한국어 학술 문헌의 본문 인용문 인식을 위한 규칙 기반 방법*

A Rule-based Approach to Identifying Citation Text from Korean Academic Literature

강인수 (In-Su Kang)**

초 록

학술 문헌 원문에서 발견되는 인용문은 인용에 기초한 학술문헌 자동 요약, 리뷰 논문 자동 생성, 인용문 감성 분석, 인용문 기반 문헌 검색 등 다양한 학술 정보 서비스의 창출을 가능케 한다. 이러한 서비스가 가능하기 위해서는 원문 텍스트로부터 인용문의 자동 인식이 선행되어야 한다. 그러나 인용문의 인식은 인용 표지가 부착되지 않은 암묵 인용문의 존재로 인해 그 처리가 용이하지 않다. 영어의 경우 최근 이에 대한 연구가 집중되고 있으나 한국어 학술 문헌 내 인용문의 자동 인식 연구는 찾기 힘들다. 이 논문은 한국어 인용문을 자동 인식하는 규칙 기반의 방법을 제시하고 다양한 베이스라인 기법들과 인용문 인식 성능을 비교하였다. 제안된 방법은 테스트 셋 내 전체 암묵 인용문의 30%를 약 70%의 정확률로 인식할 수 있었다.

ABSTRACT

Identifying citing sentences from article full-text is a prerequisite for creating a variety of future academic information services such as citation-based automatic summarization, automatic generation of review articles, sentiment analysis of citing statements, information retrieval based on citation contexts, etc. However, finding citing sentences is not easy due to the existence of implicit citing sentences which do not have explicit citation markers. While several methods have been proposed to attack this problem for English, it is difficult to find such automatic methods for Korean academic literature. This article presents a rule-based approach to identifying Korean citing sentences. Experiments show that the proposed method could find 30% of implicit citing sentences in our test data in nearly 70% precision.

키워드: 인용문, 인용문 인식, 암묵 인용문, 인용문 인식 규칙, 인용문 단서 어구
citing sentences, citing sentence identification, implicit citing sentences, rules for identifying citing sentences, cue phrases for citing sentences

* 본 연구는 2012학년도 경성대학교 학술연구비지원에 의하여 연구되었음.

** 경성대학교 컴퓨터공학부 교수(dbaiskang@gmail.com)

■ 논문접수일자 : 2012년 11월 5일 ■ 최초심사일자 : 2012년 11월 10일 ■ 게재확정일자 : 2012년 12월 16일
■ 정보관리학회지, 29(4), 43-60, 2012. 〔http://dx.doi.org/10.3743/KOSIM.2012.29.4.043〕

1. 서 론

최근 학술 문헌의 원문에 기반한 학술 정보 서비스와 관련하여 문헌 내 인용문에 대한 관심이 증가하고 있다. 다음은 한 논문 P[1]에서 발췌한 인용문(citing sentence)의 예이다.

> 가장 잘 알려진 온톨로지 디버깅 방법은 HST (hittingset tree) 기반의 블랙박스 기법을 사용하여 주어진 지식 베이스에서 특정 개념을 유도하는 최소 공리 집합을 계산하는 알고리즘이다 [7,8]. Kalyanpur와 Parsia가 제안한 이 방법은 비논리적인 개념을 유도하는 모든 공리를 찾는다고 알려져 있다.

위 인용문은 첫 문장에서 인용 표지 [7,8]로 인용된 선행 연구에 대한 논문 P의 저자의 요약된 의견을 담고 있다. 일반적으로 인용은 선행 연구와 현재 연구와의 관련성을 밝히는 것이므로 많은 경우 인용문에는 위 예에서처럼 선행 연구에 대한 부정/긍정/중립/비교/대조적 의견 등이 포함된다(Teufel et al., 2006). 따라서 많은 후행 연구들로부터 수집된 동일 선행 연구에 대한 인용문 집합은 해당 선행 연구의 질적 가치를 평가하는 기초 자료가 될 뿐 아니라 인용에 기초한 학술문헌 자동 요약(Kaplan, Iida, & Tokunaga, 2009; Abu-Jbara & Radev, 2011), 리뷰 논문 자동 생성(Nanba, Kando, & Okumura, 2000), 인용문 감성 분석(Athar & Teufel, 2012), 인용문 기반 문헌 검색(O'Connor, 1982; Bradshaw, 2003; Ritchie, Robertson, & Teufel, 2008) 등 다양한 학술 정보 서비스의 창출을 가능케 한다.

그러나, 문헌의 원문 텍스트로부터 인용문을 인식하는 일은 간단하지 않다. 그것은 인용문 중 위 예의 두 번째 문장과 같이 인용 표지가 명시적으로 사용되지 않는 암묵 인용문이 적지 않기 때문이다(Athar & Teufel, 2012). 인용 표지가 부착된 위 예의 첫 번째 문장과 같은 명시 인용문의 경우는 정규표현식을 활용한 인용 표지 매칭을 통해 그 인식이 상대적으로 용이하다. 암묵 인용문 인식을 위한 기존 연구에서는 규칙 기반 방법(O'Connor, 1982; Nanba, Kando, & Okumura, 2000; Kaplan, Iida, & Tokunaga, 2009)과 통계적 방법(Qazvinian & Radev, 2010; Athar & Teufel, 2012)이 영어에 대해 시도되었으며 54%~78%의 성능(F1 혹은 F3)을 보이고 있다.

그러나 현재까지 한국어 학술 문헌에 대한 인용문 인식 연구는 찾기 힘들다. 이 연구에서는 한국어 학술문헌의 원문 텍스트 내 명시 및 암묵 인용문을 자동 인식하는 비교사(unsupervised) 접근법 중 규칙 기반 방법을 다룬다. 이를 위해 최근 정보과학회논문지 게재 논문 35편을 대상으로 한국어 인용문 인식 말뭉치를 구축하였고, 다른 출처(강인수, 2011)의 논문 56편에 출현한 1,048개 명시 인용문들과 그 인접 문장들의 용례 분석으로부터 한국어 인용문 인식 규칙 및 규칙 기반 인용문 인식 방법을 개발하였다. 실험에서는 제안한 규칙 기반 방법의 성능을 베이스라인 방법들과 비교하며 명시 인용문 주위 암묵 인용문 후보 문장 개수의 효과, 단위 규칙의

1) "김제민, 박영택 (2012). 온톨로지의 비논리적 개념을 유발하는 공리 탐지 기법. 정보과학회논문지: 소프트웨어 및 응용, 39(6), 464-472"로부터의 발췌문.

성능 기여도, 오류 문장 분석 등을 제시한다.

논문의 구성은 다음과 같다. 2장은 관련 연구를 기술한다. 3장에서는 한국어 인용문 인식 말뭉치 구축에 대해 설명하고, 4장은 이 연구에서 시도될 규칙 기반 한국어 인용문 인식 방법을 다룬다. 5장에서는 실험 절차와 결과를 기술하고 6장에서 결론을 맺는다.

2. 관련연구

인용문은 명시 인용문과 암묵 인용문으로 나뉜다. 명시 인용문은 제한된 표기법을 갖는 인용 표지들을 정규표현식(regular expression)에 기초하여 암묵 인용문에 비해 용이하게 매칭할 수 있다. 이와 관련하여 임의의 원문 텍스트로부터 후반부 참고문헌 리스트를 자동 추출하고 각 참고문헌 항목의 표기법으로부터 본문 내 인용 표지의 가능한 표현들을 자동 생성하는 방법(Councill, Giles, & Kan, 2008)은 이미 개발되었다.

따라서 대부분의 인용문 인식 연구는 암묵 인용문 인식에 집중한다. 최초의 인용문 인식은 인용문을 결합한 문서 표현이 검색 성능에 미치는 영향을 살피기 위해 O'Connor(1982)에 의해 시도되었다. O'Connor는 명시 인용문 주위 암묵 인용문을 인식하기 위해 단서 어구(this, these, those, such, same, similar, similarly, they, their, former, latter, above-mentioned 등)와 단락 경계 제약을 중심으로 총 16개 규칙을 정의하였다. O'Connor는 인용문 인식 자체의 성능 대신 인식된 인용문 기반 검색의 성능을 제시하였고 인용문의 사용으로 검색 재현율이 50%에서 70%로 향상됨을 보고하였다.

Nanba, Kando, Okumura(2000)은 특정 분야에 대한 리뷰 혹은 서베이(survey) 텍스트를 자동 생성하기 위한 기초 자료로 사용되는 인용문들을 얻기 위해 영어 인용문 인식을 시도하였다. 그들은 명시 인용문으로부터 출발하여 이전 이후의 문장 중 단락 경계 범위 내에서 미리 정의된 86개 단서 어구(for this, but, unlike, in our work, they, such, drawback 등)를 활용하여 암묵 인용문을 결정하였고, 50개 인용 구역을 갖는 평가셋에서 재현율 80%, 정확률 76%, F1 78%의 성능을 보였다.

전술한 규칙 방법들은 미리 정의된 인용문 단서구가 발견되지 않는 암묵 인용문을 인식하는 데 어려움이 있다. 이 문제를 다루기 위해 Kaplan, Iida, Tokunaga(2009)는 암묵 인용문의 경우 명시 인용문과 하나 이상의 명사구(예: 기존 방법의 명칭, 연구자 이름 등)를 공유할 것이라고 가정하고, 입력 텍스트에 대해 동일 지시어 참조 해결(Coreference resolution)을 미리 수행하였다. 이후 명시 인용문 전후 각 다섯 문장 내에서 임계치 이상의 유사도를 갖는 명사구 공유가 발견되는 인접 문장을 암묵 인용문으로 결정하는 절차(Coreference-chain 방법)를 적용하였고, 영어 인용문 94개를 갖는 평가셋을 사용한 실험에서 재현율 65%, 정확률 74%, F1 69%의 성능을 보였다.

Qazvinian과 Radev(2010)는 각 문장에 대해 MRF(Markov Random Field)를 만들고 BP(Belief Propagation) 추론을 통해 해당 문장이 인용문인지 여부를 결정하였다. 각 문장에 대해 정의되는 MRF 내에서는 현재 문장과 현재 문장의 이전 이후 n개씩의 각 문장에 대해 인용

표지 존재 유무, 미리 정의된 단서 어구 패턴(this, previous, approach, method 등)의 매치 여부, 피인용 논문과의 유사도 값 등이 자질로 사용되었다. 실험에서는 총 203개 인용문을 갖는 평가셋에 대해 평균 54%의 F3($F_{\beta=3}$) 성능을 보였다. 그들은 또한 IR(Information Retrieval) 기반 방법을 평가하였다. 이 방법은 명시 인용문의 이전/이후 각 방향으로 이동하면서 현재 문장이 이후/이전 문장과 특정 임계치 이상의 문장 유사도를 갖는 경우 현재 문장을 암묵 인용문으로 결정하는 것인데 평균 28%의 F3 성능을 보였다.

Athar와 Teufel(2012)은, 인용문이 피인용 논문에 대해 긍정적 의견을 표현한 것인지 부정적인 것인지를 판별하는 인용문 감성 분석 작업의 전처리 단계로 SVM(Support Vector Machine) 학습에 기반하여 암묵 인용문을 인식하고자 하였다. 기계 학습 자질로 한정사-명사 패턴(예: this method, his technique), 3인칭 대명사(예: they, he), 문장 접속어(예: however, although 등), 기존 방법의 명칭(예: HMM), 기존 연구자의 성(예: Turney's approach), 장/절 경계, 1~3까지의 n-gram 등을 분류 대상인 현재 문장과 그 이전 이후 문장에 대해 추출하였다. 3,760개 영어 인용문으로 구성된 평가셋에서 제안된 방법은 2,016개 암묵 인용문에 대해 51%의 F1 성능을 보였다. 이는 명시 인용문을 배제한 순수 암묵 인용문의 인식 성능이다.

일반적으로 인용 표지의 범위는 구, 절, 문장, 다중 문장일 수 있다(Kang & Kim, 2012). Abu-Jbara와 Radev(2011/2012)는 특정 피인용 논문의 요약을 위해 해당 논문에 대한 인용 부분만을 추출할 필요가 있었으며 이를 위해 하나의 인용문을 하나 이상의 인용구나 인용절로 분리하는 시도를 하였다. 그러나 이 연구들은 명시 인용문에 대해서만 분리 시도를 하였고 암묵 인용문 인식은 다루지 않았다.

3. 한국어 인용문 인식 말뭉치 구축

한국어 인용문 태그 부착 말뭉치 구축을 위한 자료원으로 한국정보과학회 발간 정보과학회논문지(소프트웨어 및 응용 분야) 최신 호들(39권 5, 6, 7, 8호)로부터 총 35편의 논문[2]을 선택하였다. 선택된 각 논문의 서론부터 결론까지의 본문 텍스트의 전체 문장 중 장/절 제목 문장에 대해 먼저 다음 예와 같은 장/절 태그(〈Section〉, 〈/Section〉)를 부착하였다.

- 〈Section〉1. 서론〈/Section〉
- 〈Section〉2. 관련 연구〈/Section〉
- 〈Section〉3. 인용문 인식〈/Section〉
- 〈Section〉3.1 규칙 기반 인용문 인식 〈/Section〉
- 〈Section〉3.1.1 인용문 규칙〈/Section〉

다음으로 각 논문의 본문 텍스트 내 전체 문장에 대해 아래 절차에 따라 인용 태그(〈cite〉, 〈/cite〉)를 수작업 부착하였다.

2) 이 35편 논문의 본문 텍스트 발췌문들이 5장에서 직접 인용 방식으로 사용되었다.

1. 인용표지가 부착된 문장 S를 찾고 S에 인용 태그를 부착한다.
2. S가 속한 장/절 경계 범위 내에서 S 전후 문장들 중 S의 인용 논문에 대해 기술된 각 문장에 인용 태그를 부착한다.
3. S가 속한 장/절 경계를 벗어나는 전체 논문 텍스트 중 S에 출현한 키워드를 하나 이상 포함하는 문장들 중 S의 인용 논문에 대해 기술된 각 문장에 인용 태그를 부착한다.

인용 태그는 문장의 시작 어절 앞에 시작 태그 〈cite〉를 부착하고, 문장의 마지막 어절 뒤에 종료 태그 〈/cite〉를 부착함으로써 해당 문장이 인용문임을 표시하는 용도로 사용된다. 또한 인용 태그는 속성(attribute) id를 가지며 그 값(value)으로 인용 태그가 부착된 문장이 인용하고 있는 논문(들)의 (참고문헌 영역 내) 순서 번호를 둘 이상일 경우 콤마(,)로 분리된 문자열로 갖도록 정의된다. 다음 세 문장[3]은 인용 태그가 부착된 실제 문장의 예이다.

- 〈cite id="3"〉그림자 맵의 해상도 문제를 해결하기 위해서 Adaptive Shadow Map (ASM)[3]에서는 그림자의 경계부분을 찾아서 그 부분만 높은 해상도의 그림자 맵을 적용하는 방법을 사용하였으나 반복적으로 그림자 맵 트리를 구성하기 때문에 트리 구성시간이 오래 걸려서 실시간성(real-time)을 보장하지 못한다는 단점이 있다.〈/cite〉
- 〈cite id="4"〉이러한 단점을 해결하기 위해 Resolution-matched Shadow Maps (RMSM)[4]에서는 트리생성 시간을 단축하기 위해 ASM의 반복 알고리즘을 삭제하고 개선하여 결과물의 품질을 크게 해치지 않는 선에서 성능개선에 성공하였다.〈/cite〉
- 〈cite id="3,4"〉하지만 위 알고리즘들은 실시간성을 보장하기 힘들기 때문에 일반적으로 자주 사용되지 않는다.〈/cite〉

위의 첫째, 둘째 문장은 인용 표지가 명시적으로 부착된 명시 인용문들이며, 마지막 문장은 인용 표지를 명시적으로 사용하지 않으면서 암묵적으로 문헌 [3,4]에 대해 기술하고 있는 암묵 인용문에 해당한다. 위 마지막 문장이 문헌 [3,4]의 내용을 인용하고 있는가에 대해서는 사람에 따라 판단의 차이가 있을 수 있다. 이와 관련하여 본 연구에서는 기존 문헌의 내용 일부분을 직접(그대로) 및 간접(말바꿈) 기술하는 것을 포함하여 기존 문헌의 내용에 대한 저자의 의견을 포함하고 있는 문장을 인용문으로 고려하고 인용 태그 부착을 수행하였다.

〈표 1〉은 인용문 태그 부착 말뭉치의 통계치이다. 전체 35편 논문 내 6,075개 문장 중 13%에 해당하는 791개 문장에 인용문 태그가 부착되었고, 전체 인용문 중 명시 인용문과 암묵 인용문 비율은 각각 70%, 30%였다.

〈표 1〉 인용문 태그 부착 말뭉치 통계

논문	문장	인용문	명시인용문	암묵 인용문
35	6,075	791 (13%)	548 (70%)	243 (30%)

3) "김상훈, 김민철, 최원익 (2012). Cascaded Shadow Maps의 경계문제를 해결한 효율적인 분산-혼합 그림자 매핑 알고리즘. 정보과학회논문지: 소프트웨어 및 응용, 39(5), 355-366"로부터의 발췌문.

4. 규칙 기반 한국어 인용문 인식

규칙을 사용하는 암묵 인용문 인식 절차는 알고리즘 1과 같으며, 입력된 논문 한 편의 원문 텍스트 내 각 문장에 대해 그 문장이 인용 표지를 가진 명시 인용문일 경우 정해진 범위 내에서 이전 이후 문장들이 해당 명시 인용문의 암묵 인용문인지 여부를 규칙을 이용하여 판단한다. 알고리즘에서 k는 입력 논문 내 총 문장의 개수이고, s[i]는 입력 논문의 i번째 문장을 의미한다(Line 1). class[i]는 0인 경우 s[i]가 인용문이 아님을, 1인 경우 인용문임을 의미하며, 최초 모든 문장에 대해 0의 값으로 초기화되어 있다(Line 2). 이후 총 k개 각 문장 s[i]에 대해, Line 4~Line 16을 반복 수행하면서, 현재 문장 s[i]가 인용 표지를 포함하고 있을 경우(Line 4), s[i]에 (명시) 인용문 표시를 설정하고(Line 5), s[i] 이전 m개, 이후 n개 문장들 중에서 규칙에 기반하여 s[i]의 인접 암묵 인용문을 찾는 절차(Line 6~Line 10, Line 11~Line 15)를 수행한다.

규칙은 긍정과 부정 두 유형으로 나뉘며, 모든 부정 규칙에 매치되지 않으면서(Line 8, Line 13) 긍정 규칙 집합 중 하나에 매치되면(Line 9, Line 14) 현재 암묵 인용문 후보 문장 s[j]를 인용문으로 설정(class[j]=1)하는 방식으로 사용된다. 인용문 인식을 위한 규칙은 〈표 2〉에 보인 것처럼 부정 규칙 2개(N1~N2), 긍정 규칙 9개(P1~P9)로 구성되었다. 이 규칙들은 3장의 자료원과 다른 출처(강인수, 2011)의 국내 학회 56개 학술지 최신 논문들에 출현한 총 1,048개 명시 인용문의 인접 문장들을 용례 분석하여 얻어졌다. 표에서 ₩s+는 공백, 탭, 캐리지리턴, 뉴라인 문자들의 나열을, P8에서 .+은 임의의 문자열을 각각 의미한다.

명시 인용문 주위의 암묵 인용문을 탐색하는 과정에서 두 가지 가정을 사용하였다. 첫째는 특정한 명시 인용문의 암묵 인용문들은 해당 명

	Algorithm 1: Rule-based Citation Sentence Recognition
1	**Input s**[1..**k**] : an array of **k** sentences in an article
2	**Output class**[1..**k**] : an array of value **0s**
3	FOR i = 1 ~ **k**
4	IF **s**[i] has a citation marker THEN
5	**class**[i]=1
6	FOR j = i-1 ~ i-1+**m** // m≤0: Backward Window
7	IF **s**[j] is a section title THEN EXIT_FOR
8	IF **s**[j] is matched to a negative rule THEN EXIT_FOR
9	IF **s**[j] is matched to a positive rule THEN **class**[j]=1 ELSE EXIT_FOR
10	END_FOR
11	FOR j = i+1 ~ i+1+**n** // n≥0: Forward Window
12	IF **s**[j] is a section title THEN EXIT_FOR
13	IF **s**[j] is matched to a negative rule THEN EXIT_FOR
14	IF **s**[j] is matched to a positive rule THEN **class**[j]=1 ELSE EXIT_FOR
15	END_FOR
16	END_IF
17	END_FOR

〈표 2〉 인용문 인식 규칙

번호	규칙
N1	(우리)
N2	(제시 | 제안)(한다)
P1	(발견 | 발명 | 발표 | 보고 | 연구 | 주장 | 증명)(됐 | 되고 | 되었 | 된 | 한 | 했)
P2	(고 | 로 | 에서)₩s+(밝혀 | 알려)(져 | 졌 | 진)
P3	(고 | 에서 | 으로)₩s+(발표 | 보고 | 실험 | 제안 | 평가)(되 | 된)
P4	(와 관련된)₩s+(기법 | 노력 | 논문 | 발표 | 방법 | 보고 | 시도 | 실험 | 아이디어 | 알고리즘 | 연구 | 제안)
P5	(기법 | 노력 | 논문 | 발표 | 방법 | 보고 | 시도 | 실험 | 아이디어 | 알고리즘 | 연구 | 제안)(들)
P6	(그 | 그러한 | 그런 | 앞 | 앞의 | 위 | 위와 같은 | 위의 | 이 | 이러한 | 이런 | 전술한)₩s+(기법 | 노력 | 논문 | 발표 | 방법 | 보고 | 시도 | 실험 | 아이디어 | 알고리즘 | 연구 | 제안)
P7	(다른 | 다양한 | 다음과 같은 | 다음의 | 많은 | 여러 | 최근)₩s+(기법 | 노력 | 논문 | 발표 | 방법 | 보고 | 시도 | 실험 | 아이디어 | 알고리즘 | 연구 | 제안)
P8	(그러나 | 그렇지만 | 하지만).+(단점 | 문제 | 쉽지 않 | 어려움 | 어렵 | 제약 | 힘들)
P9	(그 | 그 기관 | 그 대학 | 그 연구그룹 | 그 연구소 | 그 연구팀 | 그녀 | 그들)(는 | 에서 | 의 | 은)

시 인용문이 속한 장이나 절의 경계를 벗어나지 못한다는 것이고, 둘째는 특정 명시 인용문과 관련된 암묵 인용문들은 텍스트 내에서 연속된 문장들로 출현한다는 가정이다. 첫째 가정을 반영하기 위해 알고리즘에서는 암묵 인용문 후보 문장이 장/절 제목 문장에 해당하면 암묵 인용문 탐색을 중지하도록 하였다(Line 7, Line 12). 둘째 가정인 인용문 연속성을 반영하기 위해 첫 비인용문 출현 지점에서 암묵 인용문 탐색을 중지한다(Line 8,9,13,14의 EXIT_FOR).

이 알고리즘은 영어에 대해 기존에 시도된 단서 어구 기반의 규칙 방법들(O'Connor, 1982; Nanba, Kando, & Okumura, 2000)의 절차와 크게 다르지 않으나 한국어 인용문 인식에 특화된 단서 어구들을 수집하고 그들을 긍정과 부정 규칙들로 구분하여 처리한 것은 기존 연구에서 시도되지 않은 부분이다.

5. 실 험

한국어 인용 텍스트 인식의 비교사 기법 성능 평가를 위해 3장에서 기술한 데이터 집합을 사용하였다. 인용 텍스트 인식은 한 편의 문헌을 입력으로 받아 문헌 내 각 문장에 대해 다른 논문의 내용을 인용한 것인지 여부를 결정하는 것이다. 인용문 인식을 위한 베이스라인 기법으로 선행 연구(Kaplan, Iida, & Tokunaga, 2009; Qazvinian & Radev, 2010)의 실험을 준용하여 인용 표지 방법(Citation-marker method), 윈도우 방법(Window method), 랜덤 방법(Random method)을 사용한다. 입력 텍스트에 대해 동일 지시어 참조 해결이 선행되어야 하는 Kaplan, Iida, Tokunaga(2009)의 Coreference-chain 방법은 한국어 텍스트에 대해 활용 가능한 대용어 참조 해결 모듈의 부재로 실험에서 배제하였다.

인용 표지 방법은 "(Salton et al., 1981)", "[3,4]"와 같은 인용 표지가 명시적으로 부착된

명시 인용문만을 인용문으로 인식하는 방법이다. 인용 표지 방법은, Kaplan 등의 실험에서 단서 어구 기반 규칙 방법(Nanba, Kando, & Okumura, 2000)과 Coreference-chain 방법이 그 성능을 능가하지 못할 만큼 인용문 인식에서 강한 베이스라인에 해당한다. 윈도우 방법은 입력 문헌 내 각 명시 인용문의 이전 m개 문장들과 이후 n개 문장들을 무조건적으로 암묵 인용문으로 결정하는 것이다. 랜덤 방법은 각 명시 인용문의 이전 m개 문장, 이후 n개 각 문장에 대해 암묵 인용문 여부를 무작위 결정하는 것이다.

또한 기존 비교사 인용문 인식법 중 문장 유사도 기반 방법을 평가하였다. 유사도 기반 방법(Similarity-based method)은 Qazvinian과 Radev(2010)가 시도한 IR 기반 방법과 같은 것이다. 이 방법은 명시 인용문의 이전/이후 각 방향으로 이동하면서 현재 문장이 이후/이전 문장과 특정 임계치(Threshold T) 이상의 문장 유사도를 갖는 경우 현재 문장을 암묵 인용문으로 결정하는 것이다. 문장 유사도 계산을 위해 TF-IDF 기반 용어 가중치를 사용하는 코사인 유사도 수식을 사용하였다. IDF(역문헌빈도, Inverse Document Frequency) 값은 데이터 셋 내 전체 문헌 집합에 출현한 각 문장을 하나의 문서 단위로 고려하여 계산하였고, 용어 가중치 부여 기법은 lnc.ltc를 사용하였다(Singhal, Salton, & Buckley, 1996).

유사도 기법의 문장 표현을 위해 아래 예와 같이 n-gram과 형태소 기반 용어 표현을 시도하였다. 문장의 n-gram 용어로는 음절 2-gram을 추출하였다. 형태소(morpheme) 용어 표현을 위해 먼저 문장에 대해 형태소 분석 및 품사 태깅[4]을 수행한 다음 명사, 동사, 형용사, 부사의 네 개 품사에 해당하는 실질 형태소와 미등록 명사, 영어 단어들을 추출[5]하여 사용하였다.

- 문장: Allan이 제안한 이 비교사 알고리즘은 …
- 2-gram 용어: Allan, 이, 제안, 안한, 이, 비교, 교사, 알고, 고리, 리즘, 즘은, …
- Morpheme 용어: Allan, 제안, 비교사, 알고리즘, …

평가 척도로 다음과 같이 정의되는 정확률(Precision), 재현율(Recall), F1 지표를 사용하였다.

- A = 데이터 셋 내 명시 및 암묵 인용문의 집합
- S = 시스템이 결정한 명시 및 암묵 인용문의 집합
- 재현율 R = | A ∩ S | / | A |
- 정확률 P = | A ∩ S | / | S |
- F1 = 2×P×R / (P+R)

〈표 3〉은 명시 인용문 이전과 이후 윈도우 크기 파라미터인 m, n의 값을 각각 m=-5~0,

4) 한국어 형태소 분석 및 품사 태깅에 대한 국내 대표적 연구용 분석기 중 하나인 POSTECH KLE 연구실(http://kle.postech.ac.kr/)의 KoMA(Korean Morphological Analyzer) & KLE-Tagger를 사용하였다. 이 분석기의 최신 성능은 형태소 분석과 품사 태깅에서 각각 98%, 95%이다(김세종, 2012).

5) 명사류만 추출한 경우에도 성능에는 큰 변화가 없었다.

〈표 3〉 암묵 인용문 인식 성능

(W=m~n: Window range, T: Threshold, std: standard deviation)

Method	Pre.	Rec.	F1	Parameters
Random	0.2202	0.4900	0.3038 (avg) 0.3439 (max) 0.2715 (min) 0.0182 (std)	W=-4~2
Window	0.2489	0.6790	0.3642	W=0~2
Similarity-based (2-gram)	0.2613	0.6173	0.3672	W=0~2, T=0.05
Similarity-based (morpheme)	0.2709	0.6132	0.3758	W=0~2, T=0.01
Rule-based	**0.6916**	**0.3045**	**0.4229**	**W=-2~3**

n=0~+5의 범위 내에서 변화시키면서 각 암묵 인용문 인식 방법의 최적 성능을 구한 것이다. 유사도 기반 방법의 경우 윈도우 크기 변화와 함께 추가적으로 임계치 T를 0.01, 0.02, … , 0.09, 0.1, 0.2, … , 0.9로 변화시키면서 최적 성능을 구하였다. 표에서 W=-2~3는 m=-2, n=+3을 뜻하며 이는 규칙 기반 방법에서 명시 인용문 이전 2개 문장과 이후 3개 문장들을 암묵 인용 결정의 후보문으로 고려했을 때 최고 성능을 보였음을 의미한다. T=0.01은 임계치 값이 0.01일 경우 형태소 기반 유사도 방법이 최고 성능을 보였음을 의미한다. 랜덤 방법의 경우 각 문장에 대한 임의 결정 값이 일정하지 않은 문제가 있으므로, 랜덤 방법을 100회 수행한 평균 성능을 구하였다.

인용 표지가 부착된 명시 인용문을 찾기 위해 다음과 같은 간단한 정규표현식(regular expression)을 사용하였다. 이 패턴들은 [1], [2-3], [3,4-5] 등의 인용 표지를 매치하기 위해 정의된 것으로 숫자(0-9), 콤마(,), 대쉬(-), 공백 문자 중 하나 이상이 대괄호로 감싸인 문자열을 매치하되 그 중 대괄호 내 숫자를 포함하지 않는 문자열은 제외한다는 의미이다.

- Positive pattern: ₩[[0-9,-]+₩]₩
- Negative pattern: ₩[[^0-9]+₩]

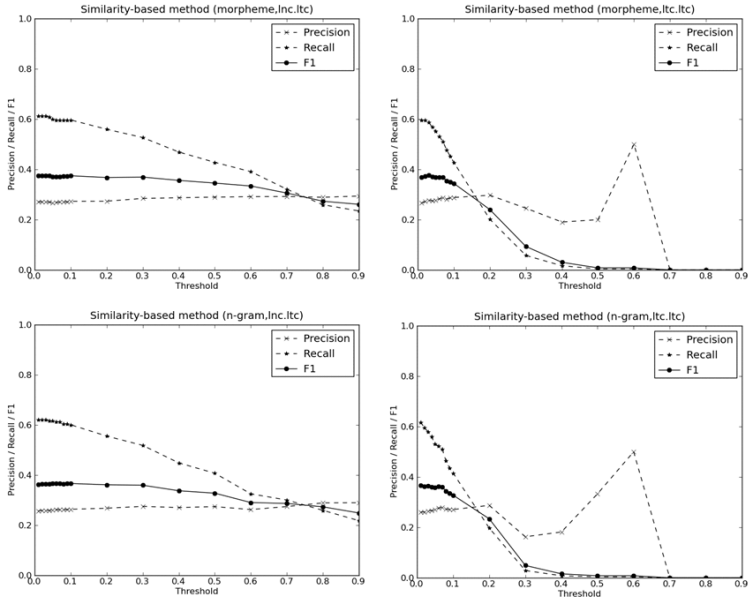
이 방법은 인용 표지 방법에 해당하며 실험 집합 내 모든 명시 인용문을 100%의 정확률로 찾아낼 수 있었다. 그러나 이것은 정보과학회논문지 게재 논문으로 이루어진 현재 평가셋에 국한된 것이다. 따라서 위 정규표현식은 향후 실용적 목적을 위해 "(Smith et al., 2011)", "홍길동 등(2012)", "(홍길동과 이영희, 2011, p. 179)", "(Kim, 2009a; Hong, 2012)"와 같은 다양한 형식의 인용 표지들을 포함할 수 있도록 일반화될 필요가 있다.

랜덤 방법에 비해 다른 베이스라인 방법들은 더 많은 암묵 인용문을 올바르게 결정함으로써 랜덤 방법의 재현율과 정확률을 동시에 향상시키고 있다. 윈도우 방법이 랜덤 방법보다 높은 성능을 보인 것은 암묵 인용문들의 출현이 연속적이라는 가정을 뒷받침하는 결과로 해석된다. 유사도 방법의 경우에도 주어진 윈도우 범위 내

에서 명시 인용문 주위의 암묵 인용문들이 연속적으로 인접해 있다는 가정을 사용한 경우의 성능이 그렇지 않은 경우보다 약간의 성능 향상이 있었다. 〈표 3〉은 랜덤 방법을 제외한 모든 방법에서 규칙 기반 방법에서 사용된 단락/절 경계 제약과 암묵 인용문의 연속성 가정을 적용하여 얻은 성능들이다.

문장 유사도 기반 방법의 경우 윈도우 방법에 비해 정확률을 향상시키면서 전체적으로 약간의 성능 향상을 보임으로써 명시 및 인접 암묵 인용문들의 문장 유사도가 암묵 인용문 결정에 긍정적으로 기능할 수 있음을 보였다. 유사도 방법의 효과를 좀 더 살펴보기 위해 〈그림 1〉에서 유사도 임계치 변화에 따른 유사도 방법의 성능 추이를 lnc.ltc, ltc.ltc의 두 가지 가중치 부여 기법(Singhal, Salton, & Buckley, 1996)에 대해 제시하였다. 그림을 통해 임계치가 증가할수록, 재현율의 완만한 감소에 따라 정확률이 미미하게 증가하거나(lnc.ltc의 경우) 재현율의 급격한 감소와 함께 오히려 정확률이 감소하여(ltc.ltc의 경우) 전체적인 F1의 감소를 보였고, 0.01(morpheme), 0.05(2-gram)와 같은 낮은 임계치에서 최고 성능을 보였다. 이는 암묵 인용문 판단에서 문장 간 다수 중요 용어들의 공유보다 일부 특정 용어(예: 〈표 2〉의 단서 어구, 기존 방법의 명칭, 연구자 성 등)의 출현이 상대적으로 더 중요할 수 있음을 간접적으로 보이는 결과이다. 〈표 3〉의 성능은 lnc.ltc를 사용한 것이다.

규칙 기반 방법은 다른 베이스라인 방법들에



〈그림 1〉 유사도 임계치에 따른 암묵 인용문 인식 성능 변화[6] (Similarity-based method)

6) ltc.ltc의 임계치 0.5, 0.6 지점들은 시스템이 추정한 암묵 인용문 개수가 극히 작아(morpheme의 경우 5개, 2개, 2-gram의 경우 3개, 2개) 높은 정확률을 보였음.

비해 정확률을 크게 높이면서 전체적인 성능을 향상시켜(〈표 3〉 참조), 규칙의 사용이 암묵 인용문 판별에 효과가 있음을 보였다. 〈표 4〉는 명시 인용문을 포함한 전체 인용문에 대해 규칙 기반 방법의 인용문 인식 성능을 보인 것이다. 명시 인용문만을 인식하는 인용 표지 방법(Citation-marker method)에 비해 규칙 기반 방법은 전체 243개 암묵 인용문의 30%를 약 70%의 정확률로 찾아낼 수 있음을 알 수 있다. 인용문 인식의 경우 자동 인식된 인용문을 활용하는 응용 분야를 고려할 때 인용문 누락보다 비인용문 포함의 문제가 더 심각할 수 있으므로 정확률을 재현율보다 더 중요하게 고려하는 $F_{\beta=0.5}$와 같은 평가척도를 동시에 사용할 필요가 발생한다. 이러한 이유로 〈표 4〉에서는 $F_{\beta=0.5}$ 수치를 함께 제시하였고, 재현율보다 높은 정확률을 보이는 규칙 기반 방법은 F1에 비해 $F_{\beta=0.5}$에서 높은 성능을 보였다.

〈그림 2〉는 암묵 인용문 결정의 후보로 고려되는 (명시 인용문 주위) 문장의 범위에 따른 규칙 기반 인용문 인식 성능을 그래프로 나타낸 것이다. 그림을 통해 $m \leq -1$, $n \geq 3$인 윈도우 범위에 대해서는 규칙 기반 방법의 성능은 큰 차이가 없음을 알 수 있다. 이는 암묵 인용문 후보

〈표 4〉 규칙 기반 방법의 전체(명시+암묵) 인용문 인식 성능

(괄호 안은 암묵 인용문 인식 성능 수치)

Method	총 인용문 수	인용문 분류 문장 수	True Positive (TP)	정확률	재현율	F1	$F_{\beta=0.5}$
Citation-marker	791 (243)	548 (0)	548 (0)	1.0000 (0.0000)	0.6928 (0.0000)	0.8185 (0.0000)	0.9185 (0.0000)
Rule-based	791 (243)	655 (107)	622 (74)	0.9496 (0.6916)	0.7863 (0.3045)	0.8603 (0.4229)	0.9118 (0.5514)

〈그림 2〉 윈도우 크기에 따른 규칙 기반 방법의 성능

로 최소한 명시 인용문의 이전 문장 1개와 이후 문장 3개를 포함시켜야 함을 의미한다. 또한 암묵 인용문 후보로 명시 인용문 이전 문장보다 이후 문장들을 더 많이 고려하는 것이 규칙 기반 인용문 인식에 더 효과적임을 의미한다. 이것은 암묵 인용문들이 대체로 명시 인용문의 이전보다 이후에 더 많이 출현한다는 기존 연구(Kang & Kim, 2012)와도 상충되지 않는 결과이다.

다음으로 4장에서 정의된 인용문 규칙 집합을 단일 규칙들로 분해하여 각 단일 규칙의 인용문 인식 성능을 구해 보았다. 다음은 4장의 한 복합 규칙을 단일 규칙들로 분해한 예이다.

- 원 규칙: (고|로|에서) (알려|밝혀) (져|졌|진)
- 단일 규칙들: (고) (알려)(져), (고) (알려)(졌), (고) (알려)(진), (고) (밝혀)(져), (고) (밝혀)(졌), (고) (밝혀)(진), (로) (알려)(져), (로) (알려)(졌), (로) (알려)(진), (로) (밝혀)(져), (로) (밝혀)(졌), (로) (밝혀)(진), (에서) (알려)(져), (에서) (알려)(졌), (에서) (알려)(진), (에서) (밝혀)(져), (에서) (밝혀)(졌), (에서) (밝혀)(진)

위와 같은 분해 절차로 4장의 긍정 규칙들을 분해하여 총 395개의 긍정 단일 규칙을 생성했다. 이후 4장의 모든 부정 규칙들과 (395개 중) 하나의 긍정 단일 규칙으로만 구성된 규칙 집합을 사용하여 규칙 기반 인용문 인식을 서로 다른 긍정 단일 규칙에 대해 총 395회 수행하였다. 윈도우 범위는 〈표 3〉에서와 같이 W=-2~3로 설정하고 실험하였다. 〈표 5〉는 그 결과이다.

실험에 사용된 데이터 셋의 경우, 총 395개 단일 규칙 중 46개(11.6%)가 적용되어 총 132개 문장을 암묵 인용문으로 분류하였으며 그 중 93개 문장이 올바른 결정이었다. 이는 전체 규칙 집합이 사용된 경우의 암묵 인용문 분류 개수 107, 올바른 암묵 인용 결정 문장 수 74와 차이가 있는데(〈표 4〉 참조), 그 이유는 하나의 문장에 매치되는 단일 규칙이 하나 이상일 수 있기 때문이다. 〈표 5〉은 전체 단일 규칙 중 "이 방법", "방법들", "연구들", "이 연구", "그러나~문제"와 같은 소수 규칙들이 규칙이 적용된 암묵 인용 후보 문장의 50%에 매치되고 있으며 이들의 평균 정확률은 80% 이상임을 보인다.

다음은 규칙 기반 방법으로 잘못 분류된 FP (False Positive) 문장들 몇 개를 보인 것이다. FP 문장은 밑줄로 표시하였고 FP 문장 내 볼드체 어구들은 해당 문장이 암묵 인용문으로 분류될 때 적용된 규칙에 해당한다. 또한 관점에 따라 인용문으로 판단될 가능성이 있는 FP 문장의 경우 해당 문장이 비인용문으로 태깅된 근거를 괄호 내 이탤릭체로 표시하였다.

- FP-1: 일반적으로 협력적 여과는 내용기반 여과 등 다른 추천 방법과 비교하면 추천 정확도 등 성능이 좋아 영화, 서적, 음반 등 여러 도메인에서 상용 서비스로 많이 활용되고 있다[2-4]. **그러나** 기존 TV에 비하여 서비스하는 콘텐츠의 양과 사용자의 시청 이력이 방대한 IPTV 환경에서는 기존 서적, 음악 등의 추천 알고리즘을 그대로 이용하기에는 여러 가지 **제약**이 있다 *(이 논문 저자의 의견인지 [2-4] 문헌의 내용인지 명확하지 않음)*.

〈표 5〉 단일 규칙 인용문 인식 성능

단일 규칙	인용문 분류 문장 수	TP (True Positive)	정확률
이 방법	23	19	0.83
방법들	13	9	0.69
연구들	13	13	1.00
이 연구	9	7	0.78
그러나.*문제	9	6	0.67
이러한 방법	4	3	0.75
이 논문	3	0	0.00
하지만.*문제	3	3	1.00
그러나.*힘들	3	3	1.00
알고리즘들	3	3	1.00
이 알고리즘	3	2	0.67
연구₩s*되었	2	1	0.50
이 제안	2	1	0.50
위의 연구	2	2	1.00
하지만.*힘들	2	2	1.00
고 알려져	2	1	0.50
그러나.*어렵	2	0	0.00
그러나.*단점	2	2	1.00
하지만.*어려움	2	1	0.50
그의	2	1	0.50
하지만.*단점	2	2	1.00
발표₩s*했	2	1	0.50
에서 제안된	1	1	1.00
위의 방법	1	1	1.00
이 기법	1	1	1.00
이런 연구	1	1	1.00
하지만.*어렵	1	1	1.00
그러나.*어려움	1	0	0.00
여러 연구	1	1	1.00
다양한 방법	1	0	0.00
발표₩s*되었	1	0	0.00
이러한 기법	1	1	1.00
와 관련된 연구	1	0	0.00
연구₩s*되고	1	0	0.00
노력들	1	0	0.00
위 알고리즘	1	1	1.00
기법들	1	1	1.00
그러나.*제약	1	0	0.00
로 알려져	1	0	0.00
이러한 연구	1	1	1.00
고 알려진	1	0	0.00
그들의	1	1	1.00
많은 연구	1	0	0.00
이 시도	1	0	0.00
연구₩s*된	1	0	0.00
연구₩s*됐	1	0	0.00
계	132	93	

- FP-2: GD Finlayson[2]은 또한 레티넥스(Retinex)를 이용한 그림자 제거 방법을 제안하였다(이후 3개 문장들이 [2]에 대해 기술하고 있음). 그림자가 제대로 검출이 된다면 광원의 방향과 물체들의 기하학적인 모양을 알 수 있게 된다[3]. **그러나** 그림자인지를 판단할 때, 종 **어려움**을 겪는 경우가 있다*(이 문장이 [3]을 인용한 것인지 [2]를 인용한 것인지 이 논문 저자의 의견인지 명확하지 않음)*.
- FP-3: 이러한 카테고리를 고려하여 Wang [8]은 식 (8)과 같이 더 정교하게 어노테이션 성능을 입증하는 방법을 제안하였다(이 위치에 식(8)이 기술됨). 위 식에서 m은 추정된 어노테이션 태**그의** 개수이고, p, r, w는 각각 "perfect", "good", "bad" 어노테이션의 개수를 의미한다*(규칙 매칭 오류)*.
- FP-4: ScottBarber는 [8,9]에서 소프트웨어의 성능 측정 및 테스트방법론에 대해 자세히 설명하였다. 소프트웨어의 성능관리**와 관련된 연구** 및 **노력들**은 최근 들어 softwareperformance engineering(SPE)이라는 이름으로 하나의 분야를 이루게 되었다. [10]에서 저자는 SPE의 현황과 다양한 이슈들을 잘 정리하였으며, 관련 연구들을 크게 두 가지 접근방법으로 나누었다*("관련된 연구 및 노력들"에 해당하는 기존 연구가 어떤 것인지 명확하지 않음)*.

FP-3 문장은 〈표 2〉의 각 규칙 적용 시 문장의 시작이나 어절의 시작 조건을 검사하지 않아 발생한 단순 오류이다. FP-3을 제외한 나머지 FP 문장들은, 제시되지 않은 나머지 대부분의 FP 문장들을 포함하여, 이 연구에서 설정한 암묵 인용문 판단 기준에 맞지 않아 비인용문으로 태깅되었으나 사람의 경우에도 인용문인지 아닌지 혹은 인용문인 경우 피인용 문헌이 어떤 것인지를 명확히 판별하기 어려운 것들이다. 전술한 내용을 고려하면 〈표 2〉의 규칙은 규칙이 적용되는 경우 인용문을 판별하는 신뢰도가 높다고 볼 수 있다. 그러나 현재의 인용문 규칙은 암묵 인용문의 30%만을 재현하고 있다. 규칙으로 재현되지 못한 FN(False Negative) 문장들(밑줄로 표시됨)의 예는 다음과 같다.

- FN-1: Cascaded Shadow Maps(CSMs) [11]에서는 … 보이게 되었다. 하지만 CSMs은 여러 장의 그림자 맵을 사용하기 때문에 생기는 문제점들을 가지고 있다. 첫째는, 화면에 그려질 물체들을 여러 번 그려야 하기 때문에 렌더링 시간이 증가 되는 것이고, 둘째는 그림 2(a)와 같이 각각의 그림자 맵이 적용되는 경계부분의 그림자가 단절된 것처럼 보이게 된다는 것이다.
- FN-2: Ashwini et al.[10] 모바일 단말기를 위한 전력관리 미들웨어를 제안하고 있다. 단말기에서 구동되는 다양한 애플리케이션에 대해 개별 하드웨어 컴포넌트의 소비전력을 최적화함으로써 전체 시스템 전력 소비를 최소화한다. 미들웨어는 가용한 배터리 잔량과 응용 프로그램의 요구를 기반으로 하드웨어 컴포넌트에 대한 전력설정 값과 최적화된 시스템 전력상태를 계산한다. 제안 모델은 컴포넌트의 동작을 전력소모에 따라 5단계로 구분하고, 시스템 전력상태를 On, User Idle, System Idle,

Suspend의 4가지 상태로 정의한다.

위 FN-1에서 둘째, 셋째 문장은 모두 첫째 문장(명시 인용문)의 암묵 인용문들이다. 둘째 문장의 경우 〈표 2〉의 규칙으로 "하지만", "문제" 등을 매치하여 암묵 인용문으로 분류되었으나 셋째 문장은 현재 규칙으로 매치되는 어구가 발견되지 않아 암묵 인용문으로 분류하지 못하였다. FN-2에서는 명시 인용문인 첫째 문장의 이후 세 개 문장들은 암묵 인용문임에도 불구하고 현재 규칙으로는 인용문으로 판단할 근거를 찾지 못한 것들이다. 이러한 FN 문장들을 올바르게 분류하기 위해서는 향후 대용량 인용문 코퍼스 구축을 통해 새로운 인용문 규칙의 추가 및 기존 규칙의 정제 과정이 필요할 것이다.

〈표 2〉의 규칙과 같은 미리 구축된 단서 어구가 발견되지 않는 암묵 인용문을 인식하기 위해 영어에 대한 기존 연구에서는 암묵 인용문 후보 문장과 명시 인용문 사이에 공유되는 어구(예: 기존 방법의 명칭, 연구자 성, 약어 등 명사구)를 활용했다(Athar & Teufel, 2012; Kaplan, Iida, & Tokunaga, 2009; Qazvinian & Radev, 2010). 그러나 대소문자 구분이 없는 한국어 문장에 전술한 방법을 적용하기 위해서는 개체명 인식이나 키워드/전문용어 인식을 선행할 필요가 있을 것이다. 그렇지 않은 경우 명시 인용문과 단순히 보통 명사(구)를 공유하는 인접 문장을 암묵 인용문으로 인식할 수도 있기 때문이다.

〈표 6〉에서는 기존 영어권 방법들과 이 연구에서 시도한 한국어에 대한 규칙 기반 방법의 성능을 비교하였다. 이는 규칙 기반 암묵 인용문 인식 방법의 한국어에 대한 처리 수준을 가늠하는 간접적 비교 자료가 될 수 있을 것이다. 기존 연구에서 사용된 평가 척도와의 비교를 위해 F-지표 값을 F1, $F_{\beta=0.5}$, $F_{\beta=3}$의 세 가지로 구분하여 제시하였다. 표를 통해 실험 집합 크기의 경우, 현재 연구가 Athar와 Teufel(2012)의 연구를 제외하고는 기존 연구들보다 많은 인용문들을 다루었음을 알 수 있다. F-지표 기준 성능의 경우, 한국어 규칙 기반 방법은 명시와 암묵 인용문을 포함한 평가에서 기존 영어에 대해 보고된 성능(Kaplan, Iida, & Tokunaga, 2009; Namba et al., 2000; Qazvinian & Radev,

〈표 6〉 기존 암묵 인용문 인식 성능과의 비교

Method	실험 집합	Pre.	Rec.	F-measure
Rule-based	한글 인용문 (명시+암묵) 791개	0.9496	0.7863	0.8603 (F1) 0.9118 ($F_{\beta=0.5}$) 0.8001 ($F_{\beta=3}$)
Rule-based	한글 인용문 (암묵) 243개	0.6916	0.3045	0.4229 (F1) 0.5514 ($F_{\beta=0.5}$) 0.3226 ($F_{\beta=3}$)
Nanba, Kando, & Okumura(2000)	영어 인용 영역 (명시+암묵) 50개	0.7630	0.7960	0.7790 (F1)
Kaplan, Iida, & Tokunaga(2009)	영어 인용문 (명시+암묵) 94개	0.6500	0.7400	0.6900 (F1)
Qazvinian & Radev(2010)	영어 인용문 (명시+암묵) 203개	n/a	n/a	0.5400 ($F_{\beta=3}$)
Athar & Teufel(2012)	영어 인용문(암묵) 2,016개	n/a	n/a	0.5100 (F1)

2010)보다 높은 성능을 보였으며, 암묵 인용문만을 대상으로 한 평가에서는 영어권의 성능(Athar & Teufel, 2012)보다 낮았다. 그러나 Athar와 Teufel(2012)의 연구는 대용량 인용문 집합에 대한 SVM 기계학습을 통해 얻어진 학습 기반 성능이므로 현재 연구에 사용된 규칙 기반의 비교사 방법과는 구별될 필요가 있다.

6. 결 론

이 연구는 규칙 기반의 한국어 인용문 자동 인식을 다루었다. 이를 위해 한국어 인용문 인식 말뭉치를 구축하였고, 긍정 및 부정 단서 어구를 활용한 규칙 기반의 인용문 인식 절차를 고안하였다. 실험에서는 인용문 인식을 위한 기존 비교사 기법들과의 비교를 통해 한국어에 대한 규칙 기반 기법의 성능 수준을 제시하였다. 규칙의 사용은 현재 데이터 셋 내에서 명시 및 암묵 인용문을 포함한 전체 한국어 인용문의 79%를 95%의 정확률로 구별해 내었으며, 전체 암묵 인용문의 30%를 약 70%의 정확률로 찾아내어 F1으로 42%, $F_{\beta=0.5}$으로 55%의 성능을 보였다. 이는 전체 인용문에 대한 평가의 경우 영어에 대한 성능보다 높으며, 암묵 인용문에 대한 평가에서는 영어에 대한 성능보다 낮은 수준이다. 또한 인용문 인식 규칙의 개별 성능 분석을 통해 한국어 학술 문헌의 경우 “이 방법”, “방법들”, “연구들”, “이 연구”, “그러나~문제”와 같은 단서 어구들이 명시 인용문에 인접하여 출현하는 전체 암묵 인용 후보 문장의 50%에 출현하고 있으며 이 규칙들의 평균 정확률은 80% 이상임을 알 수 있었다.

이 연구에서 시도된 규칙 기반 방법은 향후 인용문 자동 인식기 개발에서 통제 가능한 지식의 형태로 인용문 인식 규칙들을 제공할 수 있으며, 기계 학습 방법과의 하이브리드 결합도 가능할 것이다. 그러나 현재 연구에 사용된 규칙과 테스트 셋은 전체 인용문을 대표하는 측면에서 부족함이 있다. 향후 한국어 인용문 코퍼스를 대용량으로 구축하고 이를 바탕으로 인용문 인식 방법이 개발될 필요가 있다.

참 고 문 헌

강인수 (2011). 표절 예방을 위한 본문 인용 태깅 지침서. 대전: 한국과학기술정보연구원.

김세종 (2012). KLE 연구실의 언어처리 기반 기술 소개. 포항: 포항공과대학교 지식 및 언어 공학 연구실.

Abu-Jbara, A., & Radev, D. (2011). Coherent citation-based summarization of scientific papers. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL), 500-509.

Abu-Jbara, A., & Radev, D. (2012). Reference scope identification in citing sentences. Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational

Linguistics: Human Language Technologies (HLT-NAACL), 80-90.

Athar, A., & Teufel S. (2012). Detection of implicit citations for sentiment detection. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL), 18-26.

Bradshaw, S. (2003). Reference directed indexing: Redeeming relevance for subject search in citation indexes. Proceedings of the European Conference on Research and Advanced Technology for Digital Libraries (ECDL), 499-510.

Councill, I., Giles, C., & Kan, M. (2008). Parscit: An open-source CRF reference string parsing package. Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC), 661-667. Retrieved from http://www.lrec-conf.org/proceedings/lrec2008/pdf/166_paper.pdf

Kang, I., & Kim, B. (2012). Characteristics of citation scopes: A preliminary study to detect citing sentences. Proceedings of the 2011 International Conference on u- and e-Service, Science and Technology (UNESST), 80-85. http://dx.doi.org/10.1007/978-3-642-35603-2_11

Kaplan, D., Iida, R., & Tokunaga, T. (2009). Automatic extraction of citation contexts for research paper summarization: A coreference-chain based approach. Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries, 88-95.

Nanba, H., Kando, N., & Okumura, M. (2000). Classification of research papers using citation links and citation types: Towards automatic review article generation. Proceedings of the 11th SIG Classification Research Workshop, 117-134.

O'Connor, J. (1982). Citing statements: Computer recognition and use to improve retrieval. Information Processing and Management, 18(3), 125-131.

Qazvinian, V., & Radev, D. (2010). Identifying non-explicit citing sentences for citation-based summarization. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL), 555-564.

Ritchie, A., Robertson, S., & Teufel, S. (2008). Comparing citation contexts for information retrieval. Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM), 213-222.

Singhal, A., Salton, G., & Buckley, C. (1996). Length normalization in degraded text collections. Proceedings of the 5th Annual Symposium on Document Analysis and Information Retrieval (SDAIR), 149-162.

Teufel, S., Siddharthan, A., & Tidhar, D. (2006). Automatic classification of citation function. Proceedings of 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP), 103-110.

• 국문 참고문헌에 대한 영문 표기
(English translation of references written in Korean)

Kang, In-Su (2011). Guidelines to tag in-text citations for plagiarism prevention. Daejeon: Korea Institute of Science and Technology Information.

Kim, Se-Jong (2012). Introduction to KLE laboratory's language technology. Pohang: Knowledge and Language Engineering Laboratory, POSTECH.