Journal of Korean Language Education
한국어교육연구 제12권 1호, 2017.02. pp.57-70.

한국어 형태소 분석기 개발을 위한 알고리즘

김 현 주 · 이 영 상 · 천 승 태

한국어교육연구
Journal of Korean Language Education

『한국어교육연구』 제12권 1호, 2017.02. pp.57-70.

한국어 형태소 분석기 개발을 위한 알고리즘*

김현주·이영상·천승태(데이트스트림즈)

차 례

1. 기본구성
2. 형태소 분석기 알고리즘
3. 복합명사
4. 결론

<국문 초록>

본 연구는 규칙 기반의 한국어 형태소 분석기 개발을 위한 기본 알고리즘을 제시하고자 한다. 이를 위한 세부 사항은 다음과 같다: 1) 규칙 기반의 방법을 적용하였으며 2) 어절을 정의함으로써 분석 대상의 숫자를 줄일 수 있었으며 3) 활용빈도가 매우 높은 복합명사를 분석하기 위한 별도의 알고리즘을 제시하였다.

어절의 유형은 명사형과 동사형 및 부사 및 기타 유형으로 구분되는데 형태소 분석기의 품질(속도와 정확성)을 위하여 먼저 어절의 유형을 단순화할 필요가 있다. 이에 단순화된 어절 유형으로 이영민(2017)을 수용, 어절의 수를 15개로 한정하고자 하였다: 명사 유형 어절로는 조사가 통합한 형태를 중심으로 서술격 조사와 (선어말) 어미가 통합한 형태, 그리고 이에 다시 조사하 통합한 형태, 명사에 용언화 접사가 통합되고 다시 어미, 조사가 통합된 형태를 제시한다. 그리고 동사 유형 어절로는 동사에 (선어말) 어미가 통합한 형태, 그리고 조사가 통합한 형태를 제안한다. 나아가 부사 유형의 어절과 별도의 예외적 형태의 어절을 제안한다. 이에 더하여 선어말 어미 복합체와 조사 복합체를 구성하여 어절의 숫자를 더 단순화하고자 하였으며 불규칙 용언을 처리하기 위한 알고리즘도 염두에 두었다.

이러한 어절을 바탕으로 어절의 뒤에서부터 조사와 어미를 15개의 어절을 대상으로 분석해 나가는 방식을 적용하였다. 이에 더하여 가장 활용빈도가 높은 명사복합체를 분석할 수 있는 알고리즘을 제시하였는데 최대 14음절의 명사복합체를 분석이 가능하도록 하였다.

주제어: 형태소 분석기, 알고리즘, 문법기반, 유형 정의, 복합명사

1. 기본구성

한국어는 교착어로서 그 특성상[1] 굴절어나 고립어에 비하여 형태소 분석기를 개발하기가 쉽지가 않다. 문법범주와 어휘범주가 결합되어 과정에서 복잡한 형태론적, 음운론적 특성으로 인하여 그 원래의 형태를 분석하기가 쉽지 않기 때문이다. 이에 본 연구는 이와 같은 한국어의 특성을 고려하면서 최적의 형태소 분석기를 개발하기 위한 알고리즘을 제시하고자 한다. 기본 구성은 처리 단위와 처리 방법에 따라 다양한 방식이 적용될 수 있는데 처리 단위는 한국어의 띄어쓰기 단위인 '어절'을 처리 단위로 한다. 어절을 중심으로 어휘범주와 문법범주를 분석해 내는 데 그 방법은 규칙 기반의 방법을 적용한다.[2] 규칙 기반의 방법은 먼저 한국어 사전을 구축하고 한국어 문법을 연구, 규칙을 도출한다. 그리고 이를 통하여 한국어 형태소를 분석해 내는 방법이다. 이러한 방법은 1) 사전을 구축하고 규칙을 정립하는 데 많은 시간과 노력이 요구되며, 2) 언어 변화에 따라 사전과 규칙을 주기적으로 관리해주어야 하는 어려움이 있다. 이러한 작업이

* 이 논문은 2016년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임(R0126-15-1067, (ICBMS-1세부) CoT(Cloud of Things) 환경에서 실시간 반응성 향상을 위한 계층적 데이터 스트림 분석 SW 기술 개발).

1) 언어 유형에 관한 연구는 Palmer(1994) 참조.

2) 규칙 기반의 방법은 김성용 (1987), 강승식(1993), 권오욱 등 (1999) 참조. 한편 통계 기반의 방법은 이도길(2005), 이재성(2011), 나승훈(2012) 참조).

선/후행되지 않는 않는다면 그 성능을 보장받을 수 없다. 그렇지만 이러한 작업이 가능하다면 분석 속도가 빠르고 구조화된 문서에 대한 분석 성능을 높일 수 있다는 장점이 있다(이영민(2017) 참조).

속도를 향상시키기 위해서는 무엇보다도 분석의 대상이 되는 어절이 적절히 정의되어 있어야 한다. 대상 어절의 숫자가 너무 많으면 속도가 느려질 수밖에 없는데 이를 위해서는 필요한 만큼의 어절을 정의하는 것이다. 어절의 정의는 이영민(2017)을 수용, 15개의 어절 유형으로 정의한다.

(1) 어절 유형

구분	유 형	구 성
1	NJ	명사-조사
2	NCE	명사-서술격조사-어미
3	NCPE	명사-서술격조사-선어말어미-어미
4	NCEJ	명사-서술격조사-어미-조사
5	NCPEJ	명사-서술격조사-선어말어미-어미-조사
6	NXE	명사-용언화접사-어미
7	NXPE	명사-용언화접사-선어말어미-어미
8	NXEJ	명사-용언화접사-어미-조사
9	NXPEJ	명사-용언화접사-선어말어미-어미-조사
10	VE	동사-어미
11	VPE	동사-선어말어미-어미
12	VEJ	동사-어미-조사
13	VPEJ	동사-선어말어미-어미-조사
14	ADV	부사-조사
15	-요	어미, 조사, 부사-요

2. 형태소 분석기 알고리즘

(1)의 어절을 대상으로 어절의 뒤에서부터 조사와 어미를 확인하는 작업이 이루어진다.

(2) ㄱ. 감은 좋은 과일이다
ㄴ. 사람들이 머리를 감은 이유는

(2)에서 '감은'은 '감(N)+은(J), (2ㄱ)'과 '감(V)+은(E), (2ㄴ)'으로 분석되는데 어절의 뒷부분에서부터 분석을 적용, 조사 '-은(J)'과 어미 '-은(E)'이 모두 사전에서 확인된다. 이에 조사와 어미를 분리해내면 각각 선행하는 명사와 동사를 사정에서 확인, 명사 '감(N)'과 동사 '감-(V)'을 분석해 내는 것이다.

(2) ′ㄱ. *감는 좋은 과일이다
ㄴ. 사람들이 머리를 감는 이유는

(2)′도 (2)와 같은 방식으로 분석되는데 (2′ㄱ)의 '감(N)+는(J)'은 잘못된 분석이므로 배제하여야 한다. 이는 조사 '-은/는'의 결합이 선행 명사의 마지막 음절에 따라 다르게 실현하도록 함으로써 해결할 수 있다.

(1)의 어절 유형은 '유형15'만 제외하면 어미와 조사를 분석해 낼 수 있도록 되어 있으므로 별도의 규칙을 적용하지 않고 일관되게 결과를 도출해 낼 수 있다. 물론 표면형으로 조사가 실현되지 않은 경우는 단순 명사로 처리할 수 있다.

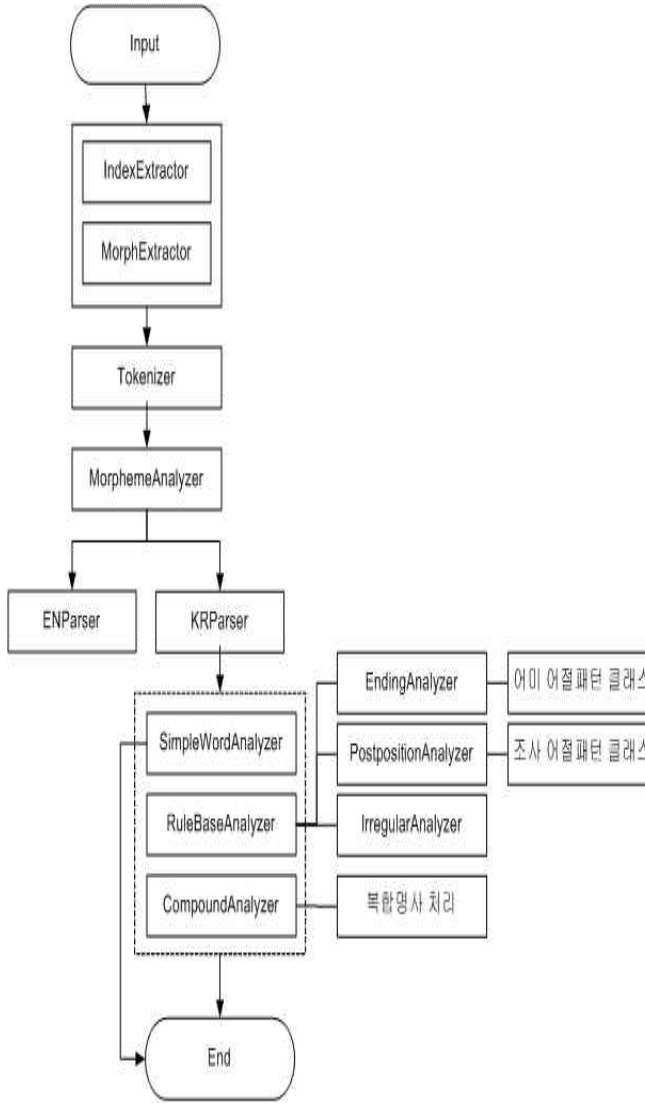
(3) ㄱ. 진주만
ㄴ. 진주(N)-만(J)/진주만(N)

(3)의 '진주만'은 ((2)와 같은 방식으로)명사 '진주'와 조사 '-만'으로 분석되지만 진주만 전체가 하나의 명사로 사전에 등재되어 있으므로 '진주만(N)'으로도 분석되는 것이다.

이에 더하여 조사가 복합적으로 실현되거나(최태성(1999)), 선어말 어미가 중복실현 되는 경우는 그 목록을 정의함으로써 유형의 숫자를 15개로 한정할 수가 있다(이영민(2017) 참조).

이를 바탕으로 형태소 분석기의 기본 알고리즘을 제시하면 [그림1]과 같다.입력(Input)된 대상은 'Index Extractor'와 'Morpheme Extractor'을 거쳐 Tokenizer되어 해당 품사[3]를 확인한다. 이후 'KRParser'에서 조사(Postposition Analyser)와 어미(Ending Analyser)가 ((2)의 예에서 설명한 방식대로) 분석된다. 불규칙 활용의 처리(이영민(2017))도 여기서 이루어진다.

3) 엄밀히 말하면 '품사 확인'이라고 할 수 없다. 다만 논의의 편의를 위하여 사전에 등재된 정보를 이르는 의미로 사용한다.



[그림 1] 형태소 분석기 알고리즘

3. 복합명사

이 절에서는 복합명사를 분석하기 위한 알고리즘을 제시한다. 알고리즘의 설계는 기본적으로 이영민(2017)을 수용한다. 이영민(2017)을 정리하면 다음과 같다: 복합명사는 파생명사와 합성명사[4]를 아우르는 개념으로서 사전에 등재된 명사를 이른다. 따라서 해당 어절에서 조사만 분리해낼 수 있으면 일반 명사와 마찬가지로 형태소 분석의 알고리즘에는 부담이 되지 않는다. 그런데 현실은 그렇지 않다. 많은 분야에서 복합명사를 명사복합체의 개념으로 사용하고 있으며 어절별로 띄어쓰기를 하지 않는 경우가 대부분이다.

(4) ㄱ. 교양소설, 인공지능, 두꺼비집, 큰아버지
ㄴ. 질소화합물,
ㄷ. 남북국시대, 복소수평면
ㄹ. 연결재무제표
ㅁ. 중거리탄도유도탄

(5) ㄱ. 계좌번호, 거래실적, 연구성과, 인접과학, 남북분단
ㄴ. 한국인류학, 항공승무원, 자기효능감
ㄷ. 세계화시대, 성분별조사
ㄹ. 전국표본조사, 회계추정방법, 국제회계기준
ㅁ. 연차별도산출가액, 대중국무기수출가

(4)의 예들은 사전에 등재된 단어들로서 복합명사로 처리된다. 이를테면 '교양소설을', '중거리탄도유도탄은'은 '교양소설(명사)+을(조사)', '중거리탄도유도탄(명사)+은(조사)'로 분석되며 문제가 되지 않는다[5][6].

---

4) complex와 compound를 각각 복합과 합성으로 이해한 것이다(이익섭·임홍빈(1983), 고영근·남기심(1991), 이익섭·채완(2012)) 참조)

5) 물론 형태소 분석으로는 결과가 다르게 나타나야 하겠지만 명사를 정확히 분석해내는 것이 목적이므로 이와 같은 처리는 문제가 되지 않는다. 현재의 설명은 국립국어원의 표준국어대사전을 대상으로 한 것이다.

반면에 (5)의 예들은 사전에 등재되어 있지 않으므로 문제가 복잡해진다. '계좌번호가'와 같이 다소 간단한 경우도 '계좌번호'가 사전에 등재되어 있지 않으므로 분석되지 않거나 조사를 분리해 낸다 하더라도 '계좌번호(UN[7])+가(조사)'로 분석될 수밖에 없다. 따라서 별도의 분석 알고리즘이 필요한데, 대부분의 4음절 복합명사는 그 내부를 확인하기가 어렵지 않다. 그렇다 하더라도 조사를 분리해 내고 별도의 알고리즘을 사용하여 내부의 명사를 확인해야 한다. 이를테면 '계좌번호를'은 일단 조사를 분리해낸 상태에서('계좌번호(UN)+를(조사)') 미확인 부분인 '계좌번호'를 분석해내는 알고리즘을 적용하는 것이다. 여기서는 4음절 이하의 복합명사는 간단한 알고리즘을 적용하기로 한다. 4음절 복합 명사의 경우 대부분이 '2+2'의 구성으로 되어있으므로 이를 확인하는 알고리즘을 적용하고 그렇지 않은 경우는 '3+1', '1+3'으로 분석하는 것이다. 이를테면 '일회용침'은 '2+2'의 구성으로 분석하면 '일회(명사)+용침(UN)'으로 분석되는데 이럴 경우는 다시 '3+1'의 알고리즘을 적용하여 '일회용(명사)+침(명사)'로 분석될 수 있도록 한다는 것이다.

5음절 이상의 복합명사는 모든 경우의 수를 따져서 분석한다. 5음절일 경우 가능한 경우의 수는 다음과 같다.

(6) '11111', '1112', '1121', '113', '1211', '122', '131', '14',
'2111', '212', '221', '23',
'311', '32'
'41',
'5'

이를 적용하면 5음절 복합명사는 24, 6음절은 25, 7음절은 26의 경우의

6) 사전에 등재된 명사의 숫자가 관건이 될 것이다. 그렇지만 언급한 바와 같이 많은 분야에서 복합명사를 명사복합체의 의미로 사용하고 있으며 어절별로 띄어쓰기를 하지 않는 경우가 대부분이므로 모든 '명사'를 등재하는 것 자체가 불가능할 것이다.

7) 확인할 수 없는(unknown) 단어나 어근.

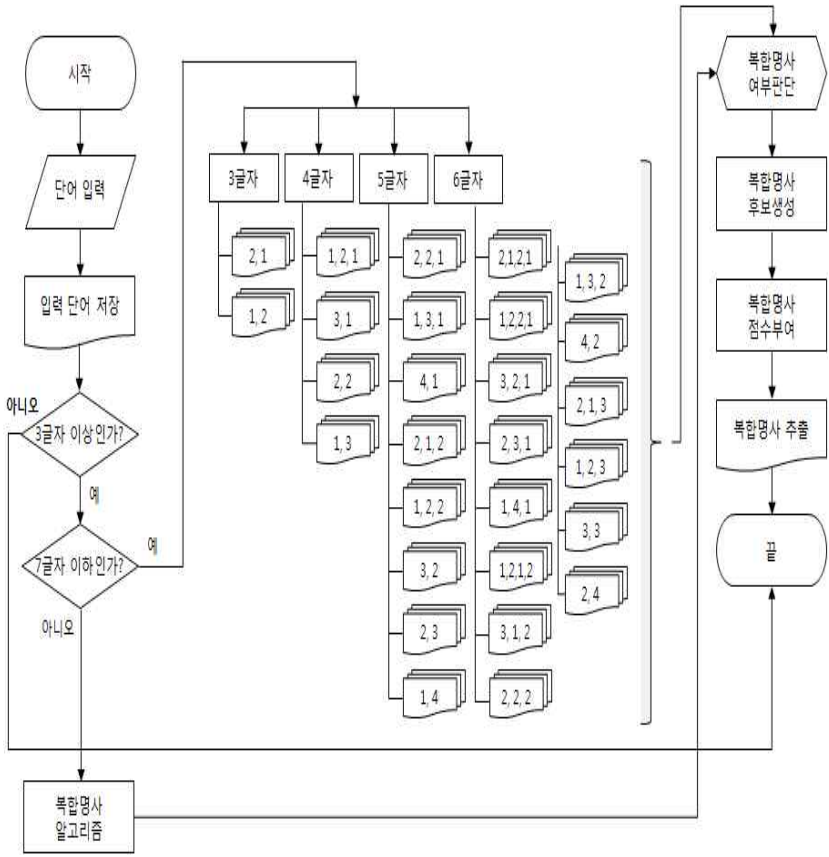
수가 산출된다. 이에 1음절로만 되는 경우('11111')와 전체가 하나의 단어로 등재된 경우(즉 (6)의 '5')를 배제하면 각각 24-2, 25-2, 26-2의 경우의 수가 산출된다. 이를 적용하여 14음절까지 분석 가능한 알고리즘을 적용한다. 5음절 복합명사에 조사 '-이'가 결합된 경우를 예로 들면 다음과 같다.

(7) 총자산가액이
ㄱ. 총자(N)산(N)가액(N)이(N):100_UK
ㄴ. 총(N)자산(N)가액(N)이(N):100_UK
ㄷ. 총자산(N)가액(N)이(N):100_UK
ㄹ. 총자(N)산가(N)액(N)이(J):100:_NJ
ㅁ. 총(N)자산가(N)액(N)이(J):100:_NJ
ㅂ. 총자(N)산(N)가액(N)이(J):100:_NJ
ㅅ. 총(N)자산(N)가액(N)이(J):100:_NJ
ㅇ. 총자산(N)가액(N)이(J):100:_NJ
ㅈ. 총자(N)산가(N)액이(UN):65:_UK
ㅊ. 총자산(N)가(N)액이(UN):60:_UK
ㅋ. 총(N)자산가(N)액이(UN):60:_UK
ㅌ. 총자산가액(UV)이(E):50:_VE
ㅎ. 총자산가액(UN)이(J):50:_NJ ...[8)]

(7ㄱ-ㅇ)은 확인되지 않은 형태가 없으므로 전부 수용가능하다(100점). 다만 (7ㄱ-ㄷ)은 유형에서 확인되지 않으므로(1장 참조) 배재할 수 있다. 그리고 (7ㄹ-ㅇ)은 일단 수용가능한 것으로 보는데 이들 중에서 음절 수가 가장 적은 것을 선택하도록 하면 (7ㅇ)이 선택된다. (7ㅈ-ㅎ)은 확인되지 않은 형태가 있으므로 배제된다. (7ㅎ)이 더 정확한 분석일 수도 있지만 '총자산가액'이 사전에 등재되어 있지 않으므로 확인이 되지 않은 것이다.

이러한 설명을 바탕으로 복합명사를 분석할 알고리즘을 제시하면 [그림 2]와 같다.

8) 더 많은 분석 결과가 산출되나 동일한 설명이 적용되므로 생략함.



[그림 2] 복합명사 알고리즘

4. 결론

이상의 논의를 통하여 한국어 형태소 분석기 개발을 위한 알고리즘을 제시하였다. 본 연구는 형태소 분석기의 성능을 향상시키고 최적의 결과물을 산출하기 위하여 1) 규칙 기반의 방법을 적용하였으며 2) 어절을 정의함으로써 분석 대상의 숫자를 줄일 수 있었으며 3) 활용빈도가 매우 높은 복합명사를 분석하기 위한 별도의 알고리즘을 제시하였다. 유형을 단순화

하기 위해서는 유형자체가 간단하고 간략하게 정의되어야 하는데 이를 위하여 조사 복합체와 선어말어미 복합체를 정의하여 그 경우의 수를 줄이고 나아가 불규칙 활용의 문제를 처리할 수 있도록 고안되었다. 더불어 영문이나 숫자, 기호 등이 사용된 어절(특수어절)의 경우도 별도의 규칙을 제시하여 처리할 수 있도록 하였다(김현주 외(2015) 참조).

<참고문헌>

강승식(1993), 음절정보와 복수어 단위 정보를 이용한 한국어 형태소 분석, 서울대학교 박사학위논문.

고영근·남기심(2014), 표준 국어문법론, 박이정.

권오욱 외(1999), 음절단위 CYK 알고리즘에 기반한 형태소 분석기 및 품사태거, 한국정보과학회 언어공학연구회 학술발표 논문집, 한국정보과학회언어공학연구회, 76-87쪽.

김성용(1987), Tabular parsing 방법과 접속 정보를 이용한 한국어 형태소 분석기, KAIST 석사학위논문.

김현주 외(2015), 문장부호를 고려한 특수어절 알고리즘, 언어정보학회 논문집 22-2, 한국언어정보학회, 1-4쪽.

나승훈 외(2012), CRF에 기반한 한국어 형태소 분할 및 품사 태깅, 한국정보과학회언어공학연구회 2012년도 제24회 한글 및 한국어 정보처리 학술대회 발표집, 한국정보과학회 12-15쪽.

이도길(2005), 한국어 형태소 분석과 품사 부착을 위한 확률모형, 고려대학교 박사학위논문.

이영민(2017), 한국어 형태소 분석기 개발을 위한 문법 전략, 배재대학교 한국어교육연구소, 한국어교육연구 12-1, 149-166쪽.

이익섭·임홍빈(1983), 국어문법론, 학연사.

이익섭·채완(2012), 국어문법론 강의, 학연사

이재성(2011), 한국어 형태소 분석을 위한 3단계 확률 모델, 소프트웨어 및 응용 38-5, 한국정보과학회, 257-268쪽.

Palmer, F.R.(1994), Grammatical Roles and Relations, Journal of Linguistics, Cambridge Press. 297-298쪽

<Abstract>

Optimal Algorithm to develop a Korean morpheme analyser.

This paper aims to present a optimal algorithm for Korean morpheme analyser. The basic strategy is 1) to adopt a rule base strategy, 2)to minimize the targets(15 asal(word + its case and ending)) by grammatical rule and analyse them. It can improve its processing speed. In addition to compress the algorithm, we define the case markers and endings and suggest irregular verbs/adjectives analyse rules. 3) To improve its quality we present a algorithm to analyse compound nouns.

To minimize the target, we adopt Lee youngmyn(2017)'s 15 asal(lexical word + grammatical word(case markers or endings)) type. 1) Noun type asal of Koran typically consist of noun+'josa'(case marker) and josa can be duplicated. 2) Verb type asal of Korean typically consist of verb+'ending' and prefinal ending can be duplicated. The 15 type of asal is : noun+josa, noun+predicate josa+ending, noun+predicate josa+prefinal ending+ending, noun+predicate josa+ending, noun+predicate josa+prefinal ending+ending+josa, noun+verbal suffix+ending, noun+verbal suffix+prefina ending+ending, noun+verbal suffix+ending+josa, noun+verbal suffix+prefinal ending+ending+josa, verb+ending, verb+prefianl ending+ending, verb+ending+josa, adverb type, extra asal type.

In addtion we present an algorithm to analyse a noun compoud of 14 syllables.

Key words: Korean morpheme analyser, algorithm, rule base, 15 asal types, noun compound.

김현주(Kim hyunjoo)
데이터스트림즈 선임연구원
주 소: 서울특별시 서초구 사임당로28, 청호나이스 빌딩 2층, 6층
전자우편: hjookim@datastreams.co.kr
전화번호: (02)3473-9077

이영상(Lee youngsang)
데이터스트림즈 대표
주 소: 서울특별시 서초구 사임당로28, 청호나이스 빌딩 2층, 6층
전자우편: yslee@datastreams.co.kr
전화번호: (02)3473-9077

천승태(Chun seungtae)
데이터스트림즈 연구소장
주 소: 서울특별시 서초구 사임당로28, 청호나이스 빌딩 2층, 6층
전자우편: stchun@datastreams.co.kr
전화번호: (02)3473-9077

■ 접 수 일: 2017.12.23.
■ 심사완료일: 2017.01.28.
■ 게재확정일: 2017.02.10.