Journal of The Korea Society of Computer and Information
Vol. 17 No. 9, September 2012
www.ksci.re.kr
http://dx.doi.org/10.9708/jksci/2012.17.9.029

문화유산정보 말뭉치 구축을 위한 개체명 및 이벤트 부착 도구

최 지 예*, 김 명 근*, 박 소 영**

Named Entity and Event Annotation Tool for Cultural Heritage Information Corpus Construction

Ji-Ye Choi*, Myung-Keun Kim*, So-Young Park**

요 약

본 논문에서는 문화유산정보 말뭉치 구축을 위한 개체명 및 이벤트 부착 도구를 제안한다. 제안하는 도구를 이용하여 말뭉치 구축자는 문화유산정보 관리에 유용한 시간, 장소, 인물, 사건을 중심으로 개체명과 이벤트를 부착할 수 있다. 이 때, 개체명과 이벤트 부착이 용이하도록, 제안하는 도구에서 줄번호나 어절번호와 같은 개체명이나 이벤트의 위치정보를 자동으로 부착하며, 구축된 개체명이나 이벤트 중에서 하나를 선택하면 해당 문자열을 원문에서 진한 이탤릭체로 표시하여 올바르게 부착되었는지 쉽게 확인할 수 있다. 그리고, 제안하는 도구는 말뭉치 구축자의 수작업을 줄이기 위해서 개체명 자동인식 패턴을 활용한다. 학습말뭉치가 거의 없다는 점을 고려하여 단순한 규칙 패턴을 학습한다. 또한, 오류 전파를 차단하기 위해서, 제안하는 개체명 자동인식 패턴은 개체명 부착 말뭉치에서 추가적인 분석처리 없이 바로 추출한다. 실험결과 제안하는 개체명 및 이벤트 부착 도구는 말뭉치 구축자의 수작업량을 절반이상 줄여주었다.

▸ Keywords : 개체명 부착 말뭉치, 이벤트 부착 말뭉치, 말뭉치 구축, 개체명 자동 인식

Abstract

In this paper, we propose a named entity and event annotation tool for cultural heritage information corpus construction. Focusing on time, location, person, and event suitable for cultural heritage information management, the annotator writes the named entities and events with the proposed tool. In order to easily annotate the named entities and the events, the proposed tool

• 제1저자 : 최지예 • 교신저자 : 박소영
• 투고일 : 2012. 07. 09, 심사일 : 2012. 07. 31, 게재확정일 : 2012. 08. 09.
* 상명대학교 디지털미디어학부(Dept. of Digital Media, SangMyung University)
* 상명대학교 게임모바일콘텐츠학과(Dept. of Game & Mobile Contents, SangMyung University)
※ 이 논문은 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2011-씨앗형-B00036).

automatically annotates the location information such as the line number or the word number, and shows the corresponding string, formatted as both bold and italic, in the raw text. For the purpose of reducing the costs of the manual annotation, the proposed tool utilizes the patterns to automatically recognize the named entities. Considering the very little training corpus, the proposed tool extracts simple rule patterns. To avoid error propagation, the proposed patterns are extracted from the raw text without any additional process. Experimental results show that the proposed tool reduces more than half of the manual annotation costs.

▸ Keywords : Named Entity Annotated Corpus, Event Annotated Corpus, Corpus Construction, Named Entity Recognition

I. 서 론

문화유산정보는 경제적 가치를 창출하는 콘텐츠로 활용될 수 있으므로, 이에 대한 체계적인 수집, 가공, 유통의 필요성이 증가하고 있다[1]. 따라서, 오랜 시간동안 많은 예산과 인력을 투입하여 대량의 문화유산에 관한 정보를 아카이브의 형태로 구축하였다. 그러나, 이러한 결과물을 올바르게 활용하여 고부가 가치를 창출하기 위해서는 콘텐츠 개발과 산업적 활용을 위한 노력을 진행할 필요가 있다[2]. 또한, 현재 한국 문화유산정보를 표현하기 위해 사용하는 국립중앙박물관의 '한국유물분류표준'과 국가기록원의 '전자기록물 보존 메타데이터'는 메타데이터의 스키마 부족 때문에 표현할 수 있는 정보량과 시소러스가 부족하다. 이를 개선하기 위해서는 2006년 국제표준으로 인증(ISO 21127:2006)을 받은 국제박물관협회의 문서화를 위한 국제 위원회(CIDOC/ICOM: The International Committee for Documentation of the International Council of Museums)에서 개발한 개념준거모델(CRM:Conceptual Reference Model)을 도입해야 한다[3].

한편, 전문 연구 분야의 텍스트 문서의 활용가치를 증대하기 위해서, 최근 정보 추출을 위한 코퍼스 구축 사업이 활발하게 진행되고 있다[4,5]. 최근 의학 및 생물학 분야의 연구에서 새로운 유전자나 의약품 정보가 매우 늘어나고 있다. 이러한 연구결과의 효용성을 높이기 위해서는 연구 문서의 텍스트 분석을 통해 유전자의 상호작용 등의 정보를 추출하여 활용할 필요가 있다. 이 때, 전문 용어 사전을 구축하여 관리하는 경우 인적 시간적 비용이 많이 소요된다. 그러므로, 개체간의 상호작용정보가 부착된 말뭉치에서 필요한 정보를 추출하여 활용하는 기계학습기법을 이용하여 텍스트 문서에서 자동으로 정보를 추출하는 연구가 활발히 진행되고 있다[5,6].

따라서, 본 논문에서는 문화유산정보 관련 문서의 활용가치를 높이기 위해서 문화유산정보 관련 문서에 개체명 및 이벤트를 부착하는 도구를 제안한다. 제안하는 도구는 말뭉치 구축자의 수작업을 줄여줄 수 있도록 구축된 소량의 말뭉치를 분석하여 개체명 자동인식 패턴을 추출하고 활용한다. 앞으로, 2장에서는 기존 말뭉치 구축도구 및 개체명 자동 인식방법에 대해 살펴본다. 3장에서는 제안하는 말뭉치 구축도구와 패턴 기반 개체명 자동인식 및 적용 방법을 설명하고, 4장에서는 제안하는 방법을 실험을 통해 평가한다. 마지막으로 5장에서는 제안하는 방법에 대한 결론을 내린다.

II. 관련 연구

최근 말뭉치에서 필요한 정보를 자동으로 추출하는 기계학습기법에 대한 연구가 많이 진행되고 있다. 그러나, 여러 명의 말뭉치 구축자가 대량의 말뭉치를 일관성있고 정확하게 구축하는 것은 매우 어렵다. 이를 고려하여, 말뭉치 구축자의 부담을 줄일 수 있는 다양한 말뭉치 구축도구가 연구되고 있다. 특히, 말뭉치 구축자의 수작업을 줄이기 위해서, 말뭉치 구축도구가 전문가가 작성한 정교한 규칙을 바탕으로 작업의 일부를 자동으로 처리하는 방법이 있다[7,8]. 이러한 방법은 정교한 규칙에 의존도가 높고 규칙의 수정이 쉽지 않다. 따라서, 규칙의 추가나 수정이 용이하도록, 이미 구축된 말뭉치에서 자동으로 규칙을 추출하여 활용하는 방법이 제안되었다[9,10]. 이 방법은 전문가가 작성한 규칙에 비해 정교함은 떨어지지만, 말뭉치의 크기가 증가함에 따라 규칙의 신뢰도가 개선되고 더 많은 양의 수작업을 줄여줄 수 있다.

한편, 기존 개체명 자동 인식에 관한 연구는 MUC(Message Understanding Conference)를 중심으로 이루어졌으며, 사람, 단체, 지역명과 같은 고유의 이름들과 시간, 날짜, 액

수, 퍼센트 표현과 같은 숫자 표현을 자동으로 인식한다[11]. 개체명 자동인식방법은 크게 지도학습기법을 사용한 접근방법과 비지도학습기법을 사용한 접근방법으로 나누어 볼 수 있다.

최대 엔트로피 모델을 이용한 개체명 인식 방법[12], 지지벡터기계를 이용한 개체명 인식 방법[13]은 개체명 인식 문제의 복잡도를 줄이기 위해 개체명 경계 인식 단계와 개체명 의미범주 분류단계를 구분한다. 또한, 풍부한 문맥정보를 활용할 수 있도록 형태소 분석기의 분석결과를 활용한다. 이러한 접근방법은 말뭉치에서 개체명 자동인식에 필요한 정보를 추출하여 사용하므로, 보다 정확하게 인식하기 위해서는 대량의 말뭉치 확보가 매우 중요하다. 그리고, 형태소 분석 결과의 오류나 개체명 경계 인식결과의 오류가 개체명 의미범주 분류에 그대로 전파될 수 있다.

비지도 학습기법을 이용한 방법에는 사전 및 패턴을 이용한 제목 개체명 인식 방법[11]이 있으며, 이 방법은 문맥 패턴 구축 단계와 사전 확장 단계로 구성된다. 제목 개체명을 부착한 말뭉치가 거의 없다는 점을 고려하여, 문맥 패턴 구축 단계에서는 미리 구축한 사전에 등록된 제목 개체명이 문장에서 어떤 문맥과 함께 나타나는지 패턴으로 추출하고 추출된 후보 중에서 신뢰도가 임계치 이상 높은 것을 패턴 저장소에 등록한다. 그리고 사전 확장 단계는 정해진 패턴을 바탕으로 제목 개체명 후보를 찾고, 이를 사전에 등록한다. 이러한 과정을 반복하여 패턴과 사전을 증가시킨다. 영문 대문자 중심으로 제목을 추출하고 자주 공기하는 문맥만을 선별하여 사용하므로, 비교적 정확률은 높은 반면 재현율을 상대적으로 낮게 나타났다.

말뭉치 구축자의 수작업을 줄이기 위해, 본 논문에서 제안하는 말뭉치 구축도구는 개체명 자동인식 패턴을 활용한다. 충분한 학습말뭉치가 없다는 점을 고려하여 통계기반 접근법이나 지도학습기법보다는 단순한 규칙패턴을 중심으로 개체명을 인식한다. 또한, 여러 리소스를 활용하거나 단계를 세분화하는 경우 이전 단계의 오류가 다음 단계로 전파할 수 있다는 점을 고려하여, 제안하는 개체명 자동인식 패턴은 말뭉치의 추가적인 분석처리 없이 원문 텍스트의 문자열에서 바로 추출한다.

III. 개체명 및 이벤트 부착 도구

문화유산정보 말뭉치 구축을 위한 개체명 및 이벤트 부착 도구는 [그림1]에 제시된 바와 같이 개체명 및 이벤트 부착 단계와 개체명 자동인식 패턴 추출단계로 구성된다. 먼저, 개체명 및 이벤트 부착 단계에서 말뭉치 구축자는 주어진 원문에서 개체명과 이벤트를 찾아서 부착한다. 말뭉치 구축자의 수작업을 줄이기 위해서, 개체명 자동인식 패턴 추출 단계에서는 그동안 구축한 개체명 부착 말뭉치에서 개체명을 자동으로 인식할 수 있는 패턴을 추출하여 제공한다. 이장의 1절에서는 개체명 및 이벤트 부착 단계에 대해 설명하고, 2절에서는 개체명 자동인식 패턴 추출 단계에 대해 설명한다. 마지막으로 3절에서는 추출된 개체명 자동인식 패턴을 어떻게 적용하는지 설명한다.

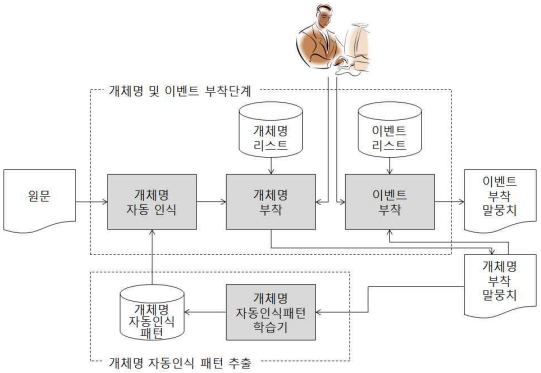

그림 1 시스템구성도
Fig. 1. System Architecture

1. 개체명 및 이벤트 부착

[그림1]에서 살펴본 바와 같이, 말뭉치 구축자는 [그림2]와 같은 말뭉치 구축도구를 이용하여 주어진 원문에 대해 개체명과 이벤트를 부착할 수 있다. 주변 문맥을 고려하여 정확하게 개체명 및 이벤트를 찾아서 부착할 수 있도록, 제안하는 말뭉치 구축도구는 [그림2]의 좌측처럼 주어진 원문을 제시한다.

먼저, 말뭉치 구축자가 좌측에 제시된 원문에 나타난 개체명을 드래그하고 추가버튼을 눌러 개체명의 종류를 선택하면,

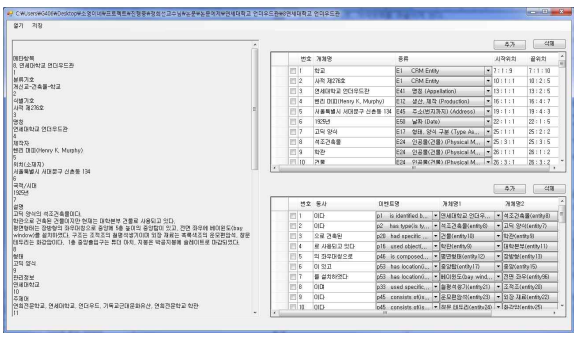

그림 2 말뭉치 구축도구
Fig. 2. Corpus Annotation Tool

표 1. 개체명 종류 일부
Table 1. Some Named Entity Categories

E10 소유권의 이전(각종 권리) (Transfer of Custody)
E11 변형 (Modification)
E12 생산, 제작 (Production)
E13 (속성, 특성, 가치, 의의 등의) 부여 (Attribute Assignment)
E14 상태, 평가 (Condition Assessment)
E15 구분, 규정 (Identifier Assignment)

[그림2]의 우측 상단과 같이 개체명, 개체명 종류, 원문에서의 시작위치 및 끝 위치를 저장한다. 이와 같이, 개체명을 드래그하면 말뭉치 구축도구에서 자동으로 개체명의 시작위치와 끝위치를 파악하므로, 말뭉치 구축자는 개체명과 그 종류를 부착하는데 집중할 수 있다.

이렇게 구축한 개체명 정보를 고려하여 말뭉치 구축자는 이벤트를 분석하여 부착한다. 즉, 제시된 원문에 나타난 이벤트를 유발하는 동사를 추가하고 이벤트 종류를 선택하면, [그림2]의 우측하단과 같이 그 이벤트와 관련된 개체명을 선택할 수 있다. 각 이벤트를 선택하면 동사와 개체명의 위치가 굵은 이탤릭체로 표시되므로, 원문에서 각 이벤트와 개체명들이 서로의 주변에 나타났는지 확인할 수 있다.

이때, 어떤 종류의 개체명과 이벤트를 중심으로 말뭉치를 구축할지를 명확하게 하기 위해서, [그림1]과 같이 개체명 및 이벤트 리스트를 바탕으로 말뭉치를 부착한다. 이 리스트는 CIDOC의 개념준거모델을 참고하여 문화유산정보관리에 적합하도록 시간, 장소, 인물, 사건을 중심으로 구성한다[14].

개체명 리스트는 총 105개로 구성되어 있으며, [표1]은 그중 일부를 보여준다. 예를 들어, 원문에 매입, 양도 등의 단어가 나타나면 개체명의 종류를 E10 소유권의 이전(각종 권리)으로 부착하고, 개조, 재건 등의 단어가 나타나면 E11 변형 개체명으로 부착하며, 건축, 준공, 신축 등의 단어는 E12 생산, 제작 개체명으로 부착한다.

표 2. 이벤트 종류 일부
Table 2. Some Event Categories

p1	is identified by(identifies)
p2	has type(is type of)
p3	has note(has type: type)
p4	has time span(is time span of)
p5	consists of(forms part of)
p6	falls within(contains)

이벤트 리스트는 총 106개로 구성되어 있으며, [표2]는 그중 일부를 보여준다. 이벤트는 개체명과 개체명간의 이벤트를 중심으로 표현되므로, 각 이벤트는 개체명을 하위논항으로 요구한다. 예를 들어, “언더우드관은 고딕 양식의 석조건축물이다”라는 문장에서 개체명 ‘언더우드관’:E41 명칭와 ‘석조건축물’:E24 인공물(건물)를 찾고, 이를 바탕으로 이벤트 ‘석조건축물’:E24 인공물(건물)-is identified by-‘언더우드관’:E41 명칭을 파악할 수 있다.

표 3. 원문 예제 일부
Table 3. Raw Text Examples

언더우드관은 고딕 양식의 석조건축물이다. 학관으로 건축된 건물이지만 현재는 대학본부 건물로 사용되고 있다. 평면형태는 장방형의 좌우대칭으로 중앙에 5층 높이의 중앙탑이 있고, 전면 좌우에 베이윈도(bay window)를 설치하였다. 구조는 조적조의 평석쌓기이며 외장 재료는 흑록색조의 운모편암석, 창문 테두리는 화강암이다. 건물 중앙출입구는 튜더 아치형으로 돌 구조 현관이 잘 보존되어 있다. 지붕은 박공지붕에 슬레이트로 마감되었다.

한편, [표3]과 같은 원문 텍스트가 주어졌을 때, 제안하는 말뭉치 구축도구를 사용하여 [표4]와 같은 텍스트 파일 형태의 개체명 부착 말뭉치를 생성할 수 있다. 먼저, 개체명 *번호*는 원문텍스트에서 개체명을 찾은 순서를 의미하고, *개체명*은 원문텍스트에 나타난 단어 문자열을 가리키며, *종류*는 그 문자열의 개체명 종류를 나타낸다. *시작위치*와 *끝위치*는 줄번호:어절번호:음절번호의 형태를 따르는데, 해당단어가 원문 텍스트에서 몇 번째줄 몇 번째 어절의 몇 번째 음절에서 시작해서 끝나는지를 나타낸다[4]. 즉, [표5]에서 ‘언더우드관’은 1번째로 찾은 개체명으로 그 종류가 *E41 명칭*이고, 1번째줄 1번째 어절의 1번째 음절에서 시작해서 1번째줄 1번째 어절 5번째 음절로 끝난다는 것을 의미한다.

표 4. 개체명 부착 말뭉치
Table 4. Named Entity Annotated Corpus

번호	개체명	종류	시작위치	끝위치
1	언더우드관	E41 명칭(Appellation)	1:1:1	1:1:5
2	고딕 양식	E17 형태,양식 구분(Type Assignment)	1:2:1	1:3:2
3	석조건축물	E24 인공물(건물)(Physical Man-Made Stuff)	1:4:1	1:4:5
4	학관	E24 인공물(건물)(Physical Man-Made Stuff)	2:1:1	2:1:2
5	건물	E24 인공물(건물)(Physical Man-Made Stuff)	2:3:1	2:3:2
6	대학본부	E15 구분, 규정(Identifier Assignment)	2:5:1	2:5:4

이렇게 구축한 개체명 부착 말뭉치가 주어지면, 말뭉치 구축자는 이벤트 리스트를 바탕으로 [표5]와 같은 이벤트 부착 말뭉치를 생성하며, 그 형식은 개체명 부착 말뭉치와 유사하다[4]. 단, 각 이벤트는 개체명을 하위논항으로 요구하므로, 해당 개체명의 번호가 추가된다. 예를 들어, [표5]의 첫 번째 이벤트 예제를 해석하면 다음과 같다. 1번째 줄 “언더우드관은 고딕 양식의 석조건축물이다”라는 문장에서 4번째 어절의 6번째 음절부터 7번째 음절까지의 동사 ‘이다’는 이벤트 p1 is identified by를 유발한다. 이때, 하위논항으로 1번 개체명과 3번 개체명을 취하는데, [표4]에 따르면 1번 개체명은

표 5. 이벤트 부착 말뭉치
Table 5. Event Annotated Corpus

번호	이벤트	종류		개체명1	개체명2	시작위치	끝위치
1	이다	p1	is identified by(identifies)	entity1	entity3	1:4:6	1:4:7
2	으로 건축된	p20	had specific purpose(was purpose of)	entity5	entity4	2:1:3	2:2:3
3	로 사용되고 있다	p16	used object(was used for; mode of use:string)	entity4	entity11	2:6:3	2:8:2

'언더우드관'이고 3번 개체명은 '석조건축물'에 해당된다. 즉, 이 예제는 '석조건축물':E24 인공물(건물)-is identified by-'언더우드관':E41 명칭의 이벤트를 나타낸다.

2. 개체명 자동인식 패턴 추출

말뭉치 구축자의 수작업을 줄이기 위해서, [표4]와 같이 이미 구축한 개체명 부착 말뭉치를 분석하여 자동으로 개체명을 인식할 수 있는 패턴을 추출한다. 대량의 개체명 부착 말뭉치를 바탕으로 통계적 방법을 이용하면 개체명 자동 인식률이 매우 높게 나타날 수 있지만, 문화유산정보와 관련된 개체명 부착 말뭉치는 거의 없기 때문에 단순한 규칙 패턴을 추출하여 활용한다. 개체명 자동인식 패턴은 개체명기반 패턴, 음절기반 패턴, 문맥기반 패턴으로 나누어 볼 수 있다.

첫째, 개체명기반 패턴은 [표4]와 같은 개체명 부착 말뭉치에서 개체명과 그 종류 및 빈도수를 분석하여 [표6]과 같이 개체명기반 자동 인식 패턴을 추출한다. 예를 들어, 개체명 '언더우드관'은 전체 말뭉치에서 5번 나타났고, 이중 3번은 E41 명칭으로 분류되고 2번은 E24 인공물(건물)로 분류되었다. 고빈도 우선순위를 적용하여, 새로운 문서에서 문자열 '언더우드관'이 나타나면 항상 E41 명칭으로 분류한다.

표 6. 개체명기반 패턴 일부
Table 6. Some Named Entity-based Patterns

빈도	개체명	종류
28	건물	E24 인공물(건물) (Physical Man-Made Stuff)
1	고딕 양식	E17 형태,양식 구분 (Type Assignment)
1	교회 건물	E24 인공물(건물) (Physical Man-Made Stuff)
1	구세군 학교 건물	E24 인공물(건물) (Physical Man-Made Stuff)
1	대학본부	E15 구분, 규정 (Identifier Assignment)
2	동대문직업학교	E41 명칭 (Appellation)
1	목조 건물	E24 인공물(건물) (Physical Man-Made Stuff)
1	문과대학 건물	E24 인공물(건물) (Physical Man-Made Stuff)
2	배화여자고등학교	E41 명칭 (Appellation)
1	병원건물	E24 인공물(건물) (Physical Man-Made Stuff)
1	사택건물	E24 인공물(건물) (Physical Man-Made Stuff)
1	서울대학교	E41 명칭 (Appellation)
5	구조	E17 형태,양식 구분 (Type Assignment)
2	건축구조	E17 형태,양식 구분 (Type Assignment)
3	언더우드관	E41 명칭 (Appellation)
1	대칭구조	E17 형태,양식 구분 (Type Assignment)
3	연세대학교	E41 명칭 (Appellation)
4	연희전문학교	E41 명칭 (Appellation)
4	영락중고등학교	E41 명칭 (Appellation)
2	이화여자대학교	E41 명칭 (Appellation)
1	벽돌 구조	E17 형태,양식 구분 (Type Assignment)
3	주일학교	E19 물리적 존재 (Physical Object)
1	지역사회 학교	E19 물리적 존재 (Physical Object)
1	철골구조	E17 형태,양식 구분 (Type Assignment)
3	학교	E75 분류기호개념객체 (Conceptual Object Appellation)

둘째, 음절기반 패턴은 개체명 기반 패턴이 주어졌을 때, 개체명의 뒤음절열을 분석하여 2음절열 이상 일치하는 경우를 패턴으로 추출한다. 예를 들어, [표6]의 '건물', '교회 건물', '구세군 학교 건물', '목조 건물', '문과대학 건물', '병원건물' 등에서 [표7]처럼 '건물'과 공백을 포함한 ' 건물'을 음절기반 패턴으로 추출한다. 개체명기반 패턴에서는 개체명 '건물'만 고려하므로 빈도가 28번이었다. 반면, 음절기반 패턴에서는 '건물'을 포함하는 모든 개체명을 고려하므로, 총 37번의 '건물' 음절열 나타났고, 이중 36번은 E24 인공물(건물)로 분류되었다. 개체명 기반 패턴과 마찬가지로 고빈도 우선순위를 적용하여, 새로운 문서에서 문자열 '건물'이 나타나면 어절 앞까지 개체명을 인식하고 그 종류를 항상 E24 인공물(건물)로 분류한다.

표 7. 음절기반 패턴 일부
Table 7. Some Syllable-based Patterns

빈도	음절	종류
36	건물	E24 인공물(건물) (Physical Man-Made Stuff)
5	건물	E24 인공물(건물) (Physical Man-Made Stuff)
6	고등학교	E41 명칭 (Appellation)
6	대학교	E41 명칭 (Appellation)
17	학교	E41 명칭 (Appellation)

셋째, 문맥기반 패턴은 원문에서 개체명 뒤쪽에 어떤 문맥이 나타났는지 분석하여 패턴으로 추출한다. 예를 들어, 원문의 문장 "건물 중앙출입구는 튜더 아치형으로 돌 구조 현관이 잘 보존되어 있다."에서 개체명 '건물'은 E24 인공물(건물)로 분류된다. 이때, '건물'에 뒤따르는 문맥 '중앙출입구는'을 [표8]과 같이 문맥기반 패턴으로 추출한다. 그리고, 문장 "좌측 중앙에 있는 건물이 주제관이다"에서 '좌측'은 E81 방위로 분

표 8. 문맥기반 패턴 일부
Table 8. Some Context-based Patterns

빈도	문맥	종류
1	중앙에	E81 방위
1	중앙출입구는	E24 인공물(건물)(Physical Man-Made Stuff)
4	중에는	E5 사건(제사, 행사, 의례, 예배, 미사, 축제)(Event)
2	중이던	E7 활동, 행위(Activity)

류되며 뒤따르는 '중앙에'를 문맥기반 패턴으로 추출한다.

이렇게 추출한 개체명 자동인식 패턴은 적용율을 높이기 위해서, 쉼표, 온점, 격조사 등을 제거하는 작업을 수행한다. 개체명의 경우 개체종류에 속하는 의미를 가지는 단어로 쉼표나 온점은 불필요한 부분이고, 단어의 뒤에 붙은 조사도 개체명에 의미를 부여하는 것과는 무관하므로, 이를 제거하여 말뭉치의 정확성을 높인다.

3. 개체명 자동인식 패턴 적용

개체명 부착 학습말뭉치를 분석하여 개체명 자동인식 패턴을 추출한 결과를 [그림1]의 개체명 자동인식 패턴으로 저장한다. 새로운 원문 텍스트가 입력되면, 말뭉치 구축자의 수작업을 진행하기 전에 미리 개체명 자동인식 패턴을 활용하여 개체명을 자동인식한다. 이 때, 개체명기반 패턴, 음절기반 패턴, 문맥기반 패턴을 이용하여 개체명 후보를 생성한다.

생성된 모든 후보에 대해 시작위치와 끝 위치가 겹치는 경우 가장 긴 개체명 후보를 선택하는 최장일치 우선순위를 적용한다. 길이가 동일할 경우 수식 (1)과 같이 개체명 추천 점수를 계산하여 가장 높은 점수를 가진 후보를 최종적으로 선택한다. 즉, 여러 후보 중에 동일한 길이의 후보는 해당 후보를 추천한 패턴의 가중치 및 빈도를 고려하여 점수를 계산한다. 이때, 개체명 자동인식 패턴의 가중치는 패턴 정확도를 바탕으로 계산되는데, 개체명기반 패턴의 가중치가 가장 크고, 문맥기반 패턴의 가중치가 작도록 조정한다.

$$Score_i = \begin{cases} \text{개체명 후보}_i\text{를 추천한 개체명 기반 패턴의 빈도수} * \lambda_1 \\ + \text{ 개체명 후보}_i\text{를 추천한 음절기반 패턴의 빈도수} * \lambda_2 \\ + \text{ 개체명 후보}_i\text{를 추천한 문맥 기반 패턴의 빈도수} * \lambda_3 \end{cases}$$

$$\text{단,} \quad \lambda_1 \gg \lambda_2 \gg \lambda_3 \qquad (1)$$

예를 들어, 문장 "기존 문과대학 건물을 증축하였다."이 입력되면, 개체명기반 패턴에서 '문과대학 건물→E24 인공물(건물)'을 적용하여 개체명 후보 '문과대학 건물'를 생성한다. 또한, '건물→E24 인공물(건물)'도 적용하여 개체명 후보 '건물'도 생성한다. 그리고, 음절기반 패턴에서는 ' 건물→E24 인공물(건물)'를 적용하여 ' 건물'부터 어절 시작부분까지 개체명후보로 인식하여 '문과대학 건물'를 생성한다. 음절기반 패턴 '건물→E24 인공물(건물)'의 경우 '건물'이 어절시작부분이 되므로, 개체명 후보 '건물'을 생성한다. 최종적으로 개체명 후보는 E24 인공물(건물) '문과대학 건물'과 E24 인공물(건물) '건물'의 두 개가 생성되었다. 최장일치 우선순위 원칙에 따라 개체명 후보 E24 인공물(건물) '문과대학 건물'를 개체명으로 선택한다.

한편, 문장 "제주시청 건물은 한국전쟁 중인 지난 1952년에 준공되어 당시에는 제주도청사로 사용되었다."가 주어지면, 개체명 기반 패턴에서는 '제주시청 건물'이라는 개체명이 학습 말뭉치에 없었으므로 이를 개체명으로 인식할 수 없다. 반면, 음절기반 패턴에서는 '건물→E24 인공물(건물)'을 적용하여 개체명 후보 '건물'을 생성한다. 이와 함께, 공백을 포함한 ' 건물→E24 인공물(건물)' 패턴을 바탕으로 뒤음절 ' 건물'부터 어절 앞까지 '제주시청 건물'을 개체명 후보로 인식하여 생성한다. 최장일치 우선순위 원칙에 따라 개체명 후보 E24 인공물(건물) '제주시청 건물'을 개체명으로 선택한다.

IV. 실험 및 평가

현재 문화유산정보 말뭉치는 서울시근대문화유산목록[15] 중 한성부 내에 존재하는 문화재 일부를 선택하고, 이 문화재에 대한 내용을 한국민족문화대백과사전[16]에서 원문 발췌 후 개체명을 부착하여 구축하고 있다[14]. 제안하는 문화유산정보 말뭉치 구축을 위한 개체명 및 이벤트 부착 도구가 말뭉치 구축자의 수작업을 어느 정도 줄여줄 수 있는지 알아보기 위해서, [표9]와 같이 소규모로 구축된 학습말뭉치에서 개체명 자동인식 패턴을 추출하였고, 이를 바탕으로 새로운 문서의 개체명을 자동인식하였다. 따라서, 말뭉치 구축자는 새로운 문서에서 전체 개체명을 모두 찾을 필요 없이 자동인식한 결과를 바탕으로 자동인식에 실패한 개체명만 추가하고, 잘못 인식한 개체명을 삭제하는 작업을 수행한다.

평가말뭉치는 구축된 문화유산정보 말뭉치 총 31개의 파일을 사용한다. 이와 같이 구축된 말뭉치의 양이 매우 적으므로, 1개의 파일만 실험집합으로 사용하고 나머지 30개의 파일은 모두 학습집합으로 사용한다. [표9]에 제시된 바와 같이 학습집합은 2,650개의 문장으로 구성되어 있고, 5,365개의 개체명이 부착되어 있다. 실험집합은 98개의 문장으로 구성되고 총 238개의 개체명이 부착되어 있다.

표 9. 평가 말뭉치 분석
Table 9. Evaluation Corpus Analysis

	개체명수	어절수	문장수
학습집합	5,365	11,795	2,650
실험집합	238	524	98

학습집합과 실험집합을 모두 포함하는 평가 말뭉치에서 하나의 문자열이 얼마나 다양한 개체명 종류를 가질 수 있는지

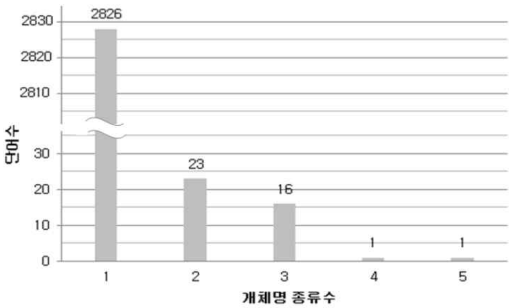

그림 3. 개체명 종류수에 따른 단어수
Fig. 3. The Number of Words According to the Number of the Corresponding Named Entity Categories

분석하였다. 예를 들어, '언더우드관'이 전체 말뭉치에서 5번 나타났고, 이중 3번은 E41 명칭으로 분류되고 2번은 E24 인공물(건물)로 분류되었다면, '언더우드관'은 총 2가지 개체명 종류를 가지는 문자열로 분류된다. 그리고, 전체 말뭉치에서 '명동성당'이 E41 명칭, E24 인공물(건물), E66 형성(조직, 단체 등 비물리적인 것)으로 여러 번 나타났다면, '명동성당'은 3가지 개체명 종류를 가지는 문자열이 된다. 대부분의 문자열이 전체 말뭉치에서 한 번만 나타나므로, 1가지 개체명 종류를 가지는 문자열이 대부분이다. 그럼에도 불구하고, [그림3]에 제시된 바와 같이 2가지 이상의 개체명 종류를 가지고 있는 문자열도 41개가 있다. 이는 단순한 규칙만으로는 정확하게 개체명을 인식하고 그 종류를 분류하는 것이 쉽지 않다는 것을 보여준다.

한편, 개체명을 자동으로 인식한 결과에 대해 얼마나 올바르게 개체명을 인식하였는지를 정량적으로 평가하기 위해서, 수식 (2)과 같이 정확률을 측정하고, 얼마나 많은 개체명을 올바르게 찾았는지를 평가하기 위해서 수식 (3)과 같이 재현율을 측정하였다. 수식 (4)는 정확률과 재현율의 조화평균인 f-척도를 나타낸다.

$$\text{정확률} = \frac{\text{올바르게 자동인식한 개체명수}}{\text{자동인식한 개체명수}} \quad (2)$$

$$\text{재현율} = \frac{\text{올바르게 자동인식한 개체명수}}{\text{말뭉치에 포함된 전체 개체명수}} \quad (3)$$

$$f-\text{척도} = \frac{2 \times \text{정확률} \times \text{재현율}}{\text{정확률} + \text{재현율}} \quad (4)$$

이러한 정확률과 재현율의 측정방법을 바탕으로 한 개체명 자동인식결과의 평과 결과는 [표10]과 같다. 실험집합의 평가결과를 살펴보면 개체명기반 패턴의 정확률은 76.37%로 음절기반 패턴의 47.98%나 문맥기반 패턴의 14.89%에 비해 매우 높다. 이는 학습집합에 있었던 개체명이 새로운 문서에서도 개체명이 되는 경우가 75%이상 많다는 것을 나타낸다. 23.63%의 오류유형 중에는 "남대문에 전도소를 개설하였고"에서 '남대문'은 개체명 종류가 E44 장소명인데, 학습집

표 10. 개체명 자동 인식 패턴에 따른 성능
Table 10. Performance According to Named Entity Recognition Patterns

패턴종류	규칙수	학습						실험					
		정확률		재현율		f-척도		정확률		재현율		f-척도	
① 개체명기반패턴	2,780	97.94	-	95.75	-	96.83	-	76.37	-	58.40	-	66.19	-
② 음절기반패턴	1,943	51.68	▽46.26	41.45	▽54.30	46.01	▽50.83	47.98	▽28.40	34.87	▽23.53	40.39	▽25.80
③ 문맥기반패턴	3,098	11.86	▽86.08	8.03	▽87.72	9.58	▽87.26	14.89	▽61.48	8.82	▽49.58	11.08	▽55.11
① + ②	4,723	85.68	▽12.26	83.90	▽11.85	84.78	▽12.05	73.87	▽ 2.50	61.76	△ 3.36	67.28	△ 1.09
① + ③	5,878	84.58	▽13.36	82.80	▽12.95	83.68	▽13.16	65.05	▽11.33	56.30	▽ 2.10	60.36	▽ 5.83
② + ③	5,041	44.62	▽53.32	39.03	▽56.72	41.64	▽55.19	43.68	▽32.69	34.87	▽23.53	38.79	▽27.41
① + ② + ③	7,821	77.33	▽20.62	75.77	▽19.98	76.54	▽20.29	66.35	▽10.02	58.82	△ 0.42	62.36	▽ 3.83

표 11. 개체명 자동인식을 이용한 수작업 감소
Table 11. The Decrease of Human Intervention by Named Entity Recognition Patterns

패턴종류	문자열 선택횟수	개체명 종류 선택횟수	삭제 횟수	총 개입횟수
사용안함	238	238	0	476
① 개체명기반패턴	99	99	43	198
② 음절기반패턴	155	155	90	310
③ 문맥기반패턴	217	217	120	434
① + ②	91	91	52	182
① + ③	104	104	72	208
② + ③	155	155	107	310
① + ② + ③	98	98	71	196

합에서 추출한 개체명기반 패턴 '남대문→E24 인공물(건물)'이 잘못 적용된 경우가 많이 발견되었다. 이는 [그림3]에서 살펴본 바와 같이 동일한 문자열에 대해 둘 이상의 개체명 종류가 허용되는 경우가 있기 때문이다.

개체명기반 패턴에 음절기반 패턴을 추가하여 함께 활용하는 경우 정확률은 76.37%에서 73.87%로 2.50%정도 떨어졌지만 재현율은 58.40%에서 61.76%로 3.36%정도 상승한 것을 알 수 있다. 이는 '구세군 학교 건물→E24 인공물(건물)', '목조 건물→E24 인공물(건물)', '문과대학 건물→E24 인공물(건물)' 등의 개체명 기반 패턴은 '병원건물'을 인식하지 못하지만, 음절기반 패턴 '건물→E24 인공물(건물)'은 패턴에 포함된 문자열 '건물'부터 어절앞까지 개체명으로 인식하므로 '병원건물'을 인식할 수 있다는 것을 보여준다. 음절기반 패턴이 자체적으로는 정확률과 재현율이 매우 떨어짐에도 불구하고, 개체명기반 패턴의 성능을 크게 떨어뜨리지 않는 이유는 최장일치 우선순위를 적용하고, $\lambda_1 \gg \lambda_2$와 같이 개체명기반 패턴이 추천한 후보의 점수가 음절기반 패턴이 추천한 후보의 점수보다 매우 높게 계산되도록 가중치가 설정되어 있기 때문이다.

개체명 부착 말뭉치 구축시 말뭉치 구축자는 원문에서 개체명에 해당하는 문자열을 드래그해서 선택하고, 드래그한 문자열의 개체명 종류를 선택한다. 개체명 자동인식 패턴을 활용하는 경우, 그 결과를 살펴보고 잘못 인식한 개체명은 제안하는 도구를 사용하여 수동으로 삭제한다. 이를 고려하여 [표10]과 같은 성능을 보이는 개체명 자동인식 패턴을 활용하여 개체명 부착 말뭉치를 구축하는 경우 실제 수작업이 얼마나 감소하는지를 [표11]과 같이 분석하였다. 말뭉치 구축자는 원래 총 238개의 개체명을 부착해야하는데, 개체명 자동인식 패턴은 총 182개를 자동 인식하였다. 이중 139개는 올바른 인식결과였고, 43개는 틀린 인식결과였다. 따라서, 말뭉치 구축자는 총 238개중 자동인식에 성공한 139개를 제외한 99개의 개체명을 찾아서 그 종류를 선택한다. 그리고, 틀리게 인식한 43개의 개체명을 선택하여 삭제해야 한다. 개체명 자동인식 패턴을 사용할 경우 [표11]에 제시된 바와 같이 말뭉치 구축자의 수작업량이 절반이상 줄어들 수 있다.

V. 결 론

본 논문에서는 문화유산정보 말뭉치 구축을 위한 개체명 및 이벤트 부착 도구를 제안하였다. 제안하는 도구를 사용하여 말뭉치 구축자는 문화유산정보 관리에 유용한 정보인 시간, 장소, 인물, 사건을 중심으로 개체명과 이벤트를 부착할 수 있다. 제안하는 방법의 특징은 다음과 같다.

첫째, 제안하는 도구는 말뭉치 구축자가 개체명과 이벤트 부착을 쉽게 할 수 있도록 설계되어 있다. 즉, 제안하는 도구에서 개체명이나 이벤트 단어의 위치정보를 자동으로 부착한다. 또한, 구축한 개체명이나 이벤트를 선택하면, 해당 문자열을 원문에서 진한 이탤릭체로 표시되므로, 올바르게 부착되고 있는지 쉽게 확인할 수 있다.

둘째, 제안하는 도구는 말뭉치 구축자의 수작업을 상당히 줄여줄 수 있다. 말뭉치 구축자가 개체명을 부착하기 전에 개체명 인식 패턴을 활용하여 60%이상의 개체명을 자동으로 미리 인식하므로, 말뭉치 구축자의 부착 부담이 상당히 줄어든다. 실험결과 제안하는 부착 도구는 말뭉치 구축자의 수작업량을 절반이상 줄여주었다.

셋째, 제안하는 도구는 개체명 자동인식 패턴의 추가 및 수정이 용이하다. 즉, 필요할 때마다 학습말뭉치에서 추가적인 분석처리 없이 패턴을 자동으로 추출하여 사용할 수 있다. 단, 말뭉치 구축 초창기에는 학습말뭉치가 거의 없다는 점을 고려하여 단순한 규칙패턴을 중심으로 개체명을 인식한다.

참고문헌

[1] Bang-Hyeon Na, “A Design of Cultural and Historical Contents Model for Web Services”, Proceedings of the Association of Korean Cultural and Historical Geographers Symposium, pp.27-35, Nov. 2010.

[2] Dong-hwan Yoo, “The current situation and the task of developing the national cultural heritage contents”, Korean Studies, Vol.12, pp.5-49, Jun. 2008.

[3] So-Young Cha, Jung-Wha Kim, “Constructing a Foundation for Semantic Structure of Korean Heritage Information : A Study on Creating a Substructure of Korean Heritage Portal by Implementing CIDOC CRM”, Proceedings of the 17th Conference on the Korean Society for Information Management, pp.177-184, Aug. 2010.

[4] Tomoko Ohta, Jin-Dong Kim, Sampo Pyysalo, Yue Wang, Jun'ichi Tsujii, “Incorporating GENETAG-style annotation to GENIA corpus”, Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing(BioNLP), pp.106-107, Jun. 2009.

[5] Özlem Uzuner, Brett R South, Shuying Shen, Scott L DuVall, “2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text”, J Am Med Inform Assoc, Vol.18, No.5, pp.552-556, Jun. 2011.

[6] Hae-Chang Rim, Young-Sook Hwang, Kyung-Mi Park, “Development of Bio Text Mining System”, Communications of KIISE, Vol.21, No.6, pp.60-68, Jul. 2003.

[7] Masaki Noguchi, Kenta Miyoshi, Takenobu Tokunaga, Ryu Iida, Mamoru Komachi, Kentaro Inui, “Multiple Purpose Annotation using SLAT -Segment and Link-based Annotation Tool-”, Proceedings of the 2nd Linguistic Annotation Workshop, pp.61-64, May. 2008.

[8] Mitchell P. Marcus, B. Santorini, and M. A. Marchinkiewicz, “Building a large annotated corpus of English : the Penn TreeBank”, Computational Linguistics, Vol.19, No.2, pp.313-330, Jun. 1993.

[9] Hye-Kyum Kim, Kyung-Mi Park, Yeo-Chan Yoon, Hae-Chang Rim, So-Young Park, “Tree Tagging Tool using Two-phrase Parsing”, Proceedings of the 17th Annual Conference on Human and Cognitive Language Technology, pp.151-158, Oct. 2005.

[10] Piek Vossen, Attila Görög, Fons Laan, Maarten van Gompel, Rubén Izquierdo, Antal van den Bosch, “DutchSemCor: Building a semantically annotated corpus for Dutch”, Proceedings of eLex, pp.286-296, Nov. 2011.

[11] Joo-Young Lee, Young-In Song, Hae-Chang Rim, “Title Named Entity Recognition based on Automatically Constructed Context Patterns and Entity Dictionary”, Proceedings of the 17th Annual Conference on Human and Cognitive Language Technology, Vol.16, No.1, pp.111-117, Oct. 2004.

[12] Chang-Ki Lee, Myung-Gil Jang, “Named Entity Recognition with Structural SVMs and Pegasos algorithm”, Cognitive Science, Vol.21, No.4, pp.655-667, Dec. 2010.

[13] Seong-Won Kim, Dong-Yul Ra, “Korean Named Entity Recognition Using Two-level Maximum Entropy Model”, Proceedings of KIISE Symposium, Vol.2, No.1, pp.81-86, Jun. 2008.

[14] Hee-Sun Chung, Hee-Sun Kim, "Database and Corpus Construction methodology for the Content of Religious architectural heritage Information", Proceedings of a Seminar Held by the Convergence Study Team of SangMyung University, pp.43-60, Jun. 2012.

[15] The Institute of Seoul Studies, “Modern Cultural Heritage Resource and Cataloging Project Report”, Jun. 2004.

[16] The Academy of Korean Studies, “Encyclopedia of Korean Culture”, Dec. 1991.

저 자 소 개

최 지 예

현 재 : 상명대학교
디지털미디어학부 재학중
관심분야 : 컴퓨터공학
Email : cgycgy@naver.com

김 명 근

현 재 : 상명대학교
디지털미디어학부 재학중
관심분야 : 컴퓨터공학
Email : kimaudrms@hanmail.net

박 소 영

1997 : 상명대학교
전자계산학과 이학사.
1999 : 고려대학교
컴퓨터과학과 이학석사.
2005 : 고려대학교
컴퓨터과학과 이학박사.
현 재 : 상명대학교
게임모바일콘텐츠학과 조교수
관심분야 : 컴퓨터공학
Email : ssoya@smu.ac.kr