한국지능시스템학회 논문지 2011, Vol. 21, No. 2, pp. 276-282
DOI : 10.5391/JKIIS.2011.21.2.276

WordNet을 매개로 한 CoreNet-SUMO의 매핑

Mapping between CoreNet and SUMO through WordNet

강신재* · 강인수** · 남세진*** · 최기선***
Sin-Jae Kang*, In-Su Kang**, Se-Jin Nam*** and Key-Sun Choi***

* 대구대학교 정보통신대학 컴퓨터 · IT공학부
** 경성대학교 공과대학 컴퓨터학부
*** KAIST 시맨틱웹 첨단연구센터

요 약

CoreNet은 한-중-일 다국어 텍스트의 분석, 언어 간 변환을 포함한 자연어처리에 유용한 자원이다. CoreNet의 보다 광범위한 분야 및 응용에의 활용을 장려하고 다국어 어휘의미망으로서의 국제적 위상을 제고하기 위해 SUMO에 연결하는 작업을 하였다. CoreNet과 SUMO를 매핑하기 위해 간접 매핑과 직접 매핑 방법을 모두 사용하였는데, CoreNet-KorLex-PWN-SUMO에 이르는 간접 매핑 작업을 통하여 한국어 중심의 CoreNet과 영어로 기술된 SUMO의 언어 간 변환의 어려움을 완화하고 CoreNet 개념에 대응하는 SUMO 클래스의 재현율을 극대화하였다.

키워드 : 온톨로지 매핑, 간접 매핑, 코아넷, 수모, 워드넷

Abstract

CoreNet is a valuable resource to use in the domain of natural language processing including Korean-Chinese-Japanese multilingual text analysis, and translation among natural languages. CoreNet is mapped to SUMO in order to encourage its application in broader fields and enhance its international status as a multilingual lexical semantic network. To do this, indirect and direct mapping methodologies are used. Through the indirect mapping among CoreNet-KorLex-PWN-SUMO, we alleviate the difficulty of translating CoreNet concept terms in Korean into SUMO concepts in English, and maximize recall of SUMO concepts corresponding to the concept of CoreNet.

Key Words : Ontology Mapping, Indirect Mapping CoreNet, SUMO, WordNet

1. 서 론

CoreNet은 한-중-일 언어에 대한 다국어 어휘의미망으로 한국과학기술원에서 10여 년의 개발 기간을 거쳐 2005년에 공개된 온톨로지이다. CoreNet은 한국(Corea) 혹은 한국어 중심(Core)의 어휘망(Network)이라는 뜻이 담겨 있으며[1], 2004년 기준으로 CoreNet의 한국어 부분은 3만 이상 어휘의 6만 이상 의미가 CoreNet 개념체계에 연결되어 있다[2].

CoreNet은 다른 어휘의미망과 구별되는 많은 특징을 갖는다. 먼저, CoreNet이 사용하는 개념체계는 명사, 동사, 형용사의 품사를 아우르는 품사 독립적 개념체계이다. 이는 자연어 텍스트의 의미 분석에서 개별 어휘의 품사에 무관하게 단일의 개념망을 통한 추론이 가능하다는 장점을 갖는다. 다음으로 한-중-일 언어를 단일 개념 체계로 포괄함으로써 동북아시아 3개 언어 간 기계번역을 위한 의미해석에 기여할 뿐 아니라 한-중-일 언어의 의미 차이, 의미-표현상의 차이 등 언어학 및 언어교육 연구에 중요한 언어자원이 될 수 있다[3]. 또한 대용량 코퍼스를 기반으로 선정된 한국어 기본 어휘들의 의미 네트워크를 구축함으로써 텍스트 분석에서 CoreNet의 높은 어휘 점유율(coverage)을 기대할 수 있다.

이처럼 서로 다른 언어의 서로 다른 품사에 속한 어휘의 의미를 단일의 개념 체계로 통합하고 있는 CoreNet은 특정 언어 및 다국어 텍스트의 분석, 언어 간 변환을 포함한 자연어처리 및 그 응용에 유용한 지식자원이다. 하지만 이러한 장점에도 불구하고 CoreNet의 국제적 활용도는 왕성하지 않은 편이다. 따라서 본 논문에서는 CoreNet의 보다 광범위한 분야 및 응용에의 활용을 장려하고 다국어 어휘의미망으로서의 국제적 위상을 제고하기 위해 CoreNet 개념체계를, 상위 온톨로지의 ISO 표준 후보인 SUMO (Suggested Upper Merged Ontology)에 연결하는 작업하였다.

CoreNet-SUMO의 연결을 위해 두 개념 체계의 대응하는 개념을 수작업으로 직접 탐색하여 연결하는 직접 연결 방식과, Princeton WordNet(PWN)-SUMO, KorLex(한국어 워드넷)-PWN의 기존 매핑에 기반하여 CoreNet-KorLex-PWN-SUMO의 매핑을 수행하는 간접 연결 방식

접수일자 : 2010년 11월 3일
완료일자 : 2011년 3월 10일
본 논문은 2009년도 지식경제부 및 한국산업기술평가관리원의 정보통신선도기반기술개발사업의 연구결과와 2010년도부터는 정부(교육과학기술부)의 재원으로 한국연구재단의 기초연구사업(No. 2010-0022444)의 연구결과로 수행되었습니다.

을 동시에 적용하는 접근법을 시도하였다. 간접연결은 한국어 중심의 CoreNet과 영어로 기술된 SUMO의 언어 간 변환의 어려움을 완화하고 CoreNet 개념에 대응하는 SUMO 개념의 재현율을 향상시키는데 기여할 수 있으므로 간접연결을 먼저 수행한 다음 그 결과를 직접 연결의 매핑 후보로 활용하는 방법을 사용하였다.

SUMO는 이미 영어에 대한 명실상부한 대표적 어휘의미망인 영어 워드넷 (Princeton WordNet, PWN)과 연결되어 있다. 따라서 CoreNet과 SUMO의 연결은 동북아시아 3개 언어에 제한된 CoreNet의 다국어 지원 폭을 국제 공용어인 영어로 확장할 뿐 아니라, 교착어인 한국어, 일본어, 고립어인 중국어와 함께 굴절어인 영어를 포함함으로써 CoreNet이 다양한 언어 유형에 대한 통합망으로 발전하는 계기가 될 것이다.[1)]

논문의 구성은 다음과 같다. 2장에서는 CoreNet, SUMO, 영어/한국어 WordNet에 대해 설명한다. 3장에서는 CoreNet-SUMO의 매핑방법론과 간접/직접 매핑 과정에서 이슈가 되는 사항들을 기술한다. 4장에서는 매핑 결과에 대해 기술하고, 5장에서 결론을 맺는다.

2. 관련 연구

2.1 CoreNet

CoreNet[4]은 한국과학기술원에서 1994년부터 구축을 시작한 다국어 어휘의미망으로 2005년 한국어, 중국어, 일본어에 대한 그간의 작업 결과가 공개되었다. 어휘의미망(lexico-semantic network)은 단어의 의미들을 상위어, 하위어, 동의어 등의 의미관계로 연결한 네트워크(망)로, 그 규모와 개별 언어의 의미적 특성을 고려하여 일반적으로 하나의 언어에 대해 구축되며 대개 품사별로 어휘의미망이 분리되어 있다. CoreNet은 기존의 다른 어휘의미망과 달리 품사별 어휘의미망이 분리되어 있지 않으며 단일의 개념체계를 명사, 동사, 형용사의 서로 다른 품사에 속한 어휘 의미들이 공유할 뿐 아니라 현재 한국어, 중국어, 일본어의 3개 언어를 포괄하는 다국어 어휘의미망으로 제작되었다.

CoreNet 개념체계는 일본 전신전화주식회사(NTT)에서 구축한 어휘대계[5]의 '단어의미속성체계'에 기반하여 한국어, 일본어의 언어 차이에서 기인하는 부적합 개념 노드의 제거 및 추가 개념 노드의 확장을 거쳐 총 2,937개 개념 노드로 구성된 최대 깊이 12의 계층적 개념망이며, 개념 노드들은 상하관계(IS-A)와 전체부분관계(HAS-A)로 연결되어 있다. 이 개념체계의 각 개념 노드는 코퍼스로부터 수집된 어휘의 의미와 연결됨으로써 전체적으로 어휘 의미의 개념 네트워크를 구성하게 된다.

언어의 표현 수단 측면에서 CoreNet은 어휘부, 의미부, 개념부로 나눌 수 있다. 어휘부는 7,000만 어절 카이스트 코퍼스에서 추출된 5,000만 어절 코퍼스에 출현한 빈도 3이상의 어휘들을 중심으로 구성되어 있다. 어휘부의 품사체계는 코퍼스 분석에 사용한 카이스트 형태소분석기의 54개 품사가 우리말큰사전의 28개 품사체계로 변환된 것이다. 의미부는 여섯 단계로 세분된 우리말큰사전의 의미부류에 기초하고 있으며 개념부는 전술한 CoreNet 개념체계에 해당한다. 2004년 기준으로 한국어 CoreNet은 31,384개 어휘의 62,632개 의미가 CoreNet 개념체계의 적어도 하나 이상의 개념 노드에 연결되어 있다[2].

2.2 SUMO

SUMO는 잘 알려진 공용 상위 온톨로지이며, IEEE 상위 온톨로지 표준 후보 중 하나이다. IEEE 표준 상위 온톨로지 (Standard Upper Ontology, SUO) 워킹 그룹[2)]에서는 데이터 상호 운용성, 정보 탐색/검색, 자동 추론, 자연어 처리 등의 컴퓨터 응용을 지원하기 위한 상위 온톨로지의 표준을 개발하고 있으며 SUMO, OpenCyc, 4D ontology 등을 후보 온톨로지로 고려하고 있다. SUMO는 공학, 철학, 정보과학을 아우르는 다양한 분야 참여자들의 메일링 리스트(SUO email list)로부터 수집된 정보를 바탕으로 기존의 많은 온톨로지들을 단일의 응집된 구조로 병합함으로써 만들어졌다[6].

SUMO는 영어 워드넷(Princeton WordNet)의 모든 품사(명사, 동사, 형용사, 부사)의 모든 신셋(synonym set, synset)과의 매핑을 가진 온톨로지라는 점에서 자연어처리, 전산언어학 측면에서의 활용도가 더 증가할 것으로 판단된다. SUMO는 현재도 MILO(MId-Level Ontology)를 포함하여 통신(Communications), 분산 컴퓨팅(Distributed Computing), 경제(Economy), 재정(Finance), 지리(Geography), 군사(Military), 운송(Transportation) 분야의 도메인 온톨로지들과 결합되면서 계속 확장되고 있다. SUMO 공식사이트[3)]에서 제공하는 자료에 따르면, SUMO는 현재 개념 1,000여 개, 공리 4,000여 개, 규칙 750개로 구성되며, 도메인 온톨로지까지 포함할 경우 개념 20,000여 개, 공리 70,000여 개를 포함하고 있다.

2.3 WordNet

워드넷(WordNet)[4)]은 미국 프린스턴대학에서 1985년부터 지금까지 계속하여 구축하고 있는 영어에 대한 공용 계층적 어휘의미망이다[7]. 1991년, 2003년, 2006년 각각 WordNet 버전 1.0, 2.0, 3.0을 발표하면서 전세계 전산언어학자, 컴퓨터과학자들 사이에 큰 관심과 활용 성과를 만들어 냈다. 이와 동시에 영어 워드넷과 같은 어휘의미망을 자국 언어에 대한 언어처리를 위해 독립적으로 구축하려는 시도가 활발히 이루어져 현재 50여 개 이상의 언어에 대해 각 언어별 워드넷이 만들어져 있다. 본 논문에서는 워드넷의 원형인 영어 워드넷을 다른 언어 워드넷과 구별하기 위해 Princeton WordNet (PWN)으로 기술한다.

워드넷은 한 언어의 단어 집합에 대해 같은 의미를 갖는 어휘들의 모음인 동의어집합(synonym set, synset, 신셋)을 명사, 동사, 형용사, 부사의 각 품사별로 별도로 정의하고, 신셋 간의 의미관계를 표현한 어휘의미망이다. 2006년 공개된 버전 3.0의 PWN은 15만여 개 단어집합이 갖는 20만여 의미에 대해 11만여 개의 신셋이 정의되어 있다.

한국어 워드넷 중 PWN을 참조모델로 사용한 것으로 포스텍과 부산대학교의 구축 시도가 있다. 이 중 부산대학교의 한국어 워드넷은 KorLex라는 이름으로 현재까지 계속 확장되고 있다. 2007년 시점 KorLex는 13만여 개 단어집합이 갖는 15만여 의미에 대해 13만여 개의 신셋이 정의되어

1) (이케하라, 2005)에서도 CoreNet에 유사한 평가를 부여하였다.

2) http://suo.ieee.org/

3) http://www.ontologyportal.org/

4) http://wordnet.princeton.edu/

있다[8]. 본 논문에서는 한국어 워드넷을 Korean WordNet(KorLex)으로 기술한다.

워드넷을 이용한 관련 연구로는 온톨로지 개체를 일반화하는 과정에서 단어의 의미를 구분하기 위해 워드넷을 이용한 연구[9] 등이 있다.

3. CoreNet-SUMO 매핑

3.1 CoreNet-SUMO 매핑 방법론

CoreNet의 각 개념을 SUMO의 대응하는 개념들과 연결하는 가장 적절한 방법은 수작업 직접 연결을 맺는 것이다. 그러나 한국어 중심으로 개발된 CoreNet을 영어로 표현된 SUMO와 관련시키는 것은 한국어와 영어 사이의 언어 간 번역 과정이 요구된다. 이는 개념 체계의 개념 노드가 그 개념을 대표하는 어휘(들)로 표현되는 것이 일반적이기 때문에 CoreNet 개념을 대표하는 한국어 어휘와 SUMO 클래스를 대표하는 클래스명 간의 대역어 선택의 어려움을 피할 수 없음을 의미하는 것이다.

이러한 언어간 변환의 문제를 완화하기 위해 영어 워드넷을 한국어로 번역하는 방식에 기초하여 제작된 한국어 워드넷 KorLex를 매개로 활용하여 CoreNet과 SUMO 사이의 간접 매핑을 시도할 수 있다. 즉 CoreNet 개념 대표 어휘와 매치되는 KorLex 신셋들에 대한 수작업 어의애매성 해결(Word Sense Disambiguation)을 통해 CoreNet과 KorLex의 연결을 맺으면, 연결된 KorLex 신셋에 대응하는 영어 워드넷의 신셋을 획득할 수 있다. 이후의 과정은 영어 워드넷 신셋과 SUMO 클래스 간의 연결인데 이는 기존 PWN과 SUMO의 매핑 결과[5]를 활용하면 자동화가 가능하다.

일반적으로 서로 다른 두 언어의 개념 체계 간 매핑에 있어 직접 연결은 두 언어 모두에 능통한 고급 전문 인력이 요구되며 두 개념 체계의 이해가 전제되어야 하나, 양질의 매핑 결과를 얻을 수 있는 장점이 있다. 반면에 전술한 바와 같은 또 다른 개념 체계를 매개로 하는 간접 매핑은 시간과 인력의 부담을 줄이는 장점이 있으나 최종 매핑 결과가 매개로 사용하는 기존 매핑의 질에 의존하는 측면이 있으며, 반복되는 개념 간 연결에서 동의관계만의 연결이 아닌 경우 과도한 일반화 문제가 발생할 수 있다. 그러나 원시 언어의 한 개념에 대한 목표 언어의 모든 가능한 대역 개념을 재현해 내는 측면에서 두 언어 간 기존 개념 체계 매핑을 활용하는 것은 매핑의 재현율 향상에 기여할 것이다.

본 논문에서는 CoreNet과 SUMO를 매핑하기 위해 전술한 간접 매핑과 직접 매핑을 동시에 사용하는 접근법을 취함으로써 직접/간접 매핑의 장점을 모두 살리고자 하였다. 먼저 CoreNet-KorLex-PWN-SUMO에 이르는 간접 매핑을 만듦으로써 CoreNet 개념에 대응하는 SUMO 클래스의 재현율을 극대화하고, 이후 간접 매핑 결과를 매핑 후보로 활용하여 CoreNet-SUMO의 직접 매핑을 수행함으로써 사람 수준의 매핑 정확률을 보장하도록 하였다. 전체적인 매핑과정을 다음 그림에 제시하였다.

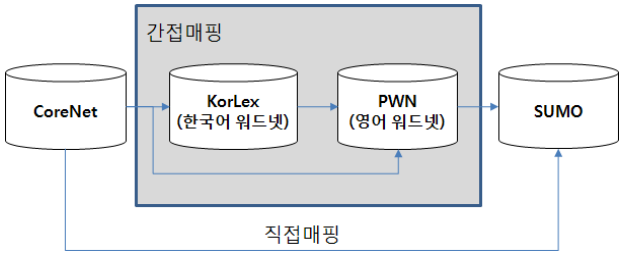

그림 1. 전체적인 매핑 방법
Fig 1. Overall mapping method

3.2 WordNet 매개 CoreNet-SUMO 간접 매핑

간접 매핑 방법을 취하는 이유는 매핑 전과정을 수작업으로 하기에는 작업의 난이도가 높고 작업 결과에 대한 객관성 및 완전성을 보장하기 어렵기 때문이다. 즉 CoreNet을 SUMO로 직접 매핑하려면 SUMO의 모든 클래스에 대한 이해가 있어야 하고, SUMO 내의 모든 매핑 후보를 철저히 찾을 수 있어야 하는데 수작업으로 감당하기에는 무리가 있다. 따라서 아래와 같은 간접 매핑 절차를 취한다.

CoreNet —②→ WordNet —①→ SUMO

① Princeton WordNet (PWN) → SUMO

Princeton WordNet(PWN)과 SUMO와의 매핑정보는 오픈 소스의 형태로 공개되어 있어서, 이 정보를 이용한다면 SUMO 클래스를 학습하는 부담을 덜 수 있게 된다. PWN의 synset과 SUMO의 클래스간 매핑 관계로는 동의(synonym), 하위-상위(hypo-hypernym), 개체화(instantiation)의 3가지가 사용되었다.

예를 들어 PWN의 synset 'man(09623892)'는 SUMO의 'Male' 클래스와 하위-상위 관계로 매핑된다.

② CoreNet → Princeton WordNet (PWN)

이제 CoreNet을 PWN에 매핑하는 방법만 확보하면 되는데, 한국어로 되어 있는 CoreNet의 클래스명을 영어로 번역하여 적합한 synset을 PWN에서 찾는 과정이 필요하다. 이러한 한-영 번역의 부담을 덜기 위해 한국어 워드넷을 다음과 같이 추가로 이용하였다.

CoreNet → Korean WordNet (KorLex) → Princeton WordNet (PWN)

CoreNet의 한국어 클래스명으로 KorLex를 검색하여 적합한 synset을 선택하기만 하면, CoreNet → KorLex → PWN → SUMO의 연결이 가능하기 때문에, 결과적으로 CoreNet과 SUMO간의 매핑이 이루어지게 된다. 하지만 경우에 따라서는 CoreNet 클래스명에 적합한 KorLex의 synset이 존재하지 않는 경우도 있다. 이러한 경우에는 한-영 번역 과정을 거쳐 PWN으로 매핑하게 된다. KorLex와 PWN간의 매핑은 동일한 워드넷 버전인 경우 동일한 synset ID를 사용하므로 자동으로 매핑할 수 있다.

예를 들어 CoreNet 클래스 '남자(11111211)'는 KorLex의 synset '남자(09623892)'를 통하여 PWN의 synset 'man(09623892)'로 자동 매핑된다.

CoreNet과 워드넷간의 매핑 관계로는 동의, 하위-상위,

5) http://sigmakee.cvs.sourceforge.net/sigmakee/KBs/WordNetMappings/

상위-하위의 3가지를 사용하였으며, 매핑 정보를 표기하기 위한 표현법은 다음과 같이 집합 개념을 도입하였다. 이는 하나의 클래스에 대응되는 대상 개념이 여럿 존재할 수도 있으며, 자식 개념을 모두 포함하는 형태의 부모 개념을 표현하기 위함이다.

{ synset ^ pos : synset_ID ^ mapping_relation } (1)

synset : 워드넷 신셋
pos : 품사
synset_ID : 워드넷 신셋 ID
mapping_relation : 동의, 하위-상위, 상위-하위 관계 중 하나

CoreNet에 존재하는 클래스 가운데 워드넷과의 매핑 과정에서 고려해야 할 특별한 유형들을 아래와 같이 4가지로 분류하여 정의하였다[10].

1. G-term (Group term)

CoreNet 클래스명 가운데 특정 클래스의 자식 클래스명들을 구분자 '/'을 이용하여 나열하는 형태로 표현된 클래스로, '규칙/법률/조약'(rule/law/treaty), '윤리/종교'(ethics/religion), '문장/구/단어'(sentence/phrase/word)와 같은 것들이 있다. '윤리/종교' 클래스의 자식 클래스는 다음 그림과 같으며, 이러한 유형의 클래스는 CoreNet에서 14.6%를 차지한다.

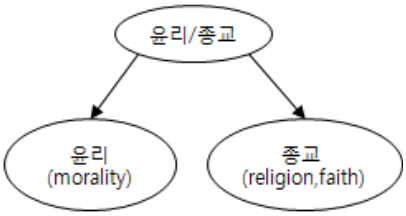

그림 2. G-term 클래스 예시
Fig 2. Example of G-term class

2. A-term (Antonym term)

영어에서는 반의어쌍의 의미가 각각의 단어로 따로 존재하지만, 한국어에서는 반의어쌍의 의미가 하나의 단어로 표현되는 클래스 유형을 뜻한다. 이는 G-term의 특별한 형태로 볼 수 있기 때문에, 매핑 시 G-term과 동일한 방법으로 처리할 필요가 있다. '빈부'(rich/poor), '주객'(host/guest) 등이 있으며, CoreNet내 구성비율은 1.0%이다.

3. C-term (Complementary term)

부모 클래스가 의미하는 개념 범위를 자식 클래스들이 나누어 가진다고 할 때, 한 자식 클래스가 다른 자식 클래스들의 개념 범위를 제외한 나머지 개념 범위를 가지는 클래스 유형이다. 주로 '기타'라는 문구가 클래스명 앞에 나타나며, '기타 근로자'(The other workers) 클래스의 계층 구조 일부분은 다음 그림에 나타나 있다. CoreNet내 구성비율은 6.3%이다.

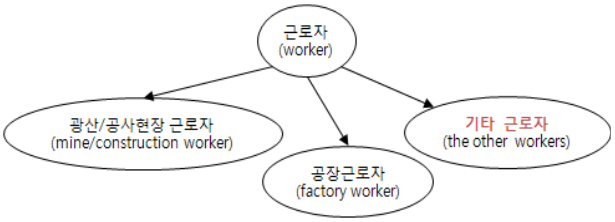

그림 3. C-term 클래스 예시
Fig 3. Example of C-term class

4. P-term (Phrasal-term)

CoreNet 클래스 명이 구 형태로 되어 있는 유형으로, '가격 인상'(price advance)과 같은 것을 예로 들 수 있다. 이러한 경우 해당 개념의 의미를 모두 포함하는 신셋을 워드넷에서 찾기 어렵다. CoreNet에 극소수가 존재한다.

CoreNet과 SUMO간의 간접 매핑 과정은 다음과 같이 3단계로 나뉘어 진행된다.

<1단계>

CoreNet 계층체계에서 모든 말단 클래스와 G-term, A-term, C-term, P-term 유형을 제외한 비말단 클래스를 수작업으로 워드넷에 매핑한다. 구체적인 수작업 절차는 다음과 같다.

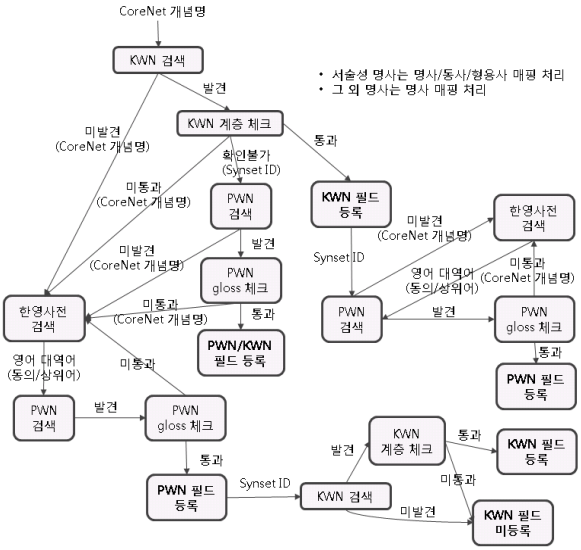

그림 4. CoreNet 클래스의 WordNet으로의 수작업 매핑 절차
Fig 4. Manual mapping steps from CoreNet classes to WordNet

수작업의 효율을 위해서 CoreNet DB, KorLex DB, PWN DB를 이용하여 CoreNet의 각 클래스에 대해 자동생성한 워크시트를 가지고 작업자가 수작업으로 KorLex와 PWN에 매핑한다. 다음은 CoreNet 클래스 '세탁'에 대한 워크시트의 일부와 매핑 결과를 예시하고 있다.

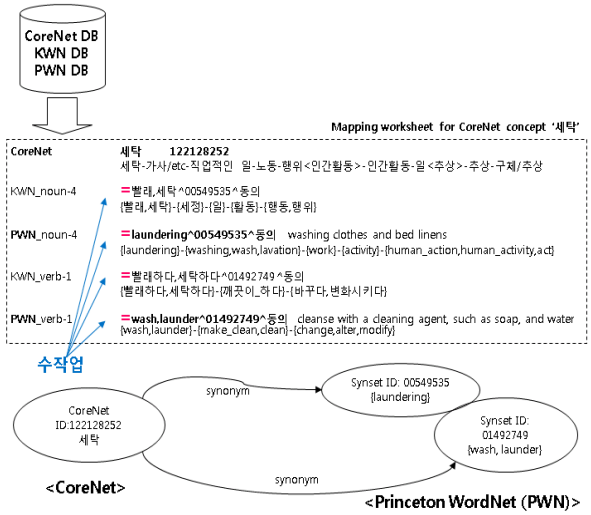

그림 5. CoreNet 클래스 '세탁'에 대한 워크시트 일부와 매핑 결과 예시

Fig 5. Worksheet about CoreNet class 'setak' and an example of mapping results

매핑 대상인 CoreNet 클래스와 워크시트에 나열된 워드넷 신셋 후보 목록 가운데 매핑이 가능하다고 판단되는 경우에는 '=' 기호를 삽입하고 해당 매핑의 종류를 기술한다. 이렇게 추가된 정보는 추후 프로그램으로 자동 추출한다.

<2단계>

1단계 후 아직 매핑되지 않은 비말단 클래스(G-term, A-term, C-term, P-term)를 대상으로 각 유형별로 워드넷에 매핑하는 방법은 다음과 같다.

1. G-term

집합 표현법을 이용하여 G-term의 자식클래스에 매핑된 synset들을 상위-하위(hyper-hyponym) 관계로 각각 매핑하고 전체를 집합으로 묶는다. '윤리/종교' 클래스의 매핑결과는 {{morality^pn:04614989^상위-하위}, {religion,faith^pn:07591116^상위-하위}} 이다.

2. A-term

G-term과 동일한 방법으로 매핑한다. '주객' 클래스의 매핑결과는 {{host^pn:09530955^상위-하위}, {guest,invitee^pn:09498008^상위-하위}} 이다.

3. C-term

C-term의 부모 클래스에 매핑된 synset과 하위-상위(hypo-hypernym) 관계로 매핑한다. '기타 근로자' 클래스의 매핑결과는 {worker^pn:09025575^하위-상위} 이다.

4. P-term (Phrasal-term)

P-term의 중심어(headword)를 찾은 후, 중심어와 매핑되는 워드넷 신셋을 찾고, 하위-상위(hypo-hypernym) 관계로 매핑한다. '가격 인상' 클래스의 경우, 중심어인 '인상'의 매핑결과는 {advance,gain^verb:00152358^동의} 이므로, {advance,gain^verb:00152358^하위-상위}이 '가격 인상' 클래스의 최종 매핑결과가 된다.

<3단계>

PWN과 SUMO 사이의 매핑관계와 CoreNet과 PWN 사이의 매핑관계를 고려하여, CoreNet과 SUMO 사이의 최종 매핑관계를 자동으로 결정하는 단계이다. <1,2단계>의 결과로 생성된 매핑관계와 기존에 공개되어 있는 PWN/SUMO간의 매핑관계 정보의 조합에 따른 최종 매핑관계는 다음과 같이 정의하였다.

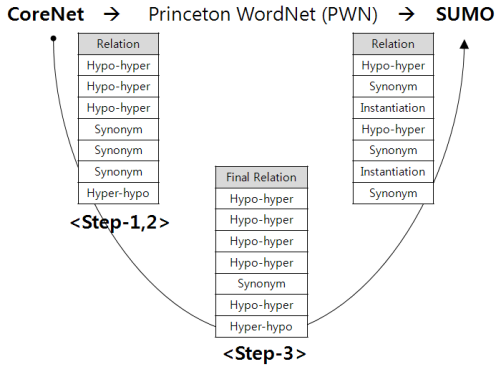

그림 6. CoreNet과 SUMO 사이의 매핑 관계 결정

Fig 6. Decision of mapping relations between CoreNet and SUMO

3.3 CoreNet-SUMO 직접 매핑

CoreNet을 SUMO에 직접 매핑 하려면 SUMO의 개별 클래스에 대한 이해뿐만 아니라 전체 클래스의 계층체계를 숙지하고 있어야 가능하다. 이 과정은 자동화하기 어렵기 때문에 전적으로 수작업에 의존해야 한다. 3.2절에서 획득한 간접 매핑 결과는 반자동으로 얻은 결과이기 때문에 그대로 최종 결과로 사용하기에는 문제가 있으나, 직접 매핑 작업의 초기 참고자료로 활용하기에는 아주 유용한 정보이다.

직접 매핑을 위한 절차는 다음과 같다.

<1단계>

SUMO에 속한 개별 클래스들의 의미를 파악하고, SUMO의 계층체계(다음 그림에서 일부예시)를 숙지한다.

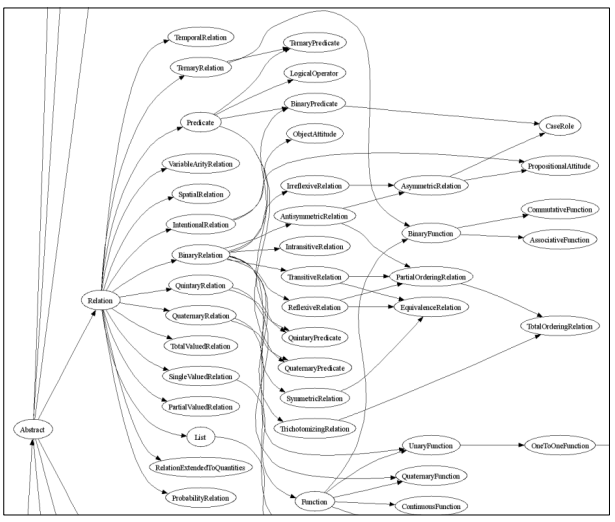

그림 7. SUMO 계층체계 일부

Fig 7. Part of SUMO hierarchy

<2단계>

이 단계는 간접 매핑 결과를 수작업으로 모두 검증하는 과정으로, 간접 매핑 결과를 보완하거나 삭제하기 위하여 SUMO를 직접 살펴보는 작업이 이루어진다. 또한 한국어로 되어 있는 CoreNet의 클래스명을 번역하여 영어 클래스명을 추가하는 작업도 같이 하였다.

4. 실험결과

총 2,937개의 CoreNet 클래스에 대해 3.2절에서 소개한 간접 매핑 작업을 한 결과, <1단계>에서는 2,295개(78.1%)의 클래스가 수작업으로 매핑되었고, <2단계>에서는 642개(21.9%)의 클래스가 자동으로 매핑되었다.

간접 매핑 절차에 따른 결과물을 검토한 결과, 3.2절 <1단계>의 결과(다음 그림의 아랫부분)는 수작업으로 이루어졌기 때문에 어느 정도 신뢰성이 있는 것으로 평가되었으나, 3.2절 <2단계>의 대상이 된 클래스들(다음 그림의 윗부분)은 주로 상위 계층체계에 위치하고 있어서 정확성이 더 요구됨에도 불구하고, 자동으로 매핑 되었기 때문에 오류가 상당수 포함되어 있었다. 그림의 첫 번째 필드는 CoreNet의 ID를 의미하며, 두 번째 필드는 CoreNet의 한국어 클래스명을, 세 번째 필드는 CoreNet 계층체계 상에서의 레벨(1~12)을, 네 번째 필드는 비단말(NT)/단말(T) 클래스 여부를, 다섯 번째 필드는 최종 매핑된 SUMO 클래스명과 그 매핑관계를, 여섯 번째 필드는 매핑된 워드넷의 신셋과 그 매핑관계 정보를 의미한다.

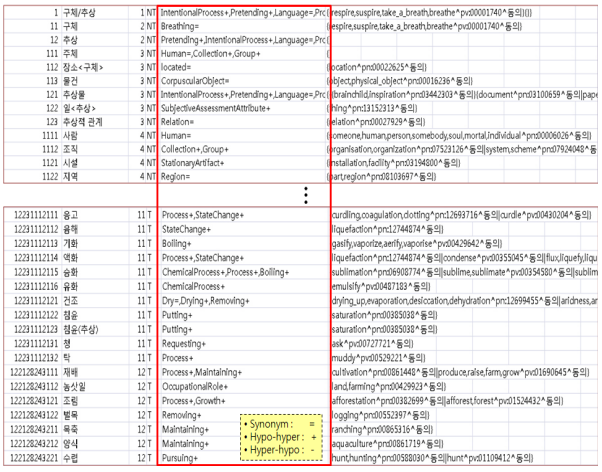

그림 8. 간접 매핑 결과 예시

Fig 8. Example of indirect mapping results

3.2절 <3단계> 과정에서는 PWN과 SUMO의 매핑 정보를 인터넷에 공개되어 있는 자원을 그대로 활용하였는데, 이 자원 또한 자동으로 생성되어서 그 품질을 완전히 신뢰할 수 없는 상태였다. 따라서 간접 매핑 결과를 최종 결과로 사용하기에는 미흡한 점이 있으므로, 이를 보완하기 위해 3.3절에서 소개한 직접 매핑 방법을 추가로 적용하였다.

총 2,937개의 모든 CoreNet 클래스에 대해 수작업으로 간접 매핑 결과의 검증 및 직접 매핑 작업을 수행하였다. 다음 그림은 직접 매핑을 거친 최종 결과의 일부(CoreNet의 최상위 계층)를 보여주고 있다.

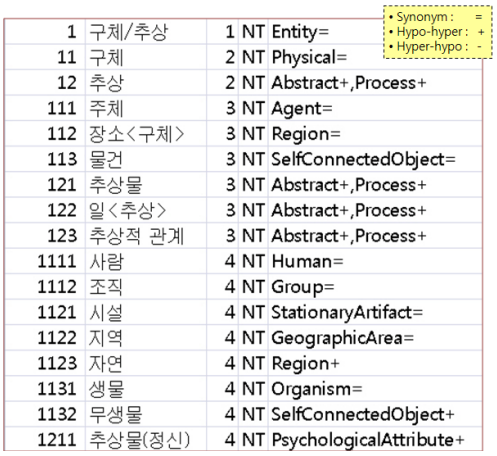

그림 9. 직접 매핑을 거친 최종 결과 예시

Fig 9. Example of final results through direct mapping

5. 결 론

CoreNet은 명사, 동사, 형용사의 품사를 아우르는 품사 독립적 개념체계이므로 자연어 텍스트의 의미 분석에서 단일의 개념망을 통한 추론이 가능하며, 다국어 텍스트의 분석, 언어 간 변환을 포함한 자연어처리 및 그 응용에 유용한 자원이다. 본 연구에서는 이러한 CoreNet 개념체계를 상위 온톨로지의 ISO 표준 후보인 SUMO에 연결하였다.

CoreNet과 SUMO를 매핑하기 위해 간접 매핑과 직접 매핑 방법을 모두 사용하였는데, CoreNet-KorLex-PWN-SUMO에 이르는 간접 매핑 작업을 통하여 한국어 중심의 CoreNet과 영어로 기술된 SUMO의 언어 간 변환의 어려움을 완화하고 CoreNet 개념에 대응하는 SUMO 클래스의 재현율을 극대화하였다. 또한 간접 매핑 결과를 매핑 후보로 활용하여 CoreNet-SUMO의 직접 매핑을 수행함으로써 사람 수준의 매핑 정확률을 얻고자 하였다.

CoreNet과 SUMO의 연결은 한국어, 일본어, 중국어에 제한된 CoreNet의 다국어 지원 폭을 국제 공용어인 영어로 확장할 뿐 아니라, 교착어인 한국어, 일본어, 고립어인 중국어와 함께 굴절어인 영어를 포함함으로써 CoreNet이 다양한 언어 유형에 대한 통합망으로 발전하는 계기가 될 것이다.

아울러 SUMO와 매핑된 CoreNet의 다국어 어휘의미망으로서의 국제적 위상을 제고하고, CoreNet의 공용화를 촉진하기 위해 NLP 사전의 국제표준언어격인 LMF로 패키징하는 작업을 수행하였다. LMF(Lexical Markup Framework)[6]는 자연어처리용 사전과 기계가독형사전을 위한 ISO 표준으로, 컴퓨터에 의한 자동 처리를 염두에 둔 사전의 제작, 사용, 병합, 정보교환 등을 위한 공통의 표준 프레임워크를 제공하는 것을 목적으로 하고 있다.

향후에는 본 연구의 결과물을 활용한 자연어처리 응용 프로그램을 개발하여 본 연구 결과물의 품질 및 활용 가능성을 객관적으로 입증할 수 있는 연구를 하고자 한다.

6) http://en.wikipedia.org/wiki/Lexical_Markup_Frame work, http://www.lexicalmarkupframework.org/

참 고 문 헌

[1] 최기선 외, *다국어 어휘의미망 제1권 어휘의미망 구축론*. 한국과학기술원 전문용어언어공학연구센터, KAIST Press, 2005.

[2] C. Biemann, S. I. Shin, and K. S. Choi, "Semiautomatic extension of CoreNet using a bootstrapping mechanism on corpus-based co-occurrences," *Proceedings of the 20th International Conference on Computational Linguistics*, Geneva, Switzerland, 2004.

[3] S. Ikehara, "Congratulatory address," *다국어어휘의미망*, 2005.

[4] K. S. Choi and H. S. Bae, "Procedures and Problems in Korean-Chinese-Japanese Wordnet with Shared Semantic Hierarchy," *Proceedings of Global WordNet Conference*, Brno, Czech Republic, pp. 91-96, 2004.

[5] S. Ikehara, et al, *The Semantic System, volume 1 of Goidaikei: A Japanese Lexicon*, Iwanami Shoten, Tokyo, 1997.

[6] I. Niles, and A. Pease, "Origins of the Standard Upper Merged Ontology: A Proposal for the IEEE Standard Upper Ontology," In *Working Notes of the IJCAI-2001 Workshop on the IEEE Standard Upper Ontology*, Seattle, Washington, August 6, 2001.

[7] C. Fellbaum, *WordNet: An Electronic Lexical Database (Language, Speech, Communication)*, MIT Press, 1998.

[8] A. S. Yoon, S. H. Hwang, E. R. Lee, and H. C. Kwon, "Construction of Korean Wordnet KorLex 1.5," *Journal of KIISE (Korean Institute of Information Scientists and Engineers): Software and Applications*, vol. 36, no. 1, pp. 92-108, 2009.

[9] 강신재, 강인수, "워드넷과 구글에 기반한 온톨로지 개체의 일반화," *한국지능시스템학회 논문지*, 제19권, 3호, pp. 363-370, 2009.

[10] I. S. Kang, S. J. Kang, S. J. Nam, and K. S. Choi, "Linking CoreNet to WordNet - Some Aspect and Interim Consideration," *Proceedings of the 5th Global WordNet Conference*, Mumbai, India, pp. 239-242, 2010.

저 자 소 개

강신재(Sin-Jae Kang)

1995년 : 경북대학교 컴퓨터공학과 공학사
1997년 : 포항공과대학교 컴퓨터공학과 공학석사
2002년 : 포항공과대학교 컴퓨터공학과 공학박사
1997년～1998년 : SK Telecom 정보기술연구원 연구원
2007년 : 오스트리아 University of Innsbruck, DERI 연구소 방문교수
2002년～현재 : 대구대학교 컴퓨터·IT공학부 부교수

관심분야 : 온톨로지, 시맨틱 웹, 자연어처리
Phone : 053-850-6584
E-mail : sjkang@daegu.ac.kr

강인수(In-Su Kang)

1995년 : 경북대학교 컴퓨터공학과 공학사
1999년 : 포항공과대학교 컴퓨터공학과 공학석사
2006년 : 포항공과대학교 컴퓨터공학과 공학박사
1995년～1997년 : 포스데이타
2006년～2008년 : 한국과학기술정보연구원
2008년～현재 : 경성대학교 컴퓨터학부

관심분야 : 자연어처리, 정보검색, 시맨틱 웹
Phone : 051-663-5147
E-mail : dbaisk@ks.ac.kr

남세진(Se-Jin Nam)

1999～2007 : K4M 기술연구소(연구소장)
2008년 : LG 데이콤 기술연구소(선임연구원)
2008년～2009년 : KAIST 시맨틱웹 첨단연구센터(팀장)
2009년～현재 : 서울대학교 치과대학 박사과정(의료경영과 정보학교실)

관심분야 : 시맨틱 웹, Literature-based Discovery, Information Retrieval, Text Mining
Phone : 010-3241-1523
E-mail : jordse@gmail.com

최기선(Key-Sun Choi)

2006년～현재 : KAIST 전산학과장
2006년 : 한국인지과학회장
2009년～2010년 : AFNLP (Asia Federation of Natural Language Processing) 회장
2002년～현재 : ISO/TC37/SC4 언어자원 운영표준 간사
2002년～현재 : AAMT(Asia Association of Machine Translation) 이사, Infoterm 부회장

관심분야 : 자연언어처리, Creative Linguistics, Wikipedia, Semantic Annotation, 문장 생성, Semantic Web, Ontology Engineering
Phone : 042-350-3501, 3525
E-mail : kschoi@cs.kaist.ac.kr