Journal of The Korea Society of Computer and Information
Vol. 23 No. 5, pp. 1-7, May 2018

www.ksci.re.kr
https://doi.org/10.9708/jksci.2018.23.05.001

# Reference String Recognition based on Word Sequence Tagging and Post-processing: Evaluation with English and German Datasets

In-Su Kang*

## Abstract

Reference string recognition is to extract individual reference strings from a reference section of an academic article, which consists of a sequence of reference lines. This task has been attacked by heuristic-based, clustering-based, classification-based approaches, exploiting lexical and layout characteristics of reference lines. Most classification-based methods have used sequence labeling to assign labels to either a sequence of tokens within reference lines, or a sequence of reference lines. Unlike the previous token-level sequence labeling approach, this study attempts to assign different labels to the beginning, intermediate and terminating tokens of a reference string. After that, post-processing is applied to identify reference strings by predicting their beginning and/or terminating tokens. Experimental evaluation using English and German reference string recognition datasets shows that the proposed method obtains above 94% in the macro-averaged F1.

▸Keyword: Reference String Recognition, Sequence Labeling, Citation

## I. Introduction

참고문헌 인용열 인식(reference string recognition)은 학술문헌으로부터 참고문헌 텍스트에 출현하는 인용열(reference string)들을 인식하는 문제이다. 다음은 참고문헌 영역에 출현한 텍스트의 가상 예를 보인 것으로 8개 인용행(reference line)들로 구성되어 있다.

> James Brown, Richard Wilson, Michael Smith, and Dale Thompson. "Semi-automatic extraction of paraphrase expressions from parallel corpora". In Eric Williams and Andrew Scott (eds.), Handbook of Corpus Processing. pp. 62-78. 2013.
> Edward Martin, Daniel Clark. "Learning to generate natural paraphrases". Information & Intelligence. 31(4). pp. 157-172. 2014.

위 예의 참고문헌 텍스트에 인용열 인식이 성공적으로 적용된다면 다음과 같이 2개 인용열이 추출될 것이다.

① James Brown, Richard Wilson, Michael Smith, and Dale Thompson. "Semi-automatic extraction of paraphrase expressions from parallel corpora". In Eric Williams and Andrew Scott (eds.), Handbook of Corpus Processing. pp. 62-78. 2013.

② Edward Martin, Daniel Clark. "Learning to generate natural paraphrases". Information & Intelligence. 31(4). pp. 157-172. 2014.

본 논문에서는 참고문헌 인용열 인식의 기존 성능을 개선하여 다양한 활용 분야로의 실용성을 높이기 위해 CRF 분류 및 후처리 규칙을 결합한 새로운 방법을 제안한다.

참고문헌 인용열 인식의 기존 연구에서는 인용열을 구성하는 인용행의 어휘적 특성이나 원문(예: PDF 원문) 내 인용행의 레이아웃 정보를 주로 활용하였다. 인용행의 어휘적 특성의 예

• First Author: In-Su Kang, Corresponding Author: In-Su Kang
*In-Su Kang (dbaisk@ks.ac.kr), Dept. of Computer Science, Kyungsung University
• Received: 2018. 03. 28, Revised: 2018. 04. 12, Accepted: 2018. 05. 16.
• 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).

로는, 인용열을 구성하는 첫 인용행은 대부분 대문자 혹은 저자명으로 시작한다거나 마지막 인용행은 숫자 출현 비율이 상대적으로 높고 마침표 문자로 종료되는 경향이 있다는 등의 규칙성을 들 수 있다. 인용행의 레이아웃 정보로는 인접 인용행 간 수직 간격이나 현재 인용행의 들여쓰기 여부와 같은 시각적 배치 정보 등이 해당된다.

인용열 인식 방법론으로 최근까지 휴리스틱 기반 방법, 군집화 기반 방법, 분류 기반 방법이 시도되었다. 휴리스틱 기반 방법 중 하나인 ParsCit 시스템에서는 인용행 종료 문자, 인용행 내 저자명 출현 정보, 인용행 길이 정보 등을 검사하여 인용행 나열을 인용열로 분해하였다[1]. 군집화 기반 방법은 각 인용행을 군집화 대상 개체로 고려하여, 인용열의 시작 행에 해당하는 군집과 나머지 행에 대응하는 두 개 군집으로 그룹화하는 방법을 활용하여 인용열을 인식하였다[2,3]. 주요 분류 기반 방법들은 CRF(Conditional Random Field) 기반의 학습 모델을 통해 인용행 나열에 태그(클래스 레이블)를 부착[4]하거나, 인용행을 구성하는 토큰들의 나열에 태그를 부착[5]한 후 그 결과로부터 인용열을 추출하였다.

본 논문에서는 Boyd의 연구[5]에서처럼 토큰 단위 CRF 기반 분류 방법을 채택한다. 그러나 Boyd의 연구와 달리 인용열의 시작 토큰들, 중간 토큰들, 그리고 마지막 토큰에 서로 다른 태그를 부착한다. 이후 각 인용행의 토큰 단위 태그 부착 결과로부터 시작 및 마지막 토큰 태그를 중심으로 인용열의 시작 인용행과 마지막 인용행을 예측하여 인용열을 인식하는 태깅-후처리 절차를 적용한다. Boyd는, 마커 토큰(예: [2], 2. 등)을 제외하면, 인용열의 첫 토큰과 나머지 토큰들을 구분하는 두 개 태그를 사용하였다.

실험에서는 전산언어학 분야 영어 논문들의 참고문헌에 출현한 인용행 및 대응하는 인용열들로 구성된 데이터셋을 구축하여 제안된 방법의 인용열 인식 성능을 평가한다. 또한 기존 독일어 인용열 인식 데이터셋을 사용하여 기존 접근법과의 성능 비교를 제시한다.

논문의 구성은 다음과 같다. 2장에서는 관련 연구를 기술한다. 3장에서는 이 연구에서 제안하는 방법에 대해 기술한다. 4장에서는 제안된 방법의 성능 평가에 대해 기술하고 5장에서 결론을 맺는다.

## II. Related Works

ParsCit 시스템은 휴리스틱 기반의 인용열 인식을 시도하였는데, 참고문헌 텍스트가 인용열 마커(예: [3], (3), 3. 등)로 표현된 것인지 여부에 따라 다른 처리를 적용하였다. 참고문헌 텍스트가 인용열 마커로 표현된 경우 마커 유형을 매치하는 정규표현식에 기반하여 인용열 분해를 수행하고, 그러한 마커가 발견되지 않는 경우 인용행 종료 문자, 인용행 내 저자명 출현 정보, 인용행 길이 정보 등을 사용하여 인용행들을 인용열로 분리하였다[1].

Pdfextract 툴에서는 참고문헌 영역의 인용행 배치 스타일을 시작 인용행 들여쓰기 유형, 시작 인용행 내어쓰기 유형, 인용열의 마지막 인용행 다음 공백행 삽입 유형, 시작 인용행 마커 표시 유형 등으로 구분하는 방법을 통해 인용열 인식을 수행하였다[6].

Kern과 Klampfl은 인용열의 첫 행과 나머지 행들이 시작 x 좌표에서 차이가 있음을 이용하여 각 인용행을 행의 시작 x 좌표로 표현하였다[2]. 이후 각 인용행을 군집 대상 개체로 고려하여 인용열의 시작 행 군집과 나머지 행 군집의 두 개 그룹으로 나누는 군집화 방식에 기반하여 인용열 인식을 수행하였으며, 군집법으로는 K-means 방법을 사용하였다.

Cermine 시스템에서도 인용행들의 군집화에 기반하여 인용열을 인식하였다[3,7]. 이를 위해 각 인용행 개체를 (1) 순번 나열 패턴으로 시작하는지 여부, (2) 이전 인용행이 마침표로 종료되는지 여부, (3) 임계치 이상의 indentation 사용 여부, (4) 직전 인용행과의 수직 간격이 임계치를 초과하는지 여부, (5) 직전 행의 zone 폭 대비 상대적 길이의 5개 자질들로 표현한 후 K-means 군집법을 적용하였다.

Grobid 시스템에서는 CRF 학습을 통해 참고문헌 텍스트에 출현한 각 토큰에 마커 시작 토큰 태그(I-<label>), 마커 중간 토큰 태그(<label>), 인용열 시작 토큰 태그(I-<reference>), 인용열 중간 토큰 태그(<reference>) 등을 부착한 후, 마커 태그나 인용열 시작 토큰 태그를 이용하여 인용행을 인용열로 분해하였다[5,8]. 마커 토큰이 출현하지 않은 경우 인용열 시작 토큰으로 인용열의 시작 위치를 인식한다.

DeepBIBX 방법은 인쇄 문헌에 대한 스캔된 이미지로부터 인용열을 추출하는 딥러닝 기반 방법으로, 이 방법은 이미지화된 인용행들의 시각적 특성을 사용함으로써 인용열 인식의 언어 독립적 활용에 있어 장점을 갖는다[9].

기존 방법들 중 가장 높은 성능을 보고한 RefExt 시스템에서는 CRF 학습된 모델을 통해 논문 텍스트 내 각 행에 인용열-시작행 태그(B-REF), 인용열-중간행 태그(I-REF), 기타 태그(O) 중 하나를 부착하고 이 태깅 결과로부터 인용열을 추출하였다[4]. 학습 자질은 행 단위의 텍스트 자질과 레이아웃 자질로 나뉜다. 텍스트 자질은 인용행 첫 문자로 대문자 출현 여부, 인용행 마지막 문자로 숫자/마침표/콤마 출현 여부, 인용행의 연도/페이지범위 등 포함 여부, 인용행 내 숫자/단어/마침표/콤마 출현 횟수 등의 자질들로 구성되었다. 레이아웃 자질로는 현재 행의 들여쓰기 여부, 이전 행과의 수직 간격의 임계치 초과 여부, 현재 행의 문자 개수가 이전 행보다 적은지 여부, 전체 문서에서 현재 행의 상대적 위치가 사용되었다.

## III. The Proposed Method

이 장에서는 본 논문에서 제안하는 토큰 단위의 CRF[10]

태깅에 기반한 인용열 인식 방법에 대해 기술한다. 인용열 인식의 입력은 논문에 출현한 참고문헌 텍스트이며 이는 인용행들의 나열로 주어진다고 가정한다. CRF 태깅을 위한 클래스 레이블은 다음과 같이 정의하며, 레이블들의 완전명은 차례로 Author, Tail, In, Year, Other이다.

A: 인용열을 시작하는 저자명 표현 토큰(들)에 부착한다. 인용열이 저자명으로 시작하지 않는 경우, 인용열의 첫 필드에 속하는 토큰(들)에 부착한다.

T: 인용열의 마지막 토큰에 부착한다. 마지막 토큰이 연도 토큰인 경우는 태그 Y 대신 태그 T를 부착한다.

I: 논문이 게재된 학술대회논문집이나 도서 등의 명칭 앞에 표기하는 'In', 'In:', 'in:' 등의 토큰에 부착한다.

Y: 연도 표현 토큰에 부착한다(예: (2015), 2015. 등).

O: A, T, I, Y에 해당하지 않는 토큰에 부착한다.

예를 들어 서론에서 예시된 첫 인용열에 A, T, I, Y, O 태그를 부착한 결과는 다음과 같으며, 토큰과 태그 구분자로 /를 사용하여 표시하였다. 아래 예에서 마지막 토큰 "2013."은 연도 표기 형식이지만 인용열의 마지막 토큰이므로 태그 T가 부착되었다.

James/A Brown,/A Richard/A Wilson,/A Michael/A Smith,/A and/A Dale/A Thompson./A "Semi-automatic/O extraction/O of/O paraphrase/O expressions/O from/O parallel/O corpora"./O In/I Eric/O Williams/O and/O Andrew/O Scott/O (eds.),/O Handbook/O of/O Corpus/O Processing./O pp./O 62-78./O 2013./T

위와 같은 태그 부착 인용열들로부터 학습된 CRF 모델을 이용하여 새로운 참고문헌 인용행들을 인용열로 분해하는 단계는 다음과 같다.

(1) CRF 모델 학습: 태그 부착 인용열들의 모음으로 구성된 학습데이터로부터 CRF 모델을 학습한다.

(2) CRF 태깅: 학습된 CRF 모델을 이용하여 인용열 인식 대상 참고문헌 인용행들의 각 토큰을 태깅한다.

(3) 후처리: 참고문헌에 출현한 각 인용행에 대해 CRF 태깅 결과를 바탕으로 다음 두 조건을 순차 검사하여 어느 하나가 만족되는 경우 현재 인용행을 인용열의 시작 인용행으로 설정한다. 특히 참고문헌의 첫 인용행은 항상 인용열의 시작 인용행으로 설정하며, 직전 인용행의 마지막 문자가 대쉬('-')인 경우는 현재 인용행을 시작 인용행으로 설정하지 않는다.

- Condition-I: 현재 인용행의 첫 토큰 태그가 'A'일 것. 그러나 직전 인용행의 모든 토큰이 'A'인 경우는 조건 만족에서 배제.
- Condition-II: 현재 인용행의 첫 토큰 태그가 'A'가 아니고, 직전 인용행의 마지막 토큰 태그가 'T'이면서, 직전 인용행과 현재 인용행의 길이 비가 임계치 α미만일 것.

위의 Condition-I은 저자명 표현으로 시작하는 인용행을 인용열의 시작 인용행으로 설정하는 조건이다. 그러나 서론 예시의 두 번째 행과 같은 인용행을 시작 인용행에서 배제하기 위해 현재 인용행의 저자명 표현이 이전 인용행의 저자명 표현과 연결되는 긴 저자명 표현의 일부인지에 대한 검사를 포함하고 있다.

Condition-II는 직전 인용행의 마지막 단어가 인용열의 마지막 토큰으로 분류되면서, 직전 인용행의 길이가 현재 인용행의 길이에 비해 허용 수준 미만으로 짧은 경우, 직전 인용행을 이전 인용열의 마지막 인용행으로 가정함으로써, 현재 인용행을 새로운 인용열의 시작 인용행으로 설정하는 조건이다.

Table 1. Example of AOIT-tagged Reference Lines

| No. | AOIT-tagged Reference Line |
|---|---|
| 01 | James/A Brown,/A Richard/A Wilson,/A Michael/A |
| 02 | Smith,/A and/A Dale/A Thompson./A "Semi-automatic/O |
| 03 | extraction/O of/O paraphrase/O expressions/O from/O |
| 04 | parallel/O corpora"./O In/I Eric/O Williams/O and/O |
| 05 | Andrew/O Scott/O (eds.),/O Handbook/O of/O Corpus/O |
| 06 | Processing./O pp./O 62-78./O 2013./T |
| 07 | Robert/O Taylor./O "Deep/O Learning"./O ABC/O press./T |
| 08 | Liang/A Huang,/A Mary/A Murphy,/A "Deep/O Learning/O |
| 09 | Approaches/O to/O Authorship/O Attribution"./O XYZ/O |
| 10 | press./O 2015./T |
| 11 | Edward/A Martin,/A Daniel/A Clark./A "Learning/O to/O |
| 12 | generate/O natural/O paraphrases"./O Information/O &/O |
| 13 | Intelligence./O 31(4)./O pp./O 157-172./O 2014./O |

표 1의 가상 텍스트를 위 단계 (2)의 CRF 태깅 결과라고 가정하고, 임계치 α=0.8을 사용하여 표 1의 텍스트에 위 단계 (3)의 후처리 절차를 적용하여 인용열을 추출하는 과정을 설명하면 다음과 같다.

Line-01은 참고문헌의 첫 인용행이므로 항상 인용열의 시작 인용행으로 설정된다. Line-02는 첫 토큰 "Smith,"의 태그가 A이지만 직전 인용행(Line-01)의 모든 토큰의 태그가 A이므로 Condition-I을 만족하지 못하며 Condition-II도 만족하지 못하므로 시작 인용행으로 설정되지 않는다. Line-03, Line-04, Line-05, Line-06은 첫 토큰 태그가 A가 아니지만 직전 인용행의 마지막 토큰 태그가 T가 아니므로 Condition-II의 조건을 만족하지 못하여 시작 인용행으로 설정되지 않는다. Line-07은 첫 토큰 "Robert"의 태그가 A가 아니고 직전 인용행의 마지막 토큰 "2013."의 태그가 T이면서 인용행 길이 비가 0.67(=28/42)로 임계치 0.8미만이므로 Condition-II를 만족하여 시작 인용행으로 설정된다. Line-08은 첫 토큰 "Liang"의 태그가 A이면서 직전 인용행(Line-07)의 모든 토큰이 A인 것은 아니므로 Condition-I을 만족하여 시작 인용행으로 설정된다. Line-09, Line-10은 첫 토큰 태그가 A가 아니지만 직전 인용행의 마지막 토큰 태그가 T가 아니므로 Condition-II의 조건을 만족하지 못하여 시작 인용행으로 설정되지 않는다. Line-11은 첫 토큰 "Edward"의 태그가 A이면서 직전 인용행

(Line-10)의 모든 토큰이 A인 것은 아니므로 Condition-I을 만족하여 시작 인용행으로 설정된다. Line-12, Line-13은 첫 토큰 태그가 A가 아니지만 직전 인용행의 마지막 토큰 태그가 T가 아니므로 Condition-II의 조건을 만족하지 못하여 시작 인용행으로 설정되지 않는다.

위 후처리 적용 결과로부터 표 1의 13개 인용행 중 인용열의 시작 인용행은 Line-01, Line-07, Line-08, Line-11이다. 따라서 최종 추출되는 인용열 및 해당 인용열을 구성하는 인용행 번호를 나열하면 다음과 같다.

- 1st 인용열: Line-01, Line-02, Line-03, Line-04, Line-05, Line-06
- 2nd 인용열: Line-07
- 3rd 인용열: Line-08, Line-09, Line-10
- 4th 인용열: Line-11, Line-12, Line-13

제안된 방법은 시퀀스 태깅을 활용한다는 점에서 기존의 Grobid[5,8] 및 RefExt[4] 방법들과 유사하다(표 2 참조). 그러나 인용행을 태깅 단위로 사용하는 RefExt 방법과 달리, 제안된 방법은 Grobid 시스템과 유사하게 인용행 내 토큰을 태깅 단위로 사용한다. 한편 Grobid 시스템과 달리, 제안된 방법에서는 인용열의 시작 부분에 출현하는 저자명 단어들에 부착하는 Author 태그와 인용열의 마지막 단어에 부착하는 Tail 태그를 구분하여 학습하고 Author 및 Tail 태그 분류 결과를 이용하여 인용열의 시작 및 마지막 인용행을 인식하는 후처리 규칙을 적용한다.

Table 2. Comparison of CRF-based Approaches to Reference String Recognition

| Method | Tag set | Tagging unit |
|---|---|---|
| Grobid[5,8] | I-<label>, <label>, I-<reference>, <reference> | Token in Reference Line |
| RefExt[4] | B-REF, I-REF, O | Reference Line |
| Proposed method | Author, Tail, In, Year, Other | Token in Reference Line |

## IV. Experiments

### 1. Evaluation of Proposed Method

본 논문에서 제안된 방법의 평가를 위한 데이터셋은 전산 언어학 분야 영어 논문 250편에서 추출된 참고문헌 텍스트로부터 구축되었다. 이 논문들은 ACL Anthology Reference Corpus (ACL-ARC)[11] 버전 2.0의 논문 목록으로부터 임의 선택되었으며 참고문헌 텍스트가 [1], 1. 등의 마커 형식으로 표현된 논문은 배제하였다. 구축된 데이터셋(이후 ACL-ARC 데이터셋으로 표기)의 통계 정보는 표 3과 같으며 학습(train), 검증(validation), 테스트(test) 집합별로 각각 100, 50, 100편의 논문에서 추출된 참고문헌 인용행 및 인용열로 구성되어 있다.

Table 3. Statistics of ACL-ARC Reference String Recognition Dataset

| | Train | Validation | Test |
|---|---|---|---|
| Number of Papers | 100 | 50 | 100 |
| Number of Reference Lines | 9439 | 4300 | 8737 |
| Number of Reference String | 2238 | 1048 | 2100 |

학습 집합은 CRF 모델 학습에 사용되므로 토큰 단위 태그가 부착된 형식으로 수작업 구축되었으며, 태그 T는 크기 2이상 인용열의 마지막 토큰에만 부착하였다. 검증 집합은 최적 CRF 태그 집합 및 임계치 α의 최적값 결정을 위해 사용된다. 이를 위한 CRF 태그 집합 후보들로 AOT, AOIT, AOYT, AOIYT를, 임계치 후보 값들로 0.1, 0.2, ..., 0.9를 평가한다. 테스트 집합은 최종 인용열 인식 성능을 평가하기 위해 사용된다.

성능 평가 지표로 정확률(Precision), 재현율(Recall), F1이 사용된다. 정확률은 시스템이 예측한 전체 인용열 중 정답 인용열들의 비율로 정의하고, 재현율은 전체 정답 인용열 중 시스템이 올바르게 예측한 인용열들의 비율로 정의한다. F1은 정확률과 재현율의 조화평균이다. 이들 평가 지표를 micro-P, micro-R, micro-F1으로 명명한다. 그러나 인용열 단위의 이러한 micro 평가 지표는 상대적으로 많은 인용열로 이루어진 일부 논문들에 대한 인용열 인식 성능이 전체 성능에 상대적으로 큰 영향을 미칠 수 있다[4]. 전술한 문제를 보완하기 위해 논문 단위로 계산된 micro 성능들의 평균에 해당하는 macro-P, macro-R, macro-F1을 함께 제시한다.

시스템이 예측한 인용열이 정답 인용열인지를 판단하는 기준으로 시스템 추출 인용열과 정답 인용열 각각을 공백과 대쉬 문자를 제외한 문자열로 변환한 후 두 문자열이 일치하는지 여부를 검사하는 방식을 사용하였다.

인용행 토큰 태깅을 위한 CRF 학습 자질로는 기존 인용 필드 인식(Citation Field Recognition) 연구들[1,12]에서 시도된 자질들을 그대로 혹은 일부 변형하여 사용하였으며 단일 자질 목록은 표 4와 같다. 표에서 인명 사전은 미인구조사국 surname 목록[13]과 영어 대문자에 대한 약자 표기(예: P. 및 P.,) 목록으로 구축하였으며, 지명 사전은 위키피디아[14]의 도시 및 국가명 목록으로부터 구축하였다.

CRF 학습 도구로 CRF++ 툴[15]을 사용하였으며, 현재 토큰의 이전 이후 각각 2개 토큰의 단일 및 2-gram 자질들을 문맥 자질로 함께 학습하였다. CRF 학습 sequence의 단위로는 학습 집합 내 인용열로부터 생성되는 인용행들을 사용하였는데, 그 생성 방식은 학습 집합 내 실제 인용행들의 평균 길이를 구한 후 각 인용열을 이 평균 길이를 초과하지 않는 인용행들로 분해하는 방식이다. CRF 분류에서는 참고문헌에 출현한 각 원시 인용행을 태깅 대상 sequence로 사용하였다. 실험용 시스템은 파이썬 코드로 작성되었으며 리눅스 환경에서 실행되었다.

Table 4. Features for CRF Learning (Figures in Parentheses Indicate the Number of Features)

| Category | Features |
|---|---|
| Word (3) | token itself, token lower-cased, token with non-alphanumeric characters removed |
| Location (1) | location of token within the reference line |
| Dictionary (2) | check if token(or with ending non-alphanumeric characters removed) is in the person/location name list |
| Initial (1) | check if token is in the form of person name abbreviation (e.g., M. or M.,) |
| N-gram (8) | 1~4 character prefixes/suffixes of token |
| First & last character (2) | check if the first/last character of token is upper-cased, lower-cased, a digit, or the other (character itself in this case) |
| Number (1) | check if token contains year, dashed pages, a digit, or the other |
| Editor (1) | check if token contains editor expression such as (eds.) |

학습 및 검증 집합을 통해 최적 CRF 태그 집합 및 임계치 α의 최적값 결정 실험을 진행하였고 그 결과 그림 1, 그림 2에 보인 것처럼 CRF 태그 집합 {A,O,I,T}와 임계치 α=0.8의 조합이 검증 집합에 대해 micro 및 macro F1 기준 가장 높은 인용열 인식 성능을 보였다.

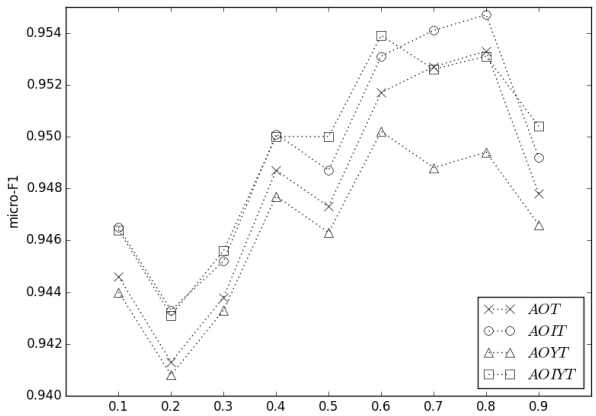


Fig. 1. Comparison of Micro-F1 Performances of Reference String Recognition with Different Tag Sets and α Values Using ACL-ARC Validation Set(X-axis Indicates α Values)

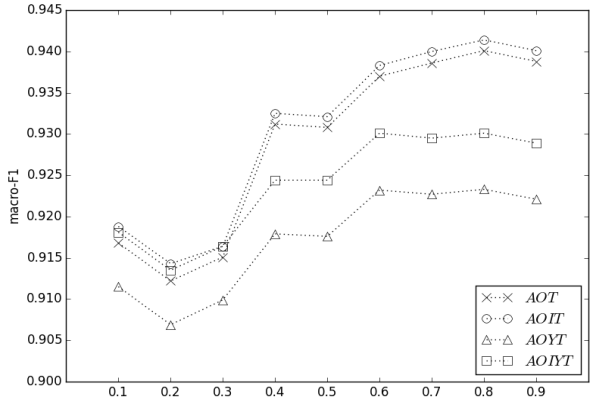


Fig. 2. Comparison of Macro-F1 Performances of Reference String Recognition with Different Tag Sets and α Values Using ACL-ARC Validation Set(X-axis Indicates α Values)

표 5는 ACL-ARC 학습 데이터 집합에 대해 10-fold 교차 검증 방식을 통해 얻은 CRF 모델의 AOIT 태깅 성능을 평가한 것으로, 제안된 방법의 시작 및 마지막 인용행 예측에 중요한 역할을 하는 두 레이블 A, T는 정확률과 재현율 모두 각각 98%, 94% 이상의 성능을 보였다.

Table 5. Performance of CRF Tagging on ACL-ARC Train Set with AOIT Tag Set

| Tag | Precision | Recall | F1 |
|---|---|---|---|
| A | 0.9815 | 0.9878 | 0.9846 |
| O | 0.9926 | 0.9910 | 0.9918 |
| I | 0.9992 | 0.9960 | 0.9976 |
| T | 0.9526 | 0.9437 | 0.9481 |

표 6은 제안된 방법의 후처리 절차를 다음 4가지 경우로 구분하여 검증 집합에 대한 인용열 인식 성능을 비교 제시한 것으로, 후처리 과정에서 사용되는 Condition-I, Condition-II의 각 요소들이 인용열 인식 성능에 기여하는 바가 적지 않음을 알 수 있다. 표 6에서 성능 표시 셀의 위/아래 수치는 각각 micro 및 macro 성능에 해당한다.

① Condition-I에서 다중 행 저자명 표현 여부 검사 미적용
② Condition-II에서 인용행 길이 비 임계치 α 검사 미적용
③ Condition-II에서 인용열 마지막 토큰 출현 검사 미적용
④ 제안된 방법 전체

Table 6. Evaluation of Reference String Recognition on ACL-ARC Validation Set with AOIT Tag Set (Upper and Lower Figure in Each Cell Correspond To Micro & Macro Performances, Respectively)

| Method | Pre. | Rec. | F1 |
|---|---|---|---|
| ① Proposed method<br>w/o multi-line author check | 0.6591<br>0.6828 | 0.8006<br>0.7948 | 0.7230<br>0.7316 |
| ② Proposed method<br>w/o α threshold | 0.8667<br>0.8599 | 0.9303<br>0.9184 | 0.8974<br>0.8863 |
| ③ Proposed method<br>w/o 'T' tag constraint | 0.9041<br>0.9106 | 0.9447<br>0.9414 | 0.9239<br>0.9248 |
| ④ Proposed method | 0.9440<br>0.9312 | 0.9656<br>0.9533 | 0.9547<br>0.9414 |

표 7은 AOIT 태그 집합으로 학습된 CRF 모델과 임계치 α=0.8을 사용하여, 제안된 방법의 최종 인용열 인식 성능을 ACL-ARC 테스트 집합에 대해 평가하여 제시한 것으로, F1 94% 이상의 성능을 보였다.

Table 7. Performance of Reference String Recognition on ACL-ARC Test Set with AOIT Tag Set and α=0.8(Upper and Lower Figure in Each Cell Correspond To Micro & Macro Performances, Respectively)

| Method | Precision | Recall | F1 |
|---|---|---|---|
| Proposed method | 0.9346<br>0.9442 | 0.9600<br>0.9572 | 0.9471<br>0.9498 |

그림 3은 테스트 집합 내 각 논문에 대한 인용열 인식 성능(F1)을 비교 제시한 것으로, 전체 논문의 70%에 대해 100%의 F1 인식 성능을 보였으며, F1 성능 80% 미만인 논문은 전체의 5%였다.

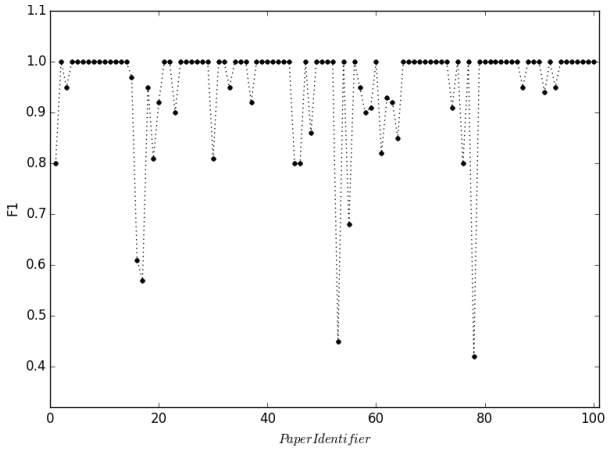


Fig. 3. Paper-wise F1 of Proposed Method Applied to ACL-ARC Test Set with AOIT Tag Set and α=0.8

## 2. Comparison to Previous Approaches

표 8은 본 논문에서 제안한 방법의 성능을 기존 인용열 인식 방법들과 비교 제시한 것이다. Körner 등은 그들이 제안한 RefExt 시스템과 함께 기존 Cermine, ParsCit, Grobid 시스템의 인용열 인식 성능을 비교 평가하였다[4]. 그들은 사회과학 분야 독일어 학술 문헌 100편으로부터 구축된 ssoar-German 데이터셋을 사용하여, 10-fold 교차 검증을 통해 macro 평균 방식의 정확률, 재현율, F1 성능을 비교 제시하였다. 표 8의 Cermine, ParsCit, Grobid, RefExt에 대응하는 행은 Körner 등의 논문에서 발췌한 수치이며, 기존 접근법 중에서는 RefExt 방법이 가장 높은 성능을 보였다.

Table 8. Comparing Performances of Previous Approaches to Reference String Recognition Using Macro-averaging Precision, Recall, F1

| Input | Method | | Pre. | Rec. | F1 |
|---|---|---|---|---|---|
| PDF File | Cermine | Figures from Körner et al.'s Article[4] | 0.303 | 0.220 | 0.235 |
| | ParsCit | | 0.617 | 0.595 | 0.590 |
| | Grobid | | 0.847 | 0.839 | 0.837 |
| | RefExt | | 0.879 | 0.906 | 0.885 |
| Reference Lines | RefExt-fair | | 0.9180 | 0.9064 | 0.9089 |
| | Proposed method | | 0.9456 | 0.9427 | 0.9434 |

표 8에서 RefExt-fair는, 본 논문의 방법과 동일한 입력 조건하에서 RefExt 시스템의 인용열 인식 성능을 ssoar-German 데이터셋에 대해 재평가한 것이다. Körner의 실험에서는 논문 전체 텍스트를 구성하는 각 행을 B-REF, I-REF, O 중 하나로 태깅한 결과로부터 인용열을 추출하였다[4]. 반면 RefExt-fair의 실험에서는 논문 전체 텍스트 대신 논문의 참고문헌 텍스트를 구성하는 인용행들만을 RefExt의 입력으로 사용하여 인용열을 추출하였다. 결과적으로 RefExt-fair의 경우 인용행의 성공적 추출을 가정함으로써 Körner의 실험에서보다 높은 인용열 인식 성능을 보였다.

표 8의 마지막 행은 본 논문의 방법을 ssoar-German 데이터셋에 대해 평가한 성능으로, 표 4의 Dictionary 자질의 인명사전 구축 부분을 제외하면 ACL-ARC 데이터셋을 대상으로 한 평가에서와 동일한 설정을 사용하여 얻은 결과이다. 즉 AOIT 기반 CRF 태깅과 임계치 α=0.8로 두고 인용열 인식을 수행하였다. Dictionary 자질의 인명사전의 경우 다음 목록들을 통합하여 구축하였다.

- 영어 대문자에 대한 약자 표기 목록
- 미인구조사국 surname 목록[13]
- 위키피디아 독일어 surname 목록[14]
- DBLP 저자명 목록[16]

위 목록에서 위키피디아 독일어 surname 목록은 ssoar-German 데이터셋이 독일어 학술문헌임을 고려하여 독일어 인명 데이터를 추가하기 위해 사용된 것이다. DBLP는 주요 컴퓨터공학 저널 및 학술대회발표논문집 게재 논문들의 서지정보를 제공하는 서비스이다. DBLP 저자명 목록은 DBLP 데이터 덤프로부터 추출하였다.

본 논문에서 제안된 방법은 ssoar-German 데이터셋에 대해 macro 평균 방식 성능 지표들에서 기존의 가장 우수한 접근법인 RefExt-fair 방법보다 높은 성능을 보였으며(표 8 참조), micro 성능 지표 관점에서도 성능 우위를 보였다(표 9 참조).

Table 9. Comparing Proposed Method and RefExt Method Using Micro-averaging Precision, Recall, F1

| Method | Pre. | Rec. | F1 |
|---|---|---|---|
| RefExt-fair | 0.9245 | 0.9182 | 0.9213 |
| Proposed method | 0.9328 | 0.9403 | 0.9365 |

# V. Conclusions

이 연구에서는 인용열 인식을 위한 새로운 방법을 제안하고 영어 및 독일어 데이터셋을 사용한 실험 결과를 제시하였다. 제안된 방법에서는 인용행의 각 토큰을 Author, Other, In, Tail 중 하나로 CRF 태깅한 후, Author 및 Tail 태그를 중심으로 인용행이 인용열의 시작 인용행인지 여부를 결정하는 후처리를 적용하여 인용열을 인식한다. 실험에서 제안된 방법은 전산언어학 분야 영어 논문들로부터 구축된 데이터셋과 기존 사회과학 분야 독일어 데이터셋에 대해 모두 94% 이상의 macro-F1 성능을 보였다.

참고문헌으로부터의 인용열 인식은 인용열 매칭 기법 및 인용필드태깅 기법과 함께 사용되어, 대용량 논문 집합으로부터

의 논문-논문 및 저자-저자 인용망 생성을 포함하여 논문 및 저자의 피인용지수 계산 등의 전처리 모듈로 활용 가능하다. 향후에는 보다 다양한 학문 분야의 참고문헌 표현 유형을 다룰 수 있도록 제안된 방법을 개선할 예정이다.

## REFERENCES


[1] I. Councill, C. Giles, and M.-Y. Kan, "ParsCit: an Open-source CRF Reference String Parsing Package," Proceedings of the 6th International Conference on Language Resources and Evaluation(LREC), 2008.

[2] R. Kern, and S. Klampfl, "Extraction of References Using Layout and Formatting Information from Scientific Articles," D-Lib Magazine, Vol. 19, No. 9/10, September/October, 2013.

[3] D. Tkaczyk, "*New Methods for Metadata Extraction from Scientific Literature*," PhD Thesis, ICM, University of Warsaw, 2015.

[4] M. Körner, B. Ghavimi, P. Mayr, H. Hartmann, and S. Staab, "Evaluating Reference String Extraction Using Line-Based Conditional Random Fields: A Case Study with German Language Publications," M. Kirikova et al. (Eds.): ADBIS 2017, CCIS 767, pp. 137-145, 2017.

[5] J. Boyd, "*Automatic Metadata Extraction The High Energy Physics Use Case*," Master's Thesis, CERN-THESIS -2015-105, 2015.

[6] Pdfextract, *https://www.crossref.org/labs/pdfextract/*

[7] D. Tkaczyk, P. Szostek, M. Fedoryszak, P. Dendek, and L. Bolikowski, "CERMINE: Automatic Extraction of Structured Metadata from Scientific Literature," International Journal on Document Analysis and Recognition(IJDAR), Vol. 18, No. 4, pp. 317-335, December, 2015.

[8] P. Lopez, "GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications," Proceedings of the 13th European Conference on Digital Libraries(ECDL), pp. 473-474, 2009.

[9] A. Bhardwaj, D. Mercier, A. Dengel, and S. Ahmed, "DeepBIBX: Deep Learning for Image Based Bibliographic Data Extraction," D. Liu et al. (Eds.): ICONIP 2017, Part II, LNCS 10635, pp. 286-293, 2017.

[10] J. Lafferty, A. McCallum, and F. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," Proceedings of the 18th International Conference on Machine Learning(ICML), pp. 282-289, 2001.

[11] S. Bird, R. Dale, B. Dorr, B. Gibson, M. Joseph, M.-Y. Kan, D. Lee, B. Powley, D. Radev, and Y. Tan, "The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics," Proceedings of the 6th International Conference on Language Resources and Evaluation(LREC), 2008.

[12] S. Anzaroot, and A. McCallum, "A New Dataset for Fine-grained Citation Field Extraction," Proceedings of the ICML Workshop on Peer Reviewing and Publishing Models, 2013.

[13] US Census Bureau, "Frequently Occurring Surnames from the 2010 Census", *https://www.census.gov/topics/population/genealogy/data/2010_surnames.html,* 2010.

[14] Wikipedia: The Free Encyclopedia. Wikimedia Foundation, Inc. 22 July 2004. Web. 10 Aug. 2004.

[15] CRF++: Yet Another CRF toolkit, *https://taku910.github.io/crfpp/*

[16] DBLP, *https://dblp.uni-trier.de/*


## 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, Kyungsung University. He is interested in natural language processing and information retrieval.