Journal of the Korea Academia-Industrial Cooperation Society
Vol. 12, No. 11 pp. 5278-5283, 2011

http://dx.doi.org/10.5762/KAIS.2011.12.11.5278

형태소 합성 기법을 이용한 형태소 패턴 사전의 반자동 구축

박인철[1*]
[1]호원대학교 컴퓨터게임학부

Semi-Automatic Construction of Morphological Pattern Dictionary using the Method of Morphological Synthesis

In-Cheol Park[1*]
[1]**Division of Computer·Game, Howon University**

요 약 초고속 한국어 형태소 분석을 위한 하나의 방법은 사전에 형태소 결과를 미리 저장해 놓고 이를 이용하는 것이다. 이러한 형태소 패턴 사전을 수작업으로 구축하려면 많은 비용이 들 뿐만 아니라 적지 않은 오류가 포함될 수 있다. 본 논문은 한국어 형태소 합성을 이용하여 자동으로 형태소 패턴을 생성하는 방법을 제안한다. 실험을 통해, 올바른 형태소 분석을 위해 사용한 형태소 패턴의 86%를 자동으로 생성함을 알 수 있었다. 형태소 패턴을 이용한 형태소 분석기가 403MB의 한국어 코퍼스를 분석하는 데 걸린 시간은 2.8GHz 윈도우 시스템에서 52.68초였다.

Abstract One approach for very high speed korean morphological analysis is to use pre-built morphological results in dictionary. It pays the high cost to build this morphological pattern dictionary manually, besides the dictionary may contain errors. This paper proposes a method to generate morphological patterns automatically using Korean morphological synthesis. The experiment shows that we automatically generate 86% morphological patterns for analyzing Korean sentences. It takes 52.68 seconds for the morphological system using the patterns to analyze 403MB Korean corpus on 2.8GHz Window system.

Key Words : Morphological analysis, Morphological patterns, Morphological synthesis

1. 서론

전통적인 한국어 문장의 형태소 분석(morphological analysis) 시 발생하는 문제점 중의 하나는 원형 복원을 위한 과도한 사전 탐색이다. 대용량 문서를 처리하기 위해 한국어 형태소 분석의 처리 속도가 중요해짐에 따라 사전에 대한 탐색을 줄이기 위한 여러 연구가 진행되어 왔으며[1-4], 이 중 형태소 분석 결과를 미리 저장하여 이를 분석에 활용하는 접근은 매우 효과적임이 입증되었다[3].

본 논문에서는 문장에 나타나는 어절과 그 어절의 형태소 분석 결과의 쌍을 형태소 패턴(morphological pattern)이라 정의한다.

형태소 패턴은 한국어 처리 시 나타나는 다양한 예외사항을 동일하게 처리할 수 있도록 한다. 예를 들어, 규칙 기반 형태소 분석에서는 어절 “뭘”을 제대로 분석할 수 없기 때문에 해당 어절을 “무엇/대명사+을/조사”와 같은 형태로 분석하기 위해서는 별도의 예외 처리가 필요하다. 그러나 형태소 패턴을 사용한 형태소 분석에서는 단순히 <뭘, 무엇/대명사+을/조사>와 같은 형태소 패턴을 추가함으로써 이를 별도의 예외 처리를 수행하지 않고도 해당 형태소 결과를 생성할 수 있다. 이와 유사하게 흔히 철자가 틀린 단어나 띄어쓰기 오류가 발생하는 어절에 대해서도 형태소 패턴을 등록하여 일괄적으로 처리할 수 있다. 이러한 예로는 동사 “헷갈리다”를 “햇갈리다”로 잘못 사용하는 경우이다.

본 논문은 2011년 호원대학교 교내학술연구비조성비 지원에 의해 연구되었음.
*교신저자 : 박인철(icpark@howon.ac.kr)
접수일 11년 10월 12일 수정일 11년 11월 02일 게재확정일 11년 11월 10일

한국어는 변형이 심한 언어이므로 활용 가능한 모든 어절에 대한 형태소 패턴을 구축하는 것은 사실상 불가능하다[1]. [1]에서는 일반적 사전과 형태소 패턴 사전을 적절히 혼합하여 이를 해결하였고 [3]에서는 용언과 'ㄴ/ㄹ/ㅁ/ㅂ/ㅆ' 받침에 대한 형태소 패턴 사전을 구축하였다.

본 논문에서는 형태소 패턴을 확장하여 모든 분석 가능한 후보들을 사전에 모두 저장한다. 현재 PC의 메모리도 보통 2GB 이상이고 형태소 사전을 구축하기 위해 사용하는 트라이 인덱스[6]는 엔트리의 개수가 탐색 성능에 영향을 미치지 않으므로 이러한 과도한 형태소 패턴의 생성은 고성능 형태소 분석기 개발에 제약을 미치지 않는다. 물론 모든 조사와 어미에 대해 형태소 패턴을 생성하는 것은 불가능하므로 단순한 규칙으로 생성할 수 없는 후보에 대해서만 자동으로 형태소 패턴을 생성할 것이다.

그러나 어절 "내"가 대명사 "나"와 조사 "의"로 분석되는 것과 같은 예외적인 경우에 대한 형태소 패턴은 수작업으로만 추가할 수 밖에 없다. 따라서 본 논문에서 자동 생성의 대상이 되는 형태소 패턴은 규칙에 의해 분석될 수 있는 경우에 한해서이다.

형태소 패턴의 자동 생성을 위해서는 한국어 형태소 합성(morphological synthesis) 기술을 이용한다. 한국어 형태소 합성은 한국어 형태소 분석과 달리 거의 대부분을 단순한 규칙을 적용하여 올바른 결과를 얻을 수 있기 때문에 오류로 인해 부적절한 형태소 패턴을 생성하는 것을 피할 수 있다.

형태소 패턴이 올바로 생성되었는지 여부를 파악하고 형태소 분석을 위해 필요한 전체 형태소 패턴 중에서 어느 정도의 비율로 자동으로 생성되었는지를 파악하기 위해 본 논문에서는 형태소 패턴을 사용한 한국어 형태소 분석기를 개발한다. 구현된 형태소 분석기는 향후 띄어쓰지 않는 문장을 분석하도록 확장하기 위해 좌우 최장일치법을 사용한다. 그러나 단순한 좌우 최장일치법은 복합명사의 처리나 미지어 처리가 어렵기 때문에 띄어쓰기 정보를 이용하여 미지어 분석 등을 처리한다.

본 논문의 구성은 다음과 같다. 2장에서는 형태소 패턴에 대해서 구체적으로 정의하고 이를 형태소 분석에 어떻게 이용할 수 있는지 살펴본다. 3장에서는 형태소 패턴을 자동 생성하는 방법을 제시하고 4장에서는 형태소 패턴을 사용한 형태소 분석 시스템의 구현에 대해 논의한다. 5장에서는 실험을 통해 자동으로 생성된 형태소 패턴의 생성 비율과 형태소 분석기의 성능 평가를 하며, 마지막으로 6장에서 결론을 맺는다.

2. 형태소 패턴

본 논문에서 형태소 패턴은 어절과 그 어절의 형태소 분석 결과의 쌍으로 정의하였다. 활용이 일어나는 용언에 대해 공통적인 변형이 일어나는 부분을 형태소 패턴으로 등록할 수 있다. 예를 들어, 1) 동사 "돕다"에 대한 형태소 패턴은 다음과 같다.

<도운, 돕/Vv+ㄴ>
<도울, 돕/Vv+ㄹ>
<도움, 돕/Vv+ㅁ>
<도우, 돕/Vv+으>
<도와, 돕/Vv+어>
<도왔, 돕/Vv+었>

2) 형태소 패턴은 용언에 국한되지 않고 체언에 대해서도 확장할 수 있다. 사전에 명사 "도"와 "도와"가 등록되어 있다면 형태소 패턴 <도와, 돕/Vv+어>는 다음과 같이 변경된다.

<도와, 돕/Vv+어 도/Nc+와 도와/Nc>

3) 마찬가지로 "지지"와 "지지도"라는 명사가 사전에 등록되어 있다면 아래와 같은 형태소 패턴도 추가되어야 한다.

<지지도, 지지도/Nc 지지/Nc+도>

4) 형태소 패턴은 예외 처리를 위해서도 정의할 수 있다. 서론에서 언급한 흔하게 잘못 표기하는 단어인 "햇갈리다"를 올바로 분석하기 위해서는 다음과 같은 형태소 패턴이 등록되어야 한다.

<햇갈린, 헷갈리/Vv+ㄴ>
<햇갈릴, 헷갈리/Vv+ㄹ>
<햇갈림, 헷갈리/Vv+ㅁ>
<햇갈립, 헷갈리/Vv+ㅂ>
<햇갈려, 헷갈리/Vv+어>
<햇갈렸, 헷갈리/Vv+었>

마지막으로, 규칙으로 처리할 수 없거나 빈번히 한 단어처럼 쓰이는 어절을 분석하기 위해 형태소 패턴을 정의할 수 있다. 예를 들어 "써내려가"에 대한 어절을 제대로 분석하기 위해서 다음과 같은 형태소 패턴이 필요하다.

<써, 쓰/Vv+어>
<어내려가,
어/Ee+내려가/Vv 어/Ee+내려가/Vv+어>

본 논문에서 형태소 패턴의 자동 생성의 대상은 용언 혹은 체언과 관련하여 1), 2)와 같은 두 경우이며 3), 4)와 같은 두 경우는 해당 예제가 발견될 때에 수작업으로 구축한다.

3. 형태소 패턴의 자동 생성

형태소 패턴의 자동 생성을 위한 기본적인 아이디어는 체언 및 용언과 같은 어근과 조사 혹은 어미를 형태소 합성하여 단순히 두 문자열을 합친 것과 비교하여 변이가 발생하면 이를 형태소 패턴으로 등록하는 것이다. 예를 들어, 동사 “먹다”와 “돕다”의 어근과 어미 “어”를 합성한 아래의 두 경우를 살펴보자.

먹+어 → 먹어
돕+어 → 도와

“먹”+“어”의 합성 형태는 두 개를 합친 형태인 “먹어”와 동일하므로 형태소 패턴으로 등록되지 않으나 “돕”+“어”의 합성 형태는 두 개를 합친 형태인 “돕어”와 다르므로 형태소 패턴으로 등록한다. 이론적으로는 모든 어미에 대해 위와 같은 작업을 수행할 수 있으나 너무나 방대한 형태소 패턴이 생성되고 실제로도 모든 어미에 대해 수행할 필요가 없다. 형태소 패턴의 자동 생성을 위해 사용하는 어미 제약은 다음과 같다.

1) “어”, “어서”, “어서도”, “어야” 등 동일한 글자로 시작하는 어미는 동일 시작 글자(즉, “어”)로만 형태소 패턴 생성을 시도한다. 비슷한 예로 “으나”, ”으고“ 등에 대해서도 ”으“가 어미는 아니지만 ”으“에 대해서만 형태소 패턴 생성을 시도한다.
2) (어)서, (어)도와 같이 “어”가 생략된 어미와 “아”, “아서”, “아서도” 등 “아”로 시작하는 어미에 대해서는 형태소 패턴 생성을 시도하지 않는다.
3) ‘ㄴ/ㄹ/ㅁ/ㅂ/ㅆ’ 등의 어미(혹은 해당 글자로 시작하는 어미)에 대해서도 형태소 패턴을 생성하지 않는다. 대신에 “은/을/음/읍/었‘ 어미와 합성을 통해 형태소 패턴을 생성한다.

3)번의 경우 어근 “먹”과 어미 “ㄴ”을 합성하면 “먹은”이 되어 불필요한 형태소 패턴인 <먹은, 먹/Vv+ㄴ>을 생성하게 되어 이러한 어미를 배제한다.

용언과 어미의 합성을 통해 기본적인 형태소 패턴을 생성할 수 있다. 하지만 형태소 패턴 사전에 전적으로 의존하여 한국어 문장을 형태소 분석하기 위해서는 아래와 같은 추가적인 작업이 필요하다.

1) 형태소 패턴이 단어로 등록되어 있거나 체언+조사의 형태인 경우 해당 후보의 분석이 가능하도록 형태소 패턴의 결과를 추가한다.
2) ‘하다’류 동사에서 ‘하+어’가 ‘하여’로 합성되는 관계로 어절 ‘해’ 또한 ‘하/Vv+어/Ee’ 형태로 분석될 수 있도록 해당 형태소 패턴을 추가한다. 비슷한 예로는 ‘보+았’이 ‘보았’으로 합성되기 때문에 ‘봤’에 대한 형태소 패턴의 추가 등이 있다.
3) 조사의 단어로 시작하는 어미에 대해 합성 시 변형이 일어나지 않아도 형태소 패턴에 추가한다. 예를 들어, 어미 “도록”은 조사 “도”로 시작하므로 이에 대한 형태소 패턴의 생성을 수행한다.
4) 보조용언 ‘보다’, ‘가다’, ‘두다’ 등의 보조 용언에 대해 어근 혹은 활용된 형태가 체언의 종료 단어이거나 조사의 시작 단어인 경우 ‘보조어미+보조용언’ 형태의 분석을 위한 형태소 패턴의 생성을 수행한다.
5) 모든 조사와 어미에 대해 형태소 합성을 시도하고 합성된 형태가 형태소 사전에 등록되어 있으면 이에 대한 형태소 결과를 추가한다.

3)의 경우는 어절 “먹도록”을 단순히 좌우 최장일치법으로 분석하였을 때 앞의 어절 “먹도”가 ‘먹/Nn+도/Jc"로 분석되고 뒤의 어절 “록”이 미지어로 분석되는 경우를 방지하기 위해서 추가된다.

4)의 경우는 어절 “사가며”가 “사가”라는 명사가 등록되어 있어 단순히 ‘사가/Nc+이/Vc+며/Ee'로만 분석되는 것을 막기 위한 목적이다. 그러나, 이러한 형태는 형태소 후보의 과다 생성을 발생할 수 있으므로 제한된 용언과 어미에 대해서만 형태소 패턴의 생성을 수행하였다.

5)의 경우 “가서”처럼 명사 “가서”가 사전에 등록되어 있을 때 ‘가서/Nc'뿐만 아니라 ’가/Vv+어서/Ee'와 같은 후보도 생성하기 위해서이다. 이 밖에도 “아름다웠다/아름다왔다”와 같이 모음조화를 따르지 않는 ㅂ불규칙 용언을 처리하기 위한 형태소 패턴, “있다/없다”류의 동사가 어미 “되”와 결합할 때 “있되/없되”는 물론 “있으되/없으되”처럼 사용하는 경우를 처리하기 위한 형태소 패턴 등을 추가로 생성하였다.

[표 1] 형태소 패턴의 이진 구조
[Table 1] Binary structure of morphological patterns

결과 개수	형태소 결과$_1$								…	형태소 결과$_n$
	유형	형태소$_1$			…	형태소$_n$	[Tail]			
		단어	태그	속성			단어	속성		
1byte	1byte	C문자열	4byte	4byte			C문자열	4byte		

4. 형태소 분석 시스템

자동으로 생성된 형태소 패턴의 유효성을 검증하는 방법 중 직접적인 것은 형태소 패턴을 사용하는 한국어 형태소 분석기를 구현하는 것이다. 이 장에서는 형태소 패턴 기반 형태소 분석기의 구현에 대해 논의한다.

4.1 형태소 패턴의 이진 표현

<어절, 형태소결과>의 쌍으로 구성된 형태소 패턴에서 어절은 DB의 key로 형태소 결과는 DB의 정보로 저장된다. 어절은 단어 사전에 통합되어 저장되며 단어 사전은 음절 단위의 트라이[6]를 사용한다. 이는 과도한 형태소 패턴의 생성에도 검색 속도에 영향을 받지 않기 위해서이다. 형태소 정보는 이진 형태로 저장되는데 그 자료구조는 표 1과 같다

맨 처음 1바이트는 형태소 패턴에 포함된 형태소 결과의 개수가 저장되며 그 값이 0인지를 비교하여 해당 단어에 형태소 패턴이 존재하는지 여부를 판단한다. 형태소 결과는 n개의 형태소와 tail로 구성되는 데 형태소는 '단어/태그'로 표현되는 부분을 말하며 tail은 태그가 없는 단어를 의미한다. 형태소 결과는 적어도 1개 이상의 형태소를 가져야 하며 tail은 나타나지 않을 수도 있다. 예를 들어 형태소 패턴 <가서, 가/Vv+어서 가서/Nc>에서 첫 번째 형태소 결과 '가/Vv+어서'는 1개의 형태소와 tail로 구성되어 있고 두 번째 형태소 결과 '가서/Nc'는 tail이 없는 1개의 형태소로만 구성되어 있다. 형태소 결과 노드에서 맨 먼저 나오는 유형은 이러한 형태소 개수와 tail의 존재 여부를 알기 위한 값으로 유형의 값은 형태소 개수에 2를 곱한 값에다가 tail이 존재하면 1값을 더한다. 즉, 형태소 결과 유형이 홀수 혹은 짝수 여부를 보면 형태소 결과에 tail이 존재하는지를 알 수 있다.

형태소의 태그 노드는 해당 형태소 단어의 태그에 대한 bit mask 값을 저장한다. 마지막으로 형태소의 속성에는 그 단어의 형태론적 속성과 형태소 분석에 필요한 기타 속성이 저장된다. 형태론적 속성은 어미 혹은 조사의 결합 관계를 조사하기 위한 것으로 종성의 유무, 양성 혹은 음성 모음 여부 등이 저장된다. 형태소 분석에 필요한 속성은 주로 어미 혹은 조사의 부착 가능 여부를 조사하기 위한 속성들이 저장되어 있다.

4.2 형태소 분석기 구현

한국어 형태소 분석기의 성능 향상을 위해 많은 방법이 제안되었다[4, 5]. 여기서 구현된 형태소 분석기는 단순히 좌우 최장일치법을 사용하며, 미지어 및 복합명사를 올바로 분석하기 위해서 띄어쓰기 정보를 이용하여 후처리로 이를 처리한다. 한국어 형태소 분석기의 기본적인 동작 방식은 다음과 같다.

1) 문장을 읽어 들여 입력 순서대로 읽어 최대로 일치하는 후보를 찾는다.
2) 형태소 패턴의 존재 여부를 파악한다. 형태소 패턴이 있으면 그 정보를 이용하여 형태소 후보를 생성하고 존재하지 않으면 태그 정보를 이용하여 형태소 후보를 생성한다.
3) 종성의 유무 등이 저장된 단어 속성과 단어의 태그 정보를 이용하여 부착어(조사, 접미사, 어미 등)와 연결 가능한지를 검사하고 연결 가능하면 사전에서 최대로 일치하는 후보를 찾는다.
4) 유효한 부착어가 발견되면 형태소 패턴의 유무에 따라 이전 형태소 후보에 새로 생성된 부착 형태소를 연결한 후 3)과 4)를 실패할 때까지 계속 반복한다.
5) 후보 중에 최대로 일치된 것을 고르고 입력 문장에서 최대로 일치된 부분을 하나의 청크(chunk)로 해서 형태소 결과를 생성한다.
6) 띄어쓰기가 없는 여러 개의 청크에 대해 적절한 구문 패턴인지 유효성 검사를 한 후 실패하면 해당 청크를 합쳐 미지어로 추정한다.

아래의 예는 좌우 최장일치를 사용하여 긍정적 결과와 부정적 결과를 보여주고 이를 미지어 추정한 결과를 보여준다.

1) "한국어형태소분석은"
   - 한국어/Nc

[표 2] 한국어 형태소 패턴의 자동 생성 결과
[Table 2] Results of the automatic generation of Korean morphological patterns

항목	단어 엔트리	총 엔트리	형태소 패턴	형태소 결과
개수	124,582	286,450	171,101	184,991

형태소/Nc
분석/Nc+은/Jc
- 한국어형태소분석/Nz+은/Jc

2) "산타페는"
- 사/Vv+ㄴ_Ee 산/Nc 살/Vva+ㄴ_Ee
타/Ncb 타/Vv+어_Ee
페/Nz + 는/Jc
- 산타페/Nz+는/Jc

이러한 미지어 추정을 하기 전에 먼저 제한적이나마 유효한 구문 패턴인지를 검사한다. 이는 형태소 분석기의 결과를 수집하여 구문 패턴에 대한 학습용으로 사용하기 위해 띄어쓰기 오류가 있는 문장에서도 최대한 용언을 분석하기 위해서이다. 예를 들어, 어절 “먹는데에도”는 단순한 구문 패턴 규칙 <동사+관형어미+(제한적)불완전명사>을 만족하기 때문에 미지어로 추정되지 않고 다음과 같은 결과를 얻을 수 있다

먹/Vv+는/Ee+데/Nb+에도/Jc

현재 올바른 구문 패턴 규칙을 찾는 것은 매우 제한적이기 때문에 아직 띄어쓰기 오류가 발생하는 모든 문장에 대해서 제대로 분석할 수가 없다. 향후 구문 패턴의 학습을 통해 궁극적으로 음성 인식 등에 사용될 수 있도록 띄어쓰기 정보를 사용하지 않는 한국어 형태소 개발을 시도해 볼 것이다.

5. 실험 및 평가

[표 2]는 형태소 합성 및 세부 조정을 통해 한국어 형태소 패턴을 자동 생성한 결과를 보여준다. 원래 한국어 사전에 있던 단어 엔트리는 124,582개이었고 형태소 패턴 등록에 의해 확장된 총 엔트리 개수는 286,450개이다. 따라서 단어가 아니면서 형태소 결과만 갖는 형태소 패턴의 개수는 모두 161,868개가 된다. 생성된 형태소 패턴의 총 개수는 171,101개이고 형태소 결과는 184,991개로 하나의 형태소 패턴 당 1.08개의 형태소 결과를 갖는 것으로 파악되었다.

형태소 패턴의 검증을 위한 테스트 문장은 웹에서 수집한 신문 기사 중 5,000개의 문장을 대상으로 하였다. 테스트를 통해 5,000개 문장을 99.9% 이상의 정확도를 가지도록 분석하기 위해 수작업으로 추가한 형태소 패턴은 3,215개였다. 5,000개 문장 90,269 개의 문장을 분석하기 위해 사용된 형태소 패턴은 22,991개 정도였다. 따라서 사용된 형태소 패턴 중 약 86% 정도를 자동으로 생성할 수 있음을 알 수 있다. 수작업으로 추가한 형태소 패턴 중에 어미, 조사, 접미사 등이 차지하는 비율이 약 35% 정도이었고 분석하는 문장이 많아질수록 이들에 대한 증가율은 급격히 낮아질 거라 기대할 수 있으므로 실제 자동 생성율은 더 높아질 가능성이 있다.

형태소 패턴을 사용한 형태소 분석기의 분석 정확도와 분석 속도를 검증하기 위해 수집된 신문 기사 문장 중 다른 1,000개 문장을 대상으로 실험하였다. 실험 환경은 2.8GHz 듀얼 코어 CPU를 장착한 노트북에서 윈도우 상에서 리눅스와 유사한 개발 환경을 제공하는 Mingw 시스템에서 작업을 수행하였다. 오류 분석 시 아래와 같은 경우에는 오류로 포함하지 않았다.

1) 복합 명사를 통합하여 하나의 미지어로 추정한 경우
- 국회사무처:
국회사무처/Nz

2) 미지어 추정 시 유효한 미지어 분석 결과가 포함되어 있는 경우
- 빼빼로
빼빼로/Nz
빼빼/Nz+로/Jc

3) 띄어쓰기 오류가 있는 경우
- 장착하여월등한
장착하여월등/Nz+하/Xv+ㄴ/Ee

4) 철자가 틀린 경우
- 쫓겨서
쫓기/Vv+어서/Ee

4)번의 경우는 단어 ‘쫓기다’를 동사 ‘쫓기다’의 빈번한 오타로 보아 형태소 패턴 사전에 ‘쫓겨’에 대한 분석 결과를 등록한 이후에 생성한 결과이다.

올바로 분석된 어절의 기준은 형태소 규칙에 의해 분석 가능한 모든 형태소 분석 후보들만을 포함하는 것이다. 그러나 미지어에 대해서는 분석 후보 중에 올바르게

분석된 미지어가 포함된 경우 이를 올바로 분석된 어절로 포함하였다.

1,000개의 평가 문장, 어절 22,226개 중 숫자, 심벌, 문장 부호 등을 제외한 16,959개의 어절을 대상으로 정확도를 분석한 결과 99.5% 정확도를 얻었다. 따라서 자동 생성한 형태소 패턴은 한국어 형태소 분석에 유용하게 사용할 수 있음을 알 수 있었다.

오류의 유형 중 하나는 형태소 패턴의 미등재에 의한 오류로 테스트 문장 중에서 12개가 발견되었다. 이러한 오류로 인해 형태소 패턴 사전에 새로 등재된 엔트리는 다음과 같다.

1) <가본, 가본/Nc 가/Vvx+어/Ee+보/Vx+ㄴ>
2) <다고밖에, 다고/Ee+밖에/Jc>
3) <뭘까, 무엇/Np+이/Vc+ㄹ까>

대부분의 오류는 미지어를 제대로 추정하지 못하여 발생하는 오류이다. 이러한 오류의 예는 다음과 같다.

1) 바이오: 바/Ncb+이/Vc+오/Ee
2) 틴들: 틴/Nz+들/Xs
3) 설법인: 설법/Nc+이/Vc+ㄴ/Ee
4) 문예창작과: 문예창작/Nz+과/Jc
5) 그린체
   그린: 그리/Vv+ㄴ/Ee 그린_Nc
   체: 체/Ncb 체/Mi

5)번 오류는 앞서 언급한 바와 같이 “먹는것”, “잡을데가”와 같이 <용언+관형어미+(제한적)불완전명사>의 형태가 띄어쓰지 않은 형태로 비교적 자주 나타나 이를 처리하는 과정에서 발생하는 오류이다.

분석 시간의 계산은 문장 입력과 형태소 분석 결과를 저장하는 시간을 제외한 형태소 분석 연산 시간만을 계산한다. 웹에서 수집한 403BM 크기의 신문 기사 코퍼스를 모두 분석한 결과 형태소 분석에 걸린 시간은 52.68초로 나타났다. 이는 1GB 문서를 134초에 분석하는 것을 의미하므로 과도한 형태소 패턴의 생성이 한국어 형태소 분석기 분석 속도 성능에 영향을 미치지 않음을 알 수가 있었다.

6. 결론

본 논문에서 형태소 패턴을 형태소 합성 기법을 사용하여 자동으로 생성하는 방법을 제안했다. 실험을 통해, 분석에 필요한 대부분의 형태소 패턴을 생성할 수 있었음을 보였다. 또한, 형태소 패턴을 이용하여 유용한 한국어 형태소 분석기를 개발할 수 있음을 보였다.

앞으로 다양한 한국어 문장의 통계적 분석을 통하여 흔히 나타나는 철자 오류 단어나 띄어쓰기 오류 어절들을 위한 형태소 패턴을 자동으로 구축하는 방안을 연구해 볼 것이다.

References

[1] J.H. Kim, C.Y. Ok, "Korean Morphological Analysis using Inflected-Word-Dictionary", Proc. of KIISE Spring Conference, Vol. 21, No. 1, pp.813-816, 1994.

[2] S.H Yang, Y.S. Kim, “A High-Speed Korean Morphological Analysis Method based on Pre-Analyzed Partial Words“, Journal of KIISE:Software and Applications, Vol. 27, No. 3, pp.290-301, 20004.

[3] K.S Shim, J.H.Yang," High Speed Korean Morphological Analysis based on Adjacency Condition Check ", Journal of KIISE:Software and Applications, Vol. 31, No. 1, pp.89-99, 2004.

[4] Y.K. Kim, M.S. Park, J.S. Choi, H.C. Kwon, "Improvement of Analysis Speed in Korean Morphological Analyzer Using Ameliorated Dictionary", Proc. of the 11th Human & Cognitive Language Technology, pp.479-483, 1999.

[5] S.S. Kang, "Korean Morphological Analysis Using Syllable Information and Multiword Unit Information", Ph. D. Thesis, Seoul National University, 1993.

[6] C.S Kim, W.J Bae, Y.S. Lee, Junichi Aoe, "Construction of Korean Electronic Dictionary Using Double-array Trie Structure", Journal of KIISE(B), Vol. 23, No. 1, pp.85-94, 1996.

박 인 철(In-Cheol Park) **[정회원]**

- 1986년 2월 : 전북대학교 대학원 전산통계학과 (이학석사)
- 1998년 8월 : 전북대학교 대학원 전산통계학과 (이학박사)
- 1990년 3월 ~ 현재 : 호원대학교 컴퓨터게임학부 교수

<관심분야>
한국어정보처리, 정보검색, 임베디드시스템, 시멘틱웹