한국어에 적합한 효율적인 품사 태깅

An Efficient Korean Part-of-Speech Tagging

김영훈
LG전자 CDMA단말연구소 선임연구원

Young-Hoon Kim
Research Engineer CDMA Lab. LG Electronics Inc.

중심어 : 태깅, 형태소 분석, 품사

요 약

본 논문에서는 형태소 분석 단계에서 발생하는 형태소 모호성을 줄이기 위해서 말뭉치를 이용한 형태소 태깅을 구현한 시스템이다. 형태소 태깅을 위한 말뭉치가 대량의 것이 아니라도 효율적인 품사 분류와 태깅 알고리즘을 가지고 올바르고 효율적인 태깅을 할 수 있도록 하였다. 어절의 올바른 품사 태깅을 위해서 어절들 간의 인접도가 아니라 품사들 간의 인접도, 그리고 품사간의 제약 정보를 추가한 품사의 태깅에 이용을 하였다. 이와 같이 함으로 효율적인 시스템을 구현할 수 있었다.

Abstract

In this paper I offer a new part-of-speech tagging method for Korean, it can solve difficulty of statistical data acquisition and ambiguities due to same part-of-speech stream input and make good use of the Corpus. This method can solve that the corpus don't have huge. This method uses pattern information about part-of-speech among eojols and constraint-rules in order to perform part-of-speech tagging. The Constraint-rule is used to select appropriate part-of-speech pattern.

I. 서론

사람이 사용하는 언어 즉 자연언어를 컴퓨터를 통하여 이용하려는 활발하게 진행되고 있다. 자연언어의 연구 방향은 크게 두 가지 분류로 연구되고 있다. 규칙을 기반으로 하는 방법과 통계를 기반으로 하는 연구가 진행되어 왔다.

규칙 기반의 연구는 자연언어의 일반적인 규칙을 찾아내고, 그 규칙을 이용하여 접근하려는 시도로 제한된 영역에 대해서는 높은 정확도를 지니고 있다. 그러나 일반적인 규칙을 찾기가 힘들며, 다른 영역으로의 확장에 대해서 좋은 결과를 얻을 수 없었다

통계기반의 연구는 사람들이 사용하는 말이나 문장을 대량으로 모아놓은 즉 말뭉치로부터 확률 정보 및 통계 정보를 추출하여 여러 가지 언어 현상을 규명하려는 방법이다. 대량의 말뭉치로부터 정보를 추출하므로 확장성이 좋으며, 어떤 영역의 자료에 대해서도 견고하게 사용할 수 있다. 통계 기반으로 하는 연구에서는 말뭉치로부터 얻은 정보를 형태소 분석의 모호성 제거, 구문구조 분석시 가장 올바른 구문구조의 생성, 또는 미지격의 명사구의 문법기능 결정 등에 이용되고 있다. 그러나 이것만으로는 실세계에 맞는 말뭉치를 구하기가 쉽지 않다. 또한 높은 확률의 태깅 시스템을 구축하기위해서는 미정의 품사에 대해서도 높은 확률의 태깅이 이루어져야 한다. 이를 위해서 품사간인접도나 제약 정보를 추가하여 사용하므로 정확성 높은 태깅 시스템을 구축 하였다.

II. 정보 구축 및 태깅 알고리즘

1. 품사 태깅

어떤 문장을 형태소 분석을 하게 되면 상당히 많은 결과가 나오게 된다. 즉 한 어절에 대해서 서로 다른 분석이 나오게 된다. 이와 같은 것을 형태소 모호성이라고 한다. 그래서 이와 같은 형태소 모호성이 일어났을 경우 그 문장 가운데서 가장 적절한 형태의 형태소 분석을 결정하여 주는 과정을 품사 태깅이라고 한다.

예를 들면, 다음과 같은 문장이 있을 때,

예) "나는 학교에 간다"
"나는"은
"나(명사)+는(보조사)",
"날(동사)+는(어미)",
"나(동사)+(어미)",
"간다"는
"가(동사)+ㄴ다(어미)",

"갈(동사)+ㄴ다(어미)"

등으로 형태소 분석이 된다. 이와 같은 경우 "나는"는 "나(명사)+는(보조사)"로, "간다"는 "가(동사)+ㄴ다(어미)"로 결정하여 주는 것이 품사 태깅이다.

형태소 분석에서는 가능한 한 정확하고 많은 형태소 분석을 보여주는 것이 목적이다. 그러나 이를 구문분석이나 다른 분야에서 이용하려고 한다면, 이와 같이 많은 분석후보의 처리는 쉬운 일이 아니다. 많은 분석후보를 일일이 처리하여야 함으로 많은 연산이 필요하다. 구문 분석시에도 많은 구문구조의 발생을 초래하게 된다. 따라서 어절이 사용된 문장 내에서 올바른 품사를 결정하여 주는 것은 여러 분야에서 효율적인 처리에 많은 도움을 준다. 즉, 구문분석시에는 연산 횟수와 구문구조의 수를 현저하게 줄일 수 있다. 이를 위해서 확률 정보 및 통계 정보를 이용하여 단어에 알맞은 품사를 태깅 하고자 한다. 품사 태깅을 위한 방법으로는 신경망에 의한 방법과 통계적인 처리에 의한 방법이 있다.

본 논문에서는 태깅된 말뭉치에서 얻은 정보를 통계적인 처리를 이용하여 품사를 태깅 하고자 한다. 말뭉치에서 얻은 품사들의 올바른 품사열을 찾는 것이 문제이다. 이를 위해서는 Viterbi 알고리즘을 이용하였다[7]. 기존의 연구를 보면 단어간의 호응 정보를 이용하여 특정 단어1과 특정 단어2 사이의 관계를 확률 정보로 얻어 이용하였다[5]. 그러나 특정 단어1과 특정 단어2가 같은 문장 내에서 발생할 수 있는 횟수는 매우 적으므로 여기서 얻은 정보는 매우 불확실할 수가 있다. 이를 위해서 대량의 말뭉치를 구축하여야 하는데 이는 현실적으로 매우 힘든 일이다. 은닉마코프 모델 방법은 말뭉치로부터 얻은 정보를 가지고 학습을 시켜야 하며, 환경의 융통성이 적은 편이다[4].

본 연구에서는 특정 단어와 주위의 품사간의 관계를 이용하고자 한다. 이는 특정 단어1과 특정 단어2가 같은 문장 내에서 같이 나타나기 어려운 것을 극복하고 꼭 대량의 말뭉치가 아닌 말뭉치로부터도 정보를 추출하여 이용하고자 함이다. 이와 같이 얻어진 정보는 형태소 분석에서 나온 결과의 올바른 품사 결정에 이용한다. 여기서 정확하게 태깅된 단어는 다시 정보 추출에 이용되도록 정보 구축에 이용되므로 품사 태깅을 하면서 점점 더 많은 정보가 쌓이게 된다. 또한 구성되는 사전을 구축하는데 있어서 어절에 적용된 품사 패턴의 개수가 여러 개일 경우는 제약 규칙을 이용하여 올바른 품사 패턴을 선택하게 된다.

제약 규칙에서 규칙 정보를 이용한 품사 태깅 방법은 품사 태깅에 적용되는 공통된 원리나 결정적 규칙을 찾아낸다. 그리고 찾아낸 규칙을 품사 태깅에 적용하는 방법으로 규칙이 적용한다. 규칙을 찾아낸 어절에 대하여 거의 100%에 해당하는 정확성을 보이나 규칙을 찾기가 어렵고 적용 범위가 제한적인 단점이 있다. 규칙 기반 품사 태깅에서 어휘적 중의성을 해소하기 위하여 사용되는 언어 지식은 크게 긍정 언어 지식(positive linguistic information)과 부정 언어 지식(negative linguistic information) 그리고 수정 언어 지식(correctional linguistic)으로 나눌 수 있다.

긍정 언어 지식은 특정 문맥에서 어떤 품사가 올바른 품사인지를 나타내는 언어 지식을 말하고, 부정 언어 지식은 특정 문맥에서 부적절한 품사를 나타내기 위한 언어 지식을 말한다. 그리고 수정 언어 지식은 품사 태거가 자주 발생시키는 오류를 수정하기 위하여 사용되는 규칙을 말한다. 긍정 언어 지식은 어절의 품사를 바로 결정짓기 때문에 가장 정확하고 효과적으로 품사 태깅을 수행할 수 있는 언어 지식이다.

부정 언어 지식은 바로 품사 태깅을 수행할 수는 없지만, 분석 후보의 개수를 줄여줌으로써 분석 후보 중 올바른 후보를 선택할 수 있는 가능성을 높여준다. 경우에 따라서는 최종적으로 하나의 분석 후보만을 남겨 줌으로써 긍정 언어 지식과 같이 정확한 품사 태깅을 수행할 수 있다.[6]

2. 품사 분류

품사 정보를 추출하기 위해서는 말뭉치로부터 필요한 품사 정보를 추출하여야 하는데 이를 위해서는 품사의 종류를 결정하여야 한다. 품사를 세부적으로 분류하면 37종류로 분류할 수 있다[3]. 이와 같이 세분류를 하게 되면 단어간의 호응 정보와 같이 대량의 말뭉치가 필요하여야만 각 품사에 대해서 정확한 결과를 얻을 수 있다. 그래서 본 논문에서는 다음과 같이 5종류의 품사를 사용하였다.

"NP(명사구),VP(동사구),AP(부사구),DP(관형사구),$(문장의 시작과 끝)".

한국어는 어절단위로 사용되고 있기 때문에 어절 전체를 보고 품사를 결정하는 것이 좋다. 그래서 어절을 전체 단위로 하여 처리하는 것이 효율적이다.

III. 정보 구축 및 태깅 알고리즘

1. 정보 구축

본 논문에서 사용한 말뭉치는 수작업에 의해서 태깅 되었으며, 품사는 위에서 정한 것 중 NP, VP, AP, DP, $로 태깅 되어 있다. 이와 같이 태깅되어 있는 말뭉치로 1차적인 정보를 얻는데, 각 표제어의 전체 빈도수와 앞쪽의 품사에 대한 자신의 품사 값을 모두 조사한다. 2차적인 정보는 1차적인 정보로부터 전체 품사의 전이 정보와 각 품사에 대한 전체 정보를 구축한다. 본 시스템에서 사용하는 정보는 2차 정보에서 얻은 사전을 이용한다. 2차정보는 1차 정보에 비하여 거의 반 정도 분량이다. 사전의 구성은 다음과 같다.

```
 길을      22 6  1  0 0 11 0  0 0 1 0 0 0  0   0 0
0 3 0 0  0
 모르는    9 0  4  0 0   2 0 0 0 2 0 0  0 0   0 0
0 0 1 0  0
 사람이 94 11 0  0 0 56 0 0 0 1 0 0  0  22 0 0 0
4 0 0  0
있으면  11  0 10 0 0  0  1 0 0 0 0 0 0   0   0 0
0 0 0 0  0
```

그림 1. 1차 사전 구조

길을	22	21	1	0	0
모르는	9	4	5	0	0
사람이	94	94	0	0	0
있으면	11	0	11	0	0

그림 2. 2차 사전 구조

10068	11659	1852	674
7697	3721	494	370
1184	1903	176	131
1501	49	3	30
4122	612	916	7740

그림 3. 품사 인접도사전

2. 태깅 알고리즘

최적 품사열을 찾기 위한 알고리즘으로는 Viterbi 알고리즘을 변형하여 사용하였다. 가장 좋은 방법은 가능한 모든 품사열을 찾아보는 것이 좋으나, 현실적으로 불가능하고 비효율적이기 때문에 이에 가깝도록 하기 위한 Viterbi 알고리즘을 선택하였다.

Viterbi 알고리즘은 Morkov 가정에 의한 방법으로 가장 좋은 품사열을 효율적으로 찾을 수 있도록 하기 위해서, I번째 단어까지 가장 좋은 품사열로 태깅 되어 있다고 가정되었을 때 I번째에서 다음으로 전이될 확률과 그 단어의 빈도 확률을 이용하여 최적의 품사를 찾고자 하는 방법이다. 그러나 이 방법의 문제점은 말뭉치에 나타나지 않은 표제어가 나타난다든지, 계산한 확률 값이 0이 되는 경우가 발생하게 된다. 이와 같은 문제점을 극복하기 위해서 본 시스템에서는 형태소 분석이 된 결과에서 나타난 품사들 가운데서 최대 전이 확률과 최대 빈도를 이용하여 품사를 태깅 하였다.

IV. 시스템 구현

1. 시스템 개요

본 논문의 시스템은 하나의 문장을 형태소 분석을 한 후 임시 파일을 통하여 중간처리를 한다. 구성은 그림 4과 같이 구성되어 있다. 우선 형태소 분석기를 통해서 분석된 품사들을 입력 받는다. 입력 받은 품사들을 2.2절에서 분류한 품사 분류 방법에 맞게 분류한다. 명사 결합, 동사구에서 보조 동사구의 결합, 수와 관련된 단어의 처리 등을 한다. 여기서 나온 결과를 품사 태깅 시스템에 입력하여 최종 품사를 결정한다. 태깅에서는 3.1절에 기술한 사전들을 이용하며, 미 분석된 것은 인접도 사전이나 품사 제약 정보를 이용하여 최종 품사를 결정한다.

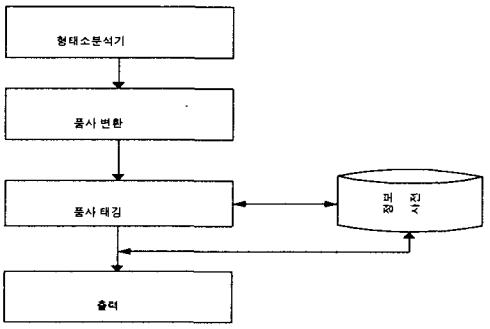

그림 4. 시스템 구성도

2. 형태소 분석기

본 시스템에서 자연언어처리 연구실에서 개발한 형태소 분석기를 이용한 것이다[1,2]. 이 형태소 분석기는 본 시스템을 초점으로 하여 개발된 것이 아니라 일반적인 형태소

분석을 위해서 개발된 것이다. 그러므로 본 시스템에서 사용하기 위해서 품사 변환 및 필요한 형태의 결과 값을 위해 변환 과정을 후처리로 하였다.

3. 확률 계산

품사 태깅을 위한 확률의 계산은 특정 단어가 나왔을 때 그것이 특정 품사가 될 확률 :

P(품사I | 어절) = P(어절 & 품사I) / P(어절)

어절의 품사가 품사 I 일 확률과 앞의 품사에서 현재 상태의 품사로 전이될 확률 :

P(현재의 품사 | 이전상태의 품사)

의 곱으로 얻어진다.
"나는" 이라는 어절이 문장의 시작에서 나올 수 있는 확률을 보면.

score(나는,i) = Max(P(NP | 나는) * P (NP | NULL),
P(VP | 나는) * P (VP | NULL),
P(AP | 나는) * P (AP | NULL),
P(DP | 나는) * P (DP | NULL))

가장 큰 값으로 태깅 된다. 이와 같이 결정된 값은 다음의 단어의 품사 태깅에 사용이 된다. Viterbi 알고리즘의 문제점인 사전에 표제어가 없을 때와 확률 계산이 0으로 나온 경우의 태깅은 형태소 분석에서 나온 결과에 있는 품사들 중에서 가장 전이 확률이 높은 것과 그 어절의 최대 빈도의 값의 곱으로 태깅한다.

P(형태소분석에서 나온 품사 | 전상태의 품사)

* 최대 빈도의 품사

본 논문에서는 bigram의 형태로 바로 인접한 품사만의 정보뿐만 아니라 인접할 수 없는 제약 정보를 추가하므로 좀 더 정확한 품사를 결정하였다.

V. 실험 결과 및 분석

실험대상 어절로는 자연언어처리 연구실에서 수집한 것으로 문장수는 700문장에, 약 5000어절을 대상으로 하였다. 확률처리에 의한 방법은 전체적으로 98.5% 정도를 보이고 있다. 이것은 미등록 어절인 경우의 확률 전이방법을 이용하였을 경우까지 포함한 것이다.

표 2. 결과표

적용방법	나타난 어절	오류 어절	오류 확률
확률 전이 (미등록어)	2000	33	1.65%
확률 계산 (등록어)	2899	17	.58%
전체	4899	50	.81%

오분석은 태깅된 문장의 잘못된 태깅으로 인하여 발생한 경우가 많으며, 잘못된 태깅이 아닌 경우도 올바르게 태깅되어야 할 품사의 확률 값이 적게 계산된 경우도 있었다. 따라서 정확한 문장을 가지고 올바른 태깅이 필요하다. 이전에 viterbi 알고리즘을 만을 이용하여 사용한 것 보다 미등록어에 대해서 인접도 사전과 품사 제약을 함께 적용하므로 더 나은 품사 태깅 시스템이 구축되었음을 알 수 있었다

중의성의 문제는 어절의 품사를 구 단위로 분류하였기 때문에 같은 품사로 형태소 분석에서 분석된 경우일 때는 둘을 모두 출력시키고 있다. 그러나 실제로 분석을 하여 본 경우는 같은 품사 내에서 중의성이 나타나는 경우가 그리 많지 않았다.

따라서 품사의 분류가 세분화 되지 않아도 품사를 태깅하는데 있어서는 큰 지장이 없었다.

VI. 결론

본 논문에서 구현한 품사 태깅 시스템은 확장성이 좋았으며, 표제어가 말뭉치에 존재하지 않는다 하더라도 가장 좋은 품사로 태깅을 할 수 있도록 구현되었다. 또한 품사 결정을 위해서 품사 제약 정보를 추가하여 사용하므로 더 나은 결과를 얻을 수 있었다. 품사의 수를 줄임으로써 적은

양의 말뭉치를 가지고도 통계정보를 구축하여 사용할 수 있게 되었다. 또한 형태소 모호성을 줄임으로써 형태소 분석을 이용하려는 다른 분야에서 이용할 때 매우 효과적으로 사용될 수 있다. 그러나 같은 품사 사이의 중의성 해결에는 아직 미약하게 되므로 같은 품사로 태깅된 경우의 중의성 해결을 위해서 좀 더 지속적인 연구가 이루어져야 하겠다.

참 고 문 헌

[1] James Allen, "Natural Language Understanding ", 2rd, 1995.

[2] 양재형외, "통계 정보를 활용한 한국어 미지격 명사구의 문법기능 결정", 정보과학회논문지, 제21권, 제5호, pp.808~815, 1994.

[3] 김재훈외, "은닉 마르코프 모델을 이용한 효율적인 한국어 품사 태깅", 정보과학회논문지, 제22권, 제1호, pp.136~146, 1995.

[4] 강유한, "어절간 주품사 정보와 제약 규칙을 이용한 한국어 품사 태깅", 충북대 컴퓨터공학과 석사학위논문, 2000.

[5] 박혜준외, "말뭉치 품사꼬리달기 시스템 구현",한국정보과학회 94 봄 학술발표논문집, 1994.

[6] 김창제, "부분적인 어절을 이용한 연어 정보에 기반 모호성 해소", 충북대 컴퓨터공학과 석사학위논문, 1996.

[7] 장동수, "음절에 기반한 한국어 형태소 분석기", 충북대 컴퓨터공학과 석사학위논문, 1993.

김영훈(Young-Hoon Kim) 정회원

1994년 2월 : 충북대학교 컴퓨터공학과 졸업(학사)

1997년 2월 : 충북대학교 컴퓨터공학과 졸업(공학석사)

1997년 2월 ~ 현재 : LG전자 CDMA연구소 선임연구원

<관심분야> : 자연언어처리, CDMA SMS, 휴먼 인터페이스, Mobile 콘텐츠, 소프트웨어공학