위키피디아를 이용한 영-한 개체명 대역어 쌍 구축

김은경○ 최기선
한국과학기술원 전산학과
{kekeeo, kschoi}@world.kaist.ac.kr

Extracting English-Korean Named-Entity Word-pairs using Wikipedia

Eun-kyung Kim○ Key-Sun Choi
Department of Computer Science, KAIST

요 약

본 논문은 공통적으로 이용할 수 있는 웹 환경에서의 한국어 정보로 획득할 수 있는 정보의 양이 영어권 정보의 양보다 상대적으로 적다는 것을 토대로, 웹정보 이용의 불균형을 해소하고자 하는 목적으로부터 출발하였다. 최근에는 지식 정보의 세계화, 국제화에 따라 동일한 정보를 각국 언어로 제공하고자하는 연구가 꾸준히 증가하고 있다. 온라인 백과사전인 위키피디아 역시 현재 다국어로 제공이 되고 있지만 한국어로 작성된 문서는 영어로 작성된 문서의 5% 미만인 것으로 조사되었다. 본 논문에서는 위키피디아 내에서 제공하는 다국어간의 링크 정보와 인포박스 데이터를 활용하여 위키피디아 문서 내에서 개체명을 인식하고, 자동으로 개체명의 영-한 대역어 쌍을 추출하는 것을 목표로 한다. 개체명은 일반 사전에 등재되지 않은 경우가 많기 때문에, 기계번역에서 사전 데이터 등을 활용하여 개체명을 처리하는 것은 쉽지 않으며 일반적으로 음차표기 방식을 함께 사용하여 해결하고 있다. 본 논문을 통해 위키피디아 데이터를 활용해 만들어진 영-한 개체명 대역어 사전을 구축하기 위해 사용된 기술은 추후 위키피디아 문서를 기계번역하는데 있어 동일한 방법으로 사용이 가능하며, 구축된 사전 데이터는 추후 영-한 자동 음차표기 연구의 사전 데이터로도 활용이 가능하다.

주제어: 위키피디아, 한국어, 개체명 대역어

1. 서 론

위키피디아[1]는 누구나 자유롭게 글을 쓸 수 있는 사용자 참여의 온라인 백과사전이다. 2001년에 시작되어 현재 전 세계 250여개 언어로 만들어가고 있으며 한국의 경우 2002년부터 시작되었다. 웹 2.0의 대표적인 모델로 거론되고 있으며 영어판이 가장 많이 이용되고 있다. 위키피디아 영어판의 경우, 2009년 8월 현재 약 300만개의 문서가 제공되고 있으며 한국어판의 경우 약 11만개의 문서가 제공되고 있다. 단순한 문서 개수 측면에서 보면 한국어로 제공되는 위키피디아 문서의 개수가 영어로 제공되는 것보다 현저히 적다는 것을 알 수 있다. 이는 공통적인 웹기반 정보를 이용하는데 있어 한국어권에서 획득할 수 있는 정보의 양이 상대적으로 적다는 것을 나타낸다. 본 논문은 이러한 웹기반 정보 획득의 불균형을 해소하고자하는 목적으로 시작된 연구의 첫 단계로 위키피디아 문서내의 '인포박스' 번역에 사용될 영어-한국어 개체명(Named Entity) 대역어 쌍을 구축하는 내용을 담고 있다.

인포박스란 위키피디아 하나의 문서에 나타난 중요 키워드에 대하여 정해진 템플릿을 이용하여 테이블 형태로 작성한 것을 나타낸다. 영어-한국어 간의 정보 균형화를 위하여 영어 위키피디아 문서 전체를 한국어로 번역하는 것이 아니라 인포박스만을 번역대상으로 한정한 것은 문서 전체를 번역 하는 것은 오역과 어색한 흐름의 문장이 많이 생성되는 문제점이 나올 수 있기 때문이다. 따라서 본 논문에서는 위키피디아의 중요한 데이터가 테이블 형태로 제공되고 있는 인포박스의 내용을 번역하는, 즉 문장 단위의 문서를 번역하는 것이 아니라 요약정보를 번역하는 것을 목적으로 하여 대역어 사전을 구축하는 내용을 담고 있다. 향후 구축된 대역어 사전을 기반으로 한 요약정보 번역 및 요약정보로부터 문서단위의 정보를 생성하는 기술에 대한 연구가 추가로 수행될 예정이다.

위키피디아의 인포박스를 번역하기 위해서는 용어 번역 기술이 있어야 하며, 이를 위해서는 용어 사전을 구축하고 이들에 대한 대역어를 부착하는 것이 첫 단계이다. 그러나 위키피디아 인포박스의 템플릿을 분석한 결과 인포박스 내에 표기된 정보는 사람이름, 장소, 조직 및 기관의 명칭등 개체명에 많이 치우친 것을 발견할 수 있었다. 따라서 인포박스 번역의 효율성을 높이기 위해서는 일반사전을 이용한 방식이나, 통계를 이용한 번역 기술을 사용한다 하더라도 개체명 번역 문제가 해결이 되지 않으며 이를 위해서 일반 용어의 대역어 사전 이외에 개체명 번역 사전이나 음차표기(Transliteration) 기술이 추가되어야 할것으로 보인다.

본 논문에서는 이러한 필요성에 의해 개체명 대역어 사전을 구축하는 것을 목적으로 하고 있으며 위키피디아 구조에서 획득할 수 있는 인터위키링크와 인포박스 템플릿 구조를 이용한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련연구에 대하여 기술하고, 3장에서는 제안하는 위키피디아 구조를 활용한 영어-한국어 개체명 대역어 사전 구축에 대하여 기술한다. 4장에는 실험에 대하여 기술하고 5장에서 결론을 맺는다.

2. 관련 연구

최근 몇 년 동안 위키피디아를 분석하는 연구가 활발히 진행되고 있다. 전 세계 사람 누구나 참여 가능한 온라인 환경 기반에서 시작된 위키피디아 커뮤니티는 급속한 세계화로 인해 현재는 하나의 독립적인 언어 뿐 아니라 다양한 언어로 문서 항목을 작성 및 검색 할 수 있는 거대한 규모로 자리 잡았다. 그러나 다양한 언어로 작성된 문서의 양이 늘어남에 따라 다국어로 작성된 동일한 내용의 문서들 사이에서도 관점, 문서의 크기, 문서에서 다루고 있는 범위, 문서를 작성한 시기 등 다양한 차이가 발견된다. 이러한 특징은 정보의 균형화를 깨트리는 문제로 발생할 수 있어서 현재 다국어로 작성된 위키피디아 문서들 사이에서의 불균형을 해소하고자하는 많은 연구가 진행되고 있다.

[4]는 위키피디아 문서의 링크, 특히 인터위키링크를 통하여 다국어 사이의 관계를 밝혔으며, [5] 역시 인터위키링크를 이용하여 개체명 번역에 사용될 사전을 구축하였다. 그러나 두 논문들은 서로 연결되어 있는 링크 구조를 분석한 것이며, 다른 구조화된 정보를 많이 사용하지 못했다. 본 논문에서는 다국어사이의 연결고리인 인터위키링크 이외에도 인포박스 구조를 활용하여 개체명 번역 사전의 효율적인 구축에 대하여 기술한다.

[6]에서는 Ziggurat 시스템을 통하여 다국어 사이에서의 위키피디아 인포박스 자동화 정렬 기법을 제안하였다. 그러나 메타데이터 형태로 구성되어 있는 인포박스 정렬에 있어서 단순히 하나의 데이터를 링크로 연결시켜주는 방식으로 문제를 간단 화하였다. 인포박스 정렬문제는 데이터베이스를 사용하는 환경에서 스키마의 맵핑과 정렬에 사용되는 기법과 유사하다고 볼 수 있다[7].

본 논문에서는 현재 인포박스 자체의 특징을 추출하여 현재 개체명 대역어 쌍을 추출하는 기법에 대해 논하고 있지만, 추후 이 기법을 확장하여 인포박스 자체의 특징을 분류하여 인포박스 정렬 문제를 보다 구조화된 입장에서 해결할 수 있을 것으로 보인다.

DBpedia[2]는 위키피디아 문서에 존재하는 구조화된 정보의 활용도를 높이기 위하여 구조화된 정보를 RDF[3] 형태로 제공하는 커뮤니티이다. 위키피디아의 구조화된 정보로는 문서의 제목을 가리키는 타이틀(Titles), 문서의 가장 첫줄에 위치한 정의문(Short Abstracts), 정의문을 포함해서 문서의 일부 내용을 추출한 요약문(Extended Abstracts), 문서 내에 표기된 이미지(Images), 하이퍼링크(Links to Wikipedia Articles), 문서가 속한 카테고리 정보(Articles Categories), 문서의 내용 중 중요한 정보를 테이블 형태로 제공하는 인포박스(Infoboxes) 등으로 구성되어 있다. DBpedia는 온라인 상에서 사람들이 만든 위키피디아의 문서 형식에서 데이터 구조를 찾아내고 이를 서로 링크해 주는 허브로 사용할 수 있다. 본 논문에서도 DBpedia에서 제공하는 RDF 형태의 위키피디아 데이터를 직접 다운 받아 사용하였다.

3. 위키피디아 구조를 활용한 대역어 추출

300만개의 영문 문서와 250여개의 언어별 정보를 제공하고 있는 위키피디아의 경우 각 위키피디아 내부 문서 사이에는 서로 연관되어있는 논리적인 링크들이 존재한다. 물론 문서의 주제어에 사용되는 링크, 카테고리 정보의 링크 등은 한 언어를 사용하고 있는 문서 집합에서의 관계를 연결시켜둔 것이다. 이외에도 위키피디아는 현재 250개 이상의 언어로 제공되고 있기 때문에 다국어 간에 같거나 혹은 비슷한 주제의 문서들이 있을 경우 인터위키링크(inter-language link, interwiki link)를 이용하여 서로 연결시켜 주고 있다. 예를 들면 영어 위키피디아에 등록된 ‘Microsoft’라는 제목의 문서와 한국어 ‘마이크로소프트’라는 문서의 제목이 인터위키링크를 통해서 상호 논리적인 연결 고리가 제공된다. 본 논문에서는 인터위키링크를 이용하여 손쉽게 영어와 한국어 용어의 대역어 쌍을 추출한다. 용어를 구축하고 이들에 대한 대역어를 부착하는 것이 용어 번역에서 최우선적으로 해결해야할 작업이기 때문이다.

본 논문에서는 기존의 유사한 논문과 다른 점으로 추출된 영어-한국어 대역어 쌍에 대하여 용어 타입을 분류하는 작업을 진행하였다. 또한 인터위키링크 정보 이외에 인포박스 구조를 활용하여 용어 타입 분류 작업에 이용했다. 현재까지의 연구에서는 개체명 타입에 한정하여 용어의 타입을 분류하였다. 이는 추후 개체명인식 기술, 개체명 번역 기술, 위키피디아 인포박스 정렬 문제, 인포박스 기계 번역 등 다양한 측면에서의 직접적인 활용도를 높이기 위해 고안되었다. 본 논문에서 사용된 개체명 타입은 일반적인 개체명 타입에서 사용되는 3가지, Person, Organization, Location 이다.

본 논문에서 제안하는 위키피디아 문서 구조를 활용한 영어-한국어 개체명 대역어 쌍을 추출하기 위해서 필요한 첫 번째 단계는 위키피디아 문서 구조에서 위에서 정의한 3가지 타입의 개체명을 인식하는 것이다. 두 번째 단계는 인식된 개체명에 따라 대역어 쌍을 구축하는 것으로 이루어진다. 본 장에서는 각 단계에 대하여 3.1과 3.2절에 나누어 자세히 기술한다.

3.1 인포박스 구조를 활용한 개체명 클루 추출

첫 번째 단계인 개체명을 인식하고 분류하기 위해서 본 논문에서는 위키피디아 문서 전체가 아닌 위키피디아 인포박스를 이용하여 개체명 후보를 추출한다. 위키피디아의 인포박스는 그림1과 같이 위키피디아 문서의 오른쪽에 위치하며 하나의 문서 내에서 중요 키워드에 대한 정보를 테이블 형태로 작성하여 제공하는 것이다. 위키피디아 그룹에서는 인포박스를 작성하기 위하여 여러 종류의 템플릿을 미리 제공하고 있으며 위키피디아에서 문서를 작성하고자 하는 유저는 해당 템플릿을 이용하여 인포박스를 쉽게 생성할 수 있다.

그림 1 위키피디아 문서의 인포박스

인포박스 템플릿은 2개의 단으로 이루어져 있으며 본 논문에서는 왼쪽 단을 'infobox-attribute'로 오른쪽 단을 'infobox-value'로 명명하여 사용한다. infobox-attribute는 데이터의 속성을 나타내며, infobox-value는 그 속성에 해당하는 실제 값을 의미한다. 예를 들어, Microsoft 인포박스의 경우 다음과 같은 인포박스 데이터를 포함한다.

- infobox-attribute: Founder(s)
- infobox-value: Bill Gates

인포박스 구조를 활용하여 개체명을 인식하기 위해서 본 논문에서는 인포박스 템플릿의 infobox-attribute를 이용한 방식을 제안한다. 즉, infobox-attribute를 통해 개체명 인식에 사용될 개체명 클루를 추출한다. 개체명 클루라는 것은 인포박스 내에서 사용된 용어가 개체명인지를 판단하기 위하여 infobox-attribute에서 개체명으로 사용될 수 있는 후보들을 나타낸다. 위키피디아 문서(인포박스 포함) 내에서 개체명 인식의 자동화를 위해서는 개체명 클루를 선별하는 작업 역시 자동화가 되어야 하지만 본 논문에서는 자동화 작업에 앞서 수동으로 개체명 클루를 선정하고 정확도와 결과 오류 분석 과정을 진행하였다. 이는 개체명 인식의 자동화 기술에 반드시 필요한 사전작업이다.

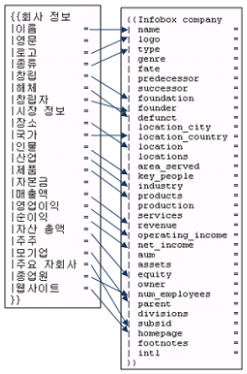

그림 2 회사 정보 템플릿의 속성 (한-영 매핑)

그림 2는 본 논문에서 사용한 인포박스 템플릿의 한 예를 보여준다. 한국어의 경우 '회사 정보 템플릿'을, 영어의 경우 한국어 회사 정보 템플릿에 대응되는 'Company 템플릿'의 infobox-attribute의 서로 대응되는 정보를 나타낸 그림이다. 이 대응되는 정보를 통하여 우리는 직관적으로 개체명으로 사용될 수 있는 개체명 클루(Clue)를 지정할 수 있다. '회사 정보 템플릿'의 경우 표 1과 같은 클루를 찾아낼 수 있었다. 앞에서도 설명하였듯이 위키피디아 인포박스에서 개체명 클루를 선정하는 작업은 현재 인포박스 템플릿 하나를 선정하여 수작업으로 진행하였으며 추후 위키피디아 링크 정보 및 연결 정보 등을 이용하여 개체명 클루 선정 방식을 자동화할 계획에 있다.

표 1. 회사 정보 템플릿에서 추출된 개체명 클루

개체명 타입	한국어 개체명 클루	영어 개체명 클루
Person	이름	name
Person	창립자	founder
Person	인물	key_people
Organization	본사	headquarters
Organization	주주	-
Organization	모기업	parent
Organization	주요 자회사	subsid
Location	장소	location

3.2 개체명 대역어 쌍 추출

두 번째 단계에서는 첫 번째 단계에서 추출한 infobox-attribute의 양국 개체명 클루를 이용하여, 이 클루들에 해당하는 infobox-value를 개체명 영-한 대역어 쌍 엔트리로 등록하는 방식으로 개체명을 인식하였다. 이때 같은 주제에 대하여 작성된 양국 위키피디아 문서의 내용이 100% 일치함을 보장하지 않기 때문에, 인포박스의 내용도 역시 100% 일치하지는 않는다. 따라서 인포박스 템플릿의 구조만을 분석하여 자동으로 대역어 후보를 추출하는 것은 잘못된 결과가 나올 수 있으며 데이터 자체의 비교를 통해서 영-한 대역어 후보를 얻어낼 수 있었다. 인포박스내의 infobox-value를 대역어 쌍 후보로 추출하는 방식은 infobox-value 데이터가 링크텍스트인 경우와, 일반 텍스트인 두 가지 경우로 나뉠 수 있다.

3.2.1 링크텍스트 기반 infobox-value 대역어 추출

infobox-value가 링크텍스트로 존재한다는 것은 위키피디아 내에서 해당 링크텍스트의 이름을 타이틀로 가지는 문서가 존재한다는 것을 나타낸다. 위키피디아는 현재 250개 이상의 언어로 제공되고 있기 때문에 다국어 간에 같거나 혹은 비슷한 주제의 문서들이 있을 경우 인터위키링크(inter-language link, interwiki link)를 이용하여 서로 연결시켜 주고 있다. 본 논문에서는 infobox-value가 링크텍스트로 존재한다면 인터위키링크를 이용하여 영어와 한국어 타이틀을 대역어 쌍에 추가하였다. 그러나 서론에서도 밝혔듯이 한국어 위키피디아 문서의 양이 영어 문서의 양의 5% 정도에 그치고 있어 영어 문서에 대해서 한국어 인터위키링크가 존재하지 않는 양이 방대하

다. 또한 반대의 경우도 발생할 수 있는데, 한국어권에서 잘 알려진 정보는 한국어 위키피디아 문서는 존재하지만, 영어 위키피디아 문서로는 제공되지 않는 것이 있었다. 이와 같은 사실은 각각의 용어에 대한 문서는 존재하지만 서로의 인터위키링크가 존재하지 않는 것으로, 영어나 한국어 문서 한쪽으로만 존재하는 것이 상당수 존재한다는 것을 나타낸다. 즉 본 논문의 초기 목적이었던 웹기반 정보 획득의 불균형 문제가 심각하다는 것을 알 수 있었다.

3.2.2 일반텍스트 기반 infobox-value 대역어 추출

infobox-value가 일반텍스트로 존재한다는 것은 위키피디아 내에서 해당 텍스트에 대한 문서가 존재하지 않는 것을 나타낸다. 이런 경우 위에서 기술된 방식인 인터위키링크를 이용하여 대역어를 추출할 수는 없다.

그림 3 “마이크로소프트” 영-한 인포박스 비교

그림 3은 같은 주제(Microsoft:마이크로소프트)에 대하여 작성된 영-한 인포박스의 내용 중 일부인데, 영어와 한국어 인포박스의 내용이 100% 일치하는 것은 아니라는 것을 알 수 있다. 본 논문에서는 영어 위키피디아 인포박스에만 존재하는 infobox-value에 해당하는 한국어 대역어를 찾기 위해 웹 검색 결과를 이용했다 웹 검색 결과를 이용하는 방식은 주로 한국어 문서에서 타언어 개체명을 표기할 때 주로 “한국어개체명(원개체명)” 또는 “원개체명(한국어개체명)” 방식을 통해 많이 표기되는 형태에서 착안하였다. 예를 들어, “마이크로소프트”의 “Don Mattrick”의 한국어 대역어를 찾고자 하면 검색어로는 다음 두 가지를 입력한다.

- 검색어1: 마이크로소프트
- 검색어2: “(Don Mattrick)“

웹검색을 통하여 나온 결과중에 “XXX(Don Mattrick)” 패턴들을 찾아내고 상위 30개의 결과 중에 가장 많이 표현된 대역어 후보를 한국어 대역어로 선정한다. 검색어1을 사용하는 이유는 웹 검색 대상 문서를 줄이고자 사용하였다.

4. 실험 및 오류 분석

영-한 개체명 대역어 사전을 구축하는 대상 용어는 3장에서 제시된 회사 정보 템플릿에서 추출된 개체명 클루를 포함하는 인포박스를 가진 한국어 위키피디아 문서의 infobox-value를 대상으로 하였다. 인포박스는 위키피디아 덤프 서비스에서 제공하는 데이터[1]를 다운받아 파싱을 해서 얻어내거나 디비피디아[2] 데이터를 이용하여 RDF 형태로 얻어낼 수 있다. 본 논문에서는 위키피디아 덤프 서비스 에서 제공하는 한국어 위키피디아 전체 문서(XML형식-2009년 8월)를 다운받아 사용하였으며 516M의 용량의 234,216 개의 문서로 구성되어 있었다. 영어 데이터는 DBpedia에서 제공하는 RDF 형태의 데이터를 다운받아 사용하였다. 버전은 DBpedia 2.0이며 2008년에 작성되었다. 위키피디아 전체 문서를 직접 다운 받아 사용하는 경우에는 각 문서에서 인포박스를 찾아내기 위하여 위키문법을 분석, 문서 내의 인포박스의 위치를 찾아낼 수 있었다.

위키피디아 한국어판의 전체 문서 중 회사 정보 인포박스를 포함한 위키피디아 한국어 문서의 개수는 총 789개였다. 이중에 영어-한국어 대역어 쌍을 추출하기 위하여 789개의 한국어 문서 중에 같은 주제로 영어 문서가 존재하는 문서들만을 대상으로 하였으며 그 개수는 총 390개였다. 390개의 문서 중 표1에 기재된 개체명 클루에 해당하여 추출된 개체명은 총 1,134개였으며 Person, Organization, Location 의 통계는 그림 4와 같이 구성되었다. 추출된 데이터의 평가를 위하여 3명이 각각 전수검사를 하였으며 개체명 분류의 평균 정확도(Precison)는 약 96%, 재현률(Recall)은 70%를 기록했다. 인포박스 템플릿을 분석하고 개체명 타입에 사용될 infobox-attribute를 미리 정해두었기 때문에 개체명 분류의 정확도는 무척 높은 것을 알 수 있었으나 위키피디아에서 제공되고 있는 현재 문서의 버전과 본 논문의 실험에서 사용된 DBpedia 데이터의 버전 차이가 존재하여 재현률이 상당히 낮은 것으로 분석되었다. 대역어 쌍을 추출하는 부분은 위키피디아 자체의 인터위키링크 및 웹 검색 결과를 이용하였기 때문에 별도의 성능 측정은 실행하지 않았다.

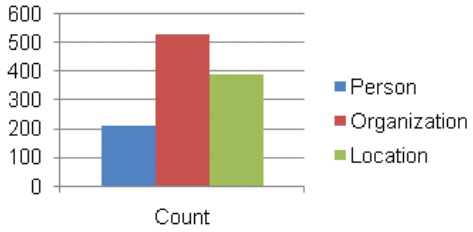

그림 4 개체명 타입 별 개수

실험 과정을 통해 인포박스를 이용하여 개체명을 분류하는데 있어서의 문제점은 크게 조직(Organization)을 구분 짓는 문제에서 발생되었다. 조직으로 판단을 하기 위해 사용된 infobox-attribute는 총 4가지였으며 본사, 주주, 모기업, 자회사 가 해당사항이었다. 그러나 “본사”의 경우 장소에 해당하는 정보가 상당수 분포되어 있었

다. 그 결과로 인해 infobox-attribute의 "본사" 경우는 개체명 타입을 조직으로 100% 판단할 수 없는 클루로 판단되며, 실제 이와 같은 문제점을 극복하기 위하여 클루를 선정하는 방식을 변경해야 할 것으로 보인다. 현재 개체명 타입을 분류하기 위한 인포박스 클루 선정을 자동화하기 위한 연구가 진행 중에 있다. 자동화하기 위해서는 infobox-attribute 에 나타난 단어 뿐 아니라, infobox-value 에 등장하는 용어의 패턴 및 링크 구조 등을 활용할 수 있다.

또 발견된 문제점으로는 사람(Person)과 조직에 중복적으로 나타날 수 있는 용어들도 발견되었다. 대표적인 예로는 "월트_디즈니(Walt_Disney)"로 이 용어는 사람의 이름이면서 회사의 이름을 가질 수 있는 것으로 판단되었다. 본 논문에서는 각 개체명의 영어-한국어 대역어 추출이었기 때문에 여러 개체명 타입으로 중복적으로 존재한다하더라도 문제가 되지 않았다. 하지만 이와 같은 문제는 개체명 인식에 있어서 아주 기본적으로 해결해야 할 과제로, 추후 문장에서 해당 용어가 나왔을 때 어떤 개체명으로 인식할지에 관한 문제를 해결하기 위해서는 문장 내에서의 주변 컨텍스트를 활용하여 개체명을 분류하는 방법으로 해결할 수 있을 것으로 보인다.

5. 결 론

본 논문에서는 위키피디아 문서 구조를 이용하여 영-한 개체명 대역어 쌍을 추출하는 방법에 대하여 기술하였다. 본 논문에서는 위키피디아 인포박스의 내용 중 infobox-attribute의 정보를 개체명을 인식하는 클루로 지정하고 개체명의 대역어 쌍을 추출하였다. 용어의 대역어 쌍을 추출하기 위해서는 위키피디아 문서 구조를 활용, 인터위키링크를 통해 영-한 대역어를 추출하는 방법을 제안하였다. 또한 인터위키링크가 존재하지 않는 용어에 대해서는 웹 검색 결과를 이용하여 새로운 대역어 쌍을 추가하는 방식으로 진행하였다.

기존에 이미 위키피디아의 인터위키링크를 이용하여 다국어 사이의 의미적인 관계를 밝히는 연구가 많이 진행되었지만, 인포박스 구조를 활용하여 용어의 타입을 분류하는 연구는 미비한 상태였다. 본 논문에서는 인터위키링크를 통하여 대역어 쌍을 추출하는 기존 연구 기법에 인포박스의 구조를 활용하여 개체명을 쉽게 분류할 수 있는 방법을 제시하였다. 하지만 개체명을 분류하기 위한 방법으로 현재는 인포박스를 수작업으로 분석하였지만 향후 자동화 할 수 있는 방법에 대한 보완이 필요할 것으로 생각된다.

본 논문의 기법은 기존 연구와 유사하게 추후 기계번역을 위한 사전 자원으로 활용가치가 높다. 기계번역 시에 사용하기 위한 사전 데이터로는 기존에 많은 영한 사전이 제공되고 있지만 개체명의 경우 사전에 등재되어 있지 않은 경우가 많은 비율을 차지했다. 본 논문에서는 기계 번역시에 개체명의 쉽고 빠른 번역을 위하여 개체명 대역어 사전을 구축하였다. 또한 추출된 영-한 대역어 쌍의 데이터를 사전으로 삼아, 영어 위키피디아 문서에 나타나는 용어를 번역할 수 있다. 번역된 한국어 용어를 바탕으로 문장을 생성할 수 있는 기반 기술로 사용한다면 한국어 위키피디아 문서 획득에 도움이 될 것으로 기대되며 현재 연구가 진행 상태에 있다. 이를 통해 한국어로 존재하지 않는 영어 문서와 같거나 혹은 비슷한 주제의 한국어 문서를 생성할 수 있는 기술을 확보할 수 있을 것으로 기대된다.

[참고문헌]

[1] http://download.wikimedia.org/

[2] http://dbpedia.org/

[3] www.w3.org/RDF/

[4] Tyers, F., Pienaar, J., 2008. Extracting bilingual word pairs from Wikipedia. In: Proceedings of the SALTMIL Workshop at Language Resources and Evaluation Conference, LREC'08. 2008

[5] Wentald, W., J. Knopp, C. Silberer, and M. Hartung, "Building a Multilingual Lexical Resource for Named Entity Disambiguation, Translation and Transliteration," LREC '08, Marrakech, Morocco, May 2008.

[6] E. Adar, M. Skinner, and D.S. Weld. Information arbitrage across multi-lingual Wikipedia. In Proceedings of the Second ACM International Conference on Web Search and Data Mining, pages 94-103. ACM New York, NY, USA. 2009.

[7] Rahm, E., and P. A. Bernstein, "A survey of approaches to automatic schema matching," VLDB Journal, 10:334-350, 2001.