Kane: 의미정보 말뭉치 구축 도구

배원식○ 차정원
국립창원대학교 컴퓨터공학과
wonsigi529@changwon.ac.kr, jcha@changwon.ac.kr

Kane: Knowledge Annotation Tool for Semantic Information

Won-Sik Bae○ Jeong-Won Cha
Department of Computer Engineering, Changwon National University

요 약

본 논문에서는 의미정보 말뭉치 구축 도구인 Kane에 대해 설명한다. 형태소 분석기나 구문 분석기, 개체명 인식기 등 자연어처리를 위한 기본이 되는 시스템에는 말뭉치가 필요하며, 말뭉치의 구축에는 많은 비용이 든다. 일반적으로 말뭉치 구축 작업은 전용 구축 도구가 없이 문서 편집기를 사용하여 이루어지는 경우가 많아 말뭉치 구축 작업 효율이 떨어지고, 자연스럽게 구축되는 말뭉치의 품질도 낮아진다. 문서 편집기를 사용할 때 발생하는 대표적인 문제는 키보드를 이용한 기계적인 작업이 반복된다는 것이며, 키보드 입력에 따른 오타 문제 또한 발생한다. Kane에서는 기계적인 작업 및 키보드 입력을 간편한 인터페이스를 통해 최소화하였으며, 마우스 조작으로도 쉽게 말뭉치를 구축할 수 있다. 또한 사전을 이용한 이전 작업 내용 참조 기능을 지원하여 작업의 효율성 및 일관성 문제를 개선하고자 하였다.

주제어: Kane, 말뭉치, 의미정보 말뭉치, 말뭉치 구축

1. 서 론

인터넷의 활발한 보급으로 인해 수많은 데이터가 생산되면서 높은 수준의 자연어처리에 대한 요구가 증가하고 있다. 높은 수준의 자연어처리를 수행하기 위해서는 형태소 분석기, 구문 분석기, 개체명 인식기 등의 기초 시스템을 필요로 한다. 기초 시스템들은 말뭉치로부터 확률이나 문맥 정보 등을 학습하는데, 말뭉치 구축 작업은 기초 시스템의 학습을 위한 말뭉치를 만드는 작업을 말한다. 기초 시스템의 성능은 학습에 사용되는 말뭉치에 의해 크게 좌우되므로 말뭉치 구축 작업의 중요도는 매우 높다고 할 수 있다. 그러나 말뭉치 구축 작업에는 많은 비용과 시간, 노력이 요구된다.

말뭉치 구축 작업은 단순하게 생각하면 텍스트 문서를 수정하는 작업과 같으므로 일반적으로 문서 편집기를 사용하여 작업을 수행한다. 그런데 문서 편집기를 사용하면 다음과 같은 문제점이 발생한다. 첫 번째 문제는 키보드 입력과 복사, 붙여넣기의 단순한 작업들이 기계적으로 되풀이된다는 것이다. 말뭉치 구축 작업은 원래 문서에는 없던 추가적인 정보(이하 태그)를 붙이는 작업(이하 태깅)을 말하며, 일반적으로 작업에 사용되는 태그들의 종류와 개수, 형태는 미리 정의되어 있다. 따라서 작업량이 많아질수록 동일한 태그를 추가하는 작업을 수행하는 횟수도 많아지기 때문에 문서 편집기를 이용한 작업은 효율적이지 못하다. 실제로 말뭉치 구축 작업에서 이와 같은 단순한 기계 작업에 소요되는 시간의 비중이 매우 높다. 두 번째 문제는 태그를 수작업으로 입력하면, 반드시 오타 문제가 뒤따르게 된다는 것이다. 작업자가 아무리 신경을 써서 작업을 하더라도 작업량이 많아질수록 오타가 발생할 확률이 높아진다. 따라서 오타 문제 해결을 위한 추가적인 작업 시간을 할애해야 한다. 세 번째 문제는 이전 작업 내용을 쉽게 참조할 수 없어 일관성이 깨지는 문제가 발생할 수 있다는 것이다. 작업을 진행하다보면 이전 작업 내용과 유사한 작업이 반복되는 경우가 있는데, 이 때 이전 작업 내용을 참조할 수 있다면 일관성이 깨지는 문제를 일부 해결할 수 있을 뿐더러 작업 효율도 높일 수 있다. 한 사람이 작업을 할 때는 크게 문제가 없을지도 모른다. 그러나 한 사람이 작업을 하는 경우보다 여러 사람이 공동으로 작업을 하는 경우가 더 많으며, 이 경우에 심각한 문제가 된다. 문서 편집기를 사용하면 이 문제에 대응하기가 어려우며, 개별 작업이 끝난 후 일관성을 맞추기 위한 작업에 많은 시간이 소요된다.

위와 같은 문제점을 해결하기 위한 방법 중 하나가 말뭉치 구축 작업을 위한 전용도구의 개발이다. 형태소 분석, 구문 분석, 개체명 인식 등 말뭉치가 필요한 여러 분야에서 전용도구들에 대한 연구가 이루어지고 있다. 본 논문에서 소개하는 Kane은 의미정보인 개체명, 관계정보를 부착하기 위한 전용도구이다. 앞에서 언급한 문서 편집기를 사용한 작업의 문제점을 해결하기 위한 간단하고 직관적인 인터페이스와 작업 효율성을 높이기 위한 기능들을 제공한다.

본 논문의 구성은 다음과 같다. 2장에서는 공개된 말뭉치 구축 도구들에 대해서 살펴보고, 3장에서는 Kane의 인터페이스와 사용 방법에 대해 설명하고, 4장에서는 결론과 향후 과제를 다룬다.

2. 기존 연구

2.1 The UAM Corpus Tool[1]

이 도구는 다양한 언어학 계층의 태깅을 지원하는 것

이 특징이다. 문서의 종류나 작성자의 특성과 같은 문서 레벨의 태깅이 가능하고, 절(Clause), 구(Phrase)와 같은 의미론(Semantic-pragmatic)적이나 구문적(Syntactic) 레벨의 태깅도 가능하다. 또한 사용자는 각 태깅 계층마다 해당 계층의 태그의 계층 구조를 GUI 도구(Graphic Tool)를 사용하여 정의할 수 있다. 그리고 문서 관리 기능도 지원하고 있다.

2.2 The Mate Workbench[2]

이 도구는 MATE 프로젝트에 의해서 개발된 대화 시스템(Spoken Dialogue System)에서 요구되는 자연어처리의 여러 분야의 학습 말뭉치들을 구축하기 위한 여러 개의 말뭉치 구축 도구로 구성되어 있다. 프로젝트 단위로 말뭉치들을 관리할 수 있는 기능도 지원하고 있으며, XML 형식으로 데이터를 저장한다. 대화 시스템에서 필요한 형태소 구조 레벨(Morpho-syntactic Level)의 태깅 도구, 음성 관련 도구 등을 지원한다.

2.3 POSBioTM/W[3]

이 도구는 단백질이나 유전자와 같은 생물할 분야의 개체명 인식 말뭉치 구축을 위한 전용도구이다. 개체명 태깅은 생물학 분야의 말뭉치 구축에 많이 사용되는 말뭉치인 “GPCR", "GENIA", "GENE"에서 사용되는 개체명 태그 집합 중에 하나를 선택하여 작업을 수행할 수 있다.

2.4 Opinion Annotation Tool(OAT)[4]

이 도구는 Opinion Mining 분야의 말뭉치를 구축하기 위한 전용도구이다. 단어나 부분 문장, 전체 문장 레벨에서 긍정/부정/중립의 견해(Opinion)을 부착할 수 있으며, 문장의 주제를 입력할 수 있는 기능도 제공하고 있다. 영어/중국어/일본어 인터페이스를 지원한다.

2.5 Swedish-Turkish Parallel Corpus Tool[5]

이 도구는 스위스어와 터키어 병렬 말뭉치(Parallel Corpus) 구축 전용도구이다. 내부적으로 어휘분석기와 문장 분리기, 품사 태거를 사용하여 서로 다른 언어로 기술된 두 문서를 언어학적으로 분석하여 문장 정렬(Sentence Alignment)을 수행하고, 단어 정렬(Word Alignment)을 수행하여 병렬 말뭉치를 자동으로 구축한다.

공개된 전용도구들의 공통적인 특징은 모두 GUI 환경의 인터페이스를 제공하고 있다는 점이다. 그리고 분야에 맞는 특별한 기능들도 제공하고 있다. Kane 또한 GUI 환경을 제공하고 있으며, 실제로 문서 편집기를 사용한 말뭉치 구축 작업의 어려움으로부터 개발되었기 때문에 직관적인 인터페이스로 쉽게 말뭉치를 구축할 수 있도록 하는 점이 특징이다. 또한 학습 기능을 통해 작업이 진행될수록 작업 효율이 증대될 수 있도록 하고 있다.

3. Kane

본 장에서는 Kane의 인터페이스와 Kane을 통해 수행할 수 있는 개체명 태깅 작업과 관계 태깅 작업에 대하여 자세히 설명한다.

3.1 인터페이스

그림 1과 그림 2는 Kane에서 제공하는 인터페이스 화면이다.

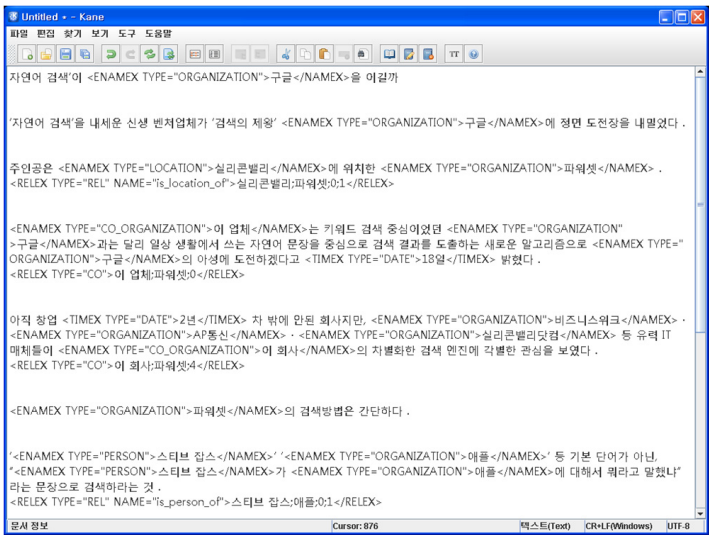

그림 1 Kane의 인터페이스(텍스트 모드)

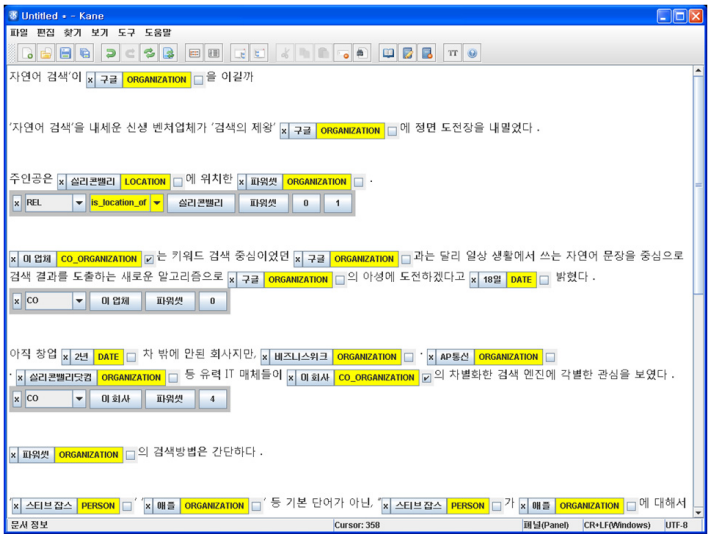

그림 2 Kane의 인터페이스(패널 모드)

우선 Kane의 전체적인 인터페이스의 구성은 일반적인 문서 편집기와 크게 다르지 않다는 것을 알 수 있다. 그리고 텍스트 모드와 패널 모드, 두 가지의 인터페이스를 제공한다. 텍스트 모드에서는 문서 편집기에서 제공되는 모든 기능을 사용할 수 있으며, 패널 모드에서 개체명 및 관계 태깅 작업을 수행할 수 있다. 패널 모드에서의 작업은 주로 마우스를 사용하여 이루어지는데, 드래그와

클릭만으로도 말뭉치 구축 작업을 할 수 있다. 또한 마우스 대신에 타블렛을 사용하면 좀 더 직관적인 인터페이스로 말뭉치 구축 작업을 수행할 수 있다. 이처럼 Kane의 친숙하고 직관적인 인터페이스는 매뉴얼 없이 처음 사용하더라도 쉽게 사용할 수 있도록 해준다. Kane에서는 비영어권 프로그램에서 문제가 되는 문자 코드 문제에 대응하기 위해 자동으로 문자 코드를 인식하여 문서를 읽어 들이는 기능을 제공한다. 자동으로 인식된 문자 코드는 프로그램 아래쪽에 작업 표시줄의 제일 오른쪽에 표시해준다. 그리고 윈도우/리눅스 간의 줄바꿈 문자가 달라 문제가 되는 경우도 있는데, 사용자가 저장할 줄바꿈 문자를 선택할 수 있도록 하고 있으며, 현재 사용 중인 줄바꿈 문자는 문자 코드 왼쪽에 표시해준다.

3.2 개체명 태깅

개체명(Named Entity)은 문서에서 의미 있는 대상을 표현하는데 사용된 단어를 의미한다. 여기서 말하는 의미 있는 대상이란 개체명 인식(Named Entity Recognition) 시스템에서 자동으로 인식을 하고자하는 대상이며, 반드시 고유 명사로만 한정되지는 않는다. 거의 모든 개체명 인식에서 공통적으로 사용하는 개체명은 인명(Person), 지명(Location), 조직명(Organization), 날짜(Date), 시간(Time), 수량표현(Quantity) 등이 있다. 일반적으로 개체명 태깅 작업은 개체명 주위에 XML 형식과 유사하게 태그를 씌우는 방식으로 진행된다. Kane에서는 그림 3의 위쪽과 같은 MUC(Message Understanding Conference)에서 정의한 “ENAMEX” 태그[6]를 지원하며, 태그는 사용자의 필요에 따라 변경하여 사용하는 것도 가능하다. 그림 3에서 개체명에 해당되는 “창원대학교”는 "ORGANIZATION"라는 개체명 태그를 부착하여 조직임을 나타낸다.

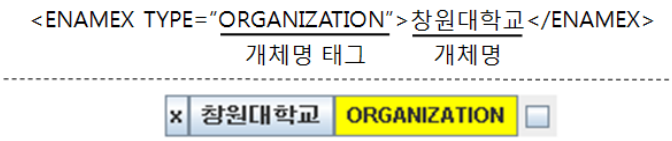

그림 3 Kane에서의 개체명 태깅

그림 3의 패널 모드에서 개체명 태깅을 수행하는 패널을 “개체명 패널”이라고 부르는데, 세 개의 버튼과 하나의 체크 박스로 구성되어 있다. 세 개의 버튼은 왼쪽부터 각각 패널을 삭제하는 버튼, 개체명을 보여주는 버튼, 개체명 태그를 선택할 수 있는 버튼이다. 또한 체크 박스는 상호 참조 해결(Reference Resolution)을 위해 사용되어지며, 체크되면 개체명 태그 앞쪽에 대명사임을 나타내는 접두사 “CO_"가 덧붙여진다.

그림 4는 Kane에서 개체명 태깅 작업을 진행하는 흐름을 나타낸 그림이다. 색이 칠해진 과정이 사용자가 수행하는 작업이며, 그 외에는 Kane이 시스템적으로 처리하는 작업이다. 사용자가 수행할 수 있는 작업은 다음과 같이 크게 3가지로 나뉜다. 첫 번째 작업은 문서에서 개체명에 해당되는 문자열을 드래그하여 개체명 패널을 생성하는 작업이다. 이때 만약 로컬 사전이나 글로벌 사전에 해당 개체명이 태깅된 적이 있다면 개체명 태그를 자동으로 부착하고, 처음 태깅된 개체명이라면 "NONE" 태그를 부착하며, 사용자에게 태깅이 되지 않았음을 표시해주기 위해 개체명 태그 버튼의 바탕색이 붉은색으로 바뀐다. 두 번째 작업은 개체명 태그를 수정/부착하는 작업이다. 개체명 태그를 부착하는 방법은 개체명 패널의 개체명 태그 버튼을 마우스로 클릭하면, 그림 5와 같은 팝업 메뉴가 나타나고, 원하는 태그를 선택하면 된다.

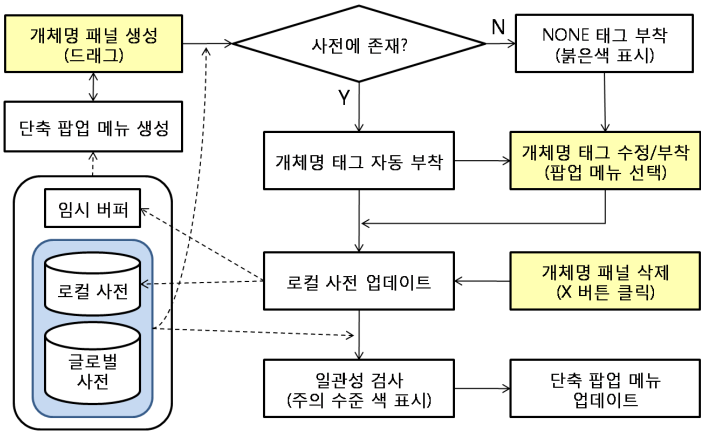

그림 4 Kane에서 개체명 태깅 작업의 흐름

그림 5의 왼쪽 팝업 메뉴가 마우스 왼쪽 클릭을 통해 호출할 수 있는 기본 팝업 메뉴이며, 모든 개체명 태그를 선택할 수 있다. 그리고 왼쪽 팝업 메뉴는 마우스 오른쪽 클릭을 통해 호출할 수 있는 “단축 팝업 메뉴”라고 부르는데, 크게 세 부분으로 나눌 수 있다. 위쪽부터 로컬 사전, 글로벌 사전, 임시 버퍼에 저장되어 있는 개체명 태그가 들어간다. 로컬 사전에는 현재 작업 중인 문서에서 태깅된 개체명이 학습되고, 글로벌 사전에는 기존에 작업한 문서에서 개체명을 학습하여 사용하며, 임시 버퍼에는 직전에 태깅된 개체명이 저장된다.

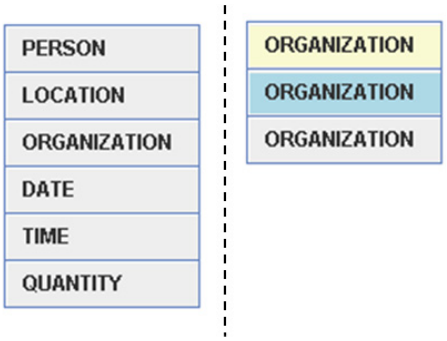

그림 5 Kane의 개체명 태깅 메뉴

개체명 태그 부착이 끝나면 로컬 사전에 해당 개체명이 학습되고, 일관성 검사를 수행한다. 일관성 검사의 결과를 버튼의 색깔을 통해 사용자에게 가시적으로 표시해준다. 개체명 버튼이 하늘색으로 바뀌면 글로벌 사전에 해당 개체명에 대한 태그가 충돌한다는 것을 의미하고,

개체명 태그 버튼이 주황색으로 바뀌면 로컬 사전에서 해당 개체명에 대한 태그가 충돌한다는 것을 의미한다. 여기서 충돌한다는 의미는 동일한 개체명이 개체명 태그가 둘 이상 학습되어 있다는 것을 의미한다.

마지막으로 세 번째 작업은 개체명 패널을 삭제하는 작업이다. 개체명 패널의 삭제 버튼(X 버튼)을 클릭하면 삭제를 할 수 있으며, 삭제의 결과로 최초 드래그를 했던 개체명만 남겨진다.

3.2 관계 태깅

관계 태깅은 태깅된 개체명들 사이의 연관관계를 문서상에 표시하는 것을 말한다. Kane에서 지원하는 관계의 종류는 총 네 가지이며, 상호 참조 해결을 위한 관계도 존재한다. 각각의 관계 종류에 대한 정의는 표 1과 같으며, 가장 기본이 되는 관계인 REL 관계 태깅 형식과 관계 패널의 형태는 그림 6과 같다.

표 1 Kane에서 지원하는 관계의 종류

관계 종류	정의
REL	한 문장 내 두 개체명 간의 연관 관계
CO	대명사와 실제 대상 간의 참조 관계
CO_REL	REL 관계와 동일하나 두 개체명 중 하나가 대명사인 경우 사용
NAME_REF	한 문서 내 동일한 대상을 지칭하는 개체명을 표현하는 관계

관계 패널을 구성하는 콤포넌트(Component)는 삭제 버튼과 관계 종류 선택 콤보 박스(Combo Box)를 제외한 나머지는 관계 종류에 따라 달라진다. REL/CO_REL 관계 패널은 그림 6과 같이 관계명 선택 콤보 박스와 주어/목적어 버튼, 주어/목적어 위치 버튼으로 구성된다. 관계명은 두 개체명 사이의 관계를 표현하기 위해 사용되며, 예제의 "located" 관계는 창원대학교가 창원시에 위치한다는 것을 의미한다. 그리고 위치 버튼은 문장에서 개체명의 인덱스 번호인데, 왼쪽부터 카운팅되며, 0부터 시작된다.

<RELEX TYPE="REL" NAME="located">창원대학교;창원시</RELEX>
관계종류 관계명 주어 목적어

그림 6 Kane에서의 관계 태깅

CO 관계는 "창원대학교"를 "그 학교"와 같이 일종의 대명사처럼 사용한 경우, 상호 참조 관계를 표현하기 위해 사용된다. 관계명이 따로 필요하지 않으므로, 관계명 콤보박스는 제거되고, 주어 자리에 대명사가 들어가고, 목적어 자리에 실제 대상에 해당되는 개체명이 들어간다. 또한 이 관계는 한 문장을 넘어서 관계가 성립되므로, 위치 버튼 중 대명사의 위치 버튼(REL 관계의 주어 위치 버튼)만 사용된다. NAME_REF 관계는 한 문서 전체에서 대명사는 제외하고 동일한 대상을 지칭하는 개체명간의 관계를 나타내기 위해 사용된다. CO 관계와 마찬가지로 관계명 콤보 박스는 사용되지 않으며, 위치 버튼도 사용되지 않는다. 그리고 주어와 목적어의 구분이 없으므로 원하는 대로 개체명 버튼을 계속 추가할 수 있다. 모든 종류의 관계가 태깅된 예제는 그림 7을 통해 살펴볼 수 있다.

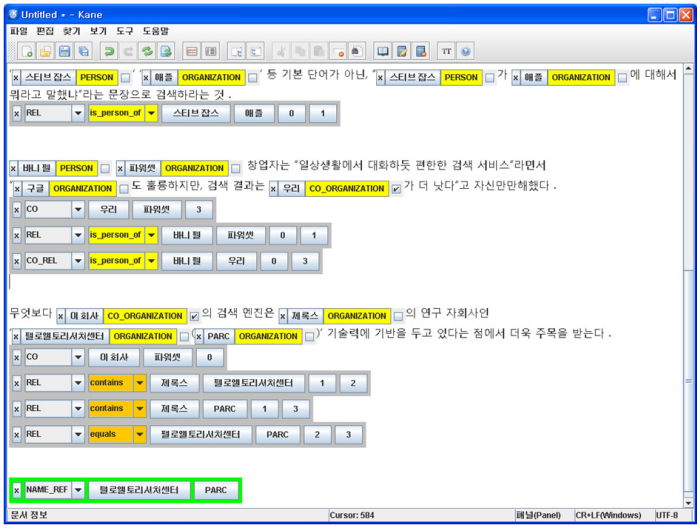

그림 7 Kane에서 관계태깅 작업 수행 화면

관계 태깅 작업의 흐름은 그림 8과 같다.

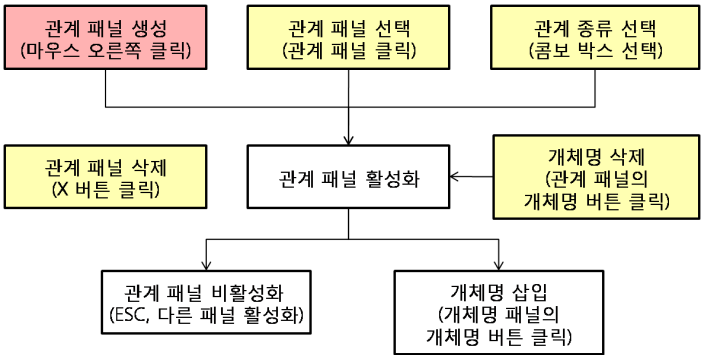

그림 8 Kane에서 개체명 태깅 작업의 흐름

먼저 관계가 성립되는 문장의 마지막이나 다음 줄에서 마우스 오른쪽을 클릭하면 관계 패널이 생성된다. 도구상자에 관계 패널 추가 버튼이나 단축키를 사용해서도 생성이 가능하다. 관계 패널을 선택하면 그림 7처럼 해당 관계 패널의 테두리가 녹색으로 변하는데, 이 상태를 관계 패널이 활성화된 상태라고 한다. 관계 패널에 개체명을 넣기 위해서는 활성화된 관계 패널이 있어야하며, 동시에 하나의 관계 패널만이 활성화 될 수 있다. 활성화를 시키려면 원하는 관계 패널을 클릭하면 되고, 활성화를 해제하려면 ESC키를 누르면 되며 개체명 패널을 클릭한다고 활성화 상태가 해제되지는 않는다. 관계 패

널에 개체명을 삽입하려면 원하는 개체명 패널의 개체명 버튼을 클릭하면 된다. 또한 관계 패널에 추가된 개체명을 제거하려면 해당 관계 패널의 개체명 버튼을 클릭하면 된다. 관계 패널을 삭제하려면 개체명 패널과 마찬가지로 삭제 버튼을 클릭하면 된다. 관계 패널에 개체명 버튼이 비어있으면 해당 개체명 버튼의 배경색을 빨간색으로 표시해주며, 관계 종류에 따라 잘못 넣어진 개체명이 있어도 해당 개체명 버튼의 배경색을 빨간색으로 표시해준다.

관계 태깅에서는 관계명을 선택하는 작업에서 문제가 발생하는데, 이 문제를 Kane에서는 규칙을 통해 처리하고 있다. 규칙에서 충돌이 발생하는 관계명일 경우 관계명 콤보 박스의 배경색을 주황색으로 표시해주며, 충돌이 없을 때는 노란색으로 표시해준다. 또한 성립할 수 없는 관계라면 관계명 콤보 박스가 빈 상태로 표시된다. 관계명에 대한 처리는 학습 기능은 구현하여 개선해야할 필요성이 있다.

4. 결 론

본 논문에서는 한국어 개체명 인식 및 관계 추출 시스템을 위한 학습 말뭉치를 구축하는 전용 도구인 Kane을 제안하고 설명하였다. Kane의 특징을 정리해보면 다음과 같다.

- 직관적인 인터페이스를 통한 기계적인 작업을 줄이고, 오타 문제 해결
- 학습을 통한 개체명 태그 자동 부착
- 자동 일관성 검사 결과를 사용자에게 가시적으로 표시
- 다양한 문자 코드에 대응 가능(다국어 지원 가능)

실험 환경을 구축하여 실험을 진행하지는 않았으나, 실제로 말뭉치 구축 작업에 Kane을 사용하고 있다. 그 결과, 문서 편집기를 사용한 이전 작업 때보다 기계적으로 반복되는 작업들이 많이 줄어 작업 속도가 향상되었으며, 오타 문제가 완전히 해결된 것을 확인할 수 있었다. 다만 사전을 공유하여 제한적으로 공동 작업환경을 구축했으나 시스템적으로 작업을 관리하는 기능이 미흡하여 일관성이 깨지는 문제가 완전히 해결되지는 못하였다. 향후 이 기능을 추가하여 일관성이 깨지는 문제에 대응할 수 있도록 할 것이다. 또한 학습 기능을 확장하고, 내부적으로 개체명 인식기와 관계 추출기를 사용하여 반자동으로 말뭉치를 구축할 수 있도록 할 계획이다. 기본적으로 말뭉치 구축 작업은 다국어로 수행이 가능하지만, 사용자 인터페이스는 한국어로 고정되어 있는데, 이 또한 다국어를 지원할 수 있도록 할 것이다. 끝으로 현재는 고정되어 있는 말뭉치의 형태도 사용자가 직접 정의할 수 있도록 하여, Kane의 확장성을 높이고자 한다.

참고문헌

[1] Michael O'Donnell, The UAM CorpusTool: Software for corpus annotation and exploration, Proceedings of the XXVI Congreso de AESLA, Almera, Spain, 3-5 April 2008

[2] Multilevel Annotation, Tools Engineering, http://mate.nis.sdu.dk/

[3] POStech Biological Text-Mining Workbench, http://isoft.postech.ac.kr/Research/BioNER/POSBIOTM/NER/main.html

[4] Tools and Annoated Corpora for Opinionated Tasks, http://nlg18.csie.ntu.edu.tw:8080/opinion/

[5] Beata B. Megyesi and Bengt Dahlqvist, The Swedish-Turkish Parallel Corpus and Tools for its Computation Linguistics NODALIDA-2007, pp. 136-143, 2007

[6] Message Understanding Conference, Named Entity Task Definition, MUC-6, 1996, http://cs.nyu.edu/cs/faculty/grishman/NEtask20.book_1.html