Let’s translate ‘b4ha5a 4laY’!

Mari kita menerjemahkan ‘b4ha5a 4laY’!

Who's Speaking?

Hi! I am Nikmatun Aliyah Salsabila. You can call me Sabila or Salsa.

I am not an expert. I'm at a beginner level. Really.

Just graduated from Universitas Al Azhar Indonesia about 2.5 months ago. Yeah! Merdeka!

"There's always first time for every thing" And yes.. this is my first time!

cantik

Terjemahkan

красивая

belle

جميلة

bonita

güzel

bella

frumoasa

Hmmmm how about translate “cantique”, “cntk”, “ntik” into “cantik”???

liihhh, km ntik bgd dech!

Machine Translation

> In a machine translation task, the input already consists of a sequence of symbols in some language, and the computer program must convert this into a sequence of symbols in another language.
>
> — Page 98, Deep Learning(Ian Goodfellow), 2016

Sequence:

makan → m-a-k-a-n

m-a-k-n-a → makna

Terjemahkan

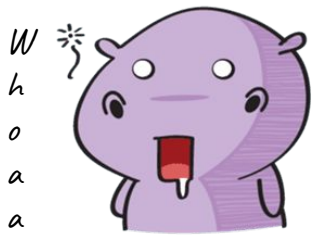
eat → e-a-t

t-e-a → tea

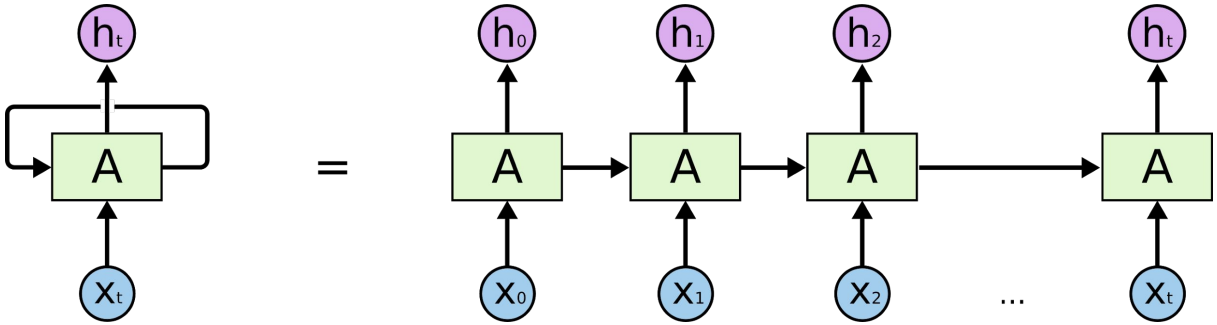
one of the most challenging AI tasks?

> - The strength of NMT lies in its ability to learn directly, in an end-to-end fashion, the mapping from input text to associated output text.
> - Three inherent weaknesses of NMT: its slower training speed, ineffectiveness in dealing with rare words, and sometimes failure to translate all words in the source sentence.
> - Finally, NMT systems sometimes produce output sentences that do not translate all parts of the input sentence – in other words, they fail to completely "cover" the input, which can result in surprising translations.
>
> — Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation, 2016

Neural Machine Translation

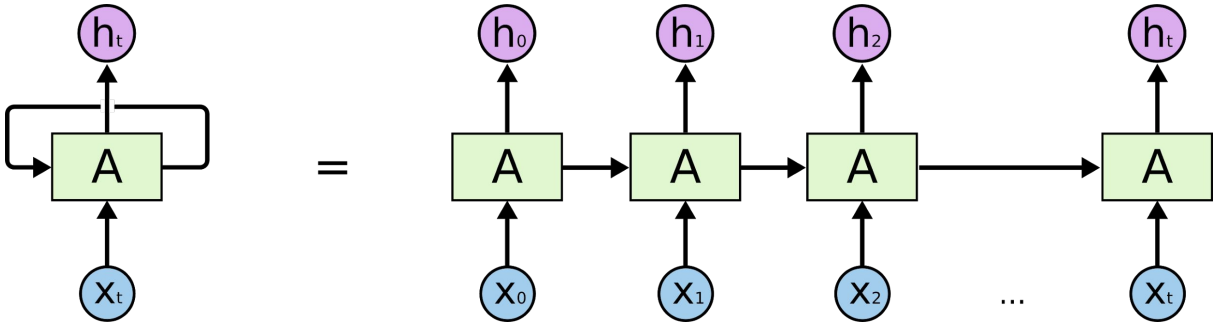


Recurrent Neural Networks



— Understanding LSTM Networks, C Olah, 2015

Recurrent Neural Networks



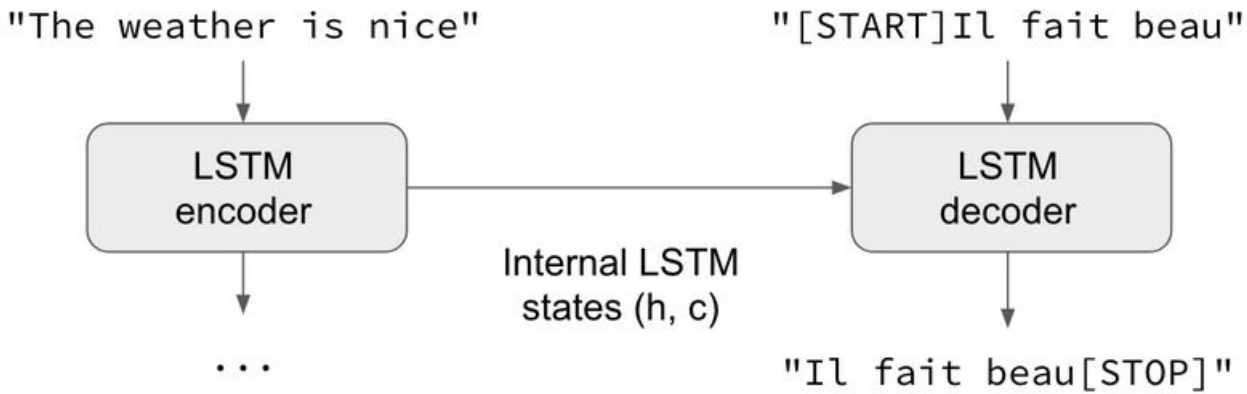
— Understanding LSTM Networks, C Olah, 2015

Long Short Term Memory

LSTMs are a special kind of recurrent neural network (RNN), capable of learning long-term dependencies

Yeah!

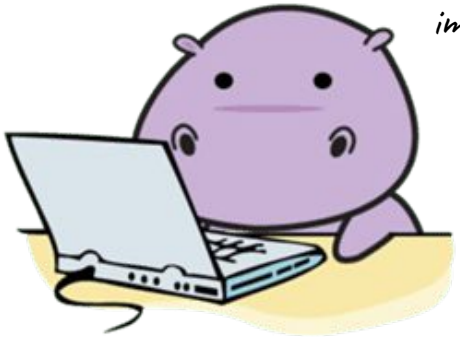
Sequence-to-sequence, Encoder Decoder Model



— A ten-minute introduction to sequence-to-sequence learning in Keras, Francois Chollet, 2017

Keras

Let's go to the implementation!



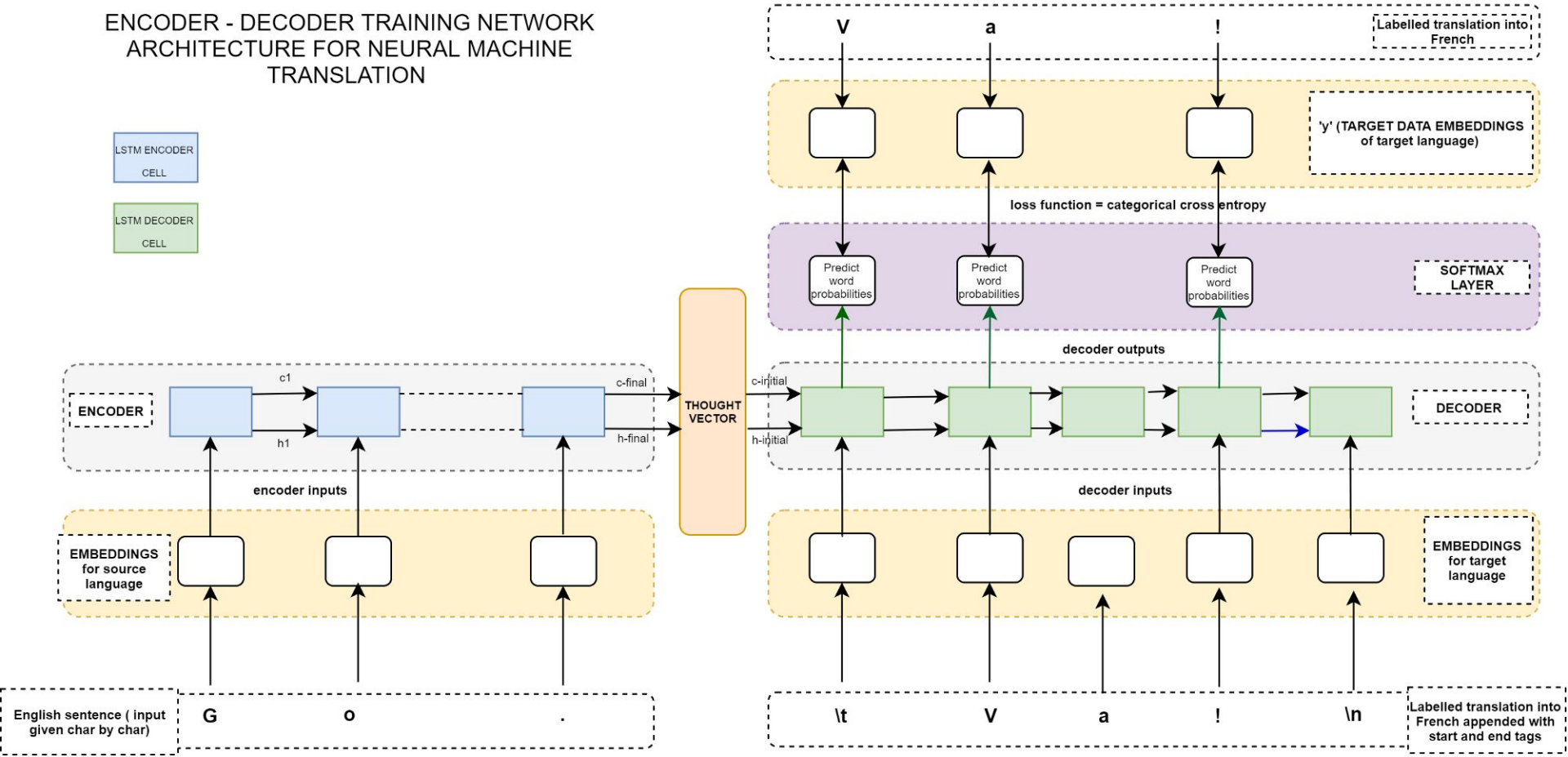
Dataset

Kamus Alay : a dataset of pairs of ‘alay’ words and their normal form

- 15,006 instances
- 4,331 alay words
- 2,005 normal words

alay	normal
kk	kakak
bgt	banget
sm	sama
gk	enggak
klo	kalo

I am going to present the details about dataset at IALP, Nov 15-18. Wish me luck!



Encoder-Decoder training architecture for NMT—image copyright@Ravindra Kompella

WHY?! WHY?! WHY?!
WHY?! WWHHYY??!!!

Loss: 0.12, Acc: ~0.25

Challenge

Language is hard

- Source language ambiguity
- Target language variation
  Yes, words with multiple meanings

- Context, culture, background, etc

km → kamu, kami
d → di, ada
banget ← bingitsss, bet, bgttt
cantik ← syantiek, canyk, cuaaantikk

> Part of the reason why translation is difficult for computers is that translation is just difficult: difficult even for humans.
>
> — Why Translation is Difficult for Computers, D.J Arnold, 2003

https://github.com/nasalsabila/PyconID2018

https://github.com/nasalsabila/kamus-alay

Thank you~