한국마린엔지니어링학회지 제36권 제8호, pp. 1123~1128, 2012. 11(ISSN 2234-8352 Online) / (ISSN 2234-7925 Print)
Journal of the Korean Society of Marine Engineering http://dx.doi.org/10.5916/jkosme.2012.36.8.1123

오류에 기반한 복합명사 좌우접속규칙 사전 구축

이공주[1] · 이성욱†

(원고접수일 : 2012년 10월 22일, 원고수정일 : 2012년 11월 7일, 심사완료일 : 2012년 11월 12일)

Error-driven Noun-Connection Rule Extraction for Morphological Analysis

Kong Joo Lee[1] · Songwook Lee†

요 약 : 본 연구의 목적은 한국어 형태소 분석기의 복합명사 분석에 이용할 수 있는 좌우접속규칙을 오류 정보를 이용하여 구축하는 것이다. 우리는 복합명사를 웹사이트로부터 수집하였고 CnuMa 형태소분석기를 이용하여 형태소를 분석하였다. 오류가 발견되면 그 오류를 수정할 수 있는 명사 접속 규칙을 구축하였으며, 명사 좌우 접속 규칙은 복합명사내의 좌우 문맥을 고려하여 작성되었다. 오류에 기반한 좌우접속규칙은 한국어 형태소 분석기인 CnuMa 형태소분석기의 정확률과 재현율을 각각 2.8%, 10.8% 향상시켰다.

주제어 : 형태소분석, 좌우접속정보, 복합명사

Abstract: The goal of this research is to develop an error-driven noun-connection rules which is used for breaking complicate nouns in Korean morphology analysis module. We collected complicate nouns from Web sites, and analyzed them by CnuMa. Whenever we find errors from outputs of the analyzer, we write noun-connection rules to correct the errors. The noun-connection rules are devised by considering left/right contexts in compound nouns. The error-driven noun-connection rules are helpful in improving precision and recall of a Korean morphology analyzer, CnuMa by 2.8% and 10.8%, respectively.

Key words: morphological analysis, nouns connection table, compound noun

1. 서 론

한국어 형태소 분석 기술은 검색엔진의 색인어 추출시스템, 자동 기계번역 시스템, 정보추출 시스템, 자동 문서 클러스터링 등 자연언어처리 기술 응용 분야에서 가장 기본이 되는 요소기술이며 형태소 분석의 오류가 그 응용프로그램의 오류에 직결되므로 상당한 정확성이 요구된다. 무한한 복합명사 생성이 가능한 한국어 특성 때문에 복합명사 분석은 한국어 형태소 분석에서 가장 어려운 문제 중의 하나이다[1].

형태소 분석 방법은 최장일치 head-tail 분석법[2], CYK 알고리즘에 기반한 tabular parsing과 접속정보를 이용한 방법[3]이 있으며, 음절 정보를 이용한 방법[4]도 있다. 대부분의 형태소 분석 방법에서 사용하는 가장 중요한 정보는 각 형태소의 배열 규칙을 표현하고 있는 접속정보이다. 이 형태소별 접속정보가 올바른 품사열을 결정하기 위해 사용된다. 일반적으로, 접속정보는 대량의 말뭉치에서 자동으로 추출하거나 수동으로 구축하여 사용한다. 이와 같은 접속정보 중, 각 형태소별로 왼쪽에 나열될 수 있는 형태소의 종류와 오른쪽에 나열될 수 있는 형태소의 종류를 정리해 놓은 것이 **Table 1**의 좌우접속 정보이다[5]. 좌우접속 정보에는 **Table 1**과 같이 접속을 제약하기 위해 형태소 분석기에서 사용하는 품사 태그보다 더 세부적인

† 교신저자(한국교통대학교 컴퓨터정보공학과, E-mail: leesw@ut.ac.kr, Tel: 043-841-5464)
1 충남대학교 정보통신공학과, E-mail: kjoolee@cnu.ac.kr, Tel: 042-821-5662

품사 범주를 사용한다.

일반적으로 한국어 형태소 분석기는 어절이 입력단위이며, CYK 알고리즘 등을 이용하여 형태소열을 분석한다. 이 때 다수의 후보열이 분석되는데, 좌우접속 정보 등의 형태소 배열 규칙을 이용하여 올바른 후보열들만 선택한다. 최종적으로 형태소 후보열은 은닉 마코프 모델 등의 가중치망을 통과하여 최적의 형태소열로 결정된다[6,7].

Table 1: Examples of noun-connection rules

대범주	소범주	세부 범주	예)	좌접속 품사	우접속 품사
명사	고유 명사		홍길동	접속불가	보통명사, 조사, 인간고유명사형 접미사
	보통 명사	한자형	분양, 실시	한자형 접두사	조사, 한자형접미사, ...
		순한글형	벼락, 노래	순한글형 접두사	보통명사, 조사,
	의존 명사	단위성	회, 마리	숫자, 양수사	조사
		기타	것, 등	접속불가	조사

복합명사의 경우 한글 맞춤법상 띄어쓰기를 원칙으로 하고 있으나, 실제로 많은 복합명사의 띄어쓰기가 다소 자유롭게 이루어지고 있다. 형태소분석 기술의 한계로 인하여 추출된 어절이 복합명사인 경우에 이를 분석하는 과정에서 과잉 분석으로 인한 오류의 정도가 분해하지 않을 때보다 더 심하게 나타나며 일반적으로 복합명사를 포함한 어절은 명사들의 조합으로 구성된 복수개의 형태소 후보열로 분석된다[1]. 명사들의 조합으로 구성된 형태소 후보열은 일반적인 좌우접속 정보로는 잘못된 분석을 제한하기 어려운 경우가 많다. 다음 어절 '비유기적'의 형태소 후보열을 보자.

(a) 비(접두사)+유기(명사)+적(접미사)
(b) 비유(명사)+기적(명사)

위의 두 형태소열은 모두 한국어 형태소 조어상 올바른 형태소열이나 (b)의 결과는 잘못된 분석결과이다. 우리는 위와 같은 복합명사의 입력이 주어졌을 때 (b)의 분석은 제한하고 (a)의 분석 결과만 남길 수 있는 좌우접속 규칙 정보를 오류 패턴으로부터 구축하고자 한다.

본 연구에서 기반이 되는 CnuMa[1] 형태소 분석기는 좌우접속정보를 사용하며 CYK 알고리즘을 기반으로 한 한국어 형태소 분석기이며 통계적 기계번역 등의 연구에 사용되었다[8]. CnuMa 형태소 분석기는 복합명사 분석을 최소 단위로 쪼개서 하도록 설계되었다.

2. 오류 기반 좌우접속 정보 구축

다음은 CnuMa의 복합명사 분석에서 나타난 오류의 예이다. 아래의 예와 같이 복합명사가 조사 등과 함께 쓰일 때, 각 형태소의 좌우접속을 의미적, 문법적 범주로만은 제한할 수 없는 경우가 발생한다. 복합명사의 조어법이나 그 쓰임을 보면 사람들이 관습적으로 단어의 의미중의성을 피하는 성향이 존재한다.

구급차도 => 구급+차도(X), 구급+차+도(O)
구급차내 => 구급+차내(X), 구급+차+내(O)
노래방안 => 노래+방안(X), 노래방+안(O)
노래방도 => 노래+방도(X), 노래방+도(O)
노래방법 => 노래+방법(O), 노래방+법(X)
단계화해서 => 단계+화해+서(X), 단계화+해서(O)
단계화방안 => 단계+화방+안(X), 단계화+방안(O)

본 연구에서는 이런 복합명사 분석 오류를 수집한 후, 그 분석 오류코퍼스에서 반복적으로 나타나는 접속 패턴을 찾아 좌우접속규칙에 반영하고자 한다. 다음 **Figure 1**은 제안 방법의 단계도이다.

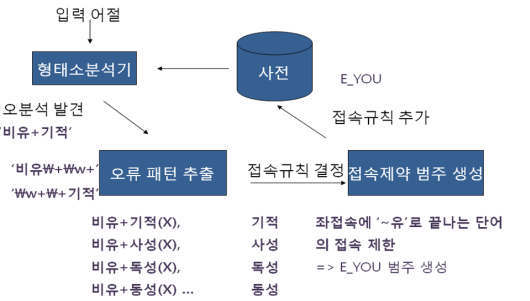

Figure 1: Steps for acquiring noun-connection rules

1) http://marble.cnu.ac.kr/~kjoolee/CNUMA/cnuma.html

Figure 1에서와 같이, 입력 어절로 ‘비유기적’이라는 단어가 주어졌을 때, 형태소 분석기가 ‘비유+기적’으로 잘못 분석하였다고 가정하자. 주어진 오분석으로부터 ‘비유\+\w+’형태와 ‘\w+\+기적’의 오류 패턴들을 추출하여 정규표현[2]으로 나타낸다.

다음 예는 오류코퍼스로부터 정규표현 ‘비유\+\w+’에 의해 추출된 오류 리스트들이다.

[비유+기적(X), 비유+사성(X), 비유+독성(X), 비유+동성(X) …]

이러한 오류들을 관찰해 보면, 오류 리스트에서 발견한 ‘기적’, ‘사성’, ‘독성’, ‘동성’등의 단어의 왼쪽에는 ‘비유’와 같이 음절‘~유’로 끝나는 명사들이 접속을 꺼리는 경향이 있다. 이는 기적’, ‘사성’, ‘독성’, ‘동성’등의 단어들이 모두‘유~’와 결합하여 의미있는 어절을 형성하게 되고 이에 따른 불필요한 의미 중의성을 피하고자 하기 때문이다. 따라서 이런 단어들의 좌접속에 ‘~유’로 끝나는 단어의 접속을 제한하는 규칙을 만들 수 있다. 그러기 위해 좌접속 정보로 제한할 단어들을 범주화한다. 현재 좌접속으로 제한할 단어는 ‘비유’인데 이 단어에 ‘E_YOU’ 범주를 생성하고 사전에 저장한다. 그런 후, 새로 만든 범주 ‘E_YOU’로 접속을 제한해야 할 단어들의 접속 정보를 수정한다.

다음 **Table 2**는 사전에 저장된 단어‘비유’와 ‘기적’, 접사 ‘비’의 좌우접속 정보와 범주 정보를 나타낸 것이다.

Table 2: New word category and its usage in the left-connection rule in the dictionary

단어	품사	범주	좌접속	우접속
비유	NNG	**E_YOU**	NNG, NNP, S(심볼)	NNG,NNP,XSV,J (각종조사), VCP (서술격조사), S
기적	NNG	NULL	**-E_YOU**, NNG, NNP, S	NNG,NNP,XSV,J, VCP, S
비	XSN	BI	NNG, S	J, VCP, XSN
비	XPN	BI	S	NNG

2) 본 논문에서는 Perl로 기술되었음.

Table 2에서 보는 바와 같이, 하나의 단어는 형태소 분석기가 사용할 3가지의 정보를 갖는다. ‘품사’는 21세기 세종계획[9]의 품사 체계를 사용하였다. ‘범주’는 형태소 분석기가 결합정보로써 사용하기 위한 단어의 세부정보를 의미하며 본 연구에서는 주로 의미적, 구문적인 분류보다는 단어의 형태에 따른 분류를 사용한다. 즉, 범주는 주로 단어의 시작과 끝 음절 정보를 이용하여 정의한다. ‘**S_**’는 시작음절 정보를 나타내며, ‘**E_**’는 끝음절 정보를 나타낸다. 그렇기 때문에 각 단어는 형태에 따른 최대 2개의 범주 정보를 가질 수 있다. **Table 2**의 접두사‘비’와 접미사 ‘비’는 시작음절과 끝음절이 모두 동일하므로 단순히 ‘BI’라는 범주를 갖고 있음을 볼 수 있다. ‘좌접속’ 및 ‘우접속’은 각 단어의 왼쪽, 오른쪽에 나타날 수 있는 단어의 부류를 의미하는데, 여기에는 품사정보와 세부 범주 정보가 함께 사용될 수 있으며 값 앞에‘-’부호가 사용될 경우, 해당 품사나 범주는 접속이 허용되지 않음을 의미한다.

앞 예제의 경우, ‘비유’에는 ‘E_YOU’ 범주를 추가하고 ‘기적’, ‘사성’, ‘독성’, ‘동성’등의 좌접속에 ‘-E_YOU’를 추가하였다 (‘-’는 접속 제약을 의미함). 이렇게 함으로써 ‘비유기적’이라는 입력에 대해‘비유+기적’의 오분석을 막을 수 있다.

우접속 제약 규칙도 위와 동일한 방법으로 찾을 수 있다.

접속정보 생성 과정을 정리해 보면, 다음과 같은 4단계를 거쳐서 수행한다.

1) 오분석 발견:
예:“경제연구도” → 경제연+구도(X)

2) 오류 패턴 추출:
예: 오류 패턴‘경제연\+\w+’을 작성하고, 이를 이용해 오류코퍼스로부터 오류 리스트 [경제연+구도(X), 경제연+구원(X), 경제연+구상(X) …]를 추출한다.

3) 접속 제약 범주 생성:
예: ‘경제연’의 우접속에 ‘구~’로 시작하는 명사의 접속 제한이 필요하므로, ‘S_GU’의 접속 제약 범주를 만든다.

4) 접속 규칙 추가:

예: **Table 3**과 같이‘구원’, ‘구도’, ‘구상’ 등의 단어에 ‘S_GU’ 범주를 추가하고, 경제연’의 우접속 정보에 ‘-S_GU’추가하여 ‘S_GU’범주의 우접속 제한한다.

Table 3: New word category and its usage in the right-connection rule in the dictionary

단어	품사	범주	좌접속	우접속
구원	NNG	**S_GU**	NNG, NNP, S	NNG,NNP,XSV,J, VCP, S
경제연	NNG	NULL	NNG, NNP, S	**-S_GU,** NNG,NNP,XSV,J, VCP, S

3. 자료 수집 및 오류에 따른 좌우접속정보 구축

본 연구에서 제안한 방법론을 적용해 보기 위한 원시말뭉치 데이터는 뉴스 포탈사이트에서 수집하였다. 수집한 말뭉치의 중복을 제외한 어절의 개수는 다음과 같다.

- 야후블로그: 931,936 어절
- 뉴스기사: 1,366,399 어절

원시 데이터에서 HTML태그를 제거한 후 CnuMa을 이용하여 형태소 분석을 수행하였다. 그 중 명사류 품사를 포함하는 어절 1,601,230개를 추출하고 오분석을 수동으로 찾아내었다.

명사류 접속정보의 불완전성으로 인한 형태소 분석의 오류 유형은 크게 다음과 같이 다섯 가지로 나눌 수 있으며, 그에 따른 각각의 해결 방안을 살펴보자.

* 명사+접사류 오류

본 논문에서 사용하는 형태소 분석기는 형태소 분석의 과분석을 막기 위해 접사의 명사 접속을 최대한 제한시켜 놓았다. 그러다 보니 명사와 접사에 대한 접속 정보가 부족하여 발생하는 분석 오류들이 많이 발견되었으며, 이런 경우 명사의 좌/우접속규칙에 해당 접사의 범주를 추가하여 오류를 수정한다.

간병비 간병비/UNKNOWN
비인도적 비인도적/UNKNOWN

==> 간병(우접속에 범주 BI 추가)
==> 인도적(좌접속에 범주 BI 추가)

* 복합명사 분석 오류

복합명사 분석 오류는 본 연구의 대상이 되는 오류로 2절에서 제시한 방법으로 해결한다.

가두판매대가 가두/NNG+판매/NNG+대가/NNG
가정통신문의 가정/NNG+통신/NNG+문의/NNG
고지마이론과 고지/NNG+마이/NNG+론/XSN+과/JC
관리부문의 관리/NNG+부/XSN+문의/NNG

* 명사+용언상당어구 오류

명사와 용언이 결합한 어절에서 발생하는 오류인데 명사와 용언의 경계를 잘못 나누거나 명사와 용언의 접속 정보가 없어 분리를 못하는 오류가 대부분이며 이런 형태소 분석 오류는 명사의 우접속규칙에 접속하는 용언의 품사나 범주를 추가하여 수정한다.

가감없는 가감없는/UNKNOWN
매일보던 매/NNB+일/XSN+보/NNG+이/VCP+더/EP+ㄴ/ETM
==> 가감 (우접속에 VA 품사 추가)
==> 매일 (우접속에 VV 품사 추가)

* 맞춤법 및 띄어쓰기 오류

한글 맞춤법 오류를 포함하고 있는 어절과 띄어쓰기 오류를 포함하고 있는 어절에서 발생하는 오류이다. 입력 어절 자체에 오류를 포함하고 있으므로 분석 결과에도 오류를 포함하고 있는 경우가 많으며 맞춤법 오류와 띄어쓰기 오류에 대한 특별한 처리 모듈이 없으면 대부분 잘못 분석되게 된다. 이런 오류 중 명사 접속규칙으로 수정할 수 있

는 오류만 규칙 생성 대상에 포함된다.

고정하실수 고/XPN+정하/NNP+실수/NNG

==> 정하 (좌접속에 -KO 추가;
우접속에 -S_SIL추가)

* 미등록어 오류

미등록어의 경우, 해당 형태소 분석기의 어휘 사전에 등록되지 않은 어휘를 말하는데, 어휘 사전에 없는 단어를 포함하고 있는 어절은 형태소 분석기에서 미등록어 추정을 별도로 수행하지 않는 한, 등록 어휘의 조합으로 잘못 분석될 수밖에 없다. 이런 미등록어로 인해 발생하는 오류는 미등록어를 사전에 등록하고 각 단어의 범주에 맞는 접속정보를 추가하여 오류를 수정한다.

개그짤방 개그짤방/UNKNOWN
가디언지 가디언지/UNKNOWN

==> 짤방(사전 등록 후, 좌접속에 품사 NNG 추가)
==> 가디언(사전 등록 후, 우접속에 범주 JI 추가)

* 규칙 충돌 오류

어떤 단어에 여러 개의 좌우 접속 규칙으로 인해 규칙의 충돌이 발생할 수도 있는데 이 때에는 새로운 범주를 만들어 규칙의 충돌을 막는다.

예를 들어,'충남'의 우접속에 '-S_DAE'가 존재하고,'대로'의 범주에 'S_DAE'가 존재한다고 가정하면 아래와 같은 분석 결과는 나올 수 없다.

충남대로 충남/NNP+대로/NNG(O)

==> 대로(범주에 DAERO 추가), 충남(우접속에 DAERO 추가), 새로운 범주를 추가하여 두 단어가 서로 접속이 가능하도록 만든다.

이와 같은 형태소분석 오류를 수정하기 위해 좌우접속 정보에 새로 추가된 범주는 1,072개이며, 이를 이용하여 수정된 좌우접속 정보의 개수는 총 4,493개이다. 좌우접속 정보의 수정 작업은 규칙의 일관성을 위해 한 사람이 작업하였으며, 규칙 구축 작업에 소요되는 시간은 15개의 좌우접속 정보를 수정하는데 약 한 시간 정도 걸린다.

4. 형태소 분석기 적용 결과

제안된 방법으로 구축된 좌우접속 정보를 한국어 형태소 분석기인 CnuMa에 적용하여 성능을 살펴보았다. 성능평가를 위한 평가셋은 좌우접속 정보 구축에 전혀 사용되지 않은 데이터로 구성되어 있으며 그 구성은 다음 **Table 4**과 같다.

Table 4: The genres and their number of Eojeols in the test set

종류	어절수
웹블로그	2,283
지식인(Q&A)	3,138
개체명	2,132
백과사전	3,348
뉴스	3,096
경제전문지	3,990
MS메신져	3,000
합계	20,987

다음 **Table 5**와 **Table 6**은 제안방법을 적용하기 전후의 CnuMa의 정확률과 재현율의 변화를 나타낸다.

Table 5: Precisions of CnuMa

정확률	전	후	향상
웹블로그	75.7%	80.2%	4.5%
지식인(Q&A)	84.3%	85.8%	1.4%
개체명	70.7%	78.5%	7.8%
백과사전	82.6%	84.7%	2.2%
뉴스	80.1%	80.7%	0.6%
경제전문지	85.2%	87.1%	1.9%
MS메신져	82.2%	83.6%	1.4%
평균	80.1%	82.9%	2.8%

Table 6: Recalls of CnuMa

재현율	전	후	향상
웹블로그	79.7%	85.0%	5.3%
지식인(Q&A)	87.1%	88.6%	1.5%
개체명	68.1%	85.2%	17.1%
백과사전	65.2%	83.5%	18.3%
뉴스	69.5%	82.9%	13.4%
경제전문지	65.6%	85.5%	19.9%
MS메신져	84.7%	84.8%	0.1%
평균	74.3%	85.1%	10.8%

Table 5와 **Table 6**에서 보는 바와 같이 제안방법을 적용하여 약 2.8%의 정확률 향상과 10.8%의 재현율 향상을 가져왔다. 특히, **Table 5**에서 보는 바와 같이 주로 명사들로 구성된 '개체명' 평가셋의 경우 형태소 분석의 '전' 결과가 다른 평가셋에 비해 현저히 낮음을 볼 수 있다. 즉, 개체명으로부터 단위 명사들을 정확히 분리해 내는 것이 어렵다는 것을 쉽게 알 수 있다. 그러나, 본 연구에서 제안한 방법을 사용하면 '개체명'이 다른 분야에 비해 월등히 높은 7.8%의 정확률 향상을 얻을 수 있었다. 즉, 본 연구에서 제안한 방법론은 복합명사를 정확히 분석하는데 매우 유용함을 알 수 있다.

5. 결론 및 향후과제

본 연구에서는 그 복합명사 분석의 정확도를 높일 수 있는 좌우접속규칙을 구축하였다. 대용량 말뭉치에서 수집된 복합명사를 CnuMa 형태소 분석기를 이용하여 분석하였으며, 오분석을 수동으로 찾아내어 이들로부터 좌우접속을 제한하는 규칙을 반자동으로 구축하였다. 구축된 좌우접속규칙을 이용하여 CnuMa의 정확률과 재현율을 각각 약 2.8%, 10.8% 향상시켰다. 본 연구에서 사용한 방법을 적용하여 다른 한국어형태소 분석기의 후처리에 사용하면 형태소 분석기의 성능을 향상 시킬 수 있을 것이다. 앞으로 GUI 사용자 인터페이스 등을 개발하여 좌우 접속 규칙 구축 작업의 효율을 높일 필요가 있다.

후 기

이 논문은 2012년도 한국교통대학교 교내학술연구비의 지원을 받아 수행한 연구임.

참고문헌

[1] Bo-Hyun Yun, Min-Jeung Cho and Hae-Chang Rim, "Segmenting korean compound nouns using statistical information and a preference rule", Journal of Computing Science and Engineering, vol. 24, no. 8, pp. 900-909, 1997.

[2] Jae-Hyuk Choi and Sang-Jo Lee, "A method for reducing dictionary access with bidirectional longest match strategy in korean morphological analyzer", Journal of Computing Science and Engineering, vol. 20, no. 10, pp. 1497-1507, 1993.

[3] Seong-Yong Kim, A Morphological Analyzer for Korean Language with Tabular Parsing Method and Connectivity Information, Master's Thesis, KAIST, 1987 (in Korean).

[4] S.S. Kang, Korean Morphological Analysis Using Syllable Information and Multi-word Unit Information, Doctoral Thesis, Seoul National University, 1993 (in Korean).

[5] Dong Un An, "A noun extractor using connectivity information", Proceddings of The 11th Annual Conference on Human & Cognitive Language Technology, pp. 173-178, 1999 (in Korean).

[6] Jae-Hoon Kim, "Korean part-of-speech tagging using a weighted network", Journal of Computing Science and Engineering, vol. 25, no. 6, pp. 951-959, 1998.

[7] Woon-Jae Lee, Design and Implementation of an Automatic Tagging System for Korean Texts, Master's Thesis, KAIST, 1993 (in Korean).

[8] Kong Joo Lee, Songwook Lee and Jee Eun Kim, "A bidirectional korean-japanese statistical machine translation system by using MOSES", Journal of the Korean society of Marine Engineering, vol. 36, no. 5, pp. 683-693, 2012 (in Korean).

[9] The 21st Century Sejong Project, http://sejong.or.kr/sejong_kr/index.html, Accessed 20. Nov. 2006.