**CHUẨN HÓA VĂN BẢN TIẾNG VIỆT**

**1. Định nghĩa bài toán**

Trong tổng hợp tiếng nói, cần đưa văn bản về dạng có thể đọc được. Văn bản trong thực tế bao gồm nhiều thành phần đặc biệt: số, ngày tháng, các từ viết tắt, tiền tệ, chuỗi chữ số, từ viết tắt, từ nước ngoài, số la mã, URL, email....

=> Được gọi chung là***NSWs*** -not standard words

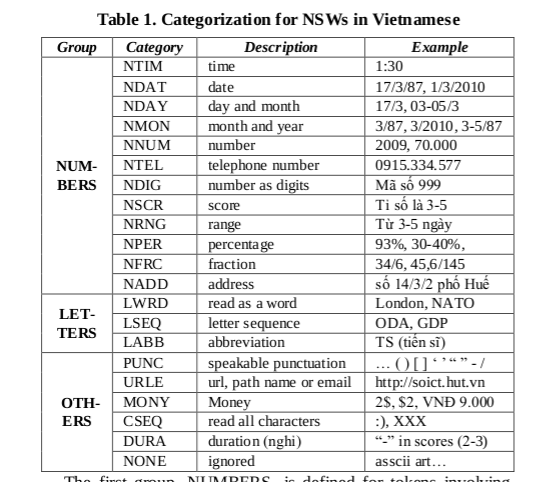
Các NSWs này không thể tìm ra cách phát âm theo luật “letter to sound”.

=> Vì vậy cần chuẩn hóa, đưa các NSWs về dạng đọc được của tiếng Việt.

Ví dụ: “ngày 29/5” => “ngày hai mươi chín tháng năm”

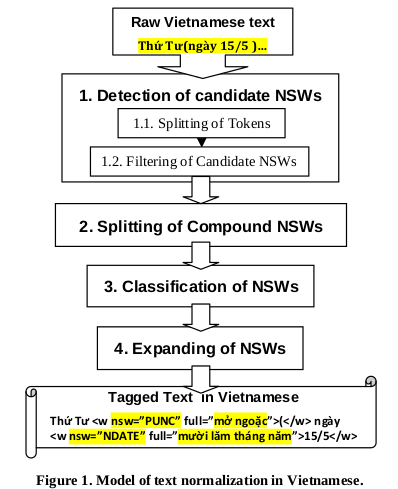
PHUƠNG PHÁP: phân loại và chuẩn hóa từng loại

**2. Các loại NSWs**



Định nghĩa NSWs dạng số được trình bày cụ thể theo đường link sau: https://docs.google.com/spreadsheets/d/1undm9qxE94KXu-OL7ET05bIF6dv1a8YapcyUyjV8n1E/edit?usp=sharing

**3. Mô hình chuẩn hóa**



**3.1. Tách câu - tách từ**

**3.1.1. Tách câu**

Tách câu theo 3 dấu có khả năng là dấu kết thúc của một câu (‘.’, ‘!’ và ‘?’), chú ý xử lý với dấu ‘–’ và ‘-’.

**3.1.2. Tách từ**

*3.1.2.1. Tách mức độ từ*

* Tách riêng các ký tự đặc biệt dạng: α, β,…
* Tách các NSWs phức tạp: các loại NSWs không phân loại được vào bảng 1 (Tách riêng dấu, chữ và số riêng biệt theo từng dạng NSWS - chú ý không tách trong dấu với số để xử lý ở giai đoạn sau)

VD: Mỹ: => Mỹ :

'không' => ' không '

1m65 => 1 m 65

* Chú ý đến URL khi phân tách punctuation

*3.1.2.2. Tách mức độ câu*

* Các loại NSWs có thể là số, số điện thoại, ngày tháng, khoảng, tỷ số, ... và có thể được phân chia bằng các dấu ./-.

VD: “17/3/2010”, “100.000”, “0914.392.492”, “40%”, “50-50”,…

* Ghép các trường hợp số và dấu câu để tổ hợp thành NSWs dạng số phức tạp

VD: 16 , 17/3/2010 => 16,17/3/2010

30 - 40% => 30-40%

7 - 8/45 => 7-8/45

* Chuyển đơn vị tiền tệ về phía sau:

VD: USD 10 => 10USD

**3.2. Phân loại NSWs**

Nếu một từ không khớp với từ điển các âm tiết tiếng việt sẽ được xét là một NSW.

**3.2.1. Phân loại nhóm NUMBERS**

Trong thực tế mỗi token có thể phân thành nhiều loại, ví dụ “3-5” có thể là “ba đến năm” hoặc “ba tháng năm”, .... Vì vậy việc phân loại tùy thuộc vào ngữ cảnh.

Phương pháp sử dụng: kết hợp mô hình phân loại sử dụng machine learning kết hợp với rule-based

*3.2.1.1. Mô hình phân loại sử dụng machine learning*

Mô hình sử dụng phương pháp tiếp cận là squence tagger và mô hình sử dụng là LSTM + CRF (trình bày cụ thể mục dưới)

Mô hình phân loại các NSWs thành 5 dạng nhãn được định nghĩa ở mục 2.

*3.2.1.2. Mô hình phân loại rule-based*

Sau khi NSW được phân loại vào 5 nhãn lớn sẽ dựa vào các đặc điểm của từng số để phân loại chính xác vào từng loại ở bảng 1, cụ thể:

* TIM => NTIM
* DAY => NDAY, NDAT
* MON => NMON
* DIG => NDIG, NTEL, NSCR, NADD
* NUM => NNUM, NPER, NRNG, NFRC, MONY

**3.2.2. Phân loại dạng LETTERS**

Dựa vào đặc điểm của dạng LETTERS có 3 dạng là UPPER (viết hoa toàn bộ), LOWER (viêt thường toàn bộ hoặc chỉ từ đầu tiên viết hoa) và MIX (dạng chứa cả viết hoa và viết thường). Dựa vào rule-based để đọc chính xác từng dạng LETTERS

* UPPER gồm 2 kiểu đọc: là từ viết tắt hoặc đọc lần lượt từng chữ cái (theo tiếng anh hoặc tiếng việt)
* LOWER chỉ xét theo trường hoặc là từ tiếng anh (nếu tồn tại dấu thì kiểu á, é,… => xóa dấu và đọc theo dạng tiếng anh)
* MIX: Tách riêng biệt từng từ theo từ viết hoa là chữ cái đầu tiên

VD: CoNLL => Co N L L

**3.2.3. Phân loại dạng OTHER**

Dựa vào việc tách từ và tách câu chỉ cần chú ý 2 dạng là PUNC và URLE

* PUNC có thể là dạng đọc hoặc là dạng nghỉ
* URLE là dạng tên website, đường link,…

**4. Biểu diễn NSWs**

Sinh ra các từ ngữ đầy đủ cho NSWs đã được phân loại. Mỗi NSW sẽ được thêm thuộc tính full để biểu diễn dạng full text của nó.

* NTIM trở thành số với “giờ” và “phút”, “giây” (1:03 => một giờ ba phút)
* NDAT trở thành số với “tháng”, năm”. (02/03/98 => hai tháng ba năm chín mươi tám)
* NDAY trở thành số với “tháng” (12/03 => mười hai tháng ba)
* NMON trở thành số với “năm” (12/1998 => mười hai năm một nghìn chín trăm chín mươi tám)
* NNUM là số (12 => mười hai)
* NTEL chuỗi các chữ số với các phân cách không được đọc (0386.151.514 => không ba tám sáu một năm một năm một bốn)
* NDIG thành chuỗi các chữ số (3925 => => ba chín hai năm)
* NSCORE (tỉ số là 2-3 => tỉ số là hai ba)
* NRNG ”-” trở thành “đến” (từ 2-3 => từ hai đến ba)
* NPER “%” là phần trăm (30-40% => ba mươi đến bốn mươi phần trăm)
* NFRC phân số (3/4 => => ba phần tư)
* LWRD sử dụng từ (facebook => phây-búc)
* LSEQ đọc từng chữ cái (WTO => vê kép tê ô)
* LABB sử dụng phương pháp mở rộng từ viết tắt (UBND => ủy ban nhân dân)
* PUNC đưa ra nghĩa đầy đủ ( ”...” => “ba chấm”)
* URLE [( van@gmail.com](mailto:(van@gmail.com) => vê a nờ a còng gờ meo chấm com)
* MONY kết hợp số và đơn vị tiền tệ, có thể thêm “/” thành “năm” (1$/người => một đô la mỗi người)
* CESQ đọc từng kí tự “:)” => “hai chấm mở ngoặc)
* DURA và NONE cỏ thể bỏ qua

**MÔ HÌNH PHÂN LOẠI DẠNG NUMBERS**

**1. Phương pháp thực hiện**

Mô hình sử dụng phương pháp sequence tagger với dữ liệu được gán nhãn theo chuẩn BIO

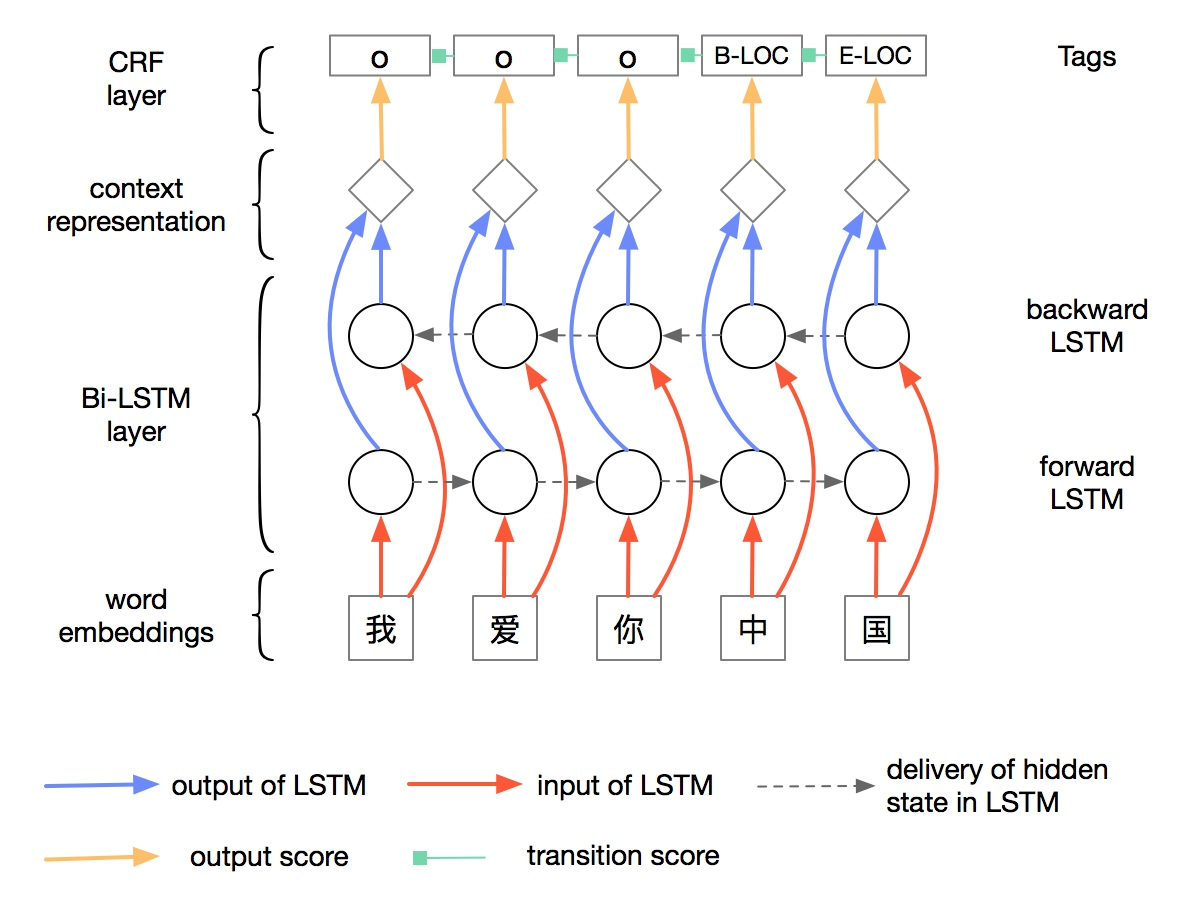
**2. Dữ liệu**

Link: <https://seafile.platform.vn/f/49f5dcd8100c41ef8128/>

**3. Mô hình**

Mô hình sử dụng bao gồm:

* Mô hình embedding sử dụng fasttext
* Mô hình phân loại LSTM-CRF

****

**PHIÊN ÂM TỪ TIẾNG ANH**

**1. Phương pháp thực hiện**

Với các từ tiếng Anh trong văn bản tiếng Việt, cụ thể là trong các bài báo, cần chuyển chúng về dạng đọc được của tiếng Việt. Các từ tiếng Anh sẽ được phân vào nhóm LWRD (đọc như một từ).

Phương pháp đề xuất:

* Xây dựng từ điển tương ứng giữa từ tiếng Anh và cách đọc tiếng Việt (1) cho một số từ thông dụng , kết hợp sử dụng từ điển đọc gần giống (2) để xử lý những từ ít phổ biến hơn.
* Với từ điển 1 sẽ có cấu trúc bao gồm nhiều dòng, mỗi dòng là một cặp từ tiếng Anh: cách đọc tiếng Việt.

Ví dụ: NATO : na tô

Paris : pa ri

London: Luân đôn

Napoleon : Na pô lê ông

* Tuy nhiên từ điển 1 sẽ không phủ được toàn bộ các từ tiếng Anh với cách đọc thường dùng. Vì vậy chúng ta cần sử dụng từ điển 2 với luật sinh cách đọc.

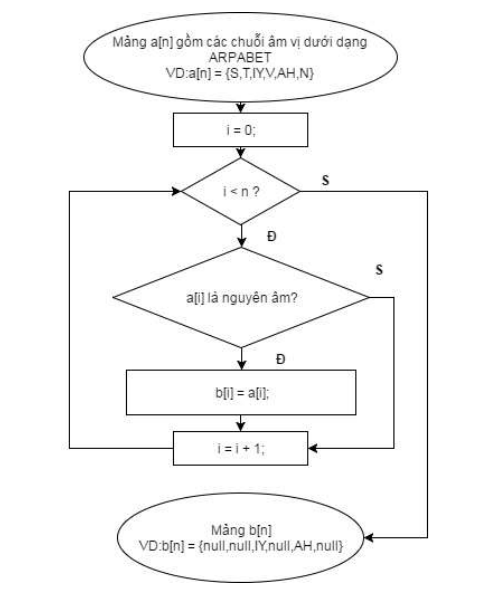
Chúng ta sử dụng từ điển CMU ([CMU pronouncing dictionary](http://www.speech.cs.cmu.edu/cgi-bin/cmudict)) để đưa ra được phiên âm tiếng Anh, sau đó chuyển về cách đọc tiếng Việt gần giống nhất. (có thể sử dụng thư viện pronouncing của python).

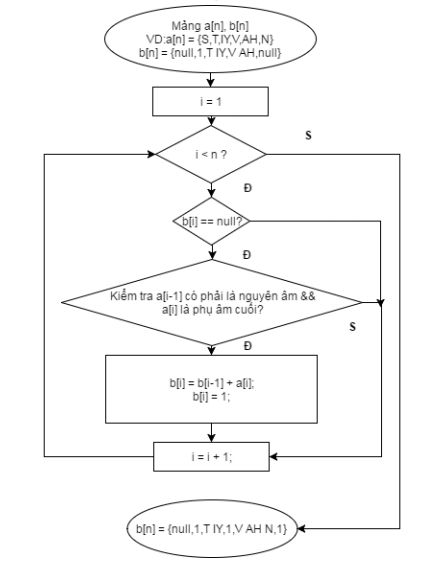
**2. Phương pháp xây dựng phiên âm**

Với các từ tiếng anh không được phiên âm từ từ điển pronouncing sẽ sử dụng rule-based theo phiên âm từ tiếng latin rồi được tham chiếu tương ứng sang phiên âm ARPABET

Thuật toán gồm 3 bước đơn giản:

* Xác định nguyên âm

****

* Thêm các phụ âm vào trước hoặc sau của nguyên âm đã xác định (chú ý xử lý các phụ âm có khả năng đứng độc lập như âm C khi đứng trước R: crawl => cờ-ro)
* Tham chiếu qua bộ từ điển để chuyển sang phiên âm tiếng việt