번역과 웹그래프를 활용한 언어 간 위키피디아 인포박스 자동생성 기법

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

An Approach to Automatically Generating Infobox for Wikipedia in Cross-languages through Translation and Webgraph

Eun-kyung Kim○, DongHyun Choi, Eun-Bi Go, Key-Sun Choi
Semantic Web Research Center, KAIST

요 약

여러 언어로 작성되는 위키피디아의 경우 언어 간에 등록되어 있는 정보의 양과 내용이 달라 언어 간 정보를 상호 추출하고 서로 통합하는 연구에 대한 관심이 증가하고 있다. 특히, 위키피디아의 요약본으로써 의미가 있는 인포박스는 위키피디아 아티클에 존재하는 구조화된 정보 중 가장 근간이 되는 정보로, 본 논문에서는 위키피디아에 존재하는 인포박스를 1)소스 언어 자원으로부터 획득하여 타겟 언어로 번역하고, 2)번역된 결과물과 웹그래프를 이용하여 타겟 언어 데이터에서 획득하는 정보와 결합하는 과정을 통해 자동으로 인포박스를 생성하는 기법에 대하여 설명한다. 웹그래프는 위키피디아에 존재하는 링크 구조를 통해 서로 다른 두 용어간의 관련도를 측정하여 인포박스에 추가될 내용을 파악하는데 사용한다. 본 논문의 기법은 언어 간 인포박스를 생성하는 측면에서, 영어 인포박스 데이터를 입력으로 하여 한국어 인포박스 데이터를 생성하는 방식으로 진행하였다. 평가를 위하여 기존 한국어에 실제 존재하는 인포박스 데이터와 비교 실험하는 방식을 사용하여 평균적으로 40%의 정확률과 83%의 재현율을 나타내었다. 하지만, 기존 한국어에 존재하는 인포박스 데이터의 내용이 인포박스에 포함될 완전한 데이터를 모두 포함했다고 볼 수 없으므로 본 논문에서 제안하는 수행한 실험의 정확률이 상대적으로 낮게 나온 것으로 분석되었다. 실제 사람이 수작업으로 새롭게 생성된 인포박스 데이터의 적합성을 판별한 경우 평균 76%의 정확률과 91%의 재현율을 나타내었다.

주제어: 위키피디아, 인포박스, 웹그래프

1. 서론

위키피디아[1]는 웹(Web)상에서 누구나 목적에 관계없이 자유롭게 사용할 수 있는 공개형 백과사전으로 위키 문법(Wiki Syntax) 기반의 오픈 커뮤니티 웹사이트이다. 위키피디아 페이지는 아티클(article)이라는 단위로 불리며 하나의 아티클은 독립적인 개체(entity)에 대한 설명을 하는 일반 텍스트 부분, 부가적인 정보를 제공하는 이미지, 하이퍼링크, 카테고리, 템플릿 등 다양한 구조화된 텍스트로 구성되어있다. 그 중 템플릿(template)은 위키에서 제공하는 유용한 기능 중 하나로 여러 페이지 내에서 반복적으로 사용되는 특정 부분을 미리 함수처럼 정의해 놓고 여러 페이지에서 각 페이지에 맞게 인스턴스만 새롭게 추가하여 동일한 템플릿을 반복적으로 사용하는 것을 말한다. 이것은 프로그래밍 언어에서 외부함수를 불러 사용하는 기능과 동일하다. 위키피디아에서 제공하고 있는 템플릿 중 가장 많이 사용되고 있는 것은 인포박스이다. 인포박스(infobox)는 아티클 상단 부분에 위치하며 하나의 아티클에 대한 요약을 제공하는 일종의 메타데이터로 각 개체에 대한 주요 속성과 그에 대한 값을 기술하는 형태로 작성한다. 예를 들어, "소크라테스1)"의 경우 '철학자'로 간주되는 인물의 정보를 요약하여 나타내기 위하여 한국어 위키피디아에서는 "소크라테스"아티클 내에'철학자 정보'라는 인포박스를 사용하고 있으며[2] 주요 속성으로는 '지역', '시대', '출생일', '학파', '연구 분야'등이 포함되어 있다.

위키피디아는 문화, 식물, 이벤트, 인물 등 다양한 주제에 걸쳐 정보가 풍부하고, 위키피디아 컨텐츠가 SQL과 XML파일 형태로 무료로 제공되어[3] 연구자들이 쉽게 접근하여 언어자원으로 활용할 수 있어 최근 많은 연구에서[4-7] 위키피디아를 바탕으로 연구가 활발히 진행되었다. 위키피디아의 또 하나의 장점은 여러 나라 언어별로 각 커뮤니티가 구성되어 제공되고 있는 것이다. 2011년 9월 현재 약 270개 이상의 언어가 등록되어 있다. 그 중 가장 대규모인 영어 위키피디아의 경우 약 373만개의 엔트리가 등록되어있으며 이는 전체 위키피디아 규모의 약 30%를 차지한다. 등록되어있는 아티클 규모 상 약 20위권인[8] 한국어 위키피디아의 경우 약 17만개의 엔트리가 제공되고 있다. 그림 1은 대표적인 언어 별 위키피디아에 등록된 아티클 개수를 비교한 그래프이다. 본 그래프가 나타내듯 언어별로 위키피디아가 포함하고 있는 정보의 양이 다르다는 것을 알 수 있다. 이는 상대적으로

1) http://ko.wikipedia.org/wiki/소크라테스

적은 양의 엔트리가 등록되어있는 중국어, 한국어, 아랍어 권의 위키피디아 사용자가 모국어로 획득할 수 있는 정보의 양이 적다는 것을 의미한다. 본 논문에서는 이러한 정보의 불균형을 해결하고자, 다른 언어로 존재하는 위키피디아 컨텐츠를 번역하여 새로운 데이터를 생성하는 기법을 제안한다. 특히, 위키피디아의 인포박스(infobox)에 대한 실험을 통해 언어 간 인포박스를 자동으로 생성하는 기법을 제안한다. 언어 간 정보 이동을 위해 사용되는 기본 작업인 번역 이외에, 이종 언어 간 편향된 정보를 해소하기 위하여 본 논문에서는 위키피디아의 링크구조를 통한 웹그래프를 형성하여 인포박스에 추가될 수 있는 관련 용어를 선별하는 작업을 추가적으로 진행하였다. 인포박스를 자동으로 생성하는 것은 하나의 개체에 대한 주요 속성에 대한 기술을 자동으로 생성하는 것으로 이는 추후 문서를 자동으로 생성해주는 기술의 근간으로 활용될 수 있다.

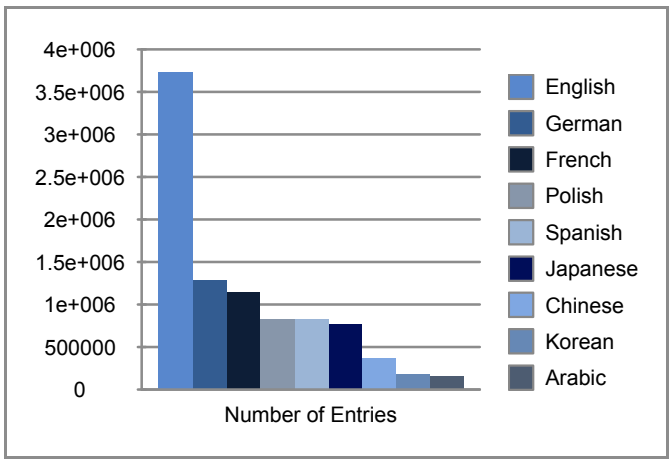

그림 1 언어 별 위키피디아 등록 아티클 수 비교

본 논문의 구성은 다음과 같다. 2장에서는 기존의 위키피디아를 활용한 다국어 활용 기술에 대하여 설명하고, 3장에서는 본 논문의 기법에 대하여 설명한다. 4장에서는 실험과 그 결과에 대하여 설명하고 5장에서는 결론을 맺는다.

2. 관련연구

DBpedia[9]는 위키피디아 내에 존재하는 다양한 구조화된 정보를 다른 연구자들이 쉽게 사용할 수 있도록 트리플 형태로 제공하는 오픈 커뮤니티이다. DBpedia가 제공하는 데이터 중 가장 근간이 되는 데이터는 인포박스 데이터이며, DBpedia에서는 인포박스 템플릿에서 추출된 정보를 <subject-property-object> 형태의 트리플 형태로 제공한다. 그러나 각 언어권별로 인포박스 내에 사용된 인코딩, 포맷(통화정보, 날짜정보, 등)의 사용이 다양하여 공통된 인포박스 추출 프레임워크를 이용하여 다양한 언어 위키피디아에 존재하는 정보를 손실 없이 추출하기에는 무리가 있다. 이에 따라 최근 DBpedia에서는 DBpedia Internationalization Committee[10]를 통해 각 나라의 언어별 특성에 따른 다양한 프레임워크를 추가 개발하여 제공한다. 2011년 현재 8개 나라가 참여하여 DBpedia 공통 추출 프레임워크에 각 언어의 특성에 맞는 확장 모듈을 추가하여 업데이트하는 방식으로 개발되고 있다. 한국어 특성에 맞게 추가된 DBpedia 추출 프레임워크는 2010년 개발되어 공통프레임워크를 이용한 추출방식보다 향상된 추출성능을 보이고 있다[11]. 본 논문에서 제안하는 기법은 각 언어별 위키피디아에서 추출된 정보를 이용하여 이종 언어별 상호 보완해주는 방법으로 활용되어 추후 DBpedia에 추가될 수 있는 새로운 정보를 찾거나 언어 간 데이터에서 누락된 정보를 찾는데 활용될 수 있다.

WikiBhasha[12] 는 마이크로소프트 연구소에서 개발한 위키피디아 다국어 컨텐츠 제작 도구이다. 이 툴은 위키피디아 아티클을 작성할 때, 다른 언어의 위키피디아 아티클이 존재하는 경우, 해당 아티클의 내용을 마이크로소프트 기계 번역 툴로 번역하여 자동으로 채우고 그 후에 사용자에게 번역된 아티클을 수정할 수 있는 에디터를 제공하여 다른 언어로 페이지를 작성할 수 있게 도와주는 일종의 컨텐츠 제작 도우미 시스템이다. 위키피디아 아티클을 작성하고자할 때 소스 언어의 데이터의 번역을 통해 타겟 언어로 아티클을 작성하는 기초를 제공하는 것은 본 논문에서 제안하는 방식의 목적과 유사하나, WikiBhasha는 번역된 초기 결과를 타겟 언어에 맞추어 사용자가 손쉽게 수정해나갈 수 있는 에디터 기능을 제공한다면, 본 논문에서는 번역을 통해 생성된 초기 데이터에 타겟 언어에서 추출된 정보를 바탕으로 두 개 이상의 언어권 데이터에서 발생되는 정보를 통합하여 보다 언어 간 관점의 균형을 이룬 데이터를 생성하는 방식에 중점을 두고 있다.

3. 인포박스 자동생성 알고리즘

인포박스는 위키피디아에 등록된 하나의 개체에 대한 주요 속성을 기술한 일종의 요약데이터이며 아티클 내에 테이블 형태로 출력된다. 그림 2는 아티클 '서울특별시'의 한국어 인포박스와 그에 대응하는 영어 인포박스의 일부를 나타낸 것으로 해당 개체를 대표하는 속성 필드들과 그에 해당하는 값이 정의되어있다. 그림 2에서 볼 수 있듯이 동일한 개체에 대한 인포박스를 구성하는 데이터는 서로 다른 언어 간 상당히 유사한 형태를 보여주는 것을 알 수 있다. 본 논문에서는 이런 특징을 이용하여 두 개의 서로 다른 언어인 소스 언어(source language), 타겟 언어(target language)가 주어졌을 때, 소스 언어로 작성된 인포박스를 입력으로 하여 타겟 언어의 인포박스로 자동 생성하는 방법에 대하여 기술한다. 이 방법을 통하여 정보가 상대적으로 부족한 타겟 언어 데이터를 자동으로 생성하는 효과를 기대할 수 있다.

Country	South Korea
Region	Seoul National Capital Area
Districts	25 [show]
Government	
- Type	Seoul Metropolitan Government
- Mayor	Oh Se-hoon
Area[2]	
- Total	605.25 km² (233.7 sq mi)
Population (2010[3])	
- Total	10,464,051
- Density	17,288.8/km² (44,777.8/sq mi)
- Demonym	Seoulite,서울시민(Seoul simin)
- Dialect	Seoul
Flower	Forsythia
Tree	Ginkgo
Bird	Korean Magpie
Website	seoul.go.kr

현황	
국가	대한민국
면적	605.25 km²
인구	10,464,051[1] 명
인구밀도	16,182.25 명/km²
행정구역	25구 0군
시장	권영규 (시장 권한대행)
상징	
시목	은행나무
시화	개나리
시조	까치
마스코트	해치
시청	
시청소재지	중구 덕수궁길 29 (서소문동 37)
웹사이트	http://www.seoul.go.kr

그림 2 인포박스 일부(예: '서울특별시')

본 논문에서 제안하는 두 개의 서로 다른 언어 간 인포박스 자동생성 알고리즘은 그림 3과 같은 단계로 진행된다. 3.1절에서는 제안하는 언어 간 인포박스 자동생성 기법의 첫 번째 단계로, 소스 언어의 인포박스 트리플을 타겟 언어로 번역하여 타겟 언어로 작성된 인포박스 트리플을 획득하는 방식에 대하여 설명한다. 3.2절에서는 제안하는 기법의 두 번째 단계인, 타겟 언어로 번역된 인포박스 트리플을 웹그래프를 이용하여 확장하는 기법에 대하여 설명한다. 번역만 이루어진 상태의 첫 번째 단계의 결과물은 소스 언어의 관점에서 작성된 트리플 정보만을 포함하고 있다. 즉 타겟 언어의 인포박스를 생성하고자할 때, 타겟 언어의 입장에서 관련된 정보를 추가적으로 추출하여 편향된 정보를 보충해주는 것이 두 번째 단계의 목적이다.

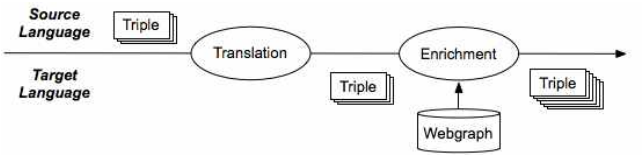

그림 3 전체 프로세스

3.1 1단계: 트리플 번역(Triple Translation)

인포박스 데이터는 SPO:<subject-property-object>형태인 트리플(triple)로 표현할 수 있다. 예를 들어 그림 2에 나타난 영어 인포박스(왼쪽)의 예제의 경우, 다음 표 1과 같은 트리플을 얻어낼 수 있다. 트리플의 subject는 위키피디아 '아티클 제목'으로, property는 '템플릿이름:템플릿속성'으로, object는 각 property에 해당하는 '값'으로 구성된다.

표 1 Seoul 아티클에서 추출된 트리플 일부

<Seoul, Settlement:Flower, Forsythia>
<Seoul, Settlement:Tree, Ginkgo>
<Seoul, Settlement:Bird, Korean Magpie>
<Seoul, Settlement:Website, seoul.go.kr>

두 개의 서로 다른 언어 간 작업을 위해서는 기본적으로 언어 간 변환, 즉 번역 작업이 필요하다. 본 논문에서는 소스 언어로 구성되어있는 인포박스 트리플을 타겟 언어로 번역하기 위하여 사전을 이용한 번역 방식을 사용한다. 사전을 이용한 번역을 위해서는 두 개의 언어 간 대응되는 대역어 쌍을 구축하여 번역 프로세스의 사전 자원으로 사용하여야 한다. 본 논문에서는 인포박스 트리플을 효과적으로 번역하기 위해서 위키피디아에서 추출할 수 있는 대역어 쌍으로 사전 자원을 구축하였다(이하 '위키대역어사전'으로 칭함).

3.1.1 위키피디아 언어 간 링크를 활용한 영-한 위키대역어사전 구축

다양한 언어로 이루어진 위키피디아에서는 서로 다른 언어로 작성된 동일한 개체에 대한 언어 간 링크(interlanguage link)를 제공하고 있다. 이 정보로 두 언어 간 대역어 쌍을 추출하여 위키대역어사전을 구축할 수 있다[13]. 특히 위키대역어사전 구축작업은 동음이의어에 대한 분류가 되어 있는[2] 위키피디아의 엔트리에 대하여 수행하므로 다른 사전을 이용한 번역시 발생할 수 있는 모호성 문제를 별도로 처리 하지 않아도 되는 장점을 가지고 있다. 위키 언어 간 링크는 위키피디아 페이지 내 사이드바 언어 섹션에 제공되며 링크를 표시하는 "[[...]]"형태의 위키문법과 국가코드를 활용하여 표시한다. 예를 들어, 영어 위키피디아 페이지 'Korean_Magpie' 의 경우 다음과 같이 총 3개의 언어 간 링크(A,B,C)가 존재한다.

표 2 언어 간 링크 예제

A. [[eo:Korea pigo]]
B. [[ko:한국까치]]
C. [[zh:朝鮮喜鵲]]

B의 경우 앞의 두 캐릭터 'ko'는 한국어를 나타내는 국가코드이고 ':' 뒤에 존재하는 용어가 영어 위키피디아 아티클인 'Korean_Magpie'에 대응되는 한국어 위키피디아 아티클 이름이 된다. 이 방법을 통해 ('Korean_Magpie', '한국까치') 가 영-한 위키대역어 사전의 엔트리 하나로 추가된다. 이와 마찬가지로 영-중 위키대역어 사전의 엔트리에는 ('Korean_Magpie', '朝鮮喜鵲')가 추가될 수 있다.

3.1.2 트리플 번역

트리플 형태로 구성된 인포박스를 번역하기 위해서 본 논문에서는 트리플의 구성요소에 따라 각기 다른 번역방식을 사용한다.

2) 예를 들어, 영어 단어 Tree의 경우 여러 뜻을 가지는 단어로 사용될 수 있으나, 일반 식물을 가리키는 단어는 "Tree"표기하며 자료 구조에 사용된 용어의 경우 "Tree_(data_structure)"로 구분해 표기하여 사용되고 있다.

- **Subject 번역:** subject는 위키피디아에 등록되어 있는 엔트리이므로 본 요소의 번역을 위해서는 3.1.1절에서 구축된 위키대역어사전이 사용된다.

- **Property 번역**: property에 사용되는 용어들은 표 1에 나타난 예와 같이 ''settlement', 'tree'등 매우 일반적으로 사용되는 간단한 단어들이 사용되고 있다. 이를 번역하기 위해서 본 논문에서는 위키대역어사전, 구글 번역[14], 다음 사전[15], 네이버 사전[16] 총 네 가지 사전 자원을 이용하는 방식으로 진행한다.

- **Object 번역**: object는 subject, property와는 다르게 다양한 형태의 정보가 표기될 수 있다. 예를 들어, 숫자, 날짜정보, 이미지, URL 등이 올 수 있다. 효과적인 번역작업을 위해 정규표현식과 패턴을 사용한 전처리과정을 거쳐 특정 데이터는 번역 작업 없이 변환만 하여 표기하며, 텍스트가 포함된 정보는 위키대역어사전을 활용하여 번역한다. 위키대역어사전에 해당하는 엔트리가 없어 번역이 되지 않는 것은 소스 언어 그대로 표기하는 것을 원칙으로 한다.

3.2 2단계: 웹그래프를 이용한 트리플 확장(Triple Enrichment with Webgraph)

3.1절을 통해 번역된 트리플은 웹그래프를 이용해 확장될 수 있다. 본 기능의 목적은 언어 간 번역을 통해 얻어진 정보는 소스 언어에서 획득된 정보이므로 타겟 언어 입장에서의 관련 정보를 부가적으로 추가하여 정보의 편향성을 해소하고자 하는 단계이다.

3.2.1 위키피디아 웹그래프 생성

웹그래프[17]는 웹에 존재하는 두 개의 페이지간의 링크정보를 바탕으로 만들어지는 일종의 다이렉티드 그래프(directed graph) 구조를 일컫는다. 웹그래프상의 노드는 웹 페이지로 구성되며 하이퍼링크를 통하여 연결되는 두 개의 서로 다른 웹 페이지간의 연결 관계가 두 개의 노드간의 에지(edge)로 구성된다. 위키피디아 웹그래프를 생성하는 방법은 위키피디아내에 존재하는 두 개의 아티클간의 인터널 링크정보를 통합하여 구성한다. 인터널 링크(internal link, 이하 링크)는 위키피디아에 존재하는 페이지간의 참조 관계 혹은 네비게이션 요소를 표현하는 하이퍼링크다. 예를 들어, 위키피디아에 세 개의 아티클 '대전광역시', 'KAIST', '포항공과대학교'가 존재하고, <'대전광역시','KAIST'>가 링크로 연결되어있고 <'KAIST', '포항공과대학교'>가 링크로 연결되어 있는 경우, 그림 4와 같은 다이렉티드 그래프를 얻을 수 있다.

그림 4 다이렉티드 그래프

본 논문에서는 위키피디아 웹그래프를 생성하기 위해 위키피디아 페이지에 존재하는 모든 링크를 추출하여 두 개의 서로 다른 페이지간의 연결관계를 구성한다. 위키피디아에서 링크는 "[[...]]"형태의 위키 문법을 이용하여 사용된다. 예를 들어, "서울특별시"라는 페이지에서 다음과 같은 문장이 표기된 경우,

> 서울특별시는 [[대한민국]] 북서부에 있는 대한민국의 최대 도시이자 [[수도]]이다.

<서울특별시, 대한민국>, <서울특별시, 수도> 두 개의 노드 페어를 추출할 수 있다. 이렇게 추출된 노드 페어를 구성하는 두 개의 노드간의 관련도, 즉 웹그래프의 에지 가중치(edge weighting)은 다음과 같은 방식을 이용하여 결정된다.

3.2.2 웹그래프의 노드간 관련도 측정

그래프상의 에지 가중치는 하나의 에지를 구성하는 두 개의 노드 간의 관련도로 계산할 수 있다. 본 논문에서 노드간의 관련도는 크게 두 가지 방식으로 계산된다.

- **컨텍스트 상의 관련도 측정:** 전체 대상이 되는 문서에서 해당 용어들이 얼마나 자주 같이 사용되었는지 여부를 바탕으로 유사도를 측정하는 방식이다. 본 논문에서는 TF-IDF와 유사한 방식의 LF-IDF를 정의한다. TF-IDF 가중치[18]는 언어 자료 내의 특정 문서에서 어떤 단어의 중요도를 평가하기 위한 통계적인 수치이다. 즉, 단어의 중요도는 문서 내에서 해당 단어가 많이 나타날수록 증가하며, 전체 언어 자료 내에서 해당 단어가 많이 나타날수록 감소한다. 이와 동일한 방식으로 LF-IDF는 위키피디아 아티클과 그에 사용된 링크 텍스트간의 관련도를 측정한다. 3.2.1절에서 구성된 웹그래프 간의 노드 페어는 <아티클 제목, 링크텍스트> 페어로 구축되기 때문에 아티클제목과 링크텍스트의 관련도를 측정할 수 있다. $lf(l,a)$는 아티클 a에서 링크 l의 중요도를 나타내며, $df(l)$은 해당 링크 l을 포함하고 있는 아티클의 수를, N은 전체 아티클의 수를 나타내며 LF-IDF 수식은 다음과 같다.

$$lfidf(l,a) = lf(l,a) \cdot idf(l)$$

$$idf(l) = \log \frac{N}{df(l)}$$

- **토폴로지 상의 유사도 측정:** 웹 그래프를 구성하는 노드 페어가 하나의 문서에 동시에 나타나지 않아도 관련도를 추가하기 위해서 사용되는 방식이다. 이는 앞서 사용된 컨텍스트상의 유사도 측정에서 링크를 포함한 문서의 개수가 충분하기 않을 때 떨어지는 정확도를 보완한다. 방식은 Jaccard 스코어 계산방식을 사용하며 이때 정의되는 $\Gamma(x)$는 x의 이웃노드를 의미한다. 이웃노드는 각 노드에서 연결되어있는 인-링크(incoming-link) 아웃-링크(outgoing-link)를 모두 포함한다. 즉, 두 개의 서로 다른 노드가 하나의 문서에서 발견되지 않더라도, 동일한 링크 구조를 포함하고 있다면 관련도 점수를 받게 되며 수식은 다음과 같다.

$$jacc(x,y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}$$

- **통합 유사도 계산 방식:** 본 논문에서는 위에 설명된 두 개의 방식을 통합하여 계산하는 다음과 같은 방식을 사용하였으며 전체 값은 0부터 1까지의 분포를 갖도록 조정되었다.

$$combi(x,y) = lfidf(x,y) \times (1 + jacc(x,y))$$

3.2.3 웹그래프를 이용한 트리플 확장

본 절에서는 앞 절에서 구축된 웹그래프 상의 두 노드 간의 유사도를 바탕으로 인포박스에 추가될 트리플 데이터를 찾는 방식을 설명한다. 웹그래프를 이용한 트리플 확장을 하는 목적은, 번역을 통한 정보 전달은 타겟 언어의 관점에서 누락된 정보를 반영하는 목적에서이다. 실제 웹 문서상의 하이퍼링크는 사용자가 수작업으로 현재의 문서와 관련된 문서를 연결하여 표시하는 것이므로 사용자의 의도가 들어가는 것으로 간주된다. 본 논문에서는 이와 같은 관점에서 사용자의 의도가 내포된 링크 정보를 이용하여 인포박스 트리플에 추가될 정보를 추출한다. 웹그래프를 이용한 트리플 확장은 그림 5의 예제와 같다. 번역된 DBpedia 트리플의 특정 property에 해당하는 object를 추가적으로 선정하는 작업이다. 3.2.2에서 구축된 웹그래프상에서 <애플컴퓨터, iMac> 페어가 특정 임계값 이상이며, iMac의 property가 주어진 property와 동일한 경우, 그림 5의 우측 블락 처리된 부분과 같이 새로운 트리플 <애플컴퓨터, 제품, iMac>이 인포박스 데이터에 추가될 수 있다. 본 논문에서 동일한 property를 판별하기 위해서는 DBpedia ontology를 이용한다.

그림 5 트리플 확장 예제

4. 실험 및 평가

본 논문에서는 한국어 인포박스를 자동으로 생성하기 위하여 소스 자원으로 영어 인포박스 데이터를 추출하여 제공하는 DBpedia 영어 인포박스 데이터[3]를 사용하였다. 사용된 DBpedia 데이터는 총 10,648,277개의 트리플을 포함하고 있다. 실험은 주어진 영어 트리플 데이터를 입력으로 하여 한국어 인포박스에 추가할 트리플 데이터를 생산하는 것으로 수행하였다. 번역 단계를 통해 얻어진 한국어 트리플은 338,195이며 이는 소스로 주어진 트리플 데이터의 subject가 한국어 위키피디아에 등록되어있는 위키피디아 엔트리에 해당하는 경우만 해당하는 수치이며, 추가적으로 한국어 위키피디아에 등록되지 않은 subject의 경우 번역된 결과를 추후 새로운 한국어 위키피디아 엔트리를 만드는데 사용할 수 있을 것으로 보인다. 번역 단계 웹그래프를 이용해 확장된 트리플은 임의의 임계값 조건($\Theta \geq 0.8$)을 사용하였을 때, 401,311개의 트리플로 번역 단계 후 약 20%정도의 정보가 추가되었다. 임계값의 변화에 따라 추가되는 정보의 양이 결정되며, 임계값의 조건을 낮추어 더 많은 트리플이 인포박스에 추가되는 경우, 노이즈 발생이 될 수 있으므로 임계값을 설정하는 이슈가 존재한다.

본 논문의 기법을 평가하기위하여 기존 한국어에 실제 존재하는 인포박스 데이터$I(O)$와 비교 실험하는 방식을 사용하였다. $I(O)$는 오리지널 인포박스(Triples of Original Infobox)의 트리플을 나타내며 $I(A)$는 본 논문에서 제안하는 방식으로 생성된 인포박스 트리플(Triples of Artificial Infobox)을 나타낸다. 인포박스 생성 실험을 위하여 다음과 같은 정확률, 재현율, F-값을 사용한다. 정확률은 자동으로 생성된 인포박스 트리플 중 옳은 비율을 의미하고, 재현율은 성공적으로 생성된 인포박스 트리플 비율을 나타낸다. F-값은 정확률과 재현율을 통합적으로 나타내는 평가 기준이다.

$$\text{정확률} = \frac{I(O) \cap I(A)}{I(A)} \quad \text{재현율} = \frac{I(O) \cap I(A)}{I(O)}$$

$$F\text{-값} = \frac{2 \times \text{정확률} \times \text{재현율}}{\text{정확률} + \text{재현율}}$$

평가는 생성된 인포박스 트리플이 포함하고 있는 약 840개의 property 중 특정 property를 임의로 선정하여 각 property별로 인포박스 생성 측면에서, 영어 인포박스 데이터를 입력으로 하여 한국어 인포박스 데이터를 생성하는 방식으로 진행하였다. 사용된 property는 총 P_1, P_2, P_3 이며 각각 전체 트리플 데이터중 5% 3% 1%를 차지하

3)http://downloads.dbpedia.org/3.7/en/mappingbased_properties_en.nt.bz2

는 비율의 property이다. 그림 6은 번역 작업만 진행한결과에 대한 3개의 property에 대한 결과이며 그림 7은 웹그래프를 통해 추가적인 정보를 인포박스 트리플로 간주하여 실험한 결과이다. 실험은 평균적으로 40%의 정확률과 83%의 재현율을 나타내었다. 번역 단계만 사용한 결과보다 웹그래프를 사용하여 트리플 확장 단계를 거친 경우, 재현율은 상당히 올라가지만 정확률이 상대적으로 낮아졌다. 이는 추가적인 정보를 찾는 본 논문의 목적에는 부합했으나, 기존 한국어에 존재하는 인포박스 데이터의 내용이 인포박스에 포함될 완전한 데이터를 모두 포함했다고 볼 수 없으므로 수행한 실험의 정확률이 상대적으로 낮게 나온 것으로 분석되었다. 실제 사람이 수작업으로 새롭게 생성된 인포박스 데이터의 적합성을 판별한 경우 평균 76%의 정확률과 91%의 재현율을 나타내었다.

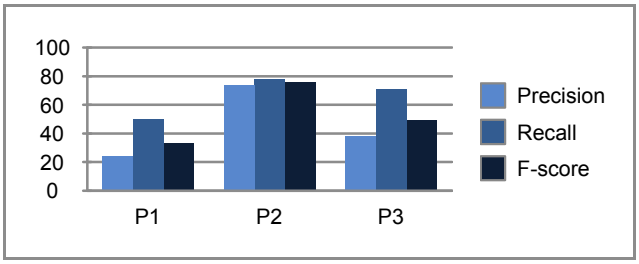

그림 6 실험 결과 (번역)

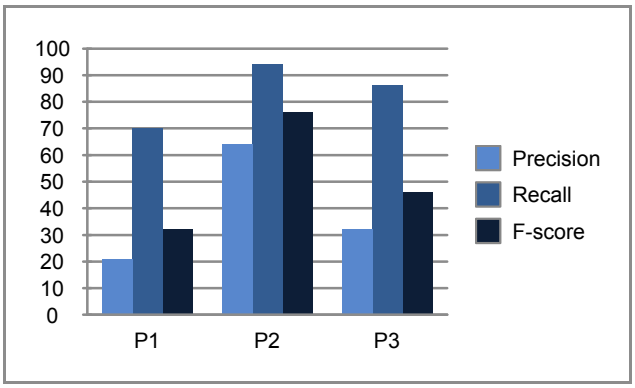

그림 7 실험 결과 (번역 + 확장)

실험을 통하여 웹그래프를 이용한 확장 방식에 의해 추가된 데이터가 실험의 재현율을 높이는 것으로 보아 본 논문의 기법이 인포박스를 자동으로 생성하는 트리플을 찾는 방식에 효과적임을 알 수 있었다. 그러나 특정 property의 경우 (예:birthDate) 해당 subject에 해당하는 object는 유일한 것으로 확장단계를 통해서도 효과를 얻을 수 없었다.

5. 결론

본 논문에서는 두개의 서로 다른 언어 간 정보의 차이를 이용하여 위키피디아 인포박스를 자동으로 생성하는 기법을 제안하였다. 본 논문의 기법은 소스 언어에 존재하는 데이터를 타겟 언어로 번역하고 번역된 데이터에 웹그래프를 이용하여 인포박스에 들어갈 수 있는 데이터를 추가적으로 발견하여 통합하는 방식을 제안하였다. 본 논문의 기법은 인포박스를 자동 생성하는 면에서 기존에 존재하는 인포박스와 비교하는 실험을 통해 상대적으로 높은 재현율을 나타내었다. 하지만, 기존 한국어에 존재하는 인포박스 데이터의 내용이 인포박스에 포함될 완전한 데이터를 모두 포함했다고 볼 수 없으므로 본 논문에서 제안하는 실험의 정확률은 비교적 낮게 나온 것으로 분석되었다.

향후 기존 소스 언어의 인포박스 데이터를 보완할 수 있는 방법이 추가적으로 수행되어 두 언어 간 상호 보완적인 방식으로 진행할 것이며 세 개 이상의 언어 간 정보의 보완을 통해 언어 간 편향된 정보의 문제점을 해소하는 방향으로 발전할 예정이다. 또한 웹그래프를 이용한 방법 이외에 인포박스를 생성하는 연구도 병행될 예정이다.

감사의 글

이 논문은 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2011-0018264)

참고문헌

[1] http://www.wikipedia.org
[2] http://ko.wikipedia.org/wiki/틀:철학자_정보
[3] http://dumps.wikimedia.org/
[4] Fei Wu, Daniel S. Weld, Automatically Refining the Wikipedia Infobox Ontology, WWW '08 Proceeding of the 17th international conference on World Wide Web, 2008
[5] Simone Paolo Ponzetto, Roberto Navigli, Large-scale taxonomy mapping for restructuring and integrating wikipedia, IJCAI'09 Proceedings of the 21st international joint conference on Artificial intelligence, 2009
[6] Eytan Adar, Michael Skinner, Daniel S. Weld, Information arbitrage across multi-lingual Wikipedia, WSDM '09 Proceedings of the Second ACM International Conference on Web Search and Data Mining, 2009
[7] Vivi Nastase, Michael Strube, Benjamin Börschinger, Cäcilia Zirn, and Anas Elghafari., WikiNet: A very large scale multi-lingual concept network, In Proceedings of the 7th International Conference on Language Resources and Evaluation, La Valetta, Malta, 17-23 May 2010
[8] http://en.wikipedia.org/wiki/Wikipedia:Multilingual_ranking_December_2009
[9] http://dbpedia.org/
[10] http://dbpedia.org/Internationalization
[11] Eun-kyung Kim, Matthias Weidl, Key-Sun Choi, Soren Auer, Towards a Korean DBpedia and an Approach for Complementing the Korean Wikipedia based on DBpedia, Proceedings of the 5th Open Knowledge Conference, 2010.
[12] http://www.wikibhasha.org/
[13] Maike Erdmann, Kotaro Nakayama, Takahiro Hara, Shojiro Nishio, An approach for extracting bilingual terminology from Wikipedia, Proceedings of the 13th international conference on Database systems for advanced applications (2008), pp. 380-392.
[14] http://translate.google.com/
[15] http://alldic.daum.net/dic/view_top.do
[16] http://dic.naver.com/
[17] http://en.wikipedia.org/wiki/Webgraph
[18] MCGILL, M. AND SALTON, G. Introduction to Modern

Information Retrieval. McGraw-Hill. 198