DOI: 10.3745/KIPSTB.2008.15-B.6.613

한국어 기준명사 추출 및 그 응용

김 재 훈†

요 약

정보검색, 문서요약 등의 분야에서 명사추출은 매우 중요하다. 본 논문은 대량의 문서로부터 기준명사를 효과적으로 추출하기 위한 한국어 기준명사 추출 시스템을 제안하고 이를 문서요약 시스템에 적용한다. 기준명사는 명사들 중에서 기본이 되는 명사이며 복합명사는 포함되지 않는다. 본 논문에서는 두 가지 기술 즉 여과기법과 분리기법을 사용한다. 먼저 여과기법을 이용해서 명사를 포함하지 않은 어절을 미리 제거하고, 그리고 분리기법을 이용해서 명사가 포함된 어절에서 명사와 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 문서요약 시스템에 적용해 보았을 때, 좋은 결과를 얻을 수 있었다.

키워드 : 명사추출, 여과기법, 복합명사 분리, 문서요약

Korean Base-Noun Extraction and its Application

Jae-Hoon Kim†

ABSTRACT

Noun extraction plays an important part in the fields of information retrieval, text summarization, and so on. In this paper, we present a Korean base-noun extraction system and apply it to text summarization to deal with a huge amount of text effectively. The base-noun is an atomic noun but not a compound noun and we use tow techniques, filtering and segmenting. The filtering technique is used for removing non-nominal words from text before extracting base-nouns and the segmenting technique is employed for separating a particle from a nominal and for dividing a compound noun into base-nouns. We have shown that both of the recall and the precision of the proposed system are about 89% on the average under experimental conditions of ETRI corpus. The proposed system has applied to Korean text summarization system and is shown satisfactory results.

Keywords : Noun Extraction, Filtering Technique, Segmentation of Compound Noun, Text Summarization

1. 서 론

웹은 많은 정보를 담고 있는 정보의 보고이나 웹에 존재하는 정보는 매우 다양하며, 그 양도 매우 빠른 속도로 증가하고 있다. 방대한 정보공간에서 유용한 정보를 찾기 위해 널리 사용되는 도구가 검색엔진이며, 검색엔진을 구축하기 위한 필수적인 도구 중 하나가 명사추출 시스템이다. 명사추출 시스템은 색인어 추출, 자연언어 질의어 분석, 시소러스 구축 등에서 널리 사용되고 있다[1]. 이 밖에도 정보추출이나 문서요약 등 대량의 자연언어 문서를 다루는 분야에서 널리 사용되고 있다[2].

자연언어에는 여러 형태의 중의성을 가지고 있으며 명사도 예외는 아니다. 이 때문에 자연언어 문장에서 명사를 정확히 추출하기 위해서는 복잡한 과정을 거쳐야 한다. 명사가 가지는 중의성의 대부분은 형태소 분석과 같은 낮은 단계의 분석으로도 쉽게 해결할 수 있으나, 일부의 중의성은 의미 분석과 같은 복잡한 과정을 통해서만 해결할 수 있다. 그러나, 정보검색과 문서요약과 같은 분야에서는 짧은 시간 내에 대량의 문장을 처리하고, 특정영역(예: 신문, 소설 등)에 무관하게 처리해야 하므로 의미 분석과 같은 복잡한 언어처리 과정을 이용하는 것은 적절한 방법이 아니다. 복잡한 언어처리 과정을 이용할 경우, 시스템 개발에 필요한 시간과 노력이 많이 요구되고, 또한 많은 실행시간이 요구되며, 시스템이 강인하지 않을 수 있다는 문제가 있다. 따라서 정보검색이나 문서요약과 같은 분야에서는 정확성에는 다소 희생되더라도 특정 영역에 무관하고 강인하며 빠른 명사추출 시스템이 필요하다. 이것이 본 논문의 목적이다.

한국어 문장은 하나 이상의 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 사전

† 종신회원 : 한국해양대학교 컴퓨터공학과 교수
논문접수 : 2008년 6월 27일
수 정 일 : 1차 2008년 7월 23일
심사완료 : 2008년 8월 4일

및 상호정보에 기반한 후방향-전방향 알고리즘을 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 상호정보에 기반한 후방향-전방향 알고리즘을 이용한 분리 기법을 사용한다. 그리고 나서 찾아진 명사구에는 많은 경우 하나의 명사로 구성되지만, 몇몇의 경우에는 여러 개의 명사가 하나의 명사구를 이루고 있는데, 이들을 분리하여 기준 명사를 찾는다.

본 논문의 구성을 다음과 같다. 2장에서 본 논문과 관련된 한국어 명사 추출 방법과 복합명사 분리 방법에 대해서 기술한다. 3장에서 상호정보를 이용한 여과 및 분리 기법을 통한 명사 추출 방법에 대해서 논하고, 4장에서 제안된 시스템의 성능을 평가하고, 5장에서 다른 한국어 명사 추출 방법들과 비교·분석하고 문서요약 시스템에 적용한다. 끝으로 6장에서 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[3,4], 둘째, 형태소 분석기를 이용하는 경우이고[5-7], 셋째, 아무런 언어분석 도구를 사용하지 않는 경우이다[8].

품사 태거를 이용하는 방법은 품사 태깅 결과에서 원하는 품사에 해당하는 단어만 출력하면 된다[3]. 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[6]. 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[8]. 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 가지게 되는 단어(예: 인공지능, 정보검색)를 말하며, 구문적으로는 단일단어와 같은 역할을 한다. 한국어 복합명사 분해는 크게 통계적 방법[9,10]과 규칙기반 방법[11,12]으로 나눌 수 있다.

통계적인 방법의 예로 [9]를 살펴보자. [9]는 통계정보(compound noun formation probability, CFP)와 선호규칙(minimal noun preference rule, MNPR)을 이용하여 복합명사를 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.

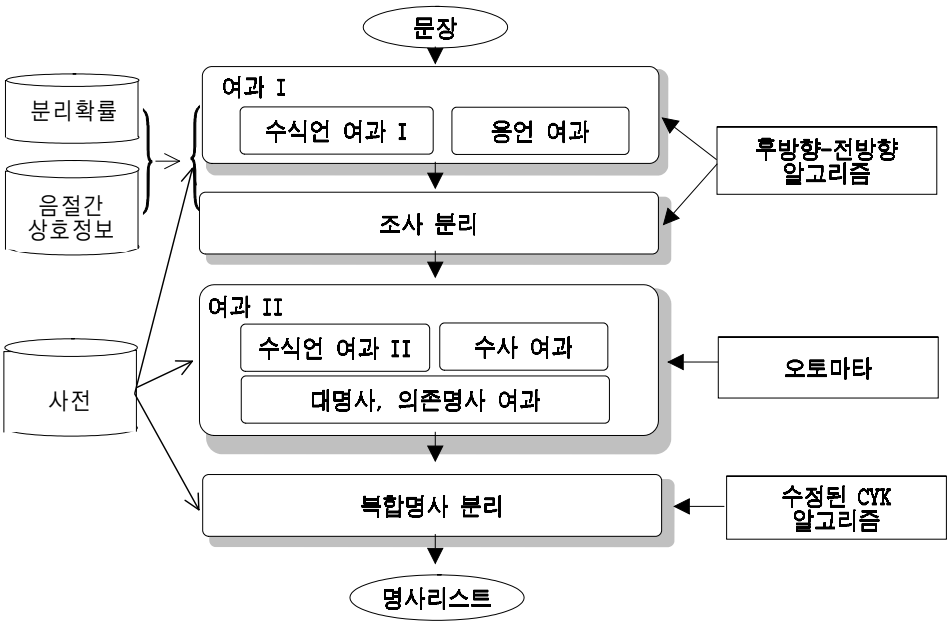
규칙기반 방법의 예로 [11]을 살펴보자. [11]은 네 종류의 복합명사 분해 규칙을 사용하고 있으며, 두 종류의 예외 규칙을 사용하고 있다. 이들 규칙에 의해서 후보 규칙을 생성하고 생성된 후보에 대해 가중치를 부여하여 가중치에 따라 최적 후보를 선정하게 된다. 가중치를 부여하는 방법도 규칙에 의해서 결정되는데, 기준명사의 유형, 사전에 수록된 음절의 길이, 접미사가 결합된 음절 길이, 음절 패턴 빈도, 중심어 빈도에 따라 적절하게 조절된다.

3. 한국어 기준명사 추출 시스템

기준명사는 명사구를 이루는 가장 최소 단위를 말한다. 명사구는 문장의 구성 성분상 체언에 속하며, 체언에는 보통명사, 고유명사, 수사, 대명사, 의존명사가 있다. 체언 중에서는 수사, 대명사, 의존명사를 제외한 보통명사와 고유명사를 구성하는 명사를 본 논문에서는 기준명사라고 한다. 예를 들면 복합명사 “경제발전”은 두 개의 기준명사 “경제”와 “발전”으로 구성되어 있다. 본 논문은 주어진 문서로부터 기준명사를 추출하기 위해 여과 기법과 분리 기법을 이용한다. 여과 기법은 비체언성 어절, 즉, 수식언, 독립언과 용언을 제거하기 위해 사용된다. 예외로 명사를 포함하는 일부 용언은 제외하지 않는다. 예를 들면, 접미어 “-하다/되다”는 명사와 결합하여 용언이 된 어절이므로 이와 같은 어절 내에는 기준 명사가 포함되어 있으므로 여과하지 않는다. 또한 문서요약에서는 체언이라고 하더라도 대명사, 수사, 의존명사는 중요한 특징으로 사용하기 않기 때문에 이들도 여과한다.[1] 분리 기법은 체언에 포함된 명사구와 조사를 분리하는 일과 명사구가 복합명사일 경우 이를 기준명사로 분리하는 일을 담당한다.

(그림 1)은 한국어 기준명사 추출 시스템의 구조이다. (그림 1)에서 사전은 가능한 모든 단어가 포함되어 있고, 각 단어의 정보는 명사, 동사, 부사, 형용사 등과 같은 품사 정보뿐이다. 용언은 일반적인 자연언어처리의 경우와 달리 변화형도 포함되어 있다. 즉, 형용사 “아름답”의 경우에는 “아름답”은 물론이고 “아름다”도 사전에 등재되어 있다. 분리확률

1) 모든 문서요약 시스템이 대명사, 수사, 의존명사를 중요한 특징으로 간주하지 않는 것은 아니다.



(그림 1) 한국어 기준명사 추출 시스템의 구조

과 음절간의 상호정보는 체언으로부터 조사를 분리하고 용언으로부터 어미를 분리하기 위해서 필요한 정보이다. 분리확률은 두 음절이 명사와 조사 그리고 용언의 어간과 어미로 분리된 확률을 의미한다. 즉 두 음절이 명사와 조사 그리고 용언의 어간과 어미 사이에서 자주 발생되면 그 분리확률이 높고 그렇지 않으면 분리확률이 낮다. 상호정보는 주어진 두 음절이 조사나 어미 부분에서 함께 나오는 정도를 말한다. 만약 두 음절이 조사나 어미 부분에서 항상 자주 나타난다면 상호정보의 값이 높고 그렇지 않으면 상호정보 값이 낮을 것이다. 후방향-전방향 알고리즘은 분리확률과 상호정보를 이용해서 용언을 인식하고 명사와 조사를 분리한다. 오토마타는 수사를 여과하기 위해서 사용된다. 각 모듈에 대한 좀더 자세한 설명은 이하의 절에서 기술될 것이다.

3.1 여과 I

여과 I 단계에서는 명사구를 포함하지 않는 어절을 여과한다. 명사구를 포함하지 않는 어절에는 용언(동사, 형용사), 수식언(부사, 관형사), 독립언(감탄사)이 있으며, 이들을 여과하기 위해서 주어진 어절이 명사구를 포함하는지를 알아야 한다. 본 논문에서는 첨가의 정도에 따라 용언(동사, 형용사)과 그렇지 않은 것(수식언, 독립언)으로 나누어서 처리한다. 용언은 첨가현상이 매우 심하게 발생되므로 모든 가능한 용언을 사전에 등재하는 것은 거의 불가능하다.[2] 반면에 수식언과 독립언은 용언에 비해 첨가정도가 아주 약하므로 거의 모든 어절을 사전에 등재한 것으로 가정한다.[3] 본 논문에서는 편의상 독립언에 대해서는 수식언 여과에 포함시켜 별도로 언급하지 않는다.

3.1.1 수식언 여과 I

수식언 여과 I 단계에서는 보조사와 결합하지 않는 부사, 관형사, 감탄사가 여기에서 여과된다. 이를 위해서 사전을 이용하며, 기본적으로는 모든 부사, 관형사, 감탄사가 사전에 포함된 것으로 가정한다. 그러나 이 가정을 실용적인 시스템에서는 올바른 가정이 아니다. 본 논문에서는 이 문제를 조금 완화시키기 아주 간단한 경험규칙[4]을 이용하여 미등록어를 처리하고 있으나, 많은 개선의 여지를 가지고 있다.

3.1.2 용언 여과

용언 여과 단계에서는 명사를 포함하지 않는 모든 용언을 여과한다. 용언은 어간(동사나 형용사)과 어미로 구성된다. 주어진 어절이 용언인지를 결정하기 위해서 본 논문에서는 후방향-전방향 알고리즘(그림 2)를 이용해서 용언의 어미를 분리하고, 어미를 제외한 부분이 동사와 형용사이면 주어진 어절을 용언으로 인식한다. 일반적으로 명사를 포함하지 않는 용언의 어간은 일반적으로 미등록어로 나타날 가능성이 적으므로 본 논문에서는 명사를 포함하지 않는 용언의 어간은 모두 사전에 있다고 가정하며, 용언 어간에 대한 가능한 모든 변화형[5]이 사전에 등재된 것으로 가정한다. 이 점이 일반적인 자연언어 처리에서 말하는 용언의 어간과는 다소 차이가 있다. 예를 들면 용언의 어간(형용사) '아름답'은 '아름답'은 물론이고 '아름다'도 사전에 등재된다.

2) 여기서 용언은 용언의 어간과 어미가 결합된 것을 의미한다.

3) 실질적으로는 수식언에도 보조사가 첨가될 수 있는데(예: '다행히도' = '다행히'+'도') 여기서는 보조사가 결합되지 않은 수식언만 등재된 것으로 가정한다.

4) 하나의 예를 들면, 명사+'히'를 부사로 간주한다. 여기에도 명사가 존재하나 이와 같은 명사는 기준명사로 간주하지 않았다. 예를 들면, "명확+히, 조용+히" 등과 같은 것이다.

5) 용언의 어간에 대한 가능한 모든 변화형은 자동으로 생성할 수 있는데, 그 규칙 중 하나를 살펴보자. 받침없는 규칙 용언의 어간일 경우에 받침 "ㄴ, ㄹ ㅁ ㅂ ㅆ"을 첨가하여 변화형을 만든다. 예를 들면, 용언의 어간 "가"는 "가, 간, 갈 감, 갑, 갔"을 추가하면 된다. 불규칙의 경우는 불규칙의 종류마다 다르다. 이것에 대한 자세한 설명은 논문의 범위를 벗어나므로 여기서는 더 이상 논의하지 않는다.

```
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},e) <θ_1);
    // 1에서 찾은 위치를 기준으로 다시
    // 오른쪽으로 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

이 알고리즘은 상호정보 $I(p_i,p_j,e)$와 분리확률 $\Pr(s|p_i,p_j)$을 이용하며, 각각 식 (1)과 (2)와 같이 정의된다. 또한 len(word)는 입력 어절 word의 길이를 구하는 함수이고, θ_1과 θ_2는 시스템 성능을 조절할 수 있는 매개변수이다.

$$I(p_i,p_j,e)=\log\frac{\Pr(p_i,p_j,e)}{\Pr(p_i)\Pr(p_j)}=\log\frac{N\cdot C(p_i,p_j,e)}{C(p_i)C(p_j)} \quad (1)$$

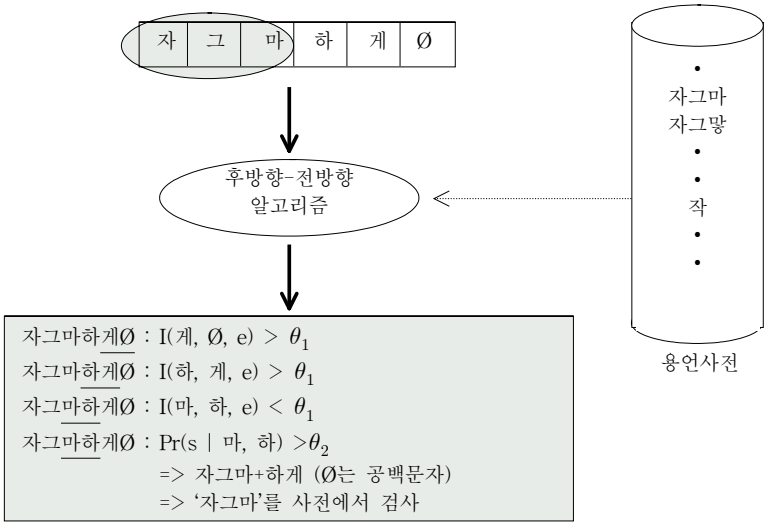
$$\Pr(s|p_i,p_j)=\frac{C(p_i,s,p_j)}{C(p_i,p_j)} \quad (2)$$

식 (1)과 (2)에서 p_i와 p_j는 음절을 표현하며, $\Pr(x)$와 $C(x)$는 각각 x가 발생할 확률과 빈도수이다. N은 학습말뭉치의 음절 수이다. 식 (1)의 상호정보 $I(p_i,p_j,e)$는 두 음절 p_i,p_j가 어미 부분에서 함께 나타날 정도를 나타낸다. 두 음절이 어미 부분에서 항상 함께 나타난다면 아주 높은 상호정보 값을 가지게 되고 어미 부분이나 그 외의 부분에서 골고루 나타내면 낮은 상호정보 값을 가지게 된다. 따라서 식 (1)의 $\Pr(p_i,p_j,e)$은 두 음절 p_i,p_j가 어미 부분에서 함께 나타날 확률이며, 품사 태깅된 말뭉치로부터 추정된다. 추정 방법은 품사 태깅 말뭉치로부터 어미 부분을 추출하고, 추출된 어미에서 나타날 수 있는 모든 음절을 빈도수 $C(p_i,p_j,e)$를 세어서 상대빈도수를 이용해서 추정된다. 식 (1)에서 $\Pr(p_i)$는 학습 말뭉치에서 음절 p_i가 나타날 확률로서 학습 말뭉치에서 p_i가 나타난 빈도수 $C(p_i)$를 세어서 상대빈도수를 이용하여 추정된다. 식 (2)의 분리확률은 두 음절 p_i,p_j 사이에서 용언의 어간과 어미가 분리될 확률을 말한다. 식 (2)에서 $C(p_i,s,p_j)$는 학습말뭉치에서 두 음절 p_i,p_j 사이에서 용언의 어간과 어미가 분리된 회수이고, $C(p_i,p_j)$는 학습 말뭉치에서 두 음절 p_i,p_j가 나타난 회수이다. 후방향-전방향 알고리즘에서 후방향 처리는 오른쪽에서 왼쪽으로 처리하면서 가능한 어미를 찾고, 전방향 처리는 왼쪽에서 오른쪽으로 처리하면서 용언과 어미의 정확한 분리 위치를 결정한다.

(그림 3)은 후방향-전방향 알고리즘을 이용해서 어절 '자그마하게'에서 용언의 어간과 어미를 분리하는 예를 보이고 있다. (그림 3)에서 문자 Ø는 공백문자를 나타내며, 이를 이용해서 한 문자로 구성된 어미에 대한 정보를 구할 수 있다.

3.2 조사 분리

대부분의 체언은 명사와 조사로 구성된다. 명사를 정확히 찾기 위해서는 명사와 조사를 분리해야 한다. 조사를 분리하기 위해서 어미를 분리하는 경우와 마찬가지로 (그림 2)의 후방향-전방향 알고리즘을 이용한다. 용언 분리와의 차이점은 크게 두 가지가 있다. 하나는 사전을 이용해서 확인을 하지 않는다는 점이고 다른 하나는 후방향-전방향 알고리즘에서 사용하는 상호정보와 분리확률이 다르다는 것이다. 전자는 명사에는 많은 미등록어가 존재할 수 있기 때문에 사전으로 확인하는 방법은 너무나 제한적이다. 후자에서 상호정보는 $I(p_i,p_j,j)$를 이용하는데, 이는 두 음절 p_i,p_j가



(그림 3) 후방향-전방향 알고리즘을 이용해서 어절 '자그마하게'에서 용언의 어간 '자그마'와 어미 '하게'를 분리하는 과정

조사 부분에서 함께 나타나는 정도를 말하며, 분리확률은 $\Pr(s|p_i,p_j)$를 이용하는데 이는 두 음절 p_i,p_j 사이에서 조사가 분리될 확률이다.

3.3 여과 II

여과 II 단계에서는 조사가 분리된 후 기준명사가 포함되지 않는 어절을 여과한다. 수식언 특히 부사에는 보조사가 첨가될 수 있기 때문에 조사를 분리한 후에 수식언 여과를 고려해야 한다. 또한 일반적인 명사구라고 하더라도 기준명사를 포함하지 않는 어절이 여기서 여과되며, 대명사와 의존명사를 포함하는 어절과 수사를 포함하는 어절이 여기에 속한다.

3.3.1 대명사와 의존명사의 여과

대명사와 의존명사는 명사구에 속하지만, 기준명사는 포함하지 않는다. 이를 여과하기 위해서 조사를 분리한 후, 나머지 문자열로 사전을 검색하여 사전에 존재하면 어절을 여과한다. 이 부류에 속하는 단어는 극히 제한적이기 때문에 미등록어를 고려하지 않는다.

3.3.2 수사 여과

수사를 여과하기 위해서 유한 상태 오토마타를 이용한다. 아래는 수사를 인식하기 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|셋|넷|다섯|여섯|일곱|여덟|아홉)
(영|일|이|삼|사|오|육|칠|팔|구)+
...
```

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 분리한 후, 명사 부분이 위의 정규표현에 일치될 때, 명사 부분은 인식되고, 수사로 인식되면 주어진 어절에는 기준명사가 존재하지 않는 것으로 가정한다. 예를 들면, 어절 "12조3억2천만원을"이 주어졌다면, 조사 분리 단계에서 "-을"을 분리하고 "12조3억2천만원"은 위에서 제시한 정규표현에 의해서 수사로 인식되므로 주어진 어절은 여과된다.

3.3.3 수식언 여과 II

3.2절에서 구체적으로 언급하지는 않았지만 조사를 분리할 때 반드시 명사와 조사만 분리하는 것이 아니라 어절에 포함된 모든 조사를 분리한다. 수식언을 제거하기 위한 기본적인 방법은 3.1.1에서 기술한 것과 같다. 단지 사전을 검사할 때 조사를 분리한 후의 나머지 문자열만을 이용한다는 점만 다르다.

3.4 복합명사 분리

복합명사를 분리하기 위해 몇 가지 경험규칙[9]을 사용하며, 아래와 같이 요약된다.

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

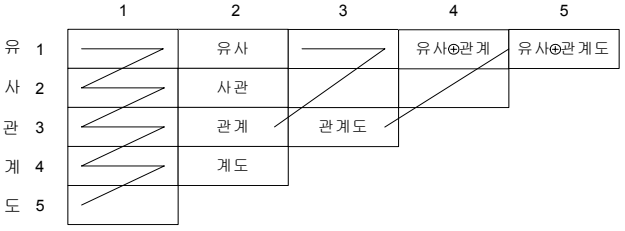
첫 번째 경험규칙은 한국어 복합명사의 대부분이 2음절 명사와 3음절 명사로 구성된다는 사실에서 기인된 것이며, 드물게 3음절 이상의 명사도 복합명사 내에 포함될 수 있다고 가정하고 본 논문에서는 5음절까지만 포함되는 것으로 가정한다. 두 번째 경험규칙은 길이가 긴 단어가 사전에 포함되어 있을 경우 이를 우선한다는 것이다. 이와 같은 경험규칙과 CYK 파싱 알고리즘[13]을 이용해서 복합명사를 분리한다. 이 알고리즘을 요약하면 (그림 4)와 같다.

여기서, 함수 select_best(.)는 위에서 언급한 두 번째 경험규칙을 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.

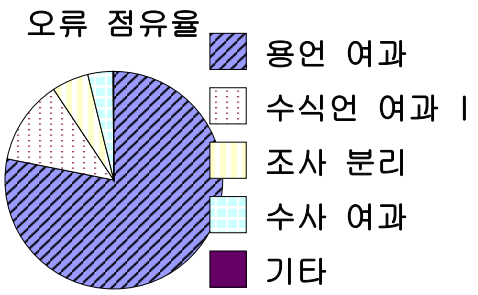
(그림 5)는 입력 어절 '유사관계도'에 대한 복합명사 분리의 과정을 테이블로 보인 것이다. 조사 분리 단계를 거쳐 조사가 제거된 단어 '유사관계도'가 입력되면, 먼저 명사 사전을 참조하여 기준명사를 테이블에 할당한다. 예에서는 2음절 명사 '유사', '사관', '관계', '계도'와 3음절 명사 '관계도'가 테이블에 할당되었다. 다음에 테이블에 할당된 기준명사의 결합 가능성을 (그림 4)의 수정된 CYK 알고리즘을 이용하여 검사한다. 예를 들면, T[1,4]에는 T[1,2]의 '유사'와 T[3,2]의 '관계'가 결합하여 '유사⊕관계'가 할당되었고, T[1,5]에는 T[1,4]의 '유사'와 T[3,3]의 '관계도'가 결합하여 '유사⊕관계

```
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 4) 복합명사 분리를 위한 수정된 CYK 파싱 알고리즘



(그림 5) 복합명사 '유사관계도'에 대한 분리 과정



(그림 6) 잘못 여과된 어절의 오류 분포

도'가 할당되었다. 최종 결과는 T[1,5]에 할당된 '유사⊕관계도'이다.

4. 실험 및 평가

4.1 실험 환경

본 논문에서 음절의 상호정보와 분리확률을 구하기 위해서 사용된 말뭉치는 KAIST 말뭉치[14]이다. 이 말뭉치는 규모는 작지만 정확성이 매우 높기 때문에 학습을 위해서 사용되었다. 평가를 위해서는 1999년에 배포한 ETRI 말뭉치[15] 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.

4.2 성능평가

4.2.1 전체 시스템의 성능

<표 1>은 평가용 말뭉치에 대한 재현율, 정확률, F-measure[16]이다. 재현율과 정확률 모두가 평균 약 89%이며, F-measure도 약 89%이다. 일반적인 내용을 다루는 뉴스에 대해서는 매우 좋은 결과를 보이고 있으나, 사람이나 대화체를 많이 사용하는 소설에 대해서는 좋은 결과를 보여주지 못했다. 또한 본 시스템에는 짧은 단어에 대한 모호성을 만족스럽게 처리할 수 없었으며, 다음절에서 논의할 여과 단계에 대한 성능을 개선할 필요가 있다. 특히 용언 여과에 대한 성능이 개선되어야 할 것이다. 구현된 시스템을 문서요약에서 명사 추출 시스템으로 사용하였을 때, 좋은 결과를 얻을 수 있었다[17].

〈표 1〉 제안된 시스템의 성능 평가

분야	재현율	정확률	F-척도
뉴스	91.71	89.99	90.84
뉴스(방송)	92.53	89.91	91.20
비소설	88.55	91.33	89.92
소설	86.62	82.35	84.43
학습서	88.48	91.68	90.05
평균	89.58	89.01	89.30

4.2.2 여과 단계의 성능

본 논문에서 제안한 기준명사 추출 모델은 여과 단계(수식언 여과(I, II), 용언 여과, 수사 여과, 대명사 여과)의 성능에 크게 의존된다. 따라서 본 절에서는 각 여과 모듈의 성능에 대해서 살펴보고자 한다. <표 2>는 여과 모듈에 대한 정확도를 보여주고 있다.

각 단계의 오류는 다음 단계로 전파되기 때문에 각 단계의 순수한 오류는 <표 2>에서 보인 것과 정확히 일치하지는 않는다. 수식언(형용사, 부사)은 사전을 이용하기 때문에 거의 100%에 가까운 정확률을 보였다. 오토마타를 이용하여 인식하는 수사의 경우에도 약 94%로 어느 정도 정확하게 여과함을 알 수 있었으며, 수사를 인식하기 위한 문맥을 오토마타에 추가한다면 더 좋은 결과를 얻을 수 있을 것이다. 그러나, 대명사나 의존명사의 경우에는 대부분이 짧은 음절(하나 혹은 둘)로 구성되어 정확하게 인식하기 어려웠으며[6], 앞으로 음절의 길이를 반영하는 모델이 추가로 개발될 경우 좋은 결과를 얻을 수 있을 것으로 기대된다. 용언의 경우에 가장 많이 틀리는 경우도 짧은 어절이며, 이 경우 많은 모호성을 지니고 있었다. 감탄사의 경우에는 그 수가 얼마 되지 않아서 본 논문에서 구현된 시스템에서 이를 충분히 고려하지 못한 점이 정확률을 떨어뜨린 요인으로 작용하였다.

4.3 오류분석

4.3.1 잘못 여과된 명사구 어절의 오류 분석

오류 분석은 모델을 개선하거나 시스템의 성능을 개선하는 데에 많은 도움을 줄 수 있기 때문에 본 절에서는 구현된 시스템의 각 단계에서 발생되는 오류를 분석하고자 한다. (그림 6)은 명사가 포함되어 있지만, 이를 여과하여 명사를 정확하게 인식하지 못한 각 모듈의 오류 점유율을 보이고 있으며, 이는 이 부류에 속하는 전체 오류(16,163개)에서 각 모듈이 야기시킨 오류가 차지하는 비율이다. 용언 여과 모듈에서 가장 많은 오류를 야기시키고 있으며, 용언 인식 모델이 개선되어야함을 간접적으로 말하고 있다. 현재 용언 인식 모델에서 사용되는 특징은 두 음절에 대한 상호정보와 분리확률이다. 이 특징만으로 주어진 어절이 용언인지를 인식하는 것은 조금 무리가 있을 것으로 판단되므로 용언 인식 모델이 개선되어야 할 것으로 판단된다. [18]에서 제안한 배제정보가 좋은 특징으로 사용될 수 있을 것이다. 조사 분리 모듈에서 야기되는 오류도 용언 인식 모듈에서

〈표 2〉 여과 모듈의 성능 평가

품사	총 어절 수	여과 어절 수	정확률(%)
수식언(I, II)	29,970	29,934	99.88
용언	76,027	65,100	85.63
대명사	13,109	11,579	88.33
의존명사	16,728	12,287	73.45
수사	6,824	6,395	93.71
감탄사	1,734	1,289	74.33

6) 본 논문에서는 대명사 등과 같이 길이가 짧은 단어들은 정확하게 인식되지 않는 경향이 있다. 예를 들면 "나는"는 대부분의 경우 용언 모듈에서 여과된다.

야기되는 오류와 비슷한 이유이다. 수식언 여과 모듈에서 야기되는 오류의 대부분은 미등록어에 해당된다. 본 논문에서 제안된 모델에서 수식언에 대한 미등록어는 간단한 경험규칙을 이용하므로 이에 대한 개선이 필요하다.

4.3.2 여과단계의 오류가 복합명사 분리에 미친 영향

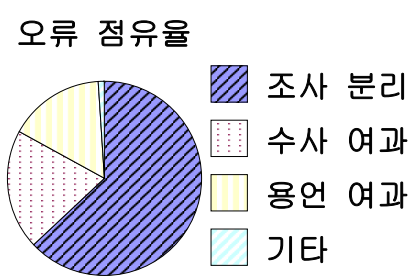
(그림 7)은 여과단계의 오류가 복합명사 분리에 미친 영향을 분석한 결과이며, 잘못 여과된 전체 오류 296개 중 각 단계에서 이들이 차지하는 비율이다. 즉 복합명사가 포함된 어절이 여과 단계에서 여과되어 복합명사로 인식하지 못한 경우에 대한 오류 분포이다. 복합명사는 조사 분리의 오류가 차지하는 비율이 전체의 약 63%를 차지한다. 수사 인식 단계의 오류가 약 20%를 차지하며, 용언 인식 모듈의 오류가 약 16%를 차지한다. 나머지 모듈의 오류는 거의 없었다. 이 경우도 용언의 여과나 조사 분리가 차지하는 비율이 상당히 높으므로 후방향-전방향 알고리즘에서 사용하는 상호정보와 분리확률에 대한 특징들이 추가되거나 수정·보완되어야 함으로 말하고 있다.

5. 비교 분석 및 응용

5.1 다른 방법과의 비교 분석

본 논문에서 제안된 명사추출 알고리즘은 문서요약을 위해서 개발되었으며, 형태소 분석기를 사용하지 않고[7], 사전과 약간의 통계적인 정보를 이용해서 명사를 추출한다. 따라서 품사 태거와 형태소 분석을 이용하는 경우보다는 정확률 면에서는 좋지 않다.

언어분석 도구를 사용하지 않는 명사추출 방법으로는 [8]이 있다. [8]은 학습데이터를 이용해서 명사추출을 위해 규칙을 생성하고, 생성된 규칙과 사전을 이용해서 명사를 추출한다. 사전을 트라이(trie)를 사용하며, 일반적인 트라이와 좀 달리 복합명사 추정을 위해 학습 동안에 명사로서 서로 겹치는 부분을 표시하는 방법으로 복합명사를 추정한다. 본 논문에서는 복합명사 분리를 위해서 수정된 CYK 알고리즘을 사용한다. 성능을 비교해보면 [8]은 재현율이 91%이고 정확률이 77%였다. [8]이 재현율 면에서는 더 좋은 결과를 보였으나, 정확률은 제안된 시스템이 훨씬 더 좋았다. 이를 F-척도로 비교해 보면 [8]은 83.42%인 반면에 제안된 시스템은 89.30%이다.



(그림 7) 여과 단계의 오류가 미친 복합명사 분리에 대한 오류 분포

언어분석 도구(형태소 분석)를 사용하지만, 본 논문에서 제안된 방법과 비슷한 개념을 가진 명사추출 방법으로 [18]이 있다. <표 3>은 본 논문과 [18]의 개념적 차이를 보이고 있다.

제안된 모델과 [18]과의 차이는 기본적으로 명사를 포함하지 않는 어절을 먼저 여과하려고 하는 시도는 매우 흡사하다. 그러나 이를 해결하는 방법 면에서는 <표 3>과 같은 차이를 보인다. 성능 면에서 보면, 같은 말뭉치에 대해서 정확률은 [18]이 91.67%로 본 논문에서 제안된 모델보다 좋았으나, 재현율은 83.37%로 본 논문에서 제안된 모델보다 못했다. 이 결과는 언어분석도구를 사용하는 정도에 따라 정확률을 높일 수 있다고 있다는 증거이기도 하다.

복합명사 분해에 관한 연구들 중에서 가장 유사한 방법은 [9]이다. [9]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 경험규칙 MNPR을 사용하고 있다. 복합명사 분해를 위해서는 사용된 경험규칙은 [9]에서 사용한 경험규칙을 그대로 사용하였으며 특별한 통계정보를 사용하지 않고 있다.

5.2 한국어 기준명사 추출 시스템의 적용: 문서요약

제안된 기준명사 추출 시스템은 실용성을 평가하기 위해 한국어 문서요약 시스템에 적용해보았다[17]. 문서요약 시스템을 평가하기 위한 두 종류의 평가용 말뭉치를 사용하는데 이들은 실험실에서 자체적으로 개발된 것이다[17]. 첫 번째 말뭉치는 문서요약 분야에서 흔히 사용하는 논문(PAPER)이고 100편의 논문으로 구성되었다. 논문의 특성을 관찰하기 위해서 논문 전체(PAPER-ALL)과 서론과 결론 부분(PAPER-InCon)으로 나누어 평가해보았다. 두 번째 말뭉치는 신문기사(NEWS)이며 KORDIC 말뭉치[20]로부터 추출된 105건의 신문기사이다. <표 4>는 기준명사 추출 시스템을 한국어 문서요약 시스템에 적용한 결과의 성능이며, 전체적으로 비교적 좋은 성능을 보였다. PAPER-InCon에서 가장 성능을 보였고 PAPER-ALL에서 가장 좋지 않은 성능을 보였다.

〈표 3〉 제안된 모델과 [18]의 비교

특성	제안된 모델	[18]의 모델
비명사구 여과	여과 기법 - 수식언 여과 - 용언 여과 - 대명사 여과 - 의존명사 여과 - 수사 여과	배제정보 이용 - 음소 배제정보 - 음절 배제정보 - 어절 배제정보
형식형태소 분리	분리 기법 - 후방향-전방향 알고리즘	후음절 분석
형태소 분석	이용하지 않음.	이용함.

7) 일부의 어절에서 형태소를 분리하고 어절의 구조를 분석하고는 있으나, 완전한 형태소 분석 기능을 수행하지 않는다.

〈표 4〉 기준명사 추출에 의한 문서요약 시스템의 성능 평가

압축률	말뭉치	정확률	재현율	F-척도
10%	PAPER-ALL	33.5	58.8	42.7
	NEWS	44.4	18.0	25.6
	PAPER-InCon	83.9	33.0	47.4
20%	PAPER-ALL	20.5	72.3	31.9
	NEWS	43.5	27.1	33.4
	PAPER-InCon	76.6	46.2	57.6

6. 결 론

본 논문은 여과 기법과 분리 기법을 이용한 한국어 기준명사 추출 시스템을 기술하였다. 여과 기법은 명사를 포함하지 않는 어절, 즉, 용언, 수식언, 독립언을 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 CYK 알고리즘을 사용한다.

본 논문에서 제안된 기준명사 추출 방법은 ETRI 말뭉치를 대상으로 약 89%의 재현율과 정확률을 보였으며 한국어 문서요약 시스템[17]에 적용했을 때, 좋은 결과를 보였다. 그러나, 아직 개선되어야 할 문제를 많이 안고 있다. 용언을 제거하기 위한 충분한 특징(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 충분히 연구할 가치가 있는 것으로 판단된다.

참 고 문 헌

[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] 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.

[4] 심준혁, 김준석, 이근배, "통계와 규칙을 이용한 강인한 품사태거", *제1회 형태소 분석기 및 품사태거 평가 워크숍 발표논문집*, pp.60-75, 1999.

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

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

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

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

[9] 윤보현, 조민정, 임해창, "통계정보와 선호 규칙을 이용한 한국어 복합 명사의 분해", *정보과학회논문지*(B), 제24권, 제8호, pp.900-909, 1997.

[10] 박혁로, 신중호, "비터비 학습 알고리즘을 이용한 한글 복합명사 분석", *1997 한국정보과학회 가을 학술 발표논문집*, Vol.24, No.2, pp.219-222, 1997.

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

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

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

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

[15] 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

[16] Manning, C. D. and Schutze, H. *Foundations of Statistical Natural Language Processing*, The MIT Press, 1999.

[17] 김준홍, 도합유사도를 이용한 추출요약 시스템, 한국해양대학교, 컴퓨터공학과, 석사학위 논문, 2000.

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

[19] Teufel, S. and Moens, M., "Argumentative classification of extracted sentences as a first step towards flexible abstracting," in Mani, I. and Maybury, M. T., editors, *Advances in Automatic Text Summarization*, pp.155-171. The MIT Press, 1999.

[20] 김태희, 박혁로, 신중호 "검색/요약/필터링을 위한 텍스트 이해 모형 연구", 제3회 소프트과학 워크숍, 1999.

김 재 훈

e-mail : jhoon@hhu.ac.kr

1986년 계명대학교 전자계산학과(학사)

1988년 한국과학기술원 전산학과(공학석사)

1996년 한국과학기술원 전산학과(공학박사)

1988년~1997년 한국전자통신연구원 선임연구원

1997년~현재 한국해양대학교 컴퓨터공학과 교수

2000년~2002년 2월 한국과학기술원 첨단정보기술연구센터 연구원

2001년~2002년 2월 University of Southern California, Information Sciences Institute 방문연구원

2007년~2008년 8월 University of Illinois at Urbana-Champaign, Beckman Institute 방문연구원

관심분야 : 자연언어처리, 한국어 정보처리, 정보검색, 정보추출