# 과도한 지식을 요구하지 않는 공통기반축에 의한 용어 번역과 한영 교차정보검색에의 응용

# Knowledge-poor Term Translation using Common Base Axis with application to Korean-English Cross-Language Information Retrieval

최 용 석* 최 기 선*
(Yong-Seok Choi) (Key-Sun Choi)

**요 약** 교차언어 정보검색은 다국어 정보검색의 일부분으로 질의어에서 사용하는 언어와 검색대상인 문서의 언어가 서로 다른 경우의 정보검색을 의미한다. 교차언어 정보검색의 성능 향상을 위해서는 양질의 언어자원이 대량으로 필요한 경우가 많기 때문에 이를 해결하기 쉽지 않다.

본 논문에서는 사전에 기반한 대역어 후보 선정 시, 가중치를 부여해 질의어를 변환하는 방식을 제안한다. 가중치 계산에 이용되는 의미거리는 영어 명사와 한국어 명사를 같은 벡터 공간에 표현하고, 두 벡터간의 관계를 이용해 거리를 계산한다. 서로 다른 두 언어의 명사를 한 공간에 표현하기 위해 "공통 기반축"의 개념을 제시하고, 구축 방법을 제안한다. 고급 자원인 온톨로지를 확보하지 않고, 제안하는 방법으로 우수한 정보검색 결과를 얻을 수 있다는 것을 실험을 통해 보여준다.

**주제어** 질의어 변환, 공기 정보, 교차언어 정보검색, 자연언어처리

***Abstract*** Cross-Language Information Retrieval (CLIR) deals with the documents in various languages by one language query. A user who uses one language can retrieve the documents in another language through CLIR system. In CLIR, query translation method is known to be more efficient. For the better performance of query translation, we need more resources like dictionary, ontology, and parallel/comparable corpus but usually not available.

This paper proposes a new concept called the Common Base Axis which is adapted to Korean-English query translation and a new weighting method in dictionary based query translation. The essential idea is that we can express Korean and English word in one vector space by Common Base Axis and use it in calculating sense distance for query weighting. The experiments show that Common Base Axis gives us good performance without ontology and is especially good for one word query translation.

## 1. 머릿글

다국어 정보검색(Multilingual Information Retrieval)의 정의는 서로 다른 언어로 이루어진 정보들로부터 원하는 정보를 검색하는 것을 말한다. 사용자가 언어에 구애받지 않고 여러 언어의 문서를 검색해서 원하는 정보를 얻게 해주는 것이다. 예를 들어 한글로 검색하면 한국어 문서뿐만 아니라 일본어, 중국어, 영어 문서를 모두 사용자에게 제시해 줄 수 있도록 하는 것이다. 교차언어 정보검색(Cross-Language Information Retrieval)은 질의어에 사용한 언어와 다른 언어로 이루어진 문서를 검색하는 것을 말한다.

교차언어 정보검색 시스템을 구성하기 위한 효율적인

* 한국과학기술원 전산학과/전문용어언어공학 연구센터
Department of Computer Science
Korea Advanced Institute of Science and Technology/KORTERM
대전시 유성구 구성동 373-1, 우:305-701
373-1 Guseong-dong, Yuseong-gu, Daejeon, 305-701
전화: 042-869-5565
FAX: 042-867-3565
angelove, kschoi@world.kaist.ac.kr
연구분야: 자연언어처리, 정보검색, 한국어정보처리

기술은 질의어 변환이다. 사용자의 질의를 데이터에 맞는 언어로 바꾸어서 다국어로 이루어진 문서 집합내에서 검색하는 방법이다. 서로 다른 언어들을 다루기 위해 사용하는 자원을 살펴보면 사전과 같은 온톨로지를 이용하는 방법, 병렬 글모둠을 이용하는 방법, 은닉의미 색인 방법 등을 사용한다. 온톨로지 기반 방식은 적용할 수 있는 단어수가 너무 적고, 신조어에 대한 적응성이 떨어진다[12, 9]. 또한, 병렬 글모둠 기반 방식은 확보하기 어려운 자원을 필요로 하고 [7], 은닉의미 색인 방법 [8]은 특정 영역에서만 우수한 성능을 발휘하는 등의 약점도 있다.

언어자원의 기본인 사전을 쓰는 것이 질의어 변환에서 가장 안정적인 방법이다. 각 질의어에 필요한 대역어를 미리 사전에 보관하여 바로 변환하는 방법이다. 이 방법은 사전에 모든 단어를 집어넣기 어렵고, 한 단어에 대해 여러 대역어가 있을 경우 대역어를 선택하는 문제가 가장 어려운 문제이다.

천정훈 [3]은 사전기반 방식의 약점인 대역어 애매성 문제를 온톨로지를 사용해 극복했다. 사전에 나오는 여러 대역어 항목 중에서 온톨로지와 공기정보를 사용해서 최적의 대역어를 선택하는 방법이다. 이 방법으로 대역어 선택에 좋은 결과를 가져올 수 있으나, 온톨로지에 많은 단어가 들어가 있지 않다는 약점이 있다. 많은 비용을 들여 온톨로지를 확장시키더라도, 온톨로지의 여러 항목 중 선택해야 하는 사전 기반 방법과 비슷한 문제점이 발생한다. 공기정보 사용으로 성능을 향상시키기 위해 공기정보의 크기를 늘릴 수 있으나, 그럴 수록 데이터 부족 문제와 대규모의 데이터를 다루어야 하는 문제점이 발생한다. 이 방법의 경우 사전에 없는 신조어에 대해서는 해결 방안이 없다.

교차언어정보검색의 성능을 향상시키기 위해서 크게 세 가지 정도의 방법을 사용할 수 있다. 첫째는 질의어의 의미구분(sense disambiguation)을 하는 것이다. 의미구분을 통해 애매한 단어의 뜻을 명확하게 변환하는 것으로 좋은 성능을 보일 수 있으나, 활용 가능 정보가 제한적인 기존 방법으로 좋은 성능을 보이기 어려우며, 온톨로지를 사용하더라도 적용범위가 좁다. 둘째는 글모둠에서 공기정보를 통한 가중치를 부여하는 것으로 많은 경우 공기정보가 좋은 역할을 한다. 셋째는 질의어를 확장하는 방법이다. 질의어를 확장해서 부족한 정보를 추가할 수도 있으나, 부작용으로 인해 관련없는 정보 추가로 성능하락을 가져오는 경우도 있다.

본 논문에서는 확률벡터를 사용해서 공통 기반축(common base axis)을 생성하는 방안을 제안한다. 술어를 축으로 하는 공간에 명사들을 표현해서 언어의 변화에 관계없이 같은 공간에서 단어를 표현할 수 있도록 한다. 공통 기반축을 이용해서 "공기"와 "air"의 유사도를 측정할 수 있는 데, 이것은 술어와의 관계를 통한 확률적 의미구분 정보와, 술어의 창을 통해 공기정보를 제공해 주는 것이며, 그리고 확장질의에 나타나는 다양한 단어에 대한 가중치를 측정할 수 있게 한다. 질의어와 대역어의 신뢰도를 수치화해서 대역어 선택에 이용할 수 있도록 한다.

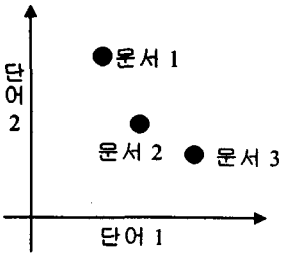
본 논문의 구성은 다음과 같다. 2장에서는 관련연구와 문제점에 관해서 다룬다. 3장에서는 공통 기반축 구축방법과 의미거리 계산방법에 대해서 다루고, 4장에서는 구축한 공통 기반축을 통한 실험에 대해 다룬다. 마지막으로 5장에서 결론과 향후 과제에 대해서 살펴본다.

## 2 관련 연구

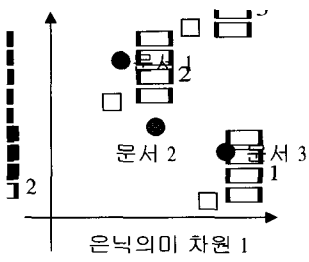
본 장에서는 교차언어 정보검색에서 제안된 방법들을 살펴본다. 교차언어 정보검색에서는 크게 언어변환으로 문제에 접근하는 방법과 언어를 변환하지 않고 원하는 문서를 검색하려는 방법이 있다. 2.1절에서는 언어변환을 하지 않는 은닉의미 색인(Latent Semantic Indexing) 방법에 대해서 다루고, 2.2절에서는 언어변환을 하는 연구에 대해서 다룬다. 2.3절에서는 본 논문에서 접근하려는 방향에 관해서 다룬다.

### 2.1 은닉의미 색인(Latent Semantic Indexing)

사람이 생각하는 언어구조로 이루어진 의미 구조를 정보검색의 기본으로 생각했었다. 하지만, 이 은닉의미 색인은 미지의 의미 구조가 있다고 생각하고 그 감춰진 의미로 색인을 하는 것이다. 따라서, 어떤 언어로 이루어진 문서라도 그것을 은닉의미 형태로 표현할 수 있다면, 같은 좌표공간에 색인해서 같은 방법으로 다룰 수 있다.




(그림 1) 기본 벡터 모델




(그림 2) 은닉의미 색인 모델

(그림 1)은 일반 벡터공간 모델의 형태를 보여준다. 벡터 공간 모델에서는 단어들이 차원이 되고 문서가 어떤 단어로 이루어졌는지 표현하게 된다. (그림 2)는 은닉의미 색인을 이용해서 새로운 차원에서 단어와 문서를 한꺼번에 표현하는 것이다.

문서와 단어를 한 공간에 표현하기 위해서 문서집합을 색인어와 문서의 행렬로 표현한다. 그 후 세 개의 직교하는 행렬로 표현한다[8]. 이 행렬은 특이값 분해(singular value decomposition) 방법을 통해 축소시켜진다. 이 과정에서 동의어나 유사한 문서들이 서로 묶여지면서 줄어든 행렬의 크기로 벡터공간을 설정해서 그 벡터공간에서 유사도를 비교하게 된다. 이 벡터공간에 다른 언어로 이루어진 문서도 같이 표현할 수 있다.

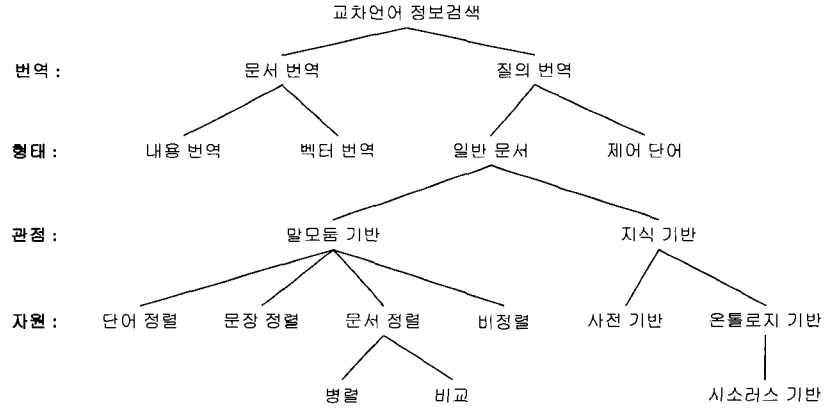
이 방법을 사용하여 벡터공간만 잘 만들어 준다면 어려운 번역과정을 거치지 않고도 사용자가 원하는 문서를 검색할 수 있다. 하지만, 이 방법을 사용하기 위해서 학습과정에 병렬 글모둠(parallel corpus)이 많이 필요한데, 이런 글모둠을 얻기가 쉽지 않다. 또한, 이 방식은 계산하는 비용이 굉장히 많이 들어가며 특정 영역에서는 잘 적용할 수 있으나 일반적인 영역에서 적용하기에는 적합치 않다.

### 2.2 언어 변환 방법

앞 절에서는 사람이 알 수 없는 은닉의미로 문서를 다루었다. 이 절에서는 사람이 이해할 수 있는 언어변환 방법을 통해 교차언어 정보검색을 하는 것을 다룬다. (그림 3)은 Oard[14]가 교차언어 정보검색에서 언어변환을 행하는 여러 방법들을 체계적으로 분류한 것이다.

먼저 번역 방법으로 분류해 보면 문서를 번역하는 것과 질의를 번역하는 것으로 나눌 수 있다. 질의를 번역하는 방법을 형태적으로 나누면 일반 문서를 다루는 것과 특정한 단어만을 사용해서 만든 문서를 다루는 방법이 있다. 일반 문서를 다루기 위해서 글모둠에 기반한 방법을 쓰는지 지식에 기반한 방법을 쓰는지에 대해서 나눌 수 있다. 글모둠에 기반한다고 하면 자원에 따라서 단어 정렬, 문장 정렬, 문서 정렬, 비정렬 글모둠을 사용하는 방법으로 나눌 수 있고, 지식기반에서는 사전을 이용하는 것과 단어들의 의미체계인 온톨로지를 이용하는 것으로 자원 구분을 할 수 있다.

사전과 병렬 글모둠을 이용하는 혼합 방법[6]은 사전의 성능에 의존적이고, 은닉의미 색인과 마찬가지로 대량 구축이 어려운 병렬 글모둠을 사용하는 단점이 있다. 다국어 온톨로지를 이용하는 유로 워드넷[12]은 독어, 이태리어, 스페인어, 영어 등 유럽 4개 언어의 단어들에 대해 서로간의 의미적 관계를 연결하는 중간언어 색인(interlingual index)을 만들어 개념기반의 문서검색을 행한다. 다국어 온톨로지를 사용함으로써 좋은 효과를 거둘 수 있으나, 다국어 온톨로지를 수동으로 구축하는데 오랜 기간이 소요된다.




(그림 3) 교차언어 정보검색 방법 분류

### 2.3 문제 정의

본 논문에서는 교차언어 정보검색 시스템에 필요한 한영 질의 변환 방법으로 가능한 대역어들에 가중치를 부여해서 중요도 순으로 변환하는 방법을 제시하려 한다.

온톨로지를 사용하는 것이 일반적으로 높은 성능을 보장해 주지만, 온톨로지와 같은 고급자원을 확보하기가 어려우며, 적용가능한 범위가 좁다. 사전 기반 방법에서는 많은 사전을 확보해 주는 것만으로 대역어를 생성시킬 수는 있으나, 사전에는 일반적인 단어들이 많이 들어가서 다른 의미의 대역어를 제시할 수 있다. 예를 들어 "자동차"라는 단어를 모든 사전에서 찾아내면 "autocar, automobile, car, motorcar" 등의 단어로 치환될 수 있다. 이는 질의를 확장하는 결과를 가져온다. 한 단어를 입력해서 질의 변환하는 과정에서 여러 단어로 바뀌면서 재현율을 높여 오히려 원 영어 질의어 보다 더 높은 성능을 보이는 경우도 있다. "(대기중의 )공기 "의 경우에는 4가지 뜻을 가지면서 사전에 나오는 단어들을 모두 사용하면 "atmosphere, bowel, empty vessel, jackstone, marble, peddle, air" 등의 단어를 질의로 사용한다. 전혀 다른 의미의 질의가 추가되면서 의미 없는 검색 결과를 얻게 된다.

천정훈[3]은 이 단어들에 온톨로지와 공기정보를 사용하여 가중치를 부여하는 모델을 제시하여 성능의 향상을 가져왔다. 하지만, 온톨로지와 공기정보가 충분치 않아 적용할 수 없는 질의가 있었다.

본 논문에서는 질의 변환에서 가중치를 부여하는 새로운 방안을 제시한다. 반자동으로 구축한 공통 기반축을 이용하여, 여러 가능한 대역어에 가중치를 부여할 수 있는 기반을 제공한다. 제안한 공통 기반축이 단어 의미 구분 정보와 공기정보, 확장 질의에 대한 신뢰도 정보 등을 포함하고 있어서 질의어 변환에서 좋은 성능을 보일 수 있는 기반 자원이 된다.

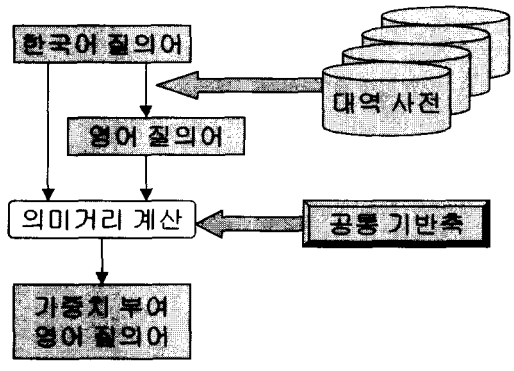
## 3 가중치 부여

본 장은 가중치 부여 모델을 제시하고, 그 모델에 필요한 요소 구축 방법을 설명한다.

### 3.1 공통 기반축에 의한 가중치 부여 모델

본 논문에서 제안하는 가중치 부여 모델은 그림 4와 같다. 입력받은 한국어 질의어를 다수의 대역 사전을 이용해서 모든 가능한 대역어로 영어 질의어를 구성한다. 각각의 질의에 나타난 한국어 단어와 영어 단어를 공통 기반축 위에 벡터로 표현한다. 사전으로 부족한 질의 단어를 보충하기 위하여, 확장 영어 질의를 사용한다. 한영 병렬 글모둠에서 먼저 한국어 질의어로 검색한 후에, 검색 결과 문서와 대응하는 영어 문서에서 단어들을 추출해서 확장 영어 질의어로 사용한다.

한국어 단어들과 영어 단어들 사이에 의미거리를 계산하고, 의미 거리가 가까운 대역어에 큰 가중치를 주어 최종적으로 가중치가 부여된 영어 질의어를 완성하는 한영 질의어 변환 모델이다.




(그림 4) 가중치 부여 모델

### 3.2 공통 기반축 구축

본 논문에서는 공통 기반축을 술어들로 구성한다. 이후에 영어와 한국어 명사들을 이 축 위에 표현한다. 기존 은닉의미 색인 방법에서는 알 수 없는 의미축에 단어와 문서들을 배열했으나, 제안하는 방법은 사람이 이해할 수 있는 술어를 기본 축으로 삼아서, 신조어가 많이 발생하는 명사들을 공통 공간에 벡터로 표현하려고 한다. 동시에 같은 공간에 표현된다고 생각하는 영어 술어들을 축으로 설정해서 영어 명사들도 벡터로 표현한다. 술어는 비교적 신조어가 적기 때문에 공간의 기준인 축으로 적합한 성질을 가지고 있다.

명사를 술어로 표현하겠다는 것은 명사의 속성은 기본적으로 술어로 표현할 수 있다는 것을 가정한다. 일반적으로 어떤 명사를 상대방에게 알아내도록 하는 방법 중의 하나로 "스무 고개"와 같은 방법을 사용한다. 즉, "먹는 것", "입는 것" 이러한 정보가 명사의 속성을 표현하며, 사람은 이러한 명사의 속성을 통하여 명사의미를 파악할 수 있다.

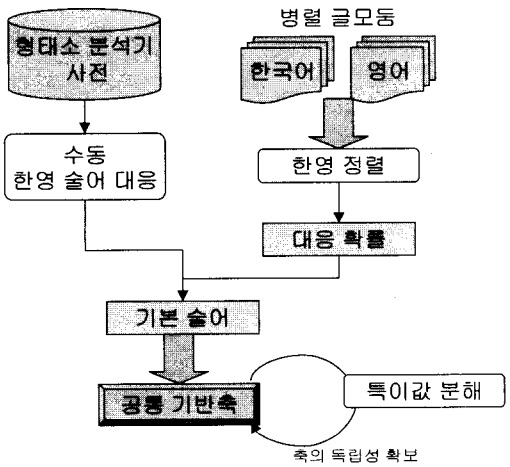
예를 들어 <표 1>과 같은 형태로 한국어 명사와 술어의 공기정보가 나타난다면, 대역어로 여겨지는 영어 명사와 술어들의 공기정보도 <표 2>와 같은 비슷한 형태

<표 1> 격틀에 들어간 단어의 공기빈도

| 술어 | 먹다 | | | 살다 | | 입다 | | | 뽀뽀하다 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 격 | 주격 | 처소격 | 목적격 | 주격 | 처소격 | 주격 | 처소격 | 목적격 | 주격 | 처소격 | 부사격 |
| **음식** | 0 | 0 | 153 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 3 |
| **도시** | 0 | 2 | 0 | 0 | 332 | 0 | 20 | 0 | 0 | 4 | 0 |
| **옷** | 5 | 0 | 10 | 2 | 10 | 3 | 0 | 200 | 0 | 0 | 8 |
| **영희** | 54 | 0 | 0 | 50 | 0 | 70 | 0 | 0 | 3 | 0 | 27 |

<표 2> 병렬 격틀 공기정보 구성

| predicate | eat | | | live | | wear | | | kiss | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| case | subject ive | location | object ive | subjecti ve | location | subjecti ve | location | object ive | subjecti ve | location | adverb |
| **food** | 0 | 0 | 153 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 3 |
| **city** | 0 | 2 | 0 | 0 | 332 | 0 | 20 | 0 | 0 | 4 | 0 |
| **cloth** | 5 | 0 | 10 | 2 | 10 | 3 | 0 | 200 | 0 | 0 | 8 |
| **Jane** | 54 | 0 | 0 | 50 | 0 | 70 | 0 | 0 | 3 | 0 | 27 |




(그림 5) 공통 기반축 구축 방법

의 공기정보로 나타나리라는 것을 가정한다. 그러나, 격까지 명확히 구분해내기 어렵기 때문에, 실제로는 격까지 구분하지 못하고 단순히 명사-술어의 공기 빈도만 조사하여 실험한다. 실험 결과 단순 공기 정보만을 가지고도 우수한 결과를 얻을 수 있음을 알 수 있었다.

공통 기반축을 구축하는 방법은 (그림 5)와 같다. 수동으로 한영 술어간의 대응 관계를 완성시키고, 병렬 글모둠으로부터 자동으로 대응 관계를 얻어낸다. 둘을 조합시켜서 기본 술어를 결정하고, 특이값 분해 방법으로 축의 독립성을 확보해서 공통 기반축을 구축한다.

### 3.2.1 기본술어 선정

술어를 기본축으로 설정하기 위해서는 기본술어 정의가 필요하다. 모든 술어들을 축으로 사용하면 보다 명확한 명사의 의미를 표현할 수도 있겠으나, 벡터의 크기가 너무 커지고, 자료 희귀문제(data sparseness)에 따른 부작용이 발생할 수 있다. 명사를 술어로 표현하기 위해서는 꼭 필요한 술어들을 파악해야 하며, 이를 기본술어라 한다. 술어들은 상태, 활동, 목적을 가진 사건의 세 부류로 나누어지는데, Dowty [5]는 이러한 모든 술어들은 상태를 나타내는 술어들과 3개의 양상 연산자 DO, BECOME, CAUSE들만 사용해서 의미속성을 표현할 수 있다고 했다. 따라서, 명확한 기본술어만의 축으로 모든 술어의 속성을 표현할 수 있다.

기본술어를 선정하기 위해서 형태소 분석기[1]의 동

사, 형용사 사전을 이용했다. 이 사전은 동사 3468개 형용사 1626개로 5097 용언류 단어가 들어있다. 형태소 분석기 사전에 들어있는 용언류가 기본적으로 필요한 단어라 할 수 있다. "하다"류의 동사를 제외하기 때문에 사전에 나오는 단어수와 많은 차이가 난다. 이 단어들에 부여한 중요도 점수[4]를 기준으로 기본술어를 선정한다.

선정한 한국어 기본술어 500개에 대하여 각각 가장 적합하다고 여겨지는 영어 술어를 다음과 같은 방법으로 대응시킨다. 이 때, 언어의 차이로 인하여 적합한 술어를 부여할 수 없는 경우 기본술어에서 제외시켰다. 품사별로 구분이 되었으나 형태적으로 같은 술어는 모두 하나로 다룬 후에, 사람이 가장 가깝다고 생각하는 의미의 영어와 1:1 대응시킨 한국어 기본술어는 총 455개로 한국어 명사와 영어 명사를 455차원 벡터로 같은 공간에 표현할 수 있다.

기본술어를 수동으로만 선정하면 "하다" 동사류가 제외되고 선택자의 의사에 따라 서로 다른 술어를 선택하게 된다. 특히, 1:1 대응만을 기본으로 하기 때문에 여러 의미를 가지는 술어를 하나의 술어로만 대응시키는 문제가 있다. 여러 의미에 대응시키는 방식을 위해 수동으로 여러 의미를 부여하면, 의미간의 중요성을 수동으로 차별해서 부여하기 어렵다. 이를 극복하기 위해서 병렬 글모둠에서 술어들에 대한 정렬[2]을 시도해 한영 술어를 대응시킨다.

단어간 대역 확률을 구하는데 있어서 대응되는 양쪽 언어 문장에 동시에 같이 자주 나타나면 대역 확률이 높다는 것이 기본 가정이다. 공기정보를 이용하여 단어간 유사도를 구하는 방법 중 많이 사용되는 것이 상호정보(mutual information)와 다이스(Dice) 계수[15]이다. 본 논문에서는 다이스 계수를 사용하여 단어간 대역 확률을 구한다.

영어 단어 $E_i$와 한국어 단어 $K_j$의 대역확률 $C_p(E_i, K_j)$는 다음 식과 같이 정의된다.

$$C_p(E_i, K_j) = \frac{2C(E_i, K_j)}{C(E_i) + C(K_j)}$$

$C(E_i)$와 $C(K_j)$는 각각 $E_i$와 $K_j$가 글모둠 내에서 나타난 빈도를 뜻하고, $C(E_i, K_j)$는 $E_i$와 $K_j$가 대응되는 문장에 동시에 나타난 빈도이다. 대역확률은 0부터 1사이의 값을 가지며, 0인 경우 두 단어가 동시에 출현한 경우가 없다는 의미이고, 1이면 두 단어가 모든 대응 문장에 동시에 출현한다는 것이다.

일단 정렬한 결과를 가지고 재추정과 복원에 쓰이는 대역 확률을 구하는 식은 다음과 같다.

$$C'_p(E_i, K_j) = \frac{2C_a(E_i, K_j)}{C_a(E_i) + C_a(K_j)}$$

$C_a(E_i)$와 $C_a(K_j)$는 각각 영어 단어 $E_i$와 한국어 단어 $K_j$가 실제 정렬에 쓰인 전체 횟수를 나타내고, $C_a(E_i, K_j)$는 $E_i$와 $K_j$가 실제로 정렬된 횟수를 나타낸다. 이렇게 수정한 대역 확률을 가지고 다시 대응되는 문장에 대해서 정렬을 수행하고 다시 그 결과를 가지고 새로운 대역확률을 구해서 대역 확률이 수렴할 때까지 반복한다.

1:1 대응의 불완전을 보충하는 자료로 정렬을 이용해 술어들의 다른 대역어를 구해 다대다(多對多) 대응을 가능하게 했다. 다대다로 출현하는 대역어를 모두 인정하지만 기본적으로 정렬은 임계값을 넘은 것만을 사용한다. 다대다 대응으로 인해 발생하는 오류는 측정하지 않는다.

수정한 대역확률은 기존 확률값(분자값)이 변함에 따라 계속 변한다. 확률이 변하면, 정렬이 다르게 되고, 정렬이 다르게 되면, 다시 확률값이 변한다. 이 과정을 반복해서 정렬 값을 수렴시킨다. 수렴시킨 값을 통하여 총 789개의 술어를 축으로 사용한다.

#### 3.2.2 축의 독립성 확보

술어를 축으로 사용한다면, 술어들은 서로 독립적이라는 틀 안에서 명사들간의 의미거리를 계산한 것이다. 그러나, 술어들은 서로 독립적이지 않고 서로간에 관련있는 의미요소들의 조합으로 이루어졌다. 예를 들어 다음과 같은 두 문장이 있을 경우 "팔다"라는 술어의 의미에는 "주다"라는 기본 의미요소는 포함되어 있고, 세부적 의미요소인 "돈같은 대가를 받았다."라는 의미까지 포함된 것이다[11].

"영희가 철수에게 시계를 주었다."
"영희가 철수에게 시계를 팔았다."

"팔다"와 "주다"는 서로 독립적인 술어가 아닌 것이다. 이럴 경우 두 술어 모두를 축으로 할 필요성이 있는가를 조사해야 한다. 축들이 서로 독립적이지 못하다면 정보의 중복뿐만 아니라, 명사간의 유사도를 계산할 때에도 서로 의존적인 정보들이 유사도 계산에 부정적인 영향을 끼친다. 본 논문에서는 이 단점을 극복하기 위해 기저벡터를 구하는 그램 슈미트 직교화(gram schmidt orthogonalization)방법을 사용하여 술어축들간의 독립성을 확보하는 방법을 제안한다. 실질적으로 컴퓨터에서

구현은 특이값 분해(singular value decomposition) 방법[10]을 사용하여 술어-명사간의 공기정보를 직교화 시켜, 기저벡터를 구하는 형식으로 이루어 진다.

기저벡터를 구한다면 서로 관련있는 의미요소들은 제거되고, 독립적인 의미요소들만 축으로 남게 된다. 기저벡터에 필요한 술어가 독립적 의미요소에 필요한 기본 술어라는 것을 알 수 있다. 기저벡터를 구하기 위해서 기본적으로 술어와 명사의 공기정보를 이용하여 구성한 공간을 전체공간으로 가정하고 계산했다. 기저벡터를 구하는 과정에서 사라지는 술어가 있다면, 그 술어의 의미요소들은 모두 다른 술어들에게 포함되어 있다는 것을 알 수 있다. 앞에서 선정한 기본술어들이 어느 정도 독립적인 축인가 조사해 본다.

기저벡터를 구하면 789개의 축에서 하나의 축이 줄어들고, 가중치가 적게 나오는 중요하지 않다고 판단할 수 있는 축이 10개 나온다. 그 줄어드는 축은 다음과 같은 술어와 가중치들로 이루어져 있다.

(나다: -0.03652, 나누다: -0.11569, 다루다: -0.120764, 매다: 0.036515, 미치다: -0.060631, 믿다: -0.144294, 싸다: -0.036515, 쓰다: -0.314025, 알리다: 0.49459, 얹다: 0.036515, 일으키다: 0.084494, 잃다: 0.060631, 짓다: 0.115686)

이 축의 의미를 인간이 파악할 수는 없다. 다른 기저축들은 더 복잡한 술어들의 조합으로 되어 있다.

이렇게 축이 거의 줄어들지 않는 것은 각 술어와 함께 출현하는 명사들이 의미 있는 수로 출현한다는 것이다. 가령 '시계'라는 명사와 같이 출현한 술어가 '주다'만 있고, '팔다'라는 술어와 같이 출현하지 않았다면 두 술어간의 관계를 공기 정보만으로 파악할 수 없다. 이를 파악하기 위해서는 더 큰 실험집합으로 명사 수를 충분히 확보하거나, 명사를 상위 개념어로 치환시켜서 공기정보를 조사하는 방법 등이 필요한데, 앞으로 독립성에 관해 더 연구해야 한다.

본 논문에서는 축이 하나 줄어든다는 것을 거의 모든 축이 의미있는 축[16]이라는 것으로 가정하고, 실험에서 기존 축을 그대로 사용했다.

### 3.3 벡터 표현 및 의미 거리 계산

명사들 간의 의미거리를 계산하기 위해서, 술어와 명사의 공기정보를 이용하여 명사의 벡터를 구성한다. 의미거리 계산에 두 가지 방법을 사용하는데, 하나는 확률벡터에 기반한 교차 엔트로피(Cross Entropy)이고, 다른 하나는 벡터 사이에 각을 사용하는 방법이다. 본 절에서는, 두 방법을 설명하고, 이어서 적용하는 가중치 기법에 대해 설명한다.

#### 3.3.1 확률 벡터 모델과 교차 엔트로피

확률 벡터(Probability Vector)는 다음과 같이 정의된다[17].

n차 벡터 $\vec{P}=(p_1, p_2, \cdots, p_n)$가 다음 식을 만족하면 n차 확률벡터이다.

$$\sum_{i=1}^{n} p_i = 1, \qquad p_i \geq 0, \quad i=1, 2, \cdots, n$$

질의어 변환 시에 질의 단어 하나를 확률벡터 $\vec{P}$로 본다면, 각 $p_i$는 그 단어의 $i$번째 술어와 공기빈도를 확률로 나타낸다.

확률벡터 $\vec{P}$에 대한 엔트로피를 다음과 같이 정의한다.

$$H(\vec{P}) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

확률벡터 $\vec{P}$의 각 요소(element) $p_i$들의 불확실성(uncertainty)을 $-\log_2 p_i$의 값으로 측정할 수 있다. 그러므로, 엔트로피는 확률벡터 $\vec{P}$의 정보 불확실성(information uncertainty)에 대한 기대값이다. $H(\vec{P})$는 모든 요소들의 확률이 같을 때 최대값을 가지며, 한 요소만이 1이고 나머지는 0일 때 $H(\vec{P})$는 최소값 0을 가진다.

확률벡터 $\vec{P_1}$와 $\vec{P_2}$가 같은 차원의 벡터일 때, $\lambda \in [0,1]$에 대해서 벡터 $\vec{P}=\lambda\vec{P_1}+(1-\lambda)\vec{P_2}$도 역시 확률벡터이다. 이 때, 확률벡터 $\vec{P}$를 $\vec{P_1}$와 $\vec{P_2}$의 복합 확률벡터(Composite Probability Vector)라고 한다. 그리고, $\vec{P_1}$와 $\vec{P_2}$를 $\vec{P}$의 구성 확률벡터(Components Probability Vector)라고 한다.

확률벡터 $\vec{P}=(p_1, p_2, \cdots, p_n)$와 $\vec{Q}=(q_1, q_2, \cdots, q_n)$가 주어졌을 때, 복합 확률벡터 $\frac{1}{2}\vec{P}+\frac{1}{2}\vec{Q}$와 이 벡터의 구성 확률벡터들 사이의 엔트로피의 차이를 교차 엔트로피(Cross Entropy)라고 하며 다음과 같이 정의한다.

$$\beta(\vec{P}, \vec{Q}) = H\left(\frac{1}{2}\vec{P}+\frac{1}{2}\vec{Q}\right) - \frac{1}{2}[H(\vec{P})+H(\vec{Q})]$$

이 때, $\beta$는 다음의 부등식을 항상 만족한다.

$$0 \leq \beta(\vec{P}, \vec{Q}) \leq 1$$

$\beta$의 값은 두 개의 확률벡터가 복합될 때, 불확실성의 증가 정도를 나타내고 있다. 만약 두 확률벡터가 관련되어 있으면 각각의 확률벡터 요소들의 확률 분포(probability distribution)가 유사하며 두 확률벡터가 관련이 많을수록 $\beta$의 값은 작아진다. 즉, 불확실성의 증가 정도가 적어진다. 이러한 $\beta$의 값은 두 확률벡터의 관계 차이를 나타낸다. 그러므로, $\beta$를 비관련도 계수(dissimilarity coefficient)로 해석할 수 있다. $\beta$를 단어간의 의미거리(concept distance) 값으로 사용한다[13].

(그림 6)은 확률벡터 모델에서 두 확률벡터간의 의미거리에 관한 예를 보인 것이다. 첫 번째 경우는 완전히 벡터가 일치할 의미 거리는 0이 된다. 두 번째 경우는 완전히 상이한 벡터인 경우로 이 때 의미거리는 1이 나온다. 세 번째는 비슷한 벡터를 비교하면 의미거리 값이 0에 가까이 나옴을 알 수 있다. 따라서, 확률벡터로 표현한 각 명사들간의 의미가 얼마나 가까운지 의미거리 값으로 측정할 수 있다.

### 3.3.2 벡터 모델과 코사인 값

확률 벡터 모델을 사용할 경우 음수의 벡터 요소값을 표현할 수 없다. 특이값 분해 방법을 사용할 경우 축이 음수 값으로 표현될 수 있기 때문에, 확률 벡터 모델을 적용할 수 없다. 두 단어를 표현하는 벡터의 의미 거리를 코사인 값을 사용해서 표현할 수 있다. 벡터 사이의 각이 작으면 의미적으로 가까운 곳에 있다는 측정치로 사용할 수 있다. 다음 식으로 의미거리를 표현한다.

$$\text{의미거리} = Dist(\vec{v}, \vec{w}) = -\cos\theta = -\frac{\vec{v}\cdot\vec{w}}{||v||||w||}$$

### 3.3.3 공통 기반축에 의한 가중치 계산

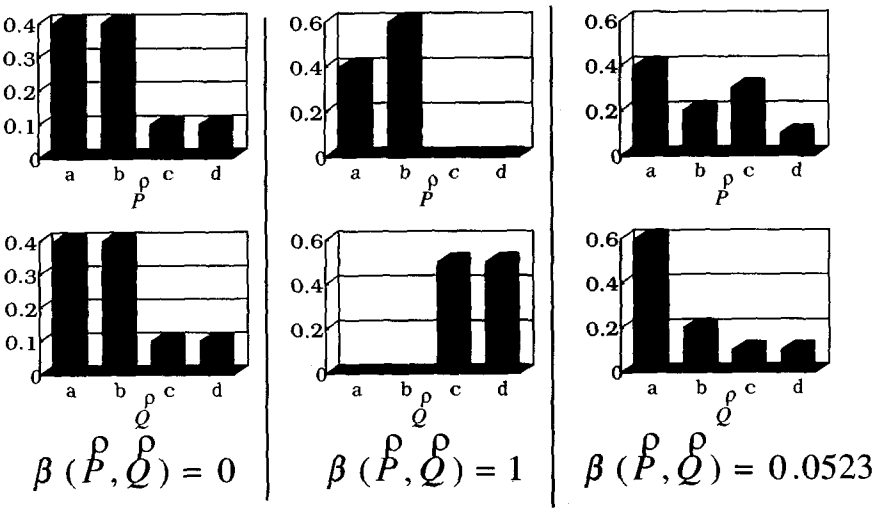
한영 질의어 변환에서 한국어 질의어 단어를 영어 단어로 변환한다. 본 논문에서는 각 영어 단어에 가중치를 부여한다. 가중치 부여는 두 가지 방법으로 이루어진다. 본 절에서는 "자동차 공기 오염"이라는 한국어 질의가 영어 질의로 변환되어 가중치가 부여되는 과정중 일부를 예로 사용한다.

첫 번째 방법은 영어 대역어와 한국어 질의와의 의미거리를 같은 비율로 합하여 가중치를 계산하는 방법으로 다음 식과 같다.

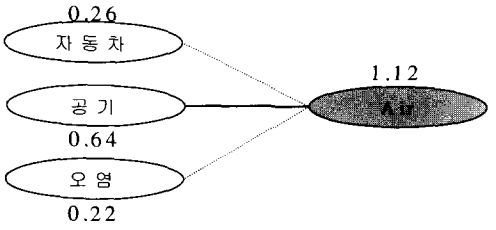
$$Weight_1(E_{ik}) = \sum_{j=1}^{||K||} Dist(K_j, E_{ik})$$

$K_j$는 한국어 질의에서 j번째 단어를 나타낸다. $||K||$는 한국어 질의 단어의 개수이다. $E_{ik}$는 i번째 한국어 질의어에 대한 영어 대역어 중에 k번째 대역어를 표현한다.

(그림 7)은 영단어 "air"와 한국어 질의어와의 의미거리 관계를 표현한 것이다. 사전을 기본 정보로 사용하면, "공기"와 "air"가 관계 있다는 정보만을 알 수 있다. 온톨로지와 같은 고급 의미정보와 공기정보를 확보해야, "자동차"와 "air"가 서로 관련이 있는지를 알 수 있었지만, 본 논문에서는 이를 단순히 공통축 위에 대응시켜서 서로간의 관계를 의미거리 수치로 파악할 수 있다. (그림 7)에 나오는 "air"의 가중치, $Weight_1(Air)$는 1.12로 결정한다.




(그림 6) 확률벡터간의 의미거리 예




(그림 7) "air"와 한국어 질의와의 관계

두 번째 방법은 다른 한국어 질의 단어의 대역어는 가중치(*Dist*)에 $\alpha$배 해주고, 영어 대역어들간의 가중치는 $\beta$배 해주어 합하여 사용하는 방법을 사용한다. $\alpha$와 $\beta$는 0.5와 0.2로 실험적으로 구하였다. 그 식은 다음과 같다. 위에 제시한 첫 번째 방법은 아래 식에서 $\alpha$와 $\beta$에 1과 0을 부여한 특수한 경우이다.

$$Weight_2(E_{ik}) = (1-\alpha)Dist(K_i, E_{ik}) + \alpha \sum_{j=1}^{||K||} Dist(K_j, E_{ik}) + \beta Con(E_{ik})$$

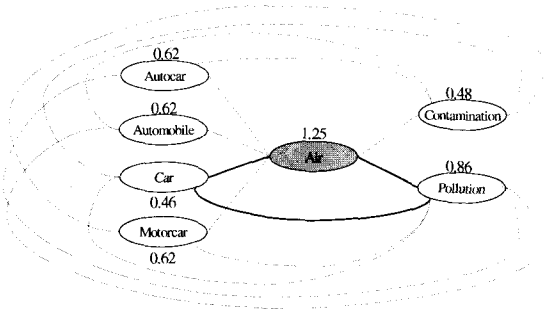
$$Con(E_{ik}) = \sum_{j}^{j \neq i} \sum_{t=1}^{||E_j||} \frac{Dist(E_{jt}, E_{ik})}{||E_j||}$$

$Con(E_{ik})$는 $E_{ik}$와 주변 다른 대역 영어단어들간의 의미거리를 이용해서 결정한 가중치 값이다. j번째 대역어들의 개수인 $||E_j||$로 나누어서 정규화 시켜준다. 주변 대역 단어들 $E_{jt}$은 i번째 질의어가 아닌 다른 한국어 질의어에 대한 대역 단어들을 표현한다.

(그림 8)은 "자동차 공기 오염"이라는 질의에서 출현하는 영단어 "air"와 주변 다른 단어들과의 의미거리를 계산하는 것을 표현해본 것이다. "atmosphere"와 같은 "공기"의 다른 대역어와는 계산하지 않고, 주변의 대역어들과 의미거리를 계산한다. 예에서는 "air"의 가중치 값, *Con(air)*가 1.25이다.

여기서 구한 가중치는 공통 기반축에 근거해서 계산되는데, 공통 기반축 위에서 표현되는 벡터는 술어의 속성으로 의미적 표현을 포함할 뿐만 아니라, 많이 공기하는 술어들을 포함하면서 일종의 공기정보 역할을 하게 된다. 천정훈은 거리 계산에 공기정보를 사용하면 좋은 결과를 얻을 수 있다는 것을 보여 주었다[3].

일반적으로 확장 질의는 재현율을 높여주며, 다국어 정보검색 성능향상에 도움을 준다. 본 논문에서는 천정훈 방법[3]처럼 한영 병렬 글모듬 집합에서 먼저 한국어 질의어로 검색한 후에, 검색 결과 문서와 대응하는 영어 문서에서 단어들을 추출해서 확장 영어 질의어로




(그림 8) 영단어 "air"와 다른 단어들간의 관계

사용한다.

확장 질의에 나오는 영어 단어에 대한 가중치는 마찬가지 방법으로 아래와 같은 식으로 계산한다. 확장 질의에 앞에서 사용한 영어 단어와 같은 단어가 있다면, 기존 가중치에 아래 식으로 구한 확장 가중치 값을 합해준다. 확장 질의에 나오는 영어 단어에 대해 원 한국어 질의와의 거리에 기존 대역 영어단어들과의 거리를 합해서 가중치를 구해준다. $\gamma$와 $\delta$는 0.3과 0.1로 실험적으로 구하였다.

$$EWeight(E_i) = \gamma \sum_{j=1}^{||K||} Dist(K_j, E_i) + \delta ECon(E_i)$$

$$ECon(E_i) = \sum_{j} \sum_{t=1}^{||E_j||} \frac{Dist(E_{jt}, E_i)}{||E_j||}$$

$ECon(E_i)$은 확장 영어 단어 $E_i$와 기존 대역 영어단어들간의 의미거리를 이용해서 결정한 가중치 값이다. j번째 대역어들의 개수인 $||E_j||$로 나누어서 정규화 시켜준다. $E_{jt}$은 j번째 한국어 질의어가 사전에 의해 대역된 영어 단어들 중 t번째 영어 단어를 의미한다.

실험에서는 각 단어의 가중치에 역문서 빈도값을 곱해보고, 길이로 나누어 정규화해서 비교해 본다.

## 4. 실험

가중치를 부여한 질의어를 이용해서 정보검색 결과를 실험한다. 실험 환경은 다음과 같다.

검색 문서 집합
TREC AP 뉴스 집합 90년
문서수: 84,306 (237M)
색인 단어 수: 129,262개
전체 문서: 17,112,648개 단어로 구성

한 문서당 평균 단어 수: 213개

한국어 질의어는 TREC-6 CLIR track의 24개 영어 Topic에서 title 필드에 나타나는 명사들을 번역하여 작성하였다. 실험에 사용된 검색 시스템은 스마트 시스템을 이용했다.

| | |
|---|---|
| 1. Waldheim Affair | 발트하임 사건 |
| 2. Marriages | 결혼 |
| 3. Drugs | 마약 |
| 4. Reusage of Garbage | 쓰레기 재활용 |
| 5. Acupuncture | 침술 |
| 6. Automobile air pollution | 자동차 공기 오염 |
| 7. Sex Education | 성교육 |
| 8. Swiss Speed Limits | 스위스 속도 제한 |
| 9. Effects of logging | 벌채 효과 |
| 10. Solar Powered Cars | 태양열 자동차 |
| 11. Organic Cotton | 유기 면화 |
| 12. Organic farming | 유기 농법 |
| 13. Middle-East Peace Process | 중동 평화 절차 |
| 14. International Terrorism | 국제 테러리즘 |
| 15. Death Penalty | 사형 |
| 16. Tuberculosis | 결핵 |
| 17. Potatoes | 감자 |
| 18. Perfume | 향수 |
| 19. Wine | 포도주 |
| 20. Effects of elephant protection on ivory trade | 코끼리 보호 상아 무역 영향 |
| 21. Child Abuse | 아동 학대 |
| 22. Effects of chocolate on health | 초코렛 건강 영향 |
| 23. Fast food in Europe | 유럽 패스트 푸드 |
| 24. Teddy bears | 테디 베어 |

24개의 한국어 질의어에 대해 단순 단어 대 단어 변환을 통해 생성된 대역어 후보들에 대한 통계는 다음과 같다.

번역된 24개 한국어 질의어의 단어 수: 47개
대역 사전에 의한 영어 대역어 수: 149개
평균 대역어수: 3.24

사용한 사전은 동아 한영 동아프라임 사전 (114,535 항목), 25개 분야 전문용어 사전 (159,979 항목), 고유명사 사전 (8,217 항목) 등이다. 24개의 한국어 질의어를 이루는 단어들 중 동아 프라임 사전에 의해 탐색된 단어는 43개이고, "패스트"와 "베어"가 전문용어 사전에 의해 탐색되었고, "발트하임"이 고유명사 사전에 의해 탐색되었다. "패스트"는 전혀 다른 뜻의 약자로 등록되어 있었다. 사전에 등록되지 않은 단어는 "푸드" 한 단어이다.

정렬을 통해 공통 기반축을 구성하기 위해서 사용한 한영 병렬 글모둠은 문서단위로 정렬된 조선일보 사설과 뉴스위크 기사를 이용하였다. 공기정보를 추출하기 위하여 과기원의 한국어 형태소 분석기와 코넥서 (Conexor) 영어 구문 분석기를 이용했다[1]. 병렬 글모둠에 대한 통계는 다음과 같다.

병렬 글모둠
문서 단위로 정렬된 조선일보 사설:
1,281쌍 (97년 - 99년) - 8M
문서 단위로 정렬된 뉴스위크:
814쌍 (95년, 99년) - 6M
한국어 색인 단어: 68,986개
한국어 전체 문서: 371,758개 단어으로 구성
한국어 한 문서당 평균 단어 수: 179개
영어 색인 단어: 30,519개
영어 전체 문서: 443,259개 단어로 구성
영어 한 문서당 평균 단어 수: 213개

<표 3> 검색 방법에 따른 11점 정확도 결과 비교

| 검색 방법 | 24개 질의어 (상대비율) | 역문서 빈도, 정규화 |
|---|---|---|
| 영어 원 질의 | 0.3176(100%) | |
| 모든 대역어 | 0.2155(67.85%) | |
| 공통축 엔트로피 가중치 | 0.2441(76.86%) | 0.2353(74.09%) |
| +대역어 관계 | 0.2431(76.54%) | 0.2309(72.70%) |
| +확장 질의 | 0.1321(41.59%) | 0.1349(42.47%) |
| +확장 대역어 관계 | 0.1176(37.03%) | 0.1194(37.59%) |
| 공통축 내적 가중치 | 0.2360(74.31%) | 0.2221(69.93%) |
| +대역어 관계 | 0.2323(73.14%) | 0.2293(72.20%) |
| +확장 질의 | 0.0997(31.39%) | 0.0961(30.26%) |
| +확장 대역어 관계 | 0.1002(31.55%) | 0.0976(30.73%) |

결과들은 11점 평균 정확도를 기본으로 비교했다. 11점 평균 정확도는 0부터 1사이의 재현율을 0.1 단위로 구분해서 11개의 점을 만들고 그 점에서의 정확도의 값들을 평균한 값이다. 정보검색의 성능 평가에 일반적으로 쓰이는 단위로 본 논문에서는 11점 정확도라 표기했다.

1) http://www.conexoroy.com/

비교의 기준치로 영어 원 질의를 사용하고, 가장 단순한 방법인 모든 대역어를 질의 변환의 방법으로 사용한 것을 기저치로 사용했다. 확장 질의는 병렬 글모둠을 사용해서 질의를 확장시킨 방법이다.

<표 3>은 가중치 변화에 따른 전체 11점 정확도를 나타내는 표이다. 가중치에 역문서 빈도값을 곱하고, 정규화를 시켜 준 값이 정규화를 하지 않은 것보다 같거나 약간 더 좋은 결과를 가지기 때문에, 정규화한 것만 고려한다. 가중치에서 역문서 빈도값을 사용하지 않은 것이 전체 결과에서는 더 높은 정확도를 보인다.

<표 4> "자동차 공기 오염" 11점 정확도 결과 비교

| 검색 방법 | "자동차 공기 오염" | 역문서 빈도, 정규화 |
|---|---|---|
| **영어 원 질의** | **0.6919(100%)** | |
| 모든 대역어 | 0.4724(68.28%) | |
| 공통축 엔트로피 가중치 | 0.5882(85.01%) | 0.5203(75.20%) |
| +대역어 관계 | 0.5870(84.84%) | 0.5073(73.32%) |
| +확장 질의 | 0.4749(68.64%) | 0.4163(60.17%) |
| +확장 대역어 관계 | 0.3980(57.52%) | 0.3123(45.14%) |
| 공통축 내적 가중치 | 0.5170(74.72%) | 0.4316(62.38%) |
| +대역어 관계 | 0.5413(78.23%) | 0.5964(86.20%) |
| +확장 질의 | 0.5206(75.24%) | 0.3829(55.34%) |
| +확장 대역어 관계 | 0.4640(67.06%) | 0.3195(46.18%) |

<표 4>는 "자동차 공기 오염" 질의에 대해 검색 방법에 따라 11점 정확도가 변화를 보여준다. 특히, "자동차 공기 오염"이라는 질의어의 대역어에 사용자가 의도하지 않은 의미를 가진 영단어가 들어가 있는 경우에, 더욱 좋은 결과를 보임을 알 수 있다. 엔트로피 가중치만을 이용한 것이 우수한 결과를 보인다. 내적 가중치와 대역어 관계를 사용해서 역문서 빈도값을 이용하는 것이 가장 좋은 결과를 보인다.

<표 5>는 천정훈의 결과[3]와 본 논문에서 제안한 방법을 비교한 것이다. 검색 시스템이 다른 관계로 직접 비교할 수는 없으나, 본 논문에서 제안한 가중치 부여 방법을 사용한 것이 상대적 성능이 높음을 알 수 있다.

<표 5> 천정훈 결과와 11점 정확도 상대 비교

| 검색 방법 | 24개 질의어 | "자동차 공기오염" |
|---|---|---|
| 천정훈 결과 | 0.2402/0.3517(68%) | 0.3182/0.4079(78%) |
| 가중치 부여 방법 | 0.2441/0.3176(76.86%) | 0.5964/0.6919(86.20%) |

한 단어로 이루어진 질의의 경우 주목할 만한 좋은 성능을 보인다. 원래 한 단어로 이루어진 질의는 별로 정보가 없으므로, 질의를 대역어로 바꾸는 것만으로 질의 확장의 효과로 인해 성능향상을 이룰 수 있으나 가중치 정보를 부여한 것이 좋은 성능을 보이고 있다. <표 6>은 한 단어 질의의 11점 정확도를 보여준다. 특히, "결혼" 질의의 경우 대역어를 모두 선택하면 성능이 떨어지지만 엔트로피 가중치를 적용할 경우 월등히 좋은 성능을 보인다. 대역어에 대한 엔트로피 가중치는 각각 "marriage (0.50404), union(0.48742), matrimony(0.00001)"이다.

<표 6> 한 단어 질의 11점 정확도 결과 비교

| 검색 방법 | "결혼" | "마약" |
|---|---|---|
| **영어 원 질의** | **0.1598(100%)** | **0.0139(100%)** |
| 모든 대역어 | 0.1258(78.72%) | 0.0223(160.43%) |
| 공통축 엔트로피 가중치 | 0.2627(164.39%) | 0.0151(108.63%) |
| 공통축 내적 가중치 | 0.1604(100.38%) | 0.0223(160.43%) |

## 5 맺음글

본 논문에서는 한국어-영어 질의어 변환을 위한 적응성을 가지는 공통 기반축 구축방법을 제안하였다. 벡터 모델의 축으로 사용할 기본 술어를 결정하는 과정, 기본 한국어 술어와 영어 술어의 수동 관계설정, 정렬을 통한 자동 관계설정, 기본 술어의 의미 중첩을 해소하기 위한 기저 벡터 추출 과정 등을 제시했다. 공통 기반축을 이용해 여러 대역어들에 가중치를 부여해 중요도가 높은 대역어를 선택할 수 있는 방법을 제시했다. 또한, 실험을 통하여 공통 기반축상에서 벡터를 표현한 후 의미거리를 계산해서 가중치를 부여하면 교차언어 정보검색의 성능을 향상시킬 수 있음을 보였다.

제안하는 방법을 사용하여 공통 기반축을 구축해서, 고급 자원인 온톨로지를 사용하지 않고도 교차언어 정보검색의 성능을 향상시킬 수 있다는 것을 보여주고 있다. 특히, 한 단어로 이루어진 질의의 경우 질의 확장 효과로 인해 높은 성능 향상이 가능하다. 성능 향상의 원인은 공통 기반축으로 표현하는 정보가 단어 의미구분 정보와 공기정보, 확장 질의에 대한 신뢰도 정보 등을 포함하기 때문이다.

공통 기반축을 사용하면 신조어에 대해서 빠르게 대응할 수 있으며, 대량의 병렬 글모둠을 확보하지 않고,

대량의 단일 글모듬만으로 공통 기반축 위에서 표현되는 벡터의 정보의 양과 질을 향상시킬 수 있을 것으로 기대한다.

향후 의미 중첩 해소를 통한 기본 술어 확보 과정에 대해 보다 상세한 연구가 필요하며, 단일 글모듬으로 정보를 확장시켜 검색결과를 살펴보는 일도 필요하다. 의미구분과 격틀 사용을 통한 표현의 명확화 연구와 병렬 글모듬을 이용한 영어 질의어 확장도 성능향상에 도움을 줄 것이다. 또한 벡터모델을 확장시켜 요소 술어들을 표현할 수 있는 방법과 축의 독립성 확보 방안을 연구할 수 있다.

## 참고문헌


[1] 이운재, 김선배, 김길연, 최기선 (1999), 모듈화된 형태소 분석기의 구현, 한글 및 한국어 정보처리 학술대회-형태소 분석기 및 품사태거 평가 워크숍, 123-136.

[2] 이주호 (1999), 자동 정렬을 통한 영한 복합어의 역어 추출, 한국과학기술원 전산학과 석사논문, 14-21.

[3] 천정훈, 최기선 (1999), 교차언어 문서검색에서 다국어 온톨로지에 기반한 한영 질의어 변환, 한글 및 한국어 정보처리 학술대회, 43-49.

[4] 최용석, 이운재, 최기선 (2000), 말모듬에서 동사분포 연구, 한글 및 한국어 정보처리 학술발표 논문집, 169-175.

[5] Chierchia, Gennaro and Sally McConnell-Genet (1993), Meaning and Grammar: An Introduction to Semantics, MIT Press, 350-360.

[6] Choi, Yong-Seok, Junghoon Chun, Key-Sun Choi (2000), A Study on Dynamic Threshold for Korean English Query Translation, The 3rd International Conference of Asian Digital Library.

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

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

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

[10] Forsythe, G.E., Malcolm, M.A., and Moler, C.B. Computer Methods for Mathematical Computations (Chapter 9: Least squares and the singular value decomposition). Englewood Cliffs, NJ: Prentice Hall, 1977.

[11] Genter, Dedre (1981), Verb Semantic Structures in Memory for Sentences: Evidence for Componential Representation, Cognitive Psychology 13, 56-83.

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

[13] Gray, R.M. (1990), Entropy and Information Theory, Springer-Verlag.

[14] Oard, Douglas W. (1997), Cross-Language Text Retrieval, SIGIR 97 Tutorial on Cross-Language Text Retrieval.

[15] Ohmori, K., J. Tsutsumi, and M. Nakanishi (1996), Building bilingual word dictionary based on statistical information, Proceedings of the Second Annual Meeting of The Association for Natural Language Processing, 49-52.

[16] Schütze, Hinrich (1992), Dimensions of Meaning, Proceedings of Supercomputing '92, 787-796

[17] Wong, S.K.M. and Y.Y. Yao (1987), A Statistical Similarity Measure, Proceeding of the Tenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 3-12.