한-영 교차언어 정보검색에서 상호정보를 이용한 질의 변환 모호성 해소 및 가중치 부여 방법

장 명길*, 맹 성현**, 박 세영*
*한국전자통신연구원 지식정보연구부 **충남대학교 컴퓨터과학과

A Disambiguation and Weighting Method using Mutual Information for Query Translation in Korean-to-English Cross-Language IR

Myung-Gil Jang*, Sung Hyon Myaeng**, Se Young Park*
*Dept. of Knowledge Information Technology, Electronics and Telecommunications Research Institute(ETRI)
**Dept. of Computer Science, Chungnam National University

mgjang@etri.re.kr, shmyaeng@cs.chungnam.ac.kr, sypark@computer.etri.re.kr

요 약

교차언어 문서검색에서는 단일언어 문서 상황을 만들기 위하여 질의나 문서를 다른 언어로 변환하게 되는데, 일반적으로 간단하면서도 실용적인 질의 변환의 방법을 주로 사용하고 있다. 하지만 단순한 대역 사전을 사용한 질의 변환의 경우에 변환 모호성 때문에 40% 이상의 검색 효과의 감소를 가져온다. 본 논문에서는 이러한 변환 모호성을 해결하기 위하여 대역 코퍼스로부터 추출한 상호 정보를 이용하는 단순하지만 효과적인 사전 기반 질의 변환 방법을 제안한다. 본 연구에서는 변환 모호성으로 발생한 다수의 후보들에서 가장 좋은 후보를 선택하는 모호성 해소 뿐 아니라 후보 단어들에 적절히 가중치를 부여하는 방법을 사용한다. 본 질의 변환 방법은 단순히 가장 큰 상호 정보의 단어를 선택하여 모호성 해소만을 적용하는 방법과 Krushall의 최소 스패닝 트리 구성과 유사한 방법으로 상호 정보가 큰 순서대로 간선들을 연결하여 모호성 해소와 가중치 부여를 적용하는 방법들과 질의 변환의 검색 효과를 비교한다. 본 질의 변환 방법은 TREC-6 교차언어 문서검색 환경의 실험에서 단일 언어 문서검색의 경우의 85%, 수작업 모호성 해소의 경우의 96%에 도달하는 성능을 얻었다.

1. 서 론

교차언어 문서검색은 사용자들에게 자신의 언어로 표현된 질의어를 사용하여 다양한 언어로 쓰여진 문서를 검색할 수 있게 한다. 교차언어 문서검색을 위하여 일반적으로 질의어 혹은 문서의 언어적 차이를 극복하기 위하여 그들간의 언어변환을 수행하는데, 언어변환을 수행하지 않는 교차언어 LSI(Latent Semantic Indexing)(Dumais et al., 1997)와 같은 방법도 도입할 수 있다.

Oard & Hackett (1997)의 문서 변환에서는 문서들을 질의 언어로 변환하는데 고품질 기계번역 시스템을 활용하고 있지만, 이 방법은 현재의 기술 수준으로 대규모로 적용되기에는 비실용적이다(Carbonell et al., 1997). 질의를 문서의 언어로 변환하는 질의 변환은 문서 변환과 비교하여 훨씬 간단하고 더욱 경제적이기 때문에 인기있는 방법으로 부각되어 왔다. 질의 변환의 세 가지 접근 방법에는 사전 기반 접근 방법, 시소러스 기반 접근 방법, 그리고 코퍼스 기반 접근 방법이 있는데, 이들 방법은 하나 혹은 둘 이상이 함께 적용될 수 있다. McCarley(1999)는 질의와 문

서 변환 모두를 포함하는 복합 방법을 통하여 가장 좋은 검색 효과를 얻음을 보였다.

본 연구에서는 코퍼스나 시소러스 기반 질의 변환 방법보다는 단순성과 실용성을 초점을 두어 사전 기반 질의 변환 방법을 채택하였다. 하지만 사전의 단순한 사용에 의한 질의 변환은 검색 성능이 단일언어 검색의 단지 40%-60% 정도에 불과한 것으로 나타났다(Ballesteros & Croft, 1997). 이것은 더 나은 성능 향상을 위하여 다른 추가적인 자원들이 사용될 필요가 있다는 것을 의미한다. 질의 변환 방법을 사용한 교차언어 문서검색에서는 해결해야 하는 세 가지 문제가 있다(Grefenstette, 1998). 첫번째 문제는 한 언어로 쓰여진 질의를 다른 언어로의 변환을 수행하는 방법에 관한 것이고, 두 번째 문제는 변환된 후보 단어들 중에서 어떤 불필요한 단어들을 제거하는 가를 결정하는 것이다. 이는 단어 모호성 해소의 문제에 해당된다. 세 번째 문제는 하나 이상의 후보 단어들의 상대적인 중요성도에 따라 적절히 가중치를 주는 방법을 결정하는 것이다.

본 논문은 마지막 두 가지 문제들에 초점을 맞추어 대역 문서 집합으로부터 추출한 상호 정보(Church & Hanks, 1990) 통계치를 이용하여 변환 모호성 해소과 가중치 부여의 문제에 대처하는 상대적으로 단순하지만 효과적인 방법을 제안한다. 다음 2 절에서는 질의 변환에서 발생하는 모호성의 정도를 알아보고 3 절에서는 질의 변환에 사용되는 상호 정보와 본 한국어-영어 교차언어 문서검색의 한국어-영어 질의 변환 과정을 개략적으로 살펴본다. 4 절에서는 질의 변환에서 상호 정보를 이용하여 단지 가장 좋은 후보를 선택하는 모호성 해소 뿐 아니라 대역 언어로 된 질의어 텀들에 가중치를 부여하는 모호성 해소 및 가중치 부여 방법들을 살펴본다. 5 절에서는 질의 변환의 검색 효과를 TREC 6 교차언어 문서검색 환경에서 실험하고 마지막으로 결론을 맺는다.

2. 변환 모호성의 분석

질의 변환을 수행하는 가장 쉬운 방법은 대역 사전을 사용하는 것이지만, 대역 사전에서는 일 대 다 매핑 때문에 변환 모호성의 문제를 격게 된다.

예를 들어, "자동차 공기 오염" 이라는 세 단어로 구성되는 한국어 질의에서 한국어-영어 대역 사전이 직접 사용될 때 각 단어는 다수의 영어 단어들로 변환될 수 있다. 질의의 첫번 째 단어 '자동차'는 'motocar', 'automobile', 'car'와 같은 의미적으로는 비슷하지만 다른 영어 단어들로 변환된다. 두번 째 동음이의어 단어 '공기'는 다른 의미를 가지는 영어 단어들인 'air', 'atmosphere', 'empty vessel', 'bowl'들로 변환 된다. 그리고 마지막 단어 '오염'은 두 개의 영어 단어인 'pollution'과 'contamination'으로 변환된다. 다수의 후보 단어들을 질의로 사용하는 것은 단일언어 문서검색에서 재현율을 증가 시키는 데 유용할 수 있는 반면에, 단어들의 의미 모호성 해소에서 실패하는 경우에는 검색 효과에 나쁜 영향을 미칠 수 있다고 이전의 연구들은 지적한다. 예를 들어, 'empty vessel'과 같은 구절은 질의의 의미를 전체적으로 변경시킬 수도 있으며, 심지어 'pollution'과 동의어인 'contamination' 같은 단어는 의미에서의 약간의 차이로 인하여 관련이 없는 문서들을 검색하게 될지도 모른다.

표 1. 모호성의 정도

	Words			Word Pairs		
	# in S. Lan.	# in T. Lang.	Average Ambiguity	# in S. Lan.	# in T. Lang.	Average Ambiguity
Title	48	158	3.29	24	212	8.83
Short	112	447	3.99	91	1459	16.03
Long	462	1835	3.97	423	6196	14.65

Jang *et al.*(1999)에서 처럼, 표 1 은 한국어-영어 대역 사전이 형태소 분석과 태깅 후에 단순히 사용되는 경우에 질의어 변환에서 모호성이 어느 정도 나타나는 지를 보여준다. *세 개의 행 title, short, long* 은 TREC 문서 집합에서 질의를 구성하는 세 가지 다른 방법들을 가리킨다. 표의 왼쪽은 각 질의에 대하여 한국어 단어에 대한 대역 영어 단어들의 평균 수를 나타내고 반면에 오른쪽은 한국어 단어 쌍으로부터 만들어질 수 있는 대역 영어 단어 쌍들의 평균 수를 나타낸다. 이것은 질의 변환의 모호성 해소 과정에서 평균적으로 9 개 이상의 가능한 쌍들로부터 하나를 선택해야 한다는 것을 의미한다.

변환 모호성에 대한 다른 통계적 분석이 중국어-영어 쌍에 대하여 최근에 이루어졌다(Chen et al., 1999). 중국어 시소러스와 영어 Roget 다국어 시소러스에 따라 중국어 단어와 영어 단어는 각각 1.397 과 1.687 개의 의미를 가진다. 상위 1000 개 빈도 단어만을 고려할 때는 중국어 1.504 개, 영어 3.527 개의 의미를 가진다. Jang *et al.*(1999)이 대역 사전에서의 변환 모호성의 정도를 나타내고 있는 반면에 이들은 단일언어 사전이나 시소러스에 대한 경우이다.

3. 상호 정보를 사용한 질의 변환

본 연구에서는 교차언어 문서검색을 위하여 병렬 코퍼스, 비교 코퍼스, 혹은 다국어 시소러스 등의 구축하기 힘든 자원들에 의존하지 않고 사전에 기반한 실용적인 질의 변환 방법을 사용한다. 마찬가지로 교차언어 문서검색 환경에서 항상 얻을 수 있는 대역 언어로 된 문서 집합에서 추출한 상호 정보를 이용하여 질의 변환의 변환 모호성을 해결하는 접근 방법을 채택하였다.

질의 변환의 모호성 문제에 대처하기 위하여 기본적으로 사전 기반의 질의 변환 방법을 채택하면서 다른 자원들을 함께 이용하는 관련 연구들이 있다. Yamabana et al.(1996)가 제안한 DMAX(Double MAXimize) 방법은 대역 사전으로 변환된 가능한 후보들로부터 모호성을 해소하기 위하여 원시 단어와 목적 단어 사이의 공기 빈도를 동시에 최대로 하는 단어 쌍을 선택하는 통계적인 단어 선택 방법을 제시하였다. Twenty-One 시스템(Kraaij & Hiemstra 1997)에서 구현된 질의 변환 방법은 네덜란드어-(독어, 불어, 영어, 스페인어) 대역 사전 뿐 아니라 표준 NLP 도구들을 사용하는데, 대역 문서 코퍼스로부터 추출된 후보 NP들에 근거하여 변환 NP들의 모호성을 해소한다. Hull(1997)은 weighted Boolean 모델을 사용하는 질의 변환 방법을 통하여 질의 단어들의 Boolean 조합을 형성하면서 단어들에 대한 가중치를 계산한다.

본 질의 변환에서 사용하는 상호 정보는 단어 공기 통계치에 근거하여 계산되는데, 단어들 사이의 상관성을 계산하는 측정 장치로 사용된다. 상호 정보 $MI(x,y)$는 다음 공식으로 정의된다(Church & Hanks, 1990).

$$MI(x,y)=\log_2\frac{p(x,y)}{p(x)p(y)}=\log_2\frac{N f_w(x,y)}{f(x)f(y)} \quad (1)$$

여기서 x 와 y 는 w 단어들의 범위 내에서 함께 나타나는 단어들이다.

확률 $p(x)$와 $p(y)$는 코퍼스에서 x와 y의 빈도 $f(x)$와 $f(y)$를 계산하고 이를 코퍼스 크기 N으로 각각을 정규화함으로써 추정된다. 확률 $p(x,y)$는 w 단어들의 범위 내에서 x 다음에 y가 순서대로 나타나는 빈도 $f_w(x,y)$를 계산하고 이를 코퍼스 크기 N으로 정규화함으로써 추정된다. 본 연구의 질의 변환의 응용에서 공기 빈도 $f_w(x,y)$는 bread and butter 와 같은 숙어의 고정된 표현 뿐 아니라 질의의 의미적 관계를 허용하기 위하여 공기 문맥의 범위를 같은 문장내에서의 여섯 단어 즉 $w=6$으로 정하였다.

앞서의 공식으로 구해진 상호 정보는 질의 변환에서 한국어 질의 단어가 하나 이상의 영어 단어들로 변환될 때 가장 가능성 있는 변환들을 선택하는 모호성 해소와 단어들의 중요도를 나타내는 가중치 부여의 기준으로 사용된다. 상호 정보 값의 사용은 두 단어들이 같은 질의에서 공기할 때 그들이 문서에서도 같은 유사성으로 공기하여 나타난다는 가정에 근거한다. 반대로 같은 유사성으로 공기하지 않는 두 단어들은 같은 질의에서 나타나지 않을 것이다. 어떤 의미에서 상호 정보는 단어들 사이의 의미적 관계의 정도를 나타낼 수 있다고 추측할 수 있다. 즉 $MI(x,y)$가 클 때 단어 관련성은 강해지고 변환 모호성 해소를 위한 신뢰할 수 있는 결과들을 산출한다. 그러나 $MI(x,y)$가 0 보다 작을 때 우리는 단어 x와 단어 y가 상보적 분포에 있다는 것을 예측한다. 본 연구의 질의 변환에 사용하는 상호 정보 값들은 116,759,540 단어들을 포함하는 1988 ~ 1990 AP 뉴스로 구성된 영어 텍스트 코퍼스로부터 추출되었다.

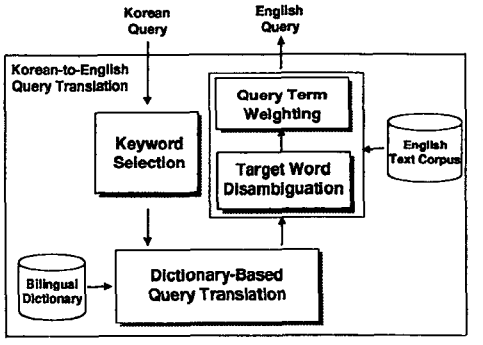


그림 1. 한국어-영어 질의어 변환을 위한 네 단계

본 논문의 한국어-영어 질의어 변환 방법은 네 단계로 수행된다. 즉 키워드 추출, 사전 기반 질의어 변환, 대역 단어 의미 모호성 해소, 그리고 질의 텀 가중치 부여이다.

사전, 시소러스, 코퍼라와 같은 언어 자원들을 단독으로 사용하여 질의 변환을 수행하는 경우는 고품질의 목적 질의를 생성하는 것이 완벽하지 않다. 본 논문에서는 두 번째 단계에서 대역 사전을 사용하고 세 번째와 네 번째 단계를 위해서는 목적 언어의 코퍼스를 사용하도록 하였

다. 이것은 본 논문의 질의 변환이 실용적인 접근 방법으로 활용하기 위하여 대부분의 언어 쌍에서 상대적으로 쉽게 구할 수 있다는 데 초점을 맞춘 것이다. 그림 1은 한국어-영어 질의 변환의 네 단계를 보여준다.

첫 번째 단계에서는 질의 변환 과정에 입력되는 유사 자연어(quasi-natural language) 질의로부터 한국어 키워드들이 추출된다. 이러한 키워드 추출은 한국어 형태소 분석기와 통계적 품사 태거(Shin *et al.*, 1996)에 의해 수행된다. 여기서 태거의 역할은 형태소 분석에 의해 생성된 다수의 후보 열들로부터 정확한 형태소 열을 선택하는 데 도움을 주는 것이다. 형태소 분석과 태거를 도입하는 이러한 과정은 한국어가 교착언어이기 때문에 질의 문들로부터 합법적인 질의어 단어들을 선택하는 데 중요한 역할을 한다. 태거 없이는 형태소 분석기로부터 생성된 모든 추가적인 후보 키워드들이 이후의 변환 과정에 들어가게 되고 그 자체는 대역 사전에서의 일대다 매핑 때문에 더욱 많은 부가적인 단어들을 생성하게 될 것이다.

두 번째 단계에서는 단어 대 단어 변환과 구 단위 변환의 적용에 의한 대역 사전 참조에 근거하여 실제적인 질의 변환을 수행한다. 한국어 질의에서 구들의 정확한 인식을 위하여 어휘적 관계들을 인식하고 Smadja(1993)에서와 같은 텍스트 코퍼스에서의 단어 쌍에 대한 통계적 정보를 이용하는 것이 도움이 될 것이다. 한국어 질의의 변환에서 대역 사전은 근본적으로 전문 용어와 외래어 같은 단어들이 결여되어 있을 수 있기 때문에 이들 미등록 단어를 인식하는 것과 외래어의 경우에는 이들을 영어 스트링으로 음역하는 것이 중요하다(Jeong *et al.*, 1997).

단어 모호성 해소 단계에서는 대역 사전 참조를 통하여 얻어진 후보 단어들을 걸러 낸다. 이때 목적 언어의 코퍼스로부터 추출된 공기 정보를 사용한 대역 단어 모호성 해소 기법을 활용한다. 4 절에서 자세히 설명한다.

마지막 단계에서는 최종 대역 질의를 산출하기 위하여 남아있는 후보 단어들에 가중치 부여 기법을 적용한다. 질의 텀 가중치 부여 기법은 기본적으로 변환된 단어들 사이의 연관성의 정도를 반영한다. 이것은 텍스트 코퍼스로부터 얻은 상호 정보를 이용하는 또 다른 영역이다. 질의 변환의 네 단계들로부터 생성된 결과는 단일언어 문서 검색의 벡터 공간 검색 모델에서 사용될 질의 텀들의 집합이다.

4. 모호성 해소와 가중치 부여 방법

변환 모호성을 해결하는 본 질의 변환 방법은 질의 단어가 얼마나 폭넓게 나타나는가(broad) 혹은 부분적으로 나타나는가(narrow)에 관계없이 상호 정보라는 단일 개념을 사용한다. 본 논문에서 사용하는 상호 정보는 모호성 해소와 가중치 부여를 위하여 구절 인식에 있어서도 가장 중요한 효과를 얻을 수 있는 바로 이웃한 단어의 단어 연관성을 반영하는 값으로 사용된다. 하지만 추가적인 계산 부담에도 불구하고 이웃한 단어를 포함한 모든 단어들을 고려하여 모호성을 해소하는 방법(Fung *et al.*, 1999)도 고려할 수 있다. 또한 이러한 단어 연관성을 측정하는 데는 상호 정보 외에 Ballesteros & Croft (1998)에서와 같은 다른 통계적인 측정치도 사용될 수 있다.

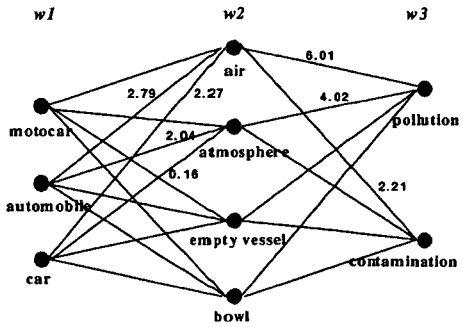


그림 2. MI 값들을 가진 단어 쌍들의 예

2 절의 예에서 한국어 단어들은 변환 모호성 때문에 다수의 영어 단어들을 가질 수 있다. 그림 2 에서 *w1*, *w2*, *w3* 밑에 있는 단어들은 원래 질의 단어들로부터 변환된 것들이고 각 단어 쌍의 선 위에는 그들 단어 쌍의 상호 정보 값들이 계산되어져 있다.

대역 단어 모호성 해소와 가중치 부여 방법에는 여러 가지 알고리즘을 적용할 수 있다. 본 연구에서는 질의 변환을 위하여 세 가지 방법의 알고리즘을 적용하여 질의 변환을 시도해보았다.

4.1 OB(One Best) 방법

OB 방법은 가장 높은 상호 정보 값만을 고려하는 모호성 해소 적용 방법이다. 즉, 각 후보 단어 쌍에서 가장 큰 상호 정보 값을 가지는 단어 쌍만을 먼저 표시하고 각 단어

의 후보 단어들 중에서 가장 큰 상호 정보 값을 가진 단어만 선택하고 나머지는 제외하는 방식으로 질의를 생성하는 단순히 모호성 해소만 적용하는 방법이다. 이 방법은 질의 변환의 모호성 해소만 적용하고 텀 가중치 부여를 적용하지 않는 경우에 어느 정도의 질의 변환 효과가 나타나는 지를 알아보기 위한 기준으로 사용된다.

4.2 MST(Maximum Spanning Tree) 방법

MST 방법은 Krushall의 최소 스패닝 트리 구성과 유사한 방법으로 상호 정보가 큰 순서대로 간선들을 연결하여 최대 스패닝 트리를 만들어 가는 데 중복 연결을 허용한다. 이 방법은 OB 방법과 마찬가지로 먼저 각 후보 단어 쌍에서 가장 큰 상호 정보 값을 가지는 단어 쌍을 표시하고 그 다음으로 각 단어의 후보 단어에 대하여 임계값을 기준으로 모호성 해소와 가중치 부여를 적용하는 방법이다. 만일 상호 정보 최대 값이 하나만 표시된 단어인 경우는 모호성 해소를 적용하고, 두 개의 단어가 표시되어 선택된 경우는 가중치 부여를 적용한다. 이때 상호 정보 값이 둘 다 임계값 보다 크거나 작으면 똑같이 0.5의 가중치를 부여 받게 되지만, 두 단어의 상호 정보 값이 하나는 임계값 보다 크고 다른 하나가 임계값 보다 작으면 작은 임계값을 가진 단어와 큰 임계값을 가진 단어에 대하여 각각 다음과 같은 가중치 Ws와 Wl을 부여한다.

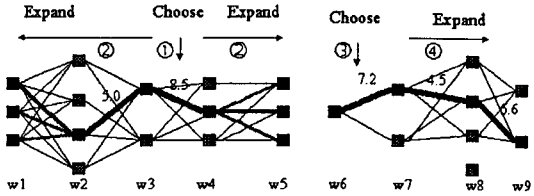
$$W_s = 1/(diff+2),\ W_l = 1 - W_s$$

여기서 *diff*는 큰 상호 정보 값과 작은 상호 정보 값의 차를 의미한다.

4.3 BFE(Best-First & Expand) 방법

BFE 방법은 Jang *et al.*(1999)에서 기술한 바와 같이 가장 큰 상호 정보 값을 가진 단어 쌍을 중심으로 그 단어 쌍의 전 후로 모호성 해소와 가중치 부여를 적용하여 확장해 나가는 방법이다. 이 방법은 상호 정보 값들의 상대적이면서 절대적인 중요도를 가장 높은 상호 정보 값을 가진 단어 쌍을 중심으로 진행한다는 것이 특징이다.

BFE 방법의 알고리즘은 먼저 가장 높은 상호 정보 값들을 가진 쌍들을 찾아 정렬한다. 그 다음 그 단어 쌍을 중심으로 선택된 쌍들과 연결된 쌍들에 대한 상호 정보 값들을 비교함으로써 그 쌍 전 후의 가장 좋은 후보들을 선택하는 모호성 해소를 적용하거나 가중치 부여를 적용한다. 이때 임계값을 기준으로 작은 경우에 이후에 설명하는 가중치 부여를 적용한다. 이러한 과정은 정렬된 임계값 이상의 상호 정보 값을 가진 단어 쌍에 대하여 모두 적용하게 된다. 아래 네트워크는 상호 정보 값을 단어 쌍들로부터 BFE 방법을 적용하는 과정을 순서대로 나타낸 것이다.



질의 텀 가중치 부여가 개념적으로 관련이 없는 텀들을 잘라내는 모호성 해소 과정에 덧붙여 의미있는 과정으로 채택하는 세 가지 이유가 있다. 첫째 모호성 해소 과정의 단어 선택은 정확한 변환을 보장하지 못한다. 이것은 두 개의 연속한 단어들이 실제 많은 문서에서 함께 나타나는 경우에 단지 신뢰할 수 있는 결과를 줄 수 있기 때문이다. 둘째 두 개 이상의 단어가 모두 큰 상호 정보 값을 가지기 때문에 서로 비슷한 정도의 큰 중요도로 나타날 수 있다. 이 경우에 다른 하나에 대한 가중치 고려가 필요하다. 셋째 임계값을 기준으로 그 이상의 단어들은 추가적인 단어들의 질의 확장 효과를 가진 재현율 향상의 수단으로 기여할 수 있다.

질의 텀 가중치 부여의 기본적인 생각은 가장 좋은 후보에게 큰 가중치를 주고 나머지 후보들에는 나머지 가중치를 똑같이 나누어 부여하는 것이다. 즉 가장 좋은 후보에 대한 가중치 W_b는 임계값 보다 크다면 1이고 혹은 그렇지 않으면 다음 식 (2)와 같이 계산된다.

$$W_b = \frac{f(x)}{\theta+1} \times 0.5 + 0.5 \qquad (2)$$

여기서 x와 θ는 각각 상호 정보 값과 임계값이다.

임의의 구간 내에 속한 상호 정보 값들을 가지는 모든 후보들이 결과적으로 같은 가중치를 가지도록 하기 위하여 정수 생성기 $f(x)$는 상호 정보 값보다 더 큰 가장 작은 정수를 준다. 이때 나머지 가중치는 식 (3)과 같이 계산된다.

$$W_r = \frac{1 - W_b}{n-1} \qquad (3)$$

여기서 n는 후보들의 수이고 $W_b + \Sigma W_r = 1$ 이어야 한다.

계산된 상호 정보 값들의 관찰에 근거하여 가장 좋은 후보를 선택하는 데 있어서 임계값은 3.0으로 정하였다. 이 임계값은 순전히 우리가 얻은 자료에 근거하여 결정되고 그것은 다른 코퍼스가 사용될 때 새로운 범위의 상호 정보 값들에 근거하여 다양하게 결정될 수 있다.

예를 들어 그림 2에서 BFE 모호성 해소와 가중치 부여 방법의 적용을 살펴본다. 상호 정보 값이 임계값 이상인 굵은 선으로 표시된 가장 강한 연관성을 표현하는 단어 쌍 <air, pollution>이 먼저 선택되었다. 그때 air를 포함하는 단어 쌍에 대하여 세 가지 상호 정보 값들은 비교되고 <automobile, air> 쌍을 선택하여 결과적으로 <automobile, air, pollution>을 결정한다. 예에서 추가적인 열들이 있다면 같은 과정이 그 네트워크의 나머지에 적용될 것이다. 다음으로 가중치 부여의 적용의 예로 *w1* 과 *w2* 사이의 후보 단어 쌍은 (automobile, air)와 (car, air)이다. 여기서 단어 쌍 (automobile, air)의 가중치는 $W_b = 0.875$ 이기 때문에 단어 'automobile'은 다른 두 단어 'motorcar'과 'car'보다 상대적으로 더 높은 텀 가중치를 가진다. 최종적으로 *w1* 의 변환 단어들에 대하여 텀 가중치를 가진 최적의 영어 질의 집합 <(motocar,0.0625), (automobile, 0.875), (car, 0.0625)>가 생성된다.

5. 실 험

본 질의 변환 방법은 한국어-영어 교차언어 문서검색을 위하여 TREC-6 의 교차언어 문서검색 환경에서 몇 가지 기본적인 실험을 수행하였다. 24 개의 영어 질의들은 3 가지 필드인 titles, descriptions, narratives 로 구성된다. 이들 영어 질의들은 한국어 질의들을 이용한 교차언어 문서검색을 위해서 수작업으로 한국어 질의들로 미리 번역되었다. 교차언어 문서검색과 단어언어 문서검색을 비교하기 위하여 Cornell 대학에서 개발한 Smart 11.0 시스템을 사용하였다.

실험의 목표는 질의 변환에서 모호성 해소와 텀 가중치 부여 방법이 어느 정도 효과적인지를 살펴보는 것이다. 실험에서 사용되는 질의는 질의 길이별로 title 필드를 가진 'title' 질의, description 필드를 가진 'short' 질의, 그리고 세 가지 필드 모두를 가진 'long' 질의로 구분하였다. 11-point 평균 정확도에 의해 측정된 검색 효과는 원래 영어 질의를 사용한 단일언어 문서검색의 성능과 비교하여 측정되었다.

표 2은 네 가지 질의 집합을 사용한 실험 결과를 보여 준다. "Translated Query I" 질의 집합은 단지 키워드 추출과 사전 기반 질의 변환에 의하여 생성되었다. "Translated Query II" 질의는 단어 모호성 해소와 텀 가중치 부여를 적용한 후에 생성된 것이다. 이들 질의는 세 가지 질의 변환 방법인 OB, MST, BFE 방법으로 생성된 질의들이다. 그리고 수작업으로 모호성이 해소된 질의 집합은 Translated Query I 으로부터 가장 좋은 후보 텀을 수작업으로 선택함으로써 생성되었다.

표 2. 실험 결과

Query Sets		Title		Short		Long	
		11pt. P	C/M(%)	11pt. P	C/M(%)	11pt. P	C/M(%)
Original Query		0.3251	-	0.3189	-	0.2821	-
Tran. Query I		0.2290	70.44	0.21443	67.20	0.1587	56.26
Tran. Query II	OB	0.2391	73.55	0.1961	61.49	0.1578	55.94
	MST	0.2652	81.57	0.2422	75.95	0.2094	74.23
	BFE	0.2675	82.28	0.2698	84.60	0.2232	79.12
M.Disam. Query		0.2779	85.48	0.3002	94.14	0.2433	86.25

Translated query set I 의 성능은 세 가지 질의에 대하여 각각 단일언어 문서검색의 약 70%, 67%, 56% 이다. Translated query set II 의 성능은 세 가지 질의 변환 방법이 BFE, MST, OB 의 순으로 좋은 검색 효과를 보이고 있는데, BFE 방법의 경우에 세 가지 질의에 대하여 각각 단일언어 문서검색의 약 82%, 85%, 79% 이다. 이것은 모호성 해소만을 적용한 OB 방법보다 훨씬 효과적이며, 모호성 해소와 가중치 부여를 함께 적용한 MST 방법보다 더욱 효과적인 방법임을 알 수 있다. 반면에 수작업으로 모호성 해소된 질의의 성능은 세 가지 질의에 대하여 각각 단일언어 문서검색의 약 85%, 94%, 86%로 나타났는데 이것은 교차언어 문서검색에서 모호성 해소의 상위 한계로써 취급될 수 있다. 그것이 100%가 아닌 이유는 여러 가지 요인들로 분석될 수 있는데, 1) 원래 영어 질의를 한국어 질의로 수작업 번역하는 데 있어서의 부정확성, 2) 질의 단어들을 생성하는 데 있어서 한국어 형태소 분석기와 태거의 부정확성, 3) 대역 사전을 사용하여 후보 텀들을 생성하는 데 있어서의 부정확성 등이 원인일 수 있다.

Translated Query I 와 Translated Query II 사이의 차이점은 상호 정보 기반 모호성 해소와 단어 가중치 부여 방법들이 검색 효과를 향상시키는 데 효과적이다는 것을 가리킨다. 덧붙여 그 결과는 이들 질의 변환 방법들의 사용이 짧은 질의 보다는 긴 질의들에 대하여 더 효과적임을 보여준다. 이것은 질의가 길수록 더 많은 문맥 정보가 질의 변환의 모호성 해소 및 가중치 부여에 사용될 수 있다는 것을 의미한다.

6. 결 론

단순한 대역 사전을 사용한 질의 변환은 변환 모호성 때문에 검색 효과에 있어서 40% 이상의 감소를 가져온다. 본 연구의 질의 변환 방법은 대역 단어 모호성 해소와 질의 텀 가중치 부여의 문제들을 해결하기 위하여 1988 ~ 1990 AP 뉴스 코퍼스로부터 추출된 상호 정보를 사용한다. 본 논문의 질의 변환 방법은 가장 중요한 문맥 정보인 이웃한 단어간의 상호 정보를 고려한 간단하면서 실용적인 여러 종류의 질의 방법을 제안하였다. TREC-6 교차언어 테스트 문서집합을 사용한 실험에서 BFE 방법에 의한 한-영 교차언어 문서검색에서의 검색 효과는 단일언어 문서검색의 경우의 85% 까지 성능을 향상 시킬 수 있었다. 더 많은 문맥 정보를 가진 긴 질의들에 대하여 가장 큰 비율의 성능 향상을 얻을 수 있었다.

실험 결과가 좋은 반면에 향후 연구해야 할 몇 가지 점들이 있다. 첫째 본 질의 변환 방법의 모호성 해소와 가중치 부여를 위한 상호 정보를 바로 이웃한 단어를 포함한 모든 단어에 대하여 구하여 적용할 필요가 있다. 둘째 모호성 해소와 가중치 부여를 위한 공기정보 통계치로 상호 정보 대신에 다른 측정치를 연구하여 질의 변환의 효과를 높이는데 적용해볼 계획이다.

덧붙여 단일언어 문서검색에서 효과적인 것으로 알려진 의사 적합성 피드백(pseudo relevance feedback) 방법을 사용하여 질의 확장을 통한 교차언어 검색 효과를 향상시킬 계획이다. 임의의 임계치 이상의 상위에 검색된 문서들에 있는 텀들은 문서의 전체는 아니라도 일부가 원래 질의에 관련되거나 혹은 사용자의 정보 필요를 적절히 표현하는 데 유용할 수 있다는 가정에서 상위 문서의 텀들을 원래 질의에 다시 사용할 수 있다. 또한 앞으로 피봇 방식의 질의 변환을 시도하여 세 개 이상의 질의 변환을 통한 교차언어 정보검색 방법을 연구할 계획이다.

참고 문헌

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

[Ballesteros & Croft 98] Lisa Ballesteros and W. Bruce Croft, "Resolving Ambiguity for Cross-language Retrieval", SIGIR' 98, 1998.

[Carbonell *et al.*. 98] J.G. Carbonell, Y. Yang, R.E. Frederking, R.D. Brown, Yibing Geng and Danny Lee, "Translingual Information Retrieval: A Comparative Evaluation". In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*, 1998.

[Chen *et al.* 99] Hsin-His Chen, Guo-Wei Bian and Wen-Cheng Lin, "Resolving Translation Ambiguity and Target Polysemy in Cross-Language Information Retrieval". In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, College Park, MD, 1999.

[Church & Hanks 90] Kenneth W. Church and Patrick Hanks, "Word Association Norms, Mutual Information, and Lexicography", Computational Linguistics, Vol. 16, No. 1, pp. 22-29, 1990.

[Fung *et al.* 99] Pascale Fung, Xiaohu Liu and Chi Shun Cheung (1999). "Mixed Language Query Disambiguation". In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, College Park, MD, 1999.

[Grefenstette 98] Gregory Grefenstette, *Cross-Language Information Retrieval*, Kluwer Academic Publishers.

[Hull 96] David Hull, "A Weighted Boolean Model for Cross-Language Text Retrieval". In *Proceedings of the 19th Annual ACM SIGIR Conference on Information Retrieval*, Zurich, Switzland, 1996.

[Kraaij & Hiemstra 97] Wessel Kraaij and Djoerd Hiemstra, "Cross Language Retrieval with the Twenty-One System". In *Proceedings of the Sixth Text Retrieval Conference (TREC-6)*, NIST, 1997.

[Jang *et al.* 99] Myung-Gil Jang, Sung Hyon Myaeng and Se Young Park, "Using Mutual Information to Resolve Query Translation Ambiguities and Query Term Weighting". In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, College Park, MD,

1999.

[Jeong *et al.* 97] Jeong, K. S., Kwon,Y. H. and Myaeng, S. H., "Construction of Equivalence Classes through Automatic Extraction and Identification of Foreign Words", In *Proceedings of NLPRS'97*, Phuket, Tailand.

[McCarley 99] J. Scott McCarley, "Should we Translate the Documents or the Queries in Cross-Language Information Retrieval?". In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, College Park, MD, 1999.

[Oard & Hackett 97)] 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.

[Shin *et al.* 96] Joong-Ho Shin, Young-Soek Han, Key-Sun Choi, "A HMM Part of Speech Tagger for Korean with Word Phrasal Relations, In *Proceedings of Recent Advances in Natural Language Processing.*

[Samdja 93] Frank Samdja, *"Retrieval Collection from Text: Xtract"*, *Computational Linguistics*, Vol. 19, No. 1, pp.143-177.

[Yamabana 96] Kiyoshi Yamabana, Kazunori Muraki, Shinichi Doi and Shin-ichiro Kamei, "A Language Conversion Front-End for Cross-Language Information Retrieval". In *Proceedings of the 19th Annual ACM SIGIR Conference on Information Retrieval*, Zurich, Switzland, 1996.