□ 기술해설 □

다국어 정보검색

한국전자통신연구원 장명길·김영길·박영찬

1. 서 론

최근 웹과 정보기술의 급속한 발전에 따라 자국 언어 외의 다른 언어로 만들어진 많은 온라인 문서들이 급격히 증가하고 있다. 또한 이들 다른 언어들의 문서 정보들을 검색하는 시스템(Multilingually Searchable System)에 대한 필요성 또한 더욱 증가되고 있는 것이 현실이다. 이러한 요구에 발맞춰 세계 각국에서는 다국어 정보검색(Multilingual Information Retrieval)에 관한 연구가 활발히 진행되고 있다.

다국어 정보검색이란 질의어의 언어와 상관없이 여러 언어로 만들어진 정보를 검색하는 것이다. 실제 다국어 정보검색에서 다루는 정보는 텍스트로 된 문서 정보뿐만 아니라 이미지, 음성 등도 검색 대상으로 다루고 있다. 본 논문에서는 다국어로 된 문서의 검색에 한정하여 논의한다. 다국어 문서검색은 검색 대상 문서의 언어가 질의어의 언어와 동일한 경우에 그 문서를 검색하는 단일언어 문서검색(Monolingual Text Retrieval)과 질의어의 언어와 다른 언어로 쓰여진 문서를 검색하는 교차언어 문서검색(Cross-Language Text Retrieval)이 함께 적용되어야 한다. 본 논문에서는 현재 많은 연구가 되어 온 단일언어 문서검색보다는 교차언어 문서검색을 중심으로 설명하고자 한다. 다국어 정보검색 관련 논문들에서는 다국어의 용어를 'multilingual', 'translingual', 'cross-lingual', 'cross-language' 등 약간 다른 의미를 가지는 여러 종류의 용어를 함께 사용해 왔다. 최근 'Cross-Linguistic Information Retrieval'이라는 SIGIR-96 워크샵에서 참석자들간에 'multilingual'라는 넓은 의미보다는 보다 명확한 'cross-language'라는 용어로 통일하자는 의견을 모은 바 있다[17].

다음 2장에서는 현재 연구되고 있는 다국어 정보검색의 방법들에 대하여 설명하고, 3장에서는 현재 시범적으로 혹은 상용으로 서비스되고 있는 다국어 정보검색 시스템에 대하여 살펴본다. 4장에서는 일반적인 다국어 정보검색의 실험 방법과 여러 방법들의 성능 평가에 대하여 설명하고 5장에서의 다국어 정보검색의 향후 연구 방향과 전망으로 결론을 맺는다.

2. 다국어 정보검색 방법들

교차언어 문서검색은 질의어의 언어와는 다른 언어로 쓰여진 문서를 검색하기 위하여 질의어 언어와 검색 대상 문서 언어사이의 언어 장벽을 해소하는 여러 가지 방법을 사용한다. 그림 1은 교차언어 문서검색의 여러 방법들의 체계적인 분류를 나타내는 그림이다[18]. 그림

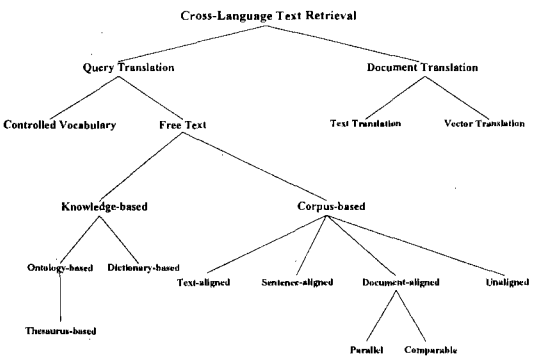


그림 1 교차언어 문서검색 방법의 분류

1에서와 같이 교차언어 문서검색은 언어 변환의 방식에 따라 두 가지로 크게 나누어 세부 방법들이 분류될 수 있다. 하나는 검색 대상 문서를 질의어의 언어로 변환하는 문서 번역(document translation) 방식이고 다른 하나는 질의어를 검색 대상 문서의 언어로 변환하는 질의어 번역(query translation) 방식이다.

2.1 문서 번역 방식

문서 번역 방식은 검색 대상 문서를 질의어의 언어로 변환하기 위하여 이에 해당하는 기계번역 시스템을 이용한다. 문서 번역 방식에서의 문서 번역은 짧은 길이의 질의어 번역보다 많은 언어적 정보를 이용할 수 있기 때문에 번역의 정확성과 검색의 정확도가 질의어 번역 방식보다 높다. 하지만 검색 대상이 웹인 경우 매우 방대한 양의 문서를 수집하고 번역하는 일은 현실적으로 매우 어려운 작업이다. 또한 현재의 기계번역 기술이 제한된 영역과 유사한 언어간의 번역은 만족할 만한 수준의 번역 결과를 내고 있으나 한국어와 영어간의 경우 일반 영역에서의 번역은 번역의 품질이 매우 낮아 기계번역 기술을 이용한 교차언어 문서검색을 채택하기에는 아직 시기 상조로 판단된다.

최근 TREC-6의 CLIR(Cross-Language Information Retrieval)에서는 문서 번역 방식에 의한 교차언어 문서검색의 성능을 알아보기 위하여 Logos 기계번역 시스템을 사용하여 독일어로 된 SDA/NZZ 문서들을 영어로 번역하는 작업을 시도하였다[20]. 이 문서 번역에는 530 MB의 독일어 뉴스 기사를 다섯대의 SPARC-20, SPARC-5 워크스테이션을 사용하여 약 2개월에 걸친 영어 문서로의 번역 작업을 수행하였다. 번역된 영어 문서에 대한 색인 후의 검색 실험 결과는 검색 성능이 단일언어 문서검색에는 미치지 못하지만 질의어 번역 방식과 비교할 때 질의어의 길이가 긴 문장들에 대해서는 분명히 좋은 검색 효과를 보임을 알 수 있었다.

한국어와 관련한 문서 번역 방식의 연구는 [16]에서 처음으로 시도되었다. 한국어 질의어를 이용하여 일본어 특허 문서를 검색하기 위하여 일본어 문서를 COBALT-JK 일한 기계번역 시스템을 이용하여 한국어 문서로 번역하였다. 이때 대역사전의 연어 패턴을 이용하여 올바른 한국어 대역어를 추출하고 효과적인 문서 검색에 필요한 색인어를 생성하는 방법을 사용하였다. 이 방법은 제한된 특허 영역을 대상으로 고성능의 일한 기계번역 시스템을 이용하여 상당히 높은 검색 효과를 얻었다.

2.2 질의어 번역 방식

질의어 번역 방식에서는 질의어가 임의의 통제 어휘(controlled vocabulary)들만으로 구성되는지 혹은 자유 어휘(free text)들을 모두 사용하는지에 따라 질의어 번역 방법들이 세분류된다. 초기의 다국어 문서검색 연구들에서는 통제 어휘들을 사용한 다국어 문서검색 방법을 시도하기도 하였다[23]. 통제 어휘를 이용한 다국어 문서검색이란 미리 정한 어휘들만을 이용하여 검색 문서들을 모두 수작업으로 색인한 후 사용자의 질의어도 같은 어휘들로 표현하여 검색하는 것을 말한다. 통제 어휘를 사용한 다국어 문서검색은 소규모의 제한된 영역에 대해서는 좋은 성능 효과를 보이나 일반 영역에 대하여 수작업으로 많은 양의 문서를 색인하여야 하는 데는 비현실적인 것으로 나타났다.

자유 어휘를 사용한 질의어 번역 방식은 질의어를 검색 대상 문서의 언어로 변환하기 위하여 활용하는 언어 자원의 종류에 따라 대략 세 가지 종류의 질의어 번역 방법으로 분류된다. 즉 질의어 변환에 사전을 이용하는 사전 기반 방법, 시소러스를 이용한 시소러스 기반 방법 그리고 코퍼스를 이용하는 코퍼스 기반 방법을 들 수 있다.

질의어 번역 방식은 최근 웹 환경의 인터넷 다국어 정보검색에서 일반적으로 사용자가 이용하는 질의어의 형태가 구절이나 문장이 아닌 세 단어 이내의 단어간에 연결성이 없는 단순 단어 나열을 많이 사용하기 때문에 질의어 번역시에 발생하는 질의어 단어의 모호성 문제를 해결하는 것이 무엇보다도 중요한 연구 논점이 되고 있다.

2.2.1 사전 기반 방법

사전 기반 방법은 질의어 번역에 대역 사전

(bilingual dictionary)을 사용하는데 대역 사전이 비교적 획득이 쉽고 번역 방식이 단순하기 때문에 가장 많이 적용되고 있는 질의어 번역 방법이다. 하지만 사전 기반 방법의 질의어 번역의 효과는 같은 질의어를 사용한 단일언어 문서검색 효과의 40~60% 정도인 것으로 나타나[3] 다른 방법에 비하여 매우 낮다고 할 수 있다.

사전 기반 방법에서 질의어 단어들에 대한 대역어 선정은 대역사전에서 질의어 단어의 대역어로 가장 많이 쓰이는 최초 대역어를 선정할 수도 있으나 보통 대역사전에 있는 모든 대역어들을 대역어로 채택하여 일종의 질의어 확장을 하는 방식을 많이 택한다[7, 13]. 예를 들어 '첨단 기술 전수의 통제'라는 질의어를 영어 질의어로 변환하는 경우, 단어 '기술'은 대역사전을 이용하여 'art', 'technique', 'skill', 'description'으로 변환된다. 이때 선택된 여러 개의 대역어들로부터 가장 적합한 대역어를 결정하는 문제(term disambiguation problem)와 질의어 단어들의 다수의 대역어들에 적절하게 가중치를 부여하는 문제(term weighting problem)가 중요한 연구 논점이다. 하지만 실제 사전 기반 질의어 번역 방법에서는 질의어에 구절(phrase)이나 숙어 등과 같은 여러 개의 단어가 하나의 의미 단위로 해석해야 하는 복합 단위어(multi-word)가 포함되어 있는 경우에 이들을 잘 인식하고 이에 해당하는 대역어를 대역 사전으로부터 잘 찾아내는 것이 검색 효과에 큰 영향을 미침을 실험을 통하여 알 수 있었다[15].

[14]의 연구에서는 질의어 단어가 여러 개의 모호한 단어들로 번역되는 경우에 이들 단어들을 벡터 공간 모델보다 가중치 불리언 모델(weighted boolean model)로 표현하여 이를 교차언어 검색에 적용하는 실험을 하였다. 비록 적은 수의 질의어를 사용하였지만 질의어 대역어들에의 가중치 부여를 통한 교차언어 문서검색의 가능성을 확인할 수 있었다.

2.2.2 시소러스 기반 방법

다국어 문서검색의 시소러스 기반 방법은 다국어 시소러스(multilingual thesauri)의 개념들을 이용하여 개념에 속한 유사 어휘의 질의어 확장을 통하여 질의어 번역을 수행하는 방법이다. 다국어 시소러스는 하나 이상의 언어들의 단어들을 언어 독립적인 의미적 관계에 의하여 표시하고 서로 연결하여 하나의 시소러스를 구성한 것이다. 다국어 시소러스를 구축하는 방법은 현존하는 시소러스를 번역하는 방법과 여러 개의 단일어 시소러스들을 통합하여 하나의 다국어 시소러스를 구성하는 방법이 있을 수 있다.

먼저 시소러스를 사용한 질의어 확장을 통한 단일언어 문서검색의 연구를 살펴본다. [25]의 연구에서는 Princeton 대학에서 구축한 WordNet을 이용하여 질의어 단어들을 WordNet의 단어 의미들로 의미를 구별하여(sense-disambiguate) WordNet에 있는 단어들로 확장한 후 확장한 그 단어들만을 사용하여 검색하였다. 실험 결과는 비록 정확도와 재현율에서 만족할 만한 수준은 아니지만 한 가지 나타난 주목할 만한 사실은 실험에서 검색된 6,501개의 적합 문서중에서 원래의 TREC-4의 어떤 질의어들로도 검색되지 않는 347개 문서가 WordNet을 이용한 질의어 확장으로 검색될 수 있었다는 것이다.

하지만 이러한 시소러스를 사용한 단일 문서검색의 효과가 전체적으로 좋지 않음에도 불구하고 다국어 문서 검색에서 다국어 시소러스를 사용한 연구는 또 다른 가능성을 보이고 있다. 특히 지금까지도 진행되고 있는 EuroWordNet (EWN) 프로젝트[12]에서는 WordNet 1.4에 독어, 이태리어, 스페인어, 영어 등의 유럽 4개 언어의 단어들 사이의 의미적 관계를 추가하여 대규모 다국어 어휘 데이타베이스를 다국어 시소러스로 구축하고 있다. EWN 데이타베이스로부터 4개 언어의 단어들에 대한 상호 번역의 연결을 제공하는 중간언어 색인(interlingual index)를 만들고 이를 이용하여 질의어와 문서를 개념 색인(conceptual indexing)하여 개념 기반의 문서 검색을 수행한다. EWN을 사용한 교차언어 문서 검색은 유사 어휘 확장을 통하여 정확도를 그대로 유지하면서 재현율을 높이는 효과를 얻게 된다.

이와 같이 일반 어휘 대상의 대규모 다국어

시소러스는 고품질 개념 기반 색인을 가능하게 하고 전문 영역 지식 획득을 위한 좋은 도구로 활용될 수 있으며 교차언어 문서검색에서 단일 언어 문서검색과 거의 유사한 우수한 검색 효과를 얻을 수 있다. 하지만 다국어 시소러스의 구축과 관리는 비용이 많이 들고 사용하기 또한 쉽지 않으며, 시소러스의 설계 단계에서 이미 적용 영역을 고려하여야 하기 때문에 지금까지는 제한된 영역에서 잘 구축되어 활용되어 왔다. 앞으로 EWN과 같은 대규모 다국어 시소러스가 질적인 면에서 완벽히 구축되고 시소러스의 개념에 의한 문서 표현에서 단어 의미 모호성 해소(word-sense disambiguation) 문제가 해결되면 향후 시소러스 기반 교차언어 문서검색 방법은 실제 응용에 활용되어 많은 성과를 얻을 수 있을 것으로 생각된다.

2.2.3 코퍼스 기반 방법

코퍼스 기반 방법은 시소러스 기반 방법과는 대조적으로 병렬 코퍼스(parallel corpus)로부터 얻은 단어 사용 통계 정보를 이용하여 교차언어 문서 검색을 수행한다. 예제 기반 기계번역 시스템에 주로 사용해온 병렬 코퍼스는 보통 문서 쌍이나 문장 쌍 혹은 단어 단위의 쌍들로 정렬되어 구성된다. 교차언어 문서검색에서 우수한 결과를 보이고 있는 단어 벡터 변환(TVT : Term Vector Translation) 방법과 잠재 의미 색인(LSI : Latent Semantic Indexing) 방법에 대하여 살펴본다.

병렬 코퍼스를 이용한 TVT 방법은 병렬 코퍼스의 문서 쌍의 단어들로부터 먼저 2차원 행렬로 색인한 공기 테이블(cooccurrence table)을 만들고 공기 빈도를 표시해 둔다. 그리고 공기 빈도값에 따른 다양한 기준의 임계값을 설정하고 이를 확률적 질의어 변환에 사용한다. 이러한 대역 코퍼스를 이용한 TVT 방법은 단순한 사전 기반 TVT 방법과 비교하여 훨씬 효과적임을 알 수 있었다[4].

Bellcore에 의하여 제안된 LSI 방법[9]은 단일 문서검색뿐만 아니라 최근 교차언어 문서검색에서도 우수한 연구 성과를 내고 있다[10, 23]. 단일 문서검색에서 LSI 방법은 다차원 벡터 공간을 특이치 분해(singular value decomposition)의 행렬 기법에 의하여 통계적으로 유도된 최적의 축소 의미 공간으로 표현하고 이 최적 공간에서 문서간 유사도 정보에 의한 검색을 수행한다. 이러한 LSI 방법은 통계적으로 유도된 '개념'에 의하여 질의어 단어와 공유하지 않는 적합 문서도 잘 검색할 수 있어 최대 30%까지의 검색 효과를 높일 수 있다고 한다. 교차언어 문서검색에 LSI 방법은 대역 코퍼스를 이용하여 적용할 수 있는데, 대역 코퍼스로부터 유도된 두 개의 행렬에 대하여 단일 문서검색에서의 LSI 기법을 같은 방식으로 적용하여 교차언어 문서검색이 이루어진다.

2.3 기타 방법

최근 전통적인 단일 문서검색의 기법들을 언어 변환 방식에 의한 교차언어 문서검색에 적용하는 방법론들이 활발히 연구되고 있다. [5]에서는 문장 단위로 정렬된 대역 코퍼스(bilingual corpus)를 사용하여 의사 적합성 피드백(PRF : Pseudo-Relevance Feedback) 방법과 일반화 벡터 공간 모델(GVSM : Generalized Vector Space Model) 방법을 교차언어 문서 검색에 적용하여 이전의 LSI 방법의 교차언어 문서 검색보다 훨씬 좋은 검색 효과를 얻음을 보이고 있다. 이들 방법을 간단히 소개한다.

일반적으로 적합성 피드백(relevance feedback)은 1차 검색된 문서에 대하여 사용자가 직접 적합한 지를 판단한 후 그 문서를 2차 검색에 이용한다. 이와 약간 다르게 PRF 방법은 사람의 판단에 의하지 않고 상위에 검색된 문서가 적합한 문서라는 가정하에 이들 문서를 2차 검색에 이용한다. PRF 방법의 교차언어 문서 검색에의 적용은 대역 코퍼스를 사용하여 설명될 수 있는데 질의어의 언어로 된 상위에 검색된 문서에 대하여 대역 코퍼스에서 일치하는 대역 문서들을 대치시키는 방식으로 검색이 이루어진다.

GVSM 방법은 기존의 벡터 공간 모델의 관점과는 달리 단어-문서 행렬을 듀얼 공간(dual space)으로 해석하여 문서들에 대한 단어의 분포 패턴을 반영한 문서 검색을 행한다. GVSM 방법에서 교차언어 문서 검색은 PRF 방법과

마찬가지로 대역 코퍼스를 사용하는데, 대역 코퍼스가 질의어 언어의 단어-문서 행렬 A와 대상 문서 언어의 단어-문서 행렬 B로 만들어지고, 이들 행렬의 열을 기준으로 대역 코퍼스의 문서 쌍들이 일치되도록 구성한다. 교차언어 문서 검색은 대역 코퍼스를 하나의 듀얼 공간으로 공유하는 행렬 A, B를 이용하여 계산되어 이루어진다.

3. 다국어 정보검색 시스템

현재 폭발적인 인터넷의 증가와 함께 다수의 일반 사용자들은 정보의 홍수속에서 AltaVista, Lycos 및 Yahoo와 같은 인터넷 검색 서비스를 이용하여 정보를 수집하고 있다. 인터넷 일반 사용자들의 주된 정보 수집 대상인 웹 문서들은 Alis Technology 사의 조사에 의하면 상위 3개국의 언어가 각각 영어(82.3%), 독일어(4.0%), 일본어(1.6%)로 구성되어 있음을 알 수 있다[1]. 이와 같은 영어의 비중을 반영하여 현재의 정보검색 엔진들도 영어에 의존적인 검색 서비스를 제공하고 있는 것이 현실이다. 그러나 비영어권의 사용자가 증가하고 있으며 아울러 비영어권 언어들의 문서들도 계속 증가하고 있는 추세에 따라 사용 언어에 구애받지 않고 신속하게 최신 정보를 검색 및 수집할 수 있는 다국어 정보검색 시스템이 속속 등장하고 있다.

표 1 다국어 정보 검색 시스템들

시스템명	대상언어	특 징
TITAN	English Spanish	• 영어 또는 일본어로 질의어 입력 • 통계 기반 자동 언어 식별, 형태소 기반 색인 • 웹 페이지의 타이틀 번역 • http://sting.navi.ntt.co.jp/titan/titan-e.html
MUNDIAL	English Spanish	• 질의어 번역 방식(Query Translation) • 다른 웹 검색 엔진과 연동 • 35,000 어휘 번역 • http://crl.nmsu.edu/users/madavis/ML/ml.html
AHOPT	English Russian	• 영어 질의어 번역, 러시아 문서 검색 • 질의어 철자 검사 • http://www.aport.ru/defeng.asp
Analogical Language Processor	English French German	• 개념 기반 질의어 해석 • 반자동적 질의어 애매성 해소 • http://www.readware.com/scrips/rwcgicli.exe
Eurospider	English French German Italian	• Relevance ranking, Word normalization • Relevance Feedback • 자동 언어 식별 • http://www.eurospider.ch/eurospider/what/high-1.html
Coronado	English French German Spanish	• 다른 웹 검색 엔진과 연동 • 온라인 문서 번역 • http://www.lhs.com/internetservices/coronado/default.asp
Search '97	. —	• 개념 기반 색인, 시소러스 이용 • 요약(Summarization) 기능 • http://www.verity.com/products/technology.html
Fast Data Finder	Arabic Asian English European Russian	• Parcel사의 텍스트 필터링 시스템 • ASIC 기술을 이용한 색인 없이 고속의 정보 검색 • 주제별 검색 지원 • 검색 프로파일 지원 • http://www.parcel.com/fdfin.htm
CLTR/JK	Japanese Korean	• 120,000 어휘 대역 사전 • 일한 번역 시스템 COBALT-J/K 이용 • 문서 번역(Document Translation) 방식 • http://madonna.postech.ac.kr/cobalt.html

현재 웹문서를 위한 다국어 정보검색 데모 시스템들이 일부 언어들에 대하여 시범 서비스를 제공하고 있는데 MUNDIAL, TITAN 등이 있다. MULDIAL은 영어 질의어를 이용하여 스페인어 문서들에 대한 검색을 수행하는데, 검색 문서들에 대한 번역 서비스도 제공한다[8]. TITAN은 사용자가 일본어나 영어 질의어를 이용하는 웹 검색 시스템으로 검색 결과에 대한 표제어 번역을 제공한다[13]. 다음 표 1은 현재 시범적으로 운영 중인 다국어 정보검색 데모 시스템과 상용화된 다국어 정보검색 시스템의 현황을 나타낸 것이다.

표 2는 유럽권을 중심으로 활발하게 진행되고 있는 다국어 정보검색 시스템 개발과 관련한 다양한 프로젝트들의 현황이다. MULINEX, TwentyOne 등의 프로젝트에서와 같이 웹 문서를 위한 다국어 정보검색 도구 세트 개발, EuroWordNet와 같은 일반적인 어휘 의미망이 유럽권 언어들을 대상으로 범국가적으로 구축되고 있다.

표 2 다국어 정보 검색 관련 프로젝트들

프로젝트명	대상언어	수행년도	특 징
CALAL/LS	English, French, German, Spanish	1995~1997	• EC의 Telematics for Libraries 프로젝트 • 다국어 도서 목록에 대한 검색 지원, 도서 자료들에 대한 언어처리 표준화 작업 수행 • 질의어 변환 방식 이용 • SGML에 기반한 클라이언트 서버 모델 개발 • http://www2.echo.lu/libraries/en/projects/canal.html
CRISTAL	English, French, Italian	1993~1996	• 신문 기사 검색을 위한 개념 기반 정보검색 시스템 개발 • 언어처리와 정보검색 기법의 접목을 시도 • 다국어에 대한 개념 사전 적용 • 색인, 검색, 대화 관리 및 다국어 질의어 인터페이스로 구성 • http://www2.echo.lu/langeng/en/lre2/cristal.html
EMIR	French, English, German	1990~1994	• 텍스트 DB에 대한 언어학적, 통계적 색인 기법 연구 • 타영역으로의 확장을 고려한 영역 의존적 방식 • 다국어 질의어의 가능성을 보임 • http://www-uk.research.ec.org/esp-syn/text/5312.html
EuroWordNet	German, Italian, Spanish English	1996~1998	• WordNet 형태의 다국어 어휘 개념망 구축 • http://www.let.uva.nl/~ewn/
MULINEX	French, English, German	1997~1998	• 키워드, 구, 개념의 조합을 이용한 질의어 사용 • 다국어 Web site 관리를 위한 Tool 제공 • 개념(Concept) 기반 검색에 의한 애매성 해소 • WWW를 위한 다국어 정보 검색기 tool set 개발 • http://www2.echo.lu/lang...le3/mulinex/mulinex.html
TRANSLIB	Greek, Spanish, English	1995~1997	• 다국어 도서목록 및 문서 검색을 위한 Tool 개발 • 도서 목록 코퍼스에서 추출한 전문 과학 용어 및 계층적 분류 정보 이용 • 대상언어 : 그리스어, 스페인어, 영어 • http://peterpan.uc3m.es/proyectos/translib/HomePage.htm
TwentyOne	English, German, Dutch French	1996~1998	• 멀티미디어(문서, 오디오, 비디오) 문서에 대한 효율적 검색을 위한 Tool 개발 • 대상언어 : 영어, 독일어, 네덜란드어, 프랑스어 • 검색 결과에 대한 부분 번역 지원 • http://twentyone.tpd.tno.nl/info/twentyone.html

4. 다국어 정보검색 평가

교차언어 문서검색 실험은 단일언어 문서 검색과 마찬가지로 질의어, 문서, 적합성 판정의 자료들을 이용하여 일반적으로 단일언어 문서 검색의 성능에 대한 교차언어 문서검색의 성능 효과의 비율을 계산하는 방식으로 이루어진다. 교차언어 문서검색의 성능 측정 기준은 단일언어 문서검색과 마찬가지로 검색 효율성(effectiveness)을 주로 사용하고 있는데, 특히 외국어에 익숙하지 않은 교차언어 문서검색의 사용자들을 고려할 때 재현율보다는 정확률을 더 중요시 하여 보통 상위에 검색된 문서의 평균 정확률을 교차언어 문서검색의 성능 측정 방법으로 사용하고 있다.

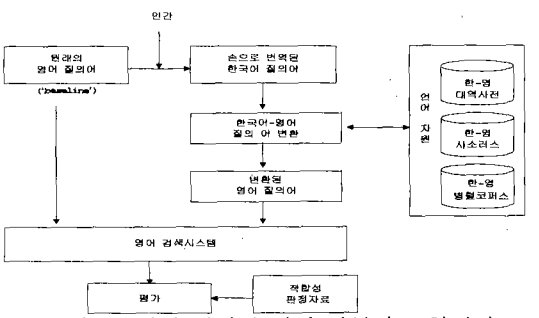


그림 2 한영 질의어 번역 방식의 교차언어 문서검색 성능 평가를 위한 실험 과정

교차언어 문서검색의 성능 평가를 위한 실험 방법은 교차언어 문서검색 방법이 질의어 번역 방식이냐 아니면 문서 번역 방식이냐에 따라 다르다. 만일 질의어 번역 방식의 교차언어 문서검색의 성능 평가인 경우에 먼저 원래의 질의어들을 전문가에 의하여 수작업으로 목적 언어로 번역한다. 그 다음 수작업으로 번역된 질의어에 대하여 질의어 번역 방식에 의한 질의어 번역을 수행하여 다시 원래 질의어 언어로 재번역한다. 마지막으로 적합성 판정 자료에 의하여 두 가지 종류의 질의어인 원래의 질의어를 이용한 문서검색 성능 측정과 재번역된 질의어를 이용한 문서검색 성능 측정을 수행한다. 이렇게 하여 원래 언어의 단일언어 문서검색의 성능과 질의어 번역 방식에 의한 교차언어 문서검색의 성능을 비교하여 교차언어 문서검색의 성능 효과를 평가하게 된다. 그림 2는 한국어 질의어를 이용한 영어 문서 검색에서 한-영 질의어 번역 방식의 교차언어 문서검색 성능 평가를 위한 실험 과정을 보여준다.

교차언어 문서검색의 여러 방법들의 검색 성능 비교는 실험 환경과 각각의 방법들의 고유한 특성으로 객관적인 평가에 다소 논란이 있을 수 있다. 하지만 각 방법들의 향후 연구 방향을 가늠해 본다는 측면에서 대표적인 실험과 그들의 성능을 알아본다.

교차언어 문서검색에 대한 최초의 실험은 60년대 후반과 70년대 초에 Gerald Salton에 의해 행해졌다[23]. 이 실험에서 Salton은 2개의 작은 단일언어의 문서집합(468개의 독일어 요약문서와 1,095개의 영어 요약문서)을 이용하여 48개의 영어 질의어에 대한 검색을 수행하였다. 이때 사용한 방법은 영어로 작성된 시소러스를 수작업으로 번역하여 사용하는 것이었고 수작업으로 영어 질의어를 독일어로 번역하여 검색을 수행하여 번역과정을 통한 정보의 손실이 적음을 밝혔다. 그러나 이러한 결과의 도출과정에서는 작은 문서집합을 사용하였고 시소러스의 번역에 상당한 노력이 소요되어 실제 적용에는 무리가 있는 것으로 나타났다.

영어와 불어의 문서집합에 대한 Radwan의 실험[21]에서는 구체적인 성능 비교가 이루어졌다. 1,400개의 영어 문서로 이루어진 Cranfield collection을 사용하였고 질의어 번역 방식을 채택하여 각 언어에 종속적인 용어사전, 번역사전을 구축하였다. 10~90%의 재현율에서의 평균 정확률에 의한 성능 비교에서는 질의어 번역 방법이 문서 번역 방법보다 우수하였고 통제어휘를 사용하지 않은 방법으로 교차언어 문서검색에 대한 가능성을 보였으며, 다국어 번역사전 구축 노력에 대한 근거를 제시하였다.

TREC의 대규모 문서를 이용한 실험은 Davis&Dunning에 의해서 행해졌다[7]. 58,000개의 스페인어 문서에 대해 25개의 질의어를 통한 실험이 이루어졌다. 이 실험을 위하여 U. N. 코퍼스(1.6GB의 영어-스페인-불어 병렬코퍼스)의 68,000개의 정렬된 문장이 테스트 알고리즘의 학습자료로 사용되었다. 질의어 번

역 방식에 의한 다양한 실험에서 질의어 번역 실험 결과가 원래의 질의어를 이용한 검색 결과와 현격한 검색 성능차이를 보임을 알 수 있었다. 이것은 TREC의 스페인어 질의가 모호하고 상당히 작으며, U.N. 코퍼스가 포함하는 영역과 TREC의 영역이 많이 다름에 기인한다고 본다. 이에 반하여 Hull&Grefenstette의 실험은 좀더 좋은 성능을 보이고 있다[15]. TIPSTER colletion과 TREC의 결과를 사용하여 성능평가를 수행하였는데, 실험에 3 종류의 번역사전을 이용한 질의 번역 방식을 채택하였다. 실험결과는 5, 10, 15, 20개의 문서가 검색되었을 때의 평균 정확율로 측정하였는데 원래의 영어 질의어의 검색 결과에 비하여 단어 기반 대역 사전을 이용한 방법보다 복합단위어 형태의 대역 사전을 이용한 교차언어 문서검색이 효과적임을 알 수 있었다.

지금까지의 대부분의 실험이 질의어 번역 방식인데 반하여 최근에 Oard&Packette는 문서번역 방식을 통한 교차언어 문서검색을 수행하였다[20]. 질의어 번역 방식이 쉽게 구현할 수 있고 단일언어 문서검색의 50~75%의 검색성능을 보이는 반면 언어적 정보가 적은 짧은 질의어에 대해서는 성능의 한계를 보이는 점을 지적하여 문서 전체를 기계번역 시스템을 통해 번역하는 방법을 채택하였다. 실험에서는 문서 제목 질의어와 짧은 질의어(한 문장), 긴 질의어(한 문서내의 모든 내용)의 세 종류의 질의어 길이에 따른 단일언어 문서검색 성능과 문서 번역 방식과 질의 번역 방식의 교차언어 문서번역의 성능을 비교하였다. 실험 결과를 통하여 Oard&Packette는 중규모 정도의 특정 영역검색을 위한 방법으로 문서 번역 방식이 효과적임을 주장하고 있다.

한국어 질의를 이용한 일본어, 영어 문서검색에 대한 연구가 국내에서도 최근에 시작되고 있다[16, 27, 28, 29]. [27]에서는 한국어 질의어의 중의성 해소를 위한 방법으로 공기 가중치 할당 방법과 한·일 대역어 사전과 카도가와 시소러스를 이용한 시소러스 개념수렴 방법을 통하여 다른 실험보다 상당히 좋은 검색 효과를 얻었다. 이러한 실험 결과는 실험 대상 문서가 제한된 영역의 특허 문서이고 한국어와 일본어간에 언어 유사성이 있음을 고려하더라도 본격적인 한국어 관련 교차언어 문서검색 연구에서 중요한 진전이라고 생각된다.

5. 결 론

웹과 인터넷 기술의 급속한 발전에 따라 여러 언어로 작성된 문서가 급격히 증가하고 있고 이들 문서에 대한 검색 요구에 따라 다국어 정보검색 관련 연구가 활발히 진행되고 있다. 본 논문은 교차언어 문서검색 관점에서 문서 번역 방식과 질의어 번역 방식의 여러 방법들을 소개하고 현재 운영중인 다국어 정보검색 시스템들, 그리고 실험 방법과 평가에 대하여 설명하였다 .

향후 다국어 문서검색의 성능 향상을 위해서는 무엇보다도 다국어 문서검색에 필요한 기계번역 시스템이나 여러 언어 자원들이 잘 구축되어야 한다. 일반 어휘에 대한 다양한 의미관계를 가진 대규모 시소러스가 잘 구축되어야 하며 구절이나 숙어 등의 복합단위어 형태를 포함한 대규모 대역 사전 또한 체계적으로 구축되어야 한다. 그리고 여러 언어에 대하여 다양한 영역의 내용을 포함하는 대규모 대역 코퍼스가 구축된다면 이를 이용한 다국어 문서검색 기법은 어떤 방법보다 효과적으로 적용되어 좋은 성과를 얻을 수 있을 것으로 생각된다. 뿐만 아니라 언어 자원 자체의 구축과 함께 이들 언어 자원을 효과적으로 구축, 관리하는 도구가 함께 개발되어야 한다. 특히 텍스트 코퍼스로부터 자동으로 단어를 추출하고 의미 관계를 생성하는 도구는 자동적인 대역 사전의 구축을 가능케 하여 다국어 문서검색의 효과를 높이는데 기여할 것이다.

다국어 문서검색은 대규모 언어 자원의 완벽한 구축이라는 과제와 함께 각각의 다국어 문서검색 방법들에 있어서 해결해야 할 연구 과제들이 있다. 사전 기반 방법에서 대역 사전에 미등록어나 구절 등의 복합단위어를 주기적으로 보완하여 구축하는 작업과 함께 질의어 번역시의 모호성 문제를 효과적으로 해결하는 기법들이 더욱 연구되어야 한다. 이를 위해서 품사 태깅이나 구절 색인 방법을 이용하여 번역

시의 모호성을 해소하는 방법을 채택할 수도 있다.

현재 다국어 문서검색은 단일언어 문서 집합에 대하여 여러 언어로 질의하여 검색하는 단계에 와있다. 하지만 향후에는 여러 언어로 구성된 문서 집합들에 대하여 여러 언어로 질의하여 검색하는 것은 물론이고 하나의 문서가 여러 언어로 구성되어 있는 경우에도 검색하는 단계로 발전해 나갈 것으로 예상된다.

참고문헌

[1] Alis Technology Inc., *Alis Technologies and the Internet Society: Web Languages Hit Parade,* 1997. http://babel.alis.com:8080/palmares.html.

[2] Lisa Ballesteros, W. Bruce Croft, "Dictionary-based methods for Cross-lingual Information Retrieval", In Proceedings of the 7th International DEXA Conference on Database and Expert Systems Applications, 1996.

[3] Lisa Ballesteros, W. Bruce Croft, "Phrasal Translation and Query Expansion Techniques for Cross-lingual Information Retrieval", SIGIR '97, 1997.

[4] Ralf D. Brown, "Corpus-Based Query Translation for Cross-lingual Information Retrieval", Position paper for SIGIR-97 workshop on Cross-Lingual Information Retrieval, July 1997.

[5] Jaime Carbonell, Yimying Yang, Rebert Frederking, Ralf D. Brown, Yibing Geng, and Danny Lee, "Translingual Information Retrieval: A Comparative Evaluation", In Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence, August 1997.

[6] Mark Davis, "New Experiments in Cross-Language Text Retrieval at NMSU's Computing Lab", The Fifth Text Retrieval Conference(TREC-5). NIST, November 1996. http://crl.nmsu.edu/users/madavis/Site/Book2/trec5.ps.

[7] Mark Davis and Ted Dunning, "A TREC Evaluation of Query Translation Methods for Multi-Lingual Text Retrieval", In Proceedings of TREC-4. 1994. http://crl.nmsu.edu/ANG/MWD/Book2/trec4.ps.

[8] Mark W. Davis and William C. Ogden, "Implementing Cross-Language Text Retrieval Systems for Large-scale Text Collections and the World Wide Web", In AAAI Spring Symposium on Cross-Language Text and Speech Retrieval, March 1997.

[9] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer and Richard Harshman, "Indexing by Latent Semantic Analysis", Journal of the American Society for Information Science, 41(6):391-407, 1990. http://superbook.bellcore.com/~std/papers/JASIS90.ps.

[10] Susan T. Dumais, Todd A. Letsche, Michael L. Littman and Thomas K. Landauer, "Automatic Cross-Language Retrieval Using Latent Semantic Indexing", In AAAI Spring Symposium on Cross-Language Text and Speech Retrieval, March 1997.

[11] Christian Fluhr, "Multilingual Information Retrieval", In Ronald A Cole, Joseph Mariani, Hans Uszkoreit, Annie Zaenen and Victor Zue, editors, *Survey of the State of the Art in Human Language Technology,* pp. 391-305, Center for Spoken Language Understanding, Oregon Graduate Institute , 1995. http://www.cse.ogi.edu/CSLU/HLTsurvey/ch8node7.html.

[12] Julio Gilarranz, Julio Gonzalo and Felisa Verdejo, "An Approach to Conceptual Text Retrieval Using the EuroWo-

rdNet Multilingual Semantic Database", In AAAI Spring Symposium on Cross-Language Text and Speech Retrieval, March 1997.

[13] Yoshihiko Hayashi, Genichiro Kikui and Seiji Susaki, "TITAN: A Cross-Linguistic Search Engine for the WWW", In AAAI Spring Symposium on Cross-Language Text and Speech Retrieval, March 1997.

[14] David A. Hull, "Using Structured Queries for Disambiguation in Cross-Language Information Retrieval", In AAAI Spring Symposium on Cross-Language Text and Speech Retrieval, March 1997. http://www.clis.umd.edu/dlrg/filter/sss/papers/.

[15] David A. Hull and Gregory Grefenstette, "Querying Across Languages: A Dictionary-Based Approach to Multilingual Information Retrieval", In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1996.

[16] O-W Kwon, I.S. Kang, J-H Lee & G.B. Lee, "Cross-Language Text Retrieval Based on Document Translation Using Japanese-to-Korean MT System", In Proceedings of NLPRS '97, pp. 101-106, 1997.

[17] Douglas W. Oard, "Alternative Approaches for Cross-Language Text Retrieval", In AAAI Spring Symposium on Cross-Language Text and Speech Retrieval, March 1997. http://www.glue.umd.edu/~oard/research.html

[18] Douglas W. Oard, "Cross-Language Text Retrieval", SIGIR '97 Tutorial on Cross-Language Text Retrieval, 1997. http://www.clis.umd.edu/dlrg/filter/papers/tutnotes.ps.

[19] Douglas W. Oard and Bonnie J. Dorr, "A Survey of Multilingual Text Retrieval", Technical Report UMIACS-TR-9619, Univ. of Maryland, 1996. http://www.ee.umd.edu / medlab / mlir / mlir.html.

[20] Douglas W. Oard and Paul Hackett, "Document Translation for the Cross-Language Text Retrieval at the University of Maryland", The Sixth Text Retrieval Conference(TREC-6). NIST, 1997.

[21] Khaled Radwan, Ver l'Acces Multiligue en Langage Naturel aux Bases de Donnees Textueles, PhD thesis, Universite de Paris-Sud, Centre d'Orsay, 1995.

[22] Bob Rehder, Michael L. Littman, Susan Dumais and Thomas K. Landauer, " Automatic 3-Language Cross-Language Information Retrieval with Latent Semantic Indexing", In Proceedings of TREC-6, 1997.

[23] Gerald Salton, "Automatic processing of foreign language documents", Journal of the American Society for Information Science, 21:187-194, 1970.

[24] Mark Sanderson, "Word Sense Disambiguation and Information Retrieval", In Proceedings of ACM-SIGIR '94, 1994.

[25] A. Smeaton, F. Kelledy and R. ODonnell, "TREC-4 Experiments at Dublin City University: Thresholding Posting Lists, Query Expansion with WordNet and POS Tagging of Spanish", In Proceedings of TREC-4, 1995.

[26] M. Wechsler, P. Sheridan and P. Schauble, "Multi-language Text Indexing for Internet Retrieval", RIAO'97, 1997.

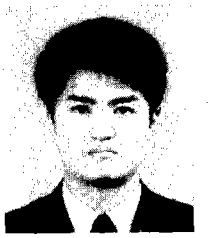
[27] 강인수, 이종혁, 이근배, "교차언어 문서검색에서 질의어의 중의성 해소 방법", 제9회 한글 및 한국어 정보처리 학술대회 논문집, pp. 52-58, 1997.

[28] 강인수, 권오욱, 이종혁, 이근배, "문서 재

순서화를 이용한 질의 변환 방식의 교차언어 문서검색", '98한국정보과학회 춘계 학술대회 논문집, 1998.
[29] 심철민, 여상화, 박동인, 권혁철, "다국어 웹 문서검색을 위한 질의어 변환 기법", '97 한국정보과학회 추계 학술대회 논문집, pp. 209-214, 1997.

장 명 길

1988 부산대학교 계산통계학과 학사
1990 부산대학교 계산통계학과 석사
1990～1996 시스템공학연구소 연구원
1997～1998 시스템공학연구소 선임연구원
1997～현재 충남대학교 컴퓨터과학과 박사과정
1998～현재 ETRI 컴퓨터.소프트웨어기술연구소 선임연구원
관심분야 : 자연어처리, 한국어 구문분석, 다국어 정보검색
E-mail:mgjang@seri.re.kr



김 영 길

1991 한양대학교 전자통신공학과 학사
1993 한양대학교 전자통신공학과 석사
1997 한양대학교 전자통신공학과 박사
1997～1998 시스템공학연구소 선임연구원
1998～현재 ETRI 컴퓨터.소프트웨어기술연구소 선임연구원
관심분야:기계번역, 대화이해, 정보검색
E-mail:ykkim@seri.re.kr

박 영 찬

1992 한국과학기술원 과학기술대학 전산학과 학사
1994 한국과학기술원 전산학과 석사
1997 한국과학기술원 전산학과 박사
1997～1998 시스템공학연구소 선임연구원
1998～현재 ETRI 컴퓨터.소프트웨어기술연구소 선임연구원
관심분야:자연어처리, 정보검색, 한국어 처리, SGML/XML
E-mail:ycpark@seri.re.kr