جامعة مولاي إسماعيل
ⵜⴰⵙⴷⴰⵡⵉⵜ ⵎⵓⵍⴰⵢ ⵉⵙⵎⴰⵄⵉⵍ
UNIVERSITÉ MOULAY ISMAÏL

MAY 26, 2023

MIAAD

Automatic Translation

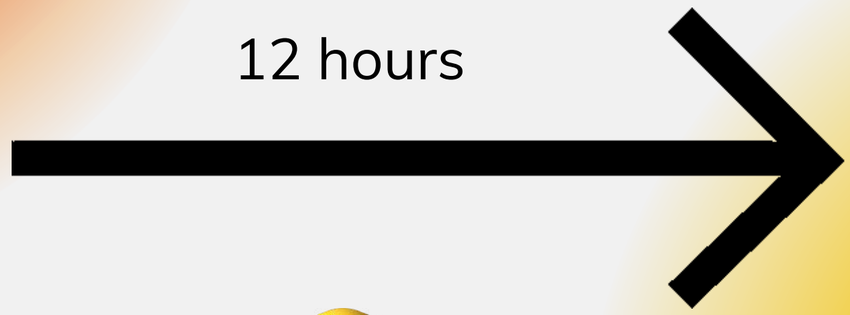
NLP

Present By: LAMYA ALAHYANE

Supervised by: Pr.Sanae Achas

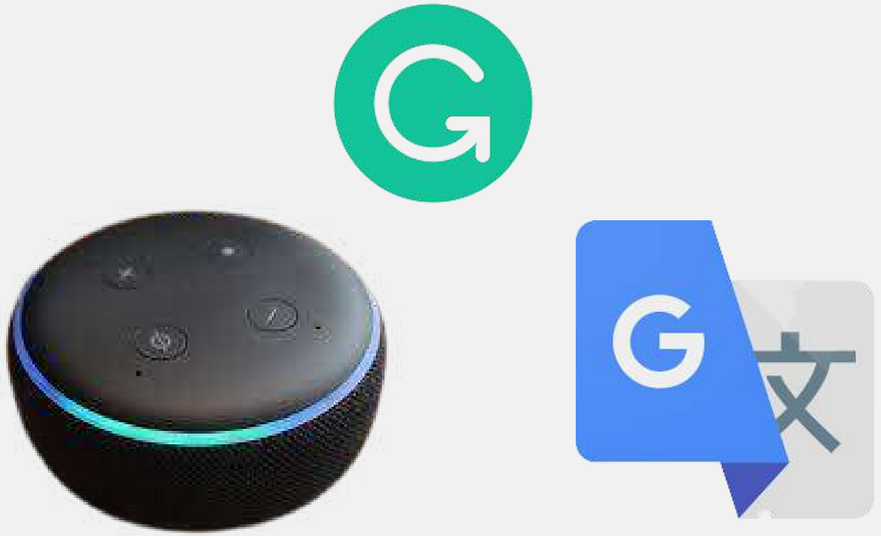
Summary

12 hours



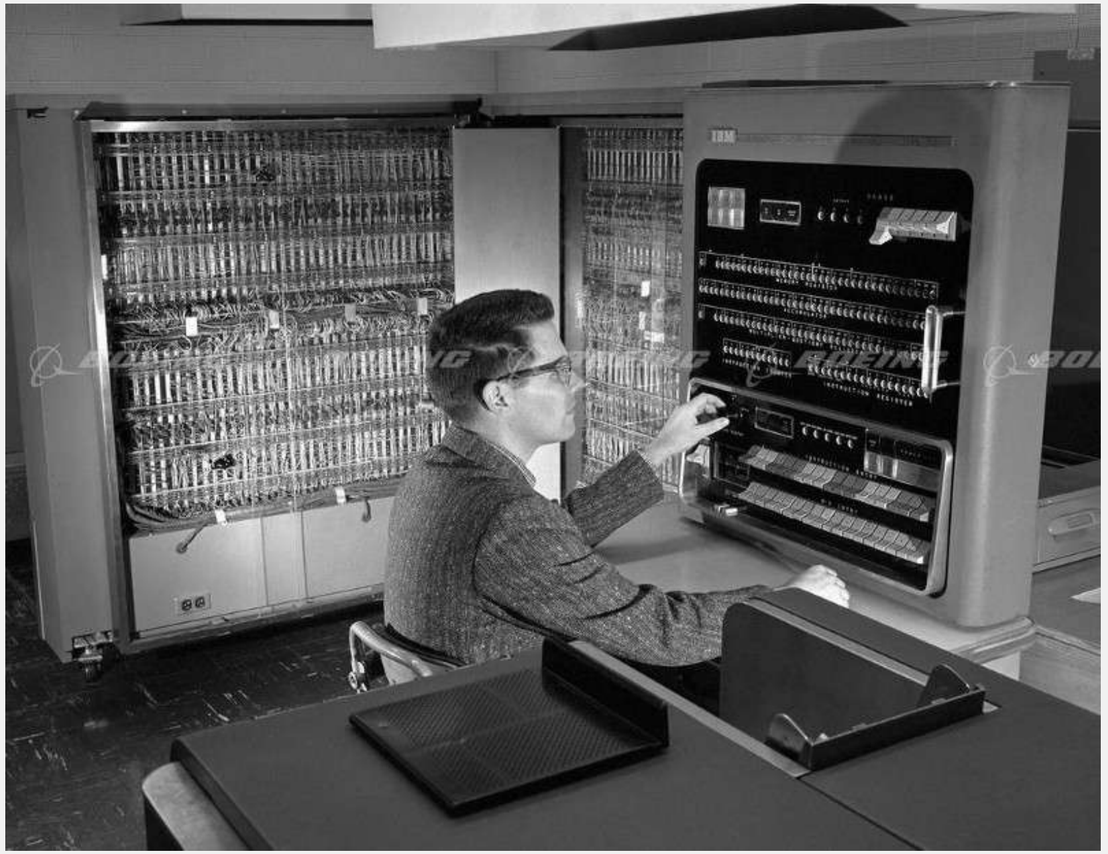
Introduction

Machine translation: sometimes referred to by the abbreviation MT is a very challenge task that investigates the use of software to translate text or speech from one language to another. Traditionally, it involves large statistical models developed using highly sophisticated linguistic knowledge.



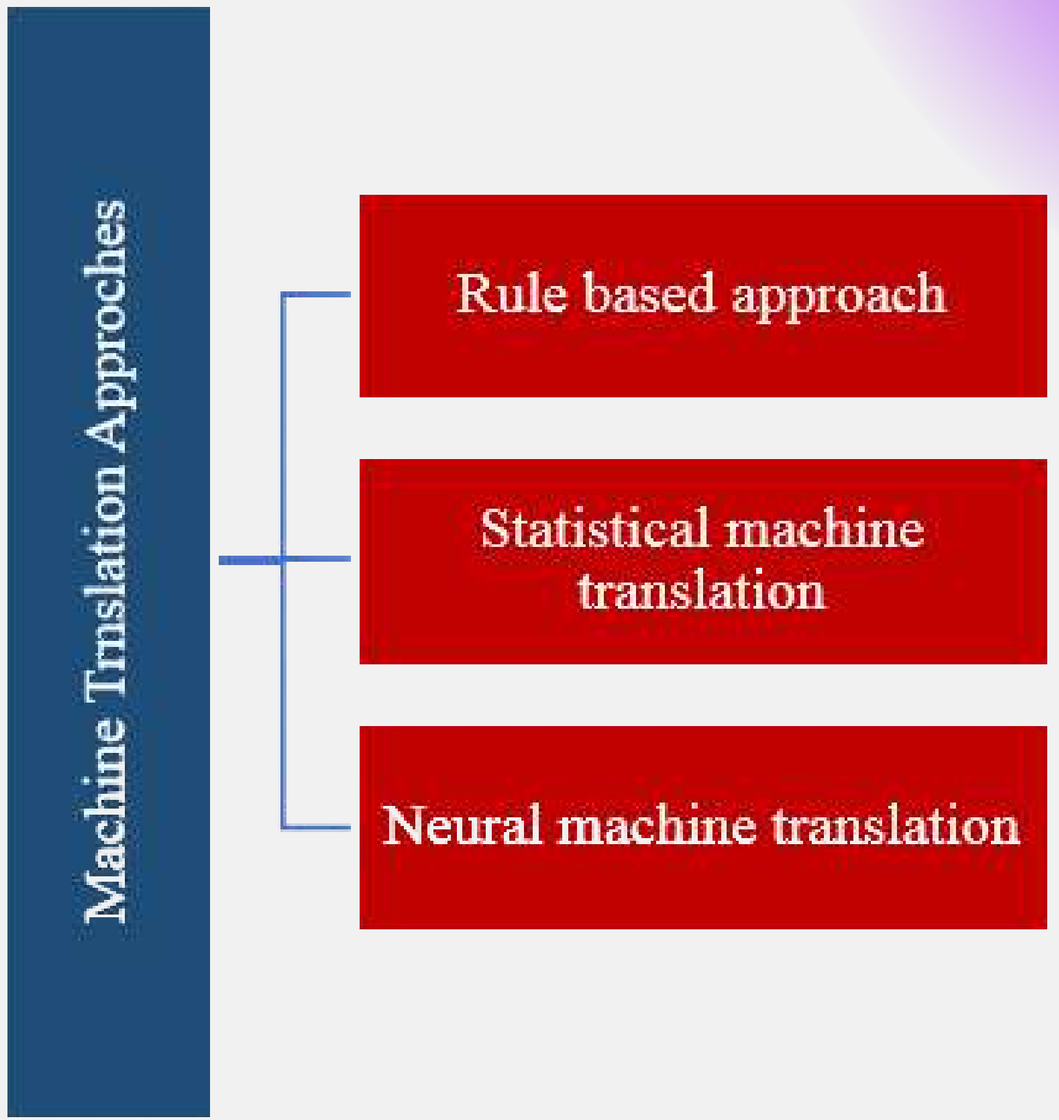
Machine translation from the Cold War to deep learning

On January 7th 1954, at IBM headquarters in New York, the Georgetown–IBM experiment started. The IBM 701 computer automatically translated 60 Russian sentences into English for the first time in history



what is Different types of machine translation?

Different types of machine translation:



Rule-Based Machine Translation

Rule-Based Machine Translation (RBMT) is an early approach to machine translation that relies on explicit linguistic rules and dictionaries. In RBMT, translation is performed based on a set of predefined linguistic rules that specify how words, phrases, and grammatical structures should be translated from the source language to the target language. Unlike Neural Machine Translation (NMT), which learns from data, RBMT relies on human-crafted rules and linguistic knowledge.

Here's an example to illustrate the rule-based approach:

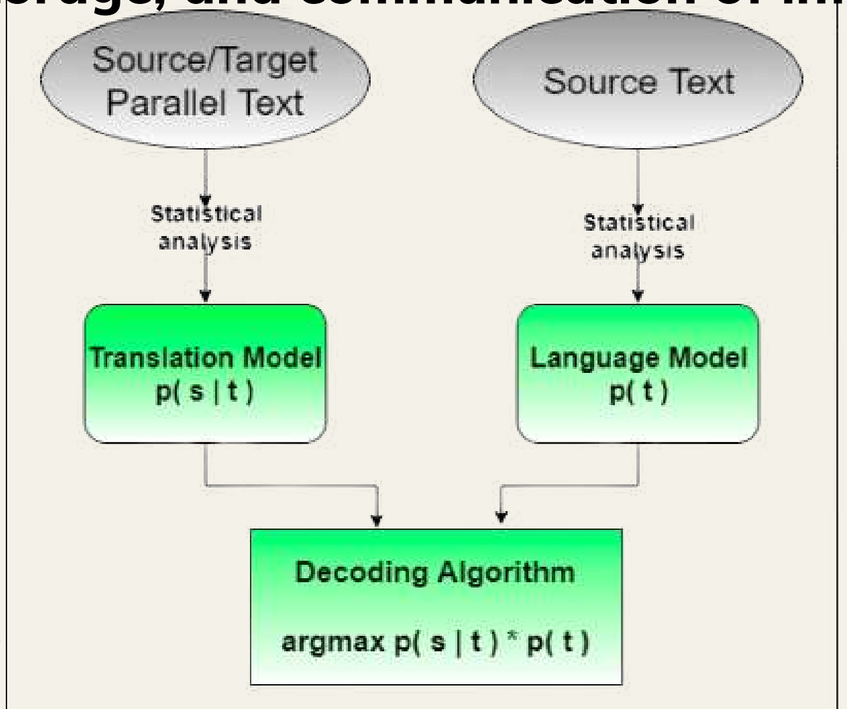
Let's say we have a rule in English-French translation that states: "If the word in English is 'cat' translate it to 'chat' in French."

Using this rule, the RBMT system would apply it when encountering the word "cat" in the source English text and translate it to "chat" in the target French text.

RBMT systems typically rely on extensive bilingual dictionaries and grammatical rules for each language pair. Linguistic experts create these rules and dictionaries to guide the translation process.

MT

Statistical Machine Translation (SMT) is a machine translation paradigm where translations are made on the basis of statistical models, the parameters of which are derived on the basis of the analysis of large volumes of a bilingual text corpus. The term bilingual text orpus refers to the collection of a large and structured set of texts written in two different languages. SMT is mostly based on Information Theory which studies the quantification, storage, and communication of information.



MT

i called ahmed but he did not .....

answer

well

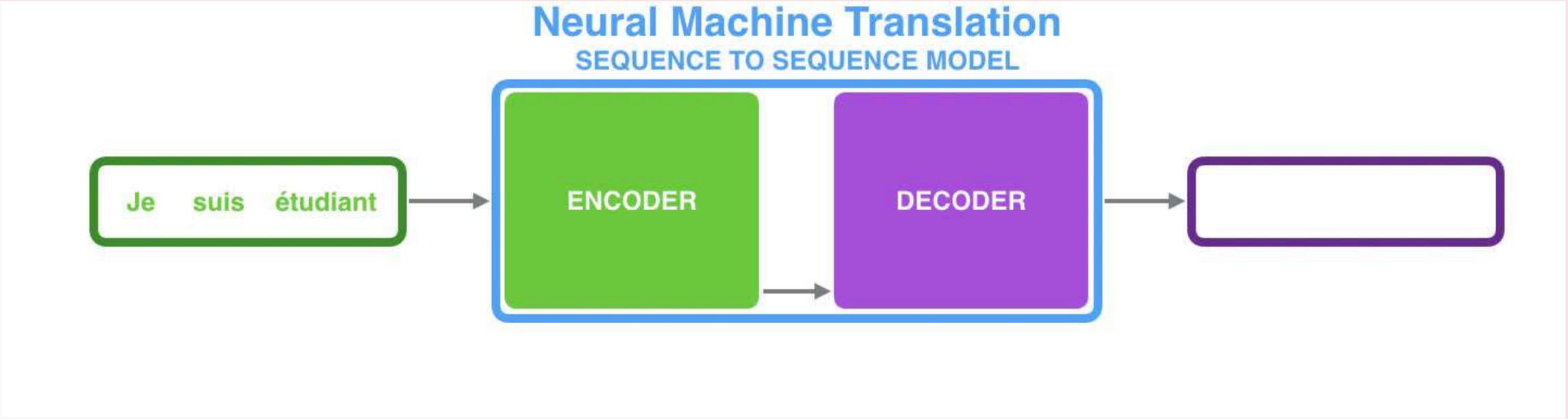
good

have

what is the solution?

Neural Machine Translation

- SMT Models are very complexes in nature
- NMT uses seq2seq Architeture which involves two RNN's
- Encoder: encode source sentences and fed this encoding to the decoder.
- Decoder:produces the target senteces in condition of the Encoded vector.

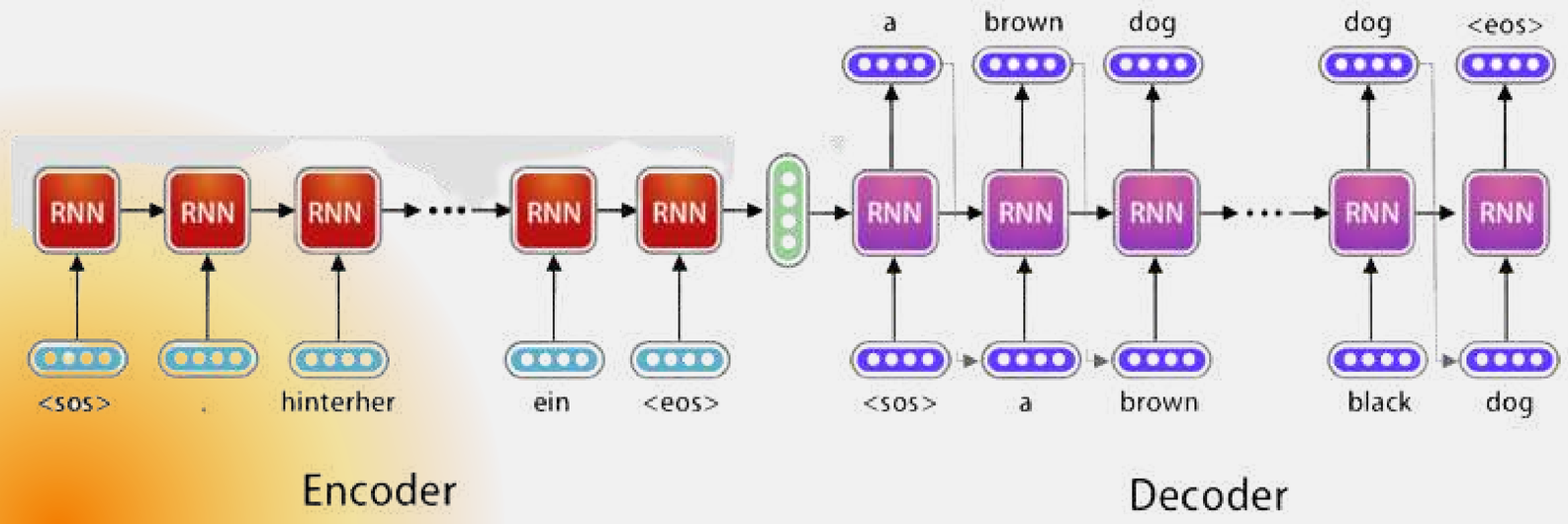


Is that all?

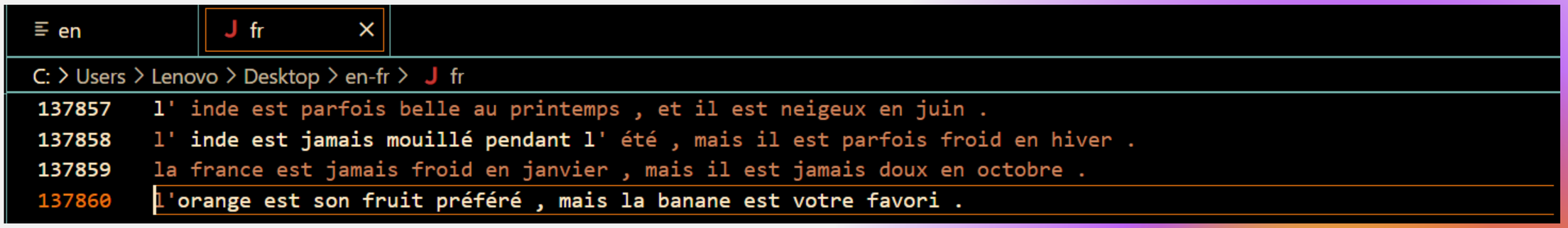
Long Short–Term Memory

LSTM (Long Short-Term Memory) is a type of recurrent neural network (RNN) architecture commonly used in Neural Machine Translation (NMT) models. NMT aims to translate text from one language to another using deep learning techniques.

LSTM is particularly well-suited for NMT due to its ability to effectively capture long-term dependencies in sequential data. In the context of language translation, sentences often have complex structures and dependencies between words that span across different parts of the sentence. LSTM's ability to retain and propagate information over long sequences makes it suitable for modeling such dependencies.

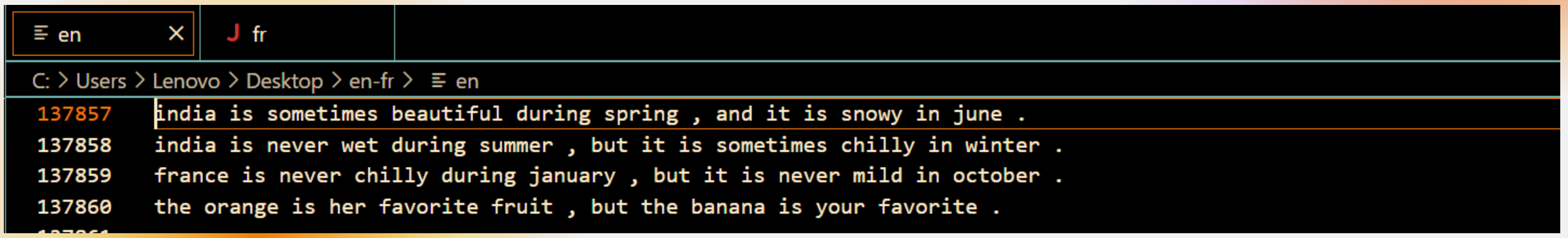


DATASET

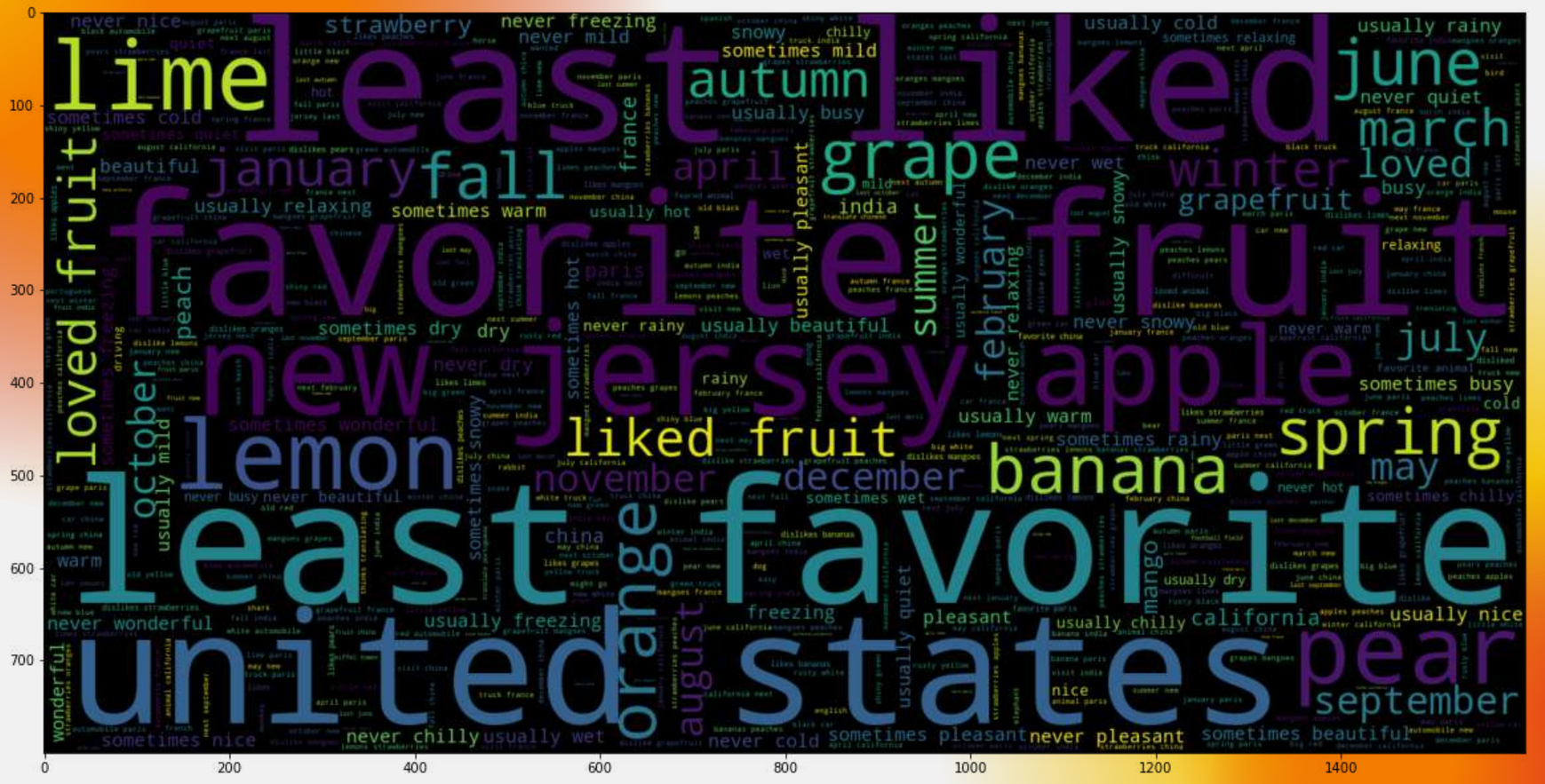


Each file contains an English sentence and its corresponding French sentence. The English sentence is the source sequence and French one is the target sequence.

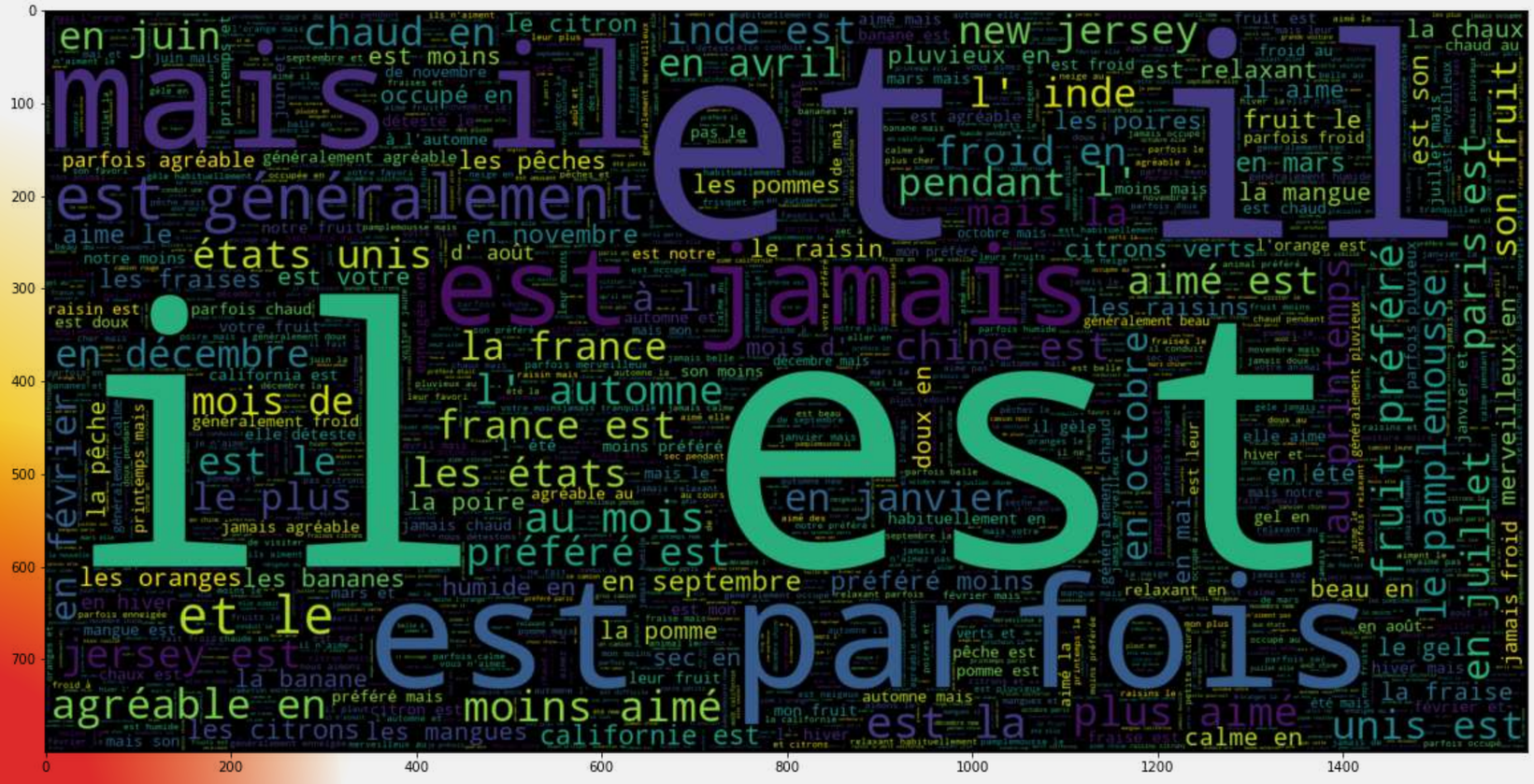
+100K line



Most frequent words representation using word cloud:



French representation:

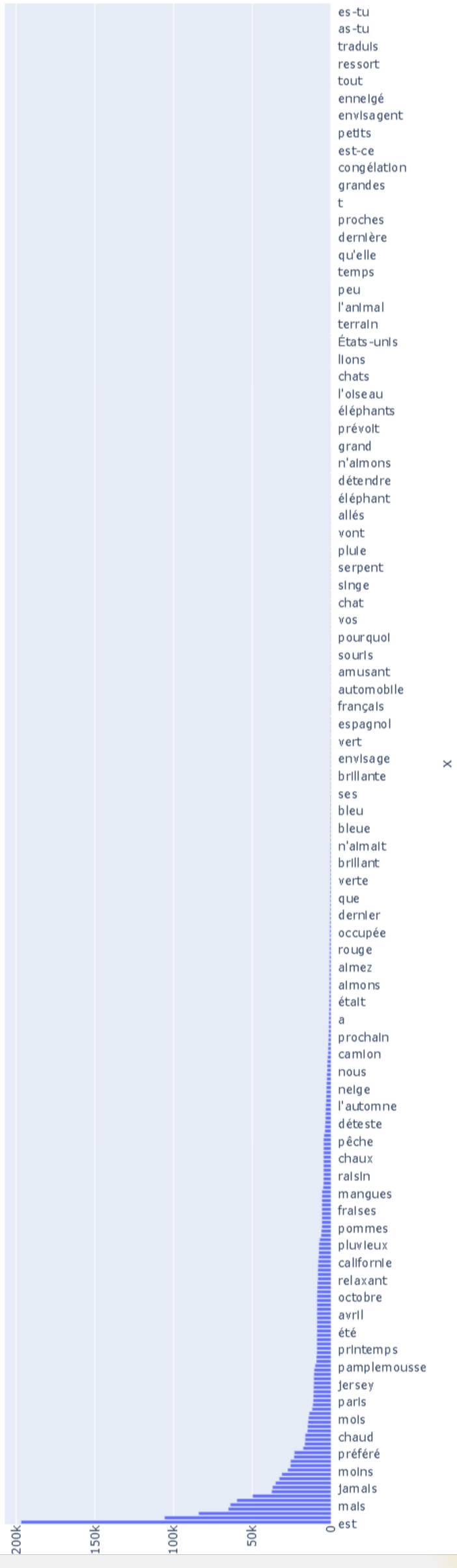


Preprocessing

Text Cleaning & tokenisation:

```python
# function to remove punctuations
def remove_punc(x):
    return re.sub('[!#?,.:";]', '', x)
```

```
['new',
 'jersey',
 'est',
 'parfois',
 'calme',
 'pendant',
 "l'",
 'automne',
 'et',
 'il',
 'est',
 'neigeux',
 'en',
 'avril',
 'les',
 'états-unis',
 'est',
 'généralement',
 'froid',
 'en',
 'juillet',
 'et',
 'il',
 'gèle',
 'habituellement',
 'en',
 'novembre',
 'california',
 'est',
 'généralement',
 'calme',
 'en',
 'mars',
 'et',
 'il',
 'est',
 'généralement',
 'chaud',
 'en',
 'juin',
 'les',
 'états-unis',
 'est',
 'parfois',
```


es-tu
as-tu
traduis
ressort
tout
ennelgé
envlsagent
petlts
est-ce
congélatlon
grandes
t
proches
dernière
qu'elle
temps
peu
l'animal
terraln
États-unls
llons
chats
l'olseau
éléphants
prévolt
grand
n'almons
détendre
éléphant
allés
vont
pluie
serpent
slnge
chat
vos
pourquol
souris
amusant
automoblle
français
espagnol
vert
envlsage
brillante
ses
bleu
bleue
n'almalt
brillant
verte
que
dernier
occupée
rouge
almez
almons
était
a
prochaln
camlon
nous
nelge
l'automne
déteste
pêche
chaux
raisln
mangues
fraises
pommes
pluvieux
calìfornie
relaxant
octobre
avril
été
printemps
pamplemousse
jersey
paris
mois
chaud
préféré
molns
jamais
mals
est
x
200k
150k
100k
50k
0

Model.

Neural Machine Translation

Model: "sequential"

Layer (type)	Output Shape	Param #
embedding (Embedding)	(None, 23, 256)	51200
lstm (LSTM)	(None, 256)	525312
repeat_vector (RepeatVector)	(None, 23, 256)	0
lstm_1 (LSTM)	(None, 23, 256)	525312
time_distributed (TimeDistri	(None, 23, 351)	90207

Total params: 1,192,031
Trainable params: 1,192,031
Non-trainable params: 0

accuracy:

```
loss: 0.6395 - accuracy: 0.8176
```

Results

Technologies Used:

- in addition to the construction of the model i used **Django** to deploy my model as web page .
- Django is a powerful and popular Python web framework that provides the necessary tools and functionality to build robust web applications.
- By utilizing Django, we will be able to develop a user-friendly and interactive website that allows users to input English text and obtain the corresponding French translation.

Predicting 100 sentences :

```
Original English word   china is pleasant during september and it is mild in april

Original French word - chine est agréable au mois de septembre et il est doux en avril

Predicted French word - chine est chaud en septembre de il est est en en avril
```

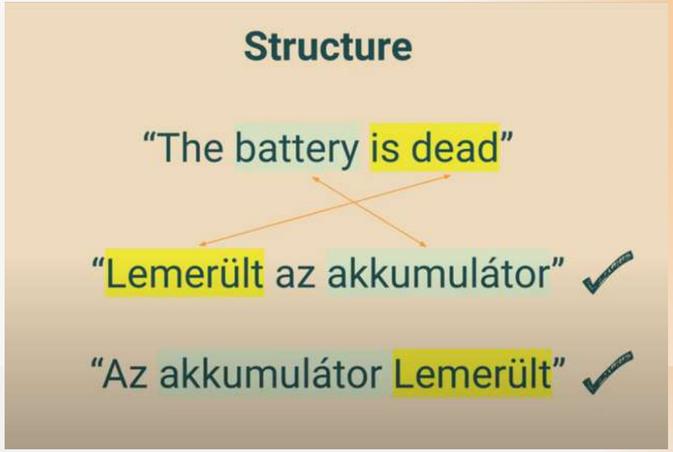
Original English word - he liked a little black car

Original French word - il aimait une petite voiture noire

Predicted French word - elle aimait une petite voiture blanche

Conclusion

the Language Translator using NLP and Neural network Model is an effective and accurate way to translate text from Language to another but we may have consequences such as complexity of models ,complexity of language (gramical level ,idioms...)



References

Visualizing A Neural Machine Translation Model (Mechanics of Seq2seq Models With Attention)

Translations: Chinese (Simplified), French, Japanese, Korean, Persian, Russian, Turkish Watch: MIT's Deep Learning State of the Art lecture referencing this post May 25th update: New graphics (RNN animation,...

github.io /

Python for NLP: Word Embeddings for Deep Learning in Keras

This is the 16th article in my series of articles on Python for NLP. In my previous article I explained how N-Grams technique can be used to develop a simple au...

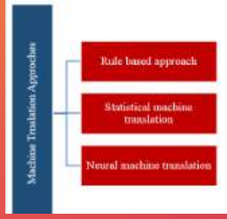
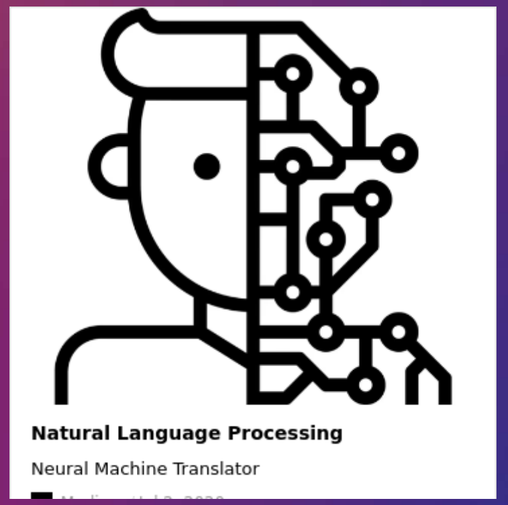


Fig. 2: Machine translation approaches

Download scientific diagram | Machine translation approaches from publication: Transferring Informal...

ResearchGate



Natural Language Processing

Neural Machine Translator

Python for NLP: Word Embeddings for Deep Learning in Keras

This is the 16th article in my series of articles on Python for NLP. In my previous article I explained how N-Grams technique can be used to develop a simple au...

Thank you!