언어 지식과 통계 정보의 보완적 특성을 이용한 품사 태깅

임희석†, 김진동, 임해창

†삼성 종합 기술원 휴먼 인터페이스 랩

고려대학교 컴퓨터학과 자연어처리 연구실

Part-of-Speech Tagging Using Complemental Characteristics of Linguistic Knowledge and Stochastic Information

Heui-Seok Lim†, Jin-Dong Kim, Hae-Chang Rim

†Human Interface Lab., Samsung Advanced Institute of Technology

NLP. Lab., Dept. of Computer Science and Engineering, Korea Univ.

Abstract

기존의 품사 태깅 방법에서 독립적으로 사용해온 언어 지식과 통계 정보는 품사 태깅의 정확도와 처리 범위의 향상을 위해서 상호 보완적인 특성을 갖는다. 이에 본 논문은 언어 지식과 통계 정보의 보완적 특성을 이용한 규칙 우선 직렬 품사 태깅 방법을 제안한다. 제안된 방법은 언어 지식에 의한 품사 태깅 결과를 선호함으로써 규칙 기반 품사 태깅의 정확도를 유지하며, 언어 지식에 의해서 모호성이 해소되지 않은 어절에 통계 정보에 의한 품사 태깅 결과를 할당함으로써 통계 기반 품사 태깅의 처리 범위를 유지한다. 또한, 수정 언어 지식에 의해 태깅 결과의 오류를 보정함으로써 품사 태깅의 정확도를 향상시킨다. 약 2만 어절 크기의 외부 평가 코퍼스에 대해 수행된 실험 결과, 규칙 우선 직렬 품사 태깅 시스템은 통계 정보만을 이용한 품사 태깅의 정확도보다 32.70% 향상된 95.43%의 정확도를 보였다.

제 1 절 서론

문장내에서 단어가 사용된 문맥에 따라 각 단어에 올바른 품사 정보를 부여하는 품사 태깅 방법은 크게 규칙 기반 품사 태깅 방법과 통계 기반 품사 태깅 방법으로 구분할 수 있다[14].

규칙 기반 방법은 언어 지식을 이용하여 결정적으로 단어의 모호성을 해소하는 방법으로 규칙이 적용되는 언어 현상에 대하여 높은 정확도를 갖는다는 장점이 있다. 하지만 규칙 기반 방법은 다음과 같은 어려움으로 처리 범위가 넓지 못하고, 대량의 코퍼스로의 확장성이 좋지 않다는 단점이 있다. 첫째, 다양한 언어 현상을 처리하기 위한 방대한 량의 규칙을 필요로 하고, 이를 획득하고 관리하기 위하여 많은 비용을 필요로 한다. 둘째, 규칙과 같은 결정적인(deterministic) 지식으로는 해결하기 어려운 자연어의 예외적인 언어 현상을 처리하기 힘들다. 셋째, 많은 언어 현상을 반영하기 위하여 규칙을 증가시킬 경우 특정 언어 현상에 효과적인 규칙이 다른 언어 현상에도 제대로 적용될 것인지 보장하기 어렵다.

규칙 기반 방법에 반하여 통계 기반 방법은 대량의 코퍼스로부터 추출한 통계 정보를 사용하여 가장 확률이 높은 결과를 선택함으로써 다양한 언어 현상에 적용할 수 있는 장점을 갖는다. 또한 확률값과 같은 비결정적인(nondeterministic) 정보를 사용함으로써 비문장(ill-sentence)이나 언어 규칙으로 설명할 수 없는 현상에도 적용할 수 있다. 하지만 통계 기반 방법은 다음과 같은 어려움으로 인하여 규칙 기반 방법에 비하여 정확도가 낮다는 단점이 있다. 첫째, 실세계의 언어 현상을 제대로 반영할 수 있는 양질의 코퍼스를 구축하기 매우 어려우므로 코퍼스로부터 통계 정보를 추출할 때 데이터 부족 문제가 심각하게 발생한다. 둘째, 통계 기반 접근법에서 사용하는 품사 태깅 모델은 데이터 부족 문제의 완화와 통계 정보량의 감소를 위하여 근거리 문맥 정보(local contextual information)만을 사용한다. 하지만 자연어에는 근거리 문맥

정보만으로는 해결될 수 없는 언어 현상들이 많이 발생한다. 셋째, 통계 기반 품사 태깅 방법은 많은 양의 파라미터를 요구하는 어휘들간의 관계를 모델링하기 어렵다. 따라서 어휘간의 관계를 고려하면서 정확하게 품사의 모호성을 해결할 수 있는 단어에 대해서도 부정확한 품사 태깅을 수행할 수 있다. 넷째, 통계 기반 품사 태깅 방법은 품사 태깅 가능한 모든 품사열 중 확률값이 최대인 품사열을 선택한다. 따라서 언어 지식을 이용하여 정확하게 품사 태깅이 가능한 단어에 대해서도 신뢰도가 낮은 통계 정보를 사용하므로 품사 태깅의 정확도가 저하된다.

자연어처리 분야의 여러 응용 분야에서 반드시 필요한 품사 태깅 시스템은 높은 정확도와 넓은 처리 범위를 가질 수 있어야 한다. 그러나 기존에 개발된 품사 태깅 시스템은 모호성 해소를 위한 언어 지식과 통계 정보를 독립적으로 사용하므로 앞서 기술한 바와 같이 처리 범위가 제한적이거나 정확도가 낮다는 문제점이 있다. 하지만 규칙 기반 품사 태깅 방법과 통계 기반 품사 태깅 방법은 각 방법이 안고 있는 처리 범위와 정확도 향상에 도움을 줄 수 있는 보완적인 특성을 갖는다. 즉, 규칙 기반 품사 태깅의 높은 정확도는 통계 기반 품사 태깅의 정확도 향상에 도움을 줄 수 있으며, 통계 기반 품사 태깅의 넓은 처리 범위는 규칙 기반 품사 태깅의 처리 범위 향상에 기여할 수 있다. 이에 본 논문은 품사 태깅의 처리 범위와 정확도 향상을 위하여 언어 지식과 통계 정보의 보완적 특성을 이용한 한국어 품사 태깅 방법을 제안한다.

제 2 절 관련 연구

품사 태깅을 위하여 언어 지식과 통계 정보를 통합하여 사용하고자 한 대표적인 기존 연구로는 변형 규칙을 이용한 방법[9, 15], Zhang의 연구[5] 그리고 Tapanainen의 연구[3]를 들 수 있다.

변형 규칙을 이용한 방법은 통계 기반 품사 태거로 품사 태깅을 수행하고, 통계 정보에 의한 오류를 수정하기 위하여 수정 언어 지식인 변형 규칙을 사용한 방법이다[9, 15]. 이 방법은 언어 지식을 코퍼스로부터 자동 학습할 수 있다는 장점을 갖지만 학습된 언어 지식이 학습 코퍼스에 종속적일 수 있으며 학습 코퍼스가 아닌 다른 코퍼스에서 높은 정확도를 가질 것인지를 보장하기 어렵다. 또한 이 방법은 초기 태거의 향상이 전체 시스템의 성능 향상에 영향을 미치므로 통계 정보만을 사용하여 품사 태깅하는 초기 태거의 성능을 향상시킴으로써 전체 시스템의 정확도를 향상시킬 수 있다.

Zhang의 연구는 정해진 신뢰 구간 사이의 확률값을 갖는 결과만을 통계 정보를 이용하여 품사 태깅하고, 언어 지식을 이용하여 통계 정보로 품사 태깅되지 않은 어절들의 품사를 결정하거나 잘못된 품사를 수정하는 방법이다[5]. 이 방법은 수작업으로 추출한 300여개의 규칙을 사용하여 중국어 품사 태깅에 적용되었고, 외부 평가 결과 98.1%의 정확도를 보였다. Zhang의 방법은 통계 기반 품사 태거가 자주 오류를 발생시키거나 정해진 신뢰 구간에 포함되지 않는 언어 현상에 대해서만 언어 지식을 사용하여 품사 태깅을 수행하므로 많은 규칙을 필요로하지 않는다. 하지만 한국어와 같이 가능한 어절의 유형과 통계 기반 품사 태거의 오류 유형이 매우 다양한 한국어의 경우 규칙으로 처리할 언어 현상을 찾기가 어려울 뿐만 아니라 규칙 획득 작업이 어렵다.

Tapanainen의 연구는 규칙 기반 품사 태깅 시스템인 ENGCG[4]와 통계 기반 품사 태깅 시스템인 Xerox 태거(XT)[2]를 통합한 방법이다. Tapanainen의 통합 방법은 ENGCG 태거와 XT를 독립적으로 품사 태깅을 수행하고, 그 결과를 비교하여 태깅 결과가 다를 경우 규칙 기반 품사 태거인 ENGCG의 결과를 선호하고, ENGCG가 처리하지 못한 단어에 대한 품사는 통계 기반 품사 태거인 XT의 결과를 따르는 것이다. Tapanainen의 방법은 26,711 단어로 구성된 실험 코퍼스에서 평가되었으며, 98.54%의 정확도를 보였다. Tapanainen의 통합 방법은 규칙 기반 품사 태깅 방법의 높은 정확도와 통계 기반 품사 태깅 방법의 처리 범위를 유지할 수 있는 통합 방법이라 할 수 있다. 하지만 이 방법은 규칙 기반 품사 태깅에서 모호성이 해결되지 않은 어절의 품사를 결정할 때 언어 지식과 통계 정보를 독립적으로 사용하므로 언어 지식에 의한 결과를 충분히 고려하지 못한다. 하지만 언어 지식에 의한 품사 태깅 결과를 충분히 반영하여 통계 기반 품사 태깅을 수행한다면 통계 정보만을 이용한 품사 태깅보다 높은 정확도로 모호성이 해결되지 않은 어절의 품사를 결정할 수 있다.

제 3 절 규칙 우선 직렬 품사 태깅

본 논문은 언어 지식과 통계 정보의 보완적 특성을 이용한 규칙 우선 직렬 품사 태깅 방법(rule-prior sequential part-of-speech tagging)을 제안한다. 규칙 우선 직렬 품사 태깅 방법은 높은 정확도를 보이는 언어 지식에 의한 결과를 선호하고, 언어 지식에 의해서 모호성이 해결되지 않은 어절에 대해서는 통계 기반 품사 태깅 결과를 할당하는 방법이다.

규칙 우선 직렬 품사 태깅은 다음과 같은 네가지 방법에 의하여 통계 기반 품사 태깅 방법의 정확도와 규칙 기반 품사 태깅 방법의 처리 범위를 확장시킨다. 첫째, 규칙에 의해서 정확하게 품사 태깅할 수 있는 언어 현상에는 신뢰도가 낮은 통계 정보를 적용하지 않고 높은 정확도를 보이는 언어 지식을 이용하여 품사 태깅한다. 둘째, 언어 지식에 의한 결과를 통계 기반 품사 태깅에 반영하여 통계 기반 품사 태깅의 정확도를 향상시킴으로써 전체 품사 태깅의 정확도를 향상시

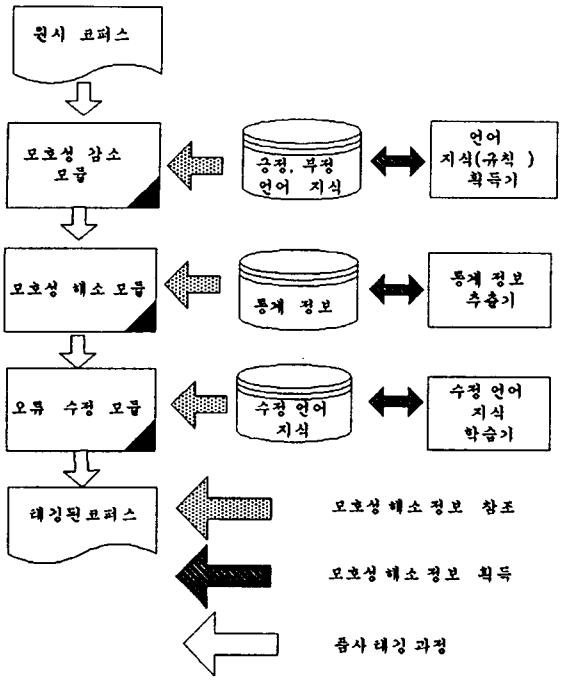

그림 1: 규칙 우선 직렬 품사 태깅 시스템

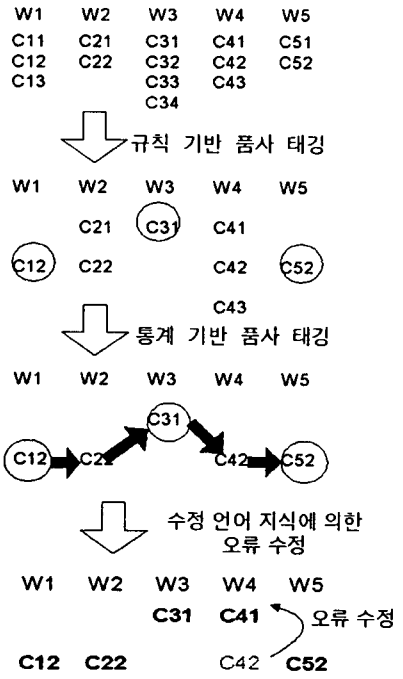

그림 2: Example of Rule-Prior sequential POS Tagging

킨다. 셋째, 언어 지식에 의해서 해결되지 않은 어절들은 통계 정보를 이용하여 품사 태깅함으로써 언어 지식만을 이용한 규칙 기반 품사 태깅 방법의 처리 범위를 향상시킨다. 넷째, 품사 태깅의 오류를 수정할 수 있는 수정 언어 지식을 자동 학습하고, 이를 이용하여 품사 태깅의 정확도를 향상시킨다.

규칙 우선 직렬 품사 태깅 방법은 언어 지식을 통계 정보보다 선호한다는 점에서 Tapanainen의 통합 방법과 유사하나 Tapanainen의 방법과는 다른 다음과 같은 특징을 갖는다. 첫째, 규칙 우선 직렬 품사 태깅 방법은 언어 지식에 의하여 모호성이 해소되었거나 부적절한 품사가 제거됨으로써 모호성이 감소된 모든 결과를 반영하여 통계 기반 품사 태깅을 수행한다. Tapanainen의 통합 방법에서는 언어 지식에 의해서 모호성을 완전히 해결할 수 있는 어절에만 품사 태깅 결과를 할당하고, 그 이외의 모든 어절에는 통계 정보에 의한 품사 태깅 결과를 할당한다. 이 때 언어 지식에 의한 모호성 감소 결과를 전혀 반영하지 않고 통계 기반 품사 태깅을 수행하므로 언어 지식으로 이미 부적절한 품사임이 밝혀진 품사들까지 고려한 품사 태깅을 수행한다. 그로 인하여 부적절한 품사를 포함한 품사 열을 통계 기반 품사 태깅 결과로 생성하고, 그 결과를 모호성이 해결되지 않은 어절의 품사로 할당하는 오류를 범할 수 있다. 하지만 규칙 우선 직렬 품사 태깅 방법은 통계 정보를 이용한 품사 태깅 시 언어 지식에 의해서 모호성이 감소된 결과를 반영함으로써 부적절한 품사를 포함한 품사 열이 품사 태깅 결과로 선택되는 오류를 최소화한다. 둘째, 통합에 사용되는 통계 기반 품사 태깅에서 언어 지식을 반영함으로써 통계 기반 품사 태깅에서 고려하기 어려운 원거리 문맥 정보와 어휘간의 관계를 고려할 수 있다.

그림 1은 본 논문이 제안한 규칙 우선 직렬 품사 태깅 시스템 구성도를 나타낸 것이다. **모호성 감소 모듈**은 긍정 언어 지식과 부정 언어 지식[1]을 이용하여 높은 정확도로 모호성을 감소시킨다. 첫째, 언어 지식으로 정확하게 모호성을 해소할 수 있는 어절에 낮은 신뢰도를 갖는 통계 정보를 적용하여 품사 태깅의 오류를 최소화한다. 둘째, 특정 단어의 모호성을 해소할 수 없는 어절에 대해서는 부적절한 품사만을 제거하여 모호성을 감소시킨다. **모호성 해소 모듈**은 모호성 감소 모듈의 결과를 입력으로 통계 정보를 이용하여 모든 어절에 대한 모호성을 해소하고, 각 어절에 품사를 할당한다. **오류 수정 모듈**은 모호성 감소 모듈과 모호성 해소 모듈에 의한 결과를 수정 언어 지식을 이용하여 보정한다.

다음은 규칙 우선 직렬 품사 태깅을 그림 2의 예를 이용하여 설명한다. 입력 문장은 5개의 어절로 이루어진 문장, "$W1$, $W2$, $W3$, $W4$, $W5$"이고, 올바른 품사 태깅 결과는 "$W1\_C12$ $W2\_C22$, $W3\_C31$, $W4\_C41$, $W5\_C52$"라고 가정한다. 이때 Cij는 i번째 어절의 j번째 형태소 결과를 나타낸다. 먼저, 모

[1]긍정 언어 지식은 특정 품사가 올바른 품사임을 나타내는 언어 지식이며 부정 언어 지식은 특정 품사가 부적절한 품사임을 나타내는 언어 지식이다.

호성 감소 모듈인 규칙 기반 품사 태거에 의하여 어절 W1, W3, W5의 품사로 각각 C12, C31, 그리고 C52가 결정되었다. 그리고 W2의 모호성은 그대로 남아 있으며 어절 W4에 대해서는 C43이 제거되었다. 이러한 규칙 기반 품사 태거의 결과를 모호성 해소 모듈인 통계 기반 품사 태거가 입력으로 받아 규칙에 의해 결정된 C12, C31, C52를 포함하는 경로 중 확률값이 최대인 경로로 품사 태깅을 수행한다. 통계 기반 품사 태깅 결과는 "W1_C12, W2_C22, W3_C31, W4_C42, W5_C52"와 같으며, 규칙에 의해서 품사가 결정되지 않은 모든 어절에 품사를 할당하였다. 그 결과, 단어 W4는 잘못된 품사가 할당되었으며, 이와 같은 오류는 수정 언어 지식에 의해서 보정되고, 최종적으로 "W1_C12, W2_C22, W3_C31, W4_C41, W5_C52"와 같은 품사 태깅 결과를 생성한다.

제 4 절 규칙 우선 직렬 품사 태깅 시스템 구현

4.1 어휘 규칙을 이용한 모호성 해소

규칙 우선 직렬 품사 태깅 방법에서 효과적으로 사용하기 위한 언어 지식은 높은 정확도를 요구한다. 따라서 본 논문은 어휘 단위의 문맥을 고려하면서 어휘 단위의 품사 태깅을 수행할 수 있는 어절 단위의 어휘 규칙과 관용어구 어휘 규칙을 모호성 감소 모듈의 언어 지식으로 사용하고자 한다.

어절 단위의 어휘 규칙은 어절의 좌우에 나타난 표층 형태만을 이용하여 어휘적 모호성을 해결하는 규칙으로, 긍정 언어 지식을 표현한 규칙이다[11]. 그림 3은 어절 단위의 어휘 규칙의 형태를 나타낸 것이다.

<P:N> <중심어> <문맥> = <태깅 결과>
<문맥> ::= $\{$앞 어절$\}_0^P$ * $\{$뒤 어절$\}_0^N$
*: 중심어를 나타내는 표시
$0 <= P <= 3,\ 0 <= N <= 3$

그림 3: 어절 단위 어휘 규칙 형태

그림 3에서 <중심어>는 모호성을 갖는 어절이며 <문맥>에서 $\{$앞 어절$\}_0^P$과 $\{$뒤 어절$\}_0^P$은 모호성 해결에 사용되는 중심어 앞에 있는 P개의 어절과 N개의 어절을 의미한다. 따라서 '[1:1] [나는] [하늘을 * 새를] = [날_동사+는_관형형어미]'는 어절 '나는'이 '하늘을'과 '새를' 사이에서 사용된 경우 '날_동사+는_관형형어미'로 품사 태깅하라는 어절 단위 어휘 규칙의 예를 나타낸 것이다.

어절 단위의 어휘 규칙은 주변 어절의 어휘 문맥을 고려하고, 어휘 단위로 적용되는 규칙이므로 정확도가 매우 높으며 규칙 추출 작업이 용이하다. 또한 서로 다른 형태의 형태소가 동일한 품사를 갖음으로써 발생하는 한국어의 이형 동품사 모호성 해소에 효과적으로 적용될 수 있다.

관용어구 어휘 규칙이란 태깅된 코퍼스에서 같은 분석 결과로 특정 횟수 이상 나타나는 어절열 또는 형태소와 어절열의 품사 태깅 결과를 규칙화한 것이다. 즉 관용어구 어휘 규칙은 연어적 속성을 갖는 단어들을 의미하며, 가중치망의 품사 태깅 방법[6]에서 어휘 정보 부가를 통한 성능 향상을 위하여 사용된 다중 단어와 [8]에서 사용한 묶인말과 유사한 의미를 갖는다. 관용어구 어휘 규칙은 품사 태깅된 코퍼스로부터 자동 학습하였다. '(아닐∧수 ∧없, 아니_형용사+ㄹ_관형형어미∧수_의존명사∧없_형용사)'는 어절 '아닐'과 '수', 형태소 '없'이 연속하여 나타날 경우의 품사 태깅을 위한 관용어구 어휘 규칙을 나타낸 것이다.

어절 단위 어휘 규칙은 20만 원시 코퍼스와 규칙 획득 도구[11]를 이용하여 문법 전문가[2]에 의해서 추출되었고, 추출된 어절 단위 어휘 규칙은 총 8,002개였다. 관용어구 어휘 규칙은 20만 어절 크기의 태깅된 코퍼스로 부터 학습하였고, 학습된 관용어구 어휘 규칙의 수는 1,346개였다.

4.2 언어 지식을 반영한 통계 기반 품사 태깅

일반적으로 통계 기반 품사 태거는 형태소 분석 결과를 입력으로 받아 가능한 모든 품사 경로 중 최대의 확률값을 갖는 품사 열을 선택한다. 하지만 규칙 우선 직렬 품사 태깅을 위한 통계 기반 품사 태깅에서는 언어 지식에 의해서 모호성 감소 모듈에서 모호성이 감소된 결과를 입력으로 통계 기반 품사 태깅을 수행한다. 따라서 어떤 통계 기반 품사 태깅 방법을 사용하든지 쉽게 규칙 우선 직렬 품사 태깅에 사용될 수 있다.

한국어의 어절은 하나 이상의 형태소로 이루어져 있고, 한국어 품사 태깅은 어절을 구성하는 각 형태소와 그에 대응하는 품사를 결정하는 작업이다. 또한 규칙 우선 직렬 품사 태깅에서는 언어 지식으로 처리되지 못한 어절을 통계 정보로 처리하므로 통계 기반 품사 태거도 높은 정확도를 갖는 것이 바람직하다. 본 논문은 규칙 우선 직렬 품사 태깅을 위한 통계 기반 품사 태거로 Twoply HMM을 사용하고자 한다. Twoply HMM은 어절 단위 한국어 품사 태깅 모델과 형태소 단위 품사 태깅 모델과의 장점을 취합하고 단점을 보완하기 위하여 제안된 모델로 어절 단위 문맥을 고려하면서 형태소 단위의 품사 태깅을 수행한다[7].

4.3 수정 언어 지식을 이용한 오류 보정

본 논문은 규칙 우선 직렬 품사 태깅을 위한 수정 언어 지식으로 [15]에서 제안된 어절 태그 변형 규칙을 사용하고자 한

[2]고려대학교 부설 연구소인 민족 문화 연구소에 근무하는 국문학과 대학원생들

표 1: 실험 코퍼스의 통계

코퍼스 종류	어절수	모호성	모호한 어절수	형태소 분석 정확도
내부평가 코퍼스	25,463개	2.54개/어절	14,909개	98.79%
외부평가 코퍼스	20,893개	2.72개/어절	12,582개	99.04%

표 2: 통계 기반 및 규칙 기반 품사 태깅의 정확도와 태깅률

코퍼스	임의 태거		통계 기반 품사 태깅		어휘 규칙을 이용한 품사 태깅	
종류	정확도	태깅률	정확도	태깅률	정확도	태깅률
내부 평가	61.79%	100%	93.70%	100%	99.13%	63.57%
외부 평가	60.98%	100%	93.68%	100%	99.01%	61.89%

다. 어절 태그 변형 규칙은 첨가어적인 한국어 특성을 고려한 변형 규칙으로 한 어절내에 하나 이상의 형태소 품사 태깅 오류를 갖을 수 있는 한국어에 적합한 수정 언어 지식으로 사용될 수 있다[15].

제 5 절 실험 및 평가

5.1 실험 결과

규칙 우선 직렬 품사 태깅 방법의 실험은 내부 평가 코퍼스와 외부 평가 코퍼스에 대해서 수행되었다. 표 1는 코퍼스의 크기 및 모호성 정도 등 실험 코퍼스의 통계를 보이고 있다.

표 2는 실험 코퍼스에서 임의의 태거, 통계 기반, 규칙 기반 품사 태거의 정확도와 태깅률을 나타낸다. 정확도는 전체 어절 중 올바르게 품사 태깅된 어절의 비율을 나타내며, 태깅률이란 품사 태깅 이후 전체 어절 중 하나의 품사만이 할당된 어절의 비율을 나타낸다.

표 2에서 '임의 태거'는 형태소 분석 결과 중 임의의 하나를 품사 태깅 결과로 할당한 결과를 의미한다. 실험 결과, 내부 평가 코퍼스와 외부 코퍼스 모두에 대해서 어휘 규칙을 이용한 규칙 기반 품사 태깅의 정확도가 가장 높았고, 통계 기반 품사 태거인 Twoply의 태깅률이 가장 높음을 알 수 있었다. 이는 규칙 기반 품사 태깅과 통계 기반 품사 태깅의 정확도와 처리 범위 향상을 위한 보완적 특성을 보이고 있다.

표 3은 언어 지식과 통계 정보를 이용한 규칙 우선 직렬 품사 태깅 결과를 보이고 있다. 표 3에서 정확도 옆에 괄호 안의 수치는 통계 기반 품사 태깅과 비교한 정확도 향상률을 나타낸다. 실험 결과, 어휘 규칙으로 표현된 긍정, 부정 언어 지식과 통계 정보만을 이용한 품사 태깅 결과도 통계 기반 품사 태깅과 비교하여 내부 코퍼스와 외부 코퍼스에 대하여 평균 28.28%의 정확도를 향상시킬 수 있음을 알 수 있었다. 그러나 수정 언어 지식까지 이용한 품사 태깅 방법은 긍정, 부정 언어 지식 그리고 통계 정보를 이용한 품사 태깅의 오류를 보정함으로써 평균 37.70%의 정확도 향상을 보였다. 최종적으로 어휘 규칙, Twoply HMM 그리고 어절 태그 변형 규칙으로 구현된 규칙 우선 직렬 품사 태깅 방법은 내부 코퍼스에서 96.39%, 외부 코퍼스에서 95.43%의 정확도를 보였다.

5.2 비교 평가

본 논문은 변형 규칙을 이용한 방법, 그리고 Tapanainen의 통합 방법을 본 논문에서 사용한 품사 집합을 이용하여 구현하고, 이를 비교하였다. 또한 기존에 개발된 한국어 품사 태깅 방법과 품사 집합, 품사 태깅의 단위, 품사 태깅에 사용한 모호성 해소 정도 등을 기준으로 비교 평가하였다. Tapanainen의 방법은 본 논문에서 사용한 어휘 규칙과 Twoply HMM을 이용하여 구현하였다. 통계 정보와 변형 규칙을 이용한 TAKTAG와의 비교를 위해서 초기 태거로는 Twoply HMM을 사용하였고[3], 어절 태그 변형 규칙을 이용하였다.

표 4는 각 방법과 규칙 우선 직렬 품사 태깅 방법과의 정확도를 비교한 결과를 보이고 있다. 표 4에서 볼 수 있듯이 Twoply HMM과 변형 규칙을 통합한 품사 태깅의 정확도가 Tapanainen의 통합 방법보다 높은 정확도를 보였고, 규칙 우선 직렬 품사 태깅 방법보다는 낮은 정확도를 보였다. 이러한 실험 결과는 긍정, 부정 언어 지식을 이용한 모호성 감소, 언어 지식을 반영한 통계 기반 품사 태깅, 그리고 수정 언어 지식을 이용한 규칙 우선 직렬 품사 태깅 방법이 가장 높은 정확도를 갖음을 보이는 것이다.

표 5는 기존에 개발된 한국어 품사 태깅 방법과 본 논문에서 제안한 규칙 우선 직렬 품사 태깅 방법과의 비교 평가 결과를 보이고 있다. 표 5에서 '정보'열은 모호성 해소를 위해 사용된 정보를 나타내며, 코퍼스 크기와 정확도에서 (어)는

[3]본 논문에서 Twoply HMM을 초기 태거로 사용한 이유는 지도 학습을 이용한 Twoply HMM이 [9]에서 초기 태거로 사용한 자율 학습을 이용한 HMM보다 높은 정확도를 갖기 때문이었다.

표 3: 규칙 우선 직렬 품사 태깅 결과

코퍼스 종류	통계 정보만을 이용한 품사 태깅	어휘 규칙과 통계 정보를 이용한 품사 태깅	통계 정보 + 어휘 규칙 수정 언어 지식 (규칙 우선 직렬 품사 태깅)
내부 평가 코퍼스	93.70%	95.62%(30.48%)	96.39%(42.70%)
외부 평가 코퍼스	93.21%	94.98%(26.07%)	95.43%(32.70%)
평균	93.46%	95.30%(28.28%)	95.91%(37.70%)

표 4: 제안된 방법과 기존의 통합 방법과의 정확도 비교

코퍼스 종류	Twoply HMM	통계 정보 + 변형 규칙	Tapanainen의 방법	규칙 우선 직렬 품사 태깅 방법
내부 평가 코퍼스	93.70%	95.37%	94.85%	95.62%
외부 평가 코퍼스	93.21%	94.32%	94.21%	94.98%

어절 단위로 계산된 값을, (형)은 형태소 단위로 계산된 값을 나타낸다. 이상호 시스템[10], 이하규 시스템[12], 김진동[7]은 HMM에 기반한 통계 기반 품사 태깅 시스템이다. 김재훈 시스템[6]은 한국어 품사 태깅에서 다입력 열을 자연스럽게 표현할 수 있는 가중치 망을 이용한 품사 태깅 방법이다.

제 6 절 결론

본 논문은 규칙 기반 품사 태깅 시스템의 정확도와 통계 기반 품사 태깅 시스템의 확장성을 향상시킬 수 있는 규칙 우선 직렬 품사 태깅 방법을 제안하였다. 규칙 우선 직렬 품사 태깅 방법은 높은 정확도를 보이는 언어 지식을 선호하고, 언어 지식에 의한 품사 태깅 결과를 통계 정보를 이용한 품사 태깅에 반영함으로써 통계 기반 품사 태깅 의 정확도와 규칙 기반 품사 태깅의 처리 범위를 향상시킨다. 제안된 방법은 2만 어절 크기의 내부 평가 코퍼스와 외부 평가 코퍼스에 대해서 수행되었으며, 각각 95.62%, 94.98%의 정확도를 보였다. 이는 통계 정보만을 이용한 품사 태깅의 정확도를 내부 평가 코퍼스에서 42.70%, 외부 평가 코퍼스에서 32.70%의 정확도 향상을 보인 결과이다.

규칙 우선 직렬 품사 태깅의 성능 향상을 위해서는 언어 지식을 이용한 모호성 감소 모듈과 언어 지식을 반영한 모호성 해소 모듈, 수정 언어 지식을 이용한 오류 보정 모듈의 성능 향상이 필요하다. 따라서 향후에는 높은 정확도를 갖는 언어 지식의 획득 방안과 미등록어 처리 문제, 어휘 간의 정보 모델링 등 규칙 기반 및 통계 기반 품사 태깅의 성능 향상을 위한 연구가 계속되어야 할 것이다.

본 논문은 규칙 우선 직렬 품사 태깅 방법을 한국어 품사 태깅에 적용하여 평가하였는데, 추후에는 영어의 품사 태깅을 위해서 규칙 우선 직렬 품사 태깅 방법을 적용하고자 한다. 또한 어휘적 모호성 해소뿐만 아니라 구문적 모호성, 의미적 모호성 해소를 위한 연구에도 본 논문이 제안한 규칙 우선 직렬 품사 태깅의 모호성 해소 방법론을 적용하고자 한다.

참고 서적

[1] E. Brill, "Transformation-based error-driven learning and natural language processing:A case study in part-of-speech tagging," *Computational Linguistics*, Vol.21, No.4, pp.543-565, 1995.

[2] D. Cutting, J. Kupiec, j. Pedersen, P. Sibun, "A Practical Part-of-Speech Tagger," *Proc. of 3rd Conference on Applied NLP*, pp.133-140, 1992.

[3] P. Tapanainen, A. Voutilainen, "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, 1994.

[4] A. Voutilainen, "A syntax-based part-of-speech analyzer." *Proc. of the 7th conference of the European chapter of the ACL*. pp.157-164, 1995.

[5] M. Zhang, S. Li, T. Zhao, "Tagging Chinese Corpus Based on Statistical and Rule Techniques," *Proc. of the Int. Conference on Computer Processing of Oriental Language(ICCPOL-97)*, pp.503-506, 1997.

[6] 김재훈, *오류-보정 기법을 이용한 어휘 모호성 해소*, 한국과학기술원 전산학과, 박사학위 논문, 1996.

표 5: 기존의 한국어 품사 태깅 시스템과의 비교

	품사 집합	정보	실험 코퍼스	정확도
이상호	51개	통계 정보	4,729(어)	93.7%(어)
김재훈	52개	통계 정보	88,683(형)	98.0%(형)
이하규	13개	통계 정보	5만(어)	98.8%(어)
김진동	59개	통계 정보	17,644(어)	93.6%(어)
신상현	25개	통계 정보+변형규칙	11,872(형)	91.5%(형)
임희석 (제안된 방법)	50개	통계 정보 + 긍정, 부정 언어 지식, 수정 언어 지식	20,893(어)	95.43%(어)

[7] 김진동, 임희석, 임해창, "어절 단위의 문맥을 고려한 형태소 단위의 한국어 품사 태깅 모델," *인지과학회 춘계 학술대회발표 논문집*, pp.97-106, 1996.

[8] 박혜준, 윤준태, 송만석, "말뭉치 품사꼬리달기 시스템 구현," *한국정보과학회 봄 학술발표논문집*, pp.829-832, 1994.

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

[10] 이상호, *미등록어를 고려한 한국어 품사 태깅 시스템 구현*, 한국과학기술원 전산과학과, 석사학위 논문, 1995.

[11] 이정규, *수작업을 최소화한 어절 규칙 기반 한국어 품사 태깅*, 고려대학교 컴퓨터학과, 석사학위 논문, 1997.

[12] 이하규, "어말-어두 공기 정보를 이용한 한국어 어휘 중의성 해소," *한국정보과학회 논문지(B)*, 제 24권, 제 1호, pp.82-89, 1997.

[13] 임철수, *HMM을 이용한 한국어 품사 태깅 시스템 구현*, 한국과학기술원 전산학과 석사학위 논문, 1994.

[14] 임해창, 임희석, 윤보현, "자연어처리를 위한 품사 태깅 시스템의 고찰," *한국정보과학회지*, 제 14권, 제 7호, pp.36-57. 1996.

[15] 임희석, 김진동, 임해창, "어절 태그 변형 규칙을 이용한 한국어 품사 태거", *한국 정보과학회 논문지(B)*, 제 24권, 제 6호, pp.673-684, 1997.