НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Кафедра прикладной математики и информатики

Компьютерная лингвистика

#8

Нижний Новгород, 2019

Лабораторная работа #9.

doc2vec

Задача.

1. Обучить модель doc2vec на собственном корпусе
2. Построить классификатор тем
3. Добиться максимального качества

```
documents_train = [
    gensim.models.doc2vec.TaggedDocument(tokens, [i]) for i, tokens in enumerate(corpus)
]

model = gensim.models.Doc2Vec(
    documents_train, vector_size=100, window=5, min_count=5, workers=4
)

...

vector = model.infer_vector(tokens)
```

Лекция #8.

- Машинный перевод

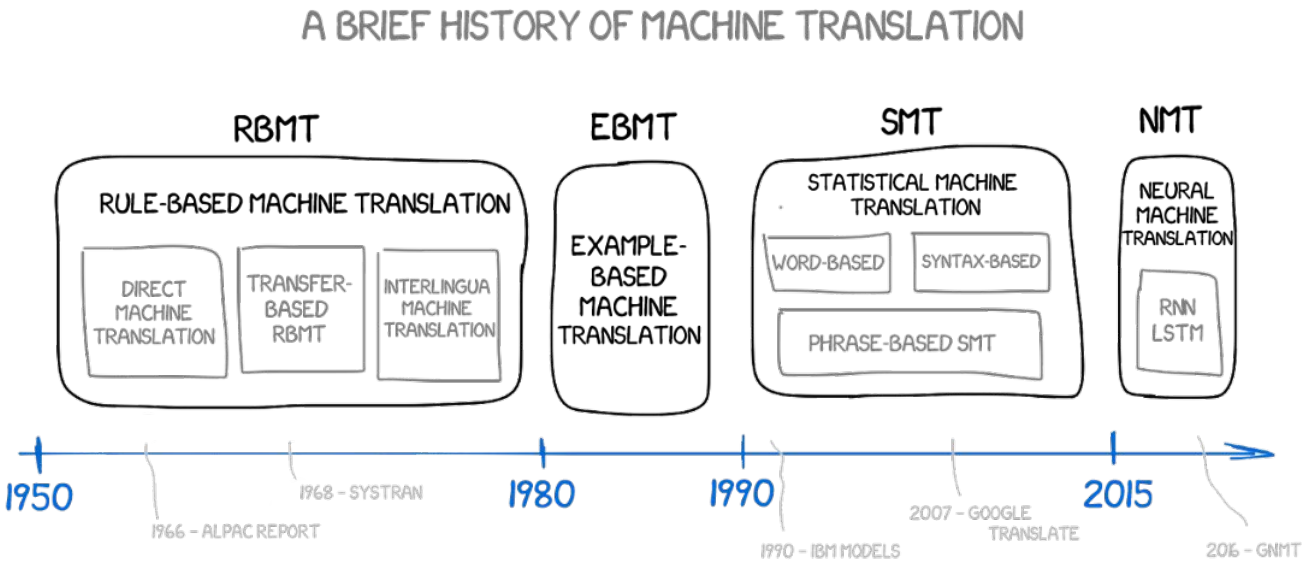


image credits:
https://medium.freecodecamp.org/a-history-of-machine-translation-from-the-cold-war-to-deep-learning-f1d335ce8b5

Rule-based machine translation.

- Direct

image credits:
https://medium.freecodecamp.org/a-history-of-machine-translation-from-the-cold-war-to-deep-learning-f1d335ce8b5

Rule-based machine translation

- Transfer-based

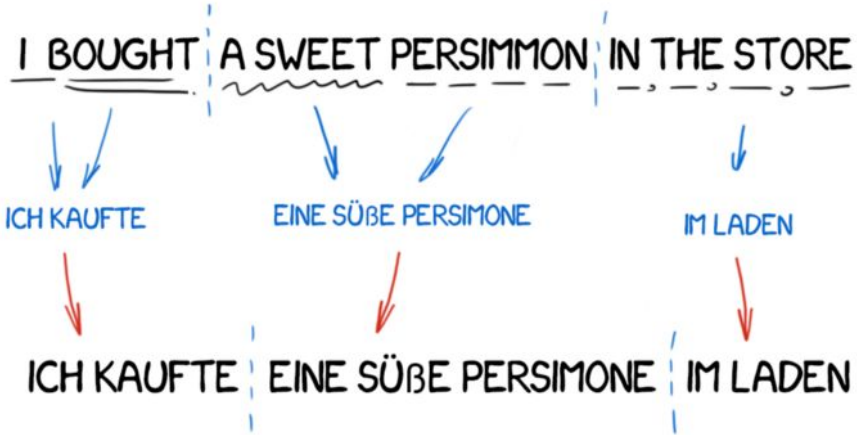


image credits:
https://medium.freecodecamp.org/a-history-of-machine-translation-from-the-cold-war-to-deep-learning-f1d335ce8b5

Example-based machine translation

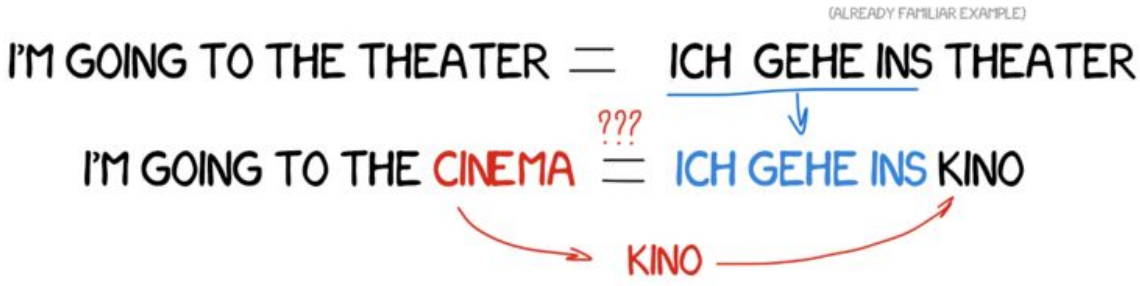


image credits:
https://medium.freecodecamp.org/a-history-of-machine-translation-from-the-cold-war-to-deep-learning-f1d335ce8b5

Syntax-based machine translation

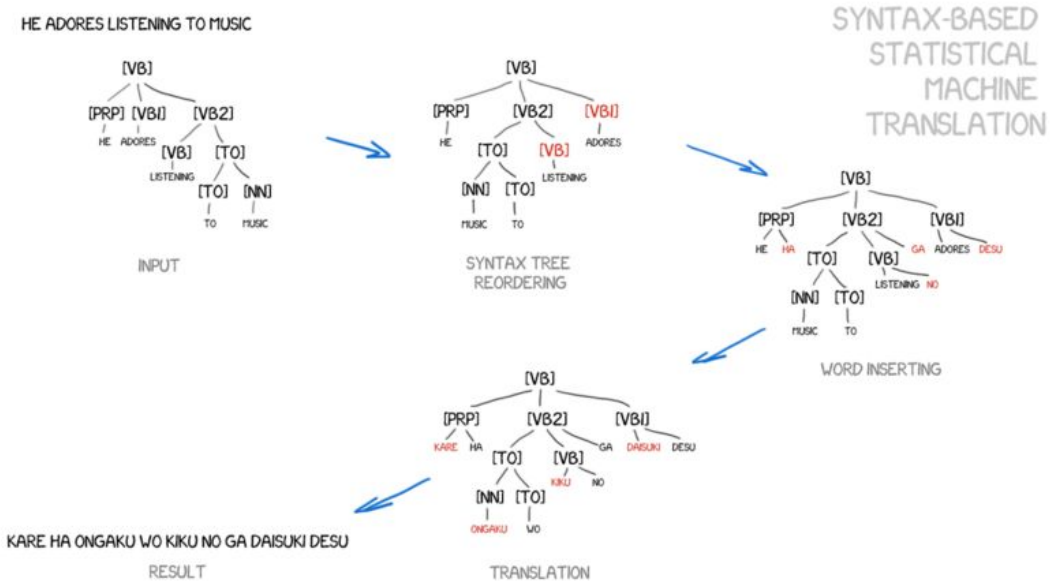
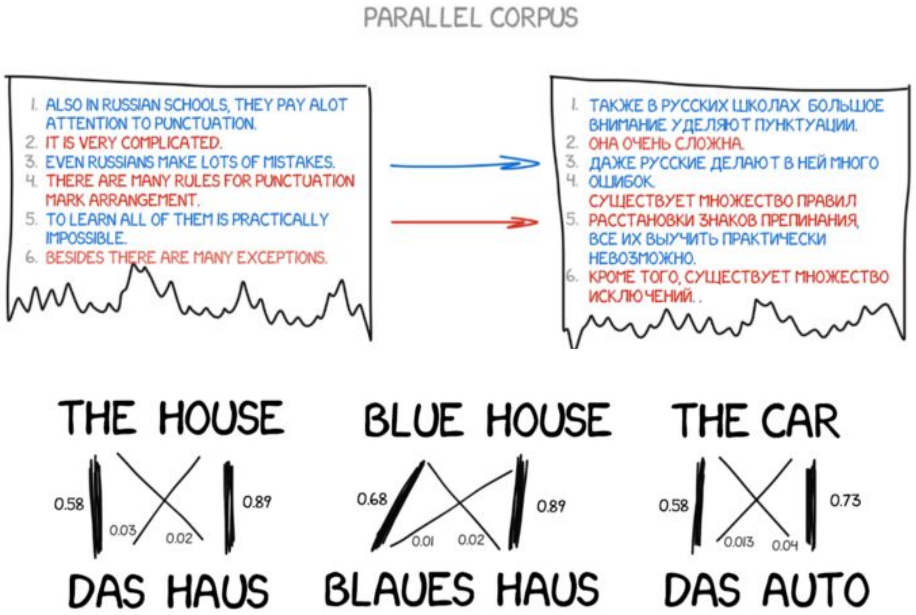
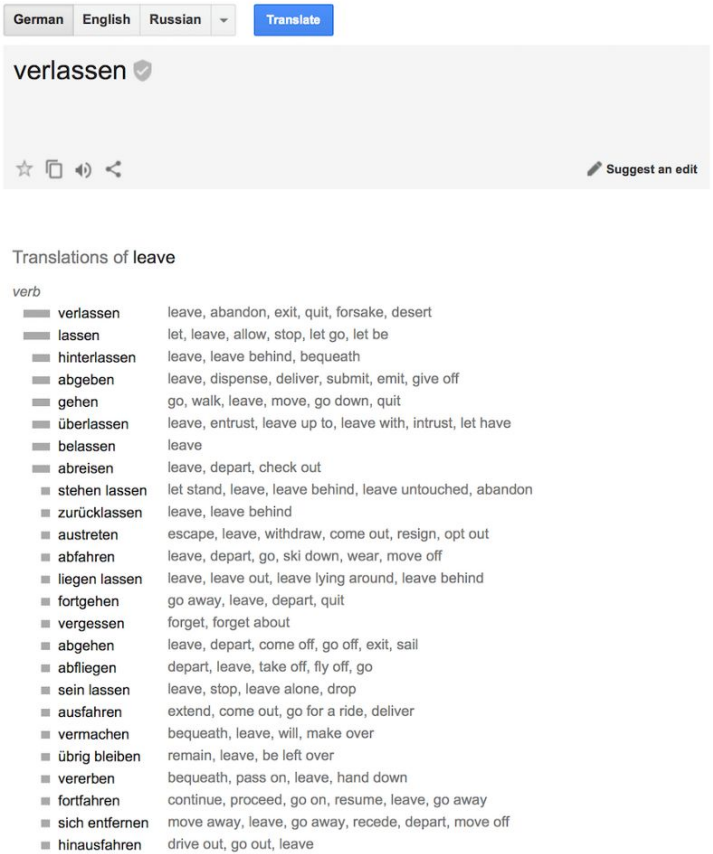


image credits:
https://medium.freecodecamp.org/a-history-of-machine-translation-from-the-cold-war-to-deep-learning-f1d335ce8b5

Statistical machine translation





Statistical machine translation

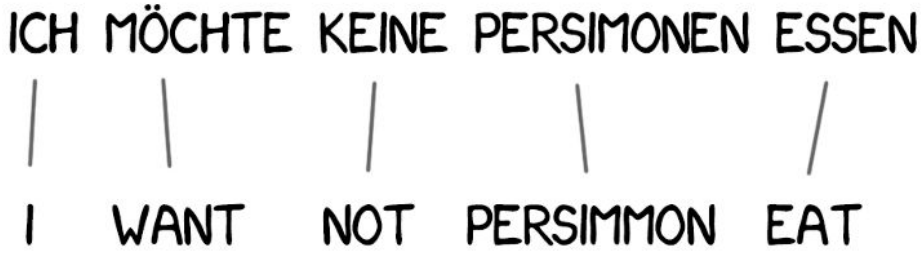


image credits:
https://medium.freecodecamp.org/a-history-of-machine-translation-from-the-cold-war-to-deep-learning-f1d335ce8b5

Statistical machine translation

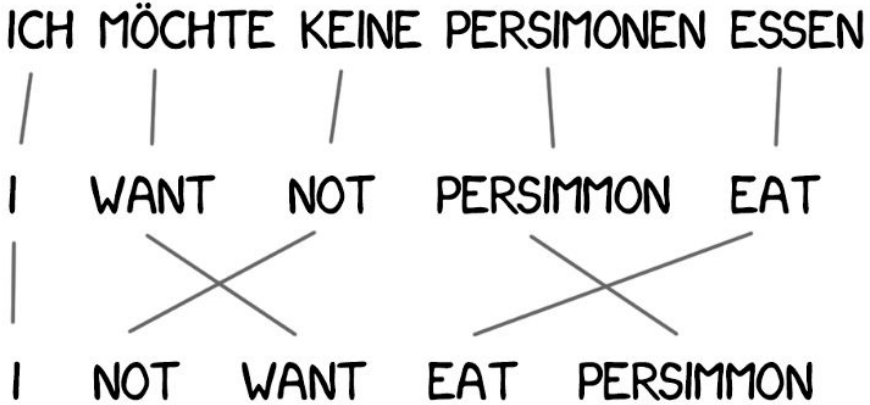


image credits:
https://medium.freecodecamp.org/a-history-of-machine-translation-from-the-cold-war-to-deep-learning-f1d335ce8b5

Statistical machine translation

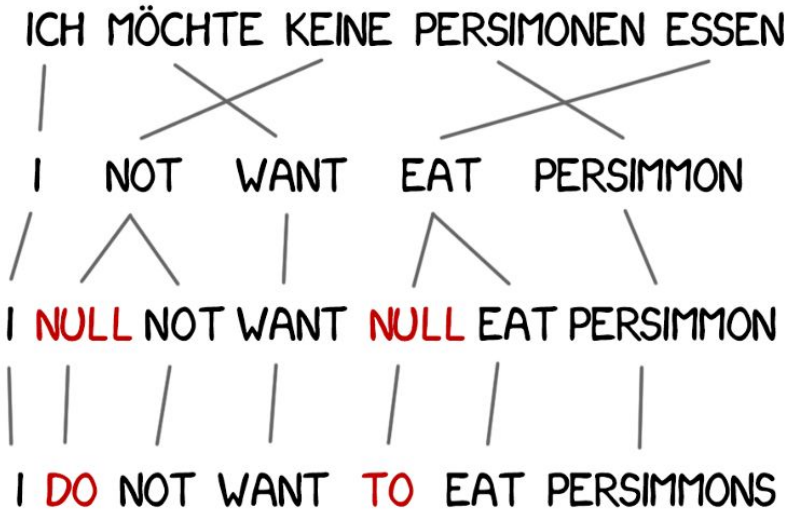


image credits:
https://medium.freecodecamp.org/a-history-of-machine-translation-from-the-cold-war-to-deep-learning-f1d335ce8b5

Statistical machine translation

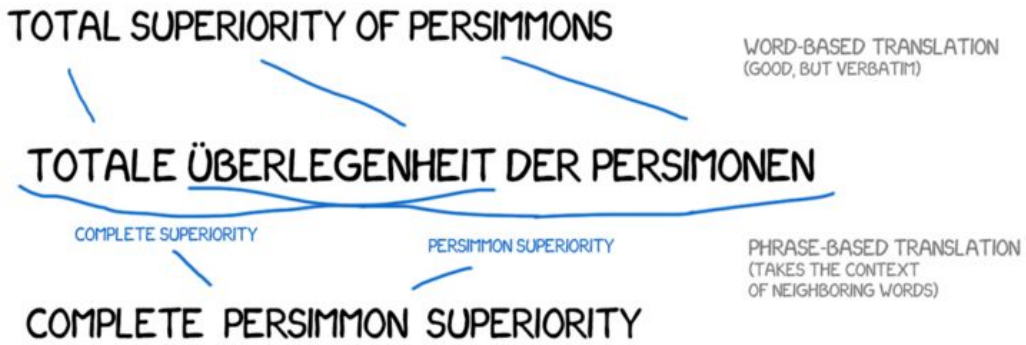


image credits:
https://medium.freecodecamp.org/a-history-of-machine-translation-from-the-cold-war-to-deep-learning-f1d335ce8b5

Neural machine translation

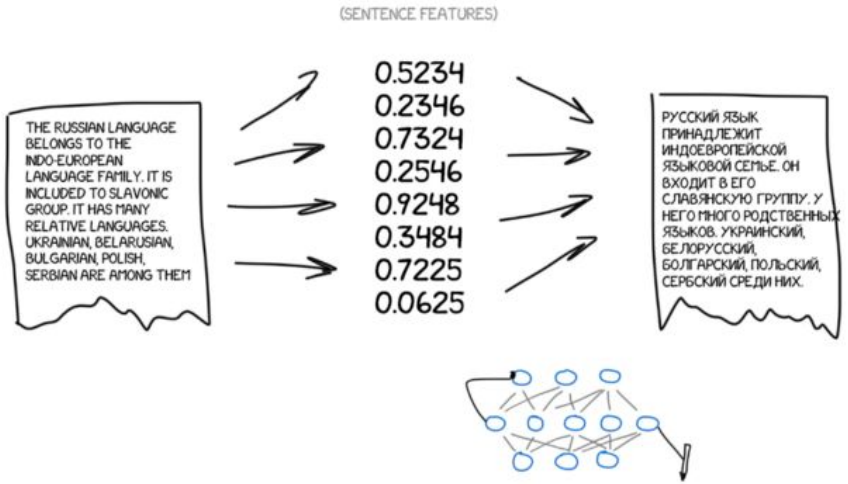
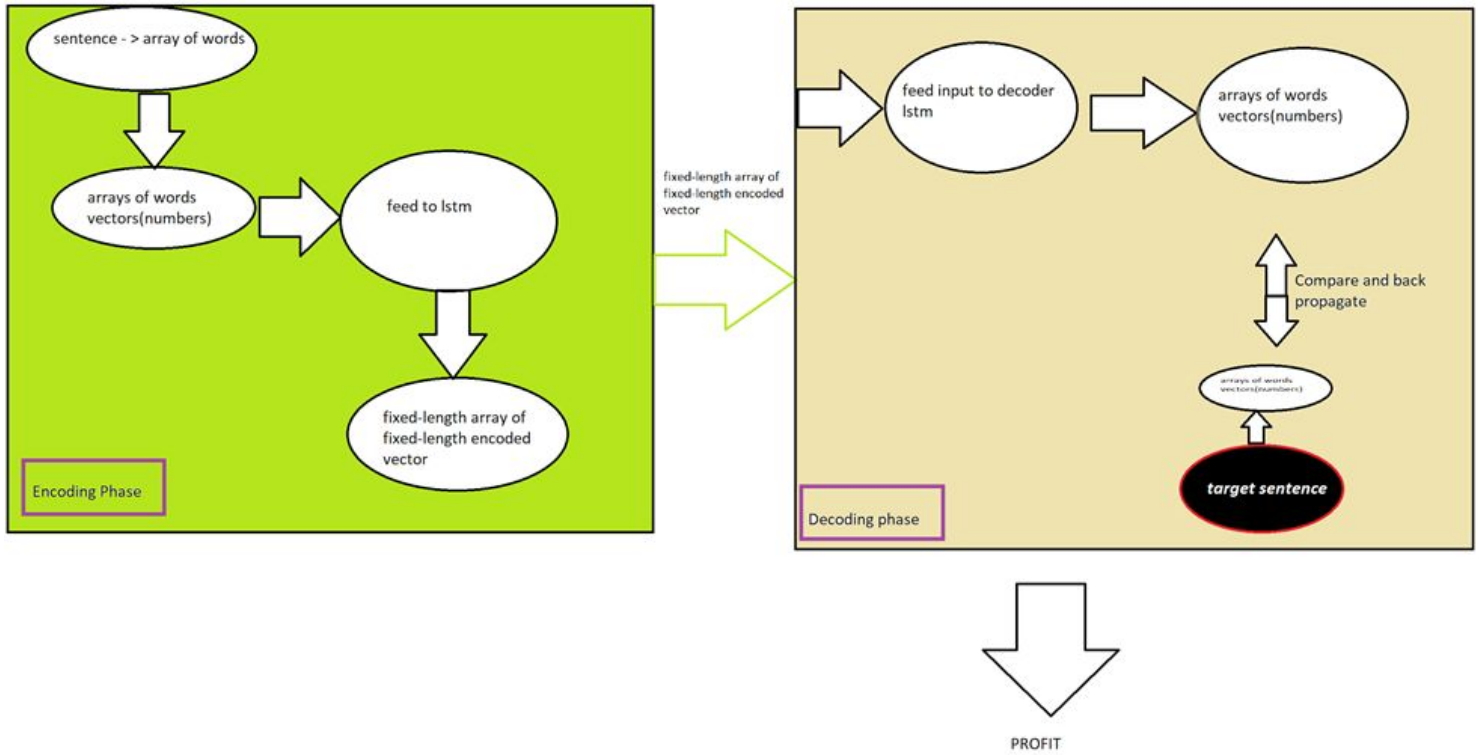
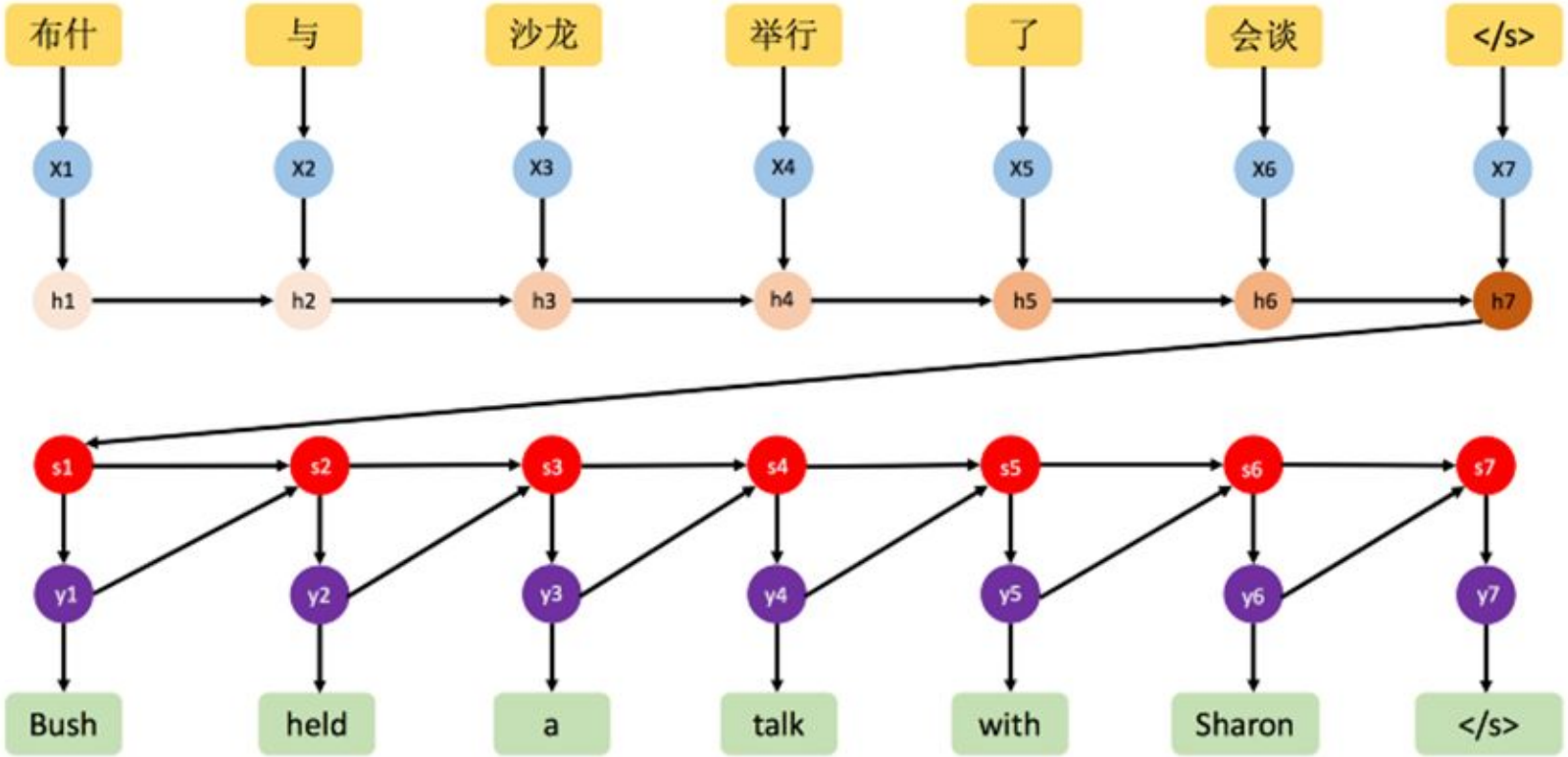


image credits:
https://medium.freecodecamp.org/a-history-of-machine-translation-from-the-cold-war-to-deep-learning-f1d335ce8b5


sentence - > array of words
arrays of words vectors(numbers)
feed to lstm
fixed-length array of fixed-length encoded vector
Encoding Phase
fixed-length array of fixed-length encoded vector
feed input to decoder lstm
arrays of words vectors(numbers)
Compare and back propagate
target sentence
Decoding phase
PROFIT



(Sutskever et al., 2014)

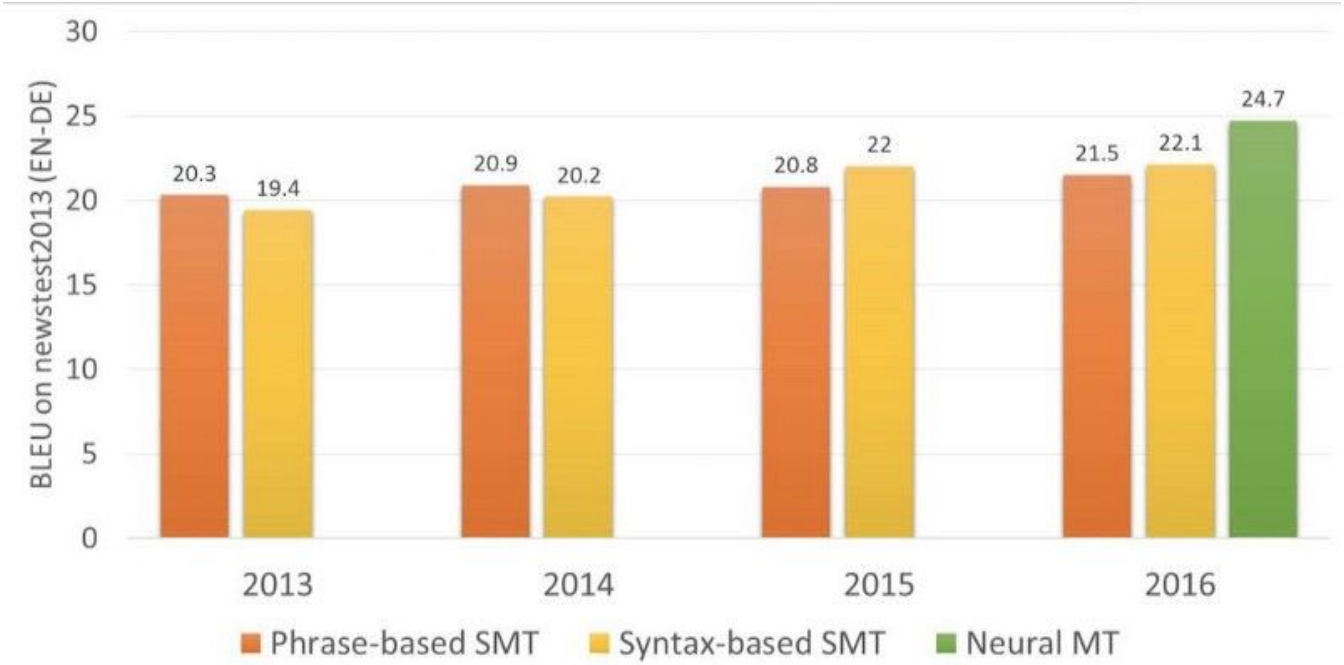


image credits:
https://medium.freecodecamp.org/a-history-of-machine-translation-from-the-cold-war-to-deep-learning-f1d335ce8b5

	Neural Machine Translation	Statistical Machine Translation
Training time	More	Less
Training data	Less	More
Translation (decoding) time	More	Less
CPU usage	More	Less
Space in disk	Less	More
Mechanism	Sentence by sentence	Word by word/ phrase by phrase
	Attentional encoder-decoder networks; optimization	Statistical analysis; probability
	Train multiple features jointly	Feature engineering required
Interpretability		👍
Long distance reordering	👍	
Morphology, syntax, and agreement errors	👍	
Translation style consistency for the same word		👍
Tolerance to noisy data	👍	
Multilingual/ multi-domain translation	👍	
Vocabulary/Rare word Problem		👍

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ