단어개념에 기반 한 한국어 복합키워드의 추출

김양선, 이상곤
전주대학교 교육대학원 컴퓨터교육학과
e-mail : nalalove@hitel.net, samuel@jeonju.ac.kr

A Study on Word Concept-based Compound Keyword Extraction

Yang-Seon Kim, and Sangkon Lee
Graduate School of Computer Education,
Jeonju University

요 약

문서를 읽고 그 내용을 개념상으로 정리해 보면, 그 문서를 대표할 수 있는 적은 수의 복합단어로 이루어진 키워드를 찾을 수 있다. 그러나, 문서 내에 키워드가 존재할 경우는 별 문제가 없지만, 존재하지 않을 때는 적당한 키워드 추출이 불가능해진다. 따라서, 본 논문에서는 문서 본문의 출현단어의 개념정보를 기초로 복합어 생성 규칙을 구축하고, 나아가 문서의미와 관련 있는 요소만을 정제하는 중요도 결정법을 사용하여 이에 대한 유용성을 확인하였다.

1. 서론

문서를 읽고 그 내용을 개념상으로 정리해 보면, 그 문서를 대표할 수 있는 적은 수의 복합단어로 이루어진 키워드를 찾을 수 있다. 그러나, 문서 내에 키워드가 존재할 경우는 별 문제가 없지만, 존재하지 않을 때는, 추출이 불가능해진다.

따라서, 본 연구에서는 문서에서 적당한 키워드가 출현하지 않은 경우에도 적당한 주제어 추출이 가능하도록 개념기반 복합 키워드 추출방법을 제안한다.

2. 키워드로 적당한 추출패턴

인간이 작성한 키워드 패턴을 조사하면 <표 1>과 같이 여섯 가지의 추출 패턴으로 구분할 수 있다. 이를 다시 종합하면, 다음과 같이 세 가지 키워드 패턴으로 분류할 수 있다[1, 2]. 첫 번째는 키워드가 문서 중에 모두 존재하는 경우, 두 번째는 키워드가 문서 중에 일부 존재하는 경우, 마지막으로 키워드가 문서 중에 전혀 존재하지 않는 경우 등이다. 각 단어의 개념을 이용하면 위의 여섯 가지의 패턴의 정확한 추출이 가능하다. 규칙에 기초한 키워드 생성의 준비단계로 복합어 키워드를 각 구성요소로 분할하고, 패턴분석을 수행한다.

<표 1> 키워드 추출패턴의 조사

경우	실례	추출패턴	비고
(1)	언어로 말하고 그것을 인식한다.	언어인식	지시대명사에 의한 추출
(2)	인간의 음성을 계산기로 처리하려고 한다. 먼저 그것을 올바르게 인식하는 것이 필요하다.	음성인식	복수문장에 존재하는 단어의 추출
(3)	단어꺼내기	단어추출	복합어변형에 의한 추출(유의어 사전 이용)
(4)	인간은 자신의 언어를 기계로 처리하기 원한다. 이를 올바르게 인식하기 위해 수 십 년 째 노력하여 왔다.	음성인식	복수문장에 존재하는 단어의 공기관계에 의한 추출
(5)	추론지식	인공지능	연상되는 분야나 추상적인 단어에 의한 추출
	품사를 부여할 수 있다.	형태소해석	
(6)	back-off	백오프	영어나 약어의 변환에 의한 추출
	문맥자유문법	CFG	

3. 키워드의 추출

3.1 복합어 생성규칙

<표 2> 예제 문서 ①

문번호	문장 예
s_1	MSLR파서에 의한 미정의**어처리**의 한가지 검사방법이다.
s_2	사전을 이용한 CFG**모델**에 기초한 **자연언어**해석**처리**에 의해 미정의**어**와 위의 사전에 존재하지 않는 입력문자열은 종단기호의 품사가 부여되지 않는 등 **처리**상의 문제가 있다.
s_3	한편, 효율이 좋은 **자연언어**해석의 **방법**인 일반화 된 LR(GLR)법이 있다.
s_4	미정의된단위가 없는 음소나 나누어 쓰기.. 단어에 대해서는 GLR법에 의한 연구는 … **한국어**를 대상으로 한 미정의**어처리**의 연구는 많이 실용화되지 않고 있다.
s_5	본 논문에서는 GLR법에 의해 **한국어**의 미지**어처리**에 대한 검사를 수행하였다.
s_6	LR테이블에 제약과 …을 가하여 GLR법에 확장을 가한 MSLR 파서와 ETRI 전자화사전을 이용하여 실험하였다.
<KYWD>>자연언어처리//미정의어처리//GLR법	

복합키워드는 구성단어가 그대로 문서에 존재하지 않고 출현하는 단어로 연상 혹은 추측하여야 하는 경우가 많다. 이 절에서는 나가타[4]의 방법을 이용하여 설명하겠다.

어떤 복합어 키워드 w가 하위개념을 가진 단어들로 복합 구성되어 있다고 가정하면, 이의 구성형태소들을 w에 대한 개념요소들이라 한다. 이들 개념요소들은 w의 **동의어**(同義語, Synonym)와 **유의어**(類義語, Related Term)로 구성되어 있다. 복합어 w는 각 구성요소들의 동의어나 유의어 집합으로 재구성할 수 있다. 이들 집합을 다음과 같이 [과]을 사용하여 구조화 할 수 있다.

개념 =	[동의어 +	유의어]
Concept(의미)=	[{의의,가치, …},	{내용, …}]
Concept(처리)=	[{처치,처분, …},	{해결,취급, …}]

또한 w_1w_2 ... w_n의 생성규칙(PR; Production Rule)은 각 개념요소들의 합으로 PR(w_1w_2 ... w_n) = Concept(w_1) + Concept(w_2) + ... + Concept(w_n)으로 정의한다. 이것은 문서 중에 Concept(w_1)에서 Concept(w_n)까지의 모든 개념이 존재하는 경우에만 복합어 w 즉, w_1w_2 ... w_n을 추출한다. 예를 들어, 문서 중에 복합어 "의미처리"의 생성규칙은 구성요소 "의미", "처리"와 양방향의 개념요소가 존재하면 복합어 "의미처리"를 키워드로 추출한다.

(예 1) PR(의미처리)=Concept(의미)+Concept(처리)

<표 3> <표 2>에서 추출된 개념어와 개념요소

개념어		해석	자연언어	처리	한국어	방법	모델
문장번호	s_1	-	어	처리	-	-	-
	s_2	해석	자연언어, 어	처리, 처리	-	-	모델
	s_3	해석	자연언어	-	-	방법	-
	s_4	-	어	처리	한국어	-	
	s_5	-	어	처리	한국어	-	
	s_6	-	-	-	-	-	
S(w_i)		2	6	5	2	1	1

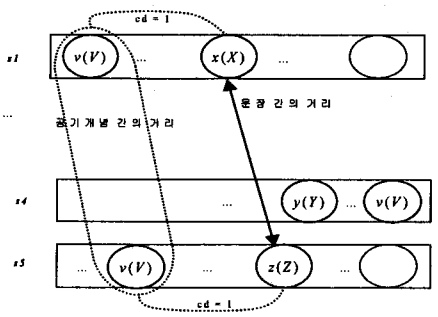
이상의 생성규칙에 의해 생성된 복합어를 키워드후보라 부른다. 생성규칙은 원문의 내용과 관계가 없는 키워드 후보가 생성되는 것을 억제시키기 위해 사용한다.

3.2 키워드후보의 중요도

추출정밀도를 향상하기 위해 새로운 중요도 계산방법이 필요하다. 또한, 중요도 계산을 하기 위해 문장간 거리(sd; sentential distance)와 개념간 거리 (cd; conceptual distance)가 필요하다.

3.2.1 개념어간 거리

중요도의 지표로 개념어의 각 요소를 포함하고 있는 문장 간의 거리를 개념어간의 거리로 이용할 수 있지만, 이 거리만으로는 개념어 사이의 관련성을 정확히 알 수 없다. 따라서 개념어의 공기관계[1]에 주목하여 개념사이의 거리를 이용한다.



(그림 1) 문장 간 혹은 개념 간의 거리

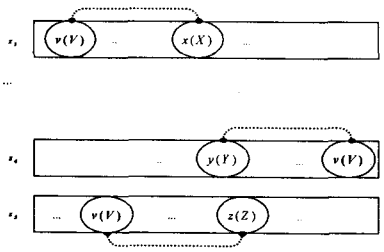
1) 동일문장 내에 동일 개념어가 복수 개 존재하면 "공기관계가 있다"라 정의하며, 개념어들 간의 거리는 1로 한다.

PX(XZ)=Concept(X)+Concept(Z)를 생각하여 보자. 문장 간 거리 sd는 (그림 1)의 화살표로 표시된 바와 같다. v, x, y, z는 문서에서 출현한 표층어이고, 영문 소문자로 표시하고 개념요소이다. 그 개념어들은 영문 대문자 V, X, Y, Z 로 표현한다.

X와 Z 사이의 개념간의 거리(XZ)는 단순히 문장간의 거리 5가 된다. 어떤 문장 i에서 j까지의 문장거리 sd는 s(j)-s(i)+1(단, j≥i)로 정의한다. 따라서, (그림 1)의 점선으로 표시한 바와 같이 XZ의 sd는 5(=5-1+1)이다. 그러나 X와 Z에 공통으로 공기하는 개념어 V에 주목하면 V를 사이에 두고 X에서 Z에 이르므로 개념간 거리 cd는 다음의 식으로 계산한다.

$$cd = \frac{(cd_1 + cd_2 + \ldots + cd_n)}{cc} \quad \cdots\cdots \text{식(1)}$$

여기서, cc는 공통개념어(common conceptual term)의 수(X), cd는 XV가 1이고, VZ가 1이므로 cd는 2(=(1+1)/1, XV와 VZ)가 된다. 이것은 XZ 사이의 의미적 관련성에 주목하여 개념간 거리 cd값을 계산한다.



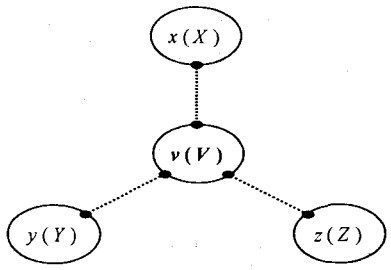
(그림 2) 개념어 V의 공기관계

3.2.2 공기관계의 수

주제를 대신하는 개념어 혹은 개념요소는 문서 중에 자주 출현한다. 또한 이들 개념어는 다른 단어들과 많은 수의 공기관계를 갖는다. 따라서 다른 단어와 많은 공기관계를 갖는 개념어가 문서의 주제를 대표하는데 가장 중요하다. (그림 2)는 문서에서 출현하는 공기관계를 나타내고, 관계가 있는 개념어를 점선으로 연결하였다. 어떤 문서에서 i번째의 복합어 w의 공기관계의 수를 $N(w_i)$이라 하면, (그림 3)과 같이 개념어 V가 갖는 공기관계 수는 $N(V)$는 3(=1+1+1)이 된다. 또한 X, Y, Z와 공기관계가 있는 개념어는 V뿐이므로 $N(X)$, $N(Y)$, $N(Z)$는 모두 1이 된다. 따라서 N(V) > N(X), N(Y) 혹은 N(Z)이므로 V가 X, Y, Z 보다 중요도가 가장 높다.

3.2.3 중요도 계산

개념어간 거리(cd)와 개념어 w_i의 공기관계수 $N(w_i)$를 고려한 키워드후보의 중요도를 계산하는 방법은 이하 식(2)로 계산할 수 있다. cd가 작을수록,

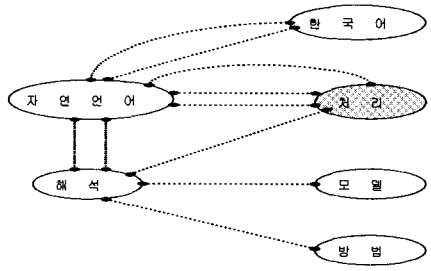


(그림 3) 개념어 V의 공기관계 수

공기관계 수가 많을수록, 개념요소에 대한 동의어(S) 및 유의어(R)의 빈도가 높을수록, 중요도(I; Importance)는 높아진다.

$$I = \left[\frac{1}{n \times cd}\right] \times \sum_{i=1}^{n}\left[\left(\frac{\{(S(W_i)\times\alpha) + (R(W_i)\times\beta)\}}{\{(S_T\times\alpha) + (R_T\times\beta)\}}\right) \times N(W_i)\right] \quad \cdots \text{식(2)}$$

여기에서, n은 키워드후보를 구성하는 개념어의 수, S(wi)는 wi에 대한 동의어의 빈도, R(wi)는 wi에 대한 유의어의 빈도, St는 동의어의 전체빈도, Rt는 유의어의 전체빈도, α와 β는 동의어와 유의어에 대한 가중치(단, α > β)를 각각 나타낸다.



(그림 4) <표 3>에서의 공기관계 수

<표 2>와 같은 예제 문서에서 언더라인으로 표시한 단어는 우리가 키워드로 관심을 갖는 단어이며, 문서에서 출현한 개념어 들이다. 그 요소분포를 테이블 형식으로 <표 3>에 표시하였다. (그림 4)는 이들 개념요소와 개념어의 공기관계를 나타낸다. 생성규칙은 추출된 개념어에 대한 키워드 후보와 동의어 집합을 사용하여 중요도를 계산한다. 예를 들면,

a) PR(자연언어처리)
= Concept(자연언어) + Concept(처리)

위에서 예(a)의 "자연언어처리"의 경우는 <표 3>에 나타난 바와 같이 n은 2('어', '자연언어'), S(자연언어)는 6, S(처리)는 5, S_T는 17(=2+6+5+2+1+1)이 된다. N(자연언어)는 5, N(처리)는 4, cd는 1인 최단거리이다. 따라서, 중요도를 계산하여 보면 다음과 같다(단, 동의어와 유의어의 가중치는 각각 α=1, β=0.5로 한다. 또한, 문서 내에 이 개념어에 대한 유의어는 문서에서 출현하지 않았다고 가정)

$$I = \left[\frac{1}{(n \times cd)}\right] \times \left[\left\{\frac{(S(\text{자연언어}) \times \alpha) + (R(\text{자연언어}) \times \beta)}{(S_T \times \alpha) + (R_T \times \beta)} \times N(\text{자연언어})\right\} + \left\{\frac{(S(\text{처리}) \times \alpha) + (R(\text{처리}) \times \beta)}{(S_T \times \alpha) + (R_T \times \beta)} \times N(\text{처리})\right\}\right]$$

$$= \left[\frac{1}{(2 \times 1)}\right]\left[\left\{\frac{(6 \times 1) + (0 \times 0.5)}{(17 \times 1) + (0 \times 0.5)} \times 5\right\} + \left\{\frac{(5 \times 1) + (0 \times 0.5)}{(17 \times 1) + (0 \times 0.5)} \times 4\right\}\right]$$

$$= \frac{1}{2} \times \frac{50}{17} \fallingdotseq 1.47$$

으로 계산된다. 나머지 키워드후보의 중요도를 모두 계산하면

(b)의 "자연언어해석"의 경우는 ≒ 1.35

(c)의 "한국어해석"의 경우는 ≒ 0.70 가 되어 중요도 부여결과 가장 많은 공기관계 수를 갖는 키워드 후보는 "자연언어처리"이며, 이 후보어가 중요도가 가장 높다. 이것은 <표 2>의 하단에 저자가 정의한 키워드(<KYWD>로 표시) "자연언어처리"가 다른 키워드후보에 비해 문서를 대표하는 가장 적당한 키워드임을 알 수 있다.

다음은 유의어 집합을 사용한 경우의 예를 <표 4>의 예제 문서 ②와 같이 유의어에 의한 키워드의 생성 예를 나타내었다. 본문은 "음환경이해"란 주제어를 갖는다. 동의어 "음향"과 유의어 "음성"의 출현으로 개념어 "음"이 생성하고, 동의어 "환경"의 출현으로 개념어의 "환경"이 생성되고, 마지막으로 "이해"의 경우, 동의어가 출현하지 않지만, 유의어의 개념요소 "인식"이 세 번 출현하여 "이해"가 생성된다. 결론적으로, "음+환경+이해"의 규칙에 의해 최종적으로 "음환경이해"(저자가 제시한 키워드<KYWD>)와 다름)의 복합 키워드가 생성된다.

<표 4> 예제 문서 ②

문번호	문장 예
S_1	음성인식 시스템을 평가하기 위해 먼저 음향스트림을 분할한다.
S_2	본 논문에서는 음향스트림 분할을 일반환경에서 음성인식 시스템의 전처리로 사용할 때의 문제점을 논의하고, 예비실험을 준비한다.
S_3	... 음향스트림 분할 결과 입력음 스펙트럼에 변형을 가한다.
<KYWD>>음성인식//음환경인식	

5. 결론

본 시스템은 가장 중요도가 큰 값을 갖는 키워드를 인간에게 알려주어, 사용자가 그 문서를 읽을 것인가를 빠르게 판단하도록 제시한다. 인간이 사용하는 몇 개의 주요단어에 의해 문서의 분야를 연상하도록 도움을 주는 분야연상어[3]나 주제어가 되는 키워드를 추출하는 점에 주목하여 개념기반 복합키워드 추출법을 새롭게 제안하였다. 추출되는 키워드의 정밀도를 향상하기 위해 개념에 기반한 생성 규칙을 정의하고, 그 개념어들 간의 출현빈도와 공기관계, 개념사이의 거리를 이용하여 중요도를 계산하는 방법을 제안하였다. 본 방법의 큰 장점은 저자가 정의한 키워드 뿐 아니라, 문서 중에 출현하지 않는 키워드도 추출할 수 있는 점이다.

참고문헌

[1] 이상곤 · 이태헌, "개념기반 복합 키워드 추출방법," 컴퓨터교육학회 논문지, 제 6권, 제 2호, pp. 23-31, 2003.

[2] 이태헌 · 박기홍, "개념 규칙을 이용한 키워드 도출방법," 한국정보과학회 학술발표 논문집(II), 제 29권, 제 2호, pp. 685-687, 2002.

[3] 이상곤, "분야연상어를 이용한 화제의 계속성과 전환성을 추적하는 단락분할 방법", 정보처리학회 논문지 B, 제 10권, 제 1호, pp. 57-66, 2003.

[4] Nagata, M. et al. (1988), "A Newspaper Keyword Generation Method Based on Key-concept Extraction." 제 37회 정보처리 전국대회 논문집, pp. 1030-1031. (in Japanese)