



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

공학석사 학위논문

# 한국어 서답형 자동채점에 적용하기 위한 형태소 분석 및 품사 부착

Morphological Analysis and Part-of-Speech Tagging  
for Applying Korean Automated Scoring of Short-Answer Questions



2016년 2월

한국해양대학교 대학원

컴퓨터공학과

천민아

본 논문을 천민아의 공학석사 학위논문으로 인준함.

위원장 박 휴 찬



위원 김 재 훈



위원 이 장 세



2015년 12월 23일

한국해양대학교 대학원

# 목 차

<b>List of Tables</b> .....	iv
<b>List of Figures</b> .....	v
<b>Abstract</b> .....	vii
<b>제 1 장 서 론</b> .....	<b>1</b>
<b>제 2 장 관련 연구</b> .....	<b>4</b>
2.1 한국어의 특성 .....	5
2.2 한국어 형태소 분석 기법 .....	9
2.3 한국어 형태소 품사 부착 .....	12
<b>제 3 장 한국어 서답형 자동채점 시스템의     형태소 분석 및 품사 부착 기법의 문제점 분석</b> .....	<b>16</b>
3.1 한국어 서답형 문항 자동채점 시스템 .....	17
3.2 기존의 형태소 분석 및 품사 부착 기법 .....	20
3.3 기존 형태소 분석기 및 품사기의 문제점 .....	23
<b>제 4 장 단어 분리와 사전 탐색 기법을 이용한     형태소 분석 및 품사 부착</b> .....	<b>25</b>
4.1 제안하는 형태소 분석 및 품사 부착 기법의 구조 .....	25
4.2 단어 분리 모델 및 사전 생성 .....	27
4.3 음절 기반의 단어 분리 .....	38
4.4 제안하는 형태소 분석 기법 .....	40
4.5 통계기반의 품사 부착 .....	51

제 5 장 실험 및 평가 .....	53
5.1 성능 평가 대상 .....	53
5.1.1 세종 말뚝치 .....	53
5.1.2 2014년 국가수준 학업성취도 평가 답안 .....	54
5.2 성능 평가 척도 .....	54
5.3 성능 평가 결과 .....	55
5.3.1 세종 말뚝치의 형태소 분석 및 품사 부착 결과 .....	55
5.3.2 2014년 국가 수준 학업성취도 평가 형태소 분석 및 품사 부착 결과 .....	57
5.4 오류분석 .....	58
제 6 장 결론 및 향후 연구 .....	60
참고문헌 .....	62
감사의 글 .....	66
부록 A 세종 말뚝치 품사 및 단순화 태그 .....	67

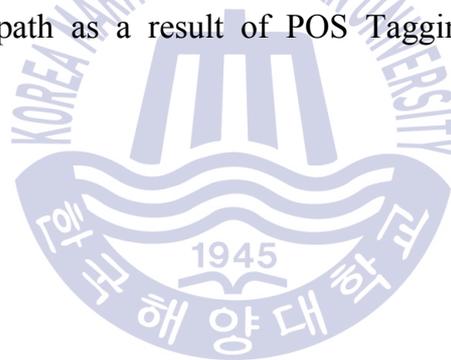
## List of Tables

<b>Table 2.1</b>	An Example of Morphological Analysis and POS Tagging .....	4
<b>Table 2.2</b>	Morphological Alternation of Korean .....	5
<b>Table 2.3</b>	Part of Speech in Korean .....	7
<b>Table 2.4</b>	Classification of Morphological Analysis in Korean .....	9
<b>Table 3.1</b>	Examples of dictionary structure of irregular verbs .....	21
<b>Table 4.1</b>	Classification of integration of morphological analysis and POS tagging in Korean .....	25
<b>Table 4.2</b>	The structure of the SEJONG corpus .....	27
<b>Table 4.3</b>	Statistics of the SEJONG corpus .....	28
<b>Table 4.4</b>	The Result of applying Fig. 3.11 algorithm to “너를 중요시해”	34
<b>Table 4.5</b>	The Results of executing the Fig. 3.11a algorithm for a given “중요시해” .....	36
<b>Table 4.6</b>	Meaning of aligning results .....	36
<b>Table 4.7</b>	The Results of executing the Fig. 3.11a algorithm for a given “중요시해” .....	37
<b>Table 4.8</b>	An example of syllable-based spacing .....	38
<b>Table 4.9</b>	The feature set for morpheme segmentation based on syllables ..	39
<b>Table 4.10</b>	An example of syllable-based word segmentation .....	40
<b>Table 5.1</b>	Statistics of the SEJONG corpus for performance evaluation .....	53
<b>Table 5.2</b>	Statistics of the 2014 NLSA for performance evaluation .....	54
<b>Table 5.3</b>	Measures for performance evaluation .....	54
<b>Table 5.4</b>	Recall of morphological analysis .....	55
<b>Table 5.5</b>	Accuracy of POS Tagging .....	56
<b>Table 5.6</b>	Performance of morphological analysis and POS tagging .....	57

## List of Figures

<b>Fig. 2.1</b> An example of Korean Part-of-Speech tagging using a lattice structure .....	13
<b>Fig. 3.1</b> A graphic user Interface of the Korean Automated Scoring System (노은희, 2014) .....	17
<b>Fig. 3.2</b> The overall structure of the Korean Automated Scoring System .....	18
<b>Fig. 3.3</b> Pseudocode for the scoring step using the self-training algorithm .....	19
<b>Fig. 3.4</b> Flow diagram for the previous morphological analysis and POS tagging .....	20
<b>Fig. 3.5</b> Pseudocode for the previous morphological analysis .....	22
<b>Fig. 3.6</b> An example of problematic student's answers .....	23
<b>Fig. 4.1</b> Flow diagram for the proposed method for morphological analysis .....	26
<b>Fig. 4.2</b> The creation process of required models and dictionaries .....	29
<b>Fig. 4.3</b> Pseudocode for extracting pairs (word, its morphological structure) by sentence .....	30
<b>Fig. 4.4</b> Alignment of syllables and a result of morphological analysis .....	32
<b>Fig. 4.5</b> Pseudocode for extracting pairs (syllable, morphological structure) by sentence .....	33
<b>Fig. 4.6</b> Python code for aligning (word, morpheme, and POS tag) using the SequenceMatcher function in difflib .....	35
<b>Fig. 4.7</b> The Results of executing the Fig. 3.11a algorithm for a given “너를 중요시해” .....	35
<b>Fig. 4.8</b> The proposed method for morphological analysis .....	40
<b>Fig. 4.9</b> The lattice structure as a result of morphological analysis .....	41
<b>Fig. 4.10</b> Adding ‘ /_SP_ ’ nodes in the lattice structure .....	42
<b>Fig. 4.11</b> Adding nodes in the lattice structure due to lookup of the pre-analyzed dictionary .....	43

<b>Fig. 4.12</b> Adding nodes in the lattice structure due to lookup of the morphological dictionary .....	45
<b>Fig. 4.13</b> Adding unknown word ( ‘word / _UK_’ ) nodes in the lattice structure .....	46
<b>Fig. 4.14</b> Adding nodes in the lattice structure due to lookup of the variant dictionary .....	47
<b>Fig. 4.15</b> Adding the first syllable and the last syllable of variant .....	48
<b>Fig. 4.16</b> The combining method of words and syllables of variants .....	49
<b>Fig. 4.17</b> Adding nodes in the lattice structure due to lookup of the morphological dictionary for combined words .....	50
<b>Fig. 4.18</b> Adding the BOS (first) node and the EOS (last) node .....	50
<b>Fig. 4.19</b> Loading weights on the edges of weighted graph .....	51
<b>Fig. 4.20</b> The shortest path as a result of POS Tagging .....	52





the BIO coding scheme. The versatile searching for morphological variants comprises four steps: The first and second steps are to look up segmented words in the pre-analyzed dictionary and morphological dictionary, respectively. For unknown words, the third step is to search for the segmented word in the variant dictionary and to concatenate the variant words with the previous words and the next words. The final step is to look up the combined words in the morphological dictionary. At each step, words in the dictionary are added into nodes on the lattice structure  $G$ , which is used for POS tagging and a weighted graph. The POS tagging is the best (shortest) path, *i.e.*, the most proper sequence for a given sentence, from the beginning node to the last node on the weighted graph.

The proposed morphological analyzer and POS tagger has demonstrated the recall and of 98.86% and the precision of 95.03% for the SEJONG corpus, and also can analyze all answers of subjects taken the 2014 National Level Student Assessment. Thus it can be said that the proposed systems are more effective than the morphological analyzer and POS tagger used for the automated scoring system of Koran short-answer questions.

KEY WORDS: Morphological Analysis; Part-of-Speech Tagging; Automated Scoring System; Short-Answer Questions Scoring

## 제 1 장 서 론

우리나라 교육제도에서 평가는 모든 학생이 교육 목표를 성공적으로 달성했는지 판단하기 위해 시행되고 있다. 평가는 선택형(객관식) 문항과 서답형 문항의 비율을 적절히 섞어 학생들의 종합적인 사고력과 창의력을 높일 수 있도록 권장하고 있다(교육과학기술부, 2009). 그러나 국가수준 학업성취도 평가와 대학수학능력시험과 같은 대규모 평가에서는 대부분 선택형 문항으로 사용하고 있다. 선택형 문항은 채점의 일관성을 유지하기에 쉽고 채점 시간이 매우 짧다는 장점이 있으나, 학생들의 종합적인 사고 능력을 측정하는데 어려움이 있다(이양락 외, 2010; 진경애, 2007). 이와 같은 선택형 문항의 단점을 보완하는 방안으로 서답형 문항이 확대되고 있다.

서답형 문항은 선택형 문항보다 학생의 종합적인 사고 능력 및 문제 해결 능력을 측정하는 데 효과적이거나, 채점자의 일관성과 신뢰성을 확보하기 어려우며 채점에 걸리는 시간과 비용 등의 문제가 있다(Chen *et al.*, 2010; Dikli, 2006; 노은희 외, 2014). 이러한 문제를 완화하기 위해서 해외에서는 다양한 형태의 자동채점 시스템이 활용되고 있다. 영어의 경우에는 ETS(Educational Testing Service)를 비롯한 많은 연구 기관에서 활발하게 연구(Attail & Burstein, 2006; Chen *et al.*, 2010; Dikli, 2006)를 진행하고 있으나 한국어의 경우에는 교육과정평가원(KICE, Korea Institute for Curriculum and Evaluation)에서 기초연구(노은희 외, 2014)로 진행하고 있으며, 그 외에는 개인차원에서 연구(강원석, 2011; 박일남 외, 2013; 이경호 & 이공주, 2014)를 진행하고 있을 뿐이다.

해외의 자동채점 시스템은 형태소 분석 및 품사 부착, 의존 구문 분석



한 모든 형태소 후보들을 생성하여 격자 구조  $G$ 를 구축한다. 품사 부착은 생성된 형태소 후보들과 전이 확률 사전을 이용하여 격자 구조  $G$ 의 간선(edge)에 전이 확률을 적재한 가중치 그래프를 생성하고, 그 가중치 그래프  $G$ 의 시작 정점(BOS)으로부터 마지막 정점(EOS)까지의 최적 경로를 찾는 것으로 해결한다. 본 논문에서 제안하는 형태소 분석 및 품사 부착 기법은 구현이 쉽고 간단하다는 특징이 있다. 실제 제안하는 기법을 구현하여 실험해 본 결과 기존의 기법보다 형태소 분석의 재현율 및 품사 부착 결과의 정확률이 향상되었다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구로 한국어의 특성과 기존의 한국어 형태소 분석 및 품사 부착 기법의 종류에 대해 살펴본다. 3장에서는 현재 연구 중인 ‘한국어 서답형 문항 자동채점 시스템’에 관한 개략적인 내용과 자동채점 시스템에서 사용 중인 기존 형태소 분석 및 품사 부착 기법의 문제점에 대해 언급한다. 4장에서는 3장에서 제시한 문제점을 해결하기 위한 새로운 형태소 분석 및 품사 부착 기법에 관해 설명한다. 5장에서는 본 논문에서 제안한 방법으로 구현한 시스템의 성능을 평가하고, 마지막으로 6장에서 결론을 맺고 앞으로의 연구 방향을 제시한다.

## 제 2 장 관련 연구

형태소(morpheme)는 어절을 구성하는 기본 요소로 최소의 의미를 가지는 가장 작은 단위이다. 여기서 의미는 어휘적 의미와 문법적 의미를 모두 포함한다(김남미, 2010). 예를 들어 “나는 간다”에서 형태소는 { ‘나’, ‘는’, ‘가’, ‘ㄴ다’ } 이다.

형태소 분석(morphological analysis)이란 주어진 어절에서 가능한 형태소 후보들을 추출하는 단계이며, 품사 부착(part-of-speech tagging)이란 형태소 후보 중, 주어진 어절에 대해 가장 적절한 품사를 선택하는 작업이다. 일반적으로 형태소 분석에서 가능한 모든 형태소 후보들을 찾아내서 중의성(ambiguity)을 생성하고, 품사 부착에서는 그 중의성을 제거한다(김재훈 외, 1995). 형태소 분석 및 품사 부착의 고려사항은 각 언어의 언어학적 특징에 달라진다(강승식, 1994; 강승식, 2002; Kim, *et al.*, 1994). Table 2.1은 한국어의 형태소 분석 및 품사 부착의 한 예를 보인다.

**Table 2.1** An Example of Morphological Analysis and POS Tagging

입력	나는 간다	
형태소 분석	나는	나(自)/NP + 는/JJ 날(飛)/VV + 는/EM 나(茁)/VV + 는/EM
	간다	가/VV + ㄴ다/EF
품사 부착	나는	나(自)/NP + 는/JJ
	간다	가/VV + ㄴ다/EF

본 장에서는 한국어 서답형 자동채점 시스템에 적용하기 위한 형태소 분석 및 품사 부착 기법을 개발하기 위한 배경 지식을 설명한다. 먼저 한

국어 형태소 분석에서 고려해야 하는 한국어의 특성에 관해 설명하고, 현재까지 연구된 한국어 형태소 분석 및 품사 부착 기법에 관해 설명한다.

## 2.1 한국어의 특성

형태소 분석 및 품사 부착은 해당 언어가 속하는 범주와 특성을 고려해야 한다(강승식, 2002). 한국어는 다양한 파생접사(접두사, 접미사)와 굴절 접사(조사, 어미)가 어근(語根)<sup>1)</sup>에 부착되어 의미가 변화되거나 문법적 관계가 표시되는 교착어(agglutinative language)에 속한다. 한국어의 형태소 분석 및 품사 부착을 위한 고려사항은 아래와 같다.

### (1) 한국어의 형태론적 변형과 이형태

한국어는 두 개 이상의 형태소가 복합명사를 구성하거나, 체언과 용언에 접사, 조사, 어미가 결합하여 하나의 어절을 구성한다는 특징이 있다. 한국어의 형태론적 변형의 종류들을 표로 정리하면 Table 2.2와 같다(김남미, 2010).

Table 2.2 Morphological Alternation of Korean

형태론적 변형 종류	예시
용언의 불규칙 활용	묻다(問) : ‘묻는-’, ‘물어-’, ‘묻고-’
모음조화	‘아/어’로 시작되는 어말어미
음운론적 이형태	‘하였다’ 등
모음 축약	‘보이+다/뵈다’, ‘가지+어/가져’
준말	‘-에서는/-에선’, ‘누구가/누가’, ‘-려고 하느냐/-려느냐’ 등

용언의 불규칙 활용은 “특정 유형의 용언들을 활용할 때 어근과 어미의 형태가 변화하는 현상”의 형태론적 변형 규칙으로 기술된다. 불규칙 활용의 예는 Table 2.2에 나타나 있는 것처럼 “묻다(問)”의 어근인 ‘묻-’이 어

1) 활용론(活用論)에서 보면 어근(語根)이 아닌 어간(語幹)이라고 하는 것이 올바른 표기이나, 본 논문에서는 형태소 분석 및 품사 부착에 초점을 맞추고 있으므로 낱말의 실질적인 의미를 나타내는 어근이라는 용어를 사용한다.

미와 결합하여 ‘물-’처럼 형태가 변형된 경우이다.

모음조화는 ‘먹어’와 ‘잡아’와 같이 두 음절 이상의 단어에서 뒤의 모음이 앞 모음의 영향으로 그와 가깝거나 같은 소리가 되는 언어 현상이다. ‘나’, ‘고’, ‘개’ 등의 양성 모음은 양성 모음끼리, ‘너’, ‘고’, ‘개’ 등의 음성 모음은 음성 모음끼리 어울리는 현상이다.

음운론적 이형태는 ‘하였다’처럼 ‘하-’가 음운론적으로는 양성모음 선어말어미 ‘-았-’이 와야 하는데도 특별한 이유 없이 ‘-였-’이 온 것처럼, 음운론적인 이유로는 설명할 수 없는 이형태를 의미한다.

모음 축약은 ‘보이+다’가 ‘뵈다’로, ‘가지+어’가 ‘가져’가 되는 것과 같이 두 형태소가 서로 만날 때 앞뒤 형태소의 두 음절이 한 음절로 줄어드는 현상을 말한다.

준말은 조사의 결합형 ‘-에서는’이 ‘-에선’과 같이 하나의 형태소가 축약되는 경우와 ‘누구가’가 ‘누가’와 같이 두 개의 형태소가 하나로 축약되는 경우, ‘-려고 하느냐’가 ‘-려느냐’로 두 어절에 걸쳐서 축약이 일어나는 경우가 있다.

## (2) 품사 체계와 문법형태소

품사의 분류 기준은 의미(meaning), 형태(form), 기능(function)이 있다. 의미는 단어가 가지는 의미의 종류별 공통성을 뜻하고, 형태는 단어 형태의 활용 여부, 기능은 문장 구성에서 단어가 가지는 역할을 의미한다. Table 2.3은 한국어의 품사를 기능, 의미, 형태에 따라 나눈 것이다(김남미, 2010).

Table 2.3 Part of Speech in Korean

기준	기능	의미	형태
품사	체언	명사	형태 변화 없음
		대명사	
		수사	
	수식언	관형사	
		부사	
	독립언	감탄사	
	관계언	조사	형태 변화 있음
		서술격 조사	
	용언	동사	
형용사			

체언은 문장에서 주어나 목적어와 같이 주체적인 역할을 하는 품사이다. 체언에는 사물의 이름을 가리키는 품사인 명사, 명사를 대신하는 품사인 대명사, 숫자나 순서를 가리키는 품사인 수사가 속한다. 수식언은 다른 단어를 꾸미는 역할을 하는 품사이다. 수식언은 체언을 꾸미는 품사인 관형사와 용언을 꾸미거나 부수적인 상황을 나타내는 품사인 부사를 포함한다. 독립언은 문장 속의 다른 성분과 관계를 맺지 않은 품사로 말하는 이의 느낌이나 부름 등을 나타내는 품사인 감탄사이다. 관계언은 명사류의 기능을 나타내거나 의미를 더하는 보조적인 품사인 조사를 포함한다. 조사는 문장에서 어떤 성분에 일정한 자격을 주는 조사인 격 조사와 두 단어를 같은 자격으로 이어 주는 접속 조사, 여러 격에 두루 쓰이면서 특정한 의미를 덧붙여 주는 보조사로 나눌 수 있다. 용언은 문장에서 주로 서술어의 기능을 하는 품사로 형용사와 동사를 포함한다. 형용사는 모양이나 상태를 나타내는 품사이고 동사는 동작을 나타내는 품사이다. 품사들에서 서술격 조사, 동사와 형용사는 활용에 따라 형태가 변화한다는 특징이 있다.

### (3) 어절 형성 규칙

한국어의 어절은 하나 이상의 형태소로 이루어져 있으며 어절을 이루고 있는 형태소들은 단어 형성 규칙 및 결합 제약 규칙(김성용 외, 1987)과

접속 정보(김성용 외, 1987; 안동언, 1993) 등의 결합 제약 조건에 따라 결합한다. 어절은 1개 이상의 형태소들로 구성되어 있으므로 어절을 단순히 연속된 형태소의 집합으로 표현하면 식 (2.1)과 같이 나타난다.

$$\langle \text{어절} \rangle ::= \langle \text{형태소} \rangle^+ \quad (2.1)$$

이처럼 어절을 정의했을 경우 형태소끼리의 결합 관계와 형태소 분리라는 문제가 발생한다. 이 문제를 해결하기 위해 어절을 어근부(stem part)와 기능어부(function part)로 이루어져 있다고 가정하면 식 (2.2)와 같이 기술할 수 있다(강승식, 1994; 강승식, 2002). 이와 같은 어절 유형 기술 방법은 형태소 분석 알고리즘에서 형태소 분리 문제를 고려하여 포괄적으로 기술했기 때문에 실제 어휘형태소의 품사 유형과 형태소 간의 결합제약 관계에 따라 달라질 수 있다(강승식, 2002).

$$\begin{aligned} \langle \text{어절} \rangle &::= \langle \text{어근부} \rangle^+ \langle \text{기능어부} \rangle^* \\ \langle \text{어근부} \rangle &::= [\langle \text{접두사} \rangle] \langle \text{어휘형태소} \rangle^+ \\ \langle \text{기능어부} \rangle &::= \langle \text{문법형태소} \rangle^+ \end{aligned} \quad (2.2)$$

#### (4) 복합어와 띄어쓰기

한글 맞춤법의 띄어쓰기 조항에는 두 개의 단어를 띄어 쓰는 것이 원칙이나, 예외적인 규칙을 두어 붙여 쓰는 것을 허용한다. 일반적으로 형태소 분석을 할 하나의 문장에는 띄어쓰기 오류가 아예 없거나 붙여 써야 할 것을 띄어 쓴 오류가 없는 것으로 간주하고 진행하거나, 띄어쓰기 오류의 유형을 고려해서 전처리를 거친 뒤 형태소 분석을 진행한다(강승식, 1994; 강승식, 2002; 김재훈 외, 1995). 대부분의 띄어쓰기 문제는 복합명사에서 발생하지만, 그 외에도 관형사(이/그/저)나 본용언과 보조용언 사이에서도 붙여쓰기가 허용된다(강승식, 1994; 강승식, 2002).

복합명사의 경우, 두 개 이상의 명사를 붙여 쓰는 것이 허용되므로 붙여 쓴 복합명사를 인식해야 하는 문제가 있다. 복합명사를 처리하는 방법

으로는 복합명사 사전을 구축하여 처리하는 방법(허정, 장명길, 2005)과 의미정보를 이용하여 추정하는 방법(김수남 외, 1998; 이용훈, 옥철영, 2011) 그리고 앞의 두 가지 방법을 절충한 방법이 있다. 복합명사는 무한히 생성될 수 있으므로 어떤 처리를 할 것인지는 개발자가 선택해야 한다. 고유명사, 외래어, 신조어와 같이 사전에 등록되지 않은 단어를 추정하는 방법은 복합명사의 처리 방법과 유사하다(박봉래, 1998).

## 2.2 한국어 형태소 분석 기법

한국어의 형태소 분석은 각 어절에서 가능한 모든 형태소 후보를 찾는 과정이다. 한국어의 형태소 분석 기법은 분류 기준에 따라 다양하게 나눌 수 있는데, 현재까지 제안된 형태소 분석 기법을 유형별로 분류하면 Table 2.4와 같다.

Table 2.4 Classification of Morphological Analysis in Korean

분석 기준	형태소 분석 유형	비고
분석 모델	언어 독립적, 언어 종속적 모델	전분석
분석 방향	상향식 & 병행 (bottom-up & parallel), 하향식 & 예측 (top-down & predictive)	
검색 방향	좌우 분석, 양방향 분석, 역방향 분석	
결합 제약	단어 형성 규칙 & 결합 제약 규칙, 접속 정보 & 접속 정보표 이용	
기계학습	지도학습	
알고리즘	규칙 기반, 사전 기반, 말뭉치 기반	
사전 탐색	단어 단위, 문장 단위, 문단 단위	무분석

형태소 분석 모델에는 언어 독립적인 방법론으로 two-level 모델을 한국어에 적용한 것(Kim, *et al.*, 1994), two-level 모델에서 기계학습으로 규칙을 학습하는 방법(장병탁 & 김영택, 1990), 어휘 변환기(lexical transducer)에 의한 방법(Kwon & Karttunen, 1994)이 있다. 언어 독립적인 방법을 사용하면 형태론적 변형이나 형태소 분리 문제를 쉽게 해결할 수 있으나 한국어의 형태론적 특성을 반영하기 어렵다. 언어 종속적인 방법론은 한국어의 형태론적 특성을 고려한 방법론으로, 언어 독립적인 방법론에 속하지 않은 모든 형태소 분석 기법이 여기에 속한다.

형태소 분석 방향은 상향식(bottom-up) 방식과 하향식(top-down) 방식으로 분류할 수 있다. two-level 모델과 최장 일치법을 비롯한 대부분의 방법론이 상향식 방식을 취하고 있는데, 하향식 방식은 탐색 공간(search space)이 커지기 때문에 백트래킹(backtracking)이 자주 일어나거나 사전 탐색의 부담이 크다는 단점이 있다(강승식, 1994; 강승식, 2002).

단어 검색 방향은 형태소 분석을 진행하는 방향에 따라 왼쪽 끝에서부터 오른쪽 끝으로 단어를 검색하는 좌우 분석법(right-to-left analysis)(안동연, 1999), 양쪽 끝에서부터 중심 방향으로 검색하는 양방향 분석법(bidirectional analysis)(최재혁 & 이상조, 1993), 양방향 분석법의 변형으로 중심부에서부터 양 끝으로 분석을 진행하는 역방향 분석(심광섭, 2007)이 있다. 좌우 분석법은 트라이(trie) 구조의 사전을 이용해서 사전 탐색의 효율을 높일 수 있으나 복합명사나 미등록어의 추정이 어렵다. 양방향 분석법은 양쪽 끝에서부터 동시에 분석을 수행한다. 역방향 분석은 구체적으로 제시되지는 않았으나 양방향 분석법의 변형으로 단어를 구성하고 있는 음절이나 자소의 특성을 이용해서 형태소 분리 위치를 먼저 추정한 뒤, 중심으로부터 양 끝으로 형태소 분석을 진행하는 방식이다.

형태소 결합 제약 조건을 기술하는 방법에는 단어 형성 규칙과 결합 제약 규칙에 따라 결합 관계를 검사하는 방법(김성용 외, 1987), 형태소들의 결합 관계를 결합 유형에 따라 접속 정보표로 기술하고 형태소마다 좌-우

결합 정보를 사전으로 구축해서 결합 조건을 기술하는 방법(김성용 외, 1987; 안동언, 1993), 그리고 두 가지 방법을 혼합하여 사용하는 인접 조건 검사 방법(심광섭 & 양재형, 2004)이 있다. 결합 제약 규칙에 따라 결합 관계를 검사하는 방법은 매번 결합 조건을 검사해야 하므로 처리 효율이 떨어진다는 단점이 있고, 접속 정보표를 이용하는 방법은 결합 여부를 쉽게 검사할 수 있다는 장점이 있으나 접속 정보를 분류하고 모든 형태소마다 접속 정보를 부여해야 하는 어려움이 있다(강승식, 1994; 강승식, 2002).

형태소 분석에서 사용하는 기계학습은 정답이 부착된 데이터를 학습 말뭉치로 사용하여 분류기(classifier)를 학습하고 정답이 부착되지 않은(학습되지 않은) 데이터를 입력받아서 정답을 부여하는 지도학습(supervised learning) 방식을 주로 사용한다(심광섭, 2011a; 전길호, 2012). 엄밀히 말하면, 기계학습은 형태소 분석을 하는 동시에 품사 부착을 위해 사용되거나 형태소 분석 과정을 생략하고 품사 부착 과정에만 사용된다.

형태소 분석 알고리즘은 형태론적 변형을 처리를 위한 방법에 따라 규칙으로 처리하는 방법과 사전을 기반으로 처리하는 방법으로 분류할 수 있다. 대부분의 방법론은 규칙을 기반으로 하는 방법을 취하고 있다. 그러나 ‘날(飛-)’의 불규칙 활용형인 ‘나(飛-)’와 같이 불규칙 활용의 경우 규칙으로 처리하기 어렵다는 점 때문에 불규칙 활용이 일어난 어근이나 형태소 분석에 필요한 여러 가지 정보를 미리 사전으로 구축하여 처리하는 방법(Kwon, *et al.*, 1991; 김재한 & 옥철영, 1994; 박영환, 1991)과 사용 빈도가 높은 단어들의 분석 결과를 기분석 사전에 수록함으로써 효율을 높이는 방법(양승현 & 김영섭, 2000)이 있다.

사전을 구축하는 방법에는 문법형태소의 최소 단위인 단위형태소만 사전에 등록하고 문법형태소끼리의 결합 관계는 접속 정보표에 기술하는 방식과 결합 가능한 모든 유형을 하나의 단위로 하는 결합형태소를 사전에 수록하는 방식이 있다. 단위 형태소만 사전에 수록할 경우에는 접속 정보

표의 유지 및 관리의 어려움이 있고, 결합형태소를 수록하는 경우에는 사전의 크기가 커지고 단위 형태소를 추출해야하는 단점이 있다. 이 외에도 자주 사용되는 어절에 대한 형태소 분석 결과를 저장하고 있는 기분석(既分析) 사전을 구축하여 사용할 수 있다.

위와 같이 분석 방식에 따라 형태소 분석 기법을 분류하는 것 외에도 형태소 분석기의 실행 시에 실제로 처리하는 형태소 분석 과정에 따라 전 분석 처리 방법과 무분석 처리 방법으로 분류할 수 있다(양승현 & 김영섭, 2000). 전분석 방법은 실행 시에 형태소 분석의 전 과정이 이루어지는 기법들을 총칭하며, Table 2.4의 형태소 분석 모델 방식에서부터 문서 입력 단위 방식까지의 기법들이 이 방법에 속한다. 무분석 방법은 기분석 어절 사전을 이용한 분석 방법으로, 어절에 대한 형태소 분석 결과를 미리 구축하여 사전에 수록해놓고, 실행 시에 사전을 탐색하여 직접 형태소 분석 결과를 출력하는 방식으로 문법형태소 사전 방식이 이 방법에 해당한다. 분석 알고리즘 기법은 알고리즘의 기준에 따라 전분석 방법이 될 수도 무분석 방법이 될 수도 있다.

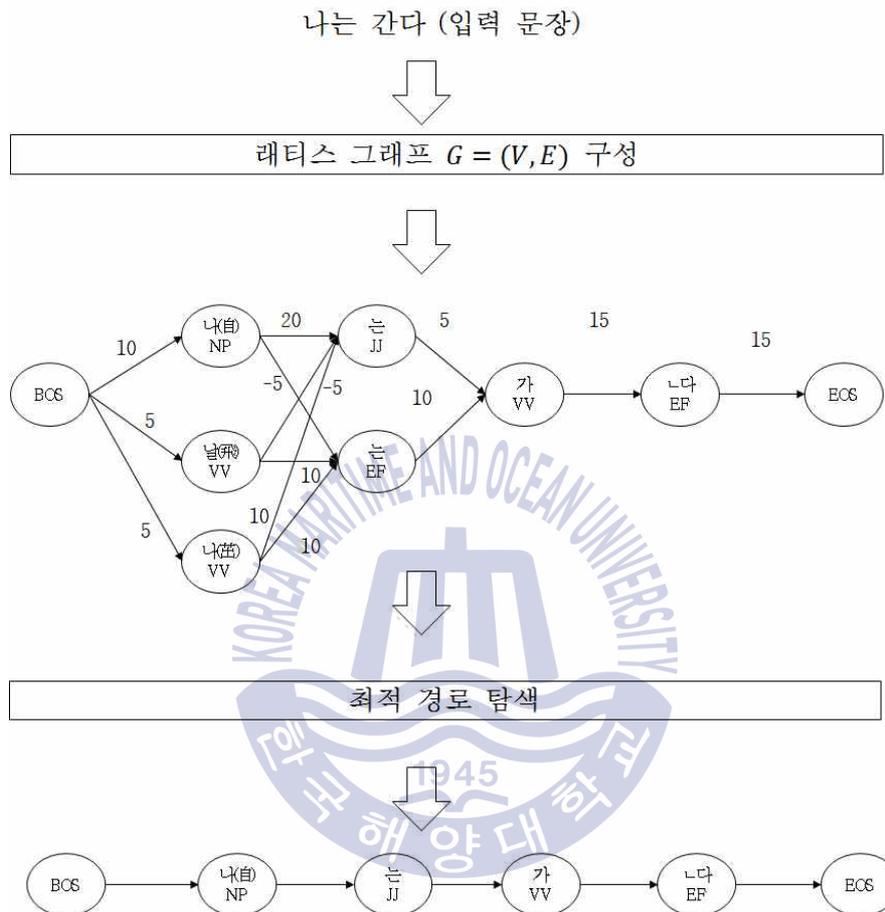
## 2.3 한국어 형태소 품사 부착

한국어의 품사 부착이란 형태소 분석을 통해 찾은 여러 개의 품사 후보 중, 주어진 문장에 가장 적합한 단어의 품사를 결정하여 어절의 형태론적 중의성을 해소하는 작업이다. 한국어 품사 부착 기법은 크게 통계 기반의 접근 방식(김재훈, 1996; 나승훈 외, 2013; 이운재, 1993)과 규칙 기반 접근 방식, 통합 접근 방식으로 구분할 수 있다(임해창 외, 1996). 통합 접근 방식은 앞의 두 방식의 장점 및 단점을 상호 보완하여 사용하는 방식이므로 자세한 설명을 생략한다.

### (1) 통계 기반 접근 방식

통계 기반 접근 방식은 원시 또는 품사 부착이 된 말뭉치를 분석하여 얻어낸 확률을 이용하는 방법으로 지도학습에 해당한다. 통계 기반 접근

방식에서의 품사 부착 기법은 각 어절 혹은 형태소의 확률정보를 추출한 뒤, Fig. 2.1과 같이 격자 구조(lattice structure)를 구성하여 적합한 품사를 결정한다(김재훈, 1996; 나승훈 외, 2013; 이운재, 1993).



**Fig. 2.1** An example of Korean Part-of-Speech tagging using a lattice structure

Fig. 2.1을 살펴보면 입력 문장 ‘나는 간다’에서 가능한 모든 형태소 분석 결과를 방향성 그래프를 통해 격자 구조  $G = (V, E)$ 로 나타낸다(김재훈, 1996; 나승훈 외, 2013).  $V$ 는 입력 문장에서 각 어절에서 가능한 모든 형태소와 품사 쌍들의 집합을 나타내며,  $E$ 는 인접하는 모든 형태소 간의 가중치를 갖는 간선(weighted edge)들의 집합이다. BOS는 문장의 시작을 가리키는 특수 형태소를, EOS는 문장의 끝을 가리키는 특수 형태소이다. 격자 구조  $G$ 에서 품사 부착 문제를 푸는 방법은 BOS에서 EOS로 가는

경로 중 가장 최적의 경로(best path)를 찾는 문제와 같다. 이렇게 격자 구조로 품사 부착 문제를 풀면 한국어 형태소 분석에서 발생하는 제반 문제들을 자연스럽게 표현할 수 있으며, 은닉 마르코프 모델, 퍼지망 모델(김재훈 외, 1993) 등의 모델에서 추출된 매개변수를 가중치로 이용할 수 있다는 장점이 있다(김재훈, 1996).

통계 기반 접근 방식에서 사용하는 확률의 종류는 문맥 확률(contextual probability)과 어휘 확률(lexical probability)이 있다. 문맥 확률  $P(w_i|w_{i-1})$ 과 어휘 확률  $P(t_i|w_i)$ 은 식 (2.3)과 식 (2.4)와 같이 단어와 품사의 빈도수로 구할 수 있다.

$$P(w_i|w_{i-1}) \approx \frac{\text{freq}(w_{i-1}, w_i)}{\text{freq}(w_{i-1})} \quad (2.3)$$

$$P(t_i|w_i) \approx \frac{\text{freq}(t_i, w_i)}{\text{freq}(w_i)} \quad (2.4)$$

식 (2.3)과 식 (2.4)에서  $w_i$ 는  $i$ 번째 단어를 의미하며,  $t_i$ 는  $w_i$ 의 품사를 의미한다. 일반적으로 문맥 확률과 어휘 확률을 결합한 확률값을 전이 확률로 많이 사용하며, 대표적으로 은닉 마르코프 모델(HMM, Hidden Markov Model)에 기반을 둔 전이 확률값이 있다(김재훈, 1996). 이를 식으로 표현하면 식 (2.5)와 같이 나타낼 수 있다.

$$T(w_{1,N}) = \operatorname{argmax}_{t_{1,N}} \prod_{i=1}^N P(w_i | w_{i-1}) P(t_i | w_i) \quad (2.5)$$

식 (2.5)에서  $T(w_{1,N})$ 은 단어의 개수가  $N$ 인 문장의 품사 부착 결과이며,  $t_{1,N}$ 은 입력 문장이 가질 수 있는 모든 품사의 종류이다. 통계 기반의 접근 방식은 모든 언어 현상에 대해 적용이 가능하다는 장점이 있지만, 실세계 언어를 충분히 대표할 만한 양과 질의 말뭉치가 없다는 단점이 있다(김재훈, 1996; 이운재, 1993).

## (2) 규칙 기반 접근 방식

규칙 기반 접근 방식은 한국어에 적용되는 공통된 원리나 결정적인 규칙을 적용하여 품사를 부착하는 방식이다. 이 방식은 일관성 있고 예외가 없는 규칙을 찾는 것이 어렵고, 새로운 환경에 대한 적응력이 낮다는 단점이 있다. 규칙을 적용한 품사 부착 기법에는 한국어의 변형규칙을 이용한 품사 부착 등이 있다(임해창 외, 1993). 이 방식을 적용한 시스템들은 여러 종류의 기능어 사전을 참조하고 예외단어와 접미어, 특수기호 등을 처리하여 사전에 정의한 품사 부착 규칙을 적용하는 방식이다.



### 제 3 장 한국어 서답형 자동채점 시스템의 형태소 분석 및 품사 부착 기법의 문제점 분석

기계학습을 이용한 자동채점 시스템은 대규모의 학생 답안을 효율적으로 처리할 수 있다는 장점이 있다. 실제로 영어권 국가에서는 GMAT, TOFLE 등의 대규모 시험의 채점을 위해 자동채점 시스템(Attali & Burstein, 2006)을 도입하고 있으며, 자동채점 결과의 정확률을 높이기 위한 연구를 진행하고 있다. 기계학습을 이용한 자동채점 시스템에서는 문장(혹은 문단)의 길이, 핵심어, n-gram 등의 문맥 정보나 구문분석 정보 등을 자질로 사용하고 있다(Attali & Burstein, 2006). 우리나라에서는 한국교육과정평가원에서 해외에서 개발된 자동채점 시스템을 참고하여 한국어 서답형 문항 자동채점 시스템을 개발하기 위한 기초 연구(노은희 외, 2014)를 진행하고 있다.

이 장에서는 한국교육과정에서 개발 및 실용화를 연구하고 있는 한국어 서답형 문항 자동채점 시스템에 대한 대략적인 구조와 해당하는 시스템에서 사용 중인 형태소 분석 및 품사 부착 기법과 그에 대한 문제점을 서술한다.

### 3.1 한국어 서답형 문항 자동채점 시스템

한국교육과정에서 연구 중인 한국어 서답형 문항 자동채점 시스템의 인터페이스는 Fig. 3.1과 같다. Fig. 3.1은 언어 분석을 끝낸 후, 해당 결과를 바탕으로 채점자가 생성한 초기 학습용 답안을 기계학습을 통해 채점한 결과를 보여주는 화면이다. 여기서 정답 예측 확률이 ‘-’로 나타난 답안들은 기계학습으로 채점하지 못한 답안들로 편의상 미채점 답안이라고 부른다.

정답 예측 확률	답안 수	예시 답안	예측 점수	점수 확인
-	3796	빙하가 퇴적작용을 하였다.	-	-
0.993	3	물이 침식 작용 하였다.	1	미확인
0.992	6	빙하가 움직여서 침식작용을 하였다.	3	미확인
0.992	4	물이 침식작용	1	미확인
0.992	88	피오르가 혼과 U자곡 작용하였다	0	미확인
0.991	9	피오르가 U자곡 작용을 했다	0	미확인

Fig. 3.1 A graphic user Interface of the Korean Automated Scoring System (노은희, 2014)

Fig. 3.1까지의 과정을 포함한 전체 자동채점 시스템의 과정을 구조도로 나타내면 Fig. 3.2와 같이 나타난다.

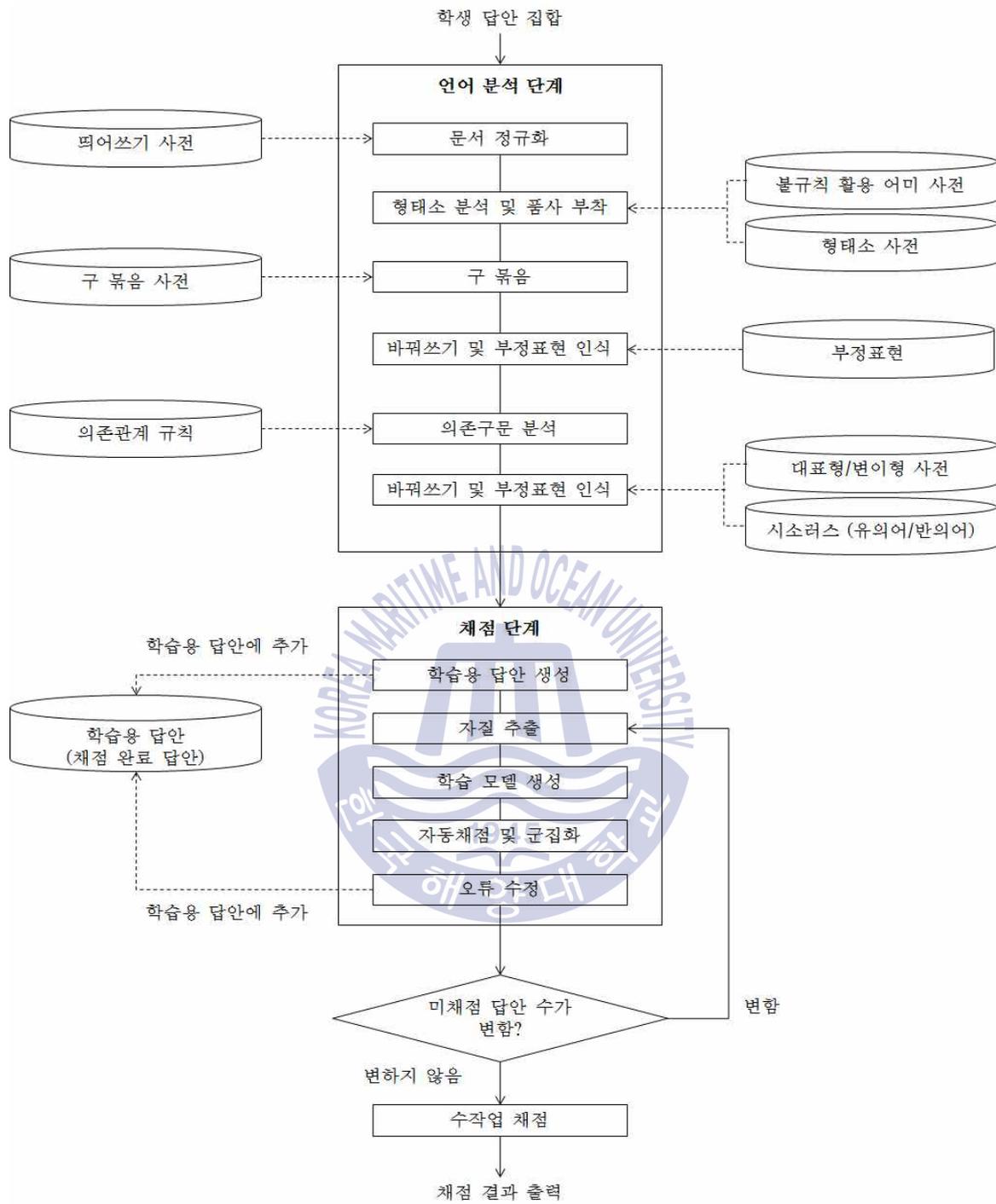


Fig. 3.2 The overall structure of the Korean Automated Scoring System

한국어 서답형 자동채점 시스템은 Fig. 3.2처럼 ‘언어 분석 단계’와 ‘채점 단계’로 나눌 수 있다. 언어 분석 단계는 학생 답안들을 입력으로 받은 후, 자동채점에 필요한 언어 정보를 분석하여 처리하는 단계이다. 언어 분석 단계의 결과가 완전히 일치하는 답안들을 유형이라고 정의한다. 채점 단계는 답안들이 많이 포함된 유형을 채점자가 수작업으로 채점하여 초기 학습용 답안을 만든 뒤, 학습용 답안들로부터 유용한 자질들을 추출하여 자동채점을 진행하는 단계이다. Fig. 3.2의 채점 단계는 Fig. 3.3에 나타난 자가 학습(self-training) 방식의 과정이다.

```

ALGORITHM 자가 학습( $L, U$ ) :
  Repeat  $U$ 의 크기가 더 이상 변하지 않을 때까지
    학습용 답안  $L$ 을 이용하여 자동채점 모델  $h$ 를 만듦
    자동채점 모델  $h$ 로 채점되지 않은 답안  $U$ 를 채점
  IF 정답 예측 확률  $\geq threshold$ 
    채점자에게  $U'$ 의 결과를 보여줌
  ELSE
     $U' = \emptyset$ 
  ENDIF
   $L = L + U'$ 
   $U = U - U'$ 
UNTIL
  
```

Fig. 3.3 Pseudocode for the scoring step using the self-training algorithm

$h$ 는 자동채점 모델(분류기)이며,  $L$ 은 학습용 답안들의 집합이다.  $U$ 는 채점되지 않은 답안들의 집합이며,  $U'$ 는 분류기를 통해 자동채점된 답안들의 집합으로  $U$ 의 부분 집합에 해당한다.  $threshold$ 는 정답 예측 확률의 기준값이다.

### 3.2 기존의 형태소 분석 및 품사 부착 기법

본 논문에서 초점을 맞추고 있는 부분은 자동채점 시스템의 ‘언어 분석 단계’ 중 ‘형태소 분석 및 품사 부착’ 부분이다. 자동채점 시스템에 적용된 형태소 분석 및 품사 부착 기법의 전체 구조는 Fig. 3.4와 같다.

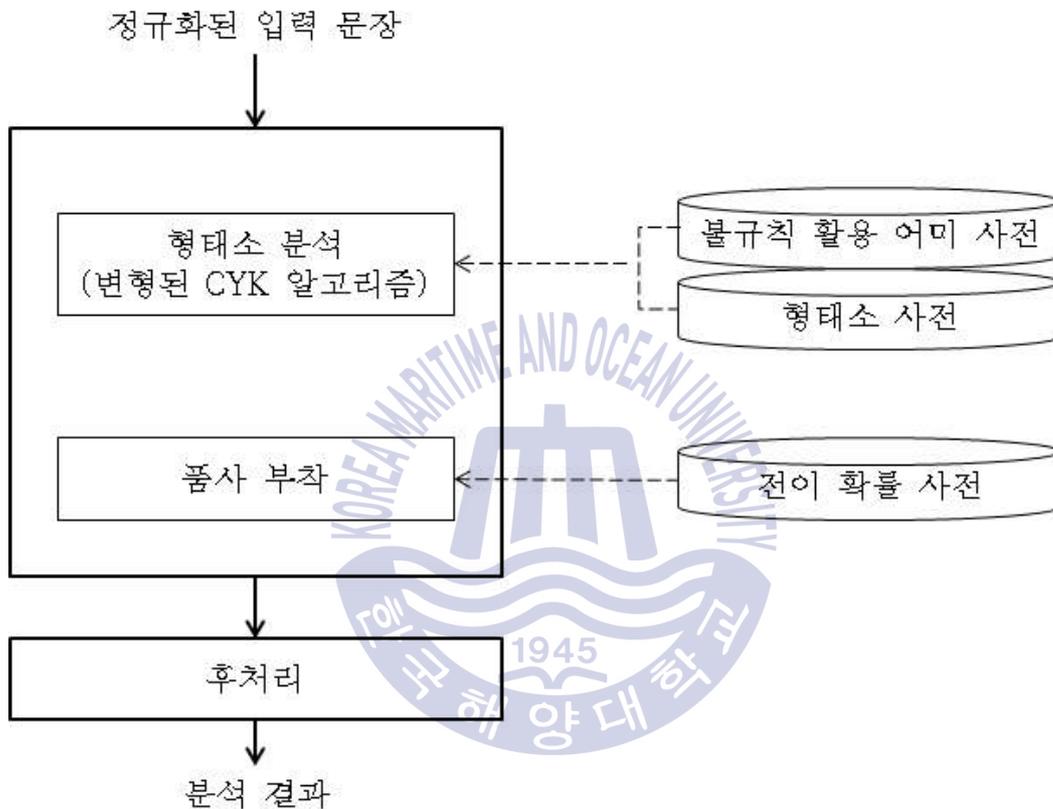


Fig. 3.4 Flow diagram for the previous morphological analysis and POS tagging

Fig. 3.4의 ‘정규화된 입력 문장’은 Fig. 3.2의 언어 처리 단계에서 문서 정규화까지 완료한 문장이다. 문서정규화는 띄어쓰기 교정, 문장 분리, 문장부호 제거와 철자 교정으로 이루어진다. 2014년에 개발된 자동채점 시스템에 적용된 형태소 분석 및 품사 부착 기법에서 형태소 분석은 Table 2.4의 형태소 분석 기법 중 상향식 분석 기법의 하나인 변형된 CYK 알고리즘(김성용 외, 1987)과 불규칙 활용 어미 사전, 형태소 사전을 결합한

방식이다. 변형된 CYK 알고리즘을 사용했을 때, 어미의 활용이 존재하지 않는 경우는 사전 정보를 이용해 가능한 모든 형태소 분석 결과를 찾을 수 있다. 그러나 “아름답+-은”이나 “아름답+-ㄴ”과 같이 동사의 원형과 어미가 결합할 때 “아름다운”과 같은 철자의 변형이 발생한다. 이와 같은 용언(동사와 형용사)의 경우, 변형된 CYK 알고리즘만으로 처리하기 힘들므로 불규칙 활용 어미 사전을 함께 사용하여 처리한다. 불규칙 활용 어미 사전의 구조는 Table 3.1과 같다.

**Table 3.1** Examples of dictionary structure of irregular verbs

표층형	형태소 분석 결과	불규칙 활용	결합 규칙
아름다	아름답/PA	ㄹ 불규칙	-
운	ㄴ/EF	-	ㄹ 불규칙

사전에 “아름다”와 “운”과 같은 용언의 표층형<sup>2)</sup>에 대한 형태소 분석 결과와 불규칙 활용 어미의 정보를 함께 추가한다. “아름다운”이라는 어절에서 “아름다”에서 ‘ㄹ 불규칙 활용’이 일어났다는 정보를 통해 “아름다”의 뒤에 올 “운”이라는 형태소 분석 후보는 앞의 단어가 ‘ㄹ 불규칙 활용’일 경우, 결합 규칙 정보를 통해 ‘ㄴ/EF’로 분석된다. 형태소 사전은 Table 3.1에서 표층형과 형태소 분석 결과만을 정보로 가지고 있는 사전이다. 불규칙 활용 어미 사전과 형태소 사전은 세종말뭉치(국립국어원, 2011)를 이용하여 구축했다. 기존 시스템의 형태소 분석 과정의 의사코드(pseudocode)는 Fig. 3.5와 같다.

2) 단어가 발화되기 이전의 상태를 어휘형, 발화된 이후의 상태를 표층형이라고 한다.

```

ALGORITHM 형태소 후보 생성(token) :
    RETURN 변형된 CYK 알고리즘(token) 결과

ALGORITHM 형태소 분석기(S) :
    T=입력 문장 S를 공백과 문자의 유형을 기준으로 분리
    Repeat i = 1부터 n까지 (n = T의 크기)
        CASE type(ti) OF
            한글      : H, 형태소 후보 생성(ti)
            영어      : E
            문장기호  : P
            숫자      : N
            자모      : J
            한자      : Hj
            공백      : Sp
        END CASE
    Until

```

Fig. 3.5 Pseudocode for the previous morphological analysis using the modified CYK algorithm

Fig. 3.5에서  $S$ 는 입력 문장이며,  $T$ 는 입력 문장을 토큰 단위로 나눈 결과이다.  $t_i$ 는  $i$ 번째 토큰을 의미하며,  $type(t)$ 는 토큰  $t$ 의 유형이다.

“9월에 학교(school)에 갔다.”를 토큰으로 분리하면 “9/N 월에/H /Sp 학교/H (/P school/E )/P 에/H /Sp 갔다/H ./P”가 된다. 여기서 토큰의 유형이 H인 경우 CYK 알고리즘과 형태소 사전을 이용해서 형태소를 분석한다. 예를 들어 “갔다.”를 분석하면 “가/PV 았/EP 다/EF ./SY”를 얻을 수 있다.

품사 부착은 통계 기반 접근 방식으로 구현한다. 형태소 분석 결과를 이용하여 격자 구조를 구성하고 구성된 격자 위에 세종 말뭉치에서 구해진 문맥 확률  $P(w_i|w_{i-1})$ 과 어휘 확률  $P(t_i|w_i)$ 을 적재하여 가중치 네트워크



부착 결과를 출력하지 못하거나, 메모리가 충분하다고 하더라도 계산량이 많으므로 결과를 출력하기까지 시간이 오래 걸린다는 문제점이 있다.

이 문제가 치명적인 이유는 Fig. 3.2처럼 ‘언어 분석 단계’ 중 형태소 부착 및 품사 부착 과정이 끝나지 않는 경우에는 그 뒤의 과정을 처리할 수 없어 채점을 진행할 수 없기 때문이다.



## 제 4 장 음절 기반의 단어 분리와 사전 탐색 기법을 이용한 형태소 분석 및 품사 부착

기존의 형태소 분석 및 품사 부착 기법은 3.3절에서 언급했던 것처럼 같은 음절이 연속되는 답안이 들어왔을 경우, 계산량의 폭주로 인해 작업을 완료하기까지 시간이 오래 걸리거나 메모리 부족 현상으로 작업을 완료하지 못한다는 치명적인 단점이 있다.

이 장에서는 이런 문제점을 해결하기 위한 형태소 분석 및 품사 부착 기법을 제안한다.

### 4.1 제안하는 형태소 분석 및 품사 부착 기법의 구조

일반적인 형태소 분석 및 품사 부착은 형태소 원형 복원, 형태소 분리, 형태소의 품사 부착 과정으로 이루어진다(이재성, 2011). 각 과정은 Table 4.1과 같이 형태소 분석 및 품사 기법에 따라 순서를 변경하거나 동시에 처리할 수 있다. ‘+’는 동시 작업을, ‘-’는 순차 작업을 나타낸다.

**Table 4.1** Classification of integration of morphological analysis and POS tagging in Korean

	형태소 분석 및 품사 부착 기법	대표적인 예
1	형태소 분리 - 복원 + 품사 부착	변형된 CYK 알고리즘 이용
2	형태소 분리 + 복원 - 품사 부착	two-level 모델
3	복원 + 형태소 분리 + 품사 부착	기분석 사전, 어절 패턴
4	복원 - 형태소 분리 + 품사 부착	2단계 확률 모델
5	복원 - 형태소 분리 - 품사 부착	3단계 확률 모델
6	<b>단어 분리 - 복원 - 품사 부착</b>	<b>본 논문의 제안 모델</b>

‘+’: parallel processing, ‘-’: sequential processing

Table 4.1의 1번의 ‘분리-복원+품사 부착’ 기법의 대표적인 예가 변형된 CYK 알고리즘을 이용한 Tabular 과상법(김성용, 1987)이다. 이 방법은 기존의 자동채점 시스템에서 사용하고 있는 형태소 분석 및 품사 부착 기법에 해당한다. 본 논문에서 제안하는 기법은 다른 기법들과 달리 형태소 분리 대신 단어 분리를 수행한다. 그 후, 형태소의 원형 복원, 형태소 품사 부착의 과정을 차례로 진행한다. Fig. 4.1은 본 논문에서 제안하는 형태소 분석 및 품사 부착 기법의 전체 구조도이다.

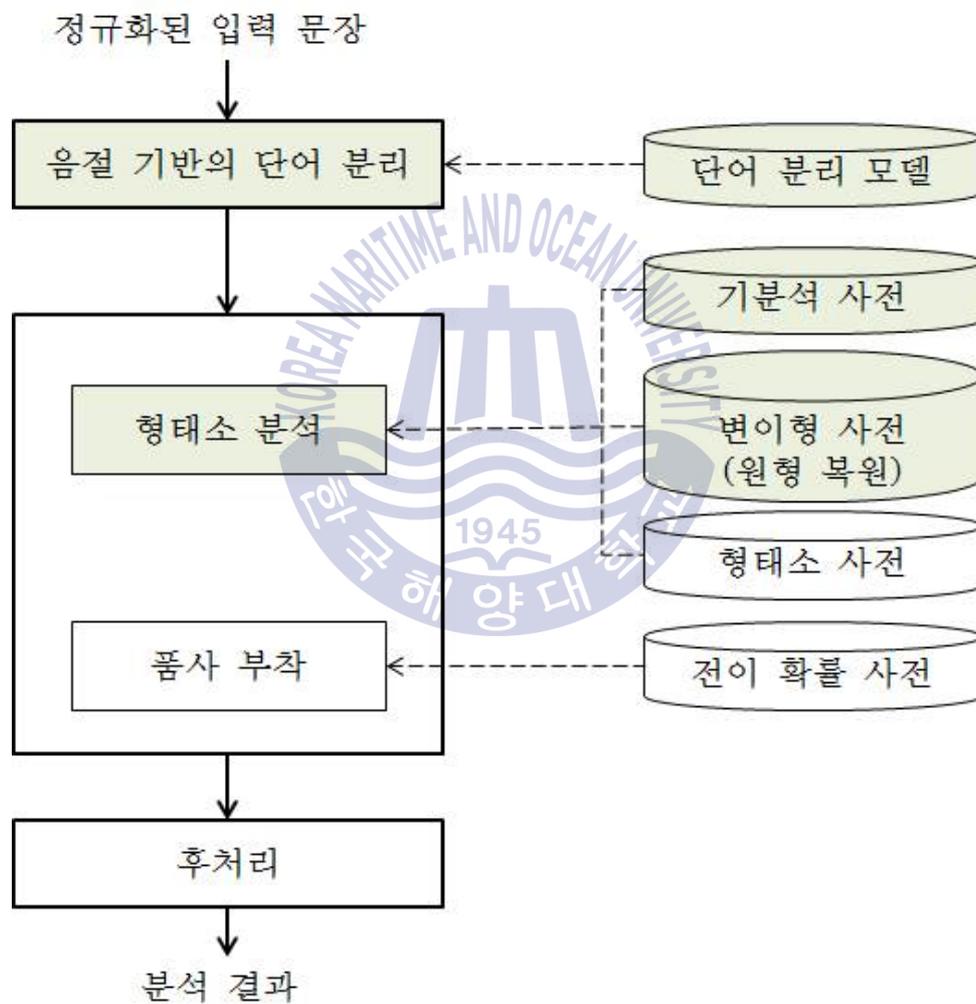


Fig. 4.1 Flow diagram for the proposed method for morphological analysis (Shading the changed part with the existing method shaded)

제안하는 형태소 분석 및 품사 부착 기법에서는 세종 말뭉치로 생성한 음절 기반의 단어 분리 모델을 이용해 기계학습 방법의 하나인 CRF로 ‘음절 기반의 단어 분리’ 작업을 진행한다. 그 후 어절 단위로 기분석 사전을 탐색하여 분석 후보를 생성한다. 이후, 분리된 단어에 대해 변이형 사전을 탐색을 수행하여 형태소 원형 복원 작업을 수행하고 형태소 사전을 이용해 분석 후보를 추가로 생성한다. 이 과정은 4.5절에서 자세히 설명한다. 형태소 분석 작업을 끝낸 후, 전이 확률 사전을 이용하여 가장 적절한 품사 부착 결과를 생성한 뒤 후처리를 통해 최종 분석 결과를 출력한다.

## 4.2 단어 분리 모델 및 사전 생성

단어 분리 모델과 형태소 분석에 필요한 사전들은 국립국어원에서 제공하는 세종 말뭉치를 이용하여 만들 수 있다. Table 4.2는 세종 말뭉치의 구조이다.

Table 4.2 The structure of the SEJONG corpus

번호	어절	형태소 분석 결과
1	집은	집__01/NNG+은/JX
2	창작의	창작/NNG+의/JKG
3	원천이라는	원천__02/NNG+이/VCP+라는/ETM
4	그는	그__01/NP+는/JX
5	옷	옷__01/NNG
6	못지않게	못지않/VA+게/EC
7	공간이	공간__05/NNG+이/JKS
8	주는	주__01/VV+는/ETM
9	미학을	미학/NNG+을/JKO
10	중요시해	중요시/NNG+하/XSV+어/EC
11	왔다.	오__01/VX+왔/EP+다/EF+./SF

Table 4.2와 같이 세종 말뭉치는 한 줄이 (어절, 형태소 분석 결과)로 이루어져 있다. Table 4.2의 1번 어절인 “집은”을 형태소 분석한 결과는 “집\_\_01/NNG+은/JX”로 나타나는데, “집\_\_01”에서 “\_\_01”은 “집”이라는 명사의 단어가 여러 개 존재하므로 해당 단어가 어떤 의미인지 명확하게 해주

기 위한 식별자이다. 이 식별자들은 기분석 사전, 변이형 사전, 형태소 사전을 구축할 때 필요 없는 정보이므로 삭제해야 한다.

Table 4.3은 음절 기반의 단어 모델 및 사전 생성에 사용한 세종 말뭉치의 통계치이다.

Table 4.3 Statistics of the SEJONG corpus

구분	통계
총 문장 수	700,014
문장 당 평균 어절 수	10.01
총 어절 수	7,007,069
어절 당 평균 형태소 수	2.24
총 형태소 수	15,661,729
단순화 한 품사 태그 수 (실제 품사 태그 수)	37 (52)

사전 구축을 위해<sup>3)</sup> 사용한 문장 수는 총 700,014개이며 문장 당 평균 어절 수는 10.01개였다. 하나의 어절 당 평균 형태소 수는 2.24개로 나타났다. 실제 세종 말뭉치에서 사용 중인 품사 태그의 종류는 52개로 조사(JJ)처럼 하나의 품사를 주격 조사(JKS), 보격 조사(JKC), 관형격 조사(JKG), 목적격 조사(JKO) 등과 같이 세분화하여 사용하고 있었다. 따라서 실제 사전 구축 및 제안하는 형태소 분석 및 품사 부착 기법에서는 세분화 된 품사 태그들의 경우 그를 대표할 수 있는 하나의 품사로 통일하여 37개의 품사만을 사용한다. 자세한 품사 태그표는 <부록 A>에 수록하였다.

3) 사전의 경우, 품사 부착기의 성능 테스트를 위해 전체 문장의 80%만 사용했다.

세종 말뚱치로부터 음절 기반의 단어 분리 모델 및 사전을 생성하는 과정은 Fig. 4.2와 같다.

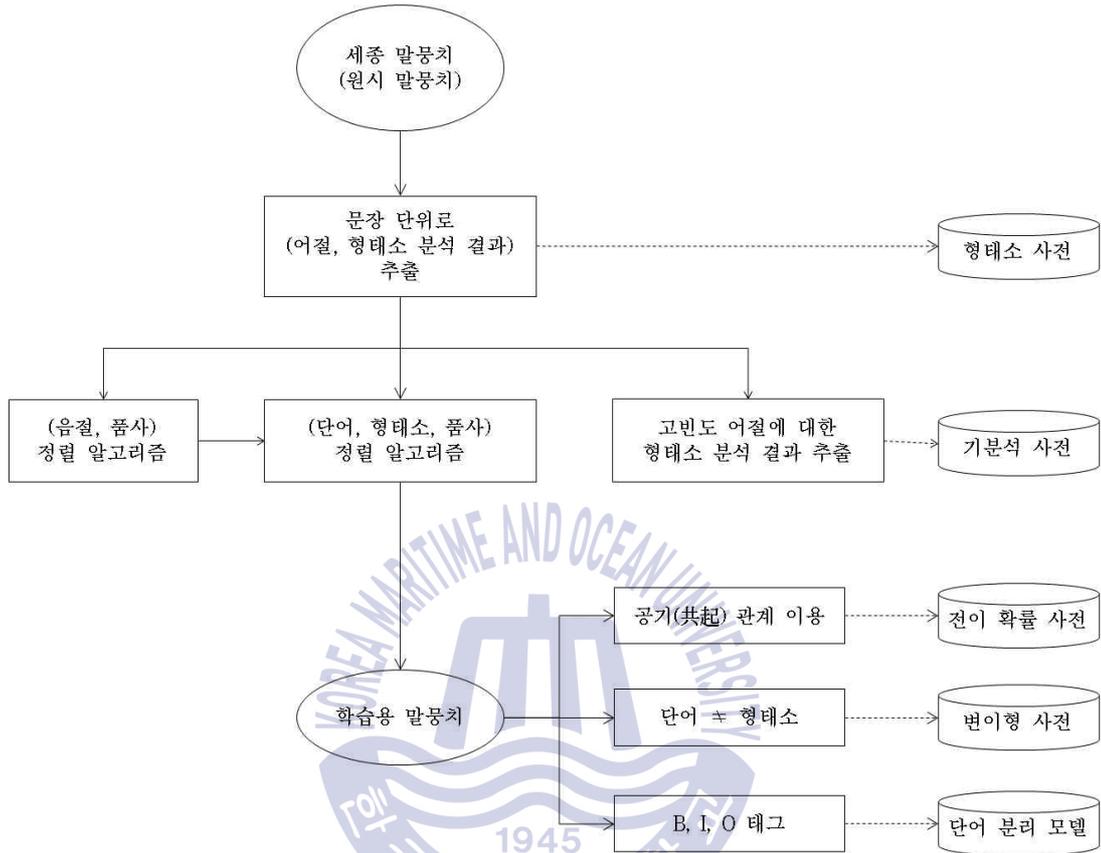


Fig. 4.2 The creation process of required models and dictionaries

세종 말뚱치로부터 ‘문장 단위로 (어절, 형태소 분석 결과)’를 추출하여 형태소 사전을 구축한다. 이때, 고빈도 어절에 대한 형태소 분석 결과를 추출하여 기분식 사전을 구축한다. (어절, 형태소 분석 결과)를 추출한 결과에서 ‘(음절, 품사) 정렬 알고리즘’과 ‘(단어, 형태소 품사) 정렬 알고리즘’을 통해 학습용 말뚱치를 생성한다. 생성한 학습용 말뚱치에서 어휘 확률과 문맥 확률을 이용하여 품사의 문맥 확률과 어휘 확률을 구하여 전이 확률 사전을 생성한다. 학습용 말뚱치에서 단어와 형태소가 일치하지 않는 결과를 변이형 사전에 추가한다. 단어 분리 모델은 학습용 말뚱치의 최종 결과에 B, I, O 태그를 부착한 결과로 생성할 수 있다.

(1) 형태소 사전 생성 : 문장 단위로 (어절, 형태소 분석 결과)를 추출

형태소 사전은 문장 단위로 (어절, 형태소 분석 결과)를 추출하는 알고리즘을 통해 생성할 수 있다. 해당 알고리즘은 Fig. 4.3과 같이 비교적 간단하게 구현할 수 있다.

```
ALGORITHM 문장 단위의 어절 형태소 분석쌍 출력 (세종 말뭉치[]):  
  eojeol[] = []  
  morphs[][] = [[]]  
  tags[][] = [[]]  
  cnt = 1  
  FOR i = 1부터 n1까지 (n1 = 세종 말뭉치의 줄 수)  
    IF 세종 말뭉치[i]가 pattern( "#[0-9]+/[0-9]+" )을 포함 THEN  
      CONTINUE  
    ELSE  
      temp[] = tokenizer(세종 말뭉치[i], '\t')  
      eojeol[cnt] = temp[1]에서 pattern( "[0-9]+" )를 지운 결과  
      temp2[][] = tokenizer( tokenizer(temp[2], '+'), '/' )  
      FOR j = 1 부터 n2까지 (n2 = temp2[]의 원소의 개수):  
        morphs[cnt][j] = temp2[j][1]  
        tags[cnt][j] = temp2[j][2]  
      cnt += 1  
    END FOR  
  RETURN eojeol[], morphs[], tags[]
```

Fig. 4.3 Pseudocode for extracting pairs (word, its morphological structure) by sentence

Fig. 4.3에서 ‘변수 이름[]’ 형태로 기술된 변수들은 배열(array) 형태의 변수이다. *sents*[]는 (어절, 형태소 분석 결과) 쌍을 문장 단위로 저장하는 변수이다. 세종 말뭉치[i]는 세종 말뭉치의 i번째 줄 수를 의미한다. *pattern*(“정규표현식”)는 “ ”안의 문장이 정규표현식임을 나타내는 함수

이다. *eojeol*[]은 세종 말뭉치에서 추출한 어절들을 저장하는 변수이며, *morphs*[][]는 *eojeol*[]의 각 원소(어절)의 형태소를 저장하는 변수이다. *tags*[][]는 *morphs*[][]에 저장된 각 형태소에 부착된 품사를 저장하는 변수이다. *tokenizer*(변수[],*sep*)는 변수[]의 각 원소를 *sep* 기준으로 분리한 결과를 배열 형태로 반환한다. 예를 들어 Table 3.2의 12번 ‘중요시해’라는 어절에 대해서 Fig 4.3의 알고리즘을 적용하면 각 변수에 저장되는 값들은 아래와 같다.

*eojeol*[12] = ‘중요시해’

*morphs*[12][1] = ‘중요시’, *morphs*[12][2] = ‘하’, *morphs*[12][3] = ‘어’

*tags*[12][1] = ‘NC’, *tags*[12][2] = ‘XV’, *tags*[12][3] = ‘EF’

Fig. 4.3의 알고리즘으로 얻은 *morphs*[][]와 *tags*[][]를 정렬한 결과를 이용하여 형태소 사전을 생성한다.

## (2) 기분석 사전 생성

(1)의 과정이 끝난 후, (어절, 형태소 분석 결과)에 대한 빈도수를 센 후, 고빈도에 해당하는<sup>4)</sup> (어절, 형태소 분석 결과)를 기분석 사전에 추가한다.

## (3) 학습용 말뭉치 생성 :

(음절, 품사) 정렬 알고리즘 + (단어, 형태소, 품사) 정렬 알고리즘

‘(음절, 품사) 정렬 알고리즘’과 ‘(단어, 형태소, 품사) 정렬 알고리즘’은 Fig. 4.4와 같이 어절의 음절과 형태소 분석 열이 일치하지 않는 경우를 고려하여 학습용 말뭉치를 생성한다. 예를 들어 “중요시해”라는 어절의 경우, 어절의 “중요시”라는 단어와 해당 단어의 형태소 분석 결과인 “중요시/NC”를 연결하는 것은 비교적 간단하다. 차례대로 음절의 일치 여부를 확인하면 되기 때문이다. 그러나 “중요시해”의 “해”라는 단어는 “하

4) 기분석 사전 생성을 위한 고빈도의 기준은 임의로 결정한다.

/XV”와 “어/EF”와 동시에 연결되어야 한다.

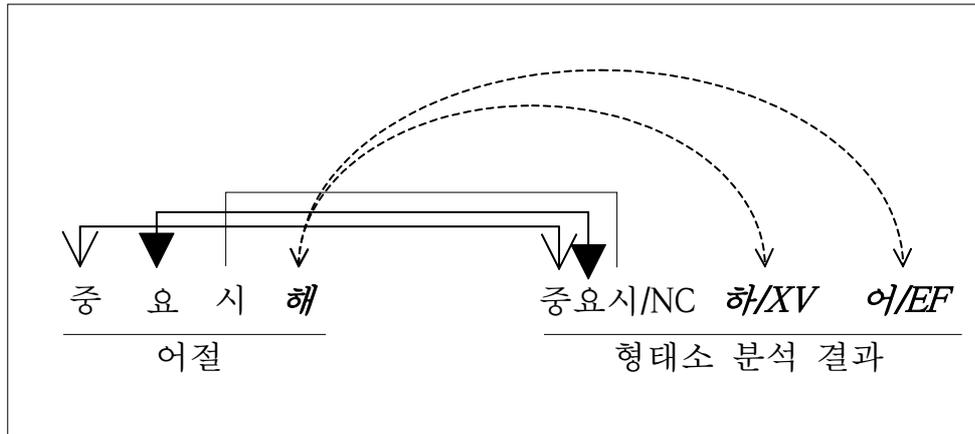


Fig. 4.4 Alignment of syllables and a result of morphological analysis

이런 사항을 고려하기 위해 Fig. 4.5의 알고리즘으로 각 문장의 어절에 대해 음절과 품사를 정렬하기 위한 전처리 작업을 수행한다.



```

ALGORITHM 음절 단위 품사 정렬 (eojeol[], morphs[[[]], tags[[[]]):
  syllable[[]] = [[]]
  syllable_tags[[]] = [[]]
  FOR i = 1부터 n1까지 (n1 = eojeol[]의 원소의 개수)
    cnt = 1
    FOR j = 1부터 n2까지 (n2 = morphs[i][[]]의 원소의 개수)
      FOR k = 1부터 n3까지 (n3 = morphs[i][j]의 음절의 수)
        syllable_tags[i][cnt] = tags[i][j]
        syllableo[i][cnt] = morphs[i][j][k]
        cnt += 1
      END FOR
      syllable_tags[i][cnt] = '+', syllable[i][cnt] = '+'
      cnt += 1
    END FOR
    syllable[i][[]]의 마지막 '+' 값 삭제 //어절의 끝이기 때문
    syllable_tags[i][[]]의 마지막 '+' 값 삭제 //어절의 끝이기 때문
  END FOR
  RETURN syllable[[[]], syllable_tags[[[]]

```

Fig. 4.5 Pseudocode for extracting pairs (syllable, morphological structure) by sentence

Fig. 4.5의 알고리즘에서  $syllable[[]] = [[]]$ 와  $syllable\_tags[[]] = [[]]$ 는 형태소 분리 모델을 만들기 위한 정보를 어절 별로 담고 있다. 차이가 있다면  $syllable[[]]$ 는 음절 정보를 저장하고,  $syllable\_tags[[]]$ 는 품사 정보를 저장한다. “너를 중요시해”라는 문장에 대해서 Fig. 4.5의 알고리즘을 적용하면 Table 4.4의 결과를 얻을 수 있다.

Table 4.4 The Result of applying Fig. 3.11 algorithm to “너를 중요시해”

<i>i</i>	어절	단어	<i>j</i>	형태소		<i>cnt</i>	음절 정보	품사 정보
				분석 결과				
				형태소	품사			
1	너를	너	1	너	NP	1	너	NP
						2	+	+
		를	2	를	JJ	3	를	JJ
						4	+	+
2	중요시해	중요시	1	중요시	NC	1	중	NC
						2	요	NC
						3	시	NC
						4	+	+
		해	2	하	XV	5	하	XV
						6	+	+
						3	어	EF
						7	어	EF

위의 테이블에서 음절 정보는  $syllable[i][cnt]$ 가 담고 있는 값이며, 품사 정보는  $syllable\_tag[i][cnt]$ 가 담고 있는 값이다.  $i$ 는 각 문장에서 어절의 순서를 나타내고,  $j$ 는  $i$ 번째 어절이 가지고 있는 형태소의 순서를 나타낸다. “중요시해”라는 어절의 경우 “중요시”와 “해”라는 단어로 분리할 수 있으며 각 단어는 다시 더 작은 단위의 형태소로 분리된다. 음절 정보의 값이 ‘+’인 곳에서 띄어쓰기하여 형태소 분리를 할 경우, “중요시해”의 형태소 분리 결과는 “중요시 하 어”가 될 것이다. 이때 형태소 “하 어”를 단어 “해”로 매칭을 시키기 위한 정렬 알고리즘이 필요하다.

위의 예시에서 “해”는 어미이나, 또 다른 단어 “해(연(年)이나 태양(太陽)의 의미를 지닌 단어 등)”는 그 자체로 다른 품사를 지닐 수 있기 때문이다. 이처럼 한국어 형태소의 중의성(重義性) 문제와 단어의 음절과 형태소의 분석 열이 일치하지 않는 경우를 해결하기 위해 단어와 형태소를 정렬해야 한다. 이를 위해 python의 difflib 라이브러리<sup>5)</sup>에서 제공하는 SequenceMatcher 함수<sup>6)</sup>를 사용했다. SequenceMatcher 함수의 동작 방식은

5) <https://docs.python.org/2/library/difflib.html#>

6) <https://docs.python.org/2/library/difflib.html#sequencematcher-objects>

diff 알고리즘(Heckel, 1978)과 유사하다. Fig. 4.6은 Table 4.4의 결과로 단어와 형태소, 품사를 정렬하기 위한 Python 코드이다.

```
def align_word2morph(eojeol, syllable, syllable_tags):
    """
    @para     eojeol   : Fig. 3.11의 eojeol[]에 해당
    @para     syllable : Fig. 3.11의 syllable[][]에 해당
    @para     syllable_tags : Fig. 3.11의 syllable_tags[][]에 해당
    @var     sentence_result : 문장의 각 어절 별로 (단어, 형태소, 품사) 정렬 결과를 저장
    @var     result : 어절 별로 (단어, 형태소, 품사) 정렬 결과를 저장
    """

    sentence_result = []

    for i in range(0, len(eojeol)):
        s = difflib.SequenceMatcher(None, eojeol[i], syllable[i])
        result = []
        for op, i1, i2, j1, j2 in s.get_opcodes():
            result.append([op, eojeol[i][i1:i2], syllable[i][j1:j2], syllable_tags[i][j1:j2]])
            """
            @var op : 정렬 결과 (equal, insert, delete, replace가 있음)
            @var i1 : eojeol[i]에서 정렬한 결과가 시작되는 위치
            @var i2 : eojeol[i]에서 정렬한 결과가 끝나는 위치
            @var j1 : syllable[i]와 syllable_tags[i]를 정렬한 결과가 시작되는 위치
            @var j2 : syllable[i]와 syllable_tags[i]를 정렬한 결과가 끝나는 위치
            """
        sentence_result.append(result)

    return sentence_result
```

Fig. 4.6 Python code for aligning (word, morpheme, and POS tag) using the SequenceMatcher function in difflib

Table 4.4의 결과에 Fig. 4.6의 알고리즘을 적용하면 Fig. 4.7과 같은 결과를 얻을 수 있고, 이를 표로 정리하면 Table. 4.5와 같다.

```
단어:   너를
음절 정보:  너+를
품사 정보:  ('NP', 'JJ')

----- SequenceMatcher() 실행 결과 -----
equal 너 너 NP
insert  + +
equal 를 를 JJ

단어:   중요시해
음절 정보:  중요시+하+머
품사 정보:  ('NC', 'XV', 'EF')

----- SequenceMatcher() 실행 결과 -----
equal 중요시 중요시 NC
replace 해 +하+머 +XV+EF
```

Fig. 4.7 The Results of executing the Fig. 4.6 algorithm for a given “너를 중요시해”

Table 4.5 The Results of executing the Fig. 4.6 algorithm  
for a given “중요시해”

음절 정보	품사 정보	SequenceMatcher 결과			
		정렬 결과	단어	형태소	품사
중	NC	equal	중요시	중요시	NC
요	NC				
시	NC				
+	+	replace	해	+하+어	+XV+EF
하	XV				
+	+				
어	EF				

SequenceMatcher 함수의 정렬 결과는 “equal”, “replace”, “delete”, “insert”가 있으며, 실제 시스템에 나타난 결과를 토대로 재정의한 의미는 Table 4.6에 정리했다. 이 정렬 결과를 이용하여 형태소 분리 모델과 변이형 사전을 만들 수 있다.

Table 4.6 Meaning of aligning results

정렬 결과	의미	비고
equal	단어와 형태소의 분석 결과가 일치	
replace	단어와 형태소의 분석 결과가 다름	분석용 사전
delete	단어에 해당하는 형태소가 없음	
insert	형태소 분리를 위한 구분자(‘+’)	

#### (4) 변이형 사전 생성

변이형 사전을 만들기 위해서는 정렬 결과가 ‘replace’인 (단어, 형태소, 품사) 쌍을 사용한다. Table 4.5의 결과처럼 “중요시해”에서 “해”와 정렬된 형태소 분석 결과는 “+하+어”이다. 여기서 나타난 ‘+’는 형태소 분리를 위한 구분자로, 음절 단위의 단어 분리에서 어절 내의 띄어쓰기를 의미하는 기호이다. 따라서 “+하+어”나 “하+어+”처럼 ‘+’가 형태소 분석 결과의 가장 앞이나 가장 마지막에 오는 경우는 삭제하는 처리 과정을 거친 후, 정렬된 단어와 형태소 분석 결과의 정보를 변이형 사전에 추가한다. “중요

시해”의 “해”를 예로 들면 “+하+어”를 “하+어”로 변환한 뒤, ( “해”, “하+어” )의 형태와 같이 변이형 사전을 만든다.

### (5) 단어 분리 모델 생성

단어 분리 모델의 경우 ‘equal’과 ‘replace’의 정보를 이용하여 Table 4.7의 형태와 같이 학습용 말뭉치를 생성한다. 단어 분리 태그는 CRF를 이용한 음절 기반의 띄어쓰기(심광섭, 2011b)에서 사용한 B, I 태그에 O 태그를 추가하여 사용한다.

Table 4.7 The Results of word-segment for a given “중요시해”

음절 정보	품사 정보	SequenceMatcher 결과				단어 분리 태그	
		정렬 결과	단어	형태소	품사		
중	NC	equal	중요시	중요시	NC	중	B
요	NC					요	I
시	NC					시	I
+	+	replace	해	+하+어	+XV+EF	해	B
하	XV					하	I
+	+					어	I
어	EF					어	I

단어 분리 태그 B는 해당 음절에서 띄어쓰기하라는 의미이며, I는 띄어쓰기하지 않는다는 의미이다. O는 어절의 구분에 쓰이는 공백 문자(space)이므로 공백 문자를 ‘\_’로 치환한 후, 띄어쓰기하여 출력하라는 의미이다.

### (6) 전이 확률 사전 생성

전이 확률 사전은 식 (2.3)을 변형한 식 (4.1)에 해당하는 품사의 문맥 확률  $P(t_i|t_{i-1})$ 과 식 (2.4)의 어휘 확률을 사용하여 생성한다.

$$P(t_i|t_{i-1}) \approx \frac{freq(t_{i-1}, t_i)}{freq(t_{i-1})} \quad (4.1)$$

단어의 문맥 확률 대신 품사의 문맥 확률을 사용하는 이유는 한국어 단어의 종류는 무수히 많으며, 계속 생성되고 있으므로 모든 단어에 대해 전이 확률을 구하는 것이 힘들기 때문이다.

### 4.3 음절 기반의 단어 분리

음절 기반의 단어 분리의 기본적인 개념은 CRF를 이용한 음절 기반의 띄어쓰기(심광섭, 2011b)와 유사하다. CRF(Condition Random Fields)는 조건부 확률을 최대화하기 위해 학습된 비방향성 그래프 모델로, 기계학습 방법의 한 종류이다(Laffery *et al.*, 2001; 전길호, 2012). 음절 기반의 띄어쓰기 문제는 주어진 문장의 각 음절에 대하여 띄어쓰기를 할 것인가 말 것인가를 나타내는 태그 부착 문제로 볼 수 있다. “너에게나를보낸다”라는 문장을 CRF를 이용하여 띄어쓰기하는 과정은 Table 4.8과 같이 정리할 수 있다.

Table 4.8 An example of syllable-based spacing

작업 내용	결과
문장을 입력	너에게나를보낸다
CRF를 통해 레이블링	너 에 게 나 를 보 낸 다 B I I B I B I I
결과를 출력	너에게 나를 보낸다

어떤 문장이 입력으로 들어가면, CRF를 통해 해당 문장에 B, I 태그가 부여된다. 여기서 B는 해당 음절에서 새로운 어절이 시작되는 곳이므로 띄어쓰기를 해야 한다는 태그이며, I는 앞의 어절에 속한 음절이 계속되므로 띄어쓰기를 하지 말아야 한다는 태그이다. B, I 태그를 이용하여 “너에게나를보낸다”라는 문장을 띄어쓰기하면 “너에게 나를 보낸다”와 같이 나타난다. 음절 기반의 단어 분리 문제도 이와 같은 방식으로 해결할 수 있으나, 띄어쓰기와 달리 Fig. 4.4처럼 어절과 형태소 분석 결과를 정렬해야 한다는 어려운 점이 있는데, 이를 위해 4.3절의 (3)에서 설명한 학습용 말뭉치를 이용한다.

CRF를 이용한 음절 기반의 단어 분리 자질은 음절 기반의 띄어쓰기 자질과 동일한 자질을 사용한다. 해당 자질의 종류들은 Table 4.9의 ‘자질 종류 사용’ 목록에서 ✓로 표시된 항목들이다.

**Table 4.9** The feature set for morpheme segmentation based on syllables

입력	너에게 나					
	음절	너	에	게	‘ ’	나
	태그	B	B	I	O	B
자질 종류	$w_{i-2}$	$w_{i-1}$	$w_i$	$w_{i+1}$	$w_{i-1}$	
사용	1		✓			
	2			✓		
	3				✓	
	4	✓	✓			
	5		✓	✓		
	6			✓	✓	
	7				✓	✓
	8	✓	✓	✓		
	9		✓	✓	✓	
	10			✓	✓	✓
출력	결과	너 에게 _ 나				

$w_i$ 는 현재의 음절을 의미하며,  $w_{i-1}$ 는 이전의 음절을  $w_{i-2}$ 는 두 번째 앞의 음절이다.  $w_{i+1}$ 는 현재 음절 뒤의 음절이며,  $w_{i+2}$ 는 두 번째 뒤의 음절이다. 이처럼 형태소 분리를 위한 자질은 현재의 음절을 중심으로 2개 앞, 2개 뒤의 음절들을 조합한 10개의 자질과 Table 4.9에 표시하지 않았으나 해당 음절이 한글, 한자, 영어, 숫자 등의 범주 중 어디에 속하는지에 대한 정보도 함께 쓴다. 이 정보는 Fig. 3.5의 의사코드의 *token*의 유형을 구하는 알고리즘에서 *token*을 음절로 바꾸는 것으로 구할 수 있다. 이렇게 학습된 CRF 모델을 이용하여 “너에게 나를 보낸다”라는 문장에 대해 음절 기반의 단어 분리를 하면 Table 4.10의 결과를 얻을 수 있다.

Table 4.10 An example of syllable-based word segmentation

작업 내용	결과
문장을 입력	너에게 나를 보낸다
CRF를 통해 레이블링	너 에 게 ‘ ’ 나 를 ‘ ’ 보 낸 다 B B I O B B O B B B
결과를 출력	너 에 게 _ 나 를 _ 보 낸 다

#### 4.4 제안하는 형태소 분석 기법

‘음절 기반의 단어 분리’가 끝나면 해당 결과를 바탕으로 ‘형태소 분석’을 수행한다. 제안하는 형태소 분석 기법은 기분석 사전과 변이형 사전, 형태소 사전을 탐색하는 방식으로 비교적 간단하게 구현할 수 있다. 형태소 분석 과정은 Fig. 4.8과 같다.

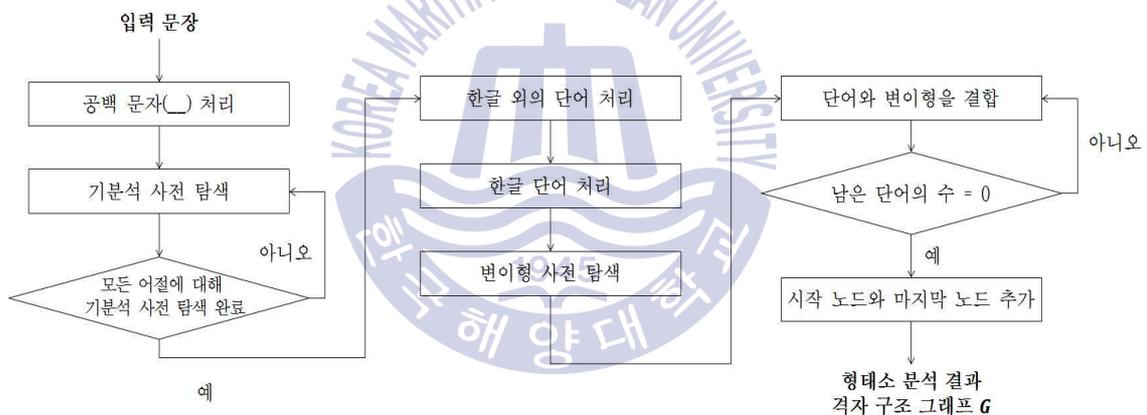


Fig. 4.8 The proposed method for morphological analysis

입력 문장이 들어오면 공백 문자( )를 처리한 후, 각 어절에 대해 기분석 사전 탐색을 수행한다. 모든 어절에 대해 기분석 사전에 대한 탐색을 완료하면 각 단어에 대해 한글 외의 단어, 한글 단어, 변이형 사전 탐색 결과, 단어와 변이형을 결합한 결합어의 형태소 사전 탐색 결과에 대해 형태소 분석 후보를 정점(node)으로 생성하여 격자 구조  $G$ 에 추가한다. 시작 정점과 마지막 정점을 추가하는 것으로 격자 구조  $G$ 를 완성하는 것으로 형태소 분석을 완료한다.

다양한 경우의 수를 고려하기 위해 예제 <4.1>을 이용하여 제안하는 형태소 분석 방법의 각 과정을 (1)~(7)로 구분하여 서술한다.

너 에게 \_ 나 를 \_ 보 낸 다

예제 <4.1>

예제 <4.1>의 문장은 “너에게 나를 보낸다”라는 문장을 4.3절에서 설명한 ‘음절 기반의 단어 분리’를 수행한 결과이다. 제안하는 형태소 분석의 최종 결과는 예제 <4.1>의 문장을 Fig. 4.9과 같은 형태의 격자 구조  $G$ 로 변환한 것이다.

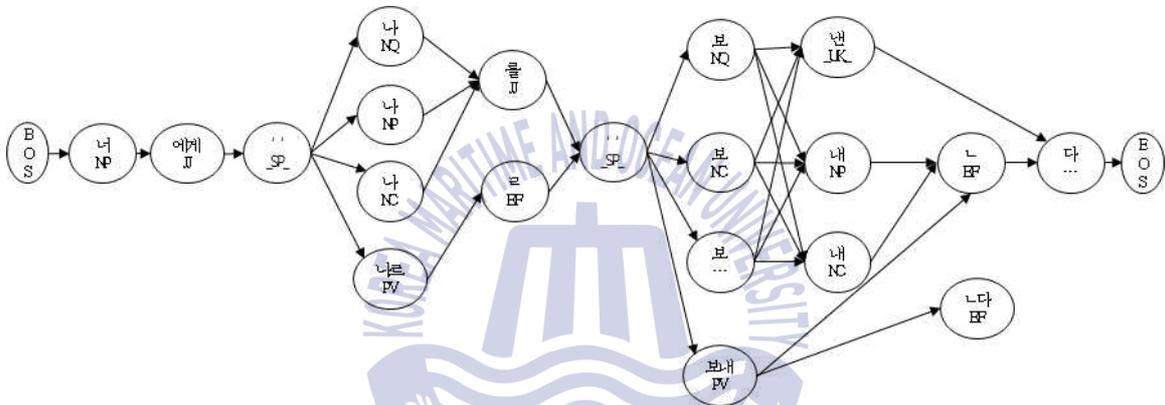


Fig. 4.9 The lattice structure  $G$  as a result of morphological analysis

(1) 공백 문자( ) 처리

형태소 분석에서 가장 먼저 수행하는 작업은 공백 문자에 해당하는 ‘ ’를 처리하는 일이다. 공백 문자의 경우 형태소 분석 결과는 ‘ /\_SP\_’밖에 없으므로, Fig. 4.10과 같이 ‘ /\_SP\_’에 해당하는 정점을 생성하여 그래프에 추가한다.

너    에    게    □    나    를    □    보    낸    다



Fig. 4.10 Adding ‘ /\_SP\_’ nodes in the lattice structure *G*

Fig. 4.10에서 진하게 표시된 부분이 추가가 완료된 정점이며, 연하게 표시된 부분은 앞으로 추가해나갈 정점과 간선들이다. 정점 추가를 완료하면 ‘ ’ 문자를 삭제하여 예제 <4.2>와 같이 어절 단위로 분리한다.

너
에
게
 
나
를
 
보
낸
다

예제 <4.2>

예제 <4.2>에서 □ 단위로 묶인 단어들의 집합이 하나의 어절이다.

## (2) 기분석 사전 탐색

기분석 사전의 탐색 단위는 어절이다. 어절 단위로 기분석 사전을 탐색하여 해당 어절이 기분석 사전의 결과에 존재할 경우, Fig. 4.11과 같이 기분석 사전의 탐색 결과들을 정점으로 만들어 격자 구조  $G$ 에 추가한다.

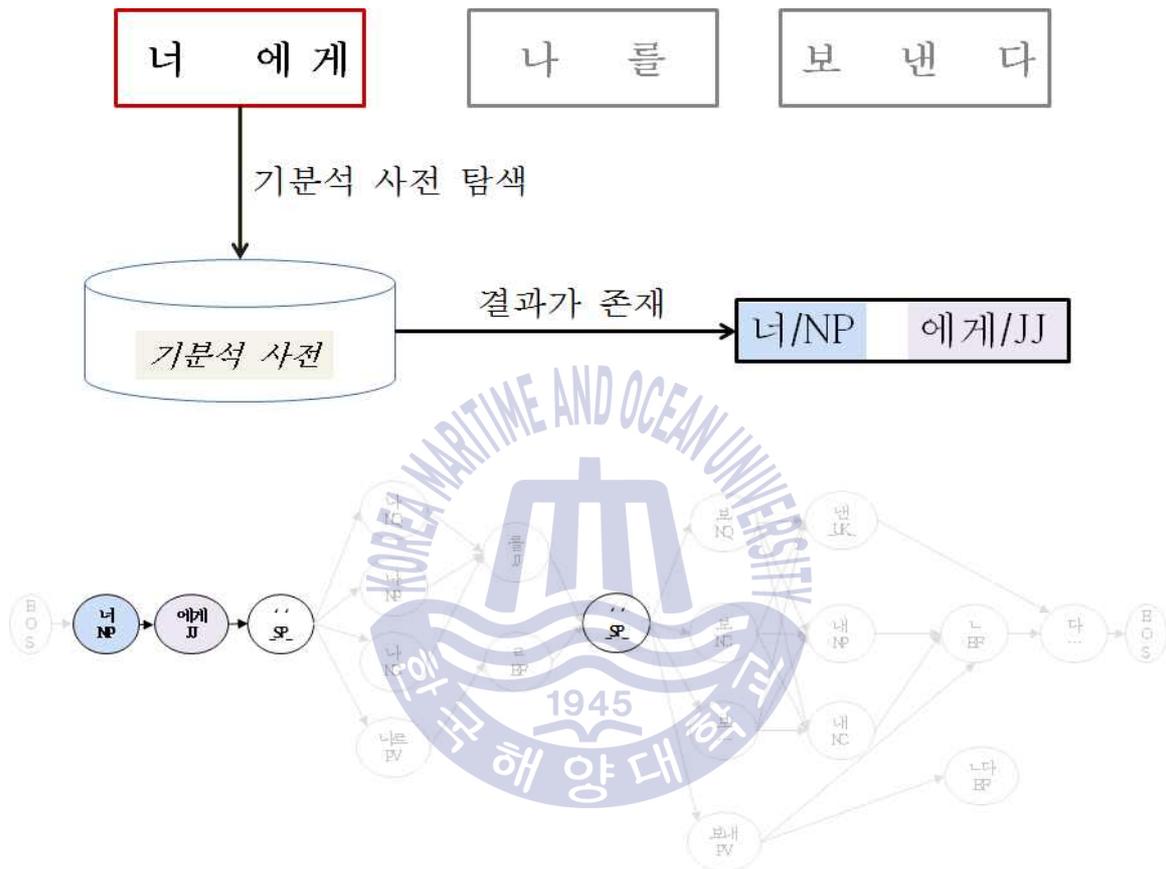
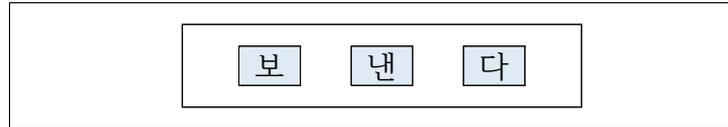


Fig. 4.11 Adding nodes in the lattice structure  $G$  due to lookup of the pre-analyzed dictionary

“너 에 게”라는 어절에 대해 기분석 사전을 탐색한 결과가 “너/NP 에 게 /JJ”일 경우, 격자 구조  $G$ 에 추가할 정점들은 탐색 결과의 형태소 단위이다. 즉, “너/NP 에 게 /JJ”를 정점으로 추가하는 것이 아니라, “너/NP”와 “에 게 /JJ”처럼 형태소 단위로 정점을 추가한 뒤, 각 정점 사이의 간선들을 추가한다. 기분석 사전의 탐색 결과가 여러 개일 경우 각 탐색 결과에 대해 정점을 추가하는 작업을 반복한다. 기분석 사전을 통해 형태소 분석 결과를 추가한 어절들은 정점 추가 작업을 완료한 후 삭제한다.

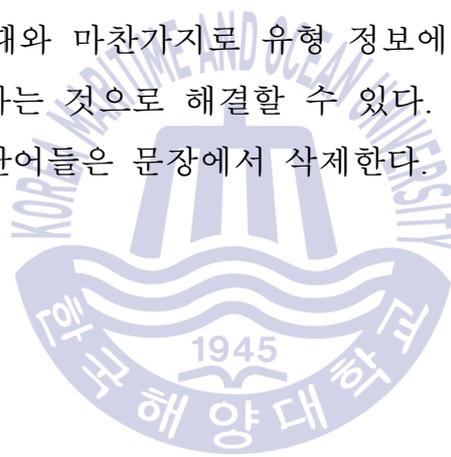
### (3) 한글 외의 단어 처리

기분석 사전 탐색을 완료한 후의 처리 단위는 예제 <4.3>과 같이 각 어절의 단어이다.



예제 <4.3>

예제 <4.3>에서 □ 단위가 단어를 나타낸다. 단어 단위로 Fig. 3.5의 의사코드 중 *token*의 유형을 구하는 것과 같은 방식으로 단어의 유형을 결정한다. 한글 이외의 단어들은 각 유형에 따라 지정된 형태소가 존재하므로 \_를 처리했을 때와 마찬가지로 유형 정보에 따른 정점을 생성하여 격자 구조  $G$ 에 추가하는 것으로 해결할 수 있다. 한글 외의 단어 처리를 통해 정점을 추가한 단어들은 문장에서 삭제한다.



#### (4) 한글 단어 처리

한글 단어의 처리 단위는 (3) 한글 외의 단어 처리 작업을 수행한 뒤, 남아 있는 단어들이다. 각 단어에 대해 형태소 사전을 탐색한다. 형태소 사전의 탐색 결과가 존재하는 경우, Fig. 4.12와 같이 탐색 결과를 정점으로 구성하여 격자 구조  $G$ 에 추가한다.

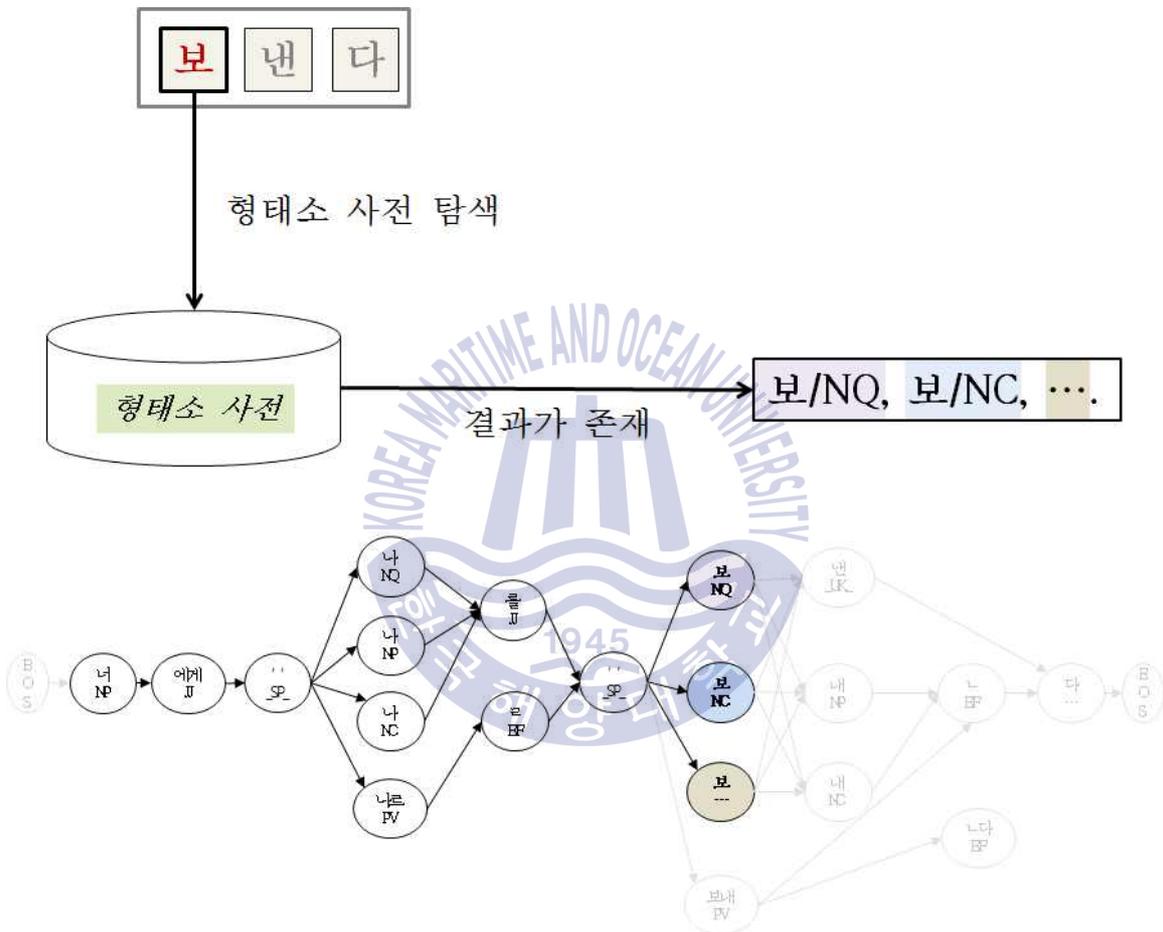


Fig. 4.12 Adding nodes in the lattice structure  $G$  due to lookup of the morphological dictionary

만약 해당 단어에 대한 형태소 사전의 탐색 결과가 존재하지 않는다면, Fig. 4.13과 같이 미등록어(Unknown word)를 뜻하는 ‘\_UK\_’를 품사로 하는 정점을 생성하여 격자 구조  $G$ 에 추가한다.

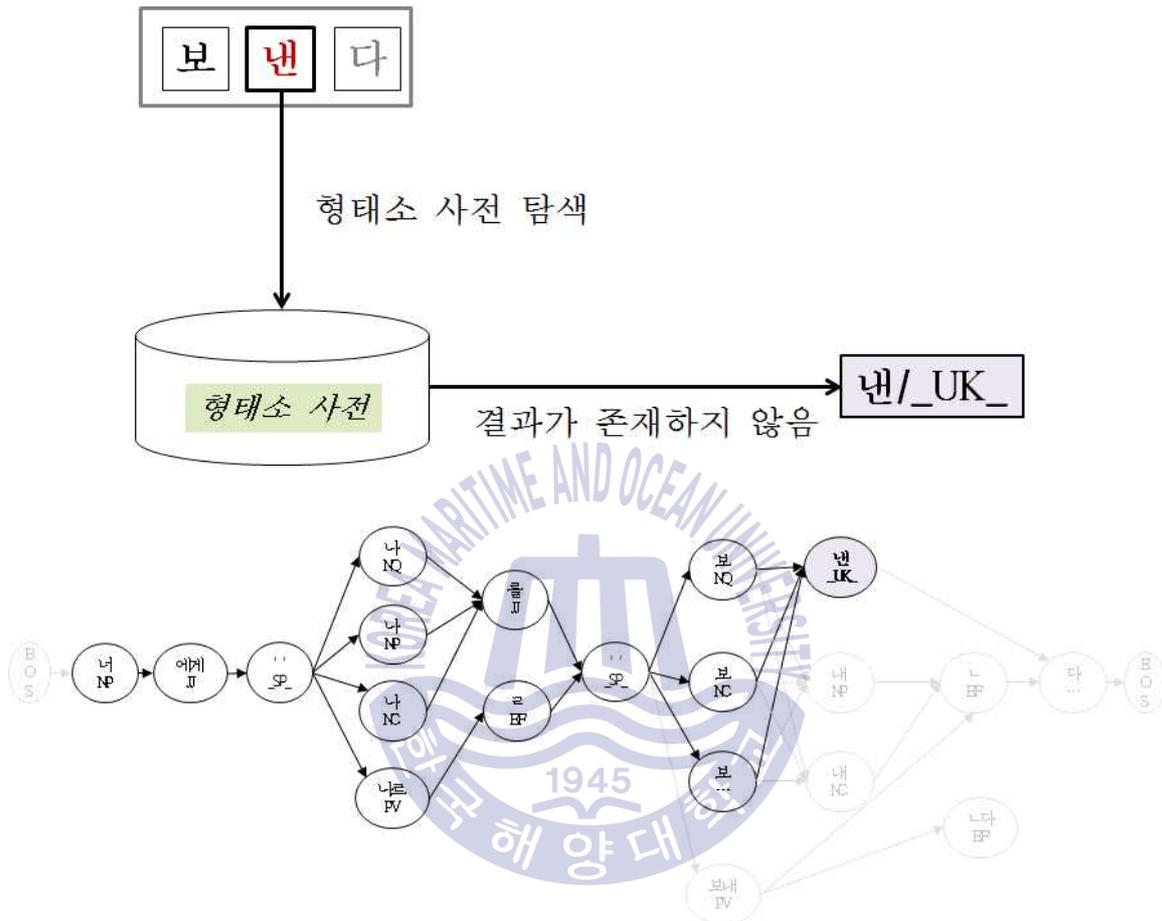


Fig. 4.13 Adding unknown word ( ‘word/\_UK\_’ ) nodes in the lattice structure  $G$

(1)-(3)까지의 과정과 달리 (4)의 경우 처리가 완료된 단어들을 삭제하지 않고 그대로 남겨둔다.

### (5) 변이형 사전 탐색

변이형 사전의 탐색 단위는 예제 <4.3>에서 표시한 단어 단위이다. 변이형 사전의 탐색도 (4)와 유사한 방식으로 진행한다. Fig. 4.14의 예제와 같이 “보 낸 다”의 “낸”의 경우 변이형 사전의 탐색 결과가 “내 ㄴ”이 된다. 변이형 사전의 탐색 결과가 존재하는 경우 해당 변이형에 대해 형태소 사전의 탐색을 진행하여 그 결과를 정점으로 추가한다.

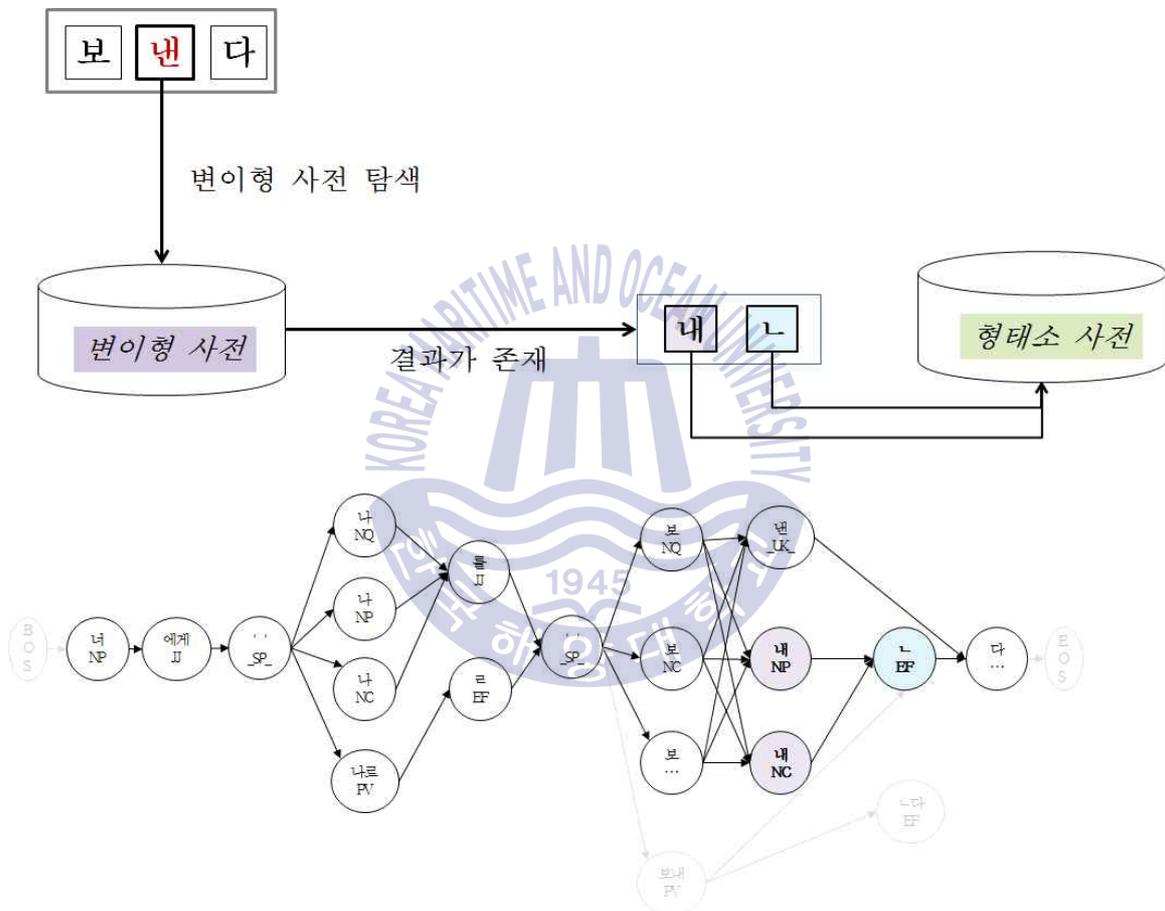


Fig. 4.14 Adding nodes in the lattice structure  $G$  due to lookup of the variant dictionary

이때, ‘(6) 단어와 변이형을 결합’하여 형태소 사전을 탐색하는 작업을 진행하기 위해, Fig. 4.15와 같이 각 단어의 변이형의 첫 번째 음절과 마지막 음절에 대한 정보를 해당 단어 정보에 추가한다.

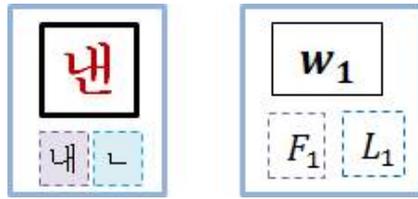


Fig. 4.15 Adding the first syllable and the last syllable of variant to word information

“넌”의 경우 변이형 “내 ㄴ”에서 첫 번째 음절인 “내”와 마지막 음절인 “ㄴ”이 단어 정보에 추가된다. Fig. 4.15의 오른쪽 그림은 변이형에 음절 정보를 추가한 일반적인 구조이다.  $w_1$ 은 단어이며,  $F_1$ 은 해당 단어의 변이형의 첫 번째 음절 정보들의 집합이다.  $L_1$ 은 해당 단어의 변이형의 마지막 음절 정보들의 집합이다.



### (6) 단어와 변이형을 결합

단어와 변이형을 결합하는 방식은 Fig. 4.16과 같이 4가지가 있다.

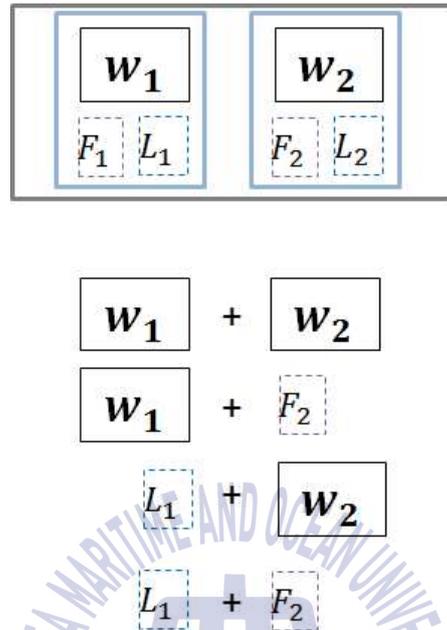


Fig. 4.16 The combining method of words and syllables of variants

결합은 두 단어 간에만 이루어진다. 현재 단어가  $w_1$ 이고 다음 단어가  $w_2$ 일 경우 결합이 일어나는 방식은 아래의 4가지이다.

- 현재 단어( $w_1$ ) + 다음 단어( $w_2$ )
- 현재 단어( $w_1$ ) + 다음 단어의 변이형의 첫 번째 음절( $F_2$ )
- 현재 단어의 변이형의 마지막 음절( $L_1$ ) + 다음 단어( $w_2$ )
- 현재 단어의 변이형의 마지막 음절( $L_1$ ) + 다음 단어의 변이형의 첫 번째 음절( $F_2$ )

각 단어와 각 변이형이 결합하여 생성된 단어들에 대해 (4)에서 설명한 방식으로 형태소 사전을 탐색하여 격자 구조  $G$ 에 결과를 추가한다. 이때, 형태소 사전에 탐색 결과가 존재하지 않는 결합어들은 정점으로 추가하지 않는다. 이 과정을 그림으로 나타내면 Fig. 4.17과 같다.

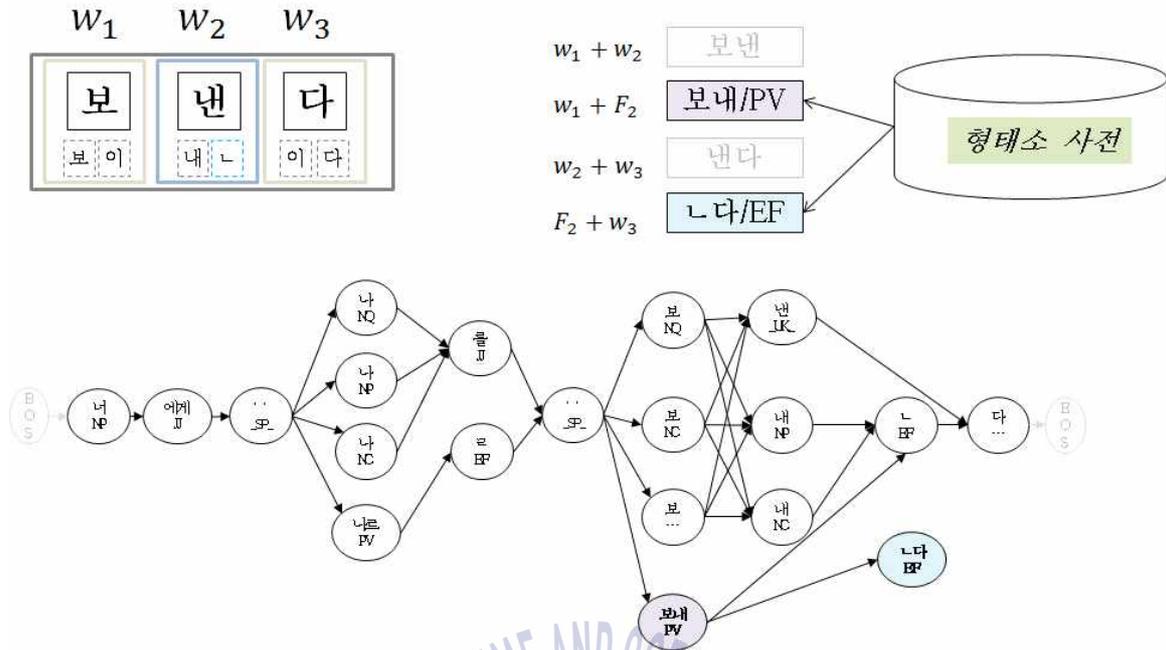


Fig. 4.17 Adding nodes in the lattice structure  $G$  due to lookup of the morphological dictionary for combined words

### (7) 시작 정점과 마지막 정점 추가

(3)-(6)의 과정을 반복하여 모든 단어에 대한 형태소 후보 생성이 끝나면 Fig. 4.18과 같이 시작 정점(BOS)과 마지막 정점(EOS)을 그래프  $G$ 에 추가한다.

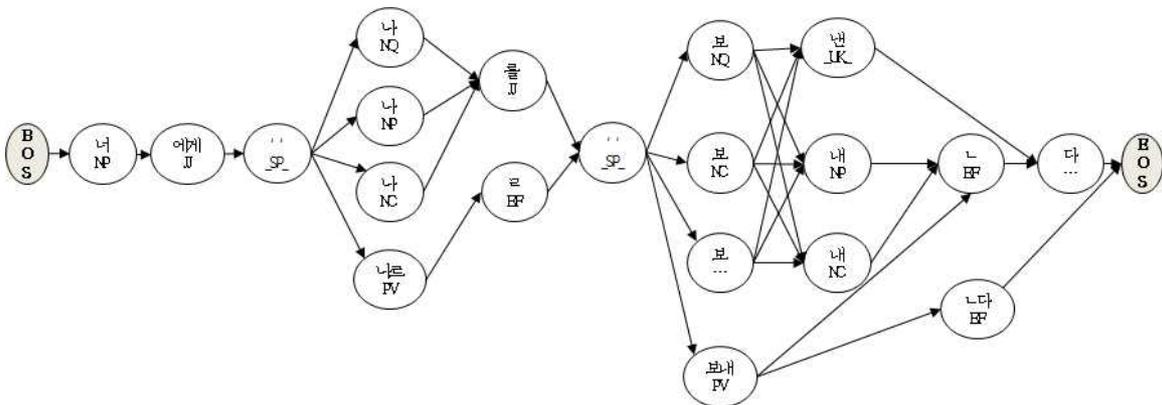


Fig. 4.18 Adding the BOS (first) node and the EOS (last) node

## 4.5 통계기반의 품사 부착

앞의 과정까지 완료하면 생성한 형태소 분석 후보 중에서 어떤 형태소가 가장 적절한지를 결정해야 한다. 이 작업을 ‘품사 부착’이라고 한다. 품사 부착 문제는 생성한 형태소 분석 후보들을 정점의 집합( $V$ )으로 하고, 현재 정점과 다음 정점을 연결하는 연결선 위에 가중치를 적재한 그래프  $G=(V,E)$ 를 구성하여 해결한다(김재훈 1996; 나승훈 외 2013).

품사 부착은 형태소 분석 과정에서 구성한 격자 구조  $G$ 의 시작 정점으로부터 마지막 정점으로 가는 가장 최적의 경로를 찾는 것으로 해결할 수 있다. 시작 정점은 패딩 문자인 ‘BOS’의 정보를 갖는 정점이며, 마지막 정점은 패딩 문자인 ‘EOS’의 정보를 갖는 정점이다. 각 정점의 간선 위에 적재할 가중치는 식 (2.4)의 어휘 확률  $P(t_i|w_i)$ 와 식 (4.1)의 품사의 문맥 확률  $P(t_i|t_{i-1})$ 을 식 (4.2)에 대입하여 구한다.

$$\begin{aligned} \text{가중치} &= P(t_i|t_{i-1}) \times P(t_i|w_i) \\ &\approx -(\log(P(t_i|t_{i-1})) + \log(P(t_i|w_i))) \\ &\approx -\left(\log\left(\frac{\text{freq}(t_{i-1}, t_i)}{\text{freq}(t_{i-1})}\right) + \log\left(\frac{\text{freq}(t_i, w_i)}{\text{freq}(w_i)}\right)\right) \end{aligned} \quad (4.2)$$

실제 태그의 문맥 확률과 어휘 확률을 곱하는 수식을 각 확률에  $\log$ 를 취하여 더하는 식으로 변환하여 사용한다.

Fig. 4.18의 격자 구조의 간선에 식 (4.2)의 가중치를 적재하여 생성한 가중치 그래프  $G$ 는 Fig. 4.19와 같다.



## 제 5 장 실험 및 평가

이 장에서는 앞에서 제안한 형태소 분석 및 품사 부착 기법의 성능을 평가하고, 오류유형을 분석하여 차후 연구 방향을 모색한다.

### 5.1 성능 평가 대상

형태소 분석 및 품사 부착 기법의 성능을 평가하기 위해 사용한 평가 자료는 국립국어원에서 제공하는 세종말뭉치(국립국어원, 2011)와 2014년에 시행된 “국가수준 학업성취도 평가”의 국어, 사회, 과학 문항의 서답형 문항이다(한국교육과정평가원, 2014).

#### 5.1.1 세종 말뭉치

성능 평가에 사용된 세종 말뭉치의 문장은 15,003개이다. 성능 평가에 사용된 말뭉치의 정보는 Table 5.1과 같다.

Table 5.1 Statistics of the SEJONG corpus for performance evaluation

구분	통계
총 문장 수	15,003
문장 당 평균 어절 수	11.05
총 어절 수	1,657,094
어절 당 평균 형태소 수	2.26
총 형태소 수 (어절 단위)	3,736,666
단순화 한 품사 태그 수	28

### 5.1.2 “2014년 국가수준 학업성취도 평가” 답안

성능 평가를 위해 사용한 “2014년 국가수준 학업성취도 평가(2014 NLSA, National Level Student Assessment)”의 대상 문항은 국어, 사회 과목의 서답형 문항을 응시한 실제 학생들의 답안 중 일부이다(한국교육과정평가원, 2014). 해당 답안은 교육과정평가원과 진행하는 “한국어 서답형 문항 자동채점 프로그램 개발 및 실용성 검증” 프로젝트를 위해 제공받은 예시 답안들이며, 보안의 문제를 위해 학생들의 정보는 블라인드 처리 되어있다. 과목별 평가 문항의 정보는 Table 5.2와 같다.

Table 5.2 Statistics of the 2014 NLSA for performance evaluation

대상 학년	고2		중3		평균
	국어	국어	국어	사회	
문항 번호	2-(2)	6-(1)	6-(2)-ㄱ	2	
답안 수	7,965	7,965	7,453	7,442	7,706
어절 수	30,498	26,935	39,010	14,713	27,789
문장 당 평균 어절 수	3.83	3.38	5.23	1.98	3.61

### 5.2 성능 평가 척도

본 논문의 성능 평가 척도는 Table 5.3에 정리된 재현율(recall), 정확률(precision)이다. 재현율은 형태소 분석 기법의 성능 평가 척도이며, 정확률은 품사 부착 기법의 성능 평가 척도이다.

Table 5.3 Measures for performance evaluation

재현율	$\frac{\sum   \text{정답 형태소} \cap \text{형태소 분석 후보}  _{\text{어절 단위}}}{\sum   \text{정답의 형태소}  _{\text{어절 단위}}} \times 100$
정확률	$\frac{\sum   \text{정답의 형태소} \cap \text{품사 부착 결과의 형태소}  _{\text{어절 단위}}}{\sum   \text{품사 부착 결과의 형태소}  _{\text{어절 단위}}} \times 100$

형태소 분석의 성능 평가 척도인 재현율은 어절 단위로 시스템의 형태소 분석 후보 결과가 정답 형태소를 포함하고 있는지를 확인하여 측정한다. 품사 부착 기법의 정확률은 어절 단위로 평가 말뭉치의 정답과 시스템의 품사 부착 결과가 얼마나 일치하는지 확인한다. 세종 말뭉치의 경우는 형태소 분석 및 품사 부착이 완료된 정답이 존재하므로 재현율과 정확률을 측정할 수 있으나, “2014 국가수준 학업 성취도 평가 답안”의 경우, 형태소 분석 및 품사 부착의 결과가 없다. 따라서 3.3절에서 언급한 문제점인 형태소 분석과 품사 부착 결과의 출력 여부를 평가 척도로 삼는다. 제안하는 형태소 분석 및 품사 부착 기법의 비교 대상(baseline)은 기존의 형태소 분석 및 품사 부착 기법<sup>7)</sup>이다.

### 5.3 성능 평가 결과

#### 5.3.1 세종 말뭉치의 형태소 분석 및 품사 부착 결과

5.1.1절에서 언급한 세종 말뭉치를 대상으로 형태소 분석을 수행한 결과는 Table 5.4이며, 품사 부착 결과는 Table 5.5이다.

Table 5.4 Recall of morphological analysis

정답 형태소	맞은 개수		재현율		시간(초)	
	기존	제안	기존	제안	기존	제안
어절 단위 3,736,666	3,192,542	3,699,673	85.44	98.86	1452.78	74.20

형태소 분석 기법의 성능 평가는 정답 문장을 어절 단위로 분리하여 해당 어절의 정답 형태소를 형태소 분석 과정에서 찾아냈는지 판단한다. 기존의 시스템은 총 3,736,666개의 정답 어절 중 3,192,542개를 찾아내 재현율이 85.44%로 측정됐다. 제안하는 형태소 분석 기법은 3,699,673개의 정답 형태소를 후보로 찾아내는 데 성공하여 98.86%의 재현율을 보였다. 재현율 자체를 봤을 때 제안하는 형태소 분석 기법이 약 13.42%p 정도 향

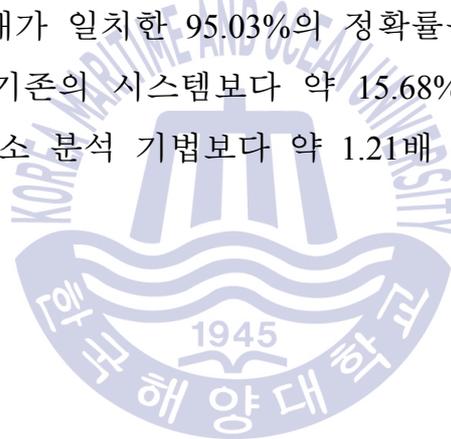
7) 기존의 한국어 서답형 문항 자동채점 프로그램에 사용된 형태소 분석 및 품사 부착 기법

상된 성능을 보였으며, 기존의 형태소 분석 기법보다 약 20배 정도 빠른 것으로 나타났다.

Table 5.5 Accuracy of POS Tagging

시스템의 형태소		맞은 개수		정확률		시간(초)	
기존	제안	기존	제안	기존	제안	기존	제안
3,325,385	3,650,892	2,638,849	3,510,208	79.35	95.03	1484.38	1230.50

품사 부착기의 성능은 시스템이 출력한 결과 문장을 어절 단위로 분리한 형태소의 수와 실제 정답과 일치하는 형태소의 비율(정확률)로 평가했다. 기존의 품사 부착기는 3,325,385개의 출력 결과 중 2,638,849개가 일치한 79.35%의 정확률을 보였다. 제안하는 품사 부착기는 3,650,892개의 출력 결과 중 3,510,208개가 일치한 95.03%의 정확률을 보였다. 제안하는 품사 부착기의 성능이 기존의 시스템보다 약 15.68%p 정도 향상된 성능을 보였으며 기존의 형태소 분석 기법보다 약 1.21배 정도 빠른 것으로 나타났다.



### 5.3.2 2014년 국가 수준 학업성취도 평가 형태소 분석 및 품사 부착 결과

Table 5.6은 5.1.2절에서 설명한 2014년 국가 수준 학업성취도 평가의 답안들을 대상으로 성능 평가를 한 결과이다. 앞서 언급한 것처럼 이 실험 대상에 대한 정답이 존재하지 않으므로 형태소 분석 및 품사 부착 결과의 출력 여부와 그에 따른 시간을 평가 대상으로 설정했다.

Table 5.6 Performance of morphological analysis and POS tagging for 2014 NLSA

학년	고2				중3				평균	
	국어				국어		사회			
과목	국어				국어		사회			
문항	2-(2)		6-(1)		6-(2)-ㄱ		2			
시스템	기존	제안	기존	제안	기존	제안	기존	제안	기존	제안
결과 출력	○	○	○	○	×	○	○	○	△	○
시간(초)	41.00	17.78	31.90	0.09	-	0.15	21.23	0.23	-	456

Table 5.6에 나타난 결과를 보면 결과 출력까지의 걸린 시간을 비교해봤을 때 제안하는 형태소 분석 및 품사 부착 기법 쪽의 성능이 더 좋은 것으로 나타났다. 특히, ‘중3 국어 6-(2)-ㄱ 문항’의 경우 기존의 시스템은 결과를 출력하지 못한 것<sup>8)</sup>에 비해 제안하는 시스템은 결과를 출력한 것을 볼 수 있다. 현재 연구 중인 한국어 서답형 문항 자동채점 시스템의 구조도인 Fig 3.2에서 확인할 수 있듯이 ‘언어 분석 단계’에서 결과를 출력하지 못하면 채점 단계를 진행할 수 없다는 치명적인 문제가 발생한다. 따라서 제안하는 형태소 분석 및 품사 부착 기법이 시간적인 측면과 어떤 입력 문장에 대해서도 결과 출력이 가능한 점을 볼 때, 기존의 형태소 분석 및 품사 부착 기법보다 더 유용하다고 판단된다.

8) 약 1.5일 정도 형태소 분석기를 돌렸으나 형태소 분석에 대한 계산을 끝내지 못해 결과를 출력하지 못했다. 컴퓨터의 사양과 시간이 충분하다면 언젠가는 계산을 완료하고 형태소 분석 결과를 출력할 것으로 예상된다.

## 5.4 오류분석

제안하는 형태소 분기 및 품사 부착 기법의 오류는 크게 음절 기반의 단어 분리 결과의 오류와 사전의 오류로 나눌 수 있다.

### (1) 음절 기반의 단어 분리 결과의 오류

이는 전처리 단계로 추가한 음절 기반의 단어 분리 결과가 정확하지 않은 경우이다. 대표적으로 같은 의미를 가지는 똑같은 어절에 대해 단어의 분리 결과가 다르게 나타나는 오류가 있다. 예를 들어, “그 사람이 밝혀”라는 문장에 대해 단어 분리를 진행할 경우 “밝혀”라는 어절은 “밝 혀”로 분리되어 사전 탐색 알고리즘을 통해 ‘밝히/PV 어/EF’라는 결과를 출력할 수 있다. 그러나 “심영희 교수가 밝혀”라는 문장에서 “밝혀”라는 어절에 대해 단어 분리를 진행하면 “밝혀”라는 잘못된 단어 분리 결과가 나왔다. “밝혀”의 경우 분석용 사전과 형태소 사전에 모두 존재하지 않으므로 ‘밝혀/\_UK\_’라는 잘못된 결과를 출력했다. 이 외에도 “와줘서”와 같은 단어의 경우 “와 줘 서”로 분리되어야 하지만 “와줘 서”로 분리되어 잘못된 출력 결과를 야기했다.

### (2) 사전의 오류

사전의 오류는 분석용 사전과 형태소 사전의 오류가 모두 포함된다. 분석용 사전의 오류는 “세계적인”에서 “인”에 대한 분석 결과인 ‘이+ㄴ’이 분석용 사전에 존재하지 않는 것과 같은 오류이다. 이 경우는 분석용 사전을 만드는 데 사용한 Fig. 4.6의 (단어, 형태소) 정렬 알고리즘의 문제이다. 향후에 Fig. 4.6의 알고리즘을 수정하여 분석용 사전을 보완하면 해당 오류는 해결될 것으로 예상된다. 형태소 사전의 오류는 특정 형태소에 대한 품사 정보 및 빈도수의 정보가 형태소 사전에 존재하지 않기 때문에 발생한 오류이다. 주로 고유명사(NQ)에서 이 같은 문제가 발생하므로, 이는 형태소 사전에 해당 형태소의 정보를 추가하는 것으로 해결할 수 있다. 이 외에도 전이 확률 사전에서 ‘JJ(조사)→XN(과생 명사의 접미사)’의

확률값이 ‘JJ(조사)→NP(대명사)’의 확률값보다 작은 값을 가져서 “나/NP  
도/JJ /\_SP\_ 네/NP 가/JJ”로 분석되어야 할 문장이 “나/NP 도/JJ /\_SP\_ 네  
/XN 가/JJ”로 분석되는 등 잘못된 품사 부착 결과를 야기했다.





향후 연구과제는 분석용 사전을 보완하기 위해 단어와 형태소의 정렬 알고리즘의 개선과 음절 기반의 단어 분리 모델의 성능 개선이 있다. 또한, 본 논문에서 사용한 사전과 모델은 한 가지의 종류의 말뭉치를 이용하여 평가했으므로 과적합(overfitting) 문제가 존재할 수 있으므로 다양한 말뭉치를 통해 사전과 모델의 성능을 개선하는 것이 요구된다.



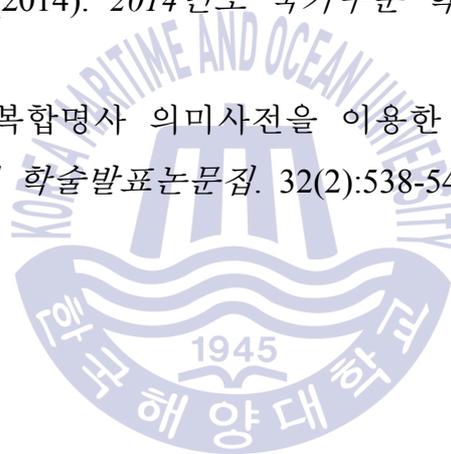
## 참고문헌

- Attail, Y., and Burstein, J., (2006). “Automated essay scoring with e-rater® v.2”, *The Journal of Technology, Learning, and Assessment*. 4(3):8.
- Cheon, Y., Liu, C., Lee, C., and Chang, T., (2010). “An unsupervised automated essay-scoring system”, *IEEE Intelligent Systems*, September/October, pp. 61-67.
- Dikli, S., (2006). “An overview of automated scoring of essays”, *The Journal of Technology, Learning, and Assessment*. 5(1):5-35.
- Heckel, P., (1987). “A technique for isolating differences between files”, *Communications of the ACM*. 21(4):264-268.
- Kim, D. B., Lee, S. J., Choi, K. S., and Kim, G. C., (1994). “A Two-level Morphological Analysis of Korean”, *Proceedings of the 15-th International Conference on Computational Linguistics (COLING-94)*. 1:535-539.
- Kwon, H. C., Chae, Y. S., and Jeong, G. O., (1991). “A Dictionary-based Morphological Analysis”, *Proceedings of Natural Language Processing Pacific Rim Symposium*. pp. 87-91.
- Kwon, H. C., and Kattunen, L., (1994). “Incremental Construction of a Lexical Transducer for Korean”, *Proceedings of the 15-th International Conference on Computational Linguistics (COLING-94)*. 2:1262-1266.
- Lafferty, J., McCallum, A., and Pereira, F., (2001), “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data”, *Proceedings of the 18<sup>th</sup> International Conference on Machine Learning*. pp. 282-289
- 강승식, (1994). “한국어의 형태론적 특성과 형태소 분석 기법”, *정보과학지*. 12(8):47-59.

- 강승식, (2002). *한국어 형태소 분석과 정보 검색*, 흥릉과학출판사.
- 강원석, (2011). “질의문 유형 분석을 통한 서답형 자동채점 시스템”, *한국콘텐츠학회논문지*. 11(2).
- 교육과학기술부, (2009). *2009 개정 교육과정 총론*. 교육과학기술부. 고시 제 2009-41호.
- 국립국어원, (2011). *21세기 세종계획*, 문화체육관광부.
- 김남미, (2010). *쉽게 배워 바로 써먹는 친절한 국어 문법*, 사피엔스21.
- 김성용, 최기선, 김길창, (1987). “Tabular Parsing 방법과 접속 정보를 이용한 한국어 형태소 분석기”, *한국정보과학회 춘계 인공지능 학술발표회 논문집*. pp. 133-147.
- 김수남, 원상연, 권혁철, (1998). “의미 정보를 이용한 한국어 복합명사 분석”, *한국정보과학회 학술발표논문집*. 25(2):195-197.
- 김재한, 옥철영, (1994). “어절 사전을 이용한 한국어 형태소 분석”, *한국정보과학회 봄 학술발표 논문집*. 21(1):813-816.
- 김재훈, 서정연, 김길창, (1995). *실용적인 한국어 형태소 해석*, 한국과학기술원. CS/TR-95-98.
- 김재훈, (1996). “가중치 망을 이용한 한국어 품사 태깅”, *정보과학회논문지 (B)*. 25(6):951-959.
- 나승훈, 김창현, 김영길, (2014). “래티스상의 구조적 분류에 기반한 한국어 형태소 분석 및 품사 태깅”, *정보과학회논문지 : 소프트웨어 및 응용*. 41(7):523-532.
- 노은희, 이상하, 임은영, 성경희, 박소영, (2014). *한국어 서답형 문항 자동 채점 프로그램 개발 및 실용성검증*, 한국교육과정평가원, 연구보고서 RRE 2014-6.
- 박봉래, 황영숙, 임해창, (1998). “용례 분석에 기반한 미등록어의 인식”, *정보과학회논문지*. 25(2):397-407.
- 박영환, (1991). *말뭉치에 기반한 형태소 분석기 및 철자 검사기의 구현*, 연세대학교 석사학위 논문.

- 박일남, 강승식, 노은희, 김명화, 성태제, (2013). “정답 템플릿 작성 방식에 의한 한국어 서답형 문항 자동채점 시스템”, *정보과학논문지 : 컴퓨팅의 실제 및 레터*. 19(12):630-636.
- 심광섭, 양재형, (2004). “인접 조건 검사에 의한 초고속 한국어 형태소 분석”, *정보과학논문지: 소프트웨어 및 응용*. 31(1):89-99.
- 심광섭, (2007). “MADE : 형태소 분석기 개발 환경”, *인터넷정보과학회 논문지*. 8(4):159-171.
- 심광섭, (2011a). “형태소 분석기 사용을 배제한 음절 단위의 한국어 품사 태깅”, *한국인지과학회 논문지*. 22(3):327-345.
- 심광섭, (2011b). “CRF를 이용한 한국어 자동 띄어쓰기”, *한국인지과학회 논문지*. 22(2):217-233.
- 안동언, (1999). “좌우접속정보를 이용한 명사추출기”, *한국정보과학회 언어공학연구회 학술발표 논문집*. pp. 173-178.
- 양승현, 김영섭, (2000). “부분 어절의 기분석에 기반한 고속 한국어 형태소 분석 방법”, *정보과학논문지: 소프트웨어 및 응용*. 27(3):290-301.
- 이경호, 이공주, (2014). “기계학습을 이용한 중등 수준의 단문형 영어 작문 자동채점 시스템 구현”, *한국정보과학회*. 41(11):911-920.
- 이양락, 조지민, 신일용, 조윤동, 이명애, 신택수, 박기범, 이광상, 김용명, 강대현, 김동영, 김현경, 김진구, 김영춘 (2010). *2014학년도 대학수학능력시험체제 개발을 위한 기초 연구*, 한국교육과정 평가원. 연구보고서 대수능 CAT-2010-3.
- 이운재, (1993). *한국어 문서 태깅 시스템의 설계 및 구현*, 한국과학기술원 전산학과 석사학위 논문.
- 이용훈, 옥철영, (2011). “의미기반 한국어 복합명사 분석”, *한국정보과학회 학술발표논문집*. 38(1C):221-224.
- 이재성, (2011). “한국어 형태소 분석을 위한 3단계 확률 모델”, *정보과학회논문지: 소프트웨어 및 응용*. 38(5):257-268.
- 임해창, 임희석, 이상주, 김진동, (1996). “자연어 처리를 위한 품사 태깅

- 시스템의 고찰”, *정보과학지*. 4(7):36-57.
- 장병탁, 김영택, (1990). “다중언어 형태소 분석 및 합성을 위한 언어 규칙의 기계학습”, *한국정보과학회 논문지*. 17(4):463-474.
- 전길호, (2012). *기계학습을 이용한 음절기반 품사 부착*, 한국해양대학교 석사학위 논문.
- 진경애, (2007). “영작문 자동채점 시스템 개발 연구”, *영어어문교육*. 13(1):236-237.
- 최재혁, 이상조, (1993). “양방향 최장일치법에 의한 한국어 형태소 분석기에서의 사전 검색 횟수 감소 방안”, *정보과학회 논문지*. 20(10):1497-1507.
- 한국교육과정평가원, (2014). *2014년도 국가수준 학업성취도 평가*, 한국교육과정평가원.
- 허정, 장명길(2005). “복합명사 의미사전을 이용한 동음이의어 중의성 해소”, *한국정보과학회 학술발표논문집*. 32(2):538-540.



## 감사의 글

먼저 석사 과정 동안 연구를 진행하면서 부족한 제가 어려움에 부딪힐 때마다 아낌없는 격려와 가르침을 주신 김재훈 지도교수님께 진심으로 감사드립니다. 바쁘신 와중에도 논문지도에 신경을 써주시고 가르침을 주신 박휴찬 교수님과 이장세 교수님께도 감사드립니다. 대학원 생활동안 제가 부족한 과목을 이해하기 쉽게 잘 가르쳐주신 손주영 교수님과 조석제 교수님께도 감사드립니다.

학부 생활부터 대학원 생활까지 항상 신경을 써주신 강근호 조교님, 김경언 조교님께도 감사의 말씀을 드립니다.

졸업한 후에도 제게 신경을 많이 써주신 자연어 처리 연구실 식구인 서형원님, 권홍석님, 이정태님을 비롯하여 석사 과정 동안 같이 연구실 생활을 했던 노경목님, 김성태님, 박호민님과 권신비님께 감사드립니다. 같은 연구실은 아니었지만 함께 수학하면서 어려울 때마다 도움을 주셨던 황훈규 박사님, 김효승 박사님, 정지은님, 김길용님, 임상우님께도 감사의 말씀을 드립니다.

마지막으로 석사 과정을 마칠 때까지 많은 어려움 속에서도 늘 저를 지지해주시고 사랑으로 보살펴주신 부모님과 늘 투덜거리는 못난 동생을 위해주는 언니에게 감사의 말을 전합니다.

## 부 록

부록 A 세종 말뭉치 품사 및 단순화 태그 (계속)

분류	세종 말뭉치 품사 태그		
	태그	설명	단순화 태그
체언	NNG	일반 명사	NC
	NNP	고유 명사	NQ
	NNB	의존존 명사	NB
	NR	수사	NR
	NP	대명사	NP
용언	VV	동사	PV
	VA	형용사	PA
	VCN	부정 지정사	
	VX	보조 용언	PX
	VCP	긍정 지정사	CO
관형사	MM	관형사	MD
부사	MAG	일반 부사	MA
	MAJ	접속 부사	
감탄사	IC	감탄사	IC
조사	JKS	주격 조사	JJ
	JKC	보격 조사	
	JKG	관형격 조사	
	JKO	목적격 조사	
	JKB	부사격 조사	
	JKV	호격 조사	
	JKQ	인용격 조사	
	JX	보조사	
	JC	접속 조사	
선어말 어미	EP	선어말 어미	EP
어말 어미	EF	종결 어미	EF
	EC	연결 어미	
	ETM	관사형 전성 어미	EN
	ETN	명사형 전성 어미	

부록 A 세종 말뭉치 품사 및 단순화 태그 (계속)

분류	세종 말뭉치 품사 태그		
	태그	설명	단순화 태그
접두사	XPN	체언 접두사	XP
접미사	XSN	명사 파생 접미사	XN
	XSV	동사 파생 접미사	XV
	XSA	형용사 파생 접미사	XA
어근	XR	어근	NC
부호	SF	마침표, 물음표, 느낌표	SY
	SP	수미표, 가운데점, 콜론, 빗금	
	SS	따옴표, 괄호표, 줄표	
	SE	줄임표	
	SO	붙임표 (물결, 숨김, 빠짐)	
	SW	기타기호 (논리수학기호, 화폐기호)	
분석 불능	NF	명사추정범주	_UK_
	NV	용언추정범주	_UK_
	NA	분석불능범주	_UK_
한글 이외	SL	외국어	SF
	SH	한자	SH
	SN	숫자	SN