'어절 정보 사전'을 이용한 형태소 분석의 중의성 (Ambiguity) 해결*

남 지순, 최 기선
한국 과학 기술원 인공 지능 연구 센터 한글 공학 연구실

Desambiguation Method based on a Lexicon of Typographical Units

NAM Jee-Sun, CHOI Key-Sun
KAIST CAIR Language Engineering Laboratory
nam@world.kaist.ac.kr, kschoi@world.kaist.ac.kr

요 약

이글은 한국어 형태소 분석시 발생하는 중의성의 유형에 대해서 논의하고, 그와 같은 여러 유형의 중의성의 발생율을 감소시키기 위한 방법으로써 '어절 정보 사전 시스템'의 구축을 강조하였다. 한국어 문서에 대한 형태소 분석시 발생하는 중의성은, 영어나 유럽어와는 달리, 어휘 형성 정보 뿐아니라 어절 형성 정보, 구문 구조에 관한 부분적인 정보까지도 제공되어야 비로소 해소될 수 있는 경우가 많아 이와 같은 정보를 얻어내기 위해서는 체계적으로 고안된 범용의 사전 (Lexicon)이 필요하다. 여기에서는 접사가 동반되어 구성될 수 있는 '파생 명사 (Affixed Noun)'들의 경우에 논의의 범위를 제한하였다. 실제로, 체계적으로 구성된 하나의 파생어 사전은, 주어진 어절에 대한 형태소 분절시 발생할 수 있는 엄청난 수의 중의적 가능성을 해소해 줄 수 있는데, 이와 같은 사전을 구축하기 위해서는 단순어와 접사 사전이 모듈화되어 완성되어야 한다. 같은 방법으로 모든 합성어 유형에 대한 사전이 구축되고, 그러한 기본 형태들에 대한 '변화형' 사전이 결합되면 어절 정보를 갖춘 대용량의 한국어 MRD의 구현이 가능해질 것이다.

1. 머릿말

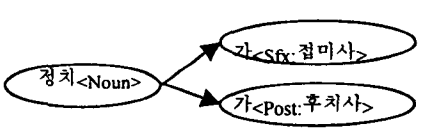
형태소 분석 (Morphological Analysis) 은 자연어 처리 시스템에 있어서 가장 우선적으로 수행되어야 하는 기본 단계로써, 구문 분석 및 의미 분석 등의 작업에 선행되는 필수적인 작업이다. 또한 정보 검색이나 기계 번역 시스템, 맞춤법 검사기 등의 응용 분야에서도 반드시 거쳐야 하는 첫번째 단계이다. 형태소 분석에 대한 연구는 이미 오랜 기간동안 진행되어 왔고, 현재는 거의 모든 시스템이 99 % 이상을 성공하고 있다는 발표 사례들이 보인다.

형태소 분석에 있어서 실제로 중요한 문제는, 하나의 어절 단위에 대한 형태소 분석의 가능성이 다양하게 제시될 때, 즉 이와 같은 '중의성 (Ambiguity)'의 문제를 어떻게 해결할 수 있느냐 하는 점이다.

1.1. 구문적 중의성 (Syntactic Ambiguity)

예를 들어, '정치가'와 같은 어절이 입력되었을 때, 이것은 '정치를 하는 사람'의 의미와 '정치라는 명사가 문장의 주어로 사용'된 의미의 두 가지 분석이 가능하다. 이 경우, 형태소 해석 시스템은 다음과 같이 '중의성'을 갖는 분석 결과를 제시하게 된다.

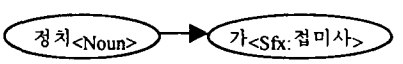
(1)



그런데, 우리가 중의성이라고 부르는 현상에는 여러 유형이 있다. (1)에서 나타난 중의성은, 어절 수준에서 문맥을 고려하지 않고 분석할 때에는 해결할 수 없는 유형이다. 바로, 구문 구조 및 문맥을 고려해야만 제거할 수 있는 '구문적 중의성 (Syntactic Ambiguity)'의 경우이다. 즉, '정치가'라는 스트링은 다음에서와 같은 구체적인 문맥속에 실현될 때, 비로소 이러한 중의성이 해소될 수 있다.

(2a) 한국 정치가 모임이 새로 구성된다
(2b) 한국 정치가 변하고 있다

즉, 위에서 (2a)는 다음 (3a)로, (2b)는 다음 (3b)로 각각 분석된다.

(3a) 

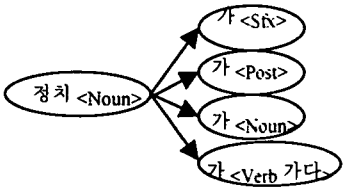
(3b) 

* 본 논문은 과학기술처의 지원을 받아 수행된 '통합국어정보베이스'의 일환으로 이루어졌다.

1.2. 품사적 중의성 (POS Ambiguity)

위의 (1)과 같은 결과는 형태소 분석 단계에서 이미 오분석의 가능성들이 제거된 상태이다. 실제로 형태소 분석기는, 접미사, 후치사로서의 '가' 뿐아니라 다음과 같이 '명사', 동사'로써의 '가'의 분석 가능성도 출력할 것이다.

(3)



이때, '가'에 대한 위와 같은 4 가지 분석의 가능성은 하나의 형태소가 갖는 '품사적 중의성 (POS Ambiguity)' 의 유형이다. 그런데, 이와 같은 결과가 갖고 있는 중의성을 해소하려면, '어휘 형성'에 대한 정보 뿐 아니라 '어절 형성'에 대한 정보. 더 나아가 부분적인 '구문 구조'에 대한 정보등이 제공되어야 한다. 그러면, 위의 4 가지 분석 가능성에 대해 살펴보자.

1. 위의 '가'가 접미사 <Sfx>로 분석될 수 있는가의 여부는, 접미사 '가'가 어떤 명사와 함께 실현될 수 있는가 하는 '어휘 형성' 정보가 있어야 결정된다. 가령,

철학가, 법률가

등은 접미사 '가'를 포함한 파생어 형태인데, 이와 같은 파생어 사전이 체계적으로 완성되지 않는 한,

그가 장가 가서, ...

에서 나타난 단순명사 '장가'의 일부인 '가'를 '장<명사>+가<접미사>'로 분석하는 오류를 막을 수가 없다.

2. 위의 '가'가 후치사 <Post>로 분석될 수 있는지는, 후치사 '가'가 어느 명사 유형과 결합 할 수 있는지, 즉 '어절 형성'에 대한 정보가 있어야 알 수 있다. 후치사 '가'는 종성으로 끝나는 명사뒤에는 실현되지 못한다는 정보를 이용하면, 다음 (4)와 같은 어절에서 '가'는 <Post>의 분석을 허용할 수 없음을 알 수 있다. 즉, 후치사 <Post>는 다음 (5)에서처럼 '이'의 형태로 실현되기 때문이다.

(4) 철학가, 법률가
(5) 철학이, 법률이

'정치'는 종성으로 끝나는 명사가 아니므로 후치사 '가'의 출현이 가능하고 따라서 <Post>의 분석이 가능한 것이다.

3. 위의 '가'가 명사 <Noun>으로 분석될 수 있는지는, '복합 명사 구성'에 대한 정보가 제공되어야 결정된다. 예를 들어,

수우미양가 모든 점수를 ...

와같이 '가'가 복합 명사의 형태속에 실현되었는지 아닌지의 여부는 'Noun-가' 유형의 모든 복합 명사의 목록이 제공되어야 결정할 수 있다. 한국어에서 복합 명사는, 다음에서 보듯이 응집된 형태로의 실현이 매우 빈번하기 때문이다.

나무토막, 나무결, 유리병, 쇠조각,

명사로서의 '가'가 어떤 형태와 결합하여 복합 명사를 이루는지 그 어휘 구성 정보가 없는 한 <Noun>의 분석을 제거시킬 수 없다.

4. 위의 '가'가 동사 <가다>의 어간 형태로 사용될 수 있는지는 '복합 동사 구성'에 대한 정보 및 '동사구의 목적어-술어 관계'에 대한 정보가 주어져야 해결된다. 가령, 다음과 같이 실현된 '내려가' 라는 어절을 보자.

그가 내려가 보니 아무도 없었다

이것은 '내려가다' 라는 복합 동사의 부사형 변화 형태이다. '가'앞에 어떤 형태가 나타날 때 그 앞 성분이 동사가 아닐지, 그래서 복합 동사의 일부를 이루는 '가다'의 어간형으로 분석될 수 있을지 '복합 동사 구성'에 대한 정보가 주어져야 한다. 또한,

어서 학교가 !

에서처럼, 동사 '가다'앞에 보어 (Complement)로 쓰인 명사가 결합되어 나타나는 경우가 있다. 이 경우 다음에서 보듯이, 대개 보어의 후치사가 생략되면서 그 보어를 취하는 술어와 결합한 형태가 나타나는데,

밥먹고 일찍 출발한 그 사람은, ...
말많은 그 여자가 왔다
춤추고 놀기에만 급급한 그 아이

이와 같은 분석의 가능성을 추정하기 위해서는 동사로 사용되는 '가다'의 보어 성분으로 어떤 명사들이 실현될 수 있는지에 대한 정보가 필요하다. 위에서, 동사 '가다'가 어떤 성분과 결합하여 복합 동사를 형성할 수 있는지, 또 어떤 성분을 보어로 취할 수 있는지에 대한 정보가 없이는 <Verb>로의 분석 가능성을 배제할 수 없다.

이상에서는 형태소 '가'의 여러 의미중 위의 4가지 품사의 경우만을 검토하였다. 위에서 살핀 바와 같이, 한국어에서 '품사적 중의성'을 해소하는 단계는 '어절 형성' 정보나 '구문 구조' 정보등이 부분적으로 반드시 필요하게 되므로, 단순히 '어휘 형성' 정보만으로 형태소 (즉 단어:Word)를 분석해낼 수 있는 유럽어나 영어의 경우와는 다르다.

1.3. 어휘적 중의성 (Lexical Ambiguity)

여기서, 위의 4 가지 품사 분석의 가능성을 보이는 '가'의 각 품사 유형을 다시 하위 분류하게 되면, '가'를 내포한 어절의 중의성은 더욱 증가하게 된다. 만일 다음과 같이, 접미사로서의 '가'를 8 가지로 분류한다면, 위의 형태소 분석 결과는 다음 <그림1>과 같이 11 가지로 늘어날 것이다.

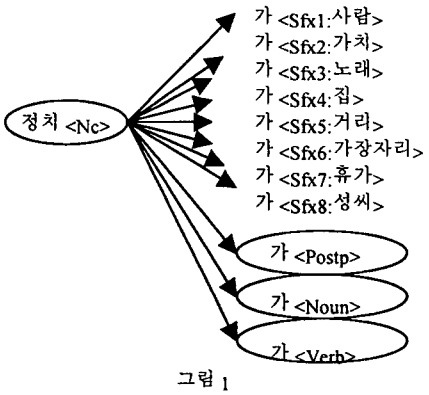


그림 1

같은 방법으로 각 품사 분류를 세분화한다면, 형태소 분석 결과는 엄청나게 증가할 것이다. 이러한 결과는 바로 한 동일 품사어의 '어휘적 중의성 (Lexical Ambiguity)'에서 기인한 것인데, 이것은 각 단어의 의미적, 어휘적 쓰임을 체계적으로 기술하지 않으면 해결하기가 어렵다. 형태소 분석 단계에서 현재는 이 부분에 대한 연구가 충분히 진전되지 못한 상태이고, 일단은 품사 분류 정도에 그치고 있으나, 다음 예에서 보듯이 접미사 '가'를 한 가지로 처리해서는 많은 문제가 발생할 것을 예상할 수 있다.

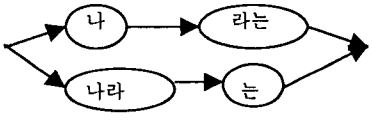
그는 사업가이다
이 빵은 영양가가 높다
그들은 응원가를 열심히 불렀다
그는 처가에 가기를 싫어한다
친구들이 대학가에 모였다
아이들이 우물가에 모여 앉았다
그는 병가를 냈다
김가, 이가, 박가등이 모였다

2. 어절의 분절 중의성 (Segmentation Ambiguity)

형태소 분석의 기본 단위가 되는 문자 스트링의 유형이 '명사_후치사'나 '명사_동사_어미'처럼 하나의 '어절'이라는 사실이, 앞서 살핀 중의성의 여러 유형에 덧붙여져서, 한국어 형태소 분석 모듈의 부담을 더욱 가중시킨다. 이것은 어절을 구성하고 있는 단어들 사이의 조합 가능성을 일일이 계산해야 하기 때문이다. 즉, 한국어 형태소 분석시에는 어절의 '분절 중의성 (Segmentation Ambiguity)'의 부담이 추가된다. 간단한 예로,

나라는

과 같은 어절은 다음과 같은 두 가지 유형의 '분절' 가능성을 갖는다.



이러한 이유로, 한국어 형태소 분석시 나타날 수 있는 '중의성'의 정도는 유럽어나 영어의 경우보다 훨씬 심각하다. <그림2>는 지금까지 논의한 한국어 형태소 분석시 발생할 수 있는 '중의성'의 유형을 영어의 경우와 비교하고 있다.

	한국어	영어
구문적 중의성	구문 구조 정보	구문 구조 정보
분절 중의성	구문 구조 정보 어절 형성 정보 어휘 형성 정보	
품사적 중의성	구문 구조 정보 어절 형성 정보 어휘 형성 정보	어휘 형성 정보
어휘적 중의성	어휘 형성 정보	어휘 형성 정보

그림 2

영어의 경우, 형태소 해석상의 중의성은 대체로 단어 자체가 갖고 있는 '어휘적 중의성' 및 '품사적 중의성'에 기인하지만 (예를 들어, 'fly'가 '파리'의 의미이거나 '바지 지퍼'의 의미를 나타내는 점, 또는 'sleep'이 Noun이거나 Verb인 점), 한국어의 경우 어절이라는 복합체가 갖게 되는 조합 가능성의 복잡도가 거기에 다시 추가된다. 가령 *abcd* 라는 어절이 존재한다면, 이때 이 어절을 이루고 있을 형태소들의 조합 가능성을 조사하기 위하여 사전 검색 모듈은 <그림3>와 같이 8 가지의 경로를 거쳐 매칭된 결과를 보인다.

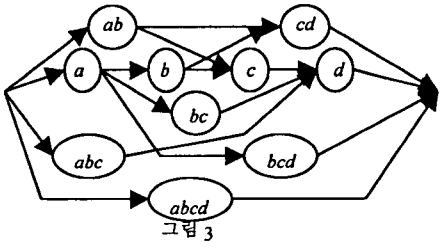


그림 3

만일 다음과 같은 스트링이 입력되었을 때,

새우편물에는

이 어절은 '품사적 중의성' 및 '어휘적 중의성'을 고려하지 않았을 경우, <그림4>과 같이 15 가지 유형으로 분할된다.

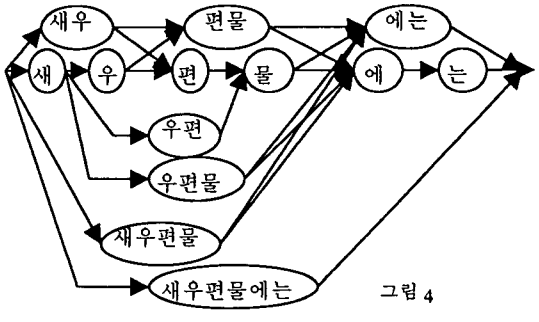


그림 4

이때, 각 단어들이 가질 수 있는 '품사적', '어휘적' 중의성까지 고려하면, 이 어절의 분석 가능성은

폭발적으로 증가하게 된다. 예를 들어 '새'의 경우를 보면, 다음 <그림5>와 같이 5가지의 품사 유형 사이의 중의성이 나타나며, 또한 그 어휘적 의미 차이까지 고려한다면, 명사의 경우 2 가지, 동사의 경우 4 가지 정도의 유형으로 더 세분화되어 모두 9 가지의 중의성을 갖게 된다.

명사	1	사이	새가 벌어졌다
	2	날짐승	새가 난다
관형사		새로운	새가방
접두사		강조	새하얀 눈송이
접미사		양태	모양새
동사	1	밝아오다	날이 새니 ...
	2	빠져나오다	불빛이 새어 나온다
	3	새우다	밤을 꼬박 새고 나니..
	4	누설되다	비밀이 새어 나갔다

그림 5

이와 같은 방법으로 모든 단어 후보들의 '품사적, 어휘적 중의성'을 고려하면, 다음 <그림6>와 같다.

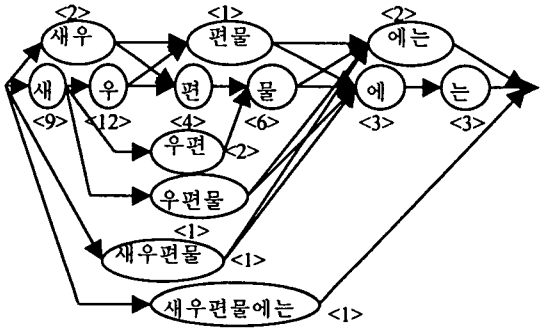


그림 6

위의 <그림6>의 전이그래프는 2 597 629 개의 분석 가능성을 보여준다.

그러면, 이와 같이 폭발적으로 나타나는 형태소 과분석의 오류를 어떻게 감소시킬 수 있는가 하는 것이 문제이다. 이러한 문제의 해결을 위해 제시되고 있는 몇 가지 모델들을 검토해 보면 다음과 같다.

3. 형태소 분석의 중의성 해결 모델

3.1. 의미망 (Semantic Network) 의 이용

첫째는, 의미망을 이용해서 관련 명사류를 구축하여 그 후보 결정을 돕는 방법이다. 예를 들어, 다음 어절들을 해석할 때 그 단어들 사이의 '관련도'를 조사하면,

(1) 새우편물
(2) 새우튀김

(1)이 '새우(바닷가재의 일종) - 편물(수예품)'과 같이 분석될 가능성은 (2)가 '새우(바닷가재의 일종) - 튀김(요리)'으로 분석될 가능성에 비해 낮을 것으로 추정될 수 있다. 그러나, 이와 같은 의미적 연관 관계를 객관적으로 구축한다는 것은 실제로 매우 어렵고, 따라서, 다음과 같은 어절들의 차이를 미리 예측하기는 어렵다.

(3) 새우등
(4) 새우눈
(5) 새우잠

(3)은 위의 (2)와 같은 맥락에서 분석되는 것이 타당하고 (즉, 바닷가재 '새우'의 '등' 부분), (4)는 (3)와 같이 '새우의 신체 일부로서의 '눈'을 나타낼 수도 있지만, '새우같이 가는 눈'을 가르키는 하나의 굳어진 표현일 수도 있다. (5)는 어원적으로 (3), (4)에서 사용된 '새우'의 의미를 내포하고 있겠지만, 실제 사용되는 의미는 '바닷가재류'와는 관계가 멀다. 이와 같은 어절들은 그것을 구성하는 어휘 자질 (lexical feature) 들을 개별적으로 검토하지 않는 한, 결코 의미망에 대한 일반적인 분류만으로 올바르게 예측할 수 없다.

3.2. 통계적 확률 계산

둘째는, 대형의 코퍼스를 이용하여 얻어진 어절들로부터 통계적인 방법을 이용하여 일정 확률값을 계산하는 방법이다. 예를 들어, '새우'라는 단어를 포함한 어절의 인접 문맥속에서 '바다'라는 단어가 나타나는 확률이 가장 높을 때, '새우'가 들어있는 어절을 모두 '바닷가재의 일종'으로서의 '새우'라는 단어가 내포되어 있는 것으로 우선 추정하는 방법이다. 즉, 이와 같이 '공기 (collocation)' 제약 및 좌우 접속 정보등을 사용하는 것은 코퍼스의 크기나 그 질에 많이 좌우되며, 어느 정도의 효율적인 추정을 돕기는 해도, 어느 단계 이상에서부터는 추가되는 자료의 양에 비해 만족할만한 향상성을 보이지 못한다. 이때, 코퍼스에서 수집된 자료를 토대로 '학습 (learning)'을 시키는 방법이 연구되고 있으나, 확률값을 자동으로 향상시키는 모델에 기초한 이상 위에서 제시된 예들과 같은 문제는 여전히 처리 하기가 어렵다.

3.3. 최장 일치,최단 일치 방법론

세째는, 여러개의 형태소 분석 가능성이 제시될 때, '최장 일치' 또는 '최단 일치'와 같은 방법을 사용하는 것이다. 위의 <그림5>와 같은 어절을 분석할 때, '최장 일치'의 방법을 취하면, '새우편물에는'을 하나의 기본 단위로, '최단 일치'의 방법을 취하면, '새', '우', '편', '물', '에', '는' 의 여섯가지 형태소로 분석 결과를 제시할 것이다. 그러나, '새우편물에는'이 사전에 등재된 하나의 단위가 아니기 때문에 다시 분절이 필요하게 되고, 이때, 그 분절 방향이 여러가지로 (좌에서 우로, 우에서 좌로, 또는 양방향에서) 제시된다. 여기서 후치사를 올바르게 떼어내기만 한다면, 부담이 큰 '명사' 사전이 따로 필요없이 '명사'를 바로 인식할 수 있다는 가정이 가능한데, 그러나 이 경우, 명사의 일부, 또는 접미사를 후치사로 잘못 인식하는 오류를 피할 수 없다. 예를 들어 다음에서,

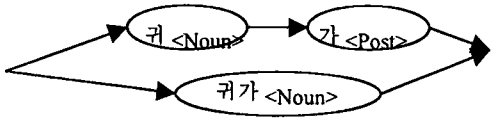
난장이 삼 형제가 모두 왔다
교장이 삼 형제를 모두 불렀다

'이'를 후치사로 가정하게 되면, '난장'을 '교장'과 같이 하나의 명사로 잘못 판정하는 결과를 가져올 것이다.

비록 명사 사전을 토대로 분석이 이루어진다 하더라도, 명사의 끝부분이 '후치사'와 동일 형태를 취할 때, 올바른 후보를 제시하지 못할 위험을 갖는다. 예를 들어,

귀가 아프다
귀가 시간이 당겨졌다

와 같은 경우, 최장 일치와 같은 방법으로는, 구문을 고려해야 비로소 선택할 수 있는 다음의 두 가지 분석의 가능성을 제시하지 못한다.



3.4. 사전의 간소화와 알고리즘의 개발

네째는, 품사 분류의 체계를 최소화하고 사전을 간소화한 후, 어절 분리 알고리즘 및 미등록어 처리등의 모듈을 강화해서 형태소 분석의 효율성을 높이는 방법이다. 실제로, 형태소 품사 분류에 있어서, 그 품사적, 어휘적 자질을 얼마만큼 자세하게 세분화 하느냐에 따라 형태소 분석의 중의성의 복잡도가 좌우된다. 사전에 그 품사별, 의미별 분류를 자세히 추가하면 할수록 형태소 분석시에 발생할 수 있는 '중의성'의 수는 폭발적으로 늘어나게 될 것이며, 반대로 분류 체계를 최소화하게 되면, 형태소 분석의 '중의성' 문제는 자연히 해소되는 듯이 보인다. 따라서, 형태소 분석시에 발생하는 이와 같은 '중의성'의 문제를 해결하기 위하여 단순히 그 품사나 의미 분류 등을 최소화시킨다는 것은 아무 의미가 없다. 이것은, 근본적으로 형태소 분석이 자연어 처리에서 왜 필요한 것인가하는 기본 문제를 잊고 있는 것과 같다.

앞서도 논의하였듯이, 하나의 단어는, 다음 문장들에서 나타난 '물'과 같이, 단일 품사로서 그 의미가 여러 가지일 수 있고,

(1) 아이가 물을 마신다
(2) 옷에 파란 물이 들었다

품사 자체가 여러 가지일 수도 있다. 다음에서 나타난 '물'은,

(3) 물이 차갑다
(4) 개가 아이를 물었다

(3)의 경우 명사로 쓰였고, (4)의 '경우 동사 '물다'의 어간 형태이다. 대개의 형태소 분석용 사전에서는 위의 (1)과 (2)와 같이 동일 품사로서 그 '의미'가 달라지는 것은 일단 고려하지 않고 (3)과 (4)에서처럼 '품사'가 달라질 때에만 분류를 표시하는 방법을 취하고 있다. 따라서 이때, 같은 동사라 하더라도, 다음 (5)에서와 같이 (4)의 '물다'와는 다른 의미의 동사 '물다'가 존재하면,

(5) 그는 벌금을 물었다

(1)-(2)의 두 가지 의미의 명사가 2 개의 구분된 코드값으로 구별되지 않은 것처럼, 마찬가지로 (4)-(5) 사이의 두 가지 동사의 의미는 형태소 분석 단계에서는 나타나지 않는 경우가 많다.

그런데, 만일 다음 예에서 나타난 '물'과 같이

(6) 그 여자가 내게 길을 물었다

기본 어형이 위의 (4), (5)의 동사들과는 달리 동사 '묻다'인 경우, 형태소 분석시 형태 변화의 정보 (Morphological Information) 를 고려할 수 있도록 (6)에서와 같은 동사는 따로 덧붙여 진다. 즉, '물'에 대한 형태소 분석 결과는 다음과 같이 나타날 것이다.

(7)

물	*Noun*
	Verb <물다>
	Verb <묻다>

여기서 이와 같은 분류가 얼마나 자의적 (Arbitrary) 이며 비체계적인 것인가 하는 점은 다음에서 쉽게 확인된다.

(8) 그 여자는 항아리를 땅에 묻었다
(9) 잉크가 옷에 묻었다

위의 (6)과 (8)/(9)를 비교해 보면, '묻다'라는 동사가 (6)에서는 '물었다'의 형태로 변화하였고, (8)/(9)에서는 '묻었다'의 형태로 변화하였다. 따라서, (7)과 같은 형태소 분석 결과에서 나타날 수 있는 'Verb <묻다>'는, (8)/(9)에서 쓰인 동사 '묻다'가 아니라 (6)에서 쓰인 의미의 동사 '묻다'인 것이다. 이때, 동사의 그 의미 차이에 따른 분류를 무시하게 되면 이와 같은 형태적 정보를 등재할 방법이 없다. '물'의 이와 같은 의미 특성까지 고려한 형태소 분석 결과는 다음과 같을 것인데,

(10)

물	*Noun* <물1>	(물을 마시다)
	Noun <물2>	(물이 빠지다)
	Verb <물다1>	(세금을 물다)
	Verb <물다2>	(개가 물다)
	Verb <묻다1>	(길을 묻다)

이것은 당연히 위의 (7)의 결과보다 더 복잡하고 많은 중의성을 보이게 된다. 이와 같은 분석이 그 다음 단계인 '구문 분석 (syntactic analysis, parsing)'이나 다른 응용 분야에서 반드시 필요하게 되리라는 점은 다시 강조할 필요가 없다.

이와 같은 용도를 고려하지 않고 고안된 형태소 분석기는, 그 중의성 발생율이 매우 낮고, 또 간소화된 품사 세트 (예를 들어, '명사', '동사', '형용사', '기타'등으로 분리된 체계) 를 거의 100 % 찾아 맞춘다고 하여도 그 자체로는 별다른 의미가 없다. 오히려, 체계적으로 기술된 형태소 정보는, 그 결과를 응용 분야들에 맞게 여러 옵션의 형태로 더 간소화하여 출력하는 데 아무 어려움이 없으므로, 위와 같이 품사 분류와 사전을 무조건 단순화하는 것은 그 다음 단계의 연구에 지장을 가져온다.

기존의 많은 형태소 분석기는, 사전 모듈보다는 알고리즘의 성능을 높이는 부분에 더 주력하였다. 이와 같은 접근의 필요성을 뒷받침해 주는 현상중의 하나가 '고유 명사 (Proper noun)' 나 '복합 명사 (Compound noun)' 등과 같은 소위 '미등록어' 처리의 문제이다. 일반 '보통 명사 (Common noun)' 들과는 달리, '고유 명사'나 '복합 명사' 사전은 쉽게 체계적으로 구축될 수 있는 부분이 아니라서 사전에 전적으로 의존하는 시스템에서는 빠른 성과를 기대하기가 어렵다. 그러나, '미등록어'의 처리는, 그것의 유형을 추정하는 데에는 한계가 있어서, 후치사 앞에서 발견된 경우, 명사의 일종으로 간주할 수 있는 정도이다. 이것이 '인명 명사'인지, '복합 명사'인지, 또는 '파생 명사'인지의 구별은 이 자체로는 불가능하다. 예를 들어,

(11) 새라토가에는
(12) 새우튀김에는
(13) 새우편물에는

에서 '에는'의 앞에 나타난 성분들이 사전에 등재되어 있지 않은 경우, 모두 미등록어 추정 알고리즘에 의하여 '명사'로 분석될 수 있다. (11)에서 나타난 '새라토가 (Saratoga)'는 미국의 한 도시 이름으로 고유 명사 사전에 등재되어야 하며, (12)와 (13)의 경우에는 앞서 본 바와 같이 각각 'Noun-Noun'구조의 복합 명사와, 'Prefix-Noun-Suffix'구조의 파생 명사로 등재되어야 한다. 더우기, 후치사의 생략이 아주 빈번한 한국어의 경우, 후치사가 생략되면 미등록어 처리 모듈은 더욱 복잡해져서, (11)에서 '에는'이 생략되어 실현되면, '가'를 조사로 추정한 '새라토'라는 미등록어가 추가되는 오류가 덧붙여져 질 것이다.

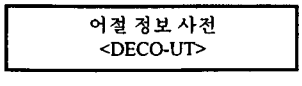
그동안, 사전 및 언어 정보를 보다 체계적으로 구축하려는 노력보다는 효율적인 알고리즘의 개발을 통해 형태소 분석의 성공율을 높히려는 노력이 더 집중되었는데, 실제로 알고리즘의 개발을 통해 기대할 수 있는 효율성의 증가폭은 이제는 그렇게 크지 못하다. 얼마만큼 잘 구축된 사전 정보를 이용하였는가 하는 점이 앞으로의 중요한 관건으로 생각된다.

3.5. 어절 정보 사전을 이용한 형태소 분석기

다섯째로, 형태소 분석시 과분석의 오류율을 저하시키기 위하여, 이글에서 강조하고자 하는 방법은, 위에서 지금까지 논의하여 온 바와 같이, 바로 '사전을 기초로 한 형태소 분석기 (Dictionary-Based Analyseur)'의 구현이다. 여기서 말하는 사전 시스템은 '어절단위의 정보'를 갖춘 형태를 일컫는 것으로, 이것은 다음에서 보이는 바와 같이 여러 개의 하위 사전들이 모듈화되어 단계적으로 구축되어 진다.

4. 한국어 어절 정보 사전 시스템 *DECO*

다음 <그림7>은 '한국어 어절 정보 사전 시스템 DECO/V01'의 기본 구조를 보여 준다.



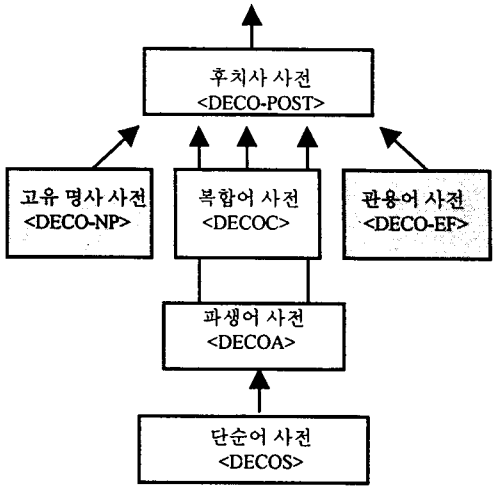


그림 7

시스템 *DECO*/V01의 세부적인 구조 및 그에 대한 소개는 다음으로 미루기로 하고, 여기서는 파생어 사전 <DECOA> 모듈에서 특히 '파생 명사'의 목록을 구축하는 단계에 국한하여 살펴보기로 한다.

4.1. 파생어 사전 <DECOA/V01>

위 시스템에서 가장 기초를 이루는 부분은 '단순어 사전 <DECOS>'으로, 이 사전의 엔트리는 '단 하나의 독립 성분 (즉 단어)'만으로 구성된다. 가령, 기존의 대사전이나 대형 코퍼스에서 추출된 명사류로부터 '접두사'나 '접미사'등을 떼어내고 그 기본 단위를 추출하는 과정을 거쳐야 한다. 따라서 '단순 명사' 사전과 '접두사'와 '접미사' 사전을 각각 체계적으로 완성해야 한다.

현재 '단순 명사 사전 <DECOS/V01>'은 15 000 개의 엔트리를 가지고 있고, 이것은 모든 파생어, 복합어, 고유 명사, 고어 등을 별도로 처리한 것이므로, 현재 사용되고 있는 '보통 명사'로서의 단일어의 성격을 가진 것들은 빠짐없이 수록되어 있다. '접사 사전 <DECO-AFX/V01>'은 접두사, 접미사 각각 900 여개씩을 함유하고 있는데, 이것은 어원적, 언어 이론적 입장에서 정의하고 있는 '접사'의 개념과는 그대로 일치하지 않음을 밝혀 둔다. 즉, 문장안에서 독립적으로 하나의 스트링을 이루지 못하며 반드시 하나의 '명사' 성분에 결합되어 실현되어야 하는 모든 형태는 일단 '접사'로 분류되었는데, 이 때, 이 '명사'와 '접사' 사전 사이의 조합 가능성을 개별 어휘에 대해 언어학적으로 검증한 결과가 바로 '파생어 사전 <DECOA/V01>'을 구성하게 된다. 이와 같은 모듈화의 필요성은 다음과 같은 이유에서 비롯된다.

4.1.1. 체계적 파생어 사전의 완성

접두사, 접미사류 중에는 생산성이 대단히 높은 것들이 상당수에 이른다. 예를 들어, 접두사 '여'와 같은 경우,

여기자, 여학생, 여선생, 여가수, ...

등에서 예상할 수 있듯이, 그 목록의 크기가 엄청나므로 기존 대사전등에 일일이 모두 수록되어 있지 않고 몇개의 경우들로 대표되어 있다. 대형 코퍼스를 검사하여도 모든 가능한 형태가 다 찾아지지는 않는다. 이와 같은 '파생어 사전'을 체계적으로 구축하기 위해서는 기본 단어 사전이 우선 구성되어야 하며, '접사 사전'이 별도로 구성되어, 개별 어휘 요소에 대한 '언어학적 조합 (Linguistical Combination)'이 이루어져야 한다.

이와 같은 단계를 거치지 않고 사전을 이용하고 있는 형태소 해석기의 경우, 위와 같이 접미사 '여'를 가진 명사가 다 수록되어 있지 않을때, 빠진 형태가 어떤 것인지 추정이 불가능하며, 따라서 체계적인 사전 확장이 어렵다. 모든 '여-Noun'의 파생 명사의 목록이 구축된다면, 더이상 접두사나 단순어라는 품사의 설정이 불필요할 수도 있다. 그러나, 파생 명사의 목록이 체계적으로 완성될 때까지는 이와 같은 단계가 반드시 필요하다.

접두사 '여'가 붙어 만들어질 수 있는 명사를 추정하기 위해서 '규칙 (Rule)'을 사용할 수도 있다. 가령, 이 접미사가 붙을 수 있는 명사는 우선 '인물성 명사' 이며, 직업을 나타내는 명사 (예를 들어, '여기자'), 또는 가족 관계등을 나타내는 명사 (예를 들어 '여동생') 일 가능성이 높다. 그런데, 다음을 보면,

(1) *여노동자, *여정치가
(2) *여언니, *여삼촌

'직업'이나 '가족 관계'의 의미를 갖는 명사들이 모두 결합 가능한 것은 아니다. 즉, (1)에서 나타난 직업 명사들은, 접두사 '여' 대신 '여성', '여자', '여류'등의 형태와 결합이 가능하다. 이와 같이 파생어의 형성은 '어휘적 자질'에 의해 결정되는 것이지, '의미적 자질'에 의존하는 것이 아니기 때문에 어휘 요소에 대한 개별적인 검증이 필요하며, 규칙으로 추정될 수 없다.

단순어 사전을 따로 분리해 내지 않고, 그렇다고 의미적 자질을 기초로 '규칙'을 가정하지도 않는 형태소 해석기의 경우, '여-Noun' 형태의 파생어의 일부만이 수록되어 있는 사전을 토대로 형태소 분석을 해야 하므로, 어절 분석 단계에서 '여'로 시작하는 단어가 실현되면, 이 부분을 일단 접사로 가정하는 절차를 거치게 된다. 예를 들어,

여가, 여독, 여물, 여우, 여행, ...

과 같이 '여'뒤의 성분이 하나의 '명사'로 인식될 수 있는 가능성이 사전에 있으면 모두 '여-Noun'의 분석 가능성을 추가로 갖게 된다. 한국어 단순 명사의 대다수가 한자어 2 음절로 구성되어 있고, 그 첫째 음절은 상당수가 단음절형 접두사와 동일 형태를 취하고 있으므로 이와 같은 분석 가능성의 과정을 고려하여야 한다면, 형태소 해석 시스템의 부담은 엄청나게 가중될 것이다.

여기서 이와 같은 부담을 덜기 위해, 접두사 '여'의 존재를 따로 고려하지 않는다면, 다음 (3)과 같은 파생어들이 사전에서 누락된 경우, 주어진 어절속에서 다른 성분으로부터 올바르게 분절이 되어도, 모두 미등록어 처리를 통하여 (4)의 복합 명사, 고유 명사들과 다름없이 동일 '명사 코드 값'을 임시로 부여받게 될 것이다.

(3)	*NewN*	여가수, 여판사, 여사장, ...
(4)		여가활동, 여우털, 여행객, ...
		여운명 (인명), 여주 (지명), ...

결국 사전을 향상시키기 위해서는 이와 같은 미등록어들에 대한 어휘적, 의미적, 구조적 분석을 시작해야 한다. 사전을 체계적으로 구축해야 한다는 이글의 입장에 뒤늦게 합류하는 셈이다.

4.1.2. 구문 분석을 위한 정보 베이스

단순어 사전과 접두사 및 접미사 사전을 토대로 하여 '파생어 사전 <DECOA/V01>'을 구축하게 되면, 일단 파생어 사전이 완성되었다 하더라도, 모든 엔트리에는 접사 정보가 들어가 있으므로, 응용 분야에 따라 매우 유용한 정보 베이스가 구축될 수 있다. 예를 들어, (5)와 같은 명사들에 (6)과 같이 접미사 '화'가 결합되면,

(5) 세계, 정보,
(6) 세계화, 정보화

다음과 같이 모두 '하다'를 수반한 동사구의 형태를 취할 수 있게 된다.

(7) *세계하다, *정보하다
(8) 세계화하다, 정보화하다

위의 (8)은 접미사 '화'가 명사에 서술성을 부여하기 때문에 나타나는 현상이다. 접사들에 있어서도 그 의미적 중의성이 반드시 분류되어 기술되어야 하는데, 예를 들어, 다음과 같은 접미사 '화'는 '하다'와 같은 조합을 허용하지 않기 때문이다.

운동화, 등산화, ...
야생화, 생화, ...
초상화, 벽화, ...

위의 단어들도 모두 '명사'에 접미사 '화'가 결합하여 형성된 것들인데, 이 접사들은 (6)의 것과는 달리 명사에 서술성을 부여하지 않는다. 이때, 어떤 접사 '화'는 서술성을 유도하여 '하다'와의 '공기 (Collocation)'를 허용하게 되고, 어떤 접사 '화'는 이와 같은 구성을 허용하지 않을지 어떠한 규칙으로써 그것을 예측하고 추정할 수 있을까 ? 이것은 사전을 구축하지 않는 한 어떠한 알고리즘의 구상으로도 불가능하다.

'N-하다' 유형의 동사구의 목록도 위의 (6)에서와 같은 '화' 접미사 파생 명사의 사전이 구축되지 않는 한 기대하기 어렵고, 따라서, '하다'를 술어로 하는 구문 분석의 체계적인 문법을 기술하기 어렵다.

4.2. 파생 접사의 의미 및 층위 분류

위에서 본 바와 같이, 접사들중에는 '동형어 (Homography)'들이 많다. 가령, 접미사 '화'의 경우, 7가지의 의미를 가질 수 있는 데, 이때, 그 의미적 차이는 과연 어떻게 기술될 수 있는가 ?

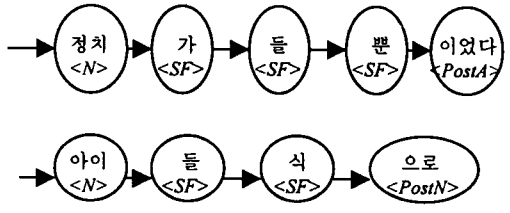
한 접미사가 가질 수 있는 의미에 대한 기술은, 결국 그 접미사가 함께 나타날 수 있는 명사들을 보여주는 일이다. 의미를 다른 표현으로써 '환언(Paraphrase)'하는 작업은 객관적으로 실현되기 어렵고 형식적인 기술 (Formal Description)이 되기 어렵다. 접미사의 의미는, 현재로는 그것이 결합될 수 있는 명사의 모든 리스트를 구축함으로써 기술되어야 한다.

다른 품사들의 의미 기술도 이와 크게 다르지 않다. 하나의 동사 '묻다'의 여러 의미의 기술은, 그 어느 환언적 설명보다도 그 동사가 나타날 수 있는 구문을 보이는 방법이 가장 정확하다. 앞서 본 바와 같이, 예를 들어 '잉크가 옷에 묻다'에서 사용된 동사 '묻다'의 의미도 이와 같이 구체적인 문장을 사용하지 않고는 그 기술이 더 복잡하고 모호할 것이다.

접사들은 또한 여러 개가 서로 결합하여 나타날 수 있다. 이때 그들간의 결합에는 여러 층위의 제약이 있다. 다음에서 보듯이,

(9) 회의에 임한 사람은 정치가들뿐이었다
(10) 그가 요즘 아이들식으로 머리를 잘랐다

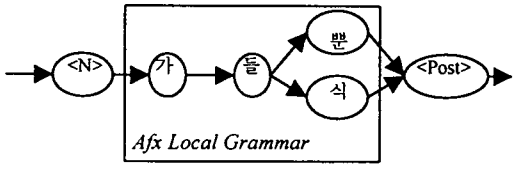
(9)와 (10)은, 각각 다음과 같이 구성된 파생명사 어절을 내포하고 있다.



위에서 접미사 '가', '들', '뿐', '식'은 서로 자유롭게 결합되어 나타날 수 없다. 예를 들어, 다음과 같은 순서로는 결합될 수 없다.

*정치-들-가-뿐
*아이-식-들

이와 같은 결합 순서상의 제약 조건은 쉽게 규칙화하여 '예측'할 수 있는 부분이 아니다. 이것은, 모든 접사들의 어휘적 속성을 파악하여, 각 접사들 사이의 결합 순서를 제어하고 있는 문법을 기술하는 방법을 취해야 한다. '부분 문법 (Local Grammar)'은 이와 같은 정보를 기술하는 데 효율적이다. 가령, 위에서 관찰된 유형은,



와 같은 부분 문법의 형태로 기술될 수 있다. 이때, 전통 문법에서 '접미사'와 '불완전 명사'를 구분 짓는 기준도 다시 검토되어야 한다. 품사의 설정이 어떤 형식적, 객관적 기준을 바탕으로 이루어지지 않았을 때, 그들간의 결합 제약에 대한 문법을 기술한다는 것은 무의미하다. 따라서, 우리는 '접미사'에 대한 그 언어 이론적 정의보다는, 위와 같은 스트링들에 대한 올바른 정보를 빠짐없이 기술하기 위하여, 단순 명사와 후치사 사이에 나타날 수 있는 모든 의존 형태들을 일단 '접미사'로 간주하였다.

5. 맺음말

하나의 형태소 분석기가 엄청나게 많은 분석의 가능성을 제시한다는 사실은, 그 자체가 시스템 성능의 평가에 장애가 되는 요인은 결코 아니다. 오히려, 형태소 분석은 모든 자연어 처리 모듈에 필연적으로 선행되어야 하는 단계이며, 그 자체로써는 실질적 존재 의미를 갖지 못하므로, 가능한 한 다양하고 세밀하게 형태소 분석이 이루어지는 것이 실은 더 바람직하다. 형태소 해석 결과는 그 응용 범위나 목적에 따라 얼마든지 그 세분성의 정도를 조절하여 출력해 줄 수 있기 때문이다.

그러나 그렇다고 해서, 자연 언어의 모든 형태적, 구문적, 의미적, 또는 담화적 정보까지 낱낱이 모두 한번에 사전에 수록할 수는 없다. 형태소 해석 단계에서 우선 필요한 사전은 형태적 정보를 빠짐없이 포함하고 있어야 한다는 점이 제일 중요하며, 각 엔트리에 대한 구문 및 의미 정보는 명시적으로 (explicitly), 그리고 체계적으로 (systematically) 등재될 수 있도록 더 많은 검토가 필요하다.

그러나, 한국어 형태소 분석기의 경우, 유럽어나 영어의 경우와는 달리 '품사 정보'나 '어휘 형성 정보' 뿐 아니라, '어절 형성 정보' 및 부분적인 '구문 구조 정보'등이 요구된다. 따라서 형태소 분석시에 발생할 수 있는 중의성 (Ambiguity)의 양이 엄청나며, 그와 같은 과분석의 수를 줄이기 위해서는 '어절 정보 사전'을 구축하는 것이 필요하다. 이글은 '어절 정보 사전 시스템 DECO' 의 한 하위 모듈을 이루는 '파생 명사' 사전의 구성에 대하여 논의하였다. 사전을 구축하는 비용이 크고, 빠른 시일내 그 결과를 기대할 수 없지만, 체계적인 사전을 구축하지 않고는 결코 해결할 수 없는 부분들이, 자연어 처리의 분야에 있어서는 분명히 존재한다는 사실을 잊지 말아야 한다.

참고 문헌

Gross, Maurice, 1987, The use of finite automata in the lexical representation of natural language, *Lecture Notes in Computer Science* 377, Springer-Verlag.

Nam, Jee-Sun, 1995, *Constitution d'un lexique électronique des noms simples en coréen*, papers in LGC-1995, UQAM, Canada.

Nam, Jee-Sun, 1996, *Construction of Korean Electronic Lexical System DECO*, Papers in Computational Lexicography Complex '96, ed. by F. Kiefer et al.: Linguistics Institute, Hungarian Academy of Sciences.

Silberztein, Max, 1993, *Dictionnaires électroniques et analyse automatique de textes, Le système INTEX*, Paris : Masson.