한국어-영어/일본어-영어 교차언어정보검색에서 클러스터 분석을 통한 성능 향상

이 경 순†

요 약

본 논문에서는 교차언어정보검색에서 점진적 클러스터링을 통해서 모호성을 묵시적으로 해소하는 방법을 제안한다. 연구 목적은 질의 번역에서 모호성이 크게 증가된 상태에서 문서 클러스터가 문서 문맥 역할과 모호성 해소 역할을 하는지를 보고자 하는 것이다. 제안하는 방법은 한국어/일본어 질의를 사전을 이용하여 영어로 번역을 하고, 번역된 영어 질의에 대해서 벡터공간검색모델이나 확률검색모델에 의해서 문서를 검색한다. 검색된 문서의 순위대로 점진적 클러스터를 동적으로 생성하고, 이 클러스터 정보를 질의에 반영해서 문서의 순위를 다시 결정하는 것이다. TREC 테스트컬렉션을 이용한 실험에서 모호성 해소를 하지 않은 질의에 대해서, 제안한 방법은 한국어-영어 교차언어정보검색에서는 벡터공간검색모델에서 39.41%의 성능향상, 확률검색모델에서 36.79%의 성능향상을 보였다. 일-영 교차언어정보검색에서는 각각 17.89%와 30.46%의 성능향상을 보였다. 적합성 피드백 방법과의 비교에서는 모호성 해소를 하지 않은 경우 확률검색모델에서 12.30%의 성능향상을 보였다. 이를 통해, 클러스터 분석은 질의 모호성 해소에 도움을 주어서 검색성능 향상에 기여하였음을 알 수 있다.

Performance Improvement by Cluster Analysis in Korean-English and Japanese-English Cross-Language Information Retrieval

Kyung-Soon Lee†

ABSTRACT

This paper presents a method to implicitly resolve ambiguities using dynamic incremental clustering in Korean-to-English and Japanese-to-English cross-language information retrieval (CLIR). The main objective of this paper shows that document clusters can effectively resolve the ambiguities tremendously increased in translated queries as well as take into account the context of all the terms in a document. In the framework we propose, a query in Korean/Japanese is first translated into English by looking up bilingual dictionaries, then documents are retrieved for the translated query terms based on the vector space retrieval model or the probabilistic retrieval model. For the top-ranked retrieved documents, query-oriented document clusters are incrementally created and the weight of each retrieved document is re-calculated by using the clusters. In the experiment based on TREC test collection, our method achieved 39.41% and 36.79% improvement for translated queries without ambiguity resolution in Korean-to-English CLIR, and 17.89% and 30.46% improvements in Japanese-to-English CLIR, on the vector space retrieval and on the probabilistic retrieval, respectively. Our method achieved 12.30% improvements for all translation queries, compared with blind feedback in Korean-to-English CLIR. These results indicate that cluster analysis help to resolve ambiguity.

키워드 : 질의 모호성 해소(Implicit Ambiguity Resolution), 교차언어정보검색(Cross-Language Information Retrieval), 점진적 클러스터링(Incremental Clustering), 문서 문맥(Document Contex), 문서 재순위화(Document Re-rank)

1. 서 론

교차언어정보검색(Cross-Language Information Retrieval)은 사용자가 질의와 다른 언어로 쓰여진 문서를 검색하는 것이 가능하도록 하는 정보검색이다. 교차언어정보검색을 위한 접근 방법은 통계적 기법과 번역 기법으로 나눌 수 있다. 통계적 기법은 언어 번역을 하지 않고 교차언어간의 연관관계 정보를 만드는데, 이 방법에서는 대량의 양국어 코퍼스가 필요하다[6]. 번역 기법은 질의어의 언어를 문서의 언어로 번역하는 질의번역 방법이나, 문서의 언어를 질의어 언어로 번역하는 문서번역 방법을 통해서 질의와 문서의 언어를 같은 언어로 만들고 검색을 수행한다. 고품질의 기계번역기가 이용가능한 경우에는 문서번역기법에 의한 교차언어정보검색[12, 14]이 가능하지만, 대량의 문서집합에 대해 검색하는 경우에는 모든 문서를 번역해야하기 때문에 실용적이지는 않다. 질의번역방법[10, 8, 7, 11, 1]은 양국어 사전, 다국어 온톨로지 또는 시소러스를 이용해서 번역을 하는데, 대부분의 연구에서는 양국어 사전이나 다국어

† 정 회 원 : 전북대학교 전자정보공학부 교수
논문접수 : 2003년 6월 30일, 심사완료 : 2004년 2월 27일

사전이 이용가능한 경우 단순하고 실용적이기 때문에, 사전을 이용한 질의번역방법(dictionary-based query translation method)을 채택해오고 있다. 그런데, 사전기반 질의번역을 통한 교차언어정보검색에서는 높은 성능을 얻기 위해서는 질의 번역에서 증가되는 어휘들의 모호성을 해결해야하는 문제가 있는데, 질의 번역 모호성을 해결하기 위해, 어휘들의 공기 발생 통계에 기반한 상호정보 방법이 이용되고 있다[3, 11]. 높은 값을 갖는 번역 어휘들쌍이 번역어로 선택된다. 공기정보는 구 번역[18]에서 좋은 효과를 보이기도 했다. 상호정보는 최적의 어휘를 선택하는 것 뿐 아니라 어휘에 가중치를 부여하는데 이용되기도 했다[11]. 적합성 피드백을 통한 자동 질의 확장은 질의에 나타난 단어의 수가 작은 경우에 특히 효과적이라고 알려져 있다. 연구[3]는 교차언어정보검색에서 적합성 피드백을 이용하였는데, 공기정보에 기반한 모호성 해소 후에 국부문맥분석[20]을 통한 번역후 확장이 효과적이라는 결과를 보였다.

문서 클러스터링은 정보검색 결과의 브라우징[9]이나 주제 탐색 등 다양하게 응용되고 있는데, 클러스터 자체는 어휘와 문서사이의 관계를 나타낸다. 연구[13]에서는 정보검색에서 벡터공간모델에 문서클러스터분석을 통해서 문서의 순위를 결정하였다. 이때, 클러스터는 문서의 문맥으로 이용되어 성능향상에 기여했다.

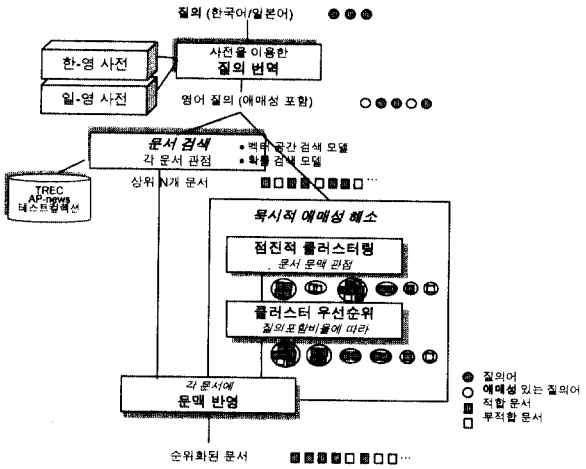
교차언어정보검색의 사전을 이용한 질의번역에서 발생하는 모호성의 정도는 단일언어 정보검색에서의 다의어에 의한 모호성의 정도보다 훨씬 더 크다.

본 연구에서는 교차언어정보검색에서 발생하는 질의 모호성을 해결하기 위해 클러스터 분석을 적용하는 방법을 제안한다. 제안하는 방법은 한국어-영어 및 일본어-영어 교차언어정보검색에서 번역된 질의에 대해 검색된 문서에 대해 점진적 클러스터링을 통해 동적으로 클러스터를 생성하고, 그 클러스터를 질의에 대한 각 문서에 대한 문맥으로 이용하여 각 문서에 대해서 순위를 다시 결정한다. 실험을 통해서 제안 방법이 질의번역에서 모호성이 많이 증가된 상태에서도 효과적임을 보인다.

본 논문의 2장에서는 제안하는 방법의 시스템 구조를 보이고, 3장에서는 묵시적인 모호성 해소방법을 설명한다. 4장에서는 TREC 교차언어 테스트컬렉션에 대한 실험을 통한 결과 분석을 하고, 5장에서 본 연구의 결론을 맺는다.

2. 클러스터 분석을 통한 모호성 해소 시스템 구조

(그림 1)은 클러스터 분석을 통해 질의 번역에서 발생한 모호성을 잠정적으로 해소하는 시스템의 전체 구조를 나타낸다. 시스템은 한국어나 일본어 질의에 대해서 한-영 사전 또는 일-영 사전을 참조해서 영어로 번역한다. 번역된 질의에 대해서 벡터공간검색기법이나 확률검색기법에 의해 문서를 검색한다. 검색된 상위 문서에 대해서 점진적인 클러스터링 기법으로 문서 클러스터를 동적으로 생성하고, 클러스터에 질의 포함 비율을 반영함으로써 우선 순위를 부여한 후, 각 문서의 문맥으로 이용함으로써, 검색된 문서의 순위를 다시 결정한다.



(그림 1) 클러스터 분석을 통한 모호성 해소 시스템 구조

제안 모델을 한국어-영어 교차언어와 일본어-영어 교차언어에 대해 적용한 것은 언어에 상관없이 적용될 수 있음을 보고자 하는 것이고, 정보검색모델로 벡터공간검색기법과 확률검색기법에 대해서 적용한 것은 검색모델에 상관없이 적용될 수 있음을 보고자 하는 것이다. 점진적 클러스터링과 묵시적 모호성 해소 모듈은 본 연구의 핵심 부분으로, 3장에서 자세히 설명하기로 한다.

2.1 양국어 사전을 이용한 질의 번역과 모호성

질의는 한국어 또는 일본어로 되어 있는데, 형태소분석과 품사 태깅을 거쳐서 키워드를 선택한다. 각 키워드에 대해서 한-영 사전 또는 일-영 사전을 참조해서 사전에 있는 영어 대역 어휘를 모두 선택한다. 한-영 질의 번역을 위해서 이용한 사전은 일반 사전과 전문 사전인데, 전체 282,511 한국어 용어와 505,003 영어 대역어를 포함하고 있다. 일-영 질의 번역을 위해서 이용한 사전은 공개된 일/영 사전인 EDICT[4]으로, 전체 161,806 일본어 용어와 283,177 영어 대역어를 포함하고 있다.

한 어휘는 여러 번역 어휘를 가질 수 있기 때문에, 번역된 영어 질의에는 동의어 뿐 아니라 서로 다른 의미의 어휘들도 포함할 수 있다. 동의어 또는 관련어일 경우에는 검색효율을 향상시키지만, 다른 의미를 갖는 어휘는 검색성능을 상당히 떨어뜨리게 된다. 이 단계에서 상호정보[5]에 기반한 통계정보를 이용해서 모호성을 해소하는 방법을 적용할 수 있다. 본 논문의 실험에서 질의 번역 단계에서 모호성을 해소한 경우와 그렇지 않은 경우 모두에 대해서 비교한다.

2.2 문서 검색

번역된 영어 질의에 대해서 벡터공간검색기법 또는 확률검색기법으로 문서를 검색한다.

2.2.1 벡터공간검색모델에 기반한 문서 검색

벡터공간검색모델은 질의에 있는 어휘가 있는지 없는지를 단순히 검사해서, 질의와 문서의 유사도를 계산하는데, 문서와 질의 벡터는 $D=\langle w_{d1}, w_{d2}, \cdots, w_{dt}\rangle$, $Q=\langle w_{q1}, w_{q2}, \cdots, w_{qt}\rangle$과 같이 표현된다. 어휘에 대한 가중치계산기법은 SMART 시스템 가중치계산기법 중에서[13, 7] 연구에서 가장 좋은 성능을 보인 것으로, 질의에 나타난 어휘 가중치 w_{qi}는 $tf\times idf$로 계산한 후 정규화 하였고, 문서에 나타난 어휘 가중치 w_{di}는 $(ln(tf)+1)\times idf$로 계산하였다. 질의-문서 유사도는 벡터 내적 곱으로 계산한다.

$$sim(q, d)=\sum_{i=1}^{t} w_{qi}\cdot w_{di} \tag{1}$$

2.2.2 확률검색모델에 기반한 문서 검색

확률검색모델에서 어떤 문서가 어떤 질의에 적합하다고 판단될 확률은 어휘들이 적합 문서와 부적합문서에서 다른 분포를 갖는다는 가정을 기반으로 한 것이다. 본 연구에서는 확률검색모델로 Okapi BM25식을 이용하였는데, Robertson-Sparck Jones 가중치계산기법[16]을 이용하고 있다.

$$sim(q, d)=\sum_{i\in Q} w^{(1)}\frac{(k_1+1)\,tf_{di}}{k_1((1-b)+b\cdot dl/avdl)\,tf_{di}}\frac{(k_3+1)\,tf_{qi}}{k_3+tf_{qi}} \tag{2}$$

여기서, Q는 질의이고, 어휘 i를 포함한다. $w^{(1)}$은 Q에 속하는 어휘 i에 대한 Robertson-Sparck Jones 가중치로, 적합 정보는 없고(R = r = 0) 역문서빈도수로 가중치가 감소된다.

$$w^{(1)}=\log\frac{(r+0.5)/(R-r+0.5)}{(n-r+0.5)/(N-n-R+r+0.5)} \tag{3}$$

여기서, N은 전체문서집합에 있는 문서 개수, n은 그 어휘를 포함하고 있는 문서 개수이다. R은 질의에 대해 적합하다고 알려진 문서 개수를 나타내고, r은 그 어휘를 포함하고 있는 문서 개수를 나타낸다. 파라미터 k_1, b, k_3는 질의와 문서집합의 특성에 의존하는데, 기본값은 $k_1=1.2, b=0.75, k_3=7$이다. tf_{di}는 문서 d에서의 어휘 i의 빈도수이고, tf_{qi}는 질의 q에서의 어휘 i의 빈도수이다. dl은 문서의 길이, $avdl$은 평균문서길이를 나타낸다.

번역된 영어 질의에는 엉뚱한 어휘가 포함될 수 있기 때문에, 문서 검색 결과에는 질의에 대해 부적합한 문서가 적합한 문서보다 높은 우선 순위로 검색될 수도 있다.

2.3 모호성 해소와 문맥 반영을 위한 점진적 클러스터링

부적합한 문서가 높은 우선 순위로 검색되는 것을 막기 위해서, 문서검색결과 상위 N개의 문서에 대해서 점진적 클러스터링(incremental clustering)을 하고, 질의와 클러스터의 유사도를 계산해서, 각 문서가 속한 클러스터 유사도를 문맥으로 반영함으로써, 문서의 순위를 다시 매긴다. 기본적인 아이디어는 다음과 같다 : 검색된 문서에 대해 생성된 클러스터는 그 클러스터에 속하는 각 문서들에 문맥 정보로 볼 수 있다. 질의와 클러스터의 유사도를 계산함으로써, 질의와 관련된 클러스터를 찾는다. 질의와 높은 유사도를 갖는 클러스터에 속하는 문서들은 질의에 보다 적합한 문서일 것이다.

정적으로 전역 클러스터링(static global clustering)을 하는 것은 문서집합이 대량인 경우에는 클러스터를 생성하는 시간이 실용적이지 못하다. 또한 생성된 클러스터를 활용하는 측면에서도 검색된 문서가 속하는 비율이 너무 빈약하기 때문에 문맥정보로 이용하기가 곤란하다(정적 클러스터링과 동적 클러스터링에 관한 비교 연구[2] 참조).

3장에서 점진적 클러스터링 방법과 클러스터를 각 문서의 문맥으로 반영하는 방법에 대해서 자세히 설명한다.

3. 점진적 클러스터링에 의한 잠정적 질의 모호성 해소

질의 번역에서 증폭된 모호성을 잠정적으로 해소하는 부분은 질의포함비율에 따라 클러스터에 중요도를 부여하고, 그 클러스터 정보를 각 문서의 문맥으로 반영하는 것이다.

3.1 점진적 클러스터링 방법

검색된 상위 문서에 대해서 동적으로 클러스터를 생성한다. 질의 중심적으로 클러스터를 생성하도록 하기 위해 질의-문서 유사도가 높은 문서를 우선적으로 처리하도록 다음과 같이 점진적 클러스터링을 하였다.

- 입력 : 문서검색 결과인 상위 문서 N개에 대해서 순서대로 처리한다 $(d_1, d_2, \cdots, d_n)$. 각 문서는 어휘의 가중치 벡터로 표현되고, 이는 문서 검색에서의 가중치와 같다.
- 단계 1 : 검색 순위가 1인 문서는, 그 문서 하나로 된 클러스터를 형성한다 $(d_1\in C_1)$.
- 단계 2 : 검색 순위가 2부터 n까지의 문서 $(d_2, \cdots, d_n)$에 대해서 단계 3에서 단계 5의 과정을 반복한다. 현재 처리할 문서를 di라고 하자.
- 단계 3 : 이미 생성된 k개의 클러스터 $(C_1, C_2, \cdots, C_k)$에 대해서, 클러스터 중심 벡터 $C_j(1\le j\le k)$와 문서 d_i의 유사도 $sim(d_i, C_j)$를 계산한다. 유사도는 코사인 계수로 계산한다.

- 단계 4 : 문서-클러스터 유사도 $sim(d_i, C_j)$가 임계치 θ 이상이면, 현재의 문서 d_i를 그 클러스터의 멤버로 한다 ($d_i, \in C_j$). 임계치 θ이하이면, 현재의 문서 d_i하나로 구성된 새로운 클러스터를 생성한다 ($d_i, \in C_{k+1}$).
- 단계 5 : 새로 생성된 클러스터나 변화가 생긴 클러스터에 대해서 클러스터 중심벡터를 생성한다. 클러스터 중심벡터는 클러스터에 속하는 문서들의 가중치의 평균으로 하였다.

여기서, 클러스터들 사이의 구분이 엄격히 구분되지 않기 때문에, 하나의 문서는 여러 클러스터의 멤버가 될 수 있도록 하였다.

관련있는 문서들은 관련없는 문서들에 비해서 보다 높은 유사도를 갖는다는 클러스터 가설[19]에 따라, 관련있는 문서들은 같은 클러스터에 속하게 될 것이다. 문서에서 같이 공유하는 어휘들이 많을수록 두 문서는 유사도가 높아진다. 그러므로, 문서 클러스터링에서, 비슷한 내용의 문서들은 하나의 클러스터로 분류되기가 쉽다.

3.2 클러스터 우선 순위

질의와 클러스터의 유사도, 즉 질의-클러스터 유사도는 클러스터 중심벡터와 질의벡터의 내적 곱과 클러스터의 질의포함비율로 다음과 같이 계산을 한다.

$$sim(q, c) = \frac{|C_q|}{|q|} \sum_{i=1}^{t} w_{qi} \cdot w_{ci} \qquad (4)$$

여기서, $|q|$는 질의에 나타난 어휘의 개수이다. $|C_q|$는 클러스터 중심벡터가 포함하고 있는 질의에 나타난 어휘의 개수이다. 따라서, $|C_q|/|q|$는 질의포함비율을 나타낸다. w_{qi}는 질의 어휘 i의 가중치이고, w_{ci}는 클러스터 중심벡터의 어휘 i의 가중치이다.

같은 클러스터에 속하는 모든 문서들은 동일한 질의-클러스터 유사도를 갖게 된다. 클러스터 선호도는 질의포함비율에 의해 영향을 받는데, 이는 하나의 질의 어휘를 많이 포함하는 것보다 다양한 어휘를 포함하는 것을 우선시하도록 한 것이다. 이와 같이, 클러스터 정보는 문서에 포함된 어휘들의 행태 뿐만 아니라, 문서들 사이의 관계가 클러스터를 통해서 다른 문서에 영향을 줄 수 있게 되어, 문맥 반영 또는 질의 모호성 해소 효과를 줄 수 있다.

3.3 각 문서에 클러스터 정보를 문맥으로 반영

질의-클러스터 유사도를 이용해서, 검색된 문서의 문맥으로 반영하여 유사도를 다음과 같이 다시 계산한다.

$$sim(q, d)' = sim(q, d) \cdot Max_{d \in c} sim(q, c) \qquad (5)$$

여기서, $sim(q, d)$는 벡터공간검색모델에 기반한 문서검색일 경우에는 식 (1), 확률검색모델에 기반한 문서 검색일 경우에는 식 (2)에 정의한 질의-문서 유사도이다. $sim(q, c)$는 문서 d가 속하는 클러스터에 대한 질의-클러스터 유사도로 식 (4)에 정의된 것이다. 하나의 문서는 여러 클러스터의 멤버가 될 수 있으므로, 그 중에서 최대값을 갖는 유사도를 취한다. 문서 d의 새로운 유사도 $sim(q, d)'$는 질의-문서 유사도와 질의-클러스터 유사도의 곱으로 계산한다. 이를 통해서, 질의에 있는 어휘 뿐 아니라 문서들 사이의 관련도를 문맥으로 반영할 수 있다.

클러스터 정보를 문맥으로 사용하여 문서의 유사도를 다시 계산해서 순위를 매김으로써 나타나는 효과는 질의-문서 유사도가 낮은 문서라 하더라도 같은 클러스터에 속하는 다른 문서들의 영향을 받아서 질의-클러스터 유사도가 높은 값을 가질 수 있다. 그 반대의 경우도 성립한다.

4. 실험 및 평가

4.1 실험 환경 설정

교차언어정보검색에서 제안한 모델의 검색 성능을 평가하기 위해, TREC-6과 TREC-8 교차언어 테스트컬렉션을 이용하여 실험하였다. 이 테스트컬렉션은 242,918개의 영어 문서(1988년에서 1990년까지의 AP뉴스기사)와 52개의 영어 질의를 포함하고 있다. 영어 질의는 사람이 직접 한국어와 일본어로 번역을 하였고, TREC 질의의 제목부분만을 질의로 사용하였다. <표 1>은 본 실험에서 이용한 테스트컬렉션에 대한 통계 정보이다. <표 2>는 사전기반 질의 번역에서 발생할 수 있는 모호성의 정도를 나타내고, <표 3>은 영어 질의에 대해 사람이 번역한 한국어 질의와 일본어 질의의 예를 보여준다. 본 실험에서는 질의에 포함된 어휘가 2개 이상인 42개의 질의에 대해서 평가하였다. 질의에 어휘가 1개만 있고, 그것이 여러 의미를 갖는 경우에는 사람이라 할지라도 정확한 의미의 대역어를 선택하기 어렵기 때문에, 2개 이상의 어휘로 구성된 질의에 대해서 제안하는 방법의 실제 효과를 살펴보았다. 1개의 어휘로 구성된 10개의 질의는 점진적 클러스터링에서 임계치를 결정하기 위한 학습자료로 이용했다.

〈표 1〉 TREC 테스트컬렉션 통계

문서 개수	242,918
질의 개수	52
평균 적합 문서 개수	45.12
문서의 평균 길이	479.08 어휘
질의의 평균 길이	2.92 어휘

〈표 2〉 질의 52개에 대한 모호성 정도

질 의	한국어	일본어
질의 어휘 개수	124	124
번역 영어 어휘 개수	585	268
평균 대역 어휘 개수	4.72	2.16

〈표 3〉 영어 질의에 대해 사람에 의해 번역된 한국어 질의와 일본어 질의 예

영어 질의	번역된 한국어 질의	번역된 일본어 질의
Swiss Speed Limits	스위스 속도 제한	スイス 制限 速度
Effects of logging	벌채 효과	伐採 影響
Solar Powered Cars	태양열 자동차	ソーラー 發電 車
Middle-East Peace Process	중동 평화 절차	中東 和平 プロセス
International Terrorism	국제 테러리즘	國際 テロリズム

벡터공간검색으로는 SMART시스템[17]을 이용하여 문서를 검색하였고, 확률검색기법으로는 Lemurs Okapi 시스템[15]을 이용하였다. 교차언어정보검색에 관한 다른 연구와의 비교를 위해서 공기정보를 이용한 모호성 해소 방법을 실험하였다. 또한, 정보검색 및 교차언어정보검색에서 높은 성능향상을 보이고 있는 방법인 적합성 피드백 방법과의 비교를 위해서, 모호성 해소를 한 경우와 그렇지 않은 경우에 대해서 번역후 자동 적합성 피드백에 의한 방법을 비교 실험으로 하였다.

공기정보에 기반한 모호성 해소에서는 질의 어휘에 나타난 어휘들에 대한 정확한 대역어는 문서에서도 같이 나타날 것이라는 가정에 따라, 모든 대역어 중에서 가장 높은 공기값을 갖는 대역어로 선택하였다[1]. 공기정보 $cooc(x,y)$는 다음과 같이 계산하였다.

$$cooc(x,y) = \sqrt{\frac{N \cdot f(x,y)}{f(x)+f(y)}} \qquad (6)$$

여기서, $f(x)$와 $f(y)$는 어휘 x와 y의 빈도수이고, $f(x,y)$는 어휘 x와 y가 윈도우 크기 6이내에서 같이 발생한 빈도수이다. AP1988 뉴스기사에 대해서 계산한 것이다. N의 값은 10,000,000으로 하였다. 정확한 대역 어휘를 선택하기 위해서, $cooc(x,y)$ 계산은 영어 어휘 a에 대한 모든 대역어 x와 영어 어휘 b에 대한 모든 대역어 y에 대해서 계산을 하고, 그 중에서 가장 높은 값을 갖는 것을 대역어로 선택한다.

벡터공간검색모델에 대한 적합성 피드백을 위해서는 검색된 상위 문서 r개에 나타나는 어휘들의 가중치를 합한 것이 가장 높은 값을 갖는 k개의 어휘를 원래 질의에 추가하였다. 확률검색모델에 대한 적합성 피드백은 덧셈에 의한 어휘 가중치 선택보다 나은 성능을 보인, Okapi에서 제공하는 피드백 함수를 그대로 이용하였다.

제안한 방법을 다각적으로 검증하기 위해서, 단일언어에서의 검색성능을 기본으로 하여 비교를 하였고, 교차언어정보검색에서 모호성 해소를 한 경우와 하지 않은 경우, 자동 적합성 피드백을 통해 질의 확장을 한 경우와 제안한 방법의 효과를 비교하였다. 또한 언어에 따른 효과를 보기 위해 한-영과 일-영 교차언어정보검색에 대해서 평가하였다.

1) *monolingual* : 원래 영어 질의에 대해 영어 문서집합에 대한 검색 성능. 단일언어 정보검색으로 비교 기준점으로 이용
2) *t_all_base* : 사전기반 질의 번역에서 모호성 해소없이 모든 대역 어휘를 선택한 경우의 검색 성능
3) *t_all_blindf* : *t_all_base*에서 검색된 문서에 대해 적합성 피드백을 한 경우의 성능
4) ***t_all_rerank*** : *t_all_base*에서 검색된 문서에 대해 클러스터분석을 통한 재순위화 성능
5) *t_one_base* : 사전기반 질의 번역에서 공기정보를 이용해서 모호성 해소 후의 검색 성능
6) *t_one_blindf* : *t_one_base*에서 검색된 문서에 대해 적합성 피드백을 한 경우의 성능
7) ***t_one_rerank*** : *t_one_base*에서 검색된 문서에 대해 클러스터분석을 통한 재순위화 성능

t_all_rerank와 t_one_rerank가 제안하는 방법이다. 문서 클러스터링을 위해 선택한 상위 문서 개수 N은 300이고, 클러스터 멤버를 결정하는 임계치는 1개 어휘로 구성된 질의 10개에서 학습한 것으로, t_all_rerank와 t_one_rerank에서 0.34로 선택하였다. 확률검색모델에서 Okapi에 의한 적합성피드백에서의 파라미터는 영어 단일언어 검색에서 가장 좋은 성능을 보이는 것으로, 10개의 문서에서 10개의 어휘로 선택하였다. 벡터공간검색모델에서의 적합성피드백에서는 파라미터에 따라 성능이 아주 민감하게 차이가 나서, 그 중에서 가장 좋은 결과를 내는 것으로 했다.

4.2 실험 결과

본 연구의 주요 목적은 교차언어정보검색에서 질의 번역에서 모호성이 증폭된 상태에서 제안한 방법의 클러스터 분석을 통한 효과를 관찰하는 것이다(t_all_base에 대한 t_all_rerank에서의 변화). 한-영과 일-영 교차언어정보검색에 대한 실험 결과는 <표 4>에 있다. 한-영 교차언어정보검색에서, 제안한 방법(t_one_rerank)은 모든 대역어를 질의로 선택한 방법(t_all_base)에 비해서 벡터공간검색모델에서는 39.41%, 확률검색모델에서는 36.79%의 성능향상을 보였다. 모호성 해소를 하지 않은 경우, 제안 방법(t_all_rerank)

〈표 4〉 한-영/일-영 교차언어정보검색에서 성능 비교

	한-영 교차언어정보검색				일-영 교차언어정보검색			
	벡터공간검색(SMART)		확률검색(OKAPI BM25)		벡터공간검색(SMART)		확률검색(OKAPI BM25)	
	11-pt 정확률	성능 향상률	집합평균 정확률	성능 향상률	11-pt정확률	성능 향상률	집합평균 정확률	성능 향상률
1) monolingual	0.267	-	0.274	-	0.267	-	0.274	-
2) t_all_base	0.170	-	0.158	-	0.190	-	0.158	-
3) t_all_blindf	0.191	12.35%	0.179	13.29%	0.206	9.47%	0.179	14.57%
4) t_all_rerank	0.204	20.00%	0.201	27.22%	0.216	13.68%	0.201	23.84%
5) t_one_base	0.210	-	0.207	-	0.202	-	0.207	-
6) t_one_blindf	0.217	3.33%	0.216	4.35%	0.211	4.46%	0.216	16.00%
7) t_one_rerank	0.237	12.86%	0.216	4.35%	0.224	10.89%	0.216	12.57%

은 적합성피드백에 의한 방법(t_all_blindf)에 비해 벡터공간 검색모델에서는 6.81%, 확률검색모델에서는 12.30%의 성능 향상을 보였다. 일-영 교차언어정보검색에서, 제안한 방법(t_one_rerank)은 모든 대역어를 질의로 선택한 방법(t_all_base)에 비해 벡터공간검색모델에서 17.89%, 확률검색모델에서는 30.46%의 성능향상을 보였다. 확률검색모델에서는 제안방법이 모호성 해소를 하지 않은 질의에 대해서 나은 성능을 보이고 있으나, 모호성 해소를 한 경우에 대해서는 비슷한 성능을 보이고 있다.

적합성 피드백에서의 파라미터에 따른 성능의 변화와 클러스터에서 임계치에 따른 성능의 변화를 살펴보았는데, 적합성 피드백에서는 파라미터에 따라 성능이 아주 민감하게 변하고 있고, 이에 비해 제안한 방법에서의 클러스터 임계치는 비교적 안정적인 성능을 보였다.

실험 결과를 통해서, 제안 방법인 클러스터 분석을 통한 잠정적 모호성 해소 기법에 공기정보에 기반한 모호성 해소 모듈을 통합한 경우 보다 더 성능을 향상시킴을 알 수 있다.

4.3 질의에 모호성이 포함된 경우에 대한 결과 분석

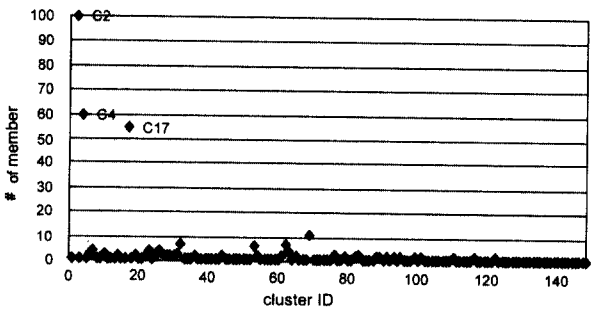
다양한 실험에서 제안한 방법의 효과를 보았는데, 질의에 모호성이 포함된 경우에 대해서 어떤 효과를 내고 있는지 분석해 보았다.

원래 영어 질의 'automobile air pollution'에 대해 번역된 한국어 질의 '자동차 공기 오염'은 한-영 사전을 이용해서 모든 대역 어휘를 질의로 하면, 다음과 같은 어휘를 질의로 갖게 된다 :

자동차	car, automobile, autocar, motorcar
공 기	air, *atmosphere, empty vessel, bowl, jackstone, pebble, marbles*
오 염	contamination, pollution

이때, 어휘 '공기'는 <air>, <atmosphere>, <jackstone>, <co-occurrence>, <bowl> 등 여러 의미를 갖는 다의어이다. 이러한 모든 어휘를 질의로 했을 경우, 정보검색시스템의 성능을 저하시키는 주요인이 된다.

이 질의에 대해 검색된 상위 문서 300개에 대해 점진적 클러스터링을 하여, 146개의 클러스터를 생성하였는데, 전체 클러스터에 속하는 문서 수는 435개였다. 이는 하나의 문서가 여러 클러스터에 속했기 때문이다. (그림 2)는 클러스터 멤버들의 분포를 나타낸다. 부적합 문서들은 하나의 멤버로 구성된 클러스터를 형성하는 경향을 보였고, 적합 문서들은 커다란 클러스터를 형성하였다.



(그림 2) 모호성이 포함된 질의에서 클러스터 분포

〈표 5〉 C4의 클러스터 중심

car	0.069
automobile	0.127
air	0.082
atmosphere	0.018
pollution	0.196
contamination	0.064

〈표 6〉 C85의 클러스터 중심

bowl	0.101
marble	0.191

클러스터가 어떻게 모호성 해소와 문맥 반영 효과를 보이는지 클러스터 내부를 살펴보았다. (그림 2)의 클러스터 C4는 60개의 문서를 멤버로 갖는데, 질의에 대해 209개의 적합 문서중에서 56개는 적합 문서였고, 4개는 부적합 문서였다. 클러스터 중심은 <표 5>와 같이 질의와 관련된 어휘들을 포함하고 있었다. 클러스터 중심에서 다른 의미인 'atmosphere'를 포함하고 있긴 하지만, 그 가중치는 비교적 작은 값이다. 다른 어휘들은 질의에 적절한 대역 어휘로, 동의어이다. 질의에 나타난 모든 어휘가 클러스터 중심에 포함되어 있으므로, 질의포함비율은 1이 되고, 다른 동의어들은 벡터내적곱에서 긍정적인 효과로 작용하므로, 질의-클러스터 유사도는 높은 값을 갖게 된다. 따라서, 클러스터 선호도는 높은 값을 갖게 되기때문에, 이 클러스터에 속하는 모든 문서들의 순위는 높아지게 된다. 클러스터는 질의에 적합한 문서들의 문맥으로서의 역할을 하게 된다. 반면에, 클러스터 C85는 1개의 문서를 멤버로 갖는 클러스터로, 그 클러스터 중심은 <표 6>과 같이 질의 어휘 3개중에서 1개를 포함하고 있다. 질의포함비율이 1/3이 되어 클러스터 선호도는 낮아지게 된다. 그러므로, 이 클러스터는 문서에 대한 영향력이 약하다.

질의에 대해 검색된 각 문서들에 대한 벡터공간모델에 의한 순위(t_all_base)와 제안한 방법에 의해 재순위화된 결과(t_all_rerank)에서의 순위의 변화에서는 대부분의 문서가 클러스터 분석을 통해서 높은 우선 순위로 변하였다. 이 질의에서의 11포인트 평균정확률은 원영어질의에 대한 검색(monolingual)은 0.6783, 질의 번역시 모호성 해소를 하지 않은 경우의 검색(t_all_base)은 0.5635, 제안한 방법에 의한 검색(t_all_rerank)은 0.6622을 나타냈다. 질의에 모호성이 증가한 상태에 대해서도 제안 방법이 단일언어검색에 대한 97.62%의 성능을 보였다.

이러한 결과를 통해 클러스터 분석은 질의 번역에서 모호성이 증가한 상태에서도 잠정적으로 모호성 해결에 도움을 주고, 문서의 관계를 문맥으로 반영하여 성능을 향상시킴을 보여준다.

5. 결 론

본 논문에서는 한-영 교차언어정보검색과 일-영 교차언어정보검색에서 사전기반 질의 번역에서 발생하는 모호성을 해소하기 위해 클러스터 분석을 제안하였다. 제안한 방법에서는 검색된 문서에 대한 클러스터를 각 문서의 문맥으로 이용하여 순위를 다시 매기는데 이용하였다. TREC 테스트컬렉션을 이용하여 제안한 방법의 유효성 평가를 하였는데, 모호성 해소를 하지 않은 경우의 질의에 대해서, 한-영 교차언어정보검색에서는 벡터공간검색모델에 대한 검색에서 39.41%의 성능향상, 확률검색모델에서 36.79%의 성능향상을 보였다. 일-영 교차언어정보검색에서는 각각 17.89%와 30.46%의 성능향상을 보였다. 제안한 방법을 교차언어정보검색에서 적용한 적합성 피드백 방법과의 비교에서는 한-영 교차언어정보검색에서는 모호성 해소를 하지 않은 경우 벡터공간검색모델에서는 6.81%의 성능향상, 확률검색모델에서는 12.30%의 성능향상을 보였다. 또한, 적합성 피드백에서는 파라미터에 따라 성능의 변화가 심한데 비해, 제안한 방법인 클러스터링에서의 임계치에 따른 성능 변화는 비교적 안정적이다. 이러한 결과를 통해, 클러스터 분석은 질의에 대해 클러스터 문맥을 제공할 뿐 아니라, 모호성 해소에 큰 도움을 주었다고 볼 수 있다. 또한 제안한 방법이 어떤 언어에도 적용될 수 있는 언어 독립적인 모델임을 보여주고, 정보검색모델로 벡터공간검색기법과 확률검색기법에 대해 적용을 하였는데, 이는 제안한 방법이 검색모델에 상관없이 적용될 수 있음을 보여준다.

제안하는 방법은 검색된 문서에 대해서 순위를 다시 결정하는 것이므로, 전체 재현율에는 변화를 미치지 않고, 정확률에 영향을 주는 것이다. 따라서, 향후 연구를 통해서 질의 확장과 같은 재현율을 향상시킬 수 있는 모델과 통합된다면 보다 높은 성능 향상을 기대할 수 있을 것이다.

참 고 문 헌

[1] 천정훈, 한영 교차언어 정보검색 시스템에서 질의어의 모호성 해소와 병렬 코퍼스를 이용한 질의어 보완, 한국과학기술원 전자전산학과 석사학위논문, 2000.

[2] Anick, P. G. and Vaithyanathan, S. Exploiting Clustering and Phrases for Context-Based Information Retrieval, In Proc. of 20th ACM SIGIR Conference, 1997.

[3] Ballesteros L. and Croft W. B. Resolving Ambiguity for Cross-language Retrieval, In proc. of 21rd ACM SIGIR Conference, 1998.

[4] Breen, J. EDICT Japanese/English Dictionary File. The Electronic Dictionary Research and Development Group, Monash University, 2003.

[5] Church, K. W. and Hanks P. Word Association Norms Mutual Information and Lexicography, Computational Linguistics, 16(1), pp.23-29, 1990.

[6] Dumais, S. T., Letsche, T. A., Littman, M. L. and Landauer, T. K. Automatic cross-language retrieval using latent semantic indexing, In Proc. of AAAI Symposium on Cross-Language Text and Speech Retrieval, 1997.

[7] Eichmann, D., Ruiz, M. E. and Srinivasan, P. Cross-Language Information Retrieval with the UMLS Meta-

thesaurus, In Proc. of the 21th ACM SIGIR Conference, 1998.

[8] Gilarranz, J., Gonzalo, J. and Verdejo, F. An Approach to Conceptual Text Retrieval Using the EuroWordNet Multilingual Semantic Database, In Proc. of AAAI Spring Symposium on Cross-Language Text and Speech Retrieval, 1997.

[9] Hearst, M. A. and Pedersen, J. O. Reexamining the Cluster Hypothesis : Scatter/Gather on Retrieval Results, In Proc. of 19th ACM SIGIR Conference, 1996.

[10] Hull, D. A. and Grefenstette, G. Querying across languages : a dictionary-based approach to multilingual information retrieval, In Proc. of the 19th ACM SIGIR Conference, 1996.

[11] Jang, M. G., Myaeng, S. H. and Park, S. H. Using Mutual Information to Resolve Query Translation Ambiguities and Query Term Weighting, In Proc. of the 37th Annual Meeting of the Association for Computational Linguistics, 1999.

[12] Kwon, O-W., Kang, I. S., Lee, J-H and Lee, G. B. Cross-Language Text Retrieval Based on Document Translation Using Japanese-to-Korean MT system, In Proc. of NLPRS '97, 1997.

[13] Lee, K. S., Park, Y. C., Choi, K. S. Re-ranking model based on document clusters, Information Processing and Management, 37(1), pp.1-14, 2001.

[14] Oard, D. W. and Hackett, P. Document Translation for the Cross-Language Text Retrieval at the University of Maryland, In Proc. of the Sixth Text Retrieval Conference (TREC-6), 1997.

[15] Paul, O. and Callan, J. Experiments Using the Lemur Toolkit, InProc. of the Tenth Text REtrieval Conference (TREC-10), 2001.

[16] Robertson, S. E. and Walker, S. Okapi/Keenbow at TREC-8, In Proc. of the Eighth Text REtrieval Conference (TREC-8), 1999.

[17] Salton, G. Automatic Text Processing : The Transformation, Analysis, and Retrieval of Information by Computer, Addison-Wesley, Reading, Pennsylvania. 1989.

[18] Smadja, F., McKeown, K. R. and Hatzivassiloglou, V. Translating collocations for bilingual lexicons : A statistical approach, Computational Linguistics, 22(1), pp.1-38, 1996.

[19] van Rijsbergen, C. J. Information Retrieval, Butterworths : London, second edition, 1979.

[20] Xu, J. and Croft, W. B. Query Expansion Using Local and Global Document Analysis, In Proc. of the 19th ACM SIGIR Conference, 1996.

이 경 순

e-mail : selfsolee@chonbuk.ac.kr

1990년~1994년 계명대학교 컴퓨터공학과 학사

1995년~1997년 한국과학기술원 전자전산학 석사

1997년~2001년 한국과학기술원 전자전산학 박사

2001년~2003년 일본 국립정보학연구소(National Institute of Informatics) 연구원

2004년~현재 전북대학교 전자정보공학부 전임강사

관심분야 : 정보검색, 지식 마이닝, 자연언어처리