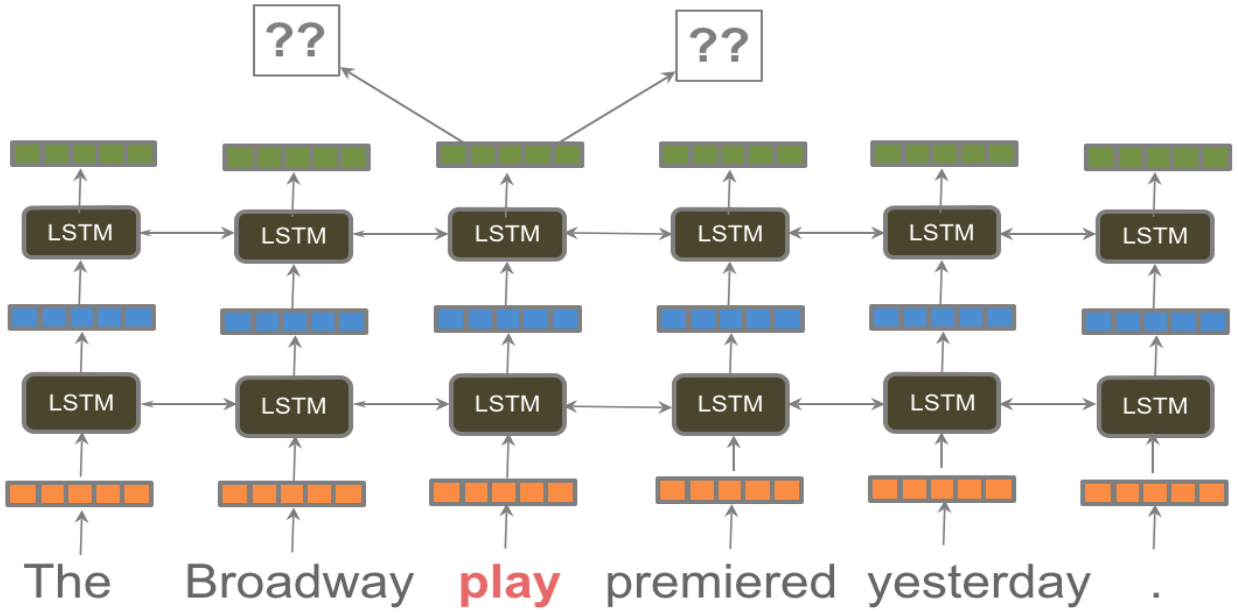
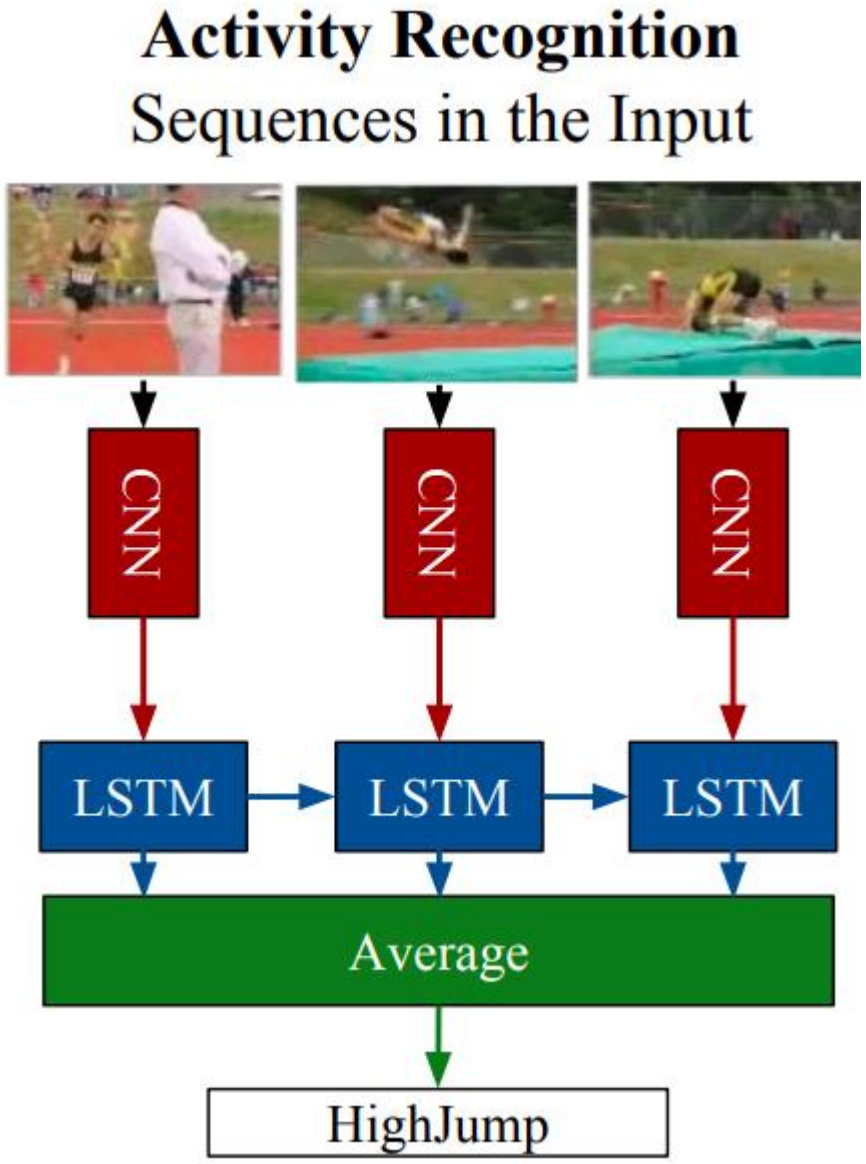
RNN

Recurrent neural network

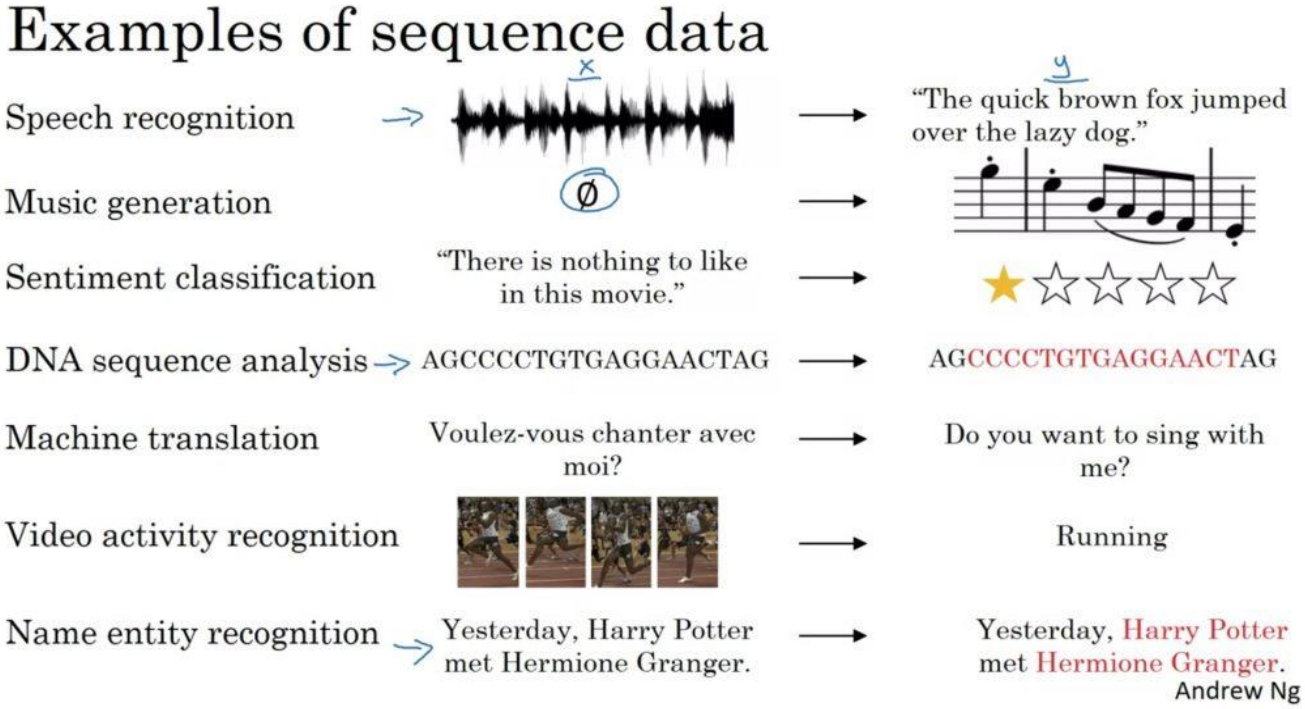
- cho bài toán dữ liệu dạng chuỗi (sequence).
- video understanding
- Natural language processing

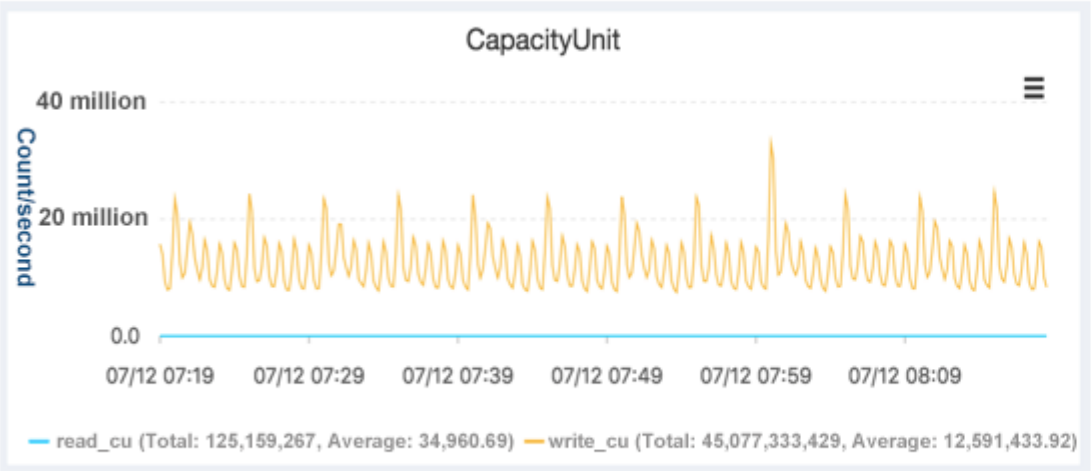




Sequence data

time-series data.





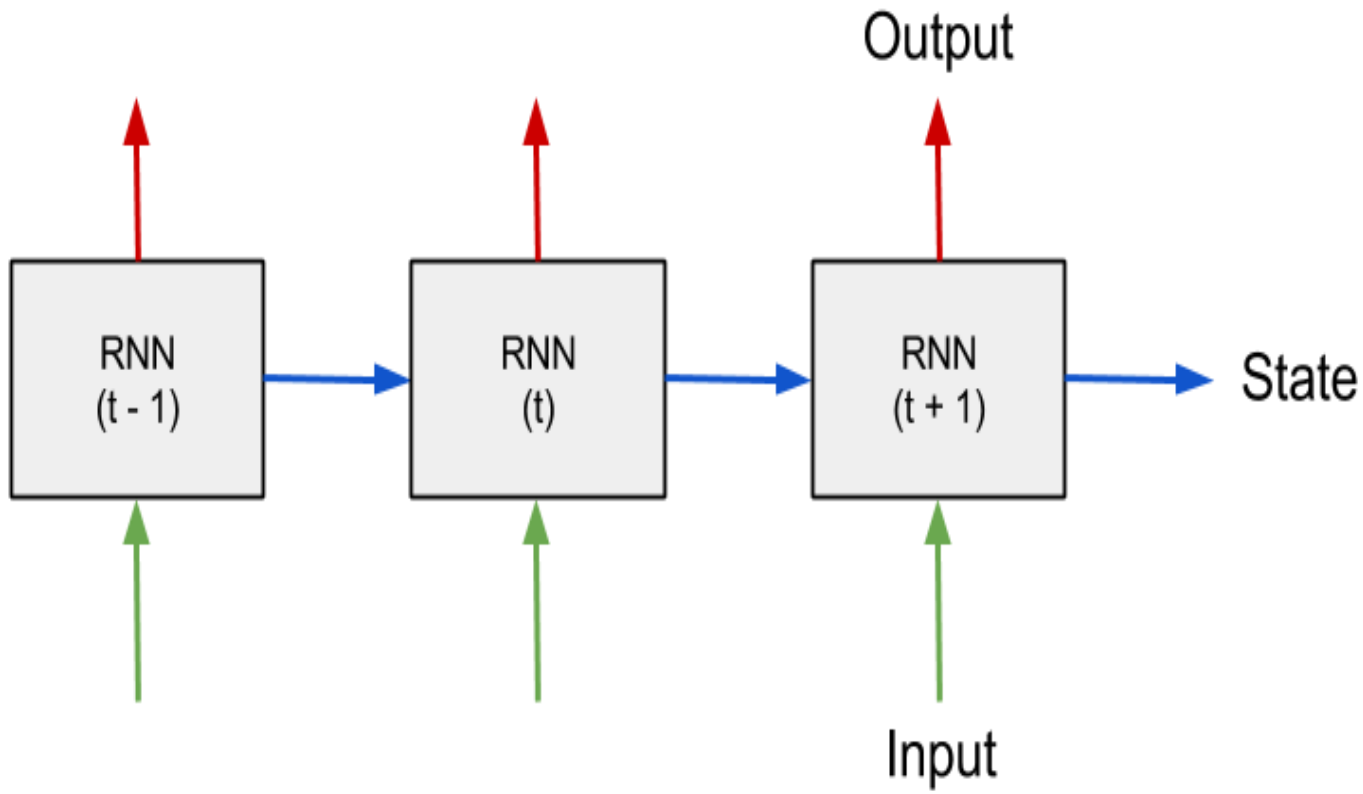
Monitoring time series data

Order No.	Time	Location	Event
77165205838	2018-07-10 10:00	Shanghai	Ship
77165205838	2018-07-11 12:00	Yuhang	Arrived in Hangzhou
77165205838	2018-07-11 14:00	Zhuantang	Start delivery and arrive at Apsara Park
77165205838	2018-07-11 16:00	Zhuantang	Accepted

Status time series data

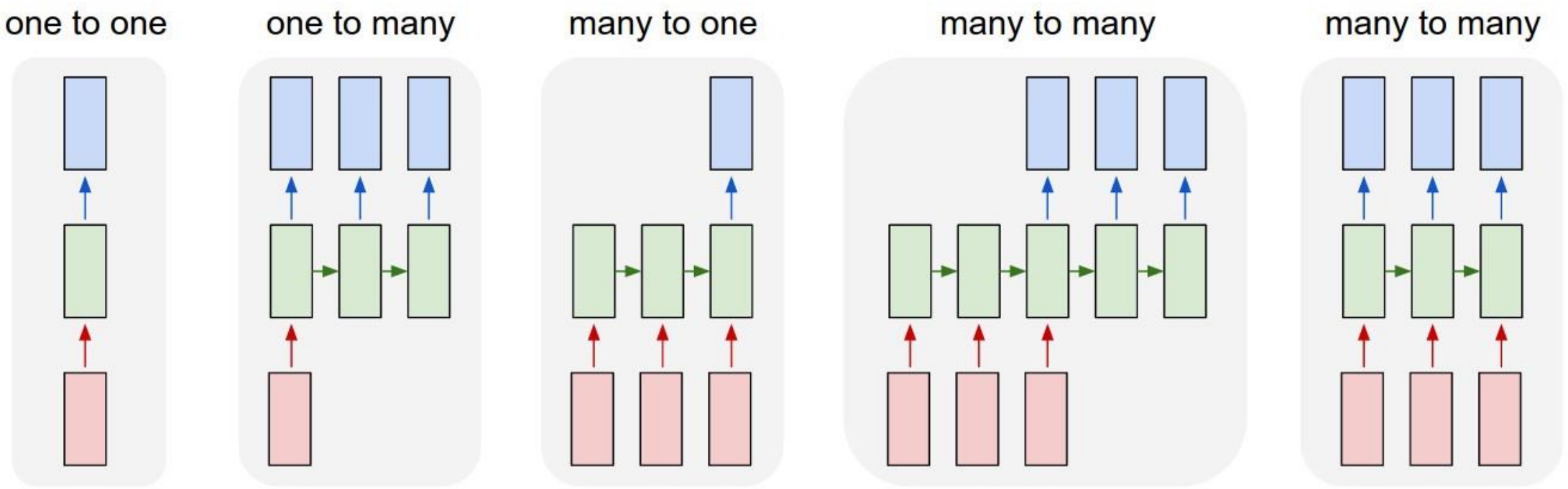
Recurrent neural network

Recurrent Neural Networks (RNNs) are a kind of neural network that specialize in processing **sequences**.



Phân loại bài toán RNN

CNN: **fixed-size inputs and fixed-size outputs**.
RNNs are useful because they let us have **variable-length sequences** as both inputs and outputs.



Process Sequences

- **One-to-one:** This is the classic feed forward neural network architecture, with one input and we expect one output.
- **One-to-many:** image captioning, one image as a fixed size input and the output can be words or sentences which are variable in length.

"man in black shirt is playing guitar."

"construction worker in orange safety vest is working on road."

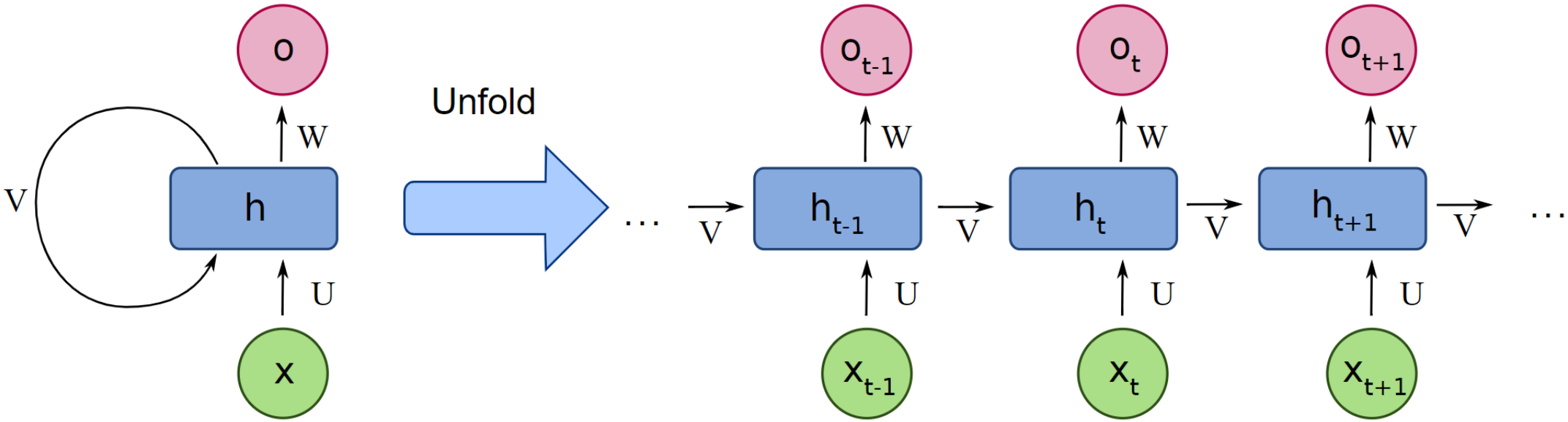
"two young girls are playing with lego toy."

Process Sequences

- **Many-to-one:** This is used for sentiment classification. The input is expected to be a sequence of words or even paragraphs of words. The output can be a regression output with continuous values which represent the likelihood of having a positive sentiment.

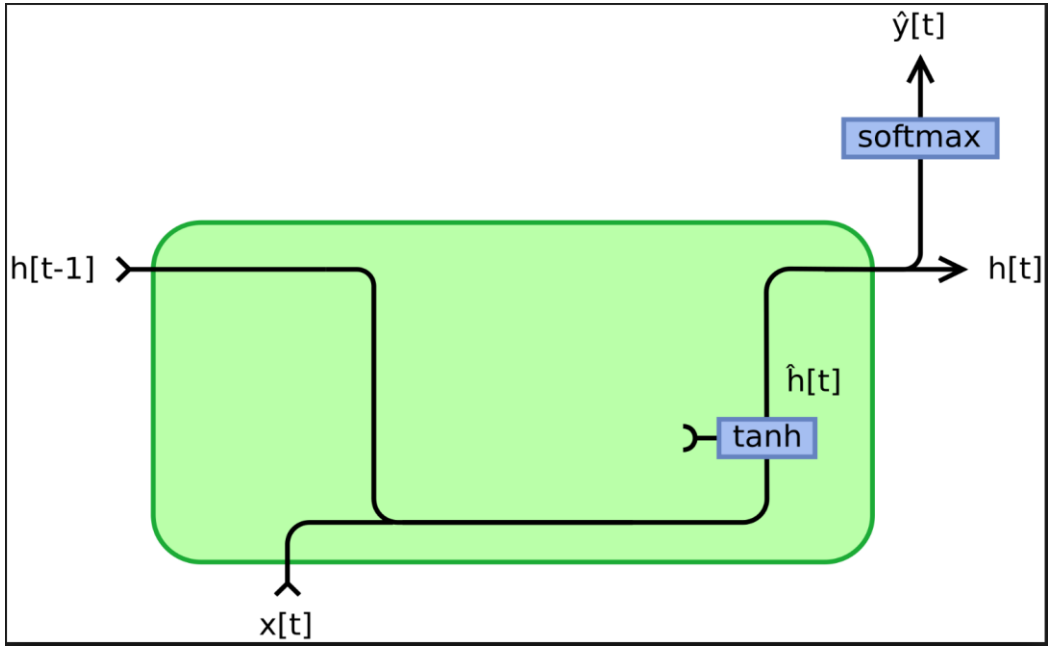
- Many-to-many: his model is ideal for machine translation like the one we see on Google translate. The input could an English sentence which has variable length and the output will be the same sentence in a different language which also has variable length.

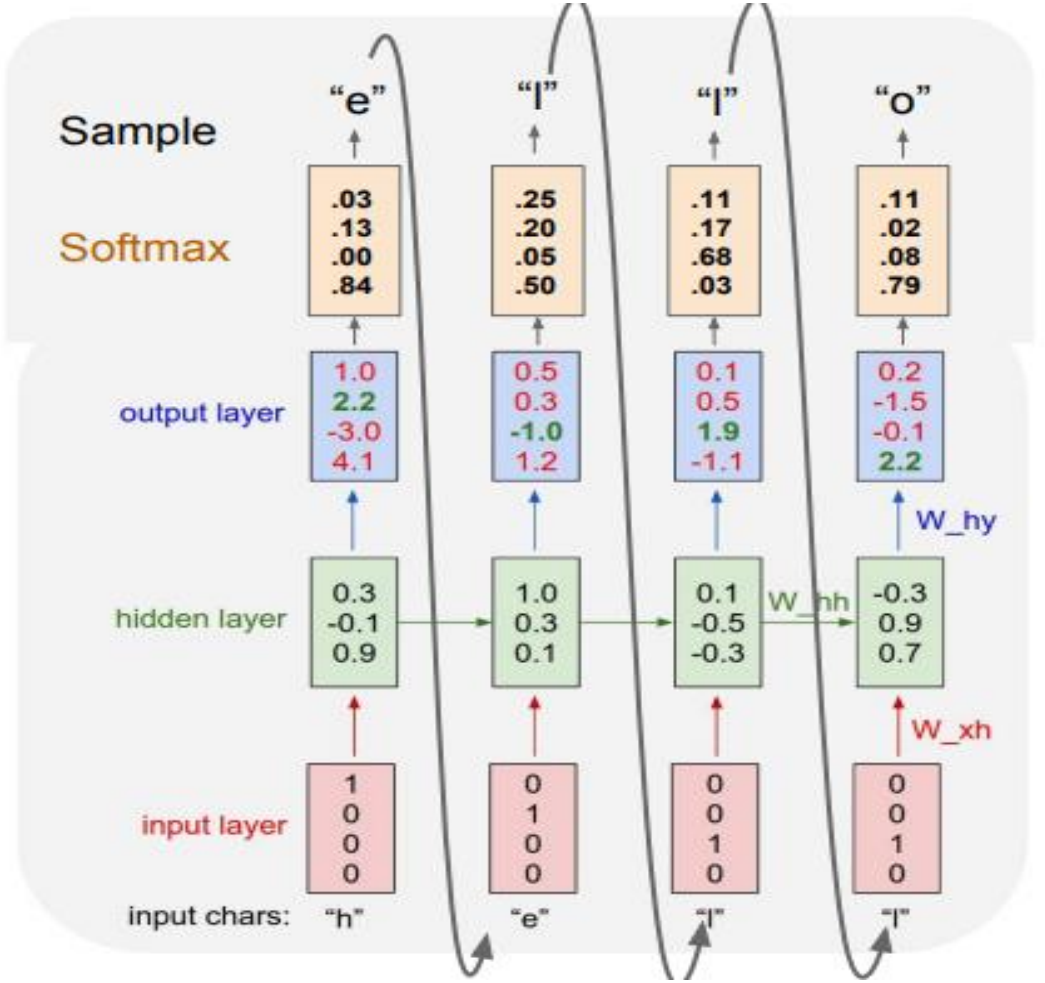
RNN model

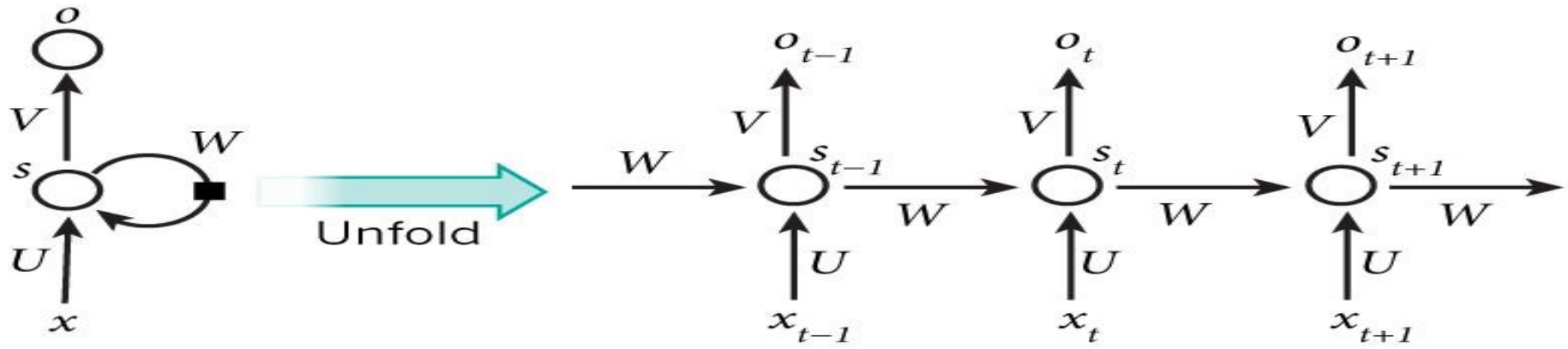


x: The input. It can be a word in a sentence or some other type of sequential data
O: The output. For instance, what the network thinks the next word on a sentence should be given the previous words
h: The main block of the RNN. It contains the weights and the activation functions of the network
V: Represents the communication from one time-step to the other.

RNN model


ŷ[t]
softmax
h[t-1]
h[t]
ĥ[t]
tanh
x[t]


Sample
Softmax
"e"
"l"
"l"
"o"
.03 .13 .00 .84
.25 .20 .05 .50
.11 .17 .68 .03
.11 .02 .08 .79
output layer
1.0 2.2 -3.0 4.1
0.5 0.3 -1.0 1.2
0.1 0.5 1.9 -1.1
0.2 -1.5 -0.1 2.2
W_hy
hidden layer
0.3 -0.1 0.9
1.0 0.3 0.1
0.1 -0.5 -0.3
-0.3 0.9 0.7
W_hh
W_xh
input layer
1 0 0 0
0 1 0 0
0 0 1 0
0 0 1 0
input chars:
"h"
"e"
"l"
"l"



$$s_t = \tanh(Ux_t + Ws_{t-1})$$

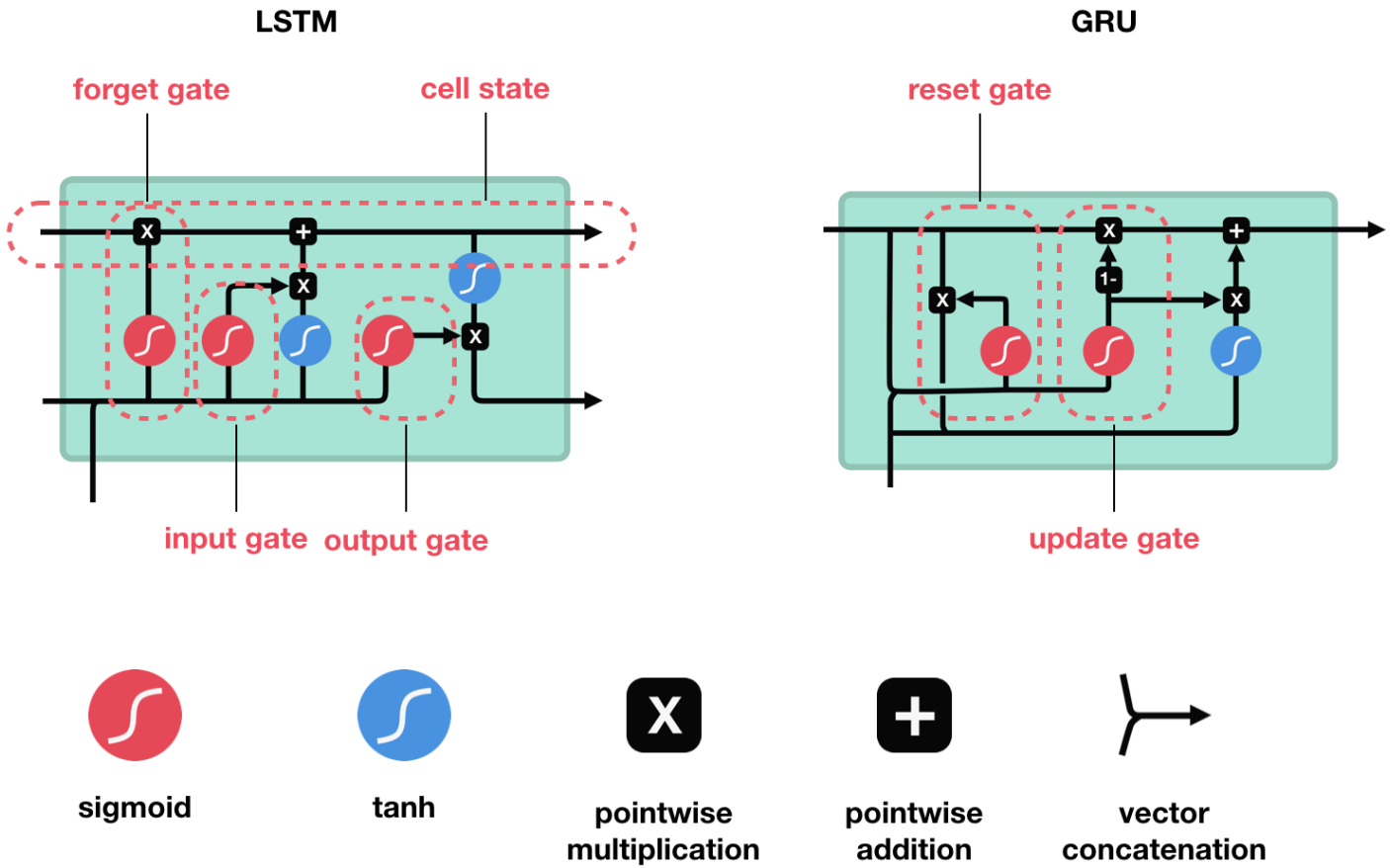
$$o_t = \mathrm{softmax}(Vs_t)$$

Cross-Entropy Loss Function :
categorical cross entropy loss

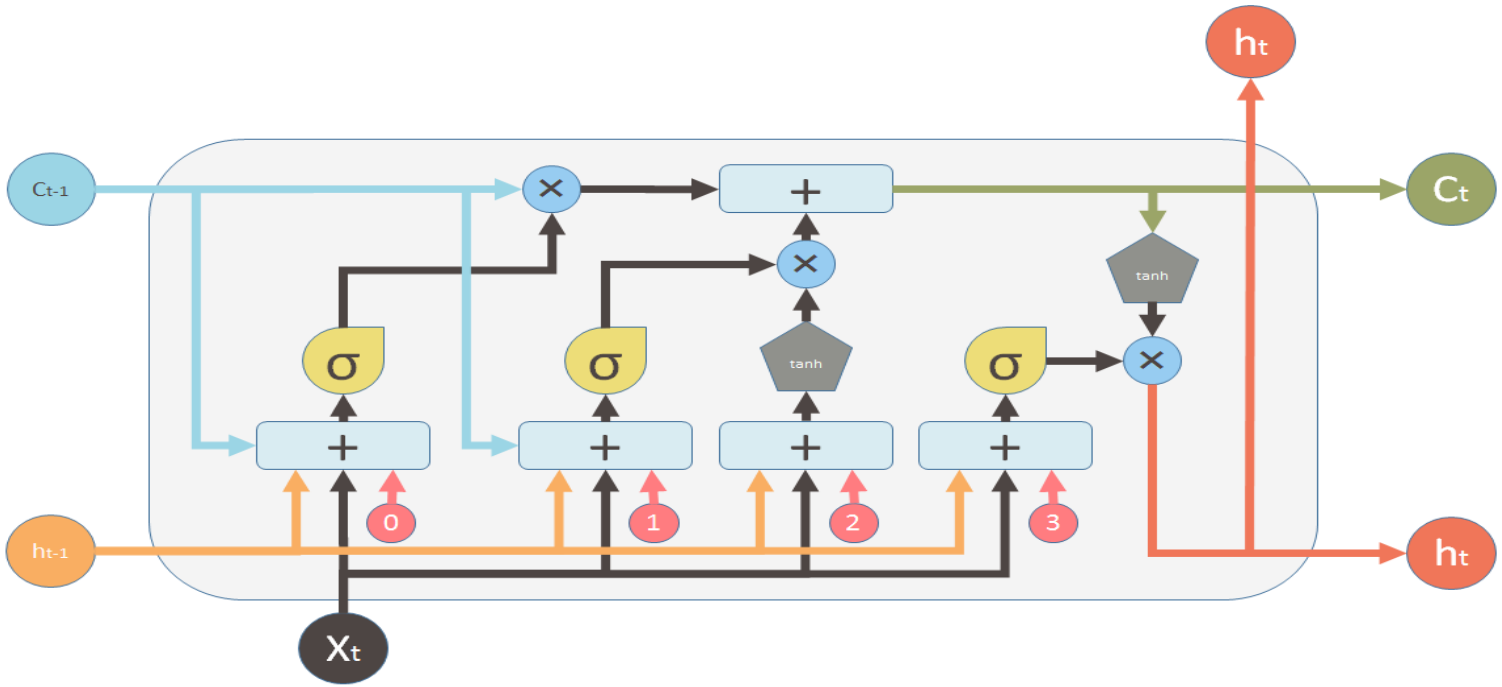
$$L(y, o) = -\frac{1}{N} \sum y_n \log o_n$$

RNN model

- Kiến trúc khá đơn giản nên khả năng liên kết các thành phần có khoảng cách xa trong câu không tốt do gradient bị thấp dần trong quá trình học(vanishing gradient)
- Không có cơ chế lọc những thông tin không cần thiết => Bộ nhớ của kiến trúc có hạn, nếu lưu tất cả những chi tiết không cần thiết thì sẽ dẫn đến quá tải
- Các kiến trúc để khắc phục các nhược điểm của RNN: LSTM và GRU



Mỗi module của 2 kiến trúc trên đều có trang bị các cổng (gate), giúp kiến trúc đánh giá được mức độ quan trọng của thông tin, từ đó đưa ra quyết định giữ lại hay bỏ đi.



Inputs:

- X_t Input vector
- C_{t-1} Memory from previous block
- h_{t-1} Output of previous block

outputs:

- C_t Memory from current block
- h_t Output of current block

Nonlinearities:

- σ Sigmoid
- tanh Hyperbolic tangent

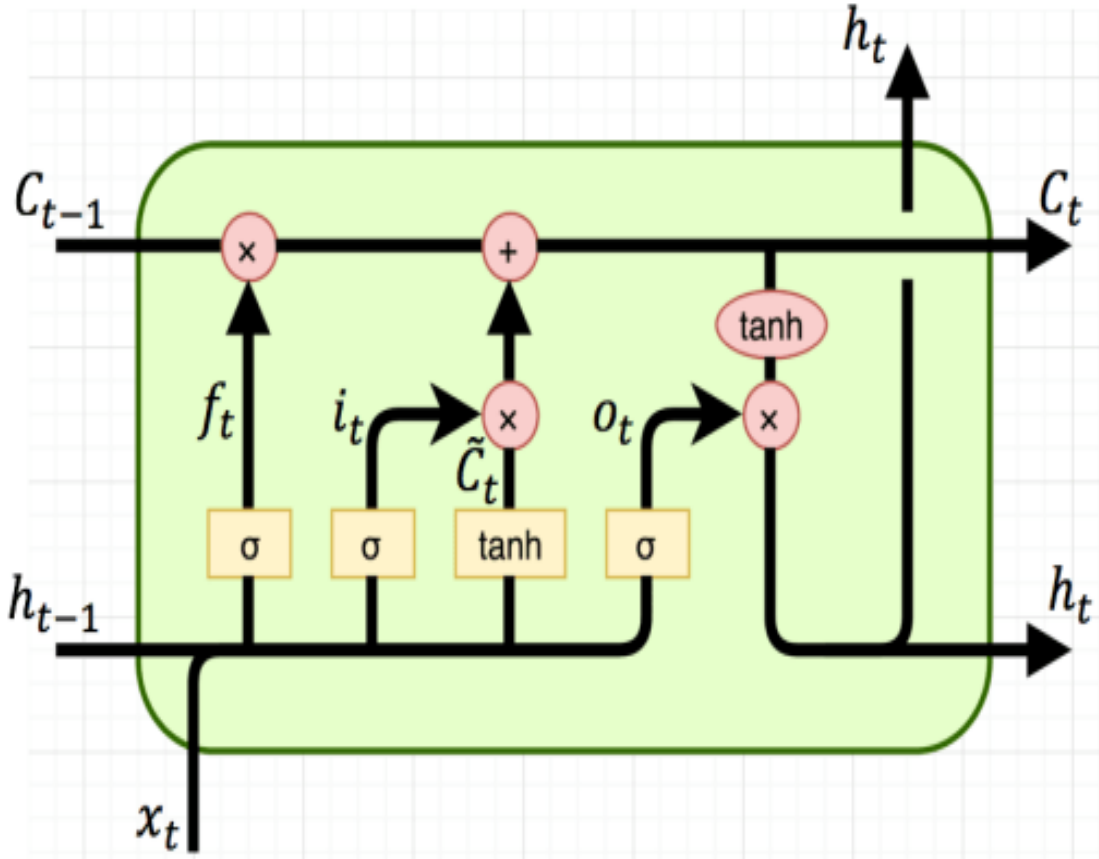
Bias: 0

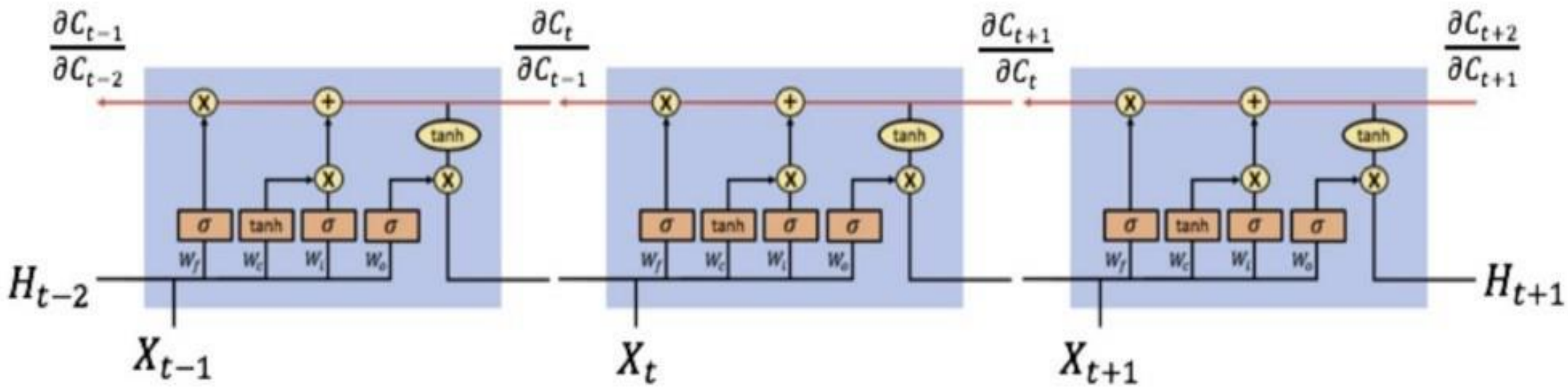
Vector operations:

- × Element-wise multiplication
- \+ Element-wise Summation / Concatenation

LSTM

Output: c_t, h_t, c cell state, h là hidden state.
Input: c_{t-1}, h_{t-1}, x_t, trong đó: x_t - input ở state thứ t , c_{t-1}, h_{t-1} - output của layer trước.
f_t , i_t , o_t tương ứng với **f**orget gate, **i**nput gate và **o**utput gate.
Forget gate: $f_t = \sigma(U_f * x_t + W_f * h_{t-1} + b_f)$
Input gate: $i_t = \sigma(U_i * x_t + W_i * h_{t-1} + b_i)$
Output gate: $o_t = \sigma(U_o * x_t + W_o * h_{t-1} + b_o)$
$0 < f_t, i_t, o_t < 1$; W, U - ma trận trọng số,
b_f, b_i, b_o là các hệ số bias.
$\tilde{c}_t = \tanh(U_c * x_t + W_c * h_{t-1} + b_c)$
$c_t = (f_t * c_{t-1} + i_t * \tilde{c}_t)$
forget gate quyết định xem cần lấy bao nhiêu từ cell state trước và **input gate** sẽ quyết định lấy bao nhiêu từ input của state và hidden layer của layer trước.
$h_t = o_t * \tanh(c_t)$
$h_t, \tilde{c}_t$ khá giống với RNN, nên model có **short term memory**. thông tin nào cần quan trọng và dùng ở sau sẽ được gửi vào và dùng khi cần => có thể mang thông tin từ đi xa=> **long term memory**.



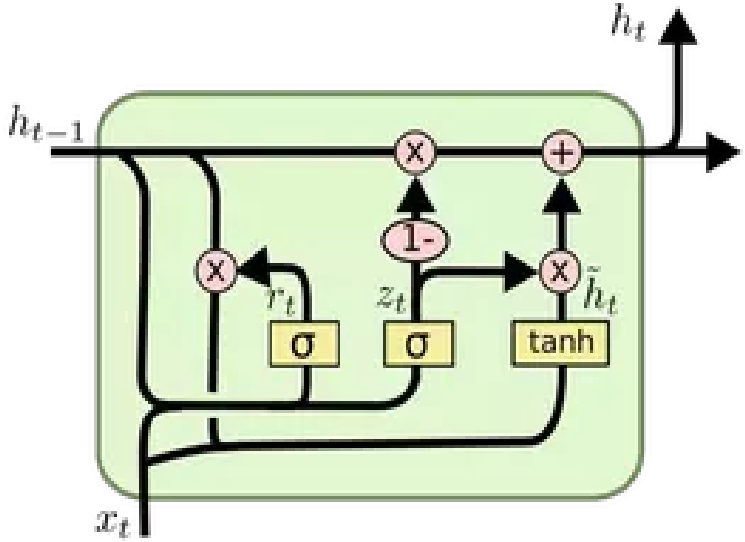


$\frac{\partial c_t}{\partial c_{t-1}} = f_t$, Do $0 < f_t < 1$, LSTM vẫn bị vanishing gradient nhưng bị ít hơn so với RNN.

khi mang thông tin trên cell state thì ít khi cần phải quên giá trị cell cũ, nên $f_t \approx 1$ => **Tránh được vanishing gradient**.

GRU

- GRU là một phiên bản của LSTM với nguyên tắc hoạt động tương tự như LSTM nhưng có cấu tạo đơn giản hơn.



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$

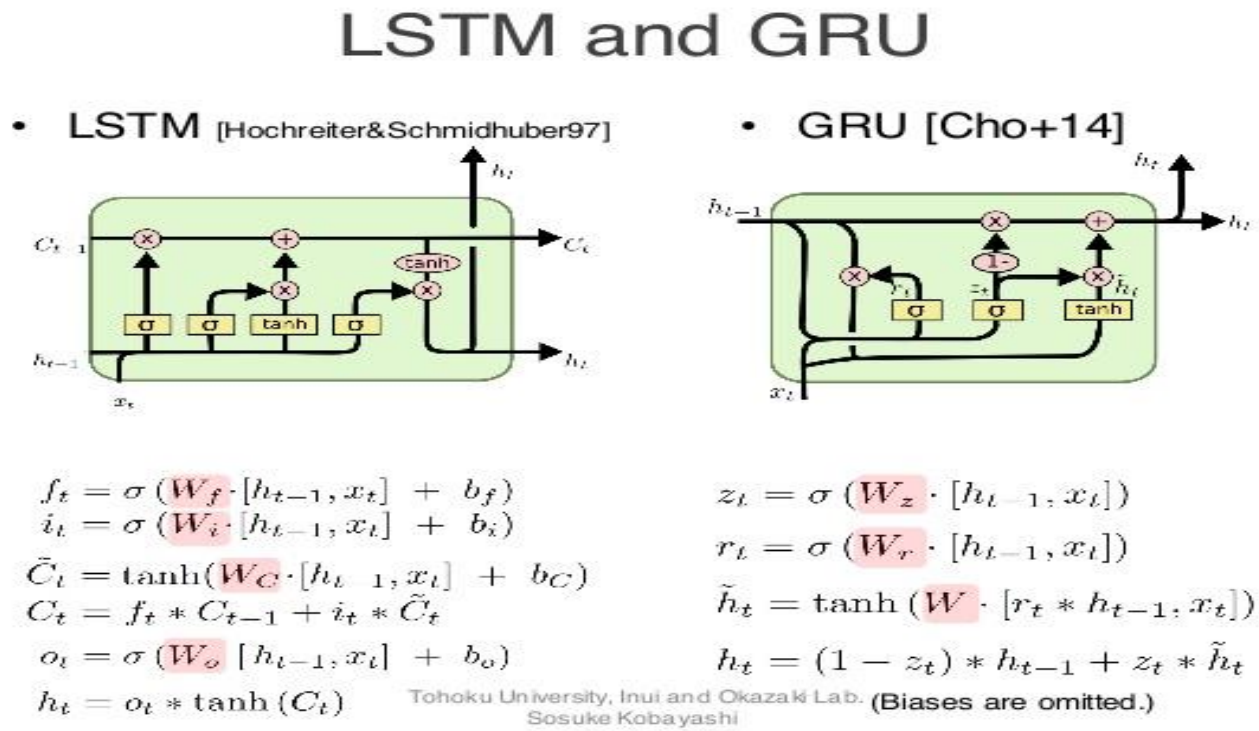
$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$

$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$

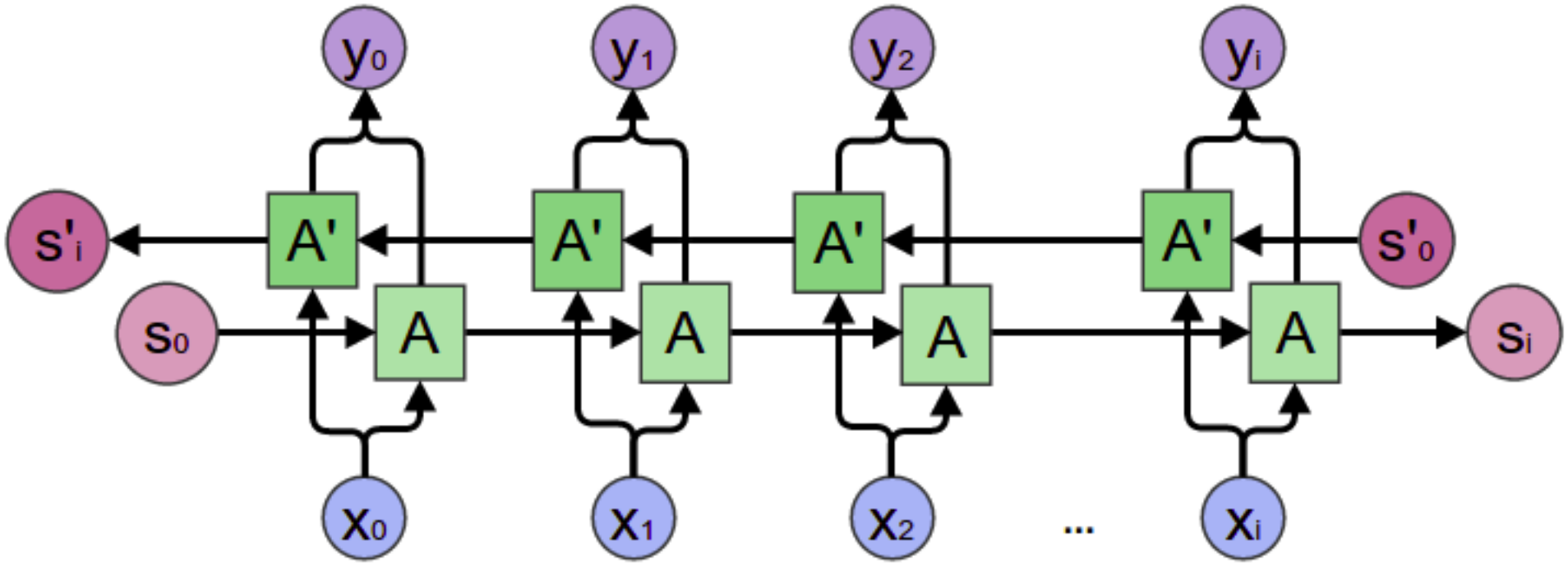
$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

LSTM and GRU

- LSTM có thể lưu trữ thông tin với dữ liệu dài hơn so với GRU.
- Do cấu tạo đơn giản của mình, GRU thường xử lý nhanh hơn LSTM và có thể dễ dàng sử dụng để xây dựng các mạng có cấu trúc phức tạp.
- Hoạt động theo một chiều nhất định (forward direction)=> chỉ mang thông tin tính tới thời điểm hiện tại.
- Trong bài toán NLP thì việc biết thông tin của các timesteps tiếp theo giúp cải thiện rất nhiều kết quả output (Translation, Speech recognition, Handwritten recognition,..)



Bi-directional RNN



Một số bài toán Xử lý ngôn ngữ tự nhiên

- Answer questions using the Web
- Translate documents from one language to another
- Help make informed decisions
- Follow directions given by any user
- Fix your spelling or grammar
- Do library research; summarize
- Manage messages intelligently
- Grade exams
- Write poems or novels
- Listen and give advice
- Estimate public opinion
- Read everything and make predictions
- Interactively help people learn
- Help disabled people
- Help refugees/disaster vicitms
- Document or reinvigorate indigenous languages

Một số bài toán Xử lý ngôn ngữ tự nhiên

Rút trích thông tin văn bản(Information extraction):

- Web mining: rút trích tên người nổi tiếng, sản phẩm đang hot, so sánh giá sản phẩm, nghiên cứu đối thủ cạnh tranh, phân tích tâm lý khách hàng
- Biomedical, Business intelligent, Financial professional : đánh giá thị trường từ các nguồn khác nhau: giá xăng dầu tăng giảm, thông tin chiến tranh, chính trị giữa các nước,
- Terrism event:sử dụng vũ khí gì, đối tượng tấn công là ai

Bài toán:

Rút trích tên thực thể (**Named entity recognition** – NER: people, organization, location)

Rút trích quan hệ giữa hai thực thể (**Relation extraction** – RE: founderOf, headQuarteredIn).

Một số bài toán Xử lý ngôn ngữ tự nhiên

Gán nhãn từ loại (Part-of-Speech tagging POS):

Con ruồi đậu mâm xôi đậu

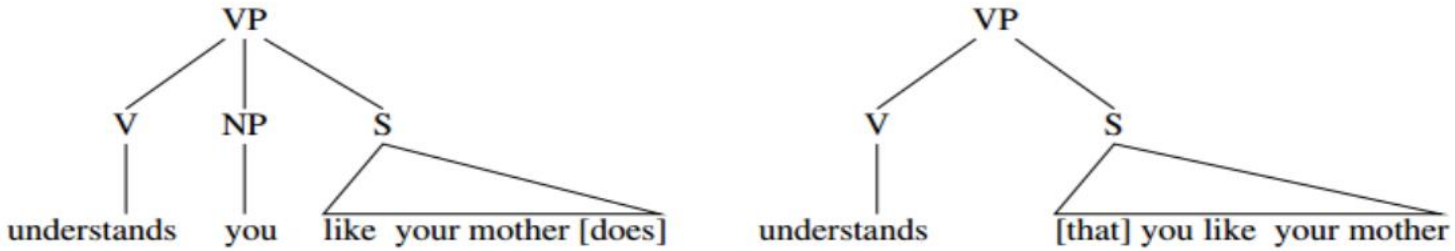
Với : D: determinator (định từ), N: noun (danh từ), V: verb (động từ). Ta có các cặp tương ứng: *Con/D ruồi/N đậu/V mâm/N xôi/N đậu/N*

Bài toán:

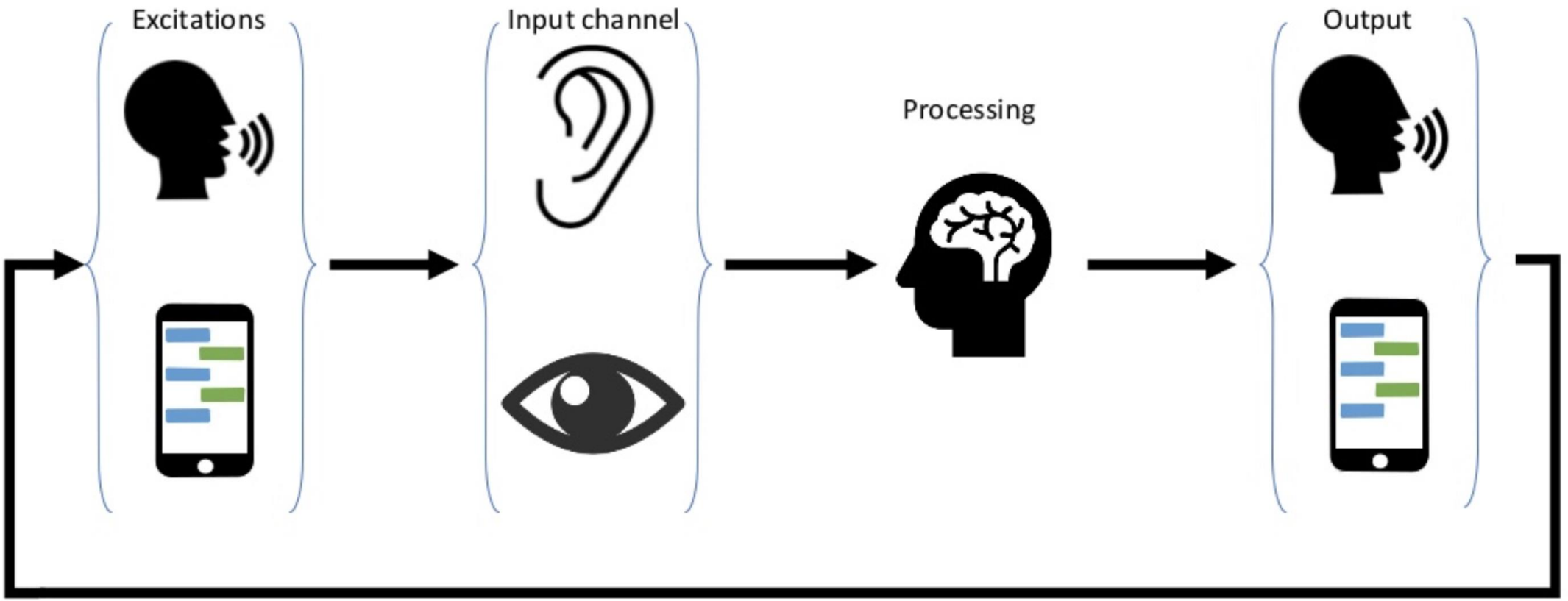
- Named-Entity recognition (gán nhãn tên thực thể).
  - *bà ba [CON NGUOI] bán bánh mì [THUC PHAM] ở phường mười ba [DIA DIEM].*
- Machine translation (dịch máy)
- Speech recognition (nhận diện tiếng nói).

Những vấn đề trong xử lý ngôn ngữ tự nhiên

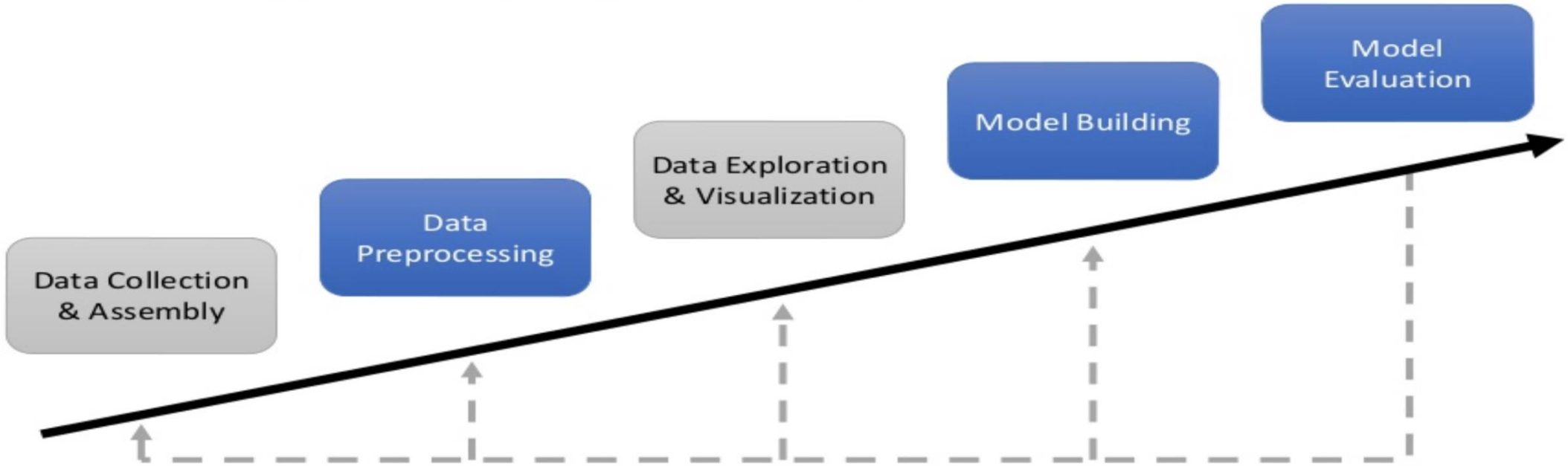
- Sự nhập nhằng trong ngôn ngữ
- Ngôn ngữ tự nhiên sử dụng ngữ cảnh một cách phức tạp và tinh tế để truyền đạt ý nghĩa.
- Ngôn ngữ tự nhiên thường gây nhầm lẫn.
- Ngôn ngữ tự nhiên liên quan tới suy luận về thế giới.
- Ngôn ngữ tự nhiên là một phần quan trọng trong việc tương tác giữa con người với nhau (một hệ thống mang tính xã hội).
- Ví dụ:
  - Ông già đi nhanh quá
  - They book that hotel. They read that book.
  - A computer understands you like your mother

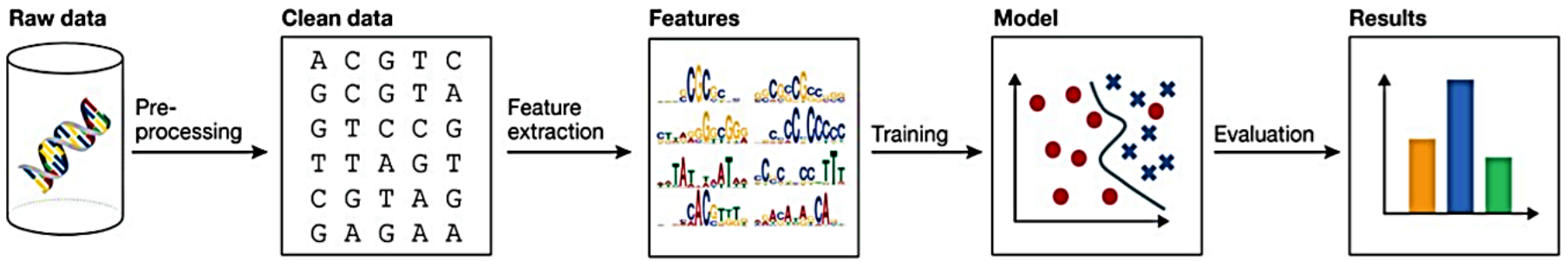


Bài toán Xử Lý Ngôn Ngữ Tự Nhiên



Bài toán Xử Lý Ngôn Ngữ Tư Nhiên



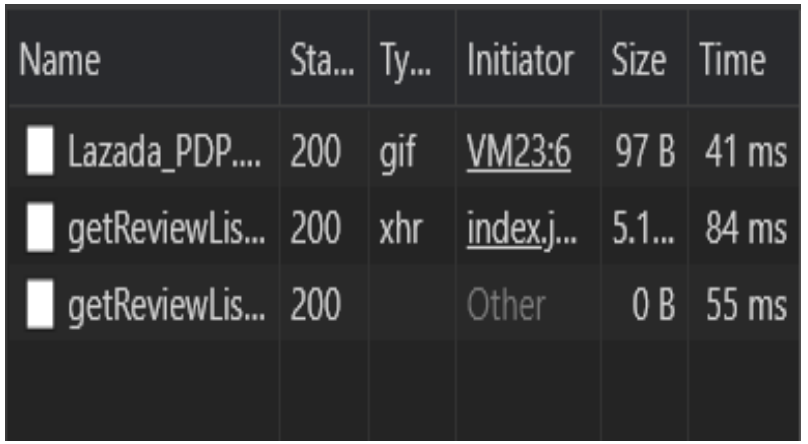


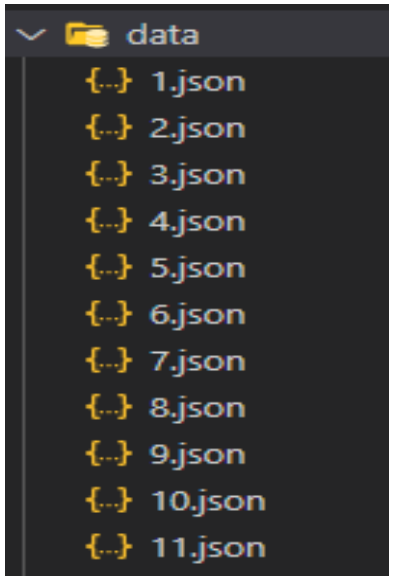
Thu thập dữ liệu

- Thu thập từ nhiều nguồn khác nhau
- Nguồn dữ liệu tin cậy
- Dữ liệu nào thực sự quan trọng cho mô hình??

=> Quá trình tốn nhiều thời gian, công sức

Thu thập dữ liệu





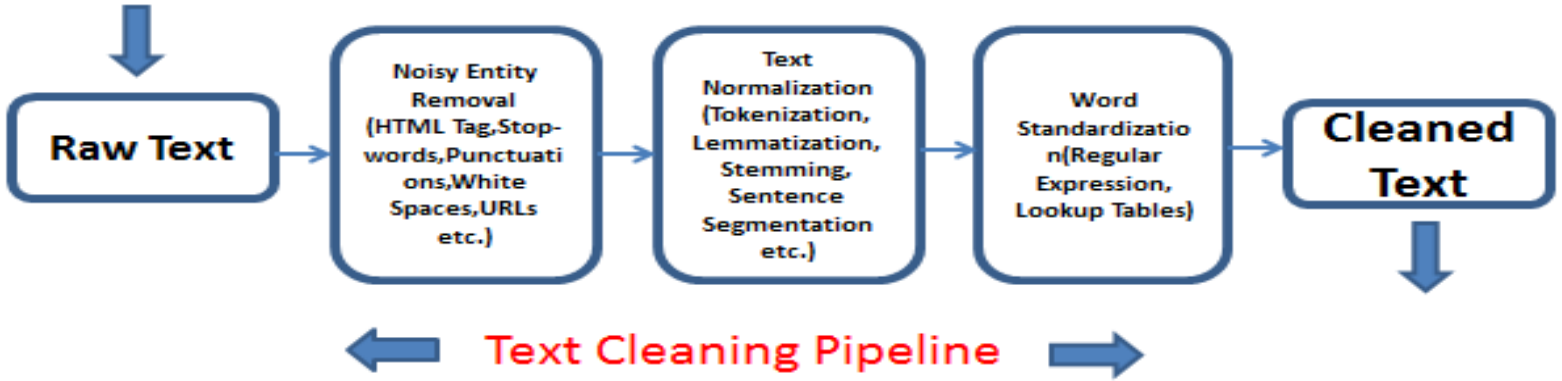
1	bt tai nghe một bên nghe được một bên không bình thường nghe tạm ổn
2	xịn sò ghê
1	hàng mẫu ma dep tam on
0	Ko ưng ý lắm, nhiều luk sd hay bị đơ
1	bphone san xuat tại VN thì chất lương như z là ổn
0	hàg tệ thẻ nhó chập chon thưa các cửa hàg khák

Tiền xử lý văn bản

Dữ liệu lớn, đa dạng, nhiều noise=> xử lý dữ liệu thô ban đầu

Ví dụ :

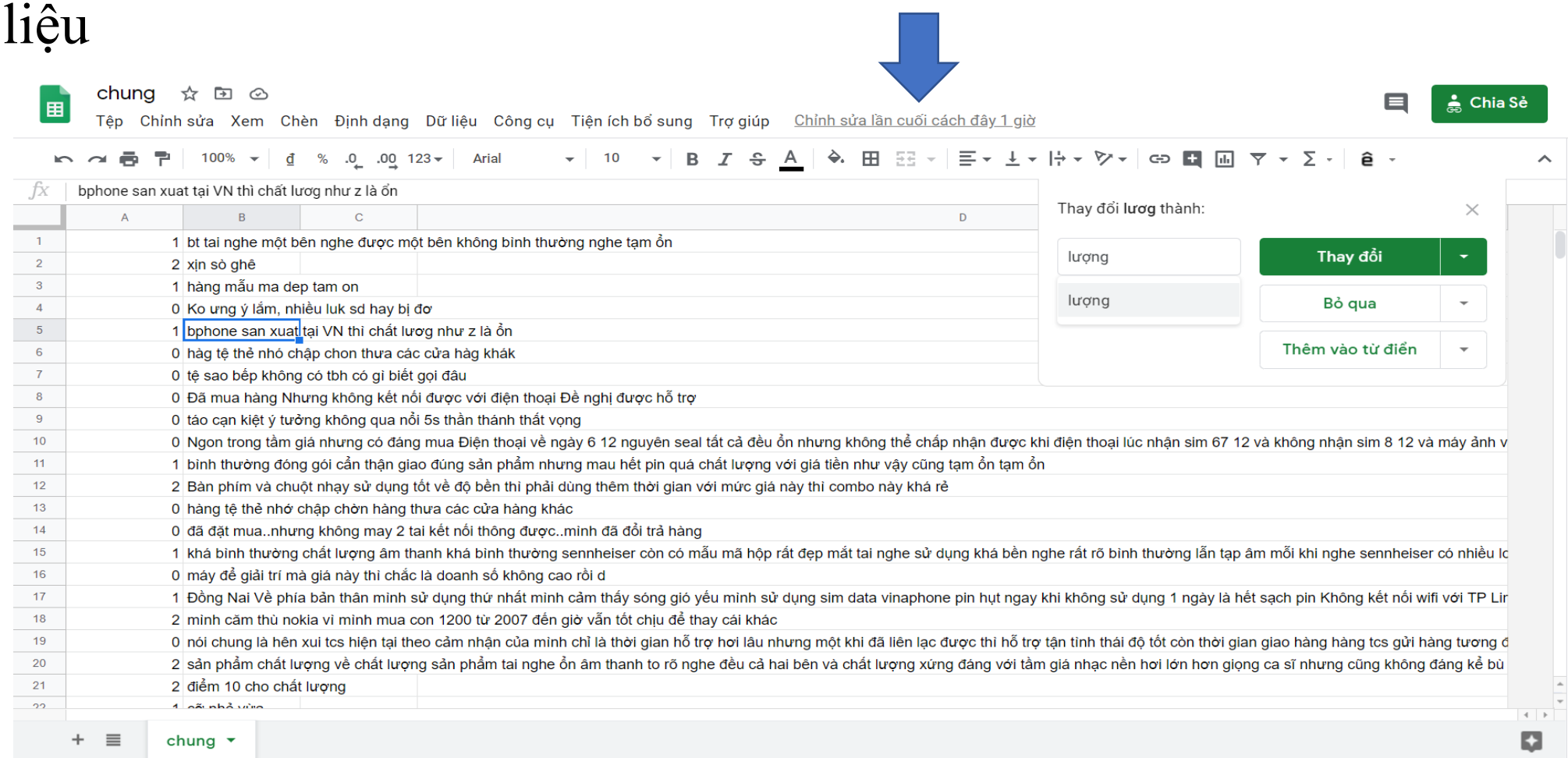
- Xử lý các ký tự lặp lại :*"goooooood" -> "good”*
- Xử lý các ký tự cho đồng nhất: *"$tupid" -> "stupid”*
- Xử lý các đầu vào đặc biệt: *"http://www.foo.com/bar" -> "[URL]”*
- Chuẩn hoá Unicode
- Chuyển chữ hoa thành chữ thường
- => **Bước làm sạch dữ liệu**



Tiền xử lý dữ liệu

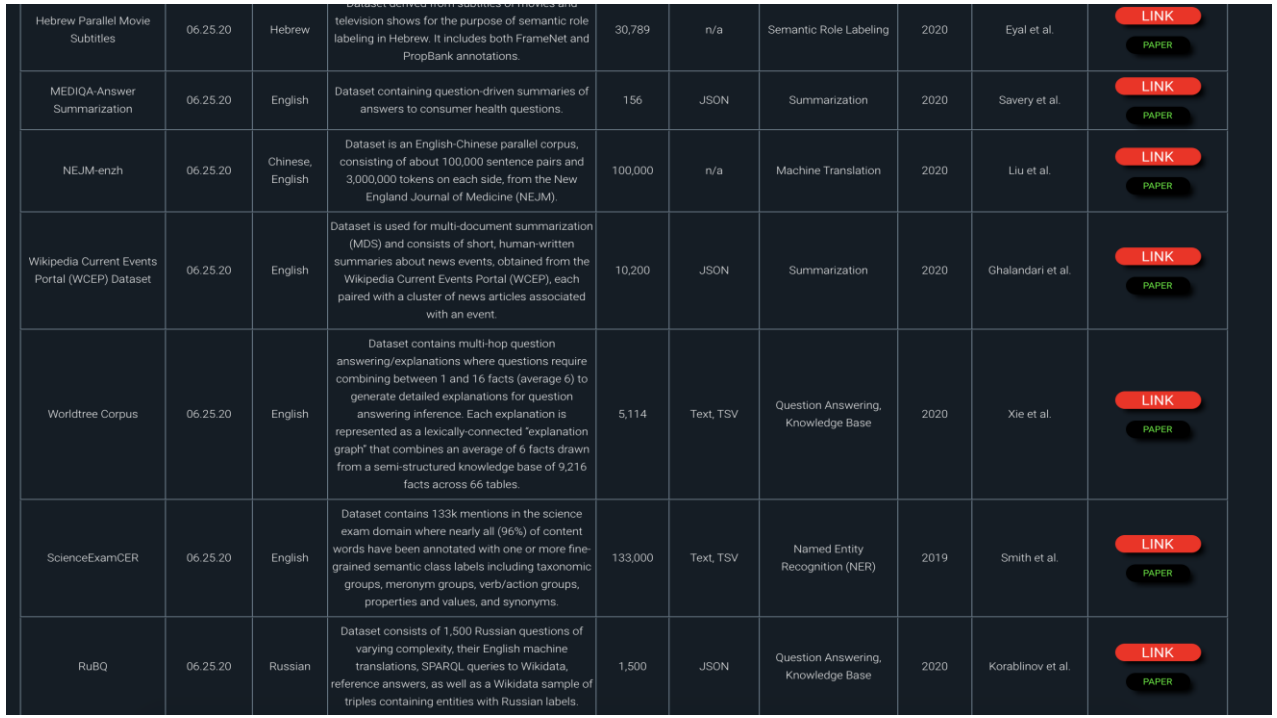
- Tiền xử lý
- Gán nhãn dữ liệu

0, Cua toi ko co dây sạc cuc pin du phong 0, Sao cục pin của mình sạc 3 tiếng mà còn chưa đầy 1 chấm xanh ….
1, Giao hàng rất chậm. Kh có cáp đi kèm
2, Wá tuyệt zờiiiiiiii
2, Rất hai long
1, Sp lỗi mỡ nguon khôg lên Xạc vào đt khôg vào Yc đổi cho mình sp khák
2, Không có dây sạc chỉ có sạc cục với bao đựng pin dự phòng



Một số dataset

- Tiếng Anh:
- https://datasets.quantumstat.com/
- https://www.kaggle.com/datasets?search=text+classification



Hebrew Parallel Movie Subtitles	06.25.20	Hebrew	Dataset derived from subtitles of movies and television shows for the purpose of semantic role labeling in Hebrew. It includes both FrameNet and PropBank annotations.	30,789	n/a	Semantic Role Labeling	2020	Eyal et al.	LINK PAPER
MEDIQA-Answer Summarization	06.25.20	English	Dataset containing question-driven summaries of answers to consumer health questions.	156	JSON	Summarization	2020	Savery et al.	LINK PAPER
NEJM-enzh	06.25.20	Chinese, English	Dataset is an English-Chinese parallel corpus, consisting of about 100,000 sentence pairs and 3,000,000 tokens on each side, from the New England Journal of Medicine (NEJM).	100,000	n/a	Machine Translation	2020	Liu et al.	LINK PAPER
Wikipedia Current Events Portal (WCEP) Dataset	06.25.20	English	Dataset is used for multi-document summarization (MDS) and consists of short, human-written summaries about news events, obtained from the Wikipedia Current Events Portal (WCEP), each paired with a cluster of news articles associated with an event.	10,200	JSON	Summarization	2020	Ghalandari et al.	LINK PAPER
Worldtree Corpus	06.25.20	English	Dataset contains multi-hop question answering/explanations where questions require combining between 1 and 16 facts (average 6) to generate detailed explanations for question answering inference. Each explanation is represented as a lexically-connected "explanation graph" that combines an average of 6 facts drawn from a semi-structured knowledge base of 9,216 facts across 66 tables.	5,114	Text, TSV	Question Answering, Knowledge Base	2020	Xie et al.	LINK PAPER
ScienceExamCER	06.25.20	English	Dataset contains 133k mentions in the science exam domain where nearly all (96%) of content words have been annotated with one or more fine-grained semantic class labels including taxonomic groups, meronym groups, verb/action groups, properties and values, and synonyms.	133,000	Text, TSV	Named Entity Recognition (NER)	2019	Smith et al.	LINK PAPER
RuBQ	06.25.20	Russian	Dataset consists of 1,500 Russian questions of varying complexity, their English machine translations, SPARQL queries to Wikidata, reference answers, as well as a Wikidata sample of triples containing entities with Russian labels.	1,500	JSON	Question Answering, Knowledge Base	2020	Korablinov et al.	LINK PAPER

Tiếng Việt:

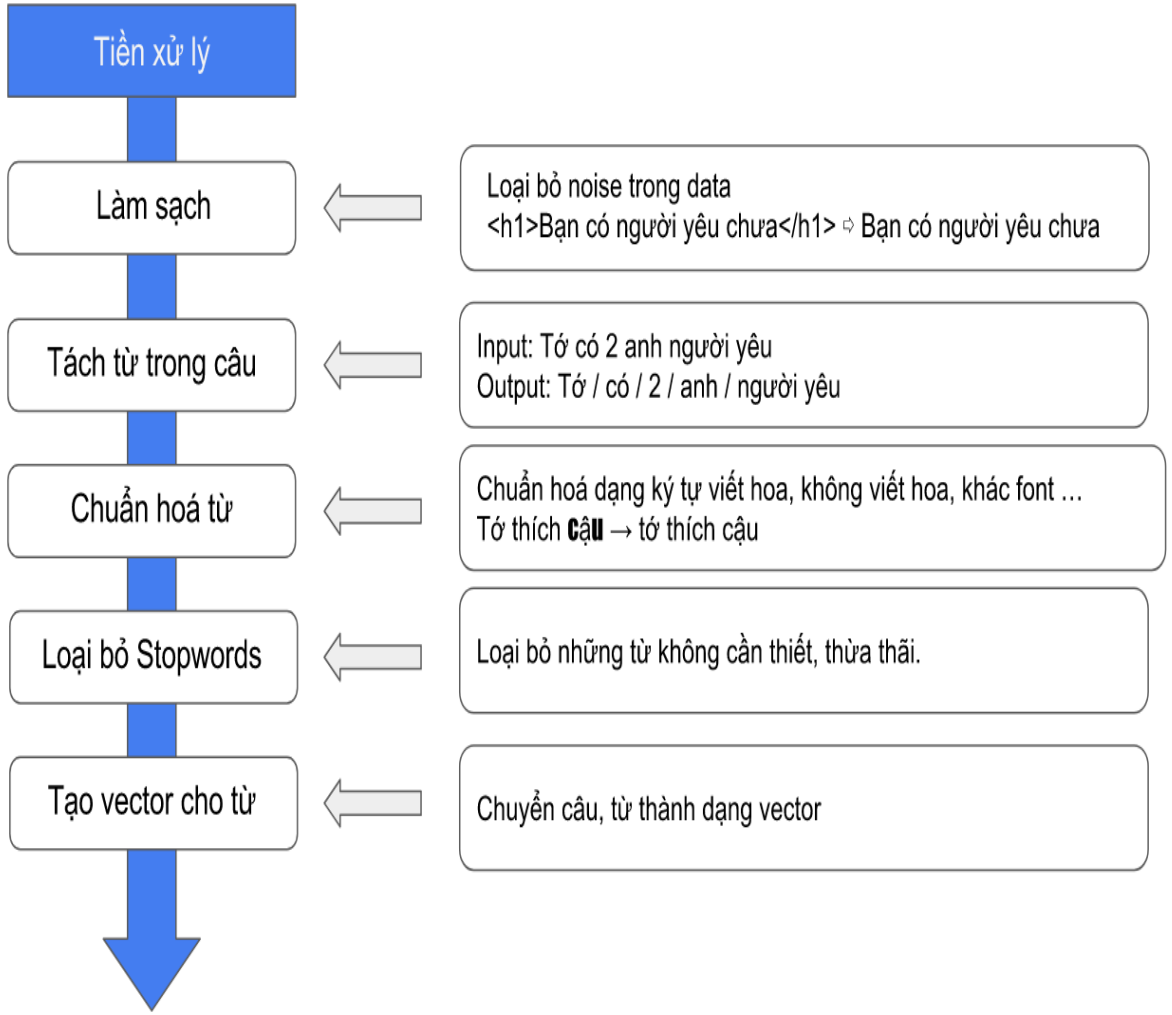
- Câu lạc bộ xử lý ngôn ngữ và tiếng nói Việt
- https://www.vlsp.org.vn/resources

- Các thư viện: underthesea,..
- Tự cralw…

Tiền xử lý văn bản

Tiếng Việt:

- Xóa HTML code (nếu có)
- Chuẩn hóa bảng mã Unicode (đưa về Unicode tổ hợp dựng sẵn)
- Chuẩn hóa kiểu gõ dấu tiếng Việt(dùng *òa úy* thay cho *oà uý*)
- Thực hiện tách từ tiếng Việt (sử dụng thư viện tách từ như pyvi, underthesea, vncorenlp,…)
- đưa về văn bản lower (viết thường)
- Xóa các ký tự đặc biệt: “.”, “,”, “;”, “)”, …

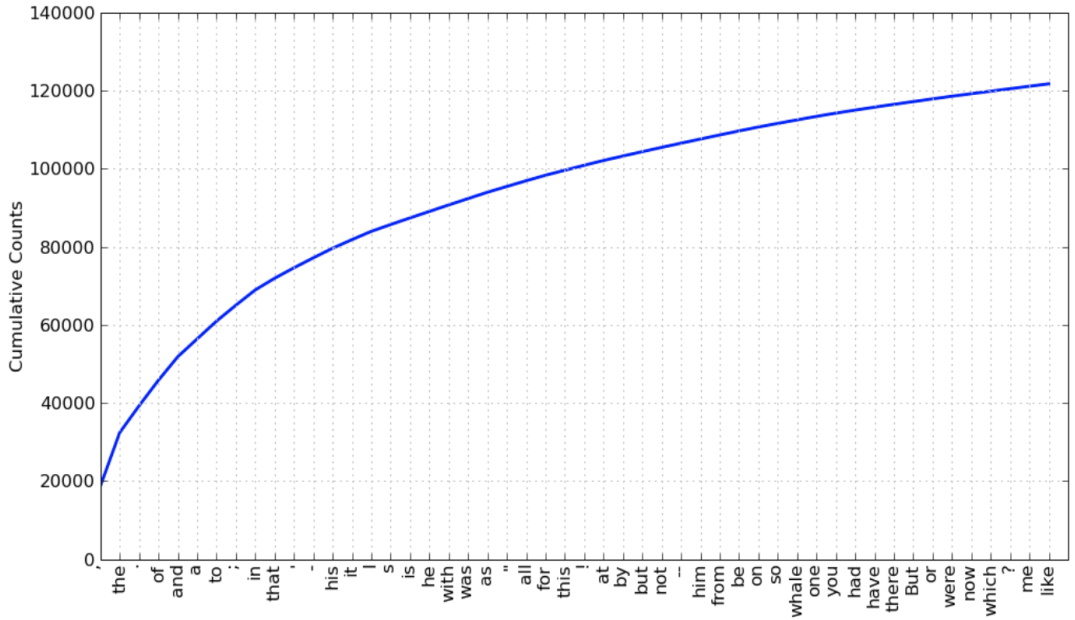


StopWords

- **StopWords** là những từ xuất hiện nhiều trong ngôn ngữ tự nhiên, tuy nhiên lại không mang nhiều ý nghĩa.
  - Tiếng Việt: để, này, kia... .
  - Tiếng Anh: is, that, this...
- Có rất nhiều cách để loại bỏ StopWords nhưng có 2 cách chính là:
  - Dùng từ điển

```
cậu
của
cứ
dù
nọ
phóc
này
kia
để
...
```

Dựa theo tần suất xuất hiện của từ



Stopwords

Tiếng Anh:

```
> stopwords("english")
  [1] "i"          "me"         "my"         "myself"     "we"
  [6] "our"        "ours"       "ourselves"  "you"        "your"
 [11] "yours"      "yourself"   "yourselves" "he"         "him"
 [16] "his"        "himself"    "she"        "her"        "hers"
 [21] "herself"    "it"         "its"        "itself"     "they"
 [26] "them"       "their"      "theirs"     "themselves" "what"
 [31] "which"      "who"        "whom"       "this"       "that"
 [36] "these"      "those"      "am"         "is"         "are"
 [41] "was"        "were"       "be"         "been"       "being"
 [46] "have"       "has"        "had"        "having"     "do"
```

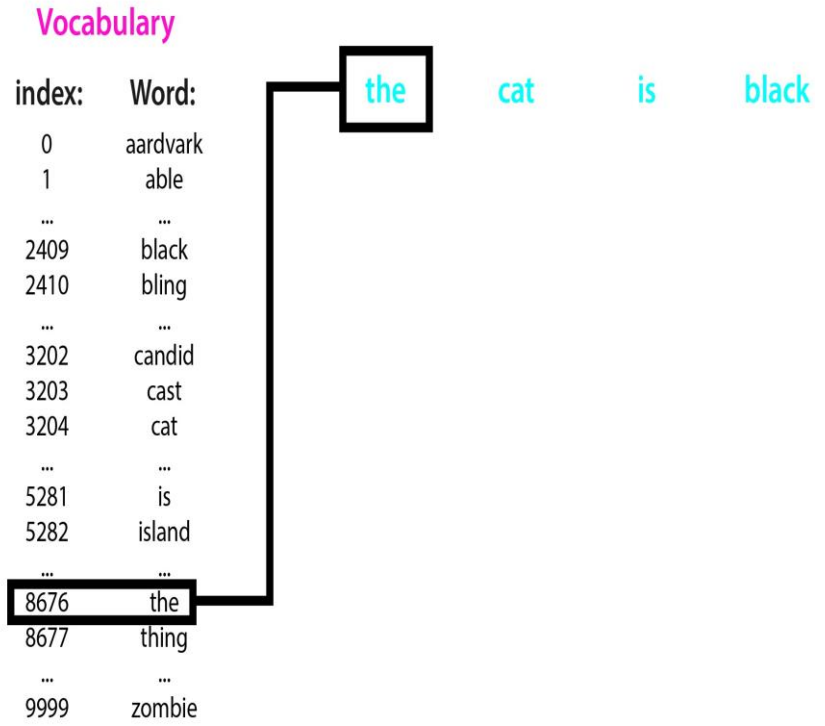
Tiếng Việt:

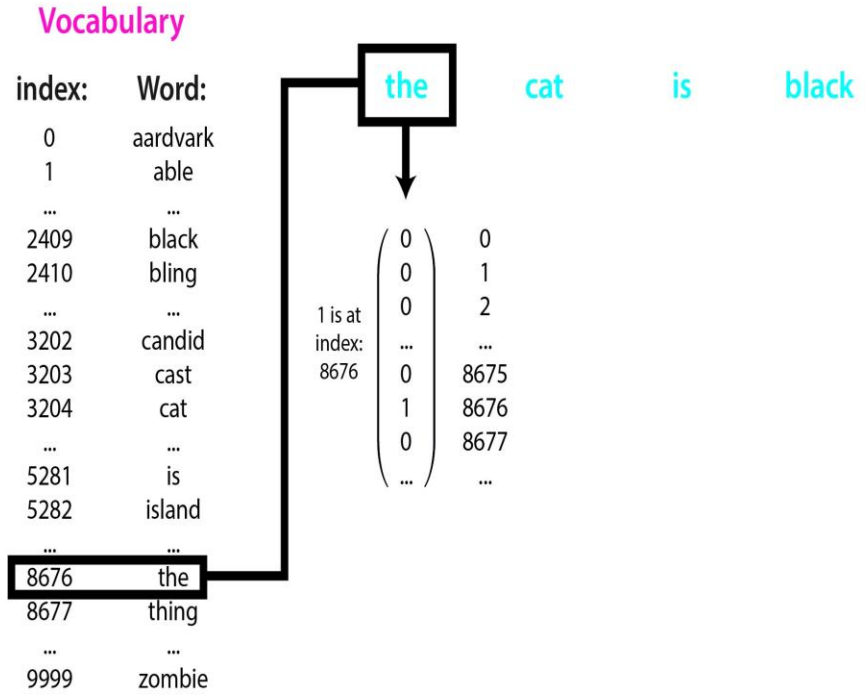
```
['bị', 'bởi', 'cả', 'các', 'cái', 'cần', 'càng', 'chỉ', 'chiếc', 'cho', 'chứ', 'chưa', 'chuyện', 'có', 'có_thể', 'cứ', 'của',
'cùng', 'cũng', 'đã', 'đang', 'đây', 'để', 'đến_nỗi', 'đều', 'điều', 'do', 'đó', 'được', 'dưới', 'gì', 'khi', 'không', 'là', 'l
ại', 'lên', 'lúc', 'mà', 'mỗi', 'một_cách', 'này', 'nên', 'nếu', 'ngay', 'nhiều', 'như', 'nhưng', 'những', 'nơi', 'nữa', 'phả
i', 'qua', 'ra', 'rằng', 'rằng', 'rất', 'rất', 'rồi', 'sau', 'sẽ', 'so', 'sự', 'tại', 'theo', 'thì', 'trên', 'trước', 'từ', 'từ
ng', 'và', 'vẫn', 'vào', 'vậy', 'vì', 'việc', 'với', 'vừa', '!', '"', '#', '$', '%', '&', "'", '(', ')', '*', '+', ',', '-',
'.', '/', ':', ';', '<', '=', '>', '?', '@', '[', '\\', ']', '^', '_', '`', '{', '|', '}', '~']
```

Word representation

One hot vector

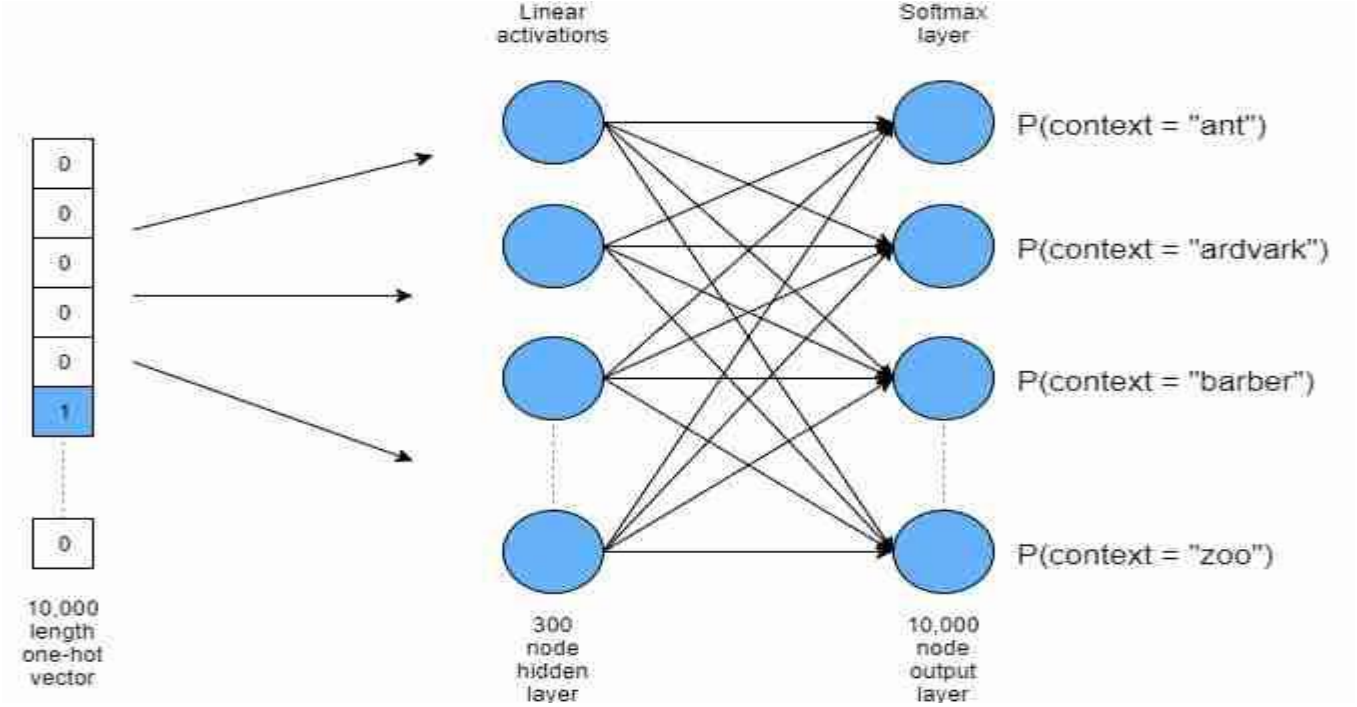
- Xây dựng một bộ từ vựng.
- Mỗi vector đại diện cho một từ có số chiều bằng số từ trong bộ từ vựng.
- Trong đó, mỗi vector chỉ có một phần tử duy nhất khác 0(bằng 1) tại vị trí tương ứng với vị trí từ đó trong bộ từ vựng.





One hot vector

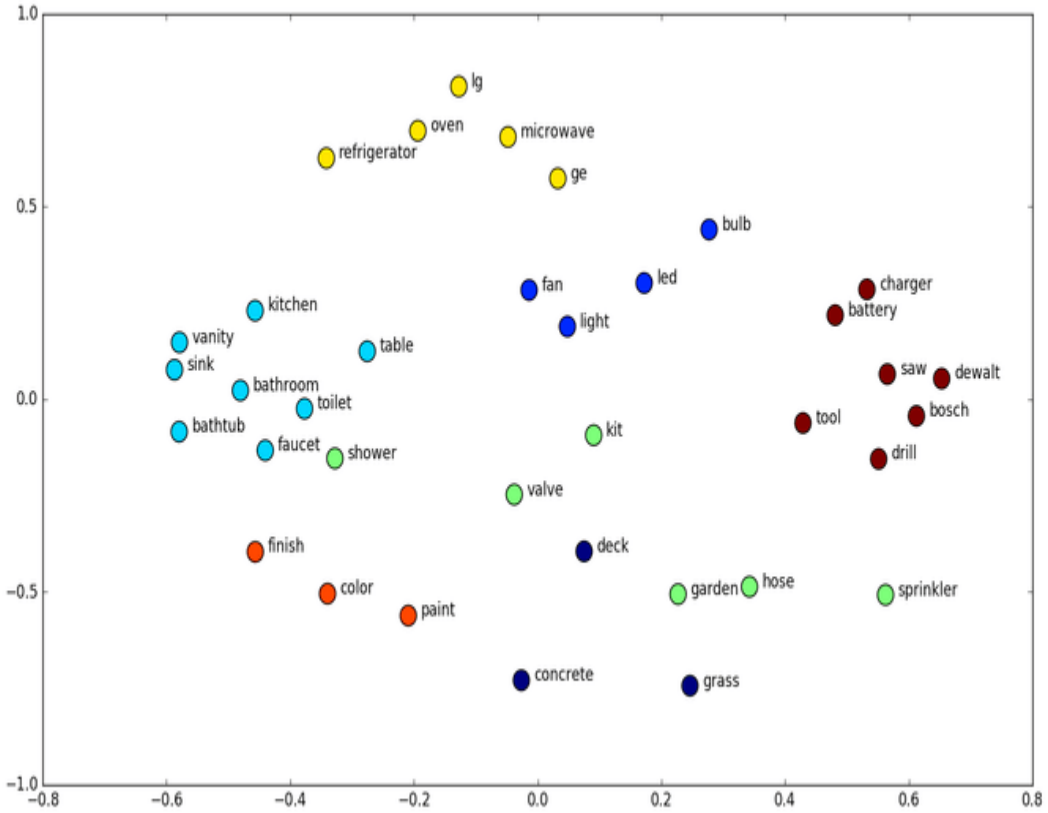
- Hạn chế
  - độ dài của vector là quá lớn
  - Không xác định được sự tương quan ý nghĩa giữa các từ
  - Tính toán với các thuật toán???

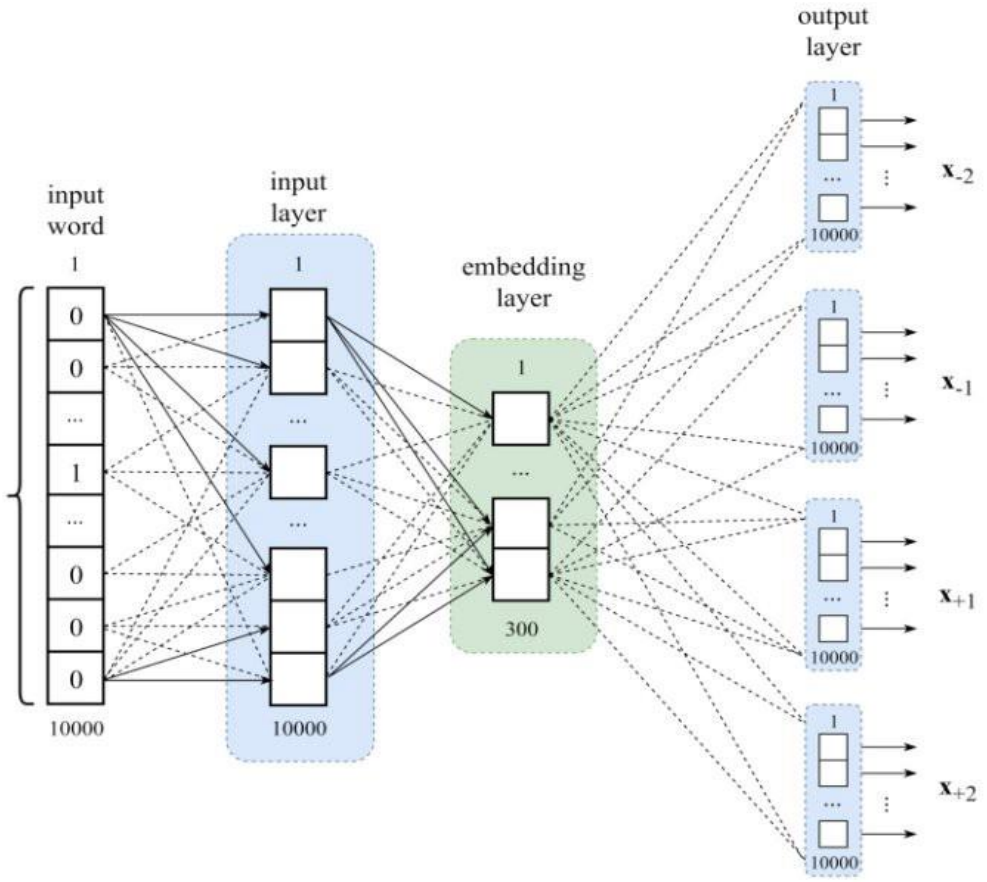


Phương pháp khắc phục

- Phương pháp truyền thống:
  - Bag of Words, IF-IDF
  - Matrix-Factorization: Giảm chiều vector SVD, PCA
- Word Embedding:
  - Word2Vec: Skip Gram, Continous Bag of Words (CBOW).
  - GloVe
  - FastText
- ..

Word Embedding

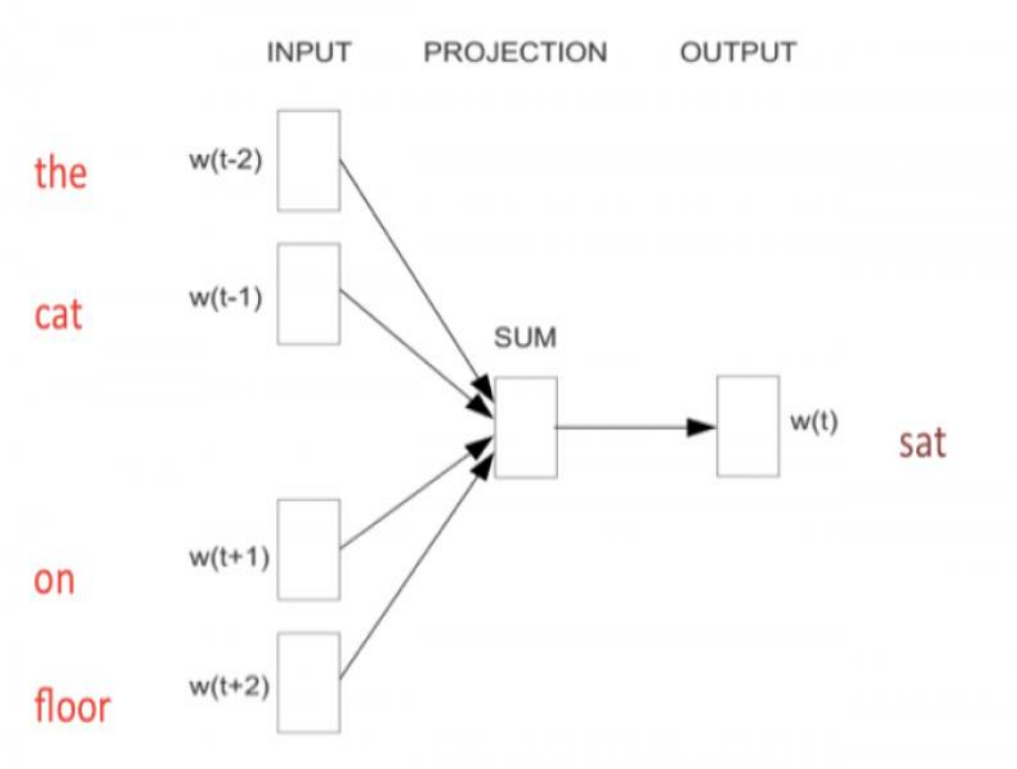


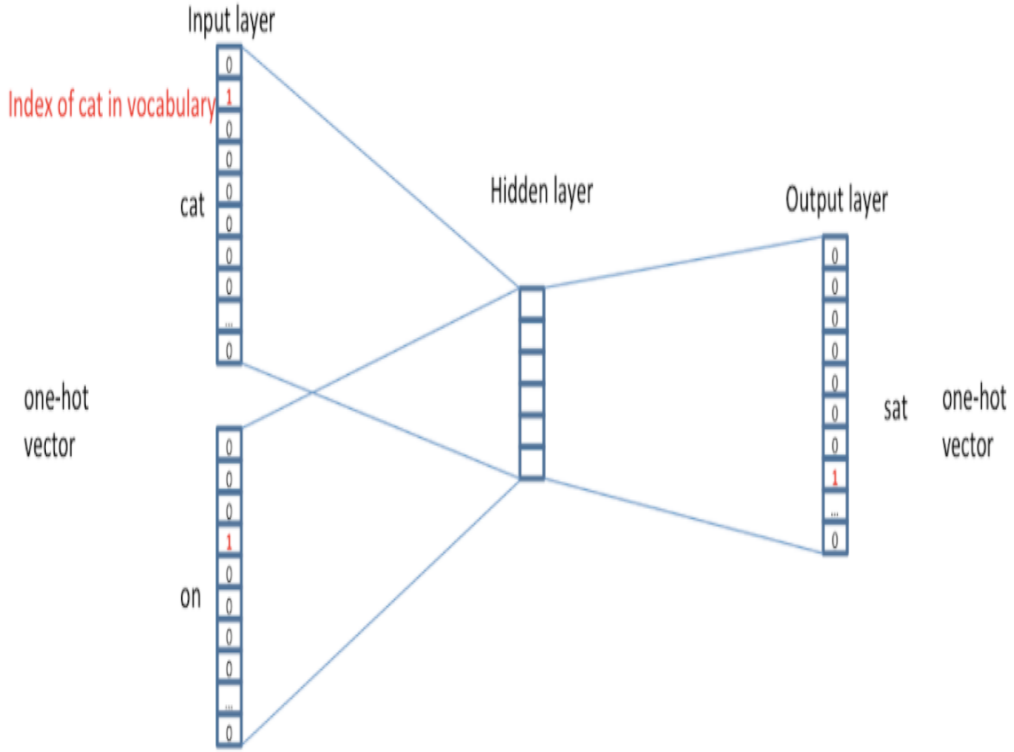


Word Embedding: Word2Vec

Continuous Bag of Words(CBOW)

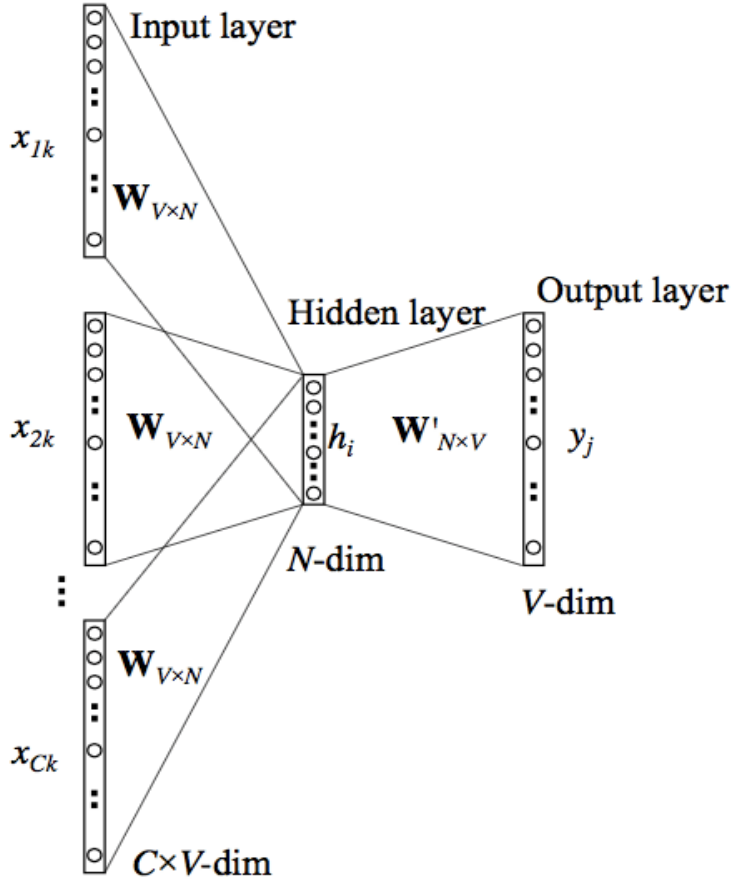
Ví dụ: *The cat sat on floor*





Continuous Bag of Words(CBOW)

- Với một corpus thật lớn, mạng CBOW học lần lượt từng từ trong corpus với context tương ứng
- 4 vector one-hot của bốn từ trong context sẽ được cộng lại thành một vector input duy nhất (có 4 giá trị 1) và đưa vào hệ thống như một input duy nhất.
- *Với ma trân* encoder W_{VxN}, ma trận decoder W'_{NxV} , *x* là one-hot vector của một từ , embedding vector tương ứng $w = x.W$.



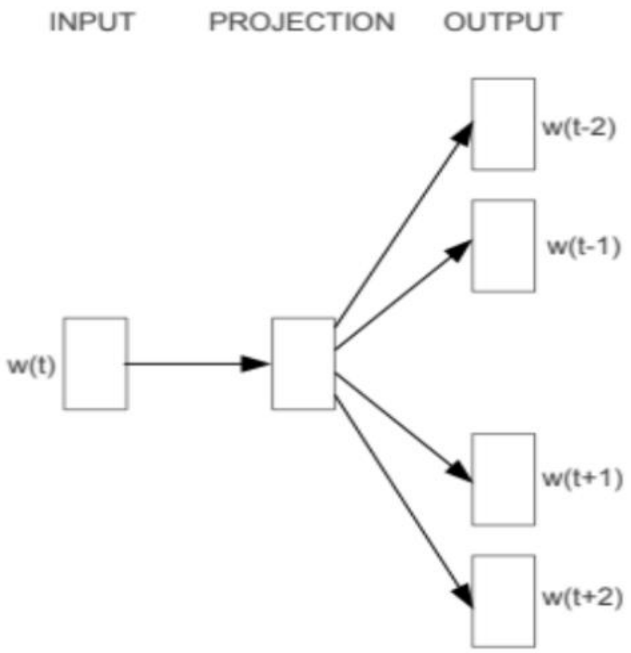
Continuous Bag of Words(CBOW)

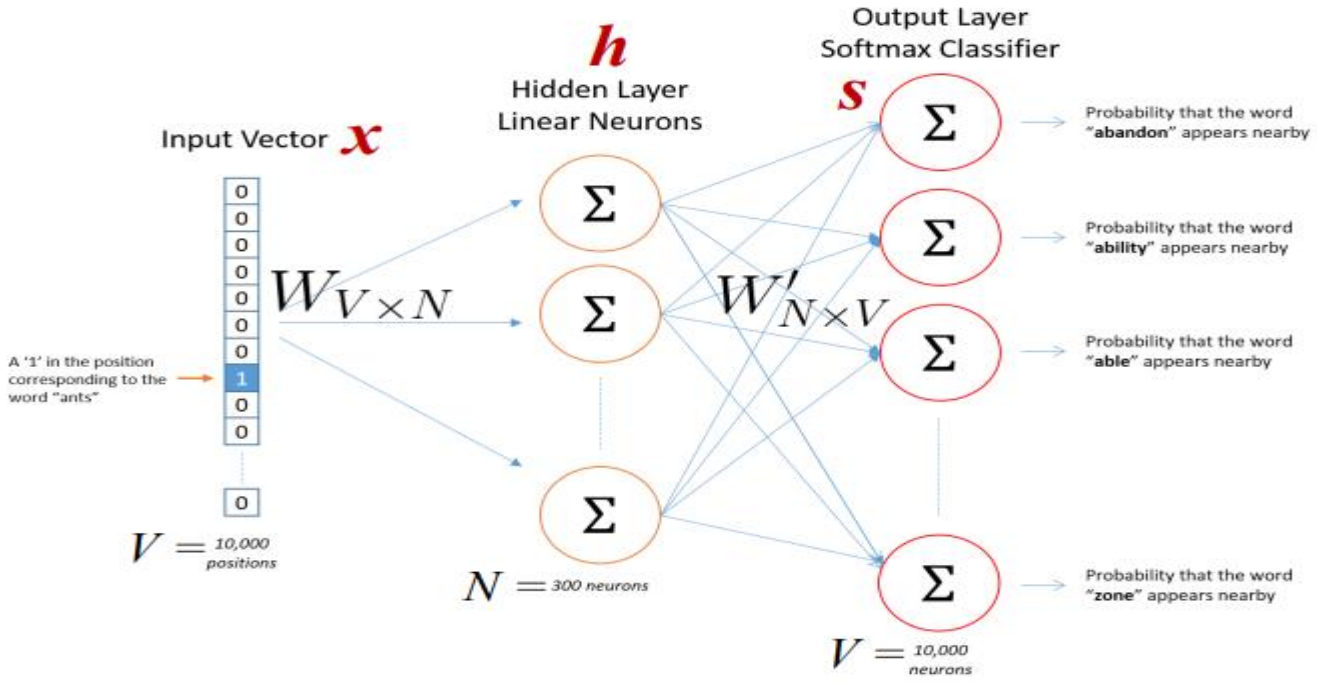
- embedding vector của một từ sẽ có các tính chất:
- Số chiều của *w* sẽ nhỏ hơn nhiều so với số chiều gốc của one-hot vector *x*.
- Một one-hot vector sẽ được encode không phải dựa vào chính nó như AutoEncoder mà dựa vào các từ *thường hay xuất hiện quanh nó* trong các văn bản.

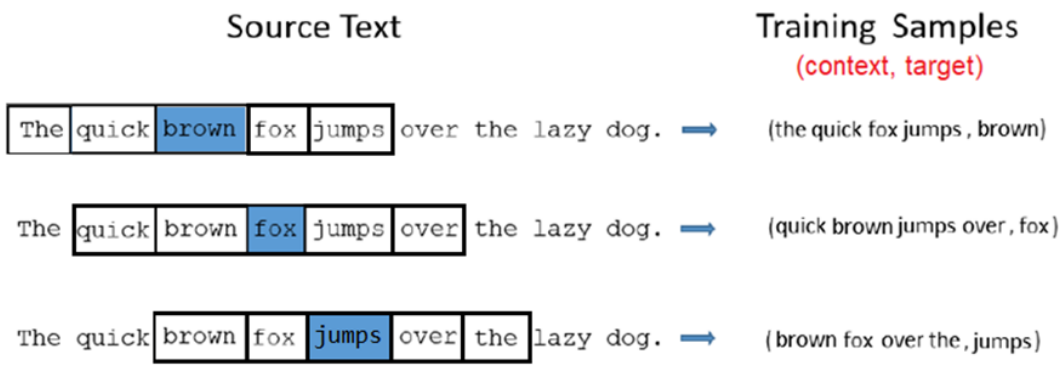
=> từ thường hay xuất hiện cạnh nhau trong các văn bản sẽ được encode thành các vector tương tự nhau.

Mô hình Skipgram

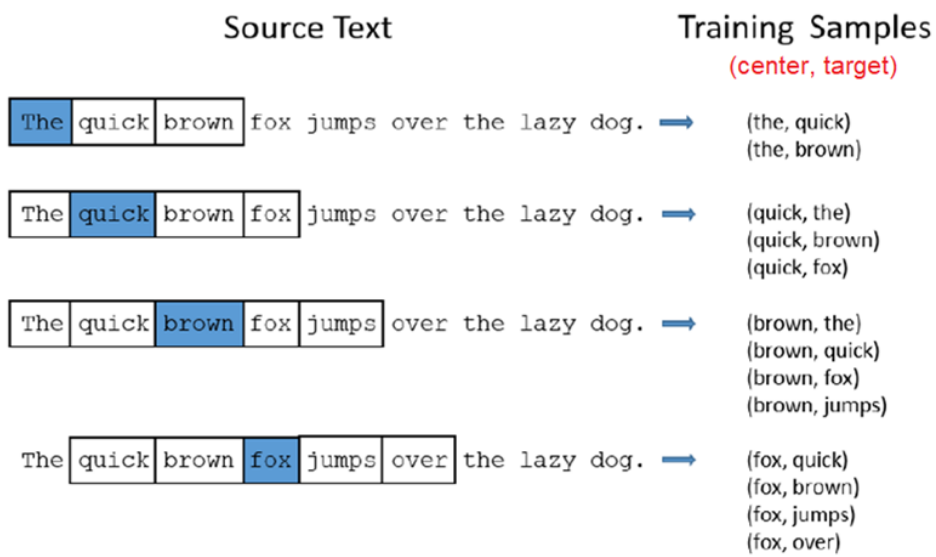
- Mô hình Skipgram gần giống như mô hình CBOW, chỉ thay đổi vai trò:
- context được dùng làm output và từ trung tâm sẽ dùng làm input => 1 ma trận encoder W và 4 ma trận decoder W‘.



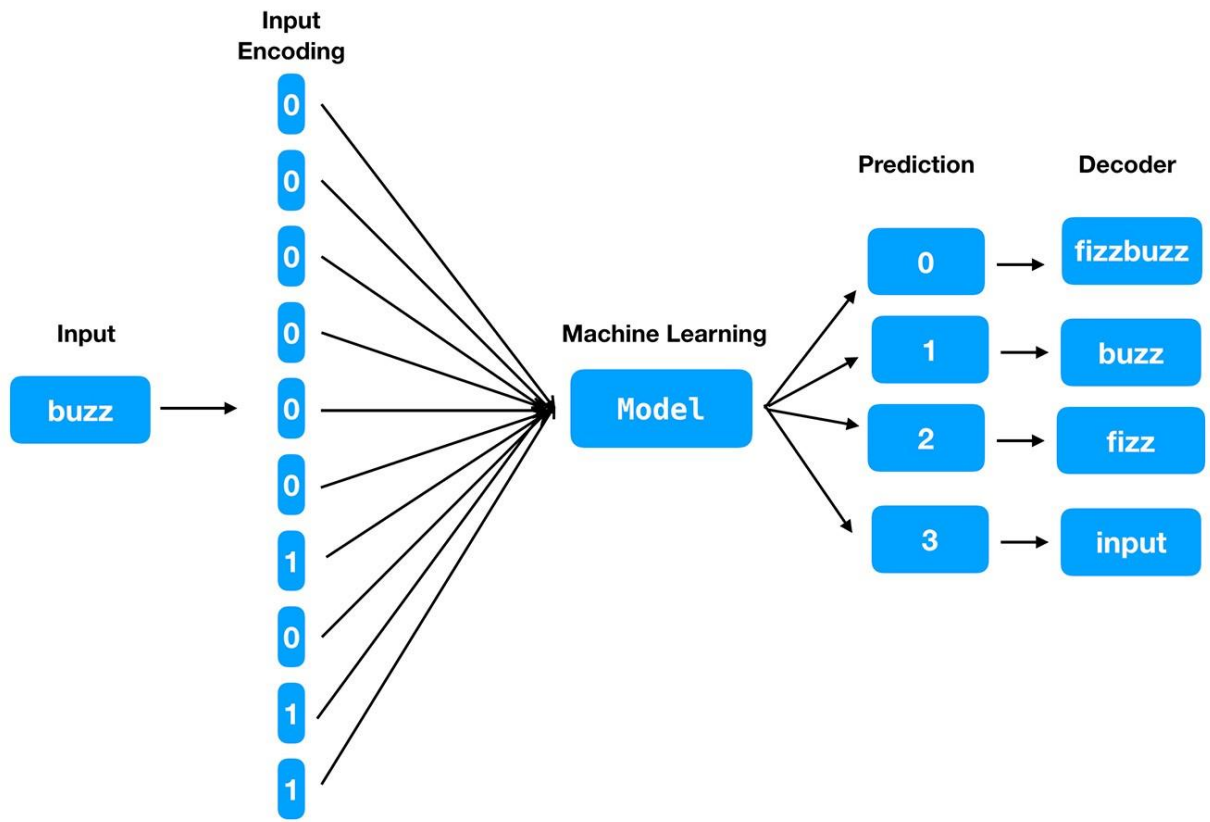




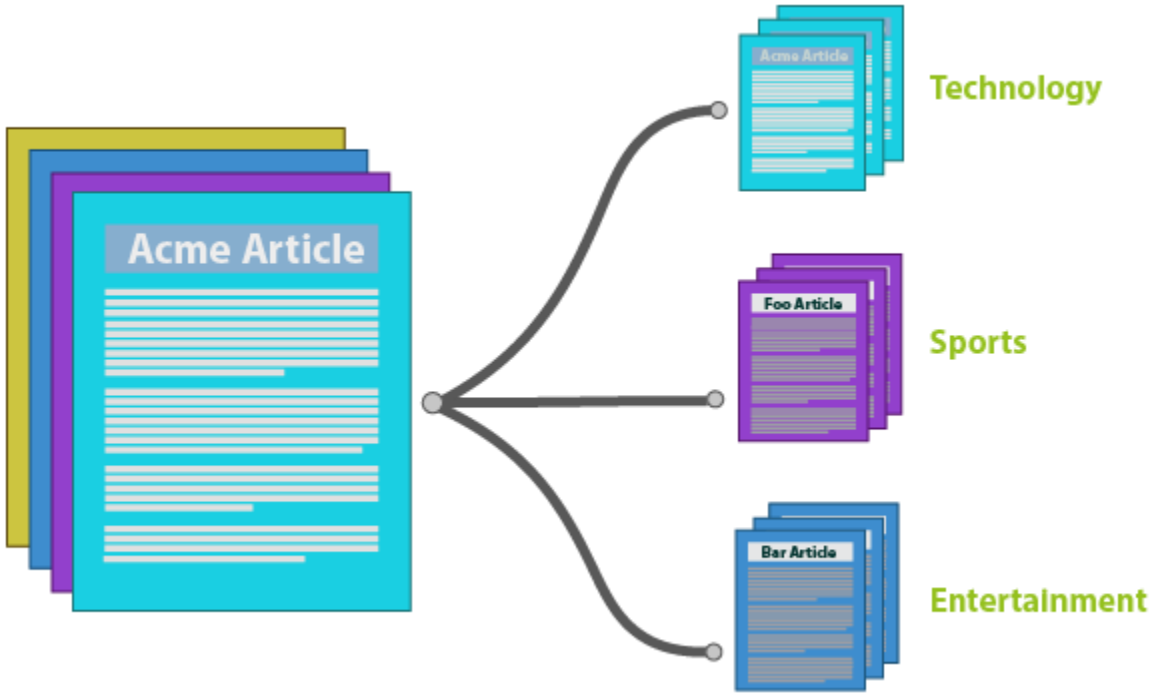
CBOW

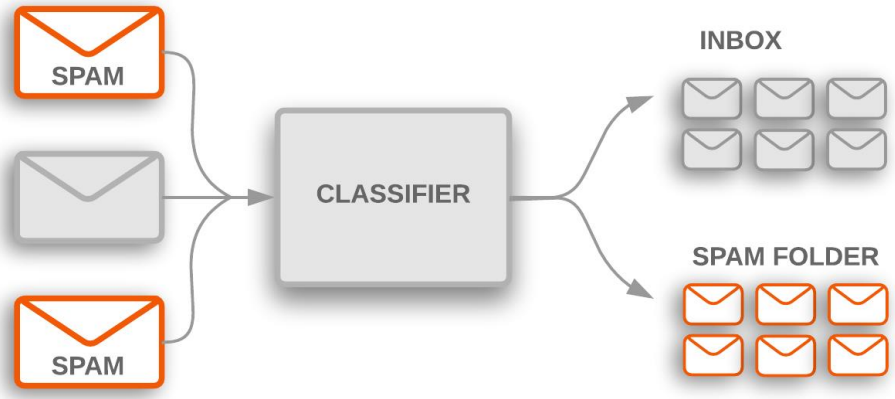


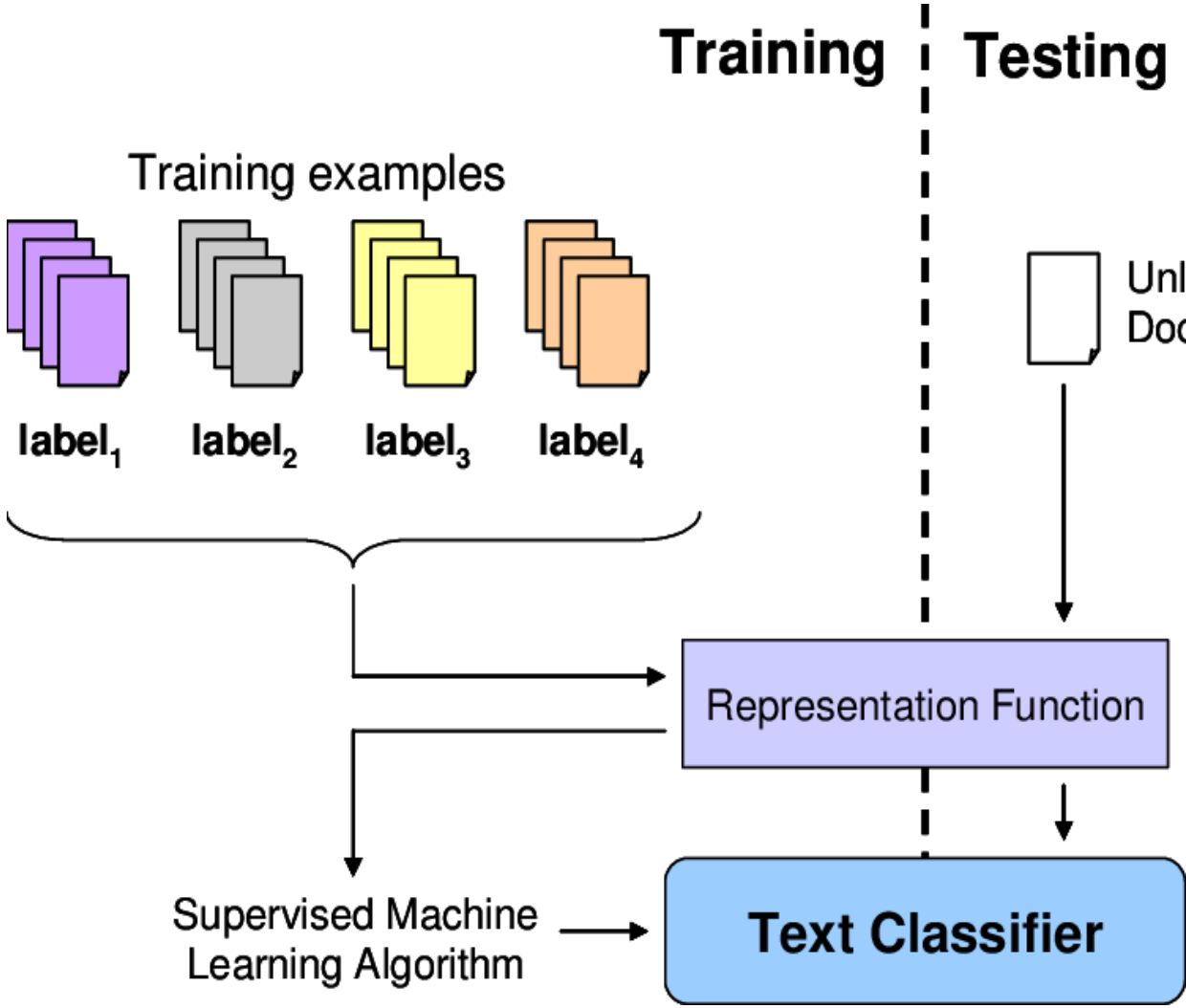
Skipgram


Input
Encoding
Input
buzz
0
0
0
0
0
0
1
0
1
1
Machine Learning
Model
Prediction
Decoder
0
fizzbuzz
1
buzz
2
fizz
3
input

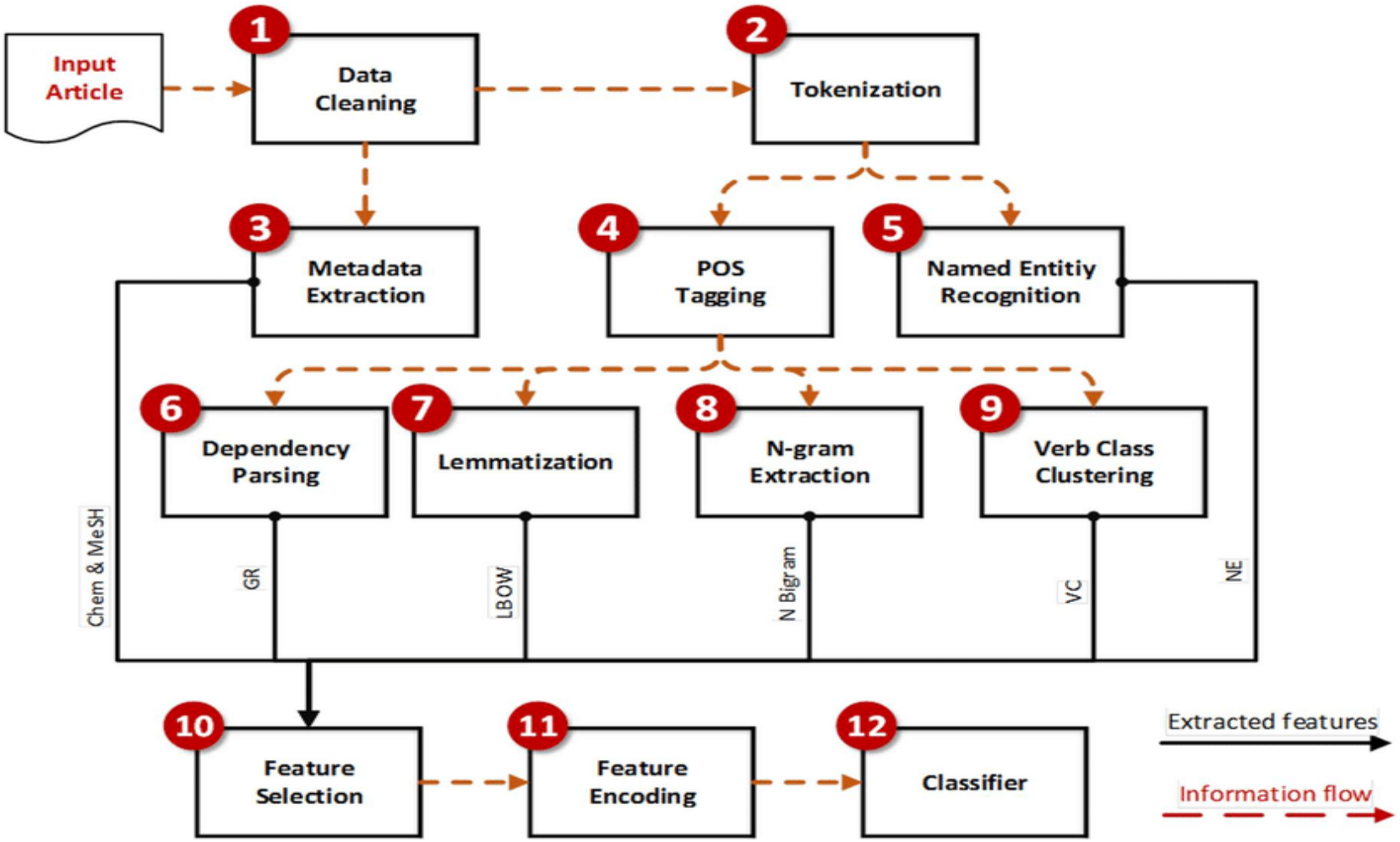
Bài toán phân loại văn bản







Phương pháp cổ điển



Deep Learning: CNN

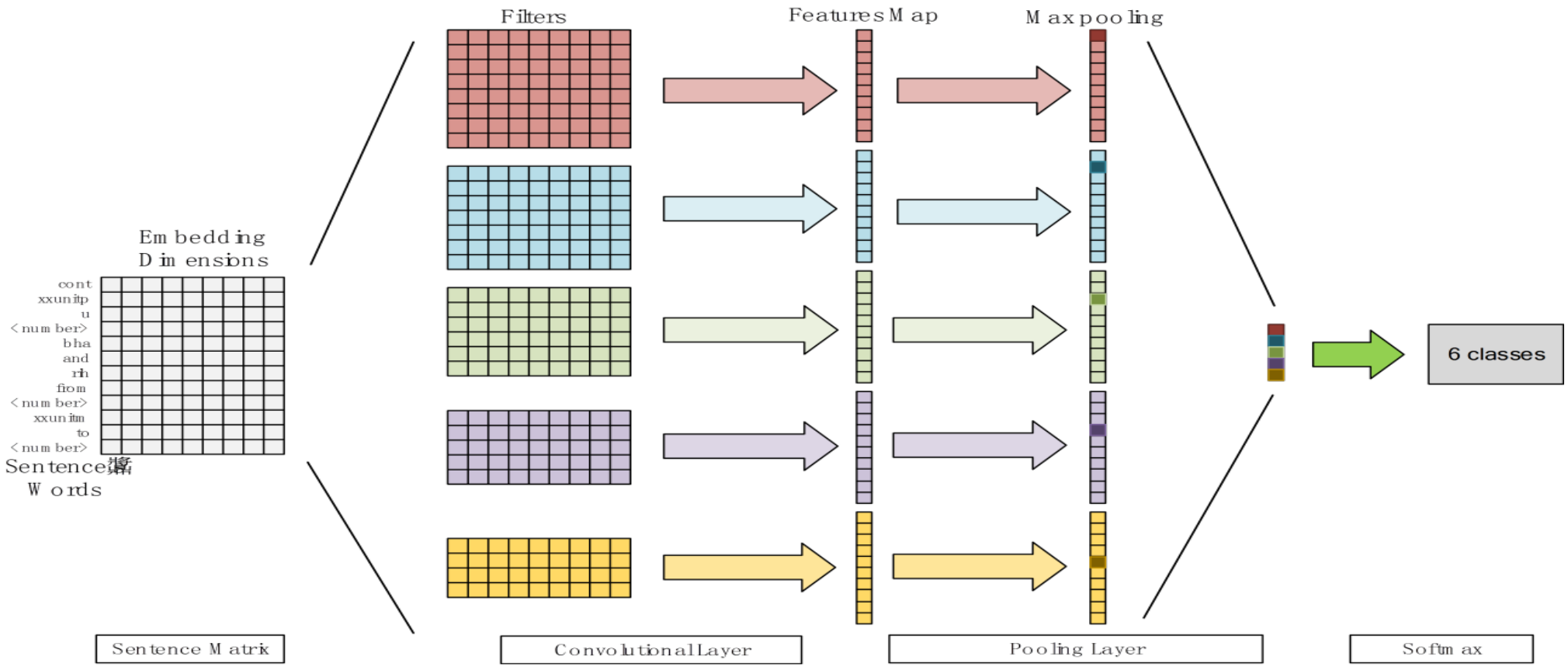
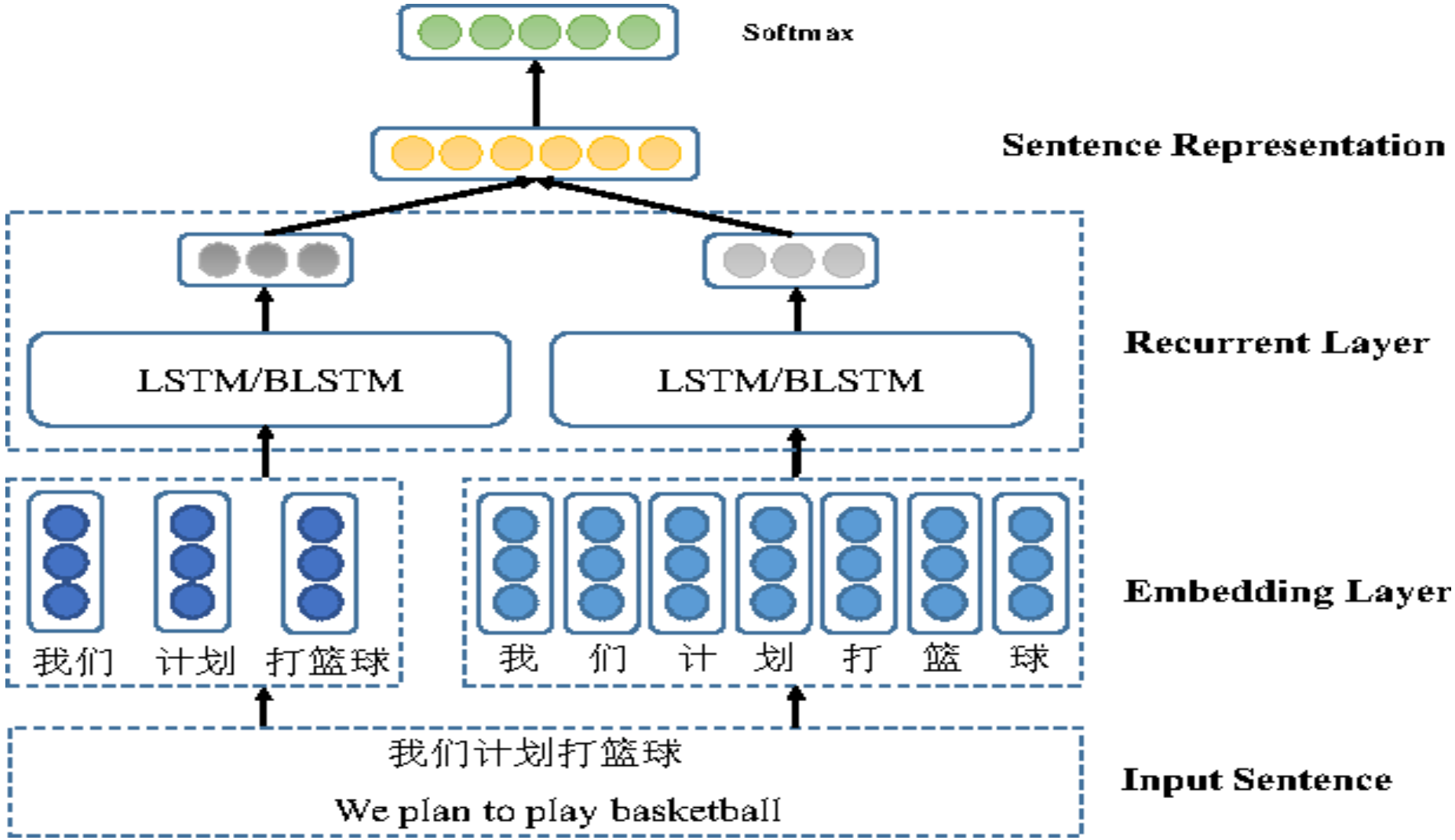
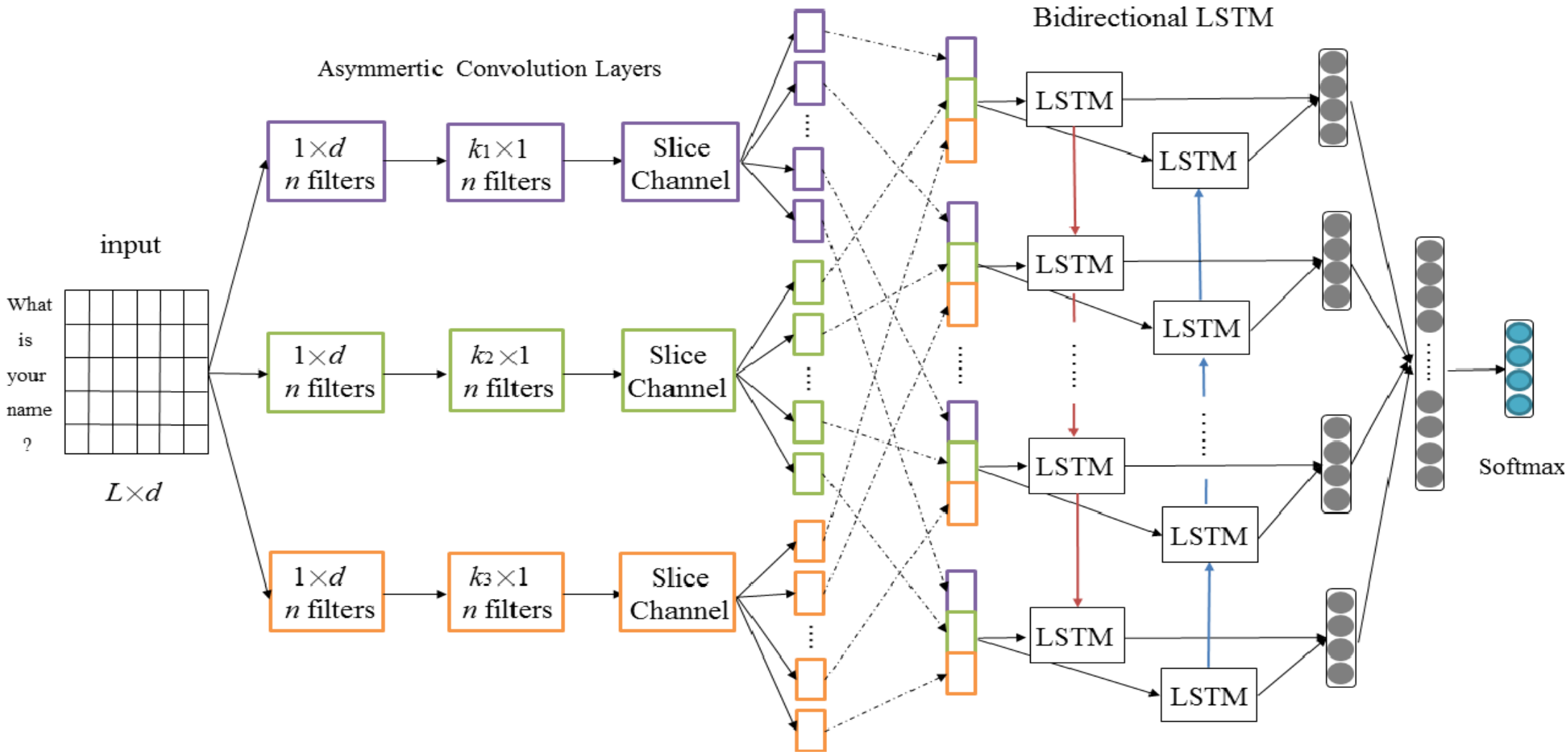


Figure 2. CNN-based text classification network.

Deep Learning: RNN



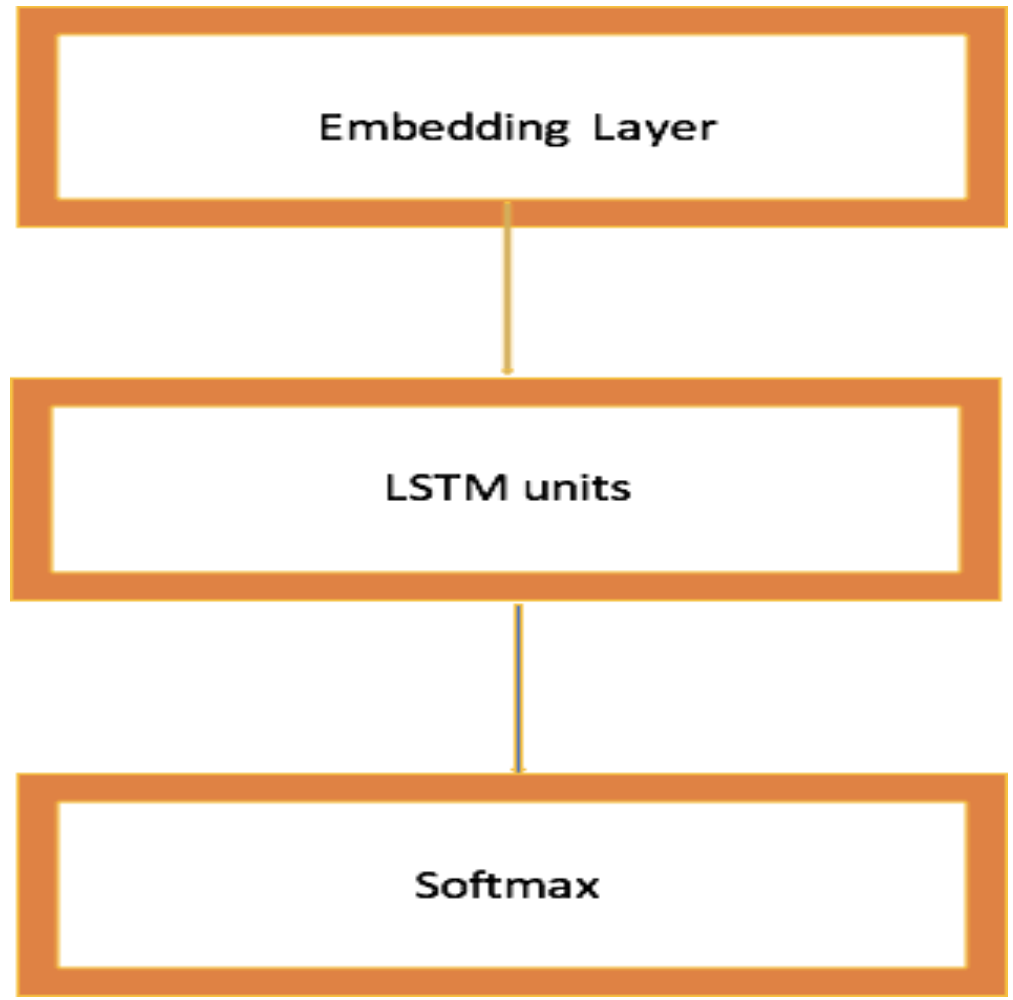
Deep Learning: CNN+RNN



Implementation of LSTM for Sentimental Analysis

Dữ liệu

	text	stars	sentiment
1411425	i went here and ordered the barbiq burger and i got the meat in medium well i had to admit it was really good although the fries got in the way it was salty and i had to dip the fries with mayo and ketchup all and all i would give the fries 35 out of 5 the service of the burger was right on time i like the look of the restaurant with the fun and relaxed look of the burger joint worth the visit especially if you just want to try a good burger	3	neg
971153	after calling 3 difference companies number one were the only one that had the least wait time for our broken ac in the middle of summer we called on monday morning and they were able to send a technician out that night at 8 pm when the technician came he determined it was our blower motor that was the problem and told us they will have to order the parts for it it will take couple days the repair will probably get done on wednesday and they will call us on tuesday to let us know the time they told us the repair will be between 11 am 2 pm on wednesday the repair guy pulled up in uhaul and after going up to the attic to check out the blower told us he needed to go get more parts for it and will be back in the afternoon\n\ncome 5 pm still no sign of the guy so we called them back and they had to call us back and track down the repair guy the receptionist called back and said he will be back soon come 7 pm still no sign so i called again and they told me they are on their way the guys did not show up until 845 pm took them about 30 minutes for the repair \n\nin the end yes my ac was repaired in a timely manner i understand summer is the busiest time for them my issue is the lack of communication we were waiting and waiting and not knowing whats going on and had to keep calling them back also we never got a receipt for the repair which they said they would email us	3	neg
1270461	went here for dinner and left very full and satisfied the family that runs the restaurant is originally from globe so the mexican food here is similar to what youd find in the globemiami area stepping into this restaurant it feels similar to how other restaurants such as serranos and rositas feels comfy atmosphere serving homestyle no frills cuisine i had the special tonight 2 chicken enchiladas with green sauce with rice and beans 850 while the enchiladas looked a bit mangled since chicken pieces were sticking out of the tortilla and the tortilla itself looked prodded and broken the enchiladas themselves were quite tasty with the green sauce and the rice and beans im usually a lightweight when it comes to finishing meals but in this case i cleaned my plate the salsa tastes good but is very watery which makes it hard to eat with chips we ordered iced teas and they were refilled promptly as needed id definitely be interested in going here again for some tasty and filling meals	4	pos


Embedding Layer
LSTM units
Softmax