Journal of the Research Institute of Industrial Technology, Vol. 18, 2001

한국어 고유명사 추출

김재훈

Proper Noun Extraction from Korean Text

J. H. Kim

Abstract

Proper nouns are highly important in information extraction, which is the problem of generating stereotypic summaries from free text. Traditional information extraction is performed on journalistic or technical documents and involves some linguistic pre-processing. In many domains, however, linguistic processing is difficult, if not impossible. Then, we propose the model for extracting proper nouns from unrestricted Korean text without linguistic processing like morphological analysis and part-of-speech tagging. The model consists of six phases: 1) eliminating non-noun word phrases (Korean words between two white spaces) using the backward-forward algorithm, which is proposed by the author, 2) segmenting a word phrase into a noun and zero or more particles using the backward-forward algorithm too, 3) filtering out nouns excluding proper nouns like numerals and pronouns using finite-state automata, 4) recognizing transcribed foreign words in Korean using hidden Markov model, 5) segmenting a compound noun into individual nouns using the modified CYK parsing algorithm, 6) recognizing proper nouns using some rules, such as regular expressions, and database for proper nouns. We have implemented the proper noun extraction system, of which performance in closed evaluations is not bad. The evaluation is not objective because we do not have the tagged corpus for proper nouns in Korean. We expect that this model is also very useful for unknown word processing in morphological analysis as well as for robust term extraction in information extraction and information retrieval.

한국해양대학교 기계·정보공학부

1. 서 론

최근 인터넷의 발달로 사이버 공간 내에는 수많은 정보(예를 들면, 개인 홈페이지, 연구보고서, 제품 설명서, 정보 발표 등)가 산재되어 있다. 사이버 공간으로부터 효율적이고 효과적으로 정보를 찾기 위해서 알타비스타(AltaVista), 야후(Yahoo) 등과 같은 정보검색 엔진이 많이 사용되고 있다. 그러나 이들 시스템의 대부분은 지나치게 많은 문서를 검색하기 때문에 정작 검색된 문서로부터 필요한 정보를 찾는 것이 또 다른 큰 문제를 야기시키고 있다. 이를 보완하기 위해 검색된 문서를 재분류하는 문서분류, 검색된 문서로부터 필요한 정보를 추출하는 정보추출, 방대한 문서를 요약하는 문서요약과 같은 기술을 사용하고 있다[1-3].

정보추출이란 문서로부터 구조화된 요약문을 생성하는 것이며, 정보추출 시스템은 특정한 형태의 정보를 추출하기 위해 제한되지 않는 문서를 분석한다[3-4]. 정보추출 시스템은 모든 문장을 이해할 필요는 없으나, 관련된 정보가 포함된 문서의 특정 부분을 집중적으로 분석한다. 관련이 있다는 것은 일반적으로 미리 정의되기 때문에 시스템이 어떤 정보를 찾아야 하는지를 알고 있다. 정보추출은 데이터마이닝과 비슷하게 자동적으로 데이터베이스를 구축할 수 있다[5]. 정보추출 시스템은 제한되지 않는 문서로부터 데이터베이스의 항목을 추출하여 데이터베이스의 특정 항목에 할당하면 된다.

제한되지 않은 문서에서 필요한 정보를 추출하기 위해서는 정보의 대상이 되는 단위를 인식해야 한다. 그 대상의 단위가 주로 명사구이며, 한국어에서 명사류에 속하는 품사는 보통명사, 고유명사, 수사, 의존명사, 대명사가 있으나, 일반적으로 정보추출에서는 의존명사와 대명사는 그 대상에서 제외된다. 명사구(주로 기준명사(base noun phrase)) 인식에 대한 연구는 어느 정도 진행되었으며, 좋은 결과를 보이고 있다[6-7]. 고유명사구 인식 문제는 보통명사구 인식 문제에 비해 훨씬 더 복잡하며, 미등록어의 대부분은 고유명사이기 때문에 자연언어처리에서 고질적인 문제 중 하나인 미등록어 처리와도 밀접한 관계를 가지고 있다.

일반적으로 미등록어를 추정하는 방법은 형태소 분석이 불가능한 어절에 대해서 미등록어 추정을 시도하고, 그 미등록어의 가능한 분석 중 하나가 고유명사가 된다. 이 방법은 형태소 분석이 실패되지 않는 한 미등록어를 발견할 수 없게 되는데 이 현상을 형태소 과분석이라고 한다. 형태소 과분석 현상은 주로 단음절 명사에서 발생된다. 예를 들면, "인터넷"이라는 단어가 사전에 없고 개개의 음절인 "인", "터", "넷"은 사전에 있다고 가정하면, "인터넷"에 대한 형태소 분석은 "인"+"터"+"넷"으로 분리된 복합명사로 분석된다. 본 논문에서 이와 같은 문제를 효과적으로 해결하기 위해서 형태소 분석을 수행하기 전에 기존의 사전과 어휘문맥에 의해서 고유명사를 미리 추정하는 방법을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서 한국어 문장에서 고유명사의 특성에 대해서 살펴보고, 3장에서 문서로부터 고유명사 추출 방법을 기술한다. 4장에서 고유명사 추출 시스템의 성능 평가하고, 다른 고유명사 추출 시스템과 비교 분석하고, 5장에서 결론을 맺고자 한다.

2. 한국어 고유명사구의 특성

고유명사는 사람, 강, 산, 나라, 도시, 단체, 상점, 상품, 명절 등의 이름을 가리키는 단어이다. 예를 들면, "이순신 장군은 우리나라를 지킨 훌륭한 인물이다."라는 문장에서 "이순신"은 사람 이름에 속하는 고유명사이다. 영어의 경우, 고유명사는 첫 번째 글자를 대문자로 시작한다(예: United States). 이것은 고유명사 인식에서 이용될 수 있는 중요한 정보이다. 반면에 한국어의 경우, 고유명사에 대한 형태론적인 특성이 거의 없기 때문에 형태론적인 정보 이외에 문맥 정보를 이용하지 않으면 안된다. 앞의 예를 보면, "이순신"이라는 단어가 사람 이름의 조건을 만족하고 다음에 직책이나 지위에 해당하는 단어가 나오기 때문에 앞 단어("이순신")는 고유명사일 가능성이 대단히 높다. 이러한 문맥 정보는 고유명사의 종류에 따라 매우 다르므로 각 고유명사의 종류에 따른 그 특성을 살펴보고자 한다. 그러나, 고유명사의 종류가 너무나 다양하기 때문에 사람 이름, 기관 이름, 지역 이름에 대해서 간단히 살펴보고자 한다.

한국어에서 사람 이름은 대부분이 3자로 구성되었으며, 보통 2자에서 4자로 구성된다. 한국에서 사용되는 성은 약 140여 개이고, 그 중에 "남궁", "선우" 등과 같은 2자로 구성된 성은 8개이다. 이와 같은 정보는 사람 이름을 인식하는 데 아주 유용한 정보이다. 또한 신문, 잡지 등에서는 사람의 이름이 처음으로 소개될 경우, 성과 이름을 모두 표기하고 다음에 직책이나 지위를 소개하거나, 괄호를 이용해서 한자명, 나이, 성별을 소개하는 것이 일반적이다. 그리고 나서는 특별한 실마리 단어 없이도 그 사람 이름을 사용한다. 아래의 예문은 고유명사인 사람 이름의 용례이다.

> "올 시즌을 '최고의 해'로 장식한 최용수(LG, 27)가 이적료와 연봉 등 총 3억엔(약 32억원)에 일본 J리그의 제프 유나이티드 이치하라로 이적한다."

또한 신문이나 잡지 등에는 외국인의 이름을 음차해서 많이 표기하는데 이 경우는 일반적으로 문맥은 비슷하나 이름의 구성이 2음절 이상일 경우도 있고, 4자 이상인 경우가 상당히 많은데, 이 경우는 문서에서 음차된 외국어를 인식한다면 많은 도움이 될 것이다. 아래의 예문에서 밑줄친 부분은 음차된 외국어의 용례이다.

> "타이거 우즈(25)가 미국의 스포츠주간지 스포츠일러스트레이티드(SI)가 선정하는 '올해의 스포츠맨'에 13일(한국시각) 뽑혔다.

기관 이름이나 회사 이름은 "한국해양대학교", "현대화재해상보험", "(주)쓰리소프트" 등과 같은 형태로 사용되는데, 여기에서도 "-대학교", "-보험", "(주)-", "-(주)", "-주식회사" 등의

실마리 단어들이 많이 있는데, 본 논문에서는 이를 위해서 461개의 실마리 단어를 사용하며, 그 일부를 부록 1에 실었다.

지역 이름에는 강, 산, 도시, 나라 등의 이름이 있다. 이 경우에는 특별한 실마리 문맥은 존재하지 않는다. 물론 의미적으로 분석이 되었을 경우에는 충분한 실마리가 될 수 있으나, 본 논문의 취지와는 잘 맞지 않는다. 그러나 "백두산", "낙동강", "서울시" 등에서 보는 바와 같이 "-산", "-강", "-시"와 같은 접미사가 실마리가 될 수 있다. 도시나 마을 동네의 이름 같은 경우에는 여러 가지의 접미사가 필요한데, 예를 들면, "-광역시", "-특별시", "-시", "-군", "-동" "-리", "-부락", "-읍" 등 여러 가지가 있다. 나라 이름의 경우에도 특별한 실마리 문맥이 없다. 그러나 전 세계의 나라 이름은 크게 많지 않기 때문에 고유명사 사전을 이용할 수 있다. 본 논문에서는 291개의 나라 이름을 사용하며 그 일부를 부록 2에 실었다. 다만 이 경우 같은 나라를 여러 가지 이름으로 표기하는 경우가 있는데 이를 위해서 본 논문에서는 유한상태오토마타를 이용해서 해결한다. 예를 들면, "남아프리카공화국"은 "남아공", "남아공화국", "남아프리카" 등으로 표기되며 심지어는 "남화공화국"이라고 잘못 표기되는 경우도 매우 자주 발생하고 있다.

그 밖의 명절 이름이나 작품 이름 등에 대해서도 실마리가 될 만한 특별한 문맥을 가지고 있지 않으며, 개별적으로 독특한 문맥을 가질 수 있다. 본 논문에서는 이들 모든 가능한 문맥으로 찾아서 이용하고 있지는 않으며 기본적인 모형을 제시하고 있다. 이 문제는 앞으로 자동 문맥 추출에 관한 연구를 수행함으로써 보완될 수 있을 것으로 생각한다. 또한 자주 사용되는 많은 고유명사는 특별한 문맥 없이 보통명사와 거의 동일하게 사용되고 있으며, 이들 대부분은 고유명사 사전을 이용해서 처리되며, 이 사전은 처리 중에 찾아진 고유명사를 즉시 반영한다(캐쉬 기능).

3. 고유명사 추출 모델

본 논문에서 제안한 고유명사 추출 과정은 그림 1과 같으며, 각 단계에 대한 상세한 설명은 이하의 절에서 기술된다. 그림 1에서 보는 바와 같이 고유명사 추출 시스템의 입력은 형태소 분석이 되지 않은 문장 자체이며(부록 5 참조), 출력은 형태소 분석기나 기타 자연언어처리 시스템의 입력이 되기 때문에 약간의 프로토콜이 필요하다(부록 5 참조). 본 논문에서는 고유명사 추출 시스템에 의해서 인식된 고유명사는 "(@Q@PROPER_NOUN@)"과 같은 프로토콜을 사용한다. 예를 들면 "동아일보사"라고 하는 고유명사가 인식되었다면, 출력은 "(@Q@동아일보사@)"가 된다. 그림 1에서 사전은 거의 모든 단계에서 사용되며 형태소 분석에서 사용되는 것과 동일하다. 그 밖의 필요한 자료에 대한 설명은 해당되는 단계를 설명할 때 자세히 기술될 것이다. 본 논문에서 제안한 고유명사 추출 모델은 개념적으로 다음과 같은 세 가지의 기능으로 이루어진다.

첫째 기능은 여과 기능(filtering function)이다. 여과 기능은 비명사구 어절을 처리 대상에서 제외시킨다. 왜냐 하면 명사구가 될 수 없는 어절에는 고유명사가 포함될 수 없기 때문이다. 한국어에서 비명사구에 속하는 어절은 용언(동사, 형용사), 수식언(부사, 관형사), 독립언(감탄사)이

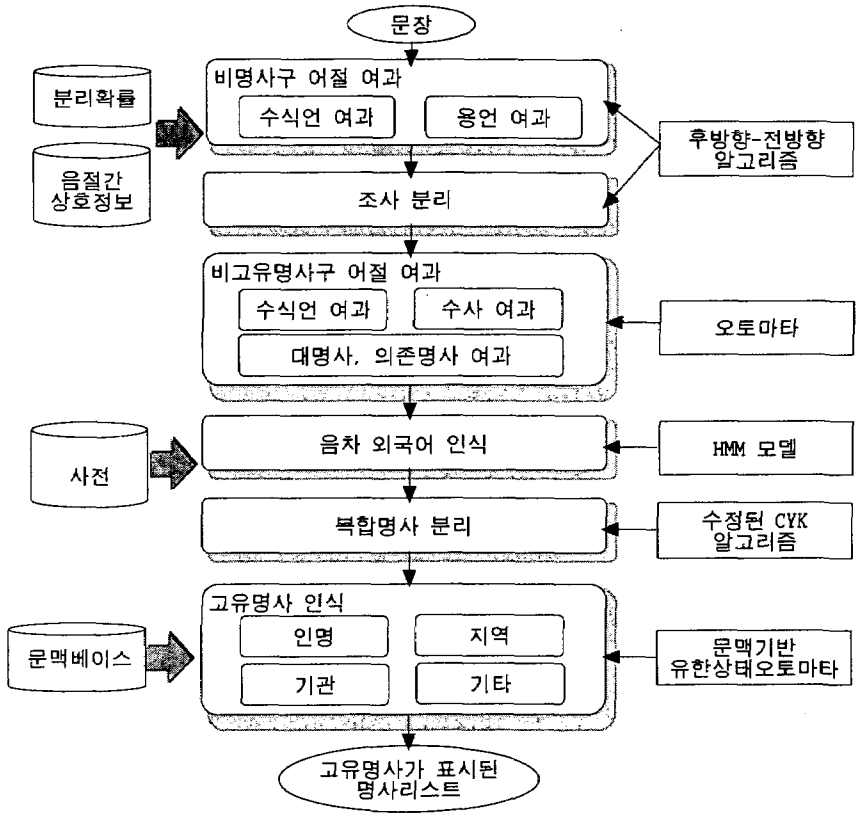

그림 1. 고유명사 추출 시스템의 흐름도

며, 이들에 속하는 어절은 일단 고유명사를 포함하지 않으므로 고려 대상에서 제외된다(비명사구 어절 여과 단계). 명사구에 속하는 어절이라고 하더라도 대명사, 의존명사, 수사를 포함하는 어절은 고유명사를 포함하지 않기 때문에 여과 단계에서 제외한다(비고유명사 어절 여과 단계).

둘째 기능은 분리 기능(segmentation function)이며, 이 기능으로 조사와 복합명사를 분리한다. 주어진 어절에 조사가 있다면 이를 먼저 분리하고, 그리고 나서 일반 사전을 이용해서 복합명사를 분리한다. 많은 고유명사는 보통명사의 결합으로 이루어진다. 예를 들면, 고유명사 "한국정신문화원"은 "한국+정신+문화원"으로 이루어진다. 이는 다음 단계인 고유명사 인식 단계에서 여러 가지 문맥 규칙으로 쉽게 적용하기 위해서 먼저 분리한다.

셋째 기능은 인식 기능(recognization function)이며, 이 기능으로 고유명사를 인식한다. 고유명사를 효과적으로 인식하기 위해서는 먼저 음차된 외국어를 인식하고, 그리고 나서 그 정보를 이용해서 사람 이름을 인식하고, 그 다음에 기타 고유명사를 인식한다. 고유명사는 문맥정보와 고유명사가 될 수 있는 여러 형태의 정규표현을 이용해서 고유명사를 인식한다.

3.1. 비명사구 어절의 여과

명사구가 될 수 없는 어절을 비명사구라고 하며, 한국어에서 비명사구 어절은 용언(동사, 형용사), 수식언(부사, 관형사), 독립언(감탄사)이 있다. 비명사구 어절은 고유명사를 포함하지 않으므로 고려 대상에서 제외된다. 비명사구 어절을 제외시키기 위해서 주어진 어절이 비명사구 어절인지를 인식해야 한다. 본 논문에서는 첨가(agglutination)의 정도에 따라 용언(동사, 형용사)과 비용언(수식언, 독립언)으로 나누어서 처리한다. 용언에는 첨가현상이 매우 심하게 발생되므로 용언에 속하는 모든 어절을 사전에 등재하는 것은 거의 불가능하다. 반면에 수식언에는 첨가현상이 거의 발생되지 않으므로 거의 모든 어절을 사전에 등재할 수 있다.

3.1.1. 수식언 여과

보조사와 결합하지 않는 모든 수식언(부사, 관형사, 감탄사)이 여기에서 제외된다. 이를 위해서 사전을 이용한다. 기본적으로 모든 수식언이 사전에 포함된 것으로 가정한다. 그러나 실질적으로는 불가능하기 때문에 현재에는 미등록어를 위해서 아주 간단한 경험규칙을 사용하고 있다.

3.1.2. 용언 여과

용언이 문장에 사용되기 위해서는 일반적으로 용언과 어미로 구성되는데(첨가현상), 본 논문에서는 용언과 어미의 경계를 명확히 분리하고, 어미를 제외한 부분이 용언이면 제외시킨다. 불규칙 용언에 대해서는 가능한 모든 변화형을 사전에 등록하여 처리한다. 용언과 어미를 분리하기 위해서 본 논문에서는 후방향-전방향 알고리즘(그림 2)을 사용한다. 이 알고리즘은 두 종류의 정보를 사용하는데, 하나는 어미를 이룰 수 있는 상호정보 $I(p_i, p_{i+1})$이며, 식 (1)과 같이 구하고, 다른 하나는 용언과 어미로 분리될 확률 $\Pr(S|p_i, p_{i+1})$이며, 식 (2)와 같이 계산된다.

$$I(p_i, p_{i+1}) = \log \frac{\Pr(p_i, p_{i+1})}{\Pr(p_i)\Pr(p_{i+1})} \tag{1}$$

$$\Pr(S|p_i, p_{i+1}) = \frac{C(p_i, S, p_{i+1})}{C(p_i, p_{i+1})} \tag{2}$$

여기서 p_i는 특정 음절을 표현하고, 확률변수 S는 이진값 $\{+, \lambda\}$를 가질 수 있다. 기호 '+'는 두 음절 p_i, p_{i+1} 사이에서 용언과 어미로 분리되는 경우이고, 기호 'λ'는 그 두 음절 사이에서 분리되지 않는 경우이다. 후방향-전방향 알고리즘에서 후방향 처리는 오른쪽에서 왼쪽으로 처리하면서 어미를 찾고, 전방향 처리는 왼쪽에서 오른쪽으로 처리하면서 용언과 어미의 정확한 분리 위치를 결정한다. 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

```
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 ( Pr(S|p_i, p_{i+1}) > θ_2);
  3. return (i);
```

그림 2. 후방향-전방향 알고리즘

θ_1은 0이고, θ_2는 0.3으로 하여 얻은 결과이다.

3.2. 조사 분리

체언은 명사와 조사로 구성된다. 명사를 정확히 찾기 위해서는 명사와 조사를 분리해야 한다. 이를 위해서도 후방향-전방향 알고리즘을 이용한다. 여기서 $I(p_i, p_{i+1})$은 조사를 이룰 수 있는 상호정보이고, $\Pr(S|p_i, p_{i+1})$은 두 음절 p_i, p_{i+1} 사이에서 명사와 조사로 분리될 확률이다.

3.3. 비고유명사구 어절 여과

비고유명사구란 명사구 중에서 고유명사를 포함하지 않은 명사구를 의미한다. 비명사구에 속하는 어절은 대명사와 의존명사를 포함하는 어절과 수사를 포함하는 어절이다. 또한 본 절에서 보조사와 결합된 수식언를 제거하는 방법도 함께 다룬다.

3.3.1. 대명사와 의존명사의 여과

대명사와 의존명사는 명사구에 속하지만 고유명사는 포함하지 않는다. 이를 제외시키기 위해서 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.

3.3.2. 수사 여과

본 절에서는 명사 중에서 수사를 포함하는 어절을 제거하는 방법에 대해서 기술한다. 수사는 매우 간단한 방법으로 유한상태오토마타(finite-state automata)를 이용한다. 아래는 유한상태오토마타를 구현하기 위한 정규표현의 일부이다.

```
([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
...
```

위와 같은 정규표현을 BASIC라고 하고, 단위성 의존명사(예: 1000원, 1개, 집 한 채 등)를 NBU라고 할 때, 수사를 제거하기 위한 정규표현은 아래와 같다.

```
"^(제)?({BASIC})+({NBU})?"
...
```

조사를 제외한 명사 부분이 위의 정규표현에 일치될 때, 수사로 인식된다. 수사로 인식된 어절에는 고유명사가 존재하지 않는 것으로 가정한다.

3.3.3. 수식언 제거

수식언을 제거하는 방법은 3.1.1에서 설명한 비명사구 여과에서 수식언을 제거하는 방법과 동일하다. 단지 사전을 검사할 때 조사를 분리한 후의 나머지 문자열만을 이용한다는 점만 다르다.

3.4. 음차 외국어 인식

최근 신문 기사 등에는 음차된 외국어가 상당히 많은 부분을 차지한다. 아래의 예문은 조선일보 1999년 11월 23일자 스포츠 면의 일부이다.

> "샌안토니오는 23일(한국시간) 필라델피아 퍼스트 유니온센터에서 열린 미프로농구(NBA) 정규리그 원정경기에서 팀 덩컨-데이비드 로빈슨 '트윈타워'를 앞세워 앨런아이버슨이 버틴 필라델피아를 94-91로 제압했다."

위의 예문에서 밑줄 친 부분은 음차 외국어로 인식되어야 하고, 음영으로 표시된 부분은 고유명사로 인식되어야 한다. 이 예에서 보는 바와 같이 음차 외국어의 대부분은 고유명사이다. 따라서 음차 외국어를 인식하는 것은 고유명사를 인식하는 데 큰 도움을 줄 것으로 생각된다. 본 논문에서는 [8]의 모델을 수정 보완하였다. [8]에서는 외래어 인식 문제를 HMM으로 모델링하였으며, 다음과 같은 세 가지 정보를 이용한다.

1. 어휘정보: bigram 어휘정보, unigram 어휘정보
2. 전이정보: trigram 전이정보, bigram 전이정보
3. 초성종성 어휘정보: bigram 초성종성 어휘정보, unigram 초성종성 어휘정보

예를 들어, "객체지향시스템에서"라는 어절에서 외래어를 인식한 최종결과는 "KKKKEEEKK"로 표시된다. 따라서 전이 bigram은 (KK) (KK) (KK) (KE) (EE) (EE) (EK)(KK)로 표현되고, trigram 정보는 (KKK), (KKK) … (EKK)와 같이 표현된다. 어휘정보

는 unigram의 경우 (객K) (체K) … (템E)(에K)(서K)로 표현되고, bigram의 경우에는 (객체KK) (체지KK) … (스템EE) (템에EK) (에서KK)로 표현된다. 또한 초성종성 정보가 (ㄱㄱ)(ㅊ0)(ㅈ0)(ㅎㅇ)(ㅅ0)(ㅌㅁ)(ㅇ0)(ㅅ0) KKKKEEEKK으로 나타내어질 수 있는데, 이에 대한 초성 종성 bigram 의 경우 (ㄱㄱ)(ㅊ0)KK … (ㅇ0)(ㅅ0)KK의 형식이 되며, unigram의 경우에선 (ㄱㄱ)K … (ㅅ0)K의 형식이 된다. 본 논문에서 모델링을 위해서 몇 가지 표기법을 사용한다. t_i는 tag(K, F, E, S, $)를 의미하고, p_i는 음절을 의미하고, h_i는 초성종성 정보를 의미한다. 본 논문에서는 이 모델에서 아래와 같은 내용들을 수정하였다.

1. 태그를 네 가지를 사용한다.
   K: 한글을 표현하는 태그이다.
   E: 음차한 한글을 표현하는 태그이다.
   S: 기호나 영어 숫자 등 ASCII문자에 대한 태그이다.
   $: 어절의 시작과 끝을 표현하는 태그이다.

2. 모델에서 bigram정보와 trigram정보를 선형결합(linear combination) 한다.
   음절정보의 선형 결합: $\prod_{i=1}^{N}\{k\Pr(p_i|t_i)+(1-k)\Pr(p_i|t_{i-1},t_i)\}$
   초성종성 정보의 선형결합: $\prod_{i=1}^{N}\{k\Pr(h_i|t_i)+(1-k)\Pr(h_i|t_{i-1},t_i)\}$
   여기서 0<k<1이며, 일반적으로 k는 0.05정도의 값을 가진다.

3. bigram 어휘정보를 온전하게 사용한다.
   음절정보의 경우: $\prod_{i=1}^{N}\{k\Pr(p_i|t_i)+(1-k)\Pr(p_{i-1}p_i|t_{i-1},t_i)\}$
   초성종성 정보의 경우: $\prod_{i=1}^{N}\{k\Pr(h_i|t_i)+(1-k)\Pr(h_{i-1}h_i|t_{i-1},t_i)\}$

이를 종합하면 음차 외국어를 인식하는 HMM(hidden Markov model)은 식 (3)과 같다.

$$P(T|W)P(W)=\prod_{i=1}^{N}\Pr(t_i|t_{i-2},t_{i-1}) \times\prod_{i=1}^{N}\{k\Pr(p_i|t_i)+(1-k)\Pr(p_{i-1}p_i|t_{i-1},t_i)\} \times\prod_{i=1}^{N}\{k\Pr(h_i|t_i)+(1-k)\Pr(h_{i-1}h_i|t_{i-1},t_i)\} \quad (3)$$

이 모델은 [8]에 비해 자료희귀성(data spareness) 문제가 더 심하나, 정확성 면에서는 더 좋은 결과를 보였다. 앞에서 언급한 예에 대해서 아래와 같은 결과를 보였다. 밑줄로 표시된 부분이 외래어 추정을 통해 고유명사로 인식된 부분이다.

"샌안토니오는 23일(한국시간) 필라델피아 퍼스트 유니온센터에서 열린 미프로농구(NBA) 정규리그 원정경기에서 팀 덩컨-데이비드 로빈슨 '트윈타워'를 앞세워 앨런아이버슨이 버틴 필라델피아를 94-91로 제압했다."

여기서 "필라델피아, 유니온센터, 팀, 데이비드, 로빈슨"의 경우에는 이미 사전을 이용해서 고유명사로 인식되었기 때문에 음차 외국어를 통한 추정 루틴을 실행하지 않았다. 그리고 "덩컨"의 경우에는 외래로 인식되어야 하는데 제대로 인식되지 않았다. 아마도 학습 데이터에 이를 위한 충분한 문맥이 없는 것으로 판단된다. 따라서 처리 대상의 단어 중에 "덩컨"을 제외하고는 정확하게 분리했을 뿐 아니라 외래어를 정확하게 찾았다.

3.5. 복합명사 분리

복합명사를 분리하기 위해 몇 가지 휴리스틱(heuristic)을 사용하며, 그 휴리스틱은 아래와 같이 요약된다.

1. 복합명사를 구성하는 기준 명사는 2-5 음절 명사이다.
2. 분리된 단어의 수가 적은 복합어를 우선한다.

첫 번째 휴리스틱은 한국어 복합명사의 대부분이 2음절 명사, 3음절 명사, 4음절 명사로 구성된다는 사실에서 기인된 것이다[9]. 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: 명사 리스트
방법:
  1. 전체 단어가 하나의 명사인지를 인식한다.
  2. 수정된 CYK 파싱 알고리즘
    2.1 2~5음절 명사를 찾아서 T[2~5, i]에 표시한다.
    2.2 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. 수정된 CYK 파싱 알고리즘

여기서, 함수 select_best(.)는 위에서 언급한 두 번째 휴리스틱을 구현한 것이다. 최종적인 결과는 T[1, N]에 존재한다. 만약 T[1, N]이 NULL이면 사전을 이용해서 복합명사를 분리할 수 없는 경우이다. 따라서 고유명사가 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.

3.6. 고유명사 인식

3.6.1. 사람 이름 인식

본 절에서는 사람 이름에 해당하는 고유명사 인식 방법에 관해서 기술된다. 사람 이름의 문맥은 정규표현으로 표현되며, 아래는 그 정규표현의 일부와 그 예를 나열하고 있다.

```
{NAME}([:blank:])?{ORGANIZATION}?{JOB}
    이의근 문화액스포 회장+이
    김우중 대우그룹 회장+이
    ...

{NAME}[씨|군|양|님]
    권+씨+가
    진승현+씨+는
    ...

{NAME}}([:blank:])?{JOB}
    올브라이트 국무장관+과
    박준규 국회의장
    ...

{ORGANIZATION}}([:blank:])?{NAME}}([:blank:])?{JOB}
    강남성모병원 남궁성+은
    국민회의  김영환 의원+도
    ...

{NAME}([:blank:])?{COUNTRY}?{JOB}
    오부치 일본 총리+와
    한 일본 대사+는
    ...
```

여기서 {.}는 정규표현에서 미리 정의된 정규표현을 참조한다는 의미이고, [:blank:]는 공백(white space)을 의미한다. 따라서 {NAME}에 대응하는 곳이 사람의 이름을 나타내는 부분이다. 사람의 이름은 성과 이름으로 나뉘어진다. 이름은 한글이나 한자가 두 자까지 올 수 있도록 정규표현에 의해서 표현되었다. {ORGANIZATION}은 기관이나 조직에 해당하며, 본 논문에서는 2349개를 사용하며, 그 일부는 부록 3에 실었다. 본 논문에서 {JOB}에 해당하는 단어는 183개이며, 그 일부를 부록 4에 실었으며, {COUNTRY}에 해당하는 단어는 283개이고 그 일부를 부록 2에 실었다.

3.6.2. 기타 고유명사의 인식

사람이름 이외의 고유명사는 실마리 단어와 약간의 패턴, 즉, 정규표현에 의해서 인식된다. 본 논문에서는 이 부분에 대한 충분한 자료를 확보하지 못해서 원형(prototype)을 구현하는 데 만족하였다. 대부분의 {ORGANIZATION}은 고유명사가 될 수 있기 때문에 사람이름을 인식할 때, 사용하던 {ORGANIZATION}을 위한 패턴(정규표현)을 그대로 사용할 수 있다. 이외에 추가로 필요한 실마리 문맥을 실마리 사전에 추가하였는데, 이 내용의 일부를 부록 1에 실었다. 실마리 단어에 의해서 인식된 단어의 예를 아래에서 보여주고 있다. 여기서 밑줄을 그은 부분이 실

마리 단어에 의해서 인식된 단어이다.

4. 실험 및 토의

본 논문에서 제시한 고유명사 추출 모델은 형태소분석이나 품사 태깅과 같은 언어처리 도구를 사용하지 않는다. 제시된 모델은 C언어로 구현되었으며, 입력은 문장이고(부록 5 참조), 출력은 고유명사가 표시된 문장이다(부록 6 참조). 출력은 자연언어처리 시스템이나 정보추출 시스템의 수정없이 그대로 사용할 수 있도록 원래 문장에 약간의 태그(tag)를 부착하여 사용한다. 시스템은 고유명사 이외에도 명사나 음차된 외국어를 선택사항(option)으로 인식할 수 있도록 하였다. 따라서 명사만을 색인어로 간주하는 정보검색 시스템이나 명사 리스트를 이용한 정보요약 시스템에서도 사용할 수 있도록 하였다. 본 장에서는 구현된 시스템에 대한 간단한 평가를 수행하고 구현된 시스템의 문제점과 기존의 시스템[12]에 대해서 기술한다.

4.1. 성능 평가

한국어에는 객관적으로 비교할 수 있는 고유명사 태깅 말뭉치가 없으므로 제안된 시스템에 대해서 객관적인 성능을 평가할 수 없었다. 다만 본 논문에서 구현된 시스템을 자연언어처리 시스템에 적용했을 경우, 성능이 개선될 수 있는지를 살펴보았다. 성능 평가를 위해 사용된 학습 말뭉치와 시험 말뭉치의 통계치는 표 1과 같다.

표 1 성능 평가를 위한 말뭉치의 통계치

학습 말뭉치				시험 말뭉치		
문장 수	어절 수	형태소수	모호성 정도	문장 수	어절 수	형태소 수
16,194	175,526	379,133	2.15	1,576	19,610	42,312

본 논문에서 개발된 시스템의 성능은 표 2와 같으며, 기본적으로 미등록어가 있다는 가정을 한다. 일반적으로 품사 태깅 시스템에 대한 평가는 미등록어를 고려하고 있지 않은 경우가 대부분이다. 그러나, 본 논문에서는 고유명사 추정이 미등록어에 어떤 영향을 끼치는지를 살펴보기 위해서 미등록어를 고려했다. 여기서, 오류축소율은 기존 시스템[11]을 기준으로 한 것이다.

표 2 고유명사 추정을 고려한 품사 태깅 시스템의 성능

시스템 종류	형태소 정확률	오류 축소율
기존 시스템	85.09%	
제안된 시스템+ 기존 시스템	88.75%	24.55%

표 2의 결과는 고유명사 추정이 품사 태깅에 좋은 영향을 주고 있음을 알 수 있다. 이 결과는 품사 태깅에 직접적인 도움을 주었다기 보다도 정확한 미등록어 추정으로 품사 태깅의 성능이 개선된 것으로 판단된다.

4.2. 기존 시스템과의 비교

한국어에서 고유명사 추출에 관한 연구는 거의 이루어지지 않았으며, 최근에 와서 [12]가 한국어 고유명사 추출 시스템의 전부이다. [12]에서는 데이터 수집기라고 하는 도구를 이용해서 여러 형태의 정보, 이름(고유명사), 접사, 실마리를 추출한다. 이 정보를 이용해서 주로 회사명을 인식하는데, 먼저 형태소 분석을 수행하고, 형태소 분석 결과에서 적절한 통계 정보를 수집하고 나서 고유명사를 추출한다. 표 3은 본 논문에서 제안된 고유명사 추출 모델과 [12]에서 제안된 모델의 특성을 비교한 것이다.

표 3 기존 시스템과의 특성 비교

특 성	[12]의 모델	제안된 모델
형태소 분석기	이용함	이용하지 않음
음차된 외국어	고려하지 않음	고려함
캐쉬 기능	이용함	이용함
단서집합	이용	이용
고유명사 사전	이용	이용
고유명사 가능 접사	이용	이용

[12]는 언어처리 도구인 형태소 분석을 사용하기 때문에 강인한 언어처리가 필수적으로 요구된다. 따라서 정보추출과 같은 영역에서 사용할 경우 약간의 문제가 발생될 수도 있다. 또한 [12]는 음차된 외국어를 고려하고 있지 않다. 제한되지 않은 문서에는 많은 외국어들이 음차되어 표기되기 때문에 이와 같은 단어에 대해서도 충분히 고려되어야 할 것이다.

4.3. 시스템의 개선 방안

본 절에서는 제안된 모델 및 시스템에 대한 문제점을 지적하고 이들에 대한 개선 방안을 살펴보고자 한다.

첫 번째 문제는 대부분의 규칙이나 문맥은 수동으로 작성된다는 것이다. 수동으로 작성된 규칙이나 문맥은 영역의 변화에 매우 민감하므로 적용 영역이 변화되면 규칙이나 문맥 심지어는 고유명사 사전까지도 많은 수정이 요구된다. 이와 같은 문제는 자동학습 기법을 통해서 문맥을 자동추출하는 연구가 계속적으로 추진되어야 할 것이다.

두 번째 문제는 문맥의 대부분은 정규표현으로 표현된다는 것이다. 그러나, 고유명사 추출을 위한 문맥은 고유명사에 상당히 원거리에 있는 단어들이 영향을 줄 수 있는데, 정규표현은 원거리 문맥을 표현하기 어렵다. 원거리 문맥이나 다양한 문맥으로 표현할 수 있는 체계적인 연구가 필요하다.

세 번째 문제는 언어학적인 근거가 미흡하다는 것이다. 본 논문은 고유명사 추출에 대한 매우 초보적인 연구로서 어떤 언어학적인 배경을 근거로 사용하고 있지 않으며, 단지 공학적인 측면에서 문제의 해결점을 찾고자 노력하였다. 언어학적으로 밑바탕이 되는 이론적 근거 없이 이 모델의 꾸준한 발전이 어려울 것으로 생각된다. 이 문제의 본질을 해석해 줄 수 있는 전산언어학적 이론을 정립하는 연구도 아울러 추진되어야 할 것이다.

네 번째 문제는 객관적인 평가가 부족하다는 것이다. 한국어에 고유명사 정보가 태깅된 말뭉치가 절대적으로 부족하다. 자동학습뿐 아니라 객관적인 시스템 평가를 위해서 고유명사 정보가 태깅된 말뭉치를 구축해야 한다.

5. 결론

본 논문에서 한국어 문서에서 고유명사 추출 모델을 제안하였다. 이 모델은 형태소 분석과 같은 언어처리 도구를 사용하지 않는다. 언어처리 도구를 사용할 경우에는 정확하게 추출할 수 있지만 언어처리 도구를 사용하기 어려운 정보추출과 같은 분야에서 잘 적용될 수 있다. 제안된 고유명사 추출 모델은 크게 3가지 가능으로 구성된다. 첫 번째 기능은 여과 기능이며, 이 기능이 고유명사를 포함하지 않는 어절을 고려대상에서 제외한다. 두 번째 기능은 분리 기능이며. 이 기능은 고유명사를 효율적으로 찾기 위해서 조사와 복합명사를 분리한다. 세 번째 기능이 인식 기능이며, 이 기능을 통해서 최종적으로 고유명사를 판단한다. 고유명사 인식을 위해서 사용되는 정보는 실마리 문맥이며 이 문맥의 대부분은 정규표현으로 표현되며, 유한상태오토마타로 구현되었다. 제안된 모델의 객관적인 평가는 수행할 수 없었지만, 품사 태깅 시스템에 적용했을 때, 성능이 개선됨을 알 수 있었고, 그 결과로 고유명사 추정이 미등록어 처리에 많은 도움을 줄 수 있었다.

고유명사 추출에 관련된 연구는 매우 초보적인 단계로서 앞으로 개선해야 할 많은 과제를 가지고 있다. 우선 고유명사에 대한 이론적인 기반을 정립해야 하고 이를 토대로 고유명사 인식에 대한 계산적인 모델을 정립해야 할 것이다. 또한 정립된 이론을 바탕으로 고유명사 정보가 부착된 말뭉치를 구축하는 일이 매우 시급한 연구 과제이다.

감사의 글

본 연구는 한국전자통신연구원 교환전송기술연구소의 지원을 받았으며, 또한 첨단정보기술연구센터를 통하여 과학재단의 지원을 받았습니다.

참고 문헌

[1] Baeza-Yates, R. and Ribeiro-Neto, B., *Modern Information Retrieval*, Addison Wesley, 1999.

[2] Mani, I. and Maybury Mark T., *Advances in Automatic Text*, The MIT Press, 1999.

[3] Pazienza, Maria T., *Information Extraction : A Multidisciplinary Approach to an Emerging Information Technology*, Springer-Verlag, 1997.

[4] Freitag, D., *Machine Learning for Information Extraction in Informal Domains*, Ph. D. Thesis, Computer Science Department, Carnegie Mellon University, 1998.

[5] Grefenstette, G. "Short query linguistic expansion techniques: palliating one-word queries by providing intermediate structures to text," in *Information Extraction : A Multidisciplinary Approach to an Emerging Information Technology*, Pazienza, Maria T. (ed), Springer-Verlag pp. 97-114, 1997.

[6] 김재훈, 김준홍, 박호진, "여과 및 분리 기법을 이용한 한국어 기준 명사 추출", 제12회 한글 및 한국어 정보처리 학술대회 발표논문집, 성공회대학교, 서울, pp. 3-10, 2000.

[7] 이도길, 류원호, 임해창, "분석 배재정보와 후절어를 이용한 한국어 명사추출", 제12회 한글 및 한국어 정보처리 학술대회 발표논문집, 성공회대학교, 서울, pp. 19-25, 2000.

[8] 오종훈, 최기선, "은닉마르코프모델(HMM)을 이용한 과학기술문서에서의 외래어 추출 모델", 제11회 한글 및 한국어 정보처리 학술대회, 전북대학교, pp. 137-141, 1999.

[9] 윤보현, 조민정, 임해창, "통계정보와 선호규칙을 이용한 한국어 복합 명사의 분해", 한국정보과학회 논문지(B), 제24권, 제8호, pp. 900-909, 1998.

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

[11] 김재훈, "가중치 망을 이용한 한국어 품사 태깅", 정보과학회 논문지, 25(6): 951-956, 1998.

[12] 김태현, 이현숙, 하유선, 이만호, 맹성현, "데이터 집합을 이용한 고유명사 추출", 제12회 한글 및 한국어 정보처리 학술대회 발표논문집, 성공회대학교, 서울, pp. 11-18, 2000.

부록 1 : 실마리 단어의 일부

(관광)?개발	(방송)?센터	(신)?학교
(전문)?대학(교)?	(정보)?통신	(종합)?건설
(종합)?식품	(주식)?회사	(초등|중|고등)?학교
(프레스)?	\(주\)	㈜
가구	가스	감독원
감호소	강관	강업
....		

부록 2 : 나라 이름의 일부

가나	가봉	가야
가이아나	감비아	고려
과테말라	괌	그레나다
그루지야	그리스	기니
기니비사우	기보스니아	기아나
나미비아	나우루	나이리지라
나이지리아	남아공	남아(프리카)?(공화국)?
...		

부록 3 : 조직 이름의 일부

DJ신당	감정원	건국위
경실련	경영자총협회	경제기획원
경제정의실천시민연합	경총	공정거래위원회
공화당	관우회	광주예술대
교우회	교총	국기원
국립오케스라	국무부	국민당
국민승리21	국민신당	국민정부
...		

부록 4 : 직업명 이름의 일부

(대|고등|중|초등)?학생	(대|기|사|부|과|관|교|의|반|회|병)장
(도|시|군|구|국회)?의원	(부장|수석|지방|평)?검사
(수|보조)?간호사	(시|지방|교육|행정)?공무원
(시간|전임|학원)?강사	(신문|잡지|TV|정치부|사회부|연예부|촬영)?기자
(영화|연극)?배우	(영화|조)?감독
(운전|택시|조명|버스|설비)?기사	(정|부|조)?교수
AD	FD

부록 5 : 시스템의 입력

1. 金대통령, 5대재벌 구조조정 직접나서…내주 정재계간담회
2. 김대중대통령이 5대그룹의 구조조정 문제를 마무리짓기 위해 다음주중 청와대에서 이들 재벌총수와 간담회를 갖는다.
3. 김대통령은 30일 청와대에서 박태준자민련총재와 주례회동을 갖고 이 문제로 더이상 뒷말이나 차질이 없도록 빠른 시일 내에 정재계간담회를 소집해 충분한 토론을 통해 결말을 짓기로 했다고 박지원공보수석이 밝혔다.
4. 이에 따라 금융감독위원회와 채권은행단은 금주중 재계와 접촉을 갖고 회의형식과 참석범위, 구조개혁의 큰 줄거리 등에 대해 사전조율할 방침이다.
5. 박수석은 "김대통령과 박총재가 이 간담회에 직접 참석해 5대재벌 구조조정 문제에 대한 완벽한 합의를 이끌어내기로 의견을 모았다"고 말했다.

부록 6 : 시스템의 출력

1. (@Q@金@)대통령, 5대재벌 구조조정 직접나서…내주 정재계간담회
2. (@Q@김대중@)대통령이 5대그룹의 구조조정 문제를 마무리짓기 위해 다음주중 (@Q@청와대@)에서 이들 재벌총수와 간담회를 갖는다.
3. (@Q@김@)대통령은 30일 (@Q@청와대@)에서 (@Q@박태준@)(@Q@자민련@)총재와 주례회동을 갖고 이 문제로 더이상 뒷말이나 차질이 없도록 빠른 시일 내에 정재계간담회를 소집해 충분한 토론을 통해 결말을 짓기로 했다고 (@Q@박지원@)공보수석이 밝혔다.
4. 이에 따라 (@Q@금융감독위원회@)와 채권은행단은 금주중 재계와 접촉을 갖고 회의형식과 참석범위, 구조개혁의 큰 줄거리 등에 대해 사전조율할 방침이다.
5. (@Q@박@)수석은 "(@Q@김@)대통령과 (@Q@박@)총재가 이 간담회에 직접 참석해 5대재벌 구조조정 문제에 대한 완벽한 합의를 이끌어내기로 의견을 (@Q@모았다@)"고 말했다.