

## Bài tập 1 (Trình phân tích từ vựng)

Trình biên dịch xây dựng CS4435 (Mùa xuân 2015)

Đại học Lahore

Maryam Bashir

Thời gian giao: Thứ Bảy ngày 14 tháng 3 năm 2015. Hạn chót: Thứ Hai ngày 23 tháng 3 năm 2015 23:59

### Phân tích từ vựng

Phân tích từ vựng là quá trình đọc dòng ký tự tạo nên mã nguồn của chương trình và chia đầu vào thành các mã thông báo. Trong bài tập này, bạn sẽ sử dụng các biểu thức chính quy và DFA để triển khai bộ phân tích từ vựng cho một tập hợp con ngôn ngữ lập trình C.

### Nhiệm vụ của bạn

Nhiệm vụ của bạn là viết chương trình đọc tệp văn bản đầu vào và xây dựng danh sách mã thông báo trong tệp đó. Chương trình của bạn có thể được viết bằng C, C++, Java hoặc bất kỳ ngôn ngữ lập trình nào khác. Giả sử rằng tệp đầu vào chứa chuỗi mã sau:

### 2.1 Đầu vào mẫu (mã C++)

Khoảng trống chính ()

```
{
    int sum = 0;
    for(int j=0; j < 10; j=j+1)
    {
        tổng = tổng + j + 10,43 + 34E4 + 45,34E-4 + E43 + 0,34;
    }
}
```

### 2.2 Đầu ra mẫu

Đầu ra của chương trình phải tương tự như sau:

Lớp: Từ khóa từ

vựng: void định

danh: chính

```
( :  
  
( ) : ) { :  
{ từ khóa : int định danh : tổng  
= : =  
số: 0  
; : ;  
từ khóa : for  
( :  
( từ khóa :  
int định danh : j  
= : =  
số: 0  
  
; : ; định  
danh : j < : <  
số : 10  
; : ;  
định danh: j  
= : =  
định danh :  
j + : +  
số : 1  
  
) : )  
{ : { định danh : tổng  
= : =  
định danh: tổng  
+ : +  
định danh :  
j + : +  
số : 10,43  
+ : +  
số : 34.E4  
+ : +  
số : 45,34E-4  
+ : +  
định danh : E43  
+ : +  
Lỗi :  
số : 34  
; : ;  
  
} : } } : }
```

Đặc tả từ vựng loại mã thông báo	
từ khóa	Một trong các chuỗi trong khi, nếu, ngược lại, trả về, ngắt, tiếp tục, , int float, void
định danh	Id mã thông báo cho số nhận dạng khớp với một chữ cái theo sau là chữ cái hoặc chữ số hoặc gạch dưới: thứ [AZ   az] chữ số [0-9] id chữ cái (chữ cái   chữ số   )
số	Số mã thông báo khớp với số không dấu: chữ số chữ số chữ số phân số tùy chọn ( . chữ số )   số mũ tùy chọn (E(+     chữ số )   num chữ số tùy chọn-phân số số mũ tùy chọn
addop	+ , -
mulop	, /
relop	< , > , <= , >= , == , !=
và	&&
hoặc	
không	!
)	)
	(
	{
	}
	[
( [ ( ) ]	]

Bảng 1: Bảng đặc tả từ vựng cho token

Mã thông báo hợp lệ

Các chương trình trong ngôn ngữ này bao gồm các mã thông báo được hiển thị trong bảng 1:

Xây dựng các sơ đồ tự động hữu hạn xác định (DFA) hoặc chuyển tiếp

Bạn có thể xây dựng DFA đơn (còn gọi là sơ đồ chuyển tiếp) để nhận dạng tất cả các mã thông báo trong này ngôn ngữ bằng cách kết hợp các DFA riêng lẻ cho từng loại mã thông báo. Ví dụ bạn có thể xây dựng sơ đồ chuyển tiếp cho các mã định danh và số và chúng hợp nhất các trạng thái bắt đầu của hai sơ đồ thành tạo một sơ đồ chuyển tiếp duy nhất nhận dạng cả số và mã định danh.

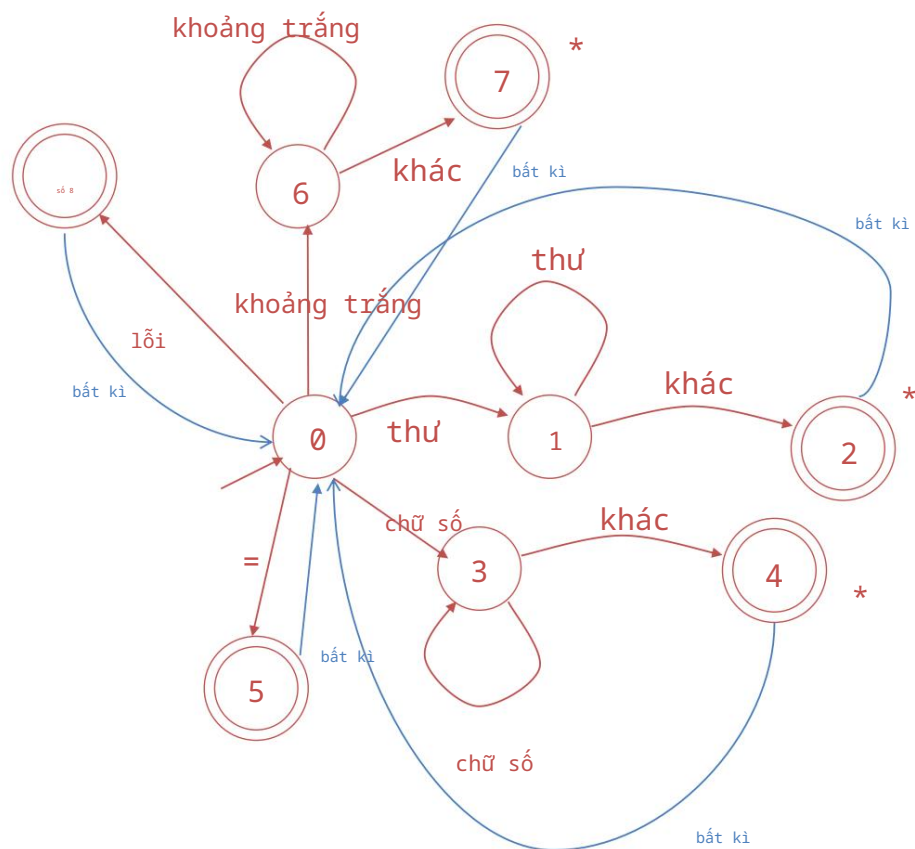
Để xác định từ khóa, bạn có thể lưu trữ từ khóa trong một số cấu trúc dữ liệu (hashmap hoặc chuỗi mảng). Bất cứ khi nào chương trình của bạn nhận ra mã thông báo nhận dạng, bạn có thể kiểm tra bản đồ hoặc chuỗi của mình

mảng nếu mã định danh này khớp với bất kỳ từ khóa nào. Nếu nó khớp với bất kỳ từ khóa nào thì hãy coi đó là từ khóa, nếu không hãy coi đó là mã định danh.

## Triển khai Máy tự động hữu hạn xác định (DFA)

Tôi đã viết một trình phân tích từ vựng nhỏ cho một DFA đơn giản được hiển thị trong hình 1. Bạn có thể tìm thấy mã nguồn của nó ("Lexer.java") trên trang Piazza. Bạn có thể thực hiện các thay đổi đối với mã này hoặc viết lại trình phân tích từ vựng của riêng mình. Bạn cũng có thể viết phiên bản C++ của mã Java này nếu bạn muốn lập trình bằng C++.

Xin lưu ý rằng DFA này không phải là DFA chính xác cho bài tập này. Đây chỉ là một DFA mẫu để cung cấp cho bạn ý tưởng về cách chuyển đổi DFA thành mã. Ví dụ: DFA này chỉ cho phép các chữ cái trong mã định danh trong khi bằng ngôn ngữ C, bạn cũng có thể sử dụng chữ số và dấu gạch dưới trong mã định danh.



Hình 1: DFA

Nộp hồ sơ

Gửi email tệp zip chứa dự án hoàn chỉnh của bạn (tất cả các tệp nguồn, cùng với bất kỳ tệp nào khác cần thiết để biên dịch chúng; tệp đầu vào và đầu ra mẫu) tới địa chỉ sau:

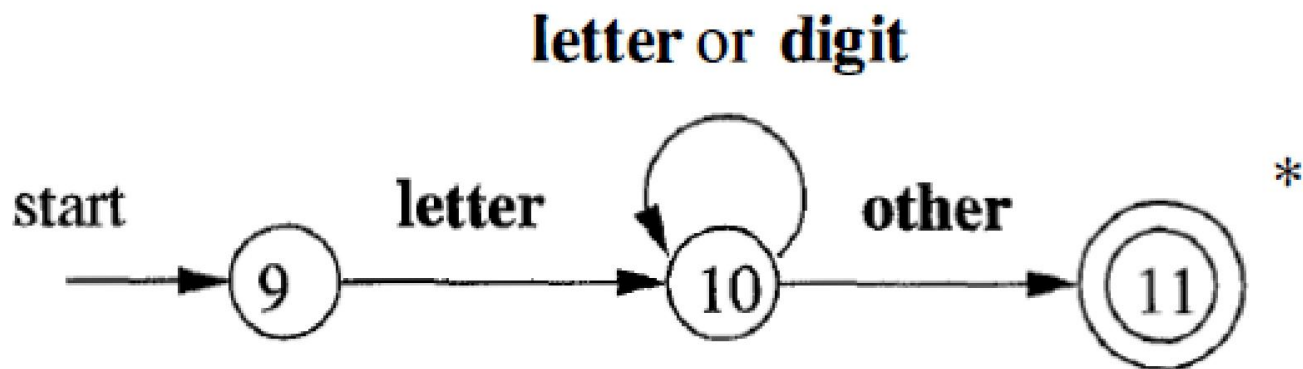
maryam.bashir@cs.uol.edu.pk

Tên của tệp zip phải là số cuộn của tất cả học sinh trong nhóm như sau:

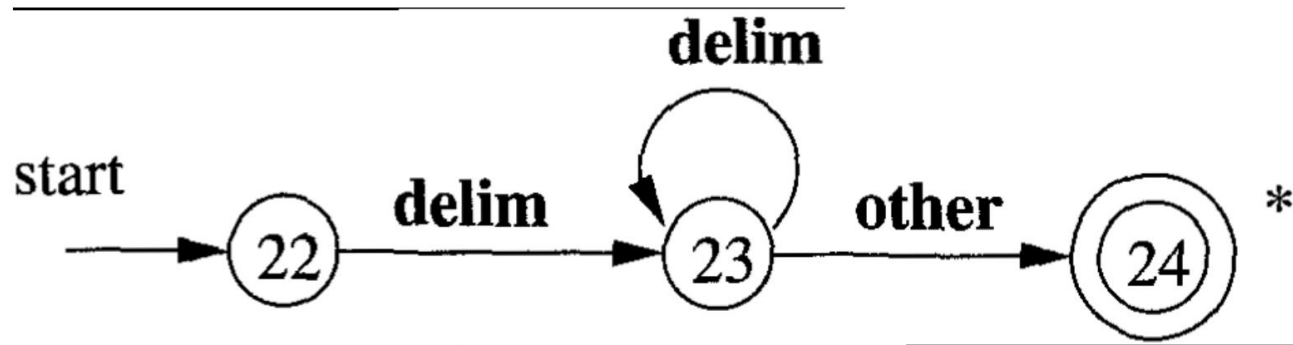
RollNumber1-RollNumber2 Bạn

nên cố gắng thực hiện bài tập này một cách riêng lẻ. Nếu bạn nghĩ rằng chương trình của bạn rất yếu thì có thể làm việc theo nhóm tối đa 2 học sinh.

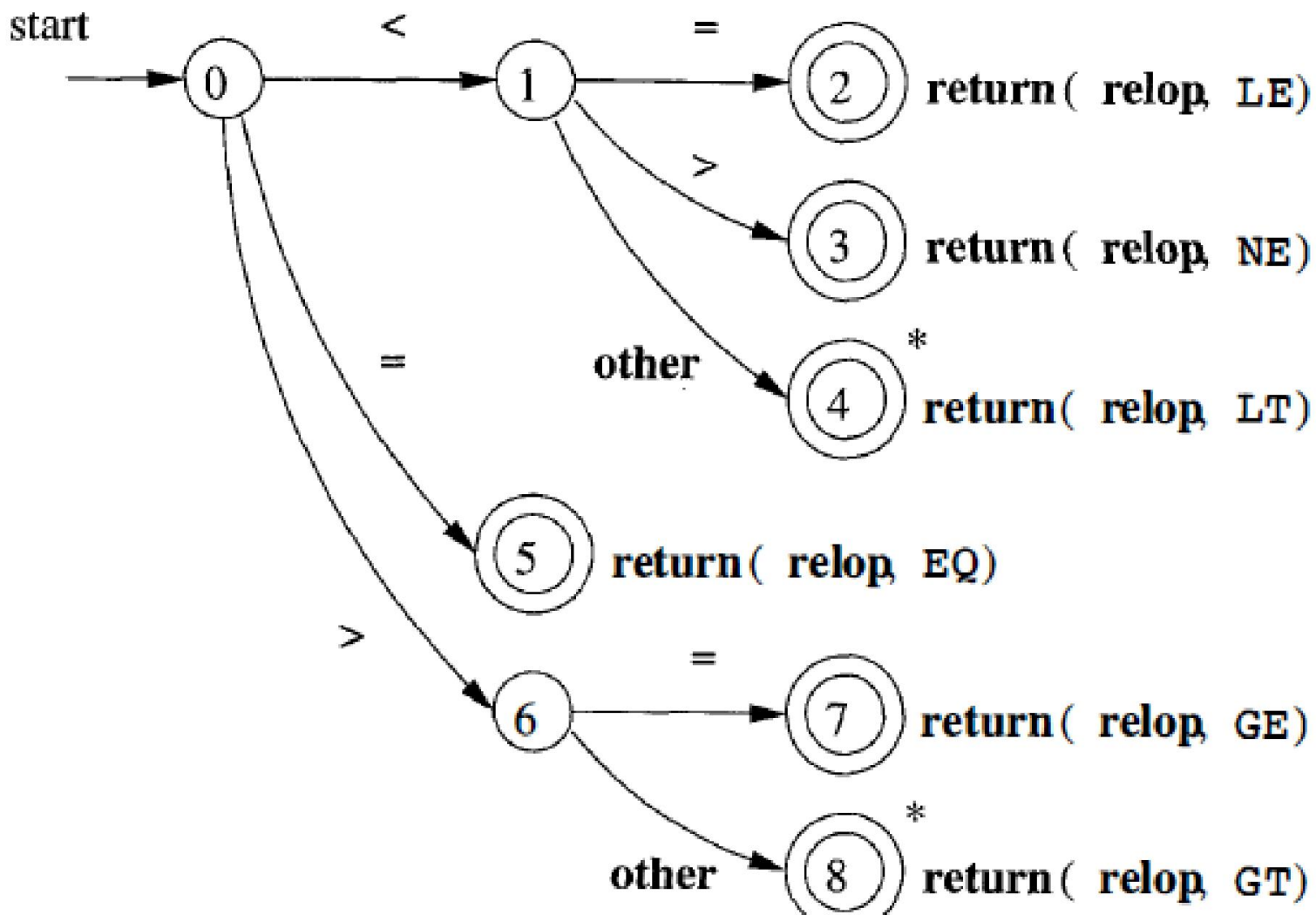
Sơ đồ chuyển tiếp cho số nhận dạng



Sơ đồ chuyển tiếp cho khoảng trắng

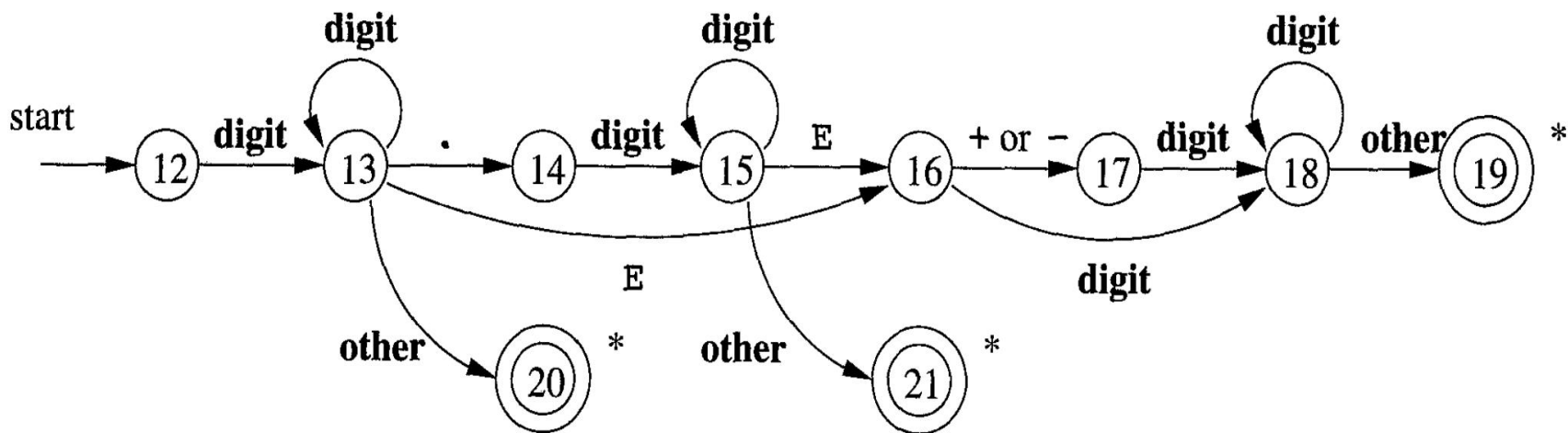


# Sơ đồ chuyển tiếp cho các toán tử quan hệ





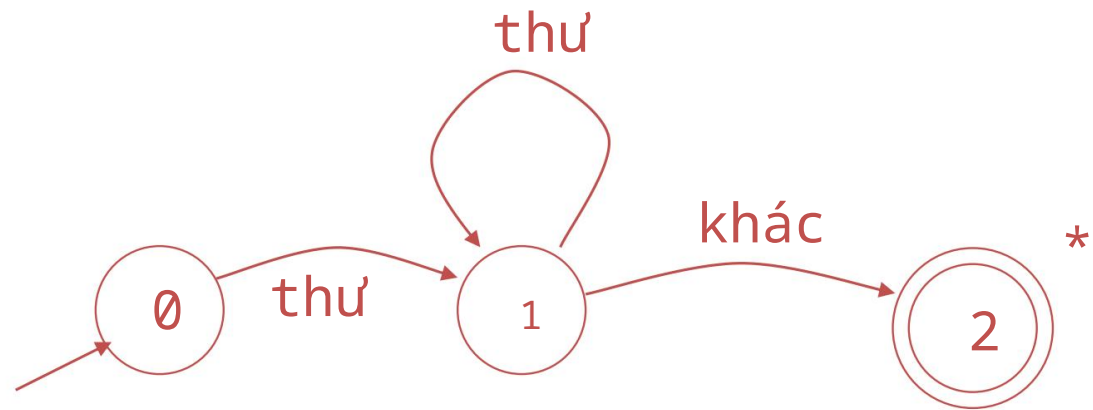
# Sơ đồ chuyển tiếp cho số không dấu



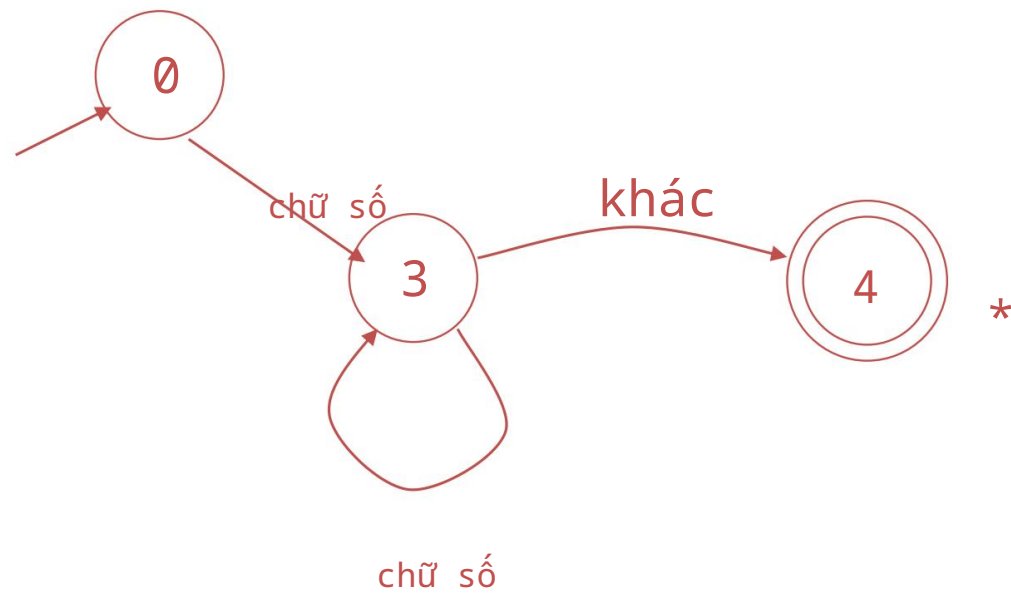
# Cách hợp nhất nhiều chuyển đổi Sơ đồ

- Bước 1: Hợp nhất các trạng thái bắt đầu của tất cả các sơ đồ chuyển tiếp
- Bước 2: Thực hiện chuyển đổi mới trên bất kỳ đầu vào nào từ trạng thái cuối cùng của mỗi sơ đồ chuyển tiếp sang trạng thái bắt đầu mới

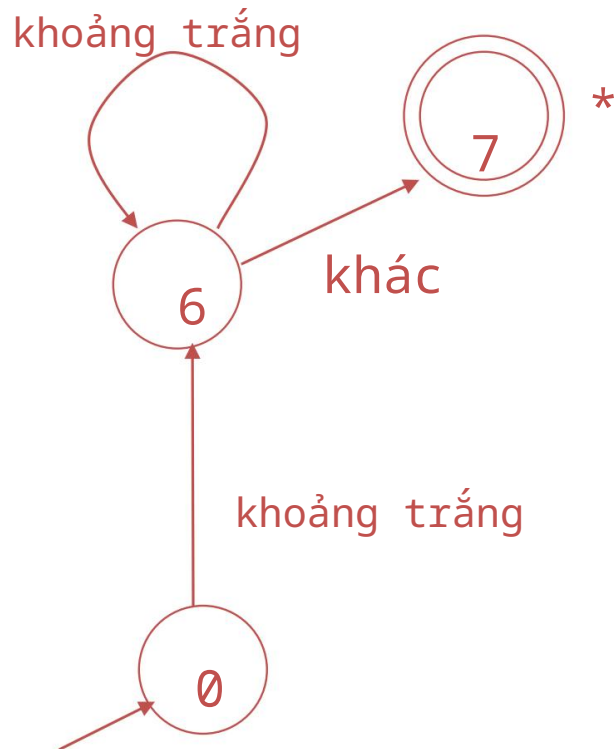
Sơ đồ chuyển tiếp cho các mã định danh chỉ gồm các chữ cái



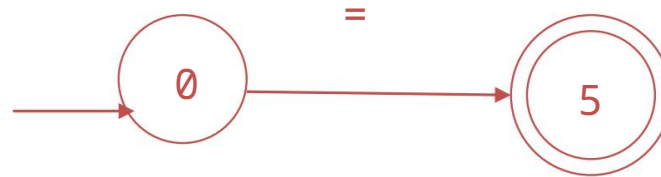
## Sơ đồ chuyển tiếp cho số nguyên



## Sơ đồ chuyển tiếp cho khoảng trắng



Sơ đồ chuyển đổi cho toán tử =



## Sơ đồ chuyển tiếp lỗi



