특집

일-한 교차언어 문서검색 시스템

권 오 욱, 이 종 혁, 이 근 배
포항공과대학교 전자계산학과

I. 서 론

정보검색에서 더 이상 사용자의 정보 요구는 자국어로 쓰여진 문서로는 만족할 수 없게 되었다. 이것은 최근에 급격히 발전한 WWW(World-Wide Web)와 CD-ROM 등의 보급을 통하여 자국어가 아닌 외국어로 쓰여진 문서를 접할 기회가 많아졌기 때문이다. 그러나 사용 언어에 구애 받지 않고 신속하고 손쉽게 최신 정보를 입수하기 위해서는 교차 언어 문서 검색(cross-language text retrieval)이 필수적이다. 교차 언어 문서 검색은 사용자의 질의(query)와 검색하고자 하는 문서(document) 간에 서로 다른 언어가 사용된 경우, 두 언어 간의 장벽을 극복하기 위한 문서 검색을 말한다.

교차 언어 문서 검색에서는 일반적으로 자국어로 쓰여진 질의를 이용하여 외국어로 쓰여진 문서들을 검색한다. 전통적으로 이러한 교차 언어 문서 검색을 위해 두 가지 접근 방법이 있다. 첫 번째는 "질의 번역(query translation)"이라 불리는 방법으로서 사용자의 질의를 문서가 쓰여진 언어로 번역하는 것이다. 두 번째 접근 방법은 검색할 문서 집합의 모든 문서들을 질의에 사용되는 언어로 번역하는 "문서 번역(document translation)" 방법이다.

질의 번역 방식은 문서 번역 방식과는 달리 문서 집합 전체를 번역할 필요 없이 사용자 질의만 번역하면 되며, 문장 위주보다는 단어 중심의 번역이므로 번역에 필요한 지식이나 방법들도 간단하다. 또한 이미 구축되어진 기존의 색인을 그대로 사용할 수 있다는 장점 때문에, 대부분의 기존 연구들은 질의 번역 방식에 집중되어 왔다[1, 2, 4, 10]. 반면에, 문서 번역 방식은 검색 대상 문서를 질의에 사용된 언어로 번역해야 하므로 기계 번역 시스템(machine translation system)의 이용이 불가피하다. 그러나 기존의 기계 번역 시스템들은 제한된 번역 영역과 문장 구조에서는 좋은 번역률을 보이지만, 제한이 가해지지 않은 대부분의 경우에서는 많은 번역 오류를 보이는 한계성을 가지고 있다.

또한 독자적으로 개발된 기계 번역 시스템을 정보 검색에 효율적으로 통합, 수용하기가 쉽지 않다. 이러한 이유들로 기계 번역 시스템을 교차 언어 정보검색에 적용하고자 했던 기존 연구들은 거의 없었다[7].

그러나, 근본적으로 교차 언어 문서 검색 시스템의 사용자는 최종 검색된 외국어 문서를 자국어로 번역 받기를 기대할 것이다. 또한 문서 검색 시의 성능 향상을 위한 단어의 의미 중의성 해소(word-sense disambiguation)를 위해서도 단어 수가 작은 질의에서의 번역보다는 많은 주변 단어들로 인해 문맥에 대한 실마리를 더 많이 이용할 수 있는 문서 번역 방식이 더욱 효과적이다. 기계번역 시스템의 번역 품질도 지난 50년 간 꾸준히 발전하여 많은 번역 영역에서 실용 수준에 도달하였다. 따라서 기계번역을 정보검색에 통합하여 사용한다면 질의 번역보다 문서 번역에서 보다 효과적일 것이다.

기계 번역의 다양한 접근 방법에도 불구하고 기본적인 기계번역 처리는 원시 언어 분석(source language analysis), 원시-목적 언어 변환(source-target transfer), 목적 언어 생성(target language generation) 등으로 구성된다.

대부분의 기계번역 시스템들은 원시 언어 해석과 원시-목적 언어 변환을 위해서 다양한 지식을 갖추고 있는데, 이러한 언어 지식을 이용하여 문서 색인 및 검색을 하는 경우, 사용자 질의 및 대상 문서의 의미 표현을 보다 정확하게 나타낼 수 있으므로 정보 검색 성능 향상에 크게 기여한다는 것이 이미 널리 알려져 있다[5].

본 논문에서는 우선 교차 언어 문서 검색에 대한 기존의 연구들을 살펴보고, 이러한 일련의 연구에서 나타나는 번역 방법들이 기존의 기계번역 방법들과 유사함을 밝히며, 기계번역의 교차 언어 문서 검색에서의 역할에 대하여 언급한다. 또한 이 같은 연구 근거를 바탕으로 하여 구현한 포항공과대학교의 일-한 기계번역 및 일-한 교차 언어 문서 검색 시스템에 대한 자세한 설명을 통해 교차 언어 문서 검색에 대한 이해를 돕는다.

본 연구에서는 한-일 양 언어간의 언어적 유사성 때문에 다른 언어 쌍 (예 : 한국어-영어)에 비해 상대적으로 높은 번역률을 보이는 일-한 기계번역시스템 COBALT-J/K (collocation-based language translator from Japanese to Korean)를 이용한 문서 번역(document translation) 방식의 교차 언어문서 검색 모델 CLTR-J/K(cross-language text retrieval from Japanese to Korean)을 제시한다. 이 시스템은 포항제철의 토털 정보 검색 시스템인 POTIS(POSCO Group Technology Information Retrieval System)에 적용되어 1996년부터 성공적으로 가동되고 있다. 제안된 교차 언어 문서 검색 모델은 일-한 기계번역 시스템의 해석 모듈을 이용하여 일본어 문서를 형태소 분석하고, 변환 사전의 연어 패턴(collocation pattern)을 이용하여 일본어 단어를 올바른 한국어 대역어로 번역한 후, 한국어 색인어를 추출한다. 검색 모델은 벡터 공간 모델(vector space model)과 확장 불리안 모델(extended Boolean model)을 기본으로 하고 있다.

본 논문의 구성은 다음과 같다. 2장에서는 교차 언어 문서 검색에 대한 지금까지의 연구들에 대하여 살펴보고, 교차 언어 문서 검색에서의 기계 번역시스템의 역할에 대해서 언급한다. 3 장에서는 본 연구에서 사용한 일-한 기계번역 시스템 COBALT-J/K에 대하여 간단히 설명하며, 4장에서는 이를 기반으로 하여 개발한 문서 번역 방식의 교차 언어 문서 검색 시스템 CLTR-J/K에 대하여 설명한 후, 마지막으로 결론을 맺는다.

II. 교차 언어 문서 검색에서 기계 번역의 역할

지금까지의 교차 언어 문서 검색에 대한 연구들은 우선 관점에 따라 두 가지 방향으로 분류될 수 있다. 첫 번째는 전통적인 정보 검색 시스템에서 어떻게 언어 장벽을 극복할 것인가에 대한 관점으로서 문서 번역 방식과 질의 번역 방식으로 구분될 수 있다. 또 다른 관점으로는 어떻게 질의/문서를 번역할 것인가에 관한 것으로서 (1) 대역 사전

(bilingual dictionary)이나 다국어 시소러스(multilingual thesaurus) 등을 이용한 다국어 지식 기반방법(multilingual knowledge based method), (2) 병렬 코퍼스(parallel corpus)[1]나 대조 코퍼스(comparative corpus)[2] 등을 이용하는 코퍼스 기반방법(corpus based method), 그리고 (3) 기계번역 모듈에 기반한 방법(machine translation modular based method) 등으로 구분될 수 있다.

질의 번역 방식에서는 이 3가지 방법을 이용한 연구가 가능하지만, 문서 번역방식에서는 문서 번역의 특성으로 인하여 기계번역 시스템을 이용한 방법만이 제안되었다[7].

Ballesteros[1]는 대역 사전(bilingual dictionary)을 이용한 질의 번역에서 단어(word) 단위의 번역보다는 구(phrase) 단위의 대역어 선정이 시스템 성능향상에 더욱 효과적이라고 언급하였다. David[4, 5]는 질의 번역을 위한 다섯가지 접근 방법을 제안하였는데, 병렬 또는 대조 코퍼스만을 이용한 대역어 선정보다는 대역 사전(Collins English-Spanish bilingual dictionary)을 함께 이용한 경우가 더욱 효과적임을 실험으로 증명하였고, 또한 [4]에서는 품사 태거(part-of-speech tagger)를 이용한 품사 추정이 대역어 선정 및 검색 성능에 더욱 효과적임을 보였다. Fluhr[2]는 질의 번역을 위해 다양한 언어 처리 방법을 도입하였는데 형태소 분석과 구문 파싱(syntactic parsing), 그리고 구문/의미적 언어 제약 규칙들을 사용하는 것이 효과적임을 보였다. 결론적으로 이상의 연구들은 자연언어처리 기법과 기계번역을 위한 다양한 지식 등이 단어의 의미 중의성 해소 및 대역어 선정에 중요한 역할을 하며, 따라서 효율적인 교차 언어 문서 검색에 매우 중요한 요소임을 보이고 있다.

그러나, Fluhr [2]은 교차 언어 문서 검색에서 단순한 기계번역 방식의 도입이 오히려 비효율적일 수 있음을 보였다. 즉, 세계적으로 잘 알려진 기계번역 시스템 SYSTRAN을 기반으로 하여 하나의 번역 결과만을 사용하는 질의 번역(query translation) 방식이 오히려 검색 효율을 감소시켰다.

이것은 기계 번역의 결과를 단순하게 그대로 문서 검색에 이용함으로써 발생한 문제이다. 그러나 대안으로 제시한 Fluhr의 질의 번역 방식도 근본적으로 기계 번역에서 사용하는 것과 같은 유형의 지식 및 처리 기법을 이용하고 있다. 단지 SYSTRAN을 이용한 방법보다 좋은 결과를 얻은 이유는 하나의 결과만을 이용하는 SYSTRAN 기반의 질의 번역 방식과는 달리, 모든 가능한 질의 번역 결과를 이용하여 검색을 행한 후 의미 필터링(semantic filtering)을 통하여 적합한 문서(relevant document)들만을 남겼기 때문이다. 따라서 독립적으로 운영 가능한 기존의 기계 번역 시스템을 문서 검색 시스템에 대한 전처리(front-end) 과정으로 단순히 도입하기보다는, 통합 시스템의 관점에서 기계번역 모듈 단위의 효율적인 통합이 매우 중요함을 보이고 있다[7].

Gachot[6]는 기계번역 시스템 SYSTRAN을 이용해서 원시언어의 문서를 번역하고 문서에 대한 색인 표현 방법으로서 언어 구조와 의미 구조를 추가하는 모델을 제안하였다. 이러한 색인 구조의 복잡성은 검색 시, 질의 표현 구조와의 매칭 부담으로 시스템의 시간 복잡성(time complexity)을 증가시키는 요인이 된다. 기존의 많은 연구에서 언급하였듯이 언어적 표현으로 색인어의 의미를 정의할 경우, 현재까지 정보검색에서 성공적이었던 통계적 방법이 배제될 수 있다는 점을 간과하고 있다. 즉 색인어에 대한 의미 협소성은 상대적으로 통계적 희귀성을 가질 수 있기 때문에 이러한 색인어에 통계적 방법을 적용할 경우 정보검색의 성능 향상은 거의 없다. 사실 정보검색에서는 정보가공의 목적인 문서 이해(text understanding) 관점에서 구문/의미의 구조적 연결성을 내부 구조로 표현할 필요가 없다. 문서 이해를 통하여 얻어진

1) Parallel Corpus : 같은 문장에 대해서 두 개 이상의 언어로 정렬된 코퍼스
2) Comparative Corpus : 각기 다른 언어로 작성되었으나 도메인과 문서의 내용이 비슷한 경우의 코퍼스, 예로 197년 영어로 쓰여진 국제 경제에 대한 신문 기사 코퍼스와 한글로 쓰여진 신문기사 코퍼스

개념을 될 수 있으면 쉽게 접근할 수 있는 간단한 구조로 표현하여야 한다.

교차 언어 문서 검색에서 사용자가 가질 수 있는 기계 번역 방법에 대한 기대는 타국어 문서를 자국어로 볼 수 있다는 것이다. Grenichiro et al.[8]은 전체 문서 번역에 드는 오류와 부담을 줄이기 위해, 문서의 제목이라도 자국어로 번역하여 사용자에게 도움을 주고자 하였다. 그리고, Yamabana et al.[10]은 영-일 기계 번역 시스템을 이용하여 전체 문서를 상당히 낮은 수준이지만 번역해서, 사용자가 검색된 문서들에서 관련 문서와 비관련 문서를 자국어로 빠르고 쉽게 판단할 수 있도록 하였다.

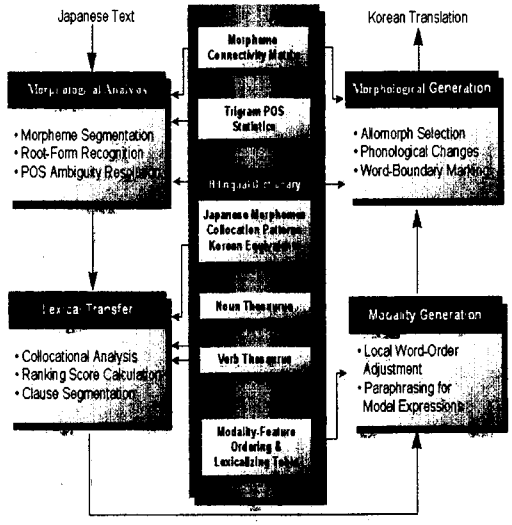
이상에서 살펴본 기존 연구를 종합하면, 교차 언어 문서 검색에서의 기계 번역의 역할 및 그 이용 방법은 다음과 같이 요약될 수 있다.

- 언어 교차(crossing languages) : 언어 장벽 해소를 위하여 한 언어를 다른 언어로 변환함으로써 질의 및 문서 간의 어휘 일관성을 가지게 한다.
- 정보검색의 성능 향상(improving the efficiency of information retrieval) : 질의/문서에 대한 언어 처리를 통하여 형태/구문/의미 등의 보다 많은 정보를 정보검색에서 사용하여 문서 검색의 성능을 향상시킬 수 있다.
- 문서 선별(screening) : 사용자가 추출된 문서의 적합성 여부를 자국어로 판단할 수 있게 하며, 자국어로 쓰여진 정보를 직접 추출할 수 있게 한다.

III. COBALT-J/K(Collocation-Based Language Translator from Japanese to Korean)

COBALT-J/K[3][9] 프로젝트는 1993년부터 POSTECH에서 번역 영역(domain)에 제한이 없는 고품질의 일-한 자동 번역을 목적으로 시작되었다. 1994년 후반부터 포항제철의 자금 지원으로 본격적으로 확장되어 현재에는 포항제철의 토탈 정보 관리 시스템 POTIS에 적용되어 성공적으로 운용 중에 있으며 1997년부터는 상용화를 위해서 삼성전자와 함께 Window 95 Version을 개발 중에 있다.

COBALT-J/K는 한국어와 일본어의 문법적 유사성으로 인하여 직접 번역 방식(direct MT strategy)으로 개발되었다. COBALT-J/K의 전체 번역 과정은 일본어 형태소 분석(Japanese morphological analysis), 어휘 변환(lexical transfer), 양상 생성(modality generation), 한국어 형태소 생성(Korean morphological generation) 등의 4 단계로 이루어진다. <그림 1>은 COBALT-J/K의 전체 처리 과정을 보이고 있다.



〈그림 1〉 COBALT-J/K의 번역 흐름도

형태소 해석 단계에서는 먼저 일본어 문장을 형태소 단위로 분할하는 동시에 여러 가지 형태론적 변형을 원형으로 복원한다. 일반적으로 하나의 일본어 단어는 하나 이상의 형태소 조합으로 해석이 가능하다. 적당하지 않은 형태소 조합을 제거하기 위해 좌우 접속 정보를 이용하여 형태소간 결합 타당성을 검사하며, 분할에 있어서는 "이문절 최장 일치법"과 "강접속 우선 법칙"이라는 두 가지 휴

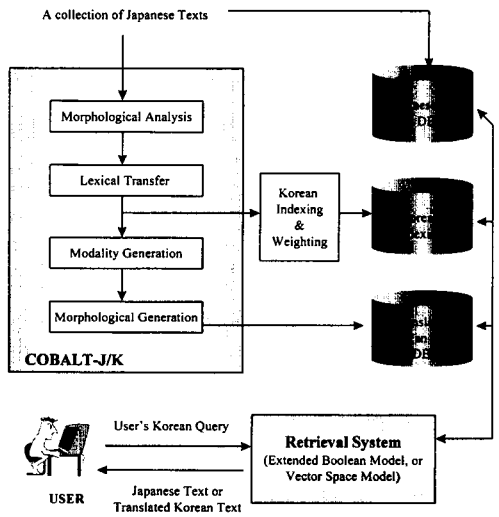
리스틱을 이용한다.

어휘 변환은 문장 중의 일본어 어휘를 한국어 대역어로 변환하는 과정이다. 원시언어에서 여러 의미를 갖는 동형이의어로 인해 또는 원시-목적 언어간 어휘적 차이로 인해 하나의 일본어 어휘가 여러 개의 한국어 대역어로 대응될 경우 이를 해결하기 위한 방법이 필요하다. 가장 적절한 대역어를 선택하기 위해서 일-한 대역 사전은 각 어휘당 여러 개의 문맥 정보를 가진 연어 패턴(collocation pattern)들로 구성된다. 각 연어 패턴은 그 어휘가 특정 의미로 쓰일 수 있는 주변 문맥 상황을 기술한 것으로서 의미 자질(semantic feature)에 의한 선택 제약(selectional restrictions)을 나타낸다. 여러 연어 패턴 중 선택 제약이 가장 적게 어긋나는 연어 패턴을 추출하기 위해 순서화 기법(ranking mechanism)을 이용한다. 순서화 기법은 명사와 동사의 시소러스 계층(thesaurus hierarchy)을 이용하여 연어 패턴과 입력 문장간의 유사도(degree of similarity)를 계산한다. 명사 시소러스는 4 계층에 약 1,000개의 의미 범주(semantic category)로 구성되어 있고, 동사 시소러스는 2~5 계층에 약 40 개의 구문-의미 범주(syntactic-semantic category)로 구성되어 있다. 더욱이 대부분의 직접 기계 번역 방식과는 달리 얕은 파싱(shallow parsing)을 하여 연어 패턴과 매칭할 원시 문장을 선택한다. 연어 패턴을 이용한 의미 중의성 해소에 대해 자세한 내용은 [3]에 기술되어 있다.

양상 생성 단계에서는 서술부에 대한 두 언어 사이의 어순 차이를 해결하고, 양상 표현에서의 존칭과 부정에 대한 의역을 한다. 그리고 마지막으로 한국어 형태소 생성 단계에서는 음운 형상과 이형태 처리를 하며, 또한 띄어쓰기 처리를 하여 완전한 한국어 문장을 생성한다. COBALT-J/K는 현재 약 120,000 단어 (일반 용어 90,000, 인명/지명 25,000, 철강 용어 15,000)의 일-한 번역 사전을 갖추고 있으며 번역 품질면에서 현재 상용화 제품들과 비교하여 높은 번역률을 보이고 있다 [9]. URL : http : //madonna.postech.ac.kr/cobalt.html에서 COBALT-J/K에 대한 보다 자세한 정보를 얻을 수 있고 X-window Version에 대한 데모를 직접 볼 수 있다.

IV. CLTR-J/K (Cross-Language Text Retrieval from Japanese to Korean)

본 논문에서는 한국어 질의를 이용하여 일본어 문서를 검색할 수 있는 일-한 교차 언어 문서 검색을 위해, 일-한 기계번역 시스템 COBALT-J/K를 이용하여 일본어 문서를 한국어로 색인하는 문서 번역(document translation) 방식의 일-한 교차 언어 문서 검색 시스템을 제안한다. <그림 2>는 본 연구에서 제안하는 전체 시스템의 구성도를 나타낸다.



〈그림 2〉 제안된 교차 언어 문서 검색 구성도

우선 일본어 문서의 한국어 색인을 위해, COBALT-J/K의 일본어 형태소 해석 및 어휘 변환 단계를 이용하여 일본어 문서에 나타나는 일본어 어휘를 한국어 대역어로 변환한 후, 대역어와 함께 관련 품사 정보 및 의미 분류 코드를 한국어 색인기에 넘겨 준다. 그 후 일본어 문서는 계속 양상 생성과 한국어 형태소 생성 단계를 거쳐 한국어 문서로 번역되며 한국어 문서 데이터베이스에

저장된다. 이는 검색 시 사용자의 편의에 따라 일본어 원문뿐만 아니라 번역된 한국어 문서도 제공하기 위해서이다. 경우에 따라서는 이러한 번역 문서 데이터베이스를 구축할 필요 없이 사용자가 보고 싶어 하는 문서를 온라인 상에서 곧바로 COBALT-J/K를 이용하여 번역, 서비스할 수도 있다.

일반적으로 한국어 색인에서는 명사만을 색인어로 고려한다. 그러므로, 본 연구의 한국어 색인 부분에서는 일본어에 대응하는 한국어 대역어와 그 품사 정보를 가지고 명사 이외의 품사들에 대해서는 불용어(stopword) 처리를 한다. 이때, 한국어 색인 관점과 일-한 번역 관점의 차이로 인하여 대역 사전에서 추출한 한국어 어휘들을 처리해야 한다. 한국어 색인 부분에서는 다음과 같은 두 가지 처리를 통하여 이러한 차이를 없앤다.

- "명사+하다" 처리 : 한국어에서는 주로 한자어에서 유래한 명사에 "하다", "되다", "시키다" 등을 붙여서 용언(동사, 형용사)으로 사용할 수 있다. 일-한 기계번역 시스템에서는 용언에 대한 연어 패턴을 구축하기 위하여 이들 명사 겸용 용언들을 일-한 대역 사전에 등록한다. 하지만, 한국어 색인에서는 이들 용언에서 명사 부분을 색인어로 사용하기 때문에, 일-한 대역 사전에서 추출한 "명사+하다" 용언에서 명사 부분만을 분리해야 한다.
- 복합 명사 처리 : 한국어에서는 단일 명사뿐만 아니라, 단일 명사로 구성된 복합 명사도 색인어로 중요한 역할을 한다. 일-한 대역 사전에는 주로 변환의 편의를 위해서 단일 명사 형태로 등록되어 있다. 혹은 의미 분류 코드를 나눌 수 없는 복합 명사에 대해서는 복합 형태로 등록되어 있다. 이러한 인접한 단일 명사들을 복합 명사로 만들어 주어 색인을 한다. 또한 복합 명사에 대해서도 단일 명사로 분리하여 복합 명사와 단일 명사 모두를 색인어로 활용한다.

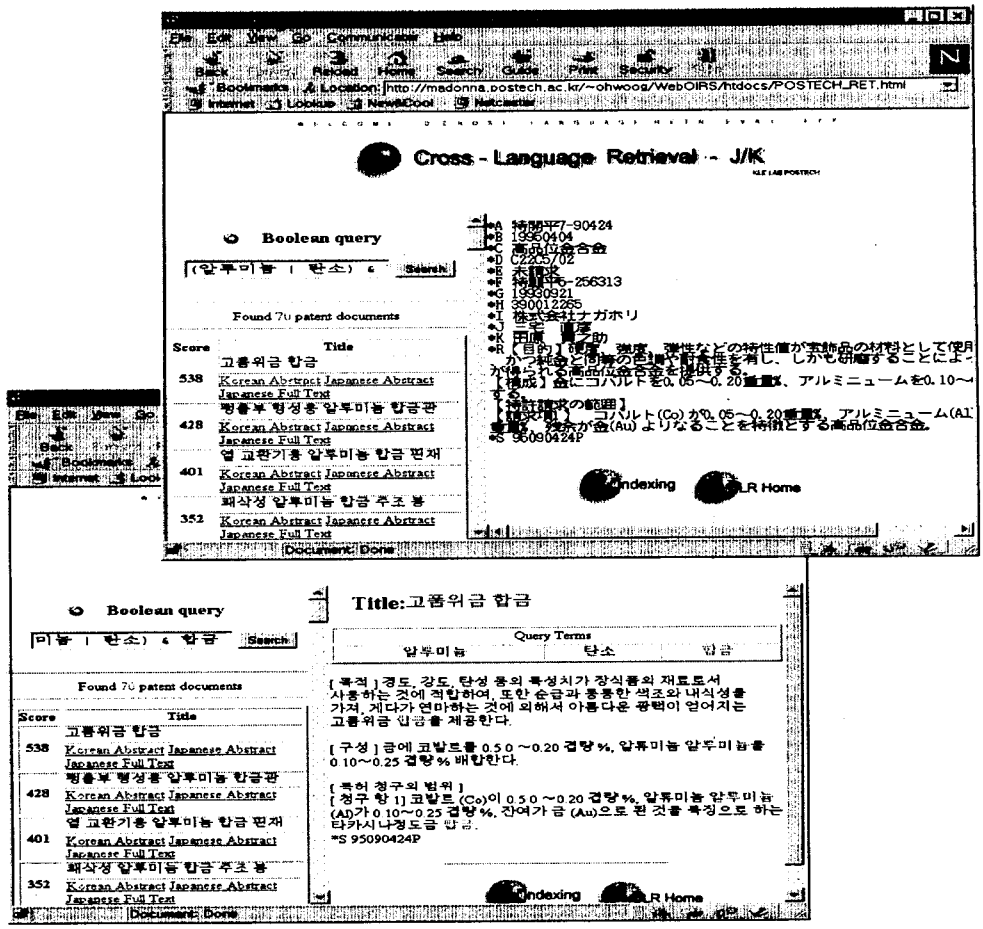
한국어 명사들을 추출한 후, 이들에 대한 가중치(weight)를 계산하다. 색인어 가중치는 전통적인 용어 빈도수(term frequency)와 역 문서 빈도수(inverted document frequency)를 이용한다. 이때, 문서의 크기에 따라 빈도의 역할이 다르므로, 용어 빈도수를 문서의 용어 수로 정규화(normalization) 한다.

검색 모델은 널리 알려진 벡터 공간 모델(vector space model)과 한국어 검색 모델로 자주 사용되는 확장 불리안 모델(extended Boolean model)을 기반으로 한다.

제안한 일-한 교차 언어 문서 검색 시스템은 2장에서 언급한 교차 언어 문서 검색에서의 기계번역 시스템의 역할 3가지를 모두 만족하도록 구성하였다. 먼저 언어 교차에 대한 역할로 일본어 문서를 한국어로 색인하여 언어 간의 일관성을 유지하였다. 또한 자국어로의 문서 선별의 용이성을 위해 번역 문서에 대한 데이터베이스를 두어 일본어 원시 문서과 함께 한국어 대역 문서도 볼 수 있게 하였다. 일-한 기계번역에서는 단어의 의미 중의성 해소를 위해서 일본어 전문가에 의하여 상당히 오랜 기간 구축된 연어 패턴을 이용하고 있다. 하지만, 기계번역에서 사용한 의미 분류 코드를 현재 교차 언어 문서 검색 시스템의 성능향상을 위해 사용하지는 않고 있다. 향후 연구에서는 이러한 의미 분류 코드를 이용한 색인과 검색 모델을 연구할 것이다.

본 연구에서 제안한 교차 언어 문서 검색 시스템 CLTR-J/K를 URL의 CGI로 구축하였다. 구축된 시스템에 대한 정보와 데모는 URL : http : //madonna.postech.ac.kr/CLTR.html에서 볼 수 있다. 데모를 위해 구축한 검색 모델은 확장 불리안 모델을 이용하였고 일본 특허 문서 2,000건에 대해 색인을 구성하였다. <그림 3>은 구축된 시스템에서 한국어 질의어에 대한 일본어 원시 문서와 번역된 한국어 문서를 보여주고 있다.

시스템의 객관적 성능 평가를 위해 '일본정보처리학회'와 '일본경제신문'이 공동으로 구축한 일본어 테스트 문서 집합을 입수하고자 했으나, 유감스럽게도 '일본경제신문'측의 국외 제공 불가 원칙으로 인해 제안된 일-한 교차 언어 문서 검색 시스템 CLTR-J/K에 대한 평가를 할 수 없었다.



〈그림 3〉 제안된 교차 언어 문서 검색 시스템의 수행 예

향후 연구에서는 독자적으로 일본어 검색 테스트 집합을 구축하여 CLTR-J/K의 성능을 평가하고, 또한 문서 번역 방식과 질의 번역 방식을 비교 평가하고자 한다. 또한 역방향의 한-일 교차 언어 문서 검색 CLTR-K/J를 위해, 질의 번역 방식에서의 일-한 기계번역 시스템의 이용과 문서 번역 방식에 대한 한-일 기계번역 시스템의 이용을 비교 평가하고자 한다.

V. 결 론

교차 언어 문서 검색에 대한 대부분의 기존 연구들은 문서 번역(document translation) 방식보다는 질의 번역(query translation)에 대한 연구가 많았다. 이러한 질의 번역들은 기존의 기계 번역 방법과는 다른 방법들로 질의의 어휘들을 번역하였다. 최근의 질의 번역 시스템들은 어휘 번역의 질을 높이기 위하여 자연 언어 처리 기법을 도입하고 있다. 하지만, 질의 번역 방식들은 정보검색 시스템에 맞게 통합하기보다는 전처리기로서의 역할을 수행하여 기존의 단일 언어 환경보다 매우 낮은 성능을 보였다.

인간만큼의 완벽한 번역은 불가능하지만 기계 번역 시스템의 꾸준히 발전으로 현재 많은 시스템들이 높은 번역률을 보이고 있는 상황이다. 따라서 본 연구에서는 주로 적은 어휘들로 구성된 질의

(query)에서 새로운 번역 방법을 모색하는 기존의 교차 언어 문서 검색 연구들과는 달리, 고품질의 번역 결과를 보이는 기계 번역 시스템을 이용하여 문서 번역(document translation) 방식의 시스템을 제안하였다. 문서 번역 방식을 교차 언어 문서 검색에 적용할 경우, 질의/문서간 어휘의 일관성 확보 및 문서 선별 기능의 확대와 더불어 정보검색 시스템과의 통합으로 인한 시스템 성능 향상을 추구할 수 있다.

참 고 문 헌

[1] L. Ballesteros and W.B., Croft "Dictionary Methods for Cross-Lingual Information Retrieval," Workshop on Cross-Linguistic Information Rretrieval following SIGIR'96, http : / / www.rxrc.xerox.com / research / mltt/DMHead/CLIR/balleste, 1996.

[2] Christian Fluhr, Dominique Schmit, Philippe Ortet, Karine Gurtner, Vera Semanova, F. Elkateb, "Distributed multilingual information retrieval," Multilinguality in the Software Industry : The AI Contribution (MULSAIC'96) Worksop, http : //www.iit.nrcps.ariadne-t.gr/~costass/muls3.html/fluhr, 1996.

[3] Chul-Jae Park, Jong-Hyeok Lee, Geunbae Lee, K. Kakechi, "Collocation-Based Transfer Method in Japanese-Korean Machine Translation," Transaction of Information Processing Society of Japan, 38(4) : pp. 707-718, 1997 (written in Japanese)

[4] M.W. Davis, "New Experiments In Cross-Language Text Retrieval At NMSU's Computing Research Lab," TREC-5 Report, http : //crl.nmsu.edu/users/madavis / Site / Book2 / book2 - toc.html / trec5.ps, 1996.

[5] M.W. Davis and T.E. Dunning, "Query Translation Using Evolutionary Programming for Multi-Lingual Information Retrieval II," In Proceedings of the Fourth Annual Conference on Evoluationary Programming, San Diego, Evolutionary Programming Society, http : / /crl.nmsu.edu/users/madavis/Site/Book2/book2-toc.html/ep96, 1996.

[6] D.A. Gachot, E. Lange and Jin Yang, "The SYSTRAN NLP Browser : An Application of Machine Translation Technology in Multilingual Information Retrieval," Workshop on Cross-Linguistic Information Rretrieval following SIGIR'96, http : // www.rxrc.xerox.com / research / mltt / DMHead/CLIR/yang, 1996.

[7] D.W. Orad and B.J. Dorr, "A Survey of Multilingual Text Rretrieval," Technical Report UMIACS-TR-96-19, Institute for Advanced Computer Studies. University of Maryland, http : //www.ee.umd.edu/medlab/filter/papers/mlir.ps, 199 6.

[8] K. Grenichiro, H. Yoshihiko, and S. Seiji, "Cross-lingual Information Retrieval on the WWW," MULSAIC 96. Multilinguality in Software Engineering : AI Contribution (in conjunction with ECAI 96), http : // isserv.tas.ntt.jp / chisho / titan - help / eng / titan-docs.html/9608KikuiMULSAIC—ps, 1996.

[9] Jung-Rak Jeong, Jung-In Kim, Kyonghi Moon, Jong-Hyeok Lee, and Geunbae Lee, "Evaluation of COBALT-J/K, Japanese to Korean Machine Translation System," in the Proceedings of the eighth Hangul & Korean Information Processing, pp.338-345, 1996. (written in Korean)

[10] K. Yamabana, K.Muraki, S. Doi, and S.

Kamei, "A Language Conversion Front-End for Cross-Linguistic Information Retrieval," Workshop on Cross-Linguistic Information Rretrieval following SIGIR'96, http : / / www.rxrc.xerox.com / research / mltt/DMHead/CLIR/yamabana, 1996.

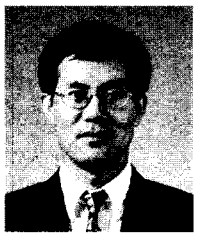
저 자 소 개

權 五 郁

1969年 4月 11日生
1992年 2月 경북대학교, 컴퓨터공학과(학사)
1995年 2月 한국과학기술원, 전자계산학과(석사)

1995年 1月～1997年 2月 포항공과대학교, 정보통신연구소(연구원)
1997年 3月～현재 포항공과대학교, 전자계산학과(박사과정)

주관심 분야 : 정보검색, 텍스트처리, 자연언어처리, 한국어정보처리



李 鐘 赫

1957年 8月 28日生
1980年 2月 서울대학교 수학교육학과(이학사)
1982年 2月 한국과학기술원(KAIST), 전자계산학과(이학석사)
1988年 8月 한국과학기술원(KAIST), 전자계산학과(공학박사)

1984年 3月～1988年 8月 고려대학교(시간강사)
1985年 3月～1989年 2月 단국대학교(사간강사)
1988年 9月～1989年 8月 한국과학기술원(박사 후 연구연구원)
1989年 11月～1991年 1月 일본전기(NEC), 중앙연구소(초청연구원)
1991年 2月～1996年 9月 포항공과대학교, 전자계산학과(조교수)
1996年 10月～현재 포항공과대학교, 전자계산학과(부교수)

주관심 분야 : 자연언어처리, 한국어정보처리, 기계번역, 정보검색

李 根 培

1961年 3月 20日生
1984年 2月 서울대학교 컴퓨터공학과(학사)
1986年 2月 서울대학교, 컴퓨터공학과(석사)
1991年 2月 UCLA(Ph.D-Computer Science)

1984年 3月~1986年 2月 RA, 서울대학교
1987年 3月~1991年 2月 RA, UCLA
1991年 3月~1991年 9月 Research scientist, Biology department, UCLA
1991年 9月~1997年 3月 포항공과대학교, 전자계산학과(조교수)
1997年 4月~현재 포항공과대학교, 전자계산학과(부교수)

주관심 분야 : Netural language processing, Human-computer interaction, Intelligent agent & text-based information retrieval, Automatic speech recognitioin