# 형태소 분석 결과의 인코딩 기법과 어절 사전 구축

강승식
국민대학교 컴퓨터학부, 첨단정보기술연구센터
sskang@kookmin.ac.kr

# Encoding of Morphological Analysis Result and Eojeol Dictionary Construction

Seung-Shik Kang
School of Computer Science, Kookmin University & AITrc

## 요 약

형태소 분석에서 사용되는 사전은 형태소와 품사 정보를 수록하고 있다. 단어가 한 개의 형태소로 구성되는 굴절어는 대부분의 단어가 어휘형태소의 기본형과 일치되기 때문에 형태소 분석 알고리즘은 사전 탐색과 형태론적 변형을 통해 입력 단어와 어휘형태소를 일치시키는 과정으로 기술된다. 이에 비해, 교착어는 입력 어절이 형태소 사전의 어휘형태소와 일치하지 않기 때문에 어절 자체가 형태소 사전에 포함되지 않아서 굴절어에 비해 상대적으로 형태소 분석 알고리즘의 복잡도가 높고 분석 시간이 오래 걸리는 단점이 있다. 본 논문에서는 고빈도 어절에 대한 기분석 어절 사전을 구축하여 형태소 분석 속도를 개선하고, 사용자가 어절 사전에 새로운 어절을 추가하거나 어절 사전에 수록된 분석 결과를 수정할 수 있는 어절 사전에 의한 형태소 분석 방법을 제안한다. 구체적인 방법론으로써 형태소 분석 결과를 저장하는 기분석 어절 사전의 크기를 최소화하기 위해 분석 결과를 생성하는데 필요한 최소한의 정보만을 인코딩하는 방법을 사용한다.

## 1. 서 론

한국어 형태소 분석은 어휘형태소와 문법형태소의 결합 제약에 따른 형태소 분리 문제를 해결하는 방법을 중심으로 분석 방법론 및 가능한 모든 분석 후보를 생성할 때 발생하는 사전 탐색의 효율성 문제를 해결하는 방법으로 발전해 왔다[1,2,3,4,9]. 형태소 분석의 정확도를 개선하는 문제가 어느 정도 해결된 후에는 분석 속도를 향상시키는 방법에 관한 연구가 수행되었으며, 그 결과로 어절 사전을 이용하는 방법, 통합형태소 단위로 분석하는 방법, 부분 어절의 기분석 방법 등이 제안되었다[5,6,7,11].

한국어는 어휘형태소에 조사, 어미, 접미사 등 문법형태소가 중첩 결합되어 어절을 구성하기 때문에 하나의 어휘형태소(lexical morpheme)로부터 생성되는 어절의 가지수가 매우 많다. 어휘형태소의 개수를 10만개로 가정하고 어휘형태소에 결합이 가능한 접미사, 조사, 어미 등 문법형태소(grammatical morpheme)가 결합되는 활용 가지수를 1,000 가지라고 할 때 이로부터 생성되는 어절수는 약 1억 어절이 된다.[1] 따라서 형태소 분석기를 구현할 때 각 어절에 대한 분석 결과를 미리 기분석 사전으로 저장하고 있거나, 혹은 어절 단위로 분석 방법에 관한 정보를 어절 사전으로 구축하는 방법을 적용하기가 쉽지 않았다.

이러한 특성은 굴절어와 교착어의 차이로 인한 것이다. 굴절어는 기본적으로 어휘형태소 자체가 단어를 구성하고 형태론적 변형에 의해 어휘형과 표층형이 달라지는 경우가 발생한다. 이에 비해, 교착어는 기본적으로 2개 이상의 형태소로 구성되기 때문에 형태소 사전에 수록되는 어휘형태소와 입력 단어(어절)가 일치하는 비율이 낮다. 따라서 굴절어는 형태소 분석 사전에 수록된 형태소와 입력 단어의 일치율이 매우 높지만, 교착어는 입력 단어와 형태소 분석 사전에 수록된 형태소의 직접 일치 비율이 낮다[10]. 이러한 이유로 인하여 굴절어의 형태소 분석은 어절 사전에 기반한 형태소 분석 방법을 쉽게 적용할 수 있으나, 교착어는 모든 어절들을 사전에 수록할 수가 없으므로 어절 사전만으로 형태소 분석을 하는 것은 쉽지 않다.

그런데 대용량 한글 문서 집합에서 어절 빈도 조사에 의하면 한국어도 영어 등 타 언어에서와 유사하게 통상적으로 자주 사용되는 어절들이 고빈도 어절의 최상위에 위치하고 있음을 알 수 있다[5,8,12]. 이러한 현상은

1) 체언의 경우 인명, 지명, 상품명, 외래어, 복합명사 등 사전 미등록어가 계속해서 생성되므로 어휘형태소의 개수는 100만개 이상이다. 이 경우에 어절수는 10억 어절 이상이 된다.

교착어의 경우에도 영어의 사용 빈도와 유사하게 빈도수에 따른 어절의 분포도가 규칙적임을 보여준다[13,14,15]. 즉, 형태소 단위의 빈도 특성보다는 약하지만 한국어의 고빈도 어절 집합이 전체 어절의 60%~80%를 차지하고 있다[8,12]. 이러한 어절 빈도의 특성에 따라 형태소 분석 속도를 개선하고 사용자가 분석 결과를 원하는 형태로 수정할 수 있는 방법을 제공하기 위하여 고빈도 어절들에 대한 형태소 분석 결과를 기분석 어절 사전으로 구축하는 방법을 사용할 수 있다.

본 논문에서는 형태소 분석 결과를 저장하는 기분석 어절 사전의 크기를 최소화하기 위해 분석 결과를 생성하는데 필요한 최소한의 정보만을 인코딩하는 기법을 제안한다.

## 2. 형태소 분석과 어절 사전

일반적으로 형태소 분석 시스템은 어휘형태소 사전과 형태론적 변형 규칙으로 구성된다. 형태론적 변형 규칙은 단지 어휘형태소의 변형에 관한 규칙이지만, 어휘형태소 사전은 각 어휘들에 대한 각종 분석 정보를 포함하고 있다. 영어의 형태소 분석이 쉬운 까닭은 영어의 변형 규칙이 단순-명료하기 때문이다. 영어는 명사의 복수형과 동사의 3인칭 단수형, 그리고 과거 및 과거분사형에 관한 규칙이 매우 단순할 뿐만 아니라 규칙이 적용되는 문맥 제약 조건이 간단하여 예외가 거의 발생하지 않는다. 다만, 불규칙 동사는 변형 규칙의 적용을 받지 않기 때문에 형태론적 변형이 아니라 사전에 관련 정보를 기술하는 사전 기반 방식으로 처리되어야 한다.

또한, 영어의 접미사(suffix)는 어휘형태소와 접미사가 융합되어 그 경계에서 변형 현상이 심하게 일어나기 때문에 변형 규칙에 의한 어휘형태소의 원형 추출이 쉽지 않다. 접미사가 분리된 정확한 형태소 분석의 생성을 위해서는 사전에 관련 정보가 기술되어 있어야 한다. 이 때 접미사 분리 대상이 되는 단어들은 대부분 이미 사전에 수록되어 있으므로 단순히 이 단어들에 대한 접미사 분리 정보만 추가하면 된다. 따라서 사전에 수록되어야 할 어휘수는 영어의 경우 통상적으로 어휘형태소 사전에 수록되는 어휘의 개수를 n이라고 할 때 여기에 불규칙 동사의 개수를 추가한 것이다. 즉, 불규칙 동사의 개수를 c개라고 할 때 영어 형태소 분석 사전에 수록되는 어휘수는 n+c개이다.

영어 형태소 분석은 형태론적 변형 규칙을 사용하지 않고 모든 표층형을 사전에 수록하는 방법이 가능하며, 이 경우에 명사의 복수형 등이 사전에 추가된다. 명사의 복수형, 동사의 3인칭 단수형 및 과거-과거분사형 등 사전에 추가될 어휘의 개수를 n′이라 할 때 사전의 크기는 N(E) = n + n′ + c 이다. 즉, 영어는 n개의 어휘형태소로부터 생성되는 모든 단어의 개수가 많지 않으므로 변형된 단어들을 모두 사전에 수록하는 것이 가능하며, 단어의 표층형과 함께 그 분석 결과를 모두 사전에 수록하는 사전 기반 분석 방법을 쉽게 적용할 수 있다.

$$\text{영어의 단어수} : N(E) = n + n' + c$$

또한, 영어 형태소 분석은 단어의 기본형과 불규칙 동사에 관한 사전만 사용하더라도 입력 단어의 대부분이 사전에 존재한다. 따라서 사전 탐색이 가장 중요한 요소이며, 형태소 분석 사전을 설계할 때 탐색 효율성을 위해 사용 빈도에 따라 초고빈도어(ultra-high frequency word), 고빈도어(high frequency word), 그리고 저빈도어(low frequency word)로 구별된 계층구조 사전을 구성할 수 있다[16]. 이 방법은 컴퓨터 자원이 충분치 않던 시기에 매우 유용하게 사용되었다.

영어의 고빈도어는 적중률(hit-ratio)이 높기 때문에 효용 가치가 높지만, 한국어의 경우 적중률이 낮아서 상대적으로 효용 가치가 크지 않다. 그 이유는 조사/어미가 결합되지 않은 어휘형태소만으로 구성된 어절이 많지 않을 뿐만 아니라, 문법형태소를 분리한 임시 중간 후보들이 생성되었을 때 이 후보들 중의 다수가 사전에 존재하지 않기 때문이다. 다시 말해서, 영어는 입력 단어 자체가 분석 사전에 수록되는 비율이 높지만, 한국어는 교착어의 특성 때문에 그 비율이 매우 낮다.

한국어의 어휘형태소 개수를 기준으로 한국어의 총 어절수를 계산하면 아래와 같다. 이 식에서 n은 통상적으로 국어사전에 수록되는 한국어의 단일 형태소 개수이고, m은 단일어의 조합으로 구성되는 복합어 및 국어사전에 수록되지 않는 전문용어, 그리고 기타 미등록어를 포함한다.[2] 상수 a는 각 형태소들이 조사, 어미, 접미사 등 문법형태소와 결합하여 생성되는 평균 어절수를 나타낸다. 이 상수는 체언과 용언, 독립언을 통틀어 하나의 상수값으로 통칭한 것으로, 실제로는 체언 어절과 용언 어절에 비해 독립언은 문법형태소와 결합하는 경우가 거의 없다.

$$\text{한국어의 어절수} : N(K) = a(n+m)$$

2) 인명, 지명 등 고유명사는 그 개수를 계산하기가 어려우므로 제외하였다.

영어의 단어수 N(E)와 한국어의 어절수 N(K)는 매우 큰 차이가 있다. 영어는 굴절 현상에 의해 생성되는 단어수 n'이 통상적으로 영어사전에 수록되는 단어수 n과 비슷한 개수이고, 불규칙 활용으로 생성되는 단어수 c는 n보다 매우 작으므로 N(E)의 근사치는 2n이다.

한국어의 어절수는 복합어와 미등록어의 개수를 어휘 형태소 개수에 비례하다고 가정하여 m=bn 이라고 하면, N(K) = abn이다. 상수 a와 b는 그 값을 구하기는 매우 어렵다. a는 문법형태소에 의해 생성되는 평균 어절수이므로 조사, 어미의 개수를 고려할 때 수백~수천으로 추정된다. b는 복합어와 미등록어의 개수로 문서의 유형에 따라 그 비율이 차이가 크며, a에 비해 매우 작은 값으로 추정된다. 따라서 a=100, b=2라고 하더라도 총 어절수는 영어에 비해 100배가 되고 a=1000 이면 1,000배가 된다.

따라서 영어는 분석 결과를 생성하여 미리 사전에 저장해 놓고, 단지 사전 탐색만으로 형태소 분석을 하는 방법을 적용하기가 쉽다. 그러나 한국어에 이 방법을 적용할 경우 사전의 크기가 매우 커지는 문제점이 발생하므로 분석 규칙을 사용하지 않고 어절 사전에만 의존하는 분석 방법만을 적용하는 것은 불가능하다. 다만, 적중률이 높은 고빈도 어절들에 대해서만 사전 기반 분석 방법을 적용하여 분석 속도 등 효율성을 높이고, 저빈도 어절들에 대해서는 기존의 방법에 의해 분석하는 혼합 방법을 적용하여 형태소 분석을 두 단계로 구분함으로써 성능을 개선할 수 있다. 김재한(1994)에 따르면 350만 어절 말뭉치에서 고빈도 어절 15만개가 전체 어절의 75% 이상을 차지하고 있으며, 윤준태(2000)는 1,000만 어절 및 3,000만 어절 말뭉치에서 고빈도 어절 6만개가 각각 전체 어절의 75%와 70%를 차지하고 있다는 사실을 이용하여 고빈도 어절 6만개에 대해 수동으로 형태소 분석 결과를 생성하였다.

강승식(2001)은 적절한 크기의 고빈도 어절 집합을 구성하고 그 효용성을 검증하기 위한 실험을 수행하였다. 그 방안으로 700만 어절 및 1,500만 어절 말뭉치로부터 각각 고빈도 어절 집합을 추출하여 전체 어절의 90%를 차지하는 어절집합에 대해 두 말뭉치에서 공통으로 나타난 고빈도 어절 18만여개를 추출하였다. 고빈도 어절 집합의 효용성을 검증하기 위한 어절 적중률 실험에 의하면, 문서 유형에 따라 적중률이 72%~87%이고 평균 81.6%의 적중률을 보이고 있다. 고빈도 어절 집합의 크기에 따른 어절 적중률은 9,484 어절이 51.5%, 22,906 어절이 61.5%, 62,044 어절이 72.1%, 99,213 어절이 76.5%, 그리고 184,128 어절이 81.6%이다. 이 고빈도 어절 집합은 문서 유형과 무관하게 통상적으로 한글 문서에서 빈번하게 사용되는 어절들이다. 이 고빈도 어절 집합에 대해 기분석 어절 사전을 구축함으로써 형태소 분석 속도 및 정확도를 향상시킬 수 있다.

## 3. 형태소 분석 결과의 인코딩

### 3.1 형태소 분석 결과의 표현 방식

고빈도 어절 집합에 대한 기분석 어절 사전을 구축할 때 문제가 되는 것은 형태소 분석 결과를 저장하는데 필요한 기억장소가 크기 때문에 기분석 어절 사전의 크기가 커진다는 점이다. 기분석 어절 사전에 수록되는 어절 개수가 적으면 적중률이 낮아지고, 어절 개수가 많으면 적중률의 개선 효과에 비해 어절 사전의 크기가 기하급수적으로 커지게 된다. 적절한 규모의 어절 사전 크기를 10만~30만 어절이라고 가정할 때 각 어절당 분석 정보가 1바이트 증가될 때마다 어절 사전의 크기가 100K~300K 바이트씩 커지게 된다. 어절에 대한 형태소 분석 결과를 단순히 {〈형태소, 품사〉}* 형태로 어절 사전에 저장할 경우에 어절 사전의 크기는 아래와 같이 계산된다.

$$(|w| + r \times m \times (|wi| + |pos|) + (r-1)) \times n$$

|w| : 입력 어절의 바이트 길이
r : 어절당 평균 분석 결과의 수
m : 분석결과에 포함된 평균 형태소 개수
|wi| : 각 형태소의 바이트 길이
|pos| : 품사 정보의 바이트 길이
r-1 : r개의 결과를 구분하는 구분자
n : 총 어절수

각 어절의 형태소 분석 결과를 저장하는 방법으로는 형태소 분석기가 생성한 결과 구조체를 어절 사전에 그대로 저장하는 방법과, 분석 결과에 대한 출력 문자열을 저장하는 방법 등이 있다. 첫 번째 방법은 형태소 분석기를 실행하여 생성된 결과를 어절 사전에 dump하는 것이다. 반면에, 결과가 binary 형태로 저장되기 때문에 수동으로 수정하기가 용이하지 않고, 분석 결과의 저장에 필요한 기억장소의 크기가 매우 커지는 단점이 있다. 그림 1의 구조체를 변형하여 각 분석 결과를 저장하는데 불필요한 기억장소 낭비를 최소화하여 30바이트를 차지하도록 수정했다고 가정할 때 20만 어절에는 6M 바이트 이상이 필요하다.

```
typedef struct ham_result { // 형태소 분석 결과
   int score;    // score of this result

   char patn;    // word pattern
   char type;    // type of input word

   char stem[STEMSIZE];// stem of input word
   char pos;     // 3 simplified stem type
   char pos2;    // pos attr. for 'pos'
   char dinf;    // POS info. in Han-dic
   char nsfx;    // index of noun suffix
   char vsfx;    // index of verb suffix

   char josa[JOSASIZE]; /* Josa string
   char *jlist;  //  unit-Josa sequence
   char eomi[EOMISIZE]; /* Eomi string
   char *elist;  //  unit-Eomi sequence
   char pomi;    // encoded prefinal Eomi

   char xverb[XVERBSIZ];  // xverb string

   /* morphological attributes */
   char vtype;  // irregular verb type
   struct jomi_variant jomi; //Josa/Eomi variant info.
} HAM_RESULT, *HAM_PRESULT;
```

[그림 1] 형태소 분석 결과의 저장 구조

두 번째 방법은 그림 2와 같이 각 어절에 대한 형태소 분석 결과를 텍스트 문자열로 기분석 어절 사전에 저장하는 것이다. 이 경우에 분석 결과가 텍스트 파일 형태로 저장되므로 사용자가 직접 분석 결과를 편집하여 수정이 가능하다. 다만, 이 방법은 각 분석 결과가 텍스트 문자열 형태로 저장되기 때문에 필요한 정보를 추출하려면 문자열로 된 결과를 파싱하는 과정을 거쳐야 한다. 그림 2의 '가는'에 대한 분석 결과는 입력 어절을 포함하여 158바이트이고, '컴퓨터로'에 대한 분석 결과는 43바이트이다. 따라서 이 방법은 어절 사전의 크기가 첫 번째 방법과 유사하거나 더 커지게 된다.

```
가는
   (V "가")〈IgW : 20〉 + (e "는")
   (V "갈")〈T : 20〉 + (e "는")
   (V "가")〈IgW : 24〉 + (e "어는")〈4〉
   (V "가늘")〈J : 24〉 + (e "ㄴ")〈13〉
   (N "가")〈N : 27〉 + (j "는")〈1〉
컴퓨터로
   (N "컴퓨터")〈N : 24〉 + (j "로")〈2〉
```

[그림 2] 형태소 분석 결과의 출력 문자열 예

그림 3은 그림 2의 형태소 분석 결과를 〈토큰, 품사〉 쌍 형태로 변환하여 단순화시켜 출력한 예이다. 그림 2의 "(V "가")〈IgW : 24〉 + (e "어는")〈4〉"에는 어휘형태소의 품사 정보가 용언 '가다'로 분석되었다는 표시 'V'와 사전에 기술되어 있는 품사 정보 "IgW"(자동사와 보조용언)로 구분되어 있는데, 이에 비해 그림 3의 "가/VV + 어는/EM"에는 상세한 품사 정보가 나타나지 않는다. 품사 정보를 단순화함으로써 품사 정보가 손실되지 않도록 하려면 세분화된 품사에 대해 각각 분석 결과를 생성하여야 한다. 즉, 용언 '가다'는 자동사 분석 결과와 보조용언 분석 결과를 각각 생성해야 하는 문제점이 발생한다. 또한, 형태소 분석 결과를 활용하는 측면에서 볼 때, 활용 분야에 따라 어미 '어는'에 대해 '어'가 복원되었다는 정보를 필요로 하는 경우가 있으므로 기분석 결과에 관련 정보가 포함되어야 한다.

```
가는
   가/VV + 는/EM
   갈/VV + 는/EM
   가/VV + 어는/EM
   가늘/VV + ㄴ/EM
   가/NN + 는/JO
컴퓨터로
   컴퓨터/NN + 로/JO
```

[그림 3] 토큰 기반 형태소 분석 결과의 예

### 3.2 분석 결과의 인코딩

기분석 어절 사전을 구성할 때는 형태소 분석 결과에 어떤 정보를 포함시킬 것인지를 결정하는 문제는 형태소 분석 결과를 사용하는 응용 시스템의 요구 사항에 따라 달라진다. 그림 1과 그림 2에서 예시한 2가지 방법론은 모두 형태소 분석 결과를 그대로 어절 사전에 수록하는 방법이다. 이에 비해, 그림 3의 토큰 기반 분석 결과는 분석 결과에 대한 모든 정보를 저장하는 것이 아니라 토큰 스트링과 품사 정보만을 분석 결과로 정의하여 완성된 형태의 분석 결과를 저장하고 있다. 이 방법에서는 형태소 분석 사전에 기술된 구체적인 정보와 형태소 복원 정보 등 분석 과정에서 생성되는 각

종 정보들이 누락되기 때문에 관련 정보를 필요로 하는 응용 시스템에는 정보량이 부족하다. 기분석 결과를 구성할 때 〈형태소, 품사〉 정보 이외에도 형태소의 변형 정보 등 가급적 많은 정보를 기술하면서 사전의 크기를 작게 유지하는 방법이 필요하다.

형태소 분석된 결과를 저장할 때 가장 많이 차지하는 기억장소는 형태소 스트링이다. 그런데 각 형태소 스트링은 직접 저장하지 않고 어휘형태소의 길이와 어근 변형 정보에 의해 각 형태소를 분리할 수 있다. 단, 선어말어미와 접미사는 별도의 항목으로 관련 정보를 저장한다. 형태소 분석 결과를 생성하는데 필요한 정보는 그림 4와 같이 7가지로 정의하였다.

'어절 패턴'은 형태소 분석기 내부적으로 정의된 어절의 유형을 구분하는 정보로써 체언 6가지, 용언 5가지, 기타 3가지로 총 14가지 유형으로 구분한다. '어근 길이'는 입력 어절에서 어근으로 분리되는 문자열의 길이이다. 사전의 '품사 정보'는 형태소 분석 사전에 수록된 어휘형태소의 품사 정보이며, '선어말 어미'는 '시/었/었/겠'이 분리되었는지 여부를 각각 1 bit씩으로 저장한다. '용언 접미사'는 접미사 테이블을 만들어 놓고 그 테이블에 대한 인덱스를 저장하고, '체언 접미사'는 접미사의 음절수(길이 정보)를 저장하여 입력 어절로부터 직접 분리할 수 있게 한다. '어근 변형 정보'는 각 변형의 유형들을 정의하여 해당 유형들에 대해 입력 어절로부터 어휘형태소와 문법형태소의 원형을 복원하기 위한 것이다.

각 항목들에 대해 할당되어야 할 값의 범위를 고려하여 인코딩되는 정보의 크기는 각 분석 결과에 대해 4바이트씩 차지한다. 즉, 기억공간의 크기가 작은 것을 2개씩 묶어서 바이트 단위로 구성되도록 4바이트로 구성한다. 그림 3에는 순서대로 각 항목들에 대한 기억공간의 크기를 명시하였다.

- 어절 패턴 : 4비트(0~15)
- 어근 길이 : 4비트(0~15)
- 사전의 품사 정보 : 1바이트(0~255)
- 선어말 어미 : 4비트(0~15)
- 용언 접미사 인덱스 : 4비트(0~15)
- 체언 접미사 길이 : 2비트(0~3)
- 어근 변형 정보 : 6비트(0~63)

[그림 4] 분석 결과 생성에 필요한 정보

## 4. 실험 및 평가

한글 어절의 적중률은 영어보다 낮지만 그렇더라도 18만개 규모의 고빈도 어절 집합이 한글 문서에서 차지하는 비율은 60%~80% 정도로 계산된다. 따라서 이 어절 집합에 대해 기분석 어절 사전을 구축하여 형태소 분석 정보를 저장하고 이 사전을 적용한 형태소 분석기의 성능을 실험하였다. 고빈도 어절 집합에 대한 기분석 어절 사전은 기존의 형태소 분석기 HAM version 6.0의 출력 결과를 수정하여 자동으로 구축하였다.

그 결과로 184,104 어절에 대해 211,254개의 분석 결과가 생성되었으며, 이는 어절당 평균 1.15개의 분석 결과가 생성된 것이다. 예를 들어, '가는'에 대한 어절 사전의 구축 내용은 아래와 같다. 이 예제는 편의상 형태소 분석 결과를 생성하는데 필요한 7가지 정보를 각각 1바이트씩 프린트 가능한 문자로 출력한 것이다. 184,104개 어절에 대해 기분석 어절 사전 구축에 필요한 기억 공간은 총 2,345K 바이트로 이 중에서 어절 스트링이 1,500K 바이트를 차지하고, 분석 정보가 845K 바이트이다.

가는 V2v000@V2v000CV2v000'V4v000cN2n000@

구축된 어절 사전의 효용성을 평가하기 위해 기분석 어절 사전에 존재하는 어절은 이 사전에 의해 분석을 하고, 어절 사전에 존재하지 않는 어절들은 기존 방식인 형태소 분석 알고리즘으로 분석하는 실험을 수행하였다. 실험 데이터는 정보통신 분야의 신문기사 458개로 문서의 크기는 약 1M bytes(약 15만 어절)이고, 이 데이터는 고빈도 어절사전을 구축하는데 전혀 사용되지 않은 문서이다. 고빈도 어절 사전을 적용했을 때와 적용하지 않았을 때의 분석 시간을 비교한 결과에 의하면 분석 시간 비율이 1.0 : 1.59 으로 어절 사전을 사용하였을 때 분석 속도가 향상되는 것을 확인하였다.

## 5. 결 론

한국어 형태소 분석에서 가장 중요시 되고 있는 것은 분석 결과의 정확도와 분석 속도 문제이다. 분석 결과의 정확도는 예외 현상이 발생하는 어절들로 인해 형태소 분석 알고리즘의 개선만으로는 해결하기가 어렵고 분석 오류를 유발하는 어절 혹은 형태소들에 대한 사전 보완이 병행되어야 한다. 본 논문에서는 기분석 어절 사전을 도입하여 분석 오류가 발생한 어절에 대해 사용자가 분석 결과를 추가하거나 수정할 수 있도록 하였고, 고빈도 어절에 대해 분석 알고리즘을 적용하기 전에 어절 사전을 적용함으로써 분석 속도를 향상시키는 효과가 있었다. 어절 사전을 도입할 때 문제가 되어 왔던

어절 사전의 크기 문제를 해결하기 위해 분석 결과 생성에 필요한 최소한의 정보만을 인코딩하는 기법을 제안하였다.

기분석 어절 사전의 목적은 고빈도 어절에 대해 매번 형태소 분석을 행하는 비효율성을 개선하기 위해 기분석 결과를 미리 생성하여 사전 탐색으로 처리하는 것과 형태소 분석기가 오분석을 하는 어절들에 대해 기분석 사전의 분석 결과를 수동으로 수정함으로써 정확도를 높이는 것이다. 향후 연구로써, 기분석 사전에 수록된 어절 정보를 이용하여 기분석 어절 사전에 수록되지 않은 미등록어의 분석 결과를 예측하는 방안으로 발전시킬 수 있을 것이다.

## Acknowledgements

본 연구는 첨단정보기술 연구센터를 통하여 과학재단의 지원을 받았음.

## 참고문헌

[1] 강승식, *한국어 형태소 분석과 정보 검색*, 홍릉과학출판사, 2002.

[2] 이은철, 이종혁, "계층적 기호 접속정보를 이용한 한국어 형태소 분석기의 구현", 제4회 한글 및 한국어 정보처리 학술발표 논문집, pp.95-104, 1992.

[3] 최재혁, 이상조, "양방향 최장일치법에 의한 한국어 형태소 분석기에서의 사전 검색 횟수 감소 방안", 정보과학회 논문지, 20권 10호, pp.1497-1507, 1993.

[4] 임희석, 윤보현, 임해창, "배제정보를 이용한 효율적인 한국어 형태소 분석기", 정보과학회논문지, 22권 6호, pp.957-964, 1995.

[5] 김재한, 옥철영, "어절 사전을 이용한 한국어 형태소 분석", 한국정보과학회 봄 학술발표 논문집, pp.813-816, 1994.

[6] 김재한, 옥철영, "통합형태소를 이용한 한국어 형태소 분석기", 한국정보과학회 가을 학술발표 논문집, pp.653-656, 1994.

[7] 양승현, 김영섬, "부분 어절의 기분석에 기반한 고속 한국어 형태소 분석 방법", 정보과학회 논문지 : 소프트웨어 및 응용, 27(3), pp.290-301, 2000.

[8] 강승식, "어절 빈도 조사에 의한 최적의 고빈도 어절 집합 추출", 제13회 한글 및 한국어 정보처리 학술발표 논문집, pp.85-88, 2001.

[9] H. Kwon, Y. Chae, and G. Jeong, "A Dictionary-based Morphological Analysis", Proceedings of NLPRS'91, pp.87-91, 1991.

[10] D. Kim, S. Lee, K. Choi, and G. Kim, "A Two-level Morphological Analysis of Korean", In *Proceedings of the 15th International Conference on Computational Linguistics (COLING'94)*, pp.535-539, 1994.

[11] K. Shim and J. Yang, "MACH : A Supersonic Korean Morphological Analyzer", In *Proceedings of the 19th International Conference on Computational Linguistics (COLING'2002)*, pp.939-945, 2002.

[12] J. Yoon, "Compound Noun Segmentation Based on Lexical Data Extraction from Corpus", In *Proceedings of the 6th Applied Natural Language Processing*, pp.196-203, 2000.

[13] H. P. Zipf, "Human Behaviour and the Principle of Least Effort", Addison-Wesley, 1949.

[14] G. A. Miller and E. B. Newman, "Tests of a Statistical Explanation of the Rank-Frequency Relation for Words in Written English", *American Journal of Psychology*, vol.71, pp.209-218, 1958.

[15] G. Leech, P. Rayson, and A. Wilson, "Word Frequencies in Written and Spoken English : based on the British National Corpus", Longman, 2001.

[16] B. Boguraev and M. Neff, "Dictionary Structure and Lexical Relations : A Study in Applied Computational Lexicography", *IBM Technical Report*, IBM T. J. Watson Research Center, 1990.