교차 언어 문서 검색에서 질의어의 중의성 해소 방법

강인수, 이종혁, 이근배
포항공과대학교 전자계산학과

Word Sense Disambiguation in Query Translation of CLTR

Insu Kang, Jong-Hyeok Lee, Geunbae Lee
Dept. of Computer Science and Engineering, POSTECH

요 약

정보 검색에서는 질의문과 문서를 동일한 표현으로 변환시켜 관련성을 비교하게 된다. 특히 질의문과 문서의 언어가 서로 다른 교차 언어 문서 검색 (CLTR : Cross-Language Text Retrieval) 에서 이러한 변환 과정은 언어 변환을 수반하게 된다. 교차 언어 문서 검색의 기존 연구에는 사전, 말뭉치, 기계 번역 등을 이용한 방법들이 있다. 일반적으로 언어간 변환에는 필연적으로 의미의 중의성이 발생되며 사전에 기반한 기존 연구에서는 다의어의 중의성 의미해소를 고려치 않고 있다. 본 연구에서는 질의어의 언어 변환시 한-일 대역어 사전 및 카도가와 시소러스 (角川 시소러스) 에 기반한 질의어 중의성 해소 방법과 공기하는 대역어를 갖는 문서에 가중치를 부여하는 방법을 제안한다. 제안된 방법들은 일본어 특허 문서를 대상으로 실험하였으며 5 % 의 정확도 향상을 얻을 수 있었다.

1. 서 론

최근 들어 인터넷과 웹 기술의 발전으로 정부, 기업체, 연구기관, 학교로부터 엄청난 양의 전자화된 정보가 쏟아져 나오고 있다. 또한 지역 및 국가간의 정보 교류도 가속화되고 있어 이를 충족시키기 위해 다양한 정보 검색의 응용 분야가 발생하고 있다. 그 중의 하나가 언어의 장벽을 뛰어 넘어 원하는 문서를 검색하고자 하는 응용이다. 즉 질의문의 언어와 다른 언어로 쓰여진 문서를 검색하는 교차 언어 문서 검색 (CLTR : Cross-Language Text Retrieval) 이다.

이것은 전통적인 정보 검색 방법론을 사용하기 전에 언어 변환 과정이 필요하므로 한층 더 어려운 문제이다. 언어 변환 과정이란 서로 다른 질의문의 언어와 검색 대상 문서의 언어를 어느 쪽으로든 변환시켜서 비교 가능한 하나의 언어로 일치시키는 과정을 말한다. 개념적으로는 질의문과 문서의 효율적인 관련도 비교를 위해 동일한 표상 (Representation Space) 으로 질의문과 문서를 변환시키는 과정으로 볼 수 있다 [Douglas96].

최근의 연구들은 사전, 말뭉치, 기계 번역 등의 자원을 이용하여 언어 변환 문제에 접근하고 있다. 기계 번역 시스템을 이용한 언어 번역 시도는 기계 번역 시스템을 통해 언어 분석에 소요된 시간과 자원과 비해 실질적인 성능이 기대에 미치지 못했다. 말뭉치를 이용하면 단일 언어 정보 검색 (MLIR : Mono-Lingual Information Retrieval) 의 성능에 근접하지만, 말뭉치의 도메인 의존성으로 인해 시스템의 확장이 어렵고 해당 분야의 말뭉치라 하더라도 구하기 어렵다는 단점이 있다. 사전에 기반한 시도 역시 단일 언어 정보 검색에 비해 40~60% 정도의 성능 저하 [Hull96, Ballesteros96] 를 가져오지만 다른 방법들에 비해 실패의 원인을 알고 있으며 사전을 통한 확장이 가능하다는 이점이 있다. 사전 기반 방법의 성능 저하 원인들은 미등록어, 어의 중의성 (word sense ambiguity), 구번역 (phrase translation) 에서의 오류에서 비롯되며 미등록어와 구번역 문제는 주기적 사전 수정을 통해 해결 가능하다. 하지만 질의어의 어의 중의성을 해결하기 위한 직접적인 시도는 거의 없었으며, 사전을 이용하는 접근은 질의 언어 변환을 위한 대역어 선정시 필연적으로 질의어의 어의 중의성 해소 (WSD : Word Sense Disambiguation) 문제에 부딪히게 된다.

현재까지 교차 언어 문서 검색의 질의어 언어 변환에서 어의 중의성 해소문제에 대한 최초의 직접적인 시도는 [Hull97]에 의한 것이었으나 이것은 질의문 작성시 사용자로부터 어의 중의성 해소를 위한 추가의 정보를 얻는 방법이었다. 본 연구에서는 질의문 작성에 사용자의 추가 노력 없이 질의어의 어의 중의성을 해소하는 공기 가중치 할당 (Cooccurrence Weighting) 방법과 시소러스 개념 수렴 (Thesaurus Grouping) 방법을 제안한다. 이것은 한-일 대역어 사전과 카도가와 시소러스를 이용한 방법이며 한-일 교차 언어 문서 검색에 적용시켜 보았다. 현재 일본어 평가 검증용 문서 집합 (Test Collection) 을 구하기 어려우므로 일본어 특허 문서를 대상으로 실험하였으

며 제안된 방법을 통해 성능 향상을 얻을 수 있었다. 본 논문의 구성은 먼저 교차 언어 문서 검색에 대한 선행 연구들을 살펴 보고, 3장에서 질의 언어 변환에서의 어의 중의성 해소방법을 제안하며, 4장에서 실험 결과와 마지막으로 향후 연구 계획과 결론을 기술하고자 한다.

2. 교차 언어 문서 검색에 대한 선행 연구

앞에서 언급했듯이 교차 언어 문서 검색에서는 질의문과 검색 대상 문서를 하나의 표현으로 일치시켜 연관성을 비교할 필요가 있다. 즉 단일 언어 정보 검색과 달리 언어간 변환 과정이 필수적으로 개입되어야 한다. 이 언어 변환은 크게 질의문을 검색 대상 문서의 언어로 변환시키는 방법과 검색 대상 문서를 질의문의 언어로 변환시켜 인덱싱하는 방법이 있다.

문서를 질의문의 언어로 변환하여 인덱싱하는 방법은 기계 번역 시스템이 없다면 현실적으로 거의 불가능하고, 다른 언어로의 확장시 해당 언어의 기계 번역 시스템이 필요하므로 자원과 시간의 부하를 견디기 힘들 것이다. 질의문을 검색 대상 문서의 언어로 변환시키는 방법은 현재 대부분의 연구에서 행해지는 방식이다. 그 이유는 질의문이 상대적으로 짧아 처리가 간단하다는 이유도 있겠지만 기존 단일 언어 정보 검색의 인덱싱 모듈을 수정없이 그대로 이용할 수 있다는 장점이 있기 때문이다. 즉 질의문의 언어를 검색 대상 문서의 언어로 변환시키는 질의 변환 모듈만 있으면 기존에 구축된 모든 단일 언어 정보 검색시스템의 검색 부분만 교체시킴으로서 교차 언어 문서 검색 시스템을 얻을 수 있는 것이다.

두가지 언어 변환 방법 중 어느 편을 택하든 질의문과 문서의 언어가 서로 다르므로 변환 과정은 필수적이며 이때 어떠한 형태로든 외부 자원 (사전, 말뭉치, 기계 번역 시스템 등) 을 이용하지 않을 수 없다[Douglas97]. 현재까지의 교차 언어 문서 검색연구는 이러한 언어 변환에 사용되는 자원의 형태에 따라 크게 3가지로 분류해 볼 수 있다.

2.1 사전에 기반한 방법

사전에 기반한 접근은 최근 쉽게 얻을 수 있는 전자화된 사전 (MRD : Machine Readable Dictionary) 의 등장으로 가능해진 방법이며 크게 세가지 시도가 있었다. 먼저 질의문의 단순 단어 변환 방식은 어의 중의성이 있는 단어의 의미 해소에 사전의 최초 의미-일반적으로 가장 많이 쓰이는 최초 대역어-를 택하는 방식이며 단일 언어 정보 검색에 비해 40~60% 의 성능 저하를 보인다 [Hull96, Ballesteros96]. 다음으로 검색 대상 문서의 언어로 번역하기 전과 후에 각각 질의어 확장 (Query Expansion) 을 하는 국부 귀환 방식 (local feedback) 은 단순 단어 변환의 성능을 반 이상 향상시킨다. 이것은 변환 전에 질의어의 개념을 강조하도록 유사 단어들을 추가시켜 변환을 위한 단어를 많이 보유함으로써 정확도를 높이고, 변환된 질의문에서 문서의 언어로 다시 질의어 확장을 함으로써 오번역 단어로 인한 영향을 감소시켜 재현율의 향상을 얻을 수 있기 때문이다 [Ballesteros96]. 세번째 시도는 질의문의 구가 대상 문서 언어의 구나 단어로 번역되는 경우에 단순 단어 변환을 행함으로서 야기되는 오류를 줄이기 위해 구 대역어 사전 (bilingual phrase dictionary) 을 이용한 시도이며 사전 기반 방법들 중 가장 좋은 성능을 보였다 [Ballesteros97]. 이러한 사전 기반 시도들은 다음과 같은 공통된 제약이 있다. 먼저 질의 단어는 여러 대역어로 번역될 수 있으며, 대역어들간 의미차가 심한 경우가 많다 [Douglas96]. 둘째 질의 단어가 사전에 없는 경우이다. 예를 들어 최신 기술 분야 용어나 약어, 속어 등이 질의문에 포함되어 있을 수 있다. 마지막으로 구가 갖는 고유한 의미 변환에 이용할 구 대역어 사전을 구하기 어려우며 구축하기도 힘들다는 것이다. 위의 둘째, 세째 제약은 주기적으로 사전을 갱신함으로서 해소될 수 있으므로 질의 언어 변환의 핵심은 질의어에 나타난 중의어의 올바른 의미를 찾는 것이 된다. 현재까지는 질의문에 나타나는 중의어에 대한 어의 중의성 해소 기법은 시도된 바 없고, 국부 귀환이나 구 변환을 통해 오번역으로 인한 영향을 감소시키려는 입장을 취해왔다. 하지만 질의 변환을 통한 교차 언어 문서 검색에서는 필연적으로 중의어의 의미를 해소해야만 한다.

따라서 본 연구에서는 공기 가중치 할당과 질의어의 개념에 기반한 시소러스 개념 수렴을 통하여 중의성을 해소하는 방법을 제안하고자 한다.

2.2 말뭉치에 기반한 방법

사전 기반 방법이 갖는 여러 제약들은 말뭉치를 통해 해소될 수 있는 여지가 많다. 특히 일상에서 사용하는 단어와 사전에 수록된 단어 간 시간적 격차는 전문 기술 용어일수록 더 크며 이것은 해당 전문 분야의 대량의 말뭉치를 통해 좁힐 수 있다. 말뭉치에 기반한 시도로는 크게 두가지로 분류할 수 있다. 첫째 수학적 접근법으로 의미 내장 인덱싱 (LSI : Latent Semantic Indexing) 이 대표적이다 [Deerwester90, Berry95, Susan97]. 의미 내장 인덱싱은 대역 문서 집합 (Parallel document) 으로부터 뽑혀진 용어-문서 행렬을 SVD (Singular Value Decomposition) 를 통해 의미 내장 구조 (latent semantic structure) 로 변환해 두고, 입력 질의문에 대해, 용어 매칭이 아닌 용어의 개념 간의 유사도에 따라 검색을 시도한다. 의미 내장 인덱싱 방법은 인덱싱되지 않은 용어들로 구성된 질의문을 던져도 대상 문서를 검색할 수 있다는 장점이 있다. 이런 수학적 시도는 시스템이 커질 경우 복잡도 (complexity) 가 문제된다는 단점이 있다. 또한 용어-문서 행렬을 얻기 위해 사용된 말뭉치에 의존적이므로 확장에 무리가 있다. 둘째로 말뭉치의 통계적 분석을 통해 자주 공기하는 단어들을 대역어로 사용하는 방식이 있으며 [Davis95, Sheridan96], 주어진 말뭉치의 주제별 문서 정렬에 따라 대역어의 질이 결정된다. 말뭉치 기반 방법들의 공통된 제약으로는 말뭉치를 구하기 어렵다는 것이다. 결국 도메인이 작고 제한된 말뭉치들을 사용한다는 것이 가장 큰 어려움이다.

2.3 통합 방법

교차 언어 문서 검색의 목표 달성을 위해 사전이나 말

뭉치만을 사용할 필요는 없다. 필요하다면 사용자 개입도 질의 변환에 이용할 수 있을 것이다. 사전에 기반한 방법은 언어에 대한 넓고 얕은 단어의 변환에 유효하며, 말뭉치에 기반한 방법은 좁고 깊은 전문 분야의 용어들을 다룰 수 있으므로 양자의 약점을 보완하는 측면에서 앞으로의 교차 언어 문서 검색 연구는 사전, 말뭉치, 시소러스, 기계 번역 시스템, 사용자 개입 등을 모두 이용하는 방법을 찾아야 할 것이다 [Hull96]. 현재까지 보고된 여러 통합 시스템들이 있지만 아직은 초보 단계이다.

3. 질의 언어 변환에서의 어의 중의성 해소 방법

3.1 대역어들의 의미 조합

질의문을 구성하는 질의어들 중 적어도 하나가 다의어일 경우 질의문을 검색 대상 문서의 언어로 변환하는 가능성은 두가지 이상 발생한다. 먼저 사용자 질의문의 질의어들을 성분으로 갖는 질의문 벡터 SQ를 식 (1)과 같이 정의하면, 식 (1)로부터 각 질의어에 대응하는 대역어 집합을 식 (2)와 같이 표시할 수 있다.

$$SQ = (SQ_1, SQ_2, \ldots, SQ_n) \quad \text{----------(1)}$$

where

n = 사용자 질의문에서 추출된 질의어 개수

SQ_i : 질의문을 구성하는 i번째 질의어

$$\left.\begin{array}{l} SQ_1\text{의 대역어 집합 } SSQ_1 = \{TQ11, TQ12, \ldots, TQ1S_1\} \\ SQ_2\text{의 대역어 집합 } SSQ_2 = \{TQ21, TQ22, \ldots, TQ2S_2\} \\ \ldots \\ SQ_n\text{의 대역어 집합 } SSQ_n = \{TQn1, TQn2, \ldots, TQnS_n\} \end{array}\right\} \text{-(2)}$$

where

TQij : 질의어 SQ_i의 j번째 대역어

$1 \le j \le S_i, 1 \le i \le n$

S_i : SQ_i의대역어 개수

이제 질의문 SQ로부터 변환되는 대역어들의 의미 조합은 식 (3)으로 정의할 수 있다.

$$K \in X_{i=1}^{n} SSQ_i \quad \text{----------(3)}$$

즉 K는 사용자 질의문 벡터 SQ를 구성하는 각 질의어들의 대역어 집합들의 집합곱 (cartisian product) 의 한 원소가 되며, 그림 3.1에서 임의의 p번째 대역어 의미 조합인 K_p에 해당한다. (그림에서 p번째 의미 조합 = K_p)

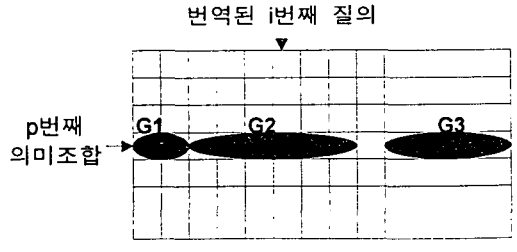

그림 3.1 사용자 질의문 SQ의 대역어들의 의미 조합
(그림에서 p번째 의미 조합 = K_p)

3.2 공기 가중치 할당 (CW : Cooccurrence Weighting)

자연 언어 질의문을 작성하는 것은 문서를 작성하는 것처럼 하나의 문장을 만드는 과정이다. 한 문서 내에 같이 나타나는 단어의 의미들은 문장을 만들 때 같이 사용되는 의미들이므로 다른 언어로 작성된 질의문에서도 질의문을 작성할 때 그 의미들로 사용되었을 것이라고 생각할 수 있다. 바꿔 말하면 어의 중의성이 있는 질의어들의 의미 조합이 검색 대상 문서내에서 공기하지 않는다면 질의문에서도 그 의미로는 사용되지 않았을 것이라고 간주할 수 있다. 즉 번역된 대역어들 중 한 문서 내에 공기하는 것들은 맞게 번역되었을 가능성이 높을 것이므로, 대역어들이 공기하는 개수에 비례하여 번역된 질의문과 검색 대상 문서와의 유사도 (similarity) 를 높혀줌으로써 맞게 의미 해소된 대역어를 가지는 문서들을 검색 시 상위 등급에 위치시켜 줄 것이다.

이러한 공기 가중치 할당방법을 벡터 모델에 적용하면 교차 언어 문서 검색의 어의 중의성 해소를 위한 새로운 검색 모델을 식 (4)과 같이 얻을 수 있다. 본 논문의 실험에서는 상수 C를 1.2로 정하였다.

$$sim(P,D) = C^{M-1} \frac{\sum_{i=1}^{t} P_i D_i}{\sqrt{\sum_{i=1}^{t} P_i^2 \sum_{i=1}^{t} D_i^2}} \quad \text{---------- (4)}$$

where

K_p : 질의문의 p번째 대역어 의미조합.

M : 벡터 P와 D의 일치하는 0이 아닌 성분의 개수

t : 인덱싱에 사용된 용어의 개수

C : 상수 (실험에서는 1.2로 잡았음)

$D = (D_1, D_2, \ldots D_t)$

D_i: i번째 질의어의 문서벡터D에서의 가중치.

$$P = (P_1, P_2, \ldots P_t), \quad \begin{array}{l} P_i = 1 \text{ if } P_i \in K_p \\ P_i = 0 \text{ if } P_i \notin K_p \end{array}$$

식 (4)의 의미는 다음과 같다. 사용자의 질의문 벡터 SQ를 구성하는 질의어들을 검색 대상 문서의 언어로 변환했을 때 대역어들의 가능한 의미 조합을 얻을 수 있다. 그 의미 조합의 하나인 K_p에 대해서 K_p를 구성하는 번역된 대역어들을 t 차 벡터 공간으로 보내어 벡터 P를 얻는다. 이 t 차 벡터 P와 t 차 문서 벡터 D사이에 두개 이상의 공유 성분이 나타날 때마다 벡터 모델의 기존 코사인 유사도 함수에 가중치 C를 곱해줌으로서 대역어의 의미 조합과 문서 사이의 유사도를 높여 주게 된다. 위의 식 (4)에서 구해진 P는 K_p로부터 얻어지는 벡터이며 모든 K_p (가능한 대역어 의미 조합) 에 대해 대응하는 벡터 P를 구하고 유사도를 계산하여 문서를 등급화를 한다.

공기하는 대역어를 갖는 문서에 더 많은 가중치를 줘야

한다는 생각은 [Hull97]에 나타난다. Hull 의 어의 중의성 해소 방법은 질의어의 대역어들을 하나의 창 (window) 에 넣은 다음 창 내의 대역어들은 OR 연산자로, 창 간에는 AND 연산자로 묶고, 사용자가 지정한 각 창의 가중치로부터 가중치에 기반한 불린 질의를 만드는 방식이었다. 이 방법을 쓰면 AND 연산자에 의한 효과는 공기 가중치 할당과 같지만 OR 연산자에 의해 AND 연산의 효력이 무색해질 수 있을 것이다. 그 이유는 하나의 창에 OR 로 묶인 대역어들은 유사 개념의 단어들이 아닌 경우가 많이 발생하기 때문이다. 특히 한국어 명사의 70% 이상을 차지하는 한자어의 경우 대부분이 한글 표기는 같지만 뜻이 다르므로 일본어명사로 변환될 때 전혀 뜻이 다른 한자어에 대응될 수 있다. 이것은 비단 한-일간 교차 언어 문서 검색에서의 문제만이 아니다. 결국 어의 중의성이 있는 단어의 대역어들을 OR 연산자로 묶게 되면 문서 간의 변별력이 약해지므로, AND 연산자에 의한 문서 변별력을 상쇄시키게 된다.

제안된 공기 가중치 할당 방법에서는 각 대역어 의미 조합에 대해서 별도의 유사도를 계산한다. 따라서 유사도를 계산하는 문서 내에 다른 대역어가 나타나더라도 그 대역어는 현재의 의미 조합에 없으므로 유사도 계산에서 고려되지 않는다. 즉 Hull 방법의 OR 연산자에 의한 상쇄 효과는 발생하지 않는다.

3.3 시소러스 개념 수렴 (TG : Thesaurus Grouping)

3.3.1 시소러스 개념 수렴에 대한 이해

사전에 기반하여 질의문을 검색 대상 문서의 언어로 변환할 때에는 필연적으로 어의 중의성을 해소해야 하는 문제가 발생한다. 질의문에서 추출된 질의어의 대역어가 둘 이상 나타날 경우 기존 사전 기반 방법에서는 최초 대역어로 변환시키는 방식을 취하고 있다. 질의어의 가능한 대역어들은 질의문 언어에서는 하나의 단어로 표현된다. 하지만 검색 대상 문서의 언어에서는 서로 다른 의미의 단어들로 나타나는 경우가 많다. 결국 최초 대역어로 변환하는 방법은 질의어의 의미와 전혀 다른 의미의 대역어로 변환될 위험이 있는 것이다. 이런 오번역이 교차 언어 문서 검색에서 성능 저하의 주요 요인의 하나이다. 결국 오번역을 최소화시키는 방법으로 질의어의 어의 중의성을 해소해야 한다.

정보 검색에서의 사용자는 용어 매칭을 통한 검색이 아니라 질의어들의 개념에 기반한 문서 검색을 원한다. 사용자가 작성하는 질의문은 정보에 대한 요구를 표현한다고 볼 수 있다. 그 요구가 넓고 좁은 차이는 있겠지만 원하는 하나의 개념으로 수렴하는 경우가 많을 것이다(개념 수렴). 질의문을 구성하는 각 단어의 개념들의 조합이 질의어들의 개념과 전혀 무관한 새로운 개념을 표현할 수도 있겠지만 - 이것은 질의문 작성시의 주위 상황을 고려하지 않는다면 사람도 이해하기 어려울 수 있다 - 정보 검색 요구자의 질의로는 적절치 않다. 앞의 경우를 배제하고, 질의문을 구성하는 각 질의어의 개념들의 공통된 하나 이상의 상위 개념을 찾을 수 있다면, 그 상위 개념들을 질의 의도를 개념적으로 넓게 표현한 질의로 볼 수 있을 것이다. 그리고 그 때의 각 질의어들의 개념이 어의 중의성이 해소된 올바른 대역어라고 간주할 수 있다. 예를 들어 질의문의 질의어들이 [지구,지각,변동]이고 가능한 의미 조합이 다음과 같다고 하자.

가. [지구 1(earth), 지각 1(perception), 변동]
나. [지구 1(earth), 지각 2(the crust), 변동]
다. [지구 2(endurance), 지각 1(perception), 변동]
라. [지구 2(endurance), 지각 2(the crust), 변동].

이 의미 조합들을 상위 개념으로 그룹화하면 지구 1 (earth) 과 지각 2 (the crust) 는 지구를 나타내는 상위 개념으로 그룹화될 것이다. 또 지구 2 (endurance) 와 지각 1 (perception) 은 경과의 상위 개념으로 지구 2,지각 1. 변동은 연동을 나타내는 상위 개념으로 각각 그룹화될 것이다. 결국 [지구,지각,변동]의 각 의미 조합들을 상위 개념으로 묶어보면 순서대로 다음과 같이 그룹화될 것이다.

가. (지구 1) (지각 1) (변동)
나. (지구 1,지각 2) (변동)
다. (지구 2,지각 1) (변동)
(지구 2,지각 1,변동)
라. (지구 2) (지각 2) (변동)

아래 3.3.2 절에서 설명할 카도가와 개념 체계에 따르면 위의 그룹들 중에서 [(지구 1,지각 2)(변동)]이 지구라는 가장 가까운 상위 개념에서 그룹화됨을 알 수 있고 사용자 질의문의 의도를 가장 잘 반영한다고 볼 수 있다. 즉 '지구'는 *earth* 로 '지각'은 *the crust* 로 의미를 할당하게 된다. 이러한 상위 개념으로의 그룹화 방법을 시소러스 개념 수렴이라 부른다.

3.3.2 카도가와 시소러스와 개념 수렴

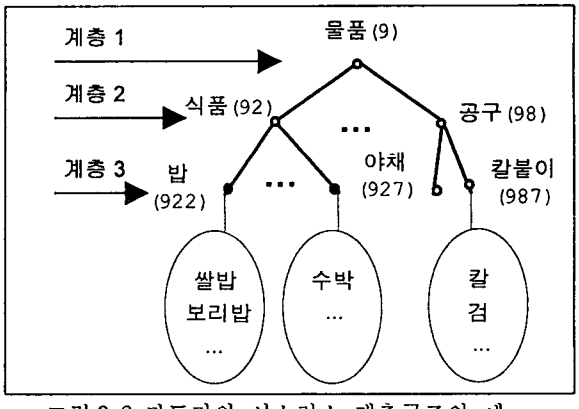

그림 3.2 카도가와 시소러스 계층구조의 예

카도가와 시소러스는 트리 구조 형태의 개념 체계로서 자연, 성상, 연동, 행동, 심정, 인물, 성향, 사회, 학회, 물품의 최상위 개념 10 개에서 출발하여, 각 노드마다 10 개씩의 하위 개념 노드들을 가지며, 계층 3 까지 개념이 구축되어 있고, 최하위 노드에는 해당 개념의 단어들이 나열되어 있다. 다 합하면 총 개념수가 1110 여개가 된다. 그림 3.2 에서 물품이라는 최상위 개념 (코드 9) 의 하위

노드로 90, 91, 92, ..., 99 가 있고, 그 중 식품 (92) 은 다시 하위 개념인 밥 (922) 을 포함한다. 최하위 개념노드 '밥'에는 그 개념에 속하는 단어들 ('쌀밥','보리밥'...) 이 나열되어 있음을 볼 수 있다. 위의 예에서 보리밥 (922) 과 쌀밥 (922) 의 공통된 상위 개념을 찾으면 세번째 계층인 개념 노드 '밥' (922) 에서 만나고, 쌀밥과 수박 (927) 의 경우 두번째 계층인 개념 노드 '식품' (92) 에서, 쌀밥과 칼 (987) 은 첫번째 계층인 개념 노드 '물품' (9) 에서 각각 상위 개념을 찾게 된다. 하위 개념 질의어들의 공통된 상위 개념이 발견된 경우 해당 계층에서 개념 수렴이 발생했다고 (그룹화되었다) 하며 이 때 질의어들이 나타내는 공통 개념들을 찾을 수 있게 된다.

3.3.3 시소러스 개념 수렴을 사용한 검색 모델

시소러스 개념 수렴은 사용자 질의문 벡터 SQ의 모든 가능한 대역어 의미 조합인 식 (1)의 임의의 원소 K_p - 검색 대상 문서의 언어로 변환된 하나의 질의 벡터 - 에 대해서 개념 수렴을 검사한다. 이 검사에 사용되는 척도에는 두 가지가 있다.(그림 3.1 참조)

1) L_{pi} : p 번째 의미 조합에서 i 번째 질의어가 속한 그룹 - 그림 3.1 에서 G2 에 해당 - 의 개념 수렴 계층을 뜻하며, 개념 수렴 계층이 낮을수록 즉 계층값이 클수록 하위 개념에서 그룹화되었을 것이므로 개념 수렴의 정확도가 크다.
2) $|g_{pi}|$: p 번째 의미 조합에서 i 번째 질의어가 속한 그룹의 구성 요소수를 의미하며, 그림 3.1 에서는 $|g_{pi}|=6$ 이 된다. 그룹 구성 요소수가 많을수록 즉 크게 그룹화 됐을수록 개념 수렴의 가능성이 크다.

위의 두가지 척도를 이용하면 다음 식 (5)과 같이 K_p의 각 질의어의 가중치를 정하는 함수를 얻을 수 있다.

$$w(t_{pi}) = \log(1 + L_{pi} \times |g_{pi}|) \quad \text{----------(5)}$$

where

t_{pi} : K_p의 i번째 질의어.

L_{pi} : g_{pi}의 시소러스 개념 수렴이 발생한 계층.

$|g_{pi}|$: g_{pi}의 구성 요소수.

g_{pi} : K_p의 i번째 질의어가 속한 그룹.

K_p : 질의어 벡터 SQ의 p번째 의미조합 $(1 \le p \le N)$

$N = \prod_{i=1}^{n} S_i$, (S_i:SQ_i의대역어개수)

$SQ = (SQ_1, SQ_1, \ldots, SQ_t)$, SQ_i: 질의문의i번째질의어

식 (5)으로부터 개념 수렴을 반영한 K_p의 각 질의어의 가중치를 얻고 나면 아래 식 (6)에서처럼 K_p에 대응하는 질의어 벡터 P를 생성하여 문서 벡터 D와의 유사도를 계산한다. 유사도 계산에서 앞에서 살펴본 공기 가중치 할당 방식과의 차이점은 번역된 질의문 벡터 P를 시소러스 개념 수렴을 통해 얻어지는 K_p의 각 질의어의 가중치로부터 생성한다는 것이다.

이제 식 (6)는 공기 가중치 할당과 시소러스 개념 수렴을 포함하는 검색 모델이 된다. 이 유사도에 따라 문서 등급화를 하면 사용자 질의문의 전체 개념을 가장 잘 반영하는 문서들을 검색 시 상위 등급에 위치시킬 수 있다. 이로써 오번역 대역어들이 검색 대상 문서에 공기할 때 발생할 수 있는 공기 가중치 할당 방식의 오류를 만회할 수 있게 된다. 본 논문에서는 각 단어에 카도가와 시소러스 의미 코드가 할당된 한-일 대역어 사전을 사용하여 시소러스 개념 수렴의 타당성을 검증해 보았으며 질의 변환 방식의 어의 중의성 해소에 효과가 있음을 알 수 있었다.

$$sim(P,D) = C^{M-1} \frac{\sum_{i=1}^{t} P_i D_i}{\sqrt{\sum_{i=1}^{t} P_i^2 \sum_{i=1}^{t} D_i^2}} \quad \text{----------(6)}$$

where

K_p : 질의문의 p번째 대역어 의미조합.

t_{pi} : K_p의 i번째 질의어

M : 벡터 P와 D의 일치하는 0이 아닌 성분의 개수

t : 인덱싱에 사용된 용어의 개수

C : 상수 (실험에서는 1.2로 잡았음)

$D = (D_1, D_2, \ldots D_t)$, D_i: i번째 질의어의 D에서의 가중치.

$$P = (P_1, P_2, \ldots P_t), \quad \begin{matrix} P_i = t_{pi} \text{ if } P_i \in K_p \\ P_i = 0 \text{ if } P_i \notin K_p \end{matrix}$$

4. 실 험

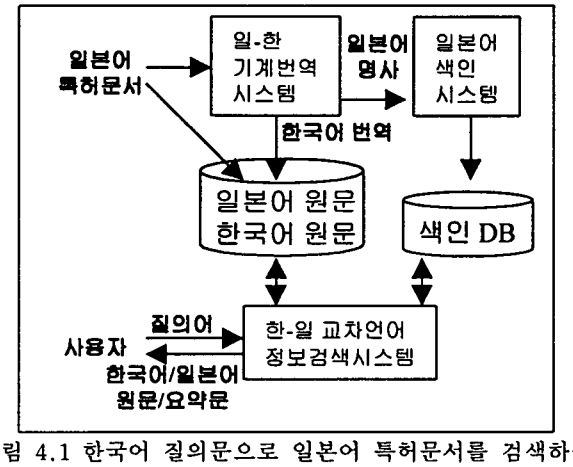

그림 4.1 한국어 질의문으로 일본어 특허문서를 검색하는 교차 언어 문서 검색 시스템

구현된 한-일 교차 언어 문서 검색 시스템(그림 4.1)은 질의어들의 나열을 입력으로 받아 문서와의 유사도에 따라 문서 등급화 목록을 출력한다. 자연 언어 질의를 입력할 수도 있으나 한국어 질의어 추출에 사용되는 형태소 사전과 한-일 대역어 사전의 표제어의 차이로 자연어 질의를 처리할 때 미등록어가 많이 발생한다. 따라서 실험을 위해 질의어들을 나열한 질의문을 입력한다. 일본어

문서 인덱싱에는 일본어 정보 검색에서 여러 장점을 보이는 단어 기반 인덱싱 (word-based indexing) 방법을 사용하였다 [Tokunaga96, Ogawa96]. 일본어 문서로부터 명사를 추출하기 위해서는 연구실에서 개발한 일-한 기계 번역 시스템 (COBALT-J/K : Collocation-based Language Translator from Japanese to Korean) 을 이용하였다. 한국어 질의어의 일본어 변환에 쓰이는 한-일 대역어 사전은 일-한 기계 번역 시스템의 일-한 대역어 사전을 수정하여 만들어졌으며 약 6600 여 표제어를 갖고 있다.

제안된 방법을 통하여 한국어 질의문에 대한 일본어 문서의 검색 성능을 평가하기 위해서는 최소한 일본어 평가 검증용 문서 집합이 필요하다. 그러나 현재 이용 가능한 일본어 문서 집합을 구하지 못한 관계로 포항제철의 지원으로 수행한 정보 검색 연구 과제에 사용되었던 일본어 특허 문서 2000 건을 대상으로 소규모 문서 집합을 만들어 실험에 이용하였다. 먼저 일본어 특허 문서를 일-한 기계 번역 시스템을 이용하여 한국어로 번역한 다음 문서 제목별로 문서를 분류하여 문서 제목을 질의어로 사용하고 동일 문서 제목을 갖는 문서들을 관련 문서들로 간주하였다.

이렇게 추출된 질의문 중에서 시소러스 개념 수렴에 성공한 질의문 241 개를 대상으로 전통적인 TF*IDF 검색 모델, 공기 가중치 할당 (Cooccurrence Weighting) 방법, 시소러스 개념 수렴 (Thesaurus Grouping) 방법들을 적용했을 경우, 각각에 대한 평가 결과는 그림 4.2 와 같다. 교차 언어 문서 검색이 실제 사용되어질 환경에서는 질의문을 작성하는 사용자는 검색 대상 언어에 대한 작문 능력뿐 아니라 문서 이해 능력도 그리 뛰어나지 않을 것이다. 즉 검색된 문서들을 하나씩 살펴보며 관련성 여부를 판단한다는 것은 무리가 있으며 문서의 내용을 검색하는 것은 기계 번역 시스템을 이용하는 경우가 많을 것이다. 결국 교차 언어 문서 검색에서는 상위 등급 문서들의 정확도가 재현율보다 더 중요하다. 따라서 본 실험에서는 질의어의 정확도 평가를 위해 3, 5, 10, 20 상위 등급 목록 지점들의 평균 정확도를 계산한다. TF*IDF 검색 모델을 기본 비교 대상으로 봤을 때 공기 가중치 할당 (Cooccurrence Weighting) 의 경우 4 %, 시소러스 개념 수렴 (Thesaurus Grouping) 의 경우 5 % 의 평균 정확도 성능 향상을 보였다. 또한 시소러스 개념 수렴 방법은 공기 가중치 할당 방식에 비해 1 % 의 성능 향상을 보였다.

각 방법의 평균 정확도의 차이를 질의문별로 분석해보면 정확도에 긍정적으로 (positive) 영향을 미친 질의문과 부정적으로 (negative) 하게 영향을 준 질의문들을 구별할 수 있다. 그렇게 얻어진 긍정 질의문과 부정 질의문의 개수를 비교해 보는 것이 실제 평균으로 나타난 정확도보다 더 정확한 성능 측정이 될 수 있다. 아래 그림 4.2 에서 살펴 보면 공기 가중치 할당 방식와 시소러스 개념 수렴 간의 정확도의 차이는 1 % (0.88 - 0.87) 불과하지만 시소러스 개념 수렴을 적용시킨 후 정확도가 바뀐 질의문은 모두 21 개이며 이 중 성능이 떨어진 질의문의 개수는 5 개, 향상된 질의문은 16 개이다. 즉 공기 가중치 할당 방식을 적용한 질의문의 76 % (16 * 100 / 21) 는 시소러스 개념 수렴에 의해 성능이 향상됨을 알 수 있다. 또 TF*IDF 를 적용한 질의문의 92 % 는 공기 가중치 할당에 의해, 88 % 는시소러스 개념 수렴에 의해 성능이 향상됨을 알 수 있다. 이러한 사실로부터 제안된 방법들에 의해 정확도의 증가를 가져온 질의문들은 많았으나 증가치가 적었으며 정확도 감소를 야기시킨 소수 질의문들의 감소의 폭은 컸음을 알 수 있다. 그림 4.2 에서 알 수 있듯이 본 실험의 정확도는 기존 교차 언어 문서 검색실험에서의 결과보다 상당히 높다. 이것은 일본에서 대부분의 명사는 뜻글자인 한자를 사용하기 때문이기도 하지만, 특허 문서의 특성상 문서 내에 사용되는 어휘들이 문서마다 서로 구별되므로 별도의 방법을 사용하지 않은 TF*IDF 에 의해서도 문서들이 서로 구별되어졌기 때문이다.

일반적으로 정보 검색에서 성능 저하의 원인은 (1) 사용자가 같은 문서에 대한 요구를 서로 다른 표현으로 나타낼 수 있다는 것과 (2) 인덱싱에 사용된 문서의 용어와 질의문의 질의어가 문자적으로 같더라도 사용된 의미가 서로 다를 수 있다는 사실에서 비롯되어진다. 전자는 동의 다형 표현의 문제 (synonymy) 로서 재현율을 감소시키는 원인이다. 후자는 동형 다의어 문제 (polysemy) 로서 정확도를 감소시킨다 [Deerwester90]. 한-일 교차 언어 문서 검색 환경에서 대역어인 일본어의 명사는 대부분 한자를 사용하므로 동형 다의어 문제는 거의 발생하지 않는다. 동의 다형 표현의 문제는 문서를 인덱싱할 때 카도가와 의미 코드별로 일본어 단어들을 모아 두었다가 질의어 확장 (Query Expansion) 기법을 사용한다면 어느 정도 해결될 수 있을 것이다. 한국어 질의어의 올바른 일본어 번역만 이루어진다면 앞에서처럼 단일 언어 정보 검색의 여러 기법들을 적용할 수 있으므로 동의 다형 표현이나 동형 다의어문제는 그렇게 중요하지 않게 된다. 결국 한-일 교차 언어 문서 검색에서의 가장 큰 어려움은 어의 중의성이 있는 질의어의 올바른 대역어 선정이며, 본 논문에서 제안된 방법은 그 의미 해소의 한 방법이 될 수 있음을 알 수 있었다. 그러나 실험에 사용된 문서 집합이 작았고, 특허 분야에 치우여 있어 실험의 객관성이 떨어지므로 향후 대규모 일본어 평가 검증용 문서 집합에 대해 공기 가중치 할당 방법과 시소러스 개념 수렴 방법들을 적용시켜 그 타당성을 확인해 볼 필요가 있을 것이다.

시소러스 개념 수렴에 성공한 질의문 개수 : 241

질의문 평균 단어수 : 8.24

질의문 단어의 평균 다의성 (Ambiguity) : 2.52

평균 정확도 (Average Precision) :

TF*IDF : 0.84

공기 가중치 할당 적용 : 0.87

시소러스 개념 수렴 적용 : 0.88

모델간 평균 정확도 차이에 영향을 준 질의문 개수

(긍정적 영향 / 부정적 영향)

TF*IDF -> 공기 가중치 할당 방법 = 2 / 24

공기 가중치 할당 / 시소러스 개념 수렴 = 5/16

TF*IDF -> 시소러스 개념 수렴 = 4 / 30

그림 4.2 실험 결과

5. 결 론

정보 검색 시스템의 사용자는 질의어의 문자적 매칭에 의한 문서 검색이 아니라 개념에 기반한 문서 검색을 원한다. 사용자의 질의문이 표현하는 전체 개념은 각 질의어들의 개념들로부터 유추할 수 있다. 이러한 전제에서 출발한 시소러스 개념 수렴은 질의문을 구성하는 단어의 수가 많은 경우 본 실험에서처럼 성능 향상을 얻을 수 있었다. 또한 공기하는 대역어가 나타나는 문서의 가중치를 증가시킴으로써 올바른 대역어들의 조합을 찾을 수 있음도 알아 보았다.

그러나 사용자의 질의문이 짧은 경우 시소러스 개념 수렴 방법은 수렴하는 상위 개념을 찾을 수 없는 경우가 많이 발생할 것이므로 사용하기 어려울 것이다. 이런 경우의 어의 중의성 해소를 위해 검색 대상 문서 집합에서의 대역어들의 분포 정보를 이용한 의미 해소 방법을 찾아내야 한다. 예를 들면 사전에 나타난 최초 대역어가 아니라 검색 대상 문서 집합 내에 가장 많이 나타나는 대역어로 변환하는 방법이 있을 수 있다. 또한 도메인 의존적 전문 용어들을 주기적으로 사전에 등록한다 하더라도 현재의 카도가와 시소러스 개념 체계에 최신 전문 용어들을 포함시키기엔 개념 체계가 너무 광범위하다. 향후 카도가와 개념 체계를 뼈대로 하여 도메인 의존적 말뭉치로부터 자동 학습에 의해 시소러스를 세분화시켜 나가는 방법에 대한 연구가 필요할 것이다.

[참고문헌]

[Ballesteros96] Ballesteros, L. and Croft, W.B., "Dictionary-Based methods for cross-lingual information retrieval," *In Proceedings of the 7th international DEXA Conference on Database and Expert Systems Applications*, pp.791-801, 1996.

[Ballesteros97] Ballesteros, Lisa, and W. Bruce Croft, "Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval," *In AAAI Symposium on Cross-Language Text and Speech Retrieval*, 1997.

[Berry95] Berry, M.W., Dumais, S.T., and O'Brian, G.W., "Using linear algebra for intelligent information retrieval," *SIAM Review*, 37(4), pp.573-595, 1995.

[Davis95] Mark Davis, and Ted Dunning, "A TREC evaluation of query translation methods for multi-lingual text retrieval," *In D. K. Harman, editor, The Fourth Text Retrieval Conference(TREC-4), NIST*, 1995

[Deerwester90] Deerwester, S., Dumais, S.T., Landauer, T.K., Furnas, G.W. and Harshman, R.A., "Indexing by latent semantic analysis," *Journal of the Society for Information Science*, 41(6), pp.391-407, 1990.

[Douglas96] Douglas W. Oard, and Bonnie J. Dorr, "A Survey of Multilingual Text Retrieval," *Technical Report UMIACS-TR-96-19, Institute for Advanced Computer Studies, University of Maryland*, http://www.ee.umd.edu/medlab/filter/papers/mlir.ps, 1996.

[Douglas97] Douglas W. Oard, "Alternative Approaches for Cross-Language Text Retrieval," *In AAAI Symposium on Cross-Language Text and Speech Retrieval*, 1997.

[Hull96] Hull, D.A. and G. Grefenstette, "A dictionary-based approach to multilingual information retrieval," *In Proc. of the 19th ACM SIGIR Conference*, pp.49-57, 1996.

[Hull97] Hull, D.A., "Using Structured Queries for Disambiguation in Cross-Language Information Retrieval," *In AAAI Symposium on Cross-Language Text and Speech Retrieval*, 1997.

[Ogawa96] Ogawa Yasushi, Kameda Masayuki and Matsuda Toru, "Inforium: A user-friendly document retrieval system," *Proceedings of the workshop on Information Retrieval with Oriental Languages*, pp.143-149, 1996.

[Sheridan96] Sheridan, P., and Ballerini, J.P., "Experiments in multilingual information retrieval using the spider system," In Proceedings of the 19th International Conference on Research and Development in Information Retrieval, pp.58-65, 1996.

[Susan97] Susan T. Dumais, Todd A. Letsche, Michael L. Littman, and Thomas K. Landauer, "Automatic Cross-Language Retrieval Using Latent Semantic Indexing," *In AAAI Symposium on Cross-Language Text and Speech Retrieval*, 1997.

[Tokunaga96] Tokunaga Takenobu, and Iwayama Makoto, "Word-based vs. Character-based indexing: An Experimental study on Japanese Text representation for text categorization," *Proceedings of the workshop on Information Retrieval with Oriental Languages*, pp.73-78, 1996.