COAT: 시맨틱 어노테이션 말뭉치 구축 지원 도구

최동현, 김은경, 고은비, 최기선[O]
Semantic Web Research Center, KAIST
{cdh4696, kekeeo, eunbi, kschoi}@world.kaist.ac.kr

COAT: Manual Semantic Annotation Support Toolkit

DongHyun Choi, Eun-Kyung Kim, Eun-Bi Go, Key-Sun Choi[O]
Semantic Web Research Center, KAIST

요 약

수동 어노테이션을 통한 말뭉치 구축 작업은 많은 시간과 노력이 필요한 작업이지만, 자동화된 정보 추출 도구의 훈련 및 실험, 평가를 위해서는 꼭 필요한 작업이기도 하다. 본 논문에서는, 수동 시맨틱 어노테이션을 통한 말뭉치 구축 작업을 지원하는 수동 시맨틱 어노테이션 지원 도구 COAT를 소개한다. COAT는 각 어노테이터의 작업 효율을 높이기 위하여 GUI 기반 인터페이스를 제공하고, 작업의 대부분을 단축키만 이용하여 수행 가능하도록 설계되었다. 또한 최종 결과로 얻어지는 데이터의 신뢰성을 높이기 위하여, 최소 두 명 이상의 어노테이터가 같은 문서에 대하여 작업하면 고참 어노테이터가 각 결과물들을 통합하는 컨쥬게이션 도구를 구축하였으며, 각 어노테이터들의 작업 및 데이터들을 관리 감독하기 위한 관리자 도구를 개발하였다. 본 도구를 직접 사용하여 어노테이션 작업을 수행한 결과, 본 도구를 사용하지 않고 작업을 수행할 때와 비교하여 약 87%의 비용 절감 효과를 얻을 수 있었다.

주제어: 시맨틱 어노테이션, 정보 추출, 말뭉치 구축

1. 서론

DBpedia[1]가 등장한 이후, Linked Data의 형태로 표현되는 Data Web은 그 규모적인 측면에서 빠르게 성장하고 있으나, 아직 상당한 양의 웹 데이터는 비구조적인 (unstructured) 형태로 남아 있다. 메릴 린치의 보고서에서는 그 양을 전체의 약 80 %로 추산[2]하고 있으며, 이 수치는 Linked Data 이후에도 그다지 변화하지 않은 것으로 생각된다[3].

비구조적인 데이터에서 정보를 얻어내는 정보 추출 시스템의 경우, 대개의 경우 학습 및 평가를 위해서 얻고자 하는 정보가 수동으로 어노테이션된 말뭉치가 필요하다. 그러나 수동으로 말뭉치를 구축하는 작업에는 세 가지 난점이 존재한다. 첫째로, 수동 어노테이션 작업은 많은 시간과 도메인에 대한 지식을 필요로 하는 작업이다. 이 때문에, 대개 수동 어노테이션 작업은 대상 도메인에 대하여 잘 훈련된 전문가를 필요로 한다.

둘째로, 얻어진 데이터에 대한 신뢰도의 문제가 있다. 악의적인 어노테이터가 고의로 잘못된 어노테이션을 수행하거나, 또는 어노테이터가 부주의하게 작업을 수행하는 경우를 제외하더라도, 수동 어노테이션 작업은 작업을 수행하는 어노테이터 개개인의 배경 지식 및 관점에 따라 어노테이션된 결과물에 어노테이터 개인의 성향이 포함된다. 이렇게 구축된 말뭉치를 이용하여 훈련된 시스템의 경우, 주어진 입력에 대하여 일반적으로 기대되는 출력이 아닌 다른 결과를 도출하도록 훈련되게 된다.

셋째는 어노테이션 작업 관리의 어려움이다. 소규모 말뭉치에 대한 어노테이션 작업을 수행할 경우에는 큰 어려움을 겪지 않을 수도 있으나, 대상 말뭉치의 규모가 커지고, 이에 따라 작업을 수행하는 어노테이터의 수도 늘어나게 되면, 이들 각각에게 작업량을 배정하고 작업 결과물들을 관리 및 정리하는 작업도 점점 힘들어지게 된다.

본 논문에서 소개된 COAT는 이러한 난점들을 해결하기 위하여, 시맨틱 어노테이션 작업을 대상으로 하여 구축된 수동 시맨틱 어노테이션 지원 도구이다. 본 논문에서 말하는 시맨틱 어노테이션은, 주어진 텍스트에서 의미 있는 단어의 부분을 찾아내고, 그 단어가 보유하는 의미에 대하여 표기하며, 이 때 찾아내어진 단어들 간에 무슨 관계가 있는지 찾아내어 표기하는 작업을 의미한다. 이 때, 찾고자 하는 단어의 종류 및 단어간의 관계는 사용자가 얻고자 하는 최종 시스템에 따라 좌우된다. 본 도구를 이용하여 얻어진 말뭉치는 개체명 인식 등의 다양한 형태의 정보 추출 시스템 훈련 및 평가에 이용될 수 있다.

그림 1은 시맨틱 어노테이션의 예를 나타낸다. 그림 1에서는 “레이텍” 및 "TeX"가 소프트웨어의 일종이며, “도널드 크누스”와 “레슬리 램포트”는 사람이며, “1984년”은 시간 표현이라는 정보가 표기되어 있다. 또한, “레슬리 램포트”가 “레이텍”을, “도널드 크누스”가 “TeX”를 만들었으며, “레이텍”은 “1984년”에 만들어졌다는 정보도 표기되어 있다. 이후 본 논문에서는 “레이텍” 또는 “도널드 크누스”와 같이, 의미를 부여받은 단어들을 “텀 (Term)”으로 지칭한다.

COAT는 데이터의 신뢰도에 관한 문제를 해결하기 위하여 컨쥬게이션의 개념을 도입하였다. 컨쥬게이션이란, 서로 다른 두 어노테이터가 동일 대상에 대하여 작업한 결과물을 제 3의 고참 어노테이터가 통합하여 최종 결과물을 만들어내는 것을 의미한다. 이 작업을 수행하게 되면, 최종 어노테이션 결과물에는 고참 어노테이터를 포함한 최소 2인 이상의 어노테이터가 동의한 내용만 포함되게 되므로 데이터의 신뢰도가 크게 향상된다. 또한, 컨

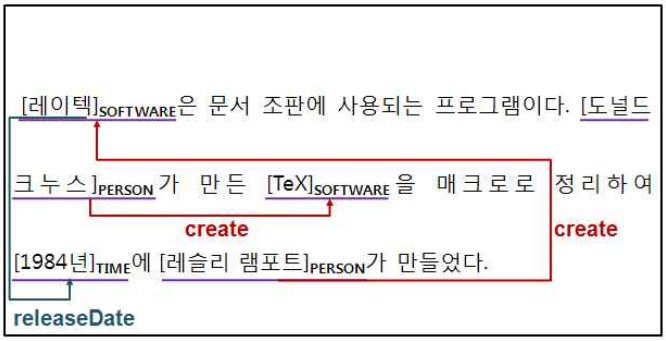

그림 1 시맨틱 어노테이션 작업의 예

쥬게이션의 개념을 도입함으로써 어노테이션 작업에는 비교적 초보적인 어노테이터를 투입하는 것이 가능해져, 추가적인 비용 절감 효과를 기대할 수 있다. 다음 그림 2는 어노테이션-컨쥬게이션 작업의 기본 개념 및 전체 작업 흐름도를 보여 준다. 컨쥬게이션 작업에 대한 좀 더 자세한 설명은 3.2장에 서술되어 있다.

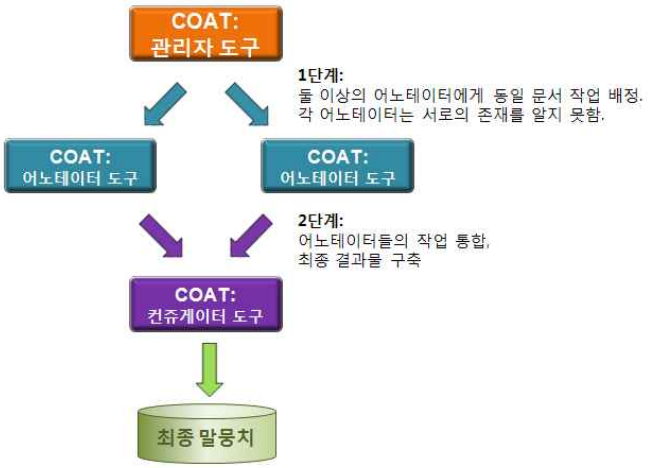

그림 2 COAT의 전체적인 워크플로우

2장에서는 관련 연구에 대하여 간략히 소개한다. 3장에서는 시스템에 대하여 좀 더 자세히 소개되고, 4장에서는 실제 시스템 사용 경험이 간략히 소개된다. 마지막으로 5장에서는 결론에 대하여 언급한다.

2. 관련연구

수동 시맨틱 어노테이션 작업을 지원하기 위하여 기존에도 몇몇 도구들이 구축되었다. KIM annotation platform[4]은 개체명 인식을 위한 어노테이션 작업에 특화된 기능을 제공하였다. XConc suite[5] 어노테이션 지원 도구는 GENIA 말뭉치에 event annotation 정보를 표현하기 위하여 구축되었으며, S-CREAM[6]은 반자동적인 요소를 어노테이션 작업에 추가하고자 시도하였다. 국내에서는 웹 페이지를 기반으로 한 OntoCS[7] 도구가 개발되어 있다.

위에서 보는 바와 같이, 현재에도 다양한 어노테이션 도구가 존재하나, 위에 소개된 도구들은 모두 얻어진 데이터에 대한 신뢰도 문제에 비교적 취약하다. 실제 이들 도구들에 의해 얻어진 데이터를 사용하기 위해서는, 서로 다른 어노테이터들간의 작업 일치도를 파악하고, 일치도가 일정 수준 이상일 경우에만 데이터를 사용 가능하다. 저자가 아는 한, COAT 도구는 기 존재하던 컨쥬게이션의 개념을 처음으로 GUI 기반 인터페이스 상에서 지원해 주는 도구이다.

3. 도구 소개

본 장에서는 COAT의 각 도구에 대하여 좀 더 자세히 소개한다. COAT는 어노테이터 도구, 컨쥬게이터 도구와 관리자 도구의 세 가지 도구로 이루어져 있다. 그림 3은 전체 시스템 구조를 나타낸다. 모든 말뭉치 관련 정보들은 공용 데이터베이스에 저장되어 있고, COAT 어노테이터 도구 및 COAT 컨쥬게이터 도구는 COAT 관리자 도구에 의해 허용된 부분의 데이터에 대해서만 접근 및 수정이 허가되며, 서로 다른 어노테이터 및 컨쥬게이터들의 데이터 접근 상황 및 작업 결과물에 대하여 접근할 수 없다. COAT 관리자 도구는 각 어노테이터 및 컨쥬게이터들에 대하여 데이터의 접근 권한 관리 및 작업 상태, 진행 상황 등을 감시할 수 있으며, 작업된 결과물을 XML형태로 파일 시스템에 저장할 수 있으나, 어노테이터 및 컨쥬게이터들이 수행한 작업 자체에 대한 수정 및 변경은 엄격히 금지되어 있다.

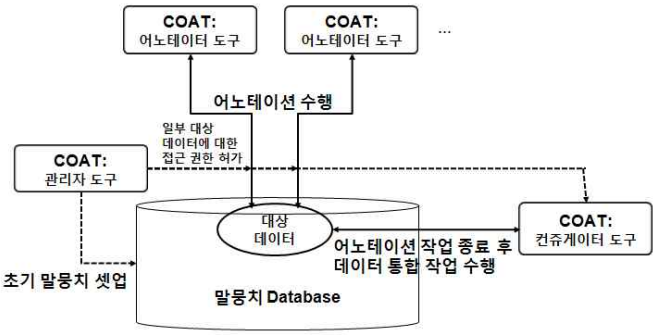

그림 3 개략적인 시스템 구조

3.1 COAT 어노테이터 도구

COAT 어노테이터 도구는 실제 어노테이션 작업, 즉 의미 있는 단위의 단어들을 찾고 의미 태그를 부여하며, 의미 태그가 부여된 단어들간의 관계를 찾아내고, 찾아낸 관계에 태그를 부착하는 작업을 수행하기 위한 도구이다. 그림 4는 COAT 어노테이터의 예시 스크린샷이다. COAT 시스템은 한국어와 영어 모두에 대하여 사용가능하다.

COAT 어노테이터는 사용자의 효율성을 증대시키기 위하여, 주요한 기능들에 대해서는 키보드 단축키를 정의하여 마우스 커서의 움직임을 최소화하였다. 예를 들어, 단축키 'a'는 의미 있는 단어의 부분을 추가하며, 'd' 또는 'del' 키는 단어 또는 단어간의 관계를 삭제한다. 도구에서 채택된 어노테이션의 정책상, 하나의 텀 내

부에 다른 텀이 존재하는 (nested term) 것을 허용하지 않는다.

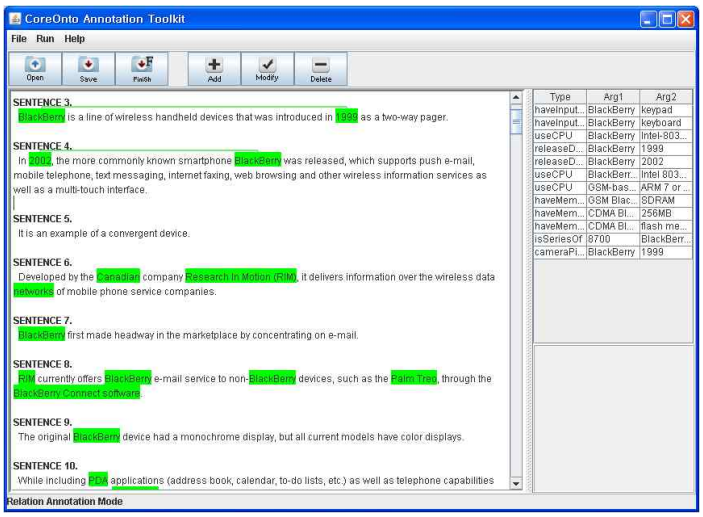

그림 4 COAT 어노테이터: 스크린샷

COAT 어노테이터 도구는 사용자에게 매우 편리한 GUI 인터페이스를 제공할 뿐만 아니라, 모든 어노테이션 결과물은 이후 3.2장에서 서술될 COAT 컨쥬게이터 도구를 사용한 작업을 통하여 추가로 정제되게 된다. 따라서, 기존 작업에서는 어노테이터로서 고도로 훈련받은 고급 전문가를 필요로 하였으나 COAT를 사용하게 되면 그다지 많이 훈련받지 않은 초보적인 어노테이터로도 작업을 수행할 수 있게 된다.

3.2 COAT 컨쥬게이터 도구

COAT 컨쥬게이터 도구는, 두 명 이상의 어노테이터가 동일 문서에 대하여 수행한 작업 결과를 통합하는 작업을 지원하기 위하여 개발된 도구이다. 이 과정은 한 명의 악의적인 어노테이터에 의하여 시스템 전체가 붕괴하는 일을 방지하고, 어노테이터의 부주의로 인한 실수가 최종 말뭉치에 미치는 영향도 최소화하게 된다. 또한, 각 어노테이터가 가지고 있는 개인적인 성향 및 편견이 말뭉치에 미치는 영향도 최소화할 수 있다. 예를 들어, 영어 문장 “Celeron processors can run ...” 과 같은 문장의 경우, 어떤 어노테이터들은 "[Celeron]$_{CPU}$ processors can run ...”와 같이 작업하지만, 일부 어노테이터들은 “[Celeron processors]$_{CPU}$ can run ...”와 같이 작업한다. 이러한 결과의 차이는 명백히 악의나 실수의 결과물이라기보다는 각 어노테이터들 간의 생각의 차이에서 기인되는 것으로 보인다.

따라서, 위와 같이 하나의 동일한 부분에 대해서 어노테이터들이 서로 다른 방식으로 어노테이션 작업을 수행하였을 경우, 컨쥬게이터는 여러 어노테이션 작업들 중에서 하나를 최종 결과물로서 선택하게 된다. 따라서, 최종 어노테이션 결과물에는 반드시 한 명 이상의 어노테이터와 컨쥬게이터가 동의한 내용만 포함되게 된다. 예를 들어, 위의 경우 컨쥬게이터가 “[Celeron processors]$_{CPU}$ can run ...”를 선택 하였을 경우, 최종 결과 말뭉치에서는 "[Celeron]$_{CPU}$ processors can run ...”는 버려지고 “[Celeron processors]$_{CPU}$ can run ...” 만 포함되게 된다.

컨쥬게이터는 최종 결과물을 결정하는 역할을 하기 때문에, 고도로 훈련받은 전문 인력이 필요하다. 그러나 컨쥬게이션 작업은 몇 가지 제시된 어노테이션 옵션들 중에서 하나를 선택하는 일이기 때문에 매우 빠른 속도로 수행할 수 있고, 따라서 전문 인력을 사용하는 데 드는 작업 시간 및 비용을 크게 절약할 수 있다. 그림 5는 COAT 컨쥬게이터의 스크린샷을 보여 준다.

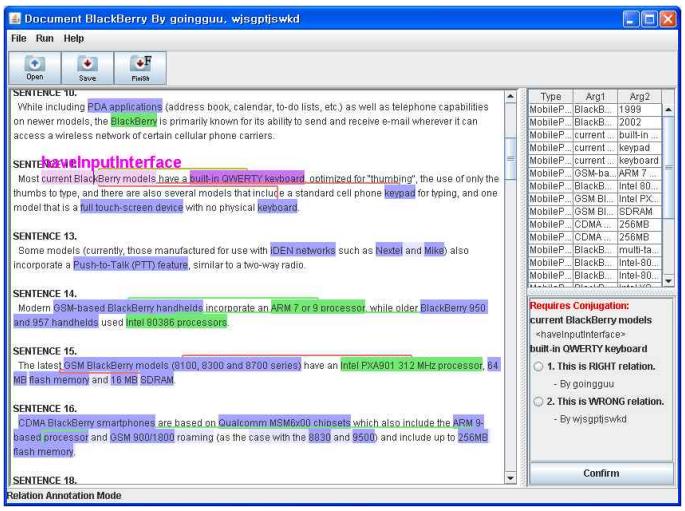

그림 5 COAT 컨쥬게이터: 스크린샷

3.3 COAT 관리자 도구

COAT 관리자 도구는 초기 말뭉치 설정 및 말뭉치에 대한 수정 권한 등을 부여한다. COAT 관리자 도구를 이용하여 여러 명의 어노테이터가 동시에 작업할 경우, 작업할 문서의 분량 및 범위를 효과적으로 조절할 수 있다. 또한 각 어노테이터가 작업한 분량 및 내용을 바로 확인 가능하며, 작업한 내용을 XML 파일 형태로 파일 시스템에 저장할 수도 있다. 다만, COAT 관리자 도구로는 실제 어노테이션 작업 또는 컨쥬게이션 작업을 수행할 수 없다.

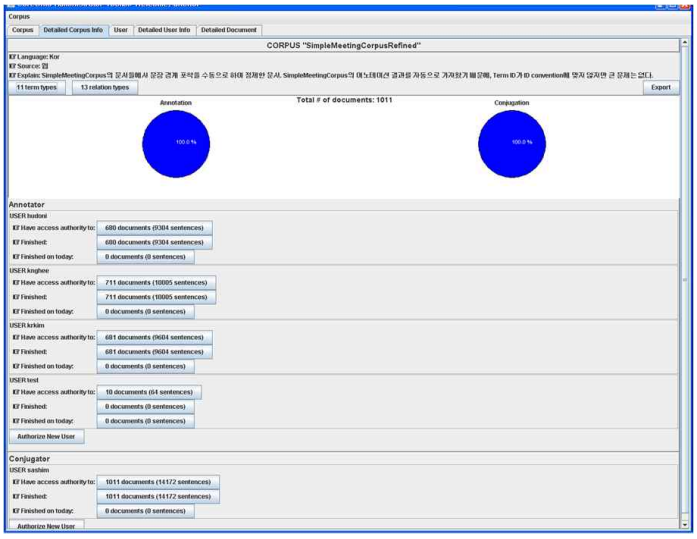

그림 6 COAT 관리자 도구: 스크린샷

그림 6은 COAT 관리자 도구의 스크린샷을 나타낸다. 각 어노테이터 및 컨쥬게이터별로 진행 상황, 말뭉치 전체에 대한 진행 상황을 확인할 수 있으며, 텀 및 관계의 타입을 정의할 수 있고, 각 어노테이터 및 컨쥬게이터에게 새로운 작업 대상의 배정 등도 가능하다.

4. 실제 사용 예: IT 분야 정보 추출 작업을 위한 말뭉치 구축

COAT 도구를 이용하여, Wikipedia에서 수집된 IT 분야 문서들에 대한 어노테이션 작업을 수행하였다. IT 분야 문서들 중에서도, 휴대폰, 운영체제, 알고리즘, 중앙처리장치, 컴퓨터 bus, 프로그래밍 언어, 소프트웨어의 7개 분야의 문서들에 대하여 작업을 수행하였다.

COAT 사용 이전에는 두 명의 고도로 훈련받은 어노테이터가 IT 분야 문서들에 대하여 따로 작업하였다. 즉, 각 문서들은 단 한 명의 어노테이터에 의하여 작업되고, 그 결과물이 최종 말뭉치에 포함되었다. 이런 방식으로, 두 명의 full-time 어노테이터가 2달 동안 총 3,000개의 문장에 대하여 어노테이션을 수행하였다. 이에 든 총 비용은 약 703만 6천원이었으며, 문장 개당 단가는 2,345원을 기록하였다.

반면에, COAT를 사용한 이후에는, 어노테이션 작업을 위하여 10명의 대학생 아르바이트를 고용하여 작업하였다. 각 아르바이트생들은 작업 시작 전 어노테이션 방법에 대하여 약 1시간의 교육을 받았고, 또한 작업 첫 주에는 매일 1시간씩 나와서 도구 사용법에 대한 교육을 받았다.

어노테이션 작업을 위하여 각 2명의 어노테이터들끼리 그룹을 만들고 동일 문서를 배당하여 작업케 하였고, 어노테이터들에게는 그룹의 존재나, 같은 문서를 작업하는 사람이 있다는 정보를 알리지 않았다. COAT 이전에 어노테이터를 수행하던 어노테이터들 중 하나는 고참 어노테이터로 임명되어 컨쥬게이션 작업을 수행하였고, 나머지 한 명은 작업에 참여하지 않았다. 아르바이트생들은 평일에 하루 3시간씩 작업할 것을 요구받았으며, 컨쥬게이션은 full-time으로 수행되었다.

한 달 동안의 작업 결과, 총 52,952 개의 문장에 대하여 어노테이션이 수행 되었다. 각 문장은 2명의 서로 다른 어노테이터들에 대하여 작업되었다. COAT를 사용하지 않았을 경우에는 숙련된 어노테이터도 한 달에 평균 750개의 문장에 대한 어노테이션을 수행하는 데 그쳤으나, COAT를 사용한 이후에는 초보 어노테이터도 한 달에 평균 10,590 개의 문장에 대한 어노테이션 작업을 수행할 수 있었다. 이는 COAT 어노테이터 도구의 편리한 사용자 인터페이스가 어노테이터의 생산성을 비약적으로 향상시킨 것으로 추정된다.

또한, 한 달 동안 총 15,646 개의 문장에 대한 컨쥬게이션 작업이 고참 어노테이터에 의하여 수행되었다. 즉, 컨쥬게이션이 수행된 15,646 개의 문장은 최소 3명의 어노테이터 – 2명의 초보 어노테이터와 1명의 고참 어노테이터 – 에 의하여 내용을 검증받음으로써, COAT를 사용하지 않았을 경우 – 1명의 고참 어노테이터에 의해서만 내용이 구축됨 – 에 비하여 신뢰도가 비약적으로 향상되었다고 생각할 수 있다.

COAT를 사용하여 어노테이션 작업을 수행하는 데 든 총 비용은 아르바이트생 고용 비용 880만 원과 고참 어노테이터의 인건비 220만 원을 합한 총 1,100만 원이다. 컨쥬게이션 작업이 완료된 15,646 개의 문장만을 토대로 계산하였을 경우 개당 단가는 703원으로, COAT를 사용하기 이전의 단가의 3분의 1 수준에도 이르지 못하는 반면 데이터의 신뢰도는 비약적으로 향상되었다. 만약 2명의 어노테이터들에 의하여 어노테이션은 수행되었으나 컨쥬게이션은 수행되지 않은 남은 37,306개의 문장에 대하여 컨쥬게이션 예상 비용을 계산하고, 이를 토대로 52,952 개 문장 전부에 대하여 컨쥬게이션 작업까지 수행하였을 때 드는 비용을 계산하면, 개당 단가는 약 307원까지 내려가게 된다. 즉 약 87 % 에 이르는 수동 어노테이션 비용 절감 효과를 기대할 수 있다. 이러한 비용 절감 효과는 COAT의 편리한 인터페이스로 인한 생산성 향상뿐만 아니라, 컨쥬게이션 개념을 도입함으로써 상당수의 어노테이터들을 단순 아르바이트 학생으로 교체할 수 있었던 것에도 크게 기인한다.

현재 COAT는 웹에 오픈 소스로서 공개되어 있으며, http://sourceforge.net/projects/coatsemantic에서 다운로드 받을 수 있다.

5. 결론

본 논문에서는 수동 시맨틱 어노테이션 도구인 COAT에 대하여 서술 하였다. COAT 도구를 실제 수동 어노테이션 작업에 적용한 결과, 수동 어노테이션의 작업 비용 및 효율이 크게 증가하였다.

본 논문의 결과물은 향후 타 자연언어처리 연구를 위한 말뭉치를 구축할 때, 신뢰도 높은 말뭉치를 빠르게 구축할 수 있도록 도와줌으로써, 자연언어처리 연구에 큰 도움을 줄 수 있을 것으로 기대된다.

감사의 글

이 논문은 2010년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. 2010-0022444)

참고문헌

[1] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak and Z. Ives, DBpedia: A Nucleus for a Web of Open Data, Lecture Notes in Computer Science, Vol. 4825/2007, pp. 722-735, 2007.

[2] C. Shilakes and J. Tylman, Enterpriseinformation portals, Technical report, Meryll Lynch, 1998

[3] http://clarabridge.com/default.aspx?tabid=137&ModuleID=635&ArticleID=551

[4] B. Popov, A. Kiryakov, A. Kirilov, D. Manov, D. Ognyanoff and M. Goranov, KIM-Semantic Annotation Platform, Lecture Notes in Computer Science, Vol. 2870/2003, pp. 834-849, 2003

[5] J. Kim, T. Ohta and J. Tsujii, Corpus annotation for mining biomedical events from literature,

BMC Bioinformatics, Vol. 9(10), 2008
[6] S. Handschuh, S. Staab and F. Ciravegna, S-Cream --- Semi-automatic CREAtion of Metadata, Lecture Notes in Computer Science, Vol. 2473/2002, pp. 165-184, 2002
[7] D. Hwang, I. Lee and J. Jung, ONTOCS: A Web-Based System For Collaborative Ontology Construction, Computing and Informatics, Vol. 28(6), pp. 781-793, 2009