교차언어 문서검색에서 중의성 해소를 위한 가중치 부여 및 질의어 구조화 방법

정의헌[0] 권오욱 이종혁
포항공과대학교 컴퓨터공학과
{mission, ohwoog, jhlee}@kle.postech.ac.kr

Weighting and Query Structuring Scheme for Disambiguation in CLTR

Eui-Heon Jeong[0] Oh-Woog Kwon Jong-Hyeok Lee
Dept. of Computer Science & Engineering, POSTECH

요 약

본 논문은 사전에 기반한 질의변환 교차언어 문서검색에서, 대역어 중의성 문제를 해결하기 위한, 질의어 가중치 부여 및 구조화 방법을 제안한다. 제안하는 방법의 질의 변환 과정은 다음의 세 단계로 이루어진다. 첫째, 대역어 클러스터링을 통해 먼저 질의어 단어의 적합한 의미를 결정짓고, 둘째, 문맥정보와 지역정보를 이용하여 후보 대역어들간의 상호관계를 분석하며, 셋째, 각 후보 대역어들을 연결하여, 후보 질의어를 만들고 각각에 가중치를 부여하여 weighted Boolean 질의어로 생성하게 된다. 이를 통해, 단순하고 경제적이지만, 높은 성능을 낼 수 있는 사전에 의한 질의변환 교차언어 문서검색 방법을 제시하고자 한다.

1. 서론

교차언어 문서검색 (CLTR: Cross Language Text Retrieval)은 사용자의 질의 언어와 문서의 언어가 서로 다른 상황에서 문서를 검색하는 방법에 대한 연구이다.

교차언어 문서검색에서는 질의의 언어와 검색 문서의 어를 서로 비교 가능한 형태로 일치시키는 언어 변화과정{Oard96}이 필요하며, 무엇을 번역하느냐에 따라서 질의어를 변환하는 방법과 문서를 변환하는 방법, 그리고 변환을 위해 어떤 자원을 이용하냐에 따라 기계번역기를 이용한 방법, 말뭉치를 이용한 방법, 그리고 사전을 이용한 방법이 있다. 이 중에서 문서변환은 그 비용이 많이 들고 또한 문서번역에 필요한 고품질의 기계번역기를 얻기 힘들기에 한계를 지니며, 질의 변환 방법에서도 병렬말뭉치(parallel corpus)나 비교 가능한 말뭉치(comparable corpus)의 경우도 그것을 구하기 어렵다는 단점이 있다.

그러기에 교차언어 문서검색의 연구는 대체로 질의 변환에 초점이 맞춰져 있으며, 본 논문에서는 질의 변환 방식 중에서도, 실용적이며 경제적인, 사전에 의한 접근 방식을 취하고 있다. 그런데 이런 사전에의 의한 질의 변환 방법은, 대역어 선택시 중의성 문제, 미등록어 문제, 구 번역 문제 등으로 인해 성능 저하를 보이고 있으며. 이러한 문제점들을 해결하기 위해 여러 연구가 진행되어 왔는데, 본 논문은 그 중에서도 대역어 중의성 문제 해결에 초점을 맞추고 이를 위해, 대역어 클러스터링을 통한 의미 추론과정, 지역정보와 문맥정보를 모두 이용한 질의어 가중치 방법, 그리고 가중치 부여 후 Boolean 질의어 생성과정, 세 단계 방법을 제시하며 이를 통해 한영 교차언어 문서검색에서 효과적으로 중의성 문제를 해소하고자 한다.

2 장에서는 사전에 의한 질의어 변환에 관한 기존 연구에 대해 알아보고 3 장에서는 대역어 사이의 지역정보와 문맥정보에 의한 관계 설정에 대해, 4 장에서는 질의어 가중치 부여 및 weighted Boolean 질의어 생성, 5 장에서는 대역어 클러스터링, 그리고 6 장에서는 이 세 단계를 합친 전체 생성과정과 예를 보여주고, 7장에서 결론을 맺는다.

2. 사전기반 방식의 기존 연구

대역어 사전에 기반한 질의 변환 방식은 최근 들어서 쉽게 구할 수 있는 전자화된 사전(MRD: Machine Readable Dictionary)의 등장으로 가능해졌다. 이는 다른 언어 자원들 보다 비교적 얻기 쉽고 그 방식이 간단하기 때문에, 질의 변환 방식에서 많이 이용되고 있다. {Hull96}에서는 이러한 사전에 의한 방법에서 성능 저하의 문제점이 무엇인지를 보이고 있는데, 그

원인은 다음과 같다.

첫째, 대역어 중의성 문제인데, 예를 들어, 한국어 '공기' 의 경우 영어의 'air', 'atmosphere', 'bowl', 'jackstone' 으로 번역될 수 있다. 즉, 대역어 중의성 문제란 원 질의어가 검색 문서의 언어로 변환될 때, 목표 언어의 가능한 대역어가 둘 이상일 경우 나타나는 어의 중의성 문제를 가리키며, 이때, 적합한 대역어를 선정하는 작업이 필요하다.

둘째. 구(phrase) 번역의 문제이다. 번역이 단순히 단어 대 단어로 이루어지지 않는 경우이다. 예를 들어, 한국어의 '중동 지역' 이 영어의 'Middle East' 의 경우를 말한다.

셋째, 사전에 나타나지 않는 미등록어 문제이다.

이렇게 사전기반 방식에는 크게 세가지 문제점들이 있는데 이를 해결하기 위한 기존 연구들을 살펴보면,

우선, 대역어 중의성 문제의 경우 우선, 목표언어의 문서에 나타나는 공기 정보를 이용하여 대역어들에 가중치를 부여함으로써 직접적으로, 중의성 문제를 해결하고자 하는 방법으로, [Jang99]와 [Lee2001]에서는 대역어 사이의 상호정보(MI: Mutaual Information)를 목표언어의 말뭉치로부터 검색 이전에 미리 구축하고, 이를 이용하여 대역어에 가중치를 부여하고 있고, [Kang99]에서는 연관 피드백을 통해 실시간으로 얻어진 공기정보를 이용하여 두 단계에 걸쳐 대역어들에 대해 가중치를 부여하고 있다. 위와 같은 방법은 단일언어 검색 시스템에 질의어를 입력하기 전에 질의어 단어들 마다 가중치를 부여함으로써 질의변환 과정에서 직접적으로 중의성 문제를 해소하고 적합한 대역어를 선정하는 방법들이다.

이와는 달리 [Hull97]에서는 질의어들을, 동일 원질의어의 대역어들은 OR로 묶고, 대역어 집합사이는 AND로 연결하여 이를, 벡터 모델이 아닌 weighted Boolean 모델에 적용함으로써 간접적으로 중의성 문제를 해결하는 것이다. 이는 Boolean 질의가 벡터 질의어보다 검색 시스템에 더 많은 정보를 제시할 수 있기 때문에 교차언어 검색 시스템에서도 좋은 효과를 얻을 수 있을 것이라고 예측할 수 있다.

또한 대역어 중의성 문제를 해결하기 위한 또 다른 방법으로는, 여러 대역어 중에서도 의미가 서로 확연히 구분됨으로써 검색 성능에 큰 악영향을 미칠 수 있는 동형이의어(homograph)를 제거함으로써, 적합한 대역어들을 선정하는 대역어 클러스터링 방법이 있다. [Sperer2000]에서는 중-영 교차언어 검색에서 이를 실험했으나, 성능 향상에 별로 도움이 되지 않았고 [Lee2001]의 한-영 교차언어 검색에서는 이러한 대역어 클러스터링 방법이 효과적임이 혀졌다. 이러한 차이는 중국어의 경우 의미를 규정짓는 단어가 한자이기 문에 실질적으로 대역어 중의성이 거의 발생하지 않는데 반해, 한국어에서는 명사의 상당부분이 한자의 음차로 사용되고 있기에 동형이의어로 인한 대역어 중의성이 많이 발생하기 때문이다.

이렇게 대역어 중의성 문제 해결을 위한 연구는 활발히 진행되고 있으나, 그 밖에 구 번역이나 미등록어 문제는, 구 번역 사전을 이용하거나 [Ball97] [Hull97], 사전확장 작업을 통해 사전의 적용범위를 넓히는 방법 외에 자동적으로 이를 해결하는 방법은 제시되지 않고 있다.

사전기반 방식의 문제점을 해결하여 성능을 높이고자 하는 방법들 이외에도, [Ball97]에서는 단일언어 검색에서 많이 이용되었던 연관 피드백을 통한 질의어 확장 방법이 교차언어 검색 환경에도 좋은 성능을 나타냄을 보이고 있다..

본 논문에서는 대역어 중의성 문제해결에 초점을 맞추고 지금까지 이루어져 왔던 방법들 중, 대역어 클러스터링 방법을 그대로 받아들이고, 공기정보를 이용하여 대역어에 가중치를 부여하는 방법과, weighted Boolean 모델에 적용하는 방법을 개선시켜, 성능 향상을 이루고자 하는 것이다.

3. 문맥 정보와 지역 정보를 이용한 대역어 사이의 관계 설정

3.1 문맥정보와 지역정보

기존의 교차언어 문서검색에서, 대역어들에 대해 가중치를 부여하는 방법들을 보면, 두 단어가 일정한 범위의 영역에서 얼마나 동시에 자주 나타나는가를 나타내는 공기정보를 바탕으로 대역어와 대역어 사이의 관계를 설정해주고 있다. 그런데 공기정보는 두 단어가 나타나는 영역의 범위에 따라 그 의미가 달라질 수 있다. 넓은 범위에 공기 하는 단어들에 의한 문맥정보(Topical Context)와 좁은 영역의 범위에서 공기하는, 하나의 구(phrase)나 관용적으로 함께 자주 쓰이는 단어들에 의한 지역정보(Local Context)가 있다. 문서 범위에서 자주 공기하는 단어는 그 두 단어가 같은 테마를 나타낸다고 할 수 있고, 한 문장과 같은 근접한 지역에서 자주 공기하는 단어는 같은 주제를 나타내기 보다는 관용적으로 많이 쓰이는 표현이라고 볼 수 있다.

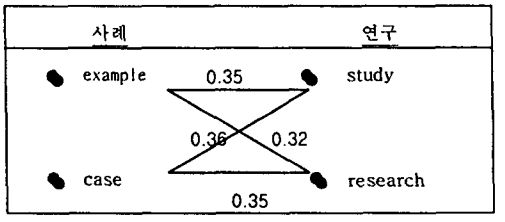
보통 의미 결정이 쉬운 단어들은 많은 문맥 정보 단어들로 둘러싸여 있으나, 의미 중의성 해소가 어려운 단어들은 결정적인 단서가 되는 정보가 매우 지역적인 경우가 많고, 주의의 다른 문맥 정보들은 오히려 잡음(noisy)이 되기 쉽다,[Leacock93]. 또 일반적으로 대상 단어에 대하여 지역적으로 가까운 단어들이 주는 정보가 원거리의 단어들이 주는 정보보다 더 신뢰성이 있다. 그러나 이러한 지역 정보들은 그 신뢰성은 매우 높지만 출현 빈도가 작아서 적용하는데 그 한계가 있어서, 이러한 점들을 고려하여 [Kim98]에서는 공기정보를 이용할 때 지역정보에 더 비중을 두고

지역정보가 없는 경우에 문맥정보에 따라 단어의 의미를 결정하고 있다.

3.2 공기정보를 이용하는 기존연구의 문제점

대역어 중의성 문제를 해소하려고 하는 교차언어 검색의 기존 방법들에서도 공기정보를 이용하고 있으나 그것을 이용할 때, 지역 정보와 문맥 정보의 이러한 차이점을 간과하고 있다. 즉, [Jang99] [Sperer200]에서는 의미 결정 지침으로, 지역 정보만을 이용하고 있으며, [Lee2001]에서는 문맥정보만을 이용하고 있다.

그러나 다음의 예를 보면, 교차언어 검색시 하나만을 이용한 방법들이 가진 문제점들을 알 수 있다.



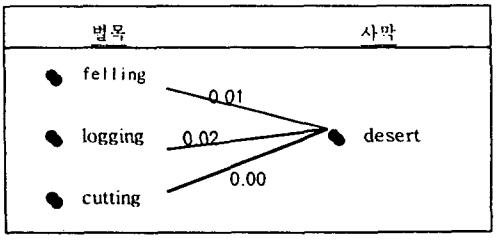
정확한 번역 : case study

부정확한 가중치 부여

그림1) 문맥 정보만을 이용

먼저 위 그림 1)에서 처럼 문맥 정보만으로 단어 사이의 의미를 결정하는 경우를 보자, 위에서 정확한 번역이 'case study' 라고 하면 'case' 와 'study' 간의 가중치가 다른 것보다 높아야 하지만, 'case research' 와 공기 정보 값이 거의 차이가 나지 않는다. 이는 문맥 정보는 같은 주제를 나타내는 단어 사이에서 그 값이 높게 나오고 'study' 와 'research' 는 서로 유의어이기 때문이다. 이렇게 문맥 정보만을 이용한 방법은 'case study' 처럼 마치 구(phrase)와 같이 쓰여 좁은 영역에서 자주 공기하는 두 단어의 상호 관계를 설정하는 지표로 쓰이기에는 부정확하다.

다음으로 지역정보만을 이용한 경우를 보자.



자료 부족 현상

그림 2) 지역 정보만 을 이용

위 그림 2)에서 '벌목' 에 대해, 3개의 대역어가 나오나 그 중 어느 하나가 'desert' 와 좁은 영역에서 더 많이 공기 한다고 할 수 없다. 왜냐하면 '벌목' 과 '사막화' 라는 두 단어는 같은 주제를 나타내는 단어이므로 같은 문서 안에 나타나는가를 보는 것이 더 정확하기 때문이다. 또, 위의 경우, 더 심각한 문제는 지역 정보의 경우 영역(window)이 좁아 자료 부족 현상이 나타날 수 있다는 것이다.

그러므로 대역어 사이의 상호 관계를 설정해 줄 때, 문맥정보만을, 또는 지역정보만을 이용하는 것은 올바른 방법이 아니며, 그것의 문제점을 해결하고 이를 개선할 필요가 있다.

3.3 개선된 상호 관계(Term relationship) 설정 방법

본 논문에서 제안하는 개선된 단어와 단어 사이의 올바른 상호 관계 설정 방법은 다음과 같다.

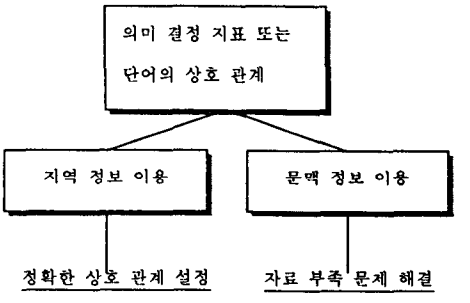


그림 3) 개선된 상호 관계 설정

즉, 공기 정보로써 대역 중의성 문제 해결에 더 신뢰할 만한 정보인 지역 정보를 이용하면서, 지역정보의 자료 부족 문제를 해결하기 위해, 문맥 정보도 함께 이용하는 것이다. 문맥 정보와 지역 정보를 동시에 사용하는 방법으로는, 지역 정보가 없거나 임계값을 넘지 못할 정도로 작을 경우, 이 때 문맥 정보를 이용하며, 지역 정보에 좀 더 비중을 두기 위해 문맥 정보의 값을 그대로 쓰지 않고 낮추어서 사용한다.

$$W_{edge} = \begin{cases} \alpha \times LC & \text{if } LC > threshold \\ (1-\alpha) \times TC & \text{if } LC < threshold \end{cases}$$

where $\alpha > 0.5$

LC : Local Context,
TC : Topical Context

수식1) 단어 사이의 상호관계 설정식

지역 정보와 문맥정보를 혼합해서 사용하는 방법에는 기존의 WSD와 관련된 연구에서 이를 신경망을 이용하여 최적화시키는 방법 등이 있으나 본 논문에서는 간단히 위와 같은 방법을 제시한다. 그리고, 일정 지역안에서 공기 정보를 구하는 방법으로는 [Jang99]

[Lee2001]에서 처럼 상호정보(MI: Mutual Information)를 사용한다.

이렇게 문맥 정보와 지역 정보를 함께 이용하면, 정확한 상호관계 설정으로 인해 신뢰할만한 가중치 부여가 이루어 질 수 있으며, 그로 인해 성능 향상이 이루어질 수 있을 것이라고 기대할 수 있다.

4. Weighted Boolean 질의어 생성

4장에서는 대역어 중의성 문제를 해결하기 위해 가중치를 부여하여 목표 언어의 질의어로 생성할 때, 각각의 대역어에 가중치를 부여하는 것이 아닌, 하나 이상의 여러 단어로 이루어진 번역후보 질의어에 가중치를 부여하고 weighted Boolean 질의어를 생성하는 것이, 단어간의 상호관계의 의미를 정확하게 반영하여,교차언어 검색에서 성능을 향상 시킬 수 있음을 설명하려고 한다.

4.1 기존의 대역어 가중치 부여 방법의 문제점

공기 정보를 통해 대역어 사이의 관계를 설정하고, 이를 이용하는 기존 방법에서는 다른 원 질의어에 나타나는 대역어들과의 공기정보 값을 지표로 원 질의어의 각각의 대역어 마다 가중치를 부여함으로써, 그것이 대역어로써의 원 질의어에 대한 충실도(fidelity)와 정확성을 나타낸다고 가정하고 있다. 즉, 각각의 연구마다 조금씩 차이는 있지만, 앞에서 보았던 예로 설명할 수 있다.

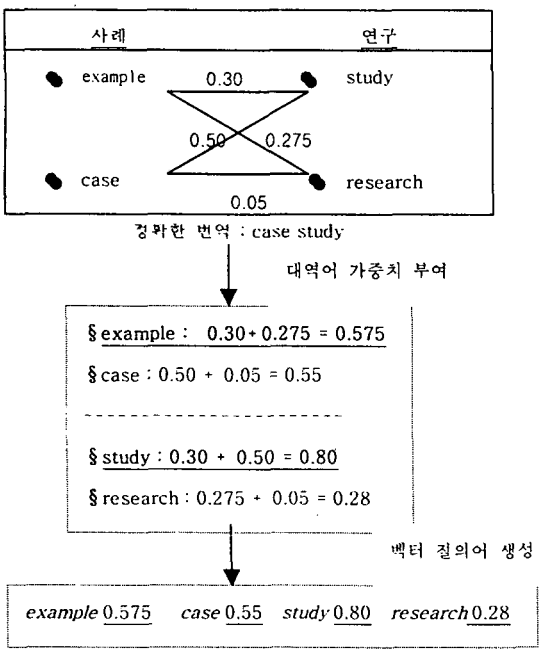


그림 4) 기존 연구에서 가중치 부여

위 그림 4)에서 각 대역어간의 상호관계가 위와 같이 설정되어 있고, 정확한 번역이 'case study' 라고 가정해보자. 각각의 대역어들은 다른 원 질의어의 대역어와의 상호관계의 값을 합하여 자신의 가중치를 갖게 된다. 그런데 여기서 example은 정확한 대역어가 아님에도 가장 높은 가중치를 갖게 되며 따라서 부정확한 번역으로 인해 검색 성능을 떨어뜨릴 수 있게 된다. 이것은 각각의 대역어 마다 독립된 가중치를 갖게 하는 기존연구의 두 가지 문제점에서 비롯된다.

첫째로는, 정확하지 않은 대역어와의 상호정보 값은 잡음(noise)이 된다는 것이다. 즉, 'example' 이라는 대역어에 주어진 0.575의 가중치 값은, 그것이 다른 원 질의어의 올바른 대역어와 자주 함께 나타난다는 의미가 아니라, 올바른 대역어이든, 그렇지 않은 대역어이든 간에, 다른 원 질의어의 가능한 모든 대역어들과 함께 자주 나타난다는 의미이다. 즉 여기서는 'example' 과 틀린 대역어인 'research' 와의 상호 관계가 잡음으로 작용해 부정확한 가중치 부여가 이루어 졌다.

둘째로, 벡터 질의어는 각각의 단어가 서로 독립이라고 가정한다. 그러나 벡터 질의어의 각 단어들에 가중치를 부여하는 과정을 보면, 가중치 부여의 지표로 단어들 간의 상호관계를 지표로 삼고 있는데, 이는 단어와 단어가 서로 독립적이라는 벡터 모델의 가정에 모순이 된다. 이러한 모순은 성능저하에도 영향을 미치게 된다. 왜냐하면, 각각의 대역어마다 독립적으로 가중치를 부여하게 되면, 그것이 다른 원 질의어의 어느 대역어와 얼마나 관계를 갖고 있었는지를 기억 할 수 없어서, 다른 원 질의어의 올바른 대역어와의 관계와, 올바르지 않은 대역어와의 관계를 구분해 낼 수가 없게 되기 때문이다. 그러므로 단어간의 상호관계에 대한 자료가 아무리 신뢰할 만하게 구축되어 있다 해도 이를 정확하게 반영함에 있어 벡터 질의어로는 한계를 갖게 된다.

4.2 Weighted Boolean 질의어의 생성.

본 논문에서는 위 4.1의 기존연구의 문제점을 해결하기 위해 weighted Boolean 질의어를 생성함으로써, 단어와 단어의 상호 관계 자료를 정확하게 반영할 수 있음을 나타내고자 한다.

교차언어 문서검색에서 원질의어를 목표언어로 질의 변환할 때 Boolean 질의어로 생성하는 방법에는 대표적으로 [Hull97]이 있었다. 그의 방법대로 그림4)의 질의를 변환해보면, 동일한 원질의어의 대역어 끼리는 OR로 묶고 다른 원질의어의 대역어 집합과는 AND로 묶어, 다음과 같은 변환된 질의어를 생성하게 된다.

(*example* OR *case*) AND (*study* OR *research*)

이러한 방법은 부정확한 대역어들끼리는 어차피 목표언어 문서에 함께 나오는 경우(즉, AND로 묶여진 경우)가 거의 없다는 가정하에 대역어들을 AND, OR의 Boolean 연산자로 묶어 줌으로써, 간단히, 간접적인 방법으로 중의성 문제를 해결하고 있다.

그러나 본 논문에서는 이러한 Boolean 질의어를 생성함으로써 간접적으로 간단히 중의성 문제를 해결하는 방법뿐만 아니라, 검색 이전에 이미 구축된 대역어들 간의 상호 관계 값을 이용하여, 각각의 단어가 아닌 후보 질의어에 가중치를 부여하는 직접적인 방법까지 병행함으로써 성능 향상에 기여할 수 있는 방법을 제시한다.

본 논문에서 제시하는 방법을 위의 예를 바탕으로 전개하면 다음과 같다.

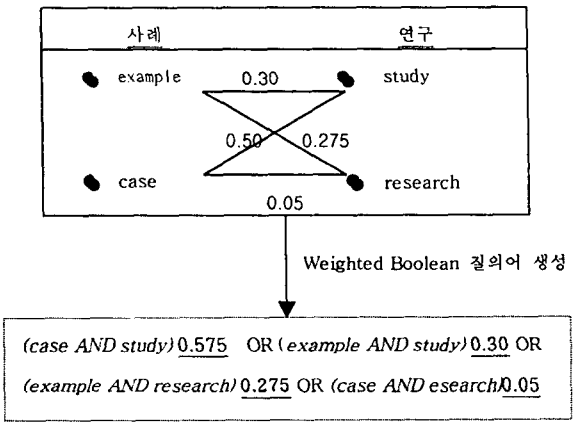


그림 5) Weighted Boolean 질의어 생성

위의 그림 5) 의 방법이 이전 방법과 구별되는 것은, 대역어들 간의 상호관계를 나타내는 그래프에서, 상호정보를 이용해서, 각각의 개별적인 대역어, 즉, 그래프의 각 노드 단위로 가중치를 계산함으로써, 각 대역어가 어느 대역어와 얼마나 상호 관련이 있는지에 대한 정보를 상실하는 것이 아니라, 이와는 달리 각각의 대역어와 대역어의 상호관계를 나타내는, 위에서 그래프의 모서리에 부여되는 가중치를 변환되는 질의어 구조에 그대로 반영한다는 것이다.

여기서 AND로 묶여진 단어의 집합들은, 각각의 변환되는 질의어 후보를 나타내며, 거기에 가중치를 부여한다는 것은 결국, 각각의 개별적인 대역어에 가중치를 부여하는 것이 아니라, 질의어 후보에 가중치를 부여함을 의미한다.

이러한 방법은, [Hull97]에서 제시했던, 교차언어 검색에 있어서, Boolean 질의어로 생성함으로써 얻을 수 있는 장점과, [Lee2000][Jang99]에서 대역어에 가중치를 부여함으로써, 얻을 수 있었던 장점들을 모두 취하게 되며, 대역어에 가중치를 부여했던 방법들의 문제점, 상호정보의 이용과 벡터 모델의 용어 독립 가정과의 모순에 따른 문제점을 해결하고 상호정보의 의미를 정확하게 반영함으로써 성능향상에 기여할 수 있게 된다.

그러나 이러한 방법은, 원 질의어의 길이가 길수록, 그 대역어 수에 따라, 후보 질의어의 수가 매우 많아지므로, 이를 처리해 줄 필요가 있다.

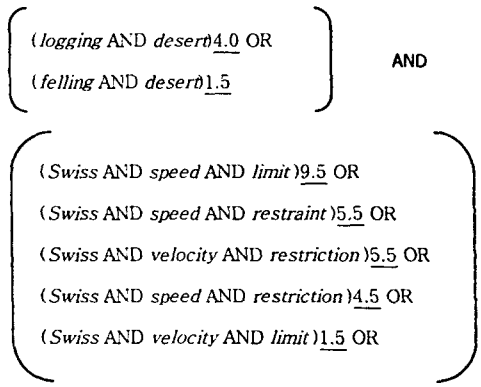


그림 6) 복잡도를 줄이기 위한 질의어구조화

위 그림 6) 은 원 질의어가 '벌목', '사막화', '스위스', '속도', '제한' 일 때 변환된 질의어 구조를 나타낸다. 후보질의어의 수가 많아짐으로써 복잡도가 커지는 것을 막기 위해서, 위에서처럼 '벌목 사막화' 와 '스위스 속도 제한' 사이에 분기점 을 두고 그것들을 AND 연산자로 묶어준다.

여기서 분기점을 정하는 기준은, 인접해 있는 원 질의어의 두 단어의 대역어들이 모두 상호정보가 약할 때, 즉, 관련 없는 두 단어가 인접하는 곳으로 정하게 된다. 그러나 세 단어가 지났는데도 이러한 경우가 생기지 않을 때는, 세 단어에서 묶어준다.

5. 대역어 클러스터링

[Pirkola98]는 사전을 번역해서 나오는 대역어들이 원 질의어의 개념과 다른 것이 거의 나타나지 않을 것이라고 가정하였지만, 실제로, 중의성 문제 해결에서 있어서, 원 질의어가 다의어가 아닌 동형이의어일 경우, 원 질의어와의 충실도가 매우 낮은, 다른 개념의 대역어가 변환된 질의어에 포함됨으로써, 교차언어 문서 검색의 성능을 크게 떨어뜨릴 수 있다.

이러한 문제점을 해결하기 위해 [Sperer2000]과 [Lee2001]에서는 원 질의어의 대역어들을 클러스터링하고 적합한 클러스터를 선택한 후, 원 질의어와 같은 개념을 나타내는 클러스터의 대역어들만을 변환된 질의어에 반영하고 있다.

클러스터링에서 단어들간의 유사도는 WordNet과 [Lin98]에서 제시한, WordNet 유사도를 구하는 방법

을 이용하였다. [Sperer2000]에서는 이러한 방법을 중영 교차언어 문서검색에 적용하였으나, 오히려 성능 감소를 가져왔고, 그 원인 클러스터에 가중치를 부여하여, 적합한 클러스터를 찾는 방법의 문제로 보고 있으나, 이는 무엇보다, 중국어에서 의미를 규정짓는 단어가 한자이기에, 실질적으로 대역어 중의성이 거의 나타나지 않기 문이다. 이는 [Lee2001]에서 한영 교차언어 검색에 적용한 결과, 성능이 향상되었다는 사실이 이를 말해준다. 즉, 한국어 명사의 상당 부분이 한자의 음차로 사용되기 문에 동형이의어로 인한 대역어 중의성이 많이 발생하기 때문에, 한국어의 경우 이러한 방법이 효율적임을 알 수 있다.

그러므로 본 논문에서는 이러한 장점을 받아들여 사용하고자 한다.

6. 질의 변환 과정 예

6.1 질의 변환 과정

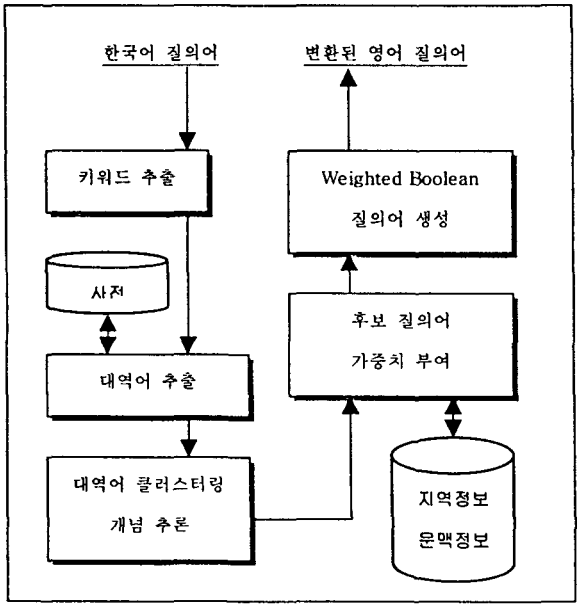


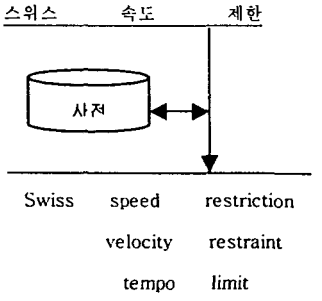
그림 7) 질의 변환 과정

본 논문에서 제시하는 질의 변환의 전체 과정은 그림 7)과 같다. 즉, 한국어 질의어가 들어오면 이중 명사만을 골라 키워드로 추출하고, 대역어 사전을 통해 대역어들을 추출하며, 5장에서 언급했던,대역어 클러스터링 방법으로 적합한 의미를 결정짓고, 그 다음으로 4장에서의 방법대로 지역정보와 문맥정보를 이용하여 후보 질의어들에 대해 가중치를 부여하며, 마지막으로 Weighted Boolean 질의어를 생성해 낸다.

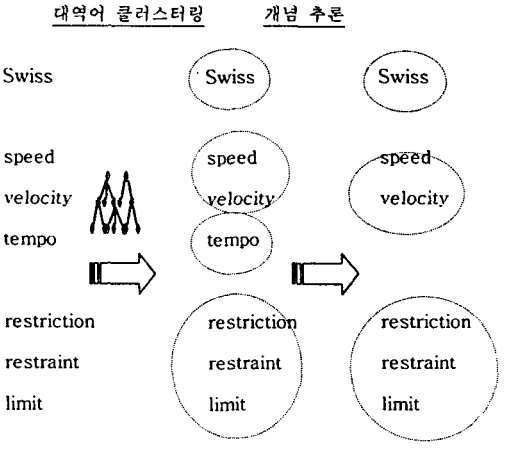
6.2 질의 변환 예

다음은 한국어 질의어로 '스위스 속도 제한' 이라는 질의어가 들어왔을 때 이를 영어 질의어로 변환하는 과정을 보여준다.

가. 사전으로부터 대역어들을 얻는다.

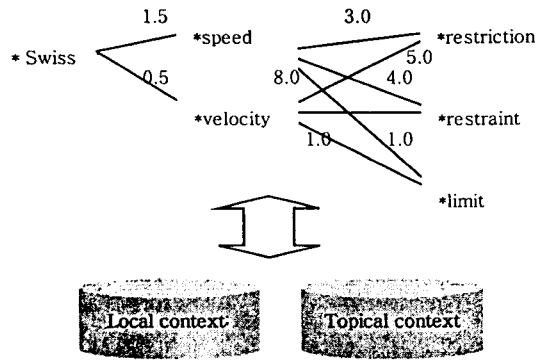


나. 대역어 클러스터링, 개념 추론



두 번째 단계에서, 동형이의어 문제를 해결하게 되는데 클러스터링시 단어간의 유사도 계산을 위해 WordNet을 사용하며 클러스터에 가중치를 주어 적합한 클러스터를 결정하기 위해서, 단어들간의 공기정보를 사용한다. 위에서 '속도' 에 대한 대역어로 'speed' ,' velocity' ,' tempo' 가 나왔으나, 그 중 'tempo' 라는 대역어가 원 질의어와의 충실도가 낮다고 판단되어 이 과정에서 제거된다.

다. 가중치 부여



대역어들간의 관계가 나타난 그래프에서 각각의 모서리에 지역정보와 문맥정보를 이용하여 가중치를 할당한다. 가중치를 할당하는 식은 다음과 같다.

$$W_{edge} = \begin{cases} \alpha \times LC & \text{if } LC > threshold \\ (1-\alpha) \times TC & \text{if } LC < threshold \end{cases}$$

where $\alpha > 0.5$

LC : Local Context, TC : Topical Context

위 식은 지역정보와 문맥정보를, 상호관계를 설정하는데 동시에 이용하나, 지역정보에 좀 더 비중을 둠을 나타낸다.

라. 후보 질의어 가중치 부여

위 다)에서 각 모서리에 가중치를 부여한 후, 복잡도를 해결하기 위해, 앞에서 언급했던 대로 분기점을 잡고 그 후 각 후보 질의어에 가중치를 부여한다.

Weight (Swiss speed limit)

= Weight(Swiss speed) + Weight(speed limit)

= 1.5 + 8.0 = 9.5

마. Weighted Boolean 질의어 생성

(*Swiss* AND *speed* AND *limit*)9.5 OR

(*Swiss* AND *speed* AND *restraint*)5.5 OR

(*Swiss* AND *velocity* AND *restriction*)5.5 OR

모든 후보 질의어에 대해 가중치를 계산하여, 그 값이 임계치를 넘지 못하는 후보 질의어들을 제거한 후 이를 위와 같이 AND, OR 의 Boolean 연산자로 묶어준다. 그리고 분기점으로 구분되는 지점에서는 그림 6) 에서 처럼 AND 연산자를 넣어준다.

7. 결론

본 논문에서는 사전에 기반한 질의변환 교차언어 문서 검색에서, 대역어 중의성 문제를 해결하기 위한 세단계 방법을 제시하였다. 이 중, 본 논문에서 처음으로 제안한 두 가지 중, 지역정보와 문맥정보를 함께 이용하는 것은, 정확하고 신뢰할 만한 상호정보를 구축하기 위한 것이고, 이를 weighted Boolean 질의어로 만드는 것은 그러한 대역어들간의 상호 관계를 변환된 질의어에 정확히 반영하기 위한 것이다.

이를 통해, 다른 자원에 비해 상대적으로 쉽게 얻을 수 있으면서도, 높은 성능을 낼 수 있는 교차언어 문서검색이 이루어질 수 있을 것으로 기대할 수 있다.

참고 문헌

[Ball97] Lisa Ballesteros and W.Bruce Croft, "Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval" In Proceedings of of the 19th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 97)

[Hull96] David A. Hull. Gregory Grefenstette, "Query Across Languages A Dictionary-Based Approach to Multilingual Information Retrieval, In Proceedings of the 19th International ACM SIGIR Conference on Research and Development in Inoformation Retrieval (SIGIR 96), pp 49-57, 1996

[Hull97] David A. Hull, "Using Structured Queries for Disambiguation in Cross-Language Information Retrieval," In Working Notes of AAAI Spring Symposium on Cross-Language Text and Speech Retrieval, Stanford, CA, pages 73-81, 1997

[Jang99] M.G.Jang, S.H.Myaeng and S.H.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, 1999

[Kim98] 김봉섭, "한-일 기계번역에서 의미 태깅된 말뭉치의 자동 생성 및 이를 이용한 명사의 의미 중의성 해소" 포항공대 전자컴퓨터공학부 (컴퓨터공학) 석사학위논문, 1998

[Kang99] In-Su Kang, Oh-Woog Kwon, Jong-Hyeok Lee, Geunbae Lee, "Cross-Language Text Retrieval by Query Translation Using ,Term Re-weighting" The 2nd International Conference on Multi Modal Interface(ICMI 99), pages IV.22-IV.27, 1999

[Leacock93] Claudia Leacock, Geoffrey Towell, and Ellen Voorhees, "Corpus-based Statistical Sense Resolution," in Proceedings of ARPA Human Languages Technology Workshop, pp.260-265, 1993

[Lee2001] 이문기. "대역어 클러스터링과 가중치 할당에 기반한 질의 변환 방식의 교차언어 문서검색" 포항공대 전자컴퓨터공학부(컴퓨터공학) 석사학위논문, 2001

[Lin98] DeKang Lin, " An Information-Theoretic Definition of Similarity" , Proceeding of International Conference on Machine Learning, Madison Wisconsin, Huly, 1998

[Oard96] Douglas W. Oard and Bonnie J.Dorr,, "A, Survey of Multilingual text Retrieval," Technical report, UMIACS-TR-96-19 CS-TR-3615,1996

[Pirkola98] Ari Pirkola, "The Effects of Query Structure and Dictionary Setups in Dictionary-Based Cross-language Information Retrieval" In Proceedings the 21th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 98), 1998

[Sperer2000] Ruth Sperer and Douglas W. Oard, "Structured Translation for Cross-Language Information Retrieval," , In Proceedings of the 23th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR2000), pages 120-127, 2000