KIPS Tr. Software and Data Eng.
Vol.2, No.12 pp.865~872 pISSN: 2287-5905

http://dx.doi.org/10.3745/KTSDE.2013.2.12.865

Automatic Construction of Korean Two-level Lexicon using Lexical and Morphological Information

Bogyum Kim† · Jae Sung Lee††

ABSTRACT

Two-level morphology analysis method is one of rule-based morphological analysis method. This approach handles morphological transformation using rules and analyzes words with morpheme connection information in a lexicon. It is independent of language and Korean Two-level system was also developed. But, it was limited in practical use, because of using very small set of lexicon built manually. And it has also a over-generation problem. In this paper, we propose an automatic construction method of Korean Two-level lexicon for PC-KIMMO from morpheme tagged corpus. We also propose a method to solve over-generation problem using lexical information and sub-tags. The experiment showed that the proposed method reduced over-generation by 68% compared with the previous method, and the performance increased from 39% to 65% in f-measure.

Keywords : Korean Two-level Morphology, Two-level Lexicon, Korean Morphotactic, Automatic Construction, Tagged Corpus

어휘 및 형태 정보를 이용한 한국어 Two-level 어휘사전 자동 구축

김 보 겸† · 이 재 성††

요 약

Two-level 형태소 분석 방법은 규칙 기반 방법 중 하나로 형태소의 변화 현상을 규칙으로 처리하고, 기본 어휘 사전을 기반으로 형태소 결합관계를 분석한다. 이는 언어에 독립적인 방법으로 한국어에 대해서도 일부 구축되어 적용됨이 증명되었다. 그러나 기존 한국어에 대한 Two-level 형태소 분석기는 사전을 수동으로 구축하여 규모가 매우 작고 실제 사용에 제한적이었으며, 과분석이 많아 효율성이 매우 떨어졌다. 본 논문은 세종 품사부착 말뭉치에서 대규모의 Two-level 어휘 사전을 자동으로 구축하여 형태소 분석기의 적용 범위를 넓히고, 형태소간의 결합관계를 어휘 정보와 어휘 형태에 따른 하위품사 정보를 이용하여 분석함으로써 형태소 분석기의 성능을 향상시킬 수 있는 방법을 제시한다. 실험 결과, 기존의 방법보다 형태소 분석기의 과분석을 68% 이상 줄여 f-measure를 25.5% point 이상 향상시킬 수 있었다.

키워드 : 한국어 2단계 형태소 분석, 2단계 형태소 분석 어휘사전, 규칙기반 형태소 분석, 자동 구축, 품사부착 말뭉치

1. 서 론

Two-level 형태소 분석 방법은 Kimmo에 의해 제안된 규칙 기반 형태소 분석 방법으로, 언어에 독립적이며, 형태소 분석에서의 철자 변형 현상을 효과적으로 처리한다[1, 2]. 이 모델은 형태소들 간 결합시 발생하는 철자의 변화를 Two-level 규칙으로 표현하고, 형태소의 원형과 그들의 결합관계는 Two-level 어휘 사전에 그래프 형태로 표현한다. 또한 이는 형태소의 원형과 활용형을 중간단계 없이 2단계로 처리하여 형태소 분석 및 생성이 동시에 처리가 가능하다[1-3].

Two-level 모델을 이용한 한국어 형태소 분석은 [4]에서 처음 시도되었다. 이 연구에서는 한국어의 철자 변형 현상에 대한 규칙을 용언을 중심으로 비교적 자세히 작성하였으며, 어휘 사전을 수동으로 일부 구축하여 작은 규모의 실험이기는 하나 한국어 형태소 분석에도 Two-level 모델이 적용될 수 있음을 증명하였다[4, 5]. 그러나 활용할 수 있는 언어자원이 충분하지 않아 어휘 사전의 규모가 작았으며, 그로 인해 매우 제한적인 분석만이 가능하였다. 또한 형태소간의 결합관계를 품사의 결합관계만으로 표현하여 결과의

※ 이 논문은 2011년도 충북대학교 학술연구지원사업의 연구비 지원에 의하여 연구되었음.
† 준 회 원 : 충북대학교 디지털정보융합학과 박사과정
†† 종신회원 : 충북대학교 디지털정보융합학과 교수
논문접수 : 2013년 7월 29일
수 정 일 : 1차 2013년 8월 29일
심사완료 : 2013년 10월 4일
* Corresponding Author : Jae Sung Lee(jasonlee@cbnu.ac.kr)

과분석이 많은 문제가 있었다.

최근에는 세종 형태부착 말뭉치[7]와 같은 언어 자원이 많이 개발되어 있어, 이를 이용한다면 신뢰성 있는 대규모의 어휘사전을 손쉽게 구축할 수 있으며, 어휘들 간의 결합관계를 추출하여 형태소 분석에 적용함으로써 기존에 개발된 한국어 Two-level 형태소 분석 모델의 성능을 더욱 향상 시킬 수 있을 것이다.

본 논문은 형태소 품사부착 말뭉치로부터 어휘를 추출하여 대규모의 Two-level 어휘사전을 자동으로 구축한다. 또한 어휘 및 품사들 간의 결합관계를 자동으로 추출하여 어휘사전을 구성한다. 특히 형태소의 결합관계를 '품사-품사(이하 품사전이)'의 결합관계가 아닌 '형태소/품사-품사(이하 어휘전이)' 결합관계로 확장하고, 어휘의 형태적 특징에 따른 세부 품사 분류(이하 하위품사)를 통해 형태소 분석 결과의 과분석을 줄이는 방법론을 제시한다.

본 논문의 구성은 다음과 같다. 2장에서는 Two-level 모델의 개요와 한국어에 적용하는 방법을 설명하고, 3장에서는 어휘 사전을 자동으로 구축하는 방법과 형태소 결합관계 추출 방법에 대해 설명한다. 4장에서는 세종 품사부착 말뭉치를 이용한 실험 및 이전 방법들과의 비교를 하고, 5장에서는 결론을 낸다.

2.관련 연구

형태소 분석이란 자연언어 처리의 기초 단계로써 어절(영어의 경우 단어)에서 형태소들을 분리해내어 그 원형을 복원하는 것을 말한다. 이를 위해서는 형태소 원형에 대한 정의, 각 원형들 간의 결합관계에 대한 정보, 형태소들 간의 결합 시 발생되는 철자의 변화 등을 처리하는 것이 필요하다[4].

Two-level 모델은 Kimmo Koskenniemi에 의해 제안된 형태소 분석 방법으로 단어의 발화에서 생기는 여러 단계의 음운론적 변형을 중간 단계 없이 어휘형(lexical form)과 표층형(surface form)을 직접 일치시킨다. 즉, 표층형 철자와 어휘형 철자 사이의 대응 관계만을 기술함으로써 형태소 분석과 생성을 모두 표현할 수 있으며, 여러 개의 규칙이 병렬적으로 적용이 가능하다(Fig. 1)[1, 6].

Two-level 모델에서는 형태소간의 결합시 발생하는 철자의 변화는 규칙으로 표현하고, 그들의 결합관계는 어휘 사전에 그래프 형태로 저장한다. Table 1은 Two-level 모델의 규칙 기술 방법이다.

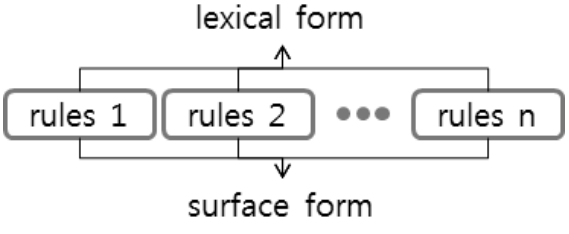

Fig. 1. Two-level transformation[2]

Table 1. Expression of Two-level rule[8]

Rule	$\alpha:\beta <=> \gamma^{l}:\gamma^{s}$ ___ $\delta^{l}:\delta^{s}$
Meaning	Lexical α (between lexical γ^{l} and δ^{l}), surface β (between surface γ^{s} and δ^{s}) correspond to each other.

규칙을 살펴보면, '<=>'를 중심으로 LHS(Left Hand Side)에는 변형이 일어나는 글자를 기술하고, RHS (Right Hand Side)에는 변형이 일어나는 글자 주변의 환경을 기술한다. '___'는 철자의 변화가 일어나는 위치를 의미하며, ':' 기호를 이용하여 어휘형과 표층형의 글자를 1:1로 대응시킨다.

단, 어휘형과 표층형을 1:1로 대응시킬 때 삽입 혹은 삭제와 같은 현상이 발생하면 글자를 1:1로 대응시키기가 힘들다. 이와 같은 현상을 해결하기 위해 Two-level 모델에서는 'Ø' 기호를 삽입 혹은 삭제가 일어나는 글자의 위치에 NULL 문자로 삽입하여 1:1 대응이 되도록 하였다. 또한 형태소간의 경계가 명확하지 않으면 어휘의 변화가 단일 형태소 내에서의 변화인지, 아니면 형태소와 형태소가 결합할 때 그 경계에서 일어나는 것인지 구분할 수가 없다. 그래서 형태적 변형이 일어나는 형태소의 경계 부분에 '+'기호를 추가하여 이를 명확히 구분할 수 있도록 하였다.

Two-level 모델을 한국어에 대해 일부 구축하여 적용시킨 연구로는 [4, 5]가 있다. 이 연구에서는 한국어의 음절을 자소단위의 N-바이트 코드로 표현하여 Two-level 규칙을 기술하고 있으며, 형태소간의 결합관계를 수동으로 작성하여 Two-level 어휘 사전을 구축하였다. 예를 들어, '도와'를 '돕+아'로 분석하는 과정을 살펴보면 Fig. 2와 같다.

(a) Two-level rule

ㅂ:ㅗ <=> ___ +:Ø ㅏ

(b) Morpheme recovery

Lexicon	ㄷ	ㅗ	ㅂ	+	ㅏ
Surface	ㄷ	ㅗ	ㅗ	Ø	ㅏ

*'Ø'는 NULL을 의미함

Fig. 2. Morpheme recovery of '도와' using Two-level rule

N-바이트 코드는 한글 음절을 기본 자소로 표기하는 방법으로써 자소단위로 변화가 일어나는 한국어 변형 현상을 Two-level 규칙으로 표현하기 적합한 코드이다. 또한 이는 중성에서 나타나는 복모음을 단모음들의 결합으로 보고 복모음 코드를 이용하여 따로 표기함으로써 모음조화, 모음축약 등의 음성학적 특징을 적절히 반영하여 처리할 수 있다. 단, 초성의 'ㅇ'은 음가가 없으므로 생략하여 표현하며 어휘형과 표층형의 자소가 동일하면 ':'을 이용한 대응관계가 아닌 하나의 자소로 표현한다.

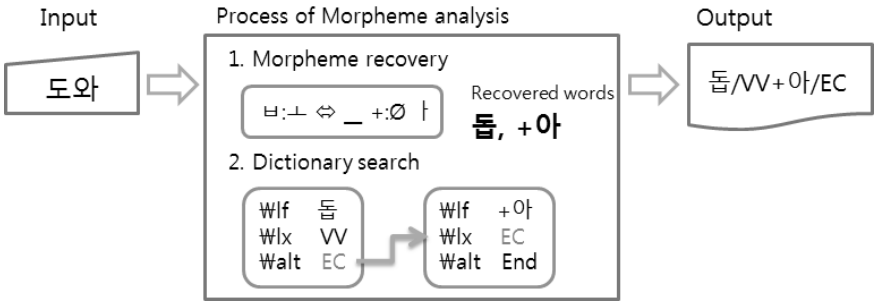

Fig. 3. An example of analyzing process

Two-level의 어휘사전은 각 형태소의 원형 어휘가 Trie 구조로 저장되어 있으며, 형태소의 품사, 품사간의 연결 관계 등도 저장되어 있어 형태소를 분석하는데 여러 정보를 제공해 준다. 이 중, 특히 품사간의 연결 관계는 Two-level 어휘사전과 규칙을 이용하여 원형 어휘를 찾은 이후에, 품사의 결합이 올바른지 검사하여 잘못된 과분석을 줄여주는 역할을 수행한다. 어절 '도와'를 분석하는 전체 수행 과정은 Fig. 3과 같다.

어휘를 사전에 기록할 때에는 한 가지 주의해야 할 사항이 있는데, 이는 어휘의 형태는 같으나 형태론적 변형 현상이 다른 어휘를 기록하는 것이다. Two-level 규칙은 어휘의 형태만을 고려하기 때문에 이러한 단어에 대한 처리를 따로 하지 않는다면, 잘못된 분석 결과를 생성해낼 수밖에 없다.

Lexical form	POS tag	Surface form	Word sense
굽	VV(verb.)	구워, 구운, 굽고	roast
	VA(adj.)	굽어, 굽은, 굽고	bent

Example of inflection
: 1. 굽/VV -> 고기를 구워 먹었다.
2. 굽/VA -> 팔이 안으로 굽어 있었다.

Fig. 4. Comparison of '굽/VV' and '굽/VA'

Fig. 4의 예제에서 '굽/VV'은 '어'로 시작하는 어미와 결합할 때 어간의 마지막 'ㅂ'이 'ㅜ'로 변하는 불규칙 용언이며 '굽/VA'의 경우는 변형이 일어나지 않는 규칙용언이다. 어휘의 형태만을 고려하면 'ㅂ'으로 끝나는 모든 용언에 대해 '어'로 시작하는 어미와의 결합시 'ㅂ'을 'ㅜ'로 대응하는 오류를 범하게 된다. 이를 위해서 규칙 용언은 '+', 불규칙 용언은 '$'를 원형 어휘의 뒤에 추가하여 구분한다[4].

이외에도 '르'로 끝나는 불규칙 용언은 어미 '어'와 결합시 어미만 변하는 '러'불규칙과 어간과 어미가 동시에 변하는 '르'불규칙의 두 가지 변형이 있는데 이는 'X'를 '러'불규칙 용언의 뒤에 추가하여 구분한다.

한국어에서 대부분의 어휘 변형 현상은 용언과 어미, 체언과 조사 사이에서 일어나며 특히 체언의 경우는 변형 현상이 거의 없으므로, 이러한 구분 기호는 용언과 어미, 조사에만 사용한다. 형태소 구분자 추가 규칙은 Table 3과 같다.

Table 2. Homonym with different inflection

Lexicon	Type of inflection
굽$/VV, 굽+/VA	irregular verb. & regular adj.
흐르$/VV, 이르X/VV	irregular verb[1]. & irregular verb[2].

Table 3. Morpheme separator using building lexicon[4]

Separator	Rules of separator addition
+	Front of Josa & Eomi Back of regular verb. & predicative Josa
$	Back of irregular verb. (except for irregular verb. '러')
X	Back of irregular verb. '러'

이외에, 체언과 조사의 결합관계에서 서술격 조사 '이'의 경우(세종 말뭉치에서는 긍정 지정사)는 일반 조사 '이'와 다르게 무종성의 체언과 결합시 탈락하는 규칙이 있다.

Table 4. An example of predicative Josa('이/VCP') missing

Eojeol	Morpheme analysis
시계다	시계/NNG+이/VCP+다/EF

[4]에서는 서로 다른 코드 값을 배정하여 이를 구분하고 있으나 본 논문에서는 형태소 구분자 '+'를 긍정지정사 뒤에 추가하여 구분한다.

3. 형태소 품사 부착 말뭉치를 이용한 한국어 Two-level 어휘사전 자동 구축

본 연구의 전체 시스템 구성은 Fig. 5와 같다. PC-KIMMO는 Two-level 모델을 컴퓨터로 수행하기 위해 개발된 공개 소프트웨어로, Summer Institute of Linguistics사에서 1985년 제작하여 현재까지 널리 사용되어 왔으며 규칙을 처리하는 부분과 사전을 구성하는 부분이 비교적 완벽하게 구성되어 있다[3]. 본 연구에서도 이미 검증이 된 공개 소프트웨어인 PC-KIMMO를 이용하여 한국어 Two-level 모델을 실험하였다.

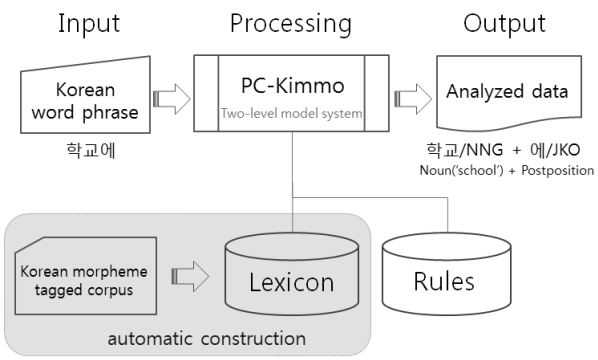

Fig. 5. Overview of Korean Two-level system

PC-KIMMO를 구성하는 규칙은 Two-level로 표현한 규칙들이 FST(Finite State Transducer)형태로 저장되며, 어휘 사전에는 각 형태소들이 변형이 일어나지 않은 원형의 형태로 저장(Lexicon field)되고, 형태소들 간의 결합관계도 표현(Alternation field)된다. 한국어에서의 형태소 변형 규칙에 대한 Two-level 규칙은 [4]에서 비교적 자세히 작성되어 본 연구에서는 이를 그대로 사용하며, 형태소의 원형과 그 결합관계는 세종 형태부착 말뭉치(이하 세종 말뭉치)[7]를 이용하여 자동으로 구축한다.

3.1 학습 자료 만들기

세종 말뭉치[7]는 약 1,000만 여개의 한국어 어절에 대한 분석 결과를 담고 있는 언어 데이터베이스이며 XML태그를 이용하여 문서 정보, 한국어 어절, 어절의 형태소-품사 분석 결과를 표현하고 있다. 이를 사용하면 대량의 어휘 사전을 손쉽게 구축할 수 있으나 Two-level 어휘 사전의 형식에 맞게 변환하기 위해서는 몇 가지 작업이 필요하다.

Two-level 어휘사전은 각 형태소의 원형과 품사, 그리고 품사간 결합관계로 구성된다. 이를 위해, 먼저 말뭉치로부터 어절과 형태소 분석 내용이 포함된 학습 자료를 만들고 이로부터 각 형태소의 원형과 품사간 결합 관계 등을 추출하여 어휘 사전을 구축한다. Fig. 6은 말뭉치로부터 추출한 어절 및 형태소 분석 내용이다.

Eojeol	Result of morpheme analysis
먹거나	먹/VV + 거나/EC
익사할	익사/NNG + 하/XSV + ㄹ/ETM
위험이	위험/NNG + 이/JKS
있다는	있/VV + 다는/ETM
지적이다.	지적/NNG + 이/VCP + 다/EF + ./SF
	.....

Fig. 6. Example of learning data

3.2 형태소 원형 어휘 및 bi-gram 정보 추출

수작업으로 어휘사전을 구축한 경우, 서로 다른 변형을 하는 동형 형태소를 구분하기 위해 '+', '$', 'X'의 기호를 각 형태소의 특성에 따라 형태소 원형의 앞 혹은 뒤에 추가하였다[4]. 그러나 말뭉치에서 이를 자동으로 구축할 때에는 학습 자료로부터 어휘의 변형 현상을 감지하고, 어떠한 변형현상이 있는지, 혹은 동형 형태소는 어떠한 것이 있는지를 찾은 후에 이를 추가해 주어야 한다. Fig. 7은 Fig. 6에서 구축된 학습 자료에서 활용형 어휘에 변화가 일어난 어절을 찾은 후 규칙 혹은 불규칙 변형 형태소에 형태소 구분자를 추가한 후 형태소의 원형과 bi-gram을 추출하는 알고리즘이다.

1. Read an Eojeol (surface form) and its analysis
2. Extract morphemes (M) from the analysis and concatenate them into C.
3. if (surface form (S) ≠ concatenated morpheme form (C)) then
   ① Align S and C using char_align program modified from [9]
   ② Find mismatched character positions
   ③ Lookup the inflection patterns using the mismatched positions
   ④ If found, add morpheme separator to the each morpheme form (M)
4. Store morpheme forms (M) and their bi-grams with tags
5. Repeat 1-4 until end of file

Fig. 7. Lexicon construction algorithm

단, Fig. 7의 4번 과정에서 저장되는 형태소 bi-gram 정보는 형태소 결합관계 추출 방법에 따라 저장되는 형태가 다르다.

3.3 형태소 결합관계 추출

Two-level 규칙에 의해 형태소의 원형이 복원된 이후 각 형태소들의 연결 가능성을 확인한 후 결합 가능한 형태소 연결만을 형태소 분석 결과로 출력한다. 형태소 결합 관계에 대한 학습은 3.2에서 작성된 형태소 bi-gram 정보를 학습 데이터로 사용한다(Fig. 8).

Front	Rear
먹+/VV	+거나/EC
+거나/EC	-/End
익사/NNG	하+/XSV
하+/XSV	+ㄹ/ETM
+ㄹ/ETM	-/End
위험/NNG	+이/JKS

Fig. 8. Morpheme bi-gram

Front	Rear
VV	EC, ETM, ...
EC	End, ...
NNG	XSV, JKS, VCP, ...
VCP	EF, ...
XSV	ETM, ...

Fig. 9. Tag-tag transition

1) 품사전이 결합관계 추출

품사전이 결합관계는 형태소 bi-gram 정보에서 각 형태소의 품사 정보만을 추출하여 Fig. 9와 같이 품사 결합관계를 획득한 다음 이를 형태소 결합관계로 사용한다. 이 방법은 [4]연구에서 사용된 방법과 동일하나, 본 연구에서는 이를 자동으로 구축한다.

2) 어휘전이 결합관계 추출

이 방법은 어휘적 특징을 고려하여 품사전이 결합관계를 확장한 방법이다. 이는 형태소 bigram 정보로부터 앞쪽에 위치하는 형태소의 경우는 어휘와 품사를 모두 추출하고, 뒤에 위치하는 형태소의 경우는 품사만을 추출하여 Fig. 10과 같은 '어휘/품사-품사' 결합관계를 만들어 형태소 결합관계로 사용한다.

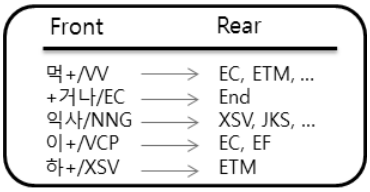

Fig. 10. Lexeme-tag transition

이 방법은 형태소 결합관계에 제약을 더 많이 줄 수 있어 품사전이 방법에서 문제가 되는 과분석 문제를 줄일 수 있다. 예를 들어, 품사전이 방법으로 '제일'과 '반사회' 어절을 분석하면 Fig. 11과 같은 결과를 얻을 수 있다.

제일	제/XPN+일/NR
반사회	반/XPN+사회/NNG
	반/XPN+사/NR+회/NNG *analysis error

Fig. 11. Example of morpheme analysis result using tag-tag transition[12]

'제일'의 경우는 올바른 분석을 하지만, '반사회'의 경우는 잘못된 분석 결과를 같이 얻게 된다. 체언 접두어 '제'는 수사와 결합하여 차례를 의미하는 형태소이지만 '반'은 일반 명사와 결합하여 반대를 의미하는 형태소이기 때문이다. 이처럼 한국어에는 같은 품사를 가지는 어휘라 할지라도 그 의미와 형태에 따라 결합할 수 있는 품사의 종류가 제한될 수 있다.

3) 하위품사(sub-tag) 전이 관계 추출

하위품사(sub-tag)는 같은 품사를 가진 형태소들을 어휘적 특징에 따라 분류한 후 이를 구분할 수 있도록 품사를 세분화하여 표현한 방법이다. 두 개의 형태소가 결합할 때에는 어휘의 형태적 특징에 따라 결합할 수 있는 형태소를 제한할 수 있기 때문에 이를 하위품사로 분류하면 더욱 정확한 형태소 분석 결과를 얻을 수 있다. 예를 들어, '사랑가는' 어절을 분석하면 Fig. 12와 같은 결과를 얻을 수 있다.

Lexicon	Following POS tags
사랑/NNG 사랑가/NNG	NNG, NNB, ... JKS, JX, ...
가/JKS	JKO, JKB, JX ...

Eojeol	Result of morpheme analysis	
사랑가는	사랑/NNG+가/JKS+는/JX	(a)
	사랑가/NNG+는/JX	(b)

Fig. 12. Morpheme analysis result of '사랑가는'

Fig. 12의 분석 결과는 어휘 전이 결합 관계에서는 모두 올바른 분석 결과이지만, 주격조사 '가/JKS'의 어휘적 특징을 고려하면 분석 (a)는 잘못된 분석이다. 왜냐하면 주격 조사 '가/JKS'에 선행하는 체언의 마지막 음절에는 종성이 없어야 하기 때문이다(음운법칙).

이러한 특징은 여러 형태소간의 결합에서 나타나는데 본 연구에서는 그 중 특징이 가장 잘 나타나는 '체언-조사'의 결합관계에 대해서만 실험을 하였다. 어휘적 특징을 반영하여 품사를 세분화한 다음 Fig. 13과 같이 어휘사전을 구축하게 되면 보다 정확하고 효율적인 한국어 Two-level 형태소 분석기를 구축할 수 있다.

Lexicon	Storage structure	
	[4]	Sub-tag
사랑/NNG	사랑/NNG	사랑/NNGC
사과/NNG	사과/NNG	사과/NNGV
이/JKS	이/JKS	이/JKSC
가/JKS	가/JKS	가/JKSV

Fig. 13. Example of sub-tag

4. 실험 및 분석

본 연구의 실험에 사용한 자료는 세종 말뭉치(문어체)이며[7], 총 10,130,363 어절 중 일부 오류(품사부착오류, 말뭉치형식오류 등)를 제외한 8,676,400 어절에서 순수 한글 어절인(부호, 숫자, 외국어 제외) 7,810,329 어절을 사용하였다. 또한 이 중 단 한 번 출현한 어절은 분석 오류의 경우가 많아 실험에서 제외하여 실제 실험에 사용한 어절은 총 7,252,908 어절이다.

학습에 사용한 어절은 전체 어절의 90%이며, 실험에 사용한 어절은 나머지 10%로 구성하여 10배수 교차 검증(10-fold-cross-validation) 실험을 하였다. 실험은 품사전이 방법과 어휘전이 방법으로 어휘 사전을 구축하여 수행하였으며, 각 방법에 하위품사 전이 방법을 적용하여 결과를 비교하였다. 또한 10%의 실험어절 중 8음절 이상의 어절은 실험에서 제외하였는데 이는 품사전이 방법으로 PC-KIMMO를 수행할 때 너무 많은 분석 결과를 생성해 내어, 프로그램 수행이 불가능하기 때문이다.

평가에 사용한 지표는 정답포함률(recall), 생성효율(precision), f-measure, 평균 분석 수(ANO, Average Number of Output), 미분석률(Failure)이며, 각 수식은 다음과 같다.

$$recall(R) = \frac{1}{N}\sum_{i=1}^{N} S(E_i) \quad (1)$$

$$Precision(P) = \frac{1}{N}\sum_{i=1}^{N} \frac{|A(E_i)|}{|O(E_i)|} \quad (2)$$

$$f-measure = \frac{2\times R\times P}{(R+P)} \quad (3)$$

$$ANO = \frac{\sum_{i=1}^{N}|O(E_i)|}{N} \quad (4)$$

$$Failure = \frac{1}{N}\sum_{i=1}^{N} NB(E_i) \quad (5)$$

$$E = Eojeol,\ S(E_i): \begin{cases} 1, \text{if } |O(E_i)\cap A(E_i)| \geq 1 \\ 0, otherwise \end{cases},$$

$$NB(E): \begin{cases} 1, |O(E_i)| = 0 \\ 0, otherwise \end{cases}$$

$$A(E_i): answers\ of\ E_i,\quad O(E_i): outputs\ of\ E_i$$

Fig. 14. Equations for evaluation

어휘 사전 구축 방법에 따라 실험을 한 결과 Table 5와 같은 결과를 얻을 수 있었다.

Table 5. Performance of Two-level morpheme analysis

	Previous work	Proposed work		
	Tag to tag	Lexeme to tag	Sub-tag to sub-tag	Lexeme to sub-tag
Recall	93.60%	92.45%	93.59%	92.32%
Precision	25.02%	45.34%	26.79%	50.11%
F-measure	39.48%	60.84%	41.66%	64.96%
ANO	12.67	3.80	9.49	2.97
Failure	1.87%	3.23%	1.91%	3.64%

결과를 살펴보면 어휘전이 방법의 f-measure 값이 품사전이 방법보다 각각 21.26% point, 23.3% point의 성능 향상이 있었다. 이는 어휘전이 방법이 품사전이 방법보다 과분석이 줄어들어(일반품사: 70%, 하위품사: 68.7%) 생성효율이 높아졌기 때문이다. 물론 어휘전이 방법을 사용하게 되면 어휘 부족에 따른 자료 부족 문제가 생기기 때문에 분석하지 못하는 어절이 품사전이 방법보다 많이 생기게 된다. 이는 학습 말뭉치의 크기를 늘려 어휘의 양을 증가시키거나, 미분석 어절에 대해 품사전이 방법으로 한 번 더 형태소 분석을 하는 등의 방법이 필요하다.

또한 하위품사를 이용하여 형태소 결합관계를 표현한 방법이 일반품사를 이용한 것보다 품사전이 방법의 경우 1.7% point, 어휘전이 방법의 경우 4.7% point 효율이 높아졌음을 알 수 있다. 본 연구에서는 '체언-조사' 결합관계에서의 어휘적 특징만을 이용하여 하위품사를 적용시켰는데, 이를 다른 형태소들 간의 결합관계에 대해 더 확장한다면 성능이 더욱 좋아질 것으로 예상된다.

본 실험에서 성능 저하의 가장 큰 원인은 미등록어에 의한 분석 실패로 분석 오류 전체에서 품사전이의 경우 평균 89.92%, 어휘전이의 경우 평균 82.64%를 차지하고 있다. Two-level 모델은 미등록어를 추정할 수 없으므로 이 문제를 바로 해결하기는 힘들다. [4]의 실험에서는 미등록어로 인한 분석 실패 어절을 실험의 결과에서 제외하였으며, 이와 같은 방법으로 본 연구에서 재 실험한 결과, Table 6과 같은 실험 결과를 얻을 수 있었다.

Table 6. Performance of Two-level morpheme analysis without unknown words

	Previous work	Proposed work		
	Tag to tag	Lexeme to tag	Sub-tag to sub-tag	Lexeme to sub-tag
Recall	99.28%	98.66%	99.40%	98.15%
Precision	26.54%	48.39%	28.45%	53.27%
F-measure	41.88%	64.93%	44.24%	69.06%
ANO	13.16	4	9.90	3.11
Failure	0.04%	0.26%	0.04%	0.61%

여전히 분석을 하지 못하는 일부 어절의 경우는 대부분 규칙으로 정의하지 못한 예외적인 경우이다. 예를 들어, Table 7과 같은 구어체 표현 중 Two-level 규칙으로 기술하지 못한 일부 어절에 대해서는 분석을 하지 못한다.

Table 7. Morpheme analysis result of '그걸'

Eojeol	Morpheme analysis result	
	Correct analaysis	Two-level model result
그걸	그것/NP+ㄹ/JKO	그거/NP+ㄹ/JKO

만약 한국어에서 발생하는 어휘변형 현상에 대한 규칙을 좀 더 상세히 기술할 수 있다면 이와 같은 오류는 해결이 가능할 것이다.

5. 결론 및 향후연구

언어에 독립적인 형태소 분석 방법인 Two-level 모델은 어휘형과 표층형의 두 가지 표현만으로도 철자의 변화를 처리할 수 있는 효율적인 모델이다. 이는 다양한 어미와 접사를 가지고 있으며, 불규칙 변형 현상, 음운 현상이 발달한 한국어에도 잘 적용됨이 이미 기존 연구에서 밝혀진 바 있다

[4]. 또한 기존의 연구에서 한국어의 다양한 변화를 처리할 수 있는 여러 규칙들과 어휘 사전이 일부 구축되기도 하였다. 그러나 기존의 연구들은 수동으로 모든 규칙과 사전을 만들어서 그 적용 범위가 매우 작고, 구축하는데 많은 시간이 소요되며 또한, 여전히 많은 오류들을 포함하고 있다. 그리고 어휘 사전을 구축할 때 형태소 간의 결합 관계를 품사의 결합 관계만으로 표현하여 과분석 문제가 발생하였다.

본 논문에서는 한국어 형태소 품사부착 말뭉치를 이용하여 어휘 사전을 자동으로 구축함으로써 수작업으로 인한 오류를 줄이고, 사전 구축에 소요되는 시간을 줄일 수 있었다.

Table 8. Elapsed time of dictionary construction

CPU	Intel core i7 3.2GHz
RAM	8GB
Dictionary size	55,292 morpheme entry
Elapsed time	40.1 sec

또한 어휘 사전 구축시 형태소의 결합관계를 품사의 결합관계에서 어휘와 품사의 결합관계로 확장하여 기존 연구에서 문제가 되었던 결과의 과분석을 68% 이상 줄일 수 있었으며, 어휘의 특징에 따라 품사를 보다 세분화하여 한국어 Two-level 형태소 분석기의 효율을 품사전이 방법의 경우 1.7% point, 어휘전이 방법의 경우 4.7% point 높일 수 있었다. 물론 본 연구에서도 문제가 되는 현상들이 있는데 이는 Table 9와 같다.

Table 9. Examples of corpus error types

Types	Example
Tag name	Undefined POS tags
Format	Mismatched morpheme-POS tag pairs
Spacing	Incorrectly spaced words or word phrases

미분석 문제의 경우는 크게 두 가지로 나눌 수 있다. 첫 번째는 미등록어에 의한 분석 실패이고, 두 번째는 규칙의 부재에 따라 분석하지 못하는 어절이 생기는 것이다. Two-level 모델의 특성만으로는 미등록 형태소에 대한 분석이 불가능하다. 이를 해결하기 위해서는 학습 말뭉치의 양을 늘리거나, 사용자 정의 사전 등을 이용하여 해결하는 것이 필요하다. 또한 어휘전이 방법의 경우에는 품사전이 방법보다 많은 정보를 필요로 하여 그에 따른 분석 실패 어절이 늘어나므로, 이를 해결하기 위해서는 분석에 실패한 어절에 대해 품사 전이 방법으로 한 번 더 형태소 분석을 수행하는 등의 방법이 필요하다.

본 연구에서 사용한 형태소 변형 규칙은 대부분 용언에서 발생하는 형태소 변형 현상으로 [4]에서 수작업으로 정제하며 만든 규칙을 사용하였다. 그러나 한국어의 형태소 변형 현상은 용언 이외의 형태소에서 발생하기도 하며, 예외적인 현상 또한 존재한다. 만약 이러한 현상에 대해 더 조사를 하고 규칙에 반영할 수 있다면 분석 가능한 어절 수가 늘어날 것으로 예상된다.

참 고 문 헌

[1] Koskenniemi, Kimmo, “Two-level Model for Morphological Analysis,” In IJCAI'83, *International Joint Conference on Artificial Intelligence,* pp.683-685, 1983.

[2] Koskenniemi, Kimmo, “A general computational model for word-form recognition and production,” In Proceedings for *COLING-84:Association for Computational Linguistics,* pp.178-181, 1984.

[3] Antworth and Evan L, “PC-KIMMO :A Two-level Processor for Morphological Analyzis,” *Occasional Publications in Academic Computing No.16. Summer Institute of Linguistics,* Dallas, TX, 1990.

[4] S. Lee, “A Two-level Morphological Analysis of Korean,” Master dissertation, Korea Advanced Institute of Science and Technology, Dept. of Computer Science, 1992. (in Korean)

[5] S. Lee, D. Kim, J. Seo, K. Choi, G. Kim, “A Two-level Approach to Korean Verb Morphology,” Proceedings of *Fall Korea Information Science Society Conference,* Vol.19, No.2, pp.993-996, 1992. (in Korean)

[6] Barton. G. Edward Berwick, Robert C. and Ristad, Eric Sven, *“Computational and Natural Language,”* The MIT Press, Cambridge, 1987.

[7] The national institute of the Korean Language, “Part-Of-Speech Tagged Corpus For Korean,” 21C Sejong Project, 2011. (in Korean)

[8] A. Arppe, L. Carlson, K. Linden, J. Piitulainen, M. Suominen, M. Vainio, H. Westerlund and A. Yli-Jyra, “*Inquiries Into Words; a Festschrift for Kimmo Koskenniemi on his 60th Birthday,”* CSLI Publications, Stanford University, pp.71-83, 2005.

[9] W. A. Gale and K. W. Church, *“A Program for Aligning Sentences in Bilingual Corpora,”* In Using Large Corpora (ed. Armstrong, S.), The MIT Press, Cambridge, Massachusettes, London, England, pp.75-102, 1994.

[10] S. Y. Kim, “A morphological analyzer for korean language with tabular parsing method and connectivity information,” Master dissertation, Korea Advanced Institute of Science and Technology, Dept. of Computer Science, 1987. (in Korean)

[11] J. W. Kang, “A design and implementation of hangul spelling and word-spacing checker using connectivity information,” Master dissertation, Korea Advanced Institute of Science and Technology, Dept. of Computer Science, 1990. (in Korean)

[12] J. S. Lee, B. Kim. “Automatic Construction of Korean Morphotactic for Two-level Lexicon,” *In LaRC2011, International Conference on Terminology, Language and Content Resources,* 2011.

김 보 겸

e-mail : bogyum@cbnu.ac.kr
2007년 충북대학교 컴퓨터교육과(학사)
2011년 충북대학교 정보컴퓨터교육과 (석사)
2011년～현 재 충북대학교 디지털정보융합학과 박사과정
관심분야 : 자연 언어 처리, 정보 검색

이 재 성

e-mail : jasonlee@cbnu.ac.kr
1983년 서울대학교 컴퓨터공학과(학사)
1985년 KAIST 전산학과(석사)
1999년 KAIST 전산학과(박사)
1985년～1988년 큐닉스 컴퓨터 과장
1988년～1993년 미국 및 한국 마이크로소프트 개발부 차장
1999년～2000년 ETRI 컴퓨터소프트웨어기술연구소 팀장
2005년～2006년 (미) 아리조나 대학 방문 교수
2000년～2011년 충북대학교 컴퓨터교육과 교수
2011년～현 재 충북대학교 디지털정보융합학과 교수
관심분야 : 자연 언어 처리, 정보 검색, 컴퓨터 교육 등