韓國 컴퓨터情報學會 論文誌
第13卷 第7號, 2008. 12.

2008-13-7-1-2

형태소 분석 및 품사 부착을 위한 말뭉치 기반 혼합 모형

이 승 욱*, 이 도 길**, 임 해 창***

A Corpus-based Hybrid Model for Morphological Analysis and Part-of-Speech Tagging

Seung-Wook Lee*, Do-Gil Lee**, Hae-Chang Rim***

요 약

한국어 형태소 분석은 일반적으로 입력된 문장의 분석 후보를 다수 생성한 후, 그 중 최적의 후보를 선택하는 과정을 거친다. 분석 후보를 많이 생성할수록 올바른 분석이 포함될 가능성이 높아지지만 동시에 모호성이 증가한다는 문제가 생긴다. 이를 해결하기 위해 본 논문은 단일 후보를 생성하는 규칙 기반 분석 모형을 제안한다. 분석 규칙은 품사 부착 말뭉치를 통해 자동으로 추출되기 때문에 규칙 구축비용을 필요로 하지 않을 뿐만 아니라 높은 분석성공률을 보인다. 분석이 성공한 경우에는 단 하나의 분석 후보만을 생성하기 때문에 최적 후보 선택 단계에서의 모호성이 효과적으로 감소되고, 계산 복잡도 역시 줄어든다. 규칙 모형으로 분석이 실패한 경우를 대비해 기존 확률 기반 모형을 결합함으로써 형태소 분석 성능을 향상시킬 수 있다.

Abstract

Korean morphological analyzer generally generates multiple candidates, and then selects the most likely one among multiple candidates. As the number of candidates increases, the chance that the correctly analyzed candidate is included in the candidate list also grows. This process, however, increases ambiguity and then deteriorates the performance. In this paper, we propose a new rule-based model that produces one best analysis. The analysis rules are automatically extracted from large amount of Part-of-Speech tagged corpus, and the proposed model does not require any manual construction cost of analysis rules, and has shown high success rate of analysis. Futhermore, the proposed model can reduce the ambiguities and computational complexities in the candidate selection phase because the model produces one analysis when it can successfully analyze the given word. By combining the conventional probability-based model, the model can also improve the performance of analysis when it does not produce a successful analysis.

▸ Keyword : 형태소 분석(Morphological Analysis), 품사 부착(Part-of-Speech Tagging), 혼합 모형(Hybrid Model)

• 제1저자 : 이승욱
• 접수일 : 2008. 9. 29, 심사일 : 2008. 11. 3, 심사완료일 : 2008. 12. 24.
* 고려대학교 컴퓨터·전파통신공학과 박사과정 ** 고려대학교 민족문화연구원 HK연구교수
*** 고려대학교 컴퓨터·전파통신공학과 교수
※ 이 연구에 참여한 연구자(의 일부)는 '2단계 BK21사업'의 지원비를 받았음.

I. 서 론

한국어 형태소 분석은 문장을 구성하는 각 어절에 대해 이를 구성하고 있는 형태소를 파악해내는 작업이다. 형태소 분석 결과는 단어 중의성 해결(word sense disambiguation), 구문 분석(syntactic parsing), 의미역 부착(semantic role labeling), 기계 번역(machine translation) 등과 같은 다양한 자연어처리 분야에서 중요한 역할을 하며, 부정확한 분석 결과는 후행하는 모듈에게 치명적인 영향을 미칠 수 있기 때문에 높은 정확도를 요한다. 하지만 특정 어절을 형성할 수 있는 형태소 조합이 하나 이상인 경우가 존재하기 때문에 형태론적 모호성이 발생한다. 이로 인해 형태소 분석의 오류가 발생한다. 더욱이 여러 형태소들이 결합하여 어절을 형성할 때 발생하는 활용과 축약과 같은 현상[2]은 형태소 분석을 보다 어렵게 한다.

이런 어려움을 해결하기 위해 일반적으로 다수의 분석 후보를 생성하는 형태소 분석 단계를 거친 후, 이 가운데 가장 적절한 후보를 선택하는 품사 부착 단계로 나눠져 분석이 수행된다. 형태소 분석 단계에서는 입력문의 각 어절에 대해 올바른 분석 결과를 포함하는 복수개의 분석 후보를 생성하는 것을 목표로 한다. 이 과정에서 후보를 과다하게 생성하면 최종 분석 결과의 정확도 하락의 원인이 된다. 따라서 최소의 분석 후보를 생성하되 올바른 분석 후보가 포함되게 생성하여야 한다. 품사 부착 단계는 형태소 분석 후보 생성 단계에서 생성한 다수의 분석 후보 가운데 가장 적절한 하나의 후보를 선택한다. 분석 후보를 생성할 때와는 달리 품사 부착 모형은 좌우문맥과 같은 추가적인 정보를 활용하여 보다 정확히 형태소 분석 후보를 선택한다.

형태소 분석에 있어서 분석 속도 역시 중요한 요소이다. 형태소 분석 과정은 주로 응용 분야의 전처리 단계로써 활용되기 때문에 빠른 속도의 분석을 보장할 수 있어야 한다. 예를 들어 정보 검색 엔진은 대량의 문서에서 색인어를 추출하기 위해 형태소 분석 결과를 활용하기도 하는데, 이 경우 형태소 분석 속도가 색인의 병목현상을 야기할 수 있다.

본 논문에서는 형태소 분석 후보 생성 단계에서 발생하는 과도한 분석 후보 생성을 방지함과 동시에 정확한 하나의 분석 후보만을 생성하는 단일 분석 후보 생성 모형을 제안한다. 하나의 후보만 생성하기 때문에 최적 후보 선택 단계에서 선택의 모호성이 해소되어 분석 속도를 향상시켰을 뿐만 아니라 보다 높은 분석정확률을 보였다.

2) [1]에서 더 자세히 다루고 있다.

II. 관련 연구

형태소 분석은 한국어 자연어처리의 오랜 연구 분야 중 하나이다. 최근까지 다양한 연구가 진행되어 왔으며 이들 대부분은 규칙 기반 접근법, 통계 기반 접근법, 혼합 접근법으로 분류 가능하다.

2.1 규칙 기반 접근법

규칙 기반 접근법은 결정적 규칙을 이용하여 입력문을 분석한다. 규칙은 다양한 방법으로 언어 지식을 반영할 수 있게 정의된다. [2]는 어휘의 패턴 정보를 규칙으로 정의하였고, [3]은 품사 정보의 패턴으로 규칙을 정의하였다. [4]는 다른 품사 부착 모형의 결과가 가지는 오류를 자동으로 수정할 수 있는 변형 기반 학습(transformation based learning)을 제안하였다.

이런 규칙 기반 방법들의 문제점은 정확한 분석 규칙을 충분히 구축해야 고성능의 분석 결과를 얻을 수 있다는 점이며, 일반적으로 고품질의 분석 규칙을 수집하는 데는 많은 비용이 요구된다. 또한 구축된 규칙 간 충돌이 발생하여 상이한 분석 결과를 도출하는 경우, 이들 간의 우선순위를 정하는 기준이 필요하다는 문제점도 있다.

2.2 통계 기반 접근법

통계 기반 모형은 확률이나 통계를 활용하여 입력문을 분석한다. 확률은 대량의 학습 말뭉치를 활용하여 추정이 된다. N-gram 모형이나 은닉마르코프 모형(hidden Markov model)은 가장 많이 쓰이는 통계 모형이다. 한국어 형태소 분석을 위한 N-gram 모형으로 [5], [6]은 특정한 수의 이전 어절을 바탕으로 그 다음 어절의 형태소를 분석하는 방법을 제안하였다. N-gram 모형에서 unigram, bigram 단위로는 원거리 정보를 고려하지 못하여 부정확하게 분석하는 경우가 발생하지만, 이를 해결하기 위해 문맥을 늘리게 되면 자료부족문제가 심화된다는 문제점이 있다. 또한 어절 단위 확률 추정에 있어서 한국어 어절의 높은 생산성으로 인한 자료부족문제로 인해 어려움을 겪는다. 자료부족문제는 한정된 양의 말뭉치를 통해 확률을 추정하기 때문에 유발되며, 부정확한 확률 추정은 최종 분석 정확도의 하락의 원인이 된다. 또한 규칙 기반 방식에 비해 부정확한 결과를 보완하기 힘들다는 단점이 있다. 규칙 기반 방법에서는 분석 결과가 부정확한 경우에 대해 해당되

는 분석 규칙을 추가 및 수정함으로써 보완 가능하지만 통계 기반 모형에서는 이와 같은 방법을 찾기 힘들다.

2.3 혼합 접근법

혼합 접근법은 규칙 기반 접근법과 통계 기반 접근법의 장점을 취합한 방법이다. 일반적인 기존 연구는 규칙 기반 모형이 확률 기반 모형의 후처리기로 이용하는 방법이다. [7]이 제안한 방법에서 규칙 기반 모형은 확률 모형이 분석한 결과에 대해 오류를 탐지하여 교정하는 역할을 한다. 교정 규칙은 확률 모형의 분석 결과와 올바른 분석 결과를 비교를 통하여 수동으로 구축되었다. [8]은 이와 유사하지만 오류를 수정하는 모형 역시 확률 모형을 이용하였다는 차이가 있다. 규칙 기반 모형을 후처리기로 사용한 혼합 접근법은 규칙 기반 모형이 확률 기반 모형의 단점을 파악하여 보완하기 위해 고안되었기 때문에 성능 향상을 가져왔지만, 특정 확률 기반 모형 결과의 특성에 종속되기 때문에 다른 종류의 형태소 분석 모형에서도 동일한 성능 향상을 보일 것을 보장하긴 힘들다.

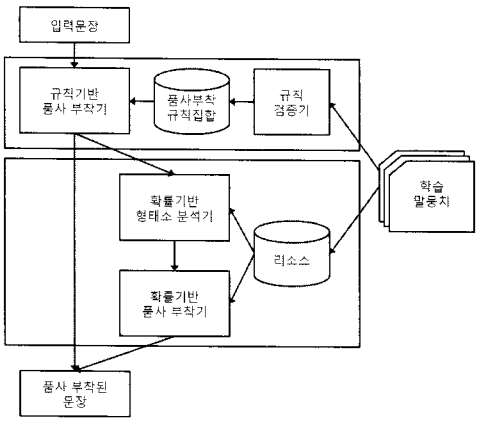

그림 3.1. 제안하는 혼합 모형의 구조도
Fig 3.1. Architecture of the Proposed Hybrid Model

반면, 규칙 기반 모형을 전처리에 활용한 혼합 접근법도 있었다. [9]는 규칙 기반 모형과 통계 기반 모형이 동시에 활용되는 혼합 형태소 분석 모형을 제안하였다. 규칙 기반 분석 모형은 분석 후보 생성 시 발생하는 비문법적인 후보를 제거하고 확률 모형이 잘못 분석한 경우를 탐지해 수정해 주는 역할을 한다. 규칙 기반 모형은 분석정확도를 향상시켰지만, 이를 위해 필요한 규칙들은 수작업을 통하여 획득되었기 때문에 많은 비용을 요구하며, 재구현이 힘들다는 단점이 있다. [10]이 제안한 혼합 모형에서의 규칙 모형은 보조용언, 숙어, 관용적 표현에 한하여 적용가능하다는 한계가 있다.

혼합 접근법에서의 규칙 기반 모형은 정확한 분석을 통해 성능 향상에 도움을 주었지만, 규칙 구축에 많은 노력을 들여야 한다는 공통점이 있다. 따라서 규칙 기반의 정확한 분석 규칙을 자동으로 구축 및 활용할 수 있는 방법이 요구된다.

III. 제안하는 모형

본 논문은 규칙 기반과 통계 기반 접근법의 장점을 취합한 혼합 모형을 제안한다. <그림 3.1>은 제안하는 혼합 모형의 구조도를 나타낸다. 규칙 기반 모형은 분석 규칙을 이용하여 하나의 어절에 대해 단 하나의 정확한 분석 후보만을 생성한다. 과도한 분석 후보를 발생하지 않고 단 하나의 분석 후보만 생성하기 때문에 품사 부착 단계에서 최적 후보 선택의 모호성이 발생하지 않는다는 장점이 있다. 분석에 사용되는 규칙들은 품사 부착 말뭉치를 통해 자동으로 추출되고 검증 단계를 거친다. 규칙 검증 단계에서 분석정확률을 높이기 위해 부정확한 분석 규칙들은 제거되는데, 이에 따라 규칙의 부재로 인한 분석에 실패하는 경우 역시 증가하게 된다. 이런 문제점을 보완하기 위해 기존의 확률 기반 모형을 활용하여 규칙 기반 모형이 분석 실패한 어절에 대해 기존의 방식대로 다수의 분석 후보를 생성한다. 확률 기반 형태소 분석 모형 및 품사 부착 모형은 한국어에 높은 성능을 보였던 [11]을 활용하였으며, 다른 어떠한 종류의 분석 모형이라도 대체 가능하다. 제안하는 혼합 모형은 규칙 기반 모형의 높은 분석정확률과 확률 기반 모형의 높은 분석성공률을 동시에 가진다. 아래는 제안하는 시스템이 입력문을 분석하는 과정을 기술하는 알고리듬이다.

입력문의 각 어절에 대해 규칙 기반의 형태소 분석을 수행
if 검색된 분석 규칙이 있다면 분석 규칙을 이용하여 분석 후보를 생성
else 확률 기반 형태소 분석을 수행하여 다수의 분석 후보를 생성
문장 전체의 분석 후보에 대해 확률 기반 품사부착 모형이 최적의 후보들을 선택

3.1 규칙 기반 형태소 후보 생성 모형

3.1.1 규칙 추출

제안하는 규칙 기반 모형은 어절 단위 규칙과 형태소 단위 규칙을 활용한다. 두 종류의 규칙 모두 품사 부착된 말뭉치에서 자동으로 추출된다. 어절 단위 규칙과 형태소 단위 규칙은

표 3.1. 어절 단위 규칙 추출에 활용된 품사 부착 말뭉치 예문
Table 3.1. An Example of Part-of-Speech Tagged Sentence for Eojeol-unit Rule Extraction

원시어절	분석 결과
어제	어제/NNG
친구를	친구/NNG+를/JKO
만났다.	만나/VV+았/EP+다/EF+./SF

표 3.2. 추출된 어절 단위 규칙 예제
Table 3.2. An Example of the Extracted Eojeol-unit Rule

좌문맥	분석 대상	우문맥	분석 결과
-	어제	-	어제/NNG
-	친구를	-	친구/NNG+를/JKO
-	만났다.	-	만나/VV+았/EP+다/EF+./SF
BOS	어제	친	어제/NNG
제	친구를	만	친구/NNG+를/JKO
를	만났다.	BOS	만나/VV+았/EP+다/EF+./SF
BOS	어제	친구	어제/NNG
어제	친구를	만났	친구/NNG+를/JKO
구를	만났다.	BOS	만나/VV+았/EP+다/EF+./SF

〈좌문맥, 분석 대상, 우문맥, 분석 결과〉로 구성된다. 좌문맥과 우문맥은 분석 대상의 좌우에 인접한 가변적인 길이의 음절열이다. 본 논문에서는 가변적인 길이의 범위를 0부터 2까지 정하였다. 분석 대상은 어절 단위 규칙에서는 어절이 해당되며 형태소 단위 규칙에서는 어절의 부분음절열이 해당된다.

· 어절 단위 규칙

학습에 이용되는 품사 부착 말뭉치는 〈표 3.1〉과 같이 원시 어절의 표층형과 그에 해당하는 형태소 분석 결과로 구성된다. 각각의 어절 단위 규칙은 좌우문맥의 가변적인 길이의 조합에 따라 추출된다. 이렇게 추출된 규칙은 구조가 단순하여 효율적인 분석이 가능할 뿐만 아니라 분석정확률까지 높다. 〈표 3.2〉는 다양한 문맥의 길이에 따라 추출된 어절 단위 규칙의 예를 보여준다.

한국어 어절의 높은 생산성으로 인해 어절 단위 규칙이 분석하지 못하는 경우가 발생할 수 있다. 이를 보완하기 위해 어절 보다 작은 단위인 형태소 단위 규칙을 이용한다.

· 형태소 단위 규칙

한국어의 어절은 하나 이상의 형태소들의 조합으로 이루어지기 때문에 어절을 분석하기 위해서 어절을 형성하는 형태소 단위로 분석을 시도하며, 제안하는 형태소 단위 규칙의 분석 대상은 〈표 3.3〉, 〈표 3.4〉에 나타난 것과 같이 어절의 부분음절열이 된다. 기존의 형태소 단위 분석 방법론들은 형태소들이 결합되면서 발생한 음운 현상을 복원한 후 분석을 시도하지만, 제안하는 형태소 단위 규칙은 복원 과정 없이 바로 분석 가능하다는 장점도 있다. 형태소 단위 규칙도 어절 단위 규칙과 마찬가지로 다양한 가변길이 문맥의 조합에 대해 각각 규칙을 추출한다.

3.1.2 규칙 검증

제안하는 규칙 기반 모형은 단일 후보를 생성하기 때문에 높은 정확률을 보장할 수 있어야 한다. 분석 시 모호성이 높은 어절은 규칙 기반 모형으로 분석하지 않고, 다수의 분석 후보를 생성하는 확률 기반 모형이 분석하게 구성하였다. 이를 위해 규칙 검증 단계에서 모호성이 높은 어절을 분석하는 규칙들은 제거하였다. 또한, 추출된 분석 규칙들은 품사 부착 말뭉치에서 자동으로 추출되었기 때문에 부정확한 분석 규칙들을 포함하고 있다. 특히 좌우문맥 정보를 포함하고 있지 않은 규칙들 중에는 부정확하고 신뢰하기 어려운 것들이 다수 포함되어 있다. 이런 규칙들을 그대로 이용하게 되면 부정확한 분석 결과로 이어지기 때문에 규칙을 검증 단계에서 제거한다.

본 논문에서 규칙 검증은 추출된 규칙의 분석정확률을 측정하여 오류를 포함하는 규칙들은 제거하는 작업으로 정의한다. 규칙의 분석정확률은 해당 규칙을 학습말뭉치에 다시 적용하였을 때의 정확률로 측정하였으며, 이 과정에서 제거되지 않고 남은 규칙들은 분석 단계에서 이용된다. 규칙 검증의 한 예로, 들어 규칙 〈"-", "나는", "-", "나는/VV+는/JX"〉의 분석정확률은 100%가 되지 않아서 제거되었지만, 규칙 〈"소리", "나는", "쪽엔", "나/VV+는/ETM"〉의 정확률은 100%

표 3.3. 형태소 단위 규칙 추출에 활용된 품사 부착 말뭉치 예문
Table 3.3. An Example of Part-of-Speech Tagged Sentence for Morphemel-unit Rule Extraction

원시어절	분석 결과
집과	집/NNG+과/JKB
가깝다해서	가깝/VA+다/EC+하/VX+아서/EC
꼭	꼭/MAG

표 3.4. 추출된 형태소 단위 규칙 예제
Table 3.4. An Example of the Extracted Morpheme-unit Rule

좌문맥	분석 대상	우문맥	분석 결과
-	집	-	집/NNG
-	과	-	과/JKB
-	가깝다	-	가깝/VA+다/EC
-	해서	-	하/VX+아서/EC
BOS	집	과	집/NNG
집	과	_가	과/JKB
과_	가깝다	해	가깝/VA+다/EC
다	해서	_꼭	하/VX+아서/EC
BOS	집	과_가	집/NNG
BOS 집	과	_가깝	과/JKB
집과_	가깝다	해서	가깝/VA+다/EC
깝다	해서	_꼭	하/VX+아서/EC

이므로 그대로 활용한다.

3.1.3 규칙 적용

제안하는 규칙 기반 형태소 분석 모형은 주어진 문장의 각 어절에 대해 해당하는 어절 단위 분석 규칙을 탐색하여 그에 해당하는 분석 결과를 할당함으로써 분석을 완료한다. 해당하는 분석 규칙이 다수개가 탐색된다면 문맥의 길이가 긴 순서대로 적용한다. 어절 단위 분석 규칙이 존재 하지 않는다면 보다 작은 단위인 형태소 단위 분석 규칙을 탐색한다. 어절 단위 분석과는 달리 부분음절열을 단위로 분석하여 그 결과를 취합하는 방식이기 때문에 여러 분석 알고리듬을 고려해 볼 수 있다. 본 연구에서는 아래와 같은 분석 전략을 취하였다.

1. 어절의 왼쪽부터 부분음절열의 크기를 증가하며 해당하는 형태소 단위 규칙을 탐색한다. 규칙이 탐색 되지 않았다면 그 시점에서 증가를 중지한다.
2. 어절의 오른쪽부터 부분음절열의 크기를 증가하며 해당하는 형태소 단위 규칙을 탐색한다. 규칙이 탐색 되지 않았다면 그 시점에서 증가를 중지한다.
3. 양방향의 부분음절열의 분석 결과를 조합하여 어절을 생성할 수 있는 모든 분석 후보를 생성한다. 두 가지 이상의 분석 후보가 생성된다면 분석 규칙이 충돌한 경우이며, 실패로 간주한다.

실제 분석 결과의 한 예로, "그래서 나는 오늘"이라는 문맥에서 어절 "나는"에 대해 기존의 확률 기반 형태소 분석 모형은 "나/NP+는/JX", "나/VV+는/ETM", "나/NNG+는/JX", "나/VX+는/ETM" 등의 여러 분석 후보를 생성했지만, 제안하는 규칙 기반 형태소 분석 모형은 단 하나의 분석 결과 "나/NP+는/JX"만을 올바르게 생성한다.

3.2 통계 기반 형태소 분석 모형

규칙 기반 모형이 분석 규칙의 부재로 인해, 혹은 분석 규칙의 충돌로 인해 분석 실패하는 경우 기존의 확률 기반 형태소 분석 모형이 분석 실패한 어절에 대해 다시 분석을 시도한다. 확률 기반 모형뿐만 아니라 다양한 종류의 형태소 분석 모형으로 대체 가능하며, 본 논문에서는 한국어에서 높은 성능을 보였고, 재구현이 가능한 [11]을 그대로 활용하였다. 이 모형은 주어진 어절 w에 대한 모든 분석 후보 R을 생성한다. 분석 후보를 생성하는 것만을 목적으로 하는 일반 형태소 분석과는 달리, 이 모형은 각각의 후보에 대해 추가적으로 확률을 추정

하여 순위화하는 것이 특징이다. 확률 추정은 어절 단위로 먼저 추정 되며, 실패할 경우 형태소 단위로 추정이 된다.

· 어절 단위 확률 모형

어절 단위 확률 모형에서 주어진 어절에 대해 해당하는 분석 결과의 확률 추정 시 아래 〈수식 3.1〉과 같이 학습 말뭉치로부터 최대우도추정법(maximum likelihood estimation)을 이용한다.

$$P(R/w) = \frac{freq(R,w)}{freq(w)} \quad \cdots\cdots (3.1)$$

R은 형태소 분석 결과, w는 어절을 나타낸다. 이 모형은 제안하는 어절 기반 규칙 모형과 유사한 점을 보인다. 두 모형 모두 단순한 구조이고 빠른 분석을 보장하지만 확률 모형은 규칙 모형과는 달리 하나 이상의 후보들을 생성한다는 점에서 차이가 있다.

· 형태소 단위 확률 모형

형태소 단위 확률 모형은 형태소들이 결합할 때 발생하는 음운 현상 고려하기 위하여 가능한 음운현상 복원 후보를 모두 생성한 후, 각각의 복원 후보에 형태소 분석을 수행하여 그 결과를 취합한다. 형태소 분석은 은닉마르코프 기반 모형과 CYK 알고리즘을 이용하여 각 분석 후보를 생성하고, 그 확률을 같이 추정한다.

3.3 확률 기반 품사 부착 모형

확률 기반 형태소 분석 모형이 생성한 다수의 분석 후보에 대해 확률 기반 품사 부착 모형은 최적의 분석 후보 하나를 선택하는 역할을 한다. 〈수식 3.2〉와 같이 주어진 어절 W에 대해 가장 높은 확률을 가진 분석 결과 R을 찾는 과정으로 정의된다.

$$\Gamma(W) \equiv \arg max_R P(R/W) \quad \cdots\cdots (3.2)$$

$$\cong \arg max_R \prod_{i=1}^{n} P(r_i|w_i) \frac{P(r_i|r_{i-1})}{P(r_2)} \quad \cdots (3.2)$$

본 논문에서는 은닉마르코프 기반 품사 부착 모형 [11]을 활용하였다. 이 모형은 〈수식 3.3〉에 나타난 바와 같이 마르코프 가정을 바탕으로 하고 있으며, 각 어절마다 분석 후보들의 분석 확률을 추정한 후, 가장 높은 확률을 가지는 후보열을 선택하는 방식이다. 입력문 각각의 어절에 대해 형태소 분석 결과가 생성될 확률과 분석 결과가 다음 분석 결과로 전이할 확률이 추정되고, 이들의 곱이 최고 확률을 가지는 분석 결과열이 최종 분석 결과로 선택된다.

Ⅳ. 실험

4.1 실험 환경

제안하는 모형의 성능을 평가하기 위해 한국어 형태소 분석 평가에 많이 활용되는 21세기 세종계획 품사 부착 말뭉치[3]를 활용하였다. 세종 말뭉치는 뉴스기사, 소설, 잡지 등의 다양한 분야에서 수년에 걸쳐 수집, 구축되었다. 이 말뭉치는 〈표 3.1〉, 〈표 3.3〉의 예에 나타난 바와 같이 원시 말뭉치와 그에 해당하는 형태소 분석 정보가 부착되어 있다. 본 논문에서는 2003년 구축 분량만을 이용하였다. 이 중 90%는 학습에 이용하였고, 나머지 10%는 실험에 활용하였다. 〈표 4.1〉은 실험에 이용된 말뭉치의 통계를 보여준다. 실험의 신뢰성을 높이기 위해 10-fold 교차검증법(10-fold cross validation)을 이용하였다[4].

표 4.1. 품사 부착 말뭉치 통계
Table 4.1. Statistics of the Part-of-Speech Tagged Corpus

전체 어절 수	1,800,000
학습집합의 어절 수	1,620,000
실험집합의 어절 수	180,000
품사집합의 수	41

4.2 평가 척도

제안하는 모형을 분석하기 위해 분석성공률, 평균분석후보수, 정답포함률, 분석정확도 등의 척도를 이용하였다. 분석성공률은 얼마나 많은 어절이 분석 후보를 생성했는지를 나타낸다. 평균분석후보수는 어절이 가지는 분석 후보의 수의 평균으로서 분석 결과의 모호성을 나타낸다. 정답포함률은 얼마나 많은 어절이 분석 후보 중에서 실제로 올바른 분석 결과를 후보로 가지고 있는지를 나타낸다. 분석 후보를 많이 생성할수록 그 중 정답이 포함될 가능성이 증가하기 때문에 평균분석후보수와 정답포함률은 상충(trade-off)관계를 가진다. 분석정확도는 품사 부착 모형이 선택한 다수의 분석 후보 중 하나의 분석 결과가 실제로 올바른 분석 결과인 어절수의 비율이

3) http://www.sejong.or.kr/
4) 실험을 위해 펜티엄 3.2GHz의 CPU, 16GB 메모리 사양의 컴퓨터를 이용하였다.

표 4.2. 형태소 분석 모형의 평가
Table 4.2. Evaluation of the Morphological Analysis Model

형태소 분석 모형	분석속도(문장/초)	분석성공률	정답포함률	평균분석후보수
확률기반 모형	79.87	98.38%	96.78%	1.87
제안하는 규칙기반 모형	173.61	82.96%	81.39%	1.00
제안하는 혼합 모형	140.25	98.83%	96.59%	1.36

표 4.3. 품사 부착 모형의 평가
Table 4.3. Evaluation of the Part-of-Speech Tagging Model

형태소 분석 모형	품사 부착 모형	분석속도(문장/초)	분석정확도
확률기반 모형	은닉마르코프 기반 품사 부착 모형	2127.66	94.10%
제안하는 혼합 모형		3030.30	95.25%

다. 각각의 척도는 다음과 같이 계산할 수 있다.

$$\text{분석성공률} = \frac{\text{분석후보를 생성한 어절수}}{\text{어절수}}$$

$$\text{평균분석후보수} = \frac{\text{생성된 분석후보수}}{\text{어절수}}$$

$$\text{정답포함률} = \frac{\text{올바른 분석후보를 포함하는 어절수}}{\text{어절수}}$$

$$\text{분석정확도} = \frac{\text{올바로 분석한 어절수}}{\text{어절수}}$$

4.3 성능 평가

성능 평가는 두 단계에 나누어 수행되었다. 우선 제안하는 규칙 기반 형태소 분석 모형의 성능을 기존의 확률 기반 형태소 분석 모형의 성능과 비교한 후, 규칙 기반 모형이 확률 기반 모형과 결합한 혼합 모형의 최종 성능을 평가하였다.

〈표 4.2〉에서 나타난 바와 같이 제안한 규칙 기반 모형은 82%가 넘는 어절들을 분석하였다. 기존의 수동으로 구축 규칙 기반 모형으로 이에 해당하는 양의 분석성공률을 달성하기 위해서는 높은 규칙 구축 비용이 필요하지만, 제안하는 모형에서는 자동으로 구축되었다는 점에 의의가 있다. 또한 분석 성공한 어절은 모두 분석 후보 하나씩 생성하였고, 그 정확률은 98%(81.39%/82.96%)의 높은 성능을 보였다. 이는 확률 기반 모형이 다수의 분석 후보를 생성하면서 발생하게 될 모호성을 효과적으로 방지한 것으로 분석할 수 있다. 기존의 확률 모형에서의 평균분석후보수는 어절별 1.87개였지만, 제안하는 규칙 기반 모형은 분석한 모든 어절에 대해 단 하나만의 정확한 분석 후보를 생성하다.

제안한 최종 혼합 모형은 확률 모형과 규칙 모형이 결합된 모형이다. 규칙 모형이 먼저 분석을 시도하고, 규칙의 부재로 인해 분석 실패한 어절에 대해 확률 모형이 분석을 시도한다. 규칙 모형으로 인해 평균분석후보수는 1.87개에서 1.36개로 27%가 감소하였으며, 분석속도 역시 초당 79.87 문장에서 140.25 문장으로 크게 개선되었다. 정답포함률은 규칙 기반의 단일 후보 생성 전략으로 인해 미미하게 감소하였다.

최적의 분석 후보를 선택하는 품사 부착 성능을 측정하기 위해 확률 기반 형태소 분석 모형과 혼합 분석 모형 각각을 은닉마르코프 기반 품사 부착 모형과 결합하였다. 〈표 4.3〉에 나타난 것과 같이 제안한 혼합 모형은 분석속도와 분석정확도 측면에서 기존의 확률 기반 모형[11]의 성능을 크게 향상시켰다. 분석정확도는 94.10%에서 95.25%로 1.22%의 성능 향상을 보였다. 이는 제안하는 규칙 기반 분석 모형의 단일 후보 생성이 품사 부착 모형의 탐색 공간을 줄여주었을 뿐만 아니라 잘못된 후보가 선택되는 경우를 방지하였기 때문이다. 분석속도 역시 초당 2127.66 문장에서 3030.30 문장으로 개선되었다.

V. 결론 및 향후 연구

통계 기반 형태소 분석 모형은 높은 분석성공률을 보인 반면 학습 말뭉치가 불충분한 경우 확률 추정이 부정확하여 잘못된 분석으로 이어진다. 규칙 기반 모형은 고성능의 규칙을 구축함으로써 통계 기반 방식보다 정확한 분석이 가능하지만 다양한 언어현상을 고려하기는 힘들다. 본 논문에서는 규칙 기반 접근법과 통계 기반 접근법의 장점만을 취합한 혼합 모형을 제안하였다. 제안한 모형은 98.83%에 이르는 분석성공률과 95.25%의 분석정확도를 보였으며, 이에 필요한 분석

규칙을 자동으로 학습하였다. 통계 기반 모형과 결합한 혼합 모형은 분석정확도를 향상시켰을 뿐만 아니라 분석속도 역시 크게 개선하였다.

제안한 규칙 모형은 확률 모형에 비해 상대적으로 높은 분석정확도를 보인 반면 낮은 분석성공률을 보였다. 보다 정련된 규칙 형태와 분석 알고리즘을 고안하여 분석성공률을 높일 수 있다면 더 높은 성능 향상을 이룰 수 있을 것으로 예상된다. 또한 제안하는 모형은 품사 부착 말뭉치만 주어진다면 자동으로 분석 정보를 학습할 수 있기 때문에 높은 유연성을 가진다. 구어체나 웹문서와 같은 새로운 도메인에 적용을 고려해 볼 수 있다.

참고문헌

[1] 임희석, 어절의 중의성 유형 분류에 근거한 한국어 형태소 분석기, 고려대학교 전산과학과 석사학위논문, 1993.

[2] Chanod J. P. and P. Tapanainen, "Tagging French-Comparing a Statistical and a Constraint-based Method", Proc. of the 7th conference of the European chapter of the ACL, Doublin, pp.149-156, 1995.

[3] Hindle D., "Acquiring Disambiguation Rules from Text", Proc. of 27th Annual Meeting of the ACL, pp.118-125, 1989.

[4] Brill E, "A Simple Rule-based Part-of-speech Tagger", Proc. of the 3rd Conf. on Applied NLP. Trento Italy. pp.153-155, 1992.

[5] 박혜준, 윤준태, 송만석, "말뭉치 품사꼬리달기 시스템 구현", 한국정보과학회 봄 학술발표논문집 제21권 제1호, pp.829-832, 1994.

[6] 이하규, "어말-어두 공기 정보를 이용한 한국어 어휘 중의성 해소"", 한국정보과학회 정보과학회논문지 제24권 제1호, pp.82-89, 1997.

[7] 심준혁, 김준석, 이근배, "통계와 규칙을 이용한 강인한 품사태거", 한국어 형태소 분석기 및 품사태거 평가 워크숍 논문집, pp.60-75. 1999.

[8] 최원종, 오류 유형별 후처리를 통한 한국어 품사 부착 성능향상, 고려대학교 컴퓨터학과 석사학위논문, 2007.

[9] 임희석, 언어 지식과 통계 정보를 이용한 한국어 품사 태깅 모델, 고려대학교 컴퓨터학과 박사학위논문, 1997.

[10] 박희근, 서영훈, "어절별 중의성 해소를 이용한 품사 태깅의 성능 향상", 한글 및 한국어 정보처리 학술대회 , pp.134-139. 2007.

[11] 이도길, 한국어 형태소 분석과 품사 부착을 위한 확률 모형, 고려대학교 컴퓨터학과 박사학위논문, 2005.

[12] http://www.sejong.or.kr/

[12] 실험을 위해 펜티엄 3.2GHz의 CPU, 16GB 메모리 사양의 컴퓨터를 이용하였다.

저 자 소 개

이승욱 (Seung-Wook Lee)

2006년: 수원대학교 컴퓨터학과 이학사
2008년: 고려대학교 컴퓨터학과 이학석사
2008년 - 현재: 고려대학교 컴퓨터·전파통신공학과 박사과정

이도길 (Do-Gil Lee)

1999년 2월 고려대학교 컴퓨터학과 학사
2001년 2월 고려대학교 컴퓨터학과 석사
2005년 8월 고려대학교 컴퓨터학과 박사
2005년 9월 ~ 2006년 3월 고려대학교 컴퓨터정보통신연구소 연구교수
2006년 4월 ~ 2008년 2월 NHN (주) 과장
2008년 3월 ~ 현재 고려대학교 민족문화연구원 HK연구교수

임해창 (Hae-Chang Rim)

1981년: Missouri 주립대학 학사
1983년: Missouri 주립대학 석사
1990년: Texas 주립대학 박사
1991년~1994년: 고려대학교 전산과학과 조교수
1994년~1999년: 고려대학교 컴퓨터학과 부교수
1999년~현재: 고려대학교 컴퓨터·전파통신공학과 교수