한국어 복합명사 분해 오류 교정 기법

강민규 ○ 강승식
국민대학교 컴퓨터공학부
pelcious@kookmin.ac.kr, sskang@kookmin.ac.kr

Error Correction Method for Korean Compound Noun Decomposition

Minkyu Kang○ Seungshik Kang
School of Computer Science, Kookmin University

요 약

복합명사의 구성요소로 미등록어, 1음절어, 접사 등이 포함된 경우에 복합명사 분해기의 분해 결과가 분해중의성을 보이게 된다. 특정 복합명사에 대한 분해 결과가 잘못된 것일 경우, 이를 분해 오류로 판단하고, 재처리과정을 통해 교정해야 한다. 본 논문에서는 복합명사의 분해 결과에서 분해 오류에 대하여 각 구성명사의 빈도 정보를 통해서 오류 여부를 판단하고, 적절한 재분해 결과를 제공하여 분해 오류를 교정하는 방법을 제안한다.

1. 서 론

음절 길이에 따른 분해 패턴을 적용하는 방식[2, 3, 5, 6]의 복합명사 분해 방법에서 나타나는 문제점 중 하나가 분해중의성이다. 분해중의성을 갖는 복합명사의 분해 결과가 잘못된 분해일 경우, 이를 분해 오류로 판단하고 재처리를 통해 정확한 분해 결과를 보일 수 있도록 해야 한다.

분해중의성을 갖는 복합명사 분해 오류의 형태와 그 특징을 분석해서 분해 오류의 패턴을 알아낸다면 그에 맞춰진 오류 교정 방법을 만들어 각 정보처리시스템에 더 정확한 복합명사 분해 결과를 제공할 수 있다.

분해중의성을 갖는 복합명사 분해 결과의 형태는 크게 두가지로 살펴 볼 수 있다.

첫째, 고유명사, 외래어의 한국어 표기, 신조어 등 복합명사 분해기의 명사사전에 등록되지 않은 미등록어가 포함된 표현인 경우이다.

복합명사 내부의 미등록 구성명사는 미등록어 부분을 따로 분해할 수도 있지만, 다른 명사로 오인될 경우 오인된 형태로 분해되기 때문에 원래의 명사형태를 유지할 수 없다. '박정희대통령'의 경우 미등록 고유명사 '박정희'가 미등록어로 분해되어야 하지만 명사사전에 등록된 '박정', '희대'로 분해되기 때문에 분해중의성을 갖는 잘못된 분해 결과가 나타난다.

미등록어 문제의 경우, 모든 명사를 사전에 등록하는 방법[6]이 있으나, 일반 명사에 비해 출현빈도가 매우 낮은 고유명사 전체를 사전에 유지하는 비용의 낭비에 비해서 전체적인 시스템의 성능 향상의 정도가 낮기 때문에 좋은 방법이라고 볼 수 없다.

둘째, 1음절 접사가 포함된 형태의 명사를 사용하는 복합명사의 경우, 1음절어 처리를 하지 않고 분해과정이 진행되기 때문에 분해중의성을 갖게 된다.

한국어는 1음절로 의미를 갖는 한자어의 사용량이 많지만, 실제 명사 표현은 2음절 이상이 많기 때문에 복합명사 분해 알고리즘에서 음절길이에 따른 주요 패턴은 1음절 패턴을 무시하고 2/3음절 패턴을 중시하게 된다.[2] 이로 인해 1음절 접사가 포함된 복합명사들은 분해중의성을 보인다. '부-위원장', '정치-경제-학-적' 등의 복합명사를 살펴보면, 1음절 패턴 '부', '학', '적'을 무시하고 2음절 패턴으로 분해 가능한 경우를 먼저 판단하여 분해 결과가 '부위-원장', '정치-경제-학적'으로 나타난다.

분해중의성은 그 원인이 어떤 것이든지 기본적으로 "내부에 다른 분해 가능성이 존재한다."라는 공통점을 찾아볼 수 있다. 그리고 그 결과물은 다음과 같은 두가지 특징을 갖는다.

1. 명사가 아닌 분해 결과 혹은 실제 사용빈도가 매우 낮은 명사가 구성명사로 사용된다.
2. 분해된 구성명사 간 조합이 올바르지 않은 형태를 보인다.

본 논문에서는 분해중의성을 보이는 복합명사 분해 오류를 해결하기 위한 방법으로 복합명사를 구성하는 각 단위명사의 빈도 정보를 통해 분해 오류를 판단하고, 발견된 오류를 재구성하여 오류를 교정하는 방법을 제시하고자 한다. 2장에서는 복합명사를 구성하는 각 단위명사들의 빈도 정보를 추출하는 방법을 설명하고, 3장과 4장에 걸쳐 복합명사의 분해 오류를 탐지 후,

교정하는 방법에 대해 설명한다.

2. 복합명사 통계자료 추출

분해중의성을 보이는 복합명사 분해 결과를 통해 나타나는 분해 오류는 복합명사를 구성하는 각각의 단위명사가 저빈도 특성을 보이거나, 단위명사 간의 연관성이 떨어지는 특성을 보이게 된다. 따라서 분해 결과의 오류 유무를 판단하는 척도로 각 구성명사의 빈도 혹은 구성명사 간 연관성을 나타내는 빈도 값을 미리 계산해서 분해 오류 탐지 및 교정에 사용하는 것이 바람직하다고 보여진다.

본 논문에서 제시하는 방법을 실험하기에 앞서 각 구성명사의 사용 빈도와 구성명사 간의 연관성을 판단하기 위해 구성명사의 단일 빈도와 bigram 공기 빈도를 계산하여 빈도 사전을 작성한다. 빈도 정보의 계산은 형태소 분석기의 분석 결과가 명사(N), 복합명사(C), 미등록어(K)인 단어들을 모두 수집해서 빈도계산 프로그램을 통해 빈도 계산을 하는 단순한 형태로 진행한다.

빈도 정보를 수집하기 위해 사용된 원시 말뭉치는 다음과 같다.

- 세종 원시 말뭉치
  - 21세기 세종 계획(2007) 현대 문어 원시 말뭉치 - 국립국어원
  - 어절 수 : 6700만 어절[1]
- 한국일보 2년치 기사 (1998-1999)
  - 한국일보 문서범주화 실험문서집합 - 한국과학기술정보연구원
  - 어절 수 : 731만 어절
- 한겨레신문 1년치 기사 (2002)
  - 자체 수집
  - 어절 수 : 543만 어절

2.1 명사 unigram 빈도 자료 추출

단일 명사의 빈도 자료는 단일 명사 빈도와 함께 구성명사 위치 빈도를 수집한다. 단일 명사 빈도는 명사가 아닌 분해 결과와 저빈도를 갖는 분해 결과 탐지에 사용되며, 위치 빈도는 bigram 공기 빈도와 함께 각 구성명사의 사용이 적절한가에 대한 판단 기준으로 사용한다.

복합명사 n1n2에서 n1과 n2는 술어-보어, 목적어-서술어 등 일정한 관계를 갖게 되고, 특정 위치에 나타나는 구성명사는 그 위치에 해당하는 역할 만을 맡게 된다. 따라서 자주 등장하지 않는 위치에 나타난 구성명사는 원래 역할에 맞지 않는 사용이므로 분해 오류의 가능성이 높다고 판단할 수 있다.

[1] 세종 원시 말뭉치 중, 현대 문어 말뭉치 만을 사용하였다.

복합명사의 위치 정보는 크게 prefix(처음), infix(중간), suffix(끝)으로 분류하고, 복합명사 내에서 각 위치에 나타나는 구성명사를 B(begin), M(middle), E(end)의 태그로 구분하여 각각 저장해서 위치 빈도를 계산하기 위한 데이터를 수집한다.

그림 1은 단일 명사 빈도와 위치 빈도를 포함한 단일 명사 빈도 사전을 작성하는 예이다. 형태소 분석기에서 추출된 띄어쓴 복합명사 데이터와 단일 명사 데이터를 각각 빈도 계산 후, 띄어쓴 복합명사 데이터(구분 기호: SS, space-space)로부터 각 구성명사를 위치 정보별로 B/M/E로 구분해 추출한다. 이렇게 추출된 단일 명사 빈도와 위치 빈도를 병합해서 단일 명사 빈도 사전에 저장한다. 빈도의 저장 순서는 “단일 빈도 / B빈도 / M빈도 / E빈도” 순으로 단일 명사 ‘문화’의 경우, 단일 명사 35,550회, B위치 2,698회, M위치 9,745회, E위치 3,414회의 등장 빈도를 보인다.

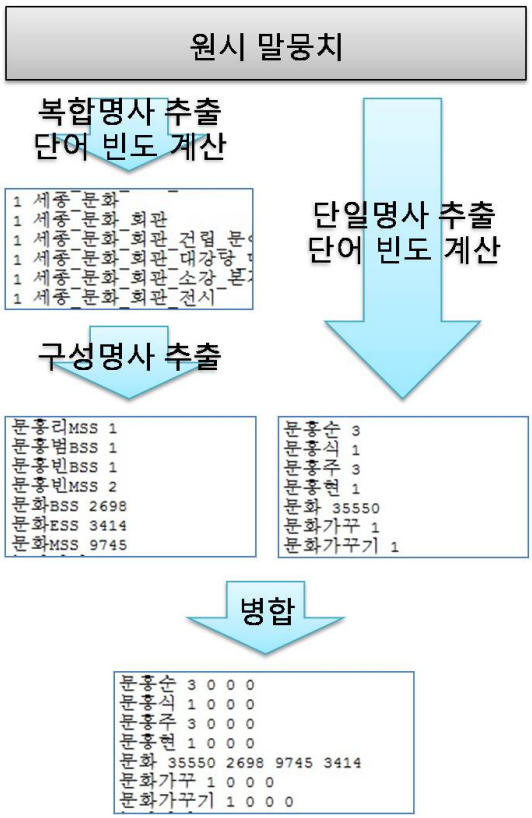

그림 1 단일 명사 빈도 추출

2.2 명사 bigram 빈도 자료 추출

구성명사와 구성명사 간의 연관성을 의미론적으로 판단하는 것은 힘들지만 두 구성명사의 공기 빈도를 통해서 간접적으로 연관성을 계산하는 것은 가능하다. 따라서 두 구성명사의 연관성을 판단하기 위한 자료로 bigram 공기 빈도 자료를 수집한다.

복합명사는 띄어쓰기와 붙여쓰기를 모두 허용하기 때문에 붙여쓴 복합명사 표현이 띄어쓰기로 표현될 가능성이 분명히 존재한다. 띄어쓴 복합명사는 분해

오류에 대한 위험도 존재하지 않기 때문에 복합명사의 분해된 구성명사들이 연관성을 갖고 있는지 판단하는 자료로 사용할 수 있다.

띄어쓴 복합명사의 bigram 공기 빈도수를 수집하는 작업은 단일 명사 사전을 작성하는 과정에서 특정 단일 명사들을 가공하는 방식으로 진행한다. 단일 명사를 수집할 때, 하나의 단일 명사와 그 다음에 검출되는 단일 명사 사이의 관계는 두가지 경우 중 하나로 볼 수 있다.

1. 단일 명사 사이에 white space만 존재
   예) … 세종 문화 회관 …
2. 단일 명사 사이에 다른 품사나 기호가 존재
   예) … 세종과 문화 …

두가지 경우 중, 1의 경우는 명사와 명사 사이에 다른 기호나 품사가 없으므로 띄어쓴 복합명사로 판단할 수 있다. 이러한 경우에 해당하는 형태소 분석 결과들을 따로 수집하여 저장하고, 빈도계산 프로그램으로 빈도값을 계산한다.

원시 말뭉치로부터 수집된 띄어쓴 복합명사를 그림 2와 같이 2어절 단위로 분해해서 bigram 데이터를 만들고 빈도 정보를 포함한 합병과 정렬을 실시해서 띄어쓴 복합명사의 bigram 빈도 사전을 작성한다.

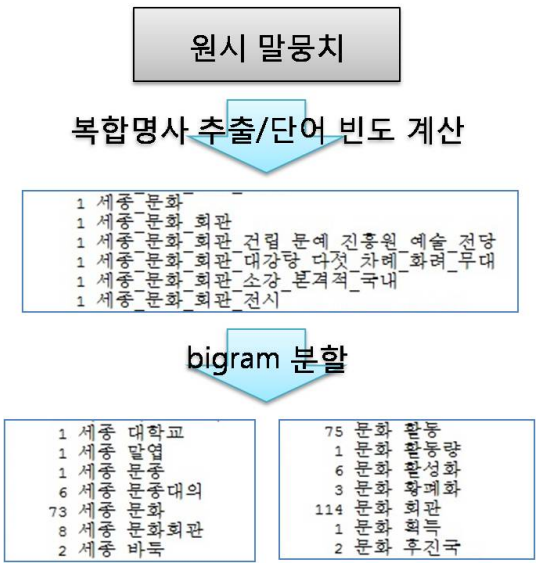

그림 2 띄어쓴 복합명사의 bigram 공기 빈도 추출

3. 복합명사 분해 오류 탐지 방법

원시 말뭉치로부터 추출된 구성명사의 빈도정보를 사용하여 복합명사의 분해 오류를 탐지하는 과정은 [1]에서 제시된 방법에 따라 진행된다. 그림 3은 분해 오류를 탐지하는 방법으로 복합명사 '박정희대통령'의 분해 결과에 대한 오류 탐지 과정을 도식화한 것이다.

구성명사 n_1, n_2, n_3, …, n_n으로 구성된 복합명사 N이 주어졌을 때, N이 분해 오류임을 판단하는 기준점은 각 구성명사 n_1, n_2, n_3, …, n_n의 분해 오류 여부이다. 다른 구성명사들이 정상적인 분해라고 하더라도 하나의 구성명사가 분해 오류로 판단될 경우, 복합명사 분해가 잘못되었다고 판단할 수 있다.

각 구성명사의 분해 오류 여부는 앞서 설명한 두가지 방법인 공기 빈도를 사용한 방법과 단일 빈도를 사용한 방법을 통해서 판단한다.

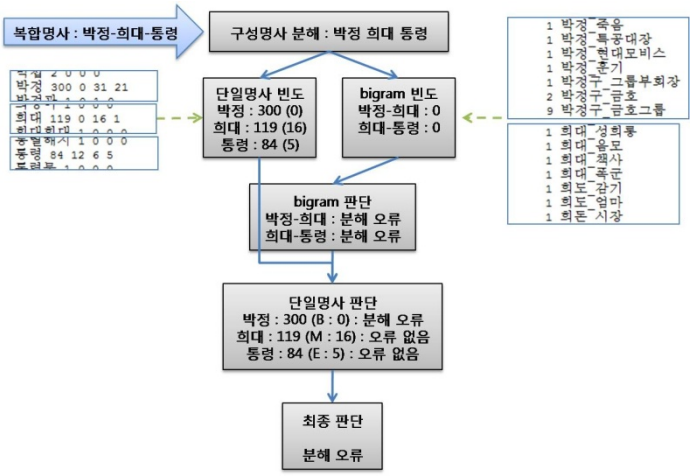

그림 3 '박정-희대-통령'의 분해 오류 탐지(임계값 = 5)

두 구성명사의 공기 빈도는 "두 구성명사가 연접해서 사용된 적이 있는가"에 대한 판단 척도로 사용한다. "잘 분해된 복합명사의 연접한 두 구성명사는 띄어쓴 복합명사 표현에도 나타난다."라는 가정에 따라, 띄어쓴 복합명사의 bigram 공기 빈도가 기준 임계값 이상일 경우 올바른 분해로 판단하고, 임계값 이하일 경우 분해 오류로 판단한다.

'박정희대통령'의 분해 결과인 '박정-희대-통령'의 경우, '박정'과 '희대', '희대'와 '통령'의 연관성 판단을 위해 각각의 bigram 공기 빈도를 분석한다. '박정-희대', '희대-통령' 모두 bigram 공기 빈도가 0이므로 공기 빈도를 사용한 방법에서는 분해 오류로 판단할 수 있다.

하지만 bigram 공기 빈도 정보를 이용한 구성명사의 분해 신뢰도 계산은 항상 옳은 값을 주지 않는다. 복합명사의 표현법이 띄어쓰기와 붙여쓰기 모두를 허용하지만, 띄어쓴 표현이 존재하지 않을 가능성이 있기 때문에 "잘 분해된 구성명사는 띄어쓰기 표현에도 나타난다."는 가정이 거짓일 수도 있다. 이러한 경우 띄어쓰기 bigram 공기 빈도에 의존한 분해 신뢰도 만을 사용할 수 없다. 따라서 bigram 공기 빈도와 함께 단일 명사 빈도도 함께 판단한다.

단일 명사 빈도는 단순 출현 빈도와 위치별 출현 빈도를 사용, 분해 오류를 두 번 판단한다. 각각의 빈도 중 하나의 빈도가 임계값보다 작은 경우, 이를 분해 오류로 판단한다.

'박정-희대-통령'을 공기 빈도로 판단한 경우는 분해 오류로 판단되었다. 단일 빈도의 경우, '박정', '희대',

'통령' 모두 단순 출현 빈도는 임계값 이상(박정:300, 희대:119, 통령:84, 임계값:5)이지만, '박정'의 begin 빈도가 0(희대의 M빈도 16과 통령의 E빈도 5는 오류가 없다고 판단됨)으로 분해 오류 판단을 내리게 된다.

결국, bigram과 단일 빈도 모두 분해 오류로 판단, 최종적으로 복합명사 '박정-희대-통령'의 분해 오류를 탐지하게 된다.

4. 복합명사 분해 오류 교정 방법

복합명사의 분해 오류가 발생하는 주된 원인은 분해 결과가 분해중의성을 갖는 경우로 그 주요한 형태는 '미등록어가 포함된 분해 오류', '1음절어 처리가 누락된 상태에서의 분해 오류', '복합명사와 조사가 결합된 상태로 분해된 경우'로 볼 수 있다. 이러한 경우들은 1음절로 의미를 갖는 단어, 즉 한자어와 같은 단어들이 좌우 어느 쪽과 결합해도 의미를 갖기 때문에 발생한다.

- 미등록어 포함 복합명사 : 박정희대통령
  분해 오류 발생 지점 : 박정-희대
  문제가 되는 1음절어 : 희
  분해 가능 형태 : 박정희-대 / 박정-희대
- 1음절어가 포함된 복합명사 : 대학생선교회
  분해 오류 발생 지점 : 대학-생선
  문제가 되는 1음절어 : 생
  분해 가능 형태 : 대학생-선 / 대학-생선
- 조사가 포함된 복합명사 : 보건복지부(가)
  분해 오류 발생 지점 : 복지-부가
  문제가 되는 1음절어 : 부
  분해 가능 형태 : 복지부-(가) / 복지-부가

위의 예와 같이 분해중의성을 보이는 분해 오류들은 오류 발생 지점에 위치한 1음절어의 분해 기준점에 따라 정확한 분해와 분해 오류로 나누어지므로 분해 지점의 1음절어를 이동하는 것만으로도 분해 오류의 교정이 가능하다. 하지만 분해 오류와 정확한 분해를 나눌 수 있는 1음절어의 발견과 재분배를 자동으로 하는 것은 힘든 작업이다. 따라서 분해 오류의 교정은 분해 오류가 발생한 지점을 기준으로 좌우에 위치한 두 구성명사간의 음절 이동을 통해 가능한 모든 형태의 분해 결과를 구성하고, 재구성된 분해 결과들을 빈도수로 비교, 가장 높은 분해 결과를 선택하는 방법을 사용한다.

4.1 분해 중의성에 대한 교정

복합명사 N의 j~k번째 음절로 이루어진 i번째 단위명사 $n_i=(s_j \cdots s_k)$가 분해 오류인 경우, n_i를 교정한 새로운 단위 명사 n_i'은 각 음절 재분해 결과 n_i'' 중에서 출현 확률이 가장 높은 것으로 한다.

$$P(n_i') = \operatorname{argmax} \frac{\mathrm{freq}(n_i'')}{\sum \mathrm{freq}(n_i'')}$$
$$\approx \operatorname{argmax} \mathrm{freq}(n_i'')$$
$$, \text{where } n_i'' = (s_j \dots s_m)$$
$$, j \le m \le l, \quad m \ne k$$
$$, n_{i+1} = (s_{k+1} \dots s_l)$$

그림 4는 분해 오류 지점의 음절 이동을 통한 분해 오류 교정 방법을 간략하게 도식화 한 것이다.

복합명사 '박정희대통령'의 분해 결과 '박정-희대-통령'은 그림 1과 같이 '박정-희대'지점에서 분해 오류가 탐지된다. 이 부분에서의 가능한 음절 이동 방법은 4가지(박정희-대, 박정희대-, 박-정희대, -박정희대)이고, 나타나는 재분해 형태는 총 3가지 이다. 나타난 3가지 재분해 형태 각각의 빈도를 파악하면 그림과 같이 박정희-대의 빈도가 가장 많다. 빈도 파악에 사용되는 명사 빈도 사전은 형태소 분석기가 미등록어(인식 기호 K)로 인식하는 결과도 모두 포함, 박정희와 같은 미등록 명사도 판단하도록 구성되어 있다.

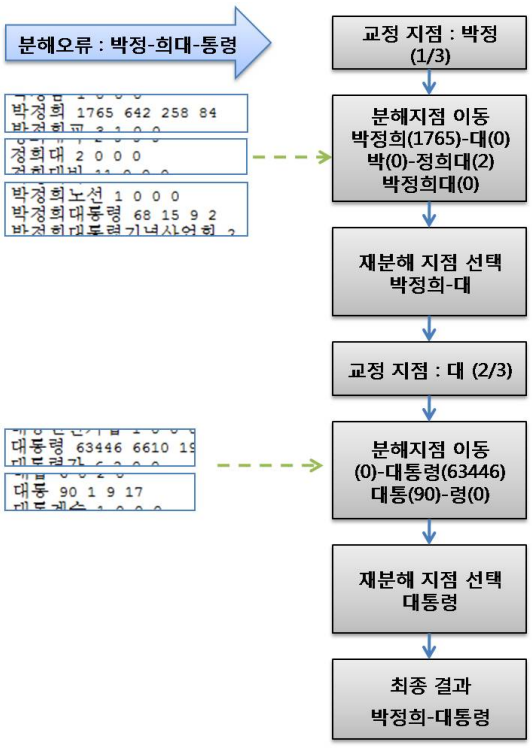

그림 4 '박정-희대-통령'의 분해 오류 교정

음절 이동을 통한 재분해 지점 검색을 통해 1차 재분해가 '박정희-대'로 선택되고, 분해 지점의 이동에 따라 이어지는 어절 역시 음절 재분해를 실시한다. 이어지는 어절에서도 위의 방식과 마찬가지로 가능한 음절 이동 방식 중 빈도가 가장 높은 '-대통령'이 선택되며, 두 개의 재분해 결과를 통해 최종 분해 결과는 '박정희-대통령'으로 교정된다.

4.2 조사가 포함된 분해 오류의 처리 방법

복합명사 표현 뒤에 조사가 이어지는 경우, '가', '과', '도', '인(이다+ㄴ)', '의' 등 특정 조사가 복합명사의 일부로 함께 분해되는 경우가 있다. 이러한 조사들은 동음이의어인 1음절 접사가 존재하고, 형태소 분석기가 조사로 판단하기 이전에 체언인 복합명사의 일부로 인식되어 분해 오류를 발생시킨다. '보건복지부가'는 복합명사+조사의 구조로 복합명사 '보건-복지부'와 조사 '가'로 분리되어야 하지만, 마지막 어절 '부가'가 명사이므로 '보건-복지-부가'로 분리된다.

이와 같이 조사가 복합명사와 함께 분해된 형태의 오류를 처리하기 위해서 복합명사의 마지막 음절이 조사로 의심되는 경우 분해 결과의 마지막 구성명사를 분해하여 조사인지 여부를 판단한다.

복합명사의 분해 결과 $N(=n_1n_2n_3\cdots n_k)$의 마지막 구성명사 n_k에서 마지막 1음절 j를 제거한 단어를 n_k'이라고 정의하고 n_k'의 형태에 따라 다음과 같이 j의 분해여부를 판단한다.

1. n_k'이 명사 사전에 등록된 단어인 경우
   - n_k' : 새로운 n_k으로 사용
   - j : 조사로 인식하고 제외시킴

   예) 기관-투자가
   (n_k=투자가, n_k'=투자, j=가)
   ➔ 기관-투자(가)
2. n_k'이 1음절 접미사인 경우
   - n_k' : n_{k-1}을 n_{k-1}+n_k'으로 대체한다.
   - j : 조사로 인식하고 제외시킴

   예) 보건-복지-부가
   (n_k=부가, n_k'=부, j=가, n_{k-1}=복지)
   ➔ 보건-복지부(가)
3. n_k'이 명사 사전 혹은 접미사 테이블에 등록되지 않은 단어인 경우
   - n_k' : 의미가 없으므로 n_k를 그대로 사용
   - j : 조사가 아니므로 분해하지 않음

   예) 기대-효과
   (n_k=효과, n_k'=효, j=과)
   ➔ 기대-효과

4.3 복합명사 분해 오류 교정 시스템

복합명사 분해 오류에 대한 처리 방법으로 제시된 탐지 방법과 교정 방법을 사용하여, 복합명사 분해 결과가 입력될 경우 분해 오류를 탐지하고 교정해주는 시스템을 개발한다. 그 구조는 그림 5와 같다.

빈도 계산시에 사용되는 bigram 공기 빈도 사전과 단일 명사 빈도 사전은 띄어쓴 복합명사의 빈도 만을 모아서 작성되었다. 붙여쓴 복합명사의 빈도 정보는 분해 오류 빈도가 포함된 정보이므로 제외했다.

분해 오류 여부에 대한 판단은 복합명사의 분해된 구성명사들 각각의 단일 빈도와 bigram 공기 빈도가 모두 임계값 이상인지를 살펴봄으로써 이루어진다.

분해 오류가 없는 복합명사는 바로 복합명사 사전에 저장되며, 분해 오류로 판단된 복합명사는 음절이동을 통한 재결합 방법과 조사 처리 방법을 통해 교정된 뒤 복합명사 사전에 저장된다.

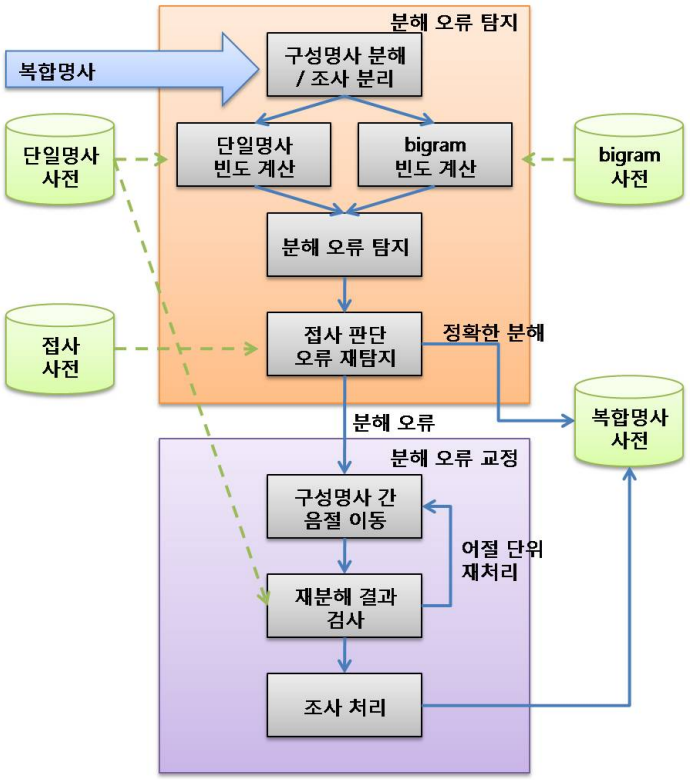

그림 5 복합명사 분해 오류 탐지-교정 시스템

5. 실험 및 결과

제시된 복합명사 분해 오류 교정 시스템에 대한 실험은 구성명사 빈도 정보를 추출한 것과 동일한 원시 말뭉치와 형태소 분석기를 사용해서 추출된 붙여쓴 복합명사를 대상으로 진행하였다.

분해 오류 탐지 방법은 붙여쓴 복합명사 중 출현 빈도가 가장 많은 5,213개(옳게 분해된 복합명사 4,963개, 잘못 분해된 복합명사 250개)를 사용해 실험하였다. 그 결과는 표 1과 같다.

표 1 복합명사 분해 오류 탐지 실험 결과

	옳게 분해된 복합명사	잘못 분해된 복합명사
옳은 분해로 판단	4,708	105
잘못된 분해로 판단	255	145

잘못 분해된 복합명사 250개 중 145개를 분해 오류로 탐지 하여 오류 탐지율(recall ratio)은 0.58이고, 전체

탐지된 분해 오류 400개 중 145개만 실제 분해 오류이므로 정확률(precision)는 0.3625가 된다. 정확도가 매우 낮으나, 이것은 분해 오류 탐지 과정에서 조사로 의심되는 마지막 음절을 강제 분리하기 때문에 발생한 것으로 교정 과정을 통해 다시 복구할 수 있다.

분해 오류 탐지과정에서 400개의 복합명사가 분해 오류로 탐지되었다. 분해 오류 교정 방법의 실험은 분해 오류로 탐지된 400개(옳게 분해된 복합명사 255개, 잘못 분해된 복합명사 145개)의 복합명사를 대상으로 진행했다. 그 결과는 표 2와 같다.

표 2 복합명사 분해 오류 교정 실험 결과

	옳게 분해된 복합명사	잘못 분해된 복합명사
교정 성공	235	96
교정 실패	20	49

옳게 분해된 복합명사를 오류로 판단했던 255개 중 235개는 원형으로 복구되었다. 앞서 설명한 것과 같이 조사로 오인된 음절이 분리되는 과정에서 발생한 오류 판단이므로 조사로 오인되어 분리되었던 원형이 복구되는 과정에서 교정이 이루어진다.

잘못 분해된 복합명사는 145개 중 96개가 교정되어 교정 정확률이 0.6621을 보였다. 탐지와 교정을 아울러 판단할 경우, 250개의 분해 오류 중에서 탐지, 교정이 모두 이루어진 것은 96개로 0.38의 탐지/교정 정확률을 보인다.

탐지/교정 전 복합명사의 분해 정확도는 0.952(=4,963/5,213), 탐지/교정 후의 복합명사 분해 정확도는 0.966(=(4,963+96-20)/5,213)으로 0.01 가량 복합명사의 분해 정확도가 올라가게 된다.

6. 결 론

본 논문에서는 복합명사의 분해 오류를 교정하기 위해 복합명사를 구성하는 각 단위명사에 대한 빈도 정보들을 이용해서 분해 오류 여부를 탐지하고 이를 재분해하는 방법을 제시하였다. 제시된 방법은 58% 정도의 탐지율과 66%정도의 교정률을 보여 아직까지는 좋은 성능을 보이지 못하고 있다.

분해 탐지의 경우, 공기 빈도와 단일 빈도를 따로 사용하는 과정에서 정확한 탐지 성능을 보여주지 못하고 있다. 두 빈도 정보를 함께 활용거나 다른 빈도 정보를 활용하는 등 탐지 성능을 높이는 작업이 추가로 요구된다.

오류 교정의 경우, 조사를 분해 후 처리하는 과정에서 많은 예외가 있고, 실제 접미사로 사용되는 음절이 조사로 오인되는 문제점도 발생했다. '농민회의'와 같은 형태의 복합명사에서는 마지막 음절 '의'가 문맥에 따라 조사로도 접미사로도 사용 가능한 형태이기 때문에 선택이 더욱 복잡해진다. 이러한 경우에 대해 조사 혹은 접미사를 처리하는 과정을 개선하여야 성능 향상에 많은 도움이 된다.

추후, 탐지와 교정 방법에서 존재하는 문제점들을 해결하는 더 나은 방법들에 대한 고찰이 진행되어야 할 것으로 보인다.

참고 문헌

[1] 강민규, 강승식, "한국어 복합명사 분해 오류 탐지 기법", 제21회 한글 및 한국어 정보처리 학술발표 논문집, pp.181-185, 2009

[2] 강승식, "한국어 복합명사 분해 알고리즘", 정보과학회 논문지(B), 25권, 1호, pp.172-182, 1998

[3] 윤보현, 조민정, 임해창, "통계 정보와 선호 규칙을 이용한 한국어 복합명사의 분해", 정보과학회 논문지(B), 24권 8호, pp.900-909, 1997

[4] 윤준태, 정의석, 송만석, "명사간 어휘 정보를 이용한 한국어 복합명사 분석", 정보과학회 논문지(B), 25권, 11호, pp.1716-1725, 1998

[5] 최재혁, "음절수에 따른 한국어 복합명사 분리 방안", 제8회 한글 및 한국어 정보처리 학술발표 논문집, pp.262-267, 1996

[6] 김응균, 서영훈, "미등록어 처리가 강화된 복합명사 분해", 제15회 한글 및 한국어 정보처리 학술발표 논문집, pp/40-46, 2003