1. VnCoreNLP

Các thuật ngữ trong NLP: <https://ongxuanhong.wordpress.com/2016/02/05/cac-thuat-ngu-trong-xu-ly-ngon-ngu-tu-nhien/>

VNCoreNLP đưa ra các chú thích trong ngôn ngữ (Tiếng Việt):

* Word segmentation: tách từ
* POS tagging: gán nhãn từ loại (part of speech tagging): một từ có nhiều từ loại (Danh từ, Động từ, Tính từ, …)
* Named entity recognition: Nhận dạng thực thể có tên (NER): Là tiến trình xác định và phân loại các phần tử trong văn bản vào các danh mục được định nghĩa trước như tên người, tên tổ chức, địa điểm, giá trị tiền tệ, tỷ lệ phần trăm
* Dependency parsing: Phân tích cây phụ thuộc tạo nên bởi các token trong câu và quan hệ hai ngôi giữa chúng. Một quan hệ phụ thuộc đơn được biểu diễn bởi relation(governor, dependent), trong đó governer và dependent là các token, và relation là một kiểu quan hệ phụ thuộc về ngữ pháp. Dạng biểu diễn này về cơ bản là một đồ thị có hướng được gán nhãn gọi là đồ thị phụ thuộc, có thể được biến đổi từ cây phân tích cú pháp
* 

Word-segmentation method: Single classification ripple down rules

<https://datquocnguyen.github.io/resources/LREC2018.pdf>

SCRDR: The data structure is similar to a [decision tree](https://en.wikipedia.org/wiki/Decision_tree). Each node has a rule, the format of this rule is IF cond1 AND cond2 AND ... AND condN THEN conclusion. Cond1 is a condition (boolean evaluation), for example A=1, isGreater(A,5) and average(A,">",average(B)). Each node has exactly two successor nodes, these successor nodes are connected to predecessor node by "ELSE" or "EXCEPT".

An example of SCRDR tree (defined recursively) is shown below:

IF (OutLook = "SUNNY" AND Temperature = "COOL") THEN PLAY="TENNIS" EXCEPT Child-1 ELSE Child-2

where Child-1 and Child-2 are also SCRDR trees. For example, Child-1 is:

IF (Wind = "WINDY" AND Humidity = "HIGH") THEN Play="SQUASH" EXCEPT NoChild ELSE NoChild

VNcoreNLP: formalize bài toán tách từ như một chuỗi các nhãn. Mỗi syllable là một labeled có giá trị B (begin of a word) hoặc I (inside of a word) rồi sử dụng mô hình SCRDR tìm nhãn của từ.

1. Sequence to sequence

Mô hình seq2seq cơ bản bao gồm hai mạng neural thành phần được gọi là mạng mã hóa (encoder) và mạng giải mã (decoder) để sinh ra chuỗi đầu ra t1:m từ một chuỗi đầu vào x1:n.Mạng neural encoder mã hóa chuỗi đầu vào thành một vector *c*có độ dài cố định. Mạng neural decoder sẽ lần lượt sinh từng từ trong chuỗi đầu ra dựa trên vector *c* và những từ được dự đoán trước đó cho tới khi gặp từ kết thúc câu. Trong mô hình seq2seq, chúng ta có thể sử dụng các kiến trúc mạng khác nhau cho thành phần encoder và decoder như RNN hay mạng neural *convolutional.*Trong bài viết này, người viết tập trung vào mô hình seq2seq sử dụng mạng RNN.

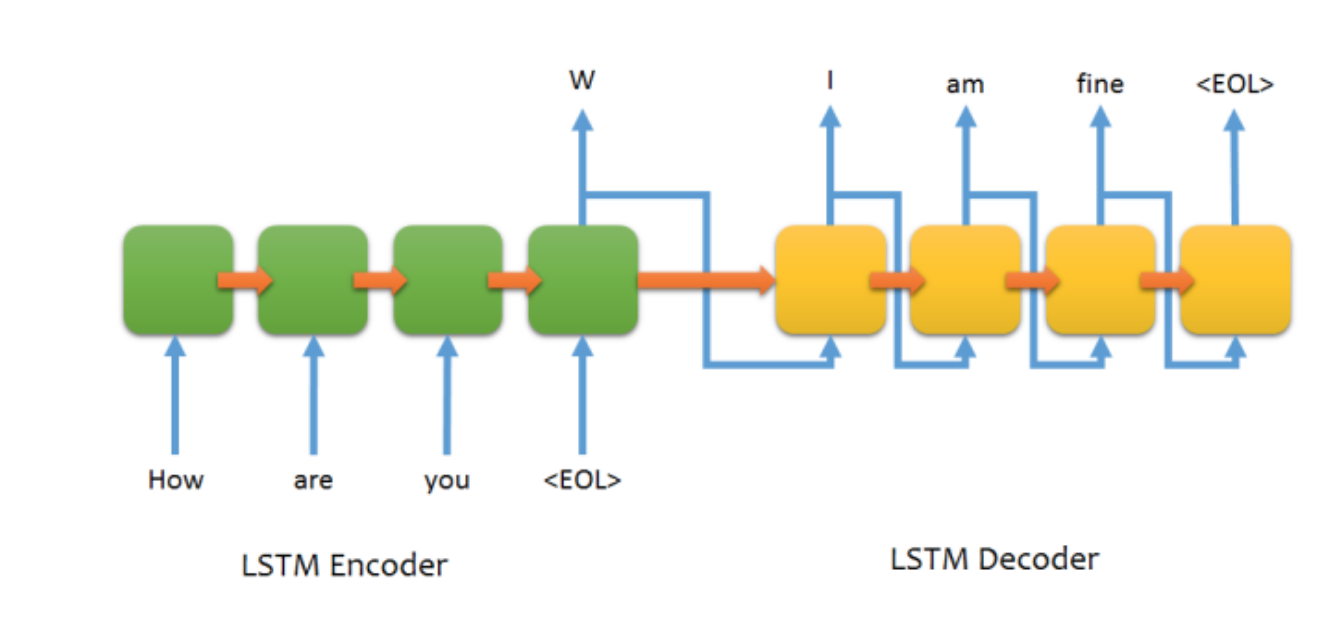
Trước hết, thành phần RNN encoder sinh ra vector c với độ dài cố định từ một chuỗi các vector đầu vào: c=RNNenc(x1:n). Sau đó, thành phần RNN decoder sẽ sinh lần lượt từng từ cho chuỗi đầu ra theo mô hình sinh có điều kiện (conditional generation).

p(tj+1=k|t̂ 1:j,c)=f(0(sj+1))

sj+1=R(sj,[t̂ j;c])

t̂ j ~ p(tj|t̂ 1:j−1,c)

Ở đây, tj là bước thời gian thứ *j*trong chuỗi output. t̂ j là giá trị dự đoán tại bước thứ *j*và sj là các trạng thái ẩn.



Mô hình seq2seq cơ bản có nhược điểm là yêu cầu RNN decoder sử dụng toàn bộ thông tin mã hóa từ chuỗi đầu vào cho dù chuỗi đó dài hay ngắn. Thứ hai, RNN encoder cần phải mã hóa chuỗi đầu vào thành một vec-tơ duy nhất và có độ dài cố định. Ràng buộc này không thực sự hiệu quả vì trong thực tế, việc sinh ra từ tại một bước thời gian trong chuỗi đầu ra có khi phụ thuộc nhiều hơn vào một số những thành phần nhất định trong chuỗi đầu vào. Ví dụ, khi dịch một câu từ tiếng nước này sang tiếng nước khác, chúng ta thường quan tâm nhiều hơn đến ngữ cảnh xung quanh từ hiện tại so với các từ khác trong câu. Kỹ thuật attention được đưa ra để giải quyết vấn đề đó.

Ở mức trừu tượng, kỹ thuật attention nới lỏng điều kiện rằng toàn bộ chuỗi đầu vào được mã hóa bằng một vector duy nhất. Thay vào đó các từ trong chuỗi đầu vào sẽ được RNN encoder mã hóa thành một dãy các vector. Sau đó RNN decoder áp dụng kỹ thuật attention mềm dẻo (soft attention) bằng cách lấy tổng có trọng số của dãy các vector mã hóa. Các trọng số trong mô hình này được tính bằng một mạng neural truyền thẳng. RNN encoder, RNN decoder và các tham số trong kỹ thuật attention được huấn luyện đồng thời từ dữ liệu.

Ngoài phương pháp soft attention, kỹ thuật attention còn có một số các kiểu attention khác khác như additive attention, multiplicative attention, self-attention, key-value attention

1. Vision

Chatbot: <https://github.com/nicolas-ivanov/tf_seq2seq_chatbot>

<http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf>

<https://arxiv.org/pdf/1506.05869v1.pdf>

Sử dụng neural conversational model cho chatbot