한·중 교차언어 검색에서 시소러스를 이용한 질의 확장

김풍[O] 강인수 이종혁
포항공대 정보통신대학원 정보처리학과[O] 포항공대 컴퓨터공학과 첨단정보기술 연구센터
{maple[O], dbaisk, jhlee}@postech.ac.kr

Query Expansion Using Thesaurus for Korean to Chinese Cross-Language Text Retrieval

Feng Jin[O] In-Su Kang Jong-Hyeok Lee
Dept. of Graduate School for Information Technology, POSTECH[O], Dept. of Computer Science &Engineering, POSTECH and Advanced Information Technology Research Center(AITrc)

요 약

본 논문은 한·중 교차언어 검색을 위한 효과적인 질의 확장에 대해 기술하고 있다. 한·중 교차언어 검색은 한국어 질의로 중국어 문서를 검색하는 것이고 본 논문에서는 대역어 사전을 이용하여 한국어 질의를 중국어 질의로 변환하는 방식을 사용한다. 질의 확장을 위한 방법으로 중국어 시소러스인 "同義詞詞林"을 사용하였다. 그리고 동의어들과 주변 단어간의 상호 정보를 비교함으로서 재현률과 정확률을 높였다. 실험을 통하여 검증한 결과 사전만 사용하여 변환하는 방법에 비하여 검색 성능이 향상되었다.

1. 서론 및 기존연구

한·중 교차언어 정보검색은 한국어 질의를 사용하여 중국어 문서를 검색하는 것을 의미한다. 교차언어 정보검색 시스템의 성능에 가장 많은 영향을 주는 단계에는 질의 변환 단계와 질의 확장 단계가 있다.

질의 변환을 하는 방법에는 질의를 문서 언어로 변환하는 방식과 문서의 언어를 질의 언어로 변환하는 방식, 그리고 질의 변환과 문서 변환 방식을 같이 사용하는 혼합 방식이 있다. 변환자원으로는 대역어 사전, 병렬코퍼스, 그리고 기계번역시스템을 이용하는 변환이 있다. 질의 변환방식에서 대역어 사전을 이용한 방식은 변환 실패의 원인을 알아 내기 쉬우며 확장하는데 어려움이 적은 장점이 있다 [강인수, 1997]. 이러한 이유로 본 논문에서는 대역어 사전을 이용한 질의 변환방식을 적용하였다.

질의 확장은 질의 변환 시에 불가피하게 발생하는 변환 중의성(translation ambiguity)을 해결하기 위한 것이다. 질의 확장의 방법으로는 크게 지역적 질의 확장과 전역적 질의 확장으로 나눌 수 있다. 지역적 질의 확장은 검색된 문서의 일부만의 정보를 사용하는 것으로 적합성 피드백을 이용한 질의 확장이 여기에 해당된다. 전역적 질의 확장은 문서 집단 전체나 대상 언어 전체의 정보를 사용하여 질의를 확장하는 방법인데 단어들의 공기정보나 시소러스를 이용한 질의 확장 방법이 가장 보편적이다[Gao, 2000][김백일, 2002]. [Rila 1998]에서는 WordNet을 보완하는 방법으로 공기정보 시소러스, 용언-논항 관계 시소러스를 구축하는 방법을 제기하였고 세가지 시소러스를 이용하여 질의를 확장하였을 때 재현률과 정확률이 많이 현저하게 하였다.

본 논문에서는 전역적 질의 확장 기법을 사용하고 있다. 단어들의 공기정보는 검색 대상 문서에서 추출하였고 중국어 질의어 확장은 중국어 시소러스인 "同義詞詞林"을 사용하였다.

"同義詞詞林"[梅家駒, 1985]은 중국어 권에서 가장 권위 있는 중국어 시소러스이다. 북경어를 기준으로 작성되었고, 수록된 단어는 약 7만개 이며, 세분화 정도에 따라 4계층으로 나누어졌다. 대분류에는 12개 카테고리, 중분류에는 94개 카테고리, 소분류에는 1428개 카테고리가 있고, 소분류 이하에는 동의어를 기준으로 3925개의 단어 집합을 만들었는데 집합마다 대표적인 표제어 하나를 선정하여 표시하였다.

2. 질의 확장

2.1 한·중 교차언어 검색 시스템 구성

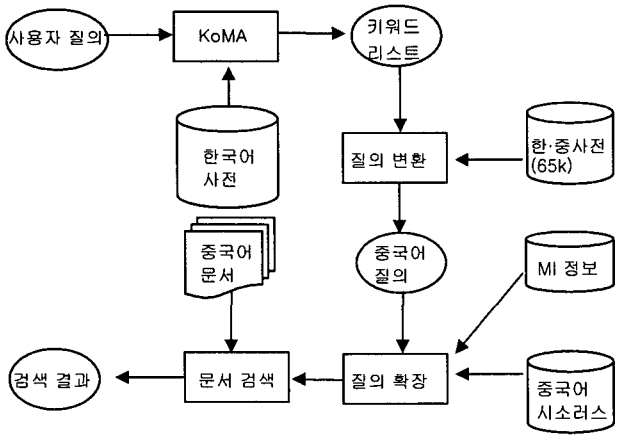
그림 1은 한·중 교차언어 검색 시스템의 구성도이다. 시스템은 표제어 추출, 질의 변환, 질의 확장, 문서 검색 등 4개의 단계로 이루어 졌고, 한·중대역어 사전, 중국어 상호정보, "同義詞詞林" 등의 자원을 사용하였다. 표제어 추출단계에서는 한국어 형태소 분석기-KoMA[1] 를 사용하였다.

본 논문에서는 한국어 질의어로부터 사전을 이용하여 변환된 중국어 대역어는 '사전 대역어'라고 정의하고 시소러스를 통하여 얻은 '사전 대역어'의 동의어는 '시소러스 대역어'라고 정의한다.

2.2 질의 확장`

질의 변환 단계를 거쳐 각 한국어 질의에 대한 중국어

1) KoMA: 포항공대 지식 및 언어공학 연구실 한국어 형태소 분석기



[그림 1] 시스템 구성도

대역어를 찾았지만 사용자가 원하는 충분한 질의를 생성시켰다고는 말하기 어렵다. 예를 들어 '여관'을 중국어로 '旅館'으로 변역하는 것은 합당하지만 '旅店'도 똑같은 의미를 지니고 있고, '客栈'으로 쓰는 경우도 있다. 그리고 생성된 대역 질의 중에 많은 필요 없는 단어들이 섞여 있을 수 있다. 예를 들어 '고도'를 중국어로 변환하면 '古都','孤島','高度'와 같은 대역어가 생성되는데 각기 다른 의미를 갖고 있는 단어로서 한 개를 제외한 나머지는 불필요한 단어들이다.

이상 질의 확장단계에서 발생하는 문제들은 "同義飼飼林"과 공기정보를 이용하여 해결하고자 한다. 중국어 질의가 충분히 생성되지 못하는 문제는 "同義飼飼林"을 이용하여 대역어들의 동의어를 추가하여 줌으로서 해결하고, 필요 없는 단어를 제거하기 위하여서는 검색 대상 문서 전체를 통계하여 얻은 각 원 질의에 대응되는 모든 시소러스 대역어와 앞뒤 원 질의의 시소러스 대역어들과의 상호정보를 사용하여 일정한 Threshold를 초과한 단어만 최종 검색 후보로 선택 함으로서 해결한다.

식(1)은 사전 대역어나 시소러스 대역어들의 가중치를 구하는 식이다. 예를 들어 사전 대역어에 더 많은 가중치를 주고 싶으면 '水道'에게는 0.7, '溝子'에게는 0.3 이라는 가중치를 각각 부여하면 된다(그림 2 참조). 본 실험에서는 가중치를 모두 1.0으로 정한다. 단어와 앞뒤 원 질의어의 시소러스 대역어안의 단어들과의 상호정보 값을 sigmoid 함수에 대입하여 결과를 구한다(식 2).

$$W(t) = \lambda(t) \cdot W_{MI}(t) \cdot W_{tf \cdot idf}(t) \quad \text{----------} \quad (1)$$

where

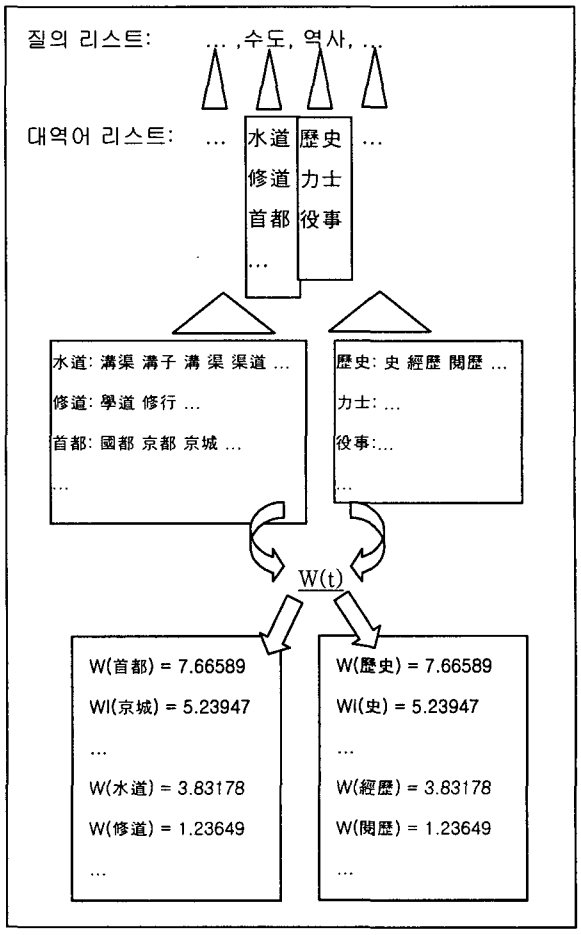
$$\lambda(t) = \begin{cases} \lambda_1 & \text{(t가 사전 대역어)} \\ \lambda_2 & \text{(t가 시소러스 대역어)} \end{cases}$$

$$W_{MI}(t) = sigmoid\left(\underset{t' \in T^{sno(t)-1} \cup T^{sno(t)+1}}{\arg\max} MI(t',t) \right) \quad (2)$$

$$sigmoid(x) = \frac{1}{1+e^{-x}}$$

$sno(t)$: t의 원 질의어의 순번

T^x : x번째 원 질의의 대역어 집합



[그림 2] 질의 확장 과정도

그림 2는 질의를 확장하는 과정을 보여주는 예이다. 본 시스템은 경험적으로 한국어 질의어 한 개 당 시소러스 대역어를 최대 10개 생성하고 W(t)의 threshold는 5.0으로 하고 있다.

3. 실험 및 결과

실험용으로 중국어 코퍼스는 NTCIR-3(The Third NII-NACSIS Test Collection for IR Systems)에서 제공한 중국어코퍼스를 사용하였는데 코퍼스 사이즈는 대략 529MB(문서 381,681건) 이다. 검색에 사용되는 질의는 NTCIR에서 제공하는 한·중 교차언어 검색용 한국어 문서

50개 중 42개이다. 질의 문서 안의 제목, 설명, 내용, 키워드를 질의로 사용하였다.

실험은 아래와 같은 7가지 경우로 나누어서 진행하였다.

1. 단일어 검색: 중국어 질의를 이용한 문서검색
2. 단순 변환: 사전을 이용한 질의 변환
3. 최적 변환: 상호 정보를 이용한 중국어 대역어(질의) 선택
4. 질의 확장1: 단순 변환 후의 질의 확장 (상호 정보 사용하지 않음)
5. 질의 확장2: 단순 변환 후의 질의 확장 (상호 정보 사용)
6. 질의 확장3: 최적 변환 후의 질의 확장 (상호 정보 사용하지 않음)
7. 질의 확장4: 최적 변환 후의 질의 확장 (상호 정보 사용)

	Avg.P.	% of Mono.IR
실험 1	0.3283	
실험 2	0.1428	43.50 %
실험 3	0.1489	45.35 %
실험 4	0.1413	43.04 %
실험 5	0.1432	43.62 %
실험 6	0.1639	49.92 %
실험 7	0.1811	55.16 %

[표 1 질의 변환 확장에 따른 실험

단일어 검색: 첫번째 실험에서는 우선 교차언어 검색 시스템의 성능에 중요한 영향을 미치며 검색 시스템에서 질의 변환 및 확장 결과를 평가할 수 있는 척도로 사용 할 수 있는 중국어 단일어 검색을 진행하였다.

단순변환: 사전을 이용한 단순 변환 시스템은 교차 언어 검색 시스템의 가장 단순한 방식이며 실험을 통하여 시스템에서 사용하는 대역어 사전의 질을 알아볼 수 있고 아울러 교차 언어 시스템의 최저 성능을 측정할 수 있다. 결과는 단일어 검색 시스템의 43%정도에 미쳤는데 6만 2천 개 표제어가 들어 있는 대역어 사전은 교차언어 검색에 사용하기에 부족함이 있다는 것을 알아 볼 수 있었다.

최적변환: 질의 변환 시에 나타나는 필요 없는 단어를 줄이는 방안으로 [실험 3]에서는 질의와 인접한 질의 사이의 상호 정보를 사용하였다. 정확률은 약 1.9%의 향상하였다.

질의 확장: 질의 확장은 4가지 경우를 나누어 진행 하였다. 사전을 이용한 질의 변환 후 시소러스를 이용하여 확장하였는데 정확률이 확장하기 전보다 떨어졌다. 그 이유는 정교하지 못한 확장으로 인하여 많은 불필요한 후보가 포함되었기 때문이다. 단순 변환 후 상호 정보를 고려한 질의 확장을 하였는데 성능이 올라가지 못했다. 변환 때 잘 못 생성된 질의 조합이 최종 질의 집합에 많이 포함 된 것이 원인이다. 최적 변환을 거친 후의 질의 확장은 보다 좋은 결과를 나타냈다. 그 중에서 상호 정보를 이용한 질의 확장은 본 시스템 성능 테스트에서 최고로 높은 정확률(0.1639)을 나타냈다.

실험 결과를 통하여 중국어 시소러스를 이용한 질의 확장은 한중 교차언어 검색에서 유용한 수단이라는 것을 알 수 있다.

4. 결론 및 향후 연구

본 논문에서는 한·중 교차언어 검색에서 질의 변환 및 확장에 관하여 살펴 보았다. 한국어 형태소 분석기-KOMA를 통하여 얻어진 한국어 단어를 사전을 통하여 얻어진 대역어가 하나 이상일 경우 중국어 단어들의 상호정보를 이용하여 가장 적합한 후보를 선택하였다. 단일어 검색이나 영어와 관련된 교차언어 검색 등에서는 시소러스를 이용한 동의어 확장 기법을 많이 써왔지만 한·중교차언어 검색에서 변환 후 확장의 기법으로 "同義詞詞林"을 이용한 질의 확장은 본문이 처음 제기하였다. 실험결과 오직 사전을 이용하여 질의 변환을 하는 기법에 비하여 성능이 많이 향상하였다.

향후 계획으로는 한·중 교차언어 검색에서 한국어 의미체계를 이용한 한국어 질의 확장을 본문에서 제기한 중국어 대역어 확장과 비교 분석하고, 시소러스를 이용한 질의 확장 외에 전통적으로 사용되고 있는 적합성 피드백 기법을 본 연구와 접목 시키는 것이 있고 교차언어 검색 성능의 저하의 주 요인인 한국어 고유명사에 대한 식별 능력을 한층 강화하고 고유명사의 변환 기법을 연구할 계획이다.

5. 감사의 글

본 연구는 첨단정보기술 연구센터를 통하여 과학재단의 지원을 받았습니다.

6. 참고 문헌

[1] 강인수, 이종혁, 이근배. 교차언어 문서 검색에서 질의어의 중의성 해소 방법, 제9회 한글 및 한국어정보처리 학술대회, 1997

[2] 김백일, 서희철, 임해창. 한영 교차언어 정보검색에서 질의 변환 및 질의 확장 방법. 제14회 한글 및 한국어 정보처리 학술발표 논문집, pp.235-242, 2002

[3] Gao, J., Nie, J.Y., Zhang, J., Xun, E., Su, Y., Zhou, M., and Huang C.. Trec-9 CLIR experiments at MSRCN. In Trec-9, pp.343-353, 2000

[4] Rila M., Tokunaga T., Tanaka H.. The Use of WordNet in Information Retrieval. Coling-ACL ' 98 Workshop, 1998

[5] 梅家駒, "同義詞詞林", 上海辭書出版社, 1985