Journal of the Korea Academia-Industrial cooperation Society
Vol. 16, No. 8 pp. 5565-5570, 2015

http://dx.doi.org/10.5762/KAIS.2015.16.8.5565
ISSN 1975-4701 / eISSN 2288-4688

어휘 자질 기반 기계 학습을 사용한 한국어 암묵 인용문 인식

강인수[1*]
[1]경성대학교 컴퓨터공학과

Recognition of Korean Implicit Citation Sentences Using Machine Learning with Lexical Features

In-Su Kang[1*]
[1]**Computer Science and Engineering, Kyungsung University**

요 약 암묵인용문 인식은 학술문헌의 본문 텍스트 내에서 명시적 인용표지가 누락된 인용문장을 자동 인식하는 것으로 인용 기반 논문 검색 및 요약의 핵심 기술이다. 기존 암묵인용문 인식의 최신 연구들은 단어 ngram, 단서어구, 명시인용문과의 거리, 기존 연구자의 성, 기존 방법의 명칭 등 다양한 자질을 활용하여 50% 이상 인식 수준을 보고하고 있다. 그러나 대부분의 기존 연구들은 영어에 대해 수행되었으며 한국어의 경우 최근 긍정/부정 단서어구 패턴을 활용한 규칙 기반 시도에서 42% 성능 수준이 보고되어 있어 추가 성능 향상이 요구되는 상황이다. 이 연구에서는 한국어 어휘 자질을 사용하여 한국어 암묵인용문의 기계학습 기반 인식을 시도하였다. 이를 위해 어절, 형태소, 음절 단위에 기반한 다양한 크기의 어휘 ngram 자질들의 인식 성능을 비교 평가하고 한국어 암묵인용문 인식에 적합한 어휘 자질로 형태소 1gram 및 음절 2gram 단위를 결정하였다. 또한 이들 어휘 자질들을 전후 명시인용문들과의 인접성을 표현한 위치 자질들과 결합하여 한국어 암묵인용문 인식 성능을 50% 이상 수준으로 대폭 향상시켰다.

Abstract Implicit citation sentence recognition is to locate citation sentences which lacks explicit citation markers, from articles' full-text. State-of-the-art approaches exploit word ngrams, clue words, researcher's surnames, mentions of previous methods, and distance relative to nearest explicit citation sentences, etc., reaching over 50% performance. However, most previous works have been conducted on English. As for Korean, a rule-based method using positive/negative clue patterns was reported to attain the performance of 42%, requiring further improvement. This study attempted to learn to recognize implicit citation sentences from Korean literatures' full-text using Korean lexical features. Different lexical feature units such as Eojeol, morpheme, and Eumjeol were evaluated to determine proper lexical features for Korean implicit citation sentence recognition. In addition, lexical features were combined with the position features representing backward/forward proximities to explicit citation sentences, improving the performance up to over 50%.

Keywords : Citation Sentence, Korean Lexical Feature, Machine Learning

1. 서론

본문 인용문(In-text citation 혹은 citation sentence)은 학술 문헌의 원문 텍스트에 출현하는 인용문장이다. 본문 인용문은 특정 선행 연구와 후행 연구와의 관련 정보를 후행 연구자의 관점에서 요약 기술하고 있어 학술 문헌에 대한 내용 기반 검색 및 응용에서 그 활용 가치가 점차 부각되고 있다[1,2,3].

본문 인용문 인식(Citing sentence identification)은 학술 문헌의 원문 텍스트 내 본문 인용문(들)을 자동 인식하는 작업이다. 본문 인용문은 “(Smith et al., 2012)”, “[4-6]” 등과 같은 구체적 인용 표지를 포함하고 있는 명

*Corresponding Author : In-Su Kang (Kyungsung Univ.)
Tel: +82-51-663-5147 email: dbaisk@ks.ac.kr
Received May 4, 2015 Revised (1st June 11, 2015, 2nd July 1, 2015)
Accepted August 6, 2015 Published August 31, 2015

시 인용문(Explicit citation)과 그러한 인용 표지가 생략된 암묵 인용문(Implicit citation)으로 나뉜다. 명시 인용문의 경우 인용 표지 표기 패턴의 매칭을 통해 그 인식이 어렵지 않으므로[4], 기존 연구들은 암묵 인용문 인식 문제에 초점을 맞추고 있다. 규칙 기반 방법들[1,3,5]은 암묵 인용문의 주요 판단 기준으로 명시 인용문 전후 n개 각 문장에 대해 기구축된 단서 어구(예: “그러나”, “이러한 방법”, “그 연구” 등) 패턴의 출현 여부 검사를 사용하였다. 최신 기법들[4,6,7,8]은 단서 어구와 함께 ngram 용어 리스트, 명시 인용문까지의 거리, 기존 방법의 명칭 및 인용된 연구자의 성 출현 여부 등의 다양한 자질을 통해 암묵 인용문 인식을 시도하고 있다.

그러나 기존 연구들은 대부분 영어로 작성된 학술 문헌을 대상으로 수행되었으며, 한국어에 대한 본문 인용문 인식 연구는 Kang[5]의 규칙 기반 시도를 제외하고는 거의 찾기 힘들다. 이 연구에서는 한국어 암묵인용문 인식의 SVM(Support Vector Machine) 기계학습을 위한 어휘 자질 표현법을 탐구한다. 영어 암묵인용문 인식의 경우 단어 중심의 ngram 자질이 사용되었다[7,8]. 그러나 형태적 언어유형론 관점에서 한국어는 여러 형태소들이 결합된 형태로 단어(어절)가 구성되는 교착어에 속하므로 영어권의 단어 단위 어휘 자질 추출을 한국어에 적용하는 것은 부적절할 수 있다. 이 연구에서는 한국어의 특성을 고려하여 단어(어절, Eojeol), 형태소(Morpheme), 음절(Eumjeol)의 세 가지 단위에 기반한 어휘 자질 생성을 시도하고 실험을 통해 한국어 암묵인용문 인식에 적합한 어휘 자질 유형을 선택한다. 또한 비어휘 자질로 인접 명시인용문까지의 거리 정보를 학습 자질로 표현하는 기존 방법[8]의 변형을 제안하고 전술한 어휘 자질과의 결합을 통해 암묵인용문 인식 성능을 향상시킨다.

논문의 구성은 다음과 같다. 2장에서는 기존 연구에 대한 기술을 다룬다. 3장에서는 한국어 암묵 인용문 인식을 위한 기계학습 자질 표현법을 기술한다. 4장에서는 실험 계획 및 평가 결과를 제시하고, 5장에서 결론을 맺는다.

2. 기존 연구

암묵인용문 인식을 위한 규칙 기반 연구들은 "they", "drawback", "nevertheless" 등 암묵인용문에 출현할 것으로 예상되는 단서어구들의 존재 유무를 검사하거나[1], 인접 명시인용문 내 명사구의 재출현 혹은 대용표현 출현 여부를 검사하는 방식[3]을 사용하였다. 한국어의 경우 Kang[5]은 “(그러나|그렇지만).+(단점|문제)”와 같은 긍정 단서 패턴과 “(제시|제안)한다”와 같은 부정 단서 패턴을 구축하여 규칙 기반의 암묵인용문 인식을 시도하였다.

최근 데이터 및 학습 기반 연구들은 단서어구들을 다중어휘패턴, 접속어구 등으로 세분하고, 피인용 논문과의 관련성을 표현하기 위해 연구자의 성, 기존 방법의 명칭, 피인용 논문과의 유사도 등을 병행 활용하여 암묵인용문 인식의 성능 향상을 시도하고 있다. Qazvinian[6]은 명시인용문 주위 문장들 간 코사인 유사도, 피인용 논문과의 유사도, 다중어휘패턴(예: "this method", "their work") 등을 활용하여 Markov Random Field 방법론에 기반한 암묵인용문 인식을 시도하였다. Athar[7]는 1gram-3gram까지 단어 ngram을 추출하여 문장의 기본 어휘 자질을 생성하고 다중어휘패턴, 접속어구, 연구자의 성, 기존 방법의 명칭 등 타자질들과 결합하여 SVM 학습을 시도하였다. Abu-Jbara[8]는 문장의 첫 2gram, 3gram을 자질 중 하나로 추출하여 CRF(Conditional Random Field) 학습에 사용하였다. Sondhi[4]는 암묵인용문 인식 문제를 문장이 인용문 상태에서 생성된 것인지 비인용문 상태에 생성된 것인지를 HMM 태깅하는 문제로 고려하고, 인용문 상태에서의 어휘 생성 확률 추정을 위해 피인용논문의 제목, 초록, 명시인용문 텍스트를 활용하였다.

3. 한국어 암묵 인용문 인식을 위한 기계학습 자질 표현

기계학습 기반의 암묵 인용문 인식에서는 학술문헌 원문에 출현한 임의 문장의 암묵 인용문 여부를 학습하기 위해 문장 단위의 자질 표현을 만들 필요가 있다. 이 장에서는 한국어 암묵 인용문 학습예제의 자질 표현을 위한 자질 집합을 정의한다. 다음 네 문장은 이후 설명의 이해를 돕기 위한 예제 문장들로 선행 명시인용문(S1)을 암묵적으로 인용하는 세 문장 S2, S3, S4을 보인 것이다.

S1: Smith(2013)는 온톨로지 생성기법 OntoGen을 ...
S2: DB로부터 온톨로지의 자동 생성을 Smith는 ...
S3: OntoGen은 최초의 온톨로지 생성 시도로 평가 ...
S4: 그러나 이 기법은 전문가가 작성한 규칙에 ...

3.1 POSITION 자질 (위치)

POSITION 자질은 현재 문장에 대해 인접 명시인용문까지의 거리를 표현한 것으로, 암묵 인용문이 명시 인용문의 전후 인접된 문장 위치에 출현하는 경향이 있음을 반영한 자질이다. 기존 영어권 연구에서는 POSITION자질을 현재 문장에서 가장 가까운 명시인용문까지의 상대 문장 거리에 해당하는 단일 정수값으로 표현하였다[8]. 예를 들어 문장 S3의 경우 이 자질의 값은 -2가 된다. 이후 설명에서 이 자질은 *Pi*로 표기한다.

이 연구에서는 현재 문장으로부터 가장 가까운 이전(backward)과 이후(forward) 명시인용문까지의 문장 거리를 두 개 자질 *Pb* (Position backward), *Pf* (Position forward)로 분리하고, 아래 식 (1)을 사용하여 임계치 거리 이내에 출현한 명시인용문과의 인접 정도를 반영한 값을 부여한다. 식 (1)에서 d_f , d_b는 현재 문장에서 가장 가까운 이후, 이전 명시인용문까지의 문장 거리로 정의하고 *Pi*와 달리 절대값을 사용한다. K_f , K_b는 각각 이후, 이전 임계치 거리 상수이다. 예를 들어 임계치 K_f=5일 때, d_f가 2 혹은 6인 경우 *Pf* 값은 각각 0.5, 0이 부여된다.

$$Pf = \begin{cases} \frac{1}{d_f} & if\ 0 < d_f \le K_f \\ 0 & othewise \end{cases}$$
$$Pb = \begin{cases} \frac{1}{d_b} & if\ 0 < d_b \le K_b \\ 0 & othewise \end{cases} \quad (1)$$

3.2 SURNAME 자질 (기존 연구자의 성)

SURNAME 자질[7]은 이전 연구를 인용할 때, 문장 S2에서와 같이, 기존 연구자의 성이 사용되는 경향을 반영한 것이다. 이 연구에서는 동일 논문 내 명시인용문들에 출현한 영문 성 문자열 출현 여부를 SURNAME 자질값으로 사용한다.

3.3 TERM 자질 (기존 연구의 주요 용어)

TERM 자질[7]은, 문장 S3에서의 OntoGen과 같이, 선행 연구의 주요 방법/도구/알고리즘 등의 용어가 후행 연구의 인용문에 사용되는 경향을 활용한 것이다. 이 연구에서는 동일 논문 내 명시인용문에 출현한 영어 약어 문자열 출현 여부를 TERM 자질값으로 사용한다.

3.4 RULE 자질 (단서 어구 패턴)

단서 어구 패턴 자질은 "이러한 방법", "전술한 연구" 등과 같이 암묵인용문 판단의 단서가 될 수 있는 어구들로 구성된다. 이러한 단서어구들은 기존 규칙 기반 접근법에서 암묵인용문 결정을 위해 사용되는 패턴 어구들에 해당한다. 이 연구에서는 Kang[5]에서 소개된 한국어 암묵인용문 단일 규칙 397개를 RULE 자질들로 사용하였다.

3.5 LEXEME 자질 (어휘)

어휘 자질은 현재 문장의 출현 어휘 집합으로부터 추출된 어휘 기반 자질을 의미한다. 이 연구에서는 한국어의 형태적 특성을 반영하여 다음과 같이 단어(어절), 형태소, 음절의 세 가지 어휘 자질을 정의한다.

(1) 어절 ngram 자질: 어절 ngram 자질 유형은 기존 영어권 연구[7,8]에서 사용된 단어 자질에 해당한다. 문장 S4에 대해 어절 ngram 자질을 생성하면 다음과 같다.

- **w1gram**: 그러나, 이, 기법은, 전문가가, 작성한, 규칙에, ...
- **w2gram**: 그러나_이, 이_기법은, 기법은_전문가가, 전문가가_작성한, 작성한_규칙에, ...

위 예에서 알 수 있듯이 한국어의 어절은 곡용/활용의 다양한 변이형을 갖는 조사/어미를 포함할 수 있어, 어절 ngram 방식은 기계학습 자질의 과다 생성을 가져올 수 있다. 암묵인용문 인식 관점에서 위 예의 어절 ngram은 단서어구 "그러나"를 자질로 추출하였으나 또 다른 주요 단서어구 "이_기법"의 경우 "이_기법은"과 같이 조사가 결합된 구체적인 형태로 추출하고 있다. 이는 어절 ngram 방식을 통해 암묵인용문 인식의 정확률은 높일 수 있으나 제한된 규모의 학습데이터를 감안하면 재현율의 저하를 피하기 어려움을 알 수 있다.

(2) 형태소 ngram 자질: 형태소 ngram 자질 유형은 어절에 포함된 모든 형태소를 분리한 다음 형태소 단위

의 ngram 자질을 생성한다. 문장 S4에 대해 생성된 형태소 ngram 자질의 예는 다음과 같다.

- **m1gram**: 그러나, 이, 기법, 은, 전문가, 가, 작성, 하, ㄴ, 규칙, 에, ...
- **m2gram**: 그러나_이, 이_기법, 기법_은, 은_전문가, 전문가_가, 가_작성, 작성_하, 하_ㄴ, ㄴ_규칙, 규칙_에, ...

이 방식은 어절 ngram과 비교하여 최소 의미 단위인 형태소 중심의 자질 표현을 생성하므로 개별 형태소 혹은 결합 형태소 단위의 중요도를 기계학습하는 것이 가능하다. 위 예는 형태소 ngram을 통해 "그러나", "이_기법"과 같은 단서어구들이 잘 추출될 수 있음을 보여준다.

(3) 음절 ngram 자질: 음절 ngram 자질 유형은 음절 단위 ngram 자질을 생성하며, 문장 S4에 대해 생성된 예는 다음과 같다.

- **e1gram**: 그, 러, 나, 이, 기, 법, 은, 전, 문, 가, 가, 작, 성, 한, 규, 칙, ...
- **e2gram**: 그러, 러나, 이, 기법, 법은, 전문, 문가, 가가, 작성, 성한, 규칙, ...

위 예에서처럼 음절 ngram은 한 어절 내의 각 음절 위치에서 연속된 n개 음절들을 추출한다. 이 방식은 "그러나"와 같은 단일 형태소를 "그러", "러나"의 두 개 자질로 분리하므로, 단일 형태소 자질 하나를 그 형태소의 음절 길이에 비례하는 수의 분산된 다중 자질들로 확장하여 표현하게 된다.

4. 실험

4.1 실험 계획

암묵인용문 성능 평가를 위해 Kang[5]에서 사용된 한국어 본문 인용문 평가세트를 사용하였다. 이 평가세트는 정보과학회논문지 2012년 게재 논문 35편의 원문 텍스트 내 각 문장에 대해 인용문 여부 태그를 수작업 부착해 둔 것으로, 총 6,075 문장에 출현한 791개 인용문(명시인용문 548개, 암묵인용문 243개)들로 구성되어 있다.

인용문 인식을 위한 기계학습 도구로 LIBSVM[9]을 사용하였다. 인식 성능은 위 평가세트에 대한 10겹 교차검증(10-fold cross-validation)을 통해 얻어진 재현율, 정확률, F1으로 제시하였다. 재현율(Recall)은 전체 정답 인용문들 중 시스템이 올바르게 판단한 인용문들의 비율이며, 정확률(Precision)은 시스템이 인용문으로 판단한 문장들 중 정답 인용문들의 비율로 정의한다. F1은 정확률과 재현율의 조화평균으로 정의된다. 특별한 언급이 없으면 실험 결과로 제시되는 인용문 인식 성능은 암묵 인용문에 대한 것이다.

3장에 소개된 명시인용문 상대 위치 자질 값 결정을 위한 파라미터 K_f , K_b는 각각 5, 5로 설정하였다. 문장의 형태소 자질 추출을 위해 문장에 대한 형태소분석과 품사 태깅을 수행하였고, 이를 위해 포항공대 지식 및 언어공학연구실의 분석기를 사용하였다.

4.2 실험 결과

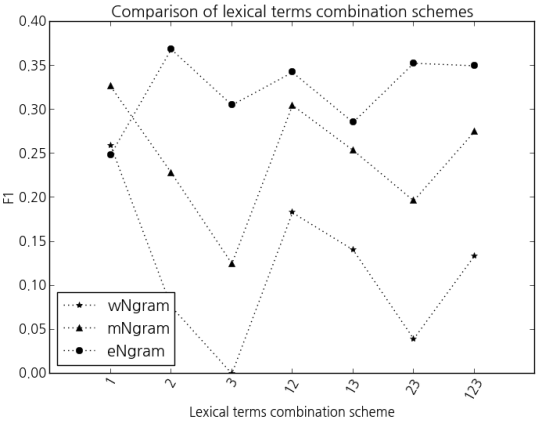

Fig. 1. Performance based on lexical features and their ngram combination schemes (wNgram: word (Eojeol) ngram, mNgram: Morpheme ngram, eNgram: Eumjeol ngram)

Fig. 1은 어절, 형태소, 음절의 세 가지 각 어휘 자질에 기반한 암묵인용문 인식 성능을 해당 자질 유형의 1-3gram 까지 각 결합 조합에 대해 제시한 것이다. 예를 들어 Fig. 1의 x축 값 2는 2gram 자질의 사용을 의미하며, 23은 2gram과 3gram 자질들의 병합 사용을 의미한다. 어절, 형태소, 음절 자질에 대해 각각 1gram (w1gram), 1gram (m1gram), 2gram (e2gram)이 동일 자

질 유형의 일곱 가지 ngram 조합들 중 가장 좋은 성능을 보였다. Table 1은 각 어휘 자질 유형의 최적 ngram 자질과 그 성능을 보인 것이다.

Table 1. Performance of the best lexical feature types

Best lexical feature types	Pre	Rec	F1
w1gram	0.4583	0.1811	0.2596
m1gram	0.3578	0.3004	0.3266
e2gram	0.4333	0.3210	0.3688

Fig. 2는 어절, 형태소, 음절의 각 최적 어휘 자질(*L*: Lexeme)과 다른 자질들을 결합한 경우의 암묵인용문 인식 성능의 변화를 도식화한 것이다. 이를 위해 SURNAME(*S*), TERM(*T*), POSITION(*P*), RULE(*R*) 자질들과의 결합을 시도하였다. POSITION 자질은 다시 *Pi*와 *Pb*, *Pf* 자질로 나뉜다.

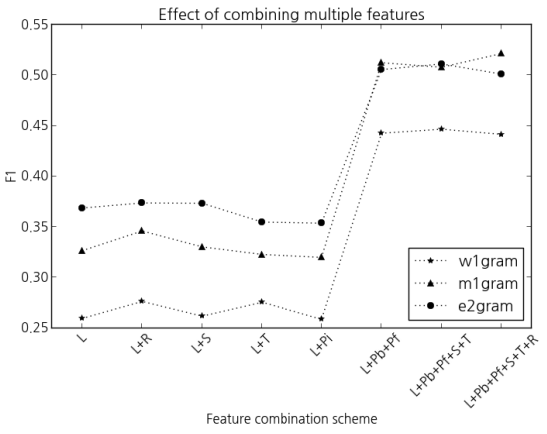

Fig. 2. Effect of combining multiple features (*L*: Lexeme, *R*: Rule, *S*: SURNAME, *T*: TERM, *Pi*: POSITION integer, *Pb*: POSITION backward, *Pf*: POSITION forward)

실험 결과 *Pb*, *Pf* 자질을 제외하고는 결합을 통한 성능 변화가 미미했다. 기존 POSITION 자질의 변형으로 제안된 *Pb*, *Pf* 자질은 세 가지 모든 유형의 어휘 자질과의 결합에서 두드러진 성능 향상을 보였다. 이는, 인접 명시인용문들의 상대 위치를 자질화함에 있어, 명시인용문의 출현 위치를 이전과 이후로 분리하고 그들의 근접성을 차별화하여 표현한 이 논문의 시도가 암묵인용문 인식에 효과적이었음을 의미한다. 그러나 어휘 자질 없이 *Pb*, *Pf* 자질만을 사용한 경우에는 정답 암묵인용문을 하나도 인식하지 못하였다. 또한 *S*, *T*, *Pi* 자질들을 각각 단독 사용한 실험에서도 정답 암묵인용문을 하나도 인식하지 못하여, 이들 자질들은 어휘 자질과 결합 사용할 필요가 있음을 확인하였다.

RULE 자질의 경우 단독 사용에서 0.4933(정확률), 0.1523(재현율), 0.2327(F1)의 성능을 보였고, 형태소 자질과의 결합을 통해 최종 암묵인용문 인식 성능 향상에 소폭 기여하였다. RULE 자질의 결합을 통한 효과가 크지 않은 이유는 어절, 형태소, 음절과 같은 어휘 자질들이 RULE 자질의 내용을 대부분 포함하고 있기 때문인 것으로 판단된다.

전체적으로 어절 자질보다 형태소, 음절 자질들이 더 우수했으며, 음절 자질과 형태소 자질은 위치 자질(*Pb*, *Pf*)과 결합된 이후에는 두 자질 간에 암묵인용문 인식 능력에 두드러진 차이가 발견되지 않았다. 한편 형태소 자질은 이 연구에서 시도된 위치 자질(*Pb*, *Pf*)과의 결합이 없는 경우 음절 자질보다 3-5% 정도 낮은 성능을 보였는데(Fig. 2의 x축 값 *L*, *L+R*, *L+S*, *L+T*, *L+Pi*에서 m1gram과 e2gram 성능 비교), 이는 형태소분석 및 품사 태깅 오류에 기인하는 것으로 판단된다. 그러나 형태소 자질과 달리 음절 자질은 의미 표현 단위의 불일치로 인해 단서어구 자질(*R*)과 결합될 때 성능 향상이 미미하거나 성능이 다소 저하되는 단점이 있었다(Fig. 2에서 e2gram의 *L*, *L+R* 간 및 *L+Pb+Pf+S+T*, *L+Pb+Pf+S+T+R* 간 성능 차이 비교).

Table 2. Comparison to other systems for the implicit citation detection

Systems		Pre	Rec	F1 or F3
Korean	Rule-based [5]	0.6916	0.3045	0.4229 (F1) 0.3226 (F3)
	Current system	0.5808	0.4733	0.5215 (F1) 0.4822 (F3)
English	Qazvinian & Radev [6]	n/a	n/a	0.5400 (F3)
	Athar & Teufel [7]	n/a	n/a	0.5130 (F1)
	Sondhi [4]	n/a	n/a	0.4950 (F1) 0.4640 (F3)

Table 2에서는 이 연구의 암묵인용문 인식 성능을 한국어 및 영어권의 기존 연구들과 비교하여 제시하였다. 한국어에 대해 시도된 기존 규칙 기반 방법[5]과 비교할 때, 이 연구의 기계학습 기반 방법은 정확률을 희생하면서 재현율을 높여 한국어 암묵인용문 인식 성능의 대폭 향상을 가져왔다. 또한 본 논문에서 시도된 기계학습 방

법의 성능은 평가지표 상 영어권 기존 연구들과 견줄만한 수준이다.

5. 결론

이 연구는 한국어 학술문헌 본문 내 암묵인용문의 기계학습 기반 인식을 위한 학습자질들을 탐구하였다. 어휘 자질로 어절, 형태소, 음절 자질을 비교 평가하였고, 형태소 및 음절 단위가 한국어 암묵인용문 인식에 효과적인 어휘 자질임을 실험적으로 제시하였다. 비어휘 자질로는, 명시인용문과의 인접성을 전후 방향의 두 개 자질로 분리하여 고안한 변형된 위치 자질(*Pb*, *Pf*)이 어휘 자질과 결합될 때 큰 폭의 성능 향상을 가능케 했다. 그러나 현재의 50% 대 인식 성능은 실용적 수준에 활용되기에 많이 부족하므로, 향후 학습데이터의 규모를 늘리고 새로운 자질 유형 개발과 함께 최신의 다양한 학습모델의 적용이 시도되어야 한다.

References

[1] H. Nanba, N. Kando, M. Okumura, "Classification of research papers using citation links and citation types: Towards automatic review article generation", Proc. of the 11th ASIS SIG/CR Classification Research Workshop, pp.117-134, 2000.

[2] A. Ritchie, S. Robertson, S. Teufel, "Comparing citation contexts for information retrieval", Proc. of the 17th ACM Conference on Information and Knowledge Management, pp.213-222, 2008.
DOI: http://dx.doi.org/10.1145/1458082.1458113

[3] D. Kaplan, R. Iida, T. Tokunaga, "Automatic extraction of citation contexts for research paper summarization: a coreference-chain based approach", Proc. of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries, pp.88-95, 2009.
DOI: http://dx.doi.org/10.3115/1699750.1699764

[4] P. Sondhi, C. Zhai, "A constrained hidden Markov model approach for non-explicit citation context extraction", Proc. of the 2014 SIAM International Conference on Data Mining, pp.361-369, 2014.
DOI: http://dx.doi.org/10.1137/1.9781611973440.41

[5] I. Kang, "A rule-based approach to identifying citation text from Korean academic literature", Journal of the Korean Society for information Management, 29(4), pp.43-60, 2012.
DOI: http://dx.doi.org/10.3743/kosim.2012.29.4.043

[6] V. Qazvinian, D. R. Radev, "Identifying non-explicit citing sentences for citation-based summarization", Proc. of the 48th Annual Meeting of the Association for Computational Linguistics, pp.555-564, 2010.

[7] A. Athar, S. Teufel, "Detection of implicit citations for sentiment detection", Proc. of ACL-12 Workshop on Discovering Structure in Scholarly Discourse, pp.18-26, 2012.

[8] A. Abu-Jbara, J. Ezra, D. R. Radev, "Purpose and polarity of citation: towards NLP-based bibliometrics", Proc. of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp.596-606, 2013.

[9] C-C Chang, C-J Lin, "LIBSVM : a library for support vector machines", ACM Transactions on Intelligent Systems and Technology, 2(3):27:1-27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm
DOI: http://dx.doi.org/10.1145/1961189.1961199

강 인 수(In-Su Kang) **[정회원]**

- 1999년 2월 : 포항공과대학교 컴퓨터공학 (공학석사)
- 2006년 2월 : 포항공과대학교 컴퓨터공학 (공학박사)
- 2006년 3월 ~ 2008년 2월 : 한국과학기술정보연구원
- 2008년 3월 ~ 현재 : 경성대학교

<관심분야>
자연어처리, 정보검색, 시맨틱웹