규칙을 적용하여 세분화한 사전기반의 한국어 지명인식 시스템 연구

장혜숙*, 정규철*, 이진관*, 박기홍*

*군산대학교 컴퓨터정보과학과

A Study on Recognition of Korean Place Names System on the Internet by Using the Rules of Dictionary Use

Hae-Suk Jang*, Kyu-Cheol Jung*, Jin Kwan Lee*, Kihong Park*

*Dept of Computer Science, Kunsan National University

e-mail: hs5486@kunsan.ac.kr

요 약

개체명 인식에 있어서 반드시 선행되어야 할 작업이 문서의 내용을 대표하는 용어의 추출이다. 높은 신뢰도의 개체명 인식은 정보추출 시스템구축을 한 차원 높일 수 있을 것이다. 지금까지 일반적인 개체명 인식이나 인명의 개체명 인식에 대한 많은 연구가 활발하게 진행되어 왔지만 세분화된 지명 인식의 연구는 다루어지지 않았다. 본 논문에서는 수작업으로 작성된 규칙을 적용하여 세분화한 사전기반의 한국어 지명인식 시스템 개발 방법을 제안한다.

ABSTRACT

A choice of a term represent the contents of the documents should be preceded in recogintion of the name of an individual. The recognition of the name of an individual which has high reliability can raise the building of information extract system into higher level. Up to now active reserch on the recognition of the name of an individual human name and the general recognition of the name of an individual have been studied, detailed research on the recogintion of the name of a place hasn't be done, however. This thesis suggests detailed methods of recognition of Korean place names system with applying a rule made manully.

키워드

개체명인식, 지명인식, 한국어지명, 사전검색, 결정규칙

Ⅰ. 서 론

자연어로 작성된 문서의 정보추출은 개체명 인식(Named Entity), Conreference Resolution, 정보 추출의 단계로 이루어진다[1]. 지금까지는 일반적인 개체명 인식에 관한 연구와[2], 특히 대용어중에서 빈도가 높은 유형인 인칭명사의 연구가 활발하게 진행되어왔다[8]. 그러나 정보 추출시 사전에 등록되지 않은 지명인식 패턴 구축이나 지명인식의 Coreference Resolution 패턴의 연구는 진행되지 못하고 있는 실정이다. 예를 들어 "군산시 에서는 올해 재배된 보리들을 오는 5일경 수확해 군산지역 각 경로당에 분배할 예정이다"의 문장에서 "군산"은 기관명, 지명으로 서로 다르게 사용되고 있다. 이러한 문제들을 고려하여 지명을 인식하는 방법은 크게 두 가지로 나눌 생각해 볼 수가 있다.

첫째, 규칙에 기반한 방법이다. 규칙을 수작업으로 구축하고 이를 기반으로 하여 새로운 문서에 대해 개체명을 인식한다[4-9]. 이 방법은 고유명사 사전이나 결합 사전 등을 이용한다. 학습 코퍼스를 만들어 자동으로 개체명 인식을 하거나 수동으로 인식 패턴을 구축하게 된다. 둘째, 통계에 기반한 방법으로 학습 코퍼스로부터 개체명 인식의 지식을 학습한 다음 HMM(Hidden Markov Model)이나 MEM(Maximum Entropy Model), 결정트리(Decision Tree) 모델 등을 이용한다. 대소문자를 구분하는 영어에 비해 문자형에 대한 정보가 부족한 한국어의 세분화된 지명 인식은 어렵다고 볼 수 있다. 이러한 점들을 고려하

여 본 논문에서는 사전을 이용한 규칙기반의 세분화된 지명인식의 방법을 제안한다.

II. 관련연구

활발한 개체명 인식의 연구에 비해 세분화된 개체명 인식의 연구는 제대로 이루어지지 않았다. 일반적인 개체명 인식의 관련연구는 수동으로 패턴을 구축하고 구축된 패턴을 기반으로 개체명을 인식하는 방법과 통계 모델에 기반한 학습을 이용하여 인식하는 방법으로 분류할 수 있다. 수동으로 작성된 규칙에 의한 개체명 인식의 방법은 일반적인 규칙을 발견하기가 쉽지 않으나 전문가에 의해 정교한 규칙을 작성하기 때문에 Coreference Resolution에 적합하다.

통계적 기법은 학습 말 뭉치로부터 대용어 관계를 결정하는 규칙을 학습한다. 학습 말 뭉치의 대용어 참조 정보는 학습 알고리즘에 따라 다르지만 <속성, 값> 쌍 형태의 문맥으로 구성된다. MLR과 RESOLVE는 통계적 학습알고리즘에 의해 학습 말 뭉치에 표시된 대용어 관계 정보로부터 임의의 두 용어들에 대한 대용어 관계여부를 좌우 문맥정보와 함께 추출하여 결정트리로 구현하는 결정트리 추론시스템이다[10,11].

두 시스템의 각각 50개 문서집합과 250개 문서집합에 대한 실험결과에 의하면, RESOLVE는 재현율과 정확률이 각각 80%~85%, 87%~92%로 나타났으며, MLR은 67%~70%, 83%~88% 였다. 그러나 이 시스템들을 MUC-6대용어 태스크의 25개 문서집합에 적용했을 때 RESOLVE는 재현율이 41%~44%, 정확률이 51%~59% 이다. 이 결과는 대용어 해결알고리즘을 수동으로 작성한 상위 5개 시스템의 재현율이 51%~63%, 정확률이 62%~72%인데 비해 매우 낮다고 볼 수 있다. 즉, 개체명 인식에서 발생되는 Coreference Resolution 알고리즘은 통계적인 방법보다는 수작업으로 규칙을 정교하게 작성했을 때 재현율과 정확률의 신뢰도가 높아진다.

III. 사전을 이용한 규칙 기반의 세분화된 한국어 지명(location name) 인식

본 논문에서는 개체명의 범주를 Coreference문제를 안고 있는 세분화된 지명의 범주로 제한하고자 한다. 구축된 지명사전을 이용하여 단어 자체의 범주를 살펴보고 이를 기반으로 제한된 주변 문맥 단어정보를 이용하여 규칙을 생성하고자 한다.

3.1 사전 구축

세분화된 지명인식을 위해서 규칙을 적용하기 전에 입력으로 들어오는 각 단어들의 사전 정보를 추출해야 한다.
이를 위해 본 연구에서는 아래의 알고리즘으로 사전 구축을 하였다.

```
main()
{
        char name[20];
        char b[5][10];
        int i,len,j,dic_len;
    FILE*fp;
        if((fp=fopen("ab.txt","r"))==NULL){
          puts("파일을 개방할수 없습니다.");
        return 0;
        }
        i=0;
        printf("사전 로딩중 ... \n");
        while((fgets(name,20,fp) != NULL)){
                len = strlen(name);
                for(j=0;j<len;j++){
                        if(name[j] == '\n') {
                                b[i][j]='\0';
                                //break;
                        }
                        else b[i][j]=name[j];
                }
                i++;
        }
        fclose(fp);
        dic_len = i;
        printf("%d 개의 데이터 로딩 완료 ...\n",
         dic_len);
          do{
        printf("검색어를 입력하세요 ? ");
        gets(name);
        j = 0;
        for(i=0;i<dic_len;i++){
                if(strcmp(name,b[i]) == 0)
        j=1;
        break;
        }
        if(j == 1) printf("지역명입니다.\n");
           else  printf("지역명이 아닙니다.\n");
        }
        while((strcmp(name,"end")==1));
        return 1;
}
```

사전은 군산지역의 우편번호부에 등록되어 있는 지역을 대상으로 구축하였다.

3.2 지역명 인식 결정 규칙

입력으로 품사 태깅 된 어절 단위의 문장을 읽어 들인 다음, 지명의 후보를 찾기 위해 고유명사

나 보통명사를 찾는다. 찾은 지명 후보에 대해 구축된 사전으로 검색을 한다. 사전 검색결과는 '지역명입니다' 와 '지역명이 아닙니다' 의 두 가지 경우 중 한가지의 결과가 나오게 된다. '지역명이 아닙니다' 의 결과가 나온 경우 지명 후보가 되는 단어의 앞이나 뒤에 나타나는 단어를 고려하는 문맥규칙을 적용시킨다.

3.2.1 문맥 규칙

사전 검색결과 '지역명이 아닙니다'의 결과가 나온 지명후보들의 문맥을 수집하였다.
지역명을 결정하는 규칙을 발견하기 위하여 문맥규칙을 두가지 유형으로 구분하여 처리하였다. 첫째 문맥규칙은 지역명 개체후보의 앞 단어를 보는 규칙과, 둘째는 지역명 개체후보의 뒷 단어를 보는 규칙이다. 이 두가지 유형에 대해 지역명을 결정하는 규칙을 발견하였다.
[규칙 1]사전 검색결과 '지역명이 아닙니다'의 결과가 나온 지명후보의 좌측(앞)단어의 품사가 동사인지를 검색한다.
[규칙 2]사전 검색결과 '지역명이 아닙니다'의 결과가 나온 지명후보의 우측(뒷)단어의 품사가 명사인지를 검색한다.
예)항동마을은 성살 이라 부르는 오성산 밑의 골과 달개가 한마을 이루어 성덕제의 맑은 물을 이용해 물 걱정 없이 미작을 주업으로 생활해오고 있다.

위의 예제에서 지명후보는 '성덕제'이다. [규칙 1]을 적용하여 성덕제의 앞단어 '이루어'의 품사가 동사임이 발견되므로 '성덕제'는 지명으로 판단한다. 또한 지명후보 단어인 '오성산'도 사전에 등록되어 있지 않지만 규칙(1)에 따라 '부르는'이 동사이므로 지명으로 인식하게 된다.
예)군산시는 지난 30일 관내 울외 주산단지인 성산면 상작마을에서 지도 공무원과 농업인등 20여명이 참석한 가운데 군산시 특산품으로 각광받고 있는 신세대용 울외 피클 상품화에 대한 평가회를 가졌다.

위의 예제에서 [규칙 2]를 적용시켜 지명후보인 '성산면' 이나 '상작' 마을의 뒷 단어 품사가 명사이므로 '성산면'이나 '상작'의 지명후보를 지명으로 판단한다.
시스템의 구성을 그림1로 나타내었다.

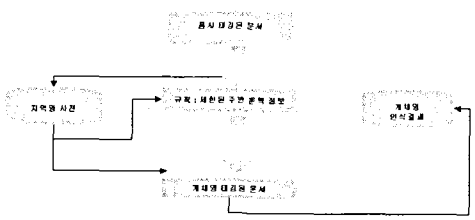

그림1. 시스템 구성도

IV. 실험 및 평가

본 논문의 실험을 위해 품사 태깅된 신문기사를 선택해서 사용하였다. 실험데이터로는 20개의 기사(40문장)을 이용하였다. 형태소 분석결과 명사로 인식되는 어절을 대상으로 사전검색한후 지역명으로 인식되지 못하는 명사들을 대상으로 규칙을 적용하는 방법으로 실험하였다. 실험 결과는 표1과 같다.

표 1. 실험결과

구분	재현율	정확율
지역명 인식	75%	81%

V. 결 론

본 논문에서는 수작업으로 작성된 규칙으로 지역명검색의 정확 율을 향상시킴으로써 지리정보의 연구에 도움이 되었다. 부족한 규칙으로 예외의 경우가 많이 발생하는 잘못된 결과는 많은 규칙의 생성으로 줄 일수 있을 것이다. 많은 규칙을 생성하는 심도 있는 연구가 필요하다.

참고문헌

[1] MUC-7(1998), Proceedings of the Seventh Message Understanding Conference(MUC-7), http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/ie_task.html

[2] Bikel,D.,Schwartz,R.,Weischedel,R., "An algorithm that learns what's in a name. Machine Learning:Special Issue on NL Learning",34,1-3,1999

[3] K.Fukuda, T. Tsunoda, A. Tamura and T. Takagi, "To-ward Information Extraction:Identifying protein names from biological papers,"In Proc.of the Pacific Symposium on Biocomputing '98(PSB '98),1998.

[4] J.Fukumoto,M.Shimohata, F. Masui and M. Saski,"De-scription of the Oki System as Used for MET-2," In Proceedings of 7th Message Understanding Conference,1998.

[5] A.Mikheev,C. Grover,M.Moens,"Description of the LTG System Used for MUC-7," In Proceedings of 7th Message Understanding Conference, 1998.

[6] C. Aone and W. Bennett, "Evaluation Automated and Manual Acquisition of Anaphora Resolution Strategies," Proceedings of the 33rd Annual Metting of

the Association for Computational Linguistics, Association for Computational Linguistics, pp.122-129,1995.

[7] J.F. McCarthy and W. G. Lehnert, "Using Decision trees for Coreference Resolution", Proceedings of the Four-teenth International Conference on Artificial Intelligence, International Joint Conferences on Artificial Intelligence, pp.1050-1055, 1995.

[8] 강승식,윤보현,우종우. Coreference Resolution 을 위한 3인칭 대명사의 선행사 결정규칙. 정보처리학회 논문지B 제11-B권 제2호(2004. 4)

[9] 김태현, 이현숙, 하유선, 이만호, 맹성현, "데이터 집합을 이용한 고유명사 추출",제 12회 한글 및 한국어 정보처리 학술대회, pp.11-18, 2000.

[10] 노태길, 이상조, "규칙 기반의 기계학습을 통한 고유명사의 추출과 분류",한국정보과학회 가을 학술발표논문집, Vol.27, No.2, pp.170-172, 2000.

[11] 이경희, 이주호, 최명석, 김길창, "한국어 문서에서 개체명 인식에 관한 연구", 한글 및 한국어 정보처리 학술대회, pp.292-299, 2000.