비감독 학습 기법에 의한 키워드 추출

신성윤* · 백정욱* · 이양원*

*군산대학교

Keyword Extraction Using Unsupervised Learning Method

Seong-Yoon Shin* · Jeong-Uk Baek* · Yang-Won Rhee*

*Kunsan National University

E-mail : syshin{ywrhee}@kunsan.ac.kr, qortnwl0326@nate.com

요 약

명사 추출이란 문서 내에 존재하는 모든 명사를 찾아내는 작업으로서, 한국어 정보검색에서는 문서를 대표하는 색인어 또는 키워드로서 명사를 사용한다. 본 논문에서는 기 구축된 사전을 이용하여 키워드를 추출하는 방법을 제시한다. 이 방법은 불필요한 연산을 줄여서 수행 시간을 단축시켰다. 그리고 대용량의 문서에서도 정확도에 크게 영향을 미치지 않으면서 명사를 추출할 수 있다. 본 논문에서는 명사의 출현 특성을 이용한 명사 추출 방법 및 비감독 학습 기법에 의한 키워드 추출 방법을 제시한다.

ABSTRACT

Noun extraction is to find all nouns presented in the document, Korean information retrieval uses noun as index terms or keywords of representing the document. In this paper, we proposes the method of keyword extraction using pre-built dictionary. This method reduces the execution time by reducing unnecessary operations. And noun, even large documents without affecting significantly the accuracy, can be extracted. This paper proposed noun extraction method using the appearance characteristics of the noun and keyword extraction method using unsupervised learning techniques.

키워드

Noun Extraction, Korean Information, Retrieval, Keyword Extraction, Unsupervised Learning Technique

Ⅰ. 서 론

본 논문에서는 기 구축된 사전[1]을 이용하여, 명사의 출현 특성을 이용한 명사 추출 방법 및 비감독 학습 기법에 의한 키워드 추출 방법을 제시한다.

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, 여기서 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 경우로 분류된다[3].

최근 연구에서는 웹 기반 접근 방법으로 검색 결과를 분석하여 단어의 쌍을 추출하는 방법[4][5]과 사전에 등록되지 않은 미등록어를 사전에 자동으로 등ㄹㄱ하는 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.

Ⅱ. 명사 및 키워드 추출

복잡한 분석을 수행하기 전에 명사가 존재하지 않는 어절을 제거하기 위해 제거 정보를 사용한

다. 한국어 어절에서 명사가 나타나지 않는 특성에 대한 정보를 제거 정보라 한다.

제거정보 검사 후 단일어 검사를 수행하고, 단일어 검사가 완료 되었으면 다음으로 명사 접미 음절열 분석을 수행한다. 그리고 다음으로 음운 현상 복원을 수행한다.

본 논문에서는, 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.

문제 분류 과정에서는 키워드와 문제간의 유사도(식 (1))를 계산하여 문제에 대한 분류 실험을 하였다. 여기서 X는 분류하고자 하는 문제에 대한 단어 벡터이고, Y는 추출된 분야별 키워드를 나타낸다.

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

문제 분류 결과에 대한 성능 평가를 위한 척도 Recall, Precision, F-measure 값을 주로 사용한다. Recall(R) 값은 카테고리 내의 전체 문제(T) 중에서 정확하게 분류된 문제(C)의 분류 비율을 의미한다.

R = C/T ×100

Precision(P) 값은 분류된 문제(S) 중에서 정확하게 분류된 문제(C)의 비율을 의미한다.

P = C/S ×100

Lewis 등은 Recall과 Precision을 결합한 F-Measure 개념을 제안하였는데 다음 식 (2)은 F-measure에 대하여 정의한 식이다.

$$F_\kappa=\frac{(\kappa^2+1)\bullet P\bullet R}{\kappa^2\bullet P+R} \qquad (2)$$

III. 실험 및 결과

제안한 기법으로 추출한 키워드에서, 다음 표 1은 F-measure 값을 이용한 제안 기법, X^2 기법, DF 기법을 비교 실험한 결과이다. 제안 기법은 평균적으로 다른 기법보다 가장 우수한 성능을 보였다.

표 1. F-measure 값 비교

F-measure	자동차	출판사	교과목	평 균
제안 기법	0.55	0.69	0.68	0.64
X^2 기법	0.5	0.63	0.57	0.57
DF 기법	0.49	0.55	0.51	0.52

IV. 결 론

본 논문에서는 명사의 등장 특성을 고려한 효율적인 명사 및 키워드 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 추출 기법 중에서 우수하다고 알려진 X^2 기법과 DF 기법보다 우수한 분류 성능을 보였다.

참고문헌

[1] 정민수, “코퍼스로부터 구문분석을 위한 사전 구성,” 군산대학교 대학원, 1999년 2월

[2] 이재성, 박재득, 차건희, 박세영, “형태소 분석기 및 품사 태거 평가대회(MATEC99) 개요,” 제1회 형태소 분석기 및 품사태거 평가 워크숍, 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, October 2006.

[6] 박소영, “웹문서에서의 출현빈도를 이용한 한국어 미등록어 사전 자동구축”, 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, 제8권 제1호, 295-302쪽, 2001년 6월.