# 표준 주소 검색을 위한 주소 보정 시스템 구현

이상윤[O] 박병준
광운대학교 컴퓨터과학과
{sylee[O], bjpark}@cs.kw.ac.kr

# Implementation of an Address Correction System for Standard Address

Sang-yun Lee[0] Byung-joon Park
Dept. of Computer Science, Kwangwoon University

**요 약**

본 논문은 형태소 분석 기법과 전문가 시스템 제작 도구를 이용하여 표준 주소를 검색하기 위한 주소 보정 시스템의 구현에 대해 기술 한다. 즉, 주소가 가지는 특성을 고려하여 표준 주소에 대한 다양한 형태의 주소들을 각각 지역 단위의 의미를 가지는 형태소로 분리하고 전문가 시스템에서 정의된 규칙에 의해 주소 요소의 원형으로 변환하게 한다. 따라서, 각각 주소 요소의 원형으로 이루어진 보정된 주소는 데이터 베이스 상에 존재하는 표준 주소가 될 것이고 정확한 검색이 이루어진다. 이는 데이터 베이스에 축적된 해당 주소에 대한 새로운 정보를 참조하는데 활용될 수 있다.

## 1. 서 론

현재 국내 우정사업은 탈규제화와 개방화에 따른 경쟁의 심화, 고객 욕구의 고도화, 다양화에 따라 표준화된 서비스의 한계를 극복하기 위한 최신의 기술 개발을 위해 노력하고 있다. 이러한 기술 연구는 Soft Science, Traditional Technology, Information Technology의 세가지 영역으로 나누어지며 특히, Information Technology의 영역은 우정사업의 정보화를 위해 필요한 기술의 연구로써 미래 우정사업의 경쟁력 재고를 위해 핵심적으로 육성하고 있는 분야이다. 이 영역에서 다루어지는 기술로서는 우정 정보 표준화, 우정 자동 구분 및 처리, 정보 통합 플랫폼 기술 등을 포함하고 있는데 이러한 기술들을 유기적으로 원활히 수행될 수 있도록 하기 위해서는 정보 표현에 따른 지능적 판독 기술이 필수적으로 선행 되어야한다. 즉, 다양한 형태로 입력되는 배달 주소에 대하여 기존의 연구를 통해 수립된 통합 정보 데이터 베이스에 접근이 가능하도록 하여야 일련의 자동화 기능이 원활히 수행될 것이기 때문이다. [1]

주소를 입력 받는 방식으로써 요즘에는 우편 번호를 이용하여 자동으로 입력되도록 하는 방식이 널리 사용되고 있다. 하지만, 여전히 오프라인상에서 이루어지는 고객들의 배달 주문과 전화 주문 등의 경우에는 접수원이 정확히 표준화된 형식으로 주소를 입력한다는 보장을 받을 수 없을 뿐만 아니라 데이터 베이스에 입력된 표준 표현을 모두 기억하고 있을 수도 없다는 문제점이 있다. 그래서, 주소 보정 시스템은 다양한 형태로 입력된 주소들에 대하여 데이터 베이스에 정의된 표준 주소로의 자동 변환을 가능하게 함으로써 우정 업무 자동화에 기여한다.

본 논문의 구성은 2장에서는 국내 우편 주소정보의 구조, 3장에서는 주소 보정 시스템의 구조, 4장에서는 실험 결과 및 평가, 5장에서는 결론과 향후 연구에 대해 기술한다.

## 2. 국내 우편 주소정보의 구조

우편물의 발송, 중계, 배달에 필요한 정보는 우편물에 기재된 수취인 주소로부터 얻기 때문에 우편물에 기재된 주소가 불분명한 경우에는 배달 불능 우편물이 되는 사례가 빈번하다. 따라서 우편물이 수취인에게 정확히 전달되기 위해서는 발송 배달을 위한 필수 주소정보가 있다고 할 수 있는데, 이의 파악이 국내 우편 주소 표기 지침안 작성에 있어 무엇보다도 시급하다고 할 수 있다. 현재 우편 주소정보로는 대체적으로 다음과 같이 최대 5단계의 구조를 갖는다고 볼 수 있다.

- 행정구역(시,군,구 + 읍,면,동 + 통,리) + 구분(일반,산) + 지번(본번,부번)
- 건물명(아파트,빌라,연립 등) + 건물번호(단지동수) + 층호수
- 입주자(상호,기관,법인명) + 부서
- 수취인 성명(직위,호칭)
- 우편번호(6자리)

여기서 발송을 위한 필수 정보는 우편번호 앞 3자리 또는 시,군,구 지역 단위까지의 주소이고 배달을 위한 필수 정보는 법정동 체계에 따른 지번이라고 할 수 있다. 법정동내에서 행정동이 세분되어 있는 경우에는 행정동명과 정확한 우편번호의 사용이 필요하며 행정구역 체계인 통, 반에 대한 정보는 사실상 무시된다. 주소 보정 시스템에서는 발송과 배달을 위한 필수 정보를 기준으로 표 1.과 같이 정형화된 주소 자료구조를 가지고 처리 작업을 하게 된다. 이 구조는 현재 국내에서 널리 사용되고 있는 법정동 주소 체계와 행정동 체계, 점차로 도입되고 있는 서양식 도로 방식의 주소를 포함하는 구조이므로 실제로 내부 연산 과정에서는 각각의 체계에 맞는 3가지 유형별로 나누어 처리한다. [2]

표 1. 주소 구조 테이블

| 구분 | 주소 정보 |
|---|---|
| DO | 특별시, 광역시, 도 |
| SI | 시, 군 |
| GU | 구 |
| DONG | 읍, 면, 동, 가(도로) |
| APT_NAME | 아파트 단지, 리 |
| DONG_NUMBER | 아파트 동 번호 |
| SAN | 산, 일반 구분 |
| BUNJI | 번지 |
| HO | 호 |

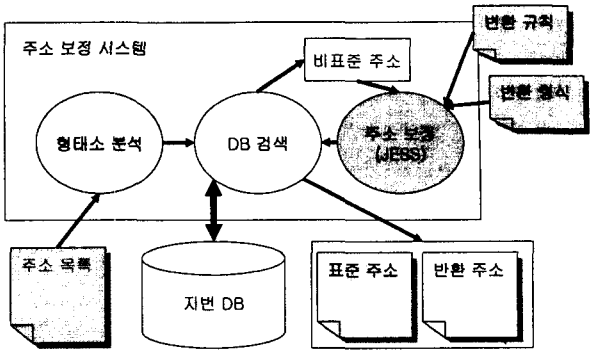


그림 1. 주소 보정 시스템의 구성

### 3. 주소 보정 시스템 구성

주소 보정 시스템은 전체적으로는 그림1.에서와 같은 구성을 가지며, 핵심적으로는 형태소 분석 모듈과 주소 보정 모듈의 두 가지로 구성된다. 우선, 입력되는 주소에 대한 주소 정보의 구조를 파악하기 위해서 주소 형태소를 분리해 낸다. 그러면 각각의 주소 형태소들은 표1.에 나타난 테이블의 각 필드에 해당 형태소들이 할당되고 주소 보정 모듈의 입력 데이터로써 사용된다. 그리고 주소 보정 모듈에서는 이 데이터를 사실들(*facts*)로 입력받고, 정의된 규칙들(Rules)을 적용하여 표준화 형식에 맞는 주소 형태소 원형으로 변환한다.

#### 3.1 주소 형태소 분석

주소 형태소 분석은 자연언어 처리에서의 형태소 분석이 하는 역할과 마찬가지로 주소를 데이터베이스에서 검색하기 위해서 필요한 정보 혹은, 보정해야 할 정보들을 추출하는 기저 단계로 주소 보정 시스템의 기본이 됨과 동시에, 주소 형태소 분석의 기능과 효율성을 고려할 때 전체 시스템의 성능에 바로 직결되는 중요한 요소이다. 그리고 여기서의 주소 형태소는 데이터베이스를 검색하거나 주소 보정 단계에서 처리해야 하는 입력 단위 집합의 원소이며 주소 요소의 의미가 부여된 것을 의미한다. 형태소 분석 모듈의 구조는 그림2.와 같다.

● 어절 분리 - 대부분의 형태소 분석기는 입력 단위를 어절로 보기 때문에 간단한 방법으로 공백을 이용한 어절 분리를 한다. 또한 특수 문자를 한글 문자열과 분리하여 독립된 어절로 나누기도 한다. 그러나 본 논문에서는 띄어쓰기 오류에 대한 문제점 해결을 위하여 주소 전체를 하나의 어절로 간주한다. 단, 아래에 설명되는 형태소 해석 방향의 차이에 따라, 숫자와 고유 명사의 조합으로 이루어지는 주소의 경우에는 각각을 구분하는 기준으로서 공백을 이용하여 어절로 나누고 처리한다.

● 어절 분리 위치 결정 - TRIE 사전을 이용하는 경우와 같이 사전 검색을 이용하여 분리 위치를 결정하게 한다. 이는 발송을 위한 주소가 제한된 고유명사를 가지는 특성과 배달 주소가 특정 키워드를 통해서 구분 가능하다는 점을 반영한 것으로 음절을 이용하여 분리하는 방식에 비해 계산 비용이 저렴하다.

● 품사 선택 - 태그 등을 이용하여 여러 형태소 분석 결과 중에 하나를 선택한다. 가능한 모든 결과를 다음 단계에서 선택하도록 하는 경우도 있으나 본 시스템에서는 주소 보정 모듈의 입력 형식이 고정되어야 한다는 점과 나뉘어진 주소 형태소가 순차적인 특성이 있다는 점을 고려하여 최대한 표1.에 가깝도록 다시 합병하는 과정을 수행한 후에 결과를 나타낸다.

● 형태소 원형 복원, 결합성 검사 - 일반적인 자연언어 처리 과정에서 필수적인 단계이지만 본 시스템에서는 이 과정을 주소 형태소 분석에서 다루지 않고 3.2절의 전문가 시스템 개발 도구(JESS)를 통해 수행한다.

이외의 주소 형태소 분석에 이용된 기법들은 한글 코드의 경우는 Symbol Code 방식, 사전은 통합 단일 사전을 사용하였으며 사전 표제어 색인의 방법은 TRIE 구조, 사전 표제어 규정은 파생어에 대한 표제어의 설정만을 허용하였다. 사용된 형태소 해석 문법은 Regular(linear) grammar를 적용하였고, 해석 방향은 형식 형태소 우선 분석과 실질 형태소 우선 분석 기법을 혼용하였는데 주소 정보 요소 중에서 발송을 위한 주소 형태소에 대해서는 형식 형태소 우선 분석 기법이 사용되었고 배달을 위한 주소 요소에 대해서는 실질 형태소 우선 분석 기법이 적용되었다. 해석 알고리즘은 Head-Tail 구분법과 Tabular 파싱법을 응용, 혼합하여 구성하였다. [4]

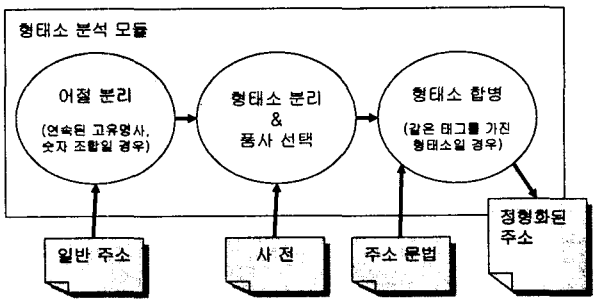


그림 2. 형태소 분석 모듈의 구성

### 3.2 주소 보정

주소 보정 시스템에서는 주소 보정을 위하여 JESS를 채용 하였다. JESS는 전문가 시스템을 구축하기 위한 도구로써 Rete 알고리즘을 사용하여 규칙들(Rules)과 사실들(facts)을 매치시키고, 사실 집합 (facts collection)에 규칙(Rule)을 반복 적용하는 방식을 사용한다. 그리고 새로운 규칙(Rules)의 지속적인 추가가 용이하다는 점과 Java 언어를 기반으로 하기 때문에 같은 기능을 수행하는 CLIPS 와는 달리 유니코드를 지원하므로 한글을 비교적 자유롭게 사용할 수 있는 장점이 있다. 그리고 무한대의 룰을 생성할 수 있고, 각각의 룰에 대해 20,000 단계의 ( -10,000 ~ 10,000 ) 우선순위를 부여할 수 있다.

주소 보정 모듈은 크게 추론 엔진, 지식 베이스, 입출력을 위한 인터페이스의 3가지 부분으로 나눌 수 있다. 추론 엔진과 인터페이스는 기본적으로 기능들을 제공하므로 본 논문에서는 지식 베이스 부분을 주소 보정 시스템에 적용하는데 필요한 작업을 수행하였다. 단, 시스템이 Java 기반이므로 전체 시스템의 구현 언어에 따라 인터페이스 부분에도 수정이 필요할 경우가 있다. [5]

지식 베이스는 다시 두 가지를 고려하는데 하나는 입력을 위한 사실들(facts)의 정의이다. 주소 보정 시스템에서는 표1.에서와 같이 정형화된 주소를 사용하고 있기 때문에 구조화되고 객체 지향적 개념의 사실 집합(fact set) 인 "Unordered facts"를 사용한다. 이 오브젝트는 사용자 정의의 name fields (slots)를 가지며 본 시스템에서는 아래와 같이 구분자의 슬롯을 가지게 된다.

```
Jess> (deftemplate addr (slot DO) (slot SI) (slot GU)
(slot DONG) (slot APT_NAME) (slot DONG_NUM)
(slot SAN) (slot BUNJI) (slot HO) (slot ADD_NAME))
```

다른 하나는 추론 엔진(Inference Engine)에서 참조하는 규칙 집합(Rule Set)이다. JESS에서 적용하는 규칙은 Procedural Language의 if ~ than 구문과 유사하다. 그러나 if ~ than 구문은 특정한 시점에서만 수행되지만 JESS의 규칙은 LHSs가 만족되면 무결성을 유지하기 위해 반복적으로 수행된다. Rete Algorithm을 사용한 이러한 구조는 if ~ than에 비해 효율적이고 안정적이다.

아래의 예는 데이터베이스의 'BUNJI' 필드가 숫자만으로 표기될 때, 'BUNJI' slot의 부가적 문자열 중에서 "번지"일 경우에 제거하는 규칙을 나타낸다.

```
(defrule MAIN::test8-1
  (declare (salience 9992))
  ?fact <- (addr(BUNJI ?temp))
  (test (> (str-length ?temp) 2))
  (test (eq "번지" (sub-string ( -(str-length ?temp) 1)
                   (str-length ?temp) ?temp)))
  =>
  (modify ?fact (BUNJI (sub-string 1 1 ?temp)))
)
```

### 4. 실험 결과 및 평가

주소 보정 시스템의 실험에는 일반인이 충청북도 청주시를 대상으로 하는 임의의 주소를 특정 형식에 구애 받지 않고 자유롭게 입력하여 주소 목록 201건을 테스트 집합으로 사용하였고, 주소 보정을 위한 규칙은 일반적으로 나타날 수 있는 주소 입력 유형을 분석하여, 표준 주소에 맞도록 변환하는 규칙 221개를 가지고 실험하였다. 분석 방법은 Recall 과 Precision 값을 비교한다.

표 2. 실험 결과

| | Total | TP | TN | FP | FN |
|---|---|---|---|---|---|
| 개수 | 201 | 152 | 40 | 0 | 9 |

Recall = 79.2 %

Precision = 75.6 %

실험 결과를 통하여 기본적인 규칙만으로도 상당수의 일반 주소를 표준 주소에서 검색할 수 있는 것을 확인할 수 있다. 이 후에 시스템 규칙의 훈련(Training)을 통하여 성능은 더욱 개선 될 것이다.

### 5. 결 론

우정 업무를 처리하는데 가장 중요한 요소가 주소이고 현재로서는 정형화된 입력 형식 및 방법을 갖고 있지 않으므로, 다양한 패턴으로 입력되는 주소들을 우정 업무 자동화를 위해 검색해야하는 데이터베이스에서 매칭되는 주소를 찾아낼 수 있는 확률은 높지 않다. 그래서, 본 논문에서는 주소의 의미를 파악해 내는 방법으로써 자연언어 처리의 형태소 분석 기법과 형태소 분석에서 원형 복원과 결합성에 대한 문제점을 보완하기 위해 전문가 시스템 개발 도구를 사용하여 주소 보정 시스템을 구현하였다. 주소 형태소 분석을 통하여 주소. 요소들을 세부적으로 파악함으로서 주소 유형에 따른 검색률을 높일 수 있게 하고, 주소 보정 모듈을 통하여 다양한 오류에 대하여 시스템이 적응할 수 있도록 한다. 본 시스템은 규칙이 지속적으로 추가, 적용될수록 정확도가 향상될 수 있으므로 이를 위해서 사용자가 쉽게 규칙을 생성할 수 있도록, 내부적인 규칙의 무결성을 자동으로 보장해주면서 규칙을 생성시킬 수 있는 사용자 인터페이스에 대한 추가 구현이 필요하다.

### 참고문헌

[1] 우정 사업 본부, "우정 사업 연구, 기술 개발 체제 강화 계획(안)", Jan. 2002.
[2] ETRI, "우편주소 표준안 보고서", Dec. 2002.
[3] 자연언어처리연구실, "Documents on Morphological Analysis in Korean", KAIST.
[4] 신효필, "한국어 형태소 분석", Seoul National University.
[5] JESS : http://herzberg.ca.sandia.gov/jess/