음절수에 따른 한국어 복합 명사 분리 방안

최재혁

부산여자대학교 컴퓨터교육과

A Division Method of Korean Compound Noun by number of syllable

Jae-Hyuk Choi

Dept. of Computer Education, Pusan Women's University

요 약

한국어 맞춤법 검사기는 문서내에서 발생되는 비표준어 오류, 띄어쓰기/붙여쓰기 오류, 조사/어미 오류, 외래어 오류, 철자 오류 등에 대해서 이에 적합한 대치어를 제시해 준다. 일반적으로 한국어의 맞춤법 오류 중 가장 빈번하게 발생되는 것이 띄어쓰기 오류이며, 이 중에서도 복합 명사에 대한 띄어쓰기 오류가 가장 많이 발생한다. 본 논문에서는 복합 명사에 대한 띄어쓰기 교정 방안으로 복합 명사의 음절 수에 따라 1개의 결과만을 출력하도록 하는 복합 명사 분리 방안을 제시하며, 또한 복합 명사 분리 시의 사전 참조 횟수를 줄이는 방법을 제안한다.

1. 서론

정보화 사회라고 일컬어지는 현대 사회에서 기하급수적으로 증가하고 있는 정보를 처리하기 위하여 많은 소프트웨어가 등장하였고, 이러한 소프트웨어 중에서 가장 많이 사용되고 있는 것이 워드프로세서이다. 초기의 워드프로세서는 그 기능이 문서 작성에만 한정되어 있었으나, 최근에는 여러 가지 다양한 기능이 추가되었는데, 그 중의 하나가 맞춤법 검색 기능이다. 한국어 맞춤법 검사기는 문서 내에서 발생되는 비표준어 오류, 띄어쓰기/붙여쓰기 오류, 조사/어미 오류, 외래어 오류, 철자 오류 등을 발견하고, 이에 적합한 대치어를 제시한다[1,2,3].

한국어의 맞춤법 오류 중에서 대부분을 차지하는 것이 띄어쓰기 오류이다[4]. 이러한 띄어쓰기 오류 중에서도 가장 빈번한 오류로 발생되는 띄어쓰기가 되지 않은 한국어 복합 명사에 대한 올바른 분리 방안은 맞춤법 교정 시스템의 질을 한 단계 높일 수 있다. 일반적인 띄어쓰기 오류에 대한 처리 방안은 형태소 분석을 행한 후, 형태소 분석에 실패한 단어에 대해 많은 어휘 사전 참조를 행하여 오류를 처리하게 된다. 그러나 본 논문은 띄어쓰기 처리를 위한 사전 참조 횟수를 줄이기 위해 형태소 분석 시의 어휘 사전 참조 때, 어휘 사전에 수록된 띄어쓰기 정보를 이용하여 더 이상의 추가적인 사전 참조를 행하지 않고도 복합 명사의 분리를 가능하게 하는 방안과 또한 어휘사전에 수록되지 않은 복합 명사에 대한 분리의 정확성을 높이기 위하여 복합 명사의 음절 수에 따른 복합 명사의 분리 특성을 조사하여 분리 순서를 정하고 1개의 복합 명사 분리 결과만을 출력하도록 하는 복합 명사 분리 방안을 제시한다.

2. 복합 명사 분리 시스템

2.1. 양방향 최장일치법과 사전 구성

띄어쓰기 오류를 처리하기 위해서는 어절에 대한 형태소 분석이 먼저 이루어져야 한다. 본 시스템의 형태소 분석은 전체적인 시스템의 처리 속도를 고려하여 기존의 형태소 분석 방법 중 형태소 분석의 정확성을 보장하면서 가장 사전 참조 횟수가 적은 양방향 최장일치법을 채택하였다[5]. 양방향 최장일치법은 입력 어절에 대해 먼저 주기억장치에 저장된 조사어미사전을 검색하는 좌방향 최장일치를 행하여 어절의 최장 조사나 어미를 구한 후, 구한 최장 조사나 어미에 대한 조사어미사전에 있는 정보를 이용하여 우방향 최장 일치의 끝을 계산하여 구함으로써 어휘 사전 검색을 행해야 하는 어절의 범위를 제안하여 어휘 사전 검색의 횟수를 감소시키는 방법이다.

본 시스템에서 사용하는 사전은 한국어 어휘 사전, 조사어미 사전, 선어말어미 사전, 접미사 사전, 접두사 사전이며, 접두사 사전을 제외한 나머지 사전은 양방향 최장일치법에서 사용하는 사전의 내용과 동일하나, 어휘 사전의 경우 수록된 단어가 띄어쓰기가 되지 않은 복합 명사일 경우 단어와 띄어쓰기 정보

를 함께 수록하여 띄어쓰기를 위한 추가적인 사전 참조를 하지 않도록 하였다[5,6].

그림 1은 본 시스템에서 사용하는 한국어 어휘 사전의 구조와 복합 명사에 대한 사전 내용이다.

0 24 25 26 27 28

단어(_띄어쓰기 정보)	품사정보	불규칙정보
.....		
가감승제	1	0
가감속전동기_4	1	0
가감계도표시_35	1	0
가로방향수정자석_357	1	0
.....		

그림 1 한국어 어휘 사전 구조 및 내용

그림 1에서 품사 정보 1은 품사가 명사임을, 불규칙 정보는 용언에 대한 불규칙 정보로써 0은 불규칙이 없음을 의미한다. 단어에서 '_'뒤의 숫자는 띄어쓰기 정보를 의미한다. 예 '가감승제'는 띄어쓰기에 대한 정보를 가지고 있지 않으므로 '가감승제'를 한 단어로 처리하며, '가감속전동기_4'에서 '_4'는 4번째 음절 앞에서 띄어쓰기를 해야 한다는 정보로, 이 정보에 의해 '가감속'과 '전동기'로 추가적인 어휘 사전 참조 없이 분리된다. '가감계도표시_35'는 3번째 음절과 5번째 음절 앞에서 띄어쓰기를 하라는 의미로 '가감 계도 표시'로 분리되며, '가로방향수정자석_357'은 '가로 방향 수정 자석'으로 분리된다. 이러한 띄어쓰기 정보는 복합 명사 그 자체를 인식할 수 있을 뿐만 아니라, 추가적인 사전 참조 없이 간단히 복합 명사를 분리할 수 있어 시스템의 전체적인 처리 속도를 향상시킬 수 있다.

2.2 어절의 음절수에 따른 복합 명사 분리 방안

일반적인 띄어쓰기 오류는 완전히 맞춤법에 어긋난 것이 대부분으로 누구나가 그 단어를 지적만 하면 쉽게 고칠 수 있으나, 복합 명사의 경우 당연히 띄어쓰야 함에도 불구하고 관례나 편의상 붙여서 쓰는 경우가 많으며, 특히 교과서를 제외한 신문 기사나 책 등에는 어떠한 규칙도 없이 무분별하게 복합 명사를 사용하고 있다.

우리가 흔히 사용하는 복합 명사를 형태별로 분류하면 표 1과 같다.

<표 1> 복합 명사의 형태별 분류

복합 명사로 오인할 수 있는 단일명사	원칙은 분리 편의상 복합 명사도 가능	의미에 따라 여러개로 분리가능한 복합 명사	반드시 분리되어야 하는 복합 명사
국민교육헌장	정형외과	원자핵분열	버스정류장
지하철역	신경외과	세포핵분열	해운대해수욕장
월임대료	...	...	조선총독부
군위생병			공원입장료
...			...

표 1의 복합 명사 형태 분류에서 복합 명사 분리시의 어려운 점은 복합 명사이면서 띄어써야 하는 경우에 분리 모호성이 발생할 때와 단일 명사이면서 복합 명사로 분리되어지는 단어에 대한 처리이다. 본 시스템에서는 이러한 복합 명사 분리시의 문제점을 가능한 해결하면서도 어휘 사전 참조 횟수를 감소시키기 위하여 복합 명사의 음절 수에 따른 특성들을 고려한 복합 명사 분리 방안을 제시하고자 한다.

그림 2는 복합 명사 분리를 위한 본 시스템의 흐름도이다.

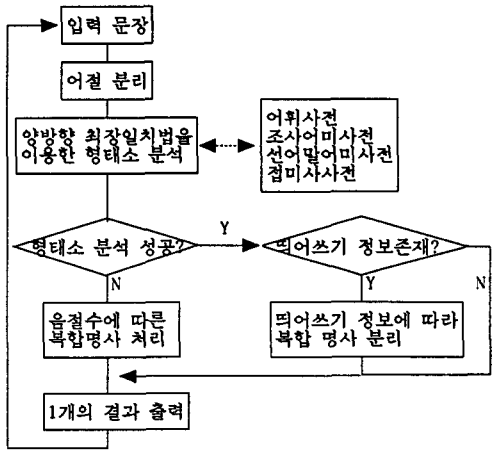

그림 2 전체적인 시스템 흐름도 (flowchart)

본 시스템은 입력 문장을 어절 단위로 분리하여 형태소 분석을 수행하는데 형태소 분석에 성공하고 어휘 사전에 검색한 단어의 띄어쓰기 정보가 존재한다면, 이 단어를 복합 명사로 추정하고 분리 작업을 행한다. 형태소 분석에 실패한 어절에 대해서는 조사 부분을 제외한 단어의 음절수에 따른 복합 명사 분리 루틴을 수행하고 1개의 분리 결과만을 출력한다.

본 시스템에서는 3음절 복합 명사에 대해서는 분리 작업을 행하지 않는다. 그 이유는 3음절의 복합 명사일 경우 3_0분리, 1_2분리 그리고 2_1분리 처리가 행해져야 하는데, 3음절 단어 자체가 사전에 존재하지 않을 경우 1_2 또는 2_1분리를 행하면 대부분의 단어가 분리되므로, 실제 3음절로 처리되어야 하는 명사까지도 분리되는 경우가 더 많이 발생하게 된다.

복합 명사의 효율적인 분리를 위하여 본 시스템의 어휘 사전에 수록된 14만 단어를 대상으로 4음절에서 12음절까지 음절수에 따른 복합 명사를 수집하고 2가지 이상으로 분리되는 복합 명사를 추출한 후, 이들에 대한 음절수에 따른 복합 명사 분리시의 특성을 분석하여 복합 명사 분리시 어휘 사전 참조를 행하는 순서에 대한 규칙을 제안하고 이 규칙을 적용함으로써 어휘 사전 참조 횟수를 감소시키면서 분리의 정확성을 보장하는 알고리즘을 고안하였다.

복합 명사 분리 규칙
5음절 이상의 어휘 형태소에 대해 좌방향 최장일치를 행하여 분리된 단어들이 동시에 어휘 사전에 존재할 때, 복합 명사를 분리된 단어들로 분리한다.

본 시스템은 복합 명사 분리의 정확성을 높이기 위하여 기존의 시스템에서는 거의 처리되지 않고 있는 접미사를 포함한 복합 명사에 대한 처리도 행한다. 예를 들어, '화훼판매장'이라는 5음절 복합 명사에서 '판매장'이 사전에 존재하지 않을 경우, 뒤 1음절 '장'을 접미사 처리하여 앞 4음절 '화훼판매'를 '화훼 판매'

로 분리한 후, 접미사 '장'을 붙여서 '화훼 판매장'으로 분리하게 된다.

2.2.1 4음절 복합 명사

표 2에 제시된 4음절 복합 명사의 분리 모호성 결과, 1_3분리와 3_1분리 모호성 단어의 대부분은 잘못 분리된 것으로, 이들 단어는 4_0분리, 즉 단일명사로 처리하는 것이 올바른 것으로 분석되었다. 따라서 5음절 이상의 복합 명사 분리시 사용하는 좌방향 최장일치법 분리 순서인 1_3, 2_2, 3_1분리의 순으로 하지 않고, 2_2, 3_1분리의 순으로 처리하되 1_3분리는 접두사 처리만 하는 것이 더 정확한 분리 결과를 얻을 수 있는 것으로 나타났다. 그러나 2_2분리가 옳은 경우가 809개로 2_2분리가 가능한 복합 명사 1348개의 60% 정도여서 이의 해결책이 4음절 복합 명사 분리 시의 가장 큰 문제점으로 나타났다. 본 시스템에서는 어휘 사전의 띄어쓰기 정보에 의해 사전에 수록된 단어에 대해서는 정확한 분리가 가능해 어느 정도는 해결 가능 하다.

<표 2> 4음절 복합 명사의 분리 모호성

	1_3, 2_2, 3_1 4_0 분리	1_3, 2_2 4_0 분리	2_2, 3_1 4_0분리	1_3, 3_1 4_0분리
총 단어수	66	617	665	93
4_0분리 맞는 경우	43	279	213	89
1_3분리 맞는 경우	0	4	0	1
2_2분리 맞는 경우	23	334	452	0
3_1분리 맞는 경우	0	0	0	3

3_1분리 시는 접미사 처리를 먼저한 후 복합 명사에 대한 분리를 행한다. 이를 위해 주기억 장치에 저장된 접미사 사전을 먼저 검색하여 접미사가 존재하지 않을 경우, 어휘 사전을 검색하게 된다. 그러나 5음절 이상의 복합 명사 분리시에 4음절 처리 루틴을 호출할 경우는 접미사 처리를 하지 않는 것이 실험 결과 더 높은 정확성을 나타내어 접미사 처리를 하지 않는다. 4음절 복합 명사의 처리 과정에서 사전에 존재하지 않는 단어에 의해 복합 명사 분리가 이루어지지 않을 경우는 default로 4음절 그대로 출력한다. 다음은 조사한 표 2를 기초로 4음절 복합 명사를 분리하는 알고리즘이다.

```
<4음절 복합 명사 분리 알고리즘 >
int devide_bokhap_noun_4(unsigened word1[])
/* F[1]: 앞 1음절, F[2]: 앞 2음절, F[3]: 앞 3음절,
   B[1]: 뒤 1음절, B[2]: 뒤 2음절, B[3]: 뒤 3음절이 저장된 배열 */
{
 if (success_search_dic(F[4])
    어휘 사전 띄어쓰기 정보에 따라 분리;
 else if (success_search_dic(F[2]) && success_search_dic(B[2]))
        if (B[2] == '주의', '나무', '자리')  4_0분리;
        else 2_2분리;
 else if (success_search_dic(F[3])) {
        if (B[1] == 접미사)  4_0분리;
        else if (success_search_dic(B[1])) 3_1분리;
      }
 else if (success_search_dic(B[3]) && F[1] == 접두사)  4_0분리;
 else 4_0분리; /* default */
}
```

위 알고리즘에서 함수 success_search_dic(F[2])는 앞 2음절을 가지고 어휘 사전을 검색하여 검색에 성공하면 true를 실패하면 false를 반환하는 함수이다.

위 알고리즘을 적용하여 신문 기사에서 발췌한 4음절 복합 명사에 대한 처리 결과는 표 3과 같다.

<표 3> 4음절 복합 명사 처리 결과

처리 단어 수		236
올바른 분리 단어 수		211 (89.41%)
잘못된 분리 단어 수		7 (2.96%)
default 처리 단어의 수	옳은 경우	3 (1.27%)
	틀린 경우	15 (6.35%)
# 잘못 분리된 예 황산 화물, 피아 노사, 재활 용품, 군위 생병, 신시 가지, 남포 동점, 알짜 배기		

표 3에서 default 처리가 틀린 경우는 분리될 단어가 어휘 사전에 존재하지 않아 분리되지 않고 4음절 단어 그대로 출력되는 경우로, 만약 어휘 사전에 단어가 수록될 경우는 정확한 분리가 가능하다. 4음절 복합 명사 처리 결과 약 91% 정도의 정확한 분리가 이루어졌으며 어휘 사전에 단어 추가시는 97%까지의 정확한 분리가 가능한 것으로 나타났다.

2.2.2 5음절 복합 명사

5음절 복합 명사는 좌방향 최장일치를 적용하여 5_0, 2_3분리 그리고 3_2분리 순서로 처리하며 1_4분리와 4_1분리는 처리하지 않는다. 이는 5음절 이상의 복합 명사에서 첫음절과 끝음절이 분리될 가능성이 거의 없기 때문이다. 그러나 뒤 1음절에 대한 접미사 처리는 행한다. 14만 어휘 사전 단어에 대한 2_3분리와 3_2분리 시의 분리 모호성에 대한 결과는 표 4와 같다.

<표 4> 5음절 복합 명사 분리의 모호성

분리 모호성 단어 수	84개
2_3분리가 옳은 단어 수	46개 (54.8%)
3_2분리가 옳은 단어 수	38개 (45.2%)

표 4에서 2_3분리의 옳은 경우가 3_2분리의 옳은 경우보다 더 많아 5음절 복합 명사 분리시는 좌방향 최장 일치의 순서인 2_3, 3_2 분리의 순으로 분리를 행하게 되나, 3_2분리가 옳은 경우 또한 45%를 차지하므로 어휘사전에 존재하지 않는 5음절 단일 단어(신한국주의, 옥수수나무, 오리온자리 등))의 분리 방지와 3_2분리 단어(관형격조사, 영문학교수, 이력서양식, 외국어학자 등)의 잘못된 2_3분리를 방지하기 위하여 이러한 단어들을 조사하여 절차적(procedural)으로 처리하였다. 다음은 조사한 표 4를 기초로 5음절 복합 명사를 분리하는 알고리즘이다.

```
int devide_bokhap_noun_5(unsigned word1[])
{
 if (success_search_dic(F[5])  사전 띄어쓰기 정보에 따라 처리;
 else  if (B[2] == '주의', '나무', '자리')
        if (success_search_dic(F[3]))  5_0처리;
 else if (success_search_dic(F[2]) && (success_search_dic(B[3]))
```

```
        if (B[3] == '격조사', '학교수', '서양식, 어학자') &&
           (success_search_dic(F[3]) && (success_search_dic(B[2]))
              3_2분리;
        else  2_3분리;
else if (success_search_dic(F[3]) && success_search_dic(B[2]))
           3_2분리;
else if (B[1] == 접미사) {
        devide_bokhapN4(F[4]);
        if (devide_success)   attach the B[1] to the last devided
              word;
    }
else  5_0분리;  /* default 처리 */
}
```

위 알고리즘에서 변수 devided_success는 단어의 음절별 분리가 성공하면 true를 실패하면 false를 값으로 가지는 boolean 변수이다. 위 알고리즘을 적용하여 신문 기사에서 발췌된 5음절 복합 명사에 대한 처리 결과는 표 5와 같다.

<표 5> 5음절 복합 명사 처리 결과

처리 단어 수		200
올바른 분리 단어 수		156 (78%)
잘못된 분리 단어 수		0 (0%)
default 처리 단어 수	옳은 경우	5 (2.5%)
	틀린 경우	39 (19.5%)

5음절 복합 명사에서 접미사로 처리된 단어의 수는 21개의 단어로 모두 올바르게 처리되었다. 5음절 복합 명사 처리 결과 약 80.5%정도의 정확성을 보장하며 어휘 사전에 수록되지 않아서 default 처리된 단어를 어휘 사전에 추가한다면 실험 대상 단어에 대해서는 100%까지의 정확한 분리가 가능한 것으로 나타났다.

2.2.3 6음절 이상의 복합 명사

N(N>=6) 음절 복합 명사는 좌방향 최장일치를 적용하여 N_0, 2_N-2, 3_N-3, ... , N-3_3, N-2_2 분리 순서로 처리한다. 6음절 이상의 복합 명사에서의 분리 모호성은 분석 결과 대부분이 복합 명사내의 4음절과 5음절에 대한 분리 모호성만 나타나므로 6음절 이상의 복합 명사에 대해서는 고려하지 않아도 되는 것으로 나타났다. 14만 어휘 사전 단어에 대한 복합 명사 분리의 결과를 나타낸 표 6에 따르면 6음절 복합 명사는 2_2_2분리가, 7음절 복합 명사는 3_2_2 분리가, 8음절 복합 명사는 2_2_2_2 분리가 많은 비중을 차지하지만, 분리시 대부분의 단어가 4음절과 5음절 복합 명사로 어휘 사전에 띄어쓰기 정보와 함께 수록되어 있어 추가적인 어휘 사전 참조 없이도 정확한 분리가 가능하게 된다. 만약 우방향 최장일치를 적용하여 N-2_2 분리를 먼저 행하면, 예 '응용프로그램/소프트웨어/데이타베이스'는 '그램/웨어/베이스'가 어휘 사전에 존재하므로 '응용 프로 그램/소프트 웨어/데이타 베이스'와 같은 잘못 분리된 결과를 얻게 된다. 표 6은 14만 어휘 사전에 있는 6-8 음절 복합 명사에 대한 분리 결과이다.

<표 6> 6-8 음절 복합 명사 분리 결과

6음절 단어수 : 1050	7음절 단어수 : 324	8음절 단어수 : 241	
2_4분리 : 97	2_5분리 : 0	2_6분리 : 1	3_2_3분리:0
3_3분리 : 87	3_4분리 : 36	3_5분리 : 1	6_2분리 :1
4_2분리 : 15	4_3분리 : 5	4_4분리 : 9	4_2_2분리:1
2_2_2분리 : 851	2_2_3분리 : 11	2_2_4분리:13	2_4_2분리:1
	5_2분리 : 1	5_3분리 : 1	3_3_2분리:1
	2_3_2분리 : 3	2_3_3분리: 1	
	3_2_2분리 : 268	2_2_2_2분리: 210	

다음은 조사한 표 6을 기초로 N 음절(N>=6) 이상의 복합 명사를 분리하는 알고리즘이다.

```
int devide_bokhap_noun_N(unsigned word1[])
{
  if (success_search_dic(F[N])
    사전 띄어쓰기 정보에 따라 처리;
  else {
    for (i=2; i<=N-2; i++) {
      if ((i==4) || (i==5) || (i==6)) {
        if (success_search_dic(B[N-i])
          if (success_search_dic(F[i])
            i_(N-i)처리;
          else {
            devide_bokhap_noun_i(F[i]);
            if (devide_success) i_(N-i)처리;
          }
      }
      else if (success_search_dic(F[i]) &&
            (success_search_dic(B[N-i]))
             i_(N-i)처리;
    }
    if ((devide_success==0) && (B[1] == 접미사)) {
      (devide_bokhap_noun_N-1(F[N-1]));
      if (devide_success)
        attach the B[1] to the last devided word;
      else N_0처리; /* default 처리 */
}
```

위 알고리즘에서 'i_(N-i)처리'는 복합 명사를 i와 N-i로 분리하되, 앞 i 음절에 대해서는 어휘 사전을 검색할 경우 어휘 사전의 띄어쓰기 정보에 따라 분리하고, 복합 명사 분리 루틴을 호출할 경우 그 루틴의 실행 결과에 따라 분리하고, 만약 분리에 성공하면 알고리즘을 종료하라는 의미의 함수이다.

본 알고리즘을 적용하여 신문 기사에서 발췌된 6/7/8 음절 복합 명사에 대한 처리 결과는 표 7과 같다.

6/7/8 음절 복합 명사 분리 시 접미사 처리를 한 단어 수는 각각 3개, 17개, 3개이며, 이 중 잘못된 처리는 7음절이 2개로 '유명 피서 지설치'와 '신세 대결 혼비용'이며, 8음절이 1개로 '올림픽 개막 식열창' 나타났지만, 분석 결과 앞 단어 중 한 개(피서지, 신세대, 개막식)가 어휘 사전에 존재하지 않음으로써

<표 7> 6/7/8 음절 복합 명사 처리 결과

처리 단어 수		6음절: 124	7음절: 119	8음절: 78
올바른 분리 단어 수		103 (83.1%)	87 (73.1%)	63 (80.8%)
잘못된 분리 단어 수		4 (3.2%)	3 (2.5%)	2 (2.6%)
default 단어 수	옳은 경우	0	1 (1.7%)	0
	틀린 경우	17 (13.7%)	28 (23.5%)	13 (16.7%)
# 잘못 분리된 예 6음절: 최종 가대 처분, 그린 벨트 지역, 온천 장취 급소, 독일 식생 맥주 7음절: 체형 교정란 제리, 방사 선암 치료제, 유명 피서 지설치 8음절: 상속세 법전면 손질, 신세 대결 혼비용				

발생한 것으로, 이들 단어의 추가시 실험 결과 모두 올바르게 분리되었다.

복합 명사 분리가 잘못 처리된 경우는 '그린벨트'와 '란제리'와 같이 어휘 사전에 단어가 수록되지 않아 발생한 경우와, '최종가, 온천장, 독일식'과 같이 3_3분리시 앞 3음절에 접미사가 붙은 경우, 그리고 '방사선암'과 '상속세법전면'과 같이 좌방향 최장일치를 행함으로써 잘못된 분리가 발생한 경우이다. 즉 '방사선암'인 경우 3_1분리 전에 2_2분리를, '상속세법전면'의 경우 4_2분리 전에 3_3분리를 먼저 행함으로써 잘못된 분리가 발생하였다. 이 중 앞 3음절에 대한 접미사 처리는 또 다른 분리 모호성을 야기하게 되므로 본 시스템에서는 이를 처리하지 않도록 하였으며, 궁극적인 해결책으로는 이러한 단어들을 어휘 사전에 추가할 수밖에 없다. 좌방향 최장일치의 적용으로 인한 잘못된 분리는 현재의 본 시스템에서는 해결이 불가능하다.

default로 잘못 처리된 것은 6/7/8 음절 모두 분리되어야 하는 단어가 어휘 사전에 수록되지 않아 발생한 것으로, 이들 단어를 어휘 사전에 수록할 경우 실험 결과 올바른 분리가 가능하였다.

6/7/8 음절 복합 명사의 처리 결과 약 83%/75%/81% 정도의 정확성을 보였으며, 어휘 사전에 단어 추가시는 약 98%까지의 정확한 분리가 가능한 것으로 나타났다. 9음절, 10음절, 11음절, 12음절 복합 명사는 14만 어휘 사전 조사 결과 각각 57개, 19개, 9개, 6개가 조사되었으나, 이들은 대부분 전공 용어를 붙여 쓴 경우이며 실험 대상 문장에서는 거의 나타나지 않아 실험 결과를 제시하지 않았다.

IV. 성능 평가 및 고찰

기존에 상용화되어 있는 워드프로세서 중에서 가장 대표적인 h 워드프로세서를 이용하여, 본 시스템의 성능을 평가하기 위해 사용했던 4음절에서 8음절까지의 복합 명사에 대한 처리 결과와 본시스템의 처리 결과를 표 8에서 제시한다.

표 8에서 757개의 복합 명사에 대한 분리 실험 결과 본 시스템은 약 83% 정도, h 워드프로세서는 약 75%의 분리 정확성을 보였다. 만약 본 시스템의 어휘 사전에 단어를 보완 수록할 경우, 실험 결과 약 98.4% 정도까지 분리 정확성을 높일 수 있었다. 이는 본 시스템이 사전의 어휘량에 의존적이며, 본 논문이 제안한 음절 수에 따른 복합 명사 분리 방안 자체는 상당히 효율적임을 알 수 있다.

<표 8> 본 시스템과 h워드프로세서와의 복합 명사 처리 결과

복합명사 음절 수	처리 시스템	올바른 분리	잘못된 분리	default 처리	
				옳은경우	틀린 경우
4음절	본 시스템	89.4%(211)	3.0%(7)	1.3%(3)	6.4%(15)
(236)	h워드	86.6%(204)	3.4%(8)		10.0%(24)
5음절	본 시스템	77.5%(156)	0%	2.5%(5)	19.5%(39)
(200)	h워드	70.5%(141)	2.0%(4)		27.5%(55)
6음절	본 시스템	83.1%(103)	3.2%(4)	0%	13.7%(17)
(124)	h워드	64.5%(80)	14.5%(18)		21.0%(26)
7음절	본 시스템	73.1%(87)	2.5%(3)	1.7%(1)	23.5%(28)
(119)	h워드	67.2%(80)	10.1%(12)		22.7%(27)
8음절	본 시스템	80.8%(63)	2.7%(2)	0%	16.7%(13)
(78)	h워드	75.6%(59)	6.4%(5)		18.0%(14)
총결과	본 시스템	81.9%(620)	2.1%(16)	1.2%(9)	12.8%(97)
(757)	h워드	74.5%(564)	6.2%(47)		19.3%(146)

h 워드프로세서에서 잘못 처리되는 경우는 복합 명사가 사전에 존재하여 복합 명사의 분리 작업을 행하지 않고 단일 명사로 처리하는 경우가 많은 것으로 추정된다. 그리고 사전에 수록되지 않은 미등록 분리 명사에 대해 h 워드프로세서에서는 바로 default로 처리하지만, 본 시스템에서는 접미사 처리를 행한후 default 처리하므로 미등록어에 대해서 본 시스템의 분리 정확성이 조금 더 높게 나타났다. 또한 h 워드프로세서의 복합 명사의 처리 방법은 우방향 최장일치를 적용하여 우선적으로 찾아지는 복합 명사 분리에 대해서는 바로 분리해 버린다. 즉 4음절 복합 명사는 4_0분리와 2_2분리만을 행하고, 5음절 복합 명사는 5_0분리, 3_2분리, 2_3분리의 순으로 처리하며 3_2분리와 2_3분리의 모호성이 있는 복합 명사에 대해서도 우방향 최장일치를 적용하여 무조건 3_2분리로 처리한다. 6음절이상의 복합 명사에 대해서도 동일한 방법을 사용하고 있다.

h 워드프로세서에서 잘못 분리된 예와 default 처리된 예를 표 9에서 제시한다.

<표 9> h 워드프로세서의 잘못 처리된 예와 default 처리된 예

복합 명사 음절수	잘못 처리된 예
4음절 복합 명사	황산 화물, 재활 용품, 남포 동점, 바젠 세일, 한국 주의, 신용카드, ... default: 건재상사, 미용타운 꽃판매장, ...
5음절 복합 명사	신자본 주의, 플라톤 주의, 신보 수주의, ... default: 변형근로제, 무료발송권, 활어판매장, 상수원개발, 버스료인상, ...
6음절 복합 명사	정전 기량 분석, 전자계 산학과, 보조장부 작성, 초염가서비스, ... default: 김영삼대통령, 전립선치료기, 농산물판매제, 쓰레기매립장, ...
7음절 복합 명사	법정전염병 지정, 초특급 문제소설, 공화당 전당대회, ... default: 항만시설이용료, 여름인기상품전, 자율절전요금제, ...
8음절 복합 명사	혈액순환 장애 예방, 국제영화제 주인공, ... default: 대통령결선투표제, 아동복실속만족전, ...

표 9에서 제시된 '황산화물'과 '남포동점'을 제외한 나머지 예에 대해서 본 시스템에서는 바르게 처리된다. 또한 표 9의 default에서 '변형근로제'처럼 밑줄 부분은 본 시스템에서 접미사 처리를 행하여 올바르게 처리가 되는 예이다.

V. 결론

한국어 맞춤법 오류에서 가장 많은 빈도 수를 차지하는 것이 띄어쓰기 오류인데, 이러한 띄어쓰기 오류 중에서도 복합 명사에 대한 띄어쓰기 오류가 대부분을 차지한다. 본 논문은 복합 명사의 띄어쓰기 오류를 효율적으로 교정하기 위해 어휘 사전에 띄어쓰기 정보를 어휘와 함께 수록함으로써 어휘 사전 참조 횟수를 감소시켜 시스템의 처리 속도를 향상시키는 방안과 복합 명사 분리의 정확성을 높이기 위하여 복합 명사의 음절 수에 따른 분리 특성을 조사하여 음절 수에 따른 분리 순서와 이를 적용하여 1개의 분리 결과만을 출력하는 복합 명사 분리 알고리즘을 제안하였다.

실험 결과 본 시스템은 약 83% 정도의 분리 정확성을 보였으며, 이는 기존의 시스템에 비해 약 8% 정도 정확성이 높은 것으로 나타났으며, 본 시스템의 어휘 사전에 수록되지 않은 단어들을 추가할 경우 약 98.4%까지의 정확성을 보장받을 수 있음을 실험적으로 알 수 있었다.

본 시스템은 한국어 키워드 추출 시스템 및 한국어 맞춤법 교정 시스템의 복합 명사 처리 루틴으로 사용될 수 있으며, 앞으로 잘못된 분리에 대한 데이터를 수집하여 100%까지의 분리 정확성을 높일 수 있도록 보완할 예정이다.

VI 참고 문헌

[1] 심철민, 김현진, 김영진, 권혁철, "연어 정보를 이용한 한국어 철자 검사기의 기능 개선," 제7회 한글 및 한국어 정보처리 학술대회 발표 논문집, pp.86-90, 1995.

[2] 이병훈, 윤준태, 송만석, "말뭉치를 기반으로 한 한국어 철자 교정기의 구현," 제5회 한글 및 한국어 정보처리 학술대회 발표 논문집, pp.285-293, 1993.

[3] 정한민, 이근배, 이종혁, "자판 특성을 이용한 Neuro-Fuzzy 한국어 철자 교정기의 구현," 제5회 한글 및 한국어 정보처리 학술대회 발표 논문집, pp.317-328, 1993.

[4] 미승우, "새 맞춤법과 교정의 실제," 어문각, 1990.

[5] 최재혁, "양방향 최장일치법에 의한 한국어 형태소 분석기의 구현," 경북대학교 공학박사학위 논문, 1993.

[6] 최재혁, "형태소 분석을 통한 자동 키워드 추출 시스템," LG정보통신(주) 최종연구보고서, 1995.

[7] 김정숙, "정보사회의 우리글 바른표기," 일진사, 1994.