(제 10회 한글 및 한국어 정보처리 학술대회)

유한 오토마타를 이용한 정보 추출 시스템의 구현 및 분석

오효정, 임정묵, 이만호, 맹성현
충남대학교 컴퓨터과학과
대전시 유성구 궁동 220번지, 우: 305-764
{dol, jmlim, mhlee, shmyaeng} @cs.chungnam.ac.kr

An Information Extraction System Using Finite State Automata

Hyo-Jung Oh, Jeong-Mook Lim, Mann-Ho Lee, Sung-Hyon Myaeng
Department of Computer Science,
Chungnam National University

요약

인터넷의 사용자가 폭발적으로 증가함에 따라, 인터넷을 이용한 다양한 정보 서비스가 생성되었으며, 이로 인해 일반 사용자들이 접할 수 있는 디지털 문서의 양은 기하 급수적으로 증가 되었다. 본 논문에서는 유사한 정보를 갖는 다량의 문서들로부터 사용자가 원하는 정보만을 추출하는 정보 추출 시스템의 개발 과정 및 결과를 기술한다. 개발된 시스템은 필요한 정보를 포함하는 문장들을 걸러 낸 후, 필요한 사실정보의 출현을 나타내는 패턴을 사용한 유한 오토마타를 통하여 사용자가 원하는 정보를 추출한다. 관광지 안내 텍스트를 대상으로 한 실험 및 분석 결과를 기술한다.

1 서 론

인터넷의 발달로 인해 사용자가 접할 수 있는 자연어 텍스트의 양이 증가됨에 따라 필요한 정보만을 추출하는 정보 추출 시스템의 필요성이 증대되고 있다.

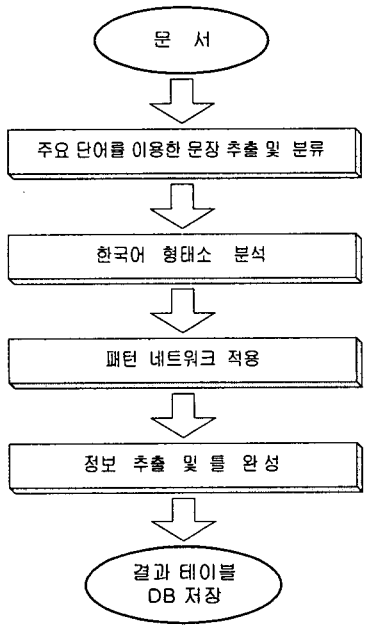
정보 추출이란 정보 검색과 다른 개념으로서, 특정 분야의 내용을 담은 자연어 텍스트로부터 원하는 정보만을 찾아내어 데이터베이스화하는 작업을 말한다[1][5]. 즉 추출할 정보의 종류를 미리 틀(template)로 정의하고, 텍스트 분석을 통해 틀을 메꾼 후 채워진 틀을 DB에 저장하는 작업이다. 틀은 해당 분야의 전문가가 정의하기도 하고, 학습 데이터로부터 문서의 주제를 학습한 후 각 주제에 해당하는 틀을 만들기도 한다.

정보 추출은 정보 검색 환경에서 그 필요성이 가장 절실하게 나타난다. 예를 들어 인터넷 상에서 검색 엔진을 사용하여 질의를 하는 경우 검색된 문서의 수가 10,000건을 넘는 경우가 보통인데 사용자가 이들을 모두 읽어 보면서 적절성을 판단하는 것은 거의 불가능하다. 이 때 필요한 정보만을 추출하여 수십 쪽의 문서를 하나의 틀로 보여 주는 기능은 정보의 과적재(information overload)를 피하기 위해 필수적이다[2][3].

본 연구에서 개발한 정보 추출 시스템은 특정 분야의 정보를 갖는 일반 문서로부터 사용자에게 필요한 정보를 추출하는 시스템으로 사용된 데이터는 관광지 정보를 갖는 일반 텍스트 파일이며, 이것으로부터 추출되는 정보는 교통정보, 숙박정보, 요금정보 등이다. 이러한 정보를 알아내기 위하여 31개의 학습 문서로부터 추출되어야 할 정보를 갖는 문장에 나타나는 어휘 패턴을 조사하였으며, 이러한 패턴을 실험 데이터에 적용 함으로서 필요한 정보를 어느 정도 정확하게 추출하는가에 대하여 실험 하였다.

본 연구에서 개발한 시스템의 흐름도는 [그림 1]이며, 각 단계에 대한 설명은 다음과 같다.

- 특정 분야의 문서 내에서 실질적으로 정보를 포함하는 문장을 걸러내어 주제별로 분류한다.
- 분류된 문장을 형태소 분석한다.
- 문장을 주제별 패턴 네트워크에 적용시킨다.
- 주제별 패턴 네트워크를 통해 추출된 정보로 틀을 완성한다.



[그림 1] 정보추출 시스템 구성도

이렇게 문장 패턴의 정의를 사용한 유한 오토마타를 통해 정보를 추출하는 방법은 영어권에서는 이미 적용되어 그 효율성이 입증되었다[11]. 그러나 한국어는 대부분의 문장 성분이 몇몇 주요 문장 성분을 제외하고 위치의 제약을 받지 않으며, 조사의 유무가 자유롭기 때문에, 영어에 비해 지역적인 규칙성을 발견하기 어렵다. 본 논문의 주안점은 유한 오토마타 기반 정보 추출 방법론이 얼마나 효율적으로 한국어 문서에 적용될 수 있는지에 대한 가능성을 타진하는데 있다. 이를 위해 비교적 제한된 영역의 문서 집합을 대상으로 추출 시스템을 구축하고 실험을 통해 그 효용성을 측정한 후 이러한 단순한 방법론의 한계성을 구명한다.

2 관련 연구

탐색 가능한 정보의 급증으로 정보검색 시 사용자의 질의에 대한 결과는 사용자가 일일이 결과 문서의 내용을 파악하기 힘들 정도로 많은 양이 제공되고 있다. 이러한 문제점을 해결하기 위해 문서의 내용을 자동적으로 요약하는 문서 요약 시스템과 문서에서 필요한 정보만을 자동적으로 추출하는 정보 추출 시스템에 대한 연구가 많이 진행되고 있다. 현재 개발되어 있는 자동 요약 시스템에는 통계적 접근 방법을 사용한 Kupiec[6], 주제기반 접근 방법을 사용한 Brasilay와 Elhadad의 연구[7]와 이를 혼합한 기법을 사용한 Hovy[8] 등이 있다.

많은 양의 문서에서 틀을 이용하여 필요한 정보만을 추출하는 정보 추출 방법론으로는 문서내의 모든 문장에 대한 종합적인 파싱과 의미 해석을 통해 문서내의 정보를 알아내는 방법[9][10]과 정보를 중심으로 부분적인 파싱을 통해 정보를 추출하는 방법[11][12][13]이 있다. 전자의 경우 문서 내용의 의미적 표현을 생성한 후 정보를 추출할 수 있으나 분석에 들인 노력에 비해 그 효과가 그리 높지 않은 편이다. 이에 비하여 부분적인 파싱을 통해 정보를 추출하는 방법은 질의 응답 등에 필요한 문서 내용의 상세분석이 이루어지지는 않지만 전체 문서에 대한 파싱을 하지 않고도 정보 추출 작업을 효과적으로 수행할 수 있다. 부분적인 파싱을 이용하여 정보를 추출하는 정보 추출 시스템에는 유한 오토마타를 사용하는 FASTUS[11], 통계적인 언어모델을 이용한 BBN's PLUM[12]과 구문 패턴 매칭을 이용한 SNAP[13] 등이 있다.

정보 추출 시스템은 그 자체로도 유용하게 사용되지만, 이를 통해 추출된 정보는 자동 요약의 목적으로 사용될 수 있다. 이러한 요약 시스템은 생성된 틀에 포함된 정보를 기반으로 문장을 생성한다[14].

이러한 정보 추출 시스템의 개발은 영어권에서는 매우 활발히 진행되고 있으나 한국어에 대한 연구는 거의 없는 실정이다.

3 정보 추출 시스템

본 연구에서 개발한 정보 추출 시스템은 주요 단어를 이용하여 문장의 추출 및 분류 단계, 형태소 분석 단계, 패턴 네트워크 적용 단계와 틀 완성 단계로 구성된다.

3.1 주요 단어를 이용한 문장 추출 및 분류

데이터 파일들로부터 필요한 정보를 추출하기 위해서, 우선 불필요한 문장을 제거하는 작업이 필요하다. 일반적으로 정보를 갖는 문장은 전체 문서 중 일부분에 불과하며, 또한 이들 중요 문장은 정보의 종류에 따라 특정 단어를 포함하는 경우가 많다. 본 단계는 정보를 포함하는 문장이 갖는 이러한 특징을 이용해서 전체 문서 중 관심 있는 정보에 따라 중요 문장을 추출 하고, 정보의 종류에 따른 문장의 패턴을 각각 만들 수 있게 하는 전처리 과정이다. 31개의 학습 문서를 이용하여 사용자가 원하는 정보를 포함하는 문장이 공통적으로 갖는 주요 단어들을 선택하였다. 틀의 정보 항목(slot)과 관련 있는 주요 단어들은 다음과 같다.

정보	주요 단어
교통정보	출항, 운항, 운행, 대중교통
요금정보	화폐단위 (백원, 천원, 만원)
숙박정보	민박, 숙박
별미정보	명물, 풍부, 특산물, 맛집 등

[표 1] 정보에 따른 주요 단어와 문장 분류

데이터 파일을 문장 단위로 분리한 후, 중요 단어를 이용하여 정보의 종류에 따라서 문장을 분류하고, 주요 단어를 포함하지 않는 문장은 이 과정에서 제외된다. 따라서 불필요한 문장에 대한 처리를 하지 않아도 되는데, 실제로 이 과정을 통해 처리해야 할 문장의 수가 40%로 줄어들었다.

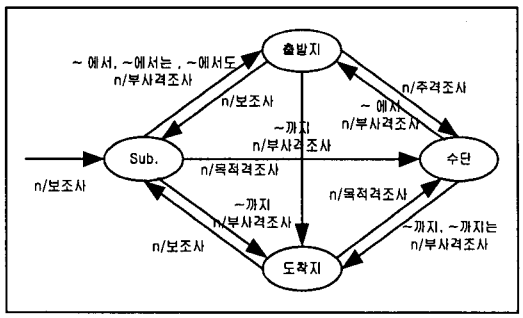
걸러진 문장은 품사 태거에 의해 문장 내의 품사가 결정된다. 본 연구에서는 한국어 품사 태거를 직접 구현하지 않고, 고려대학교에서 개발한 품사 태거를 이용하여 데이터 문서에 대한 품사 태깅을 하였다[4].

3.2 패턴 인식에 의한 정보 추출

3.2.1 유한 오토마타 적용

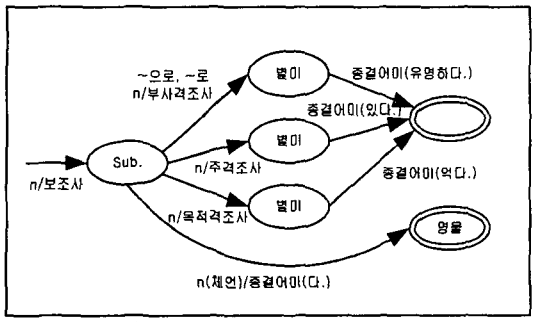
이 과정은 본 시스템에서 가장 중요한 단계로 품사 태깅된 문장이 유한 오토마타로 표현된 주제별 패턴 네트워크에 적용된다. 각 문장은 그 종류에 따라서 미리 정의된 주제별 네트워크에 의해 처리된다. 즉, 교통 정보를 갖는 문장은 교통 정보 네트워크에, 요금 정보를 갖는 문장은 요금 정보 네트워크, 숙박 정보를 갖는 문장은 숙박 정보 네트워크, 별미 정보를 갖는 문장은 별미 정보 네트워크에 각각 적용된다. 주제별 네트워크를 거쳐 추출된 정보로 틀이 완성된다.

다음은 본 시스템에서 사용되는 네트워크 중 교통 정보와 별미 정보 네트워크 및 이들 네트워크에 의해 처리되는 문장의 예이다.



[그림 2] 교통 정보 문장 네트워크

네트워크 내의 각 노드(Node)는 정보를 나타내며 노드 간의 아크(Arc)는 어절 내의 조사가 된다. 교통 정보 문장을 간단한 예로 들어보면, "대중교통은 서울에서 제주도까지 가는 비행기를 이용한다." 의 경우 '대중교통은'은 '대중교통+은'이므로 '대중교통'이 Sub. 노드에 들어간다. 마찬가지로 '서울'은 출발지 노드에, '제주도'는 도착지 노드에, '비행기'는 수단 노드에 들어간다.



[그림 3] 별미 정보 문장 네트워크

별미 정보 문장 네트워크를 통해 정보가 추출되는 과정의 예를 보면, "춘천은 춘천 닭갈비로 유명하다"의 경우 '춘천은'은 '춘천'이 명사이고 '은'이 보조사 이므로 '춘천'은 Sub노드에, '춘천 닭갈비'는 별미 노드에 들어 간다.

이런 과정을 통해 추출된 정보는 다음 단계에서 가공되어 틀에 저장된다.

3.2.2 패턴 네트워크 생성

패턴 네트워크를 생성하기 위하여 31개의 학습 문서로부터 정보를 포함하는 문장이 형태적으로 어떤 패턴을 갖는지 조사하였다. 특정 정보에 대한 문장을 조사하고, 이러한 문장의 일관된 형태를 간략한 네트워크로 표현함으로써 문장이 갖는 정보를 효과적으로 추출할 수 있다. 문장의 패턴을 나타내는 네트워크는 어절 내의 조사를 이용한 유한 오토마타로 표현된다[그림 2,3 참조].

즉, 유형별(교통, 요금, 숙박, 별미) 문장은 각각 정보를 표현할 때, 문장 내에서 이용되는 형태가 일정하고 제한적으로 사용되는 특정 조사들이 있다. 그러므로 문장 유형별로 문장을 나눈 후, 중요한 정보가 문장 내에서 표현되는 일관된 조사의 패턴을 찾아서 정보 추출에 이용한다.

데이터 파일의 문장은 체언+조사, 용언+어미, 체언만 있는 어절들로 구성된다. 문장이 갖는 대부분의 정보는 체언 부분에 속해 있으므로 이들 중 실제로 우리가 관심을 갖는 어절은 체언을 포함하는 어절들이다. 정보를 추출하기 위해서 우선 어절 단위로 문장을 나눈 후, 체언을 포함하는 어절들을 중심으로 패턴을 조사하였으며, 특별한 경우에만 용언+어미를 검사 하였다.

또한 하나의 어절이 체언, 체언+접속사, 체언+쉼표, 체언+의존명사 등으로 구성된 경우, 이 체언의 격을 정할 수 있는 조사가 생략된 경우 이므로, 이들의 격을 알 수 있을 때 까지 임시 저장소인 버퍼(buffer)에 이들을 저장한다. 버퍼에 저장된 정보는 다음 어절의 격을 공유한다.

3.3 정보 추출 및 틀 완성

패턴 네트워크를 통해 추출된 정보는 정제되지 않은 정보이므로 중복어 제거, 일부 대명사 처리와 같은 가공과정을 거칠 수 있다. 그러나 본 시스템에서는 중복어 제거는 처리하나 전체적인 파싱을 하지 않기 때문에 대명사에 대한 처리는 하지 못하고 있다.

패턴 네트워크를 통해 추출된 정보들은 다른 문장에서 추출된 정보와 합병(merge)되어 틀에 저장된다. 완성된 틀은 파일로 작성되어 DB에 저장된다. 즉, 같은 주제에 대한 문서가 여러 곳에 있다면, 패턴 네트워크를 통해 추출된 정보를 합병시켜 하나의 틀로 완성하여 DB에 저장하여야 한다.

3.4 정보 추출 예

실제 정보 추출 시스템을 통해 필요한 정보를 추출하는 과정이다. [그림 4]는 필요한 정보를 추출할 대상이 되는 문서의 예이다.

부석사

찾아가는 길

영동고속도로 원주 인터체인지에서 5번 국도를 타고 제천, 단양을 거쳐 풍기읍까지 간 다음 915번 지방도로를 따라 23km 정도 가면 된다.

대중교통편은 영주에서 풍기를 거쳐 부석사로 가는 시외버스가 1시간에 1대씩 있다.

요금

어른 8백원, 청소년 4백원, 어린이 3백50원.

문의 (0572)33-3464

....

[그림 4] 원시 데이터 파일

단계 1. 주요 단어를 이용하여 분석 대상 문장 추출

불필요한 문장을 걸러낸 후 데이터 파일의 형태는 다음과 같다.

```
부석사

/transportation 대중교통편은 영주에서 풍기를 거쳐 부석사로 가는 시외버스가 1시간에 1대씩 있다.

/cost 어른 8백원, 청소년 4백원, 어린이 3백50원.
...
```

[그림 5] 문장 추출 과정을 거친 데이터

단계 2. 품사 태깅

품사 태깅 과정을 마친 데이터 파일 내의 문장은 형태소 단위로 분리 된다.

```
부석사_NNP

/transportation 대중_NNCG+교통편_NNCG+은_PX 영주_NNCG+에서_PA 풍기_NNP+를_PO 거치_VV +어_EFC 부석사_NNP+로_PA 가_VV+는_EFD 시외_NNCG+버스_NNCG+가_PS 1_SCD+시간_NN BU+에_PA 1_SCD+대_NNCG+씩_XSNN 있_VJ+ 다_EFF+._SS.

/cost 어른_NNCG 8_SCD+백_DU+원_NNBU+,_SS, 청소년_NNCG 4_SCD+백_DU+원_NNBU+,_SS, 어 린이_NNCG 3_SCD+백_NU+50_SCD+원_NNBU+. _SS.
...
```

[그림 6] 품사 태깅 과정을 거친 데이터

단계 3. 패턴 네트워크 적용 및 틀 완성

각 패턴에 따른 네트워크를 거쳐 필요한 정보를 추출한다. 추출된 정보 중에는 중복된 정보가 있을 수 있으므로 중복어를 제거하는 과정을 거쳐 틀을 완성한다.

부석사		
1	교통수단	시외버스
2	출발 장소	영주
3	도착 장소	풍기, 부석사
5	요금	어른 : 8 백원, 청소년 : 4 백원 어린이 : 3백50원
6	숙박 시설	삼화장 여관(0572-22-1226) 명성민박(0527-33-3262) 평화민박(0572-33-3014)
7	민박 문의	
8	별미	평양냉면 한우숯불갈비 서부냉면집(0572-636-2457) 청국장 맛 부석사종점식당(0572-33-3606)

[그림 7] 완성된 틀의 모습

4 실험 및 분석

본 시스템에서 사용한 데이터 파일은 웹에서 발견한 관광지 정보 안내문이다. "Young의 여행에서 검색까지"(http://www3.shinbiro.com/~YOUNG707/home.htm) 중 테마여행 스케치라는 페이지 안에 있는 웹 문서들이 일반 텍스트 문서로 변환된 후 문장 종류에 따른 패턴에 의해서 필요한 정보가 추출되었다. 본 시스템은 JAVA 언어로 개발하였으며, compiler는 Symantec사의 Visual Café를 사용하였다

4.1 실험 방법

본 시스템의 성능을 평가하기 위해서 아래와 같은 데이터를 대상으로 실험을 수행하였다.

- 실험 1 : 학습 데이터, 31개
- 실험 2 : 오류 수정 전의 실험 데이터, 69개
- 실험 3 : 오류 수정 후의 실험 데이터, 69개

실험 1은 시스템에서 사용하는 패턴 네트워크의 적합성을 판단하기 위한 실험이고, 실험 2, 3은 학습 데이터를 통해 정의한 패턴을 이용해서 새로운 문서로부터 정보를 추출한 실험이다. 실험 3은 실험

2를 통하여 발견한 새로운 패턴 및 미등록어 등에 대한 오류를 수정한 후의 실험이다.

정보 추출에 대한 평가 방법은 일반적으로 시스템 성능을 평가하는 기준인 재현도(Recall), 정확도(Precision)와 F-score를 이용하였다. 재현도는 전체 유효한 데이터 수 대 시스템에 추출한 올바른 데이터 수의 비율을 의미하고, 정확도는 시스템이 추출한 데이터 중 유효한 데이터의 비율을 의미한다. 또한 F-score는 재현도와 정확도 모두를 이용하여 시스템의 성능을 하나의 숫자로 평가하는 방법이다. F-score를 계산 하는 방법은 다음과 같다.

$$F = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

여기서 ß의 의미는 재현도(R)와 정확도(P)의 비중을 선택할 수 있게 하는 변수로써 ß > 1 이면 정확도의 비중을, ß < 1이면 재현도의 비중을 높게 둔다는 의미이다. 일반적으로 기술적인 성능평가를 위한 ß 값으로 1,2,5를 사용한다[8].

4.2 실험 결과

4.2.1 실험 1

[표 2]는 31개의 학습 데이터를 본 시스템에 적용시킨 결과를 분석한 것이다.

	Recall	Precision	F-score		
			ß = 1	ß = 2	ß = 5
교통정보	91.4	81.2	85.9	89.1	90.9
요금정보	95.9	95.9	95.9	95.9	95.9
숙박정보	95.4	84.6	89.6	93.0	94.9
별미정보	95.3	84.7	89.6	92.9	94.8
전 체	93.7	84.6	88.9	91.7	93.3

[표 2] 학습 데이터 실험 결과 (단위 : %)

위 실험에서 교통 정보와 숙박 정보, 별미 정보에 비해 요금 정보의 정확도가 상대적으로 높은 수치로 나타났음을 알 수 있다. 다른 정보들에 비해 요금 정보의 정확도가 높은 이유는 요금 정보들을 포함한 문장들은 비교적 어절 수가 적으며, 해당 정보를 표현하는데 일관된 형태의 조사를 사용하여 표현함으로써 문장의 패턴이 분명하기 때문이다. 이러한 문장에서는 어절들 간의 관계를 따로 파악하지 않고, 조사의 패턴만을 고려해도 필요한 정보를 추출할 수 있다.

4.2.2 실험 2

본 시스템에 69개의 실험 데이터를 적용시킨 결과는 다음과 같다.

	Recall	Precision	F-score		
			ß = 1	ß = 2	ß = 5
교통정보	72.1	72.9	72.4	72.2	72.1
요금정보	66.8	71.9	69.2	67.7	66.9
숙박정보	53.4	68.6	60.0	55.8	53.8
별미정보	50.4	68.6	56.1	53.2	50.9
전 체	63.2	71.2	66.7	64.6	63.4

[표 3] 실험 데이터 실험 결과 (단위 : %)

위의 실험 결과에서 나타난 것처럼 실험 데이터를 본 시스템에 적용한 결과가 학습 데이터를 적용한 결과보다 성능이 현저히 저하되었음을 알 수 있다. 시스템의 성능이 저하되는 원인을 분석해 본 결과는 다음과 같다.

① 문장 분류 오류: 한 문장 내에 두개 이상 중요 단어가 있어서 문장 종류가 잘못 결정되는 경우
② 패턴 정의 오류: 학습데이터에서 추출하지 못한 새로운 패턴이 존재하는 경우
③ 버퍼에 저장된 정보 처리 오류
④ 부정문 처리 오류: 서술어 부분을 거의 참조하지 않기 때문에 문장의 끝에 부정의 의미가 있을 경우 이를 인식 하지 않는다.
⑤ 복잡한 어절의 형태소 분석 오류에 의한 경우
⑥ 미등록어 처리 오류
복합명사, 등록되지 않은 고유명사의 경우

위와 같은 오류를 수정하기 위하여 현재의 유한 오토마타 방법으로 가능한 다음 과정을 수행하였다.

① 문장 분류 오류 수정
문장 종류 태그를 하나 이상 붙일 수 있게 한다.

② 패턴 정의 오류 수정
학습 데이터에서 발견하지 못한 새로운 패턴을 추가하고, 기존의 패턴을 일부 수정하였다. 기존의 패턴 중 조사 만으로 어절의 격을 판단하는 것에서 일부 정보의 경우 다음 어절의 용언부를 검사하는 것으로 수정하였다.

③ 버퍼에 저장된 정보 처리 오류 수정
조사가 생략된 체언, 접속사나 쉼표에 의해 체언이 연속적으로 나타나는 경우는 버퍼에 정보를 저장한다. 새로운 데이터에서 발생한 새로운 접속사를 추가 하였다.

4.2.3 실험 3

[표 4]는 실험 2를 통해 발견한 오류를 수정한 시스템에 69개의 실험 데이터를 다시 적용시켜 본 결과이다.

	Recall	Preci-sion	F-score		
			ß = 1	ß = 2	ß = 5
교통정보	92.3	85.1	88.5	90.7	92.0
요금정보	89.8	96.5	83.0	91.0	90.0
숙박정보	87.2	89.8	88.4	87.7	87.2
별미정보	92.3	71.8	80.7	87.3	91.2
전 체	90.5	86.0	88.1	89.5	90.3

[표 4] 오류 수정 후 실험 데이터 실험 결과
(단위 : %)

[표 4]에 나타나듯이 오류 수정 과정을 거친 후의 결과에서도 적지 않은 오류가 나타난다. 오류가 나타나는 이유는 조사에 따른 명사를 추출함으로 그 명사의 하위 범주 정보가 참조되지 않았기 때문이다. 즉 어떤 명사의 하위 범주가 "지명"에 해당하는 명사인지, "탈 것"에 해당하는 명사인지를 알 수 없기 때문에, 선택된 명사가 교통 관련 정보 중 도착 정보인지 수단 정보인지를 가릴 수 없다. 이러한 오류는 추출 문서의 범위가 제한될 경우 그 분야에서 사용되는 고유 명사를 미리 시스템에 알려 줌으로써 해결될 수 있다. 실제로 실험 2에서 요금 정보의 정확도는 약 70% 정도이었지만, 지명 또는 사용료에 대한 고유 명사를 시스템에 알려 줌으로써 정확도가 90%이상으로 향상되었다.

본 시스템에서는 서술어의 정보를 참조하지 않기 때문에 부정문에 의한 오류를 수정하지 못했지만 이 문제는 다음 방법으로 해결할 수 있다. 부정 서술어가 해당 어절과 불규칙한 위치에 존재하므로, 해당 어절에서 참고할 수 있는 적당한 윈도우 크기를 정해 줌으로써, 연관된 서술어의 용언부를 검사하여 해당 어절의 의미를 알 수 있다. 즉, 패턴을 이용하여 필요한 정보뿐만 아니라 술부의 의미도 파악하여, 추출된 정보가 부정의 의미로 추출되었는지 긍정의 의미로 추출되었는지 판단할 수 있어야 한다. 보다 정확한 정보 추출을 위해서 서술어의 용언을 중심으로 문장이 갖는 어절 간의 패턴을 알아내는 연구가 필요하다.

5 결론 및 향후 연구 방향

본 연구에서는 패턴 정보에 의해 유한 오토마타로 표현된 네트워크를 사용하여 한국어 문서에서 필요한 정보를 추출하는 방법을 제시하였다. 대상 문장이 어느 정도 정형화된 형태를 갖는다면 문장에 대한 복잡한 파싱에 의하지 않아도 조사를 중심으로 어절 간의 형태적인 패턴을 찾아 적용함으로써 어절의 격을 정하고 필요한 정보를 추출할 수 있음을 알 수 있었다.

일반적인 문장에 있어서 본 시스템에서 사용한 어절 간의 패턴을 적용하기는 어렵다. 그러나 대상 문장이 갖는 정보에 따라 분류되어 있을 때 본 연구에서 제시한 방법론이 적용될 수 있음을 보였다.

본 시스템의 장점으로는 구현이 간단하면서도 추출 신뢰도가 비교적 높다는 것이다. 문장 전체에 대한 파싱이나 의미 분석에 의존하지 않고, 조사를 중심으로 어절 간의 관계를 파악하기 때문에 처리과정이 간단하며 정확한 정보를 추출할 수 있었다. 또한 일반 파싱에서 문제가 되는 어절 수에 제한 받지 않으므로 긴 문장 내의 정보도 오류 없이 추출할 수 있었다.

반면 유한 오토마타의 제한점으로 인해 정형화 되어있지 않은 형태로 작성된 문서에서는 효과적인 정보 추출을 기대하기 어렵다는 단점이 있다. 즉, 일관적이지 않은 패턴의 문장에 대해서는 새로운 패턴을 다시 정의해서 사용해야 한다.

본 시스템에서는 비교적 단순한 정보만을 추출하였지만, 좀 더 복잡한 정보를 추출하거나 추출 대상이 되는 문서가 보다 복잡한 경우에는 서술어의 용언을 중심으로 각 어절 간의 관계에 대한 패턴을 다시 알아냄으로써 보다 향상된 성능을 기대할 수 있다.

참고문헌

[1] 김준태, "제한된 분야의 자연어 분석을 위한 구문패턴의 이용방법", 1995년도 한국정보과학회 가을 학술발표논문집, Vol . 22, No. 2, pp. 631~634.

[2] 장동현, 맹성현, "자동 요약 시스템", 정보과학회지, 1997년 10월호, pp. 42~49.

[3] 장동현, 맹성현, "문서 구조 정보를 이용한 확률 모델 기반 자동요약 시스템", 1997년도 한글 및 한국어 정보처리 학술대회, pp. 15~22.

[4] 김진동, 임희석, 임해창, "Twoply HMM : 한국어의 특성을 고려한 형태소 단위의 품사 태깅 모델", 한국정보과학회 논문지(B), 제24권, 제12호, pp. 1502-1512, 1997.

[5] Mary Ellen Okurowski, "Information Extraction Overview", Tipster Text Program (Phase Ⅰ), Sep.1993, pp. 117~121

[6] Julian Kupiec, Jan Pedersen and Fracine Chen, "A Trainable Docuument Summarizer", Proc of 18th ACM-SIGIR Conference, 1995, pp. 68~73,

[7] R. Barzilay and M. Elhadad, "Using Lexical Chains for Text Summarization", Proc. Of a Workshop on Intelligent Scalable Text Summarization , July 1997, pp. 10~17.

[8] E. Hovy and C. Y. Lin, "Automated Text Summarization in SUMARIST", Proc. Of a Workshop on Intelligent Scalable Text Summarization , July 1997, pp. 18~24.

[9] Lisa F. Rau, Paul S. Jacobs and Uri Zernik, "Information Extraction And Text Summarization Using Linguistic Knowledge Acquisition", Information Processing & Management, Vol. 25, No. 4, pp. 419~428.

[10]Fujio Nishida and Shinobu Takamatsu, "Structured Information Extraction From Patent-Claim Sentences", Information Processing & Management, Vol. 18, No. 1, pp. 1~13.

[11]Jerry R. Hobbs, Douglas Appelt, John Bear, David Israel, and Mabry Tyson, "FATUS: A System for Extracting Information from Natural-Language Text", SRI International Technical Note, No. 519. November 19, 1992.

[12]The PLUM System Group, "BBN's PLUM Probabilistic Language Understanding System". Tipster Text Program (Phase Ⅰ), Sep.1993, pp. 195~207

[13]D. Moldovan, S. Cha, M. Chung, T. Gallippi, K. Hendrickson, J. Kim and C. Lin, "Description of SNAP system used for MUC-5", Proceedings of the Fifth Message Understanding Conference, 1993.

[14]K. McKeown and D. Radev, "Generating Summaries of Multiple News Articles", Proc of 18th ACM-SIGIR Conference. 1995. pp. 74~82