의존관계에 기초한 일본어 키워드 추출방법

이 태 헌[†] · 정 규 철[†] · 박 기 홍[††]

요 약

본 논문에서 일본어 키워드 추출을 목적으로 요약문서 중에 떨어진 문자열을 합성하고, 그 문장에 나타난 규칙을 가지고 단어 정보(표기, 품사)와 구문 정보를 이용하여 복합명사로 된 키워드 추출 방법을 제안한다. 저자 지신이 부여한 학술 논문의 키워드 중에서 문서 중에 그대로 존재하지 않는 키워드의 특징을 분석한 결과로 의존 관계에 의한 복합명사 생성 규칙을 구축한다. 또 문장의 내용과 다른 키워드의 추출을 억제하기 위해 생성규칙에 대한 제약과 중복 단어를 고려한 중요도 결정법을 제안한다. 자연 · 음성언어처리에 관한 일본어 논문 65파일의 타이틀과 초록부분을 이용하여 추출된 키워드의 타당성에 대한 실험을 한 결과 추출 정밀도는 중요도의 상위 1개를 출력한 경우 63%가 되어 추출 방법의 유효성을 확인 할 수가 있었다.

The Method of Deriving Japanese Keyword Using Dependence

Taehun Lee[†] · Kyu-Cheol Jung[†] · Ki-Hong Park[††]

ABSTRACT

This thesis composes separated words in text for extracting keywords from Japanese, proposes extracting indexing keywords which consist of a compound noun using words and sentences information with the rules in the sentences. It constructs generative rules of compound nouns to be based in dependence as a result of analysing character of keywords in the text not the same way as before. To hold other extracting keywords and the content of sentence, and suggest how to decide importance concerned some restriction and repetition of words about generative rules. To verify the validity of keywords extracting, we have used titles and abstracts from Japanese thesis 65 files about natural language and/or voice processing, and obtain 63% in outputting one in the top rank.

키워드 : 키워드 추출(Extracting Keyword), 복합명사(Compound Noun), 의존관계(Dependence), 생성규칙(Generative Rule)

1. 서 론

키워드 자동 추출에 관해서는 지금까지 정보 검색의 자동 색인 구축을 목적으로 하여 단어의 출현 빈도나 출현 위치 등의 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[9]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].

이와 같은 문제에 대하여, [8]은 문서 중에 출현하는 명사 단어들에 대하여 인접하는 모든 조합을 복합명사 키워드 후보로 이용하는 방법을 제안했다. 그러나 인간은 연속적인 단어를 사용하여 키워드를 생성할 뿐만 아니라, 분리된 단어에서도 중요한 키워드를 생성한다.

본 논문에서는 문서를 읽기 위해서 판단이 가능한 재료가 되는 키워드(요약 키워드)의 추출을 목적으로 하고, 특히 복합명사 생성 규칙을 이용하여 키워드의 구성 단어로 분리되어 존재하는 키워드를 추출하는 방법을 제안한다.

복합명사 생성 규칙을 구축하기 전에 학술 논문에서 저자 자신이 부여한 키워드(저자 키워드) 중 문중에 그대로 출현하지 않는 것을 분석 대상으로 하고 분석 결과를 기초로 의존 관계를 고려한 규칙(의존 규칙)을 구축한다. 또 키워드의 추출 정밀도를 향상하게 하기 위해, 의존 규칙에 대해서 추출 할 때 제약과 키워드에 대한 중요도 계산법을 제안한다.

다음 2장에서는 저자 키워드의 패턴 분석과 그 결과에 관하여 설명한다. 그리고 3장에서는 의존 규칙과 추출에 대한 제약 및 중복 단어를 고려한 중요도 계산방법, 4장에서는 시스템이 추출한 키워드의 타당성 평가에 의해서 제안 방법의 유효성을 확인하고, 5장에서는 향후 과제와 결론을 맺는다.

† 준 회 원 : 군산대학교 컴퓨터정보과학과 IT 교수요원
†† 종신회원 : 군산대학교 컴퓨터정보과학과 교수
논문접수 : 2001년 9월 14일, 심사완료 : 2003년 1월 10일

2. 키워드 구성 단어의 패턴 분석

본 장에서 문서를 읽기 위한 판단 재료가 되는 요약 키워드를 생성하기 위한 준비로서 저자가 학술 논문에서 부여한 키워드의 패턴에 따라 분류하고 그 특징을 분석한다.

2.1 저자 키워드의 분석

저자가 부여한 키워드는 요약 키워드로써 문서를 읽을 것인지 안 읽을 것인지를 판단하기 위한 지표가 된다. 그래서 본 절에서는 일본 학술 정보 센터의 정보 검색 시스템 평가용 텍스트 콜렉션[1][11]에서 자연·음성언어 처리에 관한 파일(50 파일)의 타이틀과 초록을 이용하여 저자 키워드의 특징을 추출하기 위한 패턴 분석을 하는데 특히, 인간이 부여하는 키워드에는 문서 중에 그대로 출현하지 않는 것[9]을 주목하여, 저자 키워드의 내에 그대로 출현하지 않는 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 보여준다. 「→」은 왼쪽의 문자열에서 오른쪽의 키워드가 추출되는 것을 의미한다.

2.2 키워드 추출의 패턴

(A) 문중에 전부 존재하는 경우
(A-1) 의존 관계에 의한 단어의 추출
「音声を**認識**する → 音声認識」(**음성**을 **인식**한다 → 음성 인식)
「**情報**の**検索**は → 情報検索」(**정보**의 **검색**은 → 정보 검색)
(A-2) 병렬 관계의 단어에서 추출
병렬 관계에 있는 단어에서 복수의 키워드를 추출
「音声の**認識**および合成 → 音声認識」(**음성**의 **인식** 및 합성 → 음성 인식)
「音声の認識および**合成** → 音声合成」(**음성**의 인식 및 **합성** → 음성 합성)
(A-3) 지시대명사의 대응 관계를 고려한 추출
「言語を話し、それを**習得**する → 言語習得」(**언어**로 이야기하고, 그것을 **습득**한다 → 언어 습득)
(A-4) 복수의 문에 분류하는 단어에서의 추출
「音声を計算機で処理する。そのためには正しく**認識**することが必要である。→ 音声認識」
(**음성**을 계산기로 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
「**単語切り出し** → **単語抽出**」(단어 **떼어내다** → 단어 **추출**)
「**学習方法** → **学習法**」(**학습 방법** → **학습법**)
(B-2) 복수의 문에 존재하는 단어의 공기정보에 의한 추출
「말」과 「인식」의 공기정보에서 「음성 인식」을 추출.
「人間の**言葉**を機械で処理する。そのためには、正しく**認識**させる必要がある。→ 音声認識」
(인간의 **말**을 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
「推論知識 → 人工知能」(추론 지식 → 인공 지능)
「品詞が付与できる → 形態素解析」(품사를 부여할 수 있는 → 형태소 해석)
(C-2) 영어 또는 영어의 단축 단어에서 일본어로 변환(역도 포함)되어 추출
「back-off → バックオフ」(back-off → 백 오프)
「文脈自由文法 → CFG」(문맥 자유 문법 → CFG)

2.3 분석 결과에 대한 고찰

패턴 (A-1), (A-2)는 키워드의 구성 단어가 분리한 예지만, miyazaki[12]가 제안한 복합어의 의존 규칙을 개선하여 추출이 가능하다. (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), (A-2)를 대상으로 한다 그것을 의존 관계에 기초하는 규칙이라고 부르고 3장에서 설명한다. 또 생성 규칙에 의해 추출된 복합어를 키워드 후보라고 부른다.

3. 의존 관계에 기초하는 키워드 추출 방법

인간이 문서 중에 떨어진 문자열을 합성하고, 보다 더 문장의 뜻에 따른 키워드를 추출하는 점에 주목하여 단어 정보(표기, 품사)와 구문 정보를 이용한 규칙 베이스의 복합어 키워드 추출 방법을 제안한다.

3.1 의존 관계에 기초하는 복합명사 생성 규칙

[12]는 의존 규칙을 구축하여 고정밀도의 복합명사 자동 분할을 실현했다. 본 절에서는 의존 규칙을 개선하여 요약 키워

1) 일본어 텍스트 콜렉션의 파일에는 본문 이외의 제목, 요약, 저자 키워드 등이 수록되어 있다.

드를 추출하기 위한 복합명사 생성 규칙[2]을 제안한다. 아래에 규칙의 예를 보여준다.

[규칙 1] x(보통 명사$^+$) を(을, 를) y(서술형 명사) → xy
예 : 音声を認識する → 音声認識(음성을 인식하다 → 음성 인식)

[규칙 2] x(보통 명사$^+$) の (의) y(보통 명사$^+$) → xy
예 : 情報の検索は → 情報検索(정보의 검색은 → 정보 검색)

[규칙 3] x(보통 명사$^+$) の(의) y(보통 명사$^+$) および(및) z(보통 명사$^+$) → xy, xz
예 : 音声の認識および合成は→音声認識, 音声合成
(음성의 인식 및 합성은 → 음성 인식, 음성 합성)

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 xy를 생성하는 것을 표현한다. 또 동일 장소에 대한 규칙의 적용은 1회만 한다.

3.2 의존 규칙의 적용에 대한 제약

문장의 뜻에 맞지 않는 키워드의 추출을 억제하기 위해 생성 규칙에 대한 제약을 이하에 정의한다.

3.2.1 구문에 대한 제약

생성 규칙은 구문에서 애매성이 있을 때 의미적으로 올바른 복합명사를 생성할 수 없는 경우가 있다. 예를 들면, 「人間と計算機を比較する」(인간과 계산기를 비교하다)는 「比較する」(비교하다)가 인간과 계산기를 비교하는 것이 아니라 규칙 1에 의해 계산기를 비교하기 때문에 의미적 부적절한 복합명사 「計算機比較」(계산기 비교)를 생성한다. 그래서 규칙 1에 대해서는 규칙을 적용하는 형태소의 직전에 격조사 「と」(와, 과)가 존재하는 경우 키워드를 추출 할 수 없다는 제약이 생겨나게 된다. 그밖에 수식 등 의존 관계가 애매성을 생기는 구문에 대해 규칙을 적용하여 규칙의 직전 직후에 존재하는 형태소의 품사에 주목하고 키워드의 추출을 제한한다. 이하에 제약의 예를 보여준다.

[규칙 1의 제약] 앞에 격조사 「と」(와, 과)가 있을 경우 추출 안 함.

[규칙 2의 제약] 앞에 격조사 「と」(와, 과), 「の」(의), 또는 뒤에 격조사 「と」(와, 과), 「の」(의)가 있을 경우 추출 안 함.

[규칙 3의 제약] 뒤에 격조사 「の」(의)가 있을 경우 추출 안 함.

3.2.2 서술형 명사에 의한 제약

[12]도 주목하고 있는 것처럼 인간은 복합명사를 이해할 때 복합명사의 구성 어간에 생략된 조사를 보충하면서 서술형 명사로 구문을 구성하여 전체의 의미를 파악하고 있다고 생각할 수 있다. 또 복합명사 중에 알 수 없는 단어가 존재하는 경우에서도 서술형 명사가 포함되어 있다면 알 수 없는 단어의 의미를 추측할 수 있기 때문에 복합명사를 대강 파악하는 것이 가능하다.

이상에 의해, 요약 키워드는 동사 특히 서술형 명사를 포함한 단어만 추출한다.

3.3 불용어 삭제

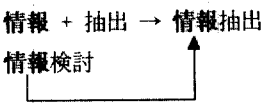
불용어란 일반적으로 키워드로써 성립되지 않는 단어를 의미한다. 제안방법으로는 이하 2개의 기준으로 불용어를 삭제한다.

[기준 1] 복합명사의 구성 단어가 안 되는 경우
예 : 「数個」(몇 개), 「有無」(유무), 「場合」(경우)

[기준 2] 논문 특유의 표현에서 생성되는 복합명사
예 : 「方法検討」(방법 검토), 「結果報告」(결과 보고), 「検討結果」(검토 결과)

3.4 키워드 후보에 대한 중요도 부여

3.4.1 중복단어를 고려한 중요도 계산



(그림 3) 중복 단어를 고려한 예

중요도 결정법으로써 키워드를 구성하는 단어는 정규화빈도[3]의 총화를 중요도라고 하는 방법이 생각할 수 있다. 예를 들면 (그림 1)에 나타나는 키워드 후보를 「정보 추출」로 하고 구성 단어 「情報」(정보), 「抽出」(추출)의 정규화 빈도를 각각 0.4, 0.3이라 하면 중요도는 0.7이 된다.

그러나 문중에 「情報」(정보)를 포함한 「情報検索」(정보 검색)이 존재할 때는 「情報抽出」(정보 추출)과 「情報検索」(정보 검색)의 사이에 관련성이 있다고 생각할 수 있어 「情報検索」(정보 검색)의 가중치를 「情報抽出」(정보 추출)에도 반영하고 있다.

전문가법으로는 키워드 후보와 문중에 있는 단어의 관련도를 형태소의 중복 비율로 하고 관련도에 따른 가중치를 키워드 후보에게 부가하는 방법을 제안한다. 문서 중에 출현하는 단어 w의 키워드 후보 k에 대한 관련도를 $R(w,k)$라고

2) 규칙의 기술에는 [13]이 제안한 다독성 규칙과 조합 엔진을 확장하여 이용했다.

3) 단어 w의 정규화 빈도 = w의 출현빈도/문서 중의 전체 보통명사 및 서술형 명사의 총 빈도

했을 때 k의 중요도 $S(k)$를 이하의 식에서 표현한다.

$$S(k) = \sum\{N(w) \times R(w, k)\}$$

$$R(w, k) = \left\{\frac{C(w, k)}{L(w)}\right\}$$

$N(w)$: w의 정규화 빈도
$C(w, k)$: k에 대한 w의 중복 형태소 수
$L(w)$: w의 형태 소수

N(情報) = 0.4, N(抽出) = 0.3, N(情報検索) = 0.2이라고 가정하면, 이 때 키워드 후보 「情報抽出」(정보 추출)에 대한 중복 형태소 수는

C(情報, 情報抽出) = 1 ;
C(抽出, 情報抽出) = 1 ;
C(情報検索, 情報抽出) = 1

이 된다. 또 「情報抽出」(정보 추출)에 대한 관련도 R은

R(情報, 情報抽出) = 1/1 = 1 ;
R(抽出, 情報抽出) = 1/1 = 1 ;
R(情報検索, 情報抽出) = 1/2 = 0.5

가 된다. 이상에 의해 중요도 S는 이하와 같이 된다.

$$\begin{aligned} S(\text{情報抽出}) &= \{N(\text{情報}) \times R(\text{情報, 情報抽出})\} \\ &\quad + \{N(\text{抽出}) \times R(\text{抽出, 情報抽出})\} \\ &\quad + \{N(\text{情報検索}) \times R(\text{情報検索, 情報抽出})\} \\ &= 0.4 \times 1 + 0.3 \times 1 + 0.2 \times 0.5 \\ &= 0.8 \end{aligned}$$

제안 방법에 의해 키워드 후보의 가중치뿐만 아니라, 키워드 후보와 중복되는 단어의 가중치가 관련도에 의해 고려할 수 있다.

3.4.2 적용 규칙에 의한 중요도의 보정

「による」(에 의하다), 「を用いた」(을(를) 이용한다) 등의 단서 표현의 전후에는 중요한 단어가 출현하는 것이 많다[1]. 또 「xをyする」(x를 y하다)라고 말하는 표현은 문의 주체를 표시하는 표현이고, 이 표현에서 생성한 복합명사 xy는 의미를 가진 것이 많다[10]. 이상의 분석에서 단서 표현을 포함하는 규칙이나, 주체를 표현하는 복합명사의 규칙은 키워드로써 중요도를 높게 한다.

구체적으로는 각 생성 규칙에 대하여 중요도의 보정 정수 α를 예비 실험에서 결정하고, 키워드 k의 보정 중요도 $S'(k)$ = k의 가중치 $S(k) \times \alpha$라고 한다

중복단어를 이용한 중요도 계산은 빈도가 적은 중요한 단어를 상위에 올라가지 않는 문제가 있지만 중요도를 보정하는 것으로 중요도를 높이는 것이 가능하다.

4. 실험과 평가

4.1 실험 데이터

실험 데이터로서는 일본어 정보 검색 시스템 평가용 텍스트 콜렉션[11]에서 자연·음성언어 처리에 관한 데이터 파일의 타이틀과 요약 부분(65개 파일[4], 총 용량 37.5KB)을 이용했다. 실험의 형태소분석의 인정에 관해서는 형태소 사전에 등록되어 있는 단어를 형태소라고 했다.

〈표 1〉 규칙에 대한 보정치

「を用いた」(을 이용한) x(보통 명사$^+$) 「の」(의) y(보통 명사$^+$)→xy	2
「による」에 의한 x(보통 명사$^+$) 「の」(의) y(보통 명사$^+$)→xy	2
x(보통 명사$^+$) 「を」(을) y(사행/어간)→xy	1.5
기타의 규칙	1

〈표 2〉 저자 키워드에 대한 실험 결과

	추출 키워드	정답 키워드
의존 관계에 근거한 규칙	386	16

실험에 이용하는 의존관계 규칙은 38개이고, 11개의 구문에 대한 제약을 정의했다. 불용어는, 3.4절에서 정의한 2개의 지표에 의해 수 작업으로 불용어 사전에 등록했다. 또 각 규칙에 대한 보정치는 2장의 분석에서 사용한 50개의 파일에서 결정했다. <표 1>에 그 값을 보여준다.

이상의 데이터를 이용하여 저자 키워드에 대한 평가와 요약 키워드로서의 타당성 평가로부터 제안방법의 유효성을 보여준다.

4.2 저자 키워드에 대한 평가

저자 키워드는 요약 키워드로서 기능을 갖기 위해 저자 키워드에 대한 재현율과 정확율에 의해 제안방법의 유효성을 평가한다. 여기에서 재현율은 키워드의 재현, 정확율은 노이즈를 보여주는 지표가 된다. 재현율 R은 정답 키워드 수 S에서 추출된 정답 키워드 수 E의 비율(E/S)을, 정확율 P는 추출된 키워드 수 T에서 추출된 정답 키워드 수 E의 비율(E/T)을 보여준다. 본 절에서는 특히 저자 키워드 내의 타이틀 및 요약 중에 출현하지 않았던 키워드를 정답 키워드로 정의했다. 65개의 파일 중에는 76개(총 키워드 갯수 263개)의 정답 키워드가 포함되고 있었다.

저자 키워드에 대한 실험 결과를 <표 2>에 보여준다.

R은 21%, P는 4%라는 낮은 값이 됐다. 특히 P가 낮은 이유는 1요약 정답에 평균 생성 키워드 수가 약 6개에 대하여 정답 키워드 수는 1요약에 약 1개 정도가 되기 때문이라고 생각된다. 또 추출되지 않았던 60개의 키워드의 내역으로는 문중에 키워드를 생성하는 단서가 전혀 존재하지 않았던

4) 실험 데이터 65개의 파일은 2장에서 저자 키워드를 분석한 50개의 파일과는 다르다.

것이 35.5%, 생성 규칙 자체가 존재하지 않는 것이 64.5%가 되었다. 이것들을 키워드 대상 범위에서 제외하고 시스템에 의해서 추출된 16개의 정답 키워드를 지표로 한 경우 P가 38%(상위 2개), R이 65%가 됐다.

4.3 요약 키워드의 타당성 평가

4.2절의 실험에 있어서 저자 키워드의 수는 1요약에 약 1개라고 적기 때문에 제안 방법의 유효성을 정확하게 판단할 수 없다. 그래서 본 절에서는 추출된 키워드(추출 키워드)가 요약 키워드로서 타당할 것인가의 판단을 하는 동시에 각 생성 규칙의 유효성도 평가한다.

먼저 5인의 피험자에 65개 파일의 타이틀과 요약을 읽게 하고, 이하의 4단계로 평가하게 했다.

A : 요약 키워드로서 적절하다
B : 키워드로서 위화감이 없다
C : 키워드로서 조금 위화감이 있다
D : 키워드로서 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.

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

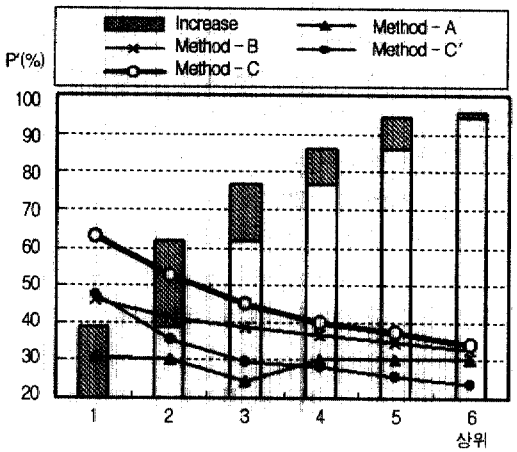
4.3.1 의존 관계에 기초하는 규칙의 평가

의존 규칙에 의해 추출된 키워드 386개 내의 약 28%가 요약 키워드가 되었다.

중요도에 의해 키워드 후보를 상위 6개까지 출력할 때의 P′을 (그림 3)에 보여준다. 또한 (그림 3)의 Method - A는 「중요도 = 키워드의 구성 단어에 대한 정규화 빈도의 총화」의 경우이고, Method - B는 Method - A + 중복단어 고려, Method - C는 Method - B + 보정치 고려이다. 또 Method - C′는 Method - C에 대하여 제약과 불용어를 적용하지 않는 경우이다.

추출된 전체 키워드에 대한 Method - C′의 P′는 24%이였지만 제약과 불용어를 적용(Method - C)하여 P′은 34%로 향상했다. 그 내역으로는 구문에 대한 제약에 의해 5.3%, 서술형 명사를 포함하지 않는 키워드 후보에 대한 제약으로 1.7%, 불용어로는 3.0%가 되었다. 이것은 구문에 의한 제약이 가장 효율적인 것을 의미한다. 한편 중요도 결정법도 정규화 빈도와 중복 단어, 보정치를 고려한 Method - C의 P′가 가장 높고, 상위 1개를 출력한 경우는 63%가 됐다.

또 그림중의 막대 그래프는 Method - C를 이용한 경우의 요약 키워드에 대한 재현율이고 사선의 부분은 그 증가율(Increase)을 보여준다. 상위 1(38.3%)과 2(23.4%)가 증가율이 높은 것을 알 수 있다 이것은 제안 방법이 요약 키워드로 상위에 출현하기 쉬운 것으로 밝혀진다.



(그림 3) 의존 관계에 기초하는 규칙에 의해 추출된 키워드에 대한 중요도 부여 결과

5. 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 기초하는 규칙을 이용한 규칙을 구축했다. 또 추출 정밀도를 향상하게 하기 위해 의존 관계에 기초하는 규칙에 대해서는 구문이나 서술형 명사에 의한 제약과 중요도 결정을 제안했다.

저자 키워드에 대한 실험을 고찰하면 앞서 말한 것처럼 추출하지 않았던 45개의 키워드의 내역은 키워드를 생성하는 단서가 전혀 존재하지 않는 것이 35.5%, 생성 규칙 자체가 존재하지 않는 것이 64.5%이였다. 특히 개념규칙을 중심으로 하는 동의어 사전이나 유의어 사전의 필요성을 알 수 있었다. 현재 번창하게 행해지고 있는 전문 용어의 추출에 관계하는 연구[14]를 이용하여 개념규칙을 구축하면 추출 정밀도가 향상할 수 있다고 생각한다. 생성 규칙이 존재하지 않는 것에 관해서는 「비유 문 생성 시스템 → 비유 생성」과 같이 복합명사의 단축에 필요한 것, 「한국어 처리와 텍스트 처리 → 한국어 텍스트 처리」와 같이 복합명사의 합성어가 키워드가 되는 경우 등 이였다. 특히 단축 단어에 관해서는 복합명사의 자동 분할 방법을 이용하고, 긴 단어를 삭제하는 방법을 고안 할 필요가 있다.

그리고 요약 키워드에 대해서 고찰하면 의존 규칙에 관해서는 상위 1의 P′가 63%, 2의 경우 52%, 3의 경우 45%라는 낮은 값이 됐다. 첫 번째 원인으로서는 저자 키워드와 동일하게 추출된 키워드가 저자 키워드보다 추출 키워드의 단어가 길기 때문에 요약 키워드로 평가되지 않는 것이 있다.

이 문제를 해소하기 위해서는 복합명사의 자동 분할 방법이나 단어 형성이론을 도입하여 복합명사의 단축을 실현할 필요가 있다. 두 번째의 원인은 비슷한 키워드가 복수 추출됐기 때문에 가장 적절한 키워드 이외는 노이즈가 된다. 또 그러한 키워드 사이에서는 중요도 또한 거의 동등하게 출현하는 경향이 있다. 인간은 첫 번째 나타났던 복합명사는 두 번째 이후에 단축해서 이용하는 경우가 많고, 그 경우도 비슷한 구문으로 이용되는 일이 있기 때문에 이런 문제가 생겼다고 생각할 수 있다. 이 해결책으로서는 추출된 키워드 사이의 관계(예를 들면, 단어의 중복 정도, 출현 구문 등)를 고려하여 후보 수를 감소할 수 있다고 생각한다.

참 고 문 헌

[1] okumura M. et al., "텍스트 자동 요약에 관한 연구 동향", 자연어처리학회, Vol.6, No.6, pp.1-26, 1999(in Japan).

[2] Hara, M. et al., "텍스트의 포맷과 단어의 범위 내 중요도를 이용한 키워드 추출", 정보처리학회논문지, Vol.38, No.2, pp. 299-309, 1997(in Japan).

[3] Ogawa, Y. et al., "복합어 키워드의 자동 추출법", 정보처리 자연언어연구회, 97-15, pp.103-110, 1993(in Japan).

[4] Kimot, H., "일본어 신문 기사에서의 키워드 자동 추출과 중요도 평가", 전자정보통신학회논문지, Vol.J74-D-I, No.8, pp. 556-566, 1991(in Japan).

[5] Tokunaga, T., "정보 검색과 언어 처리", 동경대학 출판회, 동경, 1999(in Japan).

[6] Suzuki, H. et al., "단어의 의미 분류의 출현 경향을 고려한 키워드 추출의 시험", 정보처리 자연언어연구회, 98-10, pp.73-80, 1993(in Japan).

[7] Uchiyama, K. et al., "중요 키워드 추출 방식과 그 활용 방법", 정보처리 데이터베이스 시스템연구회, 84-19, pp.151-161, 1991 (in Japan).

[8] Ito, S. et al., "이용 목적에 따른 최적 가능한 키워드 추출 방법", 전자 정보 통신학회, NLC93-53, pp.41-46, 1993(in Japan).

[9] Morohashi, M., "자동 색인 첨가 연구의 동향", 정보처리학회, Vol.25, No.9, pp.918-925, Sep., 1984(in Japan).

[10] Katoh, N. et al., "국소적 요약 지식의 자동 획득 방법", 자연언어처리논문지, Vol.6, No.7, pp.73-92, 1999(in Japan).

[11] NACSIS Test Collection for IR Systems, 학술정보센터-, 1999(in Japan).

[12] Miyazaki M. et al., "의존해석을 이용한 복합어의 자동 분할", 정보처리학회논문지, Vol.25, No.6, pp.970-979, 1984(in Japan).

[13] Andou K. et al., "일본어 정형 표현의 패턴 기술 규칙과 효율적인 조합 알고리즘", 전자정보통신학회논문지, Vol.J80-DII, No.7, pp.1860-1869, 1997(in Japan).

[14] Ohata U. et al., "연접이 다르게 되는 단어의 수에 의한 전문용어 추출", 정보처리 자연언어연구회, 136-16, pp.119-126, 2000(in Japan).

이 태 헌

e-mail : thlee@kunsan.ac.kr

1993년 군산대학교 전자계산학과(학사)

1998년 군산대학교 대학원 컴퓨터과학과 (이학석사)

2001년 일본 토쿠시마대학 대학원 지능정보과학과(공학박사)

2002년~현재 소프트웨어 진흥원 IT 교수요원

관심분야 : 지식 사전검색, 지적 문서관리 및 검색기술, 스트링 패턴 매칭, 정보검색 및 추출의 기초연구 및 응용, 자연언어처리의 기초해석(형태소, 구문해석 등)과 응용

정 규 철

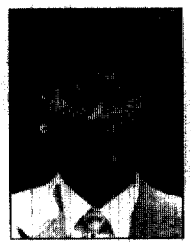
e-mail : kcjung@kunsan.ac.kr

1995년 군산대학교 컴퓨터과학과(학사)

2000년 군산대학교 대학원 컴퓨터과학과 (이학석사)

2001년~현재 군산대학교 대학원 컴퓨터정보과학과 박사과정

관심분야 : 자연언어처리, 정보검색



박 기 홍

e-mail : spacepark@kunsan.ac.kr

1982년 숭실대학교 공과대학 전자계산학과 (공학사)

1986년 숭실대학교 대학원 전자계산학과 (공학석사)

1995년 일본 토쿠시마대학 대학원 지능정보과학과(공학박사)

1997년~1998년 영국 영국 Middlesex Univ 객원교수

1999년~2000년 한국정보과학회 호남제주지부 부지부장

1987년~현재 군산대학교 컴퓨터 정보과학과 교수

2000년~현재 한국정보과학회 호남 제주지부 지부장

관심분야 : 자연언어처리, 정보검색