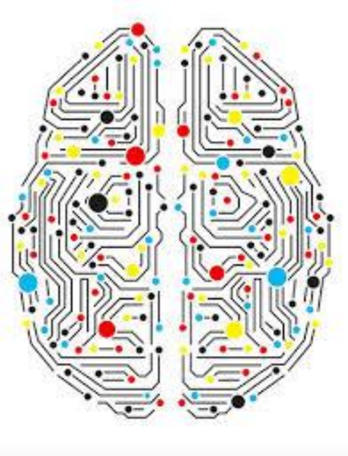
What do Neural Machine Translation Models Learn About Morphology

Yonatan Belinkov Nadir Durrani Fahim Dalvi
Hassan Sajjad James Glass



- *Presented by* Raghav Gurbaxani

FROM ACL 2017

Motivation

- In recent times Neural Machine Translation has obtained state of the art results.
- Simple and Elegant architecture.
- However, models are difficult to interpret.

Introduction

- **Goal**: analyze the representations learned by neural MT models at various levels of granularity
- In this work we analyze morphology in NMT.
- **Morphology**: study of word forms (“run”, “runs”, “ran”)
- Important when translating between many languages to preserve semantic knowledge.

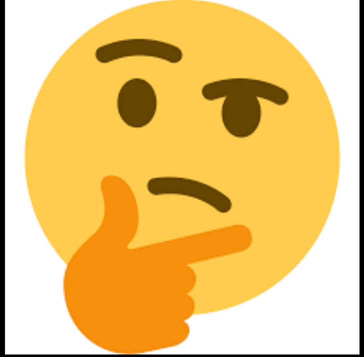
FROM ACL 2017

Questions that we need to examine?

- What do NMT models learn about word morphology?
- What is the effect on learning when translating into/from morphologically-rich languages?
- What impact do different representations (character vs. word) have on learning?
- What do different modules learn about the syntactic and semantic structure of a language?

FROM ACL 2017

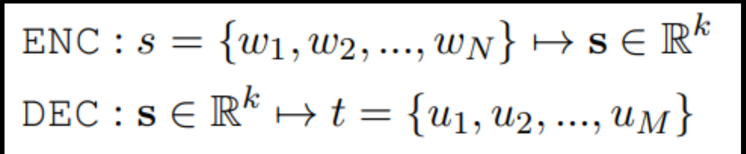
Even More Questions



- Which parts of the NMT architecture capture word structure?
- What is the division of labor between different components (e.g. different layers or encoder vs. decoder)?
- How do different word representations help learn better morphology and modeling of infrequent words?
- How does the target language affect the learning of word structure?

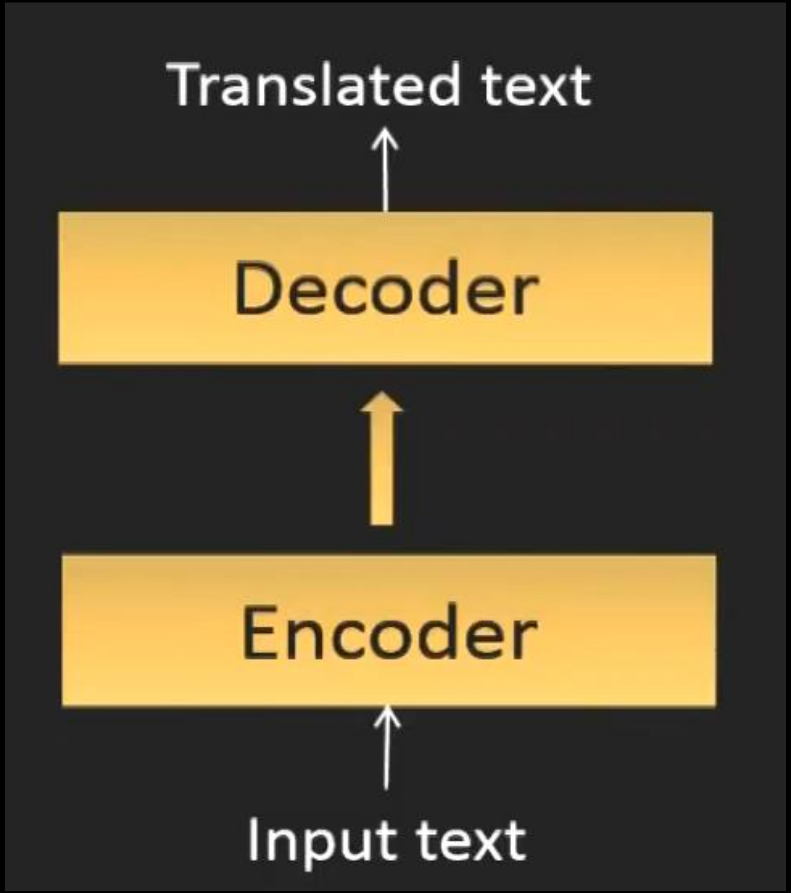
FROM ACL 2017

Generic Neural Machine Translation Architecture



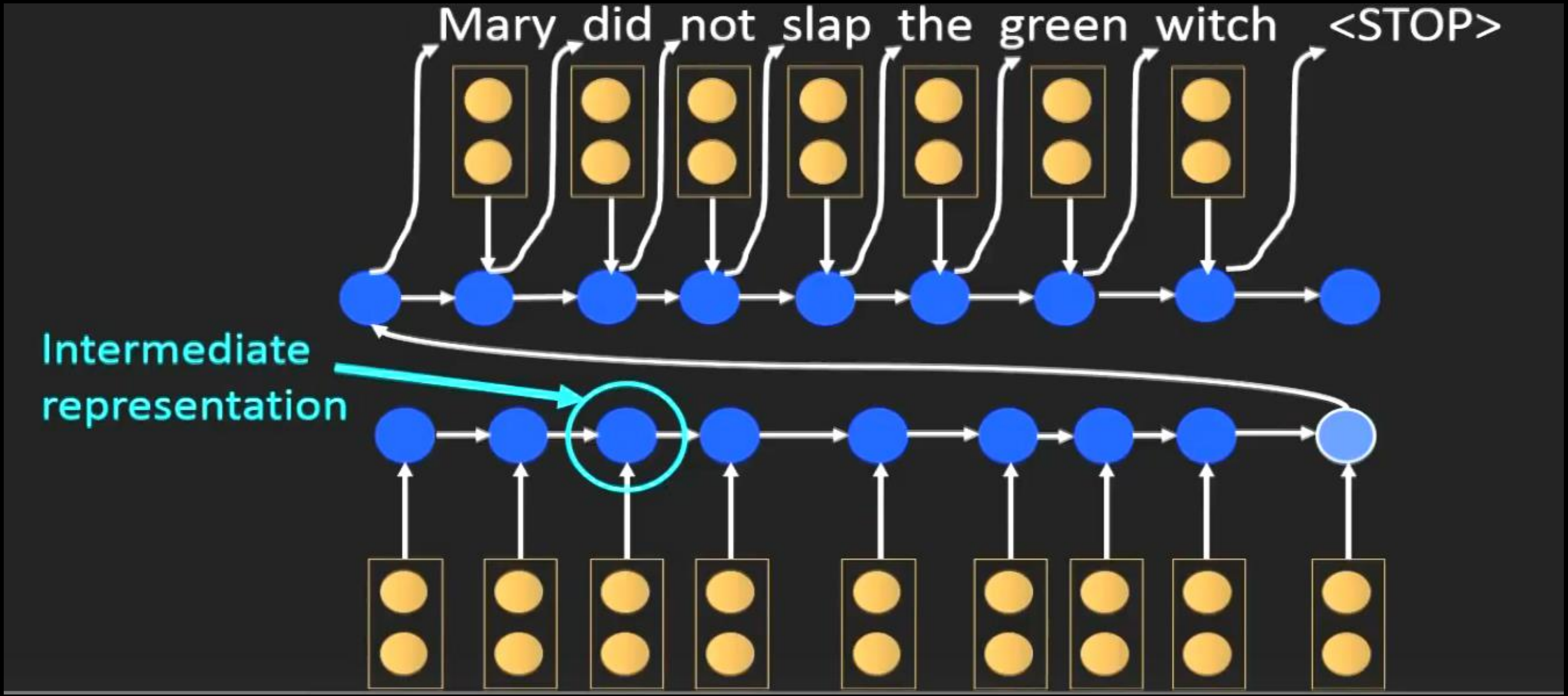
$$\mathrm{ENC} : s = \{w_1, w_2, ..., w_N\} \mapsto \mathbf{s} \in \mathbb{R}^k$$

$$\mathrm{DEC} : \mathbf{s} \in \mathbb{R}^k \mapsto t = \{u_1, u_2, ..., u_M\}$$



FROM ACL 2017

NMT Architecture (representation)



FROM ACL 2017

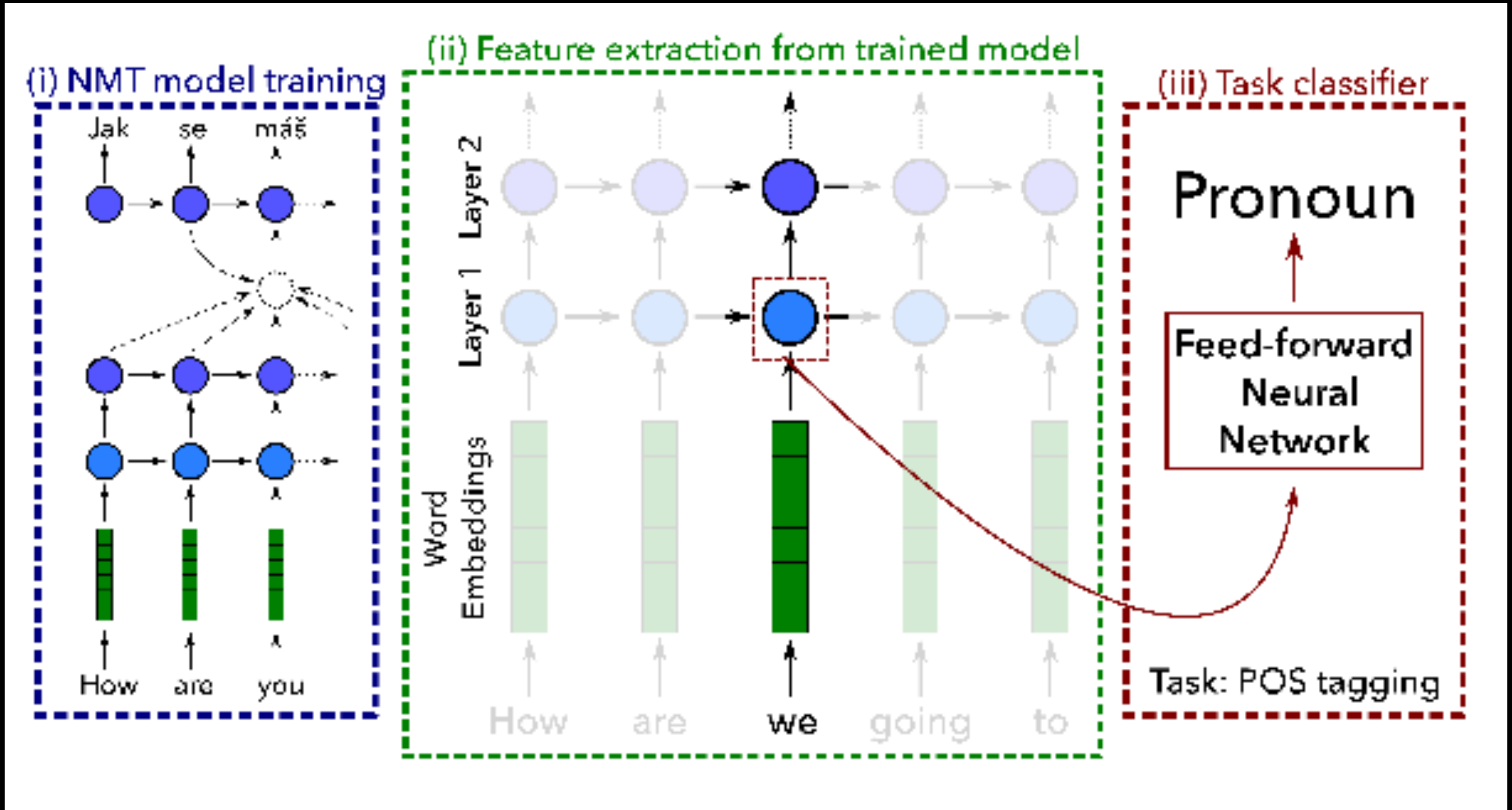
Experimental Methodology

- The experiment follows the three following steps -
    1. Train a Neural Machine Translation System.
    2. Extract feature representations using the trained model.
    3. Train a classifier using the extracted model and evaluate it on an extrinsic task.

- **Assumption –** performance of classifier reflects quality of NMT representations for a given task.

FROM ACL 2017

Model Used in the Paper



FROM ACL 2017

Experimental Setup

- Take a trained NMT model and evaluate on tasks.

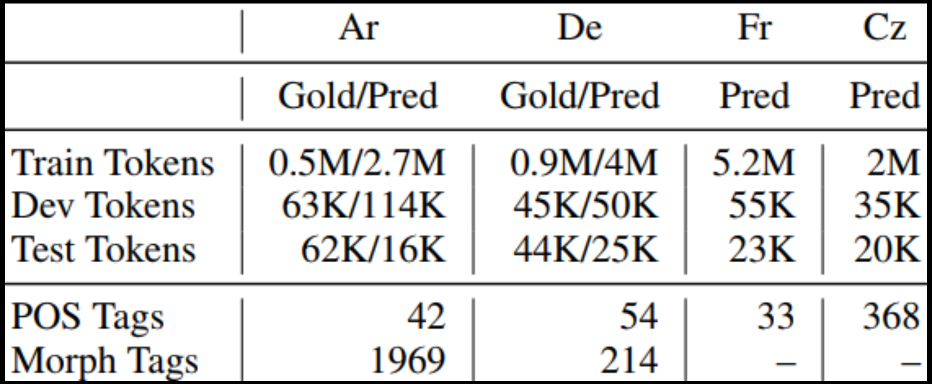
- Use features from NMT on Evaluation Tasks:
    1. Parts of Speech Tagging (“runs”=verb)
    2. Morphological Tagging (“runs”=verb, present tense, 3rd person, singular).

- Try Languages:
    1. Arabic-, German-, French-, etc.
    2. Arabic – Hebrew (rich and similar).
    3. Arabic – German (rich and different).

FROM ACL 2017

Datasets

- Experiment with language pairs, including morphologically-rich languages, Arabic-, German-, French-, and Czech-English pairs (on both Encoder and Decoder sides).
- Translation models are trained on the WIT3 corpus of TED talks made available for IWSLT 2016.
- For classification (POS tagging) they use gold annotated datasets and predicted tags used freely available taggers.



	Ar	De	Fr	Cz
	Gold/Pred	Gold/Pred	Pred	Pred
Train Tokens	0.5M/2.7M	0.9M/4M	5.2M	2M
Dev Tokens	63K/114K	45K/50K	55K	35K
Test Tokens	62K/16K	44K/25K	23K	20K
POS Tags	42	54	33	368
Morph Tags	1969	214	–	–

Statistics for annotated corpora in Arabic (Ar), German (De), French (Fr), and Czech (Cz)

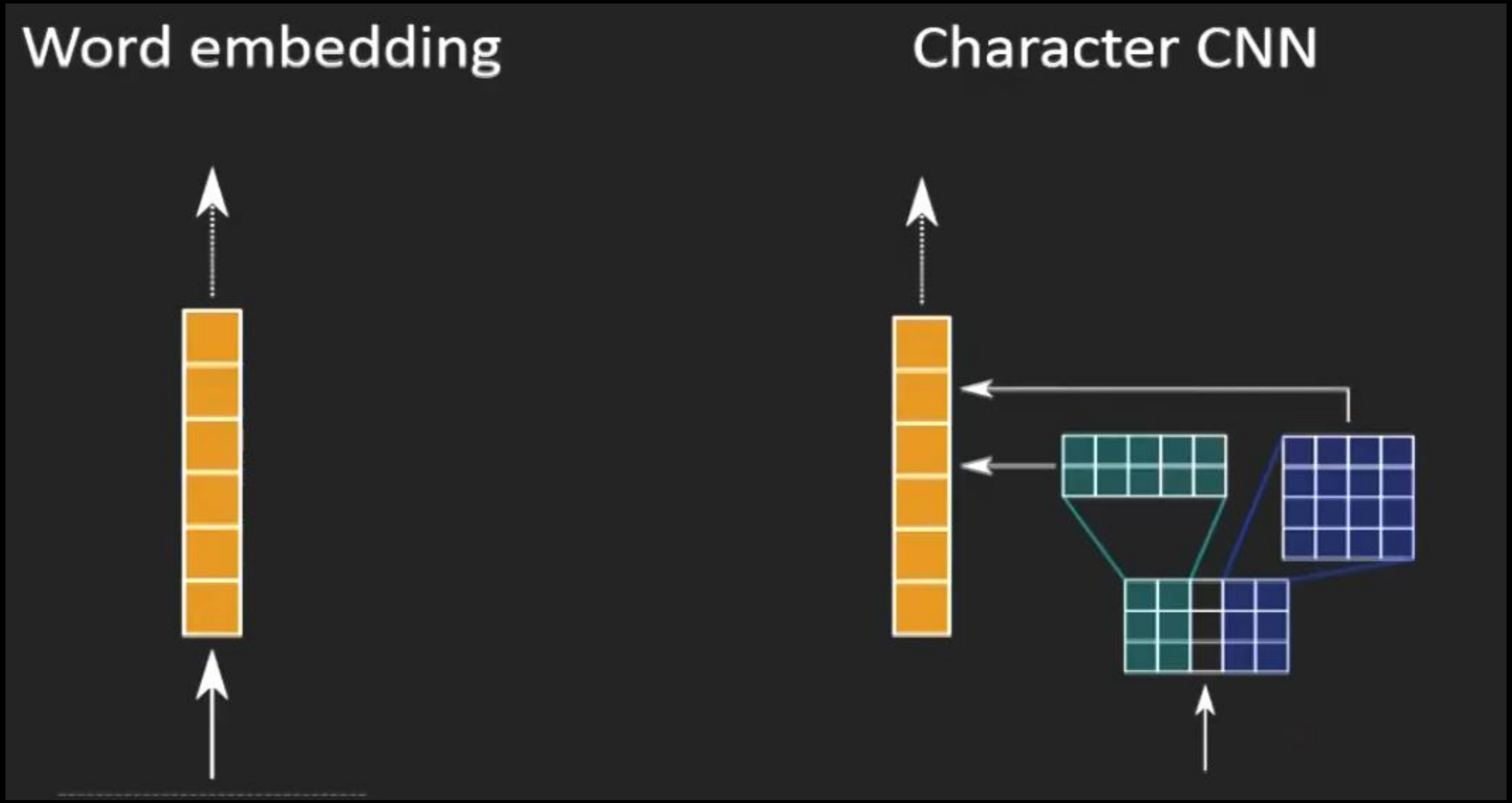
FROM ACL 2017

Encoder Analysis

- We will look at the following tasks –
  1. Effect of word representation
  2. Impact of word frequency
  3. Effect of encoder depth
  4. Effect of target language
  5. Analyzing specific tags

FROM ACL 2017

I. Effect of word representation

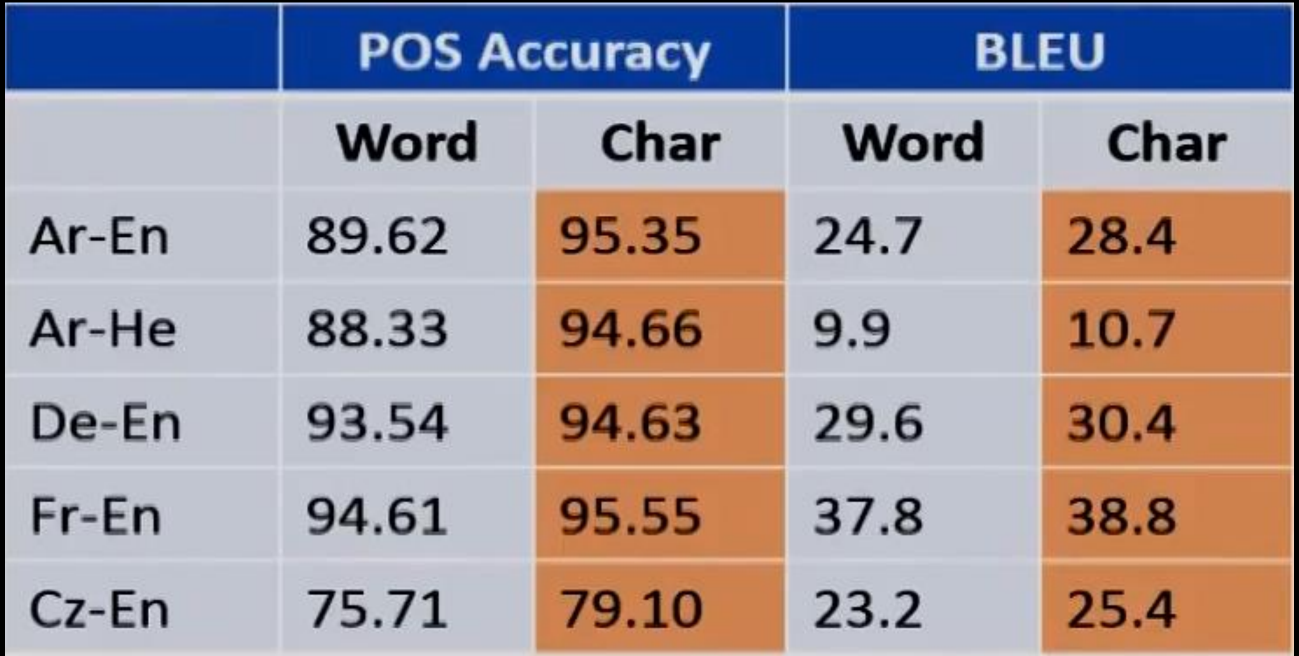


running

r u n n i n g

FROM ACL 2017

I. Effect of word representation (continued.)

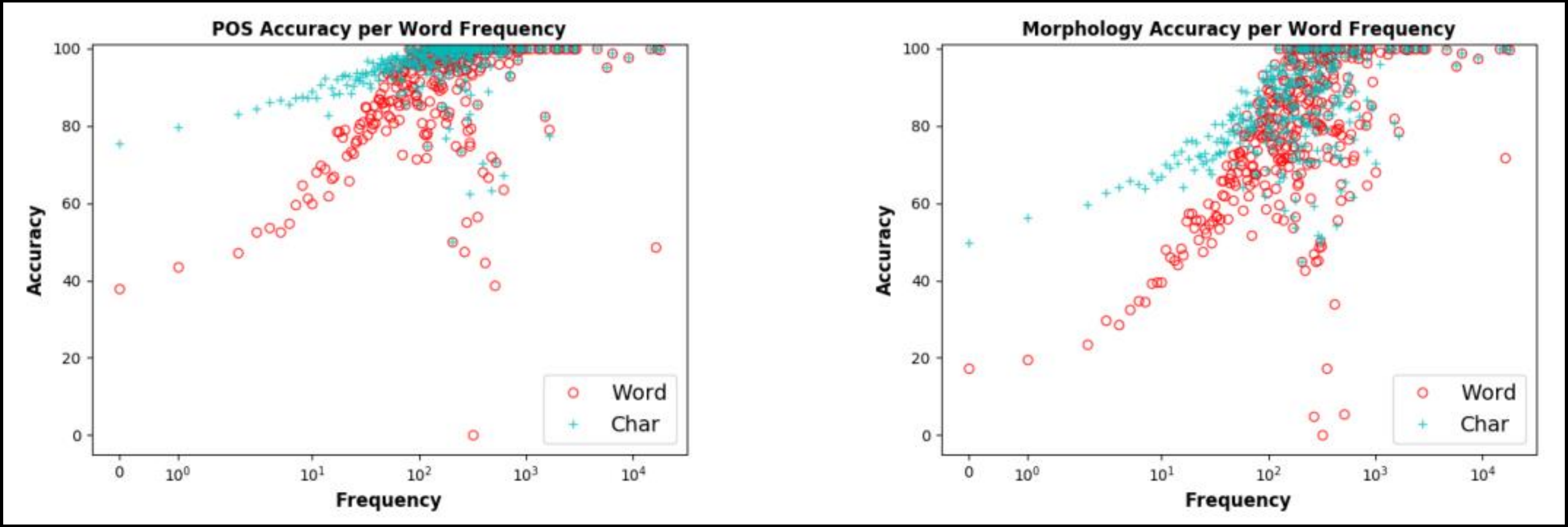


	POS Accuracy		BLEU	
	Word	Char	Word	Char
Ar-En	89.62	95.35	24.7	28.4
Ar-He	88.33	94.66	9.9	10.7
De-En	93.54	94.63	29.6	30.4
Fr-En	94.61	95.55	37.8	38.8
Cz-En	75.71	79.10	23.2	25.4

- Character based models create better representations.
- Character based models improve translation quality.

FROM ACL 2017

II. Impact of Word Frequency



POS and morphological tagging accuracy of word-based and character-based models per word frequency in the training data

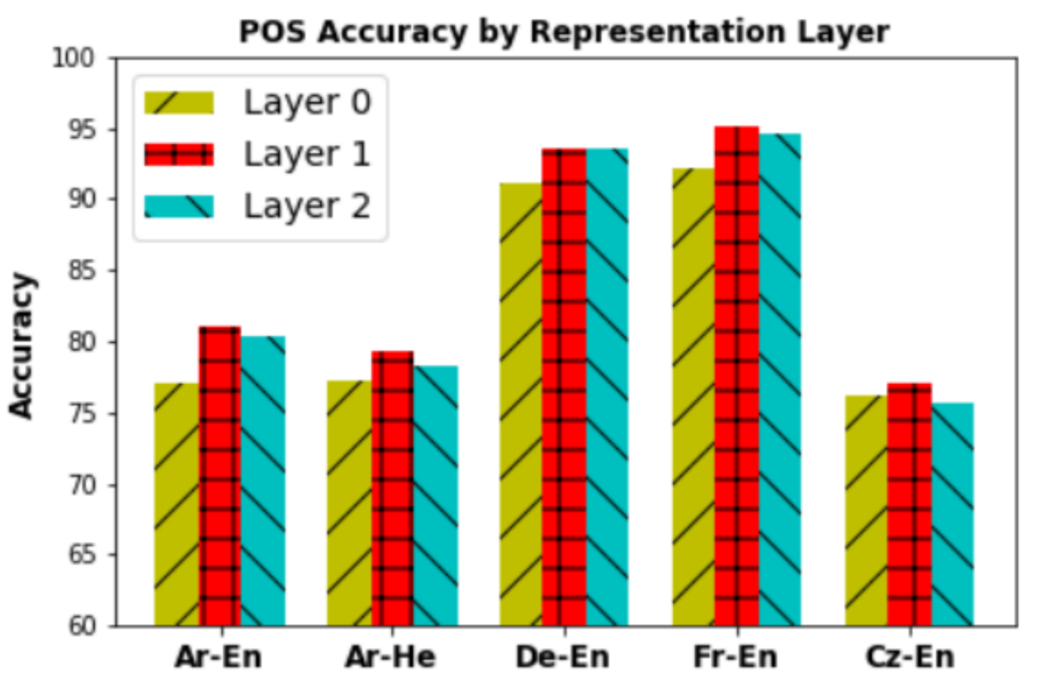
FROM ACL 2017

III. Effect of Encoder Depth

- NMT can be very deep
  - Google Translate : 8 encoder/ decoder layers.
- What kind of information is learnt at each layer ??
- They analyze a 2- layer encoder
  - Extract representations from different layers from training the classifier.

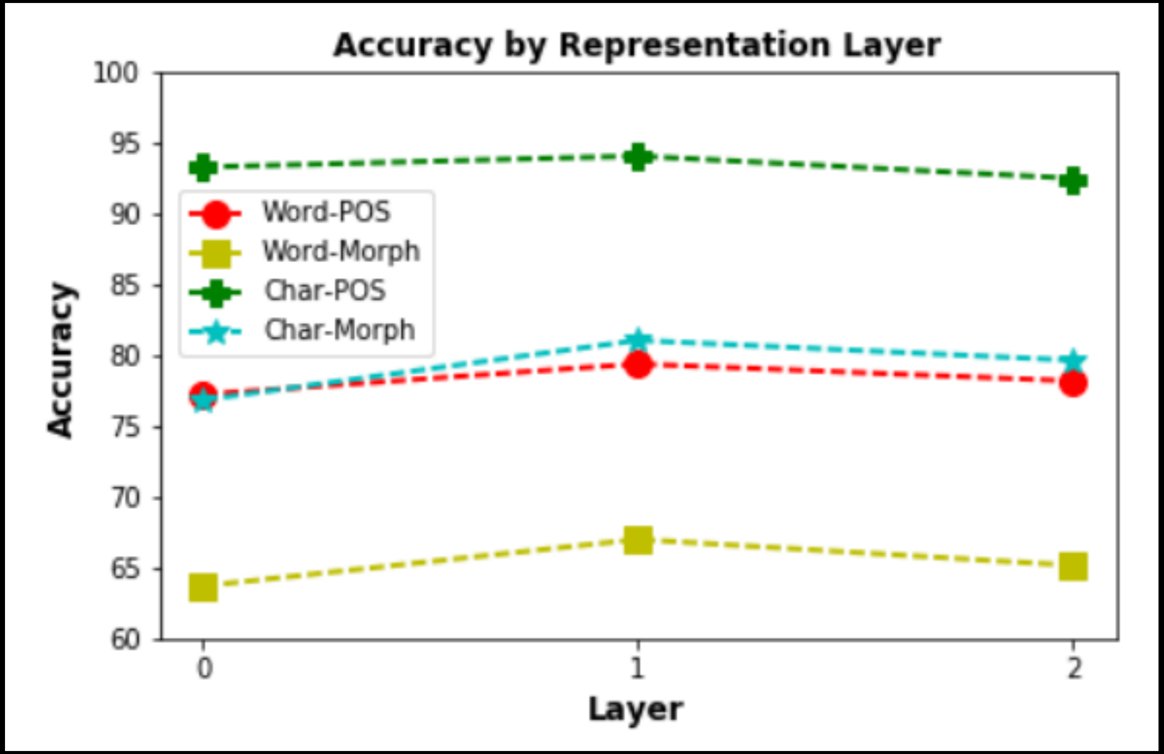
III. Effect of Encoder Depth (continued.)

- Performance on POS tagging: Layer 1 > Layer 2 > Layer 0.
- In contrast, BLEU scores increase when training 2-layer vs. 1-layer models.
- **Interpretation** : Thus translation quality improves when adding layers but morphology quality degrades.



FROM ACL 2017

III. Effect of Encoder Depth (continued.)



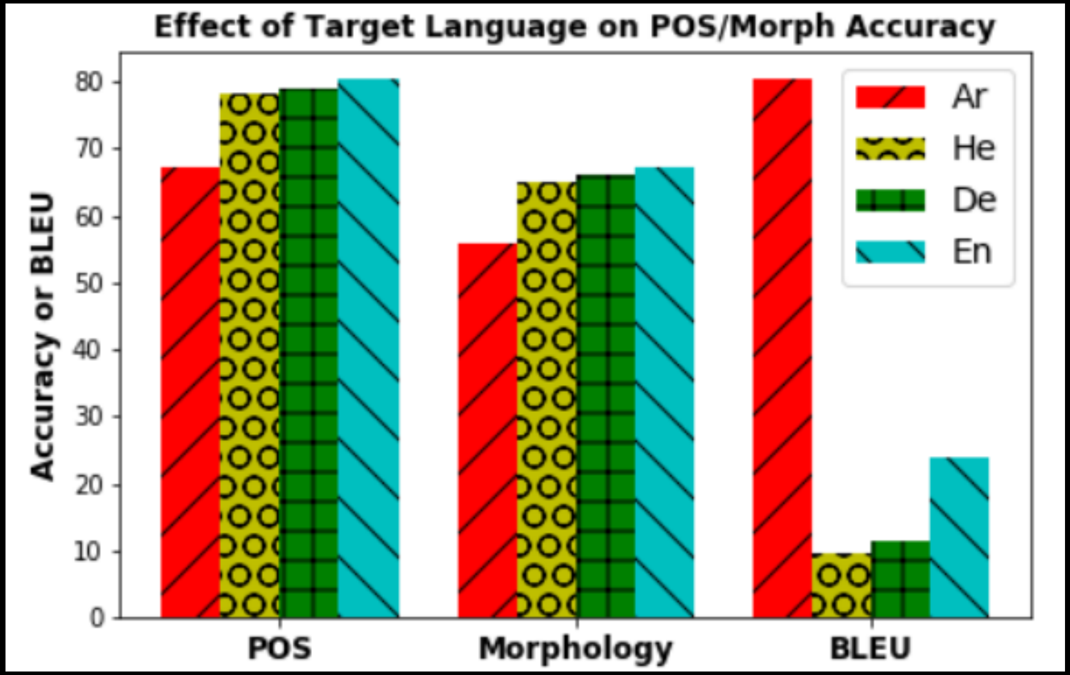
- POS and morphological tagging accuracy across layers.

FROM ACL 2017

IV. Effect of target language

- Translating from morphologically-rich languages is challenging, translating into such languages is even harder.

- The representations learnt when translating into English are better than those learned translating into German, which are in turn better than those learned when translating into Hebrew.

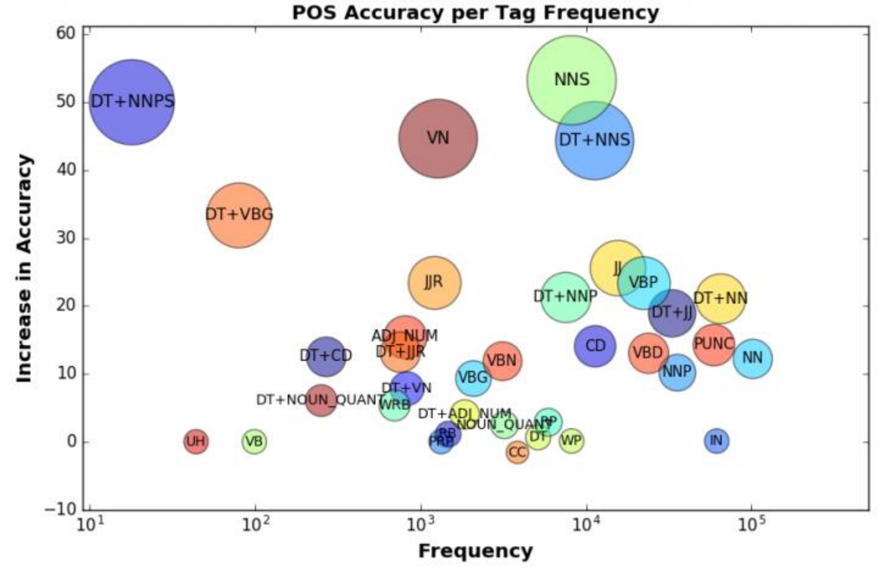


Effect of target language on representation quality of the Arabic source.

FROM ACL 2017

V. Analyzing specific tags

- The authors analyze that both char & word models share similar misclassified tags (especially classifying nouns-NN, NNP).
- But char model performs better on tags with determiner (DT+NNP, DT+NNPS, DT+NNS, DT+VBG).
- The char model performs significantly better for plural nouns and infrequent words.
- Character model also performs better for (NN, DT+NN, DT+JJ, VBP, and even PUNC) tags.



Increase in POS accuracy with char- vs. word-based representations per tag frequency in the training set; larger bubbles reflect greater gaps.

FROM ACL 2017

Decoder Analysis

- To examine what decoder learns about morphology, they train an NMT system on the parallel corpus and use features are used to train a classifier on POS.
- We then perform the following analysis-
  1. Effect of attention
  2. Effect of word representation
- **Result**: They a huge drop in representation quality with the decoder and achieves low POS tagging accuracy.

FROM ACL 2017

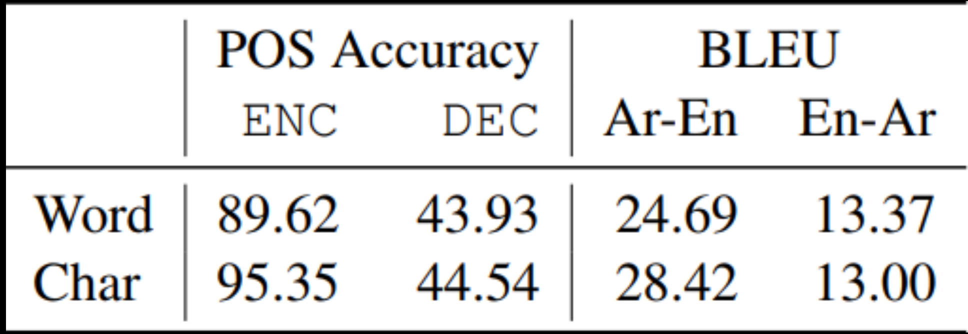
I. Effect of attention

- Removing the attention mechanism decreases the quality of the encoder representations, but improves the quality of the decoder representations.
- **Inference**: Without the attention mechanism, the decoder is forced to learn more informative representations of the target language.

FROM ACL 2017

II. Effect of word representation

- They also conducted experiments to verify findings regarding word-based versus character-based representations on the decoder side.
- While char-based representations improve the encoder, they do not help the decoder. BLEU scores behave similarly.



	POS Accuracy		BLEU	
	ENC	DEC	Ar-En	En-Ar
Word	89.62	43.93	24.69	13.37
Char	95.35	44.54	28.42	13.00

- POS tagging accuracy using word and char based encoder/decoder representations.

FROM ACL 2017

Conclusions

- NMT encoder learns good representations for morphology.
- Character – based representations much better than word based.
- Layer 1 > Layer 2 > Layer 0
- More results from paper:
    - Target language impacts more source side representations.
    - Decoder learns poor target side representations.
    - Attention based model helps decoder exploit source representations.

FROM ACL 2017

Thank You!

FROM ACL 2017