개체명 사전 기반의 반자동 말뭉치 구축 도구

노경목[o†], 김창현[‡], 천민아[†], 박호민[†], 윤호[†], 김재균[†], 김재훈[†]
한국해양대학교[†], 한국전자통신연구원[‡]
kmq7542@gmail.com[o†], chkim@etri.re.kr[‡], minah0218@kmou.ac.kr[†],
homin2006@hanmail.net[†], 4168615@naver.com[†], jgk20000@naver.com[†], jhoon@kmou.ac.kr[†]

A Semi-automatic Annotation Tool based on Named Entity Dictionary

Kyung-Mok Noh[o†], Chang-Hyun Kim[‡], Min-Ah Cheon[†],
Ho-Min Park[†], Ho Yoon[†], Jae-Kyun Kim[†], Jae-Hoon Kim[†]
Korea Maritime and Ocean University[†], Electronics and Telecommunications Research Institute[‡]

요 약

개체명은 인명, 지명, 조직명 등 문서 내에서 중요한 의미를 가지므로 질의응답, 요약, 기계번역 분야에서 유용하게 사용되고 있다. 개체명 인식은 문서에서 개체명에 해당하는 단어를 찾아 개체명 범주를 부착하는 작업을 말한다. 개체명 인식 연구에는 개체명 범주가 부착된 개체명 말뭉치를 사용한다. 개체명의 범주는 연구 분야에 따라 다양하게 정의되므로 연구 분야에 적합한 개체명 말뭉치가 필요하다. 하지만 이런 말뭉치를 구축하는 일은 시간과 인력이 많이 필요하다. 따라서 본 논문에서는 개체명 사전 기반의 반자동 말뭉치 구축 도구를 제안한다. 제안하는 도구는 크게 전처리, 사용자 태깅, 후처리 단계로 나뉜다. 전처리 단계는 자동으로 개체명을 찾는 단계이다. 약 11만 개의 개체명을 기반으로 하여 트라이(trie) 구조의 개체명 사전을 구축한 후 사전을 이용하여 개체명을 자동으로 찾는다. 사용자 태깅 단계는 사용자가 수동으로 개체명을 태깅하는 단계이다. 전처리 단계에서 찾은 개체명 중 오류가 있는 개체명들은 수정하거나 삭제하고, 찾지 못한 개체명들은 사용자가 추가로 태깅하는 단계이다. 후처리 단계는 태깅한 결과로부터 사전 정보를 갱신하는 단계이다. 제안한 말뭉치 구축 도구를 이용하여 752개의 뉴스 기사에 대해 개체명을 태깅한 결과 7,620개의 개체명이 사전에 추가되었다. 제안한 도구를 사용한 결과 사용하지 않았을 때 비해 약 57.6% 정도 태깅 횟수가 감소했다.

주제어: 개체명 사전, 말뭉치 구축, 주석 도구

1. 서론

개체명 인식은 정보추출의 한 부분으로 문서 내에서 중요한 의미를 지닌 개체명을 찾아 그 범주를 결정하는 작업이다[1-3]. 개체명은 인명, 지명, 조직명 등으로 분류할 수 있으며 질의응답, 요약, 기계번역 분야에서 핵심어로써 매우 유용하게 사용되고 있다[1-3]. 개체명은 대부분이 고유 명사이며 시간의 흐름에 따라 계속 생성되기 때문에 모든 개체명을 사전에 등록하는 것은 불가능한 일이다[2]. 또한, 같은 단어라도 문맥에 따라 개체명의 범주가 달라지는 모호성(ambiguity) 문제도 존재한다[2-3]. 따라서 개체명을 찾는 일과 그 범주를 결정하는 일은 쉽지 않다[2-3].

개체명 인식을 위해 기계학습과 사전 정보를 활용한 다양한 연구가 진행되었다[2,4-6]. [2]는 기계학습 기반 개체명 인식에서 중의성을 자질에 명확하게 표기하기 위한 사전 자질 생성에 대해 연구했다. [4]는 한국어 위키피디아로부터 개체명 사전을 자동으로 구축하는 연구를 진행했으며 [5]는 원시 말뭉치로부터 추출한 문맥 패턴 정보와 개체명 사전을 이용한 제목 개체명 인식에 대해 연구했다. [6]은 기계 학습 모델을 이용한 개체명 인식을 연구했다.

개체명은 문맥에 따라 범주가 달라지므로 개체명 범주를 명확히 하기 위해서는 문맥 정보를 고려할 필요가 있다. 즉, 문장에서 등장하는 자질을 기준으로 문장 단위로 개체명의 범주를 결정하는 것보다 문서 전체의 맥락을 파악하여 각 단어의 개체명 범주를 결정하는 것이 더 정확하다. 그러나 사람이 수작업으로 모든 문서의 개체명을 찾아 범주를 부착하는 것은 시간이 오래 걸리고 힘든 일이다. 따라서 본 논문에서는 개체명 사전을 활용하여 사용자가 개체명 범주 부착 말뭉치를 효율적으로 구축하기 위한 반자동 말뭉치 구축 도구를 제안한다.

본 논문의 구성은 다음과 같다. 2장에서 현재 사용되고 있는 말뭉치 구축 도구에 대해 살펴본다. 3장에서 제안한 말뭉치 구축 도구를 소개한다. 마지막으로 4장에서 결론 및 향후 과제를 언급하며 끝을 맺는다.

2. 관련 연구

[7]은 [8]에서 제안한 시스템의 성능이 말뭉치의 크기에 비례하는 것이라 가정하고, 이를 보완하기 위해 말뭉치 확장을 위한 반자동 의미 부착 도구를 개발하였다. JSON 형식을 따르는 말뭉치 파일을 사용하여 미리 자동 태깅한 구문 분석 정보와 의미역 정보를 화살표로 표현하고 사람이 이해하기 쉽도록 GUI 기반의 인터페이스 제공했다. 그 결과 도구를 사용하지 않았을 때보다 말뭉치

구축 속도를 약 80% 향상시켰다.

[9]은 자동화된 정보 추출 도구의 훈련 및 실험, 평가를 위한 수동 주석(annotation) 지원 도구인 COAT를 소개하고 있다. 둘 이상의 사용자에게 동일한 문서를 배정하여 각각 주석을 달게 한다. 각 사용자로부터 받은 결과물을 통합하여 최종 말뭉치를 완성하는데, 교차 검증을 통해 주석 말뭉치의 신뢰성을 높이기 위한 방법을 사용했다. [9]의 도구도 사용자의 작업 효율을 위해 GUI 기반의 인터페이스와 단축키를 지원하여 말뭉치 구축 속도를 향상시켰다.

해외의 문서 말뭉치 구축 지원 도구는 GATE[10]와 brat[11] 등이 있다. GATE는 다양한 문서 분석 및 처리에 유용한 JAVA 기반 오픈 소프트웨어이다. GATE는 전체 코어 시스템을 재사용할 수 있는 JAVA 구성 요소로 분할 가능하다. 분할된 JAVA 구성 요소는 임베디드 시스템 등에서 재사용이 가능해서 확장성이 높다. GATE의 핵심 기능은 구문 분석, 형태소 분석, 정보 검색 도구, 주석 등이 있으며 이 외에도 텍스트를 처리하기 위한 다양한 구성 요소가 포함되어 있다. brat은 웹 기반의 문서 주석 도구이다. 정해진 형식의 구조화된 주석을 위해 설계되었으며, 주석, 정보 추출, 의존 구문, 정규화, 청킹(chunking) 등 다양한 기능을 지원한다. 시각화가 잘 되어 있어서 단어와 단어의 관계를 잘 보여주기 때문에 의존 관계를 정의할 때 유용하며 드래그 앤 드롭을 지원해서 편리한 UI를 제공한다.

3. 시스템 소개

본 논문에서 제안하는 반자동 말뭉치 구축 도구 시스템은 크게 전처리, 사용자 개체명 태깅, 후처리 단계로 나뉘며 구성도는 그림 1과 같다. 그림 1에서 실선 화살표는 작업 순서를 의미하고 점선 화살표는 개체명 사전과 관련된 정보의 흐름을 나타낸다.

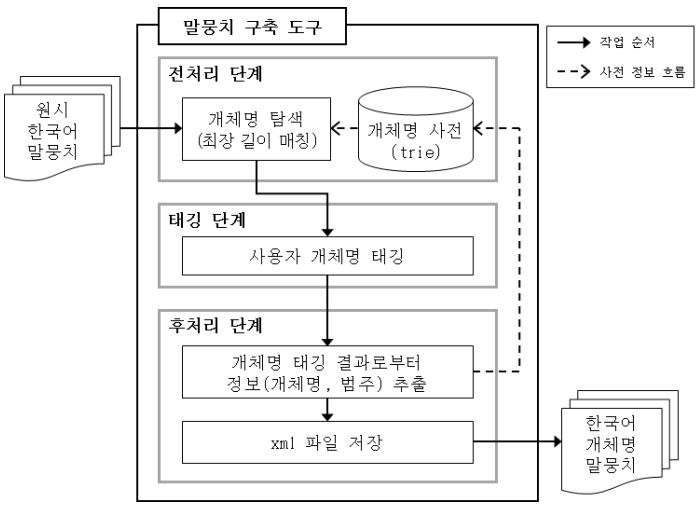

그림 1. 반자동 말뭉치 구축 도구 시스템 구성도

3.1 전처리 단계(자동)

전처리 단계에서는 개체명 사전을 이용하여 원시 말뭉치로부터 개체명을 자동으로 찾는 단계이다. 개체명 사전은 트라이(trie) 구조로 되어 있으며 개체명과 개체명의 범주, 개체명의 빈도수의 정보를 담고 있다. 사전에는 2음절 이상의 개체명만 등록되어 있다. 이러한 사전을 이용해서 문서의 시작부터 순차적으로 개체명을 탐색한다. 예를 들어, 개체명 사전에 “서울시”와 “서울시청”의 개체명이 존재할 때, “서울시는 서울시청에”란 문장의 탐색 과정은 그림 2와 같다.

그림 2. 예시 문장의 개체명 탐색 과정

3.2 사용자 개체명 태깅 단계(수동)

사용자 개체명 태깅 단계는 사용자가 개체명을 수동으로 태깅하는 단계이다. 사용자는 전처리 단계에서 잘못 부착된 개체명의 범주를 수정하거나 삭제하고 새로운 개체명을 발견하면 적절한 범주를 부착한다.

태깅 관련 기능은 세 가지가 있다. 첫 번째는 사전에 정보가 없는 개체명(unknown named entity)의 범주를 추가하는 것이고, 두 번째는 부착된 개체명의 범주를 다른 범주로 변경하는 것이다. 마지막은 부착된 개체명의 범주를 삭제하는 것이다.

3.2.1 개체명 범주 추가 기능

미등록 개체명에 범주를 추가하는 방법은 마우스나 키보드를 이용하여 개체명을 선택(드래그)하여 범주를 부착한다. 이때, 선택한 개체명과 같은 개체명이 문서 내에 존재하면 해당하는(선택한 개체명이 아닌) 모든 개체명에 같은 범주를 추가할 것인지를 “예”, “아니오” 형식으로 묻는다. “예”를 선택하면 아래의 표 1과 같

은 규칙이 적용되고 “아니오”를 선택하면 선택한 개체명만 범주를 추가한다.

표 1. 문서 내의 같은 개체명을 태깅하는 규칙

규칙	설 명
1	해당하는 개체명에 범주가 부착되어 있지 않으면 선택한 개체명의 범주를 부착한다.
2	해당하는 개체명이 어떠한 개체명의 일부분인 경우에는 태깅하지 않는다.
3	해당하는 개체명에 어떠한 개체명의 범주가 부착되어 있을 때, 선택한 개체명이 해당 개체명을 포함하거나 같다면 기존의 개체명 범주를 삭제하고 선택한 개체명의 범주를 다시 부착한다.

해당 규칙의 예시는 아래의 그림 3과 같다.

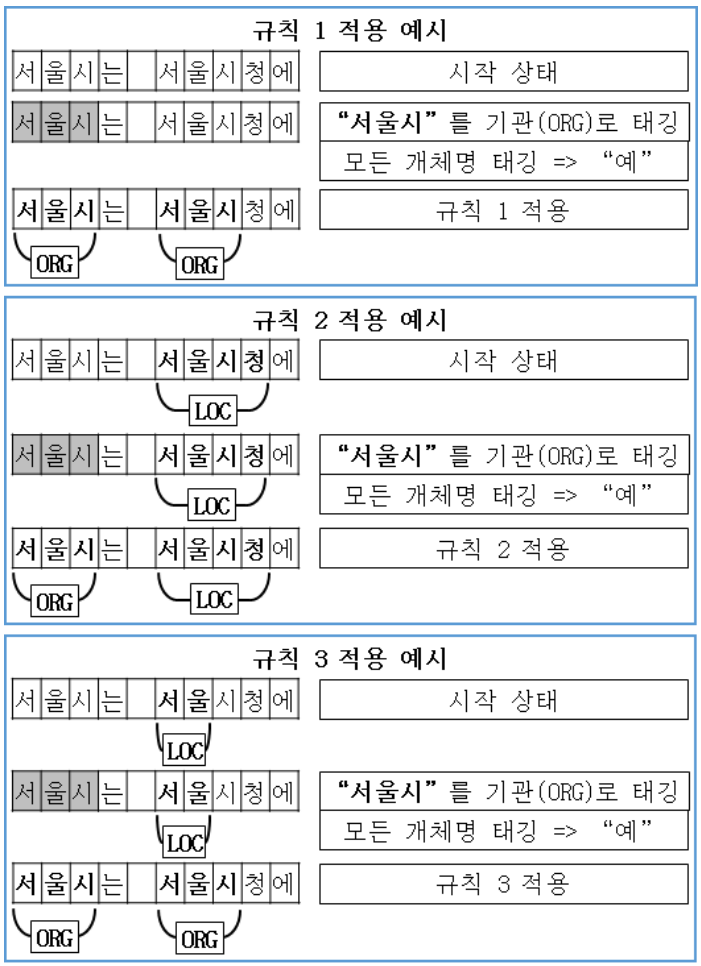

그림 3. 같은 개체명을 태깅하는 규칙 예시

3.2.2 개체명 범주 변경 기능

개체명의 범주를 변경하는 방법은 마우스로 해당 개체명을 좌클릭 또는 우클릭해서 다른 범주로 변경할 수 있다.

3.2.3 개체명 범주 삭제 기능

개체명의 범주를 삭제하는 기능은 두 가지가 있다. 첫째는 선택한 개체명의 범주만 삭제하는 것이고, 둘째는 선택한 개체명과 같은 모든 개체명의 범주를 삭제하는 것이다.

3.2.4 기타 편의 기능

그 외에 사용자 편의성을 위해 단어를 찾는 기능, 글씨의 크기를 조절하는 기능, 줄 간격을 조절하는 기능, 문서를 삭제하는 기능 등을 구현하였다.

3.3 후처리 단계

후처리 단계에서는 이때까지 진행했던 문서의 정보를 xml 파일 형식으로 저장하고 사전을 갱신하는 단계이다. 태깅을 완료한 문서로부터 2음절 이상인 개체명 정보(개체명, 범주, 빈도수)를 추출하여 개체명 사전 정보를 갱신한다.

3.4 실행 화면

그림 4는 말뭉치 구축 도구의 실행 화면 중 하나이다. 그림 4의 좌측 상단에 있는 원시 말뭉치 파일창에서 원시 말뭉치를 선택하면 전처리 단계를 거쳐 그림 4의 중앙 화면에 보이게 된다. 전처리 단계에서 찾은 개체명들은 음영처리 되어서 사용자에게 보이며 별도의 메모 없이 색상으로 구분된다. 사용자의 편의를 위해 키보드 단축키와 마우스를 이용한 태깅이 가능하다. 사용자가 개체명 태깅을 완료하면 개체명 범주 말뭉치 생성이 끝난다. 완료된 개체명 말뭉치는 그림 4의 좌측 하단에 있는 개체명 말뭉치 파일 창에 보이게 되며, 태깅한 결과를 재확인하거나 수정할 수 있다.

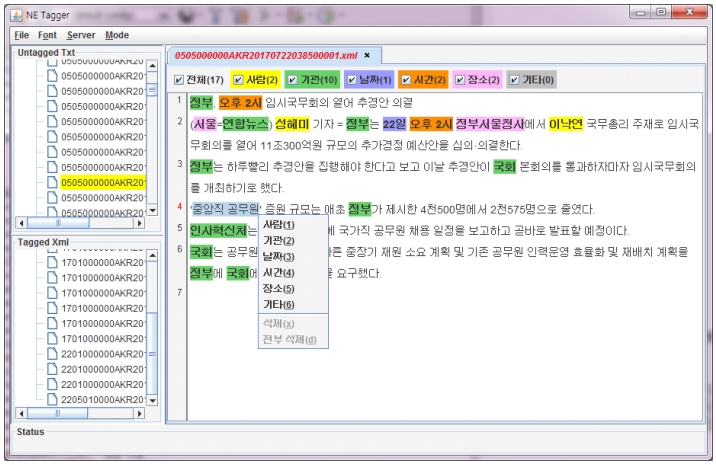

그림 4. 반자동 말뭉치 구축 도구 실행 화면

4. 결론 및 향후 과제

본 논문에서는 기존에 구현된 다양한 말뭉치 구축 도구 시스템과 개체명 인식 관련 논문을 참고하여 개체명 사전 기반의 반자동 말뭉치 구축 도구를 개발했다. 약 11만 개의 개체명이 포함된 사전 정보를 활용하여 개체명 말뭉치를 구축하였다. 원시 말뭉치는 2017년 5월~7월 기간의 뉴스 기사를 모았으며, 752개의 기사에서 개체명을 찾아 태깅하였다. 개체명의 범주는 사람, 기관, 날짜, 시간, 장소, 기타 이렇게 6가지를 선정했으며, 752개의 기사로부터 총 25,169개의 개체명을 태깅하였다. 중복을 제거한 개체명의 수는 7,620개였고 이 중 2음절 이상의 개체명은 7,546개였다. 25,169개의 개체명 중 기관이 8,365개로 가장 많았으며, 사람 5,929개, 장소 4,715개, 날짜 3,447개, 기타 2,484개, 시간 229개 순이었다. 기타는 행사명, 전쟁 명, 사건명, 프로그램명, 책 이름, 영화 제목, 노래 제목, 문화재 이름 등을 기타로 태깅했다.

개체명 말뭉치를 구축할수록 후처리 단계를 거쳐 개체명 사전이 확장되었다. 사전이 확장됨에 따라 전처리 단계에서 자동으로 태깅되는 개체명의 수가 늘어남으로써 사용자의 태깅 횟수가 감소하였다. 하지만 중의적인 표현을 가진 개체명으로 인해 사용자 태깅 단계에서 문맥에 맞지 않는 개체명을 변경, 삭제해야 하는 일이 생겼다. 예를 들어, "고려"의 경우 나라의 이름이기도 하지만 "고려하였다"와 같이 동사로 쓰이기도 하므로 오류가 발생했다. 평균적으로 한 기사에 약 33개의 개체명이 존재했으며 사용자 태깅 단계에서 사용자가 개체명 범주를 추가, 변경, 삭제해야 하는 횟수는 약 14회 정도였다.

개체명 사전을 이용하지 않고 개체명 말뭉치를 구축했을 때와 비교하면 전체적인 작업량은 약 57.6% 정도 줄었다. 하지만 불필요한 작업량이 발생하였으며, 이는 사전이 확장됨에 따라 더 늘어날 것으로 보인다. 향후 과제로는 전처리 단계에서 사용되는 개체명 사전을 기계학습 모델로 대체하여 개체명 인식의 정밀도를 높이는 방안 등 효율적으로 말뭉치를 구축할 방법을 연구할 계획이다.

감사의 글

이 논문은 2017년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임(R7119-16-1001, 지식증강형 실시간 동시통역 원천기술 개발)

참고문헌

[1] David Nadeau and Stasoshi Sekine, "A survey of named entity recognition and classification", Journal of Linguisticae Investingations, vol. 30, no. 1, pp.3-26, 2007.

[2] 김재훈, 김형철, 최윤수, "기계학습 기반 개체명 인식을 위한 사전 자질 생성", 정보관리연구, 제41권, 제2호, pp.31-46, 2010.

[3] 이경희, "한국어 문서에서 개체명 인식에 관한 연구", 제12회 한글 및 한국어 정보처리 학술대회 논문집, pp.292-299, 2000.

[4] 배상준, "한국어 위키피디아를 이용한 분류체계 생성과 개체명 사전 자동 구축", 정보과학회논문지 : 컴퓨팅의 실제 및 레터, 제16권, 제4호, pp. 492-496, 2010.

[5] 이주영, "자동 구축된 문맥 패턴과 개체명 사전에 기반한 제목 개체명 인식", 제16회 한글 및 한국어 정보처리 학술대회 발표자료집, 제16권, 제1호, pp.40-45, 2004.

[6] 이창기, "Structural SVMs 및 Pegasos 알고리즘을 이용한 한국어 개체명 인식", 인지과학, 제21권, 제4호, pp.655-667, 2010.

[7] 배장성, "한국어 의미역 말뭉치 구축을 위한 반자동 태깅 도구 개발", 한국정보과학회 2014 한국컴퓨터종합학술대회 논문집, pp. 592-594, 2014.

[8] 이창기, "Structural SVM 기반의 한국어 의미역 결정", 한국정보과학회 2014 한국컴퓨터종합학술대회 논문집, pp. 574-576, 2014.

[9] 최동현, "COAT: 시멘틱 어노테이션 말뭉치 구축 지원 도구", 제23회 한글 및 한국어 정보처리 학술대회 논문집, pp.85-89, 2011.

[10] gate [Online] https://gate.ac.uk/
[11] bart [Online] http://brat.nlplab.org/