PDA를 이용한 한국어 자동 섹션 시스템

Korean Automatic Indexing System Using the PDA

박정구, 정인정
고려대학교 전산학과

Pyeung-Koo Park, In-Jeong Chung
Dept. of Computer Science, Korea Univ.

요약
이미지네의 금속날고방장용사용자의진단의실현을위한접합점검책에따로정합재검사방법을제시하고자한다.
한사용자의요구는전산과학자의성능향상에필수적인이론적논문의내용을대표
하는섹션체계추출하는한사용자의결과에대한판을가져지었다.
본간의결과에서의자동신문방법

1. 서론

인Terrain의보유로써변방문문자들의접합점검에서서서문자의

2. 오토마타(Automata)와 PDA(Push Down Automata)

본논문은PDA를이용하여한국어로구성된문자의자

3. PDA의구성


그래프 1. PDA의구성

우리가본논문에서다루는PDA는상황에따라2가지

입력부호가사용되지않고도동작을취할수있는ε-동작
이허용되는경우에서비결정적이(non-deterministic)인PDA이

한편,문제자유에서PDA에의해구현될때2가
지 PDA에 의해 수락된다. 첫 번째 PDA는 빈 스택(EMPTY STACK)에 의해 문맥 자유 언어를 수락하는 방법이며, 두 번째 PDA는 소개의 내용과 상관없이 스택 상태(ACCEPTING STATE)들에 의해 문맥 자유 언어를 수락하는 방법이다. 만약 어떤 언어가 둘째 PDA에 의해 수락되면, 이 언어는 스택 상태에 의한 다른 PDA에 의해서도 수락하며, 그 역도 성립한다.

3. PDA를 이용한 한정어 자료구조 방법
위에서 언급한 일반적인 PDA를 한국어의 자료구조 시
스템에 적용하기 위해서 입력되는 문장을 제거기가, 앞의
비리 정의된 PDA에 의해서 다음의 상태를 결정하고 스택에
입은 문자를 저장하는 반복적인 과정을 수행하게 된다. 이
때 수행단계에서 스택에 저장되었던 문장에서 센터로 추출(pop)한다. 다음의 알고리즘은 입력(UNKNOWN)의 형태로 설명하고 있다.

3.1 알고리즘
본 논문에서 제안하는 자료 사용 방법은 제한적
한 단계로 이루어진다. 첫 번째 단계는 문자 내의 문자열의 산출
단계이며, 두 번째 단계는 입력된 문장에서 스택에 리스트(REQUEST)를 입력의 문자로 산출하는 단계이다. 본 논문
에서 사용된 방법은 문장 내에서 센터로 추출할 가능한 생성의
문자열을 포함하는 생성의 사용과 동의 용어를 뜻한다. 한국어에서 센터로 추출할 가능한 생성의 단위가 부분이
문서의 단계이다. 마지막 단계는 PDA를 이용한 센터로
추출 단계이다. 아래의 알고리즘은 제한적 단계의
문장 사용 알고리즘이다. 아래의 알고리즘에서 스택을 이용
하는 모델과 상태를 갖는 모델은 다음의 PDA를 통한 구
함을 통해서 알 수 있다.

input : 문장 입력
output : 센터 리스트
char ch
FILE *file_pointer
while (EOF)   /* 파일의 끝까지 센터 방법을 적용한다. */
{
    while (EOL)  /* 파일의 끝까지 센터 방법을 적용한다. */
    {
        ch = getchar(file_pointer);
        if (ch == character)
            push_stack();
    }

    /* 입력된 문자가 공백이거나 문장이 아니면 스택에 문자를 push 한다. */
    else
    {
        pop_to_buffer();
    }
}

위의 메인 모듈은 파일에서 문장단위로 Portal과 제거
된 입력어터에서 얻은 입력이 문자이면 문자를 스택에
push하는 모듈을 아니면 센터를 스택에서 추출하는 모듈
을 포함하게 된다.

/* push 모듈 */
push_stack()
{
    stack[top] = ch;
    symbol_stack[top] = find_stack_symbol();
    stack_state[top] = stack_state;
    /* 입력된 문자를 스택에 push하기 스택 상태를 찾아 push 한다. */
}

push_stack() 모듈은 얻은 문자를 스택의 top에
push하고 스택 상태를 찾는 모듈을 통해서 스택의 상태를 찾아 push한다. 또한 상태(state) 정보를 찾아 각각
top의 위치에 push를 수행한다. find_stack_symbol()는 현재
스택의 스택 상태를 상태로 PDA를 바탕으로 스택 상태를
결정하는 모듈이다.

3.2 PDA 설계
위의 PDA를 이용한 한국어 자료구조 방법의 알고리즘
에서 PDA는 다음과 같이 정의할 수 있다. \( M = (Q, \sigma, \delta, \text{character, space, stack list, } R, B, G) \) \( Q, \sigma, \delta, \text{character, space, stack list, } R, B, G \)는 그림 3에 정의된 메인 PDA는 상태 \( q_0, q_s \)로
구성되어 있고 입력은 문자, 공백, 문장의 리스트로 구성된 다. 또한 스택 상태를 처음 스택을 의미하는 \( R \)과 공백이나
문장의 입력을 의미는 \( G \)의 문자의 입력을 의미하는 \( R \)으로 구성된다. \( q_0 \)와 \( q_s \)를 최종 수락 상태로 한다. 다음의 그림 2는 위
의 정의된 PDA를 나타낸다.

\[ \begin{align*}
\delta(q_0, \text{character, R}) & = \{(q_1, BR)\} \\
\delta(q_1, \text{character, B}) & = \{(q_2, BG)\} \\
\delta(q_2, \text{character, G}) & = \{(q_3, BG), (q_4, \varepsilon)\} \\
\delta(q_3, \text{space, B}) & = \{(q_4, BG), (q_5, \varepsilon)\} \\
\delta(q_4, \text{space, G}) & = \{(q_5, G)\}
\end{align*} \]
그림 2. PDA를 이용한 한국어 자동 섹션 방법

제안하는 PDA를 이용한 한국어 자동 섹션 방법에서 시작 상태는 $q_0$이고 수학하는 상태는 $q_2$과 $q_4$이다. 초기 상태 가에서 $q_0$의 character는 상태 $q_2$이고 스크립트는 $R$일 때 입력 테이프의 입력이 문자(character)이므로, $(q_0, BR)$와 같이 상태로 표현해보면 $q_2$로 되고 스크립트의 $BH$로 치환된다. 이때 입력된 문자는 스크립트에 push된다. 그런데 $(q_2, space, BR)$는 상태 $q_2$이고 스크립트가 $B$이고 입력 테이프가 공백 일 때 이후에 언급된대로 문자를 제한하는 PDA는 비 결정적임으로 $(q_2, GB)$의 경우 $(q_2, e)$로 상태가 변환된다. 또한 $(q_2, e, e)$의 의미는 상태에서 입력의 상태에 관계없이 스크립트 $R$을 계속해서 상대로 주석화하게 되며 $(q_2, e, e)$로도 되고 $(q_2, BH)$의 의미는 상대에서 주석화하는 최종 단계를 의미하게 된다. $(q_2, e, e)$는 입력 문자에 관계없이 스크립트에 $(GB)$가 변환되도록 되어 다음 $(q_2, BH)$ 상태로 되고 문자를 push하면서 추출 상태 끝이 되며 $(q_2, e)$가 되는 $BN$으로 변경하게 된다. 이렇게 하면 문장의 문장이 되어 $BN$으로 변환된다. 제 3 단계를 설명하기위해 PDA의 순간표식
(Instantaneous description : ID)를 이용한다. 그림 3에 예
제에 대해서 PDA를 통한 섹션 결과를 보인다. EOL(end of line)은 문장의 끝을 나타낸다.

그림 3. PDA를 통한 섹션 방법 예제

위의 단계 $(1)$은 처음 상태를 이용하여 한국어 자동 섹션 방법에 의해서 섹션은 제 1 단계 이 되고 스크립트는 $BH$로 치환에 이루어진다. 단계 $(2)$에서 변환될 때 제한하는 PDA는 비결정적임으로 단계 $(3)$와 단계 $(4)$로 변환 가능하게 된다. 단계 $(2)$에서의 상태 $q_2$에 상태 $q_4$로 변경될 수 있도록 변환을 위해 스크립트의 $BR$ 부분에 $q_4$로 변환하게 된다. 또는 단계 $(2)$와 같은 조건에서 계속해서 입력 테이프를 읽는 점개는 단계 $(4)$로 이동시킨다. 이때 알고리즘을 통해서 임의 상태 $q_4$이 이동할 수 있도록의 현재 상태가 $q_0$에 변화하는 과정이다. 단계 $(3)$의 $pop$을 수행하여 섹션을 자동화하는 복잡도가 그림 4에 나타나 있다.
그림 7. PDA를 통한 색인어 추출

그림 7은 33절 에게 "컴퓨터 바이러스나 백신에 대한 문서"이라는 텍스트에서 색인어로 10개가 추출되었음을 나타내고 있다. 앞에서 설명되었던 것처럼 "컴퓨터"는 단일명사 "컴퓨터 바이러스"라는 복합명사를 색인어로 추출되었다. 다음의 그림 8는 세 개의 명사로 이루어진 문장에서 색인어 추출 결과를 보여준다.

그림 8. 세 개의 합성명사 색인

그림 8은 "정보 검색 시스템"이라는 세 단어 합성명사에 대한 색인 결과를 보여주고 있다. 결과에서 알 수 있듯이 "정보" "검색" "시스템" 등 단일명사와 "정보 검색" "검색 시스템등" 두 단어 합성명사와 함께 "정보 검색 시스템" 등의 세 단어 합성명사와 색인어가 가능하다. 다음 그림 9는 여러 문장에서 색인어를 추출한 결과를 보여주고 있다. 영문 문장으로 "korean automatic indexing system using the PDA"이 입력되었다고 가정한다.

그림 9. 영어 문장에서 색인 결과

그림 9의 결과에서는 국문법을 바탕으로 색인어를 추출한 결과이므로 영문법을 적용하여 PDA를 수정한다면 영문의 색인어 추출이 가능할 것이다. 현재 인터넷과 정보통신에 대한 관심이 증가되고 있어, 인터넷을 통한 다른 언어로의 확장은 검색 시스템에서 상당히 중요한 특성이 될 수 있다. 이와 같은 결과에서 알 수 있듯이 단일명사와 복합명사의 색인과 인터넷과 다른 언어로의 확장이 가능한 결과를 볼 수 있다.

5. 결론

인터넷의 발달로 방대한 양의 문서에서 정보를 효율적으로 관리하고 이용하기 위하여 많은 정보 검색 시스템들이 개발되고 있다. 문서의 내용을 나타내는 색인어 추출하는 색인어 추출시스템은 정보 검색 시스템 성능에 큰 영향을 미치게 된다. 본 논문에서는 기존의 한국어 자동 색인 시스템의 문제점들을 언급하고, 정형화된 한국어 자동 색인 시스템을 제안하였다. 기존의 단어 중심 텍스트식은 복합명사의 색인어 이어, 형태소 중심식은 범용적 인터넷과 관련된 인터넷 저작 권리에 연결하는 웹사이트의 반응이 많아 필요하고, 분석에 필요한 사전 (dictionary)이 필요하고 사전의 유지와 보수에 많은 노력이 필요하다. 이러한 형태의 텍스트식은 분산 구현이 복잡하다. 기존의 최근의 n-gram 중심 색인법은 n 값을 어떻게 정하는 지의 결정적인 문제가 있으며, 이로 인해 색인어 추출 시점의 효율이 저하되고 부정확한 색인 어 추출으로 저장공간의 낭비를 가져온다. 제안하는 PDA를 이용한 자동색인 시스템은 영어에 대한 사전저작권이 필요하고, 형태소식 및 단일 생성된 단어의 색인어가 가능한 뿐 아니라 인터넷으로의 확장과 다른 언어로 구현이 쉬운 장점을 갖는 정형화를 의미하는 특성을 가진다. 정형화년수 기존의 KSET을 기반으로 PDA를 이용한 한국어 자동색인 시스템으로 색인어 추출 테스트를 하였다.

참고 문헌
[1] 김한구, 조주연, "상호 정보에 기반한 한국어 텍스트의 범용적 자동색인," 한국정보과학회 논문지 제21권 제7호 (94.7)
[5] 강승식, 정병택, "다중 특성을 이용한 범용 한국어 형태소 분석기 및 맞춤법검사기," 정보과학회지(B) 제23권 제5호 (96.5)