개념 규칙을 이용한 키워드 도출방법

이태헌[0] 박기홍
군산대학교 컴퓨터 정보과학과
{thlee[0], spacepark}@kunsan.ac.kr

The Method of Deriving Keywords Using Concept Rules

Taehun Lee[0] Ki-Hong Park
Dept of Computer Information Science, Kunsan University

요 약

일반적으로 인간이 사용하는 몇 개의 주요단어를 이용하여, 문서의 분야나 주제어가 되는 일본어 키워드를 추출하는 점에 주목한다. 먼저, 학술논문에서 저자 자신이 부여한 키워드 중 분야 명이나 주제어가 문서 중에 출현하지 않는 경우를 분석하고, 단어의 개념정보를 기초로 복합어 생성규칙을 구축한다. 문서의 의미와 상관없는 키워드의 추출을 억제하기 위해 중요도 결정법을 새롭게 제안한다. 추출된 키워드의 타당성 검사를 위해 자연·음성언어에 관한 일본어 논문 65파일의 타이틀과 초록부분을 이용하여 추출된 키워드의 타당성에 대한 실험을 한 결과 추출 정밀도는 중요도의 상위 1개를 출력한 경우 75%가 되어 제안방법의 유효성을 확인할 수 있었다.

1. 서론

정보검색 분야의 발전에 따라 대량의 문서에서 원하는 정보를 찾아내는 방법이 필요하다. 일반적으로 종래의 방법은 "문서의 내용을 정확하게 표현하는 단어는 반드시 문서 중에 출현한다[1],[3]"은 가정 하에 원문내의 단어를 키워드로 추출하고, 이 단어나 복합어에 적절한 중요도를 부여하여 중요도가 높은 순으로 키워드를 결정하였다. 그러나, 원문에 키워드가 되는 단서가 존재하지 않고, 키워드의 구성단어가 문서의 여러 곳에 존재하는 경우 혹은 문서 내용에서 추출·추론 가능한 추상적인 단어(혹은 주제어)로 출현하는 경우는 효과적으로 대처할 수 없다[2]. 이 문제에 대하여 1988년 Nagata 등은 키워드를 구성하는 기본 개념(키개념)과 키워드간의 관계를 기술한 색인규칙을 미리 정의하고, 이를 이용하여 주제어를 생성하는 방법을 제안[2]하였다. 그러나 Nagata의 방법은 키개념을 추출할 때에 키개념의 관련성을 고려하지 않는 관계로 문서의 뜻에 적합하지 않는 키워드를 생성할 가능성이 있다.

본 연구에서는 문서의 주제에 적합한 키워드를 생성하기 위해 문서 내 주요단어에 대해 개념의 관련성에 따라 점수를 부여하는 중요도 계산법을 새롭게 정의한다.

제 2장에서는 개념을 이용한 생성규칙과 중요도 부여 방법을 제안한다. 3장에서는 추출된 복합어 키워드를 평가하고, 결론과 향후과제에 관하여 4장에서 논한다.

2. 키워드 추출

인간이 문서 중 몇 개의 주요 단어에서 분야 등의 추상어·주제어가 되는 키워드를 추출하는 점을 주목하고, 개념을 이용한 규칙 베이스의 복합어 키워드 추출방법을 제안한다.

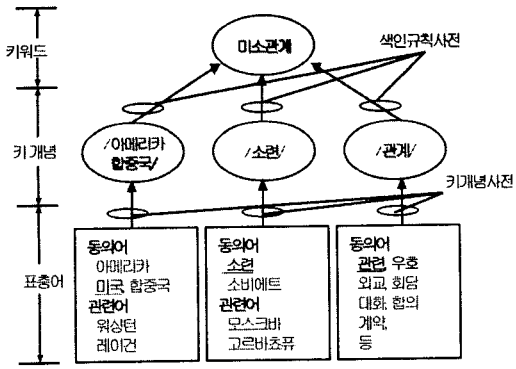

그림 1 미소 관계의 키워드 생성 규칙 예(문헌[2]으로부터 인용)

2.1 永田[2]의 복합어 생성 규칙

그림 1은 Nagata[2]의 예를 설명한다. 먼저 문장 중에서 명사를 추출하고, 그 명사의 동의어와 관련어의 모임을 표층어라고 한다. 그림 1의 예로 명사「미국」이 추출되었다면 대응하는 표층어의 동의어가 「아메리카」 등이고, 관련어는 「워싱턴」 등이다. 다음에 이 표층어의 출현 위치와 빈도의 정보를 이용하고, 키워드가 구성되는 것으로 키개념을 결정한 후 조합하여서 색인 규칙으로 키워드를 생성한다. 그림 1의 예에서는 표층어의 「아메리카」 「소련」 「관계」 등의 키개념이 결정되고, 최종적으로 키워드「미소 관계」를 얻을 수 있다. 이 방법으로는 키워드에 대한 키개념의 선정하는 방법이 형식적인 결정 식과 그 평가의 방법이 없고, 키개념의 추출할 때에 개념간의 관련성을 고려하지 않기 때문에 문장의 뜻에 적합한 키워드가 생성되지 않을 가능성이 있다.

2.2 제안한 복합어 생성 규칙

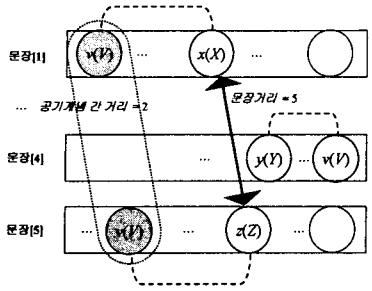

그림 2 개념간의 거리

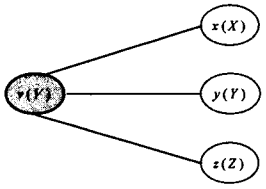

그림 3 개념의 공기정보관계

본 절에서는 Nagata의 방법을 개선하여 개념 규칙에 의해 추출된 요약 키워드는 분야 명(예를 들면 발표회의 섹션명)으로 최상위의 키워드 후보를 추출한다.

단어 w의 개념을 취득하는 함수를 CON(w)으로 표현할 때, 복합어 $w_1w_2\cdots w_n$의 생성규칙을 RULE($w_1w_2\cdots w_n$) =CON(w_1)+CON(w_2)+…+CON(w_n)라고 기술한다. 이것은 문서 중에 CON(w_1)에서 CON(w_n)의 모든 개념이 존재하는 경우만 복합어 $w_1w_2\cdots w_n$을 추출하는 것을 의미한다. 예로 복합어「意味処理」(의미처리)에 대한 생성규칙은 다음과 같다.

「意味処理」= CON「意味」(의미)+CON「処理」(처리)

또 개념을 구체화하는 단어를 Nagata의 용어를 이용 표층어라 부르고, 동의어와 유의어의 집합으로 구성되는 것으로 한다. 여기에서 개념을 직접 표현하는 단어가 동의어이고, 간접적으로 시사하는 단어가 유의어이다. 예를 들면, 개념「意味」와「処理」는 이하와 같이 된다. 예에서의 굵은 글씨는 동의어를 표현한다.

「意味」=[{**「意義」(의의)**, **「価値」(가치)**, ...},
{「内容」(내용),「文意」(문의), ...}]

「処理」=[{**「処置」(처치)**, **「処分」(처분)**, ...},
{「解決」(해결), ...}]

이상으로 복합어의 추출방법 설명했다. 먼저 문서 중에서 표층어를 검출하고, 그 개념을 추출한 후 다음에 추출된 개념과 생성 규칙의 매칭에 의해 복합어를 추출해 낸다. 이것을 키워드 후보라고 한다.

2.3 키워드 후보에 대한 중요도의 계산

2.3.1 개념간의 거리 d

개념간의 거리 d를 중요도의 지표라고 할 때, 문간의 거리[1]를 이용하는 것이 생각할 수 있지만, 문간의 거리만으로는 개념간의 관련성을 알 수 없다. 제안 방법으로는 개념의 공기정보 관계[2]에 주목하고, 공기정보 개념을 사이에 두었던 거리 d를 중요도의 지표라고 한다.

그림 2를 이용하여 개념의 거리를 설명한다. 또한 그림 2의 v, x, y, z는 표층어이고, 그 개념을 V, X, Y, Z로 표현한다. 여기에서 X와 V, V와 Z는 공기정보 관계(파선 화살표)이고, 각각의 개념 거리를 1이라고한다. X와 Z사이의 문간 거리(실선 화살표)는 5(=5-1+1)가 되지만, X와 Z에 공통되는 공기정보의 개념 V가 존재하기 때문에 X에서 V를 사이에 두고 Z에 이르는 개념 거리 d는 2(=(1+1)/1)가 된다. 단지 문간의 거리가 개념 거리 d보다 작아지는 경우만 문간의 거리로 한다.

결국 그림 2를 개념간 거리 관계로 표현한다면 그림 3이 되고, 개념 거리를 구하는 문제는 그래프 위에서 각 개념간을 연결하는 최단 경로를 찾게 된다. 이상과 같이 X와 Z에 공통해서 공기정보를 갖는 V에 주목하는 것이고, XZ 사이의 의미적인 관련성을 고려할 수 있다.

2.3.2 공기정보 관계의 수

주제를 나타내는 개념은 문서 중에 빈번하게 출현하여 많은 개념과 공기정보 관계를 갖는 것이 많다. 즉 많은 공기정보 관계를 갖는 개념이 중요성이 높다고 말할 수 있다. 그래서 개념 X에 관계하는 공기정보 개념의 수를 N(X)으로 도입한다.

그림 2의 공기정보 관계에서 얻을 수 있는 개념간의 거리가 전부 1인 그래프가 그림 3이다. 그림 3에 있어서 개념 V가 갖는 공기정보 관계의 수는 N(V)=3이 된다. 한편 X, Y, Z에 대해 공기정보 관계의 수는 전부 1이 되므로 V는 X, Y, Z보다 중요성이 높다고 말할 수 있다.

2.3.3 중요도의 계산식

개념간의 거리 d, 개념의 공기정보 관계 수 N을 고려한 $w_i(1\le i\le n)$을 포함하는 키워드 후보의 중요도 I를 이하에 보여 준다.

$$I = \left[\frac{1}{nd}\right]\sum_{i=1}^{n}\left[\frac{(D(w_i\times\alpha)+(R(w_i\times\beta)))}{(D_T\times\alpha)+(R_T\times\beta)}\times N(w_i)\right]$$

n : 키워드 후보를 구성하는 개념의 수

$D(w_i)$: w_i에 대한 표층어(동의어)의 빈도

$R(w_i)$: w_i에 대한 표층어(유의어)의 빈도

1) 문 i, j간의 거리를 j-i+1(j≥i)라고 정의한다. 개념 수가 3이상의 경우는 개념을 조합한 수의 거리를 계산하고, 그 중에서 최고 긴 문간의 거리로 한다. 단 동일 개념의 조합이 복수 출현 할 때는 상호 거리 중 최소거리를 우선으로 한다.

2) 동일 문 내에서 복수의 표층어가 존재하는 경우는 그 단어들에 대해 개념들이 공기정보 관계가 있다고 하고 그들의 개념간의 거리는 전부 1이라고 한다.

D_T : 동의어의 총 빈도

R_T : 유의어의 총 빈도

α : 동의어에 대한 가중치

β : 유의어에 대한 가중치(단, $\alpha > \beta$)

개념 거리 d가 작고, 공기정보 관계 수$N(w_i)$가 많고, 표층어의 빈도 $D(w_i)$, $R(w_i)$가 높으면 중요도 I도 높다.

3. 실험 및 평가

3.1 요약 키워드의 타당성 평가

저자 키워드의 수는 1초록에 약 1개라고 적기 때문에 제안 방법의 유효성을 정확하게 판단할 수 없다. 그래서 본 절에서는 추출 된 키워드(추출 키워드)가 요약 키워드로서 타당할 것인가의 판단을 하는 동시에 각 생성 규칙의 유효성도 평가한다.

5인의 피험자에 65개 파일의 타이틀과 초록을 읽게 하고, 이하의 4단계로 평가하게 했다. 정확율과 적합율을

A: 요약 키워드로 적절 B: 키워드로 위화감이 없음

C: 키워드로 위화감이 있음 D: 키워드로 부적절

그리고 5인 전원이 A평가를 주었던 키워드를 요약 키워드로써 타당하다고 판단했다. 또 추출 키워드의 품질을 표현하는 지표로 추출 키워드 후보 수에 대한 요약 키워드 수의 비율을 P'으로 이하와 같이 정의한다.

$$P'(\%) = \frac{\text{요약 키워드의 수}}{\text{추출 키워드의 후보수}} \times 100$$

개념을 이용한 생성 규칙의 유효성을 보이기 위해 Nagata[2]의 방법과 비교 실험을 했다. Nagata의 방법으로는 개념의 추출 빈도와 출현 위치를 이용하고 있지만 제안 방법으로는 빈도와 공기정보 관계를 이용하는 점이 다르다. 또 Nagata들은 랭킹 방법을 제안하고 있지 않기 때문에 $N(w_i)=1$, $d=$「문의 거리」라고 하는 중요도의 식 I를 이용했고 α와β는 경험적으로 1과 0.5로 했다. 개념에 이용한 규칙과 Nagata들의 방법에 의해 추출 된 요약 키워드는 총 추출 키워드 177개 중 약 54%이었다.

중요도에 의해 키워드 후보를 상위 6개 출력한 때의 P'을 그림 4에 보여 준다. 그림중의 Proposed-D는 수 작업으로 개념이 추가되지 않는 경우를 보여 준다. 그림 4에서 Nagata의 방법에 비교하여 제안 방법(Proposed-E)의 P'가 높은 것이 확인할 수 있다. 이것은 중요도 계산에 개념의 공기정보 관계를 이용한 것이 유효성이 있음을 보여 준다. 또한 상위 1개를 출력 할 때 P'는 75%로 최고치가 됐다. 또 동의어와 유의어 사전에 기재되어 있지 않는 단어를 수 작업으로 추가하지 않는 경우 P'는 39%(상위 6)가 되었다.

그림 중의 막대 그래프는 제안 방법을 이용한 경우의 요약 키워드에 대한 재현율이고, 사선 부분은 그 증가율(Increase)을 보여 준다. 개념 규칙의 경우도 재현율

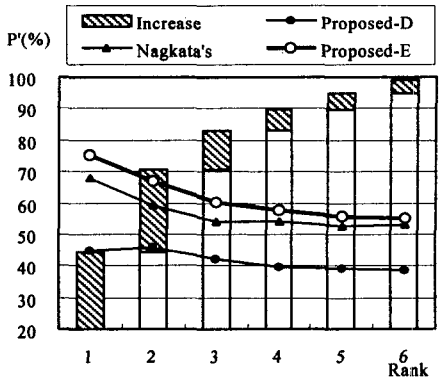

그림 4 추출 된 키워드의 중요도 계산 결과

의 증가율은 상위 1(44.2%)과 2(26.3%)일 때가 높았고 요약 키워드가 상위에 나타나는 것이 확인할 수 있었다.

4. 결론

본 논문에서는 문서를 읽기 위한 판단 재료가 되는 요약 키워드의 추출을 목적에 주고, 복합어 생성 규칙을 이용하여 문서 중에 나타나지 않는 키워드를 추출하는 방법을 제안했다. 인간이 사용하는 몇 개의 주요 단어를 편성하여서 추상어·주제어가 되는 키워드를 추출하는 점에 주목하여 개념을 이용한 규칙을 구축했다. 또 추출 정밀도를 향상하게 하기 위해 개념을 이용한 규칙에 대해서는 공기정보 관계, 개념간의 거리, 중요도 결정을 제안했다. 제안 방법의 유효성을 확인하기 위한 실험으로 Nagata의 방법과 비교하여 개념을 이용한 규칙이 상위 1개를 출력 할 때 정밀도가 75%이 됐다.

향후과제로서 추출 키워드를 미리 한정할 필요가 있기 때문에 이후는 동의어 및 유의어가 편성의 조합이 동적으로 복합어를 생성하는 기구를 생각할 필요가 있고, 전문용어에 대한 개념요소 사전을 구축하면 보다 높은 정밀도가 향상될 것으로 기대된다.

참고 문헌

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

[2] 永田昌明, 木本晴夫, "중요 개념 추출에 근거하는 신문 기사에서의 키워드 생성", 第37回情処学全国大会論文集, pp.1030-1031, 1988.

[3] 内山恵三, 中村正規, "중요 키워드 추출 방식과 그 활용 방법", 情処学DBS研報, 84-19, pp.151-161, 1991.