어절간 문맥 정보를 이용한 혼합형 품사 태깅

임 희 동†, 서 영 훈††
충북대학교 컴퓨터공학과
† hdlim@dcenlp.chungbuk.ac.kr ††yhseo@cbucc.chungbuk.ac.kr

Hybrid Part-of-Speech Tagging using Context Information among Words

Hee-Dong Lim, Young-Hoon Seo
Dept. of Computer Engineering, Chungbuk National University

요 약

본 논문에서는 규칙 정보와 통계 정보의 상호 보완적 특성을 이용한 혼합형 방법을 기반으로 규칙 정보와 통계 정보의 추출 및 적용 시에 어절간 문맥 정보를 보다 효율적으로 이용하는 혼합형 품사 태깅 시스템을 제안한다. 먼저 규칙이 적용되는 중의성들에 대해서 높은 정확률로 태깅을 수행한 후, 규칙으로 해결할 수 없는 중의성들에 대해서는 통계 정보를 이용하여 태깅을 수행한다. 규칙 정보는 중의성을 갖는 어절과 주변 어절들의 형태소 및 태그를 이용하여 정의하고 통계 정보는 문맥에 영향을 많이 미치고 많은 중의성의 원인이 되는 조사와 어미의 형태를 그대로 활용하여 추출함으로써 어절간 문맥을 보다 효율적으로 이용한다.

1. 서 론

최근 컴퓨터 하드웨어 비용의 지속적인 저하와 비약적인 성능의 향상으로 컴퓨터의 대중화와 더불어 사용자 인터페이스가 중요시되고 있다. 이에 따라 최근 컴퓨터의 자연언어처리가 차세대 사용자 인터페이스로서 각광 받으며 많이 연구되고 있는 실정이다.

자연언어처리 응용 분야들은 모두 기본 단계인 형태소 분석 과정을 거치게 되는데, 한국어의 경우 교착어라는 특성 때문에 형태소 분석 과정에서 여러 가지의 중의성이 발생하게 된다. 따라서 형태소 분석 결과를 자연언어처리 응용분야에서 그대로 사용할 수 없으며 중의성을 해소한 후 사용해야 한다. 이렇게 형태소 분석 후에 발생하는 중의성을 해소하는 시스템이 바로 품사 태깅 시스템이다. 기존의 품사 태깅 방법으로는 규칙을 이용하는 방법, 확률을 이용하는 방법 그리고 이 둘 모두를 이용한 혼합형 방법이 있고 이외에도 신경망을 이용하는 방법과 퍼지망을 이용하는 방법도 있지만 크게 확률을 이용하는 방법으로 볼 수 있다.

규칙을 이용하는 방법은 규칙이 적용되는 언어 현상에 대해 높은 정확률을 갖는 장점이 있지만 규칙으로 해결하지 못하는 예외적인 언어 현상이 존재하고 규칙의 획득과 관리에 많은 비용과 시간을 필요로 하기 때문에 처리 범위가 넓지 못하는 단점이 있다. 한국어의 경우 규칙만으로 해결하기 어려운 언어 현상이 많이 존재해서 규칙만을 한국어에 적용한 사례는 드물며, 영어권에서는 TAGGIT 시스템, Klein 시스템, Hindle 시스템 등이 있다[1].

이에 반해 확률을 이용하는 방법은 대량의 말뭉치로부터 추출한 통계 정보를 사용하므로 규칙이 적용될 수 없는 언어 현상에 적용할 수 있는 장점이 있지만 양질의 말뭉치를 대량으로 구축하기 힘들어 통계 정보 추출 시 심각한 자료 부족 문제가 발생하는 단점이 있다. 하지만 구축된 양질의 말뭉치는 다른 자연언어처리 응용분야에서도 유용한 기초 자료가 되므로 양질의 말뭉치 구축 자체로도 큰 의미를 가질 수 있다. 확률을 이용하여 중의성을 해결하는 대표적인 방법은 히든 마르코프 모델을 이용하는 방법이다[2, 3]. 이 방법은 중

의성을 갖는 주어진 형태소 분석 결과에 대해 어절 또는 형태소 단위로 확률이 가장 높은 품사열을 선택한다. 히든 마르코프 모델에 기반한 어절 단위 태깅 시스템으로는 3-gram의 상태 전이 확률과 형태소 단위의 어휘 발생 확률을 이용한 이운재의 HMM 시스템, 의사 부류를 사용하여 관측 심볼 확률을 학습한 이상주의 시스템 등이 있다[1]. 형태소 단위 태깅 시스템으로는 공유 단어열과 가상 단어 개념을 도입하여 다중 관측열에 대한 효율적인 태깅 알고리즘을 제안한 임철수의 시스템, Viterbi 알고리즘을 어절 단위로 전개시켜 최적의 형태소 결합열과 품사 결합열을 구한 이상호의 시스템이 있다[1]. 김진동의 시스템은 어절 단위 품사 태깅 시스템과 형태소 단위 품사 태깅 시스템을 결합한 2중 히든 마르코프 모델을 이용하였다[1].

최근에는 규칙 정보와 통계 정보를 모두 이용하여 태깅의 정확률을 높이는 혼합형 방법이 많이 연구되고 있다. 국내에서 혼합형 방법을 이용하여 한국어에 적용한 방법으로는 변형 규칙을 이용한 방법[4,5,6,7]과 언어 지식과 통계 정보의 보완적 특성을 이용한 방법[8] 그리고 통계 정보 적용 시에 제약 규칙을 이용한 방법[9] 등이 있다. 변형 규칙을 이용한 방법은 통계 정보를 이용하여 태깅을 수행한 후, 통계 정보에 의한 오류를 수정하기 위해 변형 규칙을 사용한 방법이다. 언어 지식과 통계 정보의 보완적 특성을 이용한 방법은 규칙 정보를 이용한 방법의 높은 정확률과 통계 정보를 이용한 방법의 넓은 처리 범위를 상호 보완적으로 이용한 방법이다. 즉, 언어 지식에 의한 결과를 선호하고, 언어 지식에 의해서 중의성이 해결되지 않은 어절에 대해서는 통계 정보를 이용하여 태깅을 수행하는 방법이다.

본 논문에서는 규칙 정보와 통계 정보의 상호 보완적 특성을 이용한 혼합형 방법을 기반으로 규칙 정보와 통계 정보의 추출 및 적용 시에 어절간 문맥 정보를 보다 효율적으로 이용하는 품사 태깅 시스템을 제안한다. 먼저 규칙이 적용되는 중의성들에 대해서 높은 정확률로 태깅을 수행한 후, 규칙으로 해결할 수 없는 중의성들에 대해서는 통계 정보를 이용하여 태깅을 수행한다. 규칙 정보는 중의성을 갖는 어절과 주변 어절들의 형태소 및 태그를 이용하여 정의하고 통계 정보는 문맥에 영향을 많이 미치고 많은 중의성의 원인이 되는 조사와 어미의 형태를 그대로 활용하여 추출함으로써 어절간 문맥을 보다 효율적으로 이용한다.

2. 규칙 정보의 추출 및 형태

규칙 정보를 추출할 대상 말뭉치를 먼저 형태소 분석기를 이용하여 형태소 분석 결과를 얻은 후, 중의성이 존재하고 말뭉치에 나타나는 빈도수가 많은 어절을 우선으로 그 어절이 출현한 문장을 모두 조사한다. 문장을 조사하는 과정에서 중의성을 가장 많이 해결할 수 있는 규칙을 중의성을 지닌 어절과 주변 어절들의 형태소와 태그를 이용하여 정의하게 된다.

대상 말뭉치로부터의 규칙 정보 추출은 수작업으로 인해 시간이 많이 소요되므로 수작업을 최소화하기 위해 규칙 정보 추출 도구를 이용하여 이루어진다. 규칙 정보 추출 도구는 형태소 분석기를 이용하여 대상 말뭉치에서 중의성을 갖는 어절들을 빈도수를 기준으로 보여주는 기능, 중의성을 갖는 특정 어절이 사용된 모든 문장을 추출하여 보여주는 기능 뿐 아니라 추출된 규칙 정보를 관리하는 기능도 제공한다.

추출되는 규칙 정보는 중의성을 갖는 어절을 중심으로 주변 어절들을 이용하여 정의되고, 규칙 정보를 추출할 때 규칙 정보는 규칙에 사용되는 모든 어절들의 형태소 및 태그를 이용하여 추출된다. 이렇게 규칙을 추출함으로서 주변의 어절 문맥을 좀 더 효율적으로 이용할 수 있고 비교적 적은 수의 일관성 있는 규칙으로 정확률이 높은 태깅을 수행할 수 있다. 그림 1은 추출되는 규칙 정보의 형태를 나타낸다.

규칙 정보의 형태를 살펴 보면 ':' 문자를 기준으로 하여 두 부분으로 구성된다. 첫 번째 부분은 현재 어절을 포함하여 앞 뒤 주변 어절들의 조건을 나타내고 한 개 이상의 조건으로 이루어진다. 두 번째 부분은 규칙의 조건이 모두 만족할 때 취해지는 행동을 나타낸다. 예를 들어, 자주 쓰이는 어구인 "~지 않~"를 살펴 보면, '하지'나 '적지'와 같이 '동사+어미'와 '명사'의 중의성을 갖는 경우가 많다. 이러한 중의성을 해결하는 규칙은 [TMT(지/EM),0][HMT(않/VV),+1]:[T(VV+EM),0,S]의 형태로 나타낼 수 있다. 즉, 중의성을 갖는 어절의 분석 결과 중, '지/EM'로 끝나는 분석 결과가 존재하고 다음 어절이 '않/VV'으로 시작되면 'VV+EM'의 분석 결과를 선택하게 된다. 또한 자주 쓰이는 관용구인 "~ㄹ 수~있/없~"는 [TMT(ㄹ/EM),0][HMT(수/NN),1][HMT(있/VV|없/VV),2]:[T(VV+EM),0,S][HMT(수/NN),1,S] 규칙으로 해결할 수 있다.

행동의 구성 요소 중, Add는 형태소 분석의 오류를 고려한 것이다. 형태소 분석의 오류로 인해 올바른 분

석 결과가 존재하지 않을 경우, Add를 이용한 규칙을 정의하여 적용함으로써 형태소 분석의 오류를 수정할 수 있다.

이렇게 문맥을 효율적으로 이용한 규칙을 정의하면 적은 수의 보다 간결하고 일관성 있는 규칙을 정의할 수 있고, 규칙이 적용되는 중의성을 갖는 어절들에 대해 높은 정확률을 얻을 수 있다.

```
([Cond,Pos])+:([Cond,Pos,Act])+
  * Cond: Condition, 조건
    - E(x): 어절이 x
    - TE(x): 어절이 x로 끝나면
    - HM(x): 첫번째 형태소가 x
    - TM(x): 마지막 형태소가 x
    - HT(x): 첫번째 태그가 x
    - TT(x): 마지막 태그가 x
    - T(x): 전체 분석 태그열이 x
    - MT(x): 형태소와 분석 결과가 x
    - TMT(x): 마지막 형태소와 분석 결과가 x
    - HMT(x): 첫번째 형태소와 분석 결과가 x
    - CT(x): 태그 x를 포함
    - 모든 조건에 쓰이는 인자 x는 '|'문자를
      기준으로 여러 개 가능
  *  Pos: Position, 위치
    - 0은 현재 어절
    - -x는 앞쪽의 x번째 어절
    - +x는 뒤쪽의 x번째 어절
  * Act: Action, 행동
    - S: Select, 선택
    - D: Delete, 제거
    - A: Add, 추가
```

그림 1 규칙 정보의 형태

3. 통계 정보의 추출 및 형태

한국어는 조사와 어미의 발달로 인해 형태소 분석 결과 조사와 어미로 인한 많은 중의성이 발생한다. 그림 2는 조사나 어미에 의해 발생하는 중의성의 예를 나타낸다. 이렇게 조사나 어미에 의한 중의성을 해결하고자 조사나 어미의 형태와 앞 뒤 주변 어절을 통계 정보로 이용한다. 조사나 어미 자체가 문맥에 많은 영향을 미치고 앞 뒤 주변 어절을 이용함으로써 문맥을 보다 효율적으로 이용하는 통계 정보를 구성할 수 있다.

```
[적지]:적지/NN:적/VV 지/EM
[수도]:수도/NN:수/NN 도/PP
[수출하고]:수출/NN 하고/PP:수출/NN 하/SF 고/EM
[젖을]:젖/NN 을/PP:젖/VV 을/EM
[적기로]:적기/NN 로/PP:적/VV 기/EM 로/PP
```

그림 2 조사나 어미에 의한 중의성

통계 정보는 신뢰성을 위해 태깅이 이미 수행된 양질의 말뭉치로부터 추출된다. 통계 정보 추출에 이용된 대상 말뭉치는 연구용으로 배포되는 29만 어절로 구성된 ETRI 품사 태그 부착 말뭉치이다. 3어절을 단위로 하여 추출하되 많은 중의성의 원인이 되고 문맥에 많은 영향을 주는 조사와 어미는 형태소 그대로, 다른 품사들은 태그 형태로 추출하게 된다. 통계 정보는 규칙 정보와는 달리 대상 말뭉치로부터 통계 정보 자동 추출 도구를 이용하여 완전히 자동으로 추출된다. 통계 정보 추출 과정에서 이미 태깅된 대상 말뭉치의 품사 태그 집합이 형태소 분석기의 태그 집합과 틀리기 때문에 태그 맵핑이 필요하게 된다. 통계 정보 자동 추출기는 한 문장을 단위로 하여 앞뒤에 각각 SOS(Start of Sentence)와 EOS(End of Sentence)를 나타내는 특수 어절을 추가한 후 3어절씩 통계 정보 추출에 이용한다.

이렇게 통계 정보 추출 시에 이미 태깅이 수행된 양질의 말뭉치를 이용하고, 많은 중의성의 원인이 되는 조사나 어미의 형태를 그대로 이용함으로써 신뢰성 있고 어절간 문맥을 효율적으로 이용한 통계 정보를 획득할 수 있다. 그림 3은 추출된 통계 정보의 형태를 나타내고 그림 4는 어절 노드 'VV+ㄴ' 대해 추출된 통계 정보를 나타낸다.

```
[CEN] (PEN NEN Count)+
  CEN: Current Eojul Node
  PEN: Previous Eojul Node
  NEN: Next Eojul Node
  * Eojul Node
    - '+'기호로 연결된 하나 이상의 태그열
      단, 조사와 어미는 형태소 그대로 이용.
    - SOS: Start Of Sentence, 문장의 시작을
      나타내는 특수 노드
```

- EOS: End Of Sentence, 문장의 끝을
  나타내는 특수 노드

* Count
- CEN 을 기준으로 앞뒤로 각각 PEN 과 NEN 이
  발견된 횟수

그림 3 통계 정보의 형태

[VV+ㄴ]
AD AD 1
AD DT 8
AD NN 112
AD NN+가 27
AD NN+과 11
AD NN+까지 2
AD NN+는 19
AD NN+다가 1
AD NN+도 12
AD NN+라고 1
AD NN+라고+는 2
...

그림 4 'VV+ㄴ'어절 노드의 통계 정보

4. 실험 및 결과

실험에 사용된 말뭉치는 ETRI 에서 연구용으로 제작된 품사 태그 부착 말뭉치 약 29 만 어절이다. 규칙 정보 추출 도구를 이용하여 약 1 만 어절로부터 35 개의 규칙 정보가 추출되었고, 통계 정보는 29 만 어절을 모두 이용하여 추출되었다. 태깅의 입력 및 규칙 정보 추출에 사용되는 형태소 분석 결과는 본 연구실이 보유하고 있는 한국어 형태소 분석기 CBKMA 2.0 을 이용하여 얻었다. 형태소 분석과 태깅에 사용된 품사 태그는 총 15 개이며 표 1 과 같다.

품사	태그
명사	NN
의존명사	NX
대명사	NP
수사	NU
동사	VV
형용사	VJ
보조용언	VX
지정사	CP
부사	AD
관형사	DT
감탄사	IJ
접미사	SF
조사	PP
선어말어미	EP
어말어미	EM

표 1 태그 집합

품사 태깅 실험은 규칙 정보와 통계 정보 추출에 이용된 내부 평가 말뭉치와 그렇지 않은 외부 평가 말뭉치에 대해 각각 수행되었고, 표 2 는 실험 말뭉치에 대한 통계 정보를 나타낸다. 표 2 에서 중의성을 가진 어절수는 어절의 형태소 분석 개수가 2 개 이상인 어절의 수를 나타내고, 중의도는 중의성을 가진 어절의 평균 형태소 분석 개수를 의미한다.

말뭉치	어절수	중의성을 가진 어절수	중의도
내부 말뭉치	11897	4240	2.70
외부 말뭉치	1681	579	2.55
계	13578	4819	2.63
말뭉치	**어절수**	**형태소 분석 오류 어절수**	**형태소 분석 오류율**
내부 말뭉치	11897	333	2.78%
외부 말뭉치	1681	52	3.09%
계	13578	385	2.83%

표 2 실험 말뭉치 통계 정보

표 3 은 본 논문에서 제안한 방법 중 통계 정보 만을 이용했을 경우의 품사 태깅 결과를 나타낸다. 정확률은 중의성을 갖는 어절에 대해 올바르게 태깅된 어절의 비율을 나타낸다. 내부 말뭉치에 대해서는 91.7%의 정확률을, 외부 말뭉치에 대해서는 90.3%의 정확률을 얻었다. 정확률 계산 시에 형태소 분석이 잘못 수행된 어절과 이 어절의 영향을 받아 잘못 태깅된 앞 뒤 어절은 제외하였고 동품사 중의성을 포함하는 어절도 제외하였다. 태깅 오류를 분석하는 과정에서 오류를 해결하기 위한 규칙을 작성하여 규칙 정보를 확장하였다.

말뭉치	중의성을 가진 어절수	맞게 태깅된 어절수	정확률
내부 말뭉치	4240	3888	91.7%
외부 말뭉치	579	523	90.3%
계	4819	4411	91.5%

표 3 통계 정보만을 이용한 품사 태깅 결과

표 4은 본 논문에서 제안한 규칙 정보와 통계 정보를 모두 이용했을 경우의 품사 태깅 결과를 나타낸다. 실험 결과, 내부 말뭉치에 대해서 95.1%의 정확률을, 외부 말뭉치에 대해서 94.3%의 정확률을 보임으로써 비교적 적은 수의 규칙과 적은 양의 통계 정보로도 품사 태깅이 비교적 정확하게 수행되었음을 알 수 있다. 일관성 있는 규칙들을 추가하고 더 많은 어절로 구성된 다른 말뭉치를 이용하여 통계 정보를 확장하면 더 높은 정확률을 얻을 것으로 기대된다.

말뭉치	중의성을 가진 어절수	맞게 태깅된 어절수	정확률
내부 말뭉치	4240	4033	95.1%
외부 말뭉치	579	546	94.3%
계	4819	4579	95.0%

표 4 규칙 정보와 통계 정보를 모두 이용한 품사 태깅 결과

5. 결 론

본 논문에서는 규칙 정보와 통계 정보의 상호 보완적 특성을 이용한 혼합형 방법을 기반으로 규칙 정보와 통계 정보의 추출과 적용 시에 어절간 문맥을 효율적으로 고려한 혼합형 태깅 방법을 제안하였다. 먼저 규칙 정보를 이용하여 규칙이 적용되는 언어 현상에 대해서 높은 정확률로 태깅을 수행한 후, 규칙으로 해결할 수 없는 언어 현상에 대해서 통계 정보를 이용하여 태깅을 수행하였다. 규칙 정보 추출 시에는 중의성을 갖는 어절과 주변 어절들의 형태소 및 태그를 이용하고, 통계 정보 추출 시에도 문맥에 많은 영향을 주고 많은 중의성의 원인이 되는 조사와 어미의 형태를 그대로 활용함으로써 어절간 문맥을 효율적으로 이용하였다. 내부 말뭉치와 외부 말뭉치에 대한 실험 결과, 적은 수의 규칙과 적은 양의 통계 정보로도 비교적 높은 정확률을 보임을 알 수 있었다.

향후 연구 과제로는 일관성을 지닌 보다 많은 규칙의 추가가 필요하고, 통계 정보의 신뢰성을 위해 더 많은 양의 말뭉치로부터 통계 정보를 확장할 필요가 있다. 보다 높은 정확률을 위해 품사 태깅에 실패한 어절들의 오류를 수정할 수 있는 규칙 정보를 정의하여 적용하는 것이 필요하다. 뿐만 아니라 동품사 중의성을 해결하기 위해 사전 정보를 이용할 수 있는 규칙의 정의 및 추가도 필요하다.

참고 문헌

[1] 임해창, 임희석, 이상주, 김진동, "자연언어처리를 위한 품사 태깅 시스템의 고찰", 한국정보과학회지, 제 14 권, 제 7 호, pp.36-57, 1996.

[2] 이상주, 임희석, 임해창, "은닉 마르코프 모델을 이용한 두단계 한국어 품사 태깅", 제 6 회 한글 및 한국어 정보처리 학술대회 발표 논문집, pp.305-312, 1994.

[3] 임철수, "HMM 을 이용한 한국어 품사 태깅 시스템 구현", 한국과학기술원 전산학과 석사학위논문, 1994.

[4] 신상현, 이근배, 홍남희, 이종혁, "확률과 규칙을 사용한 품사 태깅", 제 6 회 한글 및 한국어 정보처리 학술대회 발표 논문집, pp.318-321, 1994.

[5] 신상현, 이근배, 이종혁, "TAKTAG: 통계와 규칙에 기반한 2 단계 학습을 통한 품사 중의성 해결", 제 7 회 한글 및 한국어 정보처리 학술대회 발표 논문집, pp.169-174, 1995.

[6] 임희석, 김진동, 임해창, "한국어 특성에 적합한 변형 규칙 기반 한국어 품사 태깅", 춘계 인공지능 연구회 학술발표 논문집, pp.3-10, 1996.

[7] 임희석, 김진동, 임해창, "변형 규칙 기반 한국어 품사 태거의 개선", 제 8 회 한글 및 한국어 정보처리 학술대회 발표 논문집, pp.216-221, 1996.

[8] 임희석, 김진동, 임해창, "언어 지식과 통계 정보의 보완적 특성을 이용한 품사 태깅", 고려대학교 컴퓨터과학과 자연언어처리 연구실.

[9] 강유환, "어절간 주품사 정보와 제약 규칙을 이용한 한국어 품사 태깅 시스템", 충북대학교 컴퓨터공학과 석사학위논문, 1999.