한영 교차언어 정보검색에서 질의 변환 및 질의 확장 방법

김백일 서희철 임해창
고려대학교 컴퓨터학과
{cedar, hcseo, rim}@nlp.korea.ac.kr

Query Translation and Query Expansion Method in Korean-to-English Cross-Language Information Retreival

Baeg-il Kim Hee-cheol Seo Hae-chang Rim
Dept. of Computer Science and Engineering, Korea University

요 약

본 논문은 한영 교차언어 정보검색을 위한 질의 변환 방법과 질의 확장에 대해서 기술하고 있다. 한영 교차언어 정보 검색은 한국어 질의와 관련된 영어 문서를 검색하는 것을 말하며, 한국어 질의를 영어 질의로 변환하는 방법을 사용했다. 이를 위해 한국어 단어들에 대한 영어 대역어들의 공기 정보를 이용하며, 공기 정보로는 상호 정보를 사용했다. 또한 한국어와 영어의 연어 사전을 사용하여 성능을 향상시켰다.

추가적인 검색 성능 향상을 위한 방법으로, 기존 연구에서 많이 사용된 적합성 피드백에 의한 지역적 질의 확장 대신, 영어 워드넷을 확장하여 구축한 한영 이중언어 시소러스를 사용하여 질의 확장을 하는 전역적 질의 확장을 시도하였다. 실험결과, 정확률의 향상보다는 재현율의 향상 정도가 더 컸으며, 긴 질의보다 짧은 질의를 확장한 경우가 성능이 높았다.

1. 서론

교차언어 정보검색(Cross-Language Information Retrieval; CLIR)은 질의 언어와 다른 언어로 표현된 문서를 검색하는 것을 의미하며, 한영 교차언어 정보검색은 한국어 질의와 관련된 영어 문서를 검색한다.

교차언어 정보검색을 위한 접근 방법은 두 가지가 있다. 하나는 언어 변환을 하는 방법이 있고, 다른 하나는 언어 변환이 없는 방법이 있다. 언어 변환을 하는 방법으로는, 사용자의 질의를 문서를 표현하는 언어로 변환하는 질의 변환 방식과 사용자가 검색하고자 하는 문서를 질의를 표현하는 언어로 변환하는 문서 변환 방식 그리고 질의 변환 방식과 문서 변환 방식을 같이 사용하는 혼합 방식이 있다.

질의 변환 방식은 이중 언어 사전이나 이중 언어 온톨로지에서의 대역어 정보를 이용해서 대역 질의를 생성한다. 이 방법은 질의의 길이가 대체로 짧기 때문에 변환에 드는 비용이 적다는 장점이 있다. 그러나 변환에 사용되는 정보가 적으므로 변환의 정확도가 낮고, 질의가 입력될 때마다 언어 변환을 수행해야 된다는 단점이 있다.

문서 변환 방법[Oard1997]은 질의를 표현하는 언어로, 문서를 변환한 후에 정보 검색을 하는 방법이다. 이는 질의 변환에 비해서 변환을 위한 정보가 많다는 점에서 변환의 정확성이 다소 높을 수 있지만, 대체로 문서의 길이가 질의 길이보다 많이 길다는 점에서 변환을 위한 비용이 많이 든다는 점과, 문서 변환을 위해서는 뛰어난 성능의 기계 번역 시스템이 필요하다는 단점이 있다. 현재 영한 기계 번역 시스템의 성능이 만족스럽지 못하므로, 기계번역 시스템을 한영 교차 언어 정보 검색에 사용하기에는 무리가 있다. 더구나, 다른 언어로의 확장시 해당 언어의 기계 번역 시스템이 필요하므로 자원과 시간의 부하를 견디기 힘들다[강인수1997].

언어적인 변환을 수행하지 않는 방법으로는 잠재 의미 색인(Latent Semantic Indexing; LSI)가 있다[Dumais1997]. LSI는 질의와 문서를 하나의 공간에 투영하고, 공간에서 질의와 가까운 위치에 있는 문서를 질의와 관련된 문서로 간주하는 방법이다. 이는 언어적인 변환 과정이 필요하지 않으며, 색인되지 않은 용어들로 구성된 질의문에 대해서도 대상 문서를 검색할 수 있다는 장점이 있다. 그러나 이를 위해서 질의와 문서를 하나의 공간에 투영하게 되는데, 투영하는데 사용되는 수학적인 처리가 매우 복잡하다는 점과 하나의 공간에 투영하기 위해서는 대량의 병렬 말뭉치가 필요하다는 단점이 있다. 대량으로 구축된 한영 병렬 말뭉치가 없다는 점에서 LSI 방법을 한영 교차 언어 정보 검색에 적용하기는 곤란하다.

본 논문에서는 교차언어 정보검색에서 간단하게 언어 변환을 수행할 수 있는 질의 변환 방식에 대해서 살펴보며, 질의 변환 방식에서 정확한 대역 질의 생성의 어려움을 해결하기 위해서 연어를 사용하는 방법을 제안한다.

또한, 교차 언어 정보 검색에서는 질의 변환 과정에서의 변환 중의성으로 인해 단일 언어 정보 검색보다 성능이 떨어지게 되는데,

이를 보완하기 위해 단일 언어 정보 검색에서 널리 사용되어 왔던 여러 가지 질의 확장 방법들이 교차 언어 정보 검색에서도 최근에 많이 사용되고 있다.

관련 연구에서는 주로 적합성 피드백을 사용하여 질의를 확장했는데[McNamee2002]. 이것은 검색된 문서의 일부만의 정보를 사용하는 것으로 지역적(local) 질의 확장이라고 한다. 이러한 방법 외에도, 문서 집단 전체나 대상 언어 전체의 정보를 사용하여 질의를 확장하는 방법이 있는데, 이를 전역적(global) 질의 확장이라고 하며, 이를 위한 추가적인 언어 자원으로, 시소러스나 공기 기반 지식베이스와 같은 온톨로지를 사용하게 된다([Baeza-Yates1999][박지연2000]). 본 논문에서는 전역적 질의 확장이 교차 언어 정보 검색에서 유용한지를 알아본다.

2. 질의 변환과 중의성 해소

2.1. 관련 연구

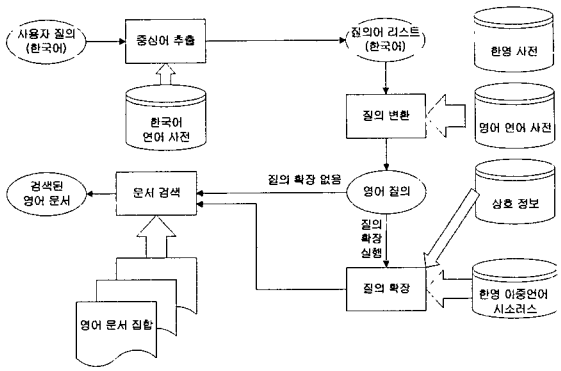
질의 변환을 위한 방법으로는 이중 언어 사전에 기반한 방법, 말뭉치에 기반한 방법, 그리고 다국어 온톨로지에 기반한 방법이 있다. 이중 언어 사전을 이용하는 경우에는 질의 단어의 대역 단어를 사전을 참조해서 획득하며, 획득된 대역단어를 사용해서 대역 질의를 생성한다[Hull1996][Ballesteros1996]. 말뭉치에 기반한 방법은 병렬 말뭉치를 사용해서 질의를 변환하거나 질의를 확장하는 방법을 사용한다[Brown1997][Davis1995]. 다국어 온톨로지는 단어간의 관계를 계층적으로 표현하고 있는 것으로 이중 언어 사전에 비해서 더 많은 정보를 포함하고 있다. 이런 다양한 어휘간의 관계 정보를 사용해서 질의를 더 견고하게 변환하고자 시도하는 방법이 다국어 온톨로지에 기반한 질의 변환 방법이다[Gilarranz1997][Eichmann1998].

한국어와 관련된 교차언어 정보검색에서 질의 변환을 위한 기존 연구들은 사전과 온톨로지를 기반으로 한다. [강인수1997]은 질의에 사용된 단어들은 유사한 개념을 가지며, 이들은 온톨로지에서 하나의 개념으로 수렴된다는 정보를 이용해서 한국어 질의를 일본어 질의로 변환했다. [장명길1999][장명길2002]는 사전에서 대역어를 추출하고, 인접한 단어의 대역어들간의 상호정보를 이용해서 한국어 질의를 영어로 변환했다. [천정훈1999]는 다국어 온톨로지에서 한국어 단어의 영어 대역어들을 이용해서 한국어 단어들의 개념을 결정하고, 결정된 개념에 속하는 영어 대역어들에 대해서 dice coefficient를 이용해서 질의를 생성했다.

2.2. 전체 구성

한영 교차언어 정보 검색의 구현을 위해서 질의 변환 방법을 사용한다. 즉, 한국어 질의를 영어 질의로 변환한 후에, 영어 정보 검색 시스템을 사용해서 영어 문서를 검색한다.

한영 교차언어 정보검색을 위한 전체 시스템 구성도는 [그림 1]과 같다. 한영 교차 언어 검색기는 중심어 추출 모듈, 질의 변환 모듈, 문서 검색 모듈, 질의 확장 모듈의 네 가지 모듈로 구성된다. 중심어 추출 모듈에서는 품사 부착기와 한국어 연어 사전을 이용해서 중심어를 추출한다. 질의 변환 모듈에서는 중심어 추출 단계의 결과물인 중심어들을 대상으로 한영 사전, 영어 연어 사전[1], 말뭉치에서 획득한 상호 정보를 사용해서 질의 변환 즉 영어 질의를 생성한다. 한영



[그림 1] 시스템 구성도

사전은 한국어 중심어들의 영어 대역어들을 구하는데 사용되고, 영어 연어 사전은 영어 대역어들간의 연어 관계를 알아보는데 사용된다. 그리고 상호 정보는 생성된 대역 질의들에 가중치를 부여하는데 사용된다. 질의 변환으로 생성된 영어 질의는 질의 확장 모듈에서 확장 된 후, 영어 문서 검색 모듈의 입력으로 영어 문서를 검색한다.

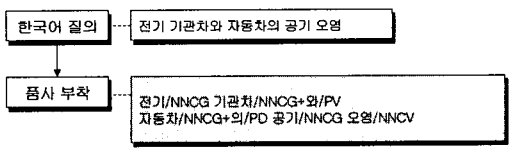
가령, "전기 기관차와 자동차의 공기 오염"이라는 한국어 질의에 대해서 중심어 추출을 통해서 "전기기관차 자동차 공기 오염"이라는 중심어가 추출된다. "전기기관차"의 경우 한국어 연어 사전에 의해서 한국어 연어로 추출된다. 추출된 중심어에 대해서 질의 변환을 수행하면, "electromotive automobile air pollution"이라는 영어 질의가 생성되며, 생성된 영어 질의로 영어 문서를 검색한다.

2.3. 중심어 추출

중심어 추출은 한국어 질의에서 의미 있는 단어만을 추출하는 과정이다. 여기서, 의미 있는 '단어'란, 질의에 있는 한국어 연어와 한국어 명사 단어를 의미한다. 기존의 교차언어 정보검색 연구에서는 중심어를 추출할 때 하나의 명사 단어만을 고려하는 방법을 사용했으나, 본 논문에서는 연어 사전을 사용하여 질의 변환의 중의성 해소에 도움을 주고자 한다.

한국어 연어와 한국어 명사 단어 추출은, 한국어 질의에 품사 부착을 하는 단계와 한국어 질의에서 연어를 인식하는 단계 그리고 한국어 질의에서 연어를 제외한 명사를 추출하는 단계의 세 가지 단계로 수행된다.

품사 부착 단계에서는, HMM 기반 한국어 품사 부착 도구[김진동1997]를 사용한다. 예를 들어, "전기 기관차와 자동차의 공기 오염"이라는 질의는 [그림 2]와 같이 품사 부착된다.



[그림 2] "전기 기관차와 자동차 공기 오염"에 대한 품사 부착

품사 부착된 한국어 질의에서 한국어 연어를 인식하게 된다. 한국어 연어 인식을 위해서는 한국어 연어 사전이 필요하며, 한영 사전에는 한국어 단어쌍이 표제어로 등록되어 있으며, 이런 단어쌍을 한국

1) 영어 워드넷(WordNet)을 사용하여 추출했다.

어 연어로 간주해서 사용했다. 가령, "전기기관차(electromotive)", "주식회사(corporation)", "과학기술(technology)"등이 한영 사전에 등록되어 있으며, 이들을 한국어 연어로 사용한다.

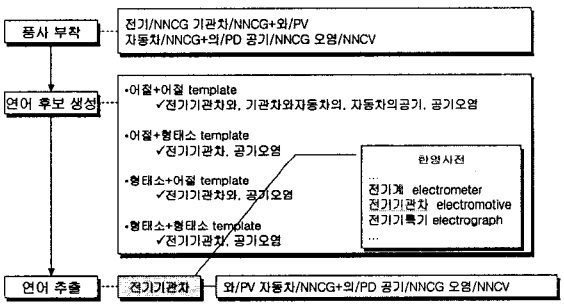
[표 1]

한국어 연어 템플릿
어 절 + 어 절
어 절 + 형태소
형태소 + 어 절
형태소 + 형태소

한국어 연어 사전을 이용해서 한국어 질의에 나타난 한국어 연어를 인식하기 위해서, 한국어 연어 템플릿을 사용해서 한국어 질의에서 연어 후보를 생성하고, 한국어 연어 사전을 참조해서 한국어 연어인지를 확인한다. 한국어 연어 템플릿은 [표 1]과 같다.

명사 이외의 품사가 부착된 형태소는 분리자 역할을 한다. 가령, "주식을 회사가"에서 '을'이 분리자 역할을 하므로, '주식회사'라는 한국어 연어 후보는 생성되지 않는다. 생성된 한국어 연어 후보가 한국어 연어 사전에 있는 경우, 연어로 추출한다.

[그림 3]은 한국어 연어 추출에 대한 예이다. 연어 후보 생성을 위해서 한국어 연어 템플릿을 사용하며, 품사 부착된 질의에서 생성 가능한 모든 연어 후보를 생성하게 된다. 생성된 연어 후보들에 대해서 한국어 연어 사전을 참조해서 최종적으로 "전기기관차"라는 연어를 추출하게 된다.



[그림 3] 한국어 연어 추출 예

한국어 질의에서 연어를 추출한 후에, 연어를 제외한 명사 단어를 추출하게 된다. 명사 단어 추출은 품사 부착된 한국어 질의에서 쉽게 획득할 수 있다. [그림 4]는 한국어 질의 "전기 기관차와 자동차의 공기 오염"에서 한국어 연어 추출 후에 한국어 명사를 추출한 결과이다.

[그림 4] 한국어 명사 중심어 추출 예

[그림 4]에서 "전기기관차"는 이미 연어로 추출되었기 때문에, 명사 추출의 대상이 아니며, 추출된 명사는 "자동차", "공기", "오염"이다. 연어 추출과 명사 추출의 과정을 거침으로써, 한국어 질의 "전기 기관차와 자동차의 공기 오염"에서 추출된 중심어는 "전기기관차", "자동차", "공기", "오염"이다.

2.4. 질의 변환

질의 변환에서는 중심어 추출 단계를 거치면서 생성된 한국어 질의의 중심어를 사용해서 영어 질의를 생성한다. 질의 분해, 대역 질의 후보 생성, 대역 질의 선정의 세 가지 단계를 거치면서 질의 변환을 수행하게 된다.

2.4.1. 질의 분해

한국어 질의의 길이가 긴 경우에 질의 분해를 수행하게 된다. 질의의 길이가 긴 경우, 생성될 수 있는 대역 질의 후보들이 너무 많아지게 되므로, 이를 방지하기 위해서 질의 분해를 수행하게 된다.

[표 2] 한국어 질의에 대한 대역 정보

	title	description	narrative
질의 수	50	50	50
질의 내의 한국어 단어의 총수	105	294	824
영어 대역어의 총수	573	2023	5524
질의당 한국어 단어의 평균 수	2.1	5.88	16.48
단어당 대역어의 평균 수	5.36	6.83	6.67
가능한 평균 조합의 수	33.98	80619.89	3.8×10^{13}

[표 2]는 TREC-8의 adhoc-retrieval task에서 사용된 질의를 한국어로 번역한 후에 살펴본 자료이다.[2] title, description, narrative는 TREC 질의의 title, description, narrative 부분을 의미하며, TREC-8의 질의 수는 모두 50개이다. '질의에 있는 한국어 단어의 총수'는 TREC-8의 질의를 한국어로 번역한 후에 추출된 중심어의 수이며, '영어 대역어의 총수'는 한국어 질의에 대해서 생성되는 영어 대역어들의 수이다. 그리고 '가능한 평균 조합의 수'는 질의 변환을 위해서 고려해야 되는 조합의 수를 의미하며, 질의당 한국어 단어수의 평균과 단어당 대역어의 평균수를 사용해서 구한다. 가능한 평균 조합의 수는 title에서는 $5.36^{2.1}$($\approx$33.98)개이고, description에서는 $6.83^{5.88}$($\approx$80619.89)개, 그리고 narrative에서는 $6.67^{16.48}$($\approx$38158715139976.62)개이다. 이런 조합들을 모두 고려해서 대역 질의를 생성하는 데에는 너무 많은 시간이 필요하게 된다.

이 문제를 해결하기 위한 방법으로, [장명길1999][장명길2002]에서는 가장 큰 상호 정보값을 가진 단어 쌍을 먼저 선택한 뒤, 그래프 이론에 기초한 one-best, maximum spanning tree, best-first & expand의 세 가지 방법을 사용하여, 상당히 적은 계산량으로 빠르게 질의를 변환하는 알고리듬을 제시하였다.

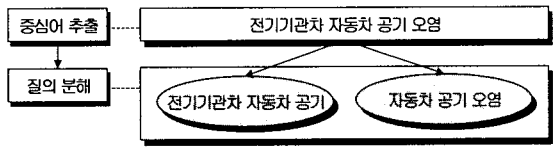
그러나 이러한 방법은 질의어의 순서 정보에 큰 영향을 받으므로, 한국어와 영어와 같이 두 언어의 어순이 유사하지 않은 경우는 성능이 저하되는 원인이 될 수 있을 것이라고 추측할 수 있다. 그러므로, 질의어의 순서 정보에 큰 영향을 받는 방법 보다는, 어순의 영향을 최소화할 수 있는 방법을 사용하는 것이 바람직하다고 판단된다.

그렇게 하기 위해서는, 한국어 질의에서 관련된 단어들을 단위로 영어 대역어를 생성하는 것이 바람직하다. 이런 단위를 생성하는 도구로 구문 분석기와 청커(chunker)가 있는데, 현재의 구문 분석기에서는 높은 정확도를 기대하기가 어렵고, 계산량도 상당히 높다. 한편, chunker는 조사없이 인접한 단어들만을 단위로 분리한다는 문제가 있다. 즉, chunker를 사용하면 "정보 검색 시스템이란"이라는 질의에 대해서는 "정보 검색 시스템"을 하나의 단위(chunk)로 분리할 수 있지만, "정보를 검색하는 시스템이란"이라는 질의에 대해서는 "정보", "검색", "시스템" 3개의 단위로 분리하게 되며,

2) 질의에서 추출한 중심어에 대해서 획득한 정보임.

이 경우에는 질의 변환을 위한 단위로서의 가치가 없게 된다. 또한 chunker에 의해서 생성되는 오류를 질의 변환에서도 물려받게 된다는 문제점이 있다.

이에 본 논문에서는 보다 간단한 방법으로서, 윈도우를 단위로 질의를 분해한다. 윈도우는 질의에서 인접해서 나타난 n(n은 윈도우의 크기)개의 단어들로 이루어지며, 두 개의 윈도우가 하나 이상의 단어를 공유할 수 있다. 즉, 하나의 단어가 두 개 이상의 윈도우에 포함될 수 있다. 이렇게 함으로써 두 가지 효과를 기대할 수 있다. 첫째는 생성된 질의에 단어의 정확한 대역어가 포함될 가능성이 높아진다는 점이다. 단어가 하나의 윈도우에만 들어가는 경우에는, 하나의 윈도우에서 발생하는 대역어 선택 오류로 인해서 정확한 대역어를 질의에 포함할 수 없게 된다. 그러나 단어를 여러 곳에 분산하는 경우에는 이런 위험이 줄어들게 된다.



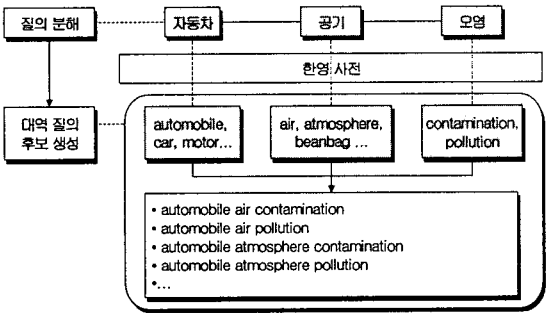
[그림 5] 질의 분해 예

[그림 5]는 중심어 추출 단계에서 추출된 한국어 질의의 중심어 "전기기관차 자동차 공기 오염"에 대한 질의 분해 결과 "전기기관차 자동차 공기"와 "자동차 공기 오염"이라는 두 가지 질의가 생성된 것을 나타낸다.

2.4.2. 대역 질의 후보 생성

대역 질의 후보 생성 단계에서는 한영 사전을 참조해서 한국어 중심어에 대한 대역어를 획득하고, 획득된 대역어들의 조합으로 가능한 대역 질의 후보를 생성한다.

대역 질의 후보 생성을 위해서는 한영 사전이 매우 중요하다. 왜냐하면, 중심어가 한영 사전에 없는 경우에는 대역 질의 후보를 제대로 생성할 수 없기 때문이다. 이런 문제를 해결하기 위해서 세 가지의 한영 사전을 통합했으며, 생성된 한영사전에는 한국어 표제어는 64727개, 영어 표제어는 55224개가 있다.



[그림 6] 대역 질의 후보 예

기존 한영 사전은 일반적인 단어만을 포함하고 있다는 문제점이 있다. 이에 본 논문에서는 특정 영역(domain)에서의 병렬 말뭉치로부터 한영 사전을 자동으로 획득해서 사용했다. 병렬 말뭉치는 한국어와 영어 문장이 정렬되어 있으며, 여기에서 한국어 단어의 대역 영어 단어를 추출한다. 병렬 말뭉치에서 한국어 단어와 자주 나타나는 영어 단어를 한국어 단어의 대역어로 간주할 수 있다.

질의 분해로 생성된 한국어 질의 "자동차 공기 오염"에 대한 대역 질의 후보를 [그림 6]에 나타냈다. "자동차", "공기", "오염"의 영어 대역어들의 조합을 통해서 "automobile air contamination", "automobile air pollution"등과 같은 대역 질의 후보를 생성한다.

2.4.3. 대역 질의 선정

대역 질의 선정 단계에서는 생성된 대역 질의 후보들 중에서 한국어 질의의 가장 적절한 대역 질의를 선정한다. 영어 연어 정보와 영어 말뭉치에서 획득한 영어 단어간의 상호 정보를 함께 사용해서 대역 질의를 선정한다.

대역 질의 후보에 가중치를 부여하고, 가중치가 가장 높은 대역 질의 후보를 대역 질의로 선정한다. 대역 질의 후보에 있는 단어간의 가중적 상호 정보와 단어간의 연어 정보를 사용해서 대역 질의 후보에 가중치를 부여한다. 대역 질의 후보에 가중치를 부여하는 방법으로 상호 정보를 사용하는 방법이 있지만[장명길1999][장명길2002], 단순한 상호 정보는 빈도가 낮은 단어에게 높은 값을 부여한다는 문제점이 있다. 이런 문제점을 해결하기 위해서 본 논문에서는 상호 정보에 가중치를 부여한 가중적 상호 정보를 사용해서 대역 질의 후보에 가중치를 부여한다.

대역 질의 후보에 가중치를 부여하는 방법은 수식 (1)과 같다.

$$weight(Q_k) = \sum_{i=1}^{n_1} \sum_{j=i+1}^{n_1} (wMI(q_{ki}, q_{kj}) + C(q_{ki}, q_{kj})) \quad (1)$$

$$C(q_{ki}, q_{kj}) = \begin{cases} 100 & \text{if } collocation(q_{ki}, q_{kj}) \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

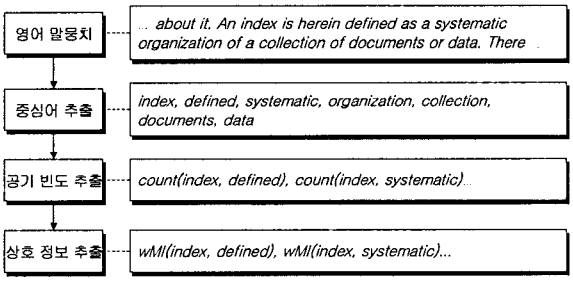
$$\begin{aligned} wMI(x, y) &= \log(count(x, y)) \times MI(x, y) \\ &= \log(count(x, y)) \times \log_2 \frac{P(x, y)}{P(x)P(y)} \end{aligned} \quad (3)$$

수식(1)에서 Q_k는 k번째 대역 질의 후보, wMI는 가중적 상호 정보, C는 연어에 대한 indicator 함수, q_{ki}는 Q_k의 i번째 단어를 의미한다. 수식 (2)에서 $C(q_{ki}, q_{kj})$는 q_{ki}와 q_{kj}가 연어이면 100을, 연어가 아니면 0이 된다. 수식 (3)에서 wMI는 상호 정보에 단어간의 공기 빈도를 가중치로 사용한 가중적 상호 정보이다.

가중적 상호 정보는 영어 말뭉치에서 추출하며, 가중적 상호 정보를 추출하는 과정을 [그림 7]에 나타냈다. 가중적 상호 정보를 추출하기 위해서, 영어 말뭉치에서 중심어 추출, 공기 빈도 추출, 상호 정보 추출의 세 가지 단계를 거치게 된다. 중심어 추출 단계에서는 영어 말뭉치에서 불용어를 제외한 모든 단어를 추출하게 된다. 공기 빈도 추출 단계에서는 추출된 중심어들간의 공기 빈도를 추출하게 되며, 마지막으로 상호 정보 추출 단계에서 단어간의 공기 빈도를 이용해서 가중적 상호 정보를 추출하게 된다.

[표 3]은 상호 정보와 가중적 상호 정보의 차이를 보여준다. 단어 a와 b가 x와 y에 비해서 공기할 확률이 낮지만, 상호 정보는 더 크다. 이는 상호 정보가 낮은 빈도의 단어들에게 높은 값을 부여하기 때문이다. 이런 문제를 해결하기 위해서 본 논문에서 사용한 가중적 상호 정보를 보면, x와 y의 가중적 상호 정보가 a와 b의 가중적 상호정보보다 더 큰 것을 알 수 있다.

대역 질의 후보에 가중치를 부여하는 수식 (1)에서 $C(q_{ki}, q_{kj})$를 계산하기 위해서는 영어 연어를 인식할 수 있어야 되는데, 이를



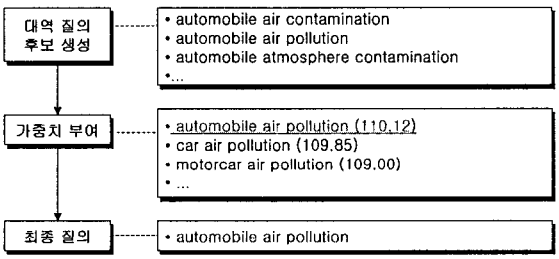
[그림 7] 가중적 상호 정보 추출 예

[표 3] 상호정보와 가중적 상호정보

count(a)	count(b)	count(a,b)	corpus size	MI(a,b)	wMI(a,b)
10	10	3	200000	12.55	5.99
count(x)	count(y)	count(x,y)	corpus size	MI(x,y)	wMI(x,y)
100	100	70	200000	10.45	19.28

위해서 본 논문에서는 영어 WordNet에 있는 언어 정보를 사용한다. 영어 WordNet에는 영어 연어 정보가 품사별로 분류되어 있으며, 이 중에서 명사와 관련된 연어만을 사용했다[3]. 영어 WordNet에서 41587개의 연어 정보 추출해서 사용했다.

영어 연어 인식을 위해서 대역 질의 후보에서 연어 후보를 생성하고, 생성된 연어 후보가 영어 연어 사전에 있는지 확인하는 과정을 거치게 된다. 대역 질의에 있는 모든 단어쌍이 영어 연어 후보가 된다. [그림 8]은 가중적 상호 정보와 영어 연어 정보를 사용해서 대역 질의 후보에 가중치를 부여하고, 대역 질의를 선정한 예를 보여준다.



[그림 8] 대역 질의 선정 예

[그림 8]의 가중치 부여 단계에서, "automobile air pollution"이 가장 높은 가중치를 할당받기 때문에 대역 질의 후보들 중에서 "automobile air pollution"이 대역 질의로 선정된다[4].

3. 질의 확장

3.1. 기존 연구

교차언어 정보검색에서는 질의 변환 과정에서의 변환 중의성으로 인해 단일 언어 정보 검색보다 성능이 떨어지게 되는데, 이를 보완하기 위해서, 단일 언어 정보 검색에서 널리 사용되어 왔던 여러 가지 질의 확장 방법들이 교차 언어 정보 검색에서도 많이 사용되어 왔다 [McNamee2002].

교차언어 정보 검색 관련 연구에서는 주로 적합성 피드백을 사용하여 질의를 확장했는데, 이것은 검색된 문서의 일부만의 정보를 사용하는 것으로 지역적(local) 질의 확장이라고 한다. 이러한 방법 외에도, 문서 집단 전체나 대상 언어 전체의 정보를 사용하여 질의를 확장하는 방법이 있는데, 이를 전역적(global) 질의 확장이라고 하며, 이를 위한 추가적인 언어 자원으로, 시소러스나 공기 기반 지식베이스와 같은 온톨로지를 사용한다. 교차언어 검색에서는 이러한 방법은 거의 없었으며, 단일 언어 검색에서는 여러 실험에서 사용되어 왔다. ([Voorhees1993] [Voorhees1998] [Mandala1998] [Baeza-Yates1999] [박지연2000]).

질의 변환에 의한 교차언어 정보검색에서 질의 확장은 두 단계로 나눌 수 있는 데, 질의 변환 전에 원본 질의를 확장하는 경우는 변환 전 확장(pre-translation expansion)이라고 하며, 질의 변환된 대상 언어의 질의를 확장하는 경우는 변환 후 확장(post-translation)이라고 한다[McNamee2002].

최근의 연구인 [McNamee2002]에서는 교차언어 정보검색에서의 적합성 피드백을 사용한 질의확장에만 초점을 두고, 주어진 언어 자원의 품질을 감소시켜 가면서, 질의 확장의 성능 향상 정도를 평가하였다. 실험 결과, 변환 전 확장과 변환 후 확장을 모두 사용한 경우가 가장 성능이 높았으며, 특히 변환 전 확장이 성능 향상 폭이 컸다. 그에 비해 변환 후 확장의 성능 향상 정도는 적었다. 또한 언어 자원의 품질이 낮고, 질의가 짧을수록 질의 향상의 성능 향상 폭이 커졌다.

일부의 예외는 있지만 많은 수의 연구에서, 변환 전 확장은 변환 후 확장에 비해 성능 향상 정도가 크다는 장점이 있지만, 원본 질의가 대상 문서 집합에 알맞게 확장되기 위해서는, 대상 문서 말뭉치와 주제가 일치하는 비교 말뭉치(comparable corpus)가 있어야 한다는 조건이 있다. [McNamee2002]에서는 CLEF-2001 [Peters2001]이라는 평가 문서 집합으로서, 영어, 독일어, 불어, 네덜란드어, 이탈리아어, 스페인어로 이루어진 약 백만 건의 문서를 포함하는 고품질의 유럽어 비교 말뭉치를 사용했다. 이 방법을 본 논문과 같은 한영 교차언어 정보검색에 적용하려면, 한영 비교 말뭉치가 필요한데, 현재는 CLEF-2001과 같은 대량의 고품질 비교 말뭉치가 구축되어 있지 못하므로, 적합성 피드백에 의한 변환 전 확장을 구현하더라도 성능 향상 폭이 크지 않을 것이라고 예상할 수 있다.

3.2. 전역적 질의 확장

본 논문에서는, 단일언어 검색의 전역적 질의 확장 실험에 비교하여, 교차 언어 정보 검색에 이중언어 시소러스를 사용하는 방법을 제시한다.

전역적 질의 확장을 위한 언어 자원으로서, 한영 이중언어 시소러스를 사용하였다. 이것은 영어 워드넷(WordNet)을 기초로, 한국어 말뭉치에서 뽑아낸 단어들을 대응시킨 구조의 시소러스이다.

본 논문에서 사용한 질의 확장 방법은 다음과 같다. 질의 변환의 중의성 해소 결과, 한국어 단어에 대응되는 하나의 영어 단어의 쌍을 결과로 얻게 된다. 이 (한국어 단어, 영어 단어)의 순서쌍을

3) 동사, 형용사, 부사와 관련된 연어가 있지만, 이런 연어에는 거의 항상 불용어가 포함되어 있기 때문에 사용하지 않았다.

4) 여기서 "air pollution"은 영어 연어이다.

한영 이중언어 시소러스에서 검색하면, 이 순서쌍이 속하는 synset(synonym set, 동의어 집합)을 얻게 된다. 이 synset 안에 있는 영어 단어들을 사용하여 질의 확장하는 방법이다.

여기서 단순히 synset 안의 모든 단어를 질의 확장에 사용하면, 대상 문서에서 상당히 저빈도로 나타나는 단어나 다른 질의어와의 공기 정보 값이 상당히 낮은 질의어를 포함할 수 있다. 이러한 단어들은 질의에서 잡음(noise)으로 작용하여 성능을 저하시키는 요인이 된다. 그러므로, 단어의 대역 가중치(상호 정보)를 임계값(threshold)로 사용하여, 잡음으로 작용할 수 있는 단어들을 제외하고 질의 확장에 사용한다.

이와 같이 질의 확장에 사용될 단어를 선택하는 방법에 따라 질의 확장의 성능이 달라질 수 있다. 다음은 이 논문에서 제안하는 4가지 선택방법에 대한 명칭과 그에 대한 설명이다.

ex_m9 (m: mutual information ≥ 9) 방법:
- 질의 확장에 사용되는 단어는, 대역 질의에 있는 단어와 동일한 synset에 있는 단어들이다.
- 기 생성된 질의에 있는 모든 단어와 질의 확장에 사용된 단어간의 상호 정보값을 구한다.
- 상호 정보 값의 합이 9이상인 단어가 질의 확장에 사용된다. (여기서 9는 실험에 의해 결정하였다.)

ex_c_m9_w3 (c: child synset, w: window size = 3) 방법:
- 질의 확장에 사용되는 단어는, 대역 질의에 있는 단어와 동일한 synset에 있는 단어와 그 synset의 바로 자식 synset에 있는 단어들이다.
- 기 생성된 질의에서, 질의 확장에 사용될 단어의 앞 3단어, 뒤 3단어에서 상호 정보값을 구한다.
- 상호 정보 값의 합이 9이상인 단어가 질의 확장에 사용된다.

위의 방법을 예를 들어 설명하면 다음과 같다.

한국어 질의 : k_1 k_2 k_3 k_4 k_5 k_6 k_7 k_8

생성된 영어 질의 : e_1 e_2 e_3 e_4 e_5 e_6 e_7 e_8
- 생성된 영어 질의에는 8개의 영어 단어(e_n)와 한개의 synset ID (s_2)이 있다.
- s_2은 e_2가 속한 synset ID이다.

ex_m9 방법 :
- 확장에 추가될 후보 단어들
  - s_2에 속한 단어들. (예를 들어, s_2에는 단어 w_{21}, w_{22}, w_{23}가 있다고 하자)
- 단어의 상호 정보 값
  - w_{21}과 e_1, e_2, e_3, ..., e_8의 상호 정보값을 구해서 합한 값을 w_{21}의 대역 가중치로 둔다.
  - w_{22}, w_{23}에 대해서도 동일한 방법으로 구한다.
- 확장에 추가할 단어 선택
  - 단어의 대역 가중치가 9이상의 단어를 확장에 사용한다.

ex_c_m9_w3 방법:
- 확장에 추가될 후보 단어들
  - (s_2에 속한 단어들) ∪ (s_1의 자식 synset에 속한 단어들.)
  - 예를 들어, s_2에는 단어 w_{21}, w_{22}, w_{23} 가 있고, s_2의 하위에는 하나의 자식 s_{21}이 있고, 이 안에는 단어 w_{211}, w_{212}, w_{213}이 있다고 하자.
- 단어의 상호 정보 값
  - s_2가 두번째 단어가 속한 synset이므로, e_1, e_2, e_3, e_4와의 상호 정보를 구한다.
  - 즉, w_{21}과 e_1, e_2, e_3, e_4의 상호 정보를 구해서 합한 값을 w_{21}의 대역 가중치로 둔다.
  - w_{22}, w_{23}, w_{211}, w_{212}, w_{213}에 대해서도 동일한 방법으로 구한다.
- 확장에 추가할 단어 선택
  - 단어의 대역 가중치가 9이상의 단어를 확장에 사용한다.

4. 실험 및 결과

4.1. 연어를 사용한 질의 변환 방법

연어를 사용한 질의 변환 방법을 평가하기 위해서 TREC-8의 adhoc retrieval task에서 사용된 영어 질의를 한국어 질의로 변환해서 사용했다[5]. 영어 대상 문서는 [표 4]와 같다.

[표 4] 질의 변환 실험에서의 검색 대상 말뭉치

문서 이름	문서 크기
Financial Times	210,158건
Federal Register(1994)	55,630건
Foreign Broadcast Information Service	130,471건
The LA Times	131,896건
총 문서 수	528,155건

검색을 위해서 사용된 질의는 TREC-8의 adhoc retrieval task의 50개 질의 중에서 교차 언어 정보 검색 평가에 적절한 35개의 질의만을 사용했다[6]. 질의 중에서 title과 description을 선택하고 narration은 사용하지 않았다. 실험은 title만을 사용하는 짧은 질의와 title과 description을 함께 사용하는 긴 질의의 두 가지를 실험했다. 그리고 연어 정보의 유용성을 평가하기 위해서 연어를 사용하는 경우와 사용하지 않는 경우에 대해서 평가하였다. 실험에 사용된 검색 방법들은 [표 5]와 같다.

실험에 사용한 검색 엔진은 [김상범2000]이 개발한 고려대학교

[표 5] 질의 변환 여부와 연어 사용에 여부에 따른 검색 방법 분류

검색 방법		설명
단일언어		영어 질의에 대해서 영어 문서 검색
교차언어		한국어 질의에 대해서 영어 문서 검색
연어 사용 여부	교차언어(1)	연어를 사용하지 않음
	교차언어(2)	한국어 연어만 사용
	교차언어(3)	영어 연어만 사용
	교차언어(4)	한국어 연어와 영어 언어 모두 사용

5) TREC-8의 adhoc retrieval task는 영어 질의에 대해서 영어 문서를 검색하는 task이다.

6) 적절하지 않은 질의로는 질의어의 수가 하나(예: "suicide; 자살")이어서 변환 중의성을 해소할 수 없는 경우와 한국어로 번역했을 때, 사람조차 제대로된 대역 질의를 생성할 수 없는 질의이다.

의 정보검색 엔진 KUIR을 사용하였으며, 실험 결과의 평가를 위해서는 평균 정확률(average precision) 평가 척도를 사용했다. average precision은 재현율을 기준으로 정확률을 평균한 값으로, 재현율을 0%에서 100%까지 11개의 단위로 세분하고, 각 범위에서의 정확률을 평균한 값이다. 검색 결과는 상위 1000개까지의 문서를 사용했다.

실험 결과는 [표 6]과 같다. Avg. P는 average precision을, % of mono. IR은 단일 언어 검색(monolingual IR)에 대한 교차언어 검색의 성능 비율을 나타낸다.

[표 6] 질의 변환 방법에 따른 실험 결과

	title		title+ description	
	Avg. P	% of mono. IR	Avg. P.	% of mono. IR
단일언어	0.1929	100%	0.1906	100%
교차언어(1)	0.1667	86.42%	0.1396	73.24%
교차언어(2)	0.1683	87.25%	0.1494	78.38%
교차언어(3)	0.1667	86.42%	0.1402	73.56%
교차언어(4)	0.1683	87.25%	0.1503	78.86%

교차언어(1)의 결과와 교차언어(2)의 결과를 비교하면, 한국어 연어의 유용성을 볼 수 있다. 한국어 연어의 경우, title만 사용하는 경우와 title과 description을 함께 사용하는 경우에 모두 성능의 향상이 있음을 볼 수 있다. 교차언어(1)의 결과와 교차언어(3)의 결과를 비교하면, 영어 연어의 유용성을 볼 수 있는데, 영어 연어의 경우 title에 대해서는 성능의 향상이 없다. 이는 title에 대해서 교차언어(1)에서 생성된 영어 질의와 교차언어(3)에서 생성된 영어 질의가 동일하기 때문이다. 그러나 title과 description을 함께 사용하는 경우에는 교차언어(3)이 교차언어(1)에 비해서 약간 더 높은 결과를 나타내는 것을 볼 수 있다. 이와 같이 영어 연어의 사용이 성능 향상이 없거나 비교적 적은 이유는, 영어에서 연어로 나타나는 단어들은 항상 큰 상호 정보 값을 갖기 때문이라고 할 수 있다.

마지막으로 교차언어(4)와 교차언어(2), 교차언어(3)을 비교함으로써 한국어 연어와 영어 언어를 함께 사용하는 경우의 성능향상을 살펴볼 수 있다. Title을 질의로 사용하는 경우에는 성능 향상을 볼 수 없지만, title과 description을 함께 사용하는 경우에는 성능 향상이 있음을 볼 수 있다. 그러므로, 연어 정보를 사용하는 경우에 성능의 향상이 있음을 기대할 수 있다.

4.2. 이중언어 시소러스를 사용한 전역적 질의 확장

이중언어 시소러스를 사용한 전역적 질의 확장의 성능을 평가하기 위한 평가 집합은, 4.1절의 실험과 유사하게 TREC-8의 adhoc retrieval task의 문서들을 사용했으나, 신속한 실험을 위하여 평가 집합을 약간 축소하여 사용하였다 [표 7].

[표 7] 질의 확장 실험에서의 검색 대상 말뭉치

문서 이름	문서 크기
Financial Times	210,158건
The LA Times	131,896건
총 문서 수	342,054건

또한 사용한 질의는, 평가 집합의 축소로 인해 단일언어(영어) 검색에서 매우 성능이 낮아진 질의를 제외한 33개를 사용하였다. 또한 4.1의 실험과 같이, title만 사용한 질의와, title과 description을 함께 사용한 질의로 나누어 실험하였다. 실험에 사용한 알고리듬은 3.2절에서 설명한 ex_m9, ex_c_m9_w3과 같다.

실험 결과의 평가를 위해서는 4.1에서 사용한 평균 정확률(average precision)에 추가하여, [Mandala1998]과 같이 재현율(recall)과 정확률(precision)척도를 사용하였다. 검색 결과는 상위 1000개까지의 문서를 사용했다. 결과는 [표 8], [표 9], [표 10]에 나타냈다(다음 페이지).

실험 결과, title 만을 사용한 짧은 질의에서는, 대역 질의의 단어와 동일한 synset의 단어를 최대한 확장한 ex_m9 방법이 성능이 가장 높았으며, title과 description을 함께 사용한 긴 질의에 대해서는 동일 synset과 자식 synset을 함께 사용하되, 상호 정보를 구하는 윈도우 사이즈를 3으로 감소시킨 ex_c_m9_w3 방법이 성능이 높았다. 성능 향상의 정도는 적합성 피드백을 사용한 연구 보다는 상당히 낮은 것으로 보인다.

긴 질의보다 짧은 질의에서 질의 확장의 효과가 더 높았는데, 이는 단일언어 검색에서 워드넷을 사용한 기존 연구([Voorhees1993] [Voorhees1998])의 결과와, 교차 언어 검색에서 적합성 피드백을 사용한 기존 연구([McNamee2002])의 결과와 일치하는 것이다. 또한 단일언어 검색에서 워드넷을 사용한 질의 확장의 실험인 [Mandala1998]에서는 재현율은 증가하고, 정확률은 감소했었으나, 이 실험에서는 실험 방법에 따라 정확률과 평균 정확률이 감소한 경우(title+ description에서 ex_m9를 사용한 경우)도 있지만, 다른 실험에서는 약간 향상되었고, 재현율의 증가 정도는 평균 정확률의 증가 정도보다는 훨씬 큰 결과를 보였다. [Mandala1998]과 같이 정확률이 감소하는 경우가 적은 이유는, 단일언어 검색에서는 다의어(polysemous word) 때문에 중의성이 발생하지만, 교차언어 검색에서는 상호정보를 사용한 질의변환 과정 중에서 중의성이 해소되는 효과가 있기 때문으로 보인다.

5. 결론 및 향후 연구

본 논문에서는 한영 교차언어 정보검색의 질의 변환과 질의 확장에 관해서 살펴보았다. 질의 변환을 통한 한영 교차 언어 검색 엔진 구현을 위해서 한영사전과 한국어 연어 사전, 영어 연어 사전을 함께 사용했다. 한국어와 영어의 각 질의에 사용된 단어는, 하나의 단어가 하나 이상의 단어와 대응되는 경우가 자주 발생할 수 있는데, 이를 처리하기 위해서 한국어 연어 사전과 영어 연어 사전을 사용했다. 실험 결과, 한국어 연어를 사용한 경우가 영어 연어를 사용한 경우보다 성능 향상 폭이 더 컸다.

또한 검색 성능을 향상시키기 위하여 질의를 확장하였는데, 기존 연구에서 많이 사용되었던 적합성 피드백에 의한 지역적 질의 확장 대신, 한영 이중언어 시소러스를 사용한 전역적 질의 확장을 수행하였다. 실험 결과, 긴 질의보다는 짧은 질의에서의 성능 향상 폭이 컸으며, 정확률의 증가율보다는 재현율의 증가율이 더 높았다.

향후 연구로는, 한국어와 영어의 적합성 피드백을 구현하여, 지역적 질의 확장의 변환 전 확장과 변환 후 확장의 결과와 비교하며, 변환 후 확장에서 지역적 질의 확장과 전역적 질의 확장과의 결합으로 성능이 향상될 수 있는지를 실험할 계획이다. 또한 한영 이중언어 시소러스를 질의 확장뿐만 아니라, 변환 중의성을 해소하는 방법에 적용하는 방법을 연구할 계획이다. 또한 한영 교차언어 정보 검색에서 성능을 크게 향상시키는 것으로 보고된, 한영 자동 음차 복원을 구현하여, 질의에 나타난 외래어에 대한 대역어를 생성하는 방법을 추가하고자 한다.

[표 8] 질의 확장 방법에 따른 실험 결과 - average prescion

	title	title+ description
질의확장 없음	0.2107	0.1806
ex_m9	0.2175	0.1736
ex_c_m9_w3	0.2043	0.1882

[표 9] 질의 확장 방법에 따른 실험 결과 - title 질의

검색한 문서수	재현률			정확률		
	확장 없음	ex_m9	ex_c_m9_w3	확장 없음	ex_m9	ex_c_m9_w3
5	0.1189	0.1174	0.0874	0.3879	0.4000	0.3758
10	0.1544	0.1589	0.1245	0.3273	0.3515	0.3242
15	0.1721	0.1789	0.1435	0.2727	0.2970	0.2747
20	0.1899	0.1961	0.1606	0.2485	0.2667	0.2485
30	0.2190	0.2337	0.1906	0.2152	0.2303	0.2162
100	0.3311	0.3528	0.3406	0.1291	0.1403	0.1303
200	0.4191	0.4485	0.4259	0.0932	0.1030	0.0932
500	0.5105	0.5468	0.5147	0.0485	0.0539	0.0490
1000	0.5817	0.6178	0.5848	0.0288	0.0319	0.0292

[표 10] 질의 확장 방법에 따른 실험 결과 - title+description 질의

검색한 문서수	재현률			정확률		
	확장 없음	ex_m9	ex_c_m9_w3	확장 없음	ex_m9	ex_c_m9_w3
5	0.1043	0.0977	0.1059	0.3576	0.3273	0.3273
10	0.1450	0.1501	0.1543	0.2818	0.2788	0.2879
15	0.1756	0.1704	0.1785	0.2384	0.2404	0.2525
20	0.2010	0.1895	0.2058	0.2227	0.2182	0.2379
30	0.2253	0.2146	0.2260	0.1879	0.1848	0.1899
100	0.3211	0.3114	0.3272	0.1018	0.1033	0.1042
200	0.4025	0.3799	0.4187	0.0685	0.0709	0.0724
500	0.5063	0.4692	0.5103	0.0387	0.0380	0.0394
1000	0.5631	0.5381	0.5690	0.0232	0.0229	0.0236

6. 참고 문헌

[강인수1997] 강인수, 이종혁, 이근배, "교차언어 문서검색에서 질의어의 중의성 해소 방법", 제9회 한글 및 한국어정보처리 학술대회, 1997.

[김진동1997] 김진동, 임희석, 임해창, "Twoply HMM : 한국어의 특성을 고려한 형태소 단위의 품사 태깅 모델", 한국정보과학회 논문지(B), 제24권, 제12호, pp.1502-1512, 1997.

[천정훈1999] 천정훈, 최기선, "교차언어 문서검색에서 다국어 온톨로지에 기반한 한영 질의어 변환", 제 11회 한글 및 한국어 정보처리 학술대회, 1999.

[장명길1999] 장명길, 맹성현, 박세영, "한영 교차언어 정보검색에서 상호정보를 이용한 질의 변환 모호성 해소 및 가중치 부여 방법", 제11회 한글 및 한국어 정보처리 학술대회, 1999.

[장명길2002] 장명길, "교차언어 정보검색에서 통계정보를 이용한 사전기반 질의변환에 관한 연구", 충남대학교 대학원 컴퓨터과학과 정보과학전공 박사학위 논문, 2002

[박지연2001] 박지연, 정영미, "질의확장에 의한 단락검색의 성능 향상에 관한 연구", 한국정보관리학회 학술대회 논문집, 2001

[김상범2000] 김상범, 한경수, 이도길, 임재수, 고명숙, 임해창, "고려대학교 정보검색엔진 KUIR의 구조 및 특징", 제5회 한국 과학기술 정보인프라 워크샵 학술발표 논문집, pp.164-174, 2000.

[Ballesteros1996] Ballesteros, L. and W.B. Croft, "Dictionary-based methods for cross-lingual information retrieval", In Proceedings of the 7th International DEXA Conference on Database and Expert Systems Applications, 1996.

[Brown1997] Brown, Ralf D. "Corpus-based query translation for cross-lingual information retrieval", SIGIR-97 workshop on Cross-Lingual Information Retrieval 1997.

[Davis1995] Davis, M. and T. Dunning, "Query translation using evolutionary programming for multilingual information retrieval", In 4th Annual Conference on Evolutionary Programming, 1995.

[Dumais1997] Dumais S.T., T.A. Letsche, M.L. Littman, and Landauer T.K. "Automatic cross-language retrieval using latent semantic indexing", AAAI symposium on Cross-Language Text and Speech Retrieval. American Association for Artificial Intelligence, March 1997.

[Eichmann1998] Eichmann, David, Miguel E. Ruiz and Padmini Srinivasan, "Cross-Language Information Retrieval with the UMLS Metathesaurus", SIGIR '98, 1998.

[Gilarranz1997] Gilarranz, Julio, Julio Gonzalo and Felisa Verdejo, "An Approach to Conceptual Text Retrieval Using the EuroWordNet Multilingual Semantic Database", In AAAI Spring Symposium on Cross-Language Text and Speech Retrieval, March 1997.

[Hull1996] Hull, David A. and Gregory Grefenstette, "Querying across languages: a dictionary-based approach to multilingual information retrieval", In the Proceedings of the 19th ACM SIGIR Conference, 1996.

[McNamee2002] P. MaNamee and J. Mayfield, "Comparing Cross-Language Query Expansion Techniques by Degrading Translation Resources", In the Procedings of the 25th International ACM SIGIR conference on Research and Development in Information Retrieval(SIGIR-2002), pp. 159-166

[Baeza-Yates1999] R. Baeza-Yates and B. Rebeiro-Neto, "Ch. 5 Query Operations" in "Modern Information Retrival", pp. 117-140, 1999

[Voorhees1993] E.M. Voorhees, "Using WordNet to disambiguate word senses for text retrieval", In the Procedings of the 16th International ACM SIGIR conference on Research and Development in Information Retrieval(SIGIR-1993), pp. 171-180

[Voorhees1998] E.M. Voorhees and D.Harman, "Overview of the fifth text retrieval conference", In the Procedings of the Fifth Text REtrieval Conference(TREC-5), pp. 1-28. NIST Special Publication 500-238

[Mandala1998] R. Mandala, T. Takenobu and T. Hozumi, "The Use of WordNet in Information Retrieval", In the Proceedings of the 16th COLING-ACL 1998 Workshop - Usage of WordNet in Natural Language Processing Systems, pp. 31-37