관광 정보 검색을 위한 자연언어 질의 해석 시스템 구현*

김 명철, 서 광준, 전 경헌, 최 기선
한국과학기술원 전산학과

A design and inplementation of query processor for travel information retrieval system

Myong-cheol Kim, Kwang-jun Seo, Kyong-hun Jeon, key-sun Choi
Department of Computer Science, KAIST

요 약

본 논문은 관광 정보 검색용 한국어 자연언어 질의 해석 시스템의 모델 정립 및 구현에 대한 것이다. 본 자연언어 질의 해석 시스템은 질의로 부터 정보 검색 시스템의 검색어들을 추출한다. 이를 위하여 1만 단어 수준의 중형사전을 구축하였으며, 불용어 사전, 전거어 사전, 유사어 사전, 복합명사 사전을 구축하였다. 사전의 어휘를 추출하기 위해서 한국어 대화체 문장에 대한 자료 수집과 분석을 하였으며, 관광 정보 검색 시스템의 텍스트를 분석하였다. 200여 자연언어 질의 문장으로 실험한 결과는 비교적 좋았다.

I. 서 론

컴퓨터의 사용이 광범위해짐에 따라 컴퓨터 사용자의 계층도 초보자에서 전문가까지 다양해지고 있다. 어떤 사용자가 컴퓨터 시스템을 처음 사용하려고 할 때, 가장 먼저 접하는 부분이 사용자 인터페이스 부분이므로 이 부분의 성패가 전체 시스템의 성패를 좌우하기도 한다. 자연언어 인터페이스 는 사용 비용이 비교적 많이 든다는 단점이 있기는 하나, 초보자의 입장에서는 새로운 시스템의 적응에 있어서 어떤 학습도 필요하지 않다는 강력한 장점이 있다. 또한 전문가 입장에서도 자기가 필요한 자연언어 명령어만을 간결하게 사용하여 원하는 결과를 얻을

---

* 본 연구는 한국전자통신 연구소의 위탁과제로 수행되었음

수 있는 등, 자연언어의 특징인 유연성(flexibility), 명료성(succintness), 풍부한 표현력(express-power)등을 그대로 간직하고 있다. 이와 같은 이유로, 자연언어 인터페이스는 차세대 사용자 인터페이스로 많은 연구가 진행되고 있으며 외국의 경우 몇몇 상용화된 시스템들*이 등장하고 있는 실정이다.

본 연구의 목표는 관광 정보 검색 시스템의 자연어 인터페이스를 위한 한국어 질의어 해석기와 이에 필요한 중형사전의 개발하는 것인데, 관광 정보로는 '세계를 간다(유럽 14개국편)'[6]을 사용하였다. 이를 위하여 한국어 질의를 포함한 대화를 수집, 분석하였으며 정보검색 corpus를 문서 분석 시스템으로 분석하여 질의 해석에 필요한 어휘를 선정하였다.

본 한국어 질의어 해석기는 형태소 해석 수준의 자연언어 처리를 하여 질의어에서 명사 검색어를 추출한다. 그리고 정보 검색의 효율성 및 정확성을 높이기 위하여 불용어(Stop Word) 정보, 전거(Authority) 정보 및 시소러스(Thesaurus) 정보를 이용한다.

본 논문의 구성은 2장에서 사전 작성을 위한 한국어 대화체 분석 및 정보검색용 corpus분석을 기술하고 3장에서는 한국어 질의 분석을 위한 중형 사전에 대하여 서술한다. 4장에서는 한국어 질의 해석기에 대하여 설명하며 5장에서는 실험 및 결과에 대하여 기술하고 6장에서 결론을 맺는다.

II. 한국어 corpus 분석

자연언어 질의해석을 위한 사전 정보를 추출하기 위하여 대화 corpus와 실제로 검색 시스템에서 사용한 '세계를 간다(유럽 14개국편)' corpus를 한국어 문서 분석 시스템(KOCP)으로 분석하였다.

1. 대화 corpus 분석

본 연구를 위하여 수집한 대화 corpus는 약 200명에게 가상의 정보 검색 시스템이 주어질 경우에 가능한 대화를 설문 형식으로 조사하였다. corpus의 크기는 1000여 개의 질의/응답 2000여 문장 정도 된다. KOCP로 corpus의 어휘를 분석한 후, 그 빈도 정보에 의해 사전 어휘 및 전거어 등의 정보를 얻을 수 있었다.

---

* 데이타 베이스 전위 시스템으로 사용된 LUNER, PLANES, EUFID, REL 등이 있으며, CAI에 사용된 UC(Unix Consultant) 등이 있다.

2. 정보 검색용 corpus 분석

정보 검색용 corpus로는 '세계를 간다(유럽 14개국편)'[6]을 사용하였다. 이 corpus의 크기는 18124 line, 34144 word이다. 이 corpus를 KOCP로 분석한 결과 약 11000단어의 명사 어휘를 얻었으며, 영어 어휘가 4000여개 포함되어 있어 이를 제외한 7000여개의 어휘를 사전에 등록할 어휘로 선정하였다.

III. 질의어 분석을 위한 한국어 중형 전자 사전

일반적으로 전자 사전이라 함은 출판물의 형태로 만들어진 언어사전을 다시 기록매체에 수록하여 어휘와 관련된 제반 언어 정보를 기계 특히 컴퓨터를 통하여 얻어낼 수 있도록 한 것을 가리킨다. 본 시스템의 사전은 크게 주 사전인 형태소 해석 사전과, 질의어 해석을 위하여 보조적으로 포함되어있는 복합 명사 사전, 불용어 사전, 전거 사전, 유사어 사전으로 이루어져 있다. 형태소 해석기는 형태소 해석 사전을 이용하여 사용자의 질의를 형태소 해석하고, 질의어 해석기는 불용어 사전, 전거어 사전, 유사어 사전에 있는 정보를 이용하여 질의어를 가공하여 적절한 검색어 집합을 생성하게 해준다.

1. 형태소 해석 사전

형태소 해석 사전은 형태소 간의 접속 가능성을 조사하기 위해 필요한 각 형태소의 좌,우 접속 범주 정보를 저장하고 있다. 사전의 한 항목는 형태소와 그에 대한 좌우 접속 정보 쌍들로 구성되어 있다. 형태소 해석 사전은 명사 사전과 기능어 사전으로 나누어져 있는데, 그 이유는 사전의 많은 부분을 차지하는 명사들의 좌우 접속 정보가 공통점이 많아 분리하면 그것들을 생략할 수가 있기 때문이다. 또한 처리의 관점에서는 필요에 따라 구분해서 접근을 하면 사전 접근의 시간을 줄일 수가 있다. 명사 사전은 일반적인 형태소 분석에서 사용되어지는 기초 명사와 특별히 '세계를 간다 : 유럽편'[6]과 대화 corpus에서 추출해낸 7,000여 개의 한글 단어로 이루어져 있다. 기능어 사전에는 명사 이외 품사의 형태소 정보를 포함하고 있다.

2. 질의어 해석을 위한 사전

질의어 해석을 위한 사전은 질의어의 올바른 처리를 목적으로 하고 있기 때문에, 올바른 질의어의 이해와 빠른 해석을 위하여, 불용어 사전, 전거어 사전, 유사어 사전이 필요로 된다.

가) 복합 명사 사전

복합 명사 사전은 초기 검색어 집합을 구성하는데 사용되는 사전으로 단순히 복합 명사만 나열된 사전이다. 이러한 복합 명사 사전으로 붙어 있는 명사들을 단일 명사 검색어로 분리한다거나 떨어져 있는 명사들을 복합 명사 검색어로 형성하는데 이용된다. 복합 명사 사전의 한 항목은 하나의 단어만으로 구성되고, 한 단어를 저장할 수 있는 충분한 크기로 고정이 되어 있다. 저장된 단어들은 정열이 되어 있다.

나) 불용어 사전(Stop Word Dictionary)

불용어는 단어 자체의 의미가 적어서, 적절한 검색어의 역할을 할 수 없는 단어를 말한다. 본 시스템에 쓰인 불용어 사전의 크기는 약 1,400여 개이다. 불용어 사전의 구조는 복합명사 사전과 동일하다.

다) 전거 사전(Authority Dictionary)

전거어란 외국어 단어에 대한 한국어 표기의 다양성으로 만들어 지는 여러개의 한국어 단어에서 표준으로 사용되는 것을 말한다. 사용자가 전거형이 아닌 변형어로 표현했을때도 같은 object를 지칭하는 것으로 처리할 수 있다면 검색의 효율이 좋아질 것이다. 즉, '프랑스'라고 지칭하는 object와 '불란서'나 '쁘랑스'라고 지칭하는 object는 같다고 처리 한다는 것이다. 본 시스템에서는 주로 인명이나 장소명의 전거어를 처리하는데, 이것을 위하여 key의 역할을 하는 전거어와 그 표준 단어로 구성된 단어 쌍을 항목으로 갖는 전거어 사전을 갖고 있다. 전거 사전의 각 항목은 변형과 전거형의 쌍으로써 이를 표현하였고, 두개의 단어를 저장할 수 있는 충분한 크기로 고정이 되어 있다. 변형 단어는 탐색의 원활함을 위하여 정렬되어 key로 사용된다.

라) 유사어 사전

유사어 사전은 질의어에 나타난 검색어가, 시스템에서 유용한 정보로 탐색되어질 수 있도록, 유사한 의미의 단어들을 같이 탐색하게 하기 위하여 사용되어진다. 유사어는 다음의 두 가지 종류로 나누어진다.

- 동의어(Synonym) : (배, 선박)
- 관련어(Related Word)

  Broad Term: (기차, 철도),

  Narrow Term: (숙소, 호텔),

  Related Term: (유명지, 관광지)

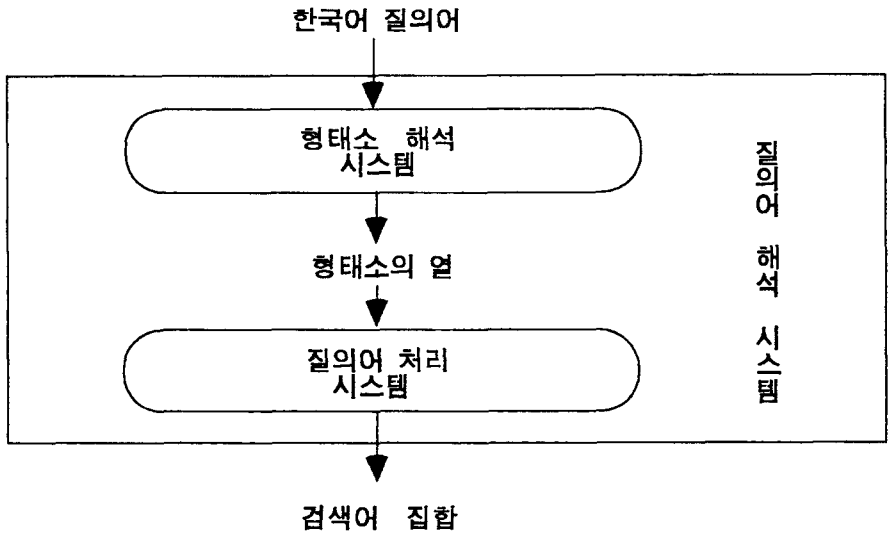
만약, '기차'가 검색어로써 추출되었다면, '철도'도 마찬가지로 검색어가 된다.

유사어 정보는 유사어 집합으로 표현이 되는데, 이러한 유사어 집합을 중복 없이 그리고 검색이 쉽게 할 수 있도록 본 시스템에서는 다음과 같은 사전 표현을 갖는다. k개의 원소를 갖고 알파벳 순으로 정렬된 유사어 집합 〈유사어1, 유사어2, . . , 유사어k〉 대한 사전 내에서의 표현은 다음과 같다.
〈유사어1, 유사어2의 포인터〉, 〈유사어2, 유사어3의 포인터〉, 〈유사어k-1, 유사어k의 포인터〉, 〈유사어k, 유사어1의 포인터〉.

이 사전의 각 항목도 전거 사전과 마찬가지로 두 단어를 충분히 저장할 수 있는 크기로 고정되어 있다.

IV. 한국어 질의어 해석 시스템

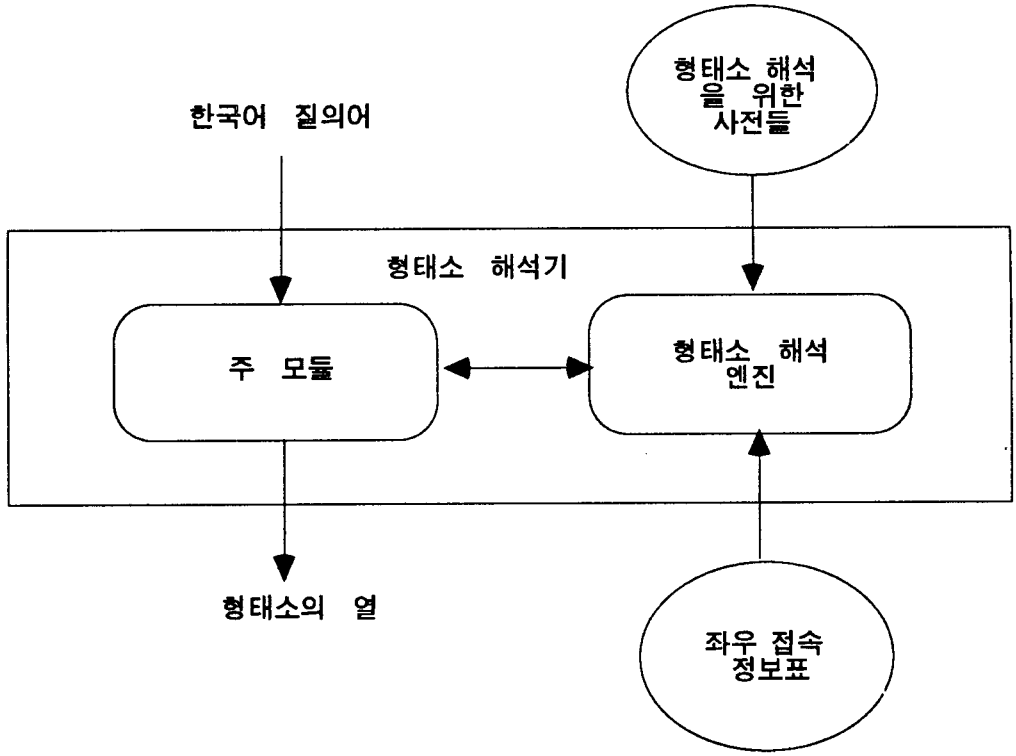
본 한국어 질의어 처리 시스템은 자연언어 질의를 입력으로 받아서 그 자연언어 질의에 대한 검색어 집합을 형성하는 시스템이다. 본 시스템은 입력 질의의 모든 어절을 형태소 해석하는 형태소 해석 시스템과 그 형태소 해석을 입력으로 하여 검색어 들을 추출하는 질의어 처리 시스템으로 구성된다.



〈그림 4. 1〉 질의어 해석 시스템 구조

1. 형태소 해석 시스템

본 형태소 해석 시스템은 질의어 해석 시스템의 전반부를 구성하며 한 개의 질의를 입력으로 받아 각 어절을 형태소 해석하여 질의어 해석기에 필요한 형태소의 열을 만들어 낸다.



〈그림 4.2〉 형태소 해석 시스템

〔그림 4.2〕에서 출력인 형태소의 열이란 문장을 구성하는 모든 형태소와 그 형태소의 품사들(한 형태소는 여러가지 품사를 가질 수 있다)이다. 예를 들면,

질의 : "불란서에서 자동차 여행 장소로 적합한 곳은 어디 입니까?"

형태소 열 : 〈 불란서(명사), 에서(격조사), 자동차(명사), 여행(명사), 장소(명사), 로(격조사), 적합(명사), 하(조용보조어간), ㄴ(관형사형어미), 곳(명사), 은(보조사), 어디(대명사), 입니까(서술격 조사) 〉

가) 좌우 접속 정보표

좌우 접속 정보표는 형태소들의 좌우 접속 형태에 따라 분류, 범주화하여 나타낸 표이다〔12,13〕. 좌 접속 분류는 형태소의 오른쪽에 붙을 수 있는 형태소들의 종류에 의해 작성되었고, 마찬가지로 우 접속 분류는 형태소의 왼쪽에 붙을 수 있는 형태소들의 종류에 의해 구분되었다.

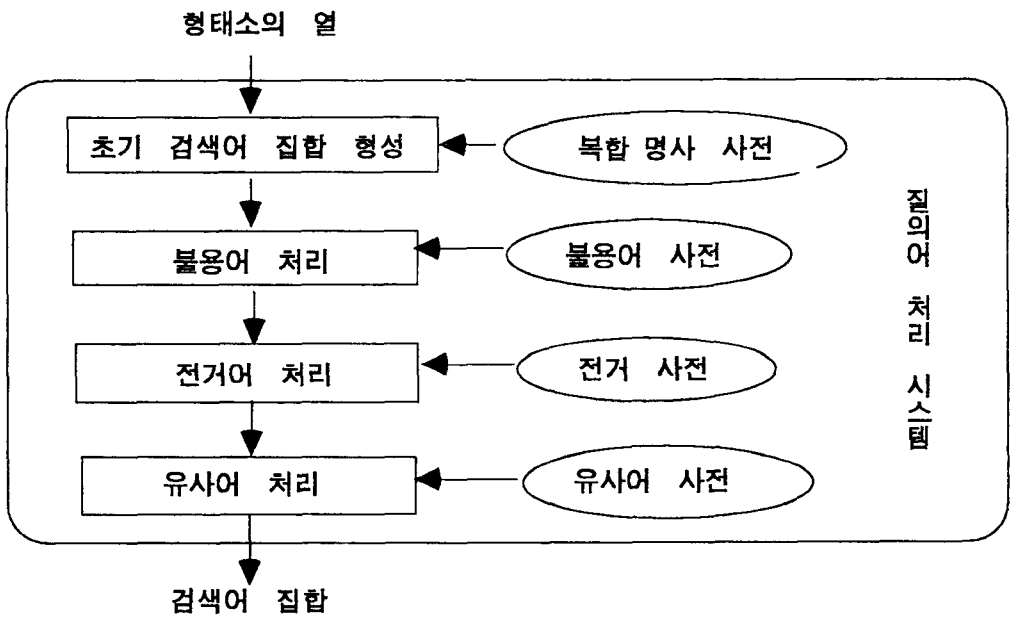
나) 형태소 해석기

형태소 해석기는 주 모듈과 형태소 해석 엔진으로 이루어져 있다. 형태소 해석 엔진은 한 어절이 주어지면 그에 대해 앞에서 부터의 최장 일치 형태소 해석 결과를 찾거나 그것에 실패하면 미 등록어 처리를 하는 모듈이고, 주 모듈은 질의를 받아들여 어절 단위로 구분한 후 형태소 엔진의 입력을 만들고 그

출력들을 모아 질의에 대한 형태소의 열을 만들어 내는 일을 한다.

2. 질의어 해석기

질의어 해석기는 형태소 해석기로부터 질의 단위로 형태소 리스트를 입력으로 받아 초기 검색어 집합을 형성하고, 불용어 처리, 전거어, 그리고 유사어 처리를 하여 향상된 검색어 집합을 만든다. 질의어 해석기는 초기 검색어 집합 생성 후에 검색어 집합을 처리하는 독립적인 여러 단계로 구성되어 있는데 각 단계 마다 하나의 처리를 한다. 따라서 필요에 의한 새로운 단계의 삽입, 삭제가 자유롭다.



〈그림 4.3〉 질의어 처리 시스템 의 구성도

가) 초기 색인어 집합 형성

이 단계에서는 형태소 열을 조사하여 독립적으로 나타나는 명사를 가지고 단일 명사 검색어, 복합 명사 검색어를 구성하여 초기 검색어 집합을 형성한다. 복합 명사 검색어는 형태소의 열에 연속으로 나타나는 명사들을 조합하여 복합 명사 사전에 존재하는가를 검사하여 형성 하는데 그것을 위하여 형태소의 열에 나타나는 연속된 명사들을 저장하는 리스트인 "명사그룹"이라는 중간 구조를 사용한다. 이 단계는 명사그룹 집합을 형성하는 단계와 그 집합으로부터 초기 검색어 집합을 형성하는 과정으로 구성되어 있다. 위에서 든 예로 보면,

명사그룹 집합 : { 〈불란서〉, 〈자동차, 여행, 장소〉, 〈적합 〉, 〈곳〉 }

초기 검색어 집합 : { 불란서, 자동차 여행, 장소, 적합, 곳 }

단, 복합명사 사전에 "자동차 여행"이 있다고 가정

나) 불용어 처리

불용어 처리는 검색어 집합의 용어들이 불용어 사전에 있으면 검색어 집합에서 제거하는 일을 한다. 불용어 사전의 검색은 사전의 레코드 사이즈가 고정되어 있고, 레코드 수도 많지 않아서 사전의 화일 포인터를 가지고 직접 이진 검색을 한다. 앞의 예에서 "적합"과 "곳"이 불용어로 제거된다.

다) 전거어 처리

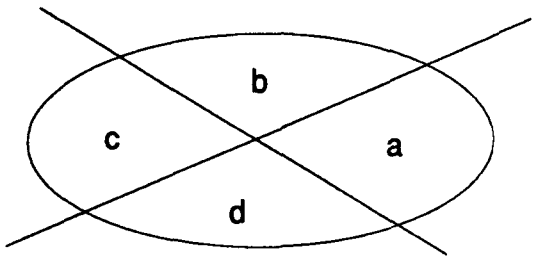
전거어 처리에서는 검색어 집합의 모든 단어에 대하여 전거어 사전 항목의 key로 사용되는가를 조사하여, key로 사용되는 단어에 대해서는 그 항목의 표준 단어로 바꾸어 주게 된다. 사전의 검색은 불용어 사전과 마찬가지 이유로 이진 검색을 한다. 위의 예에서 "불란서"가 전거형인 "프랑스"로 바뀌게 된다.

라) 유사어 처리

이러한 사전 구성에서 한 단어와 유사한 모든 단어를 찾는 방법은 먼저 사전에 그 단어를 key로 갖는 항목을 찾고, 그 항목의 유사어 포인터를 따라가서 자신의 항목으로 되돌아 올 때까지 검색하는 모든 항목의 key를 유사어로 결정한다. 유사어 사전의 검색도 주어진 key의 항목을 이진 검색으로 찾는다.

V. 실험 및 평가

본 연구에서 개발된 한국어 질의어 해석기의 실험을 수집한 대화 corpus의 질의어 약 200문에 대하여 수행하였다. 입력으로 사용된 질의어의 문장당 평균 어절 수는 5.21이고 출력으로 얻은 검색어의 수는 평균 2.45개이다. 이는 사용자가 자연언어 질의를 할때 보통 2 - 3개의 검색어를 사용한다는 것을 암시하고 있다. 시스템의 평가를 위하여 정보 검색의 유용한 평가 기준인 정확률(Precision)과 재현율(Recall)을 〔그림 5.1〕과 같이 정의한다.



a + b : 추출된 검색어
a + d : 추출해야될 검색어

$$\text{Precision} = \frac{a}{a+b}$$

$$\text{Recall} = \frac{a}{a+d}$$

〈그림 5.1〉 Precision과 Recall

즉, Precision은 시스템이 얼마나 정확하게 검색어를 추출하는가의 평가 요인이 되며, Recall은 얼마나 잘 뽑아내는 가의 척도가 된다. 209개의 질의어에 대한 실험의 결과는 〔표 5.1〕과 같다.

항	갯 수	평균 검색어의
추출해야 할 검색어 수	513	2.45
추출된 검색어 수	513	2.45
옳은 검색어 수	462	2.21

〔표 5.1〕 시스템의 결과

이와 같이 하여 Precision은 90%, Recall은 90%의 좋은 결과를 얻었다. 추출되지 않아야 할 검색어가 추출된 주된 이유는 이유는 형태소 해석 사전의 불완전에 있다. 장래에는 사전정보의 보완으로 더 좋은 결과를 얻을 수 있을 것이다.

VI. 결 론

본 연구에서는 관광 정보 검색용 한국어 자연언어 질의 해석 시스템의 설계 및 구현 하였다. 자연언어 질의 해석 시스템은 크게 질의 해석기와 이에 필요한 정보를 가지고 있는 각 종 사전으로 구성된다.

질의 해석기는 자연언어 질의문을 입력받아 검색하고자 하는 검색어들의 집합을 출력한다. 이는 다시 형태소 해석 모듈과 질의어 처리 모듈로 구성되어 있다. 형태소 해석 모듈은 질의문의 형태소 해석을 담당하며 질의어 처리 모듈은 불용어 처리, 전거어 처리, 유사어 처리 및 복합명사의 처리를 수행한다. 사전 내용으로 질의문의 형태소 해석을 위해서는 명사사전과 명사이외의 품사들의 정보를 갖는 기능어 사전으로 분리된 1만단어 수준의 중형사전을 구축하였으며 검색의 정확성을 높이기 위하여 불용어 사전, 전거어 사전, 유사어 사전 및 복합 명사 사전을 사용하였다. 사전 정보의 추출을 위해서는 한국어 대화 corpus를 수집하여 분석하였으며 본 연구의 목적시스템인 유럽 관광 정보 검색시스템의 corpus도 분석하였다. corpus 분석용 도구로는 한국과학기술원에서 만든 KOCP를 사용하였다.

실험 결과로는 200여 자연언어 질의 문장에 대하여 원하는 검색어의 추출에 대한 Precision이 90%, Recall이 90%로 나왔다. Precision 이란 추출된 검색

어의 정확성을 나타내며 Recall 은 추출되어야 할 검색어에 대한 실제 추출된 검색어의 비율이다.

앞으로의 연구과제로는 한국어의 의미해석을 위한 의미구조 모델 정립 과 컴퓨터 도움 시스템의 모델 정립 및 개발에 의한 이식성이 있는 자연언어 인터페이스의 개발 및 한국어 대화 시스템의 개발이 남아 있다.

VII. 참 고 문 헌

[1] G. Saton, "Automatic Text Processiong", Addison-Wesley Publishing Company, 1989.

[2] R. H. Fowler, W. A. L. Fowler, B. A. Wilson, "Integrationg Query, Thesaurus, and Documents Through a Common Visual Representation", Proc, of the Fourteenth Annual International ACM/SIGIR Conference, pp 142-151, Octover 1991.

[3] 한국과학기술원, "멀티미디어 입출력에 관한연구", 한국전자통신연구소, 1991.6.

[4] 한국과학기술원,"자연언어 인터페이스를 위한 도구 환경의 연구 개발", 한국과학재단, 1992.6.

[5] 한국과학기술원, "지능형 정보검색에 관한 연구", 한국통신, 1991.

[6] 중앙일보사 편집부, "세계를 간다-유럽14개국편", 중앙일보사, 1989.

[7] 정진성, "단일문서내에서의 언어및 통계정보를 이용한 자동색인", 한국과학기술원 석사학위논문, 1992.

[8] 김광해, 유사어 반의어 사전, 한샘, 1987.

[9] 최기선, "한글 문서를 위한 자동 색인어 검출 시스템 개발", 한국 데이타통신, 1991.

[10] 다국어 DB를 위한 Keyword 및 Index 생성 시스템 개발", 과학기술처, 1991.

[11] M. Bartschi, "An Overview of Information Retrieval Subjects", IEEE Computer, May 1985.

[12] 강재우, "접속 정보를 이용한 한글 철자 및 띄어 쓰기 검사기의 설계 및 구현", 한국과학기술원 석사학위 논문, 1990.

[13] 이종혁, "한국어의 접속 정보표", 1990.