2001년 한국정보처리학회 추계 학술발표논문집 제8권 제2호

의존관계에 근거한 키워드 추출방법

정규철, 이진관, 이태헌, 박기홍
군산대학교 컴퓨터정보과학과
e-mail:{kcjung, leejinkwan, thlee, spacepark}@kunsan.ac.kr

Keyword Abstraction Method to be based in Dependence

Kyu-Cheol Jung, Jin-Kwan Lee, Tae-Hun Lee, Ki-Hong Park
Dept of Computer Information Science, Kunsan National University

요약

논문의 키워드는 논문을 읽을지 여부를 알아보는 아주 중요한 요소로 존재이다. 그러나 키워드가 되는 단어가 원문 중에 존재하지 않고, 키워드의 구성 단어로 분리하여 존재하는 경우에는 대처할 수 없다. 본 논문에서는 문서를 읽기 위한 판단의 재료가 되는 키워드의 추출을 목적으로 하고, 특히 복합명사 생성 규칙을 이용하여 키워드의 구성 단어로 분리되어 존재하는 키워드를 추출하는 방법을 제안한다.

키워드 : 키워드 추출 방법, 복합명사 생성 규칙, 키워드 구성 단어

1. 서론

키워드는 정보 검색 분야뿐만 아니라 자연 언어 처리 분야에서 유용하게 이용된다. 특히 논문의 키워드는 논문을 읽을지 여부를 알아보는 아주 중요한 요소로 존재한다.

키워드 자동 추출에 대해서는 지금까지 정보 검색의 저동 색인 구축을 목적으로 하여 단어의 출현 빈도나 출현 위치등의 표면적 정보를 이용한 추출 방법[1]~[5]이나 구문 구조, 단어의 의미 분류 등, 언어 정보를 도입하는 방법[5]~[7]등이 제안되고 있다. 종래의 방법으로는 문서의 내용을 정확하게 표현하는 단어는 반드시 그 문서 중에 출현한다[8].라는 가정 아래 문서 내에 존재하는 단어 자신을 키워드로서 추출하고 있다. 그러나 이 방법들은 키워드가 되는 단어가 원문 중에 존재하지 않고, 키워드의 구성 단어로 분리하여 존재하는 경우에는 대처할 수 없다[2].

본 논문에서는 문서를 읽기 위한 판단의 재료가 되는 키워드의 추출을 목적으로 하고, 특히 복합명사 생성 규칙을 이용하여 키워드의 구성 단어로 분리되어 존재하는 키워드를 추출하는 방법을 제안한다.

2절에서는 저자가 직접 만든 키워드(저자 키워드)의 패턴을 분석하고 그 결과에 관하여 설명한다. 그리고 3절에서는 의존관계에 근거하는 복합명사 생성 규칙, 4절에서는 향후 과제와 결론을 맺는다.

2. 키워드 구성 단어의 패턴 분석

저자가 부여한 키워드는 요약 키워드로써 문서를 읽을 것인지 안 읽을 것인지를 판단하기 위한 지표가 된다. 그래서 본 절에서는 키워드가 수록된 정보처리학회 논문(2001년 1월 ~ 6월) 135개중 한글 키워드만을 가진 파일 중 65개 파일의 요약을 이용하여 저자 키워드의 특징을 추출하기 위한 패턴 분석을 하였다(표1 참조). 요약에만 추출한 이유는 본문의 내용을 모두 분석하면 보다 더 정확한 키워드를 추출할 수 있겠지만 다시 보면 엉뚱한 키워드를 추출

할 가능성도 그만큼 커지게 된다[12]. 본 논문에서는 요약의 내용만을 가지고 키워드를 추출하였다.

특히, 저자가 부여하는 키워드에는 문서 중에 그대로 출현하지 않는 것을 대상으로 하여, 키워드를 구성 단어(형태소)로 분할하면,

(A) 문중에 전부 존재하다;
(B) 문중에 일부 존재하다;
(C) 문중에 전혀 존재하지 않다;

라는 3그룹으로 분류해서 분석을 했다.

다음은 추출되는 패턴의 예를 보여 준다. 「→」은 왼쪽의 문자열에서 오른쪽의 키워드가 추출되는 것을 의미한다.

표 1 분석에 이용한 키워드와 초록 정보

데이터 파일수	65
키워드 정보	
평균키워드 수	4.0
최대 키워드 수	6
최소 키워드 수	2
키워드 총수	263
복합 키워드 수	217
기본 단어 키워드 수	46
평균 단어 구성 수	2.11
최대 구성 단어 수	6
최소 구성 단어 수	1
초록 정보	
총 문수	363
평균 문수	5.6
최대 문수	15
최소 문수	2
총 사이즈(KB)	43.2

2.1 키워드 추출의 패턴

(A) 문중에 전부 존재하는 경우
(A-1) 의존 관계에 의한 단어의 추출
「**좌표를 변환한다 → 좌표 변환**」
「추적하는 **알고리즘 → 추적 알고리즘**」
(A-2) 지시대명사의 대응 관계를 고려한 추출
「**언어**로 이야기하고, 그것을 **습득한다 → 언어 습득**」
(A-3) 복수의 문에 분류하는 단어에서의 추출
「**음성**을 컴퓨터로 처리한다. 그 때문에 올바른 **인식**이 필요하다. → **음성 인식**」

(B) 문중에 일부 존재하는 경우
(B-1) 복합어의 변형에 의한 추출
복합어의 구성 단어의 한쪽이 동의어나 유의어 또는 단축어로 변환되는 패턴
「단어 **빼어내다** → 단어 **추출**」
「학습 **방법** → 학습**법**」
(B-2) 복수의 문에 존재하는 단어의 공기정보에 의한 추출
「말」과 「인식」의 공기정보에서 「음성 인식」을 추출.
「인간의 말을 기계로 처리한다. 그 때문에 올바르게 **인식**하게 할 필요가 있다. → **음성 인식**」

(C) 문중에 전혀 존재하지 않는 경우
(C-1) 연상되는 분야 명이나 추상적인 단어의 추출
「추론 지식 → **인공 지능**」
「품사를 부여할 수 있는 → **형태소 해석**」
(C-2) 영어 또는 영어의 단축 단어에서 한국어로 변환(역도 포함)되어 추출
「back-off → **백 오프**」
「문맥 자유 문법 → **CFG**」

2.2 분석 결과에 대한 고찰

패턴 (A-1)은 키워드의 구성 단어가 분리한 예지만, [9]가 제안한 복합어의 의존 규칙을 개선하여 추출이 가능하다. (A-2), (A-3), (A-4)의 추출은 복잡한 의미 해석이 필요하다. 또 (B-1), (B-2), (C-1)에 관해서도 단어의 개념을 이용하는 규칙을 만드는 것으로 추출이 가능하다. (C-2)는 변환 사전 등을 만드는 것에 추출이 가능하다.

본 논문에서는 의존 규칙 생성을 위한 것이므로 패턴 (A-1)을 대상으로 한다 그것을 의존 관계에 근거하는 규칙이라고 부르고 3절에서 설명한다. 또 생성 규칙에 의해 추출 된 복합어를 키워드 후보라고 부른다.

3. 의존 관계에 근거하는 키워드 추출 방법

의존 규칙을 개선하여 요약 키워드를 추출하기 위한 복합명사 생성 규칙[1]을 제안한다. 아래에 규칙의

1) 규칙의 기술에는 [10]이 제안한 다독성 규칙과 조합 엔진을 확장하여 이용했다.

예를 보여 준다.

[규칙1]: x(보통 명사$^+$) 을, 를 y(서술형 명사) → xy

예: 얼굴을 인식하다 →얼굴 인식

[규칙2]: x(보통 명사$^+$)+ 의 +y(보통 명사$^+$) → xy

예: 기지국의 고장은 → 기지국 고장

[규칙3]: x(보통 명사$^+$) + 하는 + y(보통 명사$^+$) → xy

예: 추적하는 알고리즘 → 추적 알고리즘

[규칙4]: x(보통 명사$^+$) + 에 + 의한 + y(보통 명사$^+$)→ xy

예: 선형에 의한 탐색 → 선형 탐색

[규칙5]: x(보통 명사$^+$) + (으)로 + y(보통 명사$^+$)→ xy

예: 블록으로 입력 → 블록 입력

여기에서 기호 $x(a^+)$, $y(b^+)$는 품사 a, b가 1회 이상 연속적으로 구성되는 단어 x, y를 의미하고, 왼쪽의 품사패턴과 적합한 문중의 형태소에서 오른쪽의 복합명사 xy을 생성하는 것을 표현한다. 또 동일 장소에 대한 규칙의 적용은 1회만 한다.

4. 실험 및 평가

우선 실험은 저자 키워드를 정확하게 만들어내는 정확률 실험은 하지 않았다. 이유는 본 논문의 키워드 추출은 정확하게 저자가 만들어낸 키워드를 만들어 내고자 함이 아니라 요약을 이용한 주제어를 추출하는데 중점을 두었기 때문이다. 추출 키워드의 타당성을 평가하기 위해 2절에서 선정한 65개의 파일을 제외한 논문 중 키워드가 한글로 만 이루어진 65개를 선정하여 3사람의 피험자에게 추출키워드가 초록에 대한 키워드로 적합한지 아닌지를 다음과 같이 판정 받았다.

A의 평가 : 적합
B의 평가 : 부적합

판정 결과는 3인 모두가 A평가를 받은 경우에 대해서만 키워드로 타당하다고 판단하여 실험한 결과 65%가 키워드로서 타당하다고 평가를 받았고, 나머지는 키워드로써 타당하지 않다고 평가를 받았다. 키워드로써 타당하지 않다고 평가를 받은 것 중 60%는 키워드로 성립이 되지 않는 단어 예를 들면 「방법 검토」, 「결과 보고」, 「검토 결과」와 같은 논문 특유의 표현에서 생성되는 복합어와 복합어의 구성 단어가 안 되는 단어 「경우」, 「하나」이 있는 경우이다. 이를 해결하기 위해 불용어 사전의 도입으로 타당성이 증가할 것으로 생각된다.

4. 향후 과제 및 결론

본 논문에서는 문서를 읽기 위한 판단 재료가 되는 요약 키워드의 추출을 목적에 주고, 복합명사 생성 규칙을 이용하여 문서 중에 나타나지 않는 키워드를 추출하는 방법을 제안했다. 키워드의 추출에는 인간이 떨어진 문자열을 합성하고 키워드를 추출하는 점에 주목하여 의존 관계에 근거하는 규칙을 이용한 규칙을 생성하였다.

논문의 요약에서 키워드의 패턴을 분석하다 보면 개념규칙을 중심으로 하는 동의어 사전이나 유의어 사전의 필요성을 알 수 있었다. 구축 표층어의 사전 등록에 관해서는, 전문 용어가 표층어로서 사전에 등록되어 있지 않으나 현재 번창하게 행해지고 있는 전문 용어의 추출에 관계하는 연구[11]를 이용하여 개념규칙을 구축하면 추출 정밀도가 향상할 수 있다고 생각한다.

참고문헌

[1] 남영준, "색언어 형태분석에 의한 한국어 자동 색인기법 연구" 중앙대 박사학위논문, 1994

[2] 原正巳, 中島浩之, 木谷强, "텍스트의 포맷과 단어의 범위 내 중요도를 이용한 키워드 추출", 情處學論, Vol.38, No.2, pp.299-pp.309, 1997.

[3] 전영자, "표제와 초록의 정보량 분석에 의한 색인성 연구", 연세대학교 대학원 석사학위 논문, 1995

[4] 이창열, 강현규, 장호욱, 박세영, "자동 키워드 제작기 시스템 설계", 제5회 한글 및 한국어 정보처리 학술 발표 논문집, 1993

[5] 강승헌, 유재수, "문자열 부분 검색을 위한 색인 기법 및 성능 평가", 한국정보처리학회 논문지 제6권 제6호, 1999

[6] 안현수, "한글 문헌의 자동색인에 관한 실험적 연구" 연세대학교 대학원 석사학위 논문,1986

[7] 정진성, "단일문서 내에서의 언어 및 통계정보를 이용한 자동 색인" 한국과학기술원 석사학위논문, 1992

[8] 諸橋正幸, "자동 색인 첨가 연구의 동향", 情報處理, Vol.25, No.9, pp.918-925,Sep.1984.

[9] 宮崎正弘, "의존해석을 이용한 복합어의 자동 분할", 情處學論, Vol.25, No.6, pp.970-979, 1984.

[10] 安藤一秋, 辻孝子, 獅々堀正幹, 青江順一, "일본어 정형 표현의 패턴 기술 규칙과 효율적인 조합 알고리즘", 信學論, Vol.J80-DII, No.7, pp.1860-1869, 1997.

[11] 大畑博一, 中川裕志, "연접이 다르게 되는 단어의 수에 의한 전문 용어 추출", 情處學NL研報, 136-16, pp.119-126, 2000.

[12] 박정호, "문헌구조를 이용한 자동색인어 선정 시스템" 인하대학교 석사학위 논문, 1998.