HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

MIT CSAIL

Analysis of NMT Systems

Yonatan Belinkov

Guest lecture

CMU CS 11-731: Machine Translation and Seq2seq Models

10/4/2018

Outline

- Non-neural statistical MT vs neural MT
  - Previous phrase-based MT
  - Opaqueness of NMT
  - Why analyze?
- Challenge sets
- Predicting linguistic properties
- Visualization
- Open questions

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*

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

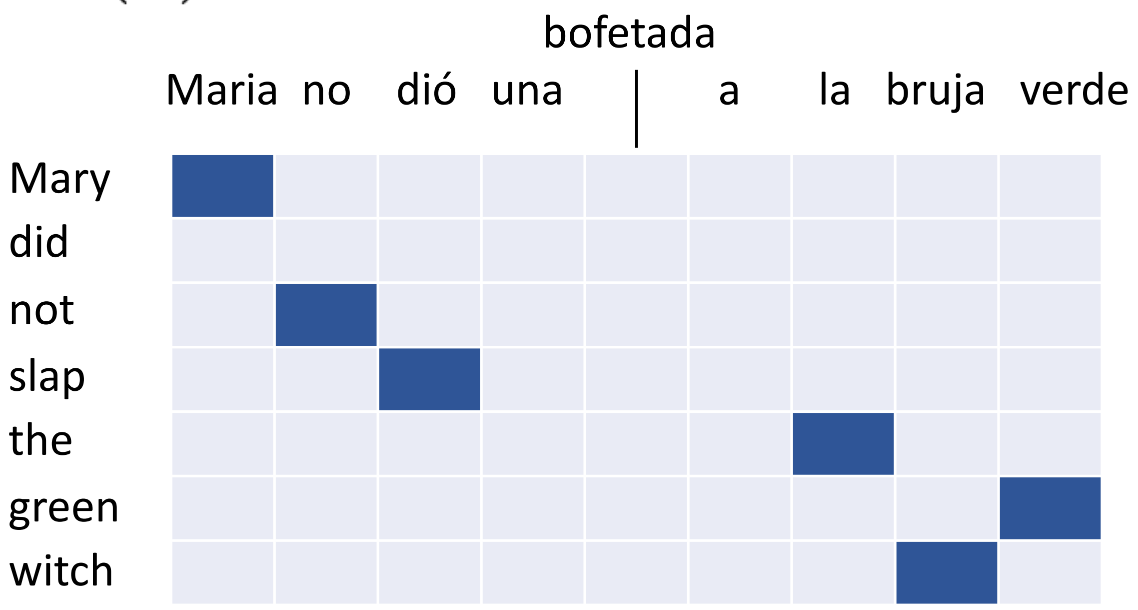
- $P(F|E)$ – Translation model
- $P(E)$ – Language model

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

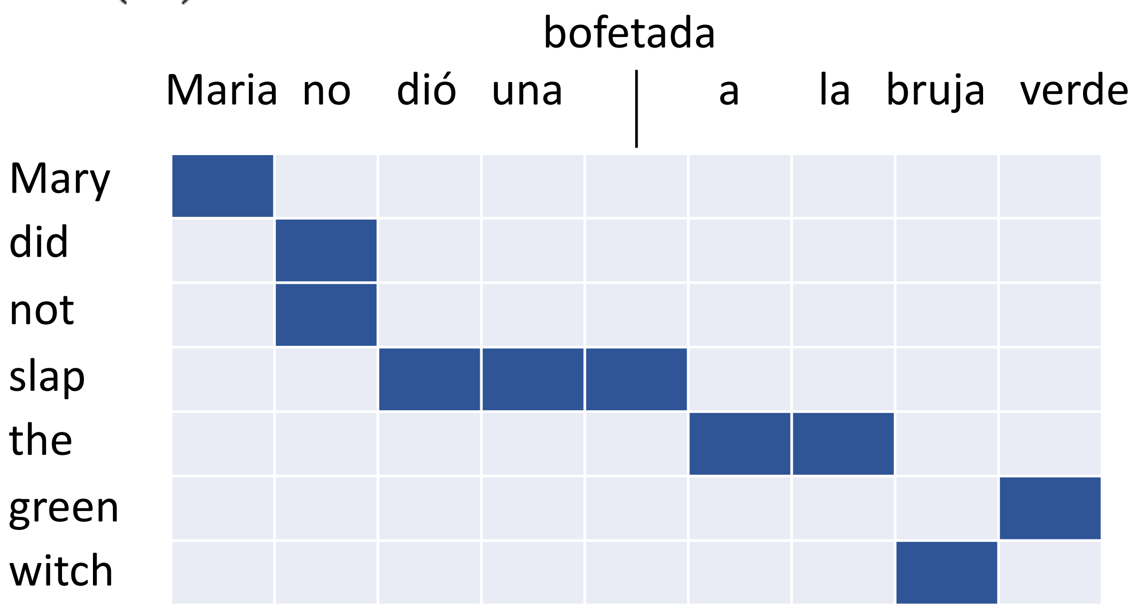
Statistical Machine Translation

- 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

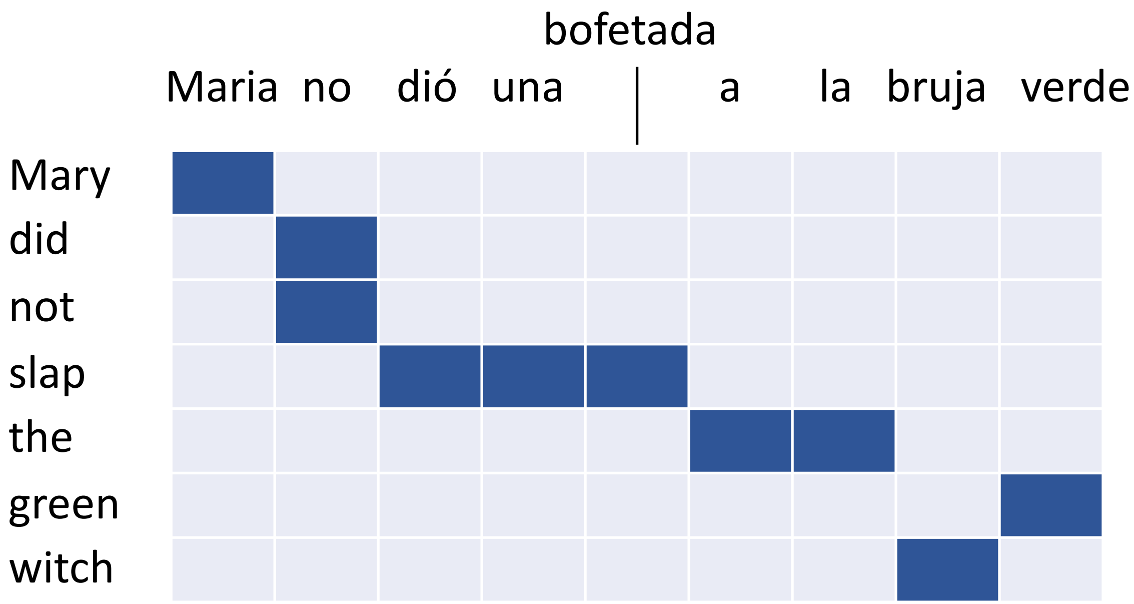
- Phrase-based MT



From: Jurafsky & Martin 2009

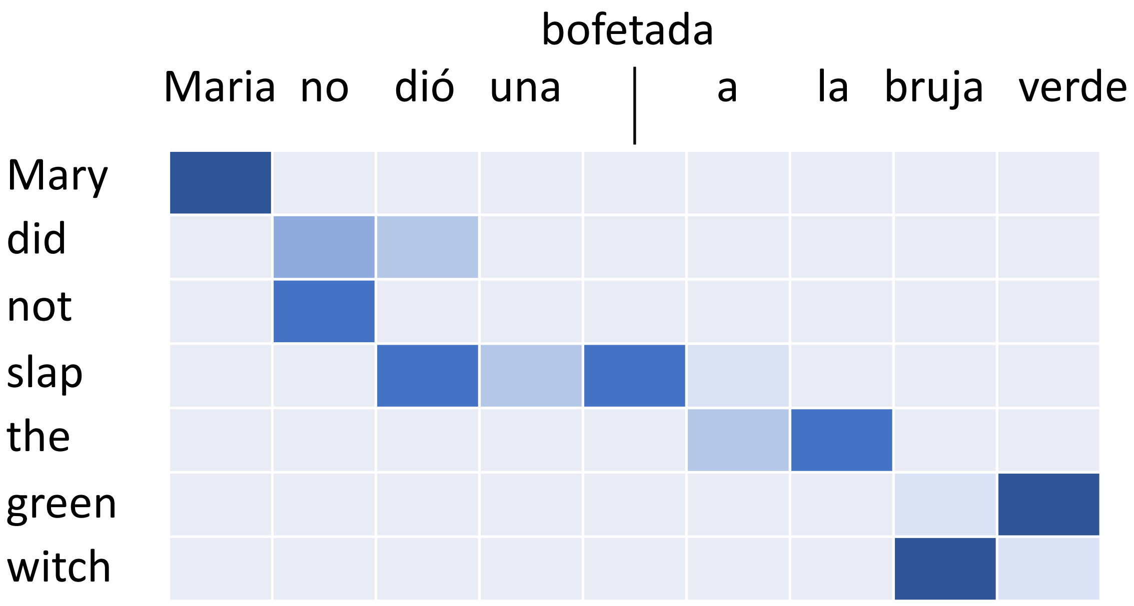
Attention as soft alignment

Phrase-based MT

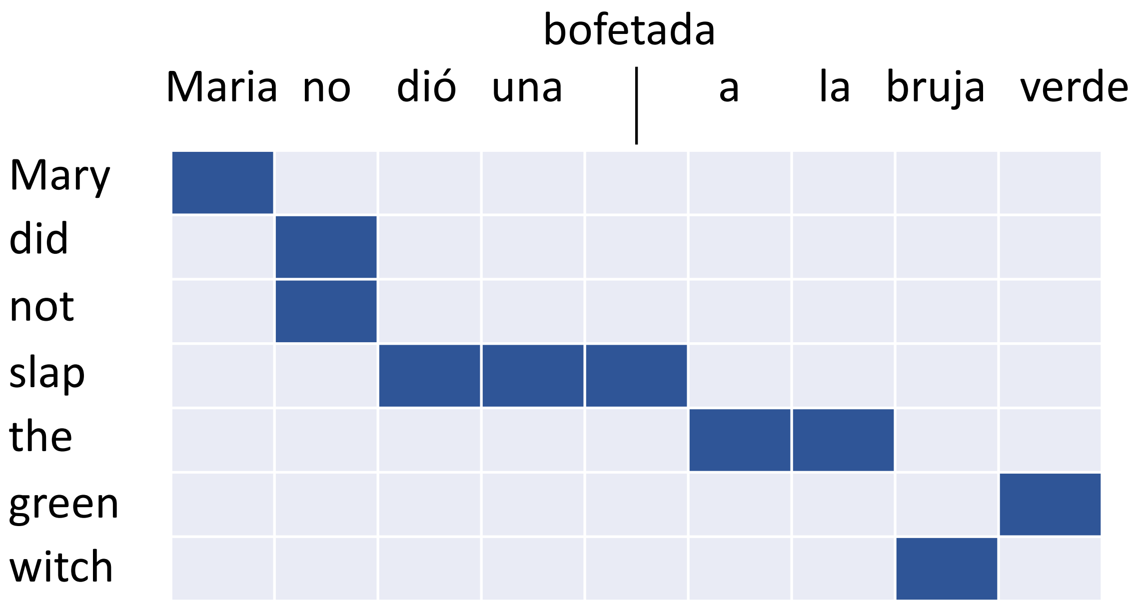


Attention as soft alignment

Neural MT



Phrase-based MT

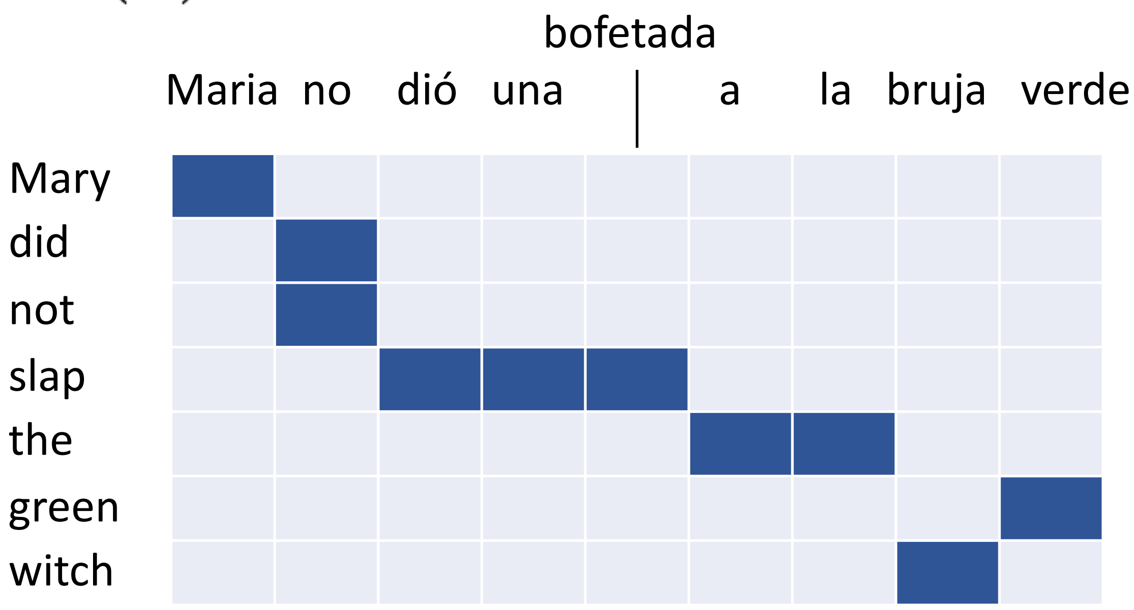


Statistical Machine Translation

- 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



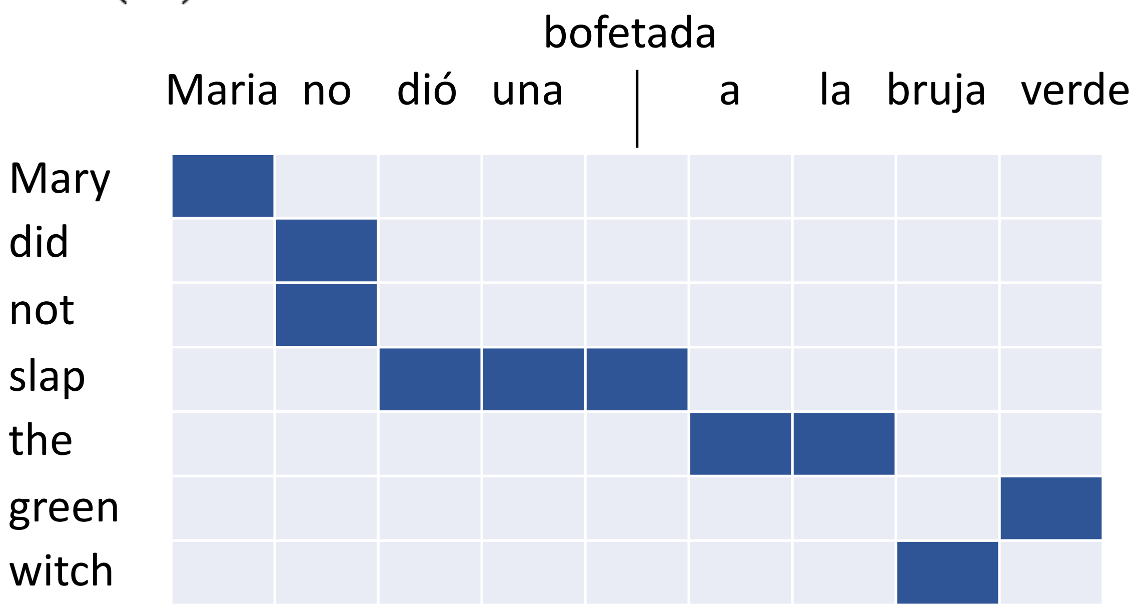
From: Jurafsky & Martin 2009

Statistical Machine Translation

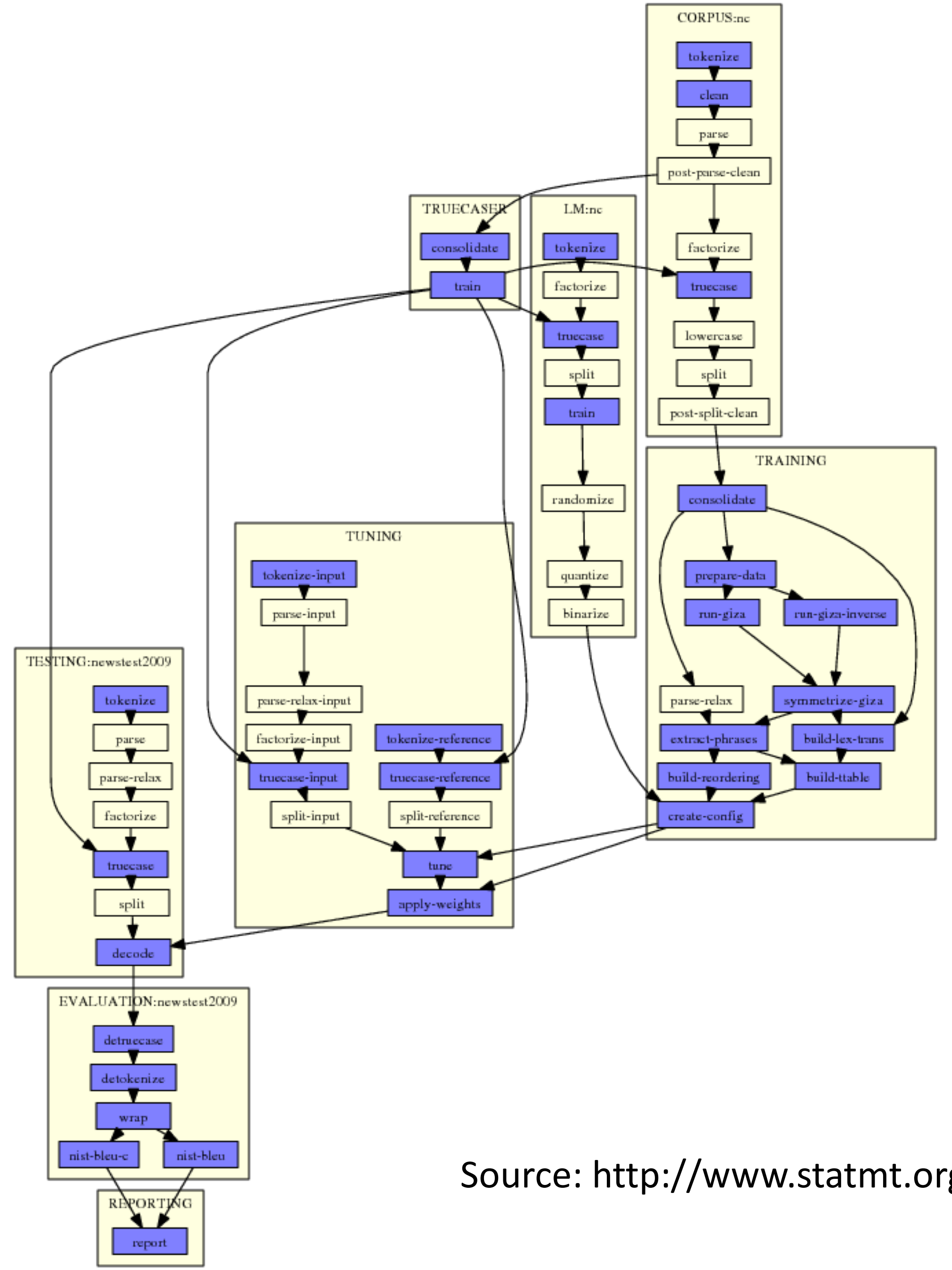
- Translate a source sentence *F* into a target sentence *E*

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

- $P(F|E)$ – Translation model
- $P(E)$ – Language model
- Additional components
  - Word order, syntax, morphology
  - Etc.

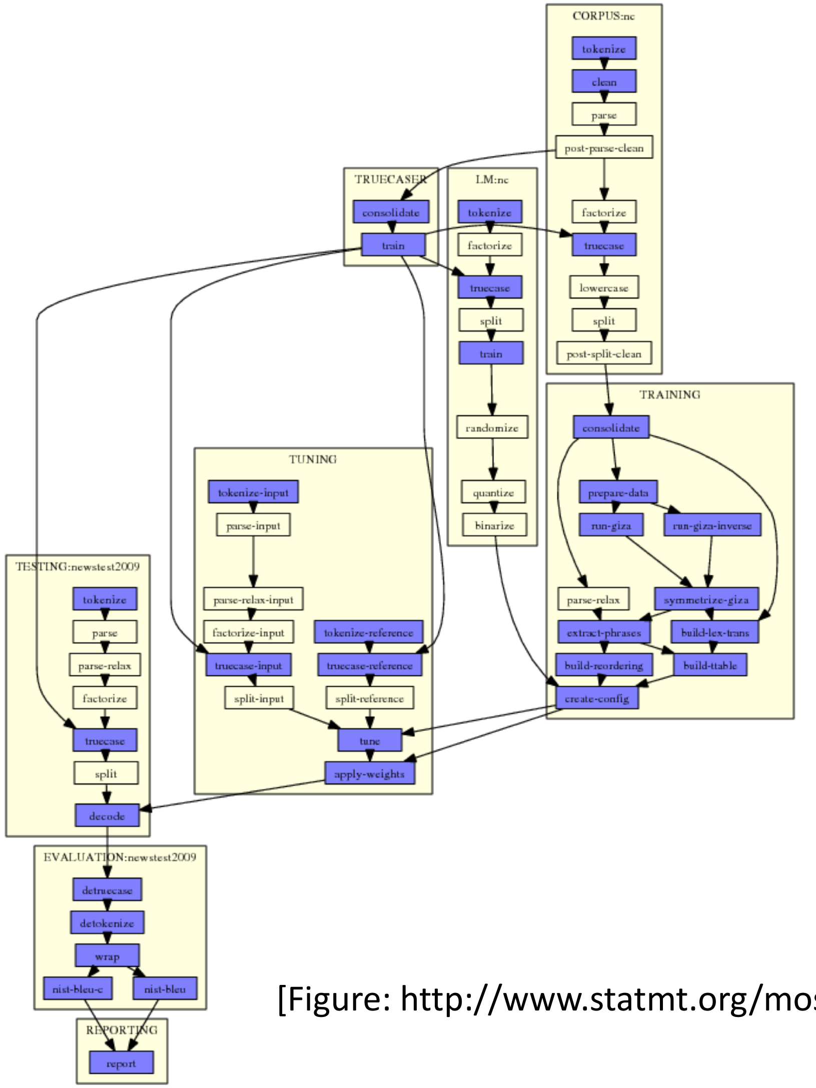


From: Jurafsky & Martin 2009

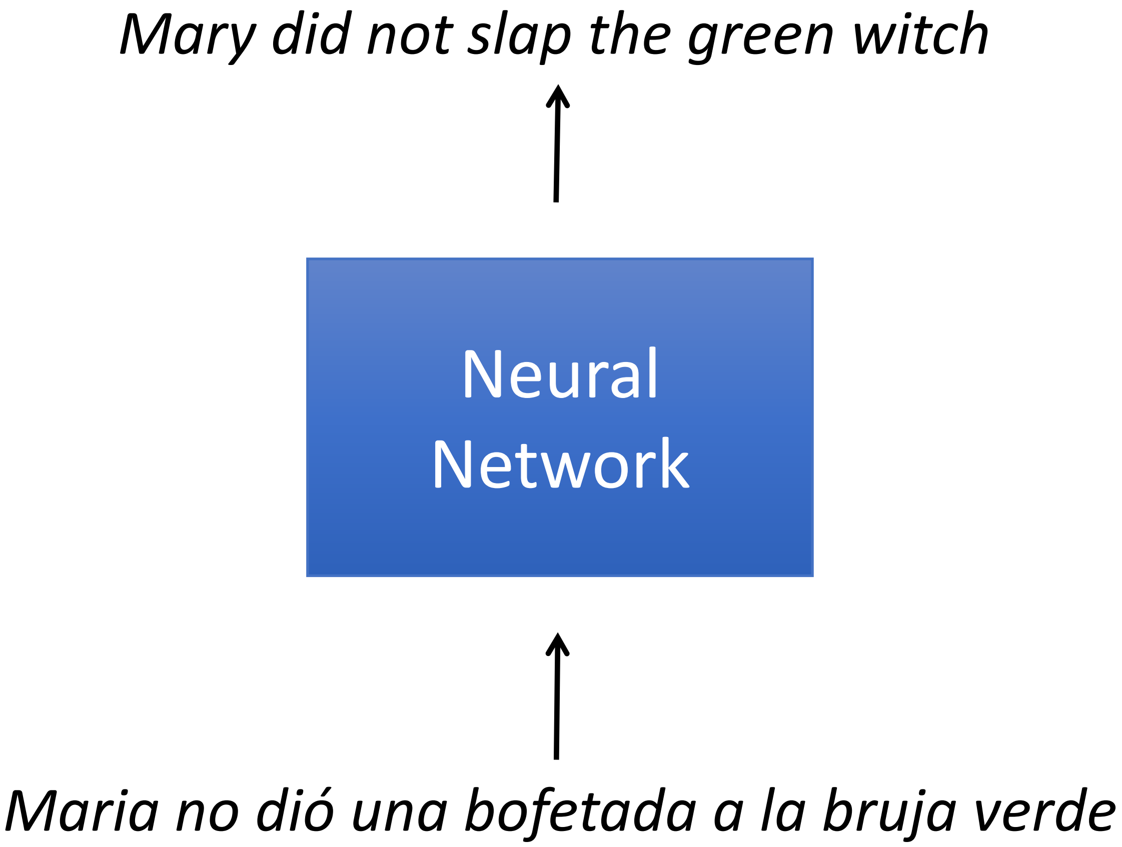


Source: http://www.statmt.org/moses

End-to-End Learning: Machine Translation

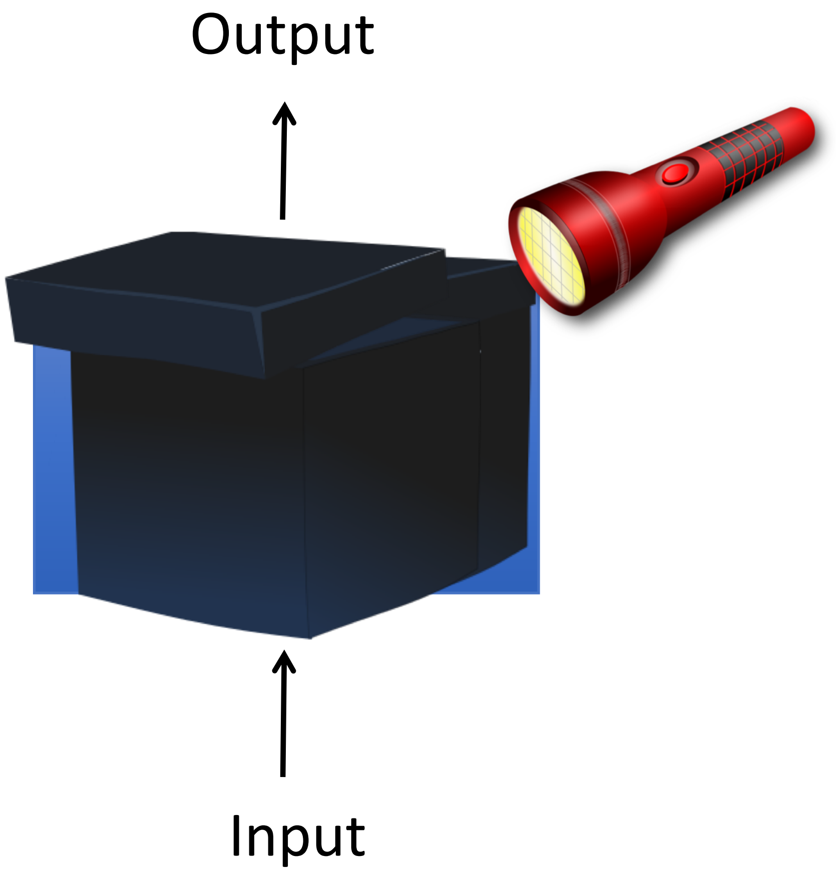


[Figure: http://www.statmt.org/moses]



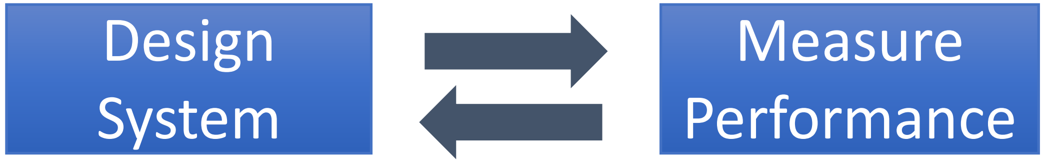
End-to-End Learning

The Black Box



Why should we care?

- Current deep learning research
  - Much trial-and-error
  - Often a shot in the dark
  - ➢ Better understanding → better systems



- Accountability, trust, and bias in machine learning
  - “Right to explanation”, EU Regulation
  - Life-threatening situations: healthcare, autonomous cars
  - ➢ Better understanding → more accountable systems

How can we move beyond BLEU?

Challenge Sets

- Carefully constructed examples
- Test specific linguistic properties
  - More informative than automatic metrics like BLEU scores
- Old tradition in NLP and MT (King & Falkedal 1990; Isahara 1995; Koh+ 2001)
  - Also known as “test suites”

- Now making a comeback in MT (and other NLP tasks)

Challenge Sets

	Phenomena	Languages	Size	Construction
Rios Gonzales+ 2017	WSD	German→English/French	13900	Semi-auto
Burlot & Ivon 2017	Morphology	English→Czech/Latvian	18500	Automatic
Sennrich 2017	Agreement, polarity, verb-particles, transliteration	English→German	97000	Automatic
Bawden+ 2018	Discourse	English→French	400	Manual
Isabelle+ 2017	Morpho-syntax, syntax, lexicon	English→French	506	Manual
Isabelle & Kuhn 2018	Morpho-syntax, syntax, lexicon	French→English	108	Manual
Burchardt+ 2018	Diverse (120)	English↔German	10000	Manual

Example: Manual Evaluation

- Isabelle et al. (2017)
  - 108 sentences to capture divergences between English and French
  - Get translations from phase-based and NMT systems
  - Ask human raters to answer questions about machine translations
  - Example:

Src	The repeated calls from his mother **should** have alerted us.
Ref	Les appels répétés de sa mère **auraient** dû nous alerter.
Sys	Les appels répétés de sa mère devraient nous avoir alertés.
Is the subject-verb agreement correct (y/n)? **Yes**	

Example: Manual Evaluation

- Isabelle et al. (2017)

Divergence type	PBMT-1	PBMT-2	NMT	Google NMT
Morpho-syntactic	16%	16%	72%	65%
Lexico-syntactic	42%	46%	52%	62%
Syntactic	33%	33%	40%	75%
Overall	31%	32%	53%	68%
WMT BLEU	34.2	36.5	36.9	—

  - NMT better overall, but fails to capture many properties
  - Example problems: agreement logic, noun compounds, control verbs, …

Example: Automatic Evaluation

- Sennrich (2017)
  - Create **contrastive translation pairs** from existing parallel corpora
  - Apply heuristics to create wrong translations
  - Compare likelihood of wrong and correct translations

category	English	German (correct)	German (contrastive)
NP agreement	[...] **of the** American **Congress**	[...] **des** amerikanischen **Kongresses**	* [...] **der** amerikanischen **Kongresses**
subject-verb agr.	[...] that the **plan will** be approved	[...], dass der **Plan** verabschiedet **wird**	* [...], dass der **Plan** verabschiedet **werden**
separable verb particle	he is **resting**	er **ruht** sich **aus**	* er **ruht** sich **an**
polarity	the timing [...] is **uncertain**	das Timing [...] ist **unsicher**	das Timing [..] ist **sicher**
transliteration	Mr. **Ensign's** office	Senator **Ensigns** Büro	Senator **Enisgns** Büro

Example: Automatic Evaluation

- Sennrich (2017)

system	agreement		verb particle	polarity (negation)		transliteration
	noun phrase	subject-verb	verb particle	insertion	deletion	transliteration
(category and size→)	21813	35105	2450	22760	4043	3490
BPE-to-BPE	**95.6**	**93.4**	**91.1**	97.9	**91.5**	96.1
BPE-to-char	93.9	91.2	88.0	**98.5**	88.4	**98.6**
char-to-char	93.9	91.5	86.7	**98.5**	89.3	**98.3**
(Sennrich et al., 2016a)	98.7	96.6	96.1	98.7	92.7	96.4
human	99.4	99.8	99.8	99.9	98.5	99.0

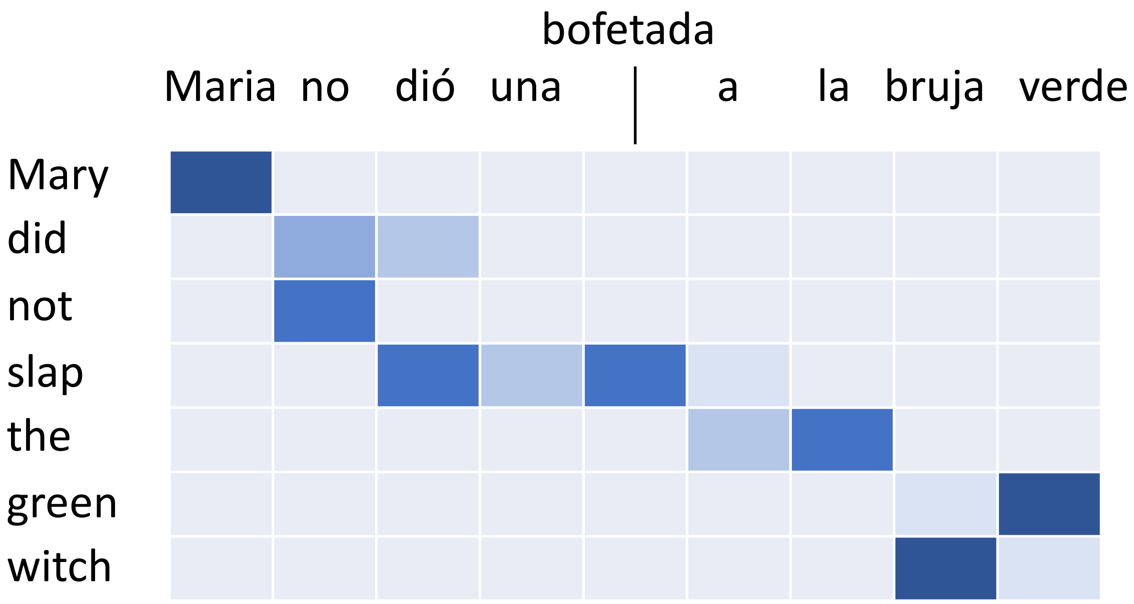
  - Char decoders better on transliteration, but worse on verb particles and agreement (especially in distant words)
  - Tradeoff between generalization to unseen words and sentence-level grammaticality

More Contrastive Translation Pairs

- Morphology (Burlot & Ivon 2017)
  - Apply morphological transformations with analyzers and generators
  - Filtering less likely sentences with a language model.
- Discourse (Bawden+ 2018)
  - Coreference and coherence
  - Manually modify existing examples
- Word sense disambiguation (Rios Gonzales+ 2017)
  - Search for ambiguous German words with distinct translations
  - Manually verify examples

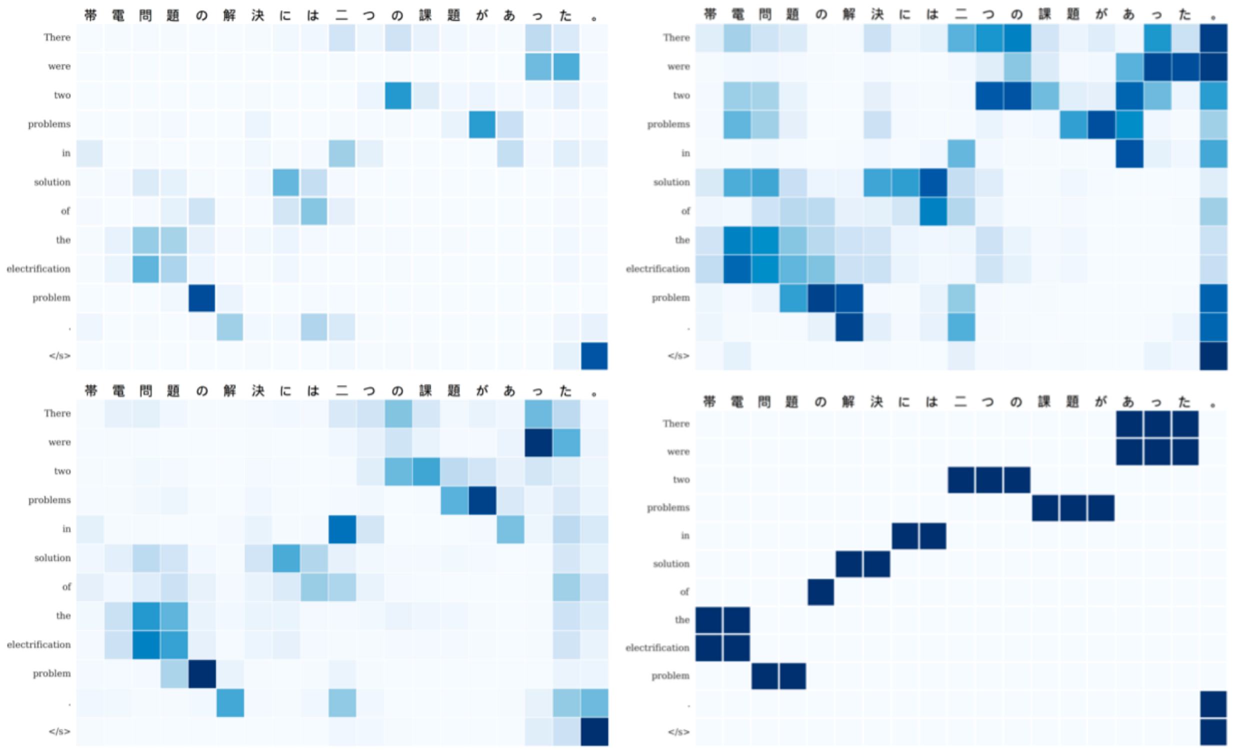
Visualization

- Visualizing attention weights



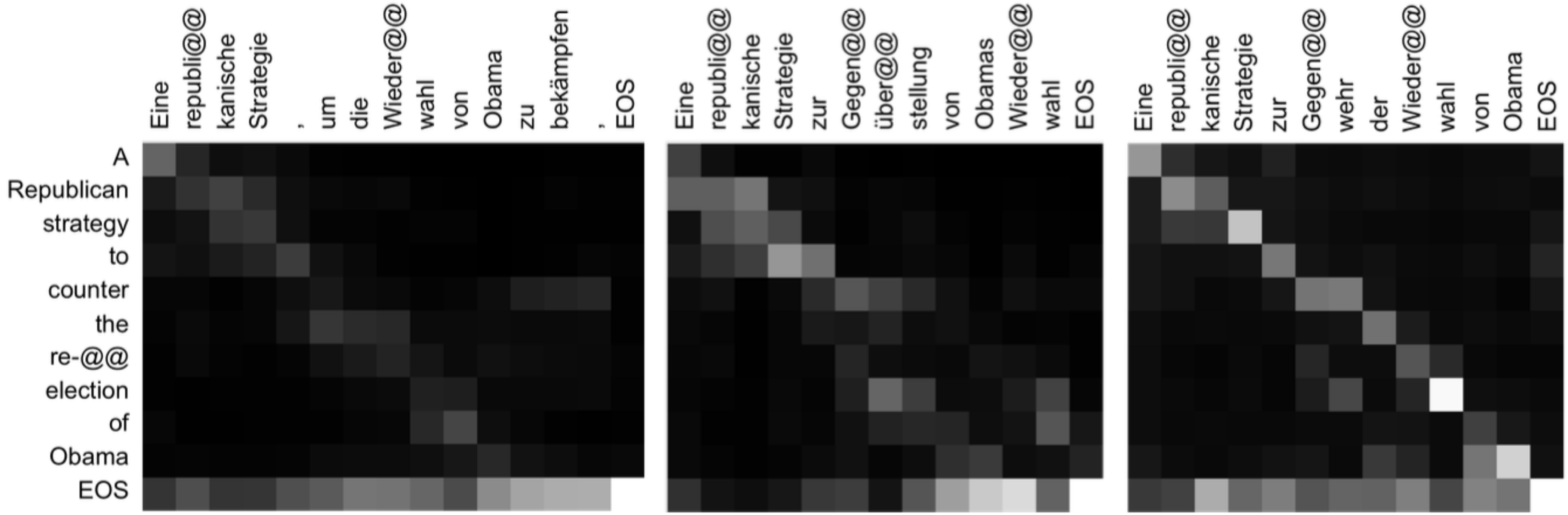
Improved attention mechanisms

- “Structured Attention Networks” (Kim+ 2017)



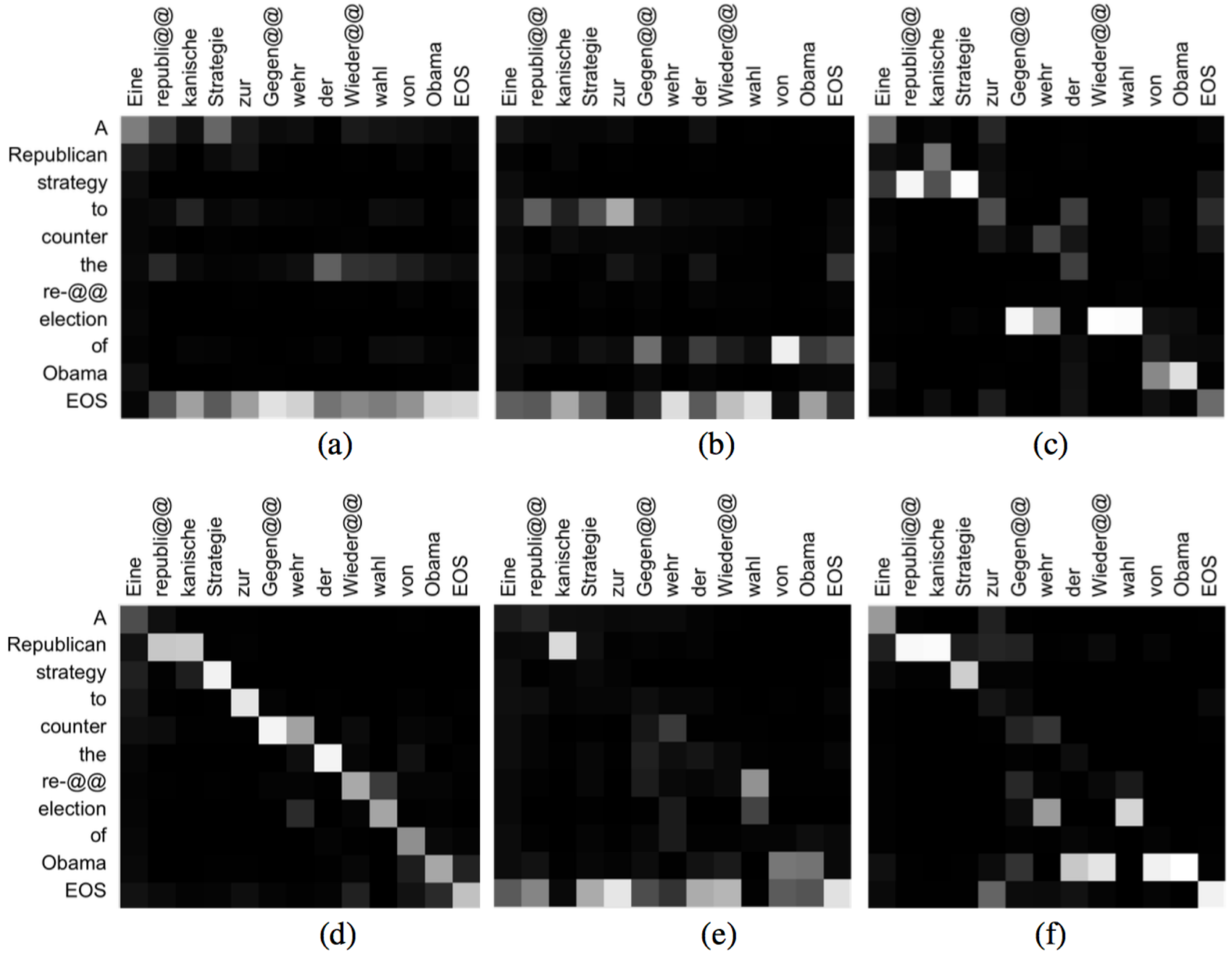
Improved attention mechanisms

- “Fine-Grained Attention for NMT” (Choi+ 2018)



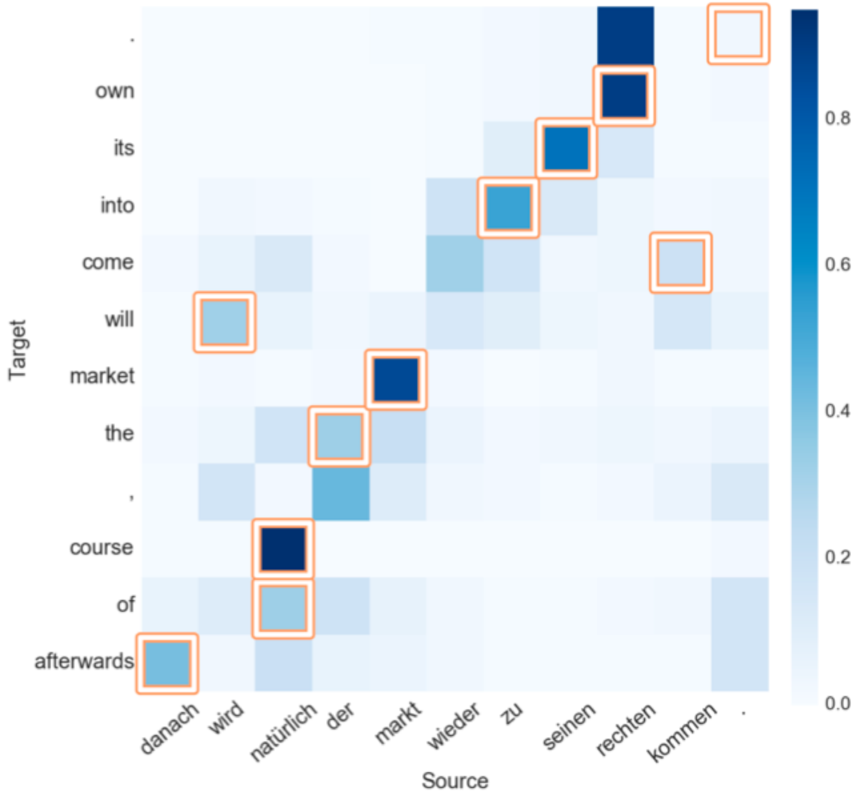
Improved attention mechanisms

- "Fine-Grained Attention for NMT" (Choi+ 2018)
- Visualizations of specific dimensions



What do these attentions do?

- "What does Attention in NMT pay attention to?" (Ghader & Monz 2017)
  - Comparing attention and alignment
  - Also looked at correlations between attention and word prediction loss
  - And which POS tags are most attended to



Visualization

- "Visualizing and Understanding NMT" (Ding+ 2017)
  - Adapt layer-wise relevance propagation (LRP) to the NMT case
  - Calculate association between hidden states and input/output

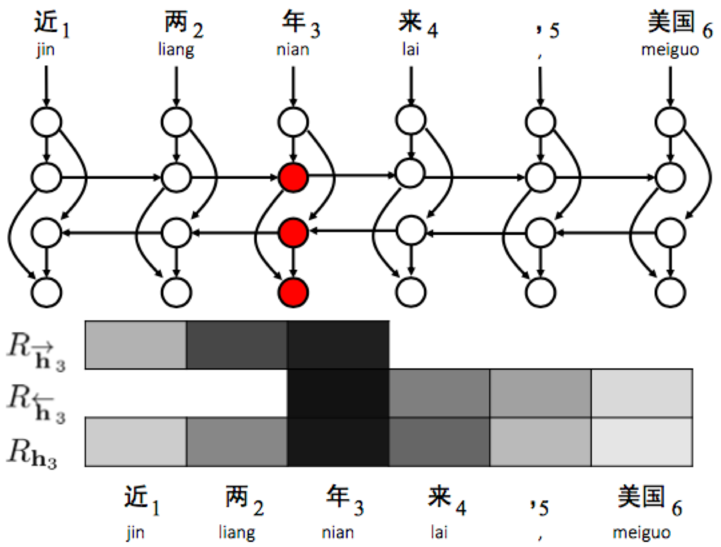


Figure 4: Visualizing source hidden states for a source content word "nian" (*years*).

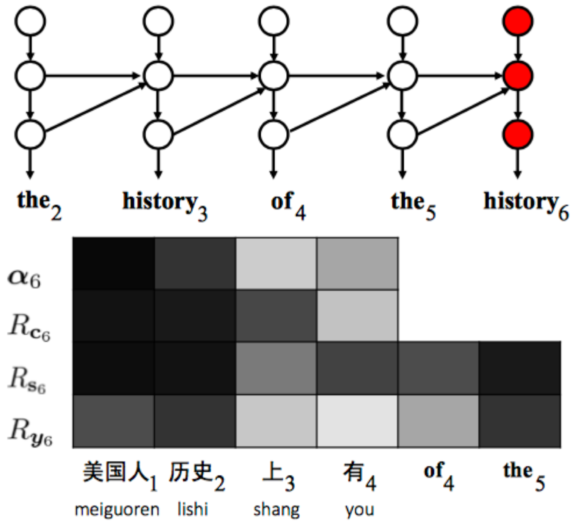


Figure 8: Analyzing translation error: word repetition. The target word "history" occurs twice in the translation incorrectly.

Looking inside NMT

- Challenge sets give us overall performance, but not
  - what is happening inside the model
  - where linguistic information is stored
- Visualizations may show input/output/state correspondences, but
  - they are limited to specific examples
  - they are not connected to linguistic properties
- Can we investigate what linguistic information is captured in NMT?

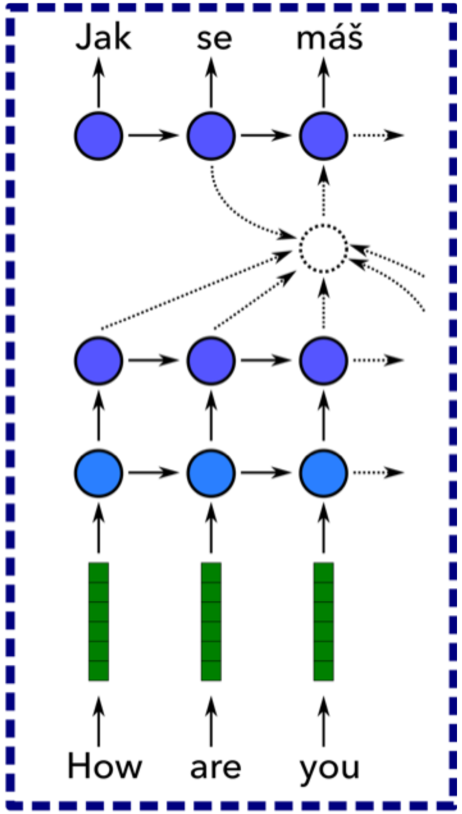
Research Questions

- What is encoded in the intermediate representations?

- What is the effect of NMT design choices on learning language properties (morphology, syntax, semantics)?
  - Network depth
  - Encoder vs. decoder
  - Word representation
  - Effect of target language
  - …

Methodology

1. Train a neural MT system



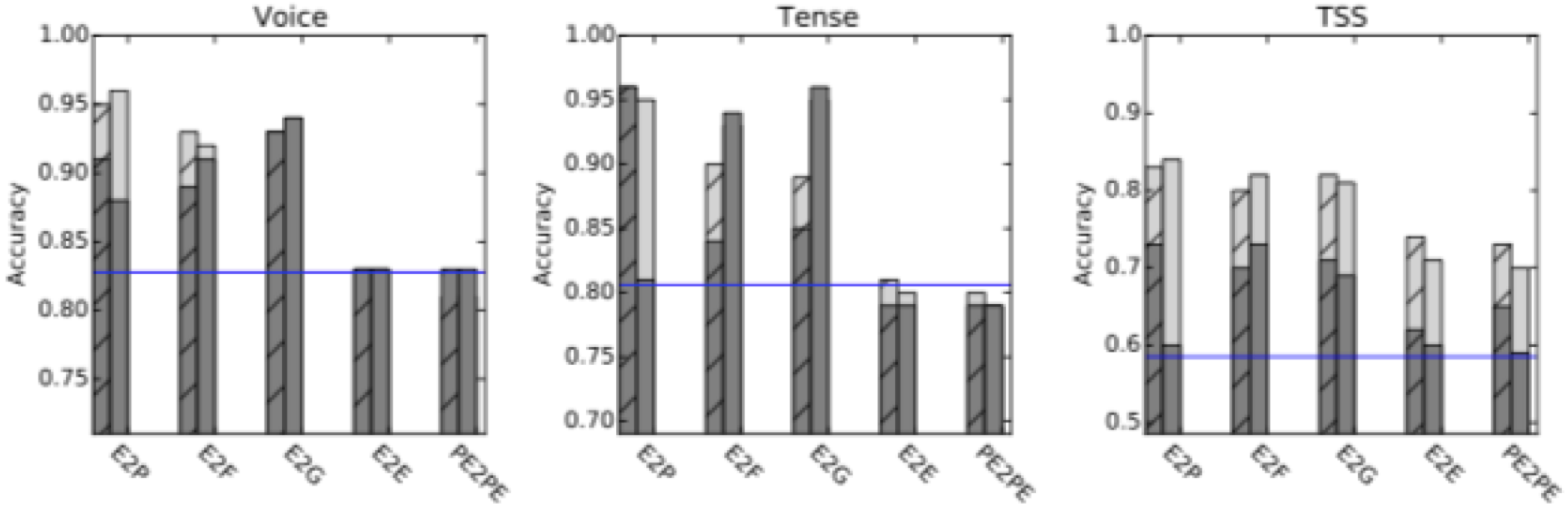
Syntax

- "Does String-Based Neural MT Learn Source Syntax" (Shi+ 2016)

- English→French, English→German
- Encoder-side representations
- Syntactic properties
  - Word-level: POS tags, smallest phrase constituent
  - Sentence-level: top-level syntactic sequence, voice, tense

Syntax

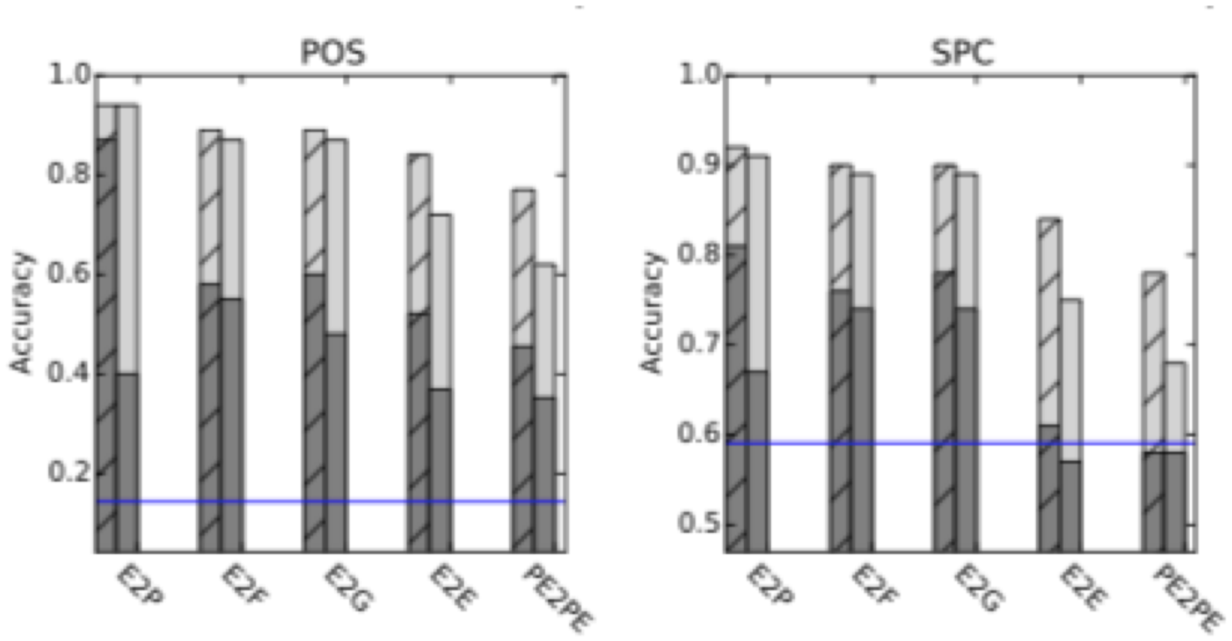
- Sentence-level tasks



  - Auto-encoders learn poor representations (at majority class)
  - NMT encoders learn much better representations

Syntax

- Word-level tasks



  - All above majority baseline, but auto-encoder representations are worse
  - First layer representations are slightly better

Syntax

- Generate full (linearized) trees from encodings

Model	Perplexity on Train	Perplexity on WSJ 22	Labeled F1 on WSJ23	# EVALB-trees (out of 2416)	Average TED per sentence	# Well-formed trees (out of 2416)
PE2PE2P	1.83	1.92	46.64	818	34.43	2416
E2E2P	1.69	1.77	59.35	796	31.25	2416
E2G2P	1.39	1.41	80.34	974	17.11	2340
E2F2P	1.36	1.38	79.27	1093	17.77	2415
E2P	1.11	1.18	89.61	2362	11.50	2415

- NMT encodings are much better (lower TED) than auto-encoders

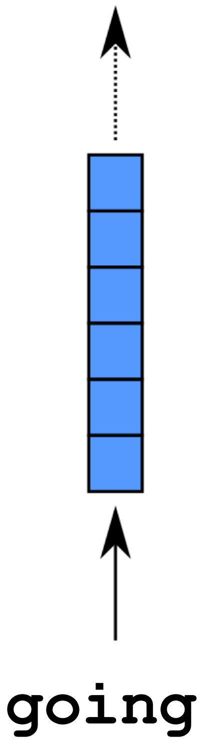
Morphology

- ”What do NMT Models Learn about Morphology?” (Belinkov+ 2017)
-

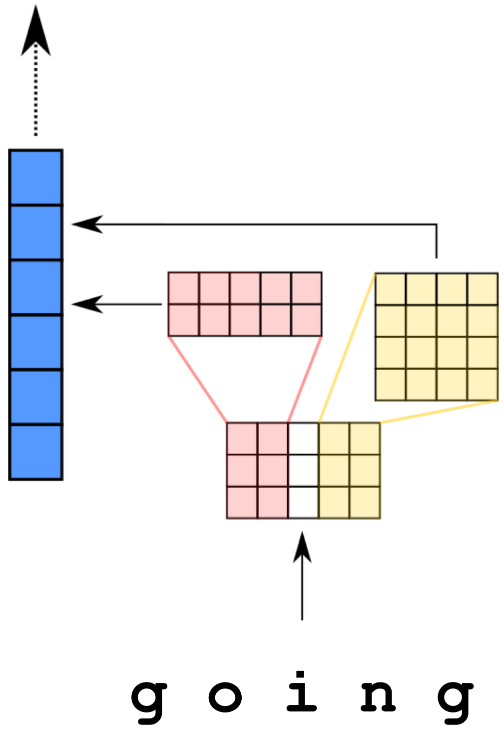
  Tasks
    - Part-of-speech tagging (“runs” = verb)
    - Morphological tagging (“runs” = verb, present tense, 3rd person, singular)
- Languages
    - Arabic-, German-, French-, and Czech-English
    - Arabic-German (rich but different)
    - Arabic-Hebrew (rich and similar)

Morphology

Word embedding



Character CNN



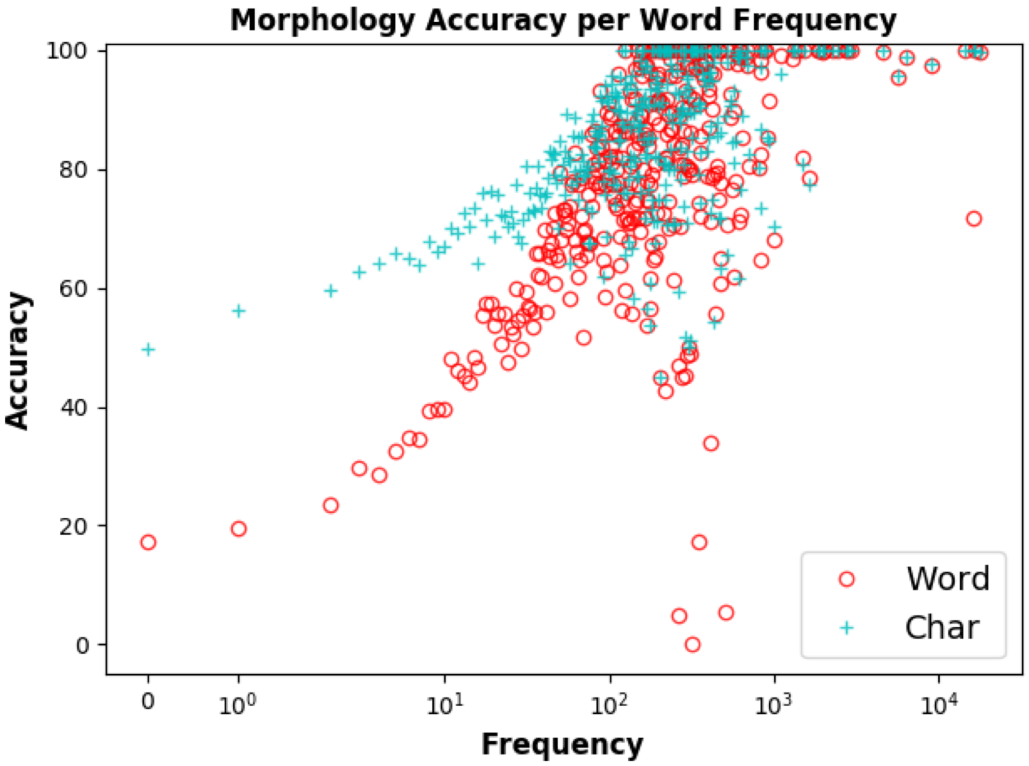
Morphology

	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 part-of-speech (and morphology)
  - Improve translation quality

Morphology

- Impact of word frequency



Morphology

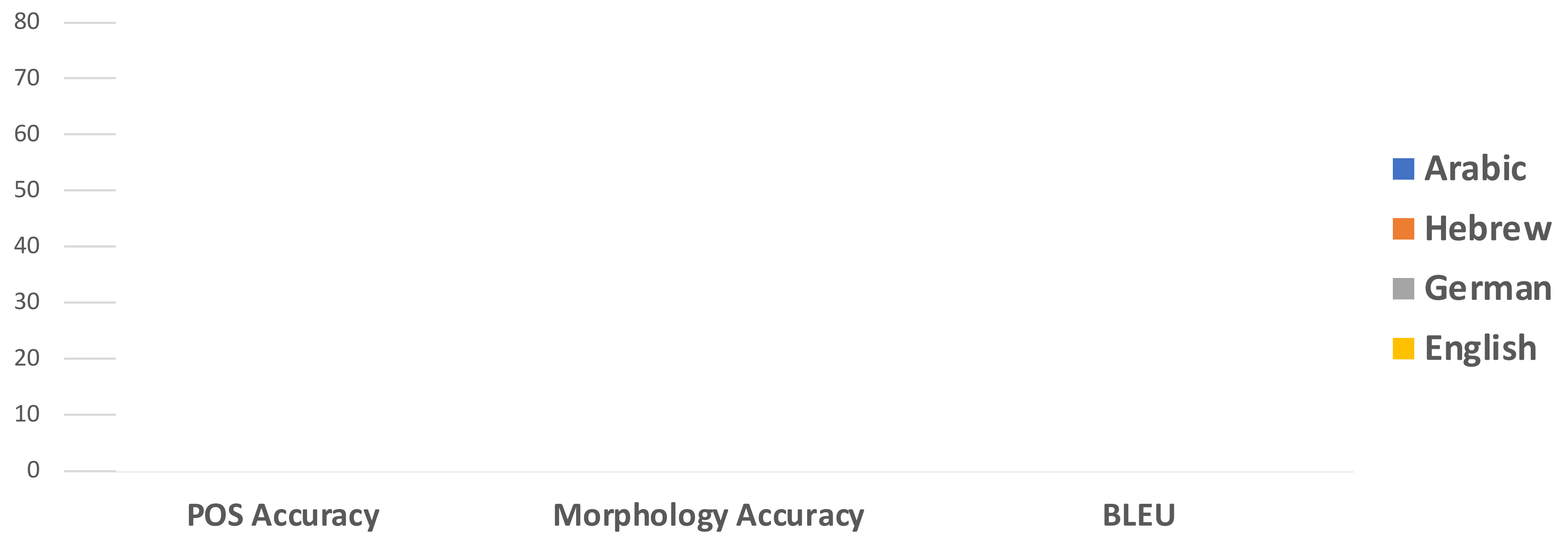
- Does the **target language** affect source-side representations?

Morphology

- Does the **target language** affect source-side representations?

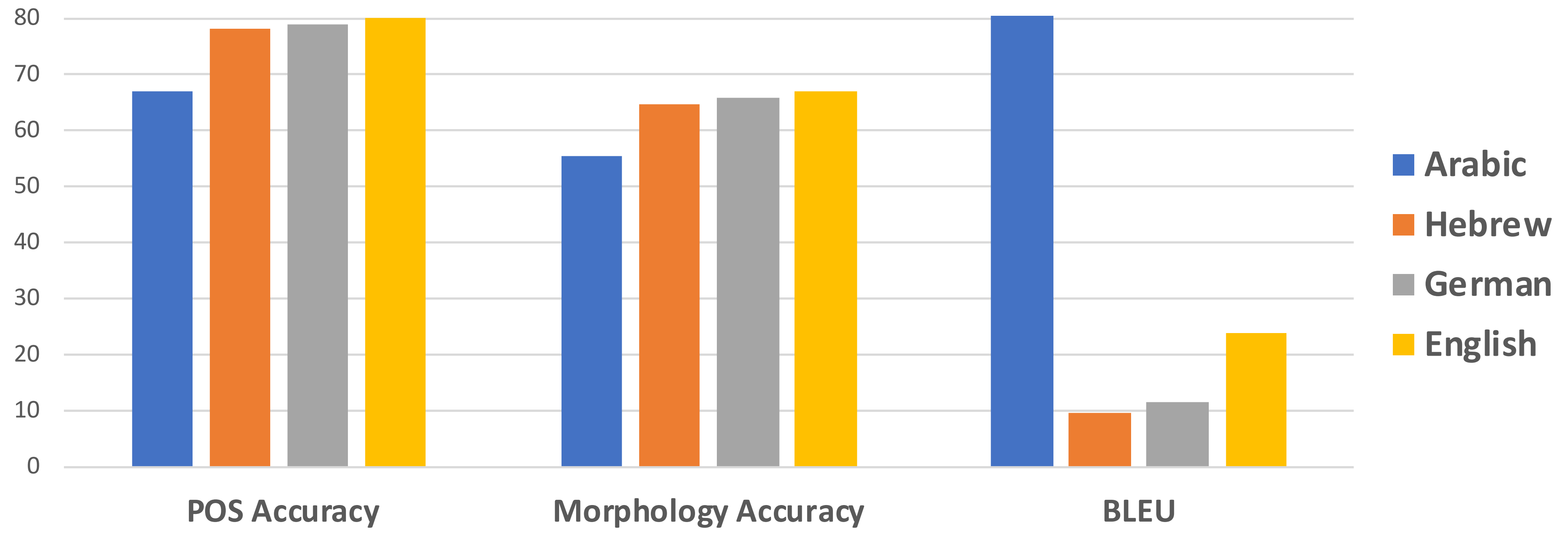
- Experiment:
  - Fix source side and train NMT models on different target languages
  - Compare learned representations on part-of-speech/morphological tagging

Morphology



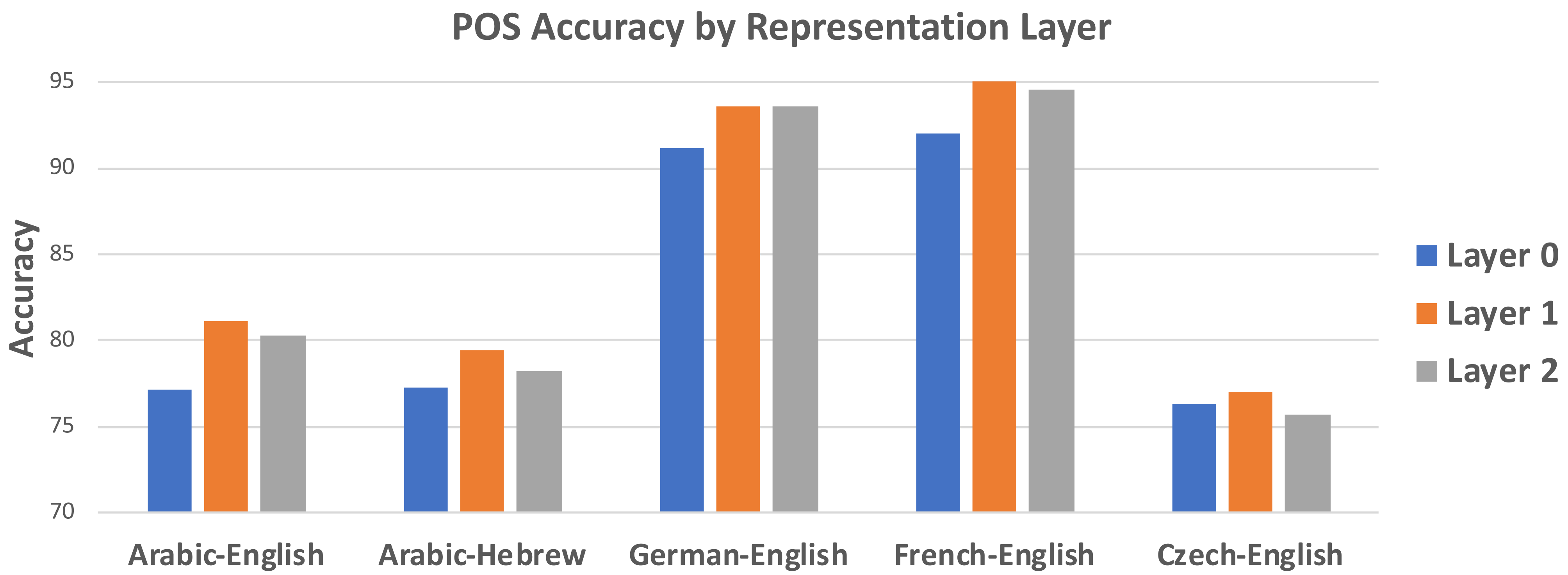
- Source language: Arabic
- Target languages: English, German, Hebrew, Arabic

Morphology



- Poorer target side morphology → better source side representations
- Higher BLEU ≠ better representations

Morphology



- Layer 1 > Layer 2 > Layer 0
- But deeper models translate better → what’s in layer 2?

(ACL 17)

Lexical Semantics

- “Evaluating Layers of Representations in NMT on POS and Semantic Tagging” (Belinkov+ 2017)
- Questions
  - What is captured in higher layers?
  - How is semantic information represented?

SEM Tagging

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

SEM Tagging

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

- Some examples
  - Determiners: *every*, *no*, *some*
  - Comma as conjunction, disjunction, apposition
  - Proper nouns: organization, location, person, etc.
  - Role nouns, entity nouns

SEM Tagging

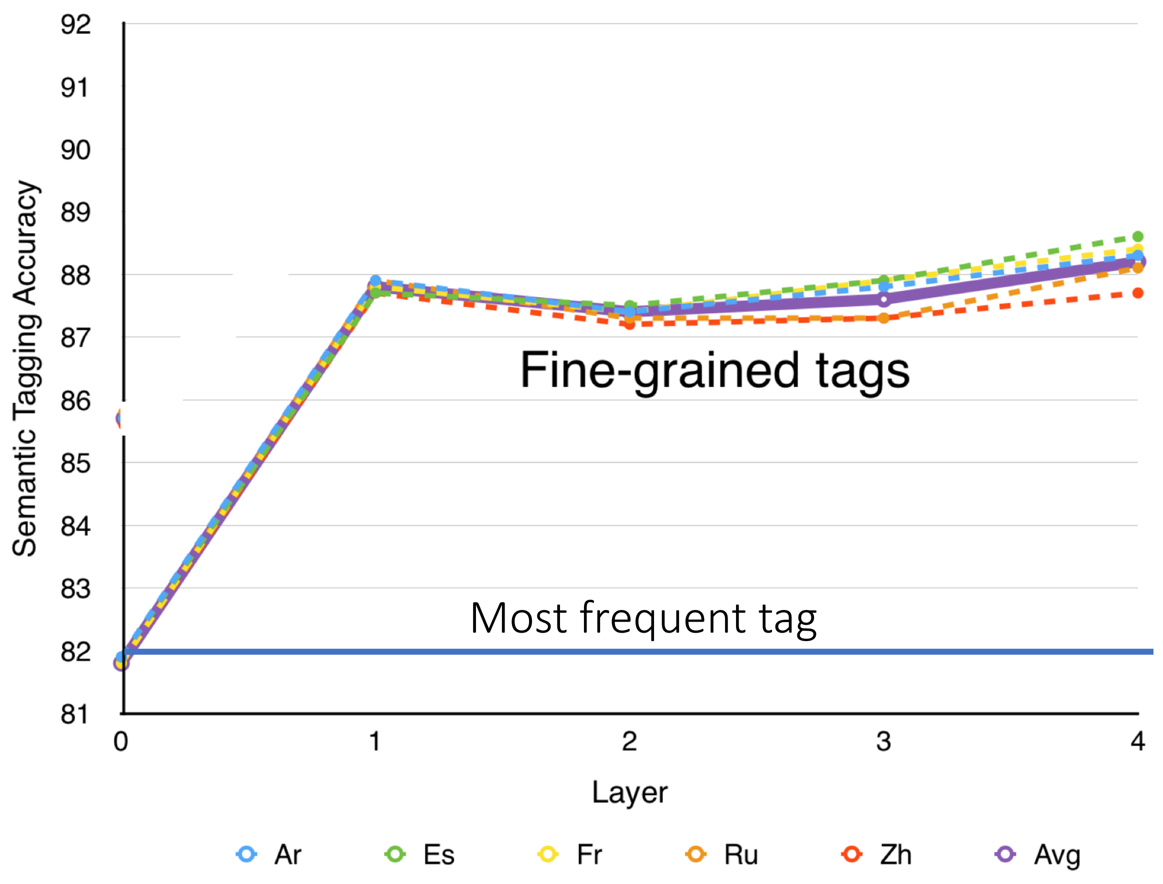
- Lexical semantics
- Abstraction over POS tagging
- Language-neutral, designed for multi-lingual semantic parsing

- Some examples
  - “Sarah bought *herself* a book”
  - ”Sarah *herself* bought a book”

  - *herself* – same POS tag but different SEM tags

SEM Tagging

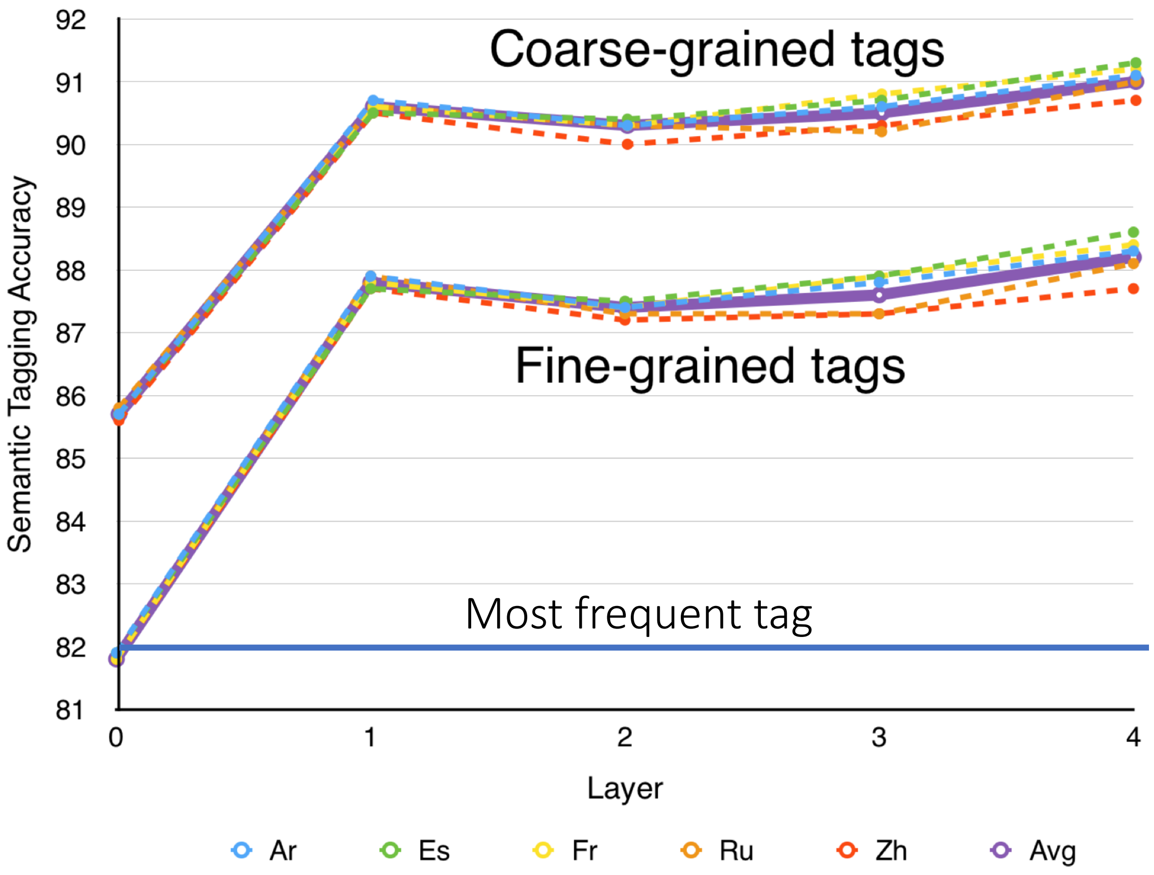
- Layer 0 below baseline
- Layer 1 >> layer 0
- Layer 4 > layer 1



SEM Tagging

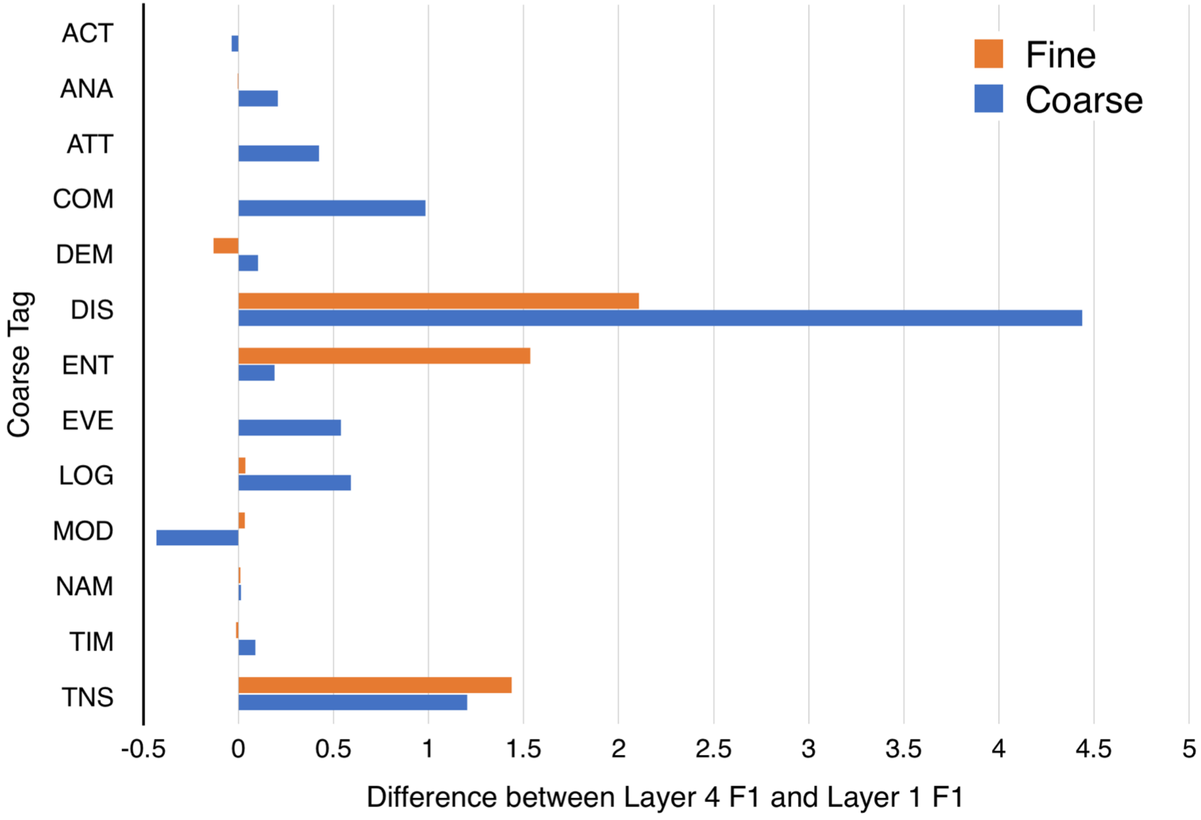
- Layer 0 below baseline
- Layer 1 >> layer 0
- Layer 4 > layer 1

- Similar trends for coarse tags



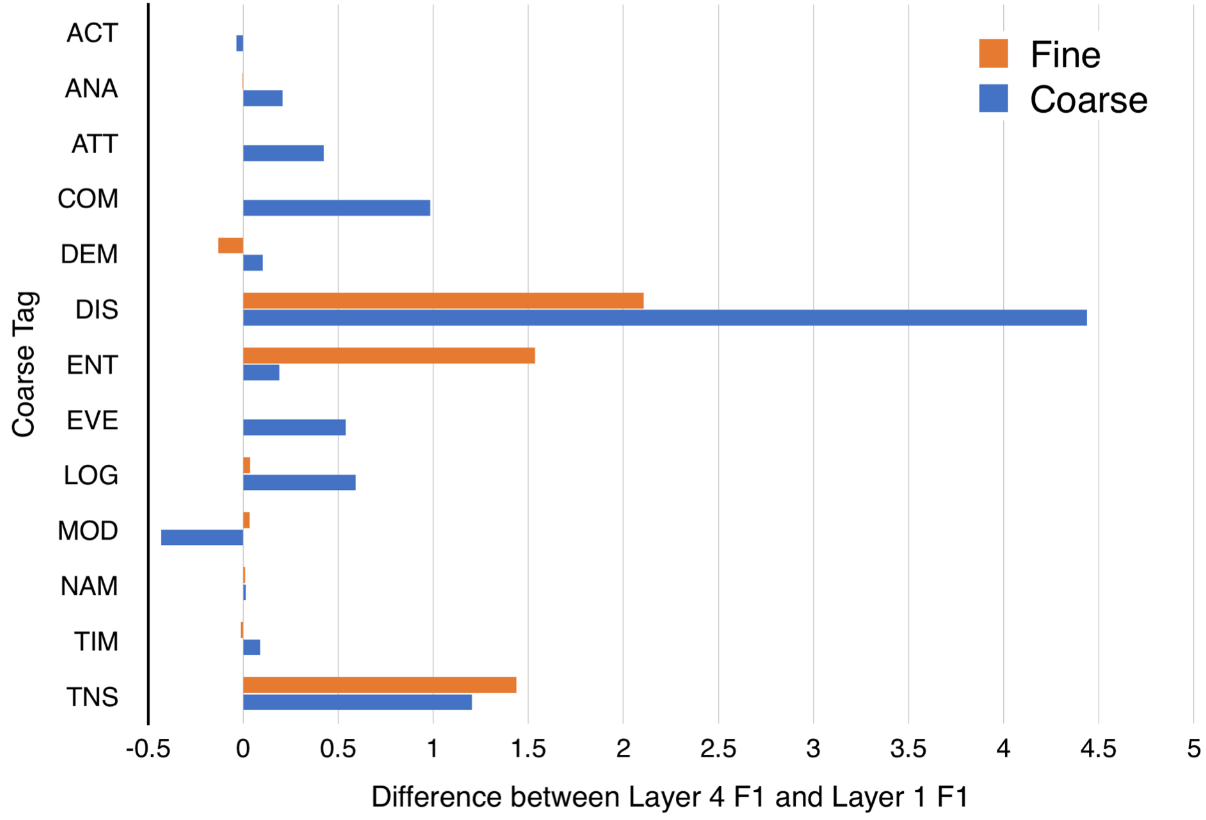
SEM Tagging

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



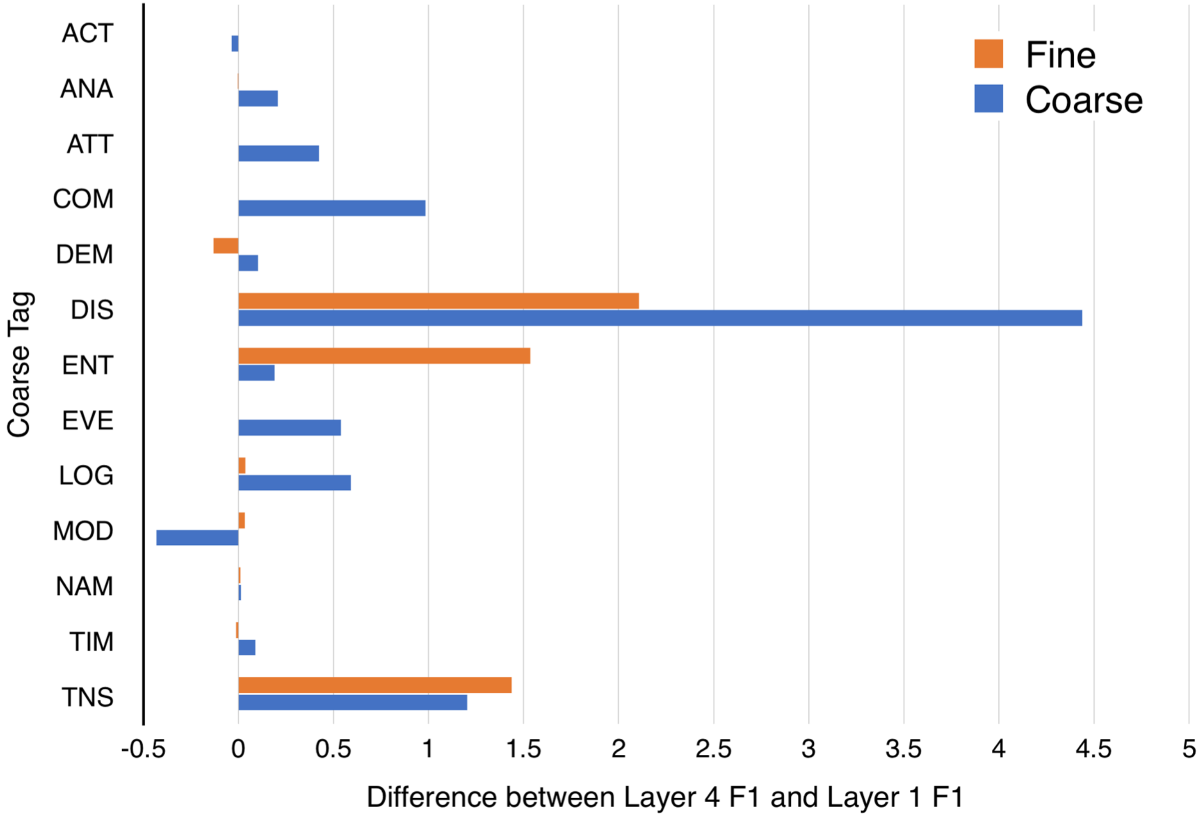
SEM Tagging

- 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*)



SEM Tagging

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



SEM tags vs. POS tags

SEM 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

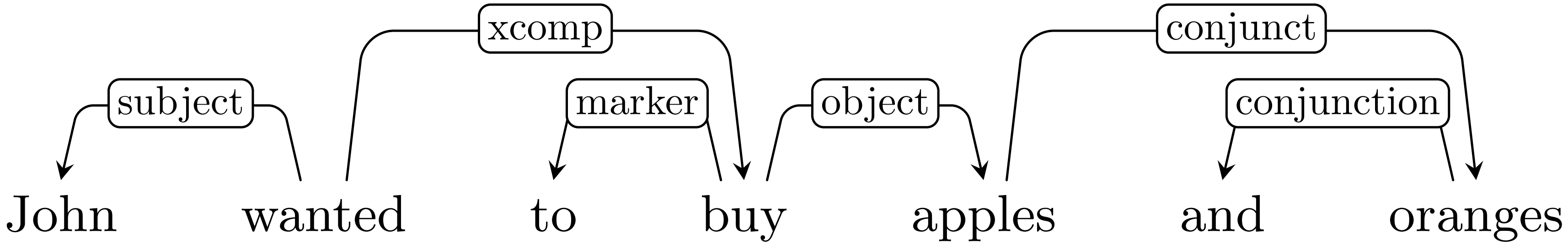
- Higher layers improve SEM tagging but not POS tagging
- Layer 1 best for POS; layer 4 best for SEM tagging

SEM tags vs. POS tags

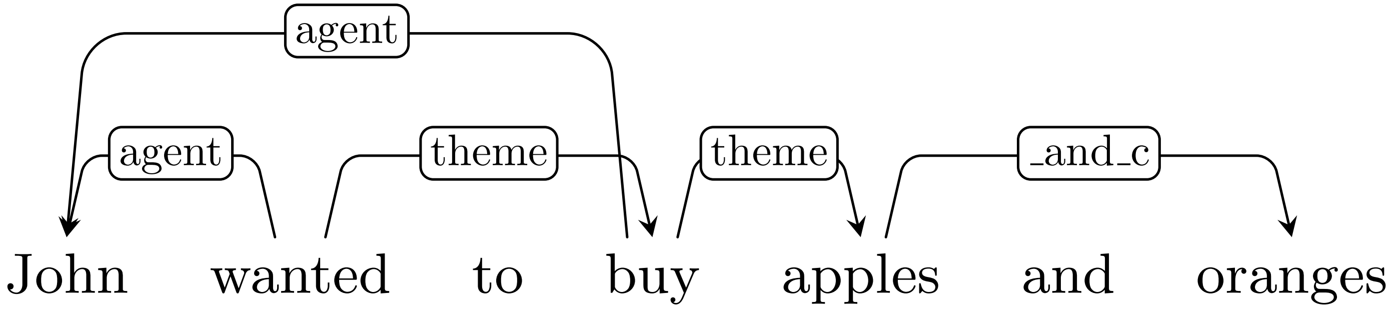
		0	1	2	3	4
Uni	POS	87.9	92.0	91.7	91.8	91.9
	SEM	81.8	87.8	87.4	87.6	88.2
Bi	POS	87.9	93.3	92.9	93.2	92.8
	SEM	81.9	91.3	90.8	91.9	91.9

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

Dependencies



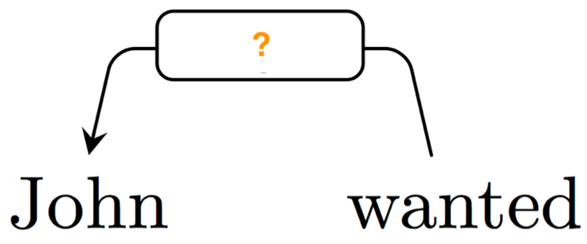
(a) Syntactic relations



(b) Semantic relations

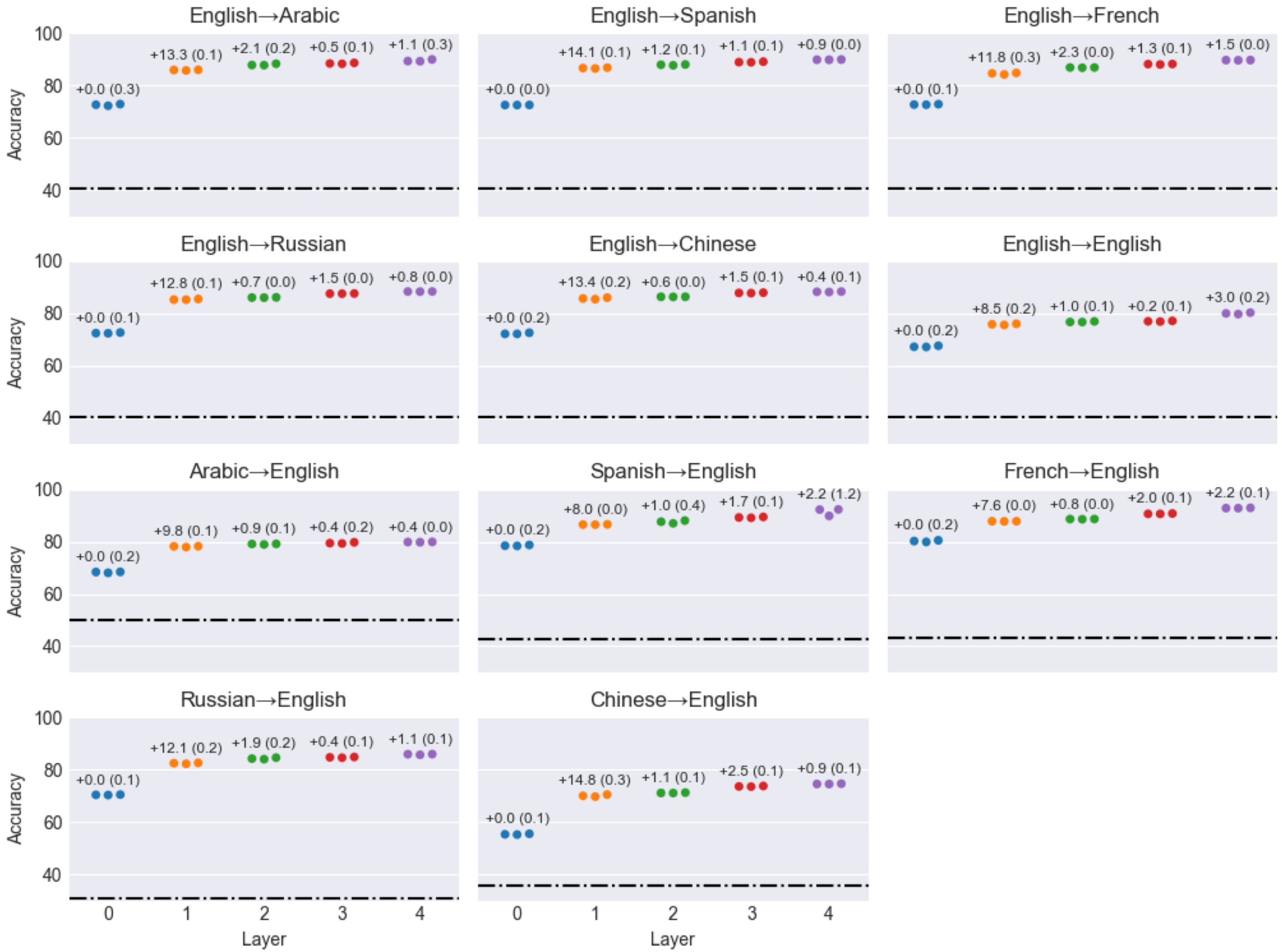
Dependencies

- Problem definition
  - Given two words, identify their relation
  - Train a classifier on NMT representations



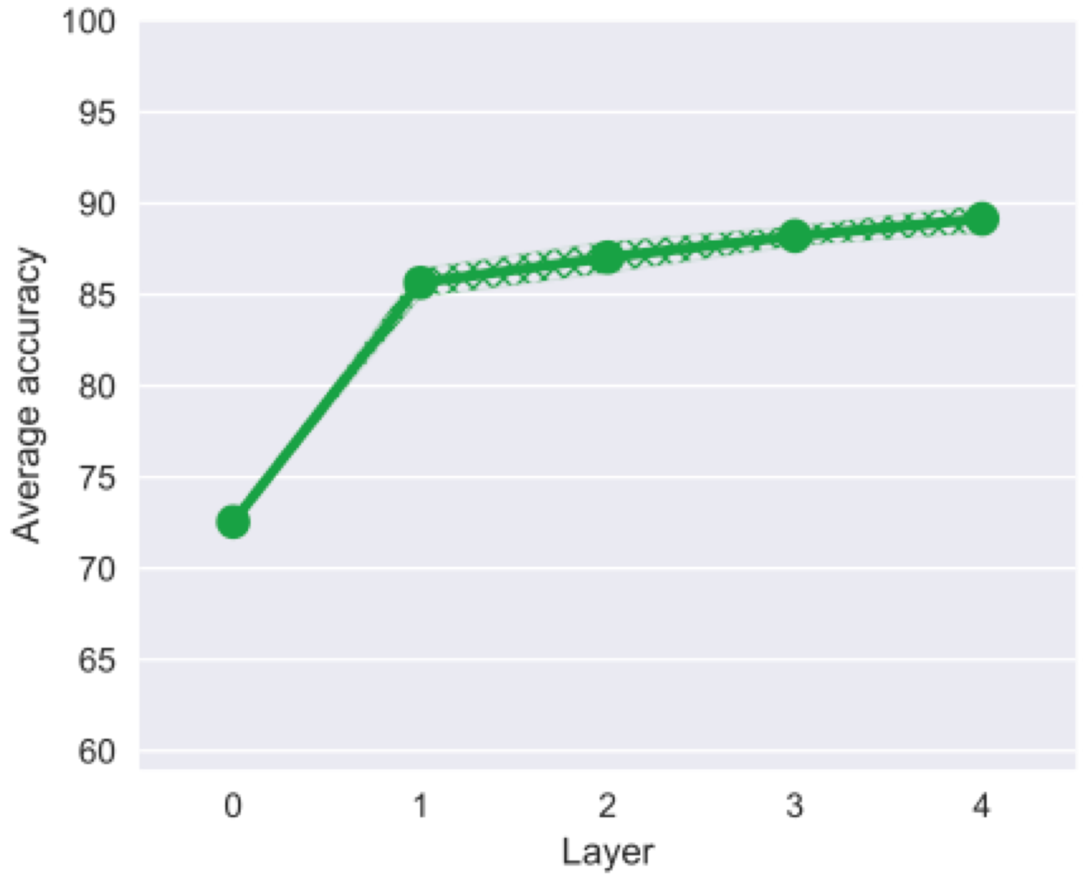
- Datasets
  - Syntax: Universal Dependencies (v2.0)
  - Semantics: Semantic Dependency parsing (Oepen+ 14-15)
  - MT data: UN corpus
  - Languages: Arabic, English, Spanish, French, Russian, Chinese

Syntactic Dependencies

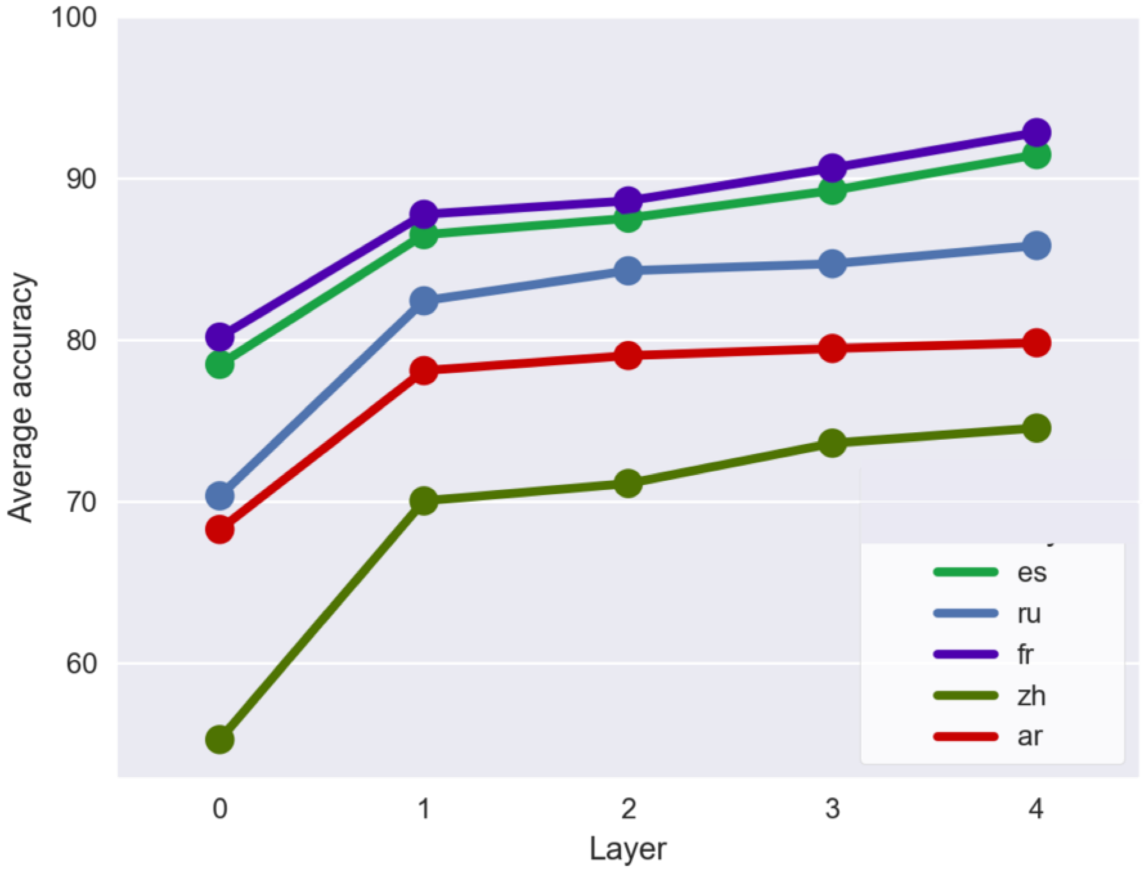


Syntactic Dependencies

English-to-*

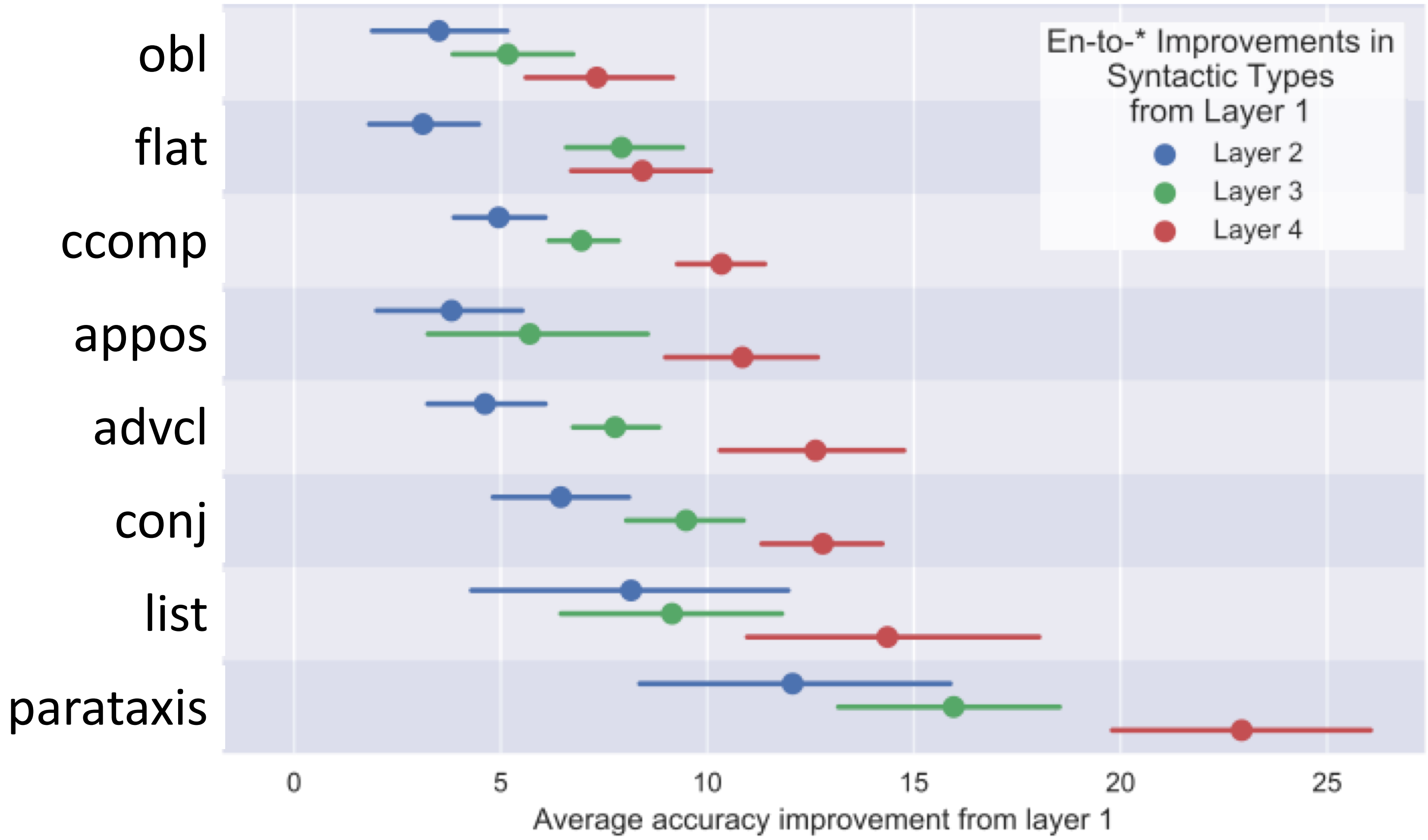


*-to-English

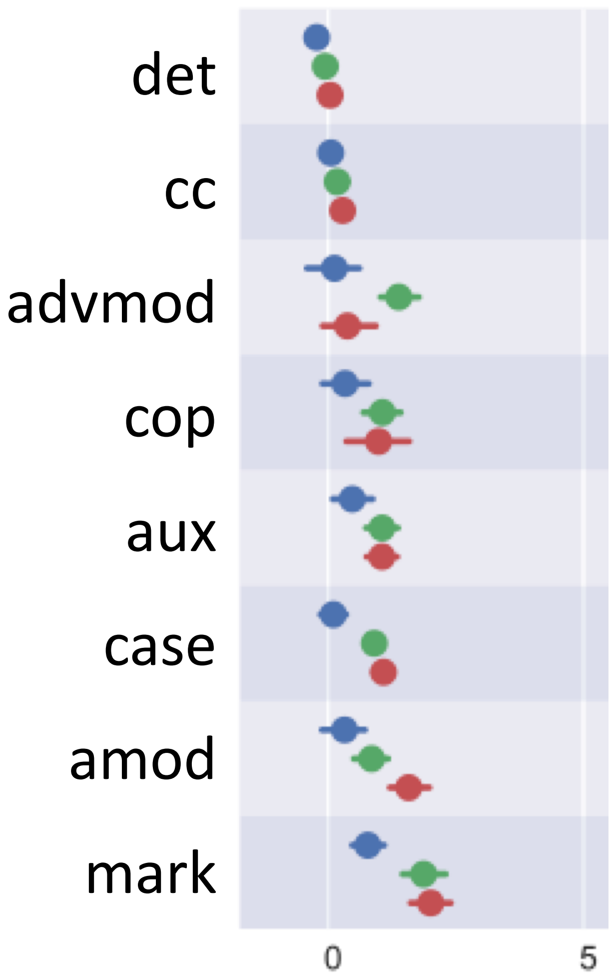


Specific Syntactic Relations

Most improvement in high layers

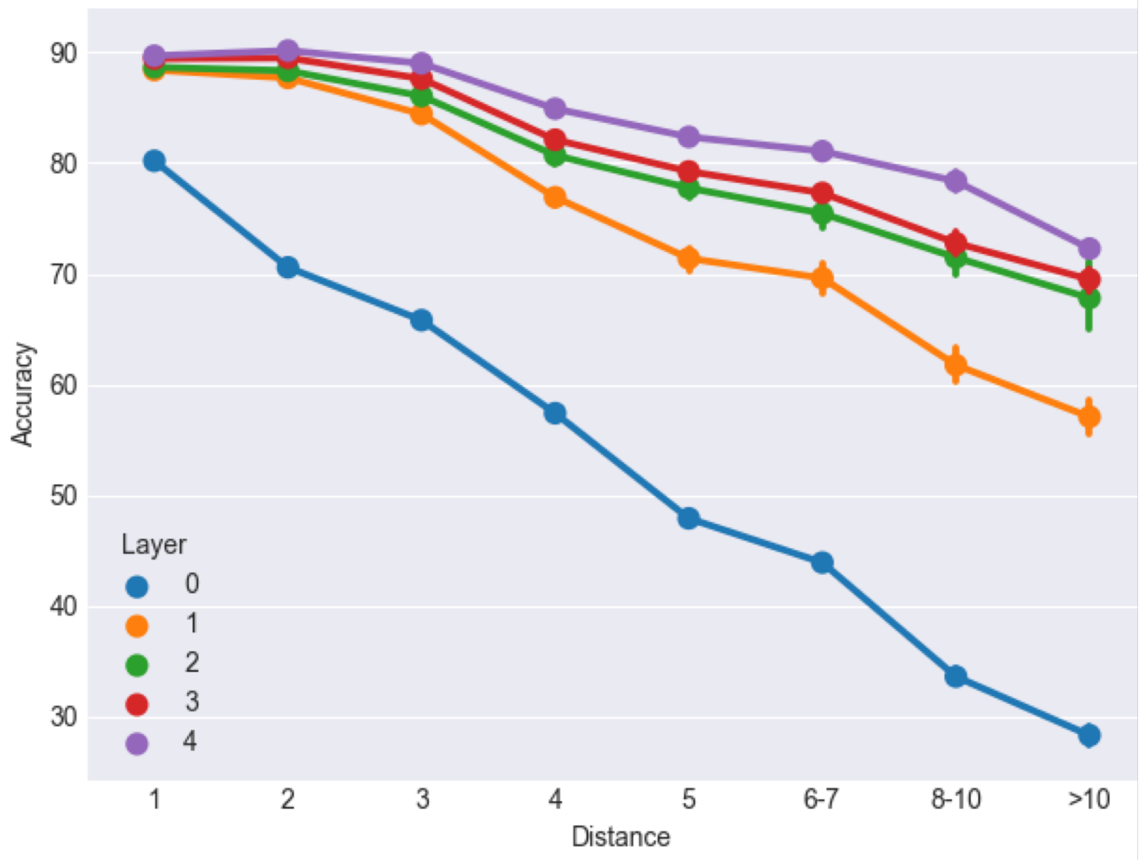


Least improvement

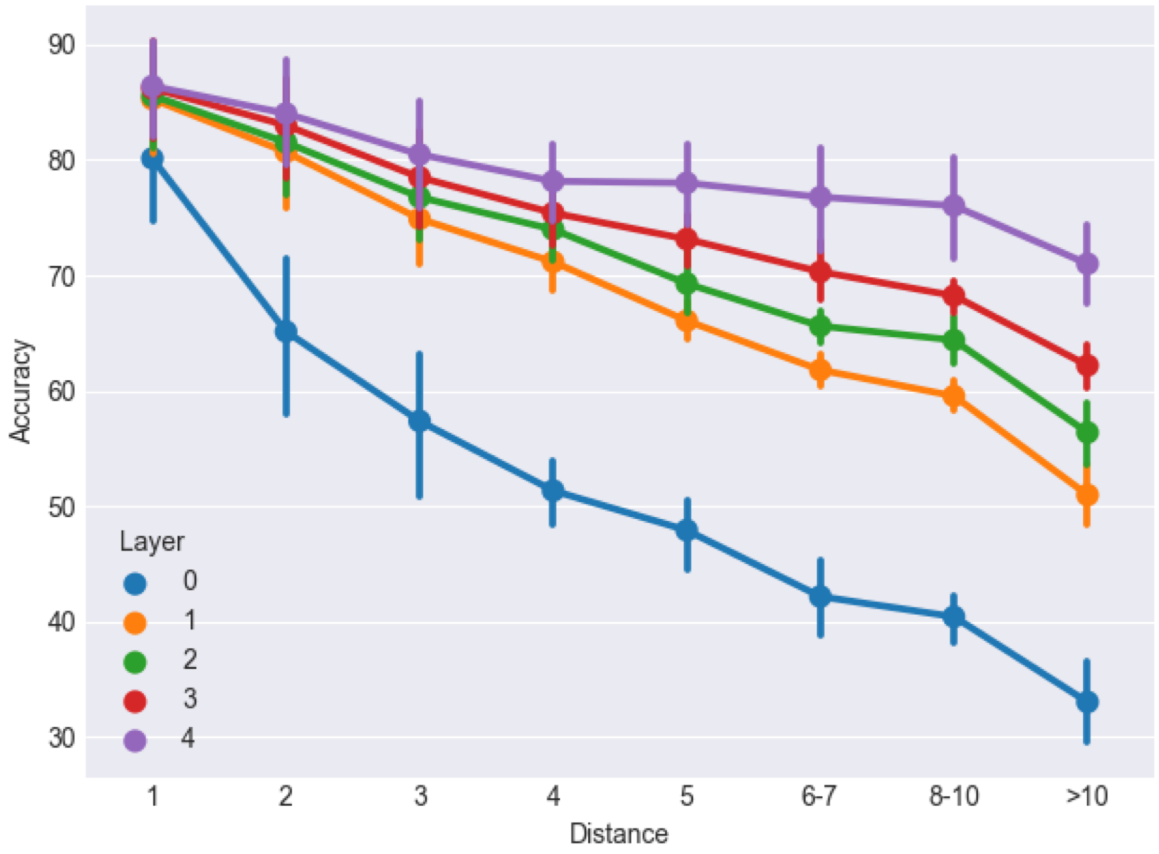


Effect of Distance

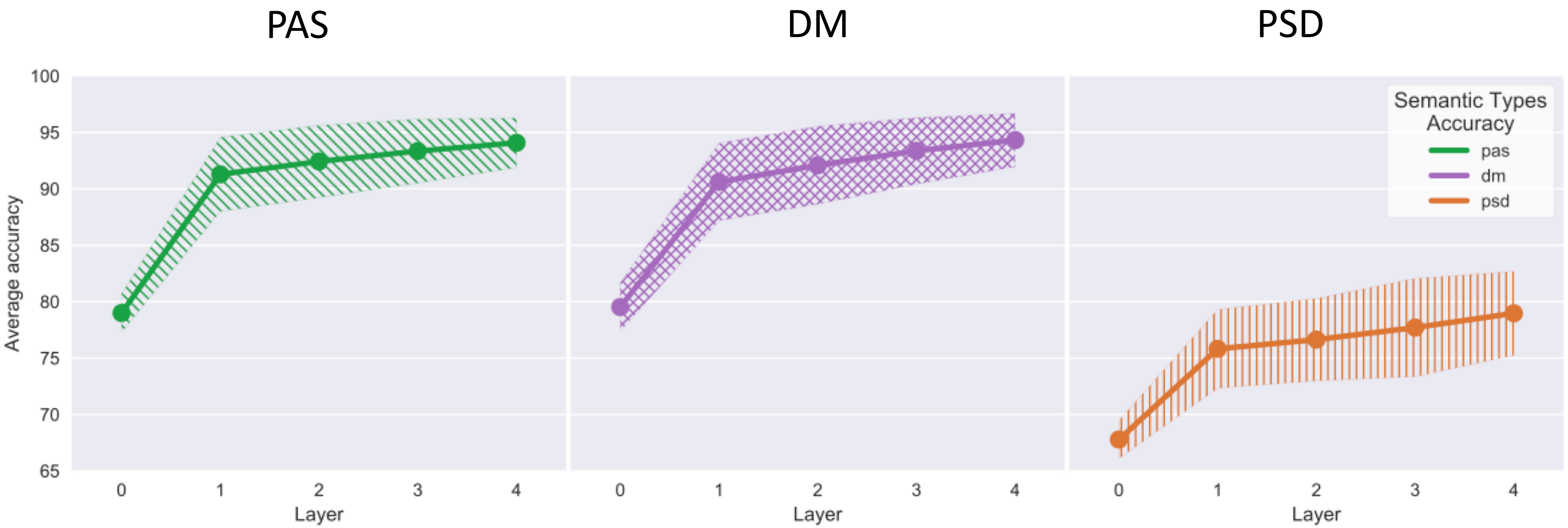
English-to-*



*-to-English



Semantic Dependencies



Open Questions

- Are individual dimensions in the vector representations meaningful?
  - We have some positive results (more on this later today)
- How much does NMT rely on the linguistic properties?
  - Can predict tense from NMT encodings at 90%, but NMT translations have correct tense only at 79% (Vanmassenhove+ 2017)
  - BLEU and sentence classification accuracy are in opposition (Cífka & Boyar 2018)
- NMT failures with adversarial examples
  - Black-box attacks (Belinkov & Bisk 2018; Higold+ 2018; Zhao+ 2018)
  - White-box attacks (Ebrahimi+ 2018; Cheng+ 2018)

Summary

- Neural MT representations contain useful information about morphology, syntax, and semantics

- Hierarchy of representations
  - Lower layers focus on local, short-distance properties (morphology)
  - Higher layers focus on global, long-distance properties (syntax, semantics)