김일성종합대학학보
(자연과학)
주체106(2017)년 제63권 제1호

JOURNAL OF **KIM IL SUNG** UNIVERSITY
(NATURAL SCIENCE)
Vol. 63 No. 1 JUCHE106(2017).

정보통합방식의 조선어필기문자인식체계 실현에 대한 연구

최 영 훈

지금까지 연구된 독립방식의 인식은 주로 정필기문서를 대상으로 연구가 진행되였으므로 개별인식기의 성능을 높이는 문제에 국한되였다. 이것을 극복하기 위하여 분리－인식－언어정보통합[2]에 기초한 인식체계들이 개발되였으나 그 성능이 아직 낮은 단계에 있다.

론문에서는 정보통합방식에 기초한 조선어필기문자인식체계실현의 한가지 방법을 제안하였다.

1. 정보통합방식의 조선어필기문자인식문제의 정식화

조선어일반필기문서는 필기자에 따라 문자정보의 변형이 다양하고 필기관습상 문자들사이에서 각이한 특성이 나타난다.

이로부터 일반필기문서에서는 개별문자인식성능이 높은 경우에도 문자분리결과에 따라 각이한 인식결과를 줄수 있으며 결국 필기문자렬인식에서 문자분리문제가 중요한 문제로 나선다.

이러한 문제를 해결하기 위하여 우리는 가능한 문자분리경로에 대하여 분리－인식－언어정보통합에 기초한 문자렬인식방법을 제안하려고 한다.

문자렬인식문제는 문자분리믿음도와 문자류사도, 어휘문맥을 결합하여 최량토막에 대응되는 최량문자렬을 찾는 문제이다.

문자렬화상 I 가 n 개의 후보패턴들인 $S=S_1\cdots S_n$ 으로 토막화되여 하나의 문자렬 $W=W_1\cdots W_n$ 에로 할당되였다고 가정하면 이때 문자렬의 사후확률은

$$P(W\mid X)=\sum_{S}P(W,\ S\mid X) \tag{1}$$

로 표시할수 있다. 여기서 후보패턴들은 특징벡토르 $X=X_1\cdots X_n$ 에 의하여 표현된다.

그리고 최량분리상태에서 최량문자렬은 다음의 식으로 표현할수 있다.

$$W^{*}=\arg\max_{W}\max_{S}P(W,\ S\mid X) \tag{2}$$

우의 식에서 볼수 있는바와 같이 최량문자렬을 구하는 문제는 결국 매 문자렬에 대하여 최량분리후보 S 를 찾는데 귀착된다.

한편

$$P(W,\ S\mid X)=\frac{P(X\mid W,\ S)P(W,\ S)}{P(X)}=\frac{P(X\mid W,\ S)P(S\mid W)P(W)}{P(X)}$$

이다.

우의 식에서 문자렬을 구성하는 매 문자들이 서로 독립이라는것을 가정한다면 $P(X|W,\ S)$는 다음과 같이 근사시킬수 있다.

$$P(X|W,\ S)\approx\prod_{i=1}^{n}P(X_i|W_i,\ S_i)\approx\prod_{i=1}^{n}P(X_i|W_i) \tag{3}$$

여기서 $P(X_i|W_i)$는 클라스 W_i에 대한 입력패턴 X_i의 분포밀도함수(우도함수)로서 개별문자식별기에 의해 추정된다.

$P(X)$가 일정하다는것을 고려하면 최량문자렬은

$$W^*=\arg\max_W\max_S P(X|W,\ S)P(S|W)P(W) \tag{4}$$

에 의하여 결정된다. 여기서 $P(X|W,\ S)$는 문자렬인식모형, $P(S|W)$는 문자분리모형, $P(W)$는 언어모형이다.

식 (4)에서 $P(S|W)=1$인 경우

$$W^*=\arg\max_W P(X|W)P(W) \tag{5}$$

로 된다.

식 (4)에서 보는바와 같이 정보통합방식의 문자렬인식은 문자분리정보, 인식정보, 언어정보 등을 어떻게 리용하는가에 따라 그 성능이 좌우된다.

따라서 우리는 조선어필기문자렬의 가능한 분리경로에 대하여 최량분리경로탐색에 기초한 정보통합실현방법을 제안한다.

2. 분리후보그라프를 리용한 문자분리경로생성

1) 분리후보그라프

정의 1 문자렬화상들에 대한 분리선들을 정점으로 하고 분리선들에 의하여 분리되는 후보문자화상을 릉으로 하는 그라프를 분리후보그라프라고 하고 $G(T,\ X)$로 표시한다. 여기서 정점모임 $T=\{t_i,\ i=\overline{0,\ n}\}$은 분리선들의 모임이며 릉의 모임 $X=\{X_{ij}\}$, $i=\overline{0,\ n}$, $j=\overline{1,\ m_i}$는 후보문자화상들의 모임이다.

2) 분리후보그라프생성

분리후보그라프생성과정은 분리선생성과 분리후보그라프생성의 두 단계로 진행된다.

① 기하학적방법 및 식별기를 리용한 분리선의 생성

평활화주변분포를 리용하여 선형분리선을, Viterbi알고리듬으로 1차비선형분리선을 생성하고 동질분리선에 의하여 2차비선형분리선을 생성한다.

두 분리선의 비교평가를 위하여 동질분리선을 다음과 같이 정의한다.

정의 2 두 분리선 p_1과 p_2에 대하여 다음의 두 조건을 동시에 만족시킬 때 두 분리선 p_1과 p_2를 동질분리선이라고 부른다.

$$\begin{aligned}&nCS_{p_1}=nCS_{p_2}\\&IDS_{p_1}[i]=IDS_{p_2}[i],\ i=\overline{1,\ nCS_{p_1}}\end{aligned} \tag{6}$$

여기서 nCS는 p와 사귀는 문자획의 개수, IDS는 p와 사귀는 문자획번호들의 모임이다.

1차비선형분리선으로부터 2차비선형분리선의 생성은 동질분리선과 조선어문자구조와의 관계를 고려하면서 불필요한 분리선들을 제거하는 방법으로 얻는다.

이렇게 얻은 1차선형분리선과 2차비선형분리선들을 최종분리선으로 한다.

② 분리후보그라프의 생성

그림 1에 분리후보그라프생성흐름도를, 그림 2에 분리후보그라프생성방법에 따르는 필기문자렬《과학기술로》에 대한 분리후보그라프를 보여주었다.

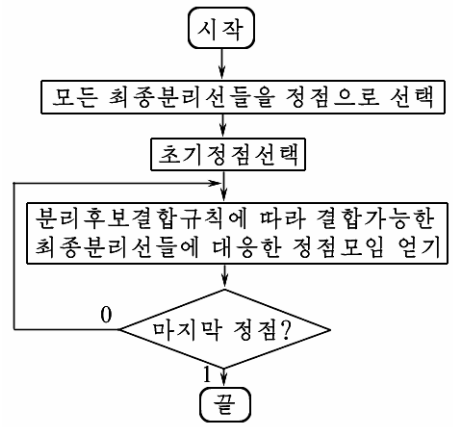

그림 1. 분리후보그라프생성흐름도

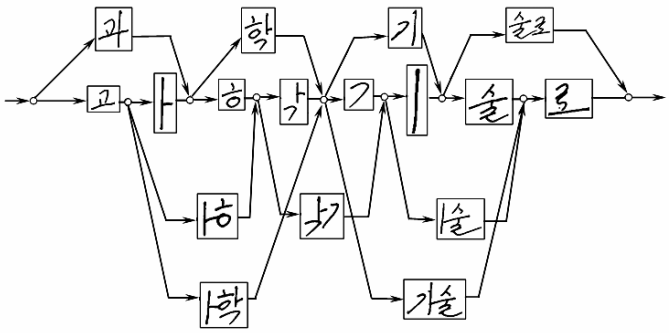

그림 2. 분리후보그라프

그림 2에서 분리후보결합규칙은 조선어일반필기문자렬에 대하여 분리선을 생성할 때 나타나는 분리특성을 반영하는 규칙으로서 분리후보화상들을 몇개까지 결합할수 있는가를 반영하는 규칙이다.

3. 정보통합방식의 조선어일반필기문자렬인식체계실현

정보통합방식의 조선어필기문자렬인식은 분리후보그라프의 분리경로모임에 대하여 분리－인식－언어정보통합에 기초한 최량분리경로를 얻는 방법으로 실현한다.

1) 최량문자렬평가값의 결정

최량문자렬의 평가는 문자분리경로믿음도, 인식믿음도, 언어우도에 의하여 결정하기로 한다. 문자분리경로믿음도는 선행한 방법[1]에 의하여 결정하며 개별문자인식믿음도와 언어우도는 문자인식부의 확률모형인 3-gram언어모형에 의하여 결정한다.

결국 문자렬인식믿음도 $V(c_1, \cdots, c_n)$은 다음과 같이 결정된다.

$$V(c_1, \cdots, c_n) = \alpha_1 \lg p_c{}^s + \alpha_2 \sum_{i=1}^{n} \lg p(x_i / w_i) + \alpha_3 \sum_{i=3}^{n} \lg p(c_i / c_{i-1}, c_{i-2}) \qquad (7)$$

식 (7)에서 1항은 문자분리믿음도, 2항과 3항은 각각 개별문자인식믿음도와 언어우도이며 α_i는 인식대상의 특성에 따라 실험적으로 결정한다.

2) 필기문자렬의 인식 및 효과성분석

문자렬인식은 우에서 얻은 분리경로모임에 대하여 최량문자렬평가값이 최대인 문자렬을 얻는 방법으로 얻는다. 이때 전탐색방법을 리용한다.

론문에서 제안한 정보통합방식과 선행한 독립체계구성방식의 인식성능을 평가한 결과를 표에 보여주었다.

표. 인식체계의 성능에 대한 실험결과

문서종류		인식률/%	인식속도/(자·s^{-1})
정필기문서	선행방법[2]	99.5	200
	정보통합	99.5	150
일반정필기문서	선행방법[2]	95.4	180
	정보통합	97.2	130
일반필기문서	선행방법[2]	87.5	150
	정보통합	92.3	110

표에서 보여주는바와 같이 정보통합방식은 일반필기문서인식에서 선행한 방법에 비하여 보다 높은 성능을 가진다.

처리속도는 정보통합방식에서 약간 떠지지만 탐색속도를 고속화하면 성능을 보다 더 개선할수 있다.

맺 는 말

필기문자렬의 가능한 분리경로에 대하여 분리－인식－언어정보통합방식에 기초한 최량분리 및 인식방법을 제안하고 그 유효성을 비교평가하였다.

참 고 문 헌

[1] 최철 등; 정보기술, **1**, 11, 주체105(2016).

[2] Hiromichi Fujisaiwa; Pattern Recognition, **41**, 2435, 2008.

주체105(2016)년 9월 5일 원고접수

Realization of Korean Handwriting Character Recognition System of Information Combination Mode

Choe Yong Hun

We proposed the optimal segmentation and recognition method based on information combination mode on possible division path of general handwriting character string and experimentally confirmed the effectiveness by applying in Koran handwriting character recognition.

Key words: character recognition, information combination mode, handwriting character string