韓國컴퓨터情報學會 論文誌
第14卷 第3號, 2009. 3.

2009-14-3-1-5

한국어 정보처리를 위한 명사 및 키워드 추출

신 성 윤*, 이 양 원*

Noun and Keyword Extraction for Information Processing of Korean

Shin Seong Yoon*, Rhee Yang Won*

요 약

언어에서 명사 및 키워드 추출은 정보처리에서 매우 필수적인 요소이다. 하지만, 한국어 정보처리에서 명사 추출과 키워드 추출은 아직도 많은 문제점을 안고 있다. 본 논문에서는 명사의 등장 특성을 고려한 효율적인 명사 추출 방법에 대해서 제시하였다. 제시한 방법은 대량의 문서를 빠르게 처리해야 하는 정보 검색과 같은 분야에서 유용하게 쓰일 수 있다. 또한 대량의 문제를 자동으로 분류하기 위하여 비감독 학습 기법에 의해 카테고리별 키워드를 구성하기 위한 방법을 제안하였다. 제안된 방법은 감독 학습 기법의 키워드 추출기법 중에서 우수하다고 알려진 X2 기법과 DF 기법보다 우수한 분류 성능을 보였다.

Abstract

In a language, noun and keyword extraction is a key element in information processing. When it comes to processing Korean language information, however, there are still a lot of problems with noun and keyword extraction. This paper proposes an effective noun extraction method that considers noun emergence features. The proposed method can be effectively used in areas like information retrieval where large volumes of documents and data need to be processed in a fast manner. In this paper, a category-based keyword construction method is also presented that uses an unsupervised learning technique to ensure high volumes of queries are automatically classified. Our experimental results show that the proposed method outperformed both the supervised learning-based X2 method known to excel in keyword extraction and the DF method, in terms of classification precision.

▸ Keyword : 명사 추출(Noun Detection), 키워드 추출(Keyword Detection), 감독학습기법(Supervised Learning)

• 제1저자 : 신성윤 교신저자 : 이양원
• 투고일 : 2009. 2. 23, 심사일 : 2009. 3. 2, 게재확정일 : 2009. 3. 10.
* 군산대학교 컴퓨터정보공학과 교수

Ⅰ. 서 론

인터넷의 발전에 따라 웹 문서가 매우 많이 늘어나고 있지만 검색엔진에서 색인화를 수행하는 웹 문서의 비율은 오히려 감소하여 사용자들이 인터넷에서 원하는 정보를 보다 빠르고 효율적으로 찾기가 어려워지고 있다. 따라서 검색엔진들은 검색의 효율을 개선하기 위해 지능형 검색 기법을 개발하고 있고, 정부기관이나 기업들이 보유한 전문 지식 문서가 매우 급격하게 증가하고 있으며, 대량의 지식 문서를 체계적으로 유지 관리하여 기업 경영에 활용하기 위한 지식관리시스템 구축이 매우 필요한 실정에 있다.

한국어 정보처리는 기초 기술이 매우 취약한 상태에서 기초 기술 보다는 응용 기술에 매진하여 다른 나라에 비해 응용 기술 분야는 크게 떨어지지는 않았다. 하지만 전자사전, 코퍼스, 구문 분석, 시소러스, 용례 사전 등 기초 기술이 매우 미약하여 응용 소프트웨어의 성능개선에 많은 문제점을 발생하고 있다. 특히 한국어가 갖는 교착어의 특성과 비정형성은 형태소 분석, 구문 분석, 사전 등의 기초 기술 개발에 커다란 문제점으로 작용하고 있다. 이러한 상태는 응용 소프트웨어 개발 및 다른 응용 분야의 연구에 결정적인 한계점으로 작용하고 있다. 이처럼 기초 기술의 미흡과 정보 자료의 부족으로 체계적이고 분석적인 방법의 연구가 어렵게 되었다. 또한 부족한 경험과 연구자의 직관에 의한 연구로 신뢰도 및 성능의 저하를 초래했다.

본 논문에서는 기 구축된 사전[1]을 이용하여, 불필요한 연산을 줄여서 수행 시간을 단축시키고, 대용량의 문서에서도 정확도에 크게 영향을 미치지 않으면서 명사를 추출할 수 있는 명사의 출현 특성을 이용한 명사 추출 방법 및 비감독 학습 기법에 의한 키워드 추출 방법을 제시한다.

Ⅱ. 관련연구

명사 추출이란 문서 내에 존재하는 모든 명사를 찾아내는 작업으로서, 한국어 정보검색에서는 문서를 대표하는 색인어 또는 키워드로서 명사를 사용한다.

한국어 정보처리에서 한국어 문장은 여러 개의 어절로 구성되고 복잡하다. 어절은 체언, 용언, 그리고 수식언 등으로 나눌 수 있으며, 대부분의 명사들은 체언에 속한다. 명사를 찾기 위해서는 어절들 중에서 일단 체언을 찾아야 한다. "형태소 분석기 및 품사 태거 평가 대회(MATEC99)"가 1999년에 열렸는데, 이 대회에서 형태소 분석기, 품사 태거, 명사 추출에 대한 평가를 수행하였다[2].

한국어 명사 추출 시스템은 크게, 품사 태거를 이용한 경우, 형태소 분석기를 이용하는 경우, 그리고 아무런 언어분석 도구를 사용하지 않는 경우의 세 가지로 분류된다[3]. 품사 태거를 이용하는 방법은 태깅 결과에서 원하는 품사에 해당하는 단어만 출력하면 되므로 품사 태거가 존재할 경우에 매우 쉽고 정확한 결과를 얻을 수 있다. 하지만 품사 태거가 존재하지 않으면 이를 구축하는데 많은 시간과 노력이 필요한 단점이 있다. 형태소 분석기를 이용하는 방법은 형태소 분석기의 결과에서 명사가 포함된 어절의 유형을 정의하고, 각 유형에 일치되는 어절은 형태소 분석 결과를 이용하여 명사 이외의 성분드를 제거하고 출력한다. 형태소 분석 방법에서 미등록어 문제 해결은 형태소 분석에 매우 의존적이며 체언 유형의 중의성이 발생될 수 있고 규칙을 이용하는 방법이므로 시스템 확장성에 문제가 발생될 수 있다는 단점이 있다. 언어분석 도구를 사용하지 않는 경우에는 사전과 규칙을 이용하여 명사를 추출한다. 이는 매우 단순하고 빠른 속도로 명사를 추출할 수 있다는 장점이 있으나 언어분석 도구를 이용하는 방법들보다 정확률이 떨어진다는 단점이 있다.

최근 연구에서는 웹 기반 접근 방법으로 검색 결과를 분석하여 단어의 쌍을 추출하는 방법[4][5]과 사전에 등록되지 않은 미등록어를 사전에 자동으로 등 ㄹㄱ하는 한국어 비등록어 사전 자동 구축 방법[6]등이 제안되었다.

본 논문에서는 명사의 출현 특성을 이용한 효율적인 한국어 명사 추출 방법[7]과 효율적인 문서 자동 분류를 위한 대표 색인어 추출 기법[8]을 각각 변형하고 하나로 통합하여 명사 및 키워드를 추출하도록 한다.

Ⅲ. 명사의 추출

복잡한 분석을 수행하기 전에 명사가 존재하지 않는 어절을 제거하기 위해 제거 정보를 사용한다. 한국어 어절에서 명사가 나타나지 않는 특성에 대한 정보를 제거 정보라 한다. 본 논문에서 사용한 제거 정보를 종류별로 분류하여 나타낸 것은 다음과 같다.

(1) 어절의 첫음절에 존재하는 특정 종성의 집합은 제거한다. 이를 음소 단위 제거 정보라 한다.

(2) 어절의 처음에 나타나는 특정 부분 어절의 집합은 제거하고, 어절의 어느 위치나 존재하는 특정 부분 어절의 집합도 제거한다. 이를 부분 어절 단위 제거 정보라 한다.

(3) 명사가 존재하지 않는 고빈도 어절의 집합은 제거한

다. 이를 어절 단위 제거 정보라 한다.

제거정보의 검사 방법은 다음 〈그림 1〉과 같다.

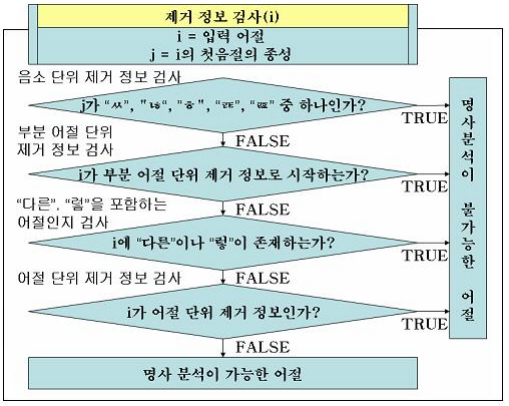

그림 1. 제거 정보 검사
Fig 1. Removal Information Test

제거정보 검사 후 단일어 검사를 수행하는데, 단일어가 명사와 다른 품사를 모두 가질 수 있는 경우에 명사와 부사인 경우는 부사로 결정, 단음절 명사와 다른 품사인 경우는 다른 품사로 결정, 2음절 이상의 명사와 다른 품사인 경우는 명사로 결정하는 순위에 따라 분석한다.

단일어 검사가 완료 되었으면 다음으로 명사 접미 음절열 분석을 수행한다. 명사 접미 음절열은 체언 뒤에 결합되는 음절의 열로서 정의하는데 명사의 출현에 대한 좋은 단서가 된다. 어절에서 명사 접미 음절열이 발견되면 바로 그 앞에 위치한 체언을 검사함으로써 복잡한 형태소 분석 과정을 거칠 필요 없이 명사를 추출할 수 있기 때문에 분석의 속도를 크게 높일 수 있다. 한국어는 교착어이므로 명사와 결합될 수 있는 형식 형태소의 조합은 이론적으로는 매우 많다. 하지만 실제 언어 현상에서는 일정한 수가 반복되어 사용되므로 명사 접미 음절열을 이용하는 것은 매우 의미 있는 일이다. 명사 접미 음절열의 유형은 조사 및 "하다", "되다", "시키다" 등과 같은 용언화 접미사의 활용형과 "밥먹다"와 같이 흔히 명사와 결합되어 복합용언으로 사용되는 용언의 활용형이 있다.

다음으로 음운 현상 복원이다. 음원 복원 정보는 품사 부착된 코퍼스에서 원시어절과 품사가 부착된 어절에서 품사를 제외한 복원된 어절이 일치하지 않을 경우, 불일치가 발생한 음절로부터 끝음절까지의 한글 부분만을 저장한다. 음운 복원 정보를 이용하여 원형을 복원할 때 고려해야 될 사항은 주어진 어절의 부분문자열 중에서 복원 대상 문자열이 둘 이상 존재할 수 있고, 복원대상 문자열에 대한 복원할 문자열도 둘 이상 존재할 수 있다는 점이다. 복원대상 문자열은 문자열이 긴 것을 먼저 적용하고, 복원할 문자열은 빈도가 높은 것을 먼저 적용한다.

Ⅳ. 키워드 추출

본 논문에서는 데이터 마이닝 기법 중 하나인 연관 규칙 탐사 알고리즘을 사용하여 비감독 학습 기법에 의한 키워드 추출 기법을 제안하였다. 즉, 사전에 분류되지 않은 대량의 문제로부터 직접 키워드를 추출하기 위한 방법이다.

전체 문제에서 연관 규칙 탐사 알고리즘을 적용하여 전문 용어들 간의 연관성을 분석하고 연관 용어 집합을 구성하였다. 그리고 핵심 키워드 별로 연관성이 높은 단어들을 하나의 집합으로 구성하였다. 여기서 핵심 키워드 집합은 각 카테고리를 대표하는 특징단어 집합이다. 연관 규칙을 발견하기 위한 트랜잭션 단위는 하나의 문제에서 추출된 전문 용어 집합이다. 전문 용어 집합은 전처리 과정에서 형태소 분석 사전에 수록된 용어를 추출하여 구성하였다. 그리고 같은 의미를 가지는 동의어를 표준화하고 불필요한 연산이나 연관규칙을 양산할 수 있는 특수용어도 제거하였다. 다음 〈표 1〉은 약 500개의 컴퓨터 분야 문제에 대하여 연관 규칙 알고리즘을 적용하여 생성된 연관 용어 집합의 예이다.

표 1. 연관 용어 집합
Table 1. Association Terms Set

연관용어	연관용어집합
운영체제	운영체제, 사용자인터페이스, 프로세스, 자원, 메모리, 보안, …
네트워트	네트워크, 전송매체, 유형, 랜, 통신장치, 중계기, 허브, 스위치, …
인터넷	인터넷, 프로토콜, 유알엘, 디엔에스, 포트, 웹페이지, 검색엔진, …

문제 분류 과정에서는 키워드와 문제간의 유사도를 계산하여 문제에 대한 분류 실험을 하였다. 모든 분야를 대상으로 계산한 유사도 중에서 최대값을 가지는 분야가 해당 문제가 속하는 분야이다. 키워드와 문제은행의 문제 사이의 유사도 계산을 위하여 코사인 계수를 사용하였다. 다음 식 (1)은 키워드와 문제간의 유사도를 계산하기 위한 코사인 계수식이다. 여기서 X는 분류하고자 하는 문제에 대한 단어 벡터이고, Y는 추출된 분야별 키워드를 나타낸다.

$$\cos\theta(X,Y)=\frac{\sum_{i=1}^{n}X_iY_i}{\sqrt{\sum_{i=1}^{n}(X_i)^2\sum_{i=1}^{n}(Y_i)^2}} \quad \cdots\cdots (1)$$

제안된 키워드 추출 기법의 성능 평가는 다음과 같다.

문제 분류 결과에 대한 성능 평가를 위한 척도 Recall, Precision, F-measure 값을 주로 사용한다. Recall 값은 카테고리 내의 전체 문제 중에서 정확하게 분류된 문제의 분류 비율을 의미한다.

Recall = 정확하게 분류된 문제/전체문제 × 100

Precision 값은 분류된 문제 중에서 정확하게 분류된 문제의 비율을 의미한다.

Precision = 정확하게 분류된 문제/분류된 문제 × 100

이러한 Recall과 Precision 값은 서로 반비례 관계에 있으므로 적절한 조정 과정이 필요하다. Lewis 등은 Recall과 Precision을 결합한 F- Measure 개념을 제안하였는데 다음 식 (2)은 F-measure에 대하여 정의한 식이다.

$$F_\kappa=\frac{(\kappa^2+1)\bullet \mathrm{Precision}\bullet \mathrm{Recall}}{\kappa^2\bullet \mathrm{Precision}+\mathrm{Recall}} \quad \cdots\cdots (2)$$

κ는 Recall 값과 Precision 값의 중요도에 따른 가중치를 나타낸다. 즉 κ=0 이면 F 값은 Recall 값과 동일하다. κ=∞이면 F값은 Recall 값과 동일하다. κ=1이면 Recall 값과 Precision 값에 동일한 가중치를 적용하여 Recall 값에 동일한 가중치를 적용하여 F 값을 계산한다. 그리고 κ=0.5이면 Recall 값에 0.5배의 가중치를 적용하여 Precision 값에 대한 중요도를 높여서 계산한다. κ=2이면 Recall 값에 2배의 가중치를 적용하여 Recall 값에 대한 중요도를 높여서 계산한다. 그러므로 Recall 값과 Precision 값의 중요도에 따라 κ의 가중치를 선택적으로 조정할 수 있다.

본 논문에서는 제안된 키워드 추출 기법의 정확성을 검증하기 위하여 κ=1 즉, Recall 값과 Precision 값에 동일한 가중치를 적용하여 분류 성능을 평가하였다.

이렇게 추출된 키워드는 키워드 DB에 저장되어 새로운 문제의 입력 시에 문제은행과 비교를 통하여 키워드가 존재하는 문제를 모두 나열하게 된다. 이를 보고 사용자는 유사 또는 비유사 문제로 판별하게 된다. 또한 새로운 키워드는 키워드 DB에 추가하여 다음 문제 입력 시에 키워드로서 활용되도록 하였다.

V. 실험

실험 환경으로, 컴퓨터 시스템은 펜티엄 4-1.3GHz의 512M RAM에서 운영체제는 Windows XP를 이용하였다. 컴파일러는 Visual C++ 6.0을 이용하였으며 DB는 MySQL을 이용하였다.

다음 〈표 2〉는 분류 대상 문제에 대하여 코사인 계수를 사용하여 분야별 키워드와 분류대상 문제간의 유사도를 계산하여 해당 문제를 가장 유사한 카테고리에 분류하였다.

표 2. 제안 기법으로 추출한 키워드 집합
Table 2. Detected Keyword Set by Proposed Method

분야	키워드
운영체제	운영체제, 사용자인터페이스, 프로세스, 자원, 메모리, 보안, 드라이버, 부팅, 커널, 유닉스, 도스, 플랫폼, 단일작업, 다중작업, 다중프로세싱, 네트워크 운영체제, 시분할, 자원, 교착상태, 폴더
네트워트	네트워크, 전송매체, 유형, 랜, 통신장치, 중계기, 허브, 스위치, 무선기술, 전송매체, 프로토콜, 게이트웨이, 이더넷, 주파수변조, 광섬유, 진폭변조, 비동기전송모드, 브리지, 위성, 라우터
인터넷	인터넷, 프로토콜, 유알엘, 디엔에스, 포트, 웹페이지, 검색엔진, 아이피주소, 자바스크립트, 에이치티엠엘, 에이에스피, 엑스엠엘, 하이퍼링크, 패킷, 티시피아이피, 인터넷서비스제공자, 백본, 웹서버, 텔넷, 엔에이티,

다음 〈표 3〉은 Recall 값, Precision 값, 그리고 F-measure 값을 이용한 제안 기법, X2 기법, DF 기법을 비교 실험한 결과이다. 제안 기법은 분야별로 Recall 값의 차이가 있지만 평균적으로 다른 기법보다 가장 우수한 성능을 보였다.

표 3. Recall, Precision, F-measure 값 비교
Table 3. Comparison of Recall, Precision and F-measure Values

Recall	운영체제	네트워크	인터넷	평균
제안기법	0.75	0.63	0.57	0.65
X2 기법	0.51	0.71	0.46	0.56
DF 기법	0.55	0.62	0.59	0.59

Precision	운영체제	네트워크	인터넷	평균
제안 기법	0.45	0.76	0.91	0.71
X2 기법	0.54	0.59	0.78	0.55
DF 기법	0.46	0.53	0.43	0.47
F-measure	운영체제	네트워크	인터넷	평 균
제안 기법	0.56	0.69	0.7	0.65
X2 기법	0.52	0.64	0.58	0.58
DF 기법	0.5	0.57	0.5	0.52

실험결과, 제안 기법은 운영체제와 네트워크에서 높은 Recall 값을 가졌고 분야별 Recall 값의 평균은 0.65로 가장 높았다. 또한, 제안 기법은 네트워크와 인터넷 분야에서 높은 Precision 값을 가지고, 분야별 Precision 값의 평균은 0.71로 가장 높았다. X2 기법은 운영체제 분야에서 높은 Precision 값을 가진다. X2 기법의 분야별 Precision 값의 평균은 0.55이다. DF 기법은 모든 분야에서 가장 낮은 Precision 값을 가지며, 분야별 평균 Precision 값은 0.47이다. 제안 기법의 분야별 평균 Precision 값은 0.71로 다른 기법에 비해 가장 우수한 성능을 보였다.

그러나 Recall 값은 잘못 분류된 것에 대해서는 고려되지 않기 때문에 Recall 값이 높다고 해서 정확하게 분류되었다는 것을 의미하지는 않는다. 또한 Precision 값은 분류된 문제 중에서만 정확도를 계산하기 때문에 Precision 값이 높다고 해서 정확하게 분류되었다는 것을 의미하는 것은 아니다. 그러므로 신뢰성 있는 분류 성능 측정을 위해서는 Recall 값과 Precision 값을 결합한 F-Measure 값에 대한 비교가 필요하다.

또한 제안 기법은 모든 분야에서 높은 F-Measure 값을 가지며, 분야별 평균 F-measure 값은 0.65이다. X2 기법은 제안 기법보다는 낮지만 DF 분야보다는 높은 F-measure 값을 가지며, 분야별 F-measure 값은 0.58이다. DF 기법은 제안 기법과 X2 기법보다 전 분야에서 낮은 값을 가진다. 분야별 평균 F-measure 값은 0.52이다. 제안 기법은 분야별 평균 F-measure 값이 0.65로 다른 기법에 비해 가장 우수한 성능을 보였다.

다음 〈그림 2〉는 명사 및 키워드 추출 실험을 나타낸다.

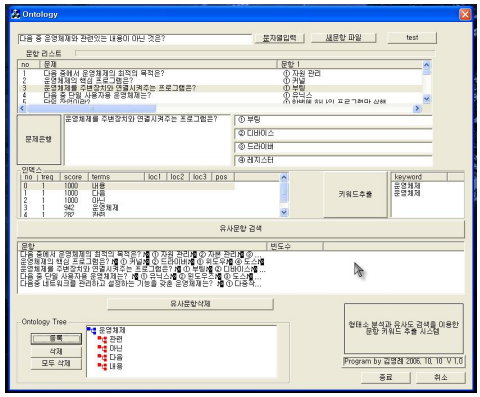

그림 2. 명사 및 키워드 추출 실험
Fig 2. Experiment of Noun and Keyword Detection

위의 〈그림 2〉에서 문제를 입력하여 문자열 입력을 선택하면 명사들이 인덱스 부분에 나타난다. 그리고 추출된 키워드가 키워드 추출 부분에 나타난다. 사용자가 유사문항 검색을 선택하면 문제 은행에서 운영체제를 키워드로 갖는 모든 문제가 나타나며 사용자는 눈으로 판단하여 유사문항을 삭제할 수 있다.

다음 〈그림 3〉은 온톨로지가 형성된 문제로서 문제를 보고 판단하여 온톨로지가 자동으로 구축된 것을 나타낸다. 문제의 온톨로지는 총 4개로서 계층형으로 구성되어 있다.

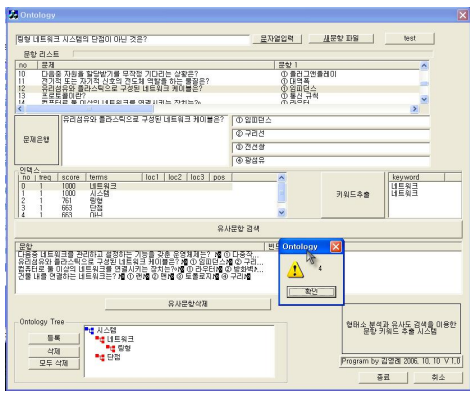

그림 3. 명사 및 키워드 추출 실험(온톨로지)
Fig 3. Experiment of Noun and Keyword Detection(Ontology)

이와 같이, 본 논문에서는 대량의 문제를 자동으로 분류하기 위하여 비감독 학습 기법에 의해 카테고리별 키워드를 구성하기 위한 방법을 제안하였다. 제안된 방법에서는 사전에 문제를 분류하지 않고 키워드를 추출하기 위하여 데이터마이닝 기법 중의 하나인 연관 규칙 탐사 알고리즘을 이용하였다.

먼저, 각 카테고리를 대표하는 핵심 키워드를 선정하고, 연관 규칙 탐사 알고리즘을 적용하여 각 핵심 키워드와 관련된 용어 집합을 추출한다. 추출된 용어 집합에서 상위 50개의 용어들로 연관 용어 집합 Ksi={T1, T2, T3, …, Tn}을 구성한다. 두 번째 단계에서는 연관 용어 집합 K_{si}의 각 원소 T_i (1<i<n)에 대해 최소 지지도 60% 이상의 상위 20개의 연관 용어를 추출하여 키워드로 구성하였다.

제안된 기법의 성능을 검증하기 위하여 컴퓨터 관련 분야를 대상으로 분류 실험을 하였다. 키워드 추출을 위한 학습용 문제는 컴퓨터 관련 서적에 발표한 문제를 500개를 사용하였고, 분류 실험에서는 키워드 추출 과정에서 사용하지 않은 분야별 30개의 문제를 사용하였다. 대표적인 감독 학습 기법인 X2 기법, DF 기법과의 비교 실험을 통하여 제안 기법의 성능을 평가하였다. 실험 결과 감독 학습 기법의 키워드 추출기법 중에서 우수하다고 알려진 X2 기법과 DF 기법보다 우수한 분류 성능을 보였다.

Ⅵ. 결론

본 논문에서는 대량의 문서를 빠르게 처리해야 하는 정보 검색과 같은 분야에서 유용하게 쓰이는 방법인 명사의 출현 특성을 이용한 명사 추출 방법에 대해 알아보았다. 또한 사전에 분류되지 않은 대량의 문제로부터 데이터 마이닝 기법중의 하나인 연관 규칙 탐사 알고리즘을 사용하여 비감독 학습기법에 의한 키워드를 추출하는 것에 대해서도 알아보았다. 실험 결과 제안된 방법은 감독 학습 기법의 키워드 추출기법 중에서 우수하다고 알려진 X2 기법과 DF 기법보다 우수한 분류 성능을 보였다.

현재 본 논문에서 사용한 제거 정보는 수작업으로 추출하였으나 코퍼스로부터 자동으로 획득하는 방법에 대해서도 연구할 계획이다. 또한 키워드 추출시에도 감독 학습 기법과 비감독 학습 기법 구분을 짓지 않고 훌륭한 성능을 보이는 방법을 찾는 것이 필요하다.

참고문헌

〔1〕 정민수, "*코퍼스로부터 구문분석을 위한 사전 구성*," 군산대학교 대학원 석사학위 논문, 1999년 2월

〔2〕 이재성, 박재득, 차건희, 박세영, "형태소 분석기 및 품사 태거 평가대회(MATEC99) 개요," 제1회 형태소 분석기 및 품사태거 평가 워크숍 논문집, 13-22쪽, 1999년 10월

〔3〕 김준홍, 김준홍, 김재훈, 박호진, "문서요약을 위한 한국어 기준명사 추출 시스템," 한국해양대학교 산업기술연구소 연구논문집, 제 19권, 169-184쪽, 2002년

〔4〕 Masaaki NAGATA, Teruka SAITO, Kenji SUZUKI, "Using the web as a bilingual dictionary," Proceedings of the workshop on Data-driven methods in machine translation, Vol. 14, pp. 1-8, July 2001.

〔5〕 QING LI, SUNG HYON MYAENG, YUN JIN, KANG Bo-Yeong, "Translation of Unknown Terms via Web Mining for Information Retrieval," Asia Information Retrieval Symposium No 3, vol. 4182, pp. 258-269, Oct. 2006.

〔6〕 박소영, "웹문서에서의 출현빈도를 이용한 한국어 미등록어 사전 자동 구축", 한국컴퓨터정보학회논문지, 제 13권, 제 3호, 27-33쪽, 2008년 5월.

〔7〕 Lee D. G., Lee S. Z., Rim H. C., "An Efficient Method for Korean Noun Extraction Using Noun Patterns," Journal of Korean Information Science Society, Vol. 30, No. 2, pp. 173-183, 2003년 2월.

〔8〕 김지숙, 김영지, 문현정, 우용태, "효율적인 문서 자동 분류를 위한 대표 색인어 추출 기법," 정보기술과 데이터베이스저널, 제 8권 제 1호, 295-302쪽, 2001년 6월.

저 자 소 개

이 양 원

1994년 8월 숭실대학교 전자계산학과 공학박사

1986년~현재 군산대학교 컴퓨터정보과학과 교수

〈관심분야〉 모바일 프로그래밍, 텔레매틱스, 가상현실

신 성 윤

2003년 2월 군산대학교 컴퓨터과학과 이학박사

2006년~현재 군산대학교 컴퓨터정보과학과 교수

〈관심분야〉 비디오 인덱싱, 비디오 요약, 멀티미디어