하였다. [7]에서는 인용문과 같은 패싯(예: aim, method, result 등)을 갖는 문장 중에서 인용문과 가장 유사한 문장을 피인용문으로 결정하였다.

그래프 기반 비교사 방법의 하나로 [8]에서는 피인용논문에 TextSentenceRank 알고리즘을 수정 적용하여 피인용문 인식을 시도하였다.

학습 기반 CTR 방법 중 [9]에서는 피인용논문을 n-sentence 청크(n=1~4)들로 표현한 후, 인용문과 n-sentence 청크의 쌍에 대해 tf-idf 기반 코사인 유사도 등의 자질을 추출하고 SVM Rank를 적용하여 최종 피인용텍스트를 결정하였다. [10]에서는 인용문과 후보 피인용문 사이에 계산되는 tf-idf 기반 유사도와 신경망 기반 유사도를 가중치 결합하여 피인용문 결정에 사용하였다. [4]에서는 인용문과 후보 피인용문의 쌍에 대해 적합/부적합 레이블을 부착하여 SVM으로 학습/분류한 후 적합 신뢰 점수에 따라 후보 피인용문을 결정하였다. [8]에서는 TextSentenceRank와 Random Forest 분류기를 순차 적용(정순 및 역순)하는 접근법을 시도하였다. [11]에서는 인용문과 후보 피인용문의 쌍에 대해 코사인, Jaccard 등의 어휘 유사도, LDA 기반 토픽 유사도, WordNet 기반 의미 유사도, TextRank 기반 문장중요도 자질을 추출하고 Learning to Rank 알고리즘을 적용하였다. [12]에서는 인용문과 후보 피인용문 쌍에 대해 코사인 유사도, 의미 유사도, 문장위치 등의 자질을 추출하여 기계학습 기반의 피인용문 결정을 시도하였다.

전술한 기존 연구들은 CL-SciSumm-2016 Shared Task에서 CTR을 위해 시도된 방법들이다. CL-SciSumm-2014 Pilot Task의 경우 문장 유사도 기반 비교사 방법들과 어휘/구문/의미 유사도 자질 기반 학습법이 CTR을 위해 시도되었다[2].

전술한 연구들과 달리 인용문 집합의 집단적 정보를 활용한 선행연구로 [13]에서는 문장 추출식 논문 요약 관점에서 피인용논문 RP의 인용텍스트 집합에 반영된 RP의 임팩트(impact)를 질의로 고려하여 RP내 각 문장에 점수를 부여하였다. [13]에서는 임팩트 언어모델을 얻기 위해 RP의 인용문 집합 전체를 하나의 단위로 사용하였다. 이 연구에서는 RP의 인용문 집합 내 각 인용문에 대응하는 피인용문 후보 정보를 집단적으로 수집하여 활용한다.

III. The Proposed Scheme

1. Overview

이 연구에서는 피인용논문 RP를 인용하는 인용논문들의 집합 $CP=\{cp_1, cp_2, \dots, cp_m\}$이 주어진다고 가정하고, $Q=\{q_1, q_2, \dots, q_n\}$를 각 인용논문에 출현한, RP를 인용하는 인용텍스트의 집합으로 정의한다. 한 편의 인용논문 내에 동일 논문을 인용하는 인용텍스트는 1회 이상 출현 가능하다고 가정한다($n \geq m$). 피인용텍스트 인식은 Q에 속한 각 인용텍스트에 대응하는 피인용텍스트를 RP 내에서 결정하는 것이다. 이 연구에서 제안하는 집단적 CTR 방법은 집단적 피인용문 후보 강도 생성 단계와 이에 기반한 피인용텍스트 인식 단계로 나뉜다.

2. Collective Cited Sentence Information

먼저 Q 내의 각 인용텍스트 q_i에 대해 q_i와 RP 내 각 문장 s 간의 문장유사도 $sim(q_i, s)$를 계산하여 순위화된 문장번호리스트 $rankedList_i$를 생성한다. 다음으로 각 인용텍스트 q_i에 대해, q_i에 대응하는 $rankedList_i$를 제외한 나머지 n-1개 $rankedList_j$ (j=1~n, i≠j)들의 상위 α 순위 내에 출현한 각 문장번호 sid(sentence id)에 식 1의 점수를 부여한 $CCSS_i$를 생성한다. CCSS는 "Collective Cited Sentence Strength"를 의미하며 집단적 피인용문 후보 강도 정보에 해당한다. 아래 수식에서 $rank_j^{\alpha}(sid)$는 $rankedList_j$의 상위 α 순위 내에 출현한 sid의 순위(rank)이며, 그러한 sid가 출현하지 않는 경우 ∞의 값을 할당한다.

$$CCSS_i(sid) = \sum_{\substack{j=1 \\ i \neq j}}^{n} \frac{1}{rank_j^{\alpha}(sid)} \quad \text{식 1}$$

3. Collective Cited Text Recognition

집단적 피인용문 후보 정보에 기반한 피인용텍스트 인식은, 인용텍스트와 RP 내 각 문장들과의 유사도 계산을 통해 얻어진 유사도 기준 상위 문장들에 대해 최종 피인용텍스트 포함 여부를 결정하는 방식으로 동작하며, 이 결정 과정에 집단적 피인용문 정보가 활용된다. 구체적인 인식 절차는 그림 1에 의사코드

로 제시하였다.

먼저 인용텍스트 q_i와 RP 내 각 문장의 유사도 $sim(q_i, s)$에 기반하여 피인용 후보 문장들의 순위화된 리스트 $rankedList_i$를 생성한다(Step-2 참조). 이렇게 생성된 순위목록의 1순위 후보문장은 최종 피인용텍스트에 기본으로 포함되도록 하고(Step-3, Step-4 참조), 2순위부터 β순위까지 각 문장들에 대해 다음 조건들을 검사하여 피인용텍스트 포함 여부를 결정한다(Step-5 ~ Step-10 참조).

- 조건-I: 다른 인용텍스트에 대한 순위화된 피인용문 목록에서 상위 α 순위 내에 포함될 것 (Step-6 참조)
- 조건-II: 조건-I의 순위화된 목록들로부터 계산된 reciprocal rank의 합이 1이상일 것 (Step-7 참조)
- 조건-III: 해당 문장의 유사도가 1순위 문장 유사도의 Θ배 이상일 것 (Step-8 참조)

CollectiveCitedTextRecognition(q_i, RP, $CCSS_i$)

INPUT
- q_i: a string of citing text
- RP: a set of sentences in Reference Paper
- $CCSS_i$: collective cited sentence strength information for q_i

PROCEDURE
1. create sidList as a null list
2. generate $rankedList_i$ based on $sim(q_i, s)$ for $s \in RP$
3. $topsid = rankedList_i[1]$
4. append topsid to sidList
5. FOR sid in $rankedList_i[2..\beta]$
6. IF sid is not in $CCSS_i$ THEN continue ENDIF
7. IF $CCSS_i(sid) < 1$ THEN continue ENDIF
8. IF $sim(q_i, s_{sid}) < \Theta \times sim(q_i, s_{topsid})$ THEN continue ENDIF
9. append sid to sidList
10. END FOR

OUTPUT sidList

Fig. 1. Collective Cited Text Recognition

4. Example

다음은 α=3, β=5, Θ=0.4로 가정하고, $Q=[q_1, q_2, q_3, q_4]$에 대한 집단적 피인용텍스트 인식 방법의 동작 과정을 예를 들어 설명한 것이다. 먼저 피인용논문 RP를 인용하는 인용텍스트 집합 Q 내 각 인용텍스트에 대해 RP 내 문장들과의 유사도 계산을 통해 문장번호리스트를 생성한다. 아래 예에서 $rankedList_2$에 해당하는 [23, 10, 14, ...]는 인용텍스트 q_2에 대해 RP 내 가장 유사한 상위 세 개 문장이 s_{23}, s_{10}, s_{14}임을 의미한다.

$Q=[q_1, q_2, q_3, q_4]$
$rankedList_1$=[53, 14, 18, 23, 56, ...]
$rankedList_2$=[23, 10, 14, ...]
$rankedList_3$=[18, 92, 23, ...]
$rankedList_4$=[56, 23, 92, ...]

위 문장번호리스트 집합으로부터 q_1에 대한 집단적 피인용문 후보 정보에 해당하는 $CCSS_1$을 생성하면 다음과 같은데, 이는 q_1에 대응하는 $rankedList_1$을 제외한 $rankedList_2$, $rankedList_3$, $rankedList_4$에 대해 상위 3순위(α=3)까지의 피인용 후보 문장들에 대해 유사도 순위 역수값들의 총합을 부여한 것이 된다.

파라미터 α=3
$CCSS_1$={23:11/6, 10:1/2, 14:1/3, 18:1, 92:5/6, 56:1}

위 예에서 23:11/6는 $CCSS_1(23)=11/6$을 의미하며, 그 계산 과정은 다음과 같다.

$$CCSS_1(23) = \frac{1}{rank_2^3(23)} + \frac{1}{rank_3^3(23)} + \frac{1}{rank_4^3(23)} = \frac{1}{1} + \frac{1}{3} + \frac{1}{2} = \frac{11}{6}$$

다음은 CollectiveCitedTextRecognition(q_1, RP, $CCSS_1$)의 동작 과정을 예시한 것으로, RP 내 s_{53}, s_{18}, s_{23} 문장들이 인용텍스트 q_1에 대해 최종 인식된 피인용텍스트에 해당한다.

파라미터 β=5, Θ=0.4
$rankedList_1$=[53, 14, 18, 23, 56, ...]
$sim(q_1, s_{53})$=1.0, ..., $sim(q_1, s_{23})$=0.4, $sim(q_1, s_{56})$=0.38
topsid=53
sidList=[53]
2순위 문서 14 => sidList=[53]
3순위 문서 18 => sidList=[53, 18]
4순위 문서 23 => sidList=[53, 18, 23]
5순위 문서 56 => sidList=[53, 18, 23]

위 예에서 보인 것처럼 먼저 q_1에 대해 RP 내 각 문장들과의 유사도 계산을 통해 문장번호리스트 $rankedList_1$를 생성한 후 1순위 문장 s_{53}이 기본으로 sidList에 포함되고, 2순위부터 5순위(β=5)까지 문장에 대해 $CCSS_1$에 기반하여 피인용문 인식 여부 결정 작업이 진행된다. 2순위 문서 s_{14}는 $CCSS_1(14)=1/3$이므로, 조건-II의 $CCSS_i(sid) \geq 1$을 통과하지 못하여 최종 피인용문으로 선정되지 못하였다. 5순위 문서 s_{56}는 조건-II는 통과하지만, 최종 피인용문의 유사도가 1순위 문서 유사도의 40%(Θ=0.4) 이상이어야 한다는 조건-III을 통과하지 못하여 최종 피인용문 선정에 포함되지 않았다.

IV. Experiments

1. Test Set and Evaluation Metrics

피인용문 인식 방법의 성능 평가를 위해 CL-SciSumm-2016 Shared Task를 통해 구축된 피인용문 인

식 정답 데이터셋(이후 SciSumm 데이터셋으로 약칭)을 사용하였다. 표 1은 SciSumm 데이터셋의 구성 및 통계 자료를 보인 것으로 학습(Train), 검증(Development), 테스트(Test) 집합이 구분되어 있다.

Table 1. SciSumm Cited Text Recognition Dataset

Type	Number of reference papers	Average number of citing papers	Number of citing text
Train	10	8.3	134
Development	10	14.6	219
Test	10	22.1	350

피인용문 인식은 각 인용텍스트에 대해 수행된다. SciSumm 데이터셋에서 하나의 인용텍스트는 하나 이상의 인용문장(들)로 구성되며, 각 인용텍스트에 대응하는 정답 피인용텍스트 역시 하나 이상의 문장들로 이루어져 있다. 표 1을 통해 SciSumm 학습 집합의 경우 총 10편의 각 피인용논문(reference paper)을 인용하는 논문들(citing papers, 평균 8.3편)에서 총 134개 인용텍스트(citing text)가 출현했음을 알 수 있다.

성능 평가 지표로 정확률(precision), 재현율(recall), F1을 사용하였다. 정확률은 시스템이 인식한 후보 피인용문(들) 중 정답 피인용문이 포함된 비율로 정의된다. 재현율은 정답 피인용문(들) 중 시스템이 올바르게 인식한 피인용문의 비율로 정의된다. F1은 정확률과 재현율의 조화평균이다.

2. Baseline System

Table 2. Baseline systems

Similarity function	Term weighting		
	Label	Document	Query
cosine	lnc.ltc	log(1+tf)	log(1+tf) × log(idf)
	ltc.lnc	log(1+tf) × log(idf)	log(1+tf)
	ltc.ltc	log(1+tf) × log(idf)	log(1+tf) × log(idf)
	ntc.ntc	log(idf)	log(idf)
	lnc.lnc	log(1+tf)	log(1+tf)
Jaccard	binary		
Jaccard Weighted	log(idf)		

피인용텍스트 인식의 베이스라인 시스템으로 문장유사도 기반 방법을 사용하였다. 이 방법은 인용텍스트와 피인용논문 내의 각 문장 간에 문장유사도를 계산한 후 가장 유사한 상위 k개 문장들을 피인용텍스트 인식 결과로 출력한다. 최종 베이스라인 시스템 결정을 위해 코사인 유사도 및 Jaccard 유사도를 tf, idf에 기반한 다양한 용어 가중치 기법들에 대해 평가하였다. 표 2는 평가에 사용된 유사도 함수와 용어 가중치 할당 기법들의 목록을 보인 것이다.

코사인 유사도의 경우 다섯 가지 가중치 할당 기법들을 시도하였다. 표에서 Query와 Document는 각각 인용텍스트와 피인용논문 내 하나의 문장에 해당하며, lnc.ltc의 경우 Document 및 Query의 용어 가중치로 각각 log(1+tf)와 log(1+tf)×log(idf)를 사용한다는 의미이다. 용어의 idf는 특정 피인용논문에 출현한 모든 문장들의 집합을 문서집합으로 고려하여 계산하였다. Jaccard 유사도[14]는 식 2와 같다. Jaccard Weighted 유사도는 [15]에서 제시된 수식을 사용하였으며 식 3과 같다. 식 2, 식 3에서 q는 인용텍스트이고, s는 피인용논문 내 하나의 문장에 해당하며, T(q)와 T(s)는 각각 q와 s의 용어 집합이다. Jaccard Weighted의 w(t)는 용어 t의 가중치로, t의 idf의 로그값을 사용하였다.

$$sim_{Jac}(q,s)=\frac{|T(q)\cap T(s)|}{|T(q)\cup T(s)|} \quad \text{식 2}$$

$$sim_{JW}(q,s)=\frac{\sum_{t\in T(q)\cap T(s)} w(t)}{\sum_{t\in T(q)\cup T(s)} w(t)} \quad \text{식 3}$$

용어 표현을 위한 전처리로 SciSumm 데이터셋 내 인용텍스트 및 피인용논문의 각 문장에 대해 괄호로 감싸인 문자열을 제거한 다음, 두 문자 이상의 영숫자 토큰(영문자 및 숫자로 구성된 토큰)을 추출하고 불용어 제거 및 포터 스테밍[16]을 적용하였다. 불용어 제거 및 스테밍을 위해 NLTK를 사용하였다. 피인용논문 및 인용논문의 본문 텍스트로부터의 문장 인식은, SciSumm 데이터셋 내 구분된 문장 단위를 사용하였다.

그림 2, 그림 3은 표 2의 각 문장유사도 기반 방법을 학습 및 검증 데이터셋에 대해 평가한 F1 성능을 제시한 것이다. 그림에서 x축은 피인용텍스트 인식 결과에 포함한 상위 (문장유사도) 순위 문장의 개수(TopK)이다. 코사인 유사도의 경우 문서검색의 표준적 가중치 할당 방식으로 알려진[17] lnc.ltc가 좋은 성능을 보였다. 전체적으로 Jaccard Weighted 문장유사도 방법이 피인용텍스트 인식에서 비교 우위의 성능을 보였다.

그림 4는 Jaccard Weighted와 lnc.ltc 방법의 피인용텍스트 인식 성능을 테스트 집합에 대해 보인 것이다.

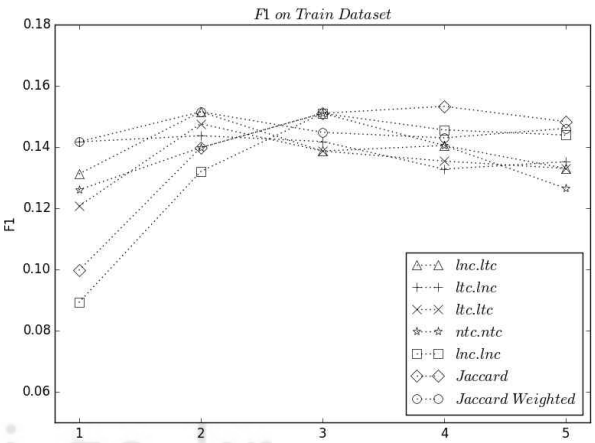

Fig. 2. CTR Performance on Train Dataset

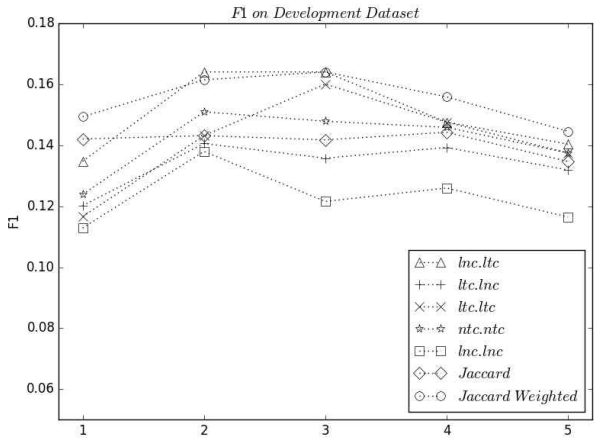

Fig. 3. CTR Performance on Development Dataset

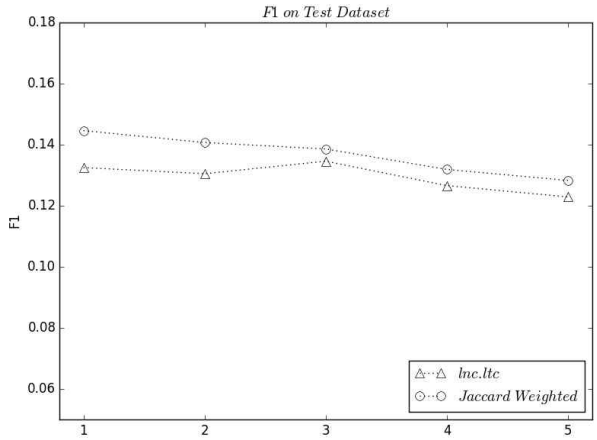

Fig. 4. CTR Performance on Test Dataset

3. Evaluation of Collective CTR

표 3은 테스트 집합에 대한 피인용텍스트 인식 방법들의 성능을 CL-SciSumm-2016 Task 1A(피인용텍스트 인식 태스크)의 최고 성능과 함께 비교 제시한 것이다. Baseline CTR의 경우 이전 실험에서 결정된 Jaccard Weighted 문장유사도 기반 방법으로 TopK의 값이 1일 때의 성능이다. Collective CTR 방법의 성능은 문장유사도 함수로 표 3의 Baseline CTR을 사용하여 얻은 것이다. 표에서 집단적 CTR 방법에 사용된 파라미터 α, β, Θ의 최적 값은 다음 범위 내 값들의 조합을 학습 및 검증 집합에 대해 각각 평가하여 결정하였다.

α = {1, 2, 3, 4, 5}
β = {2, 3, 4, 5}
Θ = {0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9}

학습 집합의 경우 같은 최고 성능을 보인 다수 파라미터 조합들이 탐색되었는데 표 3에서는 그 최적 조합들을 사용한 테스트 집합에서의 F1 성능들(0.1555~0.1576) 중 최고값을 제시한 것이다. 표 1에 제시된 것처럼 학습, 검증, 테스트 집합은 피인용논문에 대한 평균 인용논문의 수에 있어 차이를 보인다. 집단적 CTR 방법은 인용논문 집합으로부터 수집된 집단적 피인용문 정보에 의존하므로 인용논문 집합의 크기가 그 성능에 간접적 영향을 미칠 수도 있을 것이다. 이와 관련한 상관관계 유무 및 강도 확인에 대해서는 별도 연구가 필요할 것이다. 현재 실험에서는 단순히 학습 및 검증 데이터셋을 분리하여 각각 결정된 파라미터를 통한 집단적 CTR의 테스트 집합 성능을 평가하였다.

Table 3. Evaluation of Cited Text Recognition Methods on Test Dataset

Method	Precision	Recall	F1	Description
SciSumm-2016 Task 1A Best	N/A	N/A	0.13	Jaidka et al. [3,18]
Baseline CTR	0.1714	0.1250	0.1446	Jaccard Weighted TopK=1
Collective CTR	0.1283	0.2042	0.1576	α=2, β=5, Θ=0.3 (from Train data)
	0.1520	0.1729	0.1618	α=1, β=3, Θ=0.3 (from Development data)

베이스라인과 비교해 볼 때, 집단적 CTR 방법은 피인용텍스트 인식의 F1 성능 향상에 도움이 됨을 알 수 있다. 집단적 CTR은 1차 문장유사도 순위 기준으로 상위 β개 문장을 최종 피인용문 후보로 고려하므로 TopK=1인 베이스라인에 비해 정확률은 저하되고 재현율은 상승한 것으로 판단된다. 실험에 사용된 β 값인 3, 5에 대응하는 TopK의 값을 사용한 베이스라인의 성능(그림 4 참조)과 비교한 경우에도 집단적 CTR이 더 좋은 성능을 보였다. 그림 5는 테스트 집합 내 10편의 각 피인용논문에 대해 베이스라인과 제안하는 방법의 성능을 비교 제시한 것으로 과반이상의 논문들에서 제안된 방법의 성능 향상이 관찰되었다.

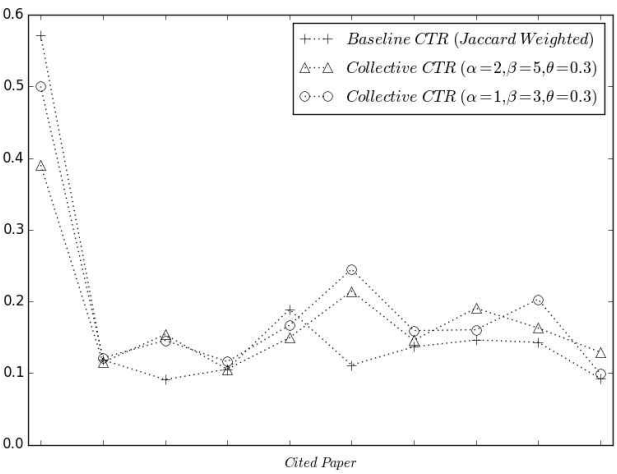

Fig. 5. Performance for Individual Cited Papers of Test Dataset

V. Conclusions

이 연구에서는 피인용텍스트 인식을 위해 피인용문장들의 집단적 정보를 수집 및 활용하는 방법에 대해 기술하였다. 이 방법은 피인용문 후보 강도를 인용텍스트 집합으로부터 집단적으로 수집하는 절차와 수집된 정보를 실제 피인용텍스트 인식에 활용하는 절차로 구성된다. 전산언어학 분야 논문들로 이루어진 데이터셋에 대한 평가에서 집단적 CTR 방법은 피인용텍스트 인식에서 문장유사도 기반 베이스라인 방법의 성능을 향상시킬 수 있음을 보였다. 향후 집단적 피인용문 강도를 활용하는 그래프 기반의 피인용텍스트 인식 방법을 시도할 계획이다.

REFERENCES

[1] A. Abu-Jbara, and D. Radev, "Coherent Citation-Based Summarization of Scientific Papers," Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pp. 500-509, 2011.

[2] K. Jaidka, M. Chandrasekaran, B. Elizalde, R. Jha, C. Jones, M. Kan, A. Khanna, D. Moll´a-Aliod, D. Radev, F. Ronzano, and H. Saggion, "The Computational Linguistics Summarization Pilot Task," Proceedings of Text Analysis Conference, Gaithersburg, USA, 2014.

[3] K. Jaidka, M. Chandrasekaran, S. Rustagi, and M. Kan, "Overview of the CL-SciSumm 2016 Shared Task," Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries, pp. 93-102, 2016.

[4] L. Moraes, S. Baki, R. Verma, and D. Lee, "University of Houston at CL-SciSumm 2016: SVMs with tree kernels and Sentence Similarity," Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries, pp. 113-121, 2016.

[5] L. Li, L. Mao, Y. Zhang, J. Chi, T. Huang, X. Cong, and H. Peng, "CIST System for CL-SciSumm 2016 Shared Task," Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries, pp. 156-167, 2016.

[6] P. Aggarwal, and R Sharma, "Lexical and Syntactic cues to identify Reference Scope of Citance," Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries, pp. 103-112, 2016.

[7] B. Malenfant, and G. Lapalme, "RALI System Description for CL-SciSumm 2016 Shared Task," Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries, pp. 146-155, 2016.

[8] S. Klampfl, A. Rexha, and R. Kern, "Identifying Referenced Text in Scientific Publications by Summarisation and Classification Techniques," Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries, pp. 122-131, 2016.

[9] Z. Cao, W. Li, and D. Wu, "PolyU at CL-SciSumm 2016," Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries, pp. 132-138, 2016.

[10] T. Nomoto, "NEAL: A Neurally Enhanced Approach to Linking Citation and Reference," Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries, pp. 168-174, 2016.

[11] K. Lu, J. Mao, G. Li, and J. Xu, "Recognizing Reference Spans and Classifying their Discourse Facets," Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries, pp. 139-145, 2016.

[12] H. Saggion, A. AbuRa'Ed, and F. Ronzano, "Trainable Citation-enhanced Summarization of Scientific Articles," Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries, pp. 175-186, 2016.

[13] Q. Mei, and C. Zhai, "Generating Impact-Based Summaries for Scientific Literature," Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, pp. 816-824, 2008.

[14] P. Jaccard, "Nouvelles recherches sur la distribution florale," Bull. Soc. Vaud. Sci. Nat. Vol. 44, pp. 223-270, 1908.

[15] O. Hassanzadeh, M. Sadoghi, and R. Miller, "Accuracy of Approximate String Joins Using Grams," Proceedings of the Fifth International Workshop on Quality in Databases, pp. 11-18, 2007.

[16] M. Porter, "An algorithm for suffix stripping," Program, Vol. 14, No. 3, pp. 130-137, 1980.

[17] C. Manning, P. Raghavan, and H. Schütze, "Introduction to information retrieval," Cambridge University Press, 2008.

[18] K. Jaidka, M. Chandrasekaran, S. Rustagi, and M. Kan, "The Computational Linguistics Summarization Pilot Task @ BIRNDL 2016," CL-SciSumm2016_sildedeck.pdf in https://github.com/WING-NUS/scisumm-corpus, 2016.

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.