공학석사 학위논문

기계학습을 이용한 음절기반 품사 부착

Korean Part-of-Speech Tagging Based on Syllables
Using Machine Learning Methods

지도교수 김재훈

2012년 2월

한국해양대학교 대학원

컴퓨터공학과
전 길 호

本 論文을 전길호의 工學碩士 學位論文으로 認准함.

위원장 신 옥 근 인

위 원 박 휴 찬 인

위 원 김 재 훈 인

2012년 1월 31일

한 국 해 양 대 학 교 대 학 원

목 차

그림 목차

표 목 차

Korean Part-of-Speech Tagging Based on Syllables Using Machine Learning Methods

Kil-Ho Jeon

Department of Computer Engineering,
Graduate School of Korea Maritime University

Abstract

Korean morphological analysis and part-of-speech (POS) tagging are more complex as compared to those for languages like English because Korean is agglutinative. In general, Korean POS taggers have, as a part of preprocessors, Korean morphological analyzers, which use very much knowledge like dictionary, but still have problems on unknown words and ambiguity. To alleviate the problems, a Korean POS tagger without a Korean morphological analyzer had proposed using machine learning techniques. The POS tagger has two problems : One is not to be able to segment compound nouns and the other one is not to be able to correctly recover root (or stem) forms from surface forms of words because of ambiguity in root form recovering rules. To solve the former problem, we propose a new syllable coding scheme for representing morphemes using machine learning methods like condition random fields (CRFs). Basically we use the BIO (Beginning, the Inside and Outside) coding scheme and additionally employ the special tag 'sp' for white spaces used between words. We also propose the new method for recovering root forms using machine learning techniques like Naïve Bayes models in order to solve the latter problem.

In this thesis, we propose the Korean POS tagging system with the two solutions. The proposed system consists of three modules. The

first module is the syllable tagger that segments words into morphemes and assigns BIO-tags coded for morphemes to syllables, the second module is the root form recoverer that recoveries the root forms of surface morphemes, and the final module is the POS recoverer that assigns POS tags to morphemes.

The proposed system has demonstrated the F_1-score of 97.5% for the ETRI tree-tagged corpus. Thus it can be decided that the proposed system is more effective than other existing Korean POS taggers using machine learning methods to handle the compound nouns and root form recovery.

제 1 장 서 론

인터넷의 급속한 발전으로 각종 포털 사이트의 게시판, 카페, 동호회, 블로그 등에는 수많은 문서가 생성되고 있다. 예를 들어 개인 블로그에는 관심분야에 따른 수많은 정보들이 게시되고 있고, 각종 동호회 게시판에는 동호회의 목적과 관련된 수많은 정보 등이 매일 게시되고 있다. 이렇게 많은 문서들은 분석과 분류를 통해 보다 많은 사람들에게 중요한 정보로 활용될 수 있고, 이러한 이유로 문서의 분석 및 분류와 같은 정보처리의 필요성이 대두되고 있다. 이러한 필요성에 따라 많은 학자들이 문서를 보다 정확하게 분석하고 분류하기 위한 방법들을 연구하고 제안하며 실제로 사용되고 있다(Manning *et al.*, 2010). 이러한 수많은 방법들 중에서 형태소 분석 및 품사 부착은 문서를 분석하고 분류하여 정보로 활용하기 위한 여러 방법들의 공통된 최하위 단계에 속한다.

형태소 분석이란 입력된 문서에 대해 형태소의 변형과 분리 경계를 결정하는 문제를 처리하는 과정으로 언어적 특성에 맞게 구현된다(Dale, *et al.*, 2000). 특히 한국어는 내용어와 기능어의 결합으로 다양한 형태의 변형이 발생된다(서정수, 1996). 이러한 이유로 한국어 형태소 분석기는 영어와 같은 외국어 형태소 분석기 보다 복잡한 구조를 가지고 있다[1]. 이렇게 복잡한 구조의 형태소 분석기를 설계하고 구현하기 위해서는 복잡한 지식과 방대한 사전정보가 요구된다(김재훈, 이공주, 2003). 뿐만 아니라 매우 까다로운 구현과정을 거치기 때문에 유지보수를 한다는 것은 형태소 분석기를 구현하는 것만큼 어려운 것이 현실이다.

그러나 일부 정보검색 시스템은 주어진 문장에서 명사만 추출하여 색인하는데 응용분야에 따라서는 모든 종류의 형태소 분석결과를 필요로 하지 않는다. 또한 품사부착은 형태소 분석에서 발생된 여러 분석 결과를 주어진 문장에 가장 적합한 분석을 선택하여 여러 응용분야에 사용된다.

이러한 문제들을 해결하국어 품사를기 위해 음절단위로 한 부착한 연구(심광섭, 2011)가 있으나 복합명사를 분석하기 어려우며 규칙을 사용하기 때문에 규칙의 모호성 문제가 존재한다.

---

1) 용언에 대한 형태소 분석은 활용 처리, 불규칙 처리, 음운현상 처리 등 매우 복잡한 과정을 포함하고 있다.

본 논문에서는 이와 같은 문제를 해결하고자 기계학습 기법을 이용한 음절기반 품사 부착 방법을 제안한다. 이 방법은 언어처리 시스템이나 대량의 사전정보를 이용하여 형태소 분석을 하지 않고 기계학습 도구를 이용하여 음절단위로 품사 부착이 가능한 학습모델을 생성하여 입력된 문장을 음절단위로 음절품사를 부착하고 어절경계를 표시하여 복합명사의 분석이 가능하다. 음절품사가 부착된 문장은 음절 복원기를 통해 음절의 원형 복원 결과를 얻는다. 음절을 복원하는 과정에서 발생하는 모호성 문제는 Naïve Bayes 분류기를 이용해서 해결한다. 본 논문에서 제안하는 형태소 분석 및 품사부착은 기계학습 기법을 이용하고 있으며, 구현이 쉽고 간단하기 때문에 단기간 내에 구현할 수 있으며, 복잡한 구조를 가진 기타 품사 부착기와 비슷한 수준의 성능을 가지고 있다.

본 논문의 구성은 다음과 같다. 2장에서 기존의 형태소 분석 및 품사 부착 방법들과 음절기반 언어 처리 방법들에 대해 살펴보고, 3장에서 기계학습에 필요한 학습말뭉치의 가공방법에 대해 살펴본다. 4장에서 기계학습을 이용한 음절 기반 형태소 분석에 대해 논하며 5장에서는 본 논문에서 제안한 방법으로 구현한 시스템의 성능을 평가한다. 마지막으로 6장에서 결론을 맺고 앞으로의 연구 방향을 제시한다.

제 2 장 관련연구

본 장에서는 본 논문에서 제안하는 품사 부착 방법에 대해 논하기에 앞서 기존 형태소 분석기 및 품사 부착 방법, 음절을 이용한 언어처리 방법 등에 대해 간단히 소개한다. 그리고 기존의 음절 기반 품사 부착 방법 및 그 문제점에 대해서 살펴본다.

2.1. 형태소 분석 및 품사 부착

형태소란 의미를 가진 가장 작은 언어의 단위를 말한다(서정주, 1996). 형태소 분석이란 문장에 포함된 형태소를 찾아내는 과정이며 형태소 분리 및 복원과 같은 복잡한 과정이 포함되어 있다. 품사부착은 형태소 분석과정에서 발생된 여러 가지 분석에서 가장 적합한 분석을 선택하는 과정이며, 이후 구묶음, 구문구조 분석 등을 수행하므로 형태소 분석 및 품사부착은 언어처리에 있어서 기초 단계이다.

<표 2.1>은 한국어 형태소 분석 및 품사부착과정의 예이다. 입력문장은 형태소 단위로 분리되고 형태소의 원형을 복원한 다음, 적용 품사를 분석하여 형태소 분석을 완료하고 품사부착을 통해 가장 적절한 품사를 결정하는 것으로 품사부착을 완료한다.

<표 2.1> 한국어 형태소 분석 및 품사부착과정의 예

구 분	분석의 예
입 력 문 장 :	아름다운 추억상자를 냈다고 해도 된다.
형 태 소 분 리 :	아름다 / 운 / 추억 / 상자 / 를 / 냈 / 다고 / 해 / 도 / 된 / 다/.
형태소 원형 분석 :	아름답 / ㄴ / 추억 / 상자 / 를 / 내 / 었 / 다고 / 하 / 어도 / 되 / ㄴ다/.
형태소 품사 분석 :	**아름답** / 형용사 + ㄴ / 전성어미
	추억 / 명사 + **상자** / 명사 + **를** / 조사
	내 / 동사 + **었** / 선어말어미 + **다고** / 연결어미
	하 / 동사 + **어도** / 연결어미
	되 / 동사 + **ㄴ다** / 연결어미 + . / 문장부호 **되** / 형용사 + **ㄴ다** / 종결어미 + . / 문장부호
품 사 부 착 :	**아름답** / 형용사 + ㄴ / 전성어미
	추억 / 명사 + **상자** / 명사 + **를** / 조사
	내 / 동사 + **었** / 선어말어미 + **다고** / 연결어미
	하 / 동사 + **어도** / 연결어미
	되 / 동사 + **ㄴ다** / 종결어미 + . / 문장부호

2.2. 한국어 형태소 분석 방법

형태소 분석은 한국어의 여러 특성을 고려하여 연구되고 있으며 여러 가지 방법(Two-level 형태론(Koskenniemi *et al.*, 1983), Head-Tail 구분법(최형석, 이주근, 1984), Tabular parsing 방법(김성용 외, 1987), 최장일치법과 최단일치법, 기계학습 기법 등)이 있다.

Two-level 형태론은 1983년 핀란드의 K. Koskenniemi가 형태소의 철자변화를 처리하기 위해 처음 제안한 일반적인 형태소 분석 및 생성 방법으로 사전표현(lexical representation)과 표층표현(surface representation)을 사용해 형태소 분석을 하는 방법이다(Koskenniemi *et al.*, 1983; Kim *et al.*, 1999).

Head-Tail 구분법은 단어를 어근(head)과 문법형태소(tail)로 분리하여 분석하는 방법이다. 한국어 처리 분야에 Head-Tail 구분법을 적용하면 인식의 모호성을 이유로 역 탐색 현상이 발생하고 효율이 떨어진다. 또한 모호성이 존재하는 분리점은 분리 가능한 모든 정보를 출력해야 하고, '사랑면서'와 같은 오류 어절은 오류로 인식해야 하는 문제점이 발생한다(Kwon *et al.*, 1994; 이근용 외, 2004). 이를 모두 해결하는 시스템을 제작하기 위해서는 고도의 기술과 많은 시간이 필요하다.

Tabular Parsing 방법은 상향식 구문분석법(bottom-up parsing) 방식을 따르는 형태소 분석 방법으로 입력 어절에 대해 가능한 모든 형태소 열을 계산하는 과정에서 중복계산을 줄이기 위해 Tabular Parsing 방법을 이용한다(김성용 외, 1987). 이는 앞서 언급한 Head-Tail 구분법과 마찬가지로 접속검사를 하기 때문에 접속검사표를 구성해야 하는 단점이 있지만 정확도가 높은 장점을 가지고 있다.

최장일치법은 가능한 모든 형태소로 분리한 형태소의 집합에서 가장 긴 형태소를 우선적으로 선택하는 방법을 말한다. (최재혁, 이상조, 1993)에서는 조어사전을 이용한 양방향 최장일치법을 사용하여 형태소 분석기 제작방법을 제안하였으며, (이기오 외, 1996)에서는 '먹어 보았다'와 같이 동사가 결합된 경우 인식하지 못하는 기존 최장일치법의 단점을 없앤 확장된 최장일치법을 제안하고 있다. 여기서는 파서(parser)를 사용하여 정확한 형태소 분석결과를 생성할 수 있도록 하고 있다. 최단일치법은 최장일치법과 반대로 가장 짧은 형태소를 우선적으로 선택하는 방법으로 기본적인 의도는 어절을 분석하는 과정에서 문법적인 형태소를 발견하면 더 이상 분석하는 과정을 하지 않고 올바른 형태소로 간주하자는 것이다(조영환 외, 1990). 이는 어절 내의 오류를 발견하기 위해서 하나의 어절에 대해 단지 형태론적인 분석으로도 족하기 때문이다.

기계학습 기법은 학습말뭉치의 문맥 정보, 품사 정보 등을 이용하여 입력된 형태소를 분석하는 기법으로 말뭉치로부터 정보를 추출하여 저장한 학습모델을 필요로 한다. 기계학습 기법은 본 논문과 밀접한 관련이 있으므로 2.5절에서 자세히 살펴본다.

2.3. 한국어 품사 부착 방법

한국어 품사 부착 방법은 크게 통계 기반 접근 방법, 규칙기반 접근 방법 등으로 구분할 수 있다(임해창 외, 1996). 통계기반 접근 방법은 말뭉치를 분석하고 확률정보를 추출하여 이용하는 방법이고, 규칙기반 접근 방법은 한국어에 적용되는 공통된 원리나 결정적인 규칙을 이용하는 방법이다.

통계 기반 접근 방법에는 어절 단위로 품사를 부착하는 모델과 형태소 단위로 품사를 부착하는 모델이 있다. 어절 단위 품사 부착 방법은 어절 열이 주어졌을 때 확률이 가장 높은 어절의 태그 열을 구하는 것이고 형태소 단위 품사 부착 방

법은 어절 단위 품사 부착 방법과 동일하게 어절 열이 주어졌을 때 확률이 가장 높은 형태소 열과 그에 대응하는 품사 열을 구하는 것이다. 확률값은 HMM모델, Tri-gram에 기반한 상태 전이 확률 등을 사용한다(김재훈, 1996).

규칙 기반 접근 방법은 한국어에 적용되는 공통된 원리나 결정적인 규칙을 적용하여 품사를 부착하는 방법인데 일관성 있고 예외가 없는 규칙을 찾는 것이 쉽지 않다. 규칙을 사용한 품사 부착 방법으로는 한국어의 변형규칙을 이용한 품사 부착(임희석 외, 1997) 등이 있다. 이 시스템들은 여러 종류의 기능어 사전을 참조하고 예외단어와 접미어, 특수기호 등을 처리하여 품사부착 규칙을 적용하는 방법이다.

2.4. 음절정보를 이용한 언어처리

음절 정보를 이용한 언어처리 방법은 단어 분리, 형태소 분석기 등 그 종류가 매우 다양하며 어절 정보를 이용한 대부분의 언어처리 방법에 필요한 형태소 분석 단계를 생략할 수 있는 장점이 있다.

2.4.1. 단어 분리 및 범주 결정

(김재훈, 이공주, 2003)에서는 형태소 분석기, 품사 부착 시스템, 대량의 사전 정보 등을 이용하지 않고 사례기반 학습을 통해 단어를 분리하고, 분리된 단어의 범주를 결정하는 방법을 제안하고 있으며 크게 말뭉치 구축과정과 학습 및 분류 과정으로 구성되어 있다.

말뭉치 구축 과정은 정답 정보를 포함하는 말뭉치를 구축하기 위해서 품사 부착 말뭉치로부터 음절 기반 단어 분리 말뭉치를 구축하는데 품사 부착 말뭉치에는 어절에 대한 올바른 형태소 분석 결과가 포함되어 있기 때문이다. 이 과정에서 구축된 말뭉치는 차후 시스템에서 사용하는 지도학습의 정답으로 이용된다.

학습 과정에서는 앞서 구축한 음절 기반 단어 분리 말뭉치로부터 자질을 추출하고 자질집합을 이용해서 학습을 수행한다. 여기서 사용된 사례기반 학습은 유사도기반 혹은 예제기반 학습이라고도 불리며 지도학습(supervised learning) 방법의 일종이다. 시스템의 구성은 품사 부착 말뭉치로부터 음절기반 단어 분리 말뭉치를 구축하는 음절기반 단어분리 말뭉치 구축 단계, 학습모델을 생성하는 학습과정 단계, 생성된 모델과 분류기를 이용해서 일반문장을 음절기반으로 단어

분리하는 분류 단계로 구성되어 있다.

2.4.2. 한국어 품사 부착

위에서 언급한 사례기반 단어 분리 및 범주결정 문제 해결 방법을 확장하여 형태소 단위 품사태깅을 한 연구도 있다(심광섭, 2011). 여기서는 기계학습을 통해 모델을 생성하고 입력된 문장에 대해 음절단위로 품사를 부착한다. 음절별로 품사가 부착된 문장은 동일한 품사가 연속된 음절들을 묶고 하나의 품사를 부착하여 형태소 단위로 품사가 부착된 문장으로 변형시킨다. <표 2.2>는 입력된 문장에 대해 음절단위로 품사부착을 하고 형태소 단위로 품사가 부착된 문장으로 변형하는 과정을 보여준다. 이 시스템은 본 논문과 유사한 점이 많기 때문에 2.5절에서 자세히 살펴본다.

<표 2.2> 음절단위 품사부착으로부터 형태소 단위 품사부착 과정

구분	부착의 예
입력문장	아름다운 추억상자를 냈다고 해도 된다
음절단위 품사 부착	아/AJ + 름/AJ + 답/AJ + ㄴ/EM 추/NN + 억/NN + 상/NN + 자/NN + 를/JO 내/VV + 었/EP + 다/EM + 고/EM 하/VV + 어/EM + 도/EM 되/VV + ㄴ/EM + 다/EM
음절단위 품사 부착으로부터 형태소 단위 품사 부착	아름답/AJ + ㄴ/EM 추억/NN + 상자/NN + 를/JO 내/VV + 었/EP + 다고/EM 하/VV + 어도/EM 되/VV + ㄴ다/EM

2.4.3. 복합명사 분해

한국어는 명사를 연속하여 사용하는 것으로 복합명사를 만들 수 있다. 이러한 복합명사는 사람과 사람 사이에서 별 문제가 되지 않지만 정보검색 시스템이나 기계번역 시스템과 같은 한국어 정보처리 시스템에서는 심각한 인식오류 문제를 야기시킨다(심광섭, 1997). 이를 해결하기 위해 조합이 가능한 모든 복합명사를 사전에 수록하는 방법이 있으나 조합이 가능한 복합명사의 수가 너무 많기 때문

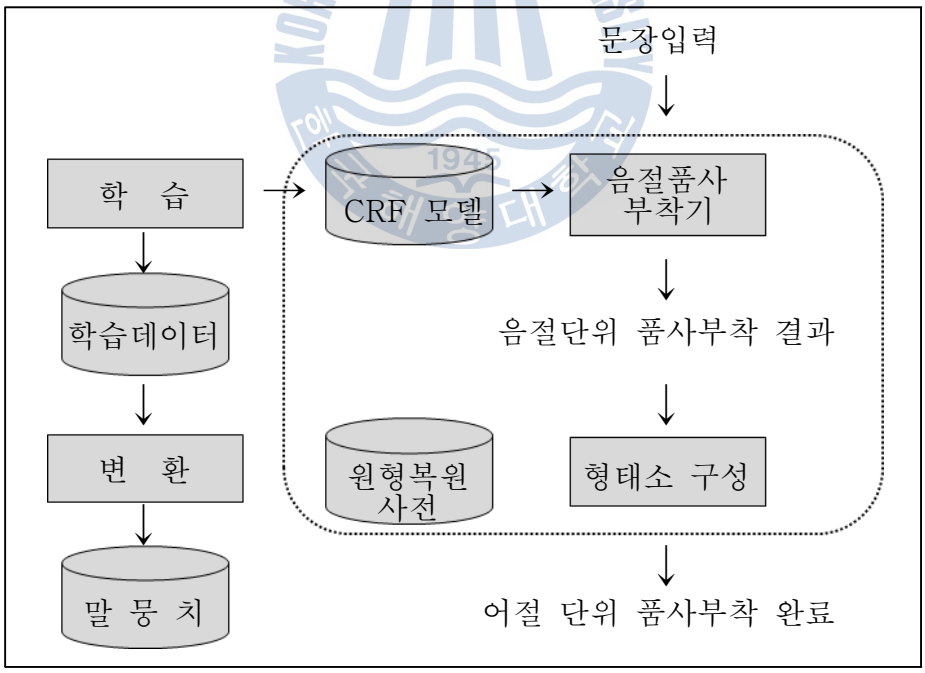
에 이 방법은 사실상 불가능하다.

(박성배, 장병탁, 2003)는 음절 정보를 이용하여 복합명사를 분해하는 방법을 제안하고 있는데 I-REP(Incremental Reduced Error Pruning) 알고리즘(Furnkranz *et al.*, 1994)에 기초하여 성능을 개선시킨 GECORAM (GEneralized Combination of Rule-based learning And Memory-based learning) 알고리즘을 개발하여 사용하고 있다.

2.5. CRF를 이용한 한국어 품사 부착

기계학습을 이용한 기존 형태소 분석 방법은 그 종류가 매우 다양하나 본 절에서는 2.4.2절에서 언급했던 기계학습을 이용한 음절기반의 한국어 품사 태깅 방법(심광섭, 2011)에 대해 살펴보고 이 시스템의 문제점 및 개선점에 대해 논하도록 한다.

이 시스템은 먼저 음절 단위로 음절품사를 부착한 뒤, 형태소를 복원하여 품사를 부착하는 방법이다. 이는 어절 단위로 품사부착을 할 때 나타나는 자료 부족 문제를 해결하기 위한 방법으로 시스템 구조는 <그림 2.1>과 같다.



<그림 2.1> CRF를 이용한 음절기반 품사부착 시스템

문장 입력 단계에서 입력된 문장은 공백을 제거하되, <그림 2.2>와 같이 어절의 시작을 B, 중간과 끝은 I로 표시하여 어절경계를 표시하고, 음절 단위로 분리

하여 미리 생성된 CRF 모델에 적용한다. CRF 모델에서는 음절단위로 입력된 문장에 음절품사를 부착하여 음절품사가 부착된 문장을 출력한다. 출력된 문장은 규칙을 이용하여 어절의 복원 및 음절의 원형을 복원하여 최종 형태소 집합을 구성한다.

2.5.1. 음절품사 부착기

이 시스템에서 음절품사 부착기는 학습말뭉치로부터 정보를 학습한 학습모델을 사용하여 입력된 문장에 음절단위로 음절품사를 부착하는 역할을 한다[2]. 학습말뭉치 제작 과정은 <그림 2.3>과 같이 어절과 어절에 대한 품사 태깅 결과가 일대일로 대응되어 있는데 이 두 부분을 음절 단위로 비교하면서 어절을 구성하는 각 음절에 적절한 태그를 부여하는 방식으로 제작된다.

품사부착 말뭉치

어절	품사부착
아름다운	아름답/AJ+은/EM
추억상자를	추억/NN+상자/NN+를/JO
냈다고	내/VV+었/EM+다고/EM
해도	하/VV+어도/EM
된다	되/VV+ㄴ다/EM

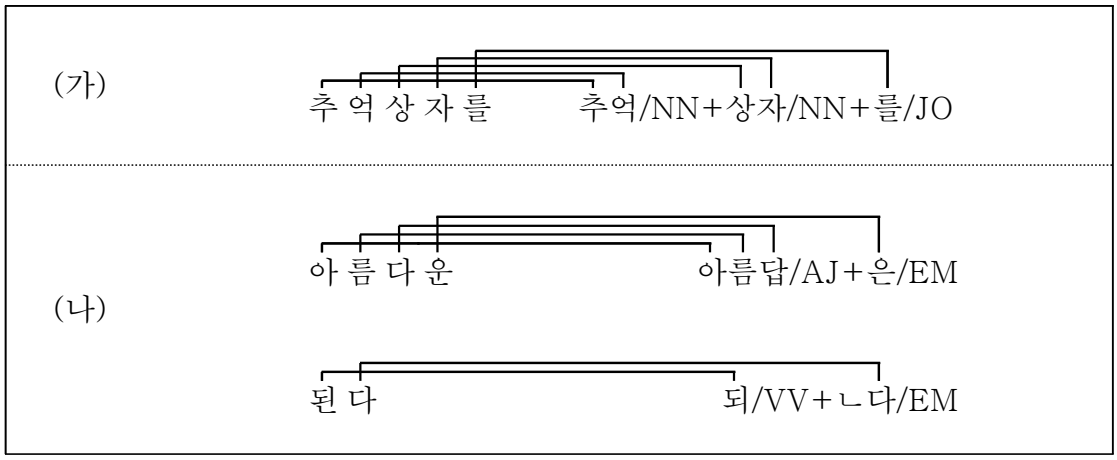
⇒

음절 품사 부착 학습 말뭉치

음절	어절 경계	품사
아	B	AJ
름	I	AJ
다	I	AJ
운	I	EM
추	B	NN
억	I	NN
상	I	NN
자	I	NN
를	I	JO
냈	B	VVEM
다	I	EM
고	I	EM
해	B	VVEM
도	I	EM
된	B	VVEM
다	I	EM

<그림 2.2> 학습말뭉치 제작과정 1

2) 품사부착에 사용된 NN, JO, EM, VV, AJ는 각각 명사, 조사, 어미, 동사, 형용사를 의미한다.



<그림 2.3> 학습말뭉치 제작과정 2

이 때 <그림 2.3> 의 (가)의 분석결과와 같이 두 부분의 음절이 일치하는 경우에는 쉽게 변환이 가능하지만 (나)의 분석결과와 같이 음절이 일치하지 않는 경우도 존재한다. 이 경우에는 한국어의 형태론적 특성을 고려하여 처리한다. 이 예에서는 'ㅂ불규칙' 용언이 'ㄴ(은)', 'ㄹ(을)', 'ㅁ(음)' 등으로 시작하는 어미 앞에서 'ㅂ'이 탈락하고 'ㄴ(은)', 'ㄹ(을)', 'ㅁ(음)'은 각각 '운', '울', '움'으로 변한다는 특성을 고려하여 '다'는 '답'과 대응하는 것으로 간주하고 '아름답'에 부착되어 있는 품사 'AJ'를 부여했으며, '운'은 '은'에 대응되는 것으로 보고 품사 'EM'을 부여했다. 또한 '된다' 에서 '된'의 오른쪽에 나오는 음절 '다'는 부착 결과 'ㄴ다'의 '다'와 일치하므로 '된'은 '되ㄴ'과 대응되는 것으로 판단하여 '된'에 대해서는 'VVEM'이라는 복합태그를 부여한다. 일반적으로 모음으로 끝나는 용언 어간에 "ㄴ, ㄹ, ㅁ, ㅂ, ㅆ, 아/어" 등이 결합된 경우에는 복합태그가 부여된다. <표 2.3>은 품사부착 중 복합태그가 부여된 경우의 예이고 'B', 'I'는 어절의 경계를 표시하는 것으로 'B'는 음절의 시작을 뜻하고, 'I'는 중간과 끝을 의미한다. 이는 차후 음절을 어절로 복원하는 단계에서 경계를 구분하는 용도로 사용된다.

<표 2.3> 품사부착 중 복합태그의 예

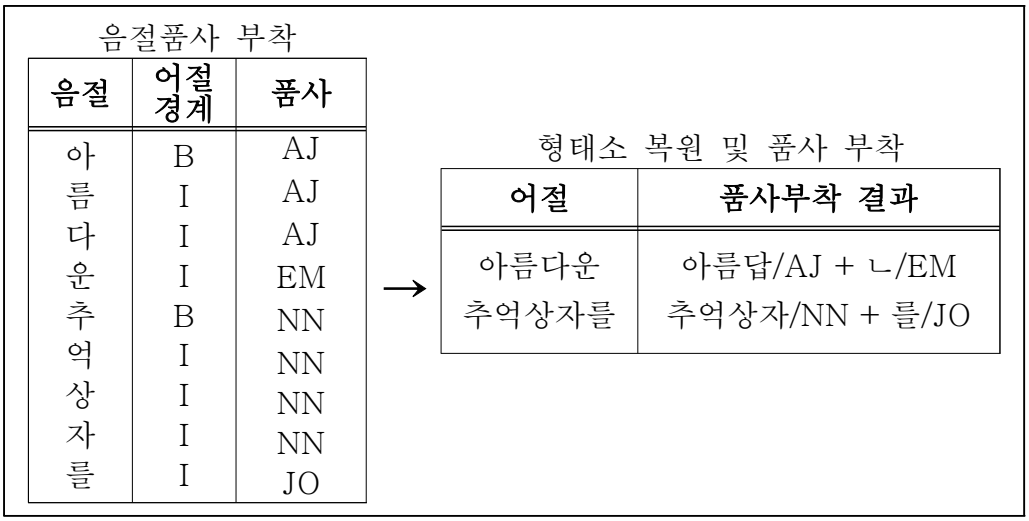
어절	품사부착 결과	음절	어절 경계	품사
다른	다르/AJ + ㄴ/EM	다 른	B I	AJ AJEM
냈	내/VV + 었/EM	냈	B	VVEM
했지만	하/VV + 었/EP + 지만/EM	했 지 만	B I I	VVEP EM EM
해	하/VV + 어/EM	해	B	VVEM

2.5.2. 규칙을 이용한 원형복원

음절품사가 부착된 문장은 사전에 정의한 규칙을 이용해 복합태그를 단순태그로 변경하고 동일한 품사를 가지는 연속된 음절을 하나의 어절로 묶어서 어절복원을 한다. 뿐만 아니라 형태론적으로 변형된 음절의 원형을 복원한다.

복합태그를 단순태그로 분리하는 방법은 음절을 “ㄴ, ㄹ, ㅁ, ㅂ, ㅆ, 아/어” 등을 제거한 음절과 “ㄴ, ㄹ, ㅁ, ㅂ, ㅆ, 아/어” 등이 포함된 음절로 분리하여 복합태그를 둘로 나눠 첫 번째 품사는 첫음절에 부착하고 두 번째 품사는 두 번째 음절에 부착하여 단순태그로 분리한다. 예를 들어 “했/SVEP → 하/SV + 었/EP”과 같이 복합품사 ‘SVEP’를 가진 ‘했’ 음절을 “ㄴ, ㄹ, ㅁ, ㅂ, ㅆ, 아/어” 등을 제거한 ‘하’ 와 “ㄴ, ㄹ, ㅁ, ㅂ, ㅆ, 아/어”가 포함된 ‘었’으로 나누고 복합품사 ‘SVEP’를 ‘SV’와 ‘EP’로 나누어 각각 음절에 부착한다.

음절을 어절로 복원하는 방법 또한 규칙을 이용하는데 <그림 2.4>과 같이 동일한 품사를 가지는 음절끼리 묶은 후 공통 태그를 부여하는 방식을 사용한다.



<그림 2.4> 어절 복원

그러나 이 방법은 음절이 형태론적으로 변한 경우에는 적용할 수 없는 문제가 있다. 이를 해결하기 위해 원형복원 사전을 이용한다. 원형복원 사전은 품사 부착 말뭉치에서 학습데이터를 제작할 때 어절과 어절에 대한 품사 부착 결과의 어절이 다른 경우 해당 음절이 포함된 형태소와 여기에 포함된 음절과 대응되는 음절열을 어절 부분에서 찾아 원형 복원 사전에 저장하는 방법으로 제작한다. 원형복원 사전은 <그림 2.5>와 같이 콜론(:) 왼쪽에는 탐색키가 위치하고 오른쪽에는 탐색키에 포함된 변형된 형태소를 대체할 원형 형태소가 위치한다.

형태소 및 품사		형태소의 원형
아름다/AJ	:	아름답
들/VV	:	듣
운/EM	:	ㄴ

<그림 2.5> 원형복원 사전의 예

원형 복원 사전을 이용하여 품사부착 및 어절복원 결과 “아름다/AJ + 운/VV 추억상자/NN + 를/EM”의 원형을 복원하면 “아름답/AJ + ㄴ/VV 추억상자/NN + ㄴ/EM”과 같은 품사부착 결과를 얻을 수 있다.

2.5.3. 시스템의 문제점

이 시스템에는 두 가지 문제점이 있다. 먼저 어절 복원 단계에서 동일한 품사를 가진 음절을 하나의 어절로 간주하는 규칙으로 어절복원을 하게 되면 복합명사를

구분해 내지 못하는 문제가 생긴다. <그림 2.6>의 (가) 예제 '추억상자를'에서 앞의 네 음절은 모두 'NN' 품사를 가지고 있기에 '추억상자/NN'로 품사부착이 되지만 실제로 '추억상자'는 복합명사이고 '추억' 과 '상자'로 분리되어 "추억/NN + 상자/NN"로 분석되는 것이 정확한 분석이다.

뿐만 아니라 규칙을 이용해 형태론적으로 변형된 음절의 원형을 복원하는 경우 "들/VV → 듣" 규칙을 <그림 2.6> 의 (나)에 적용하면 문제없이 정상적으로 원형 복원이 가능하다. 그러나 (다)의 경우 주변 음절 정보를 고려하지 않고 규칙에 의존하여 음절의 원형을 복원하기 때문에 다른 원형으로 복원되는 문제가 발생한다.

(가)

음절	어절 경계	품사
추	B	NN
억	I	NN
상	I	NN
자	I	NN
를	I	JO

→

어절	품사부착 결과	비고
추억상자를	추억상자/NN + 를/JO	복합명사오류

→

어절	품사부착 결과	비고
추억상자를	추억/NN + 상자/NN + 를/JO	

(나)

음절	어절 경계	품사
소	B	NN
리	I	NN
를	I	JO
들	B	VV
어	I	EM
보	B	VV
다	I	EM

→

어절	품사부착 결과	비고
소리를	소리/NN + 를/JO	
들어	**듣/VV + 어/EM**	
보다	보/VV + 다/EM	

원형복원 사전	비고
들 / VV : 듣	적용가능

(다)

음절	어절 경계	품사
책	B	NN
을	I	JO
들	B	VV
어	I	EM
보	B	VV
다	I	EM

→

어절	품사부착 결과	비고
책을	책/NN + 을/JO	
들어	**듣/VV + 어/EM**	원형복원오류
보다	보/VV + 다/EM	

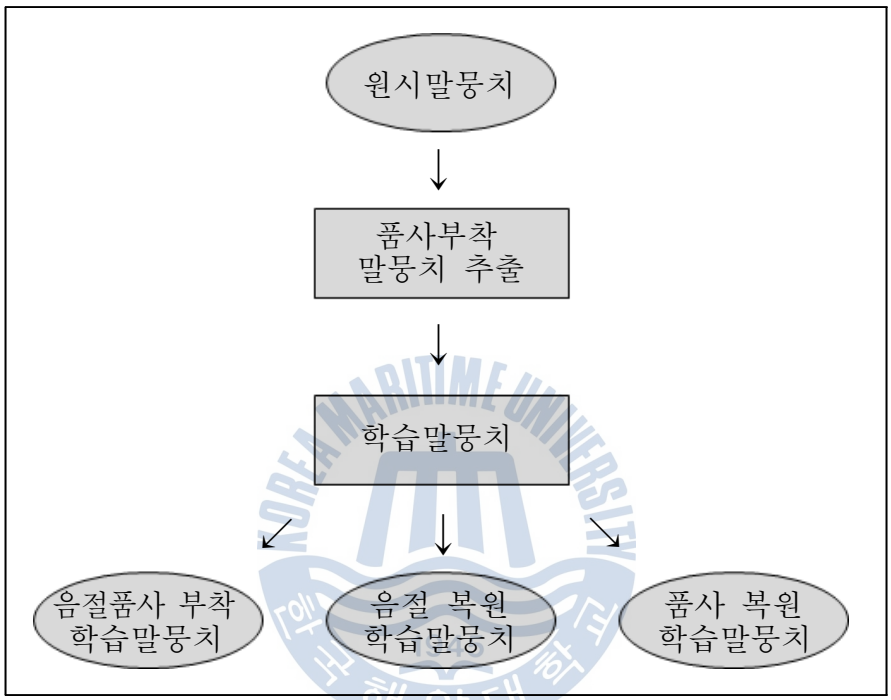
원형복원 사전	비고
들 / VV : 듣	적용불가

<그림 2.6> (심광섭, 2011)에서 제안한 품사부착 시스템의 문제점

본 논문에서는 형태소의 경계를 표시하는 '+' 기호(형태소 구분자)를 사용하여 앞서 언급한 복합명사의 형태소 분석 불가 문제를 해결한다. 형태소가 분리되는 음절의 품사는 품사 뒤에 '+' 기호를 사용하여 차후 음절을 어절단위로 복원하는 경우 정확한 구분이 될 수 있도록 하였다. 또 Naïve Bayes 분류기를 사용하여 형태소의 원형을 복원하는 과정에서 해당 전후의 문맥정보를 고려하지 않고 규칙만을 사용해서 발생하는 모호성 문제를 해결하였다. 이와 관련된 내용은 4장에서 자세히 설명한다.

제 3 장 학습말뭉치의 구성 및 가공

본 장에서는 기계학습에 필요한 학습말뭉치에 대해 살펴보고 말뭉치의 가공에 필요한 여러 알고리즘 등의 방법에 대해 기술한다. <그림 3.1>은 원시말뭉치로부터 음절품사 부착 학습말뭉치, 음절 복원 학습말뭉치, 품사 복원 학습말뭉치를 생성하는 과정이다.



<그림 3.1> 학습말뭉치의 생성 과정

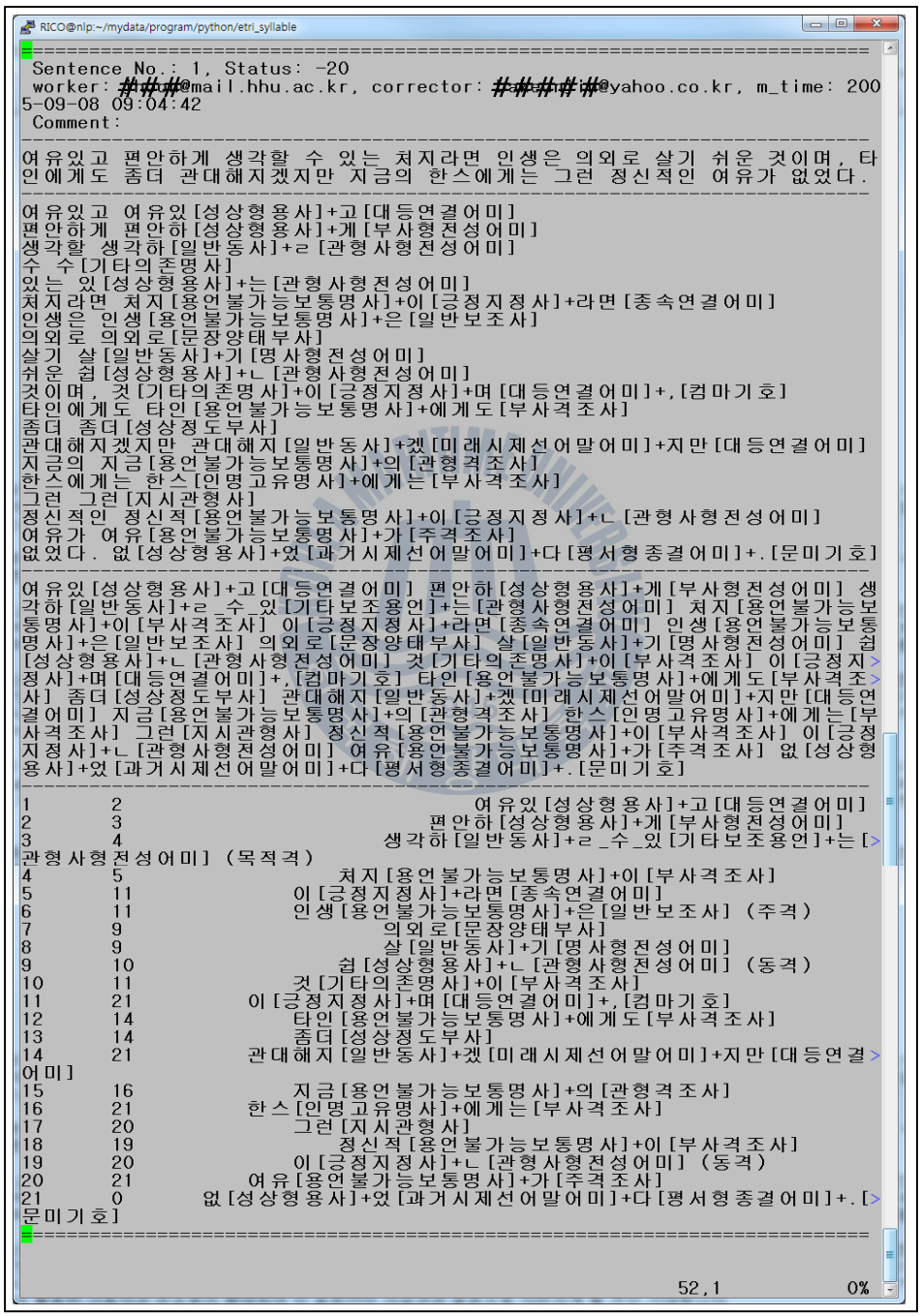
3.1. 품사 태그 집합

본 시스템에 사용된 말뭉치(김재훈 외, 2005)의 품사태그 수는 총 71개이다. 그러나 기계학습 모델의 성능을 향상시키기 위해 본 논문에서는 품사태그들을 26개의 품사태그(음절 품사 13개 + 형태소 경계 13개)로 줄여서 사용한다. 음절 품사태그는 품사의 종류[3]를 13개로 나눈 것이며, 형태소 경계는 음절 품사 태그와 동일하지만 형태소의 경계에 위치하는 음절 품사 태그로 경계를 구분할 수 있도록 음절 품사 태그 뒤에 '+'기호를 부착하여 사용한다. 자세한 사항은 <부록 1>에 기재하였다.

3) 명사, 조사, 수사, 용언, 접두/접미사, 동사, 형용사, 관형사, 전성사, 부사, 어미, 기호, 감탄사

3.2. 학습말뭉치의 구성

본 논문에서는 <그림 3.2>과 같은 ETRI 구문구조 부착말뭉치를 학습말뭉치로 사용한다. 이 학습말뭉치는 문서번호, 제작자, 수정자, 제작일자 등의 정보(A)와 함께 원본문장(B)에 대해 형태소 분석 결과(C), 구묶음 분석 결과(D), 구문구조 분석 결과(E)를 '-'의 열로 구분하고 있다. 말뭉치의 통계치는 <표 3.1>과 같다.



<그림 3.2> ETRI 구문구조 말뭉치의 일부

<표 3.1> ETRI 구문구조 말뭉치의 통계치

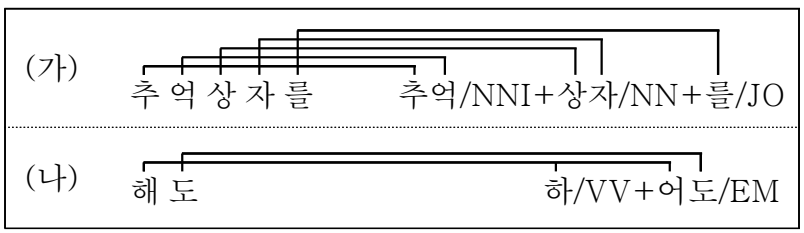
구분	수치
문장 수	101,602
문장 당 평균 어절 수	21.6
어절 수	2,200,914
어절 당 평균 형태소 수	1.98
총 형태소 수	4,375,332
전체 품사 수	71

3.3. 학습말뭉치 구축

본 시스템에 사용된 학습말뭉치는 3.2절에서 언급했던 것처럼 여러 종류의 분석결과를 포함하고 있기 때문에 필요한 분석결과만 추출하는 가공단계가 필요하다. 본 논문에서 제안하는 시스템은 음절기반의 형태소 분석이므로 형태소 분석결과를 사용하여 음절단위로 품사부착이 가능하도록 가공한다.

3.3.1. 어절과 형태소 분석 결과의 정렬

형태소 분석결과로부터 음절 단위로 음절품사를 부착하는 것은 실제 구현에 있어서 매우 까다로운 구조를 가지고 있다. <그림 3.3>의 (가)와 같이 어절과 해당 어절이 형태소 단위로 분석된 결과의 형태소열과 일치하는 경우는 순차적으로 어절의 각 음절에 해당 품사를 부착하면 된다. 그러나 (나)와 같이 어절과 형태소열이 일치하지 않는 경우에는 음절이 일치하는 부분과 일치하지 않는 부분으로 구분하여 품사를 부착해야 한다.



<그림 3.3> 음절단위 품사부착의 예

이를 위해 본 시스템에서는 어절과 형태소열이 일치하지 않는 경우 입력된 어절과 형태소열을 음절단위로 유사도를 비교하여 높은 유사도를 가진 음절끼리 같은 음절로 정렬하는 함수 align_syllable()을 제작하여 사용한다. 이 함수는 <그림 3.4>와 같다. 이 함수에서 유사도 계산은 Levenshtein Distance(Edit Distance) 알고리즘(Jurafsky and Martin, 2009)을 사용한다.

```
def align_syllable(s1, s2):
  // s1, s2 = 입력 어절
  // 입력예시 : s1 = '아름다운', s2 = '아름답ㄴ'
  // path = s1, s2내 음절의 관계 정보

  m = len(s1)  // m = 입력어절 s1의 음절길이
  n = len(s2)  // n = 입력어절 s2의 음절길이

  d = {(i,j): 0 for i in range(m) for j in range(n)}
   // d = 두 음절의 유사도 정보를 저장하는 2차원 배열
  p = {(i,j): (-1, -1) for i in range(m) for j in range(n)}
   // p = 유사도 정보에 따른 경로를 저장하는 2차원 배열

  for i in range(m): d[i, 0] = i
  for j in range(n): d[0, j] = j

  for i in range(1, m): p[i, 0] = (i-1, 0)
  for j in range(1, n): p[0, j] = (0, j-1)

  for j in range(1,n):
      for i in range(1, m):
          if s1[i] == s2[j]:
              d[i,j] = d[i-1,j-1]
              p[i,j] = (i-1, j-1)
          else:
              replace_dist = 1 - Levenshtein.ratio(s1[i], s2[j])
              delete = d[i-1,j] + 1
              insert = d[i,j-1] + 1
              replace = d[i-1,j-1] + replace_dist

              d[i,j] = min(delete, insert, replace)
              if d[i,j] == replace:  p[i,j] = (i-1, j-1)
              elif d[i,j] == insert: p[i,j] = (i, j-1)
              else: p[i,j] = (i-1, j)

      path = [(m-1, n-1)]
      i, j = m-1, n-1
      while path[-1][0] != 0 and path[-1][1] != 0:
          path.append(p[i,j])
          i, j = p[i,j]
      path.reverse()

      for i, j in path: print (s1[i], s2[j])

  return (path)
```

<그림 3.4> 어절과 형태소 분석 결과의 정렬 알고리즘

3.3.2. 원시말뭉치의 가공

본 논문에서는 학습말뭉치로 음절품사 부착 학습말뭉치, 음절 복원 학습말뭉치, 품사 복원 학습말뭉치를 사용한다. 음절품사 부착 학습말뭉치는 음절품사 부착을 위한 학습말뭉치이고, 음절 복원 학습말뭉치는 음절 복원을 위한 Naïve Bayes 분류기에 사용되는 학습말뭉치이다. 품사 복원 학습말뭉치는 원형 품사를 부착하기 위한 학습말뭉치이다. <그림 3.5>는 학습말뭉치의 생성 알고리즘이며 이 알고리즘에 의해 3가지 종류의 학습말뭉치가 생성된다. <그림 3.3>의 (가)와 같이 어절과 형태소열이 일치하는 경우에는 형태소에 부착된 품사태그를 해당 음절에 부착하고 (나)와 같이 어절과 형태소열이 일치하지 않는 경우에는 유사도를 계산하여 유사도가 높은 두 음절을 같은 음절로 간주하고 품사를 부착한다.

```
def make_learning_corpus(W):
  // W = Wi 집합
  // Wi = 어절 P와 P의 형태소 분석 결과(M, T)의 리스트, [P, M, T]
  // M = P의 형태소 분리 결과의 리스트, T = M의 품사정보 리스트
  // C1 = 음절품사 부착 학습말뭉치
  // C2 = 음절 원형 복원 학습말뭉치
  // C3 = 품사 부착 학습 말뭉치

  for i = 0 to W.size()
    for j = 0 to W[i].M.size()
      C3.append(W[i].M[j], get_reduced_form(W[i].T[j]), W[i].T[j])
      for k = 0 to W[i].M.size()-1 :
        t.append(get_reduced_form(W[i].T[j]))    //품사열 저장
      t.append(get_reduced_form(W[i].T[-1])+'+') //어절구분자 포함
      m.append(W[i].M[j])                        //형태소열 저장
    S.append(W[i].P, m, t)
    S.append('sp', 'sp', 'SP+')

  for i = 0 to S.size()
    p = S[i][0], m = S[i][1], t = S[i][2]
    if p == m
      for j = 0 to p.size()
        if j == p.size()
          C1.append(p[j], t[j], t[j]), C2.append(p[j], t[j], m[j]+)
          break
        C1.append(p[j], t[j], t[j]), C2.append(p[j], t[j], m[j])
    else
      Lp, Lm = Levenshtein_distance(p, m) // 어절과 형태소열의 음절 정렬
      Lt = t                              // 복합태그
      for k = 0 to Lm.size-1
        if Lp[k] == Lp[k+1]
          Lm[k] = Lm[k] + Lm[k+1]
          Lt[k] = Lt[k] + Lt[k+1]
          delete(Lp[k+1]), delete(Lm[k+1]), delete(Lt[k+1])
      for j = 0 to Lp.size()
        if j == p.size()
          C1.append(Lp[j], Lt[j], Lt[j]), C2.append(Lp[j], Lt[j], Lm[j])
          break
        C1.append(Lp[j], Lt[j], Lt[j]), C2.append(Lp[j], Lt[j], Lm[j])
  return (C1, C2, C3)
```

<그림 3.5> 학습말뭉치의 생성 알고리즘

알고리즘의 11번째 줄에 위치한 get_reduced_form() 함수는 품사의 원형을

입력 받아 축약된 품사를 반환하는 함수이다. 본 논문에서는 학습모델의 성능 향상을 위해 축약된 품사를 사용하는데 이와 관련된 사항은 본 논문의 3.1절에서 기술하였다. 음절품사 부착에 사용되는 학습말뭉치 C_1은 <표 3.2>와 같은 품사 부착결과에서 형태소를 음절단위로 분리하고 분석된 품사를 분리된 음절에 부착하여 학습말뭉치를 생성한다. 한 어절에 대한 분석이 끝나면 'sp'기호를 사용하여 어절을 구분하고, 하나의 음절처럼 취급한다. 또한 형태소 경계 구분자 '+'를 사용하는데 이는 차후 형태소의 경계를 구분하는 역할을 하며 복합명사를 처리하는 중요한 구분자로 사용한다. <표 3.3>은 음절품사 부착 학습말뭉치의 예이며, 음절정보와 축약된 품사정보를 포함하고 정답 또한 품사정보를 사용한다.

<표 3.2> 원시말뭉치에서 품사 부착의 예

어절	품사 부착 결과
아름다운	아름답/VAB + ㄴ/EEI
추억상자를	추억/NNI + 상자/NNI + 를/FSD
냈다고	내/VBB + 었/ERD + 다고/EEA
해도	하/VBB + 어도/EEG
된다	되/VBB + ㄴ다/EEA

<표 3.3> 음절품사 부착 학습말뭉치의 예

자질(음절)	정답(품사)
아	V
름	V
다	V
운	V+E+
sp	SP+
추	N
억	N+
상	N
자	N+
를	P+
sp	SP+
냈	B+E+
다	E
고	E+
sp	SP+
해	B+E
도	E+
sp	SP
된	B+E
다	E+

음절 복원에 사용되는 학습말뭉치 C_2 생성결과의 예는 <표 3.4>과 같고 C_1 학습말뭉치를 생성할 때 자질정보는 음절과 품사를 사용한다. 대신 정답에 대한 정보는 <표 3.2>와 같은 형태소 분석결과에서 음절의 원형 정보를 추출하여 생성하기 때문에 자질은 C_1과 동일하다.

<표 3.4> 음절 복원 학습말뭉치의 예

자질1(음절)	자질2(품사)	정답(음절원형)
아	V	아
름	V	름
다	V	답
운	V+E+	ㄴ
sp	SP+	sp
추	N	추
억	N+	억
상	N	상
자	N+	자
를	P+	를
sp	SP+	sp
냈	B+E+	내+었
다	E	다
고	E+	고
sp	SP+	sp
해	B+E	하+어
도	E+	도
sp	SP	sp
된	B+E	되+ㄴ
다	E+	다

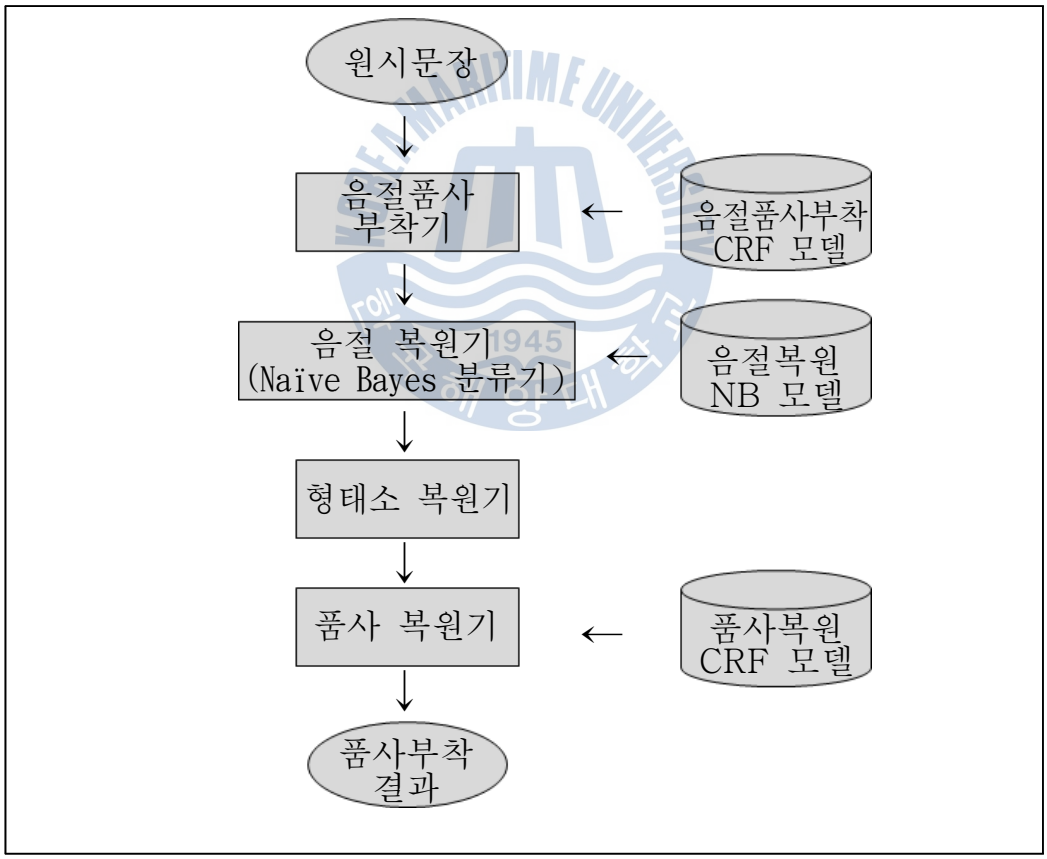
마지막으로 품사 복원 학습말뭉치 C_3는 품사의 원형을 결정하는 모델의 학습 말뭉치 이다. 이 말뭉치는 <표 3.5>와 같이 형태소와 축약된 품사정보를 자질로 사용하고 있으며, 품사의 원형을 정답으로 사용한다. 이 말뭉치에서는 형태소 경계 구분자를 제거하고 'sp'를 어절구분자로 사용한다.

<표 3.5> 품사 복원 학습말뭉치의 예

자질1(어절)	자질2(축약품사)	정답(품사원형)
아름답	V	VAB
ㄴ	E	EEI
sp	SP	SP
추억	N	NNI
상자	N	NNI
를	P	FSD
sp	SP	SP
내	B	VBB
었	E	ERD
다고	E	EEA
sp	SP	SP
하	B	VBB
어도	E	EEG
sp	SP	SP
되	B	VBB
ㄴ다	E	EEA

제 4 장 기계학습을 이용한 음절기반 품사부착

본 논문은 기계학습을 이용한 음절기반 품사 부착 방법을 제안한다. 제안하는 시스템의 구조는 <그림 4.1>과 같다. 입력된 원시문장은 품사부착기에서 음절단위로 분리하고 기계학습을 통해 생성한 학습모델을 이용하여 음절단위로 품사가 부착된다. 음절단위로 품사가 부착된 문장은 형태론적으로 변형된 형태소의 원형을 Naïve Bayes 분류기를 이용해 음절의 원형을 복원한다. 그리고 음절의 원형이 복원된 문장은 형태소 복원 알고리즘을 이용해 형태소 단위로 복원하고 기계학습 시스템의 성능 향상을 위해 축약했던 품사의 원형을 복원하기 위해 품사 복원기를 사용한다. 품사 복원기는 기계학습으로 생성한 학습모델을 이용하여 품사의 원형을 복원한다.



<그림 4.1> 본 논문에서 제안한 음절기반 품사 부착 시스템의 흐름도

4.1. 음절품사 부착기

음절품사 부착기는 기계학습 모델을 이용하여 입력된 음절에 대한 품사를 결정하는 기능을 수행한다. 음절품사 부착기에 사용되는 기계학습 모델을 생성하기 위해서는 학습말뭉치가 필요하고 생성된 학습말뭉치는 자질추출 및 기계학습을 통해 학습모델을 생성한다. <그림 4.2>는 음절품사 부착기의 학습모델 생성 과정을 보여준다.

<그림 4.2> 음절품사 부착기의 학습모델 생성 과정

4.1.1. 음절품사 부착 학습말뭉치의 자질추출

기계학습 모델은 학습말뭉치로부터 정보를 학습하여 생성하는데 학습말뭉치의 어떤 정보를 학습할 것인지 정의(자질 추출)해야 한다. 이 기계학습 모델(음절품사 부착기)은 총 7가지(w_{i-2}, w_{i-1}, w_i, w_{i+1}, w_{i+2}, w_iw_{i-1}, w_iw_{i+1})의 자질을 사용한다. 여기서 w_i 는 음절정보를 의미하고 i는 현재 음절의 위치, ±는 전후 위치를 의미한다. <표 3.3>에서 음절품사에 대한 자질추출의 예는 <표 4.1>과 같다.

<표 4.1> 음절품사 부착 학습말뭉치의 자질추출 예

w_{i-2}	w_{i-1}	w_i	w_{i+1}	w_{i+2}	w_iw_{i-1}	w_iw_{i+1}
름	다	운	sp	추	운다	운sp
다	운	sp	추	억	sp운	sp추
운	sp	추	억	상	추sp	추억

4.1.2. 기계학습 모델

음절품사 부착기의 기계학습 모델은 CRF++(Ver. 0.64)[4]를 사용하여 생성한

4) http://crfpp.sourceforge.net

다. 이 학습도구는 품사부착 문제에 있어서 우수한 성능을 보인 연구결과(Sha *et al.*, 2003)가 있는 CRFs(Conditional Random Fields) 모델을 기반으로 동작하며 앞서 생성한 학습말뭉치와 자질집합을 이용한다.

4.2. 음절 복원기

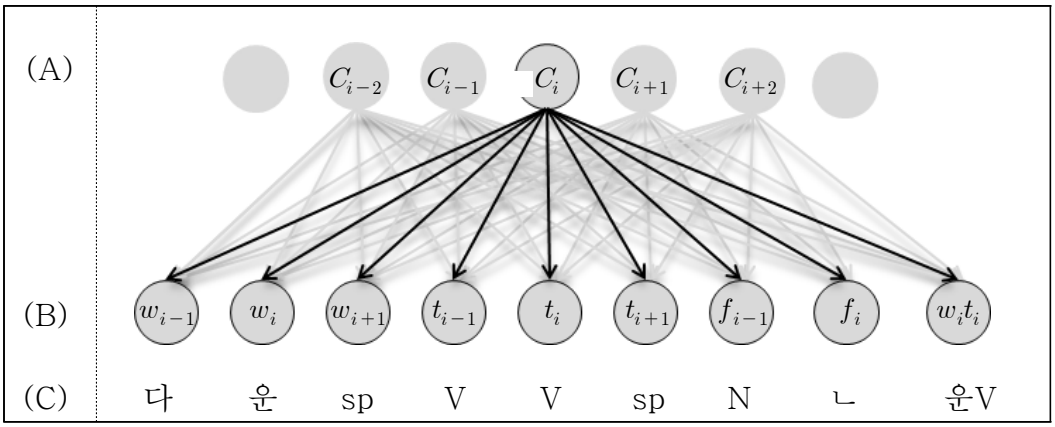
음절 복원기는 한국어의 형태론적 변형에 의해 음절이 변형된 경우, 그 음절의 원형을 복원하는 기능을 수행한다. 예를 들어 '아름다운' 은 "아름답+ㄴ" 으로 분석되고 '다', '운'의 원형인 '답', 'ㄴ'을 찾아서 복원하는 것이다. 본 논문에서는 형태소에 포함된 음절의 원형을 복원하기 위해 Naïve Bayes 분류기를 사용한다.

Naïve Bayes 분류기는 베이즈 정리에 기초하고 속성들 간의 독립성을 가정한 확률모델(Manning *et al.*, 2007)이다. 이 방법은 매우 단순한 방법이지만 자연어처리 분야에서 널리 알려진 전통적인 중의성 해소 방법이다.

Naïve Bayes 분류기의 분류식은 (식 4.1)과 같고 자질값이 주어졌을 때 가장 높은 값을 가지는 C_i 를 정답으로 결정한다. <그림 4.3>은 Naïve Bayes 분류기의 계산방법을 표현하였으며 (A)는 결정 가능한 정답후보 집합으로 가장 높은 값을 가지는 C_i 를 선택한다. (B)는 자질을 의미하며, 그 속성은 (C)에서 예로 표현하였다[5]. 각 자질이 출현했을 때 결정 가능한 후보 정답 중 가장 높은 확률을 가지는 C_i를 선택하는 것이 Naïve Bayes 분류기의 분류 방법이다.

$$\hat{C} = \arg max_{C_{i \in C}} P(C_i | w_{i-1},\ w_i,\ w_{i+1},\ t_{i-1},\ t_i,\ t_{i+1},\ f_{i-1},\ f_i,\ w_i t_i) \quad (4.1)$$

5) f는 음절의 받침정보를 의미



<그림 4.3> Naïve Bayes 분류기

Naïve Bayes 분류기는 <표 4.3>과 같이 학습말뭉치를 ARFF[6] 입력양식으로 가공하여 사용한다. '#RELATION'은 학습말뭉치의 이름을 의미하고, '#ATTRIBUTE'는 각 자질별 중복을 제거한 자질집합이다. '#DATA'는 미리 정의한 9가지 자질과 정답을 순차적으로 나열한 것이다.

<표 4.2> ARFF 입력양식의 학습말뭉치

```
#RELATION Naïve_Features
#ATTRIBUTE W1 {$, 아, 름, 다, 운, sp, 추, 억, 상, 자, 를, 냈, .....}
#ATTRIBUTE W2 {아, 름, 다, 운, sp, 추, 억, 상, 자, 를, 냈, .....}
#ATTRIBUTE W3 {$, 름, 다, 운, sp, 추, 억, 상, 자, 를, 냈, .....}
#ATTRIBUTE T1 {$, V, V+, E+, SP+, N, N+, E, B+E, .....}
#ATTRIBUTE T2 {V, V+, E+, SP+, N, N+, E, B+E, .....}
#ATTRIBUTE T3 {$, V, V+, E+, SP+, N, N+, E, B+E, .....}
#ATTRIBUTE F1 {N, ㅁ, ㄴ, sp, ㄱ, ㅇ, ㄹ, ㅆ, .....}
#ATTRIBUTE F2 {N, ㅁ, ㄴ, sp, ㄱ, ㅇ, ㄹ, ㅆ, .....}
#ATTRIBUTE B {아/V, 름/V, 다/V+, sp/SP+, 추/N, 억/N+, 냈/B+E+, .....}
#ATTRIBUTE C {아, 름, 답, ㄴ, sp, 추, 억, 상, 자, 를, 내+었, .....}
#DATA
$, 아, 름, V, $, V, $, N, 아/V, 아
아, 름, 다, V, V, V+, N, ㅁ, 름/V, 름
름, 다, 운, V, V+, E+, ㅁ, N, 다/V+, 답
다, 운, sp, V+, E+, SP+, N, ㄴ, 운/E+, ㄴ
                                    .
                                    .
                                    .
```

6) Atrribute-Relation File Format의 약어

4.3. 형태소 복원기

형태소 복원기은 음절단위로 품사가 부착된 문장의 음절을 형태소 단위로 복원하는 기능을 수행한다. (심광섭, 2011)에서는 연속된 품사를 하나의 품사로 바꾸고 동일한 품사를 가진 연속된 음절을 묶어서 하나의 형태소로 간주하였다. 이 방법은 복합명사를 정확하게 구분하지 못하는 문제가 있다. 이를 해결하기 위해 본 논문에서는 형태소 구분자 '+'를 사용하였다. 형태소 구분자는 품사열 사이에서 복합명사를 정확하게 구분할 뿐만 아니라 한 음절 내에 두 개 이상의 품사가 존재하는 경우에도 이를 정확하게 구분한다. <그림 4.4>는 음절품사가 부착된 문장을 형태소 단위로 복원하는 알고리즘이다. 이 알고리즘은 현재 음절에 부착된 음절품사에 형태소 구분자가 없다면 다음 음절은 현재 음절과 같은 어절이라는 사실을 이용한다. 알고리즘의 입력은 <그림 4.5>의 입력예시와 같이 음절과 음절품사의 집합으로 이루어져 있고, 출력예시와 같이 '+' 기호가 출현할 때까지의 음절을 하나의 형태소로 부착하여 출력한다.

```
def convert_syllable_to_morph(L)
  // L = (음절 s, 품사 t)의 집합
  // M = (음절열, 품사)의 집합
  for i = 0 to L.size()
    if not '+' in L[i].t
      m.append(L[i].s)
    else
      m.append(L[i].s)
      M.append(m, (L[i].t.pop('+')) )
      m.clear()
  return(M)
```

<그림 4.4> 형태소 복원 알고리즘

입력(L)		출력(M)
추, N 억, N+ 상, N 자, N+ 를, P+	→	추억, N 상자, N 를, P

<그림 4.5> 형태소 복원 알고리즘의 입출력 예

4.4. 품사 복원기

품사 복원기는 기계학습 모델의 성능 향상을 위해 축약하여 사용했던 품사의 원형을 복원하는 단계이다. 축약품사의 정보는 <부록 1>에서 확인 가능하다. 품사 복원기는 품사부착기와 유사한 기계학습 모델을 사용하며 있으며 기계학습 모델의 생성은 <그림 4.6>과 같다.

<그림 4.6> 품사 복원기의 학습모델 생성 과정

품사복원 학습말뭉치로부터 기계학습에 필요한 9가지 학습자질 m_{i-1}, m_i, m_{i+1}, t_{i-1}, t_i, t_{i+1}, $m_i t_i$, $o_{i-1}o_{i-2}$, $o_{i-1}t_{i-1}$을 사용하여 학습 모델을 생성하는데 m은 형태소, o는 품사원형을 의미하고 품사부착기와 동일한 기계학습 도구를 사용한다. 여기서 생성된 학습모델을 이용해 품사 복원이 완료되고 이 결과는 품사 태깅 결과로 간주한다.

제 5 장 실험 및 평가

본 논문의 학습 및 실험에 사용한 원시 말뭉치는 한국전자통신연구원 ETRI 구문구조 말뭉치(김재훈 외, 2005)이며 기계학습 도구는 CRF++을 사용한다. 또한 시스템의 성능은 각 단계별로 오류유형을 분석하고 이 유형에 대한 성능을 분석하여 차후 연구방향을 모색한다.

5.1. 기계학습 도구

품사부착기의 기계학습 모델은 CRF++(Ver. 0.64)를 사용하여 생성한다. 이 학습도구는 품사부착 문제에 있어서 우수한 성능을 보인 연구결과가 있는 CRFs 모델을 기반으로 동작하며 앞서 생성한 학습말뭉치와 자질집합을 이용하여 학습모델을 생성한다. 학습 모델이 생성되면 이 모델을 이용하여 입력된 문장에 대해 음절 단위로 품사가 부착된다. <그림 5.1>은 학습도구의 실행화면이다.

```
Number of sentences: 75467
Number of features:  13148646
Number of thread(s): 4
Freq:                1
eta:                 0.00010
C:                   1.00000
shrinking size:      20
iter=0 terr=0.99254 serr=1.00000 act=13148646 obj=26618895.64940 diff=1.00000
iter=1 terr=0.42052 serr=1.00000 act=13148646 obj=24480955.89200 diff=0.08032
iter=2 terr=0.42052 serr=1.00000 act=13148646 obj=17026015.18790 diff=0.30452
iter=3 terr=0.42574 serr=1.00000 act=13148646 obj=13554683.59103 diff=0.20388
iter=4 terr=0.35005 serr=1.00000 act=13148646 obj=9342633.14399 diff=0.31075
iter=5 terr=0.31513 serr=1.00000 act=13148646 obj=8283696.05713 diff=0.11334
iter=6 terr=0.27617 serr=0.99999 act=13148646 obj=7480153.54873 diff=0.09700
iter=7 terr=0.25073 serr=0.99999 act=13148646 obj=6447755.11366 diff=0.13802
iter=8 terr=0.22664 serr=0.99997 act=13148646 obj=5545697.54623 diff=0.13990
iter=9 terr=0.19967 serr=0.99996 act=13148646 obj=4775830.22056 diff=0.13882
```

<그림 5.1> 기계학습도구 CRF++ 실행화면

5.2. 성능 평가 척도

본 논문의 성능 평가 척도는 <표 5.1>과 같이 정확도(accuracy, Acc), 정확률(precision, P), 재현율(recall, R), F_1-척도(F_1-score, F_1)를 이용한다.

<표 5.1> 성능 평가 척도

정확도	$\frac{\text{시스템 결과의 정답 개수}}{\text{시스템 결과의 전체 개수}} \times 100$
정확률	$\frac{\text{시스템 결과의 정답 개수}}{\text{시스템 결과의 전체 개수}} \times 100$
재현율	$\frac{\text{시스템 결과의 정답 개수}}{\text{말뭉치의 분석정보 전체 개수}} \times 100$
F_1 **-척도**	$2\frac{P \times R}{P + R}$

5.3. 성능평가

5.3.1. 전체 시스템의 성능 평가

<표 5.2>는 전체 시스템의 성능평가를 나타낸다. 음절품사 부착기의 성능은 정확률과 재현율이 동일하기 때문에 정확도만 분석하였으며 성능은 97.3%이다. 다음으로 음절 복원기는 음절복원 실험말뭉치 전체의 성능과 실험말뭉치 중 적용음절(복합품사 음절과 그 이전 음절)에 대한 성능을 나누어서 평가하였다. 실험말뭉치 전체 음절수는 900,153 음절 이며 오류음절의 수는 1,499 음절로 성능평가 결과 정확률 99.8%, 재현율 99.9%, F_1 99.9%의 성능을 보이고 있으나 이는 적용되지 않는 음절의 수가 94%이상을 차지하기 때문에 정확한 성능평가로 간주할 수 없다. 이에 더욱 정확한 성능평가를 위해 실제 적용한 52,418 음절에 대한 성능평가를 하였으며 그 결과 정확률 97.2%, 재현율 98.0%, F_1 97.6%의 성능을 보였다. 끝으로 품사 복원기의 성능은 정확률 97.4%, 재현율 97.6%, F_1 97.5%의 성능을 보여 전체 시스템은 97.5%의 성능을 보였다.

<표 5.2> 전체 시스템의 성능 평가

시스템	실험 말뭉치		입력형태	시스템		Acc	P	R	F1
	종류	전체 대상수		전체 대상수	정답수				
음절품사 부착기	음절 품사 부착	900,153	음절단위 문장	861,011	838,045	97.3%	·	·	·
음 절 복원기	음절 복원 (전체음절)	900,153	음절품사 부착 결과	900,573	899,074	·	99.8%	99.9%	99.9%
	음절 복원 (적용음절)	52,418		52,838	51,339	·	97.2%	98.0%	97.6%
품 사 복원기	품사 복원	434,626	음절 복원 결과	435,509	424,064	·	97.4%	97.6%	97.5%

5.3.2. 각 시스템별 성능 평가

전체 시스템은 <표 5.2>와 같이 97.5%의 성능(F_1)을 보이지만 각 단계의 출력값에는 평균 2.53%의 오류를 가지고 있어 <표 5.2>에 표시한 각 단계의 성능은 정확한 수치라고 할 수 없다. 이에 보다 정확한 성능 파악을 위해 각 단계의 입력 값으로 오류가 없는 실험말뭉치를 사용하여 평가를 한다. <표 5.3>은 입력값으로 오류가 없는 실험말뭉치를 사용하여 각 단계의 성능을 평가한 결과이다. 음절품사 부착기는 오류가 없는 일반문장으로 입력으로 사용하기 때문에 성능이 동일하다. 음절 복원기는 오류가 없는 실험말뭉치를 입력으로 사용한 경우 98.1%의 정확률과 98.8%의 재현율을 보였으며 98.4%의 F_1을 보인다. 또한 품사 복원기의 성능에서도 실험말뭉치를 사용한 경우 정확률 98.4%, 재현율 98.9%, F_1 98.7%의 성능을 보였다. 이는 <표 5.2>의 결과와 비교할 때 성능이 향상되었다.

<표 5.3> 각 단계의 시스템 성능 평가

시스템	실험 말뭉치		입력형태	시스템		Acc	P	R	F1
	종류	전체 대상수		전체 대상수	정답수				
음절품사 부착기	음절 품사 부착	900,153	음절단위 문장	861,011	838,045	97.3%	·	·	·
음 절 복원기	음절 복원 (전체음절)	900,153	음절품사 부착	900,547	898,604	·	99.8%	99.9%	99.9%
	음절 복원 (적용음절)	52,418	실험말뭉치 (오류없음)	52,812	51,809	·	98.1%	98.8%	98.4%
품 사 복원기	품사 복원	434,626	음절 복원 결과 (오류없음)	437,042	429,936	·	98.4%	98.9%	98.7%

1) 음절품사 부착기의 성능

음절품사 부착기의 성능은 실험말뭉치의 전체 음절 개수 시스템 결과의 전체 음절 개수가 동일하여 정확도만 측정하여 성능을 평가한다. 음절품사 부착기의 성능은 97.3%이다.

2) 음절 복원기의 성능

음절 복원기의 성능은 음절품사 부착 결과가 정확한 실험말뭉치를 음절 복원기의 입력데이터로 사용한 경우와 품사 부착기의 결과를 음절 복원기의 입력데이터

를 사용한 경우로 나누어 정확률과 재현율을 평가하였다. 성능평가에 사용된 실험데이터 861,011 음절(9,827 문장) 중 실제 Naïve Bayes 분류기에 의해 복원되는 음절은 52,812 음절이다. 이는 형태소의 경계지점에서 음운변화 현상이 일어나는 점을 고려하였으며, 형태소의 경계지점 중에서도 복합품사[7]가 부착된 음절과 이 음절의 앞 음절에서 음운변화 현상이 나타나기 때문에 복합품사가 부착된 음절과 이 음절의 앞 음절에 대해서만 음절 복원을 하기 때문이다.

실험말뭉치를 사용한 시스템의 성능은 정확한 입력데이터를 사용할 때의 시스템 성능을 파악하기 위한 것으로 음절 복원기의 자체 성능을 판단하는 기준이 된다. <표 5.4>와 같이 실험 결과 입력 데이터 전체 861,011 음절에 대한 정확률은 99.9%의 성능을 보이지만 실제로 음절 복원이 필요한 52,812 음절은 입력데이터 전체의 6% 밖에 되지 않고 오류 음절은 1,003 음절에 불과하기 때문이다. 본 실험에서는 보다 정확한 성능평가를 위해 음절 복원을 해야 하는 6% 음절에 대해서만 성능평가를 하였다. 그 결과 98.1%의 정확률과 98.8%의 재현율을 보였으며 98.4%의 F_1을 보였다.

<표 5.4> 전체성능 및 실제성능 차이

구분	입력 음절 개수	정답 음절 개수	오답개수	Acc
전체 음절 입력	900,153	899,150	1,003	99.9%
적용 가능 음절 입력	52,812	51,809	1,003	98.1%

반면 음절품사 부착기의 결과에는 일부 오류가 포함되어 있어 실험말뭉치를 사용한 시스템의 성능보다 낮은 성능을 보인다. 하지만 실제 문장을 입력한 경우 품사부착기의 결과를 사용하여 음절 복원을 하는 것이 시스템의 실제 성능이므로 이와 같은 실험을 통해 시스템 전체의 성능을 판단한다. 실험 결과 97.2%의 정확률과 98.0%의 재현율을 보였으며 97.6%의 F_1을 보였다.

3) 품사 복원기의 성능

품사 복원기의 성능은 음절 복원기와 동일하게 실험말뭉치를 품사 복원기의 입력으로 사용한 시스템의 성능 분석과 음절 복원기의 결과를 품사 복원기의 입력으로 사용한 시스템의 성능 분석으로 나누어 성능을 비교한다.

---

7) 두개 이상의 품사가 한 음절에 부착된 경우를 지칭.(ex. 냈 / V+E)

실험말뭉치를 사용한 시스템의 성능은 98.4%의 정확률과 98.9%의 재현율을 보였고 98.7%의 F_1을 보였다. 반면 음절 복원 결과를 사용한 시스템의 성능은 96.4%의 정확률과 97.6%의 재현율을 보였으며 97.0%의 F_1을 보여 음절 복원기와 동일하게 실험말뭉치를 사용한 시스템의 성능이 높게 나왔다.

5.4. 오류분석

5.4.1. 음절품사 부착결과의 오류분석

음절품사 부착결과의 오류분석은 혼돈행렬을 생성하여 분석한다. 분석결과는 <표 5.5>와 같으며 경계구분 오류와 품사부착 오류를 구분하여 분석하였다. 첫 번째 열은 정답품사이고 첫 번째 행은 시스템의 결과이다. 또한 etc는 복합품사를 의미하는데 이는 한 음절이 두 가지 이상의 품사를 포함하는 경우로 형태소의 원형 복원 단계에서 분리된다.

이 혼돈행렬은 대각선을 중심으로 대칭적인 구조를 가지고 있으며 경계구분 오류와 품사부착오류의 발생빈도를 분석할 수 있다.

<표 5.5> 품사부착기의 혼돈행렬

시스템 / 말뭉치	N	N+	P	P+	Y	Y+	U	U+	F	F+	B	B+	V	V+	A	A+	C	C+	D	D+	E	E+	S	S+	I	I+	etc
N		1435	10				83	70	4	9	573	128	310	20	838	2		3	956		17	11	13	6	7		232
N+	2000		26	122			2	101	4	159	77	97	34	11	50	156		7	334	348	29	246	2	15		9	1332
P	9	35		2				2	4	2	2	3	2					60	144		215	2			2		10
P+		277						2		18						2			2	683	2	810					90
Y																					2						
Y+																					2						
U	58							21			4	6			26				7					13			
U+	41	44		7			27			5	4				4	31			4	4		13					250
F	2																										
F+	4	139		9				6	2		2					8				9	2	9					31
B	494	67						2		3		70	500		12				218		14	4			2		512
B+	115	100	20	19			13	2		2	104		4	318	20	6		7	109	5	44	33		2		2	1425
V	152	16									231	2		2	165				466	2		2					18
V+	14	3	6								5	141			16				139		3				2		46
A	503	7					19	2			38		195	2					25								2
A+		257						43		19				2						19						3	313
C																											
C+	8	6	339	22							2	12		3					2		46	260			2		7
D	703	167	90				8	3			159	88	320	301	9						57	3			8		63
D+	362			306												13						669				6	71
E	17	35	373	8							13	11		2		34		3	43			125					22
E+		167		1228				8			2	5		3	34			2		519	72			37		3	161
S	9																										
S+	10	47					11															8					45
I													2						2								
I+																				2							
etc	6	148		19						2	7	12	2			151			11	8	39	219					799

경계구분 오류는 품사는 동일하지만 형태소 경계 구분자(+)의 유무와 관련한 오류로 전체 음절오류(22,966 음절) 중 17.2%(3,952 음절)가 경계구분 오류로 분석되었다. 또한 품사부착 오류는 다른 품사가 부착되는 오류로 경계구분 오류와 중복되는 경우도 품사부착 오류로 간주하였다. 품사부착 오류는 82.8%(19,014 음절)를 차지한다.

5.4.2. 음절 복원 결과의 오류분석

음절 복원 결과의 오류는 다양한 오류를 포함하고 있다. 예를 들어 '했'의 경우 '하'와 '었'으로 복원되는 것이 정확하지만 '하'와 '았'으로 복원되어 오류로 확인된다. 또한 '슬기로운'에서 '로' 는 '롭'으로 정확히 복원되지만 '운'은 'ㄴ'이 아닌 '은'으로 복원되어 오류로 나타났다. 이 외에도 또한 '정신적인'에서 '인'은 '이'와 'ㄴ'으로 복원되어야 하지만 '은'으로 분리된 경우도 있었다. 이는 Naïve Bayes 분

류기 생성과정에서 발생한 학습오류로 판단된다.

5.4.3. 음절품사 복원결과의 오류분석

음절품사 복원결과의 오류는 종류가 다양하여 <표 5.6>과 같이 오류빈도에 대해 표시하였다. 음절품사 복원기는 동일한 분류에 속하지만 품사가 세분화 되면서 해당품사를 정확하게 결정하지 못한 경우가 대부분이며 특히 명사류에서 발생한 오류가 전체 오류의 71%를 차지하였다. 이는 입력 데이터에 명사류의 품사가 많기 때문으로 판단된다. 몇몇 다른 명사가 감탄사로 결정되는 경우도 있었으나 이는 전체 오류의 0.3%에 불과하다.

<표 5.6> 품사복원의 오류 빈도

오류품사	빈도	오류품사	빈도
D	721	U	186
A	70	P	885
E	760	S	306
F	62	V	32
N	7,490	B	6

<표 5.7>과 <표 5.8>은 품사복원 단계의 오류 중 'E'품사와 'D' 품사에 대한 혼돈행렬이다. 두 혼돈행렬에서도 알 수 있듯이 전체적인 구조는 대칭을 이루고 있으며 품사복원 오류의 유형을 쉽게 파악할 수 있다.

<표 5.7> 'E' 품사의 혼돈행렬

시스템 / 말뭉치	EEA	EEB	EEC	EED	EEE	EEF	EEG	EEI	EEJ	EEK	ERA	ERB	ERD	ERE	INA
EEA		13	19	16	10		25			1					
EEB			1				1								
EEC	2	7					7								2
EED	1						1								
EEE	1						1								
EEF			4				307		2						
EEG	26	3	15	9	2	55		54	33	25			2		3
EEI							7								
EEJ							24			1					
EEK	4	1	5	2			59								
ERA															1
ERB															
ERD											2	3		1	2
ERE															
INA															

<표 5.8> 'D' 품사의 혼돈행렬

시스템 / 말뭉치	ADA	ADB	ADC	ADD	ADE	ADF	ADG	ADH	ADI	INA
ADA		50			4			22		
ADB	76		22	103	46	101		48	51	
ADC										
ADD										
ADE	2					1				
ADF	1	2	3	1					3	
ADG										3
ADH	6	71			2	4			10	
ADI	7	27	1	3	5	13	2	29		2
INA										

제 6 장 결론 및 향후 연구과제

본 논문에서는 기계학습을 이용한 품사부착 방법을 제안하였다. 일반적으로 한국어 형태소 분석을 위해서는 형태소 분석기나 품사 부착 시스템을 사용하는 것이 대부분이지만 이들 시스템은 매우 복잡한 구조를 가지고 있다. 뿐만 아니라 구현을 위해서는 복잡한 지식과 방대한 사전 정보가 요구된다. 이러한 문제점을 해결하기 위해 규칙을 사용한 기존 연구가 있었으나 이 또한 복합명사의 분석 불가능 문제, 규칙의 모호성 등의 문제를 가지고 있었다. 이를 해결하기 위해 본 논문에서는 기계학습을 통해 품사 부착이 가능한 간단한 구조의 품사부착 방법을 제안하였다. 본 시스템은 구현이 간단하며 빠른 시간 내에 제작이 가능하고 형태소 분석을 하는 과정에 음절 단위로 품사를 부착하기 때문에 단어 분리 시스템으로 사용이 가능하다. 또한 한국어 관련 연구가 활발히 진행되면서 각종 말뭉치를 어렵지 않게 구할 수 있고, 공개된 기계학습 도구가 많기 때문에 누구라도 간단한 지식만으로 구현이 가능하다. 뿐만 아니라 간단한 구조로 설계된 시스템임에도 불구하고 복잡한 구조를 가진 기존 형태소 분석기나 품사부착 시스템과 대등한 성능을 얻을 수 있었다. 그리고 음절품사 부착, 음절과 품사의 원형 복원 단계, 형태소 복원 단계를 거치면서 여러 분야에서 활용이 가능하여 광범위한 사용영역을 가지고 있다. 물론 개인이 필요로 하는 시스템에 맞추기 위해 추가적으로 결과 값을 가공해야 하는 단점이 있지만 여러 분야에 접목하여 사용가능하다는 장점으로 생각할 수 있다.

반면 본 논문에서는 한 가지 종류의 말뭉치만으로 시스템을 구축하여 성능을 평가하였기에 특정 분야에 치우칠 가능성을 가지고 있다. 이를 해결하기 위해 향후 수집 가능한 많은 종류의 말뭉치를 이용하여 시스템을 구축하고 성능차이를 평가하여 성능을 높일 수 있는 방법을 모색한다. 뿐만 아니라 여러 종류의 실험을 통해 시스템의 각 단계에서 나타날 수 있는 오류문제를 해결하여 현 시스템의 성능을 더 끌어 올릴 수 있도록 한다.

참고문헌

Dale, R., Moisl, H., and Somers, H. (2000), Handbook of Natural Language Processing, Marcel Dekker, Inc.

Furnkranz, J. and Widmer, G. (1994), "Incremental Reduced Error Pruning", Proceedings of International Conference on Machine Learning, pp. 70-77.

Hindle, D. (1989), "Acquiring Disambiguation Rules from Text", Proc. of 27th Annual Meeting of the Association for Computational Linguistics, pp. 118-125.

Jurafsky, D., and Martin, J.H. (2009), Speech and Language Processing, Pearson Education Inc.

Kim, D.-B., Lee, S.-J., Choi, K-S., and Kim, G.-C. (1994), "A Two-level Morphological Analysis of Korean", Proceedings of Computational Linguistics, Vol. 1, pp. 535-539.

Koskenniemi, K. (1983), "Two-level Model for Morphological Analysis", Proceedings of International Joint Conference on Artificial Intelligence, pp. 683-685.

Kwon, H.-C. and Karttunen, L. (1994), "Incremental Construction of a Lexical Transducer for Korean", Proceedings of Computational Linguistics, Vol. 2, pp. 1262-1266.

Manning, C.D., Raghavan, P., Schutze, H. (2007), Introduction to Information Retrieval, Cambridge Univ Pr.

Sha, F. and Pereira, F. (2003), "Shallow Parsing with Conditional Random Fields", Proceedings of Human Language Technology Conference - North American Chapter of the Association for Computational Linguistics Annual Meetings, pp. 134-141.

김성용, 최기선, 김길창 (1987), “Tabular Parsing 방법과 접속 정보를 이용한 한국어 형태소 분석기”, 한국정보과학회 인공지능연구회 춘계 인공지능 학술발표회 발표논문집, pp. 133-147.

김재훈 (1996), “가중치 망을 이용한 한국어 품사 태깅”, 정보과학회논문지(B), 제25권, 제6호, pp. 951-959.

김재훈, 이공주 (2003), “사례기반 학습을 이용한 음절기반 한국어 단어 분리 및 범주 결정”, 정보처리학회 논문지(B), 제10-B권, 제1호, pp. 47-56.

김재훈 외 (2005), 구문구조 부착 말뭉치 구축, 모비코앤시스메타㈜, 최종보고서.

박성배, 장병탁 (2003), “음절 정보만 이용한 한국어 복합명사 분해”, 제15회 한글 및 한국어 정보처리 학술대회 발표논문집, pp. 33-39.

서정주 (1996), 현대 국어문법론, 한양대학교 출판원.

서충원, 오종훈, 최기선 (2003), “어절의 중심어 정보를 이용한 한국어 기반 명사구 인식”, 제15회 한글 및 한국어 정보처리 학술대회, pp. 145-151.

심광섭 (1997), “합성된 상호 정보를 이용한 복합명사 분리”, 정보과학회논문지(B), 제24권, 제11호, pp. 1307-1317.

심광섭 (2011), “형태소 분석기 사용을 배제한 음절 단위의 한국어 품사 태깅”, 인지과학, 제22권, 제3호, pp. 327-345.

이근용, 박기선, 이용석 (2004), “Two-level 한국어 형태소 해석에서의 복합명사처리”, 한국정보과학회 학술발표논문집, 제29권, 제1호(B), pp. 505-507.

이기오, 이근용, 이용석 (1996), “효율적인 한국어 분석을 위한 확장된 최장일치법”, 제8회 한글 및 한국어 정보처리 학술대회 논문집, pp. 255-261.

임해창, 임희석, 이상주, 김진동 (1996), “자연어 처리를 위한 품사 태깅 시스템의 고찰”, 정보과학회지, 제4권, 제7호, pp. 36-57.

임희석, 김진동, 임해창 (1997), “어절 태그 변형규칙을 이용한 한국어 품사 태거” 정보과학회 논문지(B), 제24권, 제6호, pp. 673-684

조영환, 김덕봉, 최기선, 김길창 (1990), “한글 맞춤법 오류 검사 및 교정시스템”, 정보과학회 학술발표논문집, 제17권, 제1호, pp. 195-198.

최재혁, 이상조 (1993), “양방향 최장일치법을 이용한 한국어 형태소 분석기”, 한국정보과학회 학술발표논문집, 제20권, 제1호, pp. 769-772.

최형석, 이주근 (1984), “자연어 어절의 처리 앨고리즘”, 한국정보과학회 추계 학술발표회 발표논문집, 제11권, 제2호, pp. 36-43.

홍진표, 차정원 (2008), “어절 패턴 사전을 이용한 새로운 한국어 형태소 분석기”, 한국컴퓨터종합학술대회 논문집, 제35-C권, 제1호, pp. 279-284.

부 록

<부록 1> 원시 말뭉치의 한글/영문 태그/사용예제

한글 태그	영문 태그	축약 태그	형태소 경계	사용 예제
인명고유명사	NNA	N	N+	한스[**인명고유명사**]+는[일반보조사]
기관고유명사	NNB			통계청[**기관고유명사**]+에서는[부사격조사]
지명고유명사	NNC			한반도[**지명고유명사**]+는[일반보조사]
서적고유명사	NND			삼강행실도[**서적고유명사**]+는[일반보조사]
사건고유명사	NNJ			2차대전[**사건고유명사**]+을[목적격조사]
기타고유명사	NNE			EU협정[**기타고유명사**]
외국어	NNK			SI[**외국어**]+시장[용언불가능보통명사]
용언불가능보통명사	NNF			규모[**용언불가능보통명사**]+가[주격조사]
용언가능보통명사	NNI			정보화[**용언가능보통명사**]+에[부사격조사]
단위성의존명사	NDA			4[숫자수사]+명[**단위성의존명사**]+의[관형격조사]
기타의존명사	NDF			것[**기타의존명사**]+이[보격조사]
인칭대명사	NPA			당신[**인칭대명사**]+의[관형격조사]
지시대명사	NPB			그[**지시대명사**]+들[복수접미사]+을[목적격조사]
주격조사	POA	P	P+	우리[인칭대명사]+가[**주격조사**]
목적격조사	POB			용어[용언불가능보통명사]+를[**목적격조사**]
부사격조사	POC			것[기타의존명사]+으로[**부사격조사**]
보격조사	POD			해방[용언가능보통명사]+이[**보격조사**]
관형격조사	POE			남자[용언불가능보통명사]+의[**관형격조사**]
호격조사	POH			정숙[인명고유명사]+아[**호격조사**]
접속조사	POJ			때[용언불가능보통명사]+와[**접속조사**]
일반보조사	POK			육신[용언불가능보통명사]+은[**일반보조사**]
사동보조용언	EVE	Y	Y+	부담지[일반동사]+우[**사동보조용언**]+게[종속연결어미]
피동보조용언	EVF			채우[일반동사]+어지[**피동보조용언**]+ㄴ[관형사형전성어미]
기타보조용언	EVK			고민하[일반동사]+게하[**기타보조용언**]+ㄴ[관형사형전성어미]
서수사	NUA	U	U+	일곱[**서수사**]
양수사	NUB			3[숫자수사]+천[**양수사**]
숫자수사	NUC			3[**숫자수사**]+천[양수사]
접두사	FPA	F	F+	제[**접두사**]+2[숫자수사]+차[단위성의존명사]
인명접미사	FSA			도미[인명고유명사]+네[**인명접미사**]+뿐만[일반보조사
지명접미사	FSB			·
보통명사형접미사	FSC			·
복수접미사	FSD			사람[용언불가능보통명사]+들[**복수접미사**]+이[주격조사]

<부록 1> 원시 말뭉치의 한글/영문 태그/사용예제(계속)

한 글 태 그	영 문 태 그	축 약 태 그	형태소 경계	사용 예제
숫자형접미사	FSE	F	F+	30[숫자수사]+여[**숫자형접미사**]
기타접미사	FSF			40[숫자수사]+년[단위성의존명사]+간[**기타접미사**]
지시동사	VBA	B	B+	그러[**지시동사**]+기[명사형전성어미]
일반동사	VBB			위하[**일반동사**]+어서[종속연결어미]
지시형용사	VAA	V	V+	아니[**지시형용사**]+라[종속연결어미]
성상형용사	VAB			있[**성상형용사**]+는[관형사형전성어미]
성상관형사	ANA	A	A+	주도적[**성상관형사**]
지시관형사	ANB			이런[**지시관형사**]
수관형사	ANC			여러[**수관형사**]
능동전성사	VFA	C	C+	·
수동전성사	VFB			·
사동전성사	VFC			·
긍정지정사	VFD			것[기타의존명사]+이[**긍정지정사**]+라고[종속연결어미]
성상정도부사	ADA	D	D+	대단히[**성상정도부사**]
성상상태부사	ADB			열심히[**성상상태부사**]
성상의성부사	ADC			부르르[**성상의태부사**]
성상의태부사	ADD			텅텅[**성상의태부사**]
지시처소부사	ADE			쿵쿵[**성상의성부사**]
지시시간부사	ADF			이제는[**지시시간부사**]
부정부사	ADG			아니[**부정부사**]
문장양태부사	ADH			오히려[**문장양태부사**]
문장접속부사	ADI			하지만[**문장접속부사**]
종속연결어미	EEG	E	E+	고통스럽[성상형용사]+지[**종속연결어미**]
관형사형전성어미	EEI			같[성상형용사]+는[**관형사형전성어미**]
부사형전성어미	EEJ			잦[성상형용사]+게[**부사형전성어미**]
명사형전성어미	EEK			매[일반동사]+기도[**명사형전성어미**]
의문형종결어미	EEC			되[일반동사]+는가[**의문형종결어미**]
명령형종결어미	EED			생각하[일반동사]+오[**명령형종결어미**]
청유형종결어미	EEE			하[일반동사]+자[**청유형종결어미**]+.[문미기호]
대등연결어미	EEF			얻[일반동사]+고[**대등연결어미**]
높임선어말어미	ERA			주[일반동사]+시[**높임선어말어미**]+오[청유형종결어미]
공손선어말어미	ERB			받[일반동사]+자오[**공손선어말어미**]+아[종속연결어미
현재시제선어말어미	ERC			·
과거시제선어말어미	ERD			하[일반동사]+었[**과거시제선어말어미**]+다[평서형종결어미]
미래시제선어말어미	ERE			살[일반동사]+겠[**미래시제선어말어미**]+다면[종속연결어미]

<부록 1> 원시 말뭉치의 한글/영문 태그/사용예제(계속)

한 글 태 그	영 문 태 그	축 약 태 그	형태소 경계	사용 예제
사동선어말어미	ERH	E	E+	·
피동선어말어미	ERI			·
평서형종결어미	EEA			가[일반동사]+ㄴ다[**평서형종결어미**]
감탄형종결어미	EEB			없[성상형용사]+구려[**감탄형종결어미**]+![문미기호]
문미기호	SYA	S	S+	없[성상형용사]+구려[감탄형종결어미]+![**문미기호**]
왼열림기호	SYB			'[**왼열림기호**]+세계관적[성상관형사]+'[오른열림기호]
오른열림기호	SYC			'[왼열림기호]+세계관적[성상관형사]+'[**오른열림기호**]
컴마기호	SYD			권력[용언불가능보통명사]+,[**컴마기호**]
기타기호	SYE			한[지명고유명사]+.[**기타기호**]+미[지명고유명사]
단위기호	SYH			15[숫자수사]+m[**단위기호**]+로[부사격조사]
빈칸	SYI			·
어절구분자	SYJ			한[지명고유명사]+ - [**어절구분자**]+미[지명고유명사]
감탄사	INA	I	I+	얼씨구나[**감탄사**]