한국어 및 외래어 미등록어를 포함한 복합명사 분석

김명선 나동열
연세대학교 컴퓨터정보통신공학부
mskim2272@dragon.yonsei.ac.kr dyra@ yonsei.ac.kr

Analysis of Compound Nouns Containing Korean or Foreign Unknown Words

Myoung-Sun Kim Dong-Yul Ra
Div. of Computer and Telecommunications Engineering
Yonsei University

요 약

본 논문에서는 미등록어 처리가 강화된 복합명사 분석 기법을 제시한다. 기본적으로 모든 복합명사 내에 한국어나 외래어의 미등록어가 포함되어 있을 수 있다는 가정하에 분석을 시도한다. 따라서 등록어로 구성된 복합명사에 대해서도 미등록어가 포함된 분해 후보가 생성될 수도 있다. 이는 분해 후보의 수를 크게 증가시키는 문제를 일으킨다. 이 문제에 대처하기 위하여 미등록어의 분류에 따라 미등록어로서의 가능성 여부의 판별 및 제거, 분해 후보 상호간의 견제에 의한 제거 등을 이용하였다. 이러한 과정은 정답 후보 선택시에도 영향을 미쳐 정답이 아닌 분해 후보가 선택되는 것을 방지할 수 있으며, 처리 시간을 줄일 수 있는 이점이 있다. 실험 결과 제시된 기법들이 매우 효과적임을 확인할 수 있었다.

1. 서론

한국어에 있어 복합명사는 단위명사로 띄어쓰는 것을 원칙으로 하지만 일반적으로 붙여 써도 무방 하다. 이로 인해 자연어처리의 여러 응용분야에서 많은 문제점을야기한다. 정보검색 분야에서는 정확한 색인어 추출의 실패로 검색 성능의 저하를 초래하고, 기계번역 분야에서는 적절한 대역어를 찾을 수 없으며, 맞춤법 검사 분야에서는 띄어쓰기의 오류를 유발한다.

이러한 문제의 해결방안으로 모든 단위 명사의 조합을 사전에 등록하는 방법이 있다. 하지만 한국어 단위명사의 결합은 매우 자유로우며, 결합하는 길이에 제한이 없기 때문에 이는 현실적으로 불가능하다. 또한 복합명사가 등록어만으로 구성되지 않고, 미등록어를 포함하고 있을 때는 더욱 심각한 문제를 초래한다. 예를 들어 "부르나이공화국"에서와 같이 미등록어가 포함된 복합명사는 다양한 분해 가능성이 존재 한다. 따라서 등록어만으로 구성된 복합명사 뿐만 아니라 미등록어가 포함된 복합명사도 분해할 수 있는 시스템의 개발이 필요하다.

2. 관련연구

심광섭은 말뭉치에서 추출한 음절간 상호정보를 이용한 복합명사 분해 기법을 제시하였다[2]. 이는 긍정적 상호정보, 부정적 상호정보, 머리 상호정보, 꼬리 상호정보의 네 가지 상호정보를 합성하여 임계값 이상일 경우는 띄어쓰고, 그렇지 않은경우는 붙여 쓰는 방법이다.

윤보현은 음절별 분해패턴을 이용하여분해를 시도하고, 정답 후보의 선택을 위해 명사의 개수가 같은 경우는 수식어와 중심어 빈도를 이용한 통계정보를 이용하고, 중의적 분해가 일어나는 명사의 개수가 다른 경우는 선호 규칙을 이용하는 방법을 제시하였다[3].

강승식은 네 가지 분해 규칙과 두 가지 예외 규칙을 사용하여 분해 후보들을 생성하고, 분해 후보들에 가중치를 부여함으로써 정답 후보를 선택하는 방법을 제시하였다[4].

이현민은 단위명사 사전과 접사 사전을 이용하여 오른쪽에서 왼쪽으로 최장 일치되는 단위명사를 우선으로 탐색하는 역방향 분해 기법을 제시하였다[5]. 그러나 지금까지 소개한

연구들에서는 주로 등록어로만 구성된 복합명사의 분해에 초점을 맞추었다.

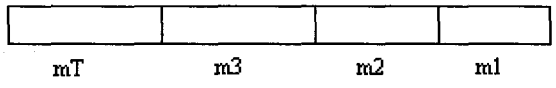
김응균은 분해패턴의 재사용성에 기반한 6음절까지의복합명사 분해를 시도하였다. 이때 음절 및 음소 출현 특성에 따라 외래어 미등록어를 판별하고, 인명 음절정보와 지명사전을 통해 이름 명사와 지명을 인식하는 방법을 제시하였다[6].

지금까지 연구된 복합명사 분해 기법은 등록어로만 구성된 경우에는 좋은 결과를 보이지만, 미등록어가 포함된 경우 만족할 만한 결과를 제시하지 못하였다. 특히, "부르나이공화국"에서처럼 미등록어(부르나이)에 등록어(나이)가 포함된 경우나, "남산동사거리"처럼 미등록어가 등록어로 나뉘어 지는 경우(남산+동사+거리), 정답후보를 생성하지 못하는 문제점을 드러냈다. 김응균은 이에 대한 해법을 제시하려 했으나, 복합명사의 길이를 6음절로 제한하고, 지명을 사전화하여 이용하는 등 실질적인 해결책을 제시하지 못하였다.

본 논문은 이처럼 미등록어가 포함된 복합명사 분해에 대한 성능향상을 목표로 한다.

3. 복합명사 분해 방법

복합명사를 분해하기 위하여 오른쪽에서 왼쪽으로 음절 단위의 사전검색을 취한다. 사전검색에 성공한 형태소는 스택에 넣고 바로 다음 좌측 음절부터 다시 사전검색을 시도하며, 실패시에는 좌측으로 한 음절을 확장하여 사전검색을 시도한다. 이러한 과정은 좌측 끝음절에 도달할 때까지 진행된다. 그 결과는 아래 그림과 같은 형태소의 열이 된다. 여기서 맨 좌측 이하의 열 m1, m2, m3 는 등록어 열로서 스택에 넣어 놓은 상황이다. 그러나 스택에 넣지 않는 맨 좌측의 부분(이것을 특별히 mT 라 부르자)은 등록어로 판명될 수도 있고 아닐 수도 있다.



전자의 경우(mT 까지도 등록어인 경우)는 모든 형태소가 등록어로서 복합명사 분해의 한 후보가 된다. 후자의 경우는 사전검색에 실패한 mT에 대해 미등록어 분석을 시도하고, 미등록어 분석 결과 생성된 형태소 열과 스택에 있는 형태소 열을 결합하여 미등록어가 포함된 후보로 생성된다.

이후 다음 분해 후보의 생성을 위하여 스택에서 한 형태소(여기서는 m3)를 꺼내서 이것을 좌측으로 한 음절씩 확장해 가며 보다 큰 명사를 찾기 위해 앞서 설명한 방법을 반복하여 진행한다.

이러한 과정은 어절 전체가 mT가 되어 이것이 하나의 등록어로 존재하거나 아니면 사전검색에 실패하여 이것에 대한 미등록어 분석을 시도한 후에 한 어절에 대한 복합명사 분석 작업이 종료된다.

그림1은 "민주연합"이라는 어절의 복합명사 분석 작업을 통한 분해 과정을 나타내고 있다. 이 그림의 맨 위의 상황은 "민(mT)+주연(m2)+합(m1)"의 등록어 열이 후보로 발견된 상황이다. 그 다음 후보를 찾기 위해 스택의 맨 위인 "주연"을 꺼내서 이것을 확장하기 위해 바로 좌측의 한음절 "민"을 붙인다. 확장된 부분인 "민주연(mT)" 이 등록어인지 사전검색을 통해 판단한다. 사전검색에 실패하므로 이부분을 미등록어 분석으로 넘겨 준다. 미등록어 분석모듈에서는 "민주연" 전체가 미등록어로 판정이 되고, 그러면 이것과 스택의 내용인 "합(m1)" 을 결합하여 한 분해 후보 "민주연/UW+합/N"으로 생성한다.

그런 다음에는 스택의 "합(m1)" 을 꺼내서 확장을 시도한다.

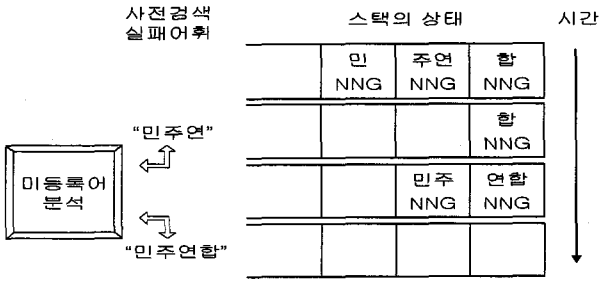


그림1. 복합명사 분석 예

4. 미등록어 분석

미등록어 분석은 3절에서 언급한 사전검색에 실패한 맨 좌측부분 mT를 대상으로 한다. 3절과 동일한 방식으로 등록어 열을 탐색하되 맨 좌측에서 우측으로(역방향으로) 진행한다. 역방향으로 분석을 진행하는 이유는 복합명사 분석에서는 우측에서부터 등록어 열을 찾고, 미등록어 분석에서는 좌측에서부터 등록어 열을 찾음으로써 나머지 사전검색에 실패한 부분을 잠정적인 미등록어로 간주하기 위함이다.

그림2는 "미국클링턴대통령"이라는 어절에 대한 분석의 예이다. 복합명사 분석과정에서 '대통령(m1)'이 등록어로 판별되어 스택에 존재하는 상태에서 어휘를 확장하다 사전검색에 실패한 '미국클링턴(mT)'에 대한 미등록어 분석 과정을 나타낸다. 스택-1은 복합명사 분석 진행

중 사용하는 스택이며, 스택-2는 미등록어 분석에서 사용되는 스택이다. 그림을 통해 알 수 있듯이 미등록어 분석은 왼쪽에서부터 등록어를 탐색해 나간다. 이 과정에서 맨 우측부분이 사전에 존재하지 않으면 잠정적인 미등록어로 간주한다. 따라서 미등록어는 항상 미등록어분석을 시도하는 부분의 가장 우측에 존재하게 된다(예로는 아래 그림의 ①의 "국클링턴").

이러한 미등록어 분석은 미등록어 분석으로 넘겨진 부분 전체가 미등록어인지 판정하는 단계까지 도달한 후 종료한다(그림의 ③의 경우).

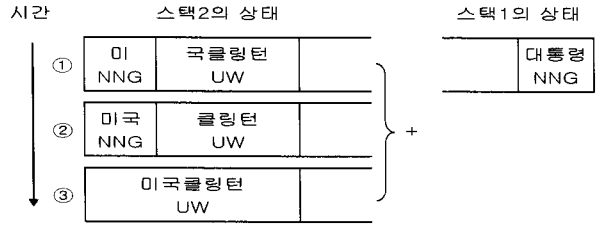


그림2. 미등록어 분석 예

4.1 미등록어 분류

위에서 설명한대로 잠정적인 미등록어가 생성되면 세부적인 처리를 위하여 한국어 미등록어인지 외래어 미등록어인지에 대한 분류를 시도한다. 미등록어가 $s_1 \ldots s_n$인 n개의 음절로 이루어졌다고 가정할 때 분류를 위하여 외래어 및 한국어 고유명사 리스트에서 추출한 음절 바이그램 및 유니그램, 중성·종성 결합 정보를 이용한다. 본 논문에서 사용된 미등록어 관련 음절 정보는 표1을 통해 얻는다.

표1. 통계정보 추출을 위한 고유명사 집합

구분		개수	구분		개수
외래어	어절	75,588	한국어 지명	어절	19,655
	음절	365,352		음절	54,696
한국어 인명	어절	772,494	한국어 기관명	어절	168,074
	음절	2,317,482		음절	336,148

먼저 음절 바이그램 정보를 이용한 분류를 시도한다. 이때 아래식에 의해 구해지는 각각의 바이그램 조합 B_{fr}과 B_{kr}의 값이 다른 경우 둘 중 1의 값을 갖는 쪽으로 미등록어가 분류된다.

$$B_{fr}(s_1 \ldots s_n) = \prod_{i=1}^{n-1} \delta(s_i, s_{i+1}) \qquad \delta(s_i, s_{i+1}) = \begin{cases} 1 & if\ C_{fr}(s_i, s_{i+1}) \neq 0 \\ 0 & otherwise \end{cases}$$

$$B_{kr}(s_1 \ldots s_n) = \prod_{i=1}^{n-1} \delta'(s_i, s_{i+1}) \qquad \delta'(s_i, s_{i+1}) = \begin{cases} 1 & if\ C_{kr}(s_i, s_{i+1}) \neq 0 \\ 0 & otherwise \end{cases}$$

- $C_{fr}(s_i, s_{i+1})$, $C_{kr}(s_i, s_{i+1})$: 주어진 바이그램이 각각 외래어와 한국어 미등록어에서 출현한 횟수.
- $B_{fr}(s_1 \ldots s_n)$, $B_{kr}(s_1 \ldots s_n)$: $s_1 \ldots s_n$ 사이 모든 바이그램이 존재하면 1, 그렇지 않으면 0을 갖는다.

그렇지 않고 B_{fr}과 B_{kr}가 같은 경우에는 아래식에 나타난 음절 유니그램과 중성-종성 결합 정보를 이용하여 미등록어 분류를 시도한다.

$$P(fr \mid uw) = \prod_{i=1}^{n} \{P_{fr}(s_i) + f(a) P_{fr}(p_i)\}$$

$$P(kr \mid uw) = \prod_{i=1}^{n} \{P_{kr}(s_i) + f(a) P_{kr}(p_i)\}$$

$$f(a) = \begin{cases} 1 & if\ P_{fr}(s_i) = 0 \,\&\&\, P_{kr}(s_i) = 0 \\ 0 & otherwise \end{cases}$$

$$C = P(fr \mid uw) - P(kr \mid uw) \qquad \text{미등록어 분류} = \begin{cases} FR & if\ C > 0 \\ KR & if\ C < 0 \\ Del & if\ C = 0 \end{cases}$$

* *exception* : P(fr|uw)와 P(kr|uw)가 0이 아니면서 같은 경우는 분류를 랜덤하게 결정.

- $P_{fr}(s_i)$, $P_{kr}(s_i)$: 음절 s_i가 각각 외래어 또는 한국어에 나타날 확률.
- $P_{fr}(p_i)$, $P_{kr}(p_i)$: 음절 s_i의 중성·종성(p_i) 결합이 각각 외래어 또는 한국어에 나타날 확률.
- FR : 외래어 미등록어, KR : 한국어 미등록어, Del : 미등록어가 아닌 것으로 판단되어 제거해야 함을 나타냄.

이때 사용되는 음절 유니그램과 중성·종성 결합 정보는 아래식에 의해 구해진다. 여기서 한국어와 외래어 고유명사 데이터의 형평성을 맞추기 위하여 두 데이터의 rate를 이용한다.

$$P_{fr}(s_i) = \frac{f_{s_i}}{f_{s_i} + k'_{s_i}} \quad P_{kr}(s_i) = \frac{k'_{s_i}}{f_{s_i} + k'_{s_i}} \quad k'_{s_i} = k_{s_i} \times \frac{1}{rate}$$

$$P_{fr}(p_i) = \frac{f_{p_i}}{f_{p_i} + k'_{p_i}} \quad P_{kr}(p_i) = \frac{k'_{p_i}}{f_{p_i} + k'_{p_i}} \quad k'_{p_i} = k_{p_i} \times \frac{1}{rate}$$

$$rate = \frac{\text{한국어 고유명사 총 음절수}}{\text{외래어 고유명사 총 음절수}}$$

- k_{pi}, f_{pi} : 음절 s_i의 중-종성 결합이 한국어/외래어 고유명사 내에서 출현한 횟수.
- k_{si}, f_{si} : 음절 s_i가 한국어/외래어 고유명사 내에서 출현한 횟수

한국어와 외래어에 대한 각각의 확률값 P(fr|uw)와 P(kr|uw)를 구하고, 두 확률값의 차C를 이용해 미등록어를 분류한다. 만약 C가 0보다

크면 외래어로 분류하고, 0보다 작으면 한국어로 분류한다. 그렇지 않고 C가 0인 경우 즉, 두 확률값이 모두 0인 경우는 미등록어를 제거한다. 만약 두 확률값이 모두 0이 아니면서 같은 경우는 극히 드문 경우로써 미등록어의 분류를 랜덤하게 결정한다.

위에서 미등록어를 제거하는 경우(Del)는 미등록어 내에 한국어에서만 사용되는 음절(그림2의 '국')과 외래어에서만 사용되는 음절(그림2의 '클', '턴')이 공존하는 경우에 해당한다. 따라서 그림2에서 ①,③번의 미등록어는 제거된다.

4.2 미등록어로서의 가능 여부 판별

미등록어의 분류가 결정되고 난 후, 이 미등록어가 포함된 하나의 분해 후보를 생성하기 전에 각각의 분류에 따라 최종적으로 미등록어로서 가능한지의 여부를 판단한다.

4.2.1 외래어 미등록어로 분류된 경우

외래어의 경우 그 특성상 미등록어의 길이에 제한을 두지 않는다. 본 논문의 복합명사 분석 기법은 그림2에서 보이는 것처럼 어휘의 계속적인 확장을 시도한다. 따라서 미등록어의 무분별한 확장을 방지하기 위한 제약이 이루어져야 한다. 미등록어가 앞이나 뒤에 2, 3음절로 구성된 단위명사를 포함하여 확장된 것일 경우 미등록어로서의 가능성 여부를 판단하기 위하여 외래어 prefix 외래어 고유명사의 앞부분에 나타난 2,3음절 단위명사 및 suffix 외래어 고유명사의 뒷부분에 나타난 2,3음절 단위명사 정보를 이용한다.

예를 들어 "거장차이코푸스키"라는 복합명사에 대해서 '차이코푸스키'가 외래어 미등록어로 제안되었다고 가정하자. 여기에는 미등록어의 앞과 뒤에 각각 '차이'와 '스키'라는 등록어가 포함되어 있다. 따라서 이들이 각각 어떤 외래어의prefix 및 suffix로 되는 경우가 있는지 판단한다. 이런 경우가 있으면 미등록어 후보로 가능하게 하고, 그렇지 않으면 이 미등록어를 제거한다. '거장차이코푸스키'가 외래어 미등록어로 제안된 경우는 '거장'이라는 등록어가 외래어prefix 리스트에 존재하지 않기 때문에 제거된다. 이 기법을 이용하면 무분별한 확장으로 생기는 외래어 미등록어를 방지한다.

4.2.2 한국어 미등록어로 분류된 경우

한국어 미등록어의 경우 그 특성상 최대 길이를 4음절로 제한한다. 그 이유는 5음절 이상의 한국어 미등록어는 매우 드물기 때문이다.

미등록어가 한국어로 분류되고 그 길이가 3음절 이하인 경우는 별도의 제약없이 미등록어로서 가능하게 하며, 4음절인 경우에는 제약조건을 두어서 결과에 따라 판단한다. 이처럼 4음절에 대해서만 제약조건을 두는 이유는 3음절 이하의 경우는 미등록어로서 출현 가능성이 높기 때문이며, 4음절의 경우는 한국어에 있어4음절 미등록어역시 출현빈도가 낮고, 4음절이 등록어로 나뉘어 지는 경우는 그것이 한국어의 특성상 더 올바를 가능성이 매우 크기 때문이다.

다음의 네 가지 경우를 적용하여 4음절 미등록어로서의 가능 여부를 결정한다. 이때 [규칙1]은 미등록어 제거 규칙으로서 모든 4음절 미등록어를 대상으로 검사하며, [규칙2,3,4]는 4음절 미등록어를 두 부분으로 나누었을 때 모두 등록어로 구성된 22, 31, 13 유형이거나, 한쪽이 2음절 이상인 등록어이고 다른쪽은 등록어가 아닌 형태소로 구성된 경우에만 적용한다. 이중 한 규칙이라도 만족하는 경우에만 미등록어로서 인정한다. 이러한 구성이 아닌 것들에 대해서는 무조건 4음절 미등록어로 인정해준다.

[규칙1] "3음절 인명 + 1음절 인명 단서어휘"의 구성으로 해석되는 경우는 미등록어로 불인정함.

이때 1음절 인명 단서어휘는 '군, 님, 옹, 양, 씨'이며, 그 우측에 명사가 나타나지 않아야 한다. 그 이유는 1음절 인명 단서어휘는 의존명사로 사용된 것이며, 의존명사 뒤에는 명사가 올 수 없기 때문이다. 3음절 인명으로서의 가능성에 대한 체크는 각 음절이 인명의 해당 위치에 사용된 적이 있는지의 여부를 이용한다. 이 경우는 이 4음절 미등록어를 제거한다.

예) 홍길동 ***균***/UW : 이 후보는 제거됨.

[규칙2] " 2음절 성씨 + 2음절 인명"의 구성으로 해석되는 경우 미등록어로 인정함.

예) ***독고영재***/UW 에서 "독고"는 성씨, "영재"는 이름이 가능하므로 미등록어로 인정된다.

[규칙3] "1음절 성씨 + 3음절 인명" 의 구성으로 해석되는 경우 미등록어로 인정함.

이때 3음절 인명은 인명사전에 나타난 4음절 인명에 대해 성을 제외한 3음절 이름이 명사로 존재하는 것만을 인정한다.

예) ***이민들레***/UW 에서 "이"는 성씨, "민들레"는 이름으로 나타난 적이 있으므로 미등록어로 인정한다.

[규칙4] "3음절 지명 + 1음절 지명 단서어휘"의 구성으로 해석되는 경우 미등록어로 인정함.

이때 1음절 지명 단서어휘는 '시, 군, 구, 읍, 면, 동, 리' 이다. 3음절 지명에 대한 판단은 음절1은 지명사전에서 지명의 음절1로 나타난 적이 있어야 하고, 음절2나 3은 지명사전에 있는 어느 한 지명의 2째 음절 이후에 사용된 적이 있으면 가능한 것으로 해석한다.

예) *갈마곡리*/UW 의 경우 : "리"는 지명단서어휘이고, "갈"은 지명의 첫 음절로 사용된 적이 있으며, "마","곡"은 지명의 2째 음절 이후에 사용된 적이 있으므로 미등록어로 인정한다.

이러한 기술을 이용함으로써 무분별한 미등록어 후보의 생성을 억제할 수 있다.

5. 분해 후보 제거기법

주어진 어절에 대한 분해 후보가 모두 생성된 후에 분해 후보 제거기법을 이용하여 정답으로서의 가능성이 낮은 후보들을 제거하는 것이 필요하다. 그 이유는 너무 많은 분해 후보가 생성되기 때문이다.

이를 위해 세가지 제거 기법을 적용함으로써 정답일 가능성이 낮은 후보를 제거한다. 이때 제거기법 1은 등록어만으로 구성된 복합명사를 대상으로 적용되고, 제거기법 2, 3은 미등록어가 포함된 복합명사 분해 후보를 대상으로 적용된다.

[제거기법1] 최소 접속 원리를 이용하여 형태소의 수가 적은 후보가 형태소의 수가 많은 후보를 제거한다[3].

[제거기법2] 단서어휘를 포함한 미등록어 후보가 단서어휘를 포함하지 않은 미등록어 후보를 제거한다. 이때 단서어휘로는 (2음절 이상인 것만 가능함) 인명, 지명, 기관명 단서어휘를 이용하며, 그 위치는 다음과 같이 한정한다.

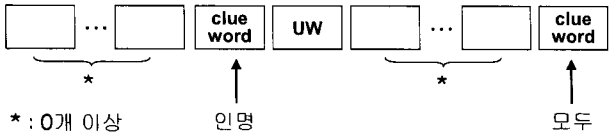


그림3. 분해 후보 내에서 단서어휘의 위치

- 인명단서어휘: 회장, 사장, 감독, ...
- 지명단서어휘: 부근, 근처, 지점, 지역, ...
- 기관명단서어휘: 강습소, 병원, 학교, ...

[제거기법3] 분해 후보들 간의 형태소 개수가 같고, 분해 후보 내에서 미등록어의 위치가 같은 경우 미등록어가 짧은 후보가 미등록어가 긴 후보를 제거한다.

제거기법3은 1의 변형된 형태로써 보다 긴 등록어를 가진 후보를 선호하는 방법이다.

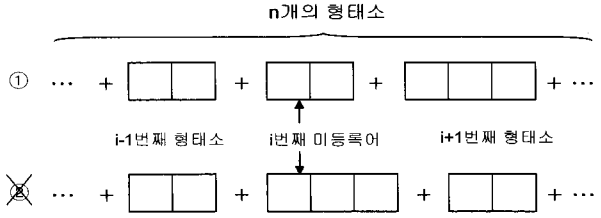


그림4. 제거기법3에 대한 예

이처럼 세가지 제거기법에 의해 정답 가능성이 낮은 분해 후보를 제거함으로써 최종단계인 통계정보에 의한 정답 후보 선택 단계에서 시간을 줄일 뿐만 아니라 정답의 선택 가능성을 높일 수 있다.

6. 정답 후보 선택

위에서 설명한 과정을 통하여 많은 후보를 제거한 후에도 살아 남은 분해 후보의 수는 매우 많다. 이중에서 최종적으로 한 후보를 선택하여 정답으로 출력하게 된다.

각 분해 후보 L 마다 다음과 같은 확률을 구한다.

$$P(L) = \prod_{i=1}^{n} P(m_i \mid m_{i-1})$$

$P(m_i \mid m_{i-1})$ 은 형태소 바이그램 확률로서 품사가 부착된 형태소 말뭉치에서 구한다. 어휘 바이그램의 부족으로 인해 데이타 부족문제가 발생하는데 이는 백오프(Back-off) 기법을 이용하여 해소한다[10].

모든 분해후보에 대해 P(L) 을 구한 후 가장 큰 값을 가지는 후보를 정답으로 출력한다.

7. 실험 및 분석

실험은 등록어만으로 구성된 복합명사 8,163개와 외래어 및 한국어 미등록어를 포함한 복합명사 900개를 대상으로 실시되었다.

분해 후보 제거기법에 따른 후보 수의 변화와 등록어 및 미등록어 복합명사에 대한 정확률을 측정하였으며, 전체 복합명사의 음절수별 정확률을 측정하였다. 그 결과 모든 경우에 대해 분해 후보 제거 기법을 적용한 경우가 그렇지 않은 경우보다 정확도가 높거나 같게 측정되었으며, 분해 후보의 수를 현저히 줄임으로써 정답 후보 선택시의 부담을 줄일 수 있었다.

복합명사 전체(9,063개)를 대상으로 실험한 결과에서는 제거기법 1,2,3을 적용하기 전에는 전체60,404개와 평균 6.3개의 후보가

생성되었지만, 제거기법을 적용한 후에는 전체39,218개와 평균 4.0개로 후보의 수가 현저히 감소함을 확인하였다.

실험은 등록어만으로 구성된 복합명사에 대해 99.33%의 정확률을 보였으며, 미등록어를 포함하는 복합 명사에 대해 94.33%의 정확률을 보였다. 또한 전체에 대해서는 98.82%의 정확률을 나타냈다.

결과적으로 본 논문에서 제시한 복합명사 분석 방법은 등록어로 구성된 복합명사를 잘 처리하면서, 일반적인 미등록어 뿐만 아니라 미등록어가 등록어를 포함하는 경우('*차이*코푸*스카*') 또는 미등록어가 등록어로 나뉘어지는 경우('*남산*+ *동*사+거리')에 대해서도 정답 후보의 생성 및 선택이 잘 이루어짐을 확인하였다.

표2. 제거기법에 따른 분해 후보 수 및 정확률

실 험 기 준 (대상)	후보 수	정확률(%)
제거기법 미적용 (등록어)	18,996	99.31
제거기법1 적용 (등록어)	9,395	99.33
제거기법 미적용 (미등록어)	6,609	93.78
제거기법2만 적용 (미등록어)	4,727	94.33
제거기법3만 적용 (미등록어)	4,764	93.78
제거기법2,3 적용 (미등록어)	3,534	94.33
제거기법 1,2,3 적용(전체)	39,218	98.82

표3. 음절수별 정확률

음절수	복합명사 개수 (한국어/외래어)	분해성공 개수 (한국어/외래어)	정확률(%)
4	5,001(4,919/82)	4,986(4,911/75)	99.70
5	1,766(1,524/242)	1,746(1,509/237)	98.87
6	1,134(915/219)	1,115(908/207))	98.32
7	572(406/166)	555(401/154)	97.03
8	316(203/113)	296(190/106)	93.67
9	150(114/36)	142(110/32)	94.67
10	79(54/25)	75(52/23)	94.94
11	33(22/11)	29(21/8)	87.88
12	8(5/3)	8(5/3)	100
13	4(1/3)	4(1/3)	100
합계	9,063(8,163/900)	8,956(8,108/848)	98.82

8. 결론

본 논문에서는 한국어 미등록어 뿐만 아니라 외래어 미등록어가 포함된 모든 복합명사에 대한 분해가 가능한 기술을 제시하였다.

이러한 모든 경우를 다룰 수 있는 복합명사 분해에서는 많은 수의 분해 후보가 가능하다. 이런 문제점을 감소시키기 위해 후보 내의 미등록어를 한국어 및 외래어 미등록어 중 어느 것인지 분류하고, 이에 따라 미등록어로서의 가능 여부를 판단하여 제거하는 기법을 개발하였다. 또한 후보 상호간의 제거기법을 도입함으로써 분해 후보의 수를 효과적으로 감소시킬 수 있도록 하였다.

실험에서 우리의 복합명사 분해 시스템은 높은 성능을 보여 주였다.

실험을 통해 발견된 문제점으로는 표1에 나타난 고유명사 집합의 데이터 부족 및 편중으로 인해 음절 통계정보들의 정확성이 다소 떨어짐으로써 오류를 발생하였다. 예를들어, "매월당문학사상연구회"라는 복합명사의 경우 정답 후보를 선택하는 과정에서 미등록어 '매월당'에 대한 점수 계산시 음절 통계정보의 부족으로 인해 점수를 낮게 받음으로써 잘못된 정답이 선택되었다. 또한 "후두스트로보스코피검사"라는 복합명사의 경우는 등록어 '후두'가 외래어 미등록어 prefix리스트에 존재하지 않아 정답 후보를 생성하지 못하는 문제점도 발견되었다.

9. 참고문헌

[1] 장동현, 맹성현, " 효율적인 색인어 추출을 위한 복합명사 분석 방법", 제8회 한글 및 한국어 정보처리 학술발표논문집, pp.32-35,1996,

[2] 심광섭, "합성된 상호 정보를 이용한 복합명사 분리", 정보과학회 논문지(B) 제24권 11호, pp.1307-1317, 1997

[3] 윤보현, 조민정, 임해창, "통계정보와 선호규칙을 이용한 한국어 복합명사의 분해", 정보과학회논문지(B) 제 24권 제8호, 1997

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

[5] 이현민, 박혁로, “복합명사의 역방향 분해 알고리즘”, 한글 및 한국어정보처리 학술발표논문집, 2000
[6] 김응균, 서영훈, “미등록어 처리가 강화된 복합명사 분해”, 제15회 한글 및 한국어 정보처리 학술대회, 2003
[7] 채영숙, 권혁철, “말뭉치로부터 추출된 통계 정보를 활용한 한국어 복합명사 분석”, 인지과학회논문지, 제 8권 2호, pp.101-108, 1997
[8] 양장모, 김민정, 권혁철, “언어 정보를 이용한 한국어 미등록어 추정”, 정보과학회지 제23권 1호, pp.957-960, 1996
[9] 박봉래, 황영숙, 임해창, "용례 분석에 기반한 미등록어의 인식", 한국정보과학회 논문지(B), 제25권 제2호, pp.397-407, 1998
[10]박재한, 김명선, 노대욱, 나동열, "백오프 통계정보를 이용한 미등록어 포함 복합명사의 분해", 제16회 한글·언어·인지 학술대회, 2004