On Learning Form and Meaning in Neural Machine Translation Models

Yonatan Belinkov

May 2017

With: Nadir Durrani, Hassan Sajjad, Fahim Dalvi, Lluis Marques, James Glass

Motivation

- Neural machine translation (NMT) obtains state-of-the-art results
- Elegant and simple end-to-end architecture

Motivation

- Neural machine translation (NMT) obtains state-of-the-art results
- Elegant and simple end-to-end architecture

- However, NMT models are difficult to interpret;
  what do they learn about the source and target languages?

Motivation

- Neural machine translation (NMT) obtains state-of-the-art results
- Elegant and simple end-to-end architecture

- However, NMT models are difficult to interpret;
  what do they learn about the source and target languages?

- Recent interest in the community (e.g. Shi+ 16 on syntax)

Motivation

- This work: analyzing morphology (and semantics) in NMT

Translation as Decoding

- Warren Weaver to Norbert Wiener, March 4, 1947:

  Also knowing nothing official about, but having guessed and inferred considerable about, powerful new mechanized methods in cryptography - methods which I believe succeed even when one does not know what language has been coded - one naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say "This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode."

Brief History of Machine Translation

- 1947: Initial ideas of MT (Weaver)
- 1950s: First MT systems
- 1960s: High-quality MT fails, cut in government funding
- 1970s-1980s: Rule-based systems, interlingua ideas
- 1990s: Statistical MT, IBM alignment models
- 2000s: Phrase-based MT, open-source toolkits
- 2014-2015: Neural MT: seq2seq + attention

Statistical Machine Translation

- Translate a source sentence *F* into a target sentence *E*

$$\hat{E} = \arg\max_{E} P(E|F)$$

Statistical Machine Translation

- Translate a source sentence *F* into a target sentence *E*

$$\hat{E} = \arg\max_{E} P(E|F) = \arg\max_{E} \frac{P(F|E)P(E)}{P(F)}$$

Statistical Machine Translation

- Translate a source sentence *F* into a target sentence *E*

$$\hat{E} = \arg\max_{E} P(E|F) = \arg\max_{E} \frac{P(F|E)P(E)}{P(F)} = \arg\max_{E} P(F|E)P(E)$$

Statistical Machine Translation

- Translate a source sentence *F* into a target sentence *E*

$$\hat{E} = \arg\max_{E} P(E|F) = \arg\max_{E} \frac{P(F|E)P(E)}{P(F)} = \arg\max_{E} P(F|E)P(E)$$

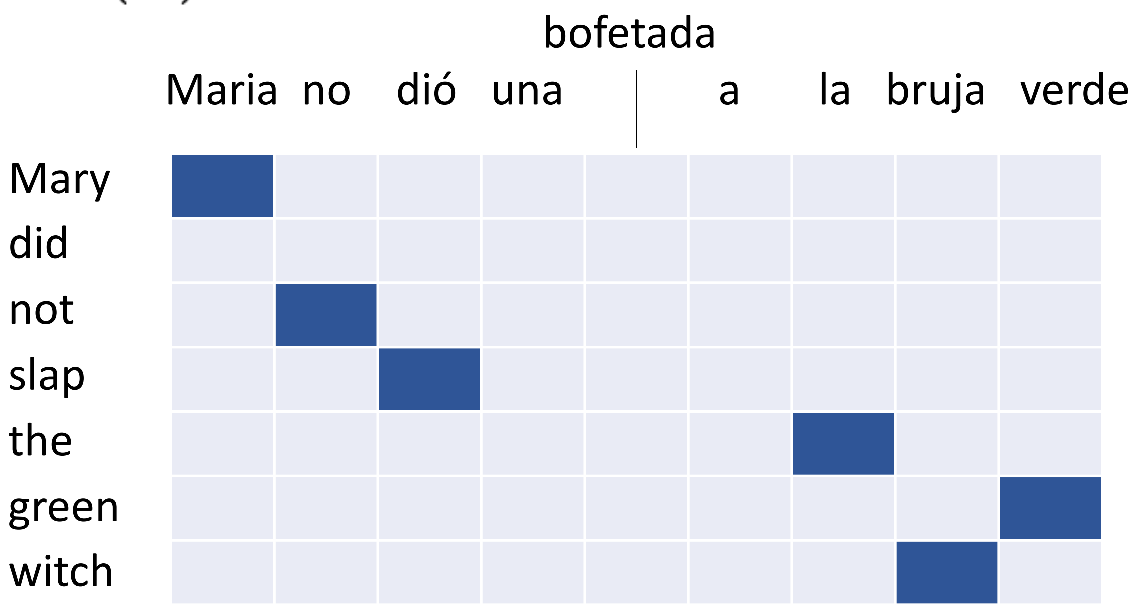
- $P(F|E)$ – Translation model
- $P(E)$ – Language model

Statistical Machine Translation

- Translate a source sentence *F* into a target sentence *E*

$$\hat{E} = \arg\max_E P(E|F) = \arg\max_E \frac{P(F|E)P(E)}{P(F)} = \arg\max_E P(F|E)P(E)$$

- $P(F|E)$ – Translation model
- $P(E)$ – Language model



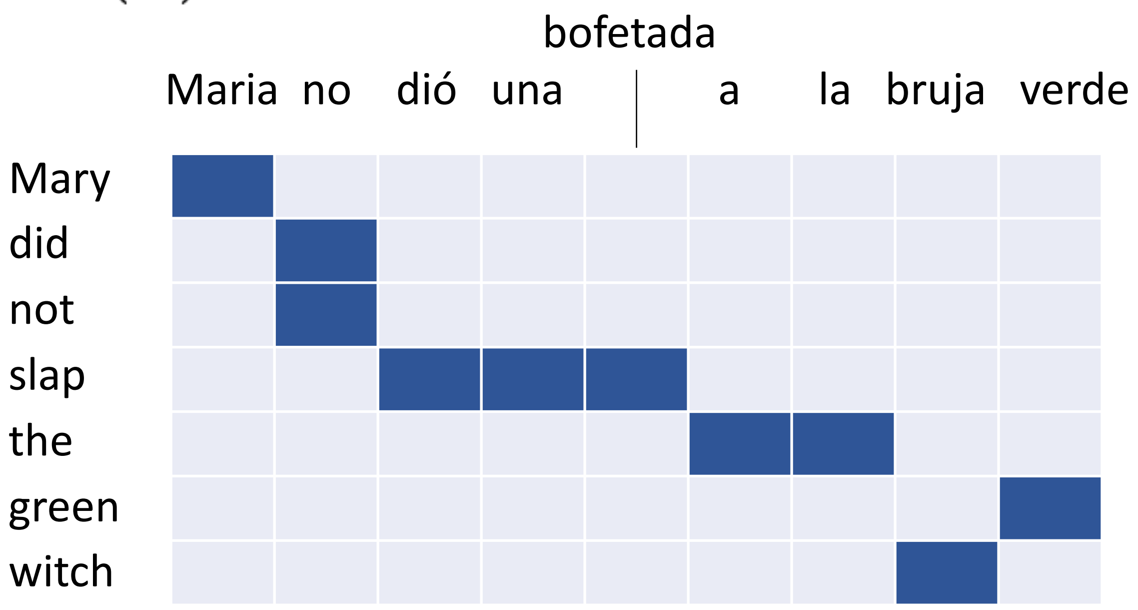
From: Jurafsky & Martin 2009

Statistical Machine Translation

- Translate a source sentence *F* into a target sentence *E*

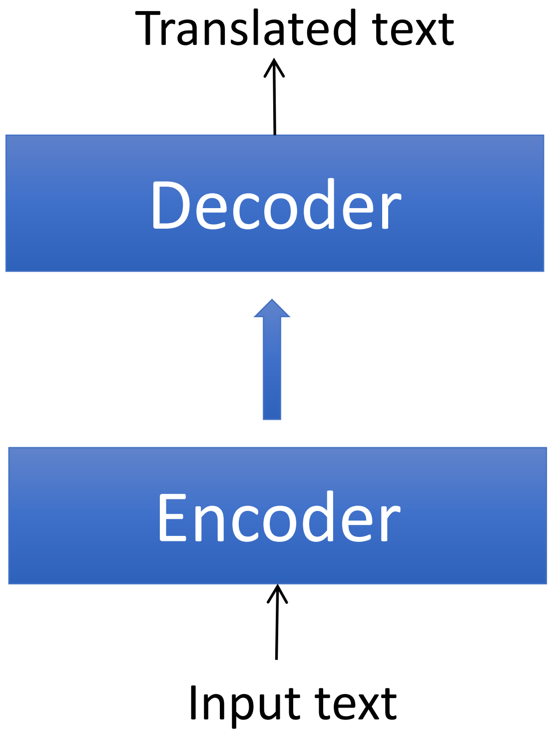
$$\hat{E} = \arg\max_{E} P(E|F) = \arg\max_{E} \frac{P(F|E)P(E)}{P(F)} = \arg\max_{E} P(F|E)P(E)$$

- $P(F|E)$ – Translation model
- $P(E)$ – Language model



From: Jurafsky & Martin 2009

Neural Machine Translation



Neural Machine Translation

$$P(E|F) = \prod_i P(e_i|e_1, ...e_{i-1}, F)$$

Neural Machine Translation

$$P(E|F) = \prod_i P(e_i|e_1, ...e_{i-1}, F)$$

- Encoder: $$h_i^F = f_F(h_{i-1}, x_i)$$
- Decoder: $$h_i^E = f_E(h_{i-1}, x_{i-1}, c)$$
$$P(e_i|e_1, ...e_{i-1}, F) = g(h_i, x_{i-1}, c)$$
- Loss: $$\frac{1}{N}\sum_{n=1}^{N}\sum_i \log P(e_i^n|e_1^n, ...e_{i-1}^n, F^n)$$

Neural Machine Translation

$$P(E|F) = \prod_i P(e_i|e_1, ...e_{i-1}, F)$$

- Encoder:

Source hidden state

$$h_i^F = f_F(h_{i-1}, x_i)$$

- Decoder:

Target hidden state

$$h_i^E = f_E(h_{i-1}, x_{i-1}, c)$$

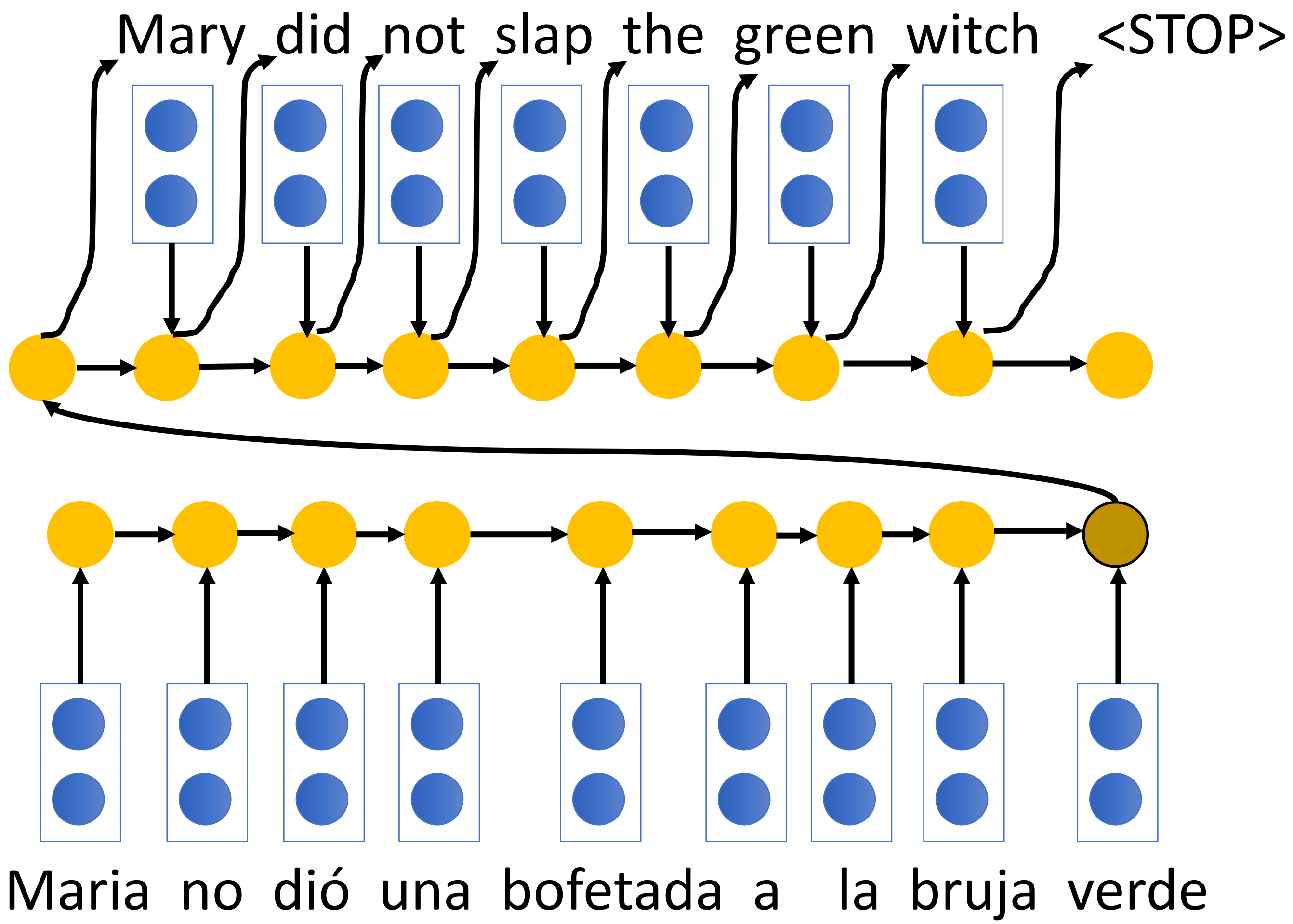
$$P(e_i|e_1, ...e_{i-1}, F) = g(h_i, x_{i-1}, c)$$

Summary vector

- Loss:

$$\frac{1}{N}\sum_{n=1}^{N}\sum_i \log P(e_i^n|e_1^n, ...e_{i-1}^n, F^n)$$

Encoder-Decoder

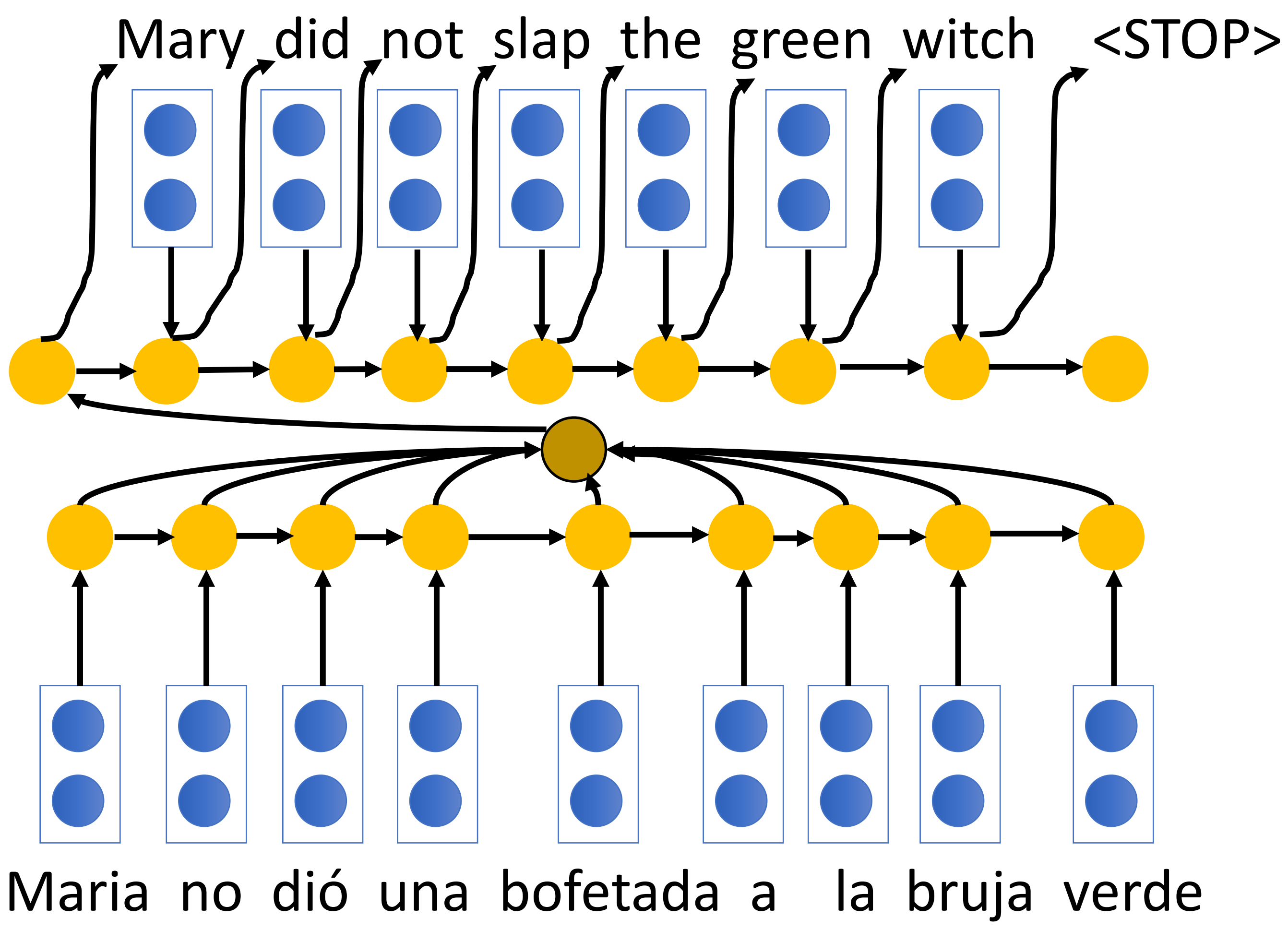


The Problem with the Encoder-Decoder

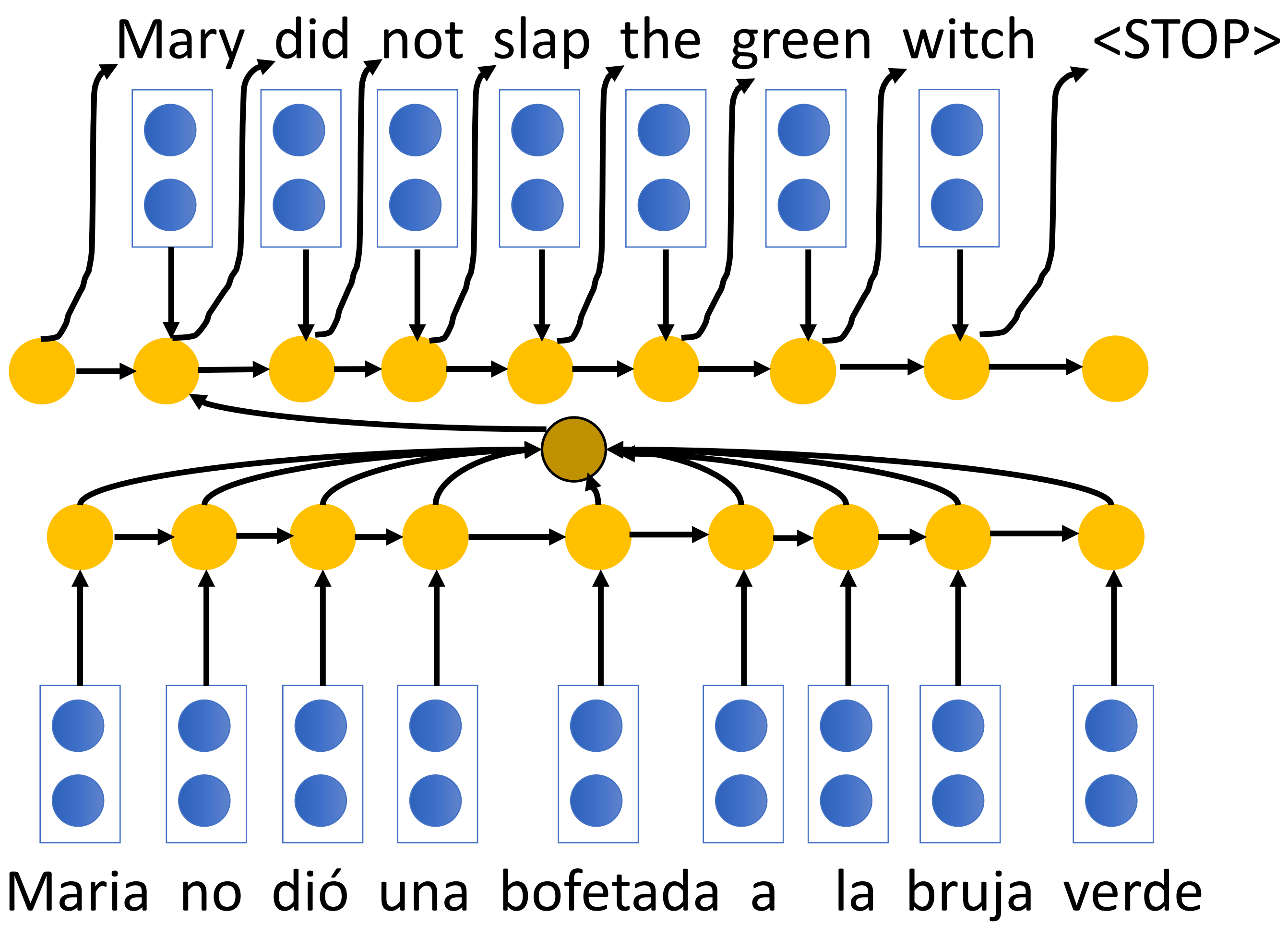
- Raymond Mooney, June 26, 2016:

“You can’t cram the meaning of a whole
%&!$# sentence into a single $&!# vector!”*

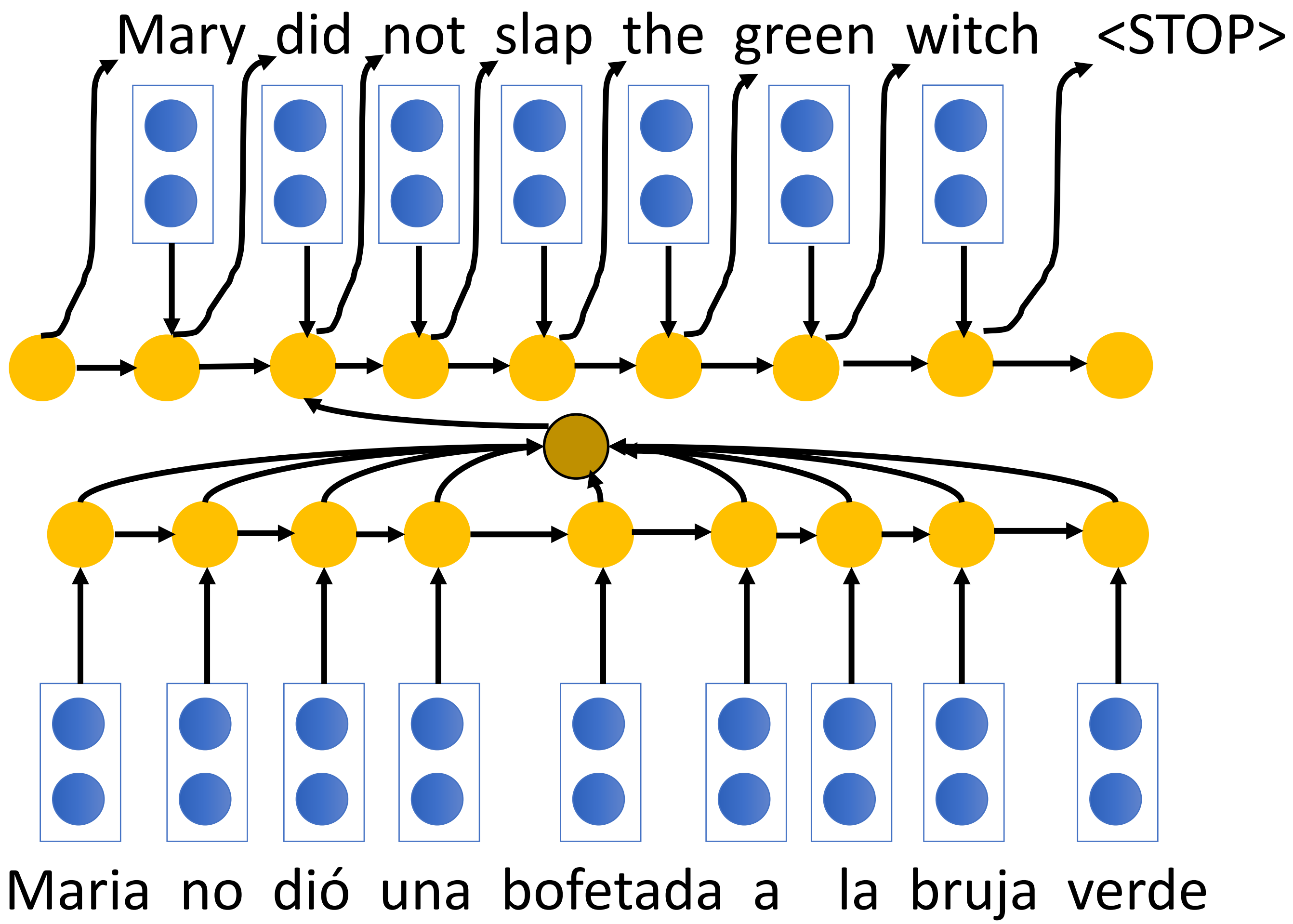
Attention Mechanism



Attention Mechanism

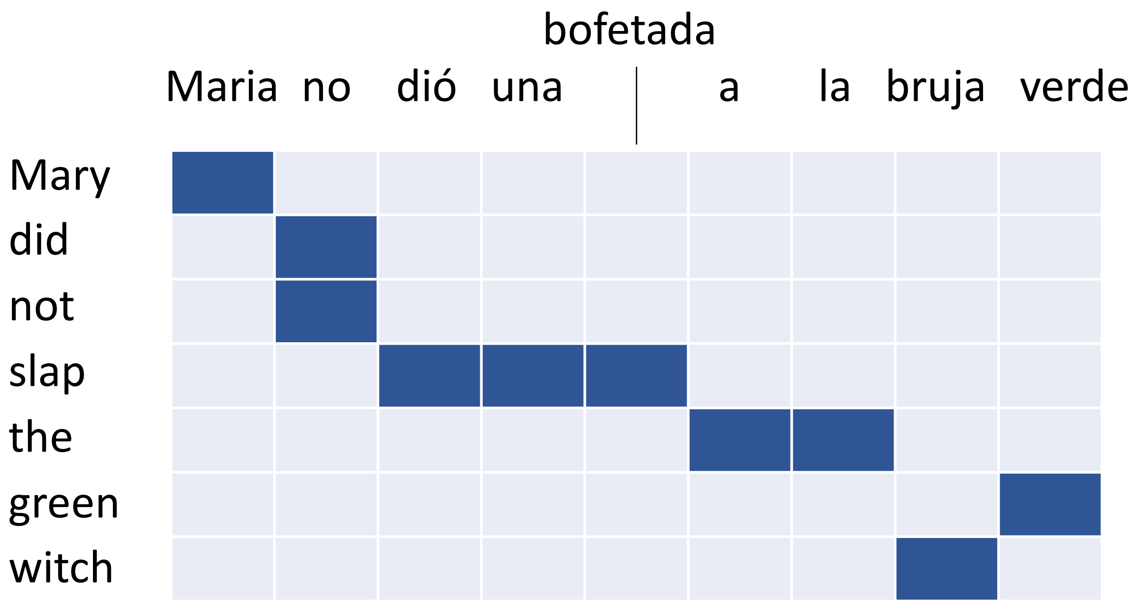


Attention Mechanism



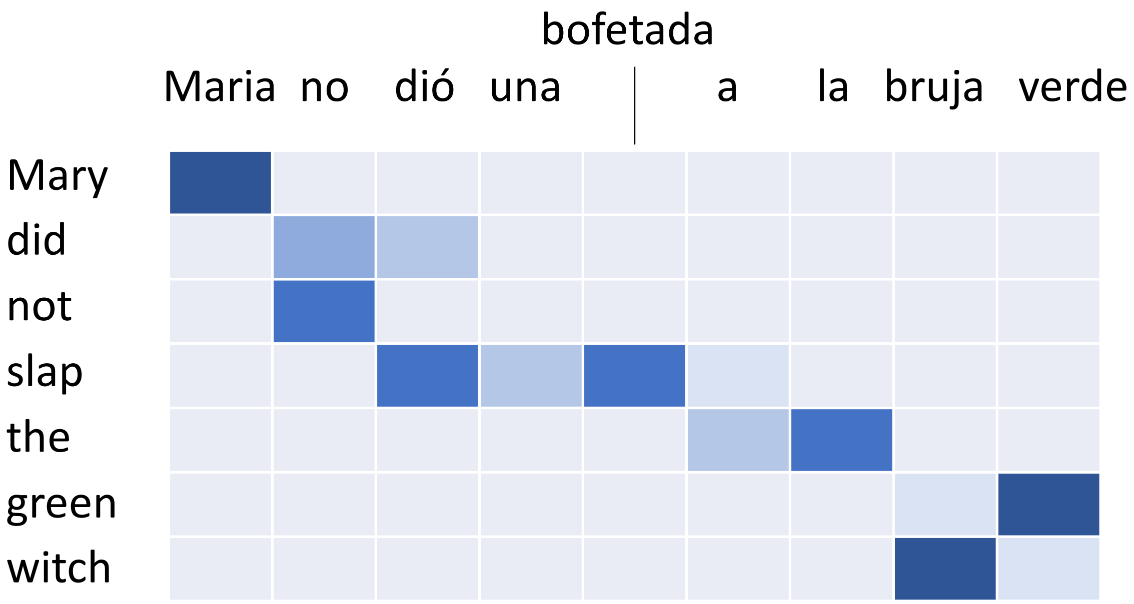
Attention as soft alignment

Phrase-based MT

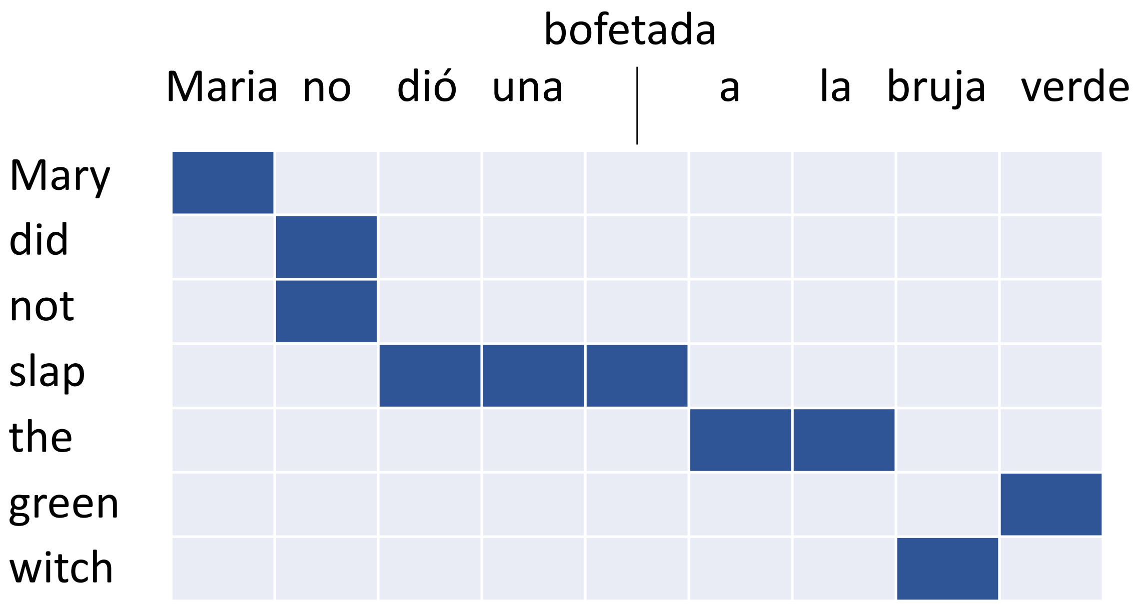


Attention as soft alignment

Neural MT



Phrase-based MT



Research Questions

Research Questions

- Which parts of the NMT architecture capture word structure? Which capture meaning?
- What is the division of labor between different components?
- How do different word representations help learn better morphology?
- How does the target language affect the learning of word structure?

Methodology

- Three step procedure:
  1. Train a neural MT system
  2. Extract feature representations using trained the model
  3. Train a classifier using extracted features and evaluate it on an extrinsic task

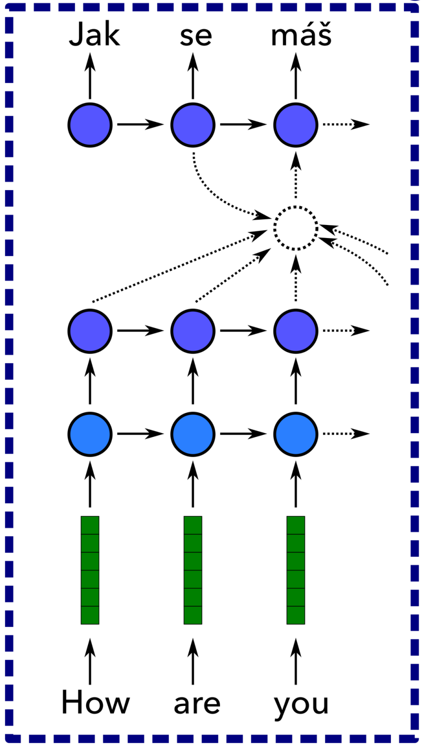
Methodology

- Three step procedure:
  1. Train a neural MT system
  2. Extract feature representations using trained the model
  3. Train a classifier using extracted features and evaluate it on an extrinsic task

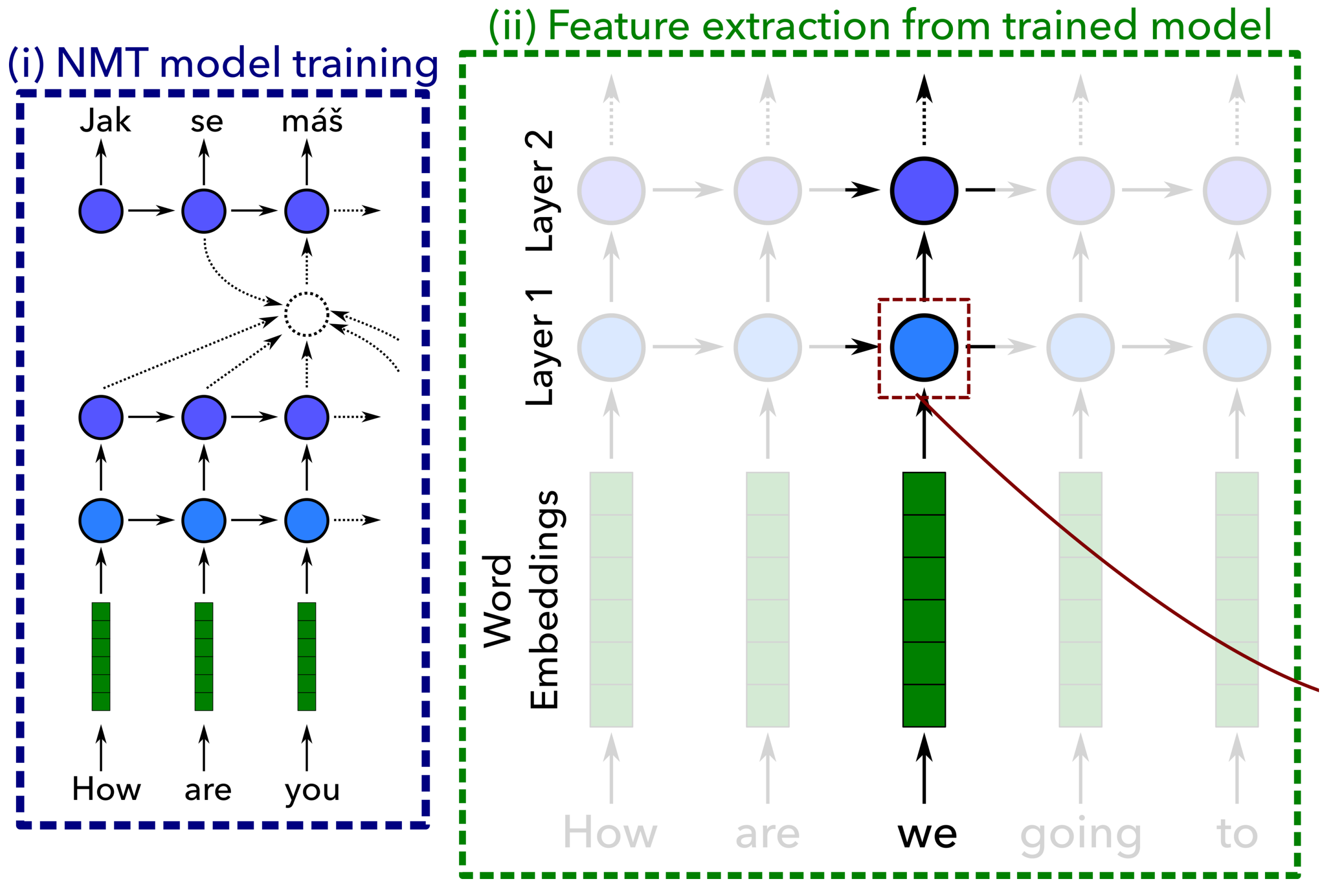
- Assumption: performance of the classifier reflects quality of the NMT representations for the given task

Methodology

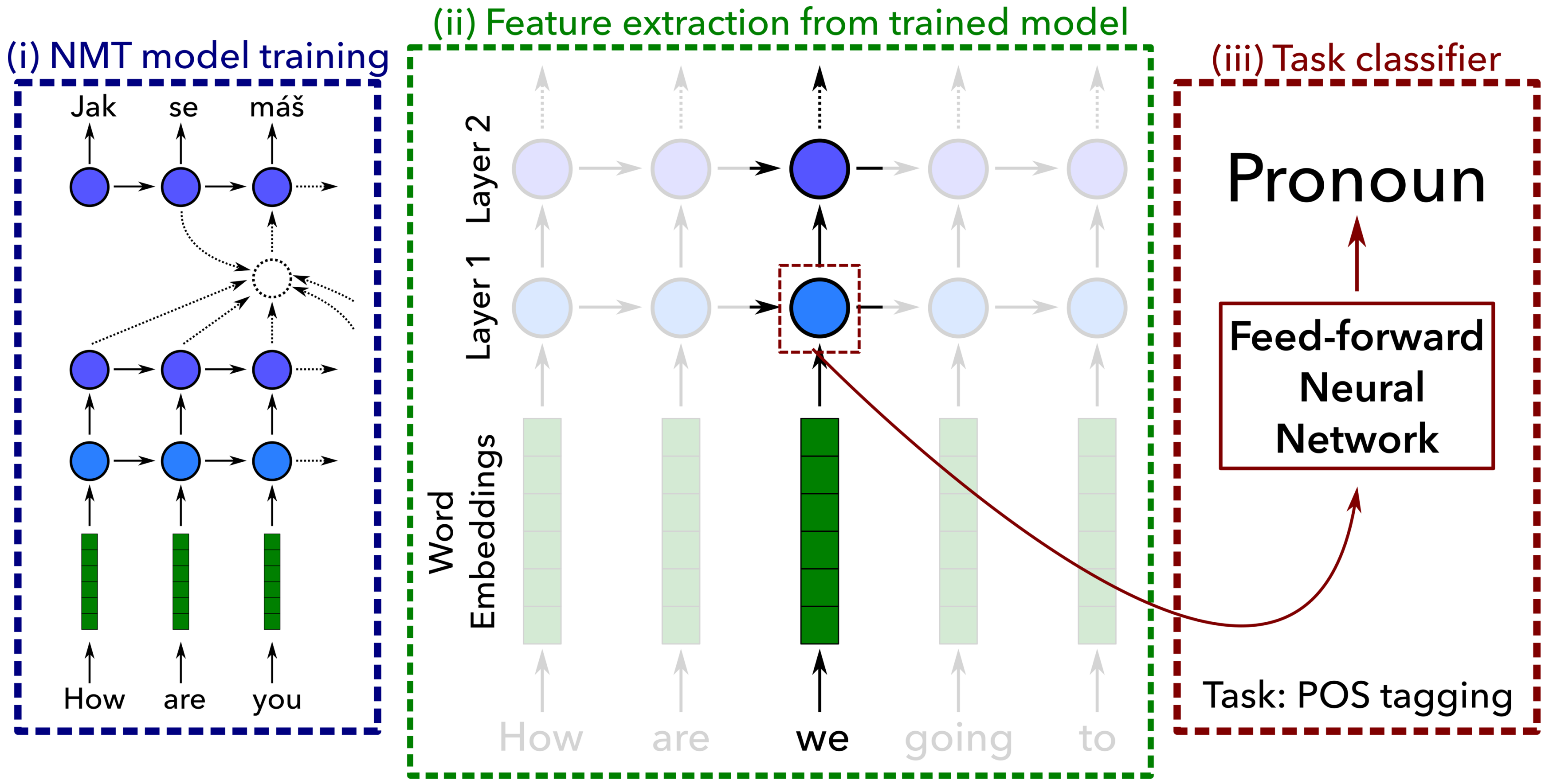
(i) NMT model training



Methodology



Methodology



Part A: Morphology

Experimental Setup

- Tasks
  - Part-of-speech tagging
  - Morphological tagging
- Languages
  - Arabic-, German-, French-, and Czech-English
  - Arabic-Hebrew (rich and similar)
  - Arabic-German (rich but different)

Experimental Setup

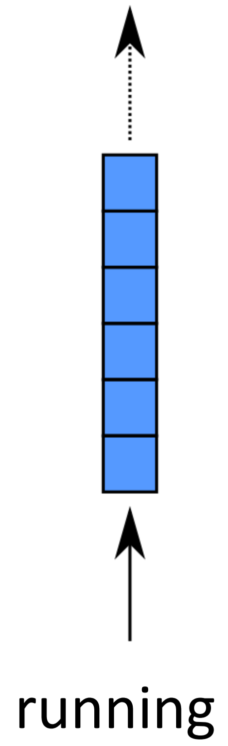
- MT data: TED talks
- Annotated data
  - Gold tags
  - Predicted tags

	Ar	De	Fr	Cz
	Gold/Pred	Gold/Pred	Pred	Pred
Train Tokens	0.5M/2.7M	0.9M/4.0M	5.2M	2.0M
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	–	–

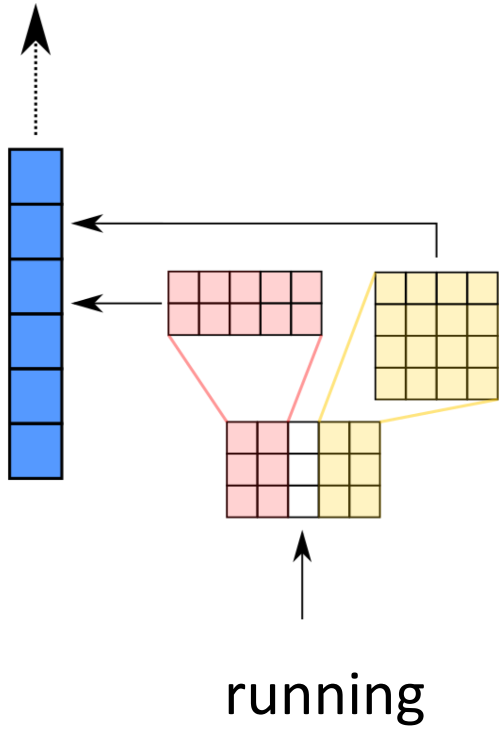
Encoder

Effect of Word Representation

Word embedding



Character CNN



Effect of Word Representation

	POS Accuracy		BLEU	
	Word	**Char**	**Word**	**Char**
Ar-En				
Ar-He				
De-En				
Fr-En				
Cz-En				

Effect of Word Representation

	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

Effect of Word Representation

	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 generate better representations for POS tagging

Effect of Word Representation

	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

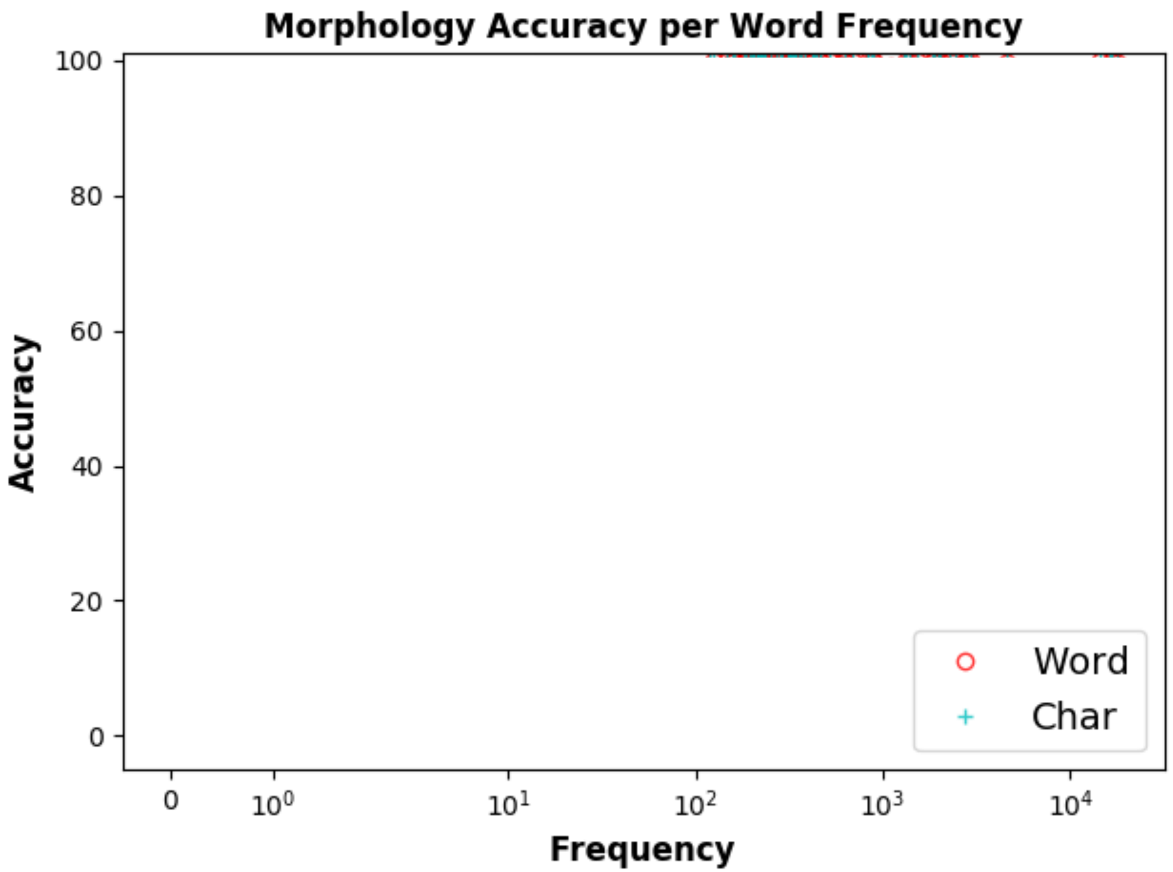
- Especially with richer morphological systems

Effect of Word Representation

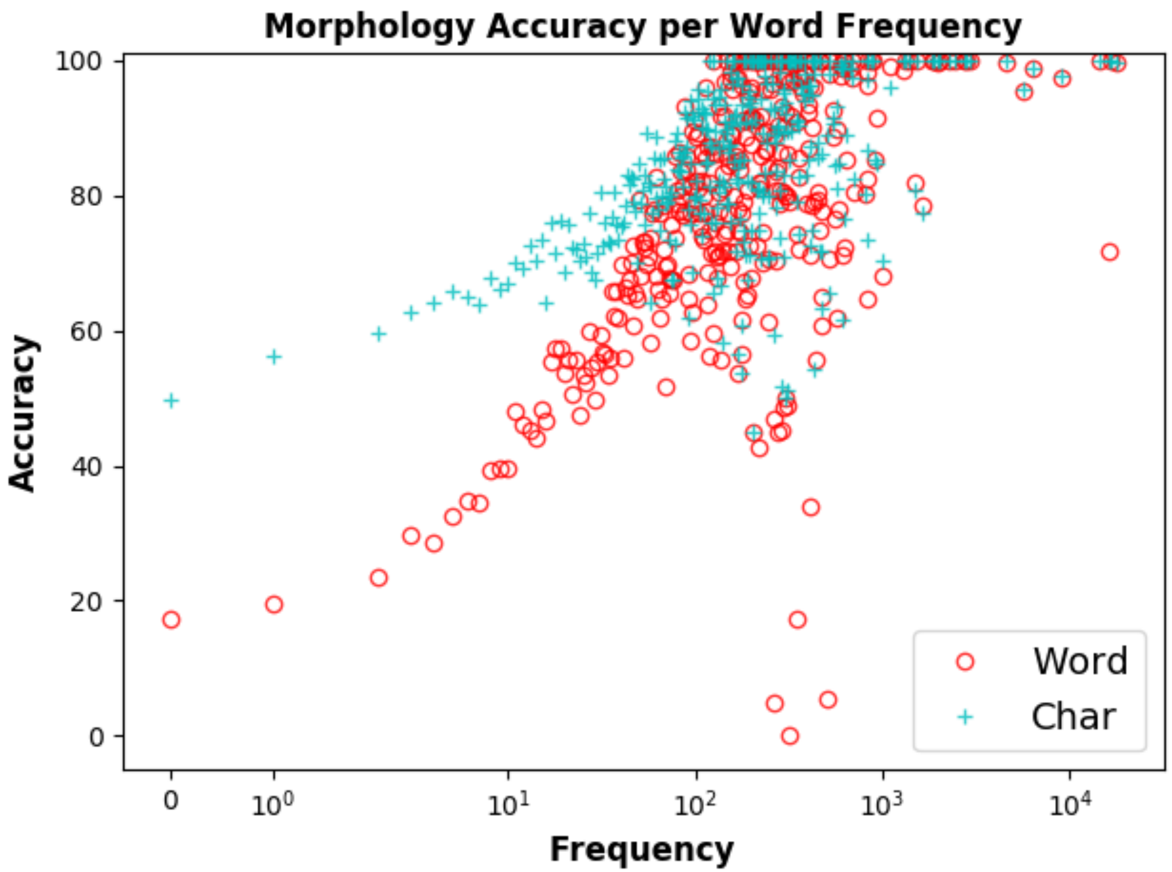
	POS Accuracy		BLEU	
	Word	**Char**	**Word**	**Char**
Ar-En	89.62	95.35	24.7	28.4
Ar-He	88.33	94.66	9.9	10.7
De-En	93.54	94.63	29.6	30.4
Fr-En	94.61	95.55	37.8	38.8
Cz-En	75.71	79.10	23.2	25.4

- Character-based models improve translation quality

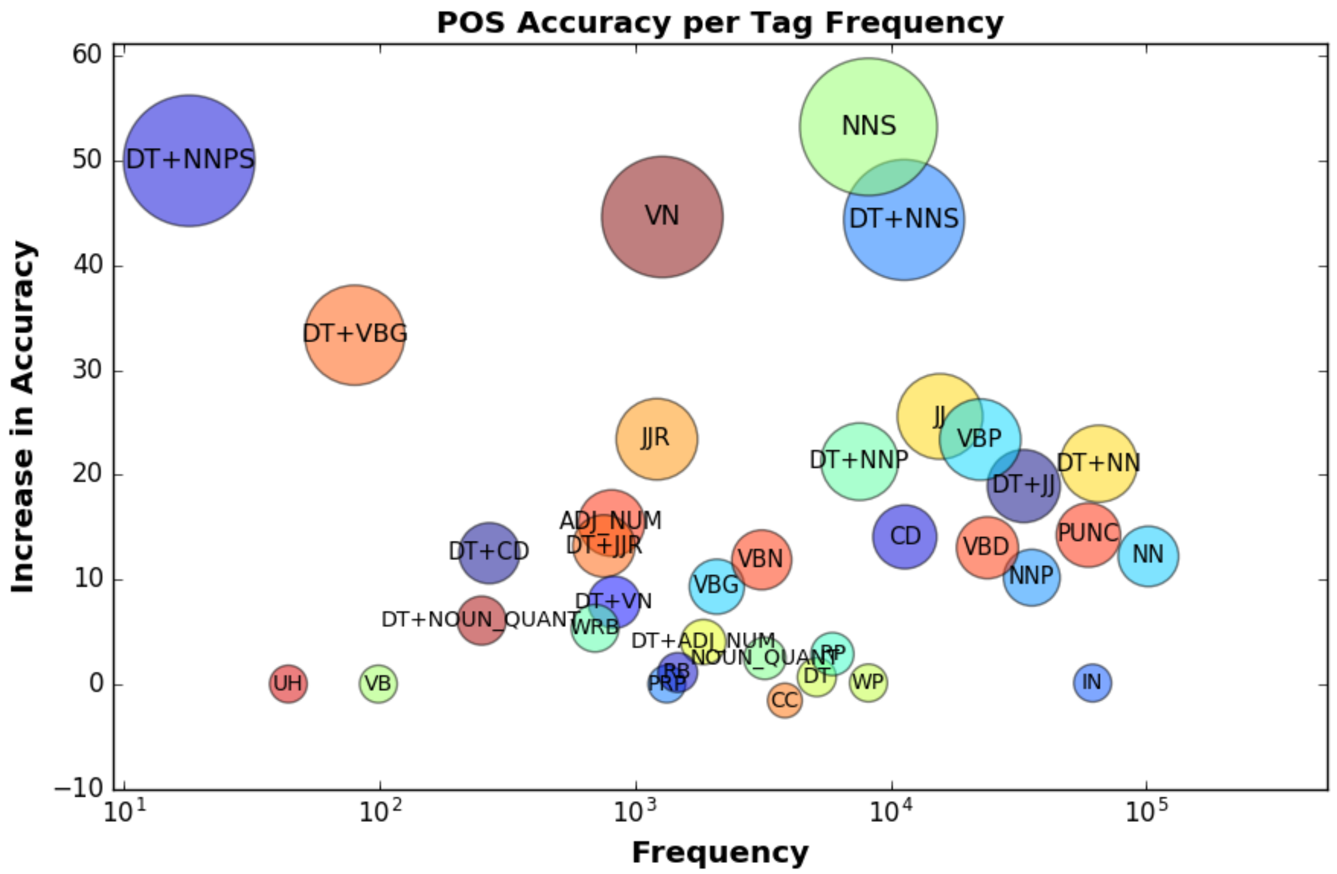
Impact of Word Frequency



Impact of Word Frequency

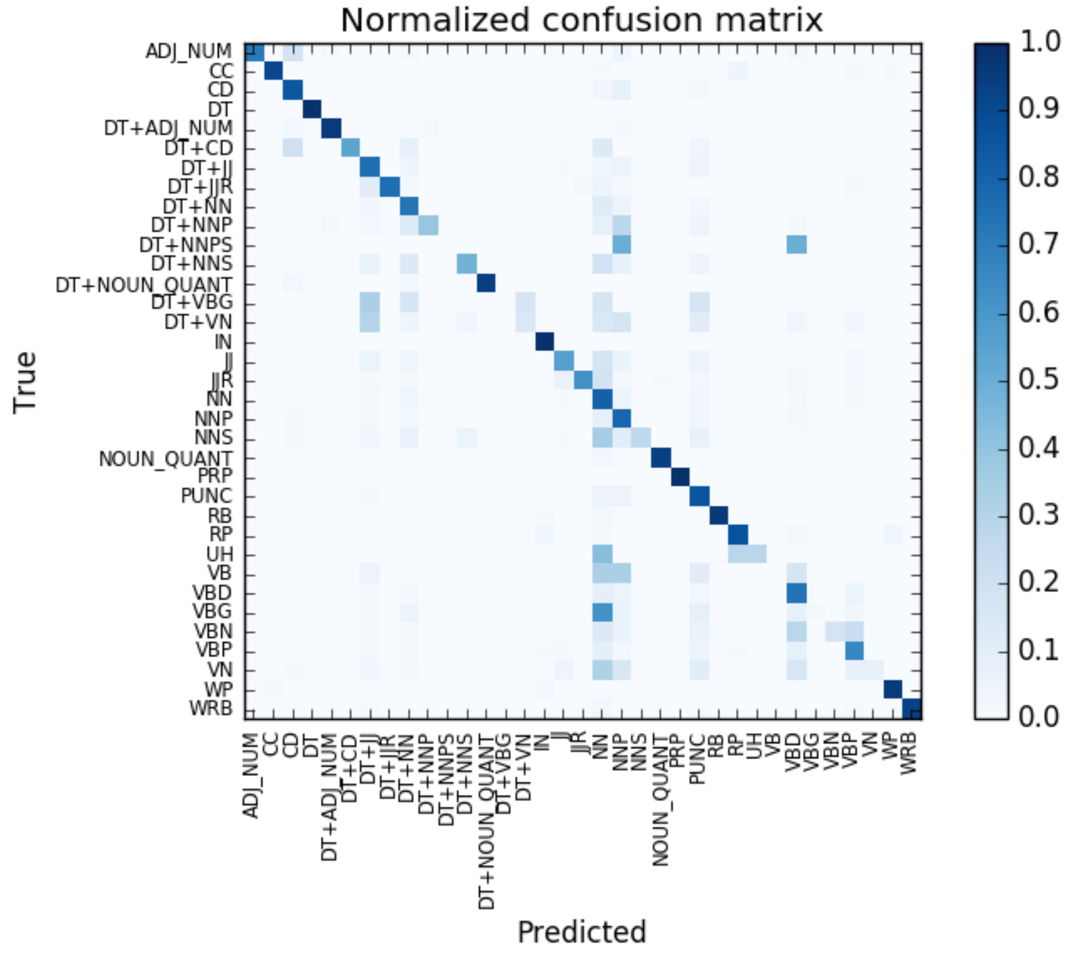


Impact of Tag Frequency

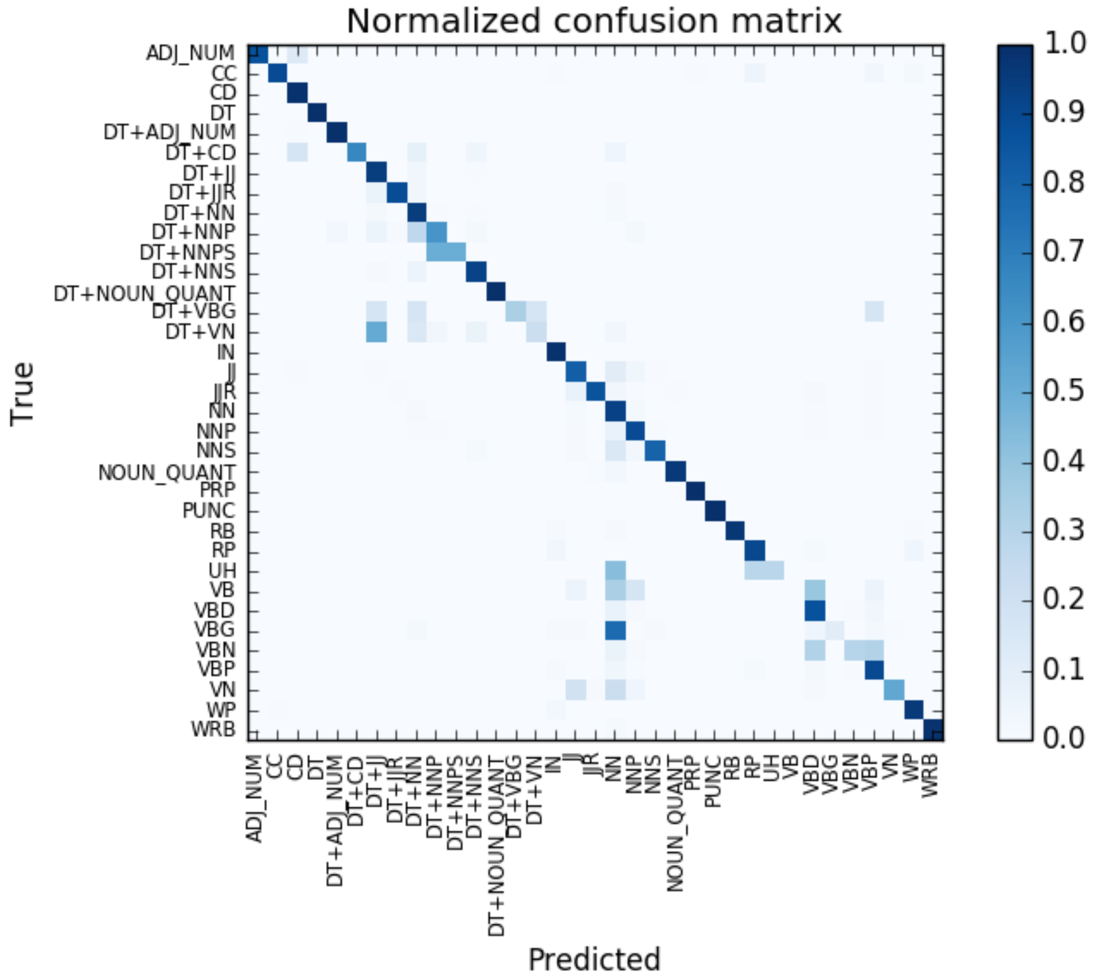


Comparing Specific Tags

Word-based



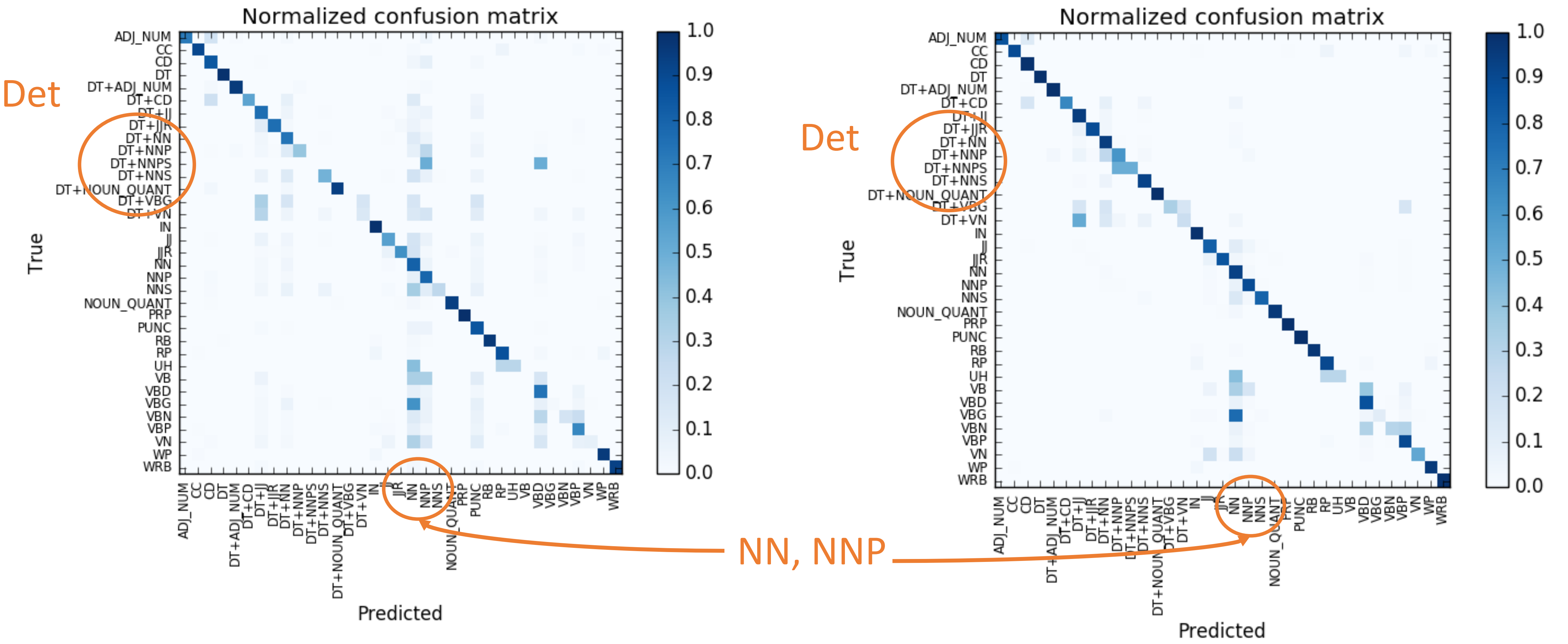
Char-based



Comparing Specific Tags

Word-based

Char-based



Effect of Encoder Depth

- NMT models can be very deep
  - Google Translate: 8 encoder/decoder layers
  - Zhou+ 2016: 16 layers

Effect of Encoder Depth

- NMT models can be very deep
  - Google Translate: 8 encoder/decoder layers
  - Zhou+ 2016: 16 layers

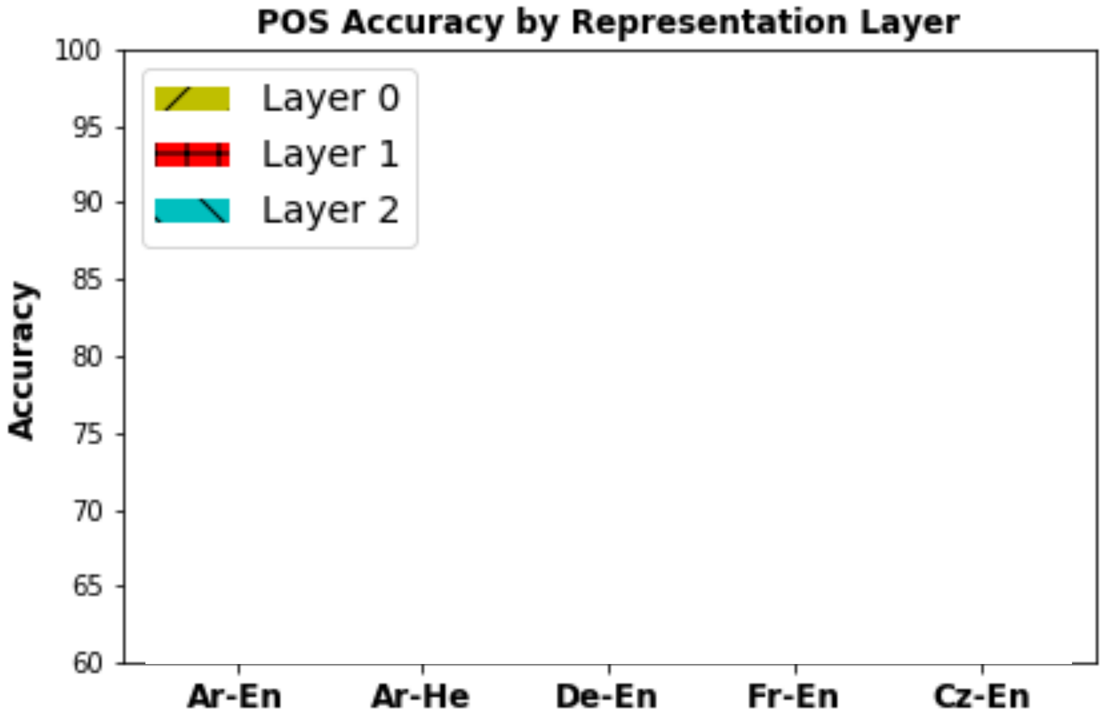
- What kind of information is learned at each?

Effect of Encoder Depth

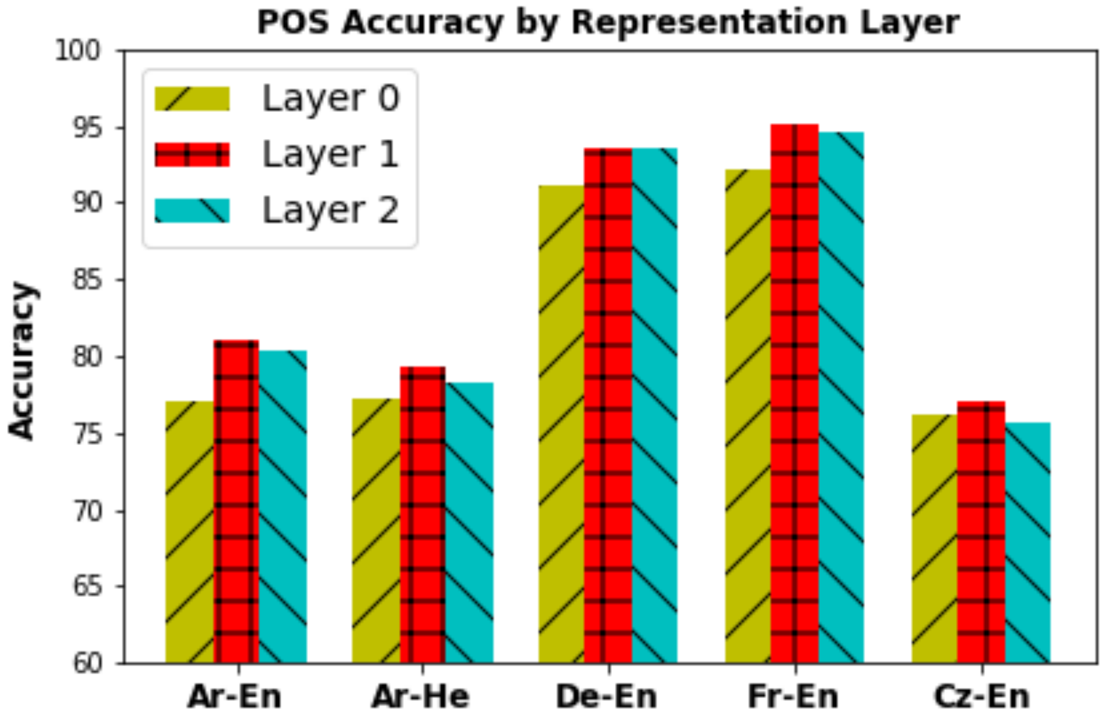
- NMT models can be very deep
  - Google Translate: 8 encoder/decoder layers
  - Zhou+ 2016: 16 layers

- What kind of information is learned at each?
- We analyzed a 2-layer encoder
  - Extract representations from different layers for training the classifier

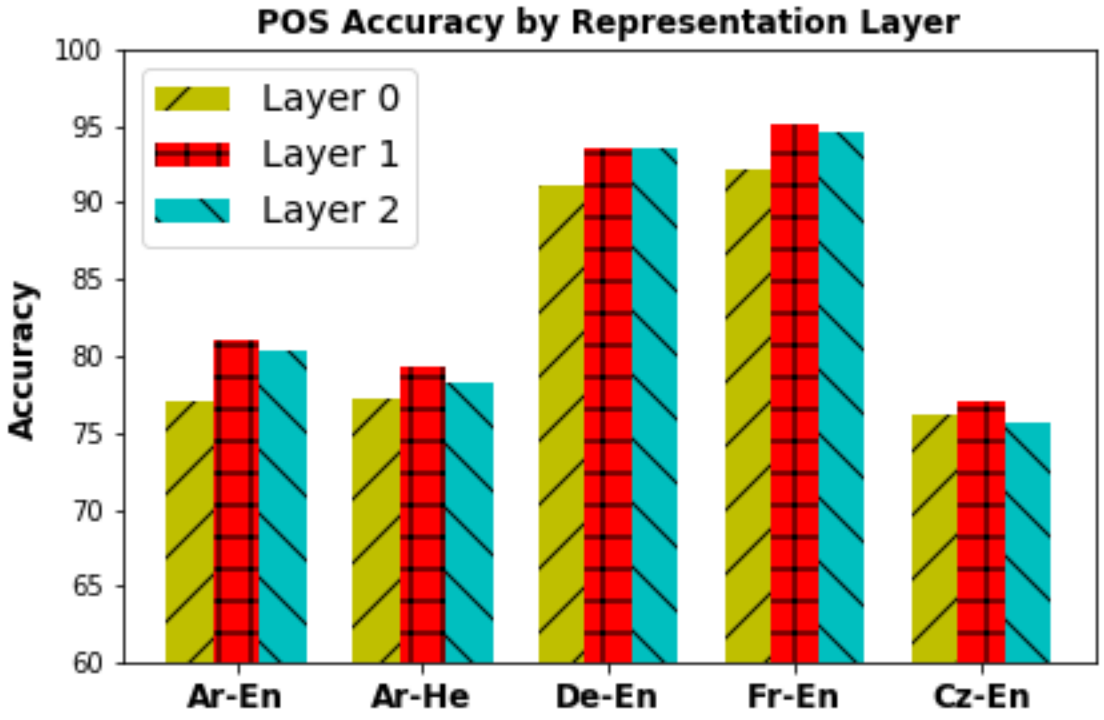
Effect of Encoder Depth



Effect of Encoder Depth

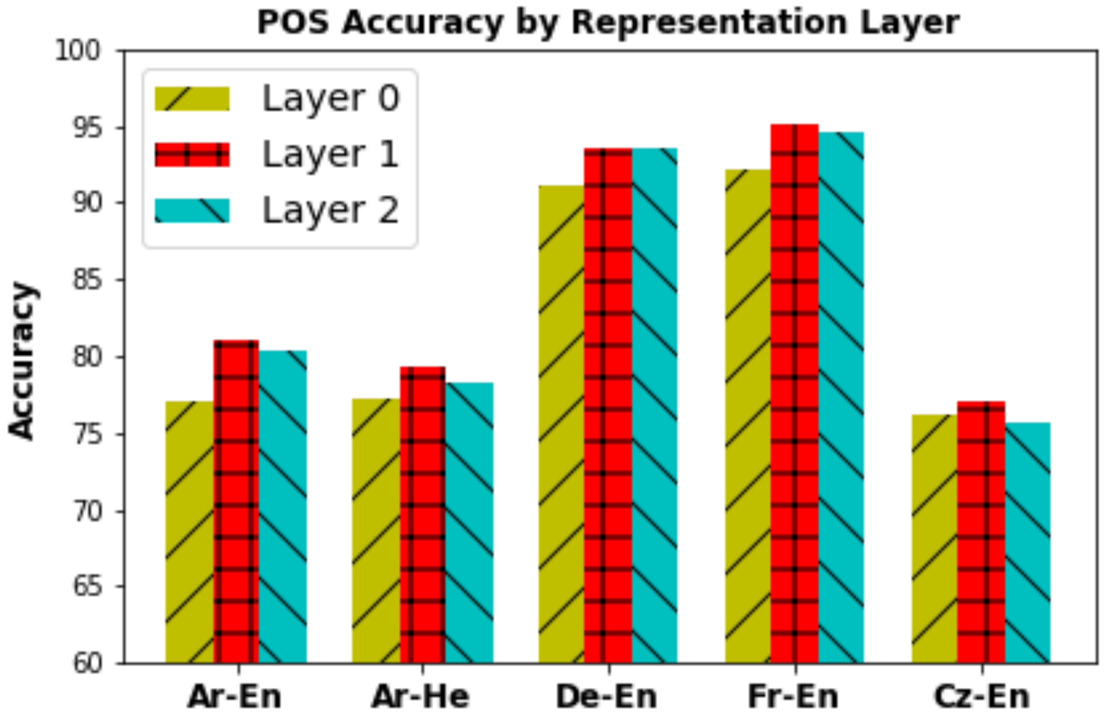


Effect of Encoder Depth



- Layer 1 > Layer 2 > Layer 0
- But deeper models translate better

Effect of Encoder Depth



- Is layer 2 learning more about semantics? More on that later…

Effect of Target Language

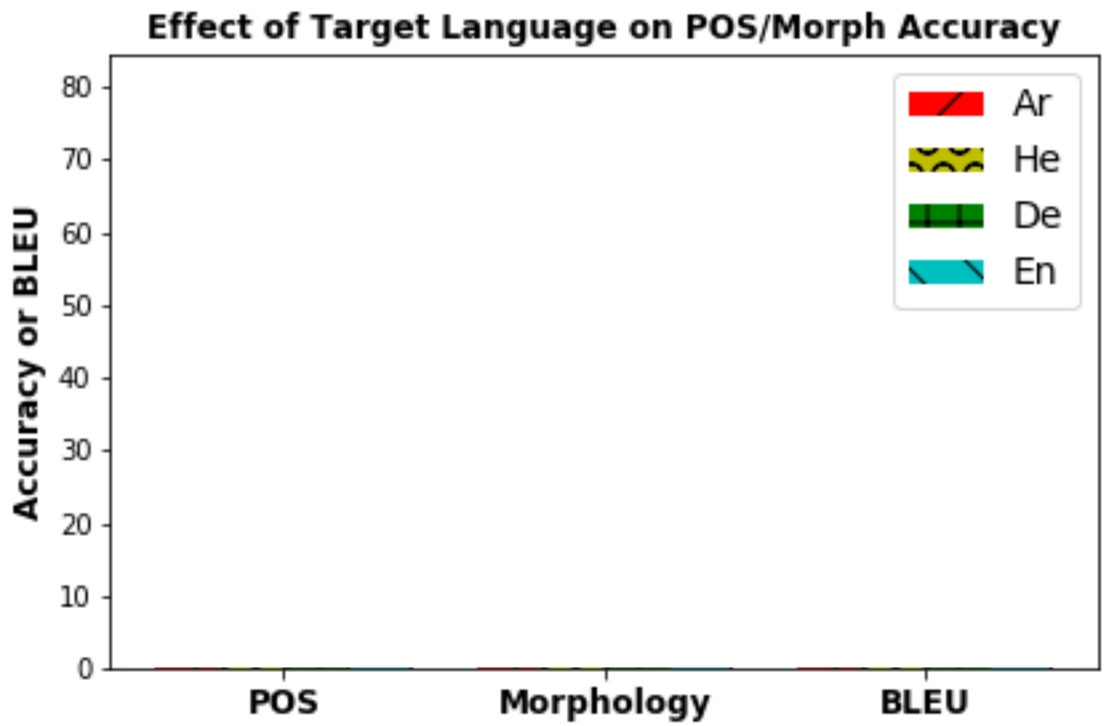
- How does the target language affect the learned source language representations?

Effect of Target Language

- How does the target language affect the learned source language representations?

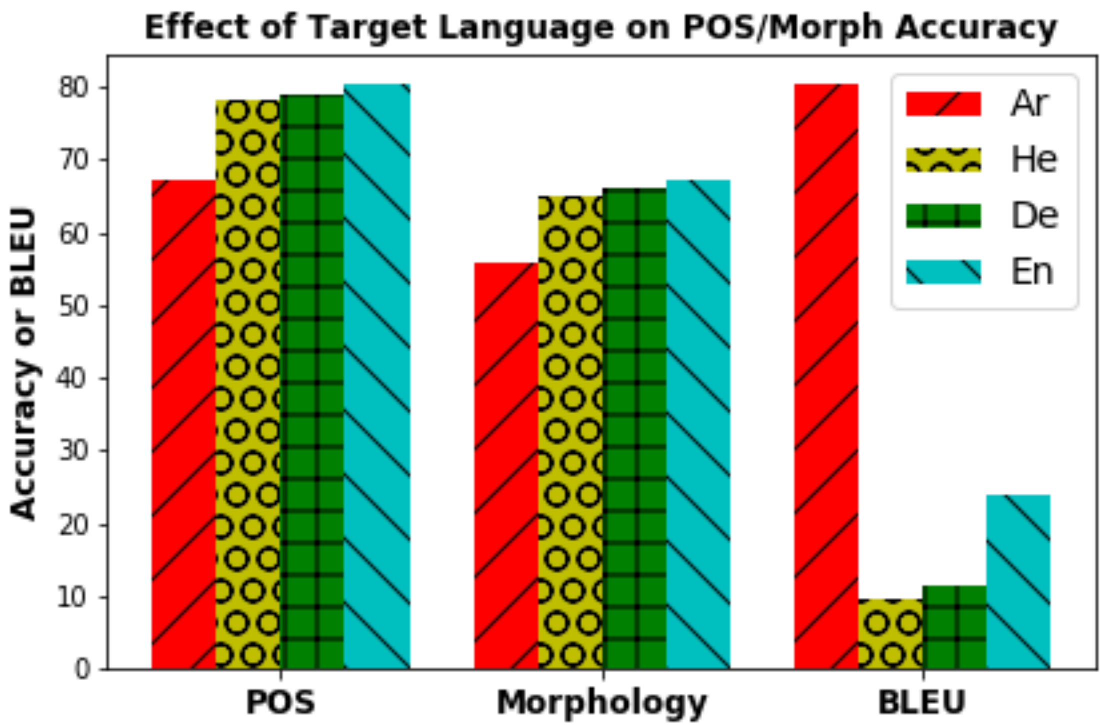
- Experiment:
  - Fix source side and train NMT models on different target languages
  - Compare learned representations on POS/morphological tagging

Effect of Target Language



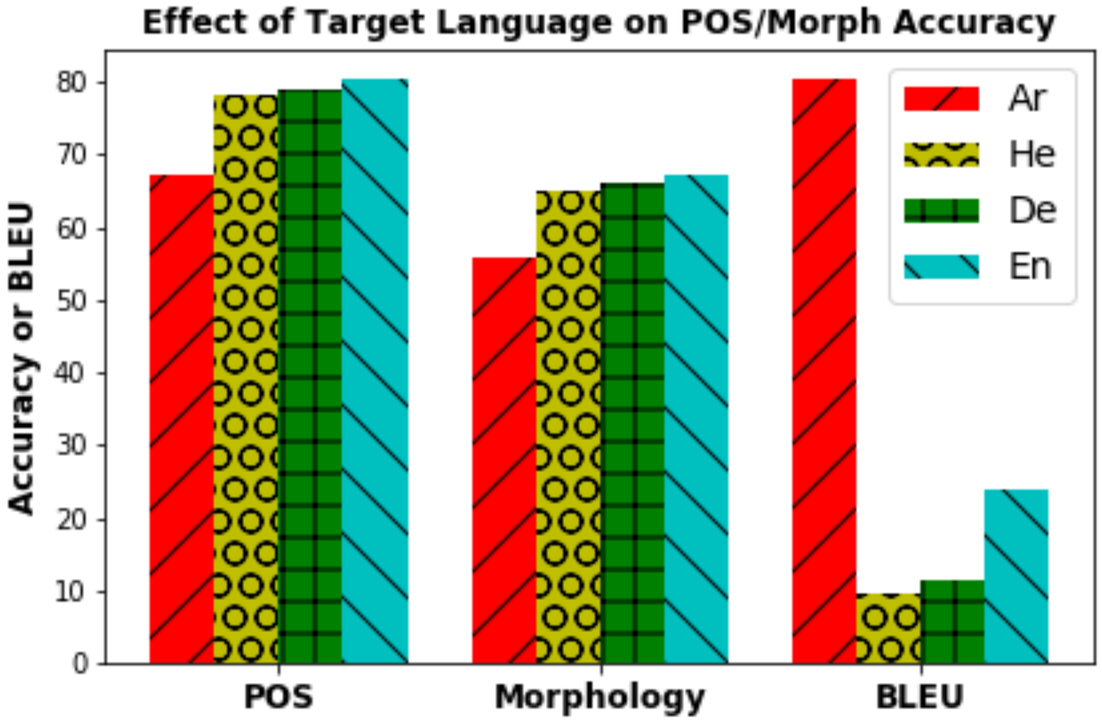
- Source language: Arabic
- Target languages: English, German, Hebrew, Arabic

Effect of Target Language



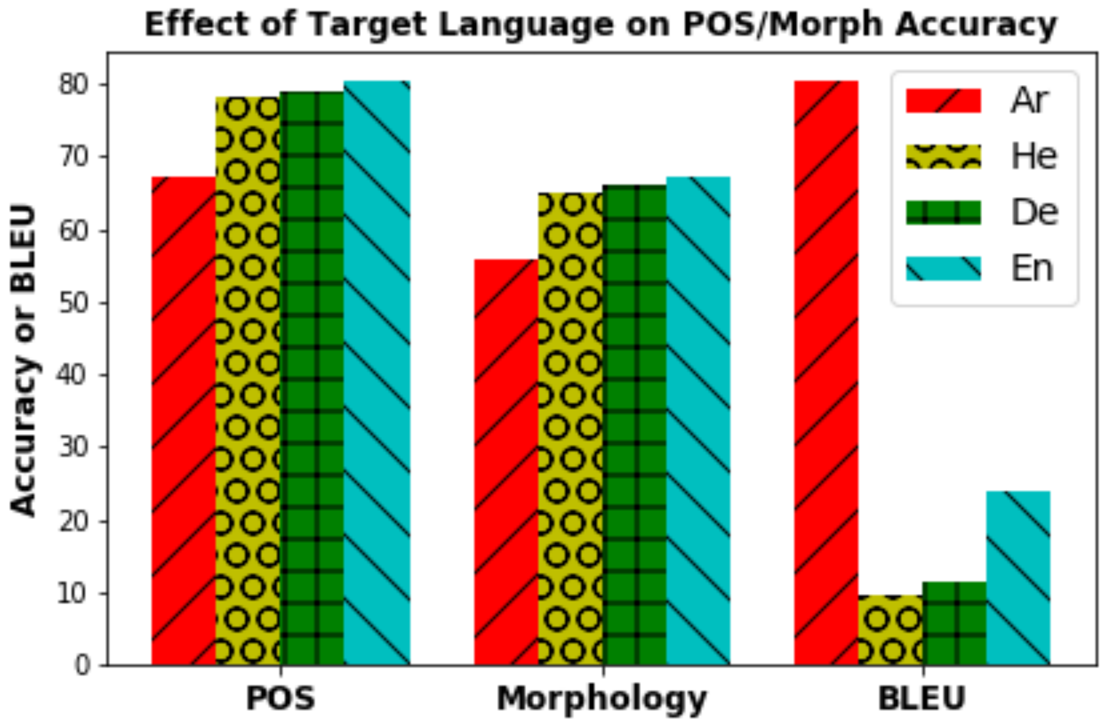
- Source language: Arabic
- Target languages: English, German, Hebrew, Arabic

Effect of Target Language



- Poorer morphology on target side,
  better source side representations for morphology

Effect of Target Language



- Higher BLEU ≠ better representations

Decoder

Encoder vs Decoder

	POS Accuracy	
	Encoder	**Decoder**
Arabic ↔ English		
German↔ English		
Czech↔ English		

Encoder vs Decoder

	POS Accuracy	
	Encoder	**Decoder**
Arabic ↔ English	89.6	43.9
German↔ English	93.5	53.6
Czech↔ English	75.7	36.3

Encoder vs Decoder

	POS Accuracy	
	Encoder	**Decoder**
Arabic ↔ English	89.6	43.9
German↔ English	93.5	53.6
Czech↔ English	75.7	36.3

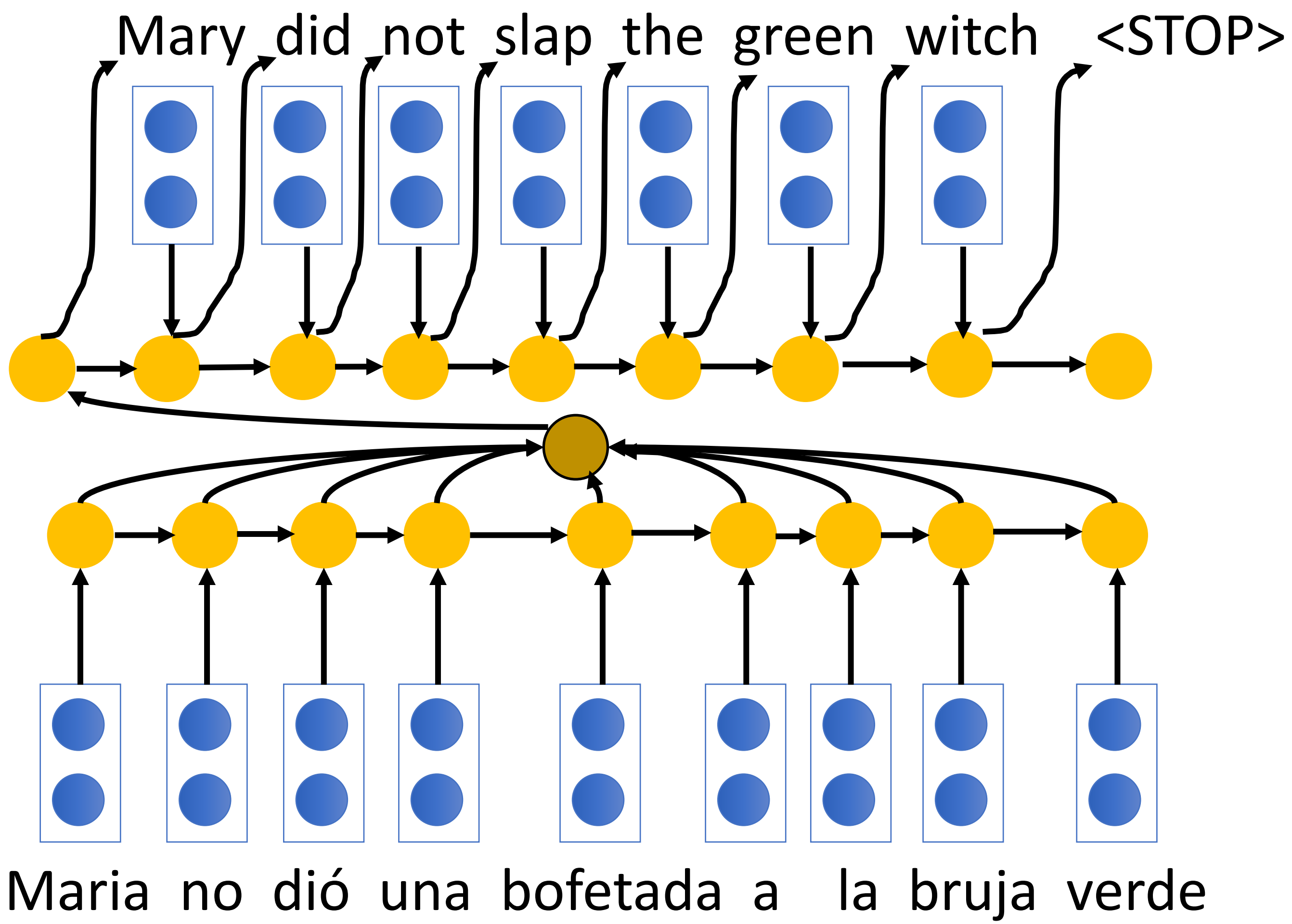
- The decoder learns very little about target language morphology

Encoder vs Decoder

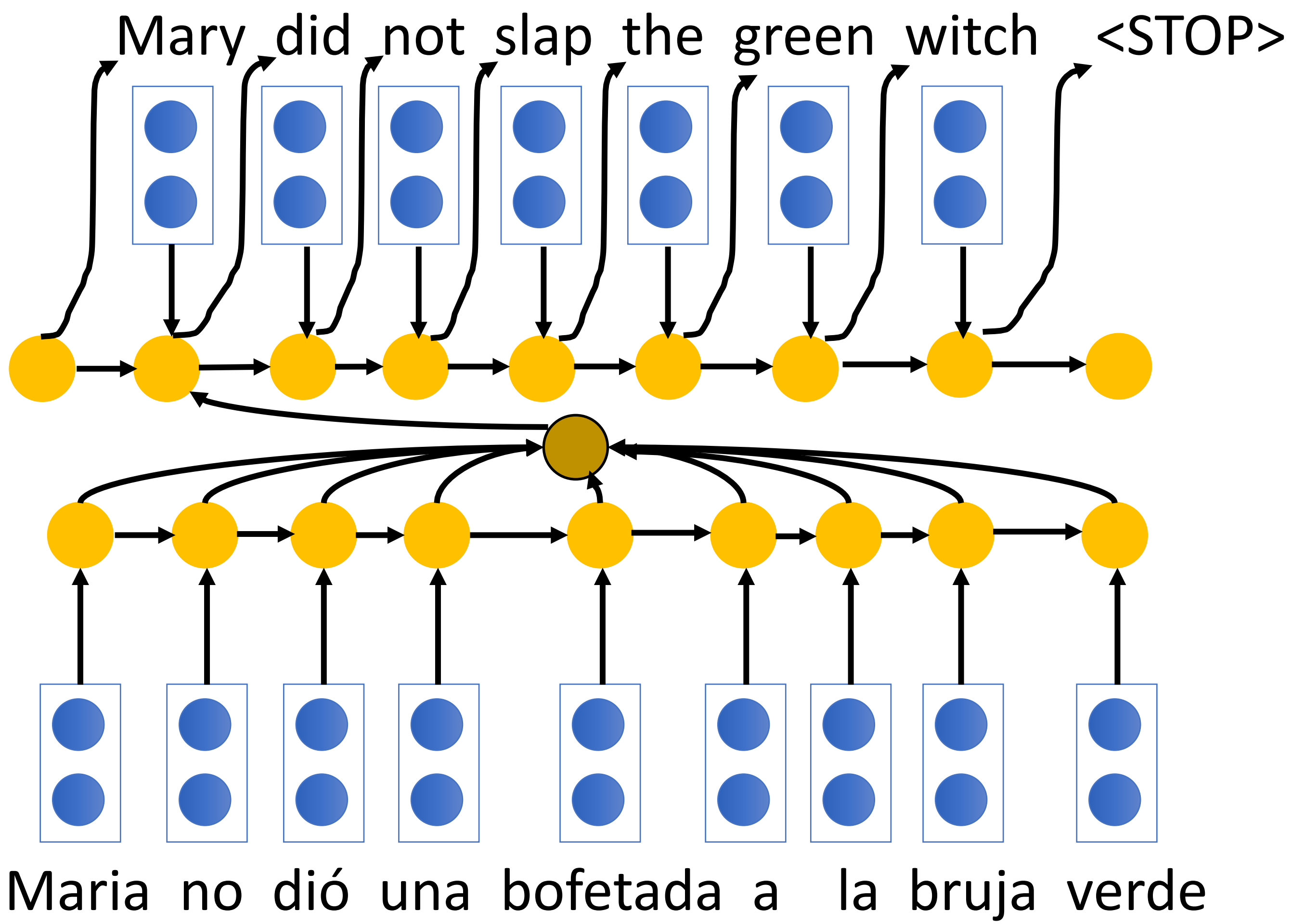
	POS Accuracy	
	Encoder	**Decoder**
Arabic ↔ English	89.6	43.9
German↔ English	93.5	53.6
Czech↔ English	75.7	36.3

- The decoder learns very little about target language morphology
- Why?

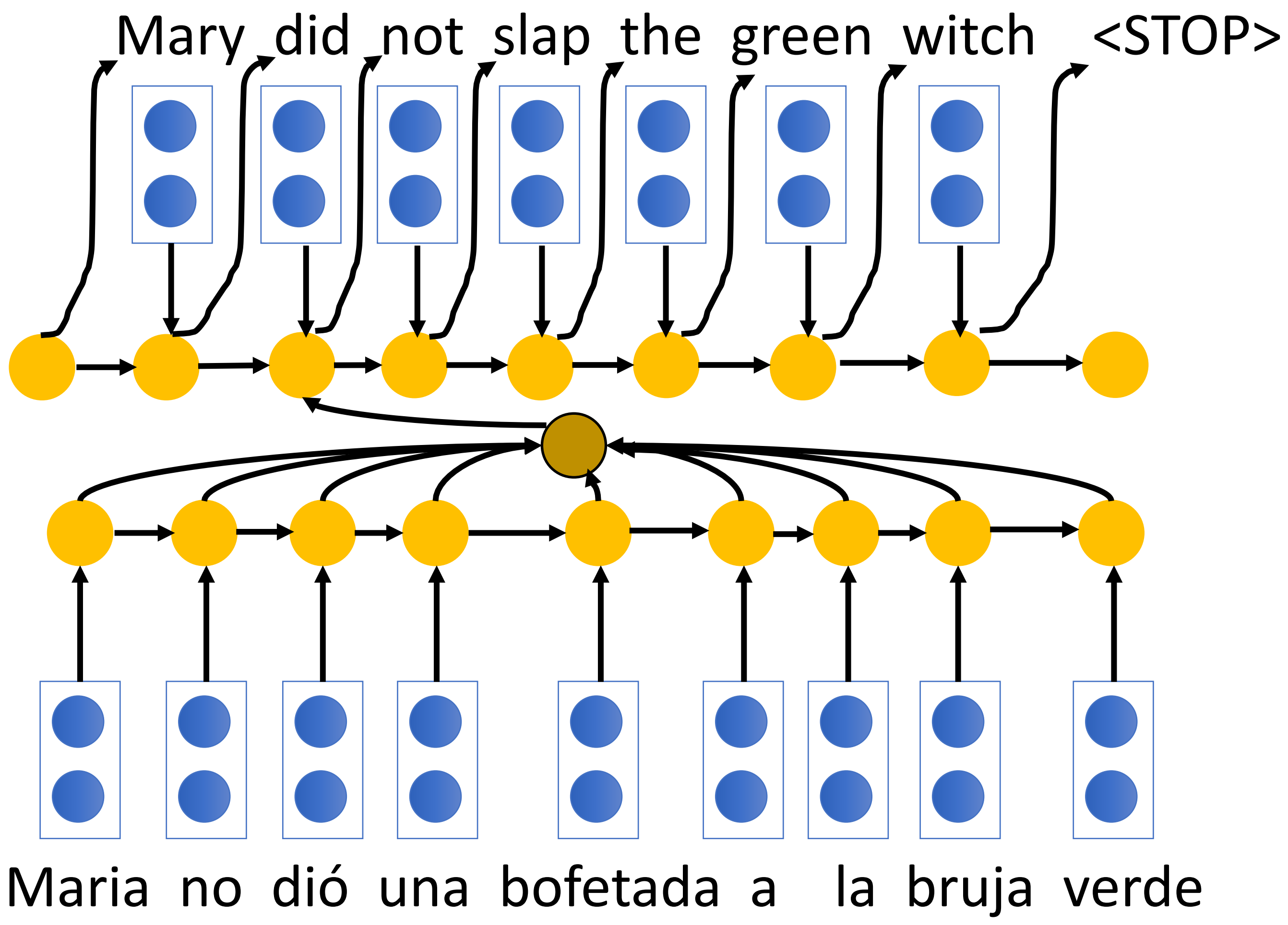
Effect of Attention



Effect of Attention



Effect of Attention



Effect of Attention

	With attention	Without attention
English → German		
English → Czech		

Effect of Attention

	With attention	Without attention
English → German	44.55	50.26
English → Czech	36.35	42.09

Effect of Attention

	With attention	Without attention
English → German	44.55	50.26
English → Czech	36.35	42.09

- Removing attention improves decoder representations
- Attention is removing burden off of the decoder
- The decoder does not need to learn as much about target words

Effect of Attention

	With attention	Without attention	With most attended word
English → German	44.55	50.26	60.34
English → Czech	36.35	42.09	48.64

- Concatenating most attended word improves performance
- Encoder representations helpful for target morphology

Effect of Attention

	With attention	Without attention	With most attended word	Only most attended word
English → German	44.55	50.26	60.34	43.43
English → Czech	36.35	42.09	48.64	36.36

- Concatenating most attended word improves performance
- Encoder representations helpful for target morphology
- But using only encoder side is not as good

Summary

- NMT encoder learns good representations for morphology
- Character-based representations much better than word-based
- Target language impacts source side representations
- Layer 1 > Layer 2 > Layer 0

- Decoder learns poor target side representations
- Attention model helps decoder exploit source representations

Summary

- NMT encoder learns good representations for morphology
- Character-based representations much better than word-based
- Target language impacts source side representations
- Layer 1 > Layer 2 > Layer 0

- Decoder learns poor target side representations
- Attention model helps decoder exploit source representations

Part B: Semantics

Recap

- We saw
  - NMT representations from layer 1 better than layer 2 (and layer 0) for POS and morphological tagging
  - Deeper networks lead to better translation performance

Recap

- We saw
  - NMT representations from layer 1 better than layer 2 (and layer 0) for POS and morphological tagging
  - Deeper networks lead to better translation performance
- Questions
  - What is captured in higher layers?
  - How is semantic information represented?

Recap

- We saw
  - NMT representations from layer 1 better than layer 2 (and layer 0) for POS and morphological tagging
  - Deeper networks lead to better translation performance
- Questions
  - What is captured in higher layers?
  - How is semantic information represented?

- Let's apply a similar methodology to a semantic task

Semantic tagging

- Lexical semantics
- Abstraction over POS tagging
- Language-neutral, aimed for multi-lingual semantic parsing

Semantic tagging

- Lexical semantics
- Abstraction over POS tagging
- Language-neutral, aimed for multi-lingual semantic parsing

- Some examples
  - Determiners: *every*, *no*, *some*
  - Comma as conjunction, disjunction, apposition
  - Role nouns, entity nouns
  - Comparison adjectives: comparative, superlative, equative

Experimental Setup

- Semantic tagging data
  - 66 fine-grained tags,
    13 coarse categories

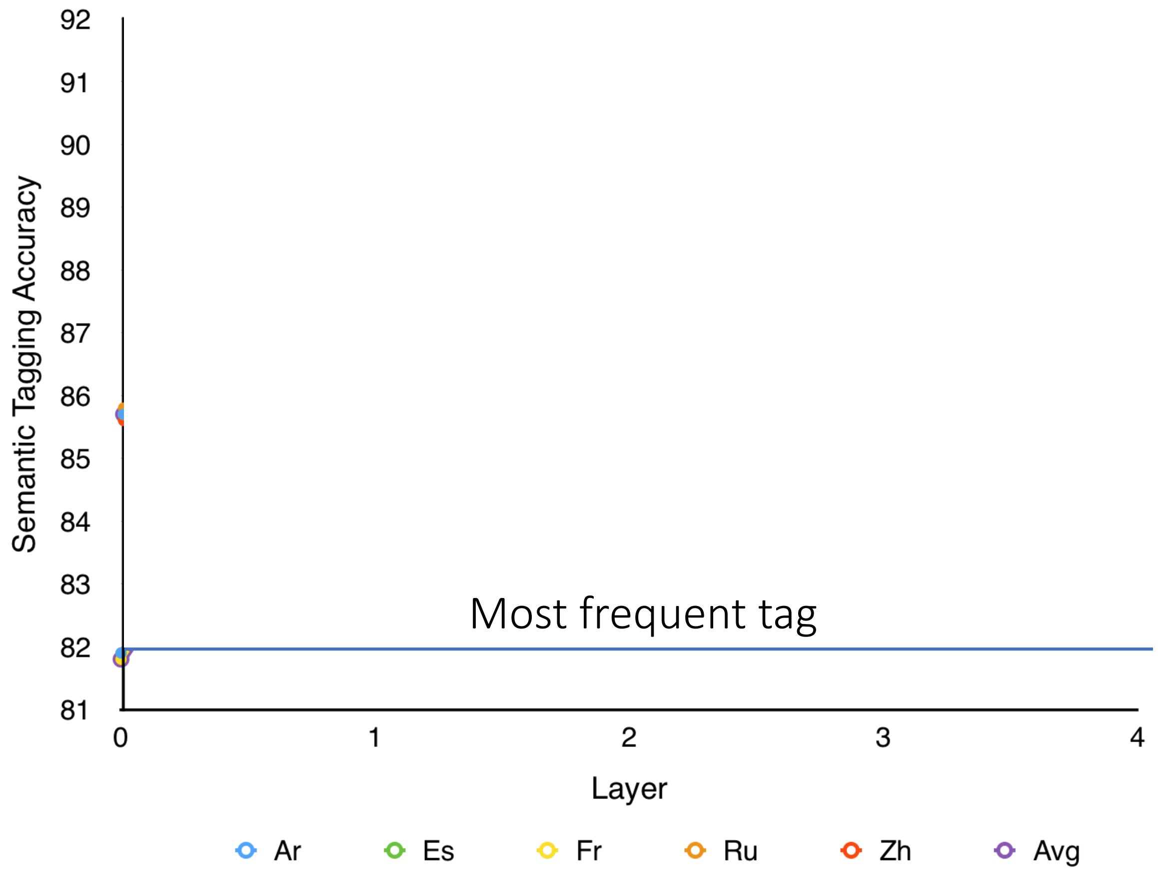
	Train	Dev	Test
Sentences	42.5K	6.1K	12.2K
Tokens	937.1K	132.3K	265.5K

- MT data – UN corpus
  - Multi-parallel
  - 11M sentences
  - Arabic, Chinese, English, French, Spanish, Russian

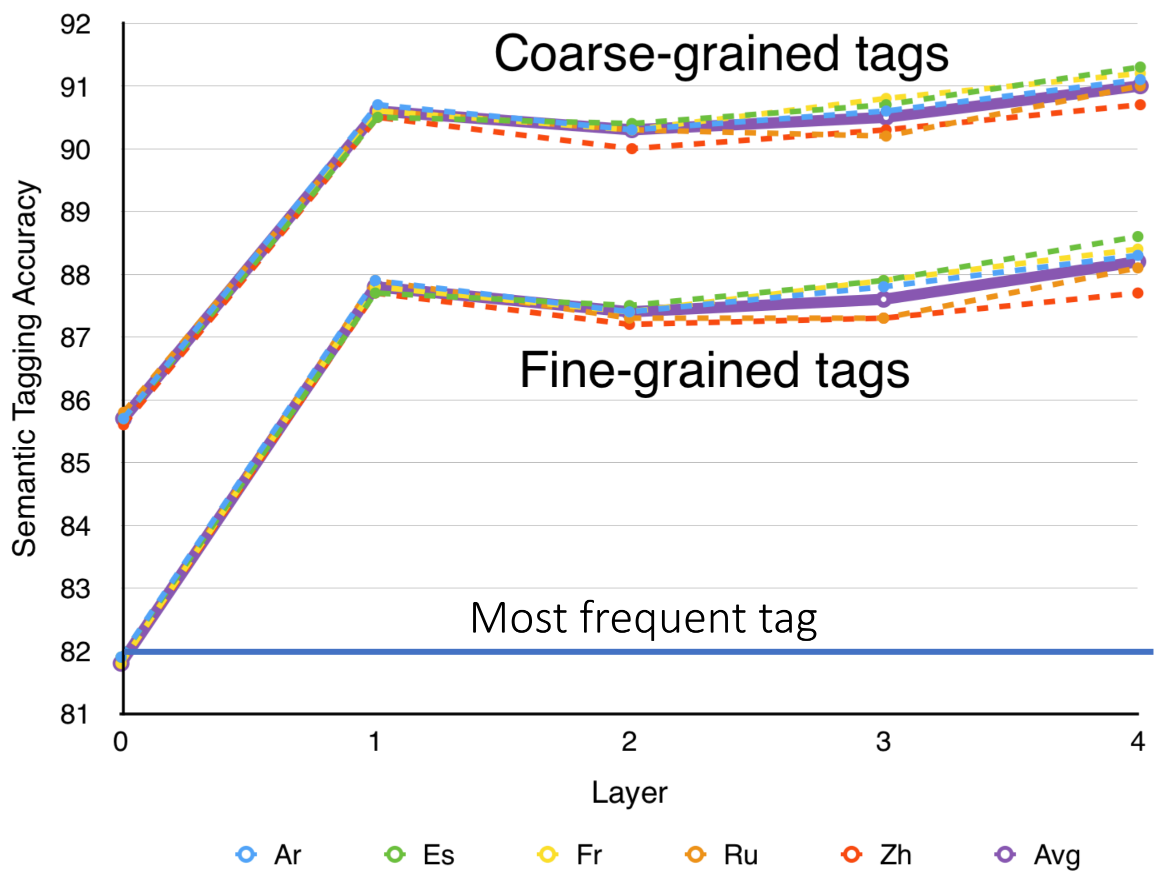
Baselines

System	Accuracy
Most frequent tag	82.0
Unsupervised embeddings	81.1
Word2Tag encoder-decoder	91.4
State-of-the-art (Bjerva+ 16)	95.5

Effect of Network Depth

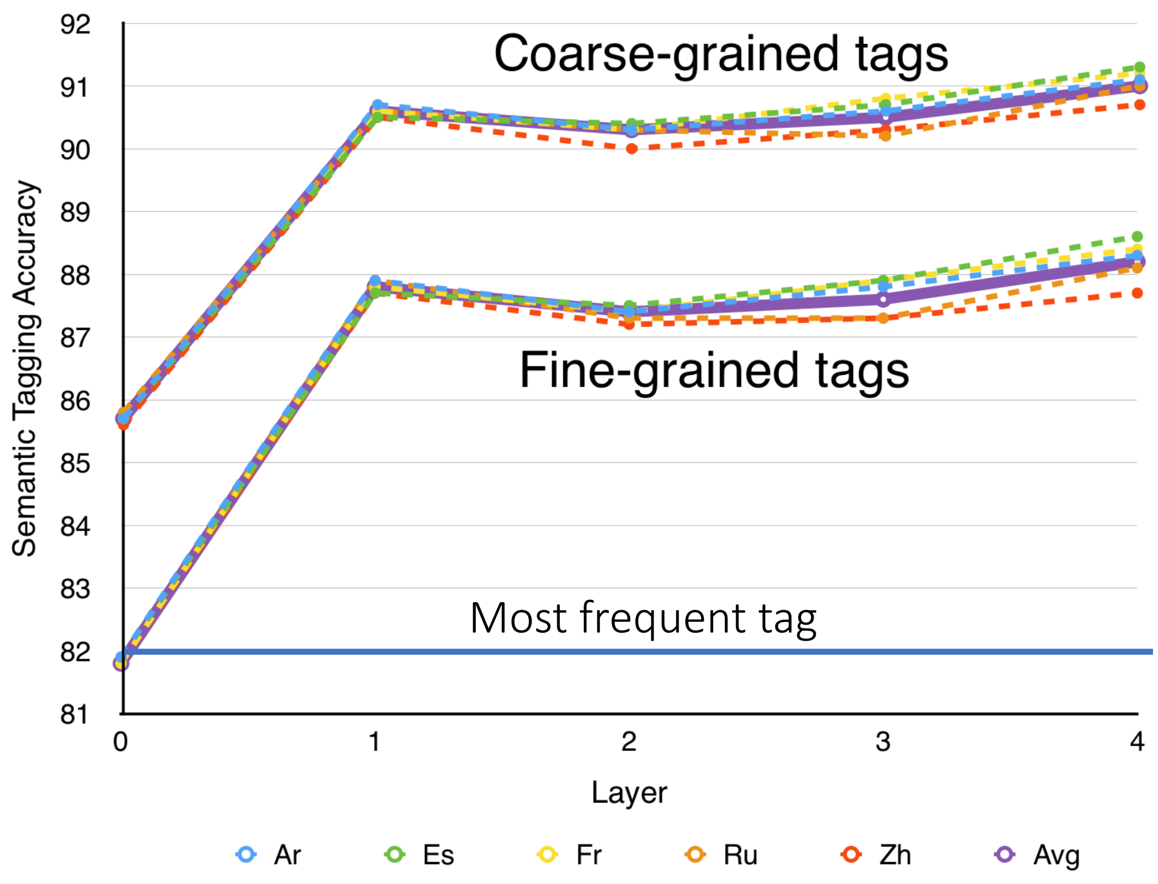


Effect of Network Depth



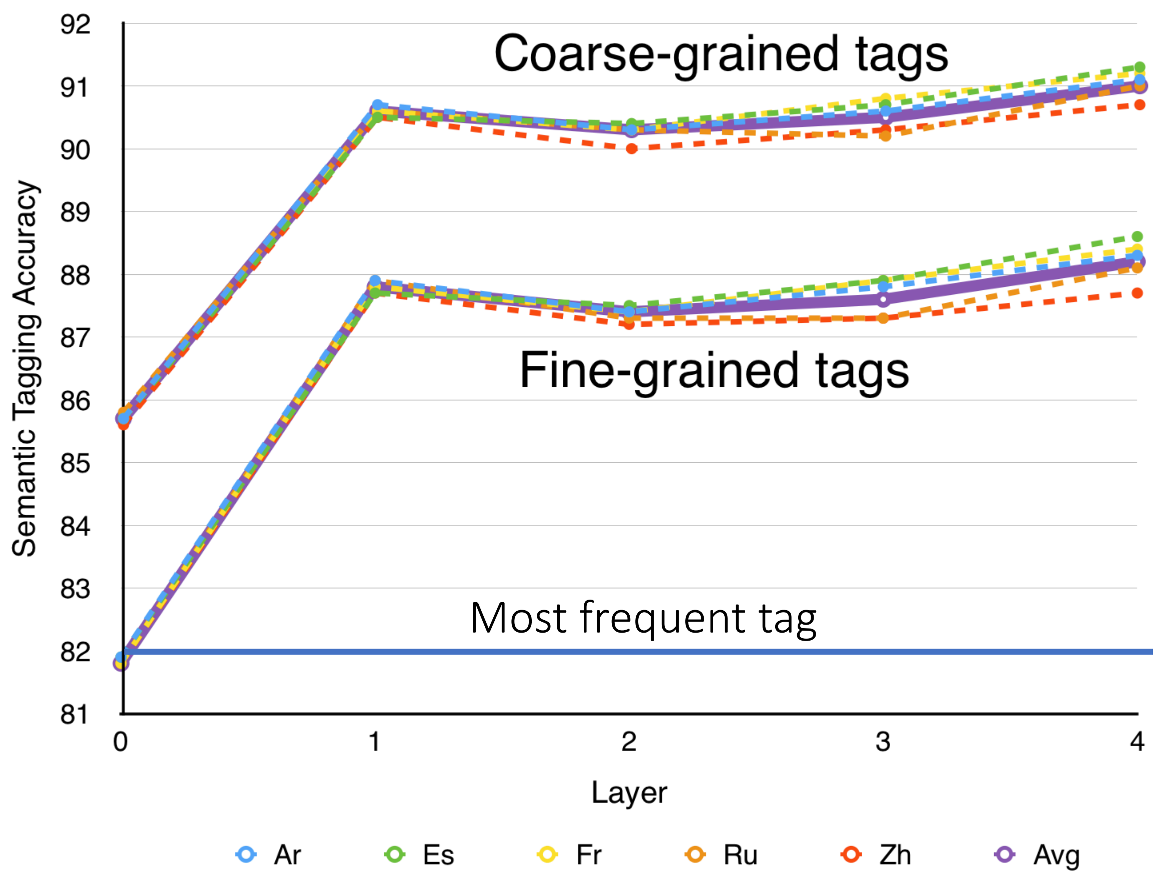
Effect of Network Depth

- Layer 0 below baseline



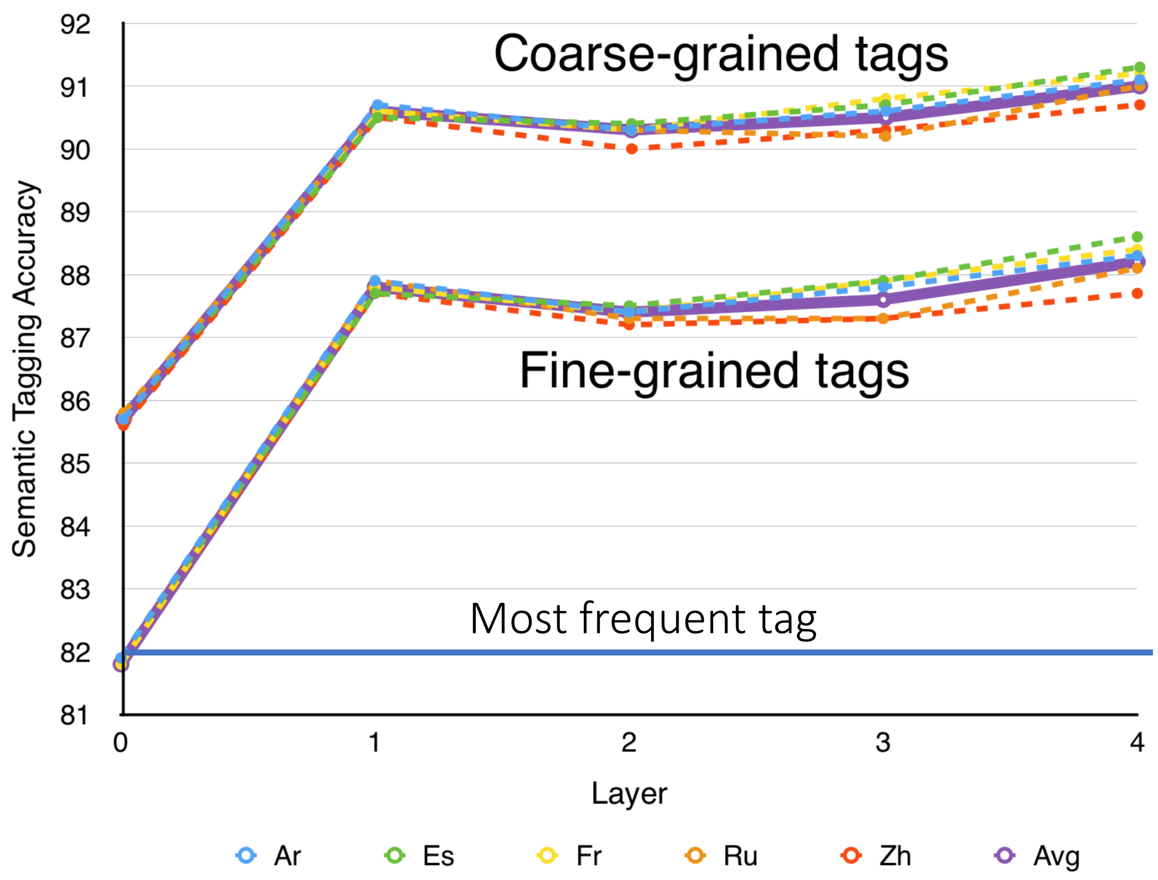
Effect of Network Depth

- Layer 0 below baseline
- Layer 1 >> layer 0



Effect of Network Depth

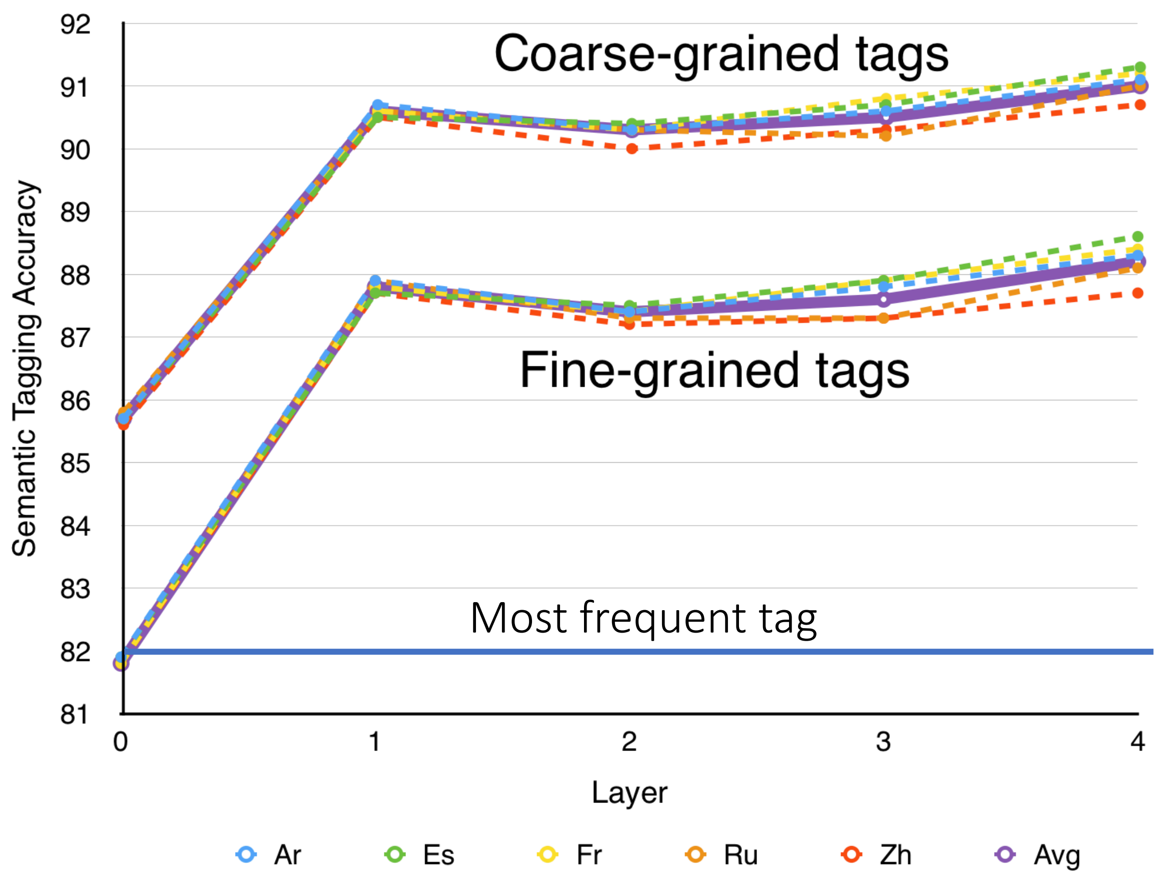
- Layer 0 below baseline
- Layer 1 >> layer 0
- Layer 4 > layer 1



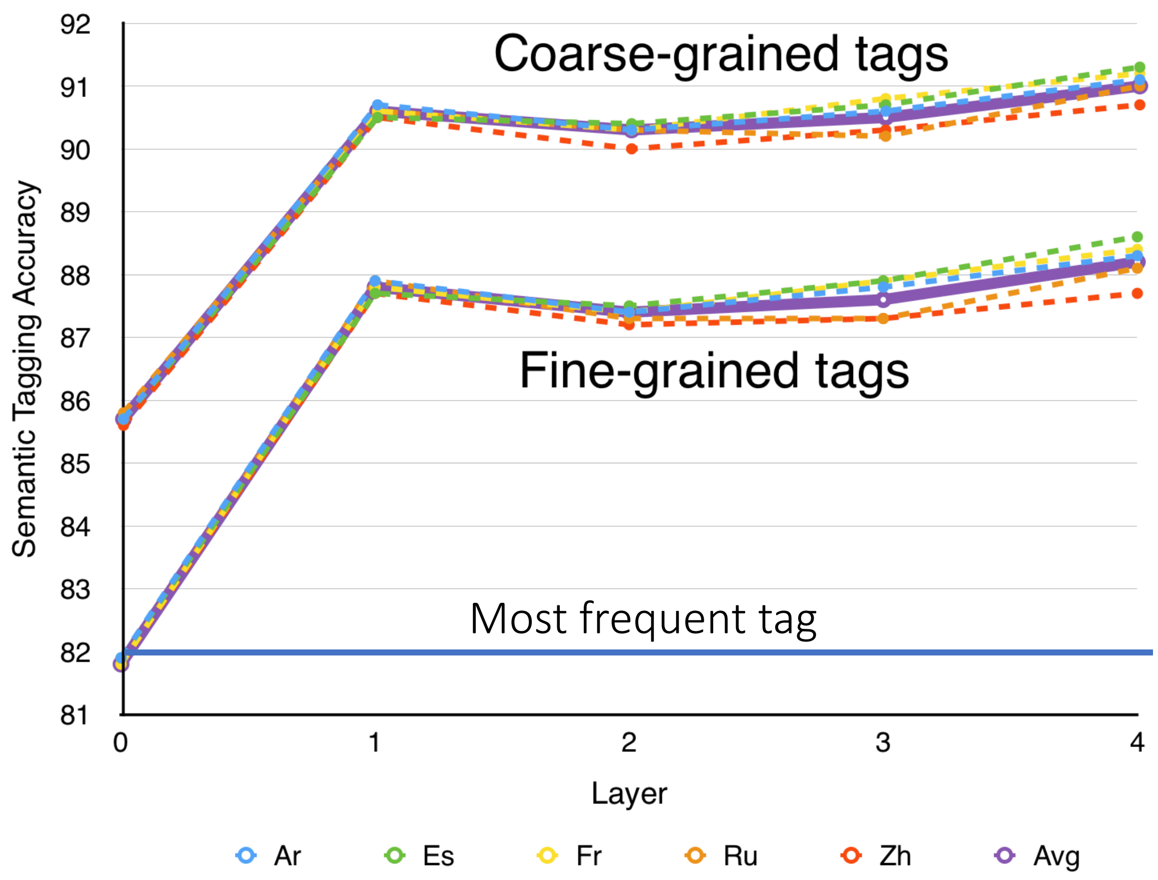
Effect of Network Depth

- Layer 0 below baseline
- Layer 1 >> layer 0
- Layer 4 > layer 1

- Similar trends
  for coarse tags

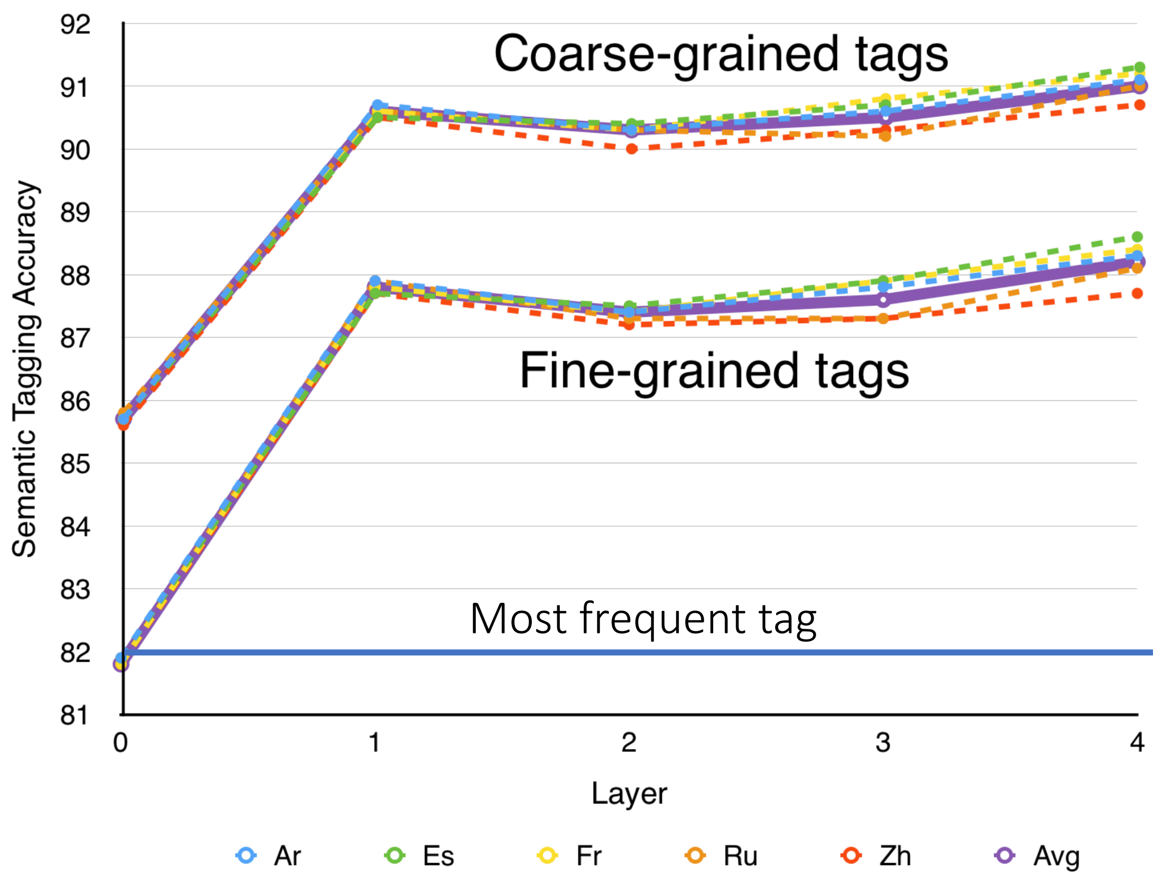


Effect of Target Language



Effect of Target Language

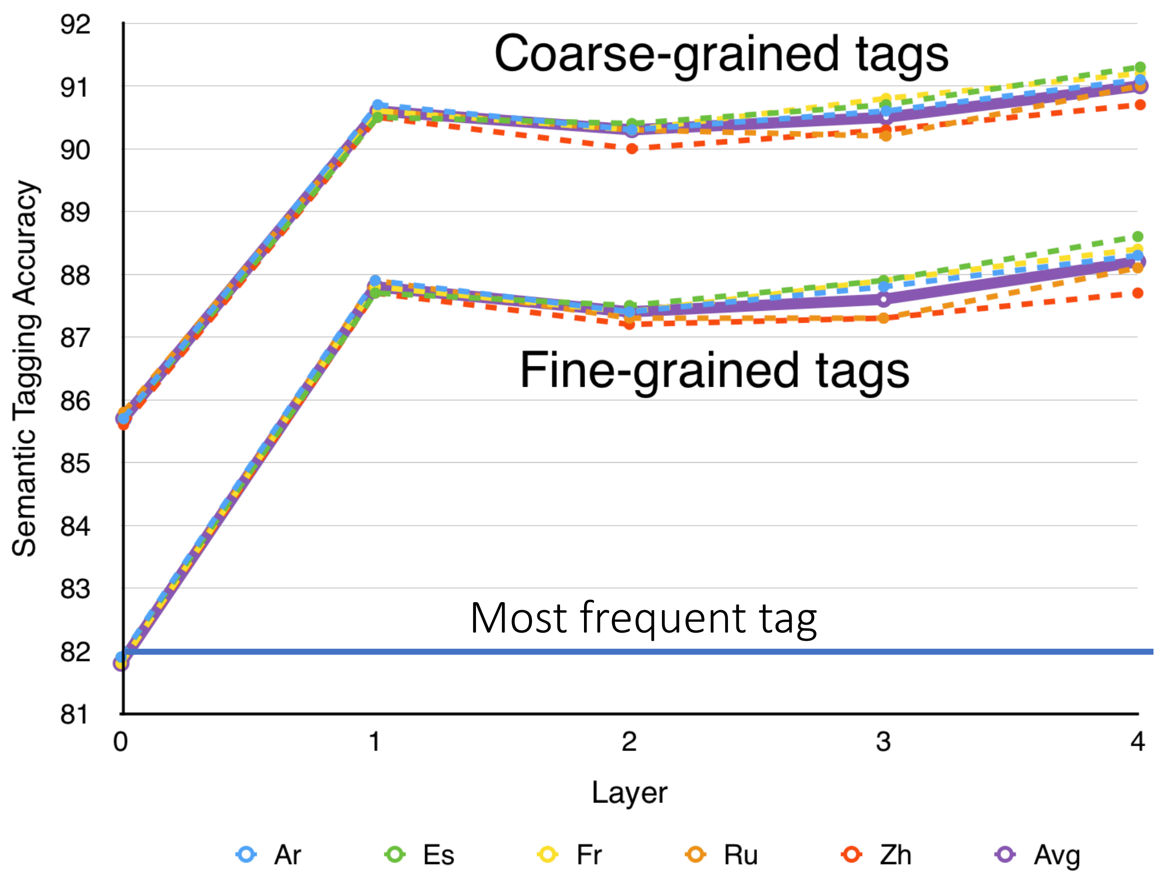
- No impact on semantic tagging



Effect of Target Language

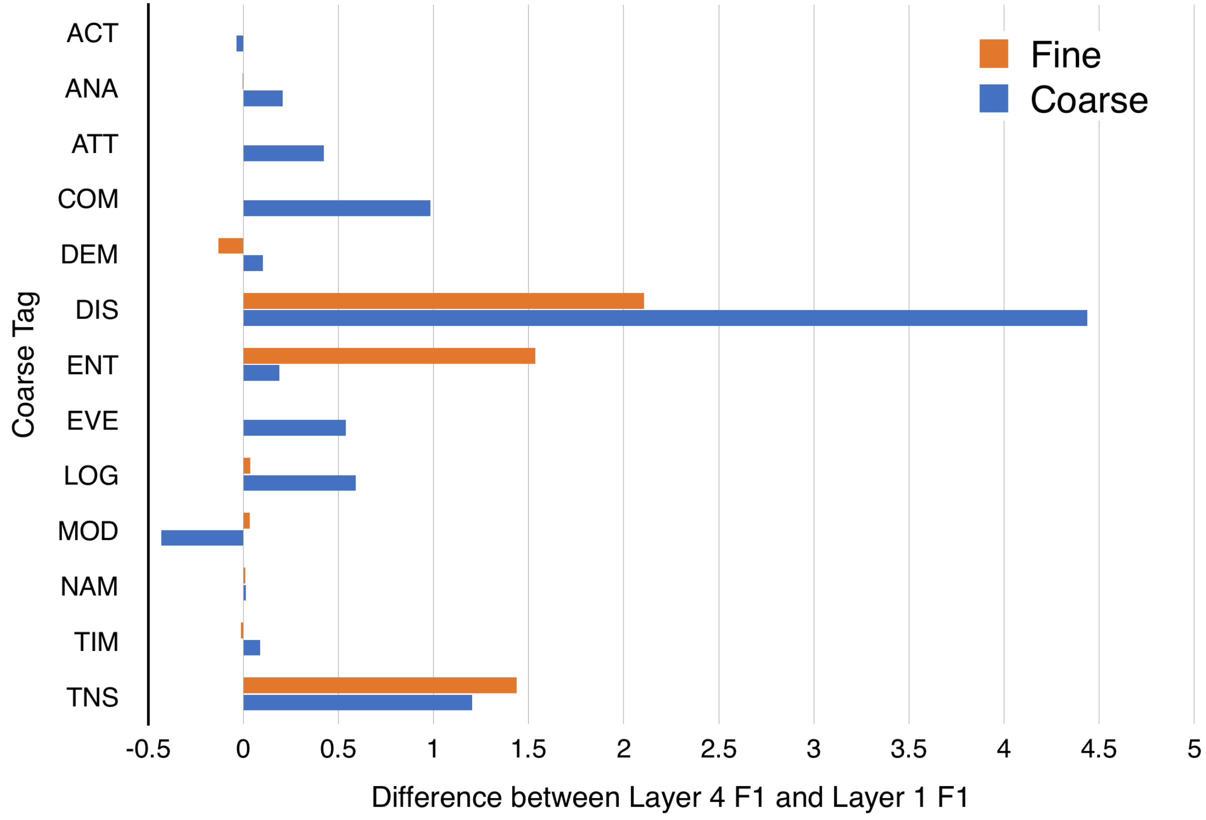
- No impact on semantic tagging
- But large impact on translation:

	BLEU
En-Ar	32.7
En-Es	49.1
En-Fr	38.5
En-Ru	34.2
En-Zh	32.1



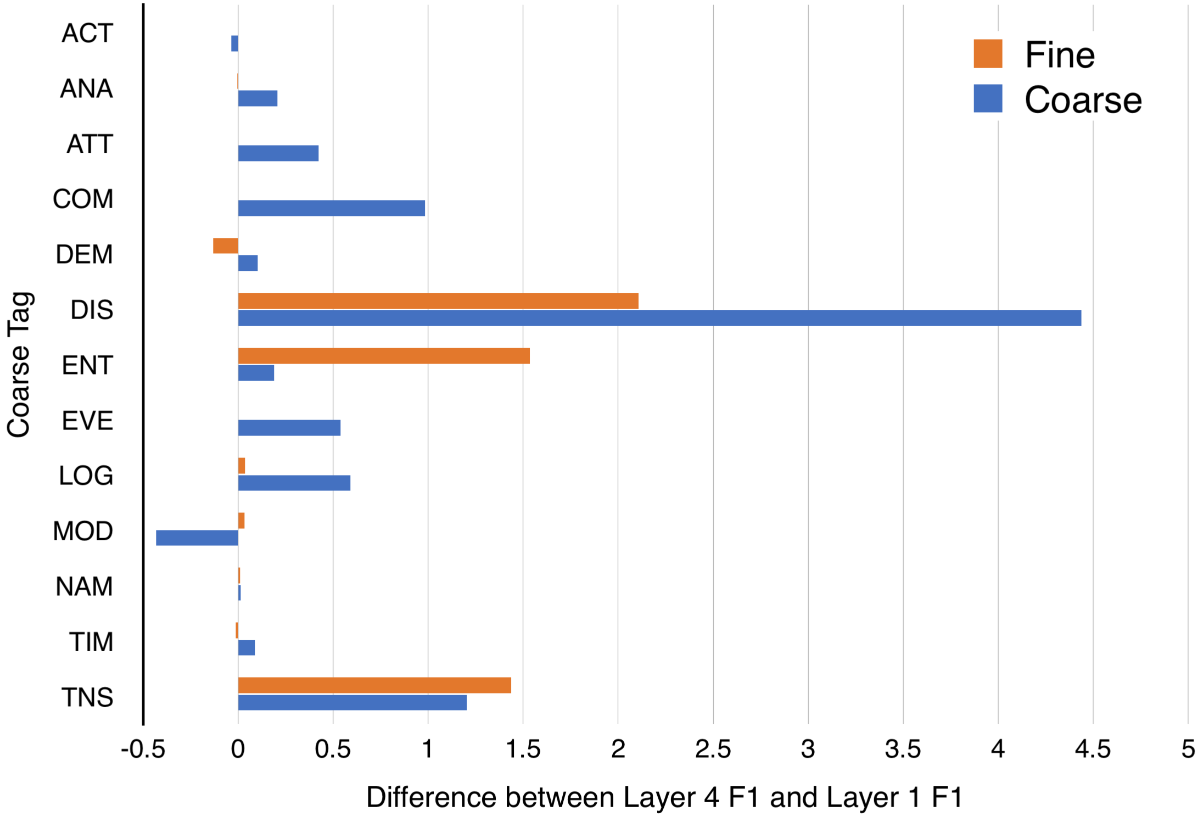
Analyzing Specific Tags

- Layer 4 vs layer 1
- Bleu: distinguishing among coarse tags
- Red: distinguishing among fine-grained tags within a coarse category



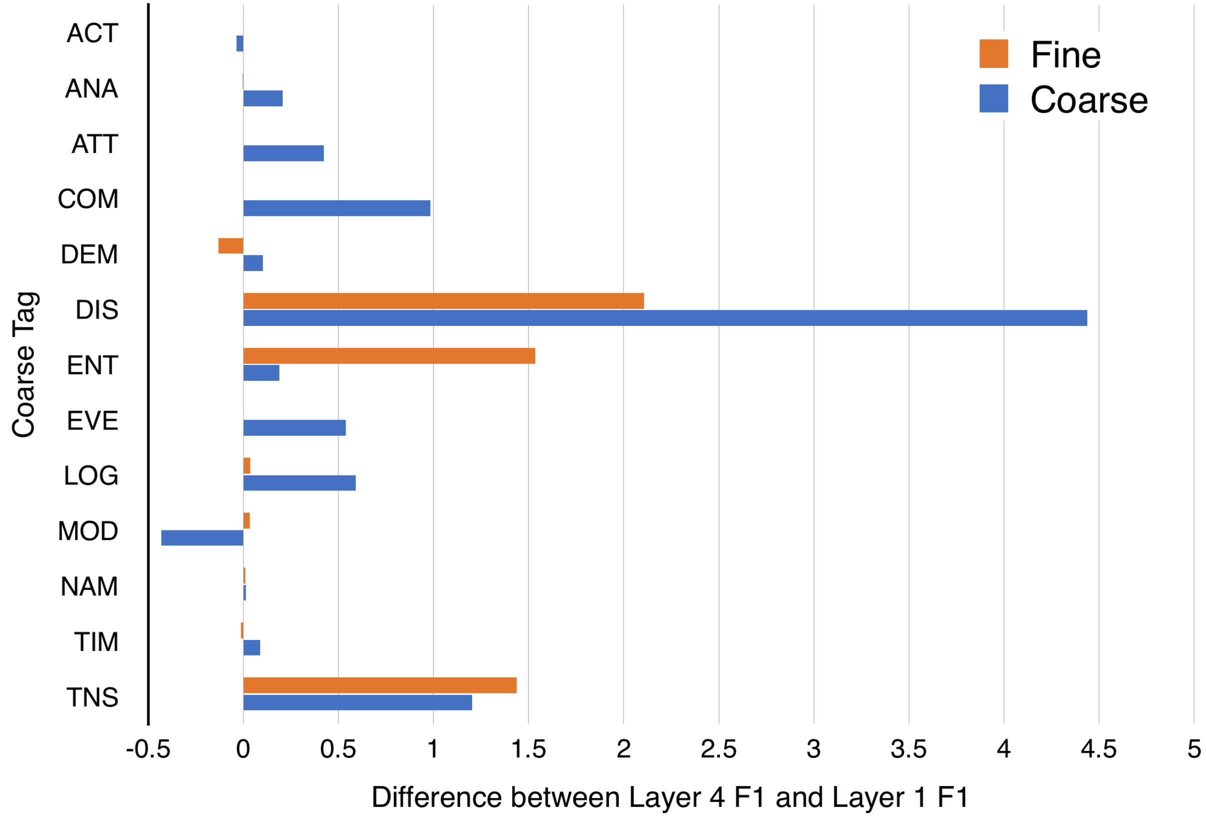
Analyzing Specific Tags

- Layer 4 > layer 1



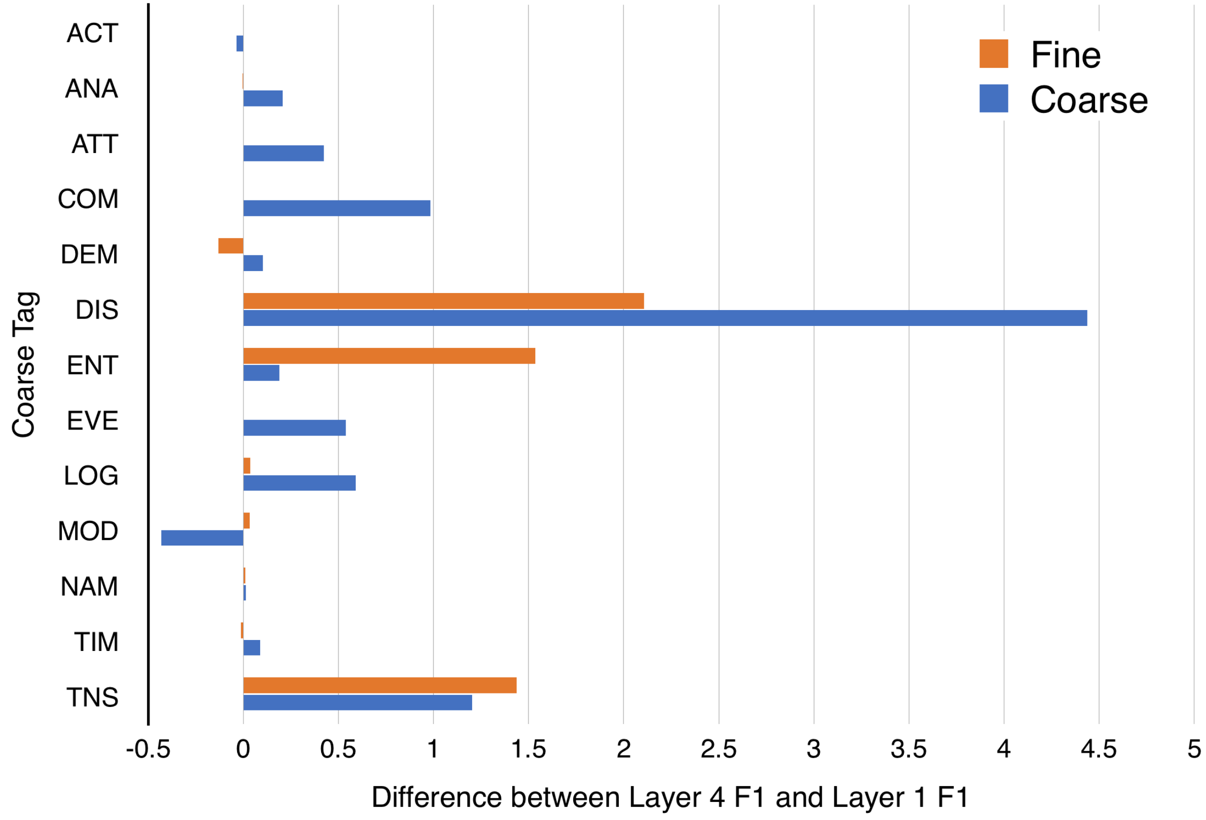
Analyzing Specific Tags

- Layer 4 > layer 1
- Especially with:
  - Discourse relations (*DIS*)
  - Properties of nouns (*ENT*)
  - Events, tenses (*EVE*, *TNS*)
  - Logic relations and quantifiers (*LOG*)
  - Comparative constructions (*COM*)



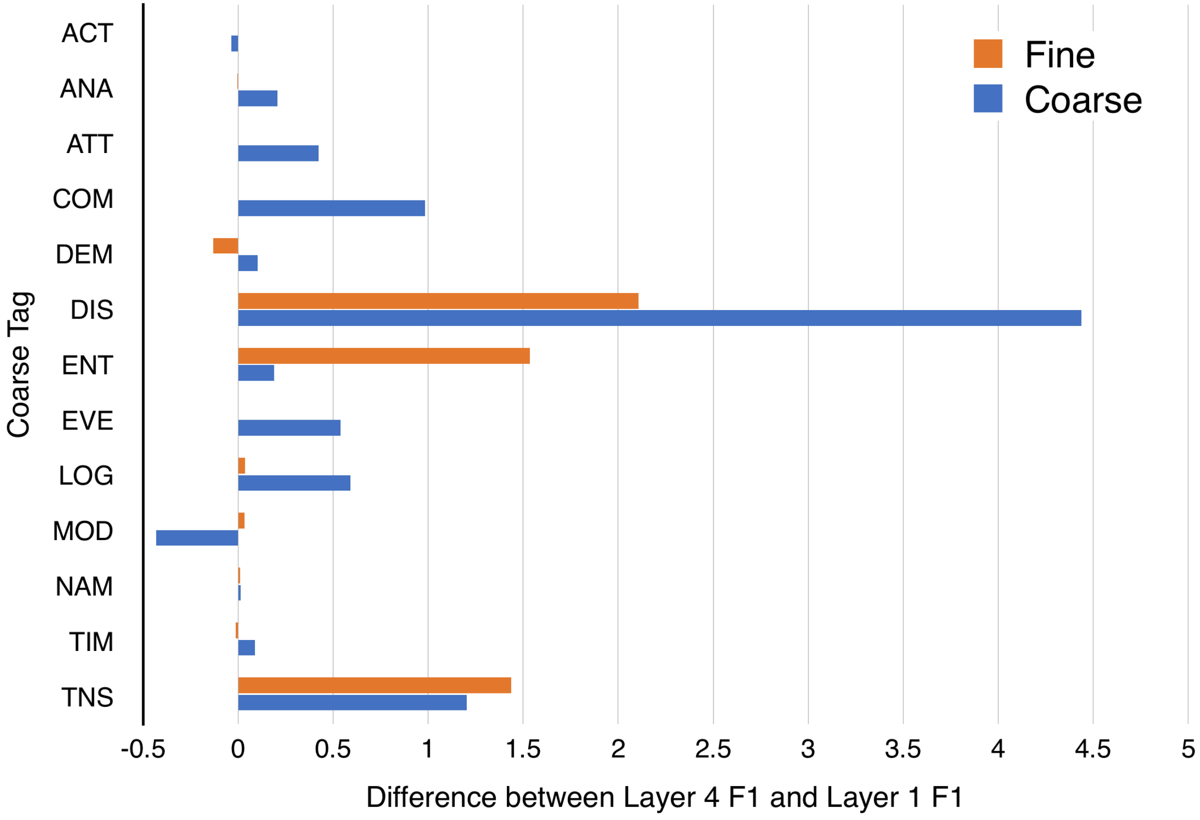
Analyzing Specific Tags

- Negative examples



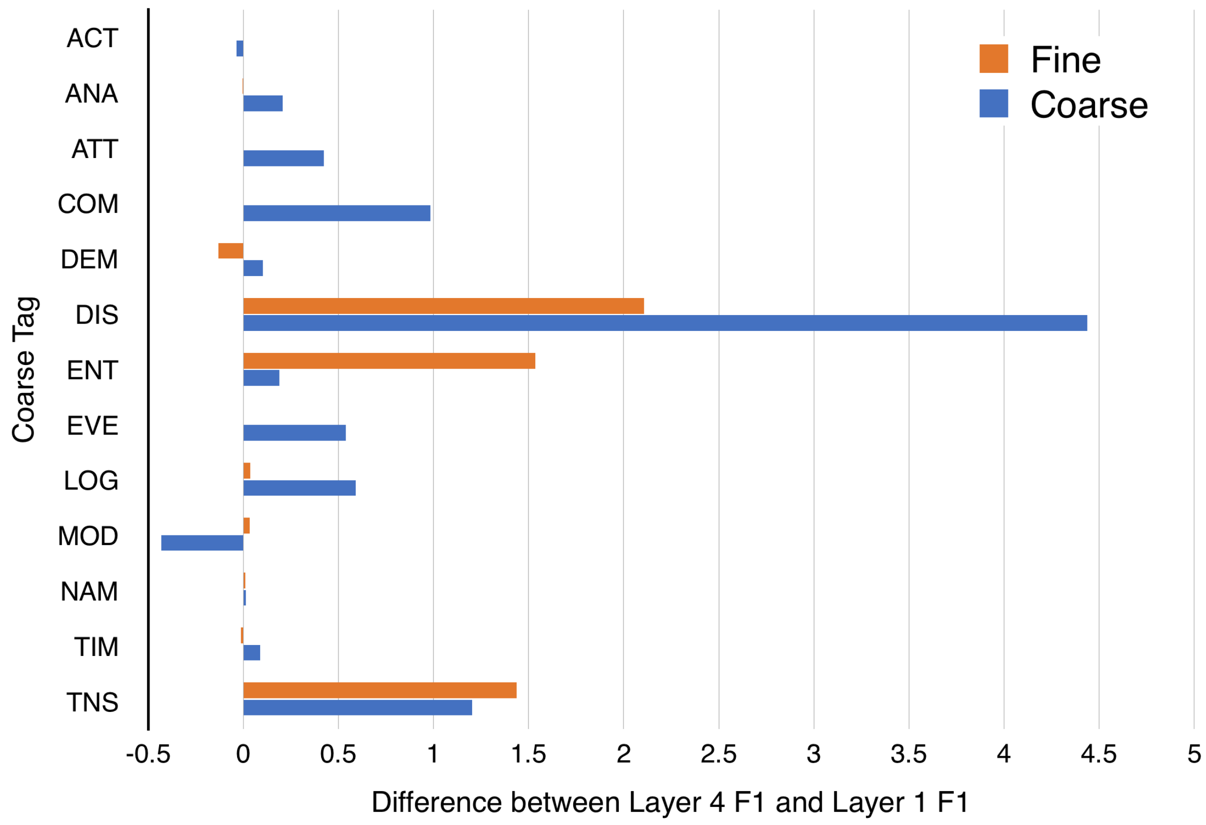
Analyzing Specific Tags

- Negative examples
- Modality (*MOD*)
  - Closed-class (“no”, “not”, “should”, ”must”, etc.)



Analyzing Specific Tags

- Negative examples
- Modality (*MOD*)
  - Closed-class (“no”, “not”, “should”, ”must”, etc.)
- Named entities (*NAM*)
  - OOVs?
  - Neural MT limitation?



Semantic tags vs. POS tags

Semantic tags vs. POS tags

	0	1	2	3	4
POS	87.9	92.0	91.7	91.8	91.9
Sem	81.8	87.8	87.4	87.6	88.2

Semantic tags vs. POS tags

	0	1	2	3	4
POS	87.9	**92.0**	91.7	91.8	91.9
Sem	81.8	87.8	87.4	87.6	**88.2**

- Higher layers improve semantic tagging but not POS tagging
- Layer 1 best for POS; layer 4 best for semantic tagging

Semantic tags vs. POS tags

		0	1	2	3	4
Uni	POS	87.9	**92.0**	91.7	91.8	91.9
	Sem	81.8	87.8	87.4	87.6	**88.2**
Bi	POS	87.9	**93.3**	92.9	93.2	92.8
	Sem	81.9	91.3	90.8	**91.9**	**91.9**

- Higher layers improve semantic tagging but not POS tagging
- Layer 1 best for POS; layer 4 best for semantic tagging
- Similar trends with bidirectional encoder

Summary

- Neural MT representations contain useful information about word form and meaning
- Lower layers focus on POS/morphology
- Higher layers focus on (lexical) semantics
- Target language does not affect semantic tagging quality

Future Work

- Other neural MT architectures
  - Word representations; multi-lingual models
- Other linguistic properties
  - Syntactic and semantic relations, complex structures
- Improving neural MT
  - Multi-task learning

- Analyzing representations in other neural models
  - End-to-end speech recognition