교차언어 문서검색에서 다국어 온톨로지에 기반한 한영 질의어 변환

천정훈, 최기선
한국과학기술원 전산학과
{gray, kschoi}@world.kaist.ac.kr

Korean-to-English Query Translation based on Multilingual Ontology in Cross-Language Text Retrieval

Jung-Hoon Chun, Key-Sun Choi
Department of Computer Science KAIST, KORTERM

요 약

본 논문에서는 교차언어 문서검색(CLTR: Cross-Language Text Retrieval)에서의 한-영 질의어 변환을 다룬다. 질의어 변환시 영어 대역어 획득과정에서는 다음 두 가지를 고려한다. 첫째, 한국어 질의어를 구성하는 단어가 한가지 개념을 기호화하지만 이에 대응되는 영어 대역어들이 하나 이상인 경우이다. 둘째, 질의어 구성 단어가 둘 이상의 개념들을 기호화하는 다의성을 지닌 경우이다. 전자의 경우는 영어 대역어들이 모두 동일한 개념, 또는 유사한 개념을 나타내므로 그대로 검색에 이용한다 해도 검색 성능을 크게 좌우하지 않지만, 후자의 경우는 모든 개념을 다 검색에 이용하게 되면 정확률(precision)이 크게 떨어지게 된다. 이에 본 연구에서는 개념 선택단계와 선택된 개념의 영어 대역어들에 가중치를 주는 가중치 부가단계로 나누어 질의어 변환을 수행한다. 본 논문의 질의어 변환에서 영어 대역어는 대역사전 대신 다국어 온톨로지인 KAIST 분류어휘표와 한영 음차복원 모듈을 통해 얻어진다.

1. 서론

정보검색에서 질의어의 언어와 검색 대상 문서의 언어가 동일한 단일언어 문서검색(MLTR: Mono-Lingual Text Retrieval)과 달리, 교차언어 문서검색(CLTR: Cross-Language Text Retrieval)은 질의어의 언어와 다른 언어로 쓰여진 문서를 검색하는 것을 가리킨다.

기존의 교차언어 문서검색에 대한 접근 방법은 변환 방식에 따라 다음 세가지로 나누어진다. 첫번째 방법은 질의어 변환(query translation) 방식으로 질의어를 문서의 언어로 변환한 후 검색하는 방법이다. 두번째는 문서 변환(document translation) 방식으로 기계번역 시스템을 이용해 문서를 질의어의 언어로 변환한 후 검색하는 방법이다. 문서 변환 방식은 제한된 영역에서 일어와 한국어와 같은 비슷한 언어들간의 번역에서는 높은 검색 결과를 보여 주었다[Kwon97]. 그러나, 일반적인 영역에서 한국어와 영어간의 경우, 현재의 기계번역 수준으로는 고품질의 번역 결과를 얻기가 힘들다는 점과 방대한 양의 문서를 변환하는데 드는 비용을 고려해 볼 때 문서변환 방식은 그 한계를 지니고 있다. 세번째 방법은 언어적 변환을 행하지 않고 검색하는 방법이다. 대표적인 예로 LSI(Latent Semantic Indexing)를 이용한 방법이 있다.[Dumais97].

질의어 변환 방식은 대역어를 획득하는데 사용되는 자원에 따라 대역사전(bilingual dictionary)에 기반한 방법, 코퍼스(corpus)에 기반한 방법, 다국어 온톨로지에 기반한 방법으로 나누어질 수 있다. 대역사전이나 다국어 온톨로지에 기반한 방법의 대역어 획득과정에서는 다음 두 가지를 고려해야한다. 첫째는 한국어 질의어를 구성하는 단어가 한가지 개념을 기호화하지만 이에 대응되는 영어 대역어들이 하나 이상인 경우이다. 둘째는 질의어 구성 단어가 둘 이상의 개념들을 기호화하는 다의성을 지닌 경우이다. 전자의 경우는 영어 대역어들이 모두 동일한 개념, 또는 유사한 개념을 나타내므로 그대로 검색에 이용한다 해도 검색 성능을 크게 좌우하지 않지만, 후자의 경우는 모든 개념을 다 검색에 이용하게 되면 정확률(precision)이 크게 떨어지게 된다.

이에 본 연구에서는 다국어 온톨로지인 KAIST 분류어휘표에 기반하여 공기 정보를 이용해 질의어의 개념 선택단계와 선택된 개념의 영어 대역어들에 가중치를 부가하는 가중치 부가 단계를 제안한다. 분류어휘표에 등록되어 있지 않은 음차표기된 고유명사나 전문용어에 대해서 한영 음차복원 모듈을 통해 대역어를 얻어 검색에 이용한다.

본 논문의 구성은 다음과 같다. 2장에서는 교차언어 문서검색에서 질의어 변환에 관한 기존의 연구들을 살펴보고, 3장에서는 다국어 온톨로지인 KAIST 분류어휘표를 소개한다. 4장에서는 분류어휘표와 공기 정보를 이용한 질의어의 개념 선택과 대역어의 가중치 부가 단계를 설명한다. 다음으로 5장에서는 실험 방법 및 결과를 알아보고 6장에서는 향후 연구 계획과 결론을 맺는다.

2. 질의어 변환에 관한 기존 연구

이 장에서는 질의어 변환에 이용되는 자원에 따라 대역사전에 기반한 방법, 코퍼스에 기반한 방법, 다국어 온톨로지에 기반한 방법에 대한 기존의 연구를 살펴본다.

2.1 대역사전에 기반한 방법

대역사전은 다른 언어 자원들보다 비교적 획득이 쉽고 번역 방식이 단순하기 때문에 많이 이용되고 있는 질의어 변환 방법이다. [Hull96]에서는 불영(French-English) 대역사전을 이용하여 질의어 변환을 하였는데, 실험 결과 질의어의 단순한 단어 대 단어(word-to word) 변환을 통해서는 검색성능이 단일언어 문서검색의 40~60% 정도인 것으로 나타났다.

사전에 기반한 방법에서 이러한 성능저하는 크게 세가지 원인에서 비롯되는데 첫째, 질의어의 의미 모호성(sense ambiguity), 둘째, 사전에서 전문용어(terminology) 등에 해당하는 단어의 부재, 셋째, 구(phrase) 단위 번역의 오류이다 [Ballesteros96, Hull96]. 둘째와 셋째 원인에 해당하는 미등록어와 구 단위 번역의 오류는 사전의 질에 따라 그 정도가 달라지고 사전의 수정을 통해서 그 오류가 완화될 수 있으나 질의어의 의미 모호성은 단순히 사전의 수정만으로는 불가피하다.

[Davis96]은 영어 질의어를 변환하여 스페인어 문서를 검색하였는데, 질의어의 모호성 해소를 위해 다음의 과정을 수행하였다. 먼저 대역사전을 탐색하여 나온 스페인어 중에 영어와 동일한 POS(Part-Of-Speech)를 갖는 스페인어 대역어들만을 추출한다. 그 다음, 영어 질의어와 추출된 스페인어 대역어들을 질의어로 하여 병렬 코퍼스(parallel corpus)를 검색한다. 영어 질의어에 의해 검색된 영어 문서들과 스페인어 질의어에 의해 검색된 스페인어 문서들 중, 공통되는 문서를 가장 많이 뽑아낸 스페인어를 대역어로 선택하여 검색에 이용한다. [Ballesteros97]에서는 질의어 변환 전과 변환 후에 각각 질의어 확장(query expansion)을 수행하여 관련도가 높은 유사 단어들을 검색에 추가함으로서 검색 성능을 향상시켰다.

2.2 코퍼스에 기반한 방법

코퍼스에 기반한 방법은 병렬 코퍼스로부터 얻은 단어 사용 통계 정보를 이용하여 변환을 수행한다. 단어 벡터 변환(Term Vector Translastion)방법은 병렬 코퍼스를 구성하는 문서쌍의 단어들로부터 이차원 행렬로 색인한 공기 테이블을 만들어 공기 빈도를 표시하고 다양한 임계값에 따라 질의어 변환을 수행한다[Brown97].

[Davis95]에서는 영어 질의어로 영어-스페인어 병렬 코퍼스를 검색한 후 그 결과로 나온 상위 100개의 대응되는 스페인어 문서들 중에서 가장 빈도수가 높은 스페인 단어들을 질의어로 선택하였다.

2.3 다국어 온톨로지에 기반한 방법

다국어 온톨로지는 둘 이상의 언어의 단어들을 언어 독립적인 의미적 관계에 의해 표시하고 서로 연결하여 하나의 온톨로지를 구축한 것이다.

EuroWordNet 프로젝트에서는 WordNet을 독어, 이태리어, 스페인어, 영어 등 유럽 4개 언어의 단어들에 대해 서로간의 의미적 관계를 연결하는 중간언어 색인(interlingual index)을 만들어 다국어 온톨로지를 구축하고 이를 이용해 개념기반의 문서검색을 행한다[Gilarranz97].

NLM(National Library of Medicine)에 의해 구축된 어휘(vocabulary) 시스템인 UMLS(Unified Medical Language System)의 메타시소러스(Metathesaurus)는 영어, 불어, 독어, 스페인어, 포르투갈어의 번역을 연결한 다국어 온톨로지이다. [Eichmann98]은 구 단위 위주로 구성된 UMLS의 메타시소러스를 이용해 스페인어, 불어 질의어를 각각 영어 질의어로 변환하여 영어 문서집합을 대상으로 실험하였다. 이 실험에서 질의어 변환을 위해 질의어를 포함하는 메타시소러스의 모든 엔트리(entry)들을 나열하고, 그 중에서 정확하게 매치되거나 가장 많이 부분 매치되는 단어를 대역어로 선택하여 검색에 이용하였다. 이 실험에서는 제한된 분야지만 유사한 언어들간이라는 특성을 고려해 POS 태거나 코퍼스, 스테머(stemmer) 등의 다른 자원을 이용하지 않고도 다국어 온톨로지만을 이용해 일반적인 사전기반 방식에 상응하는 결과를 얻었다.

3. KAIST 분류어휘표

KAIST 분류어휘표는 bottom-up 방식에 의해 단어를 의미에 의해 정리한 다국어 온톨로지로서 한국어와 이에 대응되는 영어, 중국어, 일본어의 약 60,000여쌍의 어휘를 수록하고 있다.

1 체언의 분류	1.1 추상적 관계	10,148 개
	1.2 인간활동의 주체	5,755 개
	1.3 인간활동-정신 및 행위	16,131 개
	1.4 생산물 그리고 용구	6,294 개
	1.5 자연물 그리고 자연현상	6,403 개
2 용언의 종류	2.1 추상적 관계	3,892 개
	2.3 정신 및 행위	4,041 개
	2.5 자연 현상	750 개
3 상의 종류	3.1 추상적 관계	3,275 개
	3.3 정신에 이르는 행위	2,852 개
	3.5 자연현상	862 개
4. 그 밖의 종류	4.11 접속	127개
	4.3 감동	355 개

표 1. 분류어휘표의 상위 두자리 분류체계와 수록 어휘 수

분류어휘표는 분류들이 수록된 분류목록과 어휘들이 수록된 어휘수록집으로 구성되어 있다. 분류목록을 구성하는 분류는 5자리의 분류번호로 나타내어지고, 어휘수록집에 수록된 각각의 어휘들은 자신이 속하는 분류번호와 함께 식별자(identifier)로 쓰이는 최대 5자리의 번호를 더 가지고 있다. <표 1>은 분류어휘표에서 상위 두자리의 분류 체계와 수록어휘의 수를 보여주고 있다.

<표 2>는 실제로 분류어휘표에서 분류번호 "13801", "산업(Industry)"의 분류에 속하는 어휘들 중 일부분을 한국어와 영어 어휘에 대해서 보여주고 있다. 질의어 변환시 한국어 질의어는 <표 2>의 형태로 구조화되어 있는 분류어휘표를 탐색하여 영어 대역어 후보들을 획득하게 된다. 분류어휘표를 통해서 얻어진 영어 대역어 후보들은 다음 자장에서 설명하는 개념 선택 단계와 가중치 부가 단계를 거쳐 대역어로 선정되게 된다.

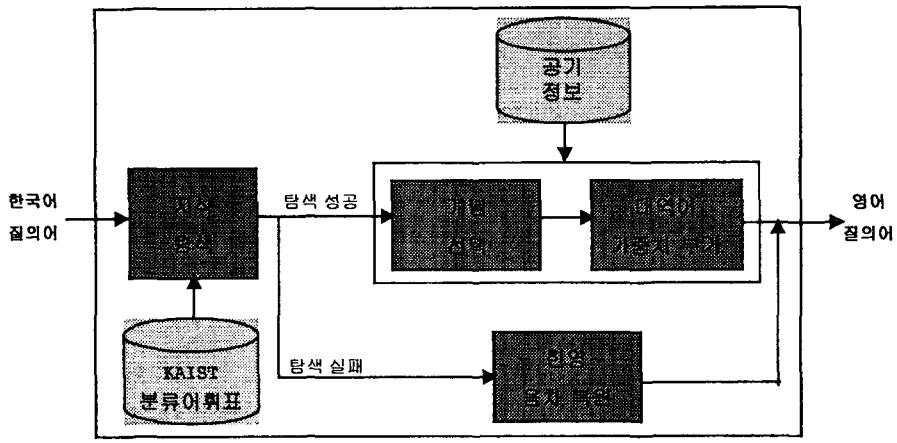
분류 번호	ID 번호	한국어 단어	영어 단어
13801	1 10	영업	Business Trade Operation
13801	2 10	상공업	Commerce and industry
13801	2 30	상업	Business Commerce Trade
13801	2 70	상사	Commercial affairs Commercial matters
13801	2 80	장사	Business Commerce Deal Trade
13801	2 90	행상	Hawking Itinerant trade Peddling
13801	3 10	공업	Industry Manufacturing industry
13801	3 20	수공업	Manual industry Handicraft Manufacturing

표 2. 분류번호 "13801", 산업-Industry에 속하는 어휘들

4. 다국어 온톨로지에 기반한 질의어 변환

본 논문의 질의어 변환 과정은 <그림 1>과 같다. 한국어 질의어가 입력으로 들어오면 먼저 분류어휘표를 탐색한다. 탐색에 성공하면 분류어휘표의 탐색결과가 개념 선택 단계로 넘겨진다. 개념 선택 단계에서는 분류어휘표에서 획득한 지식과 공기정보를 이용해 질의어의 개념을 선택한다.

그림 1. 한영 질의어 변환 시스템 구조도



이 모듈에서 선택된 개념의 대역어들은 대역어 가중치 부가 단계를 거쳐 영어 질의어로 쓰이게 된다. 입력으로 들어온 한국어 질의어가 분류어휘표의 탐색에서 실패했을 때는 한영 음차복원 모듈을 통해 영어 질의어를 생성한다.

4.1 분류어휘표에 기반한 개념 선택

"자동차 공기 오염"이라는 한국어 질의어가 주어졌을 때, 분류어휘표를 탐색하여 얻어지는 결과는 <표 3>과 같다.

한국어 단어	분류 번호	분류	영어 단어
자동차	14650	탈 것	Autocar Automobile Car Motorcar Motor vehicle
공기	11302	분위기	Atmosphere
	14520	식기	Bowl Empty vessel
	14570	장난감	Ally Jackstone Pebble Marble
	15120	공기	Air
오염	15160	물질의 변화	Pollution Contamination

표 3. 분류어휘표에서 "자동차 공기 오염"

<표 3>에서 "공기"라는 한국어 단어는 "분위기", "식기", "장난감", "공기"라는 4개의 다른 개념을 지닐 수 있음을 알 수 있다. 즉, 한국어 단어 "공기"는 4개의 다른 개념을 나타내는데 쓰이는 것이다. 그에 비해, "자동차"와 "오염"은 "탈 것"이라는 분류와 "물질의 변화"라는 분류에 속하며 각각 1개의 개념만을 지니고 있다. 하나의 개념만을 지닌 "오염"의 경우, 이 개념을 나타내는 영어 단어가 "Pollution"과 "Contamination"으로 하나 이상이 될 수 있음을 보여준다. "장난감"의 분류에 속하는 "공기"의 경우도 "Ally", "Jackstone", "Marble", "Pebble"로 개념 하나에 대해서 대역될 수 있는 단어가 하나 이상임을 알 수 있다.

"자동차 공기 오염"을 영어 질의어로 변환하여 검색할 때, "공기"가 지니는 4개의 모든 개념들에 속하는 영어 대역어들을 검색에 이용하게되면 사용자가 원하지 않는 불필요한 문서들이 지나치게 많이 검색되어 정확률을 떨어뜨리게된다. 따라서 "공기"가 지니는 4개의 후보 개념들 중 하나의 올바른 개념을 선택한다.

개념 선택은 각 질의어 단어의 개념들을 이루는 영어 대역어의 그룹(group)들간의 공기 정보를 계산하여 가장 높은 값을 가지는 그룹을 개념으로 선택한다. 개념에 따라 나누어지는 '자동차 공기 오염"의 대역어 그룹은 다음과 같이 나타낼 수 있다.

$G_{자동차1}$ = {Autocar, Automobile, Car, Motorcar, Motor vehicle}
$G_{공기1}$ = {Atmosphere}
$G_{공기2}$ = {Bowl, Empty Vessel}
$G_{공기3}$ = {Ally, Jackstone, Marble, Pebble}
$G_{공기4}$ = {Air}
$G_{오염1}$ = {Contamination, Pollution}

개념으로 나뉘어진 대역어 그룹간의 공기 정보는 EDR 영어 공기 사전(EDR English Cooccurrence Dictionary)으로부터 공기 빈도수(cooccurrence frequency)를 추출하여 아래의 식을 통해 계산된다.

$$C(A) = \sum_{i=1}^{n} C(A, B_i)$$

$$C(A,B) = \frac{1}{N_A} \cdot \sum_{i=1}^{N_A} \sum_{j=1}^{N_B} \frac{2 \cdot f(a_i b_j)}{f(a_i) + f(b_j)} \text{-----(1)}$$

C(A) : 그룹 A의 공기 정보값
C(A,B) : 그룹 A가 그룹 B에 대해 갖는 공기 정보값
A, B : 대역어 그룹
Bi : 원시 질의어 단어가 A와 다른 대역어 그룹
n : 원시 질의어 단어가 그룹 A와 다른 그룹의 수
N_A : 그룹 A의 대역어 수
N_B : 그룹 B의 대역어 수
a_i : 그룹 A의 i번째 대역어
b_j : 그룹 B의 j번째 대역어
$f(a_i)$: a_i의 발생 빈도수
$f(b_j)$: b_i의 발생 빈도수
$f(a_i b_j)$: a_i와 b_j의 공기 빈도수

식 (1)은 개념에 따라 나뉘어진 대역어 그룹 A와 B가 있을 때, 공기 사전에서 A의 모든 대역어들과 B의 모든 대역어들간의 공기 빈도수와 각각의 빈도수를 추출하여 Dice Coefficient 값을 구해 더해준다. 그리고 정규화를 위해 그룹 A의 대역어 수로 나누어준다. 따라서 "자동차 공기 오염"의 $G_{공기1}$의 공기 정보값은 다음과 같이 구해진다.

$$C(G_{공기1}) = C(G_{공기1}, G_{자동차1}) + C(G_{공기1}, G_{오염1})$$

한국어 질의어가 분류어휘표를 통해 하나 이상의 개념을 지니게 되면 각각의 개념을 나타내는 영어 대역어 그룹이 형성된다. 그리고 그 대역어 그룹들은 식 (1)에 의해 공기 정보값을 얻게되고 그 중 가장 높은 값을 가지는 그룹을 그 단어의 개념으로 선택한다.

4.2 대역어들의 가중치 부가 단계

개념 선택 단계에서 질의어의 여러 개념들 중 하나의 개념만이 선택됨으로서 가중치 부가 단계에서는 하나의 대역어 그

룹만을 가지게 된다. 공기 정보를 이용한 기존의 연구들에서는 개념 선택을 하지 않고 다른 모든 대역어들간의 공기 정보값을 계산하였다. 그 결과, 대역어의 가중치를 계산할 때, 이미 고려되지 않아야 할 개념에 속하는 대역어들과의 공기 정보값이 더해짐으로서 대역어의 선택이나 가중치 부가시에 노이즈(noise)로 작용하여 잘못된 대역결과를 가져올 수 있다.

예를 들어, "자동차 공기 오염"의 "공기"가 개념 선택 단계에서 $G_{공기4}$로 선택되었다고 하자. 그러면 가중치 부가 단계에서는 다음과 같은 대역어 그룹들이 존재할 것이다.

$G_{자동차1}$ = {Autocar, Automobile, Car, Motorcar, Motor vehicle}
$G_{공기4}$ = {Air}
$G_{오염1}$ = {Contamination, Pollution}

이제 질의어 단어들이 모두 하나의 개념만을 지니므로, 한 개념에 속하는 다수의 대역어들에 각각의 가중치를 부가한다. "오염"의 대역어들은 "Contamination"과 "Pollution"인데, "Pollution"의 가중치를 결정하기 위해서는 "Pollution"과 $G_{자동차1}$, "Pollution"과 $G_{공기4}$의 공기 정보값을 이용한다. 즉, 이미 제외된 $G_{공기1}$, $G_{공기2}$, $G_{공기3}$과의 공기 정보값은 계산에 들어가지 않는다.

각각의 대역어들에 부가할 가중치의 계산은 식 (1)과 맥락을 같이 하여 다음과 같이 계산된다.

$$W(a) = \sum_{i=1}^{n} W(a, B_i)$$

$$W(a,B) = \sum_{i=1}^{N_b} \frac{2 \cdot f(ab_i)}{f(a) + f(b_i)} \text{------(2)}$$

W(a) : 대역어 a의 공기 가중치
W(a,B) : 대역어 a가 그룹 B의 대역들에 대해 갖는 공기 정보값
B : 대역어 그룹
Bi : 원시 질의어 단어가 a의 대역어 그룹과 다른 대역어 그룹
n : 원시 질의어 단어가 a의 대역어 그룹과 다른 그룹의 수
Nb : 그룹 B를 이루는 대역어 개수
a : 그룹 A의 i번째 대역어
b_j : 그룹 B의 j번째 대역어
f(a) : a의 발생 빈도수
$f(b_j)$: b_i의 발생 빈도수
$f(ab_j)$: a와 b_j의 공기 빈도수

4.3 한영 음차 복원 모듈을 이용한 대역어 획득

불영 대역사전에 기반한 [Hull96]의 실험에서는 질의어 변환시 미등록어가 나타나면 불어 질의어를 그대로 영어 질의어로 사용했다. 이 방법은 미등록어에 대해 대역어를 주지 못하고 검색한 것보다는 어느 정도 나은 결과를 얻었는데 이는 같은 문자를 공유하고 있는 불어와 영어사이에서 가능한 방법이다. 한국어와 영어를 고려해 볼 때, 한글이 영어 문서에서 나타나는 경우는 매우 드물다.

질의어가 "패스트 푸드"나 "테디 베어"처럼 음차표기된 고유명사이거나 전문용어일 경우 분류어휘표의 탐색은 실패하게 된다. 이를 보완하기 위해 [이재성99]의 자동정렬 방식과 피봇 방식을 혼합한 한영 복원 모델을 사용한다. 자동정렬 방식은 학습 단어쌍에서 자동으로 발음단위를 선택하여 정렬하는 방식이고, 피봇 방식은 발음기호에서 외래어 표기법으로 표기된 한글로 학습한 후 이를 이용하여 복원하는 방식이다.

5. 실험 및 평가

TREC-6의 CLIR Track에서 90년도 AP news를 test-collection으로 하여 실험하였다. 한국어 질의어는 24개 CLIR 영어 Topic에서 title 필드에 나타나는 명사들을 수동으로 번역하여 생성하였고 실험을 위해 SMART.11.0 시스템을 사용하였다.

<표 4>에서는 질의어 "패스트 푸드 유럽"의 원 영어 질의어와 분류어휘표의 단순 탐색 결과 얻어진 대역어들, 그리고 음차 복원 모듈에 의해 얻어진 대역어들을 나열하였다. <표 4>는 따로 떨어진 "패스트"와 "푸드"는 분류어휘표에 수록되어 있지 않음을 보여준다.

원 질의어	Fast, Food. Europe
분류어휘표	Europe
음차 복원	Europe, Fascet, Past, Fascete, Paste, Fast, Pood, Food, Pund, Pook, Fook

표 4. "패스트 푸드 유럽"의 대역어 리스트

<표 5>는 <표 4>의 영어 단어들을 질의어로 하여 검색된 결과를 정확률과 11-point 평균 정확률로 보여준다. <표 5>를 보면 음차복원을 통해서 성능이 50% 이상 향상되었음을 알 수 있다. 이는 교차언어 문서검색의 질의어 변환에서 음차 복원 또는 음차표기 모듈의 중요성을 단적으로 보여준다.

	원 질의어	분류어휘표	음차 복원
11-pt 평균 정확률	0.0511	0.0049	0.0315
%	-	9.59	61.64
정확률 at 5 docs	0.2000	0.0000	0.0000
at 10 docs	0.1000	0.0000	0.1000
at 15 docs	0.0667	0.0000	0.0667
at 30 docs	0.0667	0.0000	0.0667

표 5. "패스트 푸드 유럽" 결과

<표 6>은 한국어 질의어 "자동차 공기 오염"의 원 영어 질의어(O)와, 분류어휘표의 단순 탐색 결과(T1) 얻어진 대역어들, 그리고 개념 선택 단계(T2)를 통해 얻어진 대역어들이다. 가중치 부가 단계(T3)에서 쓰이는 대역어 리스트는 T2와 같으므로 생략하였다.

O	Automobile, Air, Pollution
T1	Autocar, Automobile, Car, Motocar, Motor vehicle, Atmosphere, Empty vessel, Bowl, Ally, Jackstone, Pebble, Marble, Air, Pollution, Contamination
T2	Autocar, Automobile, Car, Motocar, Motor vehicle, Air, Pollution, Contamination
C	Autocar, Automobile, Car, Motocar, Motor vehicle, Atmosphere, Air, Pollution, Contamination

표 6. "자동차 공기 오염"의 대역어 리스트

비교를 위해 개념 선택 모듈을 사용하지 않고, 단지 다른 질의어의 대역어들과 공기 빈도가 없는 대역어는 제거해주는 방법(C)으로 얻어진 대역어 리스트도 함께 나타내었다. 이 방법(C)에서는 만약 질의어에 대한 대역어들의 공기 빈도수가 모두 '0'일 때는 그 단어의 모든 대역어들을 검색에 이용한다.

	O	T1	T2	T3	C
11-pt	0.6918	0.0315	0.5966	0.6768	0.6079
%	-	64.98	86.23	97.80	87.87
5 docs	1.0000	0.2000	0.8000	0.8000	0.8000
10 docs	0.8000	0.4000	0.9000	0.7000	0.9000
15 docs	0.8000	0.6000	0.7333	0.6667	0.8000
30 docs	0.7667	0.4667	0.7333	0.7667	0.7333

표 7. "자동차 공기 오염" 결과

<표 7>은 <표 6>의 대역어들로 검색하였을 때 얻어진 결과이다. <표 6>에서 T2와 C의 대역어 리스트를 비교해 볼 때, 다른 것은 "Atmosphere"가 T2에는 없지만 C에는 있다는 것이다. 이것은 C가 개념 선택 모듈을 수행하지 않기 때문에, 보다 공기 정보값이 높은 "Air"에 의해 개념이 제거되지 않아서이다. 그런데, 흥미로운 점은 개념 선택을 거친 T2와 그렇지 않은 C의 결과를 비교해보면 오히려 C가 1.64%의 향상을 보이고 있다. 이것은 "분위기"의 개념으로 쓰인 "Atmosphere"가 "Air"와 의미가 유사한 "대기"라는 개념도 가지고 있기 때문이다. 즉, "Atmosphere"가 노이즈(noise)로 작용하지 않고 오히려 의미있는 검색어로 쓰여 성능을 향상시킨 셈이다. 이것은 개념 선택을 통해 원시 언어 상에서 다른 개념들을 제거하더라도 다시 목적언어인 대역어가 의미 모호성을 가지는 경우이다.

본 실험에서는 영어 명사들 간의 공기 빈도수를 얻기 위해 EDR 공기 사전을 이용하였다. 그런데 공기 사전에 나타나는 명사들의 공기 정보는 대부분 복합명사나 구를 이루는 단어들의 공기만을 수록하고 있다. 이는 공기 사전의 정보가 질이 높은 정보라는 것이긴 하나 적은 양의 정보로 인해 data sparness 문제를 가져왔다.

6. 결론

본 논문에서는 다국어 온톨로지인 KAIST 분류어휘표에 기반하여 한영 질의어 변환을 수행하였다. 입력으로 들어온 한국어 질의어는 분류어휘표를 통해 가능한 개념들과 이에 대한 대역어들을 얻게된다. 질의어의 개념이 둘 이상일 때는 공기 정보를 이용하여 하나의 개념만을 선택한다. 그리고 선택된 개념에 속하는 대역어들에 대해 가중치를 부가하는 단계를 거침으로서 영어 질의어로 변환된다. 분류어휘표에 등록되어있지 않은 음차 표기된 고유명사와 전문용어 등에 대해서는 한영 복원 모듈을 통해 영어 질의어를 생성한다.

교차언어 문서검색에서 질의어 변환에 관한 현재까지 연구의 대부분은 대역사전과 코퍼스에 기반한 방법에 관한 것이었고 다국어 온톨로지(multilingual ontology)나 다국어 시소러스(multilingual thesauri)를 이용한 변환에 대해서는 많은 연구가 이루어지지 않았다. 이러한 주원인은 다국어 시소러스의 구축과 관리에 비용이 많이 든다는 이유 때문이었는데 대역사전이나 활용할 수 있는 제대로 된 병렬 코퍼스를 구축하는데도 다국어 온톨로지의 구축만큼이나 비용이 든다는 것을 감안한다면 이에 대한 연구도 활발히 이루어져야 할 것이다.

참고문헌

[이재성99] 이재성, "다국어 정보검색을 위한 영한 음차 표기 및 복원 모델", 한국과학기술원 전산학과 박사학위 논문, 1999.

[Ballesteros96] L. Ballesteros 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.

[Ballesteros97] L. Ballesteros and W.B. Croft, "Phrasal translation and query expansion techniques for cross-language information retrieval", In Proceedings of the 20th International ACM SIGIR Conference on Research and Development in Information Retrieval, July 1997.

[Brown97] Ralf D. Brown, "Corpus-based query translation for cross-lingual information retrieval", SIGIR-97 workshop on Cross-Lingual Information Retrieval, 1997.

[Davis95] M. Davis and T. Dunning, "Query translation using evolutionary programming for multilingual

information retrieval", In 4th Annual Conference on Evolutionary Programming, 1995.

[Davis96] M. Davis, "New experiments in cross-language text retrieval at NMSU's computiing research lab", In the fifth Text Retrieval Conference(TREC-5), November 1996.

[Dumais97] S.T. Dumais, T.A. Letsche, M.L. Littman, and Landauer T.K., "Automatic cross-language retrieval using latent semantic indexing", 1997 AAAI Symposium on Cross-Language Text and Speech Retrieval. American Association for Artificial Intelligence, March 1997.

[Eichmann98] David Eichmann, Miguel E. Ruiz, and Padmini Srinivasan, "Cross-Language Information Retrieval with the UMLS Metathesaurus", SIGIR '98, 1998.

[Gilarranz97] Julio Gilarranz, 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.

[Hull96] David A. Hull and Gregory Grefenstette, "Querying across languages: a dictionary-based approach to multilingual information retrieval", In Proceedings of the 19th ACM SIGIR Conference, 1996.

[Kwon97] O-W Kwon, I.S. Kang, J-H Lee and 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.