온톨로지 자동 구축을 위한 개념/인스턴스 분류 시스템 개발

배영준[O1] 임지희[1] 옥철영[1] 최호섭[2]
울산대학교 컴퓨터정보통신학과[1], 한국과학기술정보연구원 정보기술개발단 정보시스템개발팀[2]
{young4862[O], arisu80, okcy}@ulsan.ac.kr, hschoe@kisti.re.kr

Development of Concept & Instance Classification System for Automatic Construction of Ontology

Young-Jun Bae[O1] Ji-Hui Im[1] Cheol-Young Ock[1] Ho-Seop Choe[2]
Dept. of Computer Engineering and Information Technology, University of Ulsan[1],
Information System Development Team, Korean Institute of Science and Technology Information[2]

컴퓨터를 똑똑하게 만들기 위해서는 사람의 개념체계를 컴퓨터에게 심어주고 개념들 간의 관계나 속성을 설정해 주며 각 개념에 관한 예를 알려주어야 한다. 이러한 일들을 하기 위해서 사람이 가진 개념체계를 컴퓨터가 인식할 수 있는 형태로 바꾸어 줄 필요가 있는데, 그 형태가 온톨로지로 나타나게 된다. 이런 생각을 기반으로 현재 온톨로지에 대한 연구들이 활발히 이루어지고 있다.

지금까지 온톨로지에 대한 연구는 크게 온톨로지 구축, 온톨로지 인스턴스 생성, 온톨로지를 이용한 추론, 온톨로지를 이용한 응용서비스 등이 있다.

국내에서 구축된 몇 가지 온톨로지는 CoreNet, U-WIN, ETRI 어휘개념망, 세종 의미부류체계, 국가과학기술 기반정보 온톨로지, 스마트메신저 영화정보 온톨로지 등이 있다. 구축된 온톨로지는 서로 분야도 다르고 각 온톨로지마다 개념명이나 관계명, 또는 매핑된 관계, 속성의 정보 등 같은 것도 있을 수 있지만 다른 것들이 많다. 그래서 기구축된 온톨로지 간의 통합 등은 많은 어려움을 야기 시킨다. 그리고 각 온톨로지에 맞게 개발된 시스템 및 응용서비스는 그 온톨로지에 종속적이기 때문에 다른 온톨로지에 그 시스템 및 응용서비스를 재사용하기 어렵다. 그래서 분야별로 개념, 관계, 속성 등이 명확히 정의된 하나의 온톨로지가 필요하고 현재 국가 IT 기반 온톨로지 구축 프로젝트 등을 통해 이를 해결하려고 하고 있다.

온톨로지를 수동으로 구축하면 자동으로 구축하는 것보다 어느 정도 정확성이 향상되겠지만 많은 시간과 인력이 필요하다. 우선 기반이 되는 말뭉치나 사전 및 시소러스 또는 이미 구축된 온톨로지 등의 자원이 필요하며 말뭉치에서 텍스트 정제, 용어 추출 및 검증, 용어의 정의문·대역어·용례 추출, 추출된 용어의 개념·인스턴스 검증을 통한 개념화, 개념 간의 관계·속성 추출, OWL로 변환, 분류체계 구축, 관계 정의 및 설정, 인스턴스 부착, 온톨로지 통합 등 많은 단계의 작업을 거쳐야 한다. 이 모든 공정을 사람이 수행 하려면 많은 인력과 자원이 소모된다. 시간이 지날수록 개념 및 인스턴스가 기하급수적으로 늘어날 것인데 그때마다 새로운 개념 및 인스턴스 구축을 위해 대규모의 인력과 자원을 투입할 수는 없을 것이다. 그리고 새로 생성된 개념이나 인스턴스 및 관계를 일일이 추가하는데도 한계가 있다. 그래서 온톨로지 자동 구축이 필요하고 자동구축을 위한 도구 및 시스템들이 필요하다.

추출한 용어가 인스턴스임에도 불구하고 온톨로지 자동 구축 과정을 통해 개념으로 할당된다면, 온톨로지의 용량 증가·신뢰성 저하 등의 문제가 발생한다. 그래서 보다 정확한 온톨로지 구축을 위해 본 논문에서는 앞에서 살펴본 자동 구축 시스템들 중 하나인 개념과 인스턴스를 분류하는 시스템을 구축하였다.

본 시스템을 구축하기 위해 가장 기본 바탕이 되는 개념과 인스턴스에 대한 정의 및 분류 기준을 국어사전, 백과사전, 영영사전, 위키피디아(www.wikipedia.org), 온톨로지 관련 논문등을 참고하여 설정하

였다. 개념의 정의는 '추상적', '일반적', '보편적', '공통적인 요소', '집합' 이라는 속성들을 기반으로 하고 있고, 인스턴스의 정의는 '구체적', '특정한', '개별적인 요소', '사례, 예'와 같은 속성들을 기반으로 하고 있음을 알 수 있었다. 이런한 특성들은 우리가 일반적으로 사용하는 명사들의 분류에서도 나타나는데 개념이 보통·일반명사를 포함하고, 인스턴스가 고유명사를 포함한다고 볼 수 있다. 보통·일반명사의 특성이 개념과 같이 일반적이고 포괄적인 의미의 특징들을 가지고 있다. 그리고 고유명사의 특성을 보면 특정하고 개별적이고 유일한 특징들을 가지고 있는데 특히 명칭이 많다. 이런 용어들이나 패턴들을 추출하고 분석을 통해 개념 및 인스턴스에 대한 규칙으로 사용할 수 있다.

개념과 인스턴스를 분류하기 위해, 말뭉치 내의 분류할 용어(한글 표제어)와 용어의 대역어(영어 표제어), 그리고 용어의 정의문 또는 용어의 설명(뜻풀이)들이 필요하다. 그리고 세밀한 개념 및 인스턴스 규칙 설정을 위해 기국축된 시소러스나 온톨로지의 개념 집합, 그와 관련된 동의어, 유의어의 자원과 국가명, 지역명, 인명, 기관명, 단체명, 사건명, 책명 등의 특정한 명칭을 나타내는 고유명사들을 확보하였다.

본 논문에서는 규칙을 크게 '개념 설정 규칙'과 '인스턴스 설정 규칙'으로 나누고, 이것을 세부적으로 '한글 표제어 규칙', '영어 표제어 규칙', '뜻풀이 규칙'으로 구분하여 총 6가지 규칙을 설정하였다. 그리고 위의 6가지 규칙으로 분류하지 못하는 용어가 발생할 경우에는 '예외 규칙'과 '혼합 규칙'을 통해 분류하였다.

표제어의 개념/인스턴스 규칙에는 영어 고유명사의 특성인 제일 앞 글자의 대문자 표기, 숫자 포함 여부, 파생어, 합성어 등 주로 용어 및 대역어의 형태정보를 이용하고 뜻풀이의 개념/인스턴스 규칙에는 주로 문맥정보를 분석하여 규칙 설정에 이용하였다. 이 규칙들을 바탕으로 인스턴스 규칙이 적용되었다면 양(+)의 가중치를 더해주고 개념 규칙이 적용되었다면 음(−)의 가중치를 더해주는 가중치 계산방법을 설정하고 가중치가 임계치 이상이면 인스턴스, 그렇지 않으면 개념으로 분류하는 시스템을 개발하였다.

개념/인스턴스 분류 시스템의 성능 평가를 위해 사용하는 실험 대상 분야는 IT839 정책의 9대 신성장동력 분야 중 이동통신, 디지털TV, 지능형로봇 3가지 분야로 설정하고 실험 대상 말뭉치는 백과사전과 TTA용어사전의 한글 및 영어 표제어, 뜻풀이를 가진 용어 중 2,000개를 대상으로 실험하였다. 그리고 정확률 측정을 위해 미리 2,000개의 용어에 대해 개념(1,000개) 및 인스턴스(1,000개)를 분류하였다. 개념과 인스턴스 분류 실험결과 정확률은 97.7%, 86.4%로 측정됐고, 전체 정확률은 92.1%였다. 실험 결과를 살펴보면, 개념에 비해 인스턴의 분류 정확률이 좀 더 낮은 것을 볼 수 있다. 이것은 뜻풀이에 확실한 정보가 없거나, 아니면 뜻풀이나 영어표제어가 비어 있어서 제대로 분석해 내지 못한 경우이거나, 적절한 인스턴스 규칙이 부족한 경우이다.

향후 기구축된 온톨로지를 참고하여 인스턴스 규칙을 세밀화하고, 문법 규칙의 확장 및 통계적 또는 기계학습 방법을 통해 정확률을 향상시킬 수 있을 것이다.

감사의 글

본 논문은 정통부 및 정보통신연구진흥원의 정보통신선도기반기술개발사업의 연구결과로 수행되었습니다.