여과 및 분리 기법을 이용한 한국어 기준명사 추출

김재훈° 김준홍 박호진
컴퓨터공학과, 한국해양대학교
첨단정보기술연구센터

jhoon@hanara.kmaritime.ac.kr, {rainmk, hanwool}@nlplab.kmaritime.ac.kr

Base-Noun Extraction with Filtering and Segmentation in Korean

Jae-Hoon Kim° Jun-Hong Kim Ho-Jin Park
Department of Computer Engineering, Korea Maritime University
and
Advanced Information Technology Research Center

jhoon@hanara.kmaritime.ac.kr, {rainmk, hanwool}@nlplab.kmaritime.ac.kr

요 약

웹의 등장으로 방대한 양의 문서를 다루는 정보검색, 정보추출, 정보요약 등의 분야에서 명사 추출은 대단히 중요한 역할을 담당하는 한 모듈이다. 본 논문에서는 대량의 문서에서 효과적으로 명사를 추출하기 위해 여과기법과 분리기법을 이용한 한국어 기준명사 추출 시스템을 기술한다. 기준명사는 명사들 중에서 기본이 되는 명사로서 복합명사는 제외된다. 본 논문의 기본적인 개념은 먼저 여과기법을 이용해서 명사를 포함하지 않은 어절을 미리 제거하고, 그리고 분리기법을 이용해서 명사가 포함된 어절에서 명사어구와 조사를 분리하고, 복합명사에 해당할 경우에는 각 명사를 분리하여 기준명사를 추출한다. ETRI 말뭉치를 대상으로 실험한 결과, 재현율과 정확률 모두 약 89% 정도의 성능을 보였으며, 제안된 시스템을 한국어 정보요약 시스템에 적용해 보았을 때, 좋은 결과를 얻을 수 있었다

1. 서론

웹은 전세계를 통하여 많은 정보를 쉽게 얻을 수 있는 정보의 보고이다. 가상공간에 존재하는 정보들은 매우 다양하며, 그 양도 매우 빠른 속도로 증가하고 있다. 방대한 정보공간에서 유용한 정보를 찾기 위해 널리 사용되는 도구가 웹 검색 엔진이다. 명사 추출기는 검색 엔진을 구축하기 위해 필수적인 도구 중 하나이며, 색인어 추출, 자연언어 질의어 분석, 시소러스 구축 등에서 사용되고 있다. 이 밖에서 정보추출이나 정보요약 등 대량의 자연어 문서를 다루는 분야에서 널리 사용되고 있다.

자연언어 문장에서 정확하게 명사를 추출하기 위해서는 복잡한 과정을 거쳐야 한다. 왜냐 하면, 많은 명사들은 여러 형태의 중의성을 가지고 있기 때문이다. 대부분의 중의성은 간단하게는 형태소 분석과 같은 낮은 단계의 분석으로도 쉽게 해결할 수 있으나, 일부의 중의성은 의미 분석과 같은 복잡한 과정을 통해서만 해결할 수 있다. 그러나, 정보검색과 정보요약과 같은 분야에서는 짧은 시간 내에 방대한 양의 문장을 처리하고 특정영역(예: 신문, 소설 등)에 무관하게 처리해야 하는 경우에는 의미 분석과 같은 복잡한 과정을 이용하는 것은 적절한

방법이 아니다[1]. 왜냐 하면, 개발 측면에서 보면, 시스템 개발에 필요한 시간과 노력이 많이 요구되고, 실행 측면에서도 역시 많은 시간이 요구되고, 시스템이 강인하지(robust) 못하기 때문이다. 따라서 정보검색이나 정보요약과 같은 분야에서는 정확성에는 다소 희생이 따르더라도 특정 영역에 무관하고 강인하며 빠른 명사 추출기를 필요하게 된다.

한국어 문장은 여러 개의 어절로 구성된다. 어절은 체언, 용언, 수식언 등으로 나눌 수 있다. 대부분의 명사들은 체언에 속한다. 명사를 찾기 위해서는 어절들 중에서 일단 체언을 찾아야 한다. 본 논문에서 체언을 찾기 위해서 상호정보를 이용한 여과 기법을 사용한다. 즉, 체언이 아닌 다른 어절을 문장으로부터 제거한다. 많은 체언은 내용어에 해당하는 명사구와 기능어에 해당하는 조사로 구성된다. 따라서 명사구를 정확히 찾기 위해서는 체언으로부터 조사를 분리해야 한다. 본 논문에서 조사를 분리하기 위해서도 상호정보를 이용한 분리 기법을 사용한다. 그리고 나서 찾아진 명사구에는 많은 경우에는 하나의 명사로 구성되었지만, 몇몇의 경우에는 여러 개의 명사가 하나의 명사구를 이루고 있다. 본 논문의 최종적인 목표는 기준 명사를 찾는 것으로 하고 있기 때문에 복합명사는 기준 명사로 분리해야 한다.

본 논문의 구성을 다음과 같다. 2절에서 본 논문과 관련된 한국어 명사 추출 방법과 복합명사 분리 방법에 대해서 기술한다. 3절에서 상호정보를 이용한 여과 및 분리 기법을 통한 명사 추출 방법에 대해서 논하고, 4절에서 제안된 시스템의 성능을 평가하고 5절에서 다른 한국어 명사 추출 방법들과 비교· 분석하고자 한다. 끝으로 6절에서 결론을 맺고 앞으로의 연구 방향에 대해서 기술한다.

2. 관련 연구

2.1. 한국어 명사 추출

한국어 명사 추출 시스템은 크게 세 가지로 분류된다. 첫째, 품사 태거를 이용한 경우이고[2], 둘째, 형태소 분석기를 이용하는 경우이고[3-5], 셋째, 아무런 언어분석 도구를 사용하지 않는 경우이다[1].

품사 태거을 이용하는 방법은 품사 태깅 결과에서 원하는 품사에 해당하는 단어만 출력하면 된다[2]. 이 방법은 이미 품사 태거가 존재할 경우에 아주 쉽고 정확한 결과를 얻을 수 있다. 그러나 품사 태거가 존재하지 않는다면 품사 태거를 구축하는데 많은 시간과 노력이 필요하다. 인터넷 문서를 처리하기 위해서는 미등록어 문제를 잘 처리할 수 있어야 한다. 그러나, 이 방법에서 미등록어 문제 해결은 형태소 분석에 매우 의존적이다.

형태소 분석기를 이용하는 방법은 형태소 분석기의 결과에서 명사가 포함된 어절의 유형(체언 유형)을 정의하고, 각 유형에 일치되는 어절은 형태소 분석 결과를 이용해서 명사 이외의 성분(예: 조사 등)들을 제거하고 출력한다[4]. 체언 유형의 중의성이 발생될 수 있고, 규칙을 이용하는 방법이기 때문에 시스템의 확장성에 문제가 발생될 수도 있다. 이 방법도 미등록어 문제 해결은 형태소 분석에 매우 의존적이다.

언어 분석 도구를 사용하지 않는 경우에는 사전과 규칙을 이용해서 명사를 추출한다[1]. 이 방법은 비교적 단순하고 매우 빠른 속도로 명사를 추출할 수 있다. 그러나 언어 분석 도구를 이용하는 방법들보다 정확률이 낮을 수 있다.

2.2. 복합명사 분리

복합명사란 두 개 이상의 기준명사가 결합하여 새로운 의미를 갖게 되는 단어(예: 인공지능, 정보검색)를 말하며, 구문적으로는 단일단어와 같은 역할을 한다. 한국어 복합명사 분해는 크게 통계적 방법[6][7]와 규칙기반 방법[8][9]으로 나눌 수 있다.

통계적인 방법은 예로 [6]을 살펴보자. [6]은 통계정보(compound noun formation probability, CFP)와 선호규칙(minimal noun preference rule, MNPR)을 이용하여 복합명사

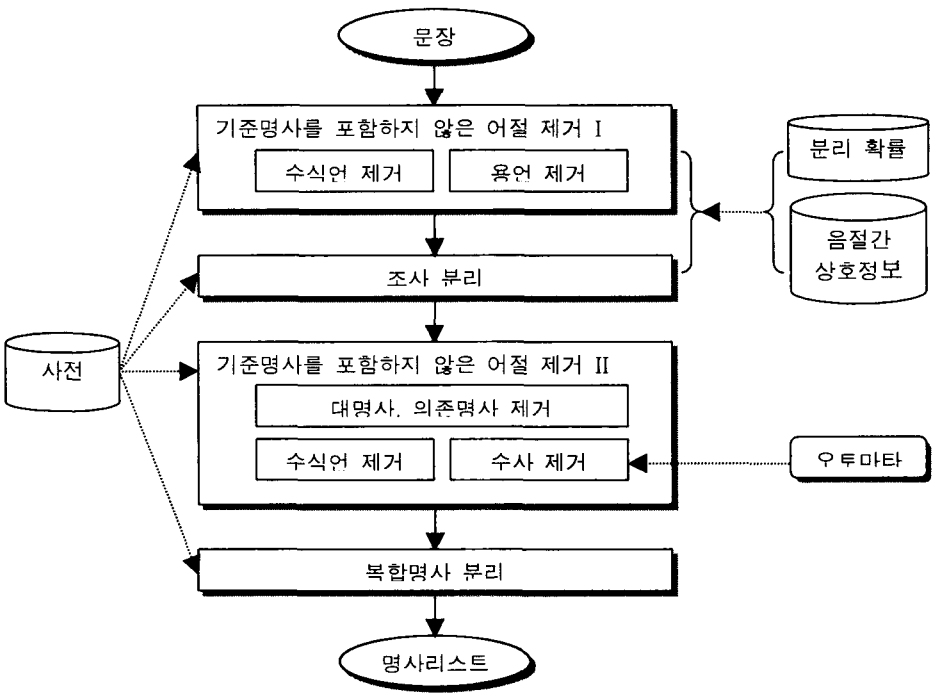

그림 1. 한국어 명사 추출 시스템의 구조

를 단일 명사로 분해하는 방법을 제안하였다. 여기서 통계정보란 1음절 접사 빈도수, 그리고 2음절 또는 3음절 단위 명사가 복합 명사 내에서 사용된 위치 정보와 빈도수를 이용한 것이고, 선호 규칙이란 중의적 분해, 즉 둘 이상의 방법으로 분리가 가능한 복합명사의 분해패턴이 있을 때 분해되어 생기는 단일 명사의 개수가 최소로 되는 분해 패턴을 선호하는 규칙을 의미한다.

규칙기반 방법의 예를 [8]을 살펴보자. [8]은 네 종류의 복합명사 분해 규칙을 사용하고 있으며, 두 종류의 예외 규칙을 사용하고 있다. 이들 규칙에 의해서 후보 규칙을 생성하고 생성된 후보에 대해 가중치를 부여하여 가중치에 따라 최적 후보를 선정하게 된다. 가중치를 부여하는 방법도 규칙에 의해서 결정되는데, 기준명사의 유형, 사전에 수록된 음절의 길이, 접미사가 결합된 음절 길이, 음절 패턴 빈도, 중심어 빈도에 따라 적절하게 조절된다.

3. 한국어 기준명사 추출 시스템

기준명사는 체언을 구성하는 가장 최소 단위를 말한다. 체언 중에서는 수사, 대명사, 의존명사를 제외한 보통명사와 고유명사의 최소 단위만을 본 논문에서는 기준명사라고 한다. 본 논문에서는 여과 기법과 분리 기법을 이용한 한국어 기준명사 추출 시스템을 제안한다. **여과 기법(filtering)**은 비체언성 어절, 즉, 수식언, 독립언과 용언을 제거하기 위해 사용된다. 한국어에서 일부 용언에 명사를 포함하고 있다. 예를 들면 접미어 "-하다/되다"와 결합하여 용언이 된 어절이다. 이들은 제거하지 않는다. 또한 체언의 경우에도 대명사, 수사, 의존명사를 포함하는 경우에는 제거한다. **분리 기법(segmentation)**은 체언에 포함된 명사구와 조사를 분리하는 일과 명사구가 복합명사일 경우 이를 기준명사로 분리하는 일을 담당한다.

이와 같은 기능을 포함하는 한국어 명사 추출 시스템의 구조가 그림 1과 같다. 그림 1에서 사전은 가능한

모든 단어가 포함되어 있고, 각 단어 정보는 명사, 동사, 부사, 형용사 등과 같은 품사 정보 뿐이다. 명사와 조사 그리고 용언과 어미의 분리확률과 음절의 상호정보는 조사 분리 모듈과 용언 제거 모듈에서 필요한데 자세한 설명은 다음 절에서 기술된다. 조사가 분리된 후에 사전과 오토마타를 이용해서 대명사, 부사, 수사 등이 제거된다. 각 모듈에 대한 자세한 설명은 다음 절에서 기술된다.

3.1. 기준명사를 포함하지 않는 어절 제거 I

본 절에서는 사전을 검사하여 기준명사에 속하지 않는 모든 어절을 제거한다. 이 부류에 속하는 어절은 크게 수식언과 용언으로 나눈다.

3.1.1. 수식언 제거

수식언의 경우에는 보조사와 결합하지 않는 모든 수식언(부사, 관형사, 감탄사)이 여기에서 제거된다. 이 모듈에서는 사전을 이용해서 제거한다. 기본적으로 모든 수식언이 사전에 포함된 것으로 가정한다. 그러나 실질적으로는 불가능하기 때문에 현재에는 미등록어를 위해서 아주 간단한 경험규칙을 사용하고 있다.

3.2. 용언 제거

용언이 문장에 사용되기 위해서는 일반적으로 용언과 어미로 구성된다. 대부분의 용언에는 명사를 포함하고 있지 않다. 그러나 접미어로 '-하다/되다'를 동반하는 용언에는 명사를 포함하고 있다. 본 논문에서는 용언과 어미의 경계를 명확히 분리하고, 어미를 제외한 부분이 용언이면 제거하고, 명사이면 어미만 제거한다. 불규칙 용언에 대해서는 가능한 모든 변화형을 사전에 등록하여 처리하고 있다. 용언과 어미를 분리하기 위해서 본 논문에서는 후방향-전방향 알고리즘(그림 2)을 사용한다. 이 알고리즘에서는 두 종류의 정보가 이용된다. 하나는 어미를 이룰 수 있는 상호정보 $I(p_i, p_{i+1})$이고, 다른 하나는 용언과 어미로 분리될 확률정보 $P(S \mid pi, p_{i+1})$이며, 식 (3)과 같이 계산된다.

$$P(S \mid p_i, p_j) = \frac{C(p_i S\, p_j)}{C(p_i, p_j)} \qquad (3)$$

여기서 p_i는 특정 음절을 표현하고, p_i는 p_{i+1}에 바로 앞에 오는 음절이다. 확률변수 S는 이진값으로 {+, λ}를 가질 수 있다. , 기호 '+'는 두 음절이 분리됨을 표시하고, 기호 'λ'는 두 음절이 분리되지 말아야함들 표시하는 것이다.

그림 2는 후방향-전방향 알고리즘이다. 후방향-전방향 알고리즘이란 후방향은 오른쪽에서 왼쪽으로 처리함을 의미하는데, 여기서는 어미를 찾기 위해서 후방향으로 처리한다. 전방향은 왼쪽에서 오른쪽으로 처리함으로 의미하고, 여기서는 용언과 어미의 정확한 분리 위치를 파악하기 전방향으로 처리한다. 그림 2에서 len(word)는 입력어절 word의 길이를 구하는 함수이다. 또한 θ_1과 θ_2는 시스템 성능을 조절할 수 있는 매개변수이다. 본 논문에서 언급된 몇 가지의 결과는 θ_1은 0이고, θ_2는 0.3으로 하여 얻은 결과이다.

```
입력: word // 어절
출력: i  // 분리 위치
방법:
    // 오른쪽에서 왼쪽으로 상호정보 값이 어떤
    // 임계값 이하가 되는 위치를 찾는다.
  1. for (i = len(word); i >= 0; i--)
       last if ( I(p_i, p_{i+1}) < θ_1 );
    // 1에서 찾은 위치를 기준으로 다시
    // 오른쪽으로 조사 분리 확률이 어떤
    // 임계값 이상인 위치를 찾는다.
  2. for ( ; i <= len(word); i++)
       last if ( P(S | p_i, p_{i+1}) > θ_2 );
  3. return (i);
```

그림 2. 후방향-전방향 알고리즘

3.3. 조사 분리

체언은 명사구와 조사로 구성된다. 명사를 정확히 찾기 위해서는 명사구와 조사를 분리해야 한다. 이 뤼 위해서도 후방향-전방향 알고리즘을 이용한다. 여기서 $I(p_i, p_{i+1})$은 조사구를 이룰 수 있는 상호정보이고, $P(S \mid p_i, p_{i+1})$은 조사와 명사로 분리될 확률정보이다.

3.4. 기준명사를 포함하지 않는 어절 제거 II

본 절에서는 조사를 분리한 후에 기준명사에 포함되지 않는 어절을 제거하는 방법에 대해서 기술한다. 이 분류에 속하는 어절은 대명사과 의존명사를 포함하는 어절과 보조사와 결합된 수식언, 그리고 수사를 포함하는 어절이다.

3.4.1. 대명사와 의존명사의 제거

대명사와 의존명사는 명사류에 속하지만 기준명사에는 속하지 않는다. 이를 제거하기 위해서는 조사를 분리한 후 나머지 문자열로 사전을 검색하여 사전에 존재하면 어절 전체를 제거한다. 이 부류에 속하는 단어는 극히 제한적이기 때문에 별도의 미등록어 처리 모듈을 사용하지 않는다.

3.4.2. 수식언 제거

수식언을 제거하는 방법은 3.1.1에서 설명한 "기준명사를 포함하지 않는 어절 제거 I"에서 수식언을 제거하는 방법과 동일하다. 단지 사전을 검사할 때 조사를 분리한 후의 나머지 문자열만을 이용한다는 점만 다르다.

3.4.3. 수사의 제거

본 절에서는 명사 중에서 수사를 포함하는 어절을 제거하는 방법에 대해서 기술한다. 수사는 매우 간단한 방법으로 유한 상태 오토마타(finite-state automata)를 이용하였다. 이를 위한 정규표현은 아래와 같은 부분 정규표현을 이용하고 있다.

```
([0-9]+|영|일|이|삼|사|오|육|칠|팔|구)(조|억|만|천|백|십)?
(수)?(천조|백조|십조|조|천억|백억|십억|억|천만|백만|십만|만|천|백|
십)
(열|스물|스무|서른|마흔|쉰|예순|일흔|여든|아흔|백)
(하나|둘|두|셋|넷|다섯|여섯|일곱|여덟|아홉)?
(열|스물|스무|서른|마흔|쉰|예순|일흔|여든|아흔|백)?
(하나|둘|셋|넷|다섯|여섯|일곱|여덟|아홉)
(영|일|이|삼|사|오|육|칠|팔|구)+
네
네댓
두서너
두세
서너
세
한
한둘
.....
```

위와 같은 정규표현을 BASIC라고 하고 단위성 의존명사(nbu)를 NBU라고 할 때, 수사를 제거하기 위한 정규표현을 아래와 같다.

"^ (제)?([0-9]|{BASIC})+({NBU})?"

조사를 제외한 명사 부분이 위의 정규표현에 일치될 때, 수사로 인식한다. 최근 인터넷에는 여러 형태의 단위성 의존명사들이 등장하고 있는데, 이 모듈에서는 이와 같이 새롭게 등장하는 미등록어에 대하서 원활히 대처할 수 없는 실정이다.

3.5. 복합명사 분해

복합명사를 정확하게 분해하기 위해 아래와 같은 휴리스틱(heuristics)을 사용한다.

1. 복합명사를 구성하는 단일 명사는 2-5음절 명사로 가정한다. 1음절은 접사만 가정한다.
2. 분리된 단어의 수가 적은 복합어를 우선한다.

첫번째 휴리스틱은 한국어 복합명사의 대부분이 2음절 명사, 3음절 명사, 4음절 명사로 구성된다는 사실에서 기인된 것이다[6]. 두 번째 휴리스틱은 한국어의 기준명사가 2음절이고 3음절의 대부분은 2음절에 접사와 결합된 명사들이다. 본 논문에서는 이와 같은 특성을 충분히

이용한 것이다. 물론 잘못된 경우도 있었다. 이와 같은 휴리스틱이 반영된 수정된 CYK 파싱 알고리즘[10]을 이용해서 복합명사를 분리한다. 이 알고리즘을 요약하면 그림 3과 같다.

```
입력: 복합명사
출력: 기준명사 리스트
방법:
  1. 전체 단어가 하나의 명사인지를 인식한다.
  2. 수정된 CYK 파싱 알고리즘
    2.1 2음절 명사를 찾아서 T[2, i]에 표시한다.
        여기서 i ≥ 2이다.
    2.2 3음절 명사를 찾아서 T[3, i]에 표시한다.
        여기서 i ≥ 3이다.
    2.3 for j = 2, N
          for i = 1, N-j+1
            for k = 1, j-1
              T[i,j]=select_best(T[i,k]⊕
                    T[i+k, j-k], T[i,j])
  3. T[1,N]을 출력한다.
```

그림 3. 복합명사 분해를 위한 수정된 CYK 알고리즘.

여기서 함수 select_best(•)는 위에서 언급한 두 번째 휴리스틱을 구현한 것이며, 기호 ⊕는 연결연산자(concatenate operator)를 나타낸다. 최종적인 결과는 T[1, N]에 존재한다. 만약 T[1, N]이 NULL이면 사전을 이용해서 복합명사를 분리할 수 없는 경우이다. 따라서 미등록어가 포함될 가능성이 높은 어절 중에 하나이다. 이와 같은 방법의 수정된 CYK 파싱 알고리즘은 정보검색이나 기타 복합명사를 분리해야 하는 곳에서 많이 사용될 수 있을 것으로 생각된다.

4. 실험 및 평가

본 논문에서 음절의 상호정보와 분리확률을 구하기 위해서 사용된 말뭉치는 KAIST 말뭉치[12]이다. 이 말뭉치는 규모는 작지만 정확성이 매우 높기 때문에 학습을 위해서 사용되었다. 평가를 위해서는 1999년에 배포한 ETRI말뭉치[13] 전체를 이용하였다. 본 시스템에 사용한 사전에는 기준명사가 45,060개, 수식언(관형사, 부사 등) 3,911개, 용언이 33,221개를 포함하고 있다. 평가용 말뭉치인 ETRI 말뭉치의 전체 어절 수는 288,291개이고, 그 중에 명사를 포함하는 어절 수는 143,482개이다. 평가용 말뭉치에는 22,651개의 명사와 2,067개의 용언이 미등록어로서 존재한다. 표 1은 평가용 말뭉치에 대한 성능이다. 평균적으로 약 89%의 재현율과 정확율을 보이고 있다. 일반적인 내용을 다루는 뉴스에 대해서는 매우 좋은 결과를 보이고 있으나, 사람이나 대화체를 많이 사용하는 소설에 대해서는 좋은 결과를 보여주지 못했다. 또한 본 시스템에는 한 단어로 구성된 명사에 대해서는 아주 좋지 않은 결과를 가져왔다.

표 1. 제안된 시스템의 성능 평가

분야	재현율	정확율
뉴스	91.71	89.99
뉴스(방송)	92.53	89.91
비소설	88.55	91.33
소설	86.62	82.35
학습서	88.48	91.68
평균	89.58	89.01
F-measure		89.30

5. 다른 방법과의 비교 분석

본 논문에서 제안된 명사추출 알고리즘은 정보요약을 위해서 개발되었으며, 형태소 분석 없이 사용하지 않고 사전과 약간의 통계적인 정보를 이용해서 명사를 추출한다. 따라서 품사 태거와 형태소 분석을 이용하는 경우보다는 정확률면에서는 좋지 않다.

언어분석 도구를 사용하지 않는 명사추출 방법으로는 [1]이 있다. [1]은 학습데이타를 이용해서 명사추출을 위해 규칙을 생성하고, 생성된 규칙과 사전을 이용해서 명사를 추출한다. 사전을 트라이(trie)를 사용하며, 일반적

인 트라이와 좀 달리 복합명사 추정을 위해 학습 동안에 명사로서 서로 겹치는 부분을 표시하는 방법으로 복합명사를 추정한다. 본 논문에서는 복합명사 분리를 위해서 수정된 CYK알고리즘을 사용한다. 성능을 비교해보면 [1]은 재현율이 91%이고 정확률이 77%였다[1]. 재현율 면에서는 더 좋은 결과를 보였다. 그러나 정확률은 제안된 시스템이 훨씬 더 좋았다. 이를 F-measure로 비교해보면 [1]은 83.42인 반면에 제안된 시스템은 89.30을 보였다.

복합명사 분해에 관한 연구들 중에서 가장 유사한 방법은 [6]이다. [6]은 1음절 접사 빈도수, 2음절 또는 3음절 기준명사가 복합명사 내에서 사용된 위치정보와 빈도수를 이용한 CFP라고 하는 통계정보와 중의성이 발생되었을 경우 명사의 개수가 최소가 되는 분해를 선호하는 휴리스틱 규칙 MNPR을 사용하고 있다. 복합명사 분해를 위해서는 사용된 휴리스틱은 [6]에서 사용한 휴리스틱을 그대로 사용하였으며 특별한 통계정보를 사용하지 않고 있다.

6. 결론

본 논문은 여과 기법과 분리 기법을 이용한 한국어 명사 추출 방법을 제안하였다. 여과 기법은 명사를 포함하지 않는 어절, 즉, 용언, 수식언, 독립언들을 미리 제거하는 방법이다. 특히 용언을 제거하기 위해서는 어절의 마지막 두 음절정보를 이용해서 결정하고, 다른 나머지는 사전에 의해서 결정된다. 분리 기법은 체언에서 명사구와 조사를 분리하기 위한 방법과 복합 명사를 분리하기 위해서 사용된다. 명사구와 조사를 분리하기 위해서는 후방향-전방향 알고리즘을 사용하고, 복합명사를 분리하기 위해서는 수정된 CYK알고리즘을 사용한다.

본 논문에서 제안된 기준명사 추출 방법은 ETRI 말뭉치를 대상으로 약 89%의 재현율과 정확률을 보였으며 한국어 정보요약 시스템[11]에 적용했을 때, 좋은 결과를 보였다.

그러나, 아직 개선되어야 할 문제를 많이 안고 있다. 용언을 제거하기 위한 충분한 자질(feature) 개발과 수식언 및 독립언을 제거하기 위한 새로운 자질을 구하는 문제에 대해서도 충분히 연구할 가치가 있다고 생각된다. 또한 복합명사 분리하는 방법도 통계적인 방법을 CYK 알고리즘에 적용하는 방법도 충분이 연구할 가치가 있는 것으로 판단된다.

[1] 이 결과는 ETRI 말뭉치에서 평가용으로 분류된 약 33,000개의 어절에 대해서 수행한 결과이다.

7. 감사의 글

본 연구는 첨단정보기술 연구센터를 통하여 과학재단과 지원을 받았으며, 또한 과학기술부 STEP2000 프로젝트에 의해 지원되고, 전문용어언어공학연구센터에 의해 수행중인 "대용량 국어정보 심층처리 및 품질관리 기술개발" 연구과제의 일환으로 수행되었습니다.

참고 문헌

[1] 장동현, 맹성현, "학습데이터를 이용하여 생성한 규칙과 사전을 이용한 명사추출기," 제1회 형태소분석기 및 품사태거 평가 워크숍 발표논문집, pp. 151-156, 1999.

[2] 김재훈, 선충녕, 홍상욱, 이성욱, 서정연, 조정미, "KTAG99: 새로운 환경에 쉽게 적응하는 한국어 품사 태깅 시스템," 제1회 형태소분석기 및 품사태거 평가 워크숍 발표논문집, pp. 99-105, 1999.

[3] 안동언, "좌우접속정보를 이용한 명사추출기," 제1회 형태소분석기 및 품사태거 평가 워크숍 발표논문집, pp. 173-178, 1999.

[4] 이중영, 신병훈, 이공주, 김지은, 안상규, "COM기반의 다목적 형태소 분석기를 이용한 명사추출기," 제1회 형태소분석기 및 품사태거 평가 워크숍 발표논문집, pp. 167-171, 1999.

[5] 최재혁, "형태소 분석을 통한 한영 자동 색인어 추출," 정보과학회논문지(B), 제23권, 제12호, pp. 1279-1288, 1996.

[6] 윤보현, 조민정, 임해창, "통계정보와 선호 규칙을 이

용한 한국어 복합 명사의 분해," 정보과학회논문지 (B), 제24권, 제8호, pp. 900-909, 1997.

[7] 박혁로, 신중호, "비터비 학습 알고리즘을 이용한 한글 복합명사 분석," 1997 한국정보과학회 가을 학술 발표논문집, vol 24, no. 2, pp. 219-222, 1997.

[8] 강승식, "한국어 복합명사 분해 알고리즘," 정보과학회논문지(B), 제25권, 제1호, pp. 172-182, 1998.

[9] 최재혁, "음절수에 따른 한국어 복합명사 분리 방안," 제8회 한글 및 한국어 정보처리 학술대회 발표논문집, pp. 262-267, 1996.

[10] Aho, V. A. and Ullman, J. D. (1972) *The Theory of Parsing, Translation, and Compiling*, Prentice-Hall.

[11] 김재훈, 김준홍, 도합유사도를 이요한 한국어 문서 요약 시스템, 한국해양대학교, 컴퓨터공학과, 기술문서 KMU-NLP-TR-2000-003, 2000.

[12] 김재훈, 김길창, 한국어에서의 품사 부착 말뭉치의 작성 요령 : KAIST 말뭉치, 한국과학기술원, 전산학과, 기술문서, CS/TR-95-9, 1995.

[13] 이현아, 이원일 임선숙, 허은경, 이재성, 차건희, 박재득, 표준안에 따른 품사 부착 말뭉치 구축, 제1회 형태소 분석기 및 품사 태거 평가 워크숍 발표 논문집, pp. 40-43, 1999.