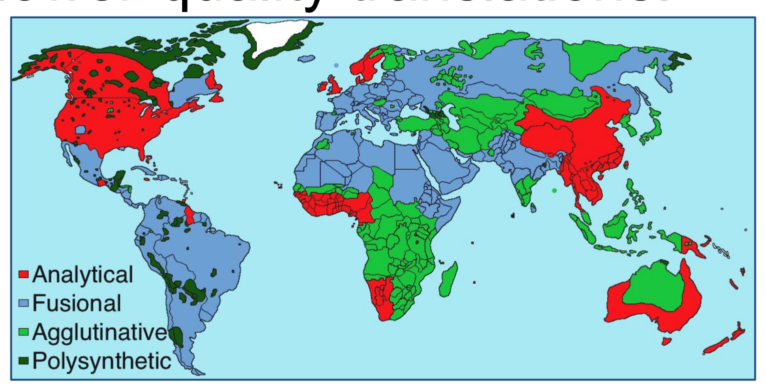
Hybrid Word-Character Neural Machine Translation

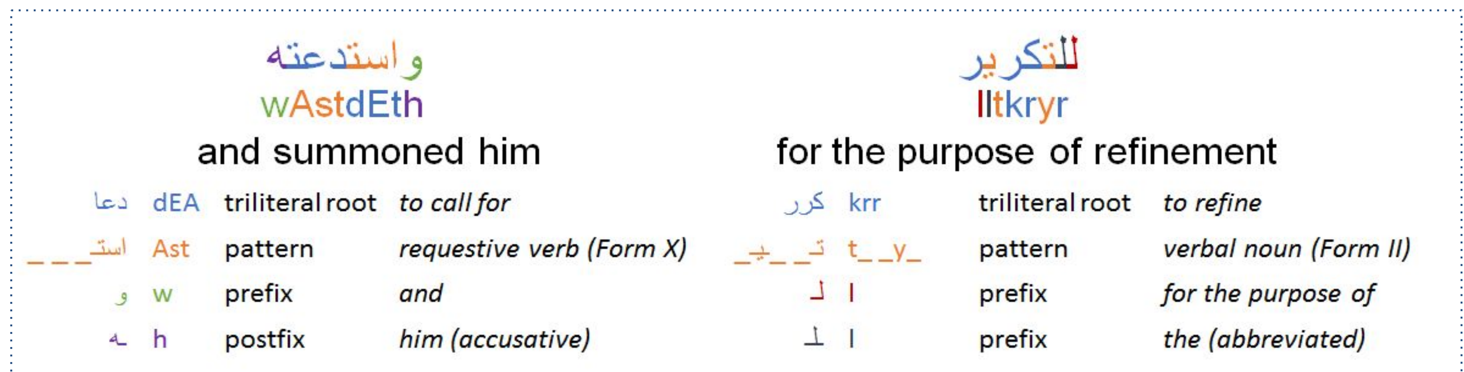
Pamela Toman, Sigtryggur Kjartansson
Department of Computer Science, Stanford University

Motivation & Background

- Words are **not** the primary level of meaning in all languages.
- Because traditional NMT is word-level, non-analytic languages tend to have lower-quality translations.
- Want a universalist architecture that performs well for all language pairs.

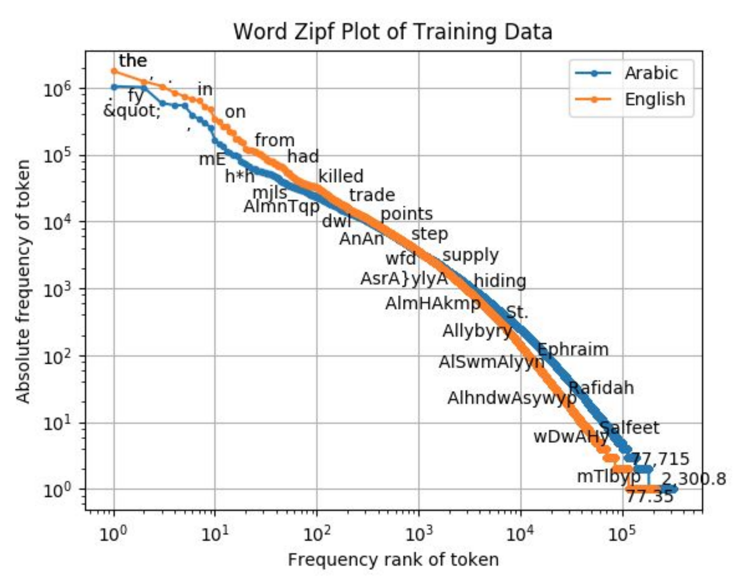


Focusing on **Arabic**: 5th largest language by number of speakers, understudied, and has a variety of clitics, affixes, spelling ambiguities, and the root-and-pattern morphology of Semitic languages.

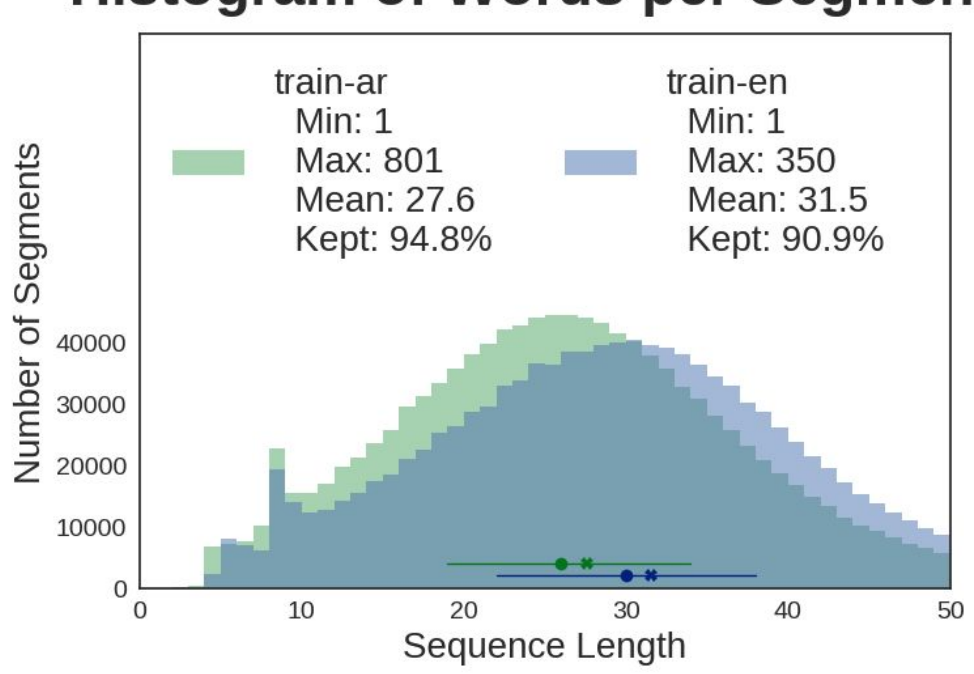


Dataset

- Modern Standard Arabic news articles.
- We follow Almahairi et al. 2016[1]:
  - *Train*: 1,087,343 sentence pairs
  - *Dev:* 915 sent. pairs
  - *Test:* 946 sent. pairs



Histogram of Words per Segment



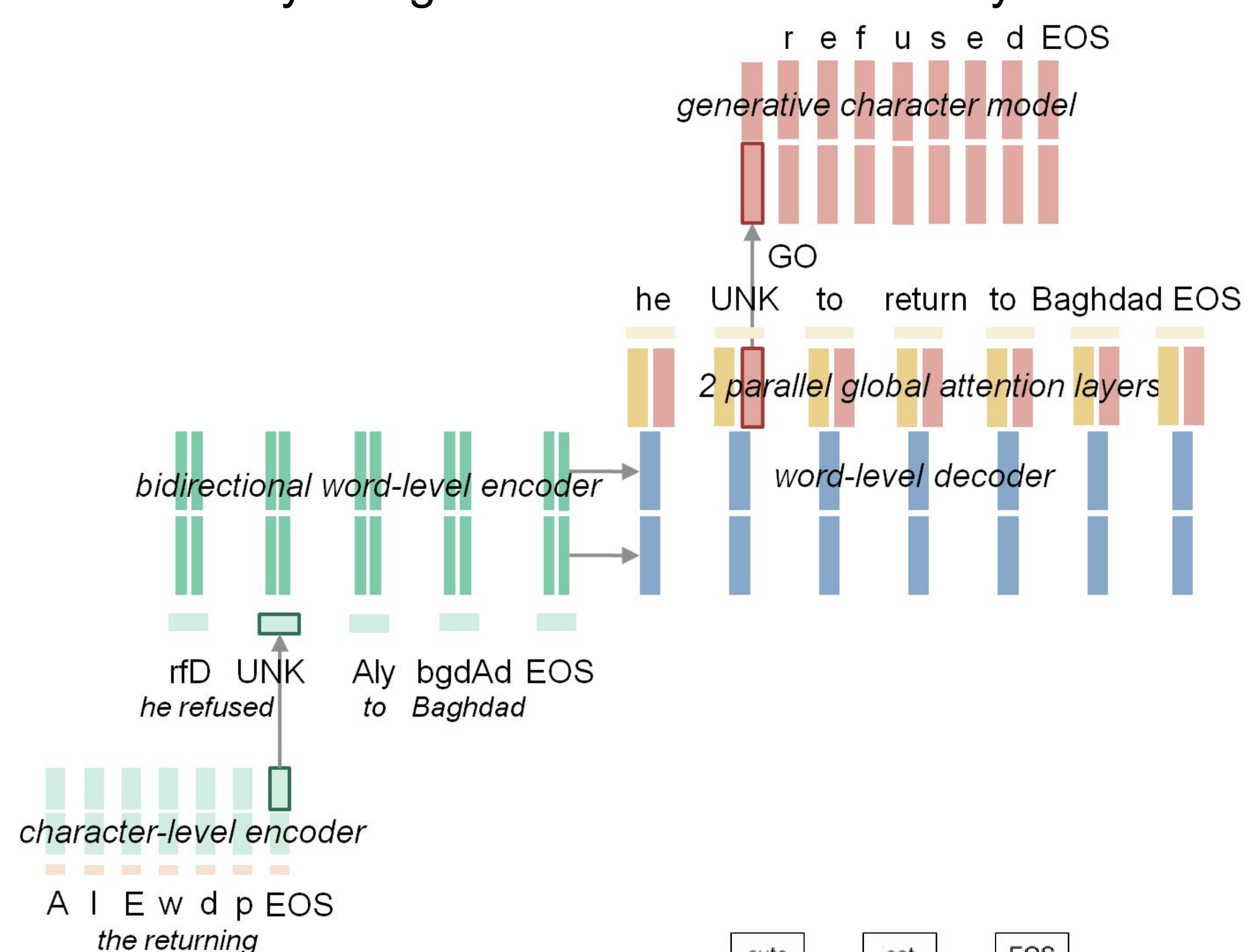
Types	AR	322k
	EN	216k
Tokens	AR	27.0m
	EN	30.3m

Problem Statement

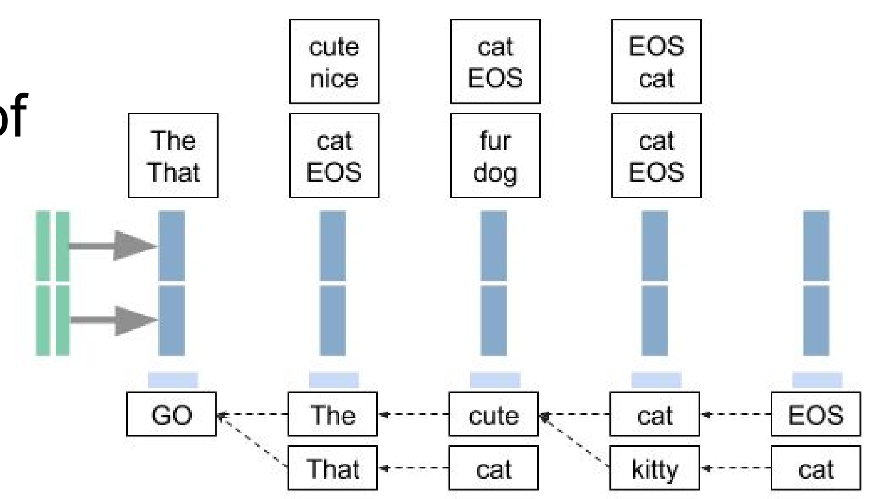
Goal: Improve Arabic-English Machine Translation.
Benchmark: Almahairi et al. 2016[1] using BLEU metric.
Approach: Neural Machine Translation model that achieves open vocabulary.

Method

- We unite three models to create a hybrid word-character model, based on Luong and Manning 2016:[2]
  - Source Character-Level Encoder.
  - Word-Level Sequence-to-Sequence with global bilinear attention.[3]
  - Target Character-Level Generator.
- Backs off to the character level when word-level representations do not exist, in order to achieve open vocabulary using a small known vocabulary.

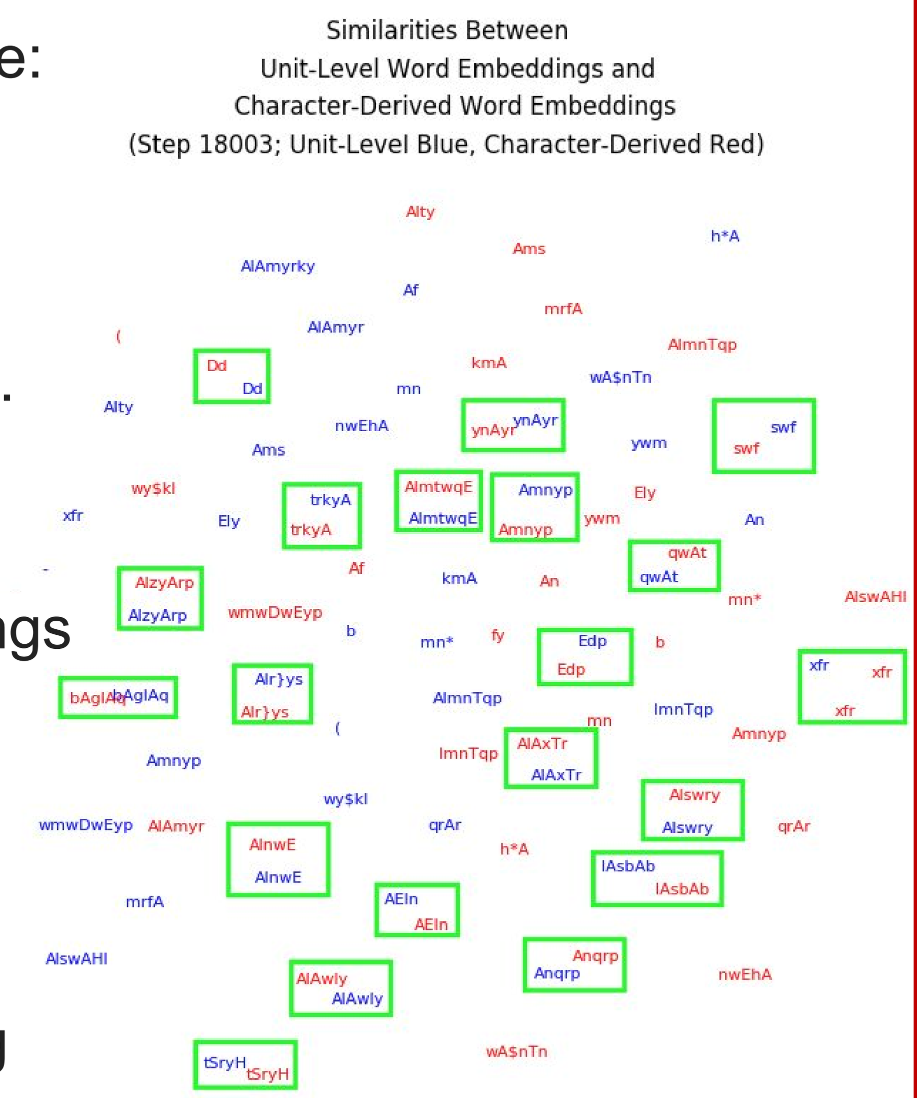


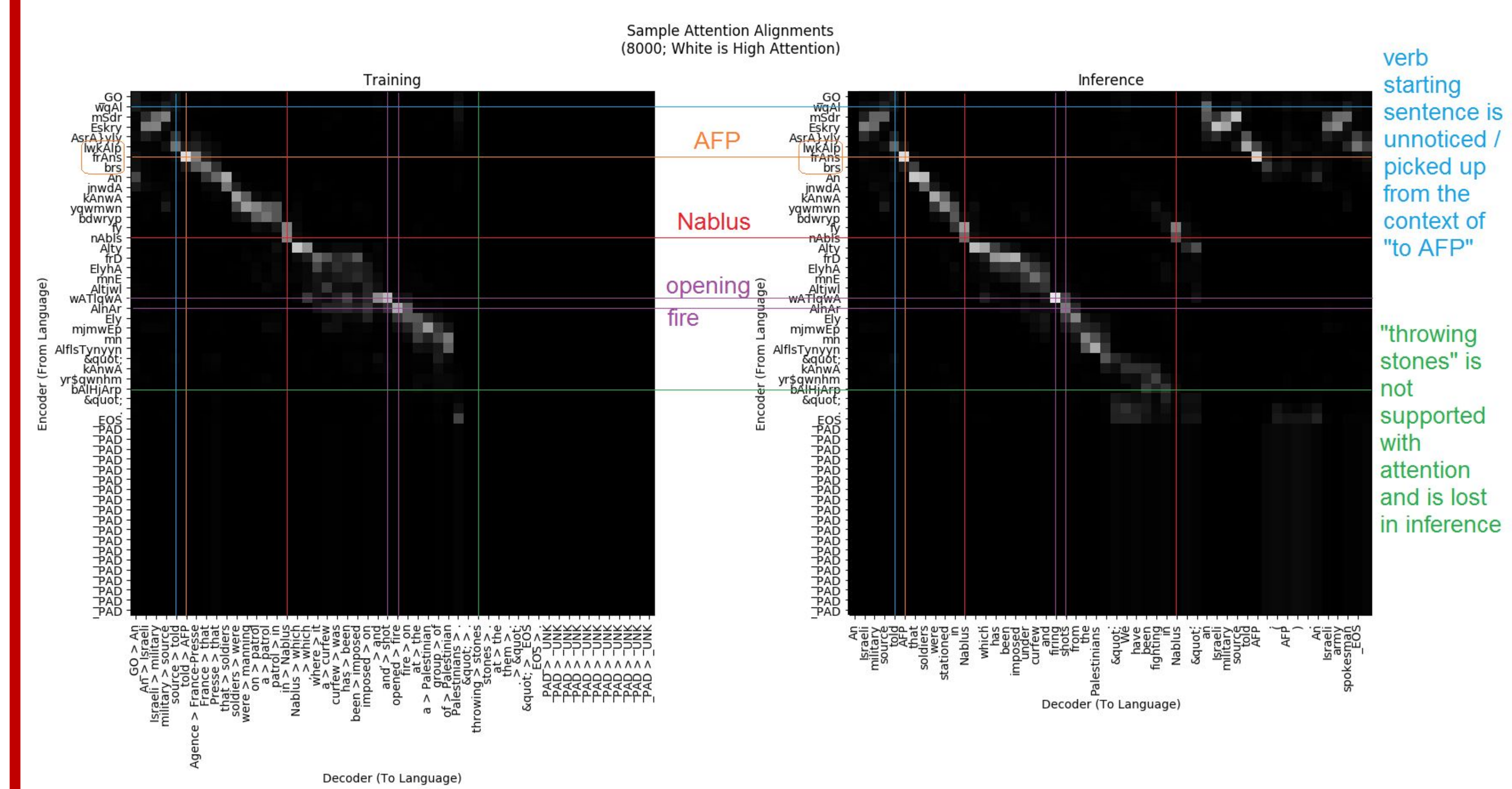
- Improves training of character models with frequent word sampling.
- Uses an in-graph beam search decoder.



Experimental Evaluation & Results

- Achieved BLEU score: 42.10
- Evaluate model by:
  - Train with different hidden states sizes.
  - Train with different vocabulary sizes.
  - Visualize embeddings using t-SNE.
  - Inspect attention alignments.
  - Use alternative frequency sampling techniques.





Conclusion & Future Direction

- Hybrid architecture combines the strength of both word- and character-based models; fast to train and offer high-quality translation; and achieves open vocabulary.
- Want to add a character-level attention mechanism and/or a convolutional layer to facilitate interactions with Arabic's complex morphology.

[1] Almahairi et al., 2016; arXiv:1606.02680
[2] Luong and Manning, 2016; arXiv:1604.00788
[3] Luong, Pham and Manning, 2015; arXiv:1508.04025
Acknowledgements: Ignacio Cases, Microsoft Azure, CS224N staff