온톨로지 인스턴스 생성을 위한 상호참조 해결 연구

최미란 이창기 왕지현 장명길
한국전자통신연구원 지식마이닝연구팀
{miranc, leeck, jhwang, mgjang}@etri.re.kr

Reference Resolution for Ontology Population

Miran Choi, Changki Lee, Jihyun Wang, Muyng-Gil Jang
ETRI, Knowledge Mining Research Team

요 약

시맨틱 웹 기술의 주축을 이루는 온톨로지의 구축시에 인스턴스를 생성하기 위하여 대상 문서를 구성하는 자연어 문장을 텍스트 마이닝 기술을 이용하여 트리플을 추출한다. 인스턴스를 생성할 때 보다 많은 정보를 추출하기 위해서 문장에 나타나는 상호참조 해결이 필요하다. 본 연구에서는 문서에서 많이 나타나는 명사구로 이루어진 대용어를 해석하기 위하여 언어 분석된 다양한 결과 정보를 이용한다. 본 연구에서는 계층적인 의미구조와 청킹을 이용한 규칙기반의 상호참조 해결 방법을 제안하고 실험을 통해 알고리즘의 정확도를 제시한다.

1. 서론

최근 국내외로 많은 관심을 받고 있는 시맨틱 웹 기술의 주축을 이루는 온톨로지는 지난 몇 년간 다양한 연구기관에서 구축되어왔다. 온톨로지 구축에는 클래스와 더불어 온톨로지의 자질을 결정하는 인스턴스 생성이 중요하기 때문에 응용 영역에서 사용하기에 충분한 인스턴스를 대량으로 자동 생성하는 온톨로지 인스턴스 시스템에 대한 연구가 많이 수행되고 있다.[8] 온톨로지 인스턴스 생성 시스템은 텍스트 마이닝 기술과 온톨로지 인스턴스 매칭 기술을 이용하여 문서로부터 인스턴스를 자동 생성하는 시스템이다.

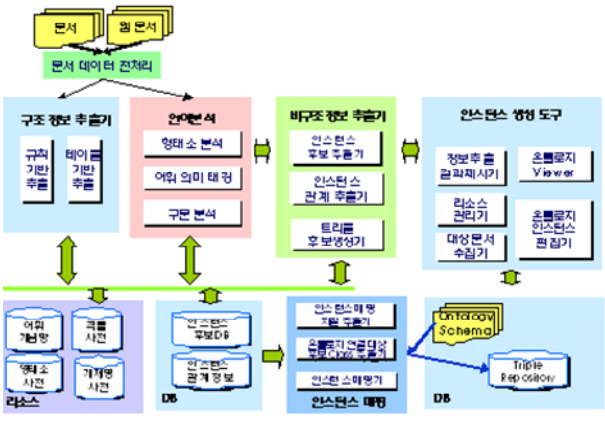

그림 1. 온톨로지 인스턴스 생성 시스템

온톨로지 인스턴스 생성 시스템은 웹 문서와 일반 텍스트 문서를 처리 대상으로 한다. 그림 1에서와 같이 전체 시스템의 흐름은 크게, 1) 웹문서의

구조적인 부분을 처리하는 구조 정보추출기 2) 웹 문서와 일반 텍스트 문서에서 비구조 부분을 처리하는 비구조 정보추출기 3) 각 구조 및 비구조 정보추출기로부터 생성된 후보 인스턴스를 기반으로 실제 온톨로지 스키마를 연결해 주는 인스턴스 매핑기 4) GUI 기능을 부여하는 인스턴스 생성도구 부분으로 구성된다..

비구조 정보추출기는 일반 문서로부터 인스턴스를 생성해 내는 모듈로 개체명 인식기, 관계 추출기, 그리고 상호참조 해결기 등으로 구성된다. 인스턴스가 자동생성되는 과정은 일반 자연어 문서로부터 형태소 분석을 거친 후에 개체명 인식과 상호참조 해결 기능이 수행되고, 인스턴스 관계정보를 추출한 후에 Triple 후보를 생성하여 온톨로지에 인스턴스로 추가된다.

상호참조 모듈은 온톨로지 인스턴스 생성시에 문장내의 상호참조를 해결함으로써 인스턴스 생성을 보다 정확하게 할 수 있고 누락되는 인스턴스를 추가하여 트리플 생성의 재현률을 높일 수 있다.

자연어 문서에서 나타나는 상호참조는 문서내의 응집력을 높이기 위한 목적을 위해 사용되는데 하나의 주제어에 대하여 설명을 하는 context를 제공함으로써 문장간의 지속성을 제공한다.

상호참조해결에 관한 기존 연구에서는 대용어와 참조대상간의 단순한 의미 매핑을 통하여 해결하거나([1],[3],[5]) 참조대상을 하나의 단어에 제한하여([2],[4],[5]) 해결하는 방식을 사용하였기 때문에 정확한 참조해결이 되지 않았다. 본 연구에서는 계층적인 의미 구조와 청킹을 이용한 상호참조 알고리즘을 제안한다.

본 논문의 구성은 다음과 같다. 먼저 2장에서 상호참조 해결의 접근 방법을 설명하고 3장에서는 상호참조 알고리즘을 제안한다. 4장에서는 제안된 알고리즘을 이용하여 실험 및 평가를 하고 5장에서는 결론과 향후 연구 방향을 소개한다.

2. 상호참조 해결의 접근 방법

상호참조는 자연어 문장에서 반복을 피하고 문맥의 통일성을 이루기 위하여 많이 사용되는 현상이며 상호참조를 위해 사용되는 대용어의 종류가 다양하게 발생되어서 전산언어학 분야에서 오랫동안 다루어졌으나 해결이 쉽지 않은 문제이다. 대용어란 하나의 문장이나 여러 개의 문장 사이에서 같은 요소가 되풀이 될 때 되풀이 되는 요소를 대명사나 지시사, 재귀대명사, 생략, 일반 명사를 사용하여 대치시키는 현상이다. 본 연구에서는 이러한 대용어를 인식하고 그 대용어가 지시하는 참조 대상을 파악하여 연결해주기 위한 방법을 제시하며 해결 대상은 지시사와 동반된 체언으로 구성된 대용어이다. 즉, 다음과 같은 예에서 “그 회사”와 “마이크로소프트”를 연결하여 참조해결을 하고 대용어를 복원하는 방법을 사용한다. 즉 참조해결 방법을 통하여 “그 회사”는 “마이크로소프트”로 복원된다.

예) 얼마 전에 **마이크로소프트**는 새로운 운영체계인 윈도우비스타를 판매하기 시작했다. **그 회사**는 이전의 운영체계를 개선하여 사용자에게 보다 편리한 제품을 제공하는 것을 목표로 하고 있다.

상호참조를 해결하는 방식은 청취자가 문장간의 연결관계와 주변 컨텍스트를 이용하여 추론을 통하여 지시어와 참조 대상을 연결하는 방법이다. 이러한 과정을 기계가 이해하는 자동화 방법을 개발하기 위하여 지시어와 참조 대상이 사용되는 문장들의 의미와 구조를 분석하여 연결고리를 찾는 알고리즘을 사용한다. 즉 문장 구조 정보와 의미 정보를 이용하여 후보를 정하고 가장 적합한 참조어를 선택하는 방식을 사용한다.

지시어의 종류에는 명사구, 대명사, 재귀대명사가 있으며 문서에 따라 나타나는 비율이 차이가 있다. IT 문서의 경우에는 명사구가 많이 나타나며 대명사나 재귀대명사는 거의 나타나지 않는다.

현재의 상호참조 모듈은 대상을 명사구로 하였으며 목적은 시맨틱 관계 트리플 생성시 추가 정보를 확보하기 위하여 사용되며 지식 융합 (knowledge consolidation)을 위하여 사용된다.

상호참조 모듈은 언어분석기의 추가 모듈로 개발되었다. 다음 예는 상호참조가 IT 문서에서 나타난 예이다. 첫번째 예에서 “이 회사”라는 명사구는 “삼성코닝”을 참조한다. 두번째 예에서는 “그들”이라는 대명사가 “퀄컴 사람들”을 참조한다.

예) **삼성코닝**(대표 송용로)은 최근 새로운 CI 작업에 들어간 상태다. **이 회사**는 종합 소재·재료 업체의 느낌을 줄 수 있는 새 CI를 발표할 예정이다

예) 마침 **퀄컴 사람들**이 홍콩을 거쳐 한국의 동정을 살피려고 나타났다. **그들**에게 전화를 건네주었다.

위와 같은 상호참조 해결 방법으로 첫번째 예문에 대하여 그림과 같은 규칙기반의 접근 방법을 사용하였다. 즉 의미레벨에서 개체명(NE:회사)의 의미와 지시어에 나타난 명사의 의미(회사)를 기준으로 상호참조를 해결하는 방식이다. 실제로 상호참조를 할 때에는 하나의 개체명이 하나의 지시어 명

사구에 의해 참조되는 경우 보다는 개체명과 추가의 단어로 이루어진 명사구가 참조 대상이 되는 경우가 더욱 많이 나타난다. 또한 의미의 1대1 매핑보다는 의미의 계층을 이용하면 상호참조를 더욱 효과적으로 할 수 있다. 3장에서는 단어의 계층적인 의미 구조와 청킹을 이용한 규칙기반의 상호참조 알고리즘을 상세하게 설명한다.

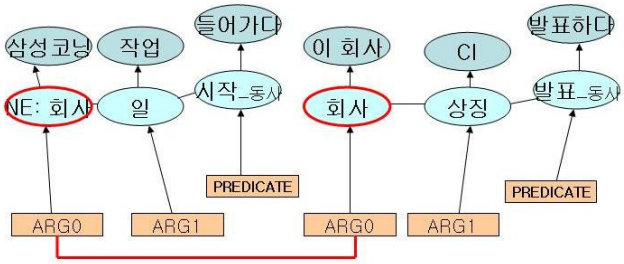

그림 2, 규칙기반 상호참조 해결 예

3. 상호참조 알고리즘

본 연구에서 제안하는 상호참조 해결방법의 핵심 요소는 계층적 의미구조를 이용한 대상 후보의 순위화와 청킹을 이용한 최장일치 후보를 찾아 내는 방식이다. 계층적 의미 구조를 이용한 대상 후보의 순위화란 다음과 같다. 모든 명사를 ETRI에서 개발한 400 여개의 계층적인 의미코드를 사용하여 분류를 하는데, 대용어와 대용어의 참조대상을 연결할 때 이 의미코드를 기준으로 하되 한 의미의 상위 의미와 매핑되는 경우도 대상에 포함하여 후보로 선정한다. 이때 하위 의미와 매핑되는 후보가 보다 더 구체적이므로 선호하여 가중치를 주는 방식을 사용하여 후보간의 순위를 정한다. 이렇게 전체 의미 구조 정보를 이용함으로써 순위에 따라 상호참조의 정확률을 높일 수 있고 후보 대상의 범위를 넓혀서 재현율을 높일 수 있다. 상세한 예는 아래 예문에 제시되어 있다.

청킹을 이용한 방법은 대용어가 지시하는 참조대상이 하나의 개체명일 경우도 있지만 복합명사나 복합절 전체가 대상이 될 수 있다. 본 연구에서는 주어나 목적어와 같이 하나의 문장구성요소를 이루는 최장 후보에 가중치를 두고 참조대상으로 선정한다.

상호참조 해결은 그림3과 같은 순서로 진행된다.

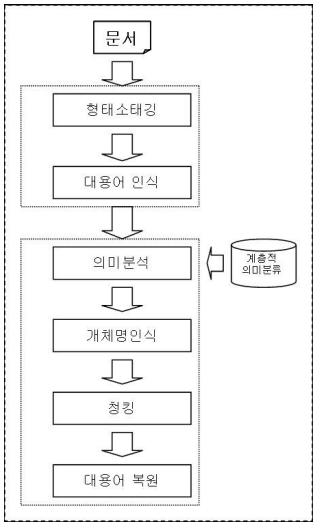

그림 3. 상호참조 모듈 구성도

입력된 문서는 우선 형태소 태깅이 되어 대용어 대상 태그를 분석하여 대용어가 인식된다. 이때 대명사와 “지시사+명사”가 대용어로 인식된다. 다음으로 의미분석을 통해서 대용어가 가지고 있는 의미를 파악한다. 의미는 계층적 의미 구조에 의해 상위 의미까지 모두 태깅된다. 예를 들어서 “고객”이라는 명사는 “사람”이라는 의미가 부여되는데 이때 사람의 상위 의미인 “생물>동물>사람”의 전체 의미 구조가 배정된다.

다음으로 개체명 인식을 통해 문서에 나오는 모든 개체명이 태깅된다. 다음으로 자연어 문장들은 청킹을 거치면서 복합명사나 수식절 등이 청크의 형태로 구성되며 하나의 청크에는 하나의 개념이 배정된다. 여기에서 “개념”은 청크 안에 나타나는 개체명이나 명사의 의미를 분석하여 정해지는데 계층적 의미 구조에 있는 의미 코드를 이용한다. 이때, 개체명과 청크 개념과의 매핑 테이블을 이용하여 개념을 정한다.

대용어 복원 단계에서는 대용어에 배정된 계층적인 의미와 청크에 배정된 개념을 분석하여 후보자들 중에서 가장 적합한 청크를 찾아낸 후에 대용어를 복원한다.

후보자의 랭킹을 매기는 방법은 대용어에 배정된 계층적 의미 중에서 가장 하위에 있는 의미가 가장 구체적인 의미이기 때문에 가장 큰 가중치를 부여한다. 계층적 의미구조의 가장 상위에 있는 의미는 가장 작은 가중치를 부여한다.

랭킹을 매길 때 적용되는 다른 규칙은 최근에 나온 정보를 우선으로 하며 같은 조건일 때는 수식절보다는 주어나 목적어처럼 문장의 필수적인 구성

요소에 해당되는 참조 대상을 우선으로 한다. 언어 분석을 거쳐서 수행되는 대용어 인식 및 대용어 복원의 예는 다음과 같다.

원문: 불교는 아시아 문화에서 많이 믿는 종교이다. 그 종교는 살생을 금하는 규율을 가지고 있다.

형태소태깅: 불교/nc는 아시아/nc문화/nc에서 많이/mag 믿/pv는 종교/nc이다. 그/mm 종교/nc 는 살생/nc 을 금하/pv 는 규율/nc 을 가지/pv고 있다.

대용어인식: 불교/nc는 아시아/nc 문화/nc에서 많이/mag 믿/pv는 종교/nc이다. **그/mm 종교/nc** 는 살생/nc 을 금하/pv 는 규율/nc 을 가지/pv고 있다.

의미분석: 불교/nc[종교] 는 아시아/nc 문화/nc[양식]에서 많이/mag 믿/pv는 종교/nc[종교]이다. **그/mm 종교/nc**[모양>양식>문화>종교] 는 살생/nc[행위] 을 금하/pv 는 규율/nc[규범] 을 가지/pv고 있다.

개체명인식: <불교:OGG_RELIGION>는 <아시아:LCG_CONTINENT> 문화에서 많이 믿는 종교이다. **그 종교**는 살생을 금하는 규율을 가지고 있다.

청킹: [<불교:OGG_RELIGION>는:종교] [<아시아:LCG_CONTINENT> 문화에서:양식] [많이] [믿는] [종교이다: 종교]. **그 종교**는 살생을 금하는 규율을 가지고 있다.

대용어복원: <불교:OGG_RELIGION>는 <아시아:LCG_CONTINENT> 문화에서 많이 믿는 종교이다. <불교:OGG_RELIGION>는 살생을 금하는 규율을 가지고 있다.

위의 단계를 통하여 복원된 대용어는 인스턴스를 생성하기 위한 관계추출 모듈에서 사용된다. 관계추출 모듈에서는 개체명과 인스턴스 관계 구조 정보를 이용하여, 한 문장안에 있는 임의의 두 개체명간에 미리 정의된 관계(predicate)가 존재할 확률을 구하여, 일정 값을 넘는 경우 트리플 후보로 생성한다.

아래 예는 대용어에 의미를 배정할 때 사용하는 계층적인 의미 구조이다. 예에서와 같이 왼쪽으로 계층이 올라갈수록 의미의 추상성이 높아지며 오른쪽으로 계층이 내려갈수록 의미의 구체성이 높아진다. 대용어에 배정되는 의미는 전체 의미 계층이 모두 배정되며 대용어에 배정된 의미와 청크의 개념을 비교할 때 구체적인 의미가 보다 높은 가중치를 받게 되어 랭킹시에 더 나은 후보가 될 수 있다. 위의 예에서 대용어 "그 종교"는 "모양> 양식> 문화> 종교"라는 전체 의미구조가 배정되며 청크에 이 의미구조 중의 하나의 의미에 해당되는 개념이 있으면 후보가 되고 예문에서와 같이 다수의 후보인 불교(종교)와 아시아 문화(양식)가 나오면 이 중에 보다 구체적인 의미인 종교와 매치되는 후보인 "불교"가 아시아 문화(양식) 보다 더 높은 가중치를 갖기 때문에 대용어 복원에 선정되어 사용된다.

대상>산물>창작물>저작물>글
물건>기계>기구>도구>문방구
모양>겉모양>모습>경치
물건>먹을거리>음식물>장
기준>조리>이치>원리

계층적인 의미구조의 예

4. 실험 및 평가

실험 대상 문서는 전자신문의 IT관련 기사 150개로 구성된 코퍼스를 이용하였다. 대상 어휘는 대명사와 "지시사+명사"를 대용어로 하여 출현빈도를 계산하였다. 수동으로 한 문서 분석 결과 출현빈도는 150개의 문서에서 120 개의 대용어가 나타났으며 (약80%) 그 중 대명사로 된 대용어는 4개가 나타났다. 평가대상에는 대명사는 제외하고 평가를 수행하였다. 평가는 실험셋을 이용하여 규칙을 튜닝하였으며 정보검색에 사용되는 정확률과 재현율을 사용하여 평가를 수행하였다. [표1]에서와 같이 재현율은 120개의 대용어 중에 118개를 찾아내어 98.3%이며, 그중에서 110개를 처리하여 정확률은 93.2%이다.

표1 상호참조 대용어 해결 실험 결과

전체 기사수	150건
대용어 등장수	120개 (약80%)
찾아진 대용어 수 (재현율)	118개 (98.3%)
처리된 대용어 수 (정확률)	110개 (93.2%)

제안된 상호참조 해결 알고리즘이 처리에 실패한 대용어의 원인을 살펴보면 다음과 같다.

- 대용어 인식 단계에서 "지시사+명사"로 구성된 대용어만 인식했는데 실제 기사에 나오는 대용어 중에는 "명사+명사"로 구성된 경우가 나타났다. (예) 이번 전시회)
- 대용어 처리 단계에서 참조대상이 청크 안에 내재되어 있을 경우에 전체 청크의 개념과 상이할 때에는 복원 후보로 추출되지 않았다.

- 의미분별에 실패했을 경우에 추출되지 않았다.(언어분석의 오류)
- “지시사+ 명사”로 구성된 대용어이지만 의미특성상 참조대상 추출이 어려운 경우가 나타났다. (예)양측 회사, 한 장소, 각 시스템)

5. 결론

본 연구에서 제안한 계층적인 의미구조와 청킹을 이용한 상호참조 해결 알고리즘은 기존의 단순 의미 매핑 방식의 한계점을 해결하여 대용어 인식과 복원의 정확률을 높일 수 있는 정교한 방식이다. 향후 연구에서는 상호참조 해결 모듈 이전에 언어분석시에 생기는 오류에도 대응할 수 있는 강건한 방법의 연구가 필요하다. 또한 빈도수는 적지만 대명사의 대용어 연구도 대량의 인스턴스 자동 생성을 위하여 필요할 것으로 생각된다. 또한 현재의 규칙기반의 알고리즘의 다양한 특성들을 이용하여 통계적인 방식으로 대용어를 해결하고 두 가지 방식의 성능을 비교하는 연구도 필요하다.

참고문헌

[1] 다중모드 대화 시스템에서 이중 캐시 모델의 센터링 알고리즘을 이용한 명사 대용어구 처리. 김학수, 서정연. 한국정보과학회논문지 2000, 27(11),pp,1133-1140

[2] Coreference Resolution을 위한 3인칭 대명사의 선행사 결정 규칙. 강승식 외. 한국정보처리학회논문지B, 2004.11B(2),pp,227-232

[3] 한국어 복합문에서의 제로 대용어 처리를 위한 분해 알고리즘과 복원규칙. 김미진 외. 한국정보과학회논문지:소프트웨어및응용, 2002, 29(10),pp.1133-1140

[4] 한국어 문장내 체언류 조응대용어의 해결방안. 김정해 외. 대한전자공학회논문지B, 1996, 33B(4), pp. 183-190

[5] 문서요약을 위한 조응 대용 해결. 김상수 외.한국정보과학회 가을 학술발표논문집 29(2), pp.679-681

[6] Fine-grained Named Entity Recognition and Relation Extraction for Question Answering. Chang-Ki Lee et. al., SIGIR 2007.

[7] Fine-grained named entity recognition using Conditional random fields for question answering. Chang-Ki Lee et. al., AIRS, 2006.

[8] 차세대 지능형 시맨틱 웹과 온톨로지. 현승순, 정보통신연구진흥원 기술동향, 2006.