What do Neural Machine Translation Models Learn About Morphology

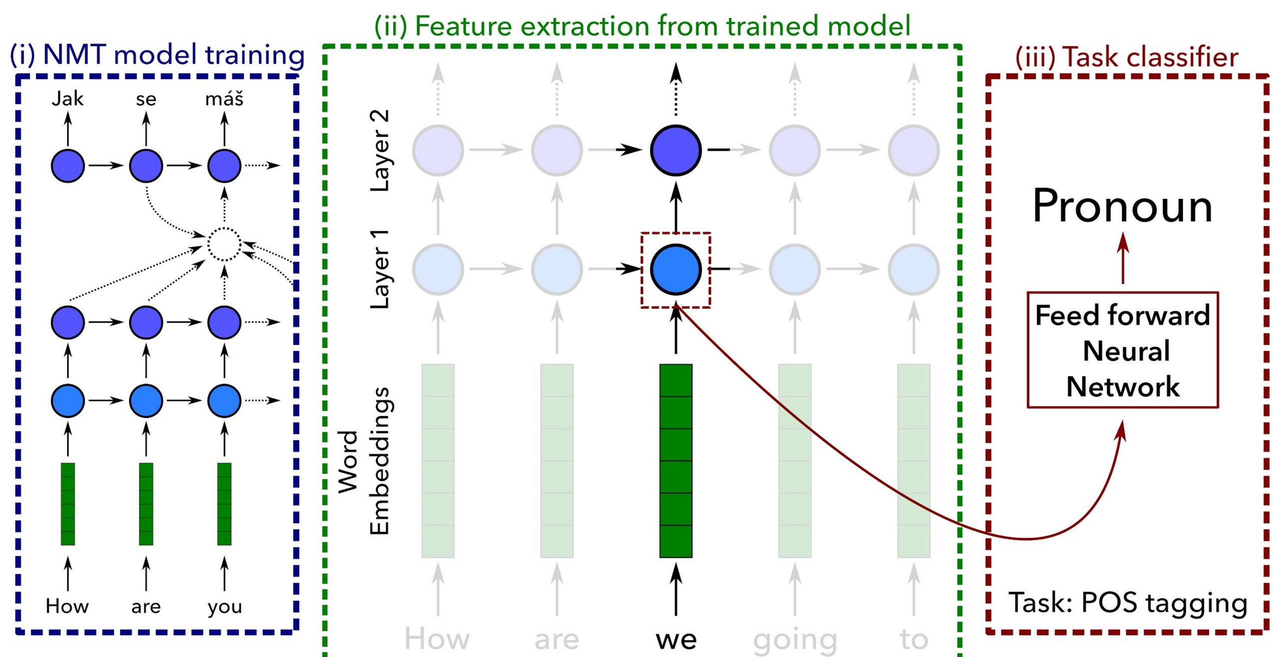
Yonatan Belinkov[1], Nadir Durrani[2], Fahim Dalvi[2], Hassan Sajjad[2] and James Glass[1]

[1]MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, USA [2]QCRI, Doha, Qatar

Motivation

- Neural machine translation (NMT) obtains state-of-the-art performance with a simple end-to-end architecture.
- Little is known about what these models learn about source and target languages during training.
- We analyze intermediate representations learned by NMT and evaluate their quality for learning morphology in different morphologically-rich languages.
- **Research questions:**
  - Which parts of the NMT architecture capture word structure?
  - What is the division of labor between NMT components (encoder, decoder, attention)?
  - 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?

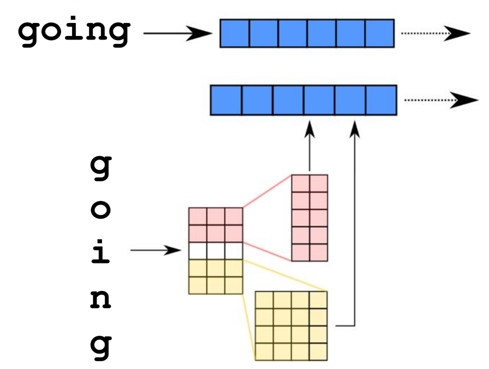
Methodology



- 3-step procedure to evaluate morphology learned in different parts of the network.
  1. Train NMT model on parallel data
  2. Extract features from pre-trained model
  3. Train classifier on supervised data
- Quality of trained classifier reflects quality of extracted representations.
- Extrinsically evaluate on POS/morph. tagging

Encoder Analysis

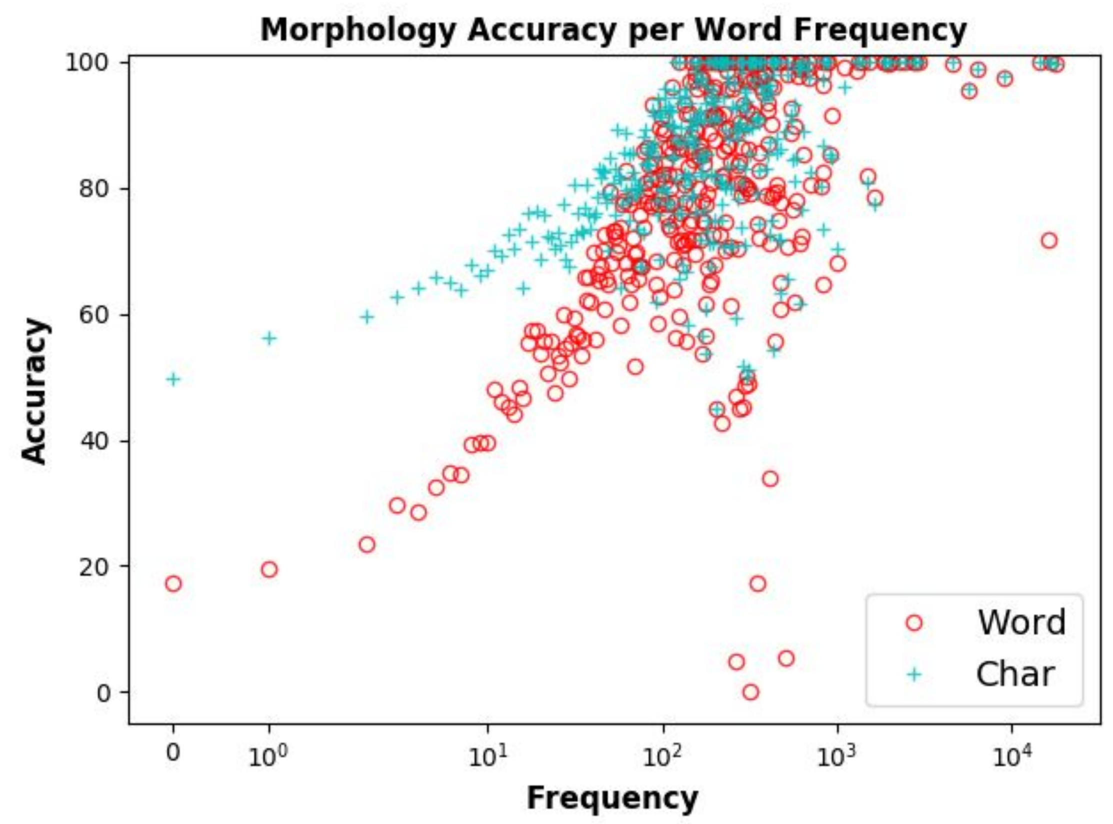
- **Effect of word representation**
  i. Word
  ii. charCNN

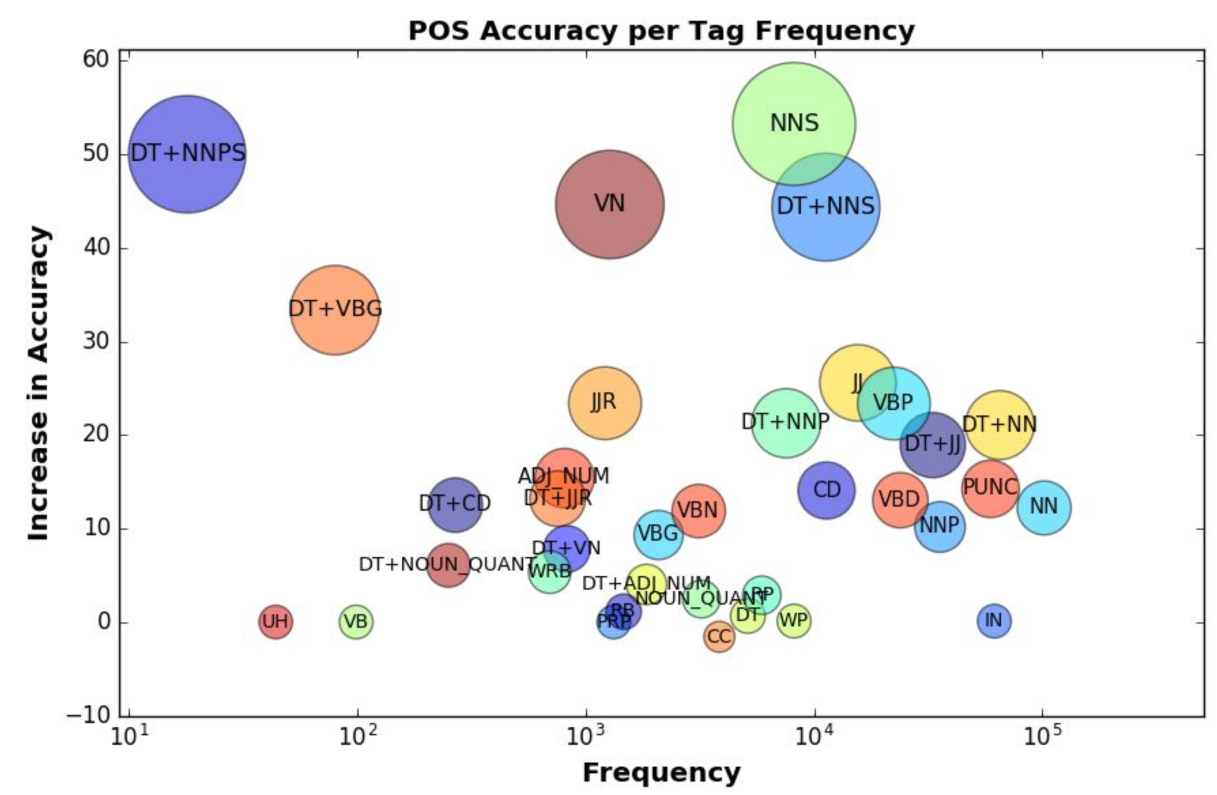


  - Representations learned with char-based models give better BLEU scores and POS accuracy
  - Infrequent words benefit most.
  - Certain tags are more sensitive to character information.

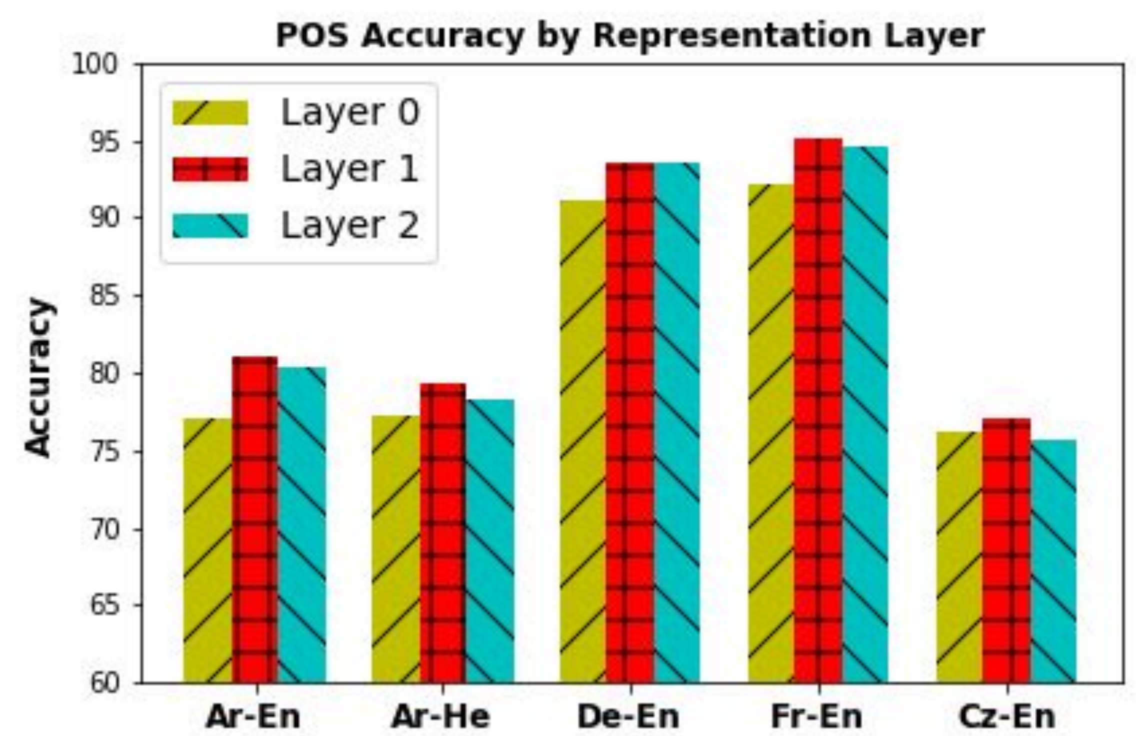
	Gold	Pred	BLEU
	Word/Char	Word/Char	Word/Char
Ar-En	80.31/93.66	89.62/95.35	24.7/28.4
Ar-He	78.20/92.48	88.33/94.66	9.9/10.7
De-En	87.68/94.57	93.54/94.63	29.6/30.4
Fr-En	–	94.61/95.55	37.8/38.8
Cz-En	–	75.71/79.10	23.2/25.4

Table 1: POS accuracy on gold and predicted tags using word-based and character-based representations, as well as corresponding BLEU scores.

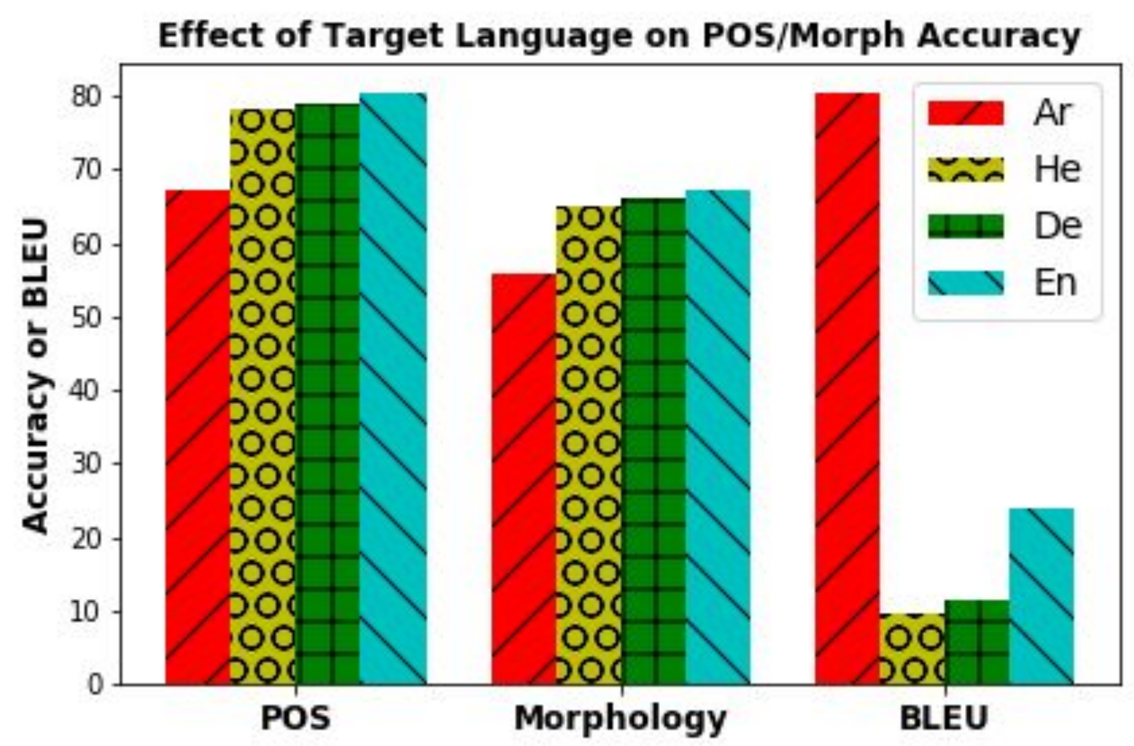




- **Effect of encoder depth**
  - Representations from lower layers are better for POS/morphology.
  - But deeper networks improve BLEU.
  - Do higher layers capture semantics?



- **Effect of target language**
  - Translating to morphologically-poorer languages leads to better representations.
  - BLEU scores do not always entail better morphological representations.



Decoder Analysis

- **Encoder vs. decoder representations**
  - Decoder representations are much worse for POS/morphology.
- **Effect of attention mechanism**
  - Attention mechanism takes load from the decoder when learning target morphology.

Attn	POS Accuracy ENC	POS Accuracy DEC	BLEU Ar-En	BLEU En-Ar
✓	89.62	43.93	24.69	13.37
✗	74.10	50.38	11.88	5.04

Table 2: POS tagging accuracy using encoder and decoder representations with/without attention.

- **Effect of word representation**
  - Character representations do not help the decoder.
  - Possible explanation: charCNN cannot generate unseen words.

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

Table 3: POS tagging accuracy using word-based and char-based encoder/decoder representations.

Conclusion

- We investigate what neural machine translation models learn about morphology.
- We evaluate NMT representation quality on POS and morphological tagging.
- Our insights can guide further development of NMT systems, for example by guiding joint learning of translation and morphology.
- Future work can extend the analysis to other representations, deeper networks, and semantic tasks.

Acknowledgments: This work was supported by the Qatar Computing Research Institute (QCRI). Any opinions, findings, conclusions, or recommendations expressed in this paper are those of the authors, and do not necessarily reflect the views of the funding organizations.