개념기반 복합키워드 추출방법

이상곤† · 이태헌† †

요 약

인간은 문서를 읽고 그 내용을 머릿속에서 개념적으로 정리하여 적은 수의 복합단어를 이용하여 문서를 대표하는 적당한 키워드로 정리한다. 본 논문은 이러한 점에 착안하여 문서를 대표하는 키워드를 추출하는 방법을 제안한다. 학술논문을 실험 예로 사용하여 저자가 자신의 문서에 부여한 키워드가 문서의 본문 중에 출현하지 않는 경우에도 동작하도록 출현단어의 개념정보를 기초로 복합어 생성규칙을 구축한다. 문서의미와 상관없는 키워드의 추출을 억제하기 위해 중요도 결정법을 새로 제안한다. 추출된 키워드의 타당성 여부는 자연언어와 음성언어에 관한 논문의 제목과 요약을 수집하여 실험하였다. 또한 저자가 부여한 키워드와 본 시스템이 출력한 키워드를 비교 한 결과, 상위 한 개의 정확율이 96%가 되어 제안방법의 유용성을 확인하였다.

Concept-based Compound Keyword Extraction

Samuel Sangkon Lee† · Taehun Lee† †

ABSTRACT

In general, people use a key word or a phrase as the name of field or subject word in document. This paper has focused on keyword extraction. First of all, we investigate that an author suggests keywords that are not occurred as contents words in literature, and present generation rules to combine compound keywords based on concept of lexical information. Moreover, we present a new importance measurement to avoid useless keywords that are not related to documents' contents. To verify the validity of extraction result, we collect titles and abstracts from research papers about natural language and/or voice processing studies, and obtain the 96% precision in a top rank of extraction result.

1. 서 론

일반적으로 키워드 추출방법은 원문(原文)에서 출현하는 단어 중에서 그 중요성을 평가하여 문서의 내용을 대표하는 키워드로 추출한다. 이러한 키워드 추출에 관한 연구는 정보검색 분야의 색인어 결정[1-8, 11, 13-14], 정보추출[5, 10-11, 13] 등의 분야에 널리 적용할 수 있는 기술이다. 문서에서 키워드 자동추출은 단어의 출현빈도나 위치 등의 표층정보를 이용한 추출[2, 5, 7, 13, 14]과 단어의 구문구조나 의미분류 등의 자연언어 처리기법을 적극적으로 이용하는 방법[3, 4, 6, 13] 등이 제안되어 있다.

키워드 추출에 관한 종래의 연구는 문서내용을

† 정 회 원: 전주대학교 정보기술컴퓨터공학부 전임강사
†† 정 회 원: 소프트웨어 전용원 IT교수요원
논문접수: 2003년 1월 13일, 심사완료: 2003년 4월 11일

정확하게 표현하는 단어는 반드시 문서 중에 출현한다[8]는 가정 하에 원문 내의 단어를 키워드로 추출하고, 중요단어에 적절한 중요도를 부여하여 중요도가 높은 순으로 적당한 키워드를 추출하였다. 그러나, 원문에 적당한 키워드가 존재하지 않거나, 문서 내에 키워드로 사용될 수 있는 단어가 존재하지 않거나, 키워드의 구성단어가 문서의 여러 곳에 존재하는 경우나, 문서의 내용으로 추측([15]에서는 도출(導出), Derivation, 유도(誘導), 파생(派生) 등의 의미로 도출이란 용어를 사용하였다. 이것은 단지 문서에서 단어의 추출 의미가 아닌 문서의 생각, 판단, 결론 따위를 이끌어 낸다는 의미), 혹은 추론이 불가능한 추상적인 단어(혹은 주제어)로 키워드가 출현하는 경우는 효과적으로 대처할 수 없다[6, 7].

이 문제에 대하여 1988년 나가타 등은 키워드를 구성하는 기본개념(키개념)과 키워드간의 공기관계를 기술한 색인규칙을 미리 정의하고, 이를 이용하여 주제어를 생성하는 방법을 제안[6]하였다. 먼저, 문장 중에서 명사를 추출하고, 그 명사의 동의어나 유의어와 같이 개념이 비슷한 단어를 모아 적당한 키워드로 추출한다. 그러나, 키개념을 추출할 때 개념어들 간의 관련성을 고려하지 않기 때문에 문서의 뜻에 적합한 키워드가 추출되지 않을 수 있다.

따라서 본 연구는 문서에서 출현하는 몇 개의 주요어를 이용하여 개념간의 관련성을 고려한 새로운 주제어 추출을 시도한다. 또한, 문서에서 적당한 키워드가 출현하지 않는 경우에도 적당한 주제어 추출이 가능하도록 개념기반 복합 키워드 추출법을 제안한다. 특히, 원문의 저자가 부여한 키워드가 문서의 해당분야나 주제어로 출현하지 않는 경우에는 복합어 생성규칙을 이용한다. 이들 생성규칙의 개념적 관련성에 주목하여 중요도 계산법을 새롭게 제안한다.

제 2장에서는 저자 자신이 정의한 저자 키워드에 대해 간략하게 분석하고, 생성규칙의 구축을 준비한다. 이에 입각하여 3장에서는 개념에 기초한 생성규칙과 각 개념간 공기관계를 고려한 중요도 계산방법을 제안한다. 제 4장에서는 저자키워드의 타당성에 대해 실험하고, 추출된 복합어 키워드를 평가하고, 결론과 향후과제에 관하여 5장에서 논의한다.

2. 키워드의 패턴

인간이 직접 부여한 몇 개의 키워드를 이용하여 문서의 해당 분야나 주제에 해당하는 정확한 키워드를 추출하기 위해 인간이 작성한 키워드의 패턴을 조사하여 보면 <표 1>과 같은 특징이 있다. 이 표에 여섯 가지의 실례(문장 혹은 문자열)와 추출패턴의 예를 표시하였다. 적당한 키워드를 추출하기 위해서 먼저, 다음과 같이 세 가지의 키워드 패턴으로 분류할 수 있다. 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. 위의 여섯 가지의 추출패턴은 각 단어의 개념을 이용하면 정확한 추출이 가능하다. 규칙에 기초한 키워드 생성의 준비단계로 복합어 키워드를 각 구성요소로 분할하고, 패턴분석을 수행한다. 단, <표 1>의 (6)의 경우는 변환사전의 생성이 필요하기 때문에 본 논문에의 설명은 생략한다. 다음 장에서 개념을 이용하여 키워드를 추출하는 방법을 제안한다.

<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	

<표 2> 예제 문서 ①

문번호	문장 예
S_1	MSLR파서에 의한 미정의**어처리**의 한가지 검사 방법이다.
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	

3. 키워드의 추출

이 장에서는 개념에 기반 한 키워드 추출을 시도한다. 3.1절에서는 복합어의 생성규칙에 대하여 논하고, 3.2절에서는 키워드 후보의 중요성을 계산하기 위해 개념 간 거리, 공기관계의 수, 중요도 계산법 등에 대해 논의한다.

3.1 복합어 생성규칙

복합키워드는 구성단어가 그대로 문서에 존재하지 않고 출현하는 단어로 연상 혹은 추측하여야 하는 경우가 많다. 이 절에서는 나가타의 방법을 기초로 키워드의 구성형태소가 갖는 개념과 키워드와의 관계를 규칙으로 정의하여 추상적인 단어나 주제를 포괄하는 단어를 추출하도록 시도한다.

어떤 복합어 w가 키워드일 때 그 복합어를 구성하고 있는 구성형태소들이 w의 하위개념 단어들로 복합 구성되어 있다고 가정하면, 이들 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. 동의어는 첫 번째, 유의어는 두 번째의 {과 }내에 표시하였다. 복합어 w와 각 구성요소들을 w_1w_2 … w_n으로 분할하고, w의 개념을 얻는 함수를 Concept()이라 하면 아래와 같이 각 형태소에 대한 개념요소로 나타낼 수 있다.

Concept(의미)=[{의의, 가치, …}, {내용, …}]
Concept(처리)=[{처치, 처분, …}, {해결, 취급, …}]
Concept(음성)=[{보이스, …}, {음색, 성조, …}]
Concept(대화)=[{회화, 회의, …}, {대담, 좌담, …}]

또한 w_1w_2 … w_n의 생성규칙(PR; Production Rule)은 각 개념요소들의 합으로 $PR(w_1w_2 \dots w_n) = Concept(w_1) + Concept(w_2) + \dots + Concept(w_n)$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

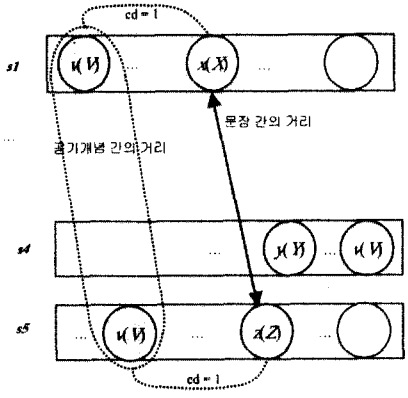
<표 3> <표 2>에서 추출된 개념어와 개념요소

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

생성규칙은 구성요소 "의미", "처리"와 양방향의 개념요소가 존재하면 복합어 "의미처리"를 추출하고, 동일한 방법으로 "음성대화처리"의 생성규칙은 "음성", "대화", "처리" 등의 세 가지 개념요소가 문서 내에 동시에 존재하면 복합어 "음성대화처리"를 아래와 같이 키워드로 추출한다.

PR(의미처리)=Concept(의미)+Concept(처리)
PR(음성대화처리)=Concept(음성)+Concept(대화)+Concept(처리)

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



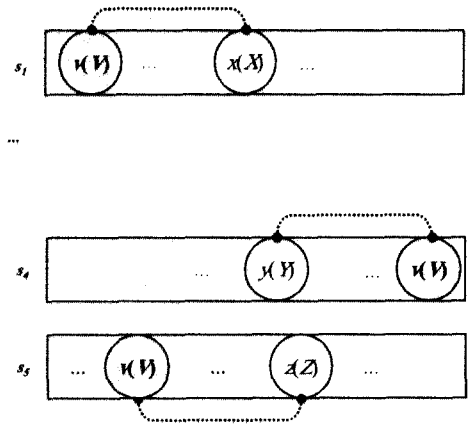
(그림 1) 문장 간 혹은 개념 간의 거리

3.2 키워드후보의 중요도

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

3.1. 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[2]에 주목하여 개념사이의 거리를 이용한다. PR(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 개념어들을 영문 대문자 V, X, Y, Z로 표현한다. X와 Z 사이의 개념간의 거리($\overline{XZ}$)는 단순히 문장 간의 거리 5가 된다. 어떤 문장 i에서 j까지의 문장 거리 sd는 s(j)-s(i)+1(단, j≥i)로 정의한다. 따라서, (그림 1)의 점선으로 표시한 바와 같이 $\overline{XZ}$의 sd는 5(=5-1+1)이다. 그러나 X와 Z에 공통으로 공기하는 개념어 V에 주목하면 V를 사이에 두고 X에서 Z에 이르므로 개념 간 거리 cd는 다음의 식으로 계산한다.



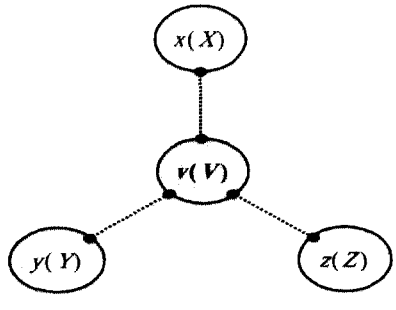
(그림 2) 개념어 V의 공기관계

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

여기서, cc는 공통개념어(common conceptual term)의 수(X), cd는 $\overline{XV}$가 1이고, $\overline{VZ}$가 1이므로 cd는 2($= \frac{(1+1)}{1}$, $\overline{XV}$와 $\overline{VZ}$)가 된다. 이것은 $\overline{XZ}$ 사이의 의미적 관련성에 주목하여 개념 간 거리 cd 값을 계산한다.

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 문서



(그림 3) 개념어 V의 공기관계 수

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

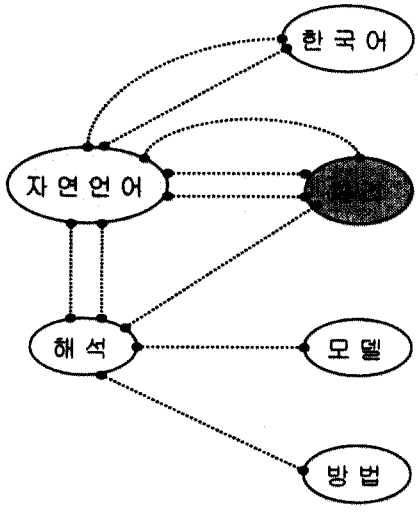
에서 출현하는 공기관계를 나타내고, 관계가 있는 개념어를 점선으로 연결하였다. 어떤 문서에서 i번째의 복합어 w의 공기관계의 수를 $N(w_i)$이라 하면, (그림 3)과 같이 개념어 V가 갖는 공기관계 수는 $N(V)$는 3(=1+1+1)이 된다. 또한 X, Y, Z와 공기관계가 있는 개념어는 V뿐이므로 $N(X)$, $N(Y)$, $N(Z)$는 모두 1이 된다. 따라서 N(V) > N(X), N(Y) 혹은 N(Z)이므로 V가 X, Y, Z 보다 중요도가 가장 높다.

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, 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, $S(w_i)$는 w_i에 대한 동의어의 빈도, $R(w_i)$는 w_i에 대한 유의어의 빈도, S_T는 동의어의 전체빈도, R_T는 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 개념어들이다. 그 요소분포를 테이블 형식으로 <표 3>에 표시하였다. (그림 4)는 이들 개념요소와 개념어의 공기관계를 나타낸다. 다음에 생성규칙은 추출된 개념어에 대한 키워드 후보와 동의어 집합을 사용하여 중요도를 계산한다. 예를 들면,

a) PR(자연언어처리)
= Concept(자연언어) + Concept(처리)
b) PR(자연언어해석)
= Concept(자연언어) + Concept(해석)
c) PR(한국어해석)
= Concept(한국어) + Concept(해석)

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).

$$I = \left[\frac{1}{(n\times cd)}\right] \times \left[\left\{\frac{(S(\text{자연언어})\times\alpha)+(R(\text{자연언어})\times\beta)}{(S_T\times\alpha)+(R_T\times\beta)}\times N(\text{자연언어})\right\} + \left\{\frac{(S(\text{처리})\times\alpha)+(R(\text{처리})\times\beta)}{(S_T\times\alpha)+(R_T\times\beta)}\times N(\text{처리})\right\}\right]$$
$$= \left[\frac{1}{(2\times1)}\right]\left[\left\{\frac{(6\times1)+(0\times0.5)}{(17\times1)+(0\times0.5)}\times5\right\}+\left\{\frac{(5\times1)+(0\times0.5)}{(17\times1)+(0\times0.5)}\times4\right\}\right]$$
$$= \frac{1}{2}\times\frac{50}{17} \fallingdotseq 1.47$$

으로 계산된다. 나머지 키워드후보의 중요도를 모두 계산하면

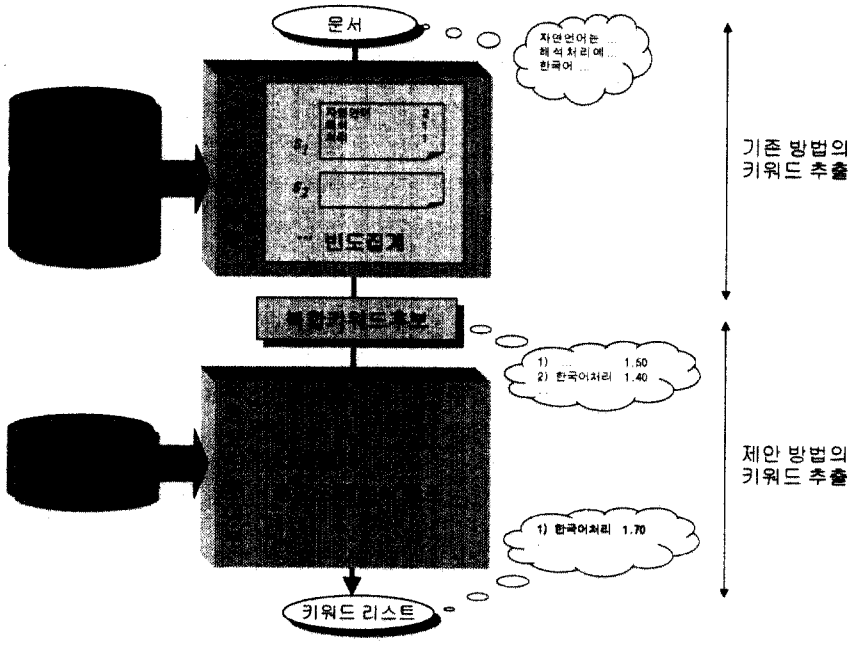
(b)의 "자연언어해석"의 경우는

$$I = \left[\frac{1}{(2\times1)}\right]\left[\left\{\frac{6}{17}\times5\right\}+\left\{\frac{2}{17}\times4\right\}\right] = \frac{23}{17} \fallingdotseq 1.35$$

(c)의 "한국어해석"의 경우는

$$I = \left[\frac{1}{(2\times1)}\right]\left[\left\{\frac{2}{17}\times2\right\}+\left\{\frac{5}{17}\times4\right\}\right] = \frac{12}{17} \fallingdotseq 0.70$$

가 되어 중요도 부여결과 가장 많은 공기관계 수를 갖는 키워드 후보는 "자연언어처리"이며, 이



(그림 5) 시스템 개요

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. 또한 위의 (b) 계산결과에서 중요도 1을 넘는 키워드 "자연언어해석"도 키워드후보로 적당함을 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.

4. 실험 및 평가

<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.
S_4	변형은 조파구조 추출, 두음부 전달함수, 또는 그룹핑 등이 있다.
S_5	분산형 단일 코드북형 HMM-LR을 위해 스펙트럼 변형을 조사하며, 음파구조 추출은 음성인식에 거의 영향을 주지 않는다. ...
<KYWD>>음성인식//음환경인식	

4.1. 시스템 개요

생성규칙에 기초한 키워드 생성 시스템의 개요를 (그림 5)에 표시하였다. 본 시스템은 키워드 생성부와 중요도 부여부 등 두 가지 핵심모듈로 구성되어 있다. 키워드 생성부는 사전에 존재하는 단어에서 명사만을 추출하여, 개념요소가 되는 동의어와 유의어를 파악하고, 생성규칙을 이용하여 복

합어로 된 키워드 후보를 생성한다. 중요도 부여부는 동의어 및 유의어의 가중치를 적용하고, 공기정보를 3.2.3절에서 언급한 중요도 계산식에 대입하여 값이 큰 순서로 키워드를 추출한다.

<표 5> 저자키워드(<KYWD>)와 비교한 결과

키워드의 종류	정답키워드 수
추출된 키워드	19
생성규칙에 의해 추출된 키워드	14
사전에 의해 생성된 키워드(경우 6)	5
추출되지 않은 키워드	35
전체 키워드 수	54

4.2. 저자 키워드의 평가

본 논문에서 사용한 실험 데이터는 KTSET과 전주대학교 학술정보관에서 제공하는 석·박사 학위논문 데이터베이스에서 자연언어와 음성언어 처리에 관련된 논문의 제목과 요약 50개(약 28.5KB)를 선정하였다. <표 5>에 시스템이 추출한 키워드와 추출되지 못한 키워드의 결과를 표시하였다. 저자가 정의한 키워드가 제목 및 요약 중에 출현하지 않았던 키워드를 정답키워드로 간주하여 추출한 결과, 저자가 정의한 키워드(저자키워드)의 총 수 201개(제목 혹은 요약 한 개에 평균 4개의 저자키워드가 존재) 중에서 정답키워드 수는 54개이었다. 전체 키워드의 약 70%가 복합키워드이고, 평균 두 개의 형태소로 구성되어 있다. 이것은 인간은 복합어를 키워드로 사용하여 문서의 내용을 대표한다는 것을 의미한다. 규칙에 의해 추출된 키워드의 수가 14개이고, 사전을 이용해 생성된 키워드(<표 1>의 경우 6)가 다섯 개가 되어 추출된 정답키워드의 수는 모두 19개이었다. 추출되지 않은 키워드는 35개이다. 이것은 문서에 키워드를 생성할 수 있는 단서가 전혀 존재하지 않는 논문이 17.5%, 개념요소(동의어, 유의어)가 사전에 존재하지 않는 것이 47.5%, 생성규칙 자체가 존재하지 않는 것이 35% 이었다.

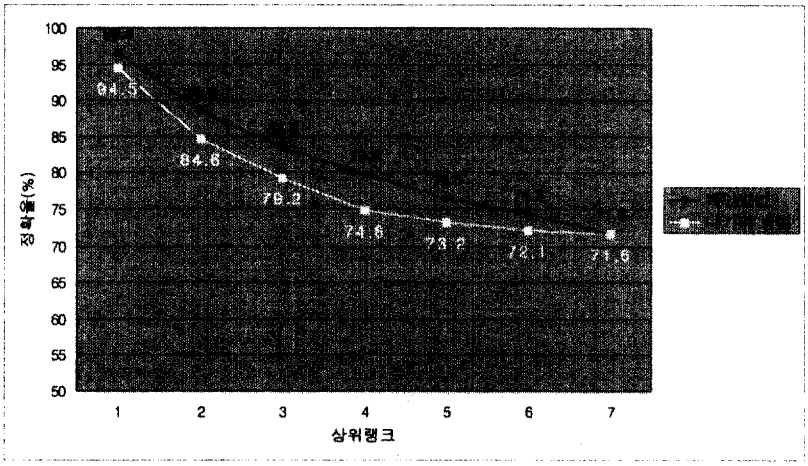
$$\text{재현율} = \frac{\text{추출결과에 포함된 정답키워드의 수}}{\text{정답키워드 수}} \quad \cdots (3)$$

$$\text{정확율} = \frac{\text{추출결과에 포함된 정답키워드의 수}}{\text{추출된 키워드 수}} \quad \cdots (4)$$

저자키워드에 대한 재현율과 정확율의 평가를 식(3)과 (4)로 계산하여 재현율 35%, 정확율 11%를 얻었다. 정확율이 대단히 낮은 이유는 요약 당 저자가 기술한 키워드 수가 네 개인 것에 반해, 정답키워드 수는 오직 한 개인 것이 원인으로 생각된다. 따라서 다음과 같이 키워드의 타당성 검사가 필요하다.

4.3. 키워드의 타당성

본 논문에서는 추출된 키워드후보의 타당성을 평가하기 위해 규칙에 의해 추출된 키워드에 대해



(그림 6) 제안방법과 나가타 방법의 실험결과 비교

다섯 명의 실험자가 <표 6>과 같이 네 가지 단계로 평가하였다. 실험자 중 네 명 이상이 A 혹은 B로 평가한 키워드후보는 키워드로써 타당하다고

<표 6> 키워드후보의 타당성 평가

평가단계	타당성 여부
A	키워드로 적절하다
B	키워드로 약간 적절하다
C	키워드로 약간 부적절하다
D	키워드로 부적절하다

판단하고, 이 후보를 정답키워드로 간주한다. 추출된 키워드 후보수에 대한 정답키워드 수의 비율을 정밀도라 정의하고, 식 (5)로 계산하였다. 또한 결과를 나가타의 실험결과와 비교하였다.

$$\text{정밀도}(\%) = \frac{\text{정답키워드수}}{\text{추출된키워드후보수}} \times 100 \quad \cdots\cdots (5)$$

먼저, (그림 5)에서와 같이 나가타의 실험(기존 방법의 키워드 추출 부분이라 표시)을 간단히 소개하면, 미리 키워드를 구성하는 기본개념(키개념)과 키워드의 관계를 기술한 색인사전을 정의하고, 문서 중에 존재하는 개념을 이용하여, 주제어를 생성하여 다음의 식 (6)과 같이 중요도를 계산[6]하였으나, 본 논문은 3장에서 소개한 바와 같이 키워드후보를 구성하는 개념어의 수, 공기관계를 이용한 개념간 거리, 빈도 등을 이용하여 계산하였다.

$$\text{중요도} = \frac{1}{\text{키워드후보를구성하는개념어의수}} \times \frac{\text{빈도}}{\text{문장간의거리}} \quad \cdots\cdots (6)$$

(그림 6)에 표시한 바와 같이 출력된 키워드 중 상위에 랭크된 7개까지의 키워드후보를 나가타의 출력과 비교한 결과 위에서 언급한 정밀도를 최고 96%까지 향상할 수 있었다. 실험에 사용한 생성규칙은 분야어·주제어를 기초로 30개를 선정하여 실험하였고, 개념요소 사전은 가도가와가 분류한 분류사전[12]을 한국어로 번역하여 이용하였다. 단, 분류사전에 기재되어 있지 않은 자연언어와 음성언어에 관련된 전문용어에 대해서는 분류표에 있는 용어 중에서 가장 의미가 가까운 단어를 인간이 수작업으로 선정하고, 동의어와 유의어를 그 단어의 개념요소로 사용하였다. 동의어와 유의어의 가중치는 여러 번의 예비실험을 통하여 얻어진 최적의 수치인 α는 1, β는 0.5로 실험하였다.

5. 결 론

본 시스템은 가장 중요도가 큰 값을 갖는 키워드를 인간에게 알려주어, 사용자가 그 문서를 읽을 것인가 혹은 읽지 않을 것인가를 빠르게 판단하도록 적당한 키워드를 제시한다. 인간이 사용하는 몇 개의 주요단어에 의해 문서의 분야를 연상하도록 도움을 주는 분야연상어[16]나 주제어가 되는 키워드를 추출하는 점에 주목하여 개념기반 복합키워드 추출법을 새롭게 제안하였다. 추출되는 키워드의 정밀도를 향상하기 위해 개념에 기반 한 생성규칙을 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 있다.

향후과제로서 본 논문의 방법은 전문용어가 개념요소 사전에 등록되어 있지 않는 키워드는 추출할 수 없다. 따라서 전문용어에 대한 개념요소 사전을 구축하면 보다 높은 정밀도가 향상될 것으로 기대되며, 이를 개발중인 정보검색시스템에 적용할 예정이다.

참 고 문 헌

[1] Kimoto, H. (1991), "Automatic Indexing and Evaluation of Keyword for Japanese Newspapers," *Transactions of IEICE of Japan*, Vol. J74-D-I, No. 8, pp. 556-566. (in Japanese)

[2] Kimoto, H., (1992), "Automatic Indexing of an Integrated Large Scale Text Database and Its Evaluation," *The SIG Notes of IPSJ*, DBS-90-9, pp. 73-81. (in Japanese)

[3] Suzuki, H., Masuyama, S., and Naito, S. (1993), "Examination of Keyword Extraction Using Thesaurus in Japanese Text," *The SIG Notes of IPSJ*, NL-133-33, 98-10, pp. 73-80. (in Japanese)

[4] Uchiyama, K., et al. (1991), "Development of an Automatic Keyword-extracting System on the Basis of Content Analysis and an Application

System," *The SIG Notes of IPSJ*, DBS-84-19, pp. 151-161. (in Japanese)

[5] Okumura, M. et al. (1999), "Automated Text Summarization: a Survey," *Journal of Natural Language Processing of Japan*, Vol. 6, No. 6, pp. 1-26. (in Japanese)

[6] 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)

[7] Hara, M., Nakajima, H. and Kitani, T. (1997), "Keyword Extraction Using a Text Format and Word Importance in a Specific Field, *Journal of IPS of Japan*," Vol. 38, No. 2, pp. 299-309. (in Japanese)

[8] Morohashi, M., (1984), "Automatic Indexing Survey," *Magazine of IPS of Japan*, Vol. 25, No. 9, pp. 918-925. (in Japanese)

[9] NACSIS (1999), "NACSIS Test Collection for IR Systems," *National Institute of Informatics*. (in Japanese)

[10] Katoh, N., Uratani, N. (1999), "A New Approach to Acquiring Linguistic Knowledge for Locally Summarizing Japanese News Sentences," *Journal of Natural Language Processing of Japan*, Vol. 6, No. 7, pp. 73-92. (in Japanese)

[11] Ito, S., et al. (1993), "Parametric Keyword Extraction Algorithm and Adaptation Method," *Technical Report of IEICE of Japan*, NLC93-53, pp. 41-46. (in Japanese)

[12] Ono, S., et al., (1981), Kadokawa Ruigo Shin Jiten, Kadokawa Syoten, Inc. (in Japanese)

[13] Tokunaga, T., et al. (1999), "Information Retrieval and Natural Language Processing," University Of Tokyo Press. (in Japanese)

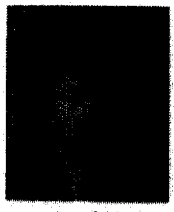
[14] Ogawa, Y., et al. (1993), "Compound Keyword Assignment Method for Japanese Texts," *The SIG Notes of IPSJ*, NL-133-33, 97-15, pp. 103-110. (in Japanese)

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

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

이 상 곤

1996년 전북대학교 컴퓨터과학과(학사)
1998년 전북대학교 전산통계학과(이학석사)
2001년 일본 도쿠시마대학교 지능정보공학과(공학박사)
2001년~2002년 원광대학교 음성정보 기술산업 지원센터 연구원
2002년~현재 전주대학교 정보기술 컴퓨터공학부 전임강사
관심분야 : 한국어 정보처리, 한글공학, 정보검색, 문서분류, 컴퓨터교육
E-Mail : samuel@jeonju.ac.kr



이 태 헌

1993년 군산대학교 전자계산학과 졸업(학사)
1998년 군산대학교 대학원 컴퓨터과학과(이학석사)
2001년 일본 토쿠시마대학 대학원 지능정보과학과(공학박사)
2001년~2002년 군산대학교 시간강사
2002년~현재 소프트웨어 진흥원 IT교수요원
관심분야 : 지식사전검색, 지적문서검색, 패턴매칭, 정보추출 및 응용
E-Mail : thlee@kunsan.ac.kr