위키피디아로부터의 자동 병렬 문장 추출 기법을 이용한 영어-한국어 교차언어 정보검색의 번역 성능 개선

천주룡○, 고영중
동아대학교
balendia@gmail.com, youngjoong.ko@gmail.com

Improving Query Translation by Extracting Parallel Sentences from Wikipedia for Cross-Language Information Retrieval

Juryong Cheon○, Youngjoong Ko
DongA University, Computer Engineering

요 약

본 논문은 영어-한국어 교차언어 정보검색의 질의어 번역에 대한 중요한 자원으로 활용되는 병렬 말뭉치의 품질 향상을 위해서, 위키피디아의 비교 말뭉치로부터 자동으로 병렬 문장을 추출하여 활용하는 기법을 제안한다. 기존 연구에서 질의어 번역을 위해 위키피디아의 이중 어휘 사전 및 동의어, 다의어 정보를 구축하고, 기 기축된 병렬 말뭉치와 함께 활용하여 여러 의미를 가진 번역 후보 단어들 중, 최적의 단어를 선택하는 방법을 이용하고 있다. 여기서 활용되는 병렬 말뭉치는 질의어 번역에서 가장 중요한 자원이다. 하지만, 기 구축된 병렬 말뭉치는 양이 적거나, 특정 영역을 중심으로 구성되어 있는 문제가 있다. 이러한 문제를 해결하기 위해, 본 논문은 위키피디아로부터 자동 병렬 문장 추출 기법을 이용, 대량의 영어-한국어 간 병렬 말뭉치를 구축하고, 이를 교차언어 정보검색을 위한 질의어 번역에 적용하여 개선을 보인다. 실험의 성능 비교를 위해서 NTCIR-5 데이터를 이용하였으며 기 구축된 세종 병렬 말뭉치를 활용한 질의어 번역의 성능이 MAP 31.5%, R-P 33.0%에서, 새롭게 구축한 위키피디아 병렬 말뭉치를 활용한 질의어 번역의 성능이 MAP 34.6%, R-P 34.6%로, 각각 MAP 3.1%와 R-P 1.6%의 성능 향상을 보였다.

주제어: 병렬 문장, 정보검색, 질의어 번역, 위키피디아

1. 서론

정보검색 시스템의 궁극적인 목적 중 하나는 사용자의 의도를 만족시키는 문서들을 검색하는 것이다. 하지만, 단일 언어로 쓰인 문서를 검색하는 전통적인 정보검색 시스템은 현재 인터넷의 특성에 부합하지 않는 측면이 있다. 최근 웹 환경에서 정보의 양은 끊임없이 늘어나고 있고, 국가 간 경계가 허물어짐에 따라 교차언어 정보검색의 연구가 활발하게 진행되고 있다.

교차언어 정보검색은 원본 언어로 표현된 질의어를 기반으로 목적 언어로 쓰인 문서들을 검색하는 시스템을 말한다. 이와 같이 원본 언어로 표현된 질의어를 목적 언어로 번역하기 위해서는 번역을 위한 사전이나 혹은 병렬 말뭉치(parallel corpus)와 같은 자원이 필수적이다. 그러나, 세종 병렬 말뭉치와 같은 한국어가 포함된 병렬 말뭉치는 양이 적고, 특정 영역을 중심으로 구성되었거나, 저작권 및 지적 소유권 등의 문제가 존재한다.

본 논문은 영어-한국어 교차언어 정보검색 시스템을 구현하기 위해 비교적 널리 연구되어지는 질의어 번역에서, 가장 중요한 자원으로 활용되는 병렬 말뭉치의 품질 향상을 통한 질의어 번역 성능 개선에 초점을 맞춘다. Sungho Kim[1]은 질의어 번역을 위해, 위키피디아로부터 이중 어휘 사전 및 동의어, 다의어 정보와 같은 언어 자원을 구축하여 활용하였다. 그리고 질의 번역 과정에서 발생하는 번역의 모호성을 해결하기 위해서 여러 후보 단어들 중, 최적의 단어를 선정, 최종적으로 번역된 질의어를 선택하였다. 여기서, 질의어를 번역 및 선택하는 과정에서 가장 큰 영향을 미치는 자원은 병렬 말뭉치이다. 따라서 질이 좋고 양이 풍부한 병렬 말뭉치를 구축할 수 있다면 질의어 번역에 대한 더 긍정적인 성능 개선을 기대할 수 있다.

하지만, 병렬 문장으로 구성된 병렬 말뭉치를 구축하는 작업은 쉽지 않으며 시간과 비용이 많이 소요되는 작업이다. 효과적으로 병렬 말뭉치를 구축하기 위해서 위키피디아와 같은 언어 자원의 비교 말뭉치(comparable corpus)로부터 병렬 문장만을 자동으로 식별하고 추출하기 위한 연구가 많이 이루어지고 있다. 천주룡[2]은 위키피디아로부터 영어-한국어 간 양질의 병렬 말뭉치를 추출하는 유사 문장 계산 방법을 제안했다. 이를 위해, 영어-한국어 간 위키피디아 문서들 중 인터링크(Interlink)로 이루어진 문서 쌍에 있는 문장들을 비교 말뭉치로 이용하였으며, 언어 자원들의 순차적인 매칭을 통해 문장 간 유사도를 계산하였다. 이는 병렬 말뭉치 추출에 대해, 누구에게나 공개된 자원이며 시간이 흘러감에 따라 정보를 가진 문서의 수가 기하급수적으로 늘어나는 위키피디아의 특성을 활용한 장점을 가진다.

* 이 논문은 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. NRF-2013R1A1A2009937)

본 논문에서는 앞서 설명한 선행 연구를 기반으로 질의어 번역 시스템에 활용되는 병렬 말뭉치의 구축을 위해 위키피디아로부터의 자동 병렬 문장 추출 기법을 이용한다. 그리하여, 양질의 병렬 말뭉치를 구축하고 이를 활용하여 질의어 번역의 개선을 통한 교차언어 정보검색의 성능을 향상시킨다.

실험을 위해, 제안하는 시스템을 NTCIR-5 데이터를 이용하여 영어-한국어 간의 성능을 평가하였으며, 기 구축된 세종 병렬 말뭉치를 활용한 질의어 번역의 성능이 MAP 31.5%, R-P 33.0%에서, 새롭게 구축한 위키피디아 병렬 말뭉치를 활용한 질의어 번역의 성능이 MAP 34.6%, R-P 34.6%로, 각각 MAP 3.1%와 R-P 1.6%의 성능 향상을 보였다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구들을 소개하며, 3장에서는 논문에서 제안하는 방법에 대해 상세히 기술한다. 4장에서는 실험 결과를 분석, 비교하며, 5장에서 결론 및 향후 연구 계획을 기술한다.

2. 관련 연구

교차언어 정보검색 연구는 세계화에 따른 정보의 교환과 지식 공유의 측면에서 점차 중요해지고 있는 연구 분야이다. 지금까지의 교차언어 정보검색은 문서 번역과 질의어 번역으로 크게 두 가지 측면에서 연구되어져 왔다. 문서 번역의 경우에는 검색하고자하는 모든 문서를 목적 언어로 번역한 후에 검색을 수행하기 때문에 느리다는 단점을 가지고 있다. 반면에 질의어 번역의 경우 질의어 자체만을 원본 언어에서 목적 언어로 번역하기 때문에 상대적으로 간편하고 효율적이다. 두 방법 모두 번역 모호성을 해결하기 어렵다는 문제를 수반하지만, 질의어 번역에 경우 문서 번역보다 번역 모호성에 더 직접적인 관련을 가진다. 하지만, 질의어 번역이 높은 번역 정확성을 가진다면 매우 효율적인 성능을 보이며, 빠르다는 장점 때문에 현재의 교차언어 정보검색은 질의어 번역에 대한 연구가 가장 많이 이루어지고 있다.

번역은 크게 3가지 접근법을 통해 연구되어져 왔다. 첫 번째는 기계 번역을 사용하여 문서나 질의어를 번역하는 기법이다. Douglas[3]의 연구에서는 이러한 기계 번역 시스템을 이용하여 문서와 질의어를 각각 번역하여 성능을 비교하였다. 두 번째는 기계 판독 사전을 기반으로 질의어를 번역하는 방법이다. 이는 기계 판독사전을 이용하여 원본 언어로 표현된 질의어를 목적 언어로 단순히 번역하는 방법이다. Gina-Anne Levow[4]는 교차언어 정보검색을 위한 사전 기반의 다양한 기법에 대한 연구를 소개하였다.

마지막으로 세 번째는 말뭉치 기반 접근법으로 하나 또는 여러 언어에 대한 자연어 데이터를 활용, 서로 다른 언어 간의 똑같거나 비슷한 문장, 문단, 문서 등의 단위로 묶어 말뭉치를 구축하고, 이러한 데이터를 활용하여 기계 번역 기법에 적용해 질의어를 번역하는 방법이다. Jianqiang Wang[5]의 연구에서는 영어와 프랑스어, 영어와 중국어 간의 구축된 병렬 말뭉치를 바탕으로 통계적인 번역 모델인 GIZA++ toolkit[6]을 활용하여 단어 간의 번역 확률을 뽑고, 이를 이용하여 질의어를 번역하였다.

양질의 말뭉치를 구축하기 위해서는 각 언어 간의 많은 데이터가 필요할 뿐만 아니라 병렬, 혹은 비슷한 문장, 문단, 문서 간의 데이터를 구축하기 위한 많은 노력이 필요하다. 유사한 문장을 추출하는 연구로 Jessica Ramirez[7]는 원시 언어의 문서와 대상 언어의 문서를 각 언어에 맞는 형태소 분석을 한 뒤 비교 문장이 가지는 형태소 규칙을 세우고 후보 문장이 규칙에 매칭이 된다면 유사한 문장이라고 판단하였다. Masao Utiyama[8]는 정보 검색 방법으로 접근을 했으며, 원시 언어 문서와 대상 언어 문서의 한 문장 당 단어의 빈도수와 전체 문서에서 단어가 출현한 문장의 빈도수를 구한 후, 원시 언어 문장과 대상 언어 문장이 유사할 경우 문장들이 가지는 단어 벡터들이 유사할 것으로 가정하였고, 이는 코사인 유사도와 같은 유사도 측정 방식으로 유사도를 계산할 수 있다. Sisay Fissaha Adafre[9]에서는 위키피디아 문서들이 가지는 링크 자질을 이용했다. 링크들이 인터-위키를 가지고 있고 원시 언어 문장 내에 링크와 대상 언어 링크들이 연결되어 있다면 두 문장은 같은 내용을 말하는 유사한 문장이라고 판단했다.

3. 제안 방법

이 장에서는 교차언어 정보검색의 질의어 번역[1]의 전체적인 과정을 설명한다. 그리고 질의어 번역의 개선을 위해, 가장 핵심적인 자원으로 활용되는 위키피디아로부터의 자동으로 병렬 문장을 추출하는 기법[2]의 과정을 설명한다.

3.1 질의어 번역 시스템

질의어 번역은 원본 언어로 표현된 질의어를 목적 언어로 표현하는 것이다. 질의어를 번역하기 위한 시스템의 전체적인 구성도는 그림 1과 같다.

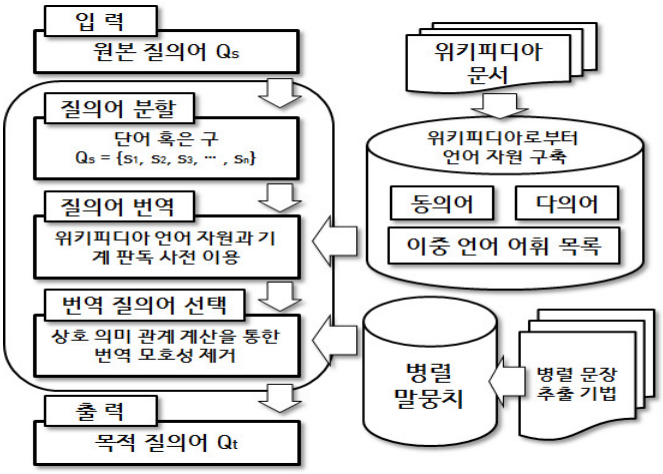

그림.1 질의어 번역 시스템의 구성도

먼저, 입력된 원본 언어 표현된 질의어가 들어오면 번역 가능한 단위로 질의를 분할하는 과정을 거친다. 질의어가 분할되면, 미리 구축되어있는 위키피디아를 포함한 언어 자원들을 이용하여 번역 후보군들을 추출한다. 그리하여 만들어진 번역 후보군들에 대해 병렬 말뭉치와 목적 문서들을 통한 상호 의미 관계를 계산하고, 최적의 번역을 선택하는 작업을 거쳐 번역 모호성을 제거한다. 최종적으로 목적 언어로 번역된 질의어를 가지고, 검색을 수행한다.

3.1.1 언어 자원 구축

질의어 번역을 위한 언어 자원으로는 먼저, 위키피디아에 포함된 정보를 바탕으로 이중 언어 어휘 목록, 동의어 집합, 다의어 집합을 생성하고 이를 결합한다. 위키피디아는 인명, 지역명, 영화 제목 등 많은 개체명과 같은 용어들의 정보를 포함하며 단어 혹은 구의 형태를 가지고 있다. 이중 언어 어휘 목록은 위키피디아의 인터링크 정보를 통해 추출한다. 인터링크 정보는 한국어 위키피디아에서 하나의 개체를 설명하고 있는 일반 문서가 같은 주제에 대해 한국어 이외의 다른 언어로 표현된 문서 주제를 제공하는 것이다. 이밖에, 동의어 집합과 다의어 집합은 각각 위키피디아의 넘겨주기 문서 정보와 동음이의어 문서 정보를 이용하여 추출한다. 이와 같이 구축한 이중 언어 어휘 목록에 대해 다의어 집합으로 확장한 사전을 위키 사전(Wikipedia based Lexicon)이라 부른다. 동의어 집합은 검색기의 동의어 정보를 입력으로 적용하여 사용한다. 구축한 위키 사전의 예를 표 1에 보인다.

표.1 위키 사전의 예

영어	한국어
Andre Agassi	안드레 애거시
Apache Software Foundation	아파치 소프트웨어 재단
President of South Korea	대한민국의 대통령
United States Secretary of Defense	미국의 국방부 장관

또한, 일반적인 용어들을 다양하게 다루기 위한 어휘 집합으로는 방대한 양의 온라인 사전을 이용하며, 이를 기계 판독 사전(machine readable dictionary)라 부른다. 이 두 가지 언어 자원은 자동 병렬 문장 추출 기법에도 동일하게 구축되어 사용한다.

3.1.2 번역 후보군 추출 및 최적 번역 선택

먼저, 입력된 원본 질의어의 불용어(stopword)를 제거하고, 번역 가능한 단어나 구로 분할하는 작업을 거친다. 이는 질의어가 단어뿐만 아니라 구나 개체명 단위로 번역될 수 있기 때문이다. 미리 구축된 위키 사전 및 기계 판독 사전을 포함한 언어 자원을 이용하여, 분할된 질의어의 형태에 따라 번역 후보군을 생성하게 된다. 즉, 구나 개체명의 단위인 경우에는 위키 사전에 의해 번역 후보가 생성되며, 일반 단어인 경우에는 위키 사전과 기계 판독 사전을 모두 이용하여 번역 후보가 생성된다.

번역이 확정된 단어와 번역 후보군을 가지고 있는 단어들과의 상호 의미 관계 계산을 통해서 최종적으로 하나의 번역 단어를 선택하기 위한 방법은 [1]의 WTDM(Weighted Total divergence to the mean)을 따르게 된다. 이 방법은 단어 간의 두 가지 확률을 먼저 추출하여야 한다. 먼저, 이웃한 모든 두 번역 후보 단어 간 목적 문서 내의 단어 확률 분포를 이용하여 통계적인 유사도를 계산한다. 이를 전이확률(transition probability)라 부른다. 또한 기 구축된 병렬 말뭉치로부터 GIZA++ toolkit을 이용하여 번역확률을 추출한다. 추출한 두 확률을 이용하여, 다음과 같은 수식을 통해 적절한 번역 단어들을 선택하기 위한 스코어를 계산한다.

$$Q_s = \{s_1, s_2, s_3, \ldots, s_n\} \tag{1}$$

$$\varnothing(t_i) = P(t_1|s_1)\prod_{j=1}^{n-1} P(t_{ij+1}|t_{ij})P(t_{ij+1}|s_{j+1}) \tag{2}$$

$$Q_t = \arg max_{ci}\varnothing(c_i) \tag{3}$$

Q_s는 입력되는 원본 질의어이며, Q_t는 질의어가 번역된 목적 질의어이다. $P(t_{ij+1}|t_{ij})$는 번역된 단어 간의 전이 확률에 해당하며, $P(t_{ij+1}|s_{j+1})$는 원본 질의 단어와 번역된 단어 간의 번역 확률에 해당한다. 단어들에 대한 전이 확률과 번역 확률은 각 위치에 대해 모두 더해 1이 될 수 있도록 정규화 작업을 거치며, 번역 확률은 너무 낮은 확률과 높은 확률에 대한 영향을 줄이기 위한 스무딩(smoothing) 작업을 거친다.

그림 2는 적절한 번역 단어를 선택하기 위해, 원본 질의어로부터 생성된 번역 후보 단어들이 포함된 경로의 스코어들을 계산하는 예시이다.

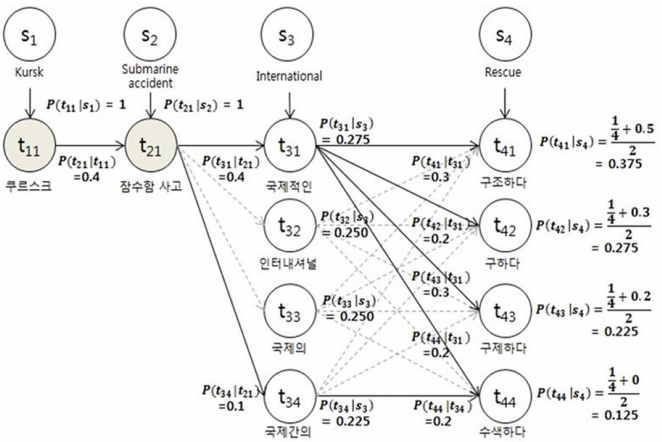

그림.2 번역 후보 단어들에 대한 스코어 계산

각 위치에 따른 번역 단어들이 포함된 경로들에 대한 스코어를 전이 확률과 번역 확률을 이용, 모두 계산하여 최종 스코어를 비교한다. 따라서 가장 높은 스코어를 가지는 경로에 포함된 단어들이 최종적인 목적 질의어로 선택되며, 검색기의 질의어로 입력된다. 여기에, 최종 번역 단어가 동의어 집합에 존재한다면 동의어들이 검색기의 입력에 추가된다.

가장 적절한 번역 단어들을 선택하기 위해서는 병렬 말뭉치로부터 추출된 정확한 번역확률이 필요하다. 따라서 본 논문은 자동으로 병렬 문장을 추출하는 기법을 통해 대량의 병렬 문장으로 구성된 병렬 말뭉치를 구축하고, 이를 질의어 번역 시스템이 적용하여 성능의 개선을 보인다.

3.2 자동 병렬 문장 추출 기법

그림 3은 위키피디아로부터 자동 병렬 문장 추출을 위한 전체적인 과정이다.

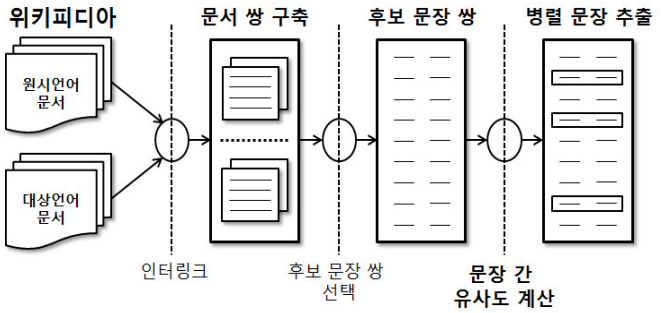

그림.3 병렬 문장 추출을 위한 전체적인 과정

병렬 문장 추출을 위해서는 먼저 영어-한국어 간 위키피디아 문서들 중 인터링크로 이루어진 문서 쌍에 존재하는 문장들을 비교 말뭉치로 활용한다. 비교하려는 문장들 간의 유사도를 계산하기 위해, 사전과 같은 언어 자원과 토픽 모델(topic model)을 이용한다. 언어 자원은 앞서 설명한 질의어 번역 시스템을 위해 구축한 위키 사전, 기계 판독 사전과 동일하다.

문장 유사도 계산은 영어와 한국어 문장에 존재하는 단어들 간의 매칭을 기본적으로 이용한다. 단어 매칭에는 다양한 언어 자원을 활용하며, 이는 우선순위를 고려한 순차적인 단어 매칭에 이용된다. 이용하는 언어 자원의 우선순위는 위키 사전, 숫자, 기계 판독 사전, 위키 사전을 통해 추출한 번역 확률, 그리고 토픽 모델을 통한 단어의 토픽 분포의 순서이다.

영어 및 한국어 문장의 유사 문장 계산은 자카르트(Jaccard) 유사도를 변형한 수식 2와 같으며, 총 5단계의 순서로 진행된다.

$$J_{DT}(A,B) = \frac{M_1 + M_2 + M_5}{M_1 + M_3 + M_4} \qquad (4)$$

단계1 : 영어와 한국어 문장 안의 각 단어에 대해, 위키 사전을 통한 단어 매칭을 수행한다. 영어 단어 기준으로 위키 사전을 통해 매칭되는 단어를 한국어 단어로 번역한다. 이러한 방식으로 매칭된 단어 수를 식 (4)의 M_1에 추가, 매칭된 단어들을 두 문장에서 제거한다.

단계2 : 한국어 문장에서 숫자를 추출하고, 영어 문장에서는 서수나 날짜를 숫자 매칭을 이용하여 숫자로 변환, 추출한다. 위와 마찬가지로 추출된 두 문장의 숫자 단어 간의 매칭 수를 식 (4)의 M_1에 추가하며, 매칭된 숫자 단어를 제거한다.

단계3 : 남아 있는 두 문장의 단어에 대해, 기계 판독 사전을 통해 단어 매칭을 수행한다. 마찬가지로 영어 단어를 한국어 단어로 번역하고, 두 문장에서 매칭하는 단어 수를 식 (4)의 M_1에 추가한다. 매칭된 단어는 제거하며, 매칭되지 않고 남아있는 영어와 한국어 단어의 수를 각각 식 (4)의 M_3과 M_4에 추가한다.

단계4 : 남아 있는 두 문장의 단어 중, 위키 사전으로부터 단어 단위로 추출한 번역 확률에 매칭되는 단어들이 있다면, 영어 단어에서 한국어 단어로 번역될 확률 값을 매칭한다. 단, 매칭되는 단어의 후보가 2개 이상 존재한다면, 가장 높은 번역 확률로 매칭하고, 매칭된 번역 확률 값의 합을 식 (4)의 M_2에 추가한다.

단계5 : 마지막으로 남아 있는 두 문장의 단어에 대해, 토픽 모델을 통한 단어 토픽 확률 분포를 이용하여 단어 간 코사인 유사도를 계산한다. 하나의 영어 단어 기준으로 한국어 단어에 대해 모든 코사인 값을 구한 뒤, 가장 높은 코사인 값을 식 (4)의 M_5에 추가한다. 적용된 값에 해당하는 영어 단어와 한국어 단어를 제거하며, 이 단계는 단어가 모두 제거될 때까지 반복한다.

4. 실험

이 장에서는 질의어 번역 시스템에 대한 실험 환경 및 데이터와 위키피디아로부터의 자동으로 병렬 문장을 추출하는 기법을 적용한 실험 데이터에 대해 설명한다. 본 논문에서는 질의어 번역 시스템에 대해 인드리(Indri)[10] 검색기를 이용하여 문서들을 색인하고, 번역된 질의어를 인드리를 이용하여 문서를 검색한다.

4.1 실험 데이터

질의어 번역에 대한 성능을 평가하기 위해서는 교차언어 정보검색을 위한 실험 데이터인 NTCIR-5 교차언어 정보검색 테스트 데이터를 사용한다. NTCIR 테스트 데이터는 중국어, 일본어, 한국어, 영어 간의 교차언어 정보검색을 위한 데이터이다. 본 논문의 질의어 번역 시스템은 영어 질의어를 입력하였을 때, 한국어 문서를 얼마나 잘 찾아줄 것인가가 평가의 주요 목적이 된다. 테스트 데이터로부터 색인된 한국어 문서의 수는 총 220,374개이며, 영어 질의의 수는 50개이다.

병렬 문장 추출 기법을 적용할 실험 데이터는 비교 말뭉치로 사용한 영어-한국어 위키피디아 문서 쌍이 총 106,582개이다. 각 문서는 서론에 해당하는 문장들만 사용하며, 길이가 충분히 길고 양쪽 문서의 길이가 차이가 많이 나지 않는 문서만을 필터해서 사용한다. 병렬 문장 추출 기법의 유용성을 평가하기 위해, 임의의 100개의 문서 쌍을 선택, 5명의 어노테이터(annotator)가 총 3,100개의 정답 병렬 문장 쌍을 수작업으로 태깅하였다.

문장 간 유사도 계산에 사용된 토픽 모델 학습은 mallet toolkit[11]에서 제공한 오픈 소스를 사용하여 학습하였고, 학습에 사용된 한국어-영어 위키피디아 문서 쌍의 수는 총 7400여개이다. 이는 문서 간 길이가 거의 비슷한 문서 쌍만을 추출한 수이다. 토픽 수는 가장 좋은 성능을 보인 1000개를 사용, 파라미터 α와 β는 각각 0.01과 50/토픽 수로 지정해서 사용하였다.

4.2 실험 결과

질의어 번역 시스템에 대한 실험 평가에 앞서, 질의어 번역에서 가장 중요한 자원으로 활용되는 병렬 말뭉치를 위한 위키피디아로부터의 자동으로 병렬 문장을 추출하는 기법의 실험 결과를 먼저 보인다. 병렬 문장 추출 기법의 평가를 위해 사용된 성능 평가 도구로는 정확률(precision), 재현율(recall), F1-score이다. 표 2는 병렬 문장 추출 기법의 실험 결과이다.

표.2 병렬 문장 추출 기법의 실험 결과

정확률	재현율	F1-score
54.9%	64.1%	59.1%

표 3과 표 4는 기존에 구축되어 있는 병렬 말뭉치인 세종 병렬 말뭉치와 위키피디아로부터의 자동 병렬 문장 추출 기법을 통해 추출한 병렬 말뭉치의 문장 수와 단어 수의 비교를 보여준다.

표.3 세종 병렬 말뭉치의 문장 및 단어 수

개수	영어	한국어
총 문장 수	33,764	33,764
총 단어 수	181,690	247,810
문장 당 평균 단어 수	5.38	7.34

표.4 자동 병렬 문장 추출 기법을 통해 추출한 병렬 말뭉치의 문장 및 단어 수

개수	영어	한국어
총 문장 수	54,381	54,381
총 단어 수	619,889	628,513
문장 당 평균 단어 수	11.40	11.56

세종 병렬 말뭉치에 비해, 병렬 문장 추출 기법을 통해 추출한 병렬 말뭉치의 총 문장 쌍의 수가 약 2만 개가 증가하였으며, 총 단어의 수는 영어가 약 44만 개, 한국어가 약 38만 개 증가하였다. 문장 쌍은 비교적 크게 증가하지 않았으나, 단어의 수가 매우 크게 증가한 점으로 볼 때, 풍부한 병렬 말뭉치를 구축하였다고 생각된다.

질의어 번역 시스템의 성능 평가는 정보검색 분야에서 가장 보편적으로 쓰이는 평가 방법인 MAP(mean average precision)과 R-P(mean R-precision)을 이용하여 자동 병렬 문장 추출 기법을 이용한 개선된 질의어 번역 시스템을 평가한다.

표 5는 질의어 번역 시스템에 대한 실험 결과이다.

표.5 질의어 번역 시스템의 실험 결과

방법	MAP	R-P
mono	35.7%	35.1%
base(% mono)	31.5%(88%)	33.0%(94%)
proposed(% mono)	34.6%(97%)	34.6%(99%)

먼저 mono는 단일 언어 실험 결과로써 질의어 번역 작업을 거치지 않고 사용자가 직접 손으로 질의어를 번역한 실험 결과이다. 이는 질의어 번역 시스템의 목표 성능이 된다. base는 기존에 구축되어 있는 세종 병렬 말뭉치를 이용하여 번역 확률을 추출하고, 이를 질의어 번역 시스템에 적용한 결과이다. 반면, proposed는 위키피디아로부터 자동 병렬 문장 추출 기법을 통해 대용량의 병렬 말뭉치를 구축한 뒤, 이를 질의어 번역 시스템에 적용한 결과이다. 또한, 세종 병렬 말뭉치와 위키피디아로부터 추출한 병렬 말뭉치를 결합하여 적용한 실험도 수행하였으나, 성능의 향상을 얻지는 못하였다.

base와 proposed, 둘 다 성능이 mono를 넘어서지는 못하지만, proposed의 경우 MAP이 97%, R-P가 99% 근접한 성능 결과를 보이고 있다. 또한, base의 비해서 proposed의 성능이 MAP 31.5%에서 34.6%로 3.1%의 개선을 보이고, R-P 33.0%에서 34.6%로 1.6%의 개선을 보이고 있다. 이는 자동 병렬 문장 추출 기법을 통해 새로이 구축한 병렬 말뭉치는 59.1%의 성능으로 완벽한 병렬 문장 쌍을 가진 병렬 말뭉치는 아니지만, 내용이 비슷한 문장 쌍으로 이루어져 있으며 기존에 구축되어 있는 세종 병렬 말뭉치보다 훨씬 풍부한 양의 문장으로 구성되어 있어 질의어 번역에 긍정적인 영향을 미치고 있다고 해석할 수 있다.

5. 결론

본 논문에서는 교차언어 정보검색에 대한 질의어 번역을 위해, 위키피디아로부터 자동 병렬 문장 추출 기법을 이용한 대용량의 병렬 말뭉치를 구축하고, 이를 질의어 번역에 적용함으로써 성능의 개선을 보였다. 계속 성장하는 원천 자원인 위키피디아를 이용하여 추출한 병렬 말뭉치는 시간이 흘러감에 따라 더욱 개선될 수 있는 여지가 남아있다.

향후 과제로는 교차언어 정보검색의 영어-한국어 간

질의어 번역의 개선뿐만 아니라, 일본어-한국어, 중국어-한국어 등 다양한 다른 언어들에 대해서도 연구를 지속할 예정이다. 또한, 교차언어 정보검색의 또 다른 핵심인 질의어 확장에 대한 연구를 질의어 번역 시스템과 결합하여, 더욱 좋은 성능을 얻을 수 있도록 연구를 지속할 생각이다.

참고문헌

[1] Sungho Kim, Youngjoong Ko and Douglas W. Oard, "Combining Lexical and Statistical Translation Evidence for Cross-Language Information Retrieval." Journal of the American Society for Information Science and Technology, Wiley-Blackwell, Vol.66, No.1, pp.23-39, 2015.

[2] 천주룡, 고영중, "언어 자원과 토픽 모델의 순차 매칭을 이용한 유사 문장 계산 기반의 위키피디아 한국어-영어 병렬 말뭉치 구축." 정보과학회논문지:(KIISE):소프트웨어 및 응용, 제 42권, 제7호, pp.901-909, 2015년.

[3] Douglas W. Oard, "A comparative study of query and document translation for cross-language information retrieval." In Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas, pp.472-483, 1998.

[4] GA Levow, DW Oard and P Resnik, "Dictionary-based techniques for cross-language information retrieval." Information Processing and Management: Special Issue on Cross-language Information Retrieval, Vol.41, No.3, pp.523-547, 2005.

[5] J Wang, DW Oard, "Combining bidirectional translation and synonymy for cross-language information retrieval." In Proceedings of the ACM SIGIR 2006, pp. 202-209, 2006.

[6] GIZA++ statistical translation models toolkit, http://code.google.com/p/inc-giza-pp/

[7] Ramirez Jessica C and Yuji Matsumoto, "A Rule-Based Approach For Aligning Japanese-Spanish Sentences From A Comparable Corpora." International Journal on Natural Language Computing, Vol.1, No.3, 2012.

[8] Utiyama Masao and Hitoshi Isahara, "Reliable measures for aligning Japanese-English news articles and sentences." In proceedings of ACL '03, Vol.1, pp.72-79, 2003.

[9] Adafre Sisay Fissaha and Maarten De Rijke, "Finding similar sentences across multiple languages in wikipedia," In Proceedings of ACL '06, p.62-69, 2006.

[10] Indri search engine, http://sourceforge.net/projects/lemur/

[11] Mallet toolkit, http://mallet.cs.umass.edu/download.php