주제어자동추출을 위한 본문의 예비처리에 대한 리해

최 영 희

위대한 령도자 **김정일**동지께서는 다음과 같이 교시하시였다.

《콤퓨터가 처음에는 단순한 계산수단으로 출현하였지만 오늘은 방대한 정보량을 처리하는 만능의 정보처리기로 발전하여 사람들의 로동과 생활에서 필수적인 수단으로 되고있습니다.》(《**김정일**선집》 증보판 제20권 352페지)

오늘 세계적으로 과학기술은 매우 빠른 속도로 발전하고있으며 사회발전과 인간생활에서 과학기술의 역할은 날을 따라 더욱더 커가고있다.

이 글에서는 콤퓨터에 의한 본문의 주제어추출을 위한 예비처리방법을 보기로 한다.

자동색인을 하자면 대상문헌을 이루는 본문에서 주제어를 추출하여야 한다.

일반적으로 본문에 있는 모든 단어가 다 주제어로 되는것은 아니다.

정보학적으로 주제어는 정보적가치가 높은 명사 또는 명사적결합만이 된다. 그러므로 본문에 대한 예비처리를 한다.

일반적으로 조선어본문에는 우리 글자를 비롯하여 외국글자와 수자, 문장구별기호 등이 포함되여있다. 그러므로 조선어본문을 공백기호로 구획할 때 얻어지는 단어들가운데는 조선어형태단어로서의 체모를 갖춘것도 있고 갖추지 못한것도 있다.

이로부터 이 글에서는 조선어본문을 공백기호로 구획할 때 얻어지는 단어를 형태단어라고 하지 않고 형태단어후보라고 한다.

형태단어후보들가운데서 조선어형태단어로서의 체모를 갖추지 못한 단어들을 제거하는것을 주제어자동추출을 위한 본문의 예비처리라고 한다. 다시말하여 형태단어후보에서 주제어로 될수 있는 형태단어들을 확정하는것이 본문에 대한 예비처리이다.

콤퓨터처리에서 조선어본문을 글자들의 모임이라고 보고 글자들을 구분하여보면 다음과 같다.

첫째로, 음절을 단위로 하여 이루어진 조선어글자이다.

둘째로, 수자이다. 여기에는 0부터 9까지의 아라비아수자와 로마수자가 속한다.

셋째로, 기타 기호이다. 여기에는 문장부호(., ?, !, ※, △, ㅇ, -, …)들과 공백, 도형기호, 단위기호들이 속한다.

넷째로, 문자이다. 여기에는 조선어자모와 외국문자, 한자들이 속한다.

조선어본문을 공백기호에 의해 구획할 때 얻어지는 형태단어후보의 류형들을 갈라보면 다음과 같다.

첫째로, 조선어음절글자들만으로 이루어진 형태단어후보가 있다.

ㅇ 문법적오유의 검사와 교정

ㅇ 개념분류의 정확성을 보장하기 위한 방도

ㅇ 조선어형태단어해석

둘째로, 수자나 기타기호, 문자들이 조선어의 형태부와 어울려 이루어진 형태단어후보가 있다.

이러한 현상은 과학기술문헌에서 많이 볼수 있다.

ㅇ 어떤 문헌에는 425개의 금지어가 제시되여있다.

ㅇ KMP알고리듬은 BF보다는 약간 빠른 선형검색알고리듬이다.

ㅇ 식 3의 결과는 88이다.

ㅇ T=t_i가 주어질 때 c와 d는 독립이다.

ㅇ 파라메터 α와 β에 값을 입력한다.

ㅇ 알고리듬 1을 수행한 다음 알고리듬 5를 수행하여야 한다.

ㅇ 실례로 《Bank》와 《bank》는 의미가 전혀 다르다.

ㅇ 형태단어→어간+토

ㅇ 《ㄷ. ㄷ》강령

이 류형의 형태단어후보들에서 조선어음절글자가 아닌 수자와 문자, 기타 수기호가 차지하는 위치는 서로 다르다. 다시말하여 조선어음절글자가 아닌 수자, 문자, 기타 기호들은 형태단어표기에서 첫번째 위치에 놓인것도 있고 가운데에 놓인것도 있으며 마지막에 놓인것도 있다. 그러므로 형태단어후보를 이루는 글자렬에서 조선어음절글자가 아닌 수자, 문자, 기타 기호들이 차지하는 위치에 따라 3가지 형태로 갈라볼수 있다.

1형태에는 조선어음절글자가 아닌 수자, 문자, 기타 기호들이 첫번째 위치에 있는 형태단어후보들이 속한다.

ㅇ x선, B.C.2세기, B-52전략폭격기

ㅇ ㄱ에서, a와

2형태에는 조선어음절글자가 아닌 수자, 문자, 기타 기호들이 가운데위치에 있는 형태단어후보들이 속한다.

ㅇ 형태단어→어간+토

ㅇ 평양－혜산행급행렬차

ㅇ 제2준위 금지어

ㅇ 자료 ①, ②에서, 수 N_k는, 질문목록 ㄱ, ㄴ에서

3형태에는 조선어음절글자가 아닌 수자, 문자, 기타 기호들이 마지막위치에 있는 형태단어후보들이 속한다.

ㅇ 형태부?

ㅇ 진행된다.(27; 271~302)

ㅇ 진행된다.(ㄱ; 4~10)

셋째로, 일부 수자, 문자, 기타 기호만으로 된 형태단어후보가 있다.

ㅇ ㄱ

ㅇ △

ㅇ a)

ㅇ Ⅳ, (3)

ㅇ S→Φ

ㅇ $S = W_1 W_2 W_3 \ldots W_n$

우의 3가지 류형에서 조선어형태단어를 규정하는 기준에 따르면 첫번째 류형의 형태단어후보는 명백히 조선어형태단어로 되며 두번째 류형은 조선어형태단어의 특수한 경우,

세번째 류형은 조선어형태단어가 아니다.

조선어음절글자를 X, 조선어음절글자가 아닌 다른 문자를 Y라고 하면 우에서 본 3가지 류형의 형태단어후보들을 X^*, $X^*Y^*X^*$, Y^*X^*, Y^*, X^*Y^*로 표시할수 있다.(여기서 기호 *은 문자가 반복될수 있다는것을 표시한다.)

결국 형태단어를 확정한다는것은 X^*, $X^*Y^*X^*$, Y^*X^*, Y^*, X^*Y^* 형태의 형태단어후보에서 X^*, $X^*Y^*X^*$, Y^*X^* 형태의 문자렬들을 얻어낸다는것을 의미한다.

형태단어후보모임을 AW, 형태단어모임을 BW 라고 하면 형태단어후보모임에서 형태단어들을 식별하는 과정을 수학적으로 다음과 같이 표시할수 있다.

즉 f_1

$$AW \rightarrow BW$$

AW 를 BW 로 넘기는 과정을 구체적으로 보면

$X^* \rightarrow X^*$

$X^*Y^*X^* \rightarrow X^*Y^*X^*$

$Y^*X^* \rightarrow Y^*X^*$

$Y^* \rightarrow \Phi$

$X^*Y^* \rightarrow X^*$

이때 X^*, $X^*Y^*X^*$, Y^*X^*, Y^*, X^*Y^* 형태의 형태단어후보들은 AW에 속하며 X^*, $X^*Y^*X^*$, Y^*X^* 형태의 형태단어후보들은 형태단어로 인정되여 BW 에 속한다.

형태단어후보에서 형태단어를 확정하기 위하여 다음과 같은 문자처리규칙을 설정한다.

규칙 1. 조선어음절글자가 아닌 다른 문자들로만 이루어진 형태단어후보(세번째 류형)는 삭제한다.

규칙 2. 특수한 경우의 형태단어로 되는 형태단어후보에서 마지막위치에 있는 조선어음절글자가 아닌 다른 문자 또는 문자렬들은 삭제한다.

조선어본문에서 쓰이는 문자가운데서 조선어음절글자가 아닌 다른 문자들의 모임을 $T1$, 형태단어모임을 BW 라고 하자. 이때 $BW = \{BW_i | 1 \le i \le n\}$이다.

형태단어후보에서 $T1$의 식별은 형태단어후보를 이루는 문자렬을 오른쪽 또는 왼쪽 방향으로 한글자씩 이동하면서 얻어지는 문자를 $T1$와 비교하면서 진행한다.

그러므로 형태단어확정과정은 형태단어후보모임 AW의 매 형태단어후보의 마지막위치에 있는 문자가 $T1$에 속하는가 하는것을 확인하고 속한다면 그것을 삭제하는 과정이라고 말할수 있다. 형태단어후보 WS_i에 $T1$중의 어느 하나를 포함하고있는가를 판정하는 함수 $Symbola(Rtrim(WS_i, 1))$을 다음과 같이 정의한다.

$$Symbola(Rtrim(WS_i, 1)) = \begin{cases} 1 & Rtrim(WS_i, 1) \in T1 \\ 0 & Rtrim(WS_i, 1) \notin T1 \end{cases}$$

그러면 조선어형태단어를 확정하는 알고리듬을 작성하여보자.

[형태단어확정알고리듬]

1.변수들을 초기화하고 $i = 1$, $n = Count(AW)$로 한다.

2. i의 값을 판정한다.

1) $i \le n$인 조건에서 다음의 조작을 반복한다.

(1) $b = Len(WS_i)$로 한다.

(2) $b \geqq 1$인 조건에서 다음의 조작을 반복한다.

① $Symbola(Rtrim(WS_i,\ 1)) = 1$이면 ③의 조작에로 넘어간다.

② $Symbola(Rtrim(WS_i,\ 1)) = 0$이면 WS_i를 BW에 보관하고 (3)의 조작수행에로 넘어간다.

③ $b = b - 1$

④ $BW_i = Ltrim(WS_i,\ b)$

(3) $i = i + 1$

2) 작업을 끝낸다.

[알고리듬 끝]

실례로 본문《① 오유글자렬 x나 y에 대한 교정글자렬들을 a, b, c라는 기호로 약속하겠는가?》에 대한 알고리듬의 적용과정을 들수 있다.

본문《① 오유글자렬 x나 y에 대한 교정글자렬들을 a, b, c라는 기호로 약속하겠는가?》를 형태단어후보모임으로 변환하면 다음과 같다.

①
오유글자렬
x나
y에
대한
교정글자렬들을
a,
b,
c라는
기호로
약속하겠는가?

얻어진 형태단어후보들에 대하여 우의 알고리듬을 적용한다.

알고리듬 1의 조작에 의하여 모든 변수들은 초기화된다. 알고리듬 2의 조작에 의하여 형태단어후보에서 문자의 위치를 표시하는 변수 i는 1로 설정되고 형태단어후보모임에 들어있는 형태단어후보의 개수는 n에 등록된다. 이때 n의 값은 11이다. 알고리듬 3의 조작에 의하여 i의 값을 판정한다. i의 값이 n에 도달될 때까지 알고리듬 2-1)의 조작을 반복수행한다. 알고리듬 2-1)-(1)의 조작에 의하여 첫번째 형태단어후보《①》의 길이가 b에 등록된다. 첫번째 형태단어후보《①》의 길이가 1이기때문에 b의 값은 1로 설정된다. 즉 $b = 1$이다. 조건 $1 \le b$이 만족하므로 $Symbola(Rtrim(WS_1,\ 1))$ 함수의 값을 판정한다. 함수의 값이 1이기때문에 알고리듬 2-1)-(2)-③의 조작수행에로 넘어간다. b의 값은 하나 감소되고 알고리듬 2-1)-(2)-④의 조작이 수행된다. 결국 첫번째 형태단어후보《①》에서 그자체는 삭제되고 알고리듬 2-1)-(2)의 조작수행에로 넘어간다. b의 값이 0이기때문에 알고리듬 2-1)-(3)의 조작수행에로 넘어간다. i의 값이 하나 증가되여 2가 된다. 따라서 두번째 형

태단어후보《오유글자렬》에 대한 해석을 진행한다. 두번째 형태단어후보《오유글자렬》의 문자개수는 5이다. 조건 $b \geqq 1$이 만족하므로 $Symbola(Rtrim(WS_2, 1))$ 함수값을 판정한다. 함수값이 0이기때문에 알고리듬 2-1)-(3)의 조작을 수행한다. 그리하여 WS_2 즉《오유글자렬》이 BW에 등록된다. i의 값이 하나 증가된다. i의 값은 3이 되여 세번째 형태단어후보《x나》의 해석에로 이행한다. 세번째 형태단어후보《x나》에 대한 $Symbola(Rtrim(WS_3, 1))$의 값이 0이기때문에 WS_3 즉《x나》도 BW에 보관되고 변수 i의 값은 4가 된다. 네번째 형태단어후보《y에》에 대해서도 $Symbola(Rtrim(WS_4, 1))$의 값이 0이기때문에 네번째 형태단어후보《y에》도 BW에 보관되고 변수 i의 값은 5가 된다. 같은 방법으로 다섯번째 형태단어후보《대한》에 대한 $Symbola(Rtrim(WS_5, 1))$의 값도 0이기때문에 형태단어후보《대한》도 BW에 보관되고 위치상태변수 i의 값은 6이 된다. 여섯번째 형태단어《교정글자렬들을》에 대해서도 $Symbola(Rtrim(WS_6, 1))$의 값은 0이기때문에 형태단어후보《교정글자렬들을》은 형태단어로 BW에 보관되고 위치상태변수 i의 값은 7이 된다. 일곱번째 형태단어후보《a,》에 대하여 $Symbola(Rtrim(WS_7, b, 1))$의 값이 1이기때문에 2-2)-(2)-①의 조작에 의하여 WS_7에서 오른쪽 마지막글자인《,》은 삭제되고 2-2)-(2)-③의 조작수행에로 넘어간다. 그리하여 b의 값이 하나 감소되여 1이 된다. 다시 $Symbola(Rtrim(WS_7, 1))$ 값을 판정한다. 판정결과에 함수의 값이 1이기때문에《a》도 삭제된다. 다시 b의 값이 하나 감소되여 b의 값은 0이 된다. 그리하여 알고리듬 2-1)-(3)의 조작이 수행되여 위치상태변수 i의 값은 8이 된다. 즉 여덟번째 형태단어《b,》에 대한 해석에로 넘어간다. 일곱번째 형태단어와 마찬가지로 두개의 문자가 모두 $T1$에 속하기때문에 이 형태단어후보에 대해서는 형태단어가 얻어지지 않는다. 위치상태변수 i의 값이 9가 된 조건에서 우와 같은 조작들을 반복수행하면 아홉번째 형태단어후보에 대해서 $Symbola(Rtrim(WS_9, 1))$의 값이 0이기때문에 그자체가 형태단어로 된다. 열번째 형태단어후보《기호로》에 대해서도 같다. 따라서 열번째 형태단어후보《기호로》도 형태단어로 확정된다. 열한번째 형태단어《약속하겠는가?》에 이르러 함수 $Symbola(Rtrim(WS_{11}, 1))$의 값은 1이다. 그러므로 알고리듬 2-1)-(2)-③의 조작에 의하여 열한번째 형태단어후보《약속하겠는가?》에서 마지막문자《?》이 삭제되고 WS_{11}에《약속하겠는가》가 등록된다. b의 값이 1에 도달되지 못하였기때문에 함수 $Symbola(Rtrim(WS_{11}, 1))$의 값이 다시 판정된다. 판정결과에 함수의 값이 0이기때문에 WS_{11}인《약속하겠는가》는 BW에 보관된다. 그리고 알고리듬 2-1)-(3)의 조작에 의하여 위치상태변수 i의 값이 하나 증가되여 12이 된다. i의 값이 b보다 크기때문에 작업은 중지된다. 결국 알고리듬의 수행결과에 형태단어후보렬 AW로부터 다음과 같은 형태단어모임 BW가 얻어진다.

오유글자렬, X나, Y에, 대한, 교정글자렬들을, c라는, 기호로, 약속하겠는가.

조선어음절글자가 아닌 문자들을 금지어로 규정하고 금지어처리단계에서 삭제할수도 있다.

자동색인에서는 본문에 대한 형태단어후보들에 대한 형태단어확정이 끝나면 주제어추출단계에로 넘어간다.

우리는 도서관의 물질기술적토대가 더욱더 훌륭히 갖추어지는데 맞게 자동색인에서 제기되는 문제들에 대한 연구를 심화시켜 사회주의강국건설에 적극 이바지해나가야 할것이다.