KAIST 전산학과 최광무

한글은 정규언어이므로 DFA로 한글을 판별하고 Mealy machine으로 모아쓸 수 있다.

(한글의 기본문자) 한글은 자음 14자와 모음 10자 모두 24개의 기본자모로 이루어져 있음은 잘 알고 있을 것이다. 그러나 초성과 받침에 모두 ㄲ과 ㅆ을 허용하므로 기본자모 24개만 허용하면 'ㄱㅏㄱㄱㄱㅗ'를 치는 경우에 '각꼬'로 받아들일 수도 있고 '갂고'로 받아들일 수도 있는 모호함(ambiguity)이 있다. 따라서 기본자모 24자에 초성에 올 수 있는 겹자음 'ㄲ, ㄸ, ㅃ, ㅆ, ㅉ' 5자를 한글 기본문자에 추가하여 총 29자를 한글기본문자로 내 석사학위논문[1)]에서 주장하였다.

구분	내용	갯 수	총갯수
자음	ㄱ, ㄴ, ㄷ, ㄹ, ㅁ, ㅂ, ㅅ, ㅇ, ㅈ, ㅊ, ㅋ, ㅌ, ㅍ, ㅎ	14 자	
모음	ㅏ, ㅑ, ㅓ, ㅕ, ㅗ, ㅛ, ㅜ, ㅠ, ㅡ, ㅣ	10 자	
겹자음	ㄲ, ㄸ, ㅃ, ㅆ, ㅉ	5 자	

표 2.4.1 한글의 기본자모

(한글의 구성) 한글은 초성과 중성 받침에 세 가지 요소로 구성되어 있다. 초성은 'ㄱ'부터 'ㅎ'까지 기본자음 14개와 겹자음 'ㄲ, ㄸ, ㅃ, ㅆ, ㅉ' 5개를 합하여 19가지가 있고, 중성은 기본모음 'ㅏ'부터 'ㅣ'까지 10개와 복합모음이 'ㅐ, ㅒ, ㅔ, ㅖ, ㅘ, ㅙ, ㅚ, ㅝ, ㅞ, ㅟ, ㅢ' 11개를 합하여 21가지가 있다. 받침은 종성은 초성을 또 쓴다는(두 개 혹은 세 개의 복자음을 허용한다) 훈민정음 서문의 규칙에 따르면 상당히 많은[2)] 받침이 허용되나[3)], 현대 한글 맞춤법 통일안에서는 용례(用例)를 두어 이를 제안하고 있었다. 한글에서는 가장 빈도수가 높은 받침이 없는 경우를 포함하여(ε) 모두 28가지의 받침을 나는 석사학위논문[4)]에서 가정하였다.

구분	내용	갯 수
초성	ㄱ, ㄴ, ㄷ, ㄹ, ㅁ, ㅂ, ㅅ, ㅇ, ㅈ, ㅊ, ㅋ, ㅌ, ㅍ, ㅎ, ㄲ, ㄸ, ㅃ, ㅆ, ㅉ	19 개
중성	ㅏ, ㅑ, ㅓ, ㅕ, ㅗ, ㅛ, ㅜ, ㅠ, ㅡ, ㅣ ㅘ, ㅝ ㅙ, ㅞ ㅐ, ㅒ, ㅔ, ㅖ, ㅚ, ㅟ, ㅢ	21 개
받침	ㄱ, ㄴ, ㄷ, ㄹ, ㅁ, ㅂ, ㅅ, ㅇ, ㅈ, ㅊ, ㅋ, ㅌ, ㅍ, ㅎ, ㄲ, ㅆ ㄳ, ㄵ, ㄺ, ㄻ ㄶ, ㄽ, ㄾ ㄿ, ㅀ, ㄼ, ㄽ, ㄾ, ㄿ ε	28 개

표 2.4.2 한글의 구성

이렇게 되면 한글 = (초성 • 중성 • 받침)*의 정규식으로 표현된다. 따라서 한글은 19 × 21 × 28 = 1,1172개 서로 다른 음절로 정의된다. 1만여 개의 음절을 기본문자 29자 Σ_{29} =

1) 최광무, "한글 모아쓰기에 관한 연구," 한국과학원, 석사학위논문, 1978.
2) 훈민정음에서 허용한 초성 26자 외에 수많은 많은 복자음들(예 ㅥ, ㅦ, ㅩ 등)
3) 훈민정음 서문에 있는 종성복용초성(終聲複用初聲)을 문자 그대로 해석하면 4)의 해석이 가능하나, 훈민정음 해례본(解例本) 종성해(終聲解)에서는 ㄱ, ㅇ, ㄷ, ㄴ, ㅁ, ㅅ, ㅂ, ㄹ의 8개의 받침만 허용한다고 나와 있으므로, 중세 한글과 현재 한글은 소리문자를 포기한 과도한 허용이라고 볼 수 있다.
4) 3)과 같음.

{ㄱ,ㄴ,ㄷ,ㄹ,ㅁ,ㅂ,ㅅ,ㅇ,ㅈ,ㅊ,ㅋ,ㅌ,ㅍ,ㅎ,ㄲ,ㄸ,ㅃ,ㅆ,ㅉ,ㅏ,ㅑ,ㅓ,ㅕ,ㅗ,ㅛ,ㅜ,ㅠ,ㅡ,ㅣ}에서 정의한 한글의 정규식(초성$_{19}$ · 중성$_{21}$ · 받침$_{29}$)*과 이 정규식으로 정의한 현대한글을 받아들이는 minimal DFA는 아래와 같다.

Σ_{29} = {ㄱ,ㄴ,ㄷ,ㄹ,ㅁ,ㅂ,ㅅ,ㅇ,ㅈ,ㅊ,ㅋ,ㅌ,ㅍ,ㅎ,ㄲ,ㄸ,ㅃ,ㅆ,ㅉ,ㅏ,ㅑ,ㅓ,ㅕ,ㅗ,ㅛ,ㅜ,ㅠ,ㅡ,ㅣ}
(초성$_{19}$ · 중성$_{21}$ · 받침$_{29}$)* =
{ (ㄱ|ㄴ|…|ㅎ|ㄲ|ㄸ|ㅆ|ㅃ|ㅉ) ·
(ㅏ|ㅑ|…|ㅣ|ㅏㅣ|ㅑㅣ|ㅓㅣ|ㅕㅣ|ㅗㅏ|ㅗㅏㅣ|ㅗㅣ|ㅜㅓ|ㅜㅓㅣ|ㅜㅣ|ㅡㅣ) ·
(ϵ|ㄱ|ㄴ|…|ㅎ|ㄲ|ㅆ|ㄱㅅ|ㄴㅈ|ㄴㅎ|ㄹㄱ|ㄹㅁ|ㄹㅂ|ㄹㅅ|ㄹㅌ|ㄹㅍ|ㄹㅎ|ㅂㅅ) }*

식 2.4.1 Σ_{29}에서 정의한 현대한글의 정규식

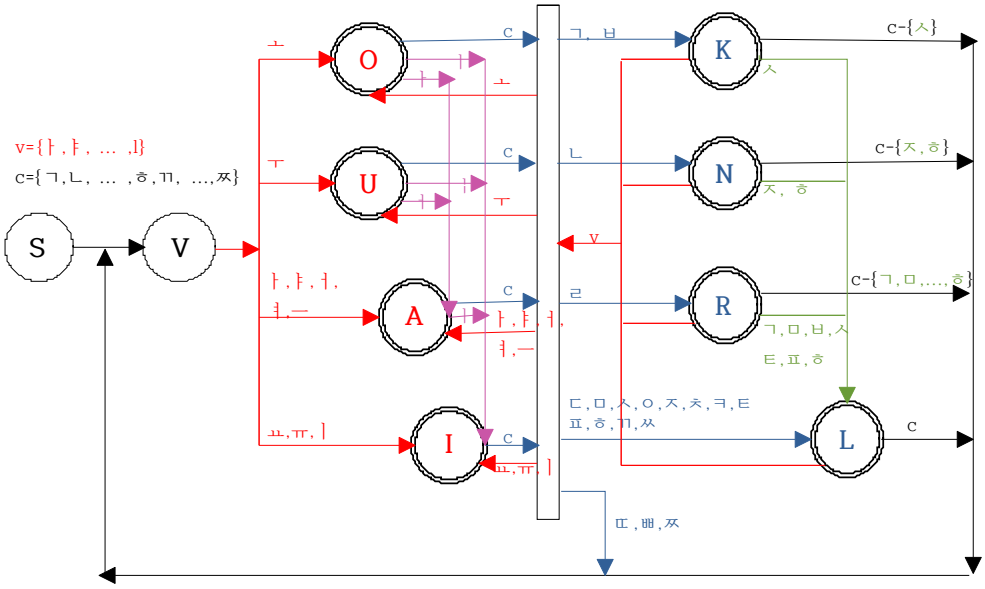

그림 2.4.1 한글 모아쓰기 오토마타(기본문자 29자)

(정의) Moore machine $Mo = (Q, \Sigma, \Pi, \delta, \lambda, q_0)$ 은 6 요소로 구성된다. Q, Σ, δ, q_0는 DFA와 같고[5], Π는 출력문자집합[6], λ: $Q \rightarrow \Pi$ 는 **상태 출력함수**이다.

(정의) Moore machine $Mo = (Q, \Sigma, \Pi, \delta, \lambda, q_0)$ 에서 임의의 입력문자열 $x = a_1 a_2 \cdots a_n \in \Sigma^*$ 에 대한 상태 변화가 $n \geq 1$: $1 \leq \forall i \leq n$: $\delta(q_{i-1}, a_i) = q_i$ 라 하자.

Moore machine Mo의 출력 $\Lambda(M_o(x)) \in \Pi^*$는 아래와 같이 정의한다.

$$\Lambda(M_o(x)) = \lambda(q_0)\lambda(q_1)\cdots\lambda(q_n) .$$

(정의) Mealy machine $Me = (Q, \Sigma, \Pi, \delta, \lambda, q_0)$ 는 6 요소로 구성된다. Q, Σ, Π, δ, q_0는

5) F가 정의되지 않음에 주의하라.
6) Π의 원소로 프로그램들(program segments) 또는 함수를 허용할 수 있다.

Moore machine과 같고, λ: $Q \times \Sigma \rightarrow \Pi$ 는 **상태변환 출력함수**이다.

(정의) Mealy machine $Me = (Q, \Sigma, \Pi, \delta, \lambda, q_0)$ 에서 임의의 입력문자열 $x = a_1 a_2 \cdots a_n \in \Sigma^*$ 에 대한 상태 변화가 $n \geq 1$: $1 \leq \forall i \leq n$: $\delta(q_{i-1}, a_i) = q_i$ 라 하자.

Mealy machine Me의 출력 $\Lambda(M_e(x)) \in \Pi^*$는 아래와 같이 정의한다.

$$\Lambda(M_e(x)) = \lambda(q_0, a_1)\lambda(q_1, a_2) \cdots \lambda(q_{n-1}, a_n).$$

입력문자열 $x \in \Sigma^*$ Mealy machine의 출력 $\Lambda(M_e(x))$ 의 길이가 $|x|$로 Moore machine의 출력 $\Lambda(M_o(x))$ 의 길이 $|x|+1$보다 하나 작지만, **상태변환** 출력함수 λ가 다양하므로 상태 수를 줄일 수 있어서 한글모아쓰기 오토마타 같은 경우 Mealy machine을 이용하는 것이 좋다.