# 변형 규칙 기반 영어 품사 태깅 시스템의 설계및 구현

이태식 ° 이상윤 최병욱 김한우 *
한양대학교 전자통신공학과, 한양대학교 전자계산학과 *

# Design and Implementation of English part of speech tagging system by transformation rule base.

Taesik Lee, Sang-yun Lee, Byunguk Choi, Hanwoo Kim *
Dept. of Electronic Communication Engineering, Hanyang University
* Dept. of Computer Science & Engineering, Hanyang University
Email : skyhawk@hymail.hanyang.ac.kr

## Abstract

In this paper, a transformation-based English part of speech tagging system is designed and implemented. The tagging system tags raw corpus at first and the transformation rule correct the errors. Apart from traditional rule based tagging system, this system makes rules automatically. Using 60,000 words of corpus as a training corpus, the transformation rules are generated automatically by iterative training. The idea how to calculate positive effect of transformation and select transformation rules is proposed to generate more effective and correct transformations.

In this paper, part of the Brown corpus and English text is used for experimental data. And the performance of transformation based tagging system is demonstrated by the calculation of accuracy.

## 1. 서 론

품사 태깅 시스템은 통계 기반 품사 태깅 시스템과 규칙 기반 품사 태깅 시스템으로 크게 분류할 수 있다. 통계 기반 품사 태깅 시스템은 은닉 마르코프 모델[1]과 Viterbi 알고리즘을 사용하여 구현한다.[2] 대량의 말뭉치에서 추출한 통계 정보를 바탕으로 최적의 태그를 얻어내는 방식이다. 그런데 대량의 통계 자료를 얻기 위해서는 많은 노력이 필요하다. 규칙 기반 품사 태깅 시스템은 수많은 규칙들을 기반으로 태깅을 하는 방법이다.[3] 그런데, 규칙 기반 시스템은 규칙의 생성과 적용에 많은 문제점이 존재한다.

본 논문에서는 개선된 규칙기반 품사 태깅 시스템을 제안한다. 변형 규칙을 도입해 자동적으로 규칙을 생성하고, 선별하여 태깅을 수행하는 방법을 제안한다.

본 논문에서는 미등록어가 없다는 전제하에 시스템을 구현하였으며, 규칙의 생성, 선별 알고리즘을 제시하고 높은 정확도를 나타낼 수 있음을 실험 결과를 바탕으로 보여준다.

## 2. 시스템 개요

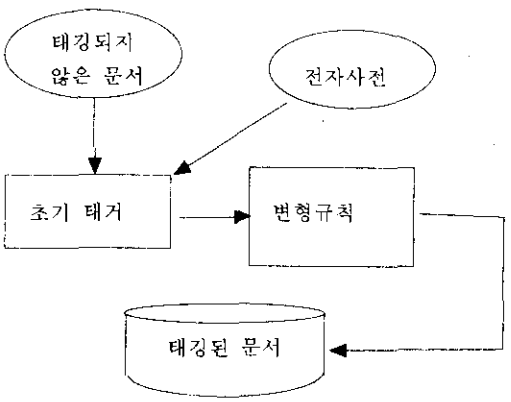


그림 1. 변형 규칙 기반 품사 태깅 시스템

우선, 초기 태거로 태깅을 수행한 후, 변형 규칙을 사용하여 오류를 바로잡아 최종 태깅 결과값을 얻게 된다. 초기 태거로는 여러 가지 태거를 이용할 수 있는데, 모든 단어를 명사로 태깅하는 태거나, N-gram 태거를 사용할 수 있다. 본 논문에서 사용된 초기 태거는 각 단어의 품사 중 가장 빈번하게 쓰이는 품사를 선택한다.

## 3. 변형 규칙(Transformations)의 생성

### 3.1 변형 규칙의 생성 과정

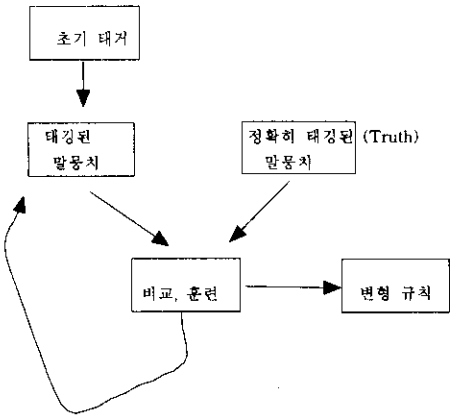


그림 2. 변형 규칙의 생성 과정

초기 태거에서 태깅된 말뭉치와 정확히 태깅되어진 말뭉치를 하나 하나 비교해 나간다. 오류가 발생하게 되면 발생한 오류를 수정하는 변형 규칙들이 생성되고 비교가 끝나면, 초기 태거로 태깅한 말뭉치에 변형 규칙들을 적용한다. 다시 변형 규칙을 적용한 말뭉치와 정확히 태깅된 말뭉치와 비교해 오류를 찾고 또 이를 처리할 수 있는 변형 규칙들을 생성한다. 이 과정을 더 이상 오류 개선의 효과가 없을 때까지, 즉 오류가 더 이상 줄지 않을 때까지 계속하여 최종적으로 변형 규칙을 결정한다.

### 3.2 변형틀(Transformation Template)

변형 규칙은 변형식과 변형 조건으로 구성된다. 이러한 형식에 맞추어 변형틀을 구성하여 변형 규칙 훈련 과정에서는 변형틀(Transformation Template)에 맞추어 변형 규칙을 생성한다.

| |
|---|
| 변형식<br>태그 A를 태그 B로 바꾼다<br>변형조건<br>왼쪽 단어가 Z로 태깅되었을 경우<br>왼쪽 두 단어중 하나가 Z로 태깅되었을 경우<br>오른쪽 단어가 Z로 태깅되었을 경우<br>. |

표 1. 품사 관계를 고려한 변형틀

| |
|---|
| 변형식<br>태그 A를 태그 B로 바꾼다<br>변형조건<br>왼쪽 단어가 W일 경우<br>왼쪽 두 단어중 하나가 W일 경우<br>오른쪽 단어가 W일 경우<br>. |

표 2. 단어 관계를 고려한 변형틀

변형틀은 주위 태그 관계를 고려하는 것과 단어 관계도 고려하는 것, 두 가지로 이루어진다. 기존의 태거가 태그 관계만을 태깅 정보로 이용했던 것에 비해, 본 논문에서 제안한 변형틀은 단어 관계까지도 고려하는데, 그 이유는 특정 단어에 관련된 현상까지 처리하기 위함이다.

**as**/IN tall/JJ as/IN ⟶ **as**/RB tall/JJ as/IN

위의 예에서 볼 수 있듯이 태그 관계만을 고려하면 "오른쪽 단어의 태그가 JJ일 경우 IN을 RB로 바꾼다"는 변형 규칙이 생성될 수 있다. 하지만 IN JJ의 태그열이 얼마든지 존재할 수 있으므로 생성된 변형 규칙은 적당하지 못하다. 따라서 단어 관계를 고려한 변형틀로 "오른쪽으로 두 번째 단어가 as일 경우, IN을 RB로 바꾼다"라는 변형 규칙을 생성시키게 되면 이 문제를 해결할 수 있다.

### 3.3 변형 규칙의 훈련

초기 태거로 태깅된 문서와 정확히 태깅된 문서를 비교할 때, 변형틀에 맞추어 오류를 수정하는 변형 규칙을 생성한다. 이 때 변형틀에 의해 만들어질 수 있는 가능한 모든 규칙을 생성하게 된다.

| 변형틀 |
| --- |
| 태그 A를 태그 B로 바꾼다 |
| 왼쪽 단어가 Z로 태깅되었을 경우 |
| 왼쪽 두 단어 중 하나가 Z 경우 |

표 3. 변형틀의 가정

위와 같이 변형 조건이 단 두 개인 변형틀이 있고 다음같은 태깅 오류가 있다고 하면,

초기 태깅 : The/DT can/MD rusted/VBD
참 (Truth) : The/DT can/NN rusted/VBD

| 1. 왼쪽 단어가 DT로 태깅되었을 경우<br>2. 왼쪽 두 단어 중 하나가 The인 경우<br>3. 왼쪽 두 단어 중 하나가 Blank인 경우<br><br>태그 MD를 태그 NN으로 바꾼다 |
| --- |

표 4. 생성된 변형 규칙의 예

위와 같이 변형 규칙이 생성된다. 이 중에서 적절하다고 생각되는 것은 첫 번째 변형 규칙이다. 변형 규칙은 많이 생성되지만 올바른 것은 그 중 한 두개에 불과하기 때문에 생성된 수많은 변형 규칙들을 선별해 내는 장치가 필요하게 된다.

## 4. 변형 규칙의 선별

적절한 변형 규칙만을 선별해 내는 작업은 태깅에 얼마나 도움을 주느냐에 따라 달라진다. 본 논문에서는 세 가지의 선별 기준을 제시한다.

### 4.1 변형 규칙의 생성 빈도수

태깅 정확도에 도움을 줄 수 있는 규칙은 많은 오류를 수정할 수가 있다. 따라서 올바른 규칙은 생성 빈도수가 높다고 볼 수 있다. 한계치(Threshold)를 설정하여 규칙을 선별해 낼 수 있다. 그러나 적당한 규칙이 생성 빈도수는 높지만, 올바르지 못한 규칙도 높은 생성 빈도수를 가질 수 있기 때문에 절대적인 선별 기준은 될 수가 없다. 따라서 생성 빈도수가 적은 변형 규칙만을 제외하도록 하였다.

### 4.2 변형 규칙의 태깅 기여도

생성된 변형 규칙이 얼마나 많은 오류를 수정할 수 있는가 하는 것은 변형규칙 선별에 중요한 기준이 된다. 생성된 변형 규칙을 변형틀에 따라 분류, 데이터화하여 기여도가 높은 것을 찾게 된다. 변형 규칙이 올바른 태깅이 했다면 태깅 기여도를 증가시키고, 올바르지 못하게 태깅을 했다면 태깅 방해도를 증가시켜 (태깅기여도 - 태깅 방해도)가 가장 높은 규칙을 최종적으로 선택한다.

```
바로 앞의 단어가 Z로 태깅된 경우 태그 A를 태그 B로 바꾼다
① 초기태거 태깅을 한다.
② 조건(앞의 단어가 Z로 태깅)이 만족되는 동안
③   태그A = 태그1부터 태그n까지
④     태그B = 태그1부터 태그N까지
⑤       말뭉치의 처음부터 끝까지
⑥         현재태그 = 태그 A, 정확한 태그 = 태그 B 이면
                  태그N에 대한 태깅 기여도 증가
⑦         현재태그 = 태그 A, 정확한 태그 = 태그 A 이면
                  태그N에 대한 태깅 방해도 증가
⑧     ( 기여도 - 방해도 )의 값이 가장 큰 태그를 찾음
⑨     이 값이 지금까지의 값중 최대이면 다음 변형 규칙을 선택
  목록에 추가
    "왼쪽 단어가 Z로 태깅된 경우 태그 A를 태그 B로 바꾼다"
```

표 5. 기여도 계산 알고리즘

### 4.3 잘못된 훈련에 의한 규칙의 제거

초기 태거로 태깅한 문서에는 연속적인 오류가 생길 수 있다.

F F F T T F F (T : 참 F : 오류)

연속적인 오류가 생긴 경우에는 주위 태그가 모두 잘못된 것이므로 변형 규칙 역시 잘못 생성되는 결과를 초래한다. 그러므로, 연속적인 오류에 대해 훈련을 하는 경우는 미리 제거하여 잘못 생성되는 것을 방지하였다.

## 5. 실험 및 평가

### 5.1 실 험

Brown corpus에서 60,000 단어 정도의 문장을 추출하여 훈련시켰다. 그 결과 200여개의 변형 규칙이 생성되었다. 태깅 정확도를 측정하기 위하여 Brown corpus에서 4,000여 단어의 문장을 추출하였고, 고1 영어 교과서에서 1,000여 단어의 문장을 추출하여 정확도를 구했다. 본 태거는 미등록어 처리를 고려하지 않았으므로, 미등록어는 정확도 계산에 포함시키지 않았다.

| | Brown corpus | 교과서 |
|---|---|---|
| 총 단어수 | 4,142개 | 1,023개 |
| 정확히 태깅된 단어 | 3,952개 | 990개 |
| 잘못 태깅된 단어 | 123개 | 23개 |
| 말뭉치자체의 오류 | 51개 | 0개 |
| 태깅 정확도 | 97.0 % | 97.7 % |

표 6. 실험 결과

태깅 정확도

$$= \frac{\text{정확히 태깅된 단어}}{\text{정확히 태깅된 단어} + \text{잘못 태깅된 단어}}$$

Brown corpus의 경우에 품사 태그를 변환하는 과정에서 모호성이 생겨 정확한 태그를 알 수 없는 경우가 발생하였는데, 이를 말뭉치 자체의 오류로 설정하고 태깅 정확도의 계산에는 포함시키지 않았다.
영어 교과서의 문장이 평이하고 단어 수준도 쉬워서 Brown corpus보다 높은 정확도를 보인 것으로 파악된다.

5.2 평가

| 태거 | 태깅방법 | 정확도 |
|---|---|---|
| Klein & Simmons [4] | 규칙기반 | 90 % |
| Xerox [5] | 통계기반 | 96 % |
| Church [6] | 통계기반 | 95-99 % |
| Pasi Tapanainen [7] | 통합기반 | 97-99 % |
| **본 논문의 태거** | **규칙기반** | **97-98 %** |

표 7. 다른 태거와의 비교

통합 기반 품사 태깅 시스템은 규칙 기반 품사 태깅 시스템과 통계 기반 품사 태깅 시스템을 병렬적으로 결합한 것이다. 태깅 정확도는 높지만 두 개의 태거를 별개로 구성하기 때문에 태깅 속도가 떨어지고 개발하기에도 어려움이 크다. 실제로 Tapanainen의 태거는 기존의 규칙 기반 태거와 통계 기반 태거를 단순 결합한 것이다.[7]

본 논문에서 구현한 태거는 미등록어 처리를 고려하지 않았고 실험 말뭉치도 서로 다르기 때문에 절대적인 비교는 될 수가 없다. 앞으로 미등록어 처리를 고려한 연구가 진행되어야 한다고 본다.

## 5. 결 론

본 논문에서는 변형 규칙 기반 영문 품사 태깅 시스템을 제안하였다. 변형틀을 구성하여 말뭉치를 통한 학습에 의해 자동적으로 변형 규칙을 생성하였고, 선별 기준을 두어 생성된 규칙 중 올바른 것을 선별하였다. 이렇게 선별된 변형 규칙과 초기 태거를 조합하여 태깅해 본 결과 97%이상의 좋은 태깅 정확도를 얻을 수 있었다.

그러나 보다 안정적이고 정확한 태깅을 위해서는 미등록어 처리에 대한 고려가 이루어져야 하고, 변형 규칙을 선별하는 기준에 대한 연구가 좀 더 이루어져야 할 것이다.

## 참 고 문 헌

[1 L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition" Proc. of the IEEE Acoust. Speech Signal Processing, vol. 77. no. 2, pp. 257-286 Feb. 1989

[2] G. D. Forney, "The Viterbi Algorithm," Proc. of the IEEE, vol. 61, pp. 268-278, Mar. 1973

[3] E. Brill "A Simple Rule Based Part of Speech Tagger," Proc. of the 3rd Conf. on Applied Natural Language Processing, Tzento, Italy, pp153-155, April. 1992

[4] Klein, S. and Simmons, R. F. "A Computational Coding of English Words", Journal of ACM pp. 334-347, 1963

[5] Doug Cutting, Julian Kupiec, Jan Pedersen and Penelope Sibun "A Practical Part-of-Speech Tagger" Proceedings of ANLP-92 1992

[6] K. W. Church, "A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text" Proceedings of Applied Natural Language Processing, Austin, Texas, pp. 136-143, 1988

[7] Pasi Tapanainen, "Tagging accurately - Don't guess if you know" Proc. of the 7th conference of the European chapter of the Association for Computational Linguistics, pp.149-156 Aug. 1994