CS 11-747 Neural Networks for NLP

Model Interpretation

Danish

Feb 28, 2019

Why interpretability?

Example from Caruana et al.

Why interpretability?

- **Task:** predict probability of death for patients with pneumonia

- **Why**: so that high-risk patients can be admitted, low risk patients can be treated as outpatients

Example from Caruana et al.

Why interpretability?

- **Task:** predict probability of death for patients with pneumonia
- **Why**: so that high-risk patients can be admitted, low risk patients can be treated as outpatients
- $\text{AUC}_{\text{Neural networks}} > \text{AUC}_{\text{Logistic Regression}}$

Example from Caruana et al.

Why interpretability?

- **Task:** predict probability of death for patients with pneumonia

- **Why**: so that high-risk patients can be admitted, low risk patients can be treated as outpatients

- $AUC_{\text{Neural networks}} > AUC_{\text{Logistic Regression}}$

- Rule based classifier

  HasAsthma(X) —> LowerRisk(X)

Example from Caruana et al.

Why interpretability?

- **Task:** predict probability of death for patients with pneumonia

- **Why**: so that high-risk patients can be admitted, low risk patients can be treated as outpatients

- $AUC_{\text{Neural networks}} > AUC_{\text{Logistic Regression}}$

- Rule based classifier

HasAsthma(X) —> LowerRisk(X)

more intensive care

Example from Caruana et al.

Why interpretability?

Why interpretability?

- Legal reasons: uninterpretable models are banned!
  - GDPR in EU necessitates "right to explanation"

Why interpretability?

- Legal reasons: uninterpretable models are banned!
  - GDPR in EU necessitates "right to explanation"

- Distribution shift: deployed model might perform poorly *in the wild*

Why interpretability?

- Legal reasons: uninterpretable models are banned!
  - GDPR in EU necessitates "right to explanation"
- Distribution shift: deployed model might perform poorly *in the wild*
- User adoption: users happier with explanations
- Better Human-AI interaction and control
- Debugging machine learning models

Dictionary definition

interpret verb

in·ter·pret | \ in-ˈtər-prət, -pət\

interpreted; **interpreting**; **interprets**

Definition of *interpret*

transitive verb

1 **:** to explain or tell the meaning of **:** present in understandable terms

// *interpret* dreams

// needed help *interpreting* the results

As per Merriam Webster, accessed on 02/25

Dictionary definition

interpret verb

in·ter·pret | \ in-ˈtər-prət, -pət\

interpreted; **interpreting**; **interprets**

Definition of *interpret*

transitive verb

1 **:** to explain or tell the meaning of **:** present in understandable terms

// *interpret* dreams

// needed help *interpreting* the results

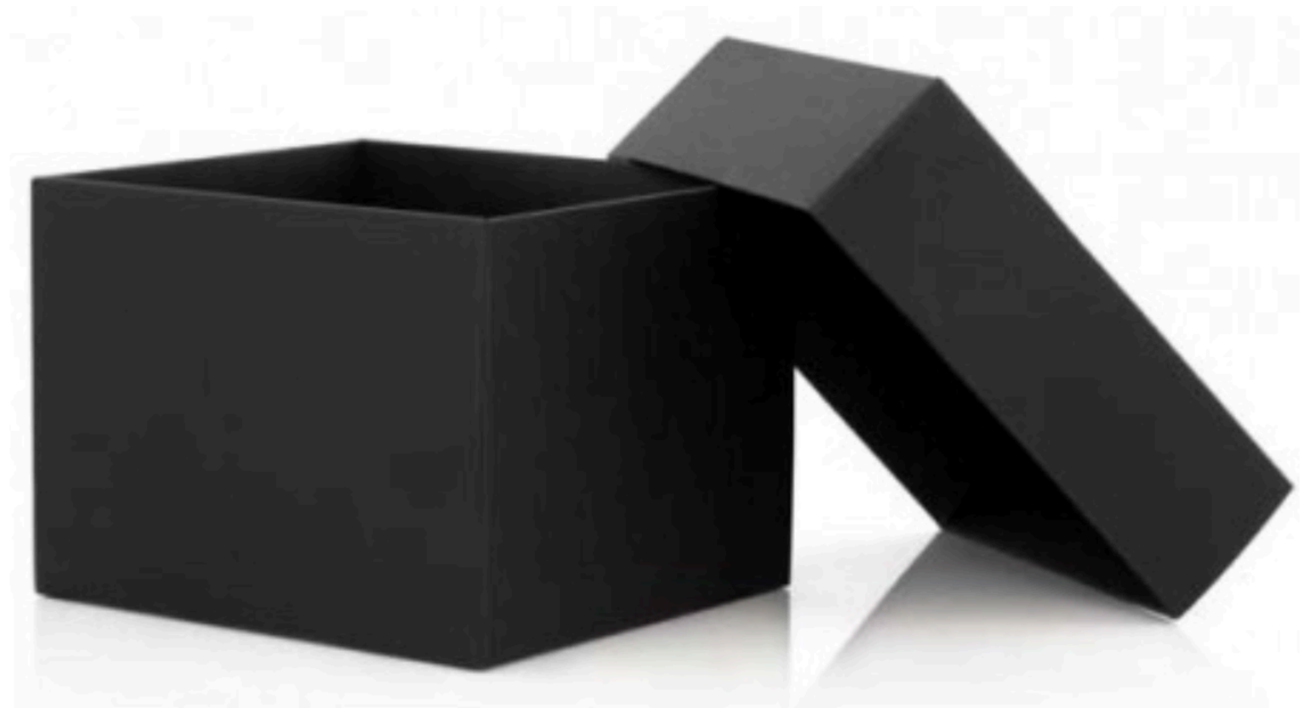
Only if we could understand

`model.ckpt`

As per Merriam Webster, accessed on 02/25

Two broad themes

- What is the model learning?
- Can we explain the prediction in "understandable terms"?



Comparing two directions

What is the model learning?

- Input: a model M, **a (linguistic) property P**
- Output: extent to which M captures P
- Techniques: classification, regression
- Evaluation: implicit

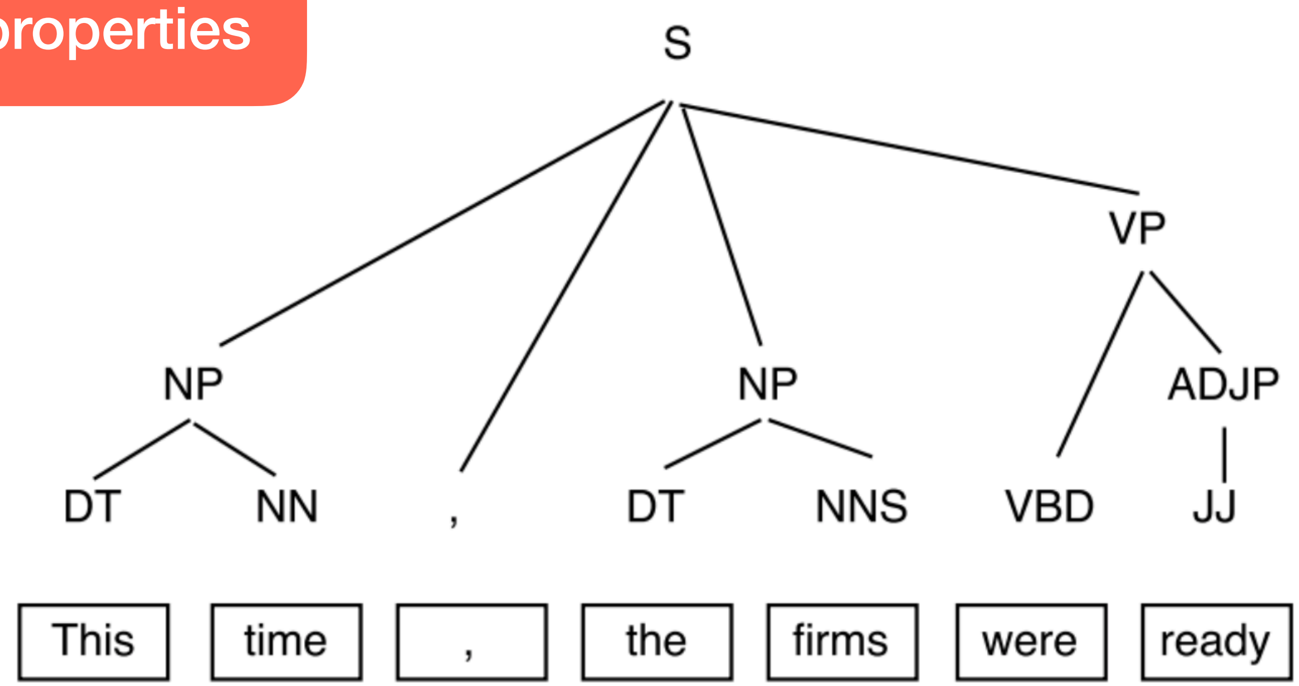
Explain the prediction

- Input: a model M, **a test example X**
- Output: an explanation E
- Techniques: varied …
- Evaluation: complicated

What is the model learning?

Source Syntax in NMT

5 syntactic properties



POS	DT	NN	,	DT	NNS	VBD	JJ
Smallest Phrase Constituent	NP	NP	-	NP	NP	VP	ADJP
Top-level Syntactic Sequence	(NP, NP, VP)						
Voice	Active						
Tense	Past						

Does String-Based Neural MT Learn Source Syntax? Shi et al. EMNLP 2016

Source Syntax in NMT

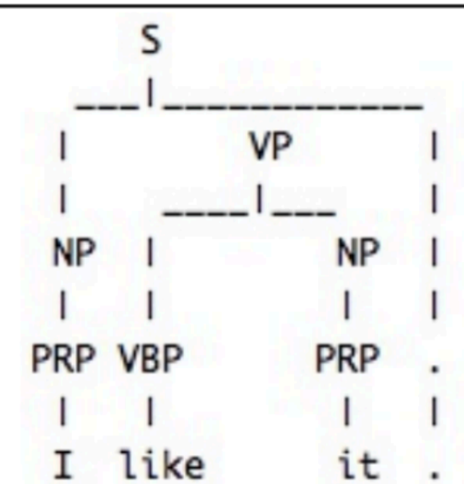
Model	Source	Target
E2E	I like it .	I like it .
PE2PE	it I . like	it I . like
E2F	I like it .	J'aime ça.
E2G	I like it .	Ich mag das.
E2P	I like it .	 (S (NP PRP $)_{NP}$ (VP VBP (NP PRP $)_{NP})_{VP}$. $)_{S}$

Figure 1: Sample inputs and outputs of the E2E, PE2PE, E2F, E2G, and E2P models.

Does String-Based Neural MT Learn Source Syntax? Shi et al. EMNLP 2016

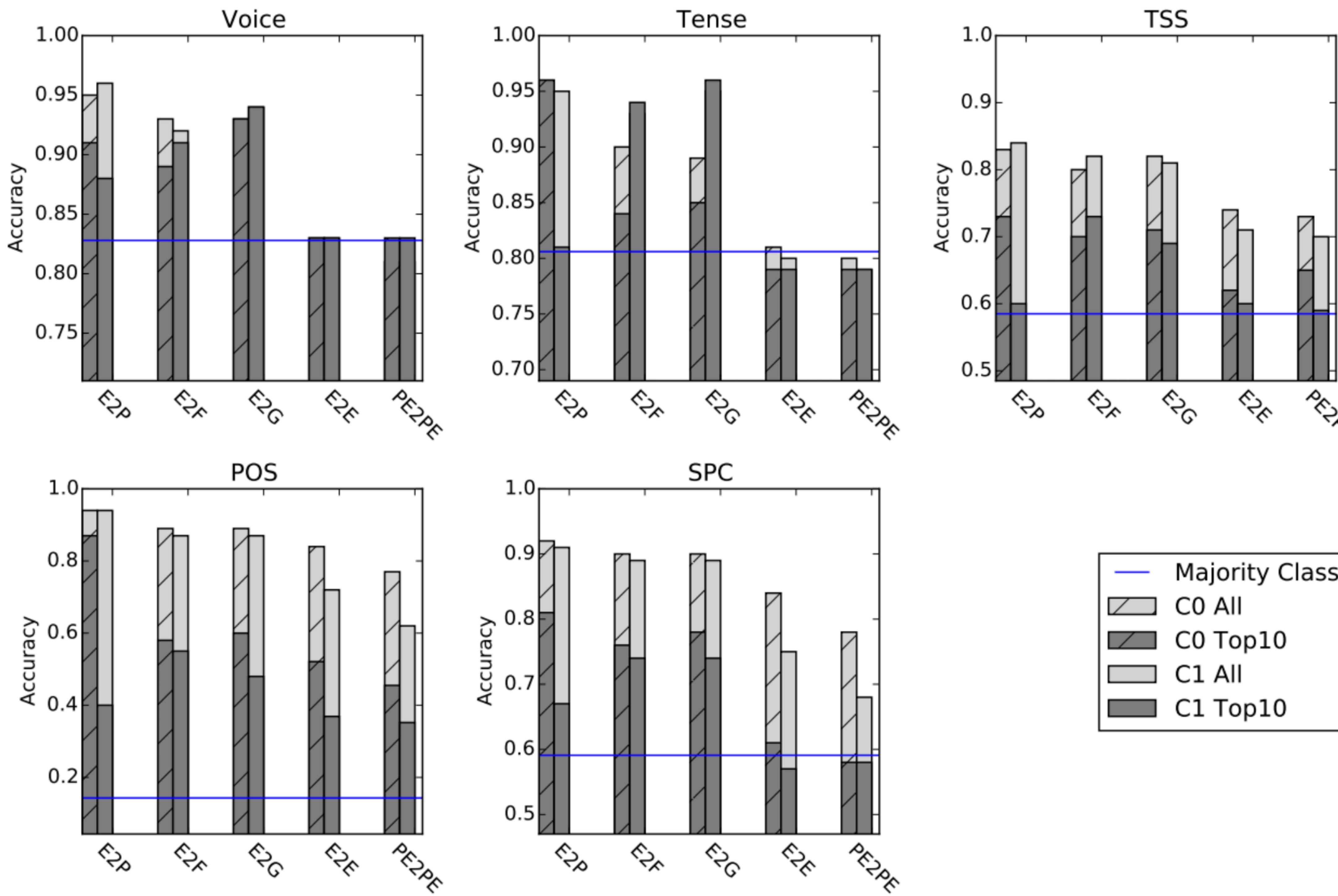
Source Syntax in NMT

Model	Accuracy
Majority Class	82.8
English to French (E2F)	92.8
English to English (E2E)	82.7

Table 1: Voice (active/passive) prediction accuracy using the encoding vector of an NMT system. The majority class baseline always chooses active.

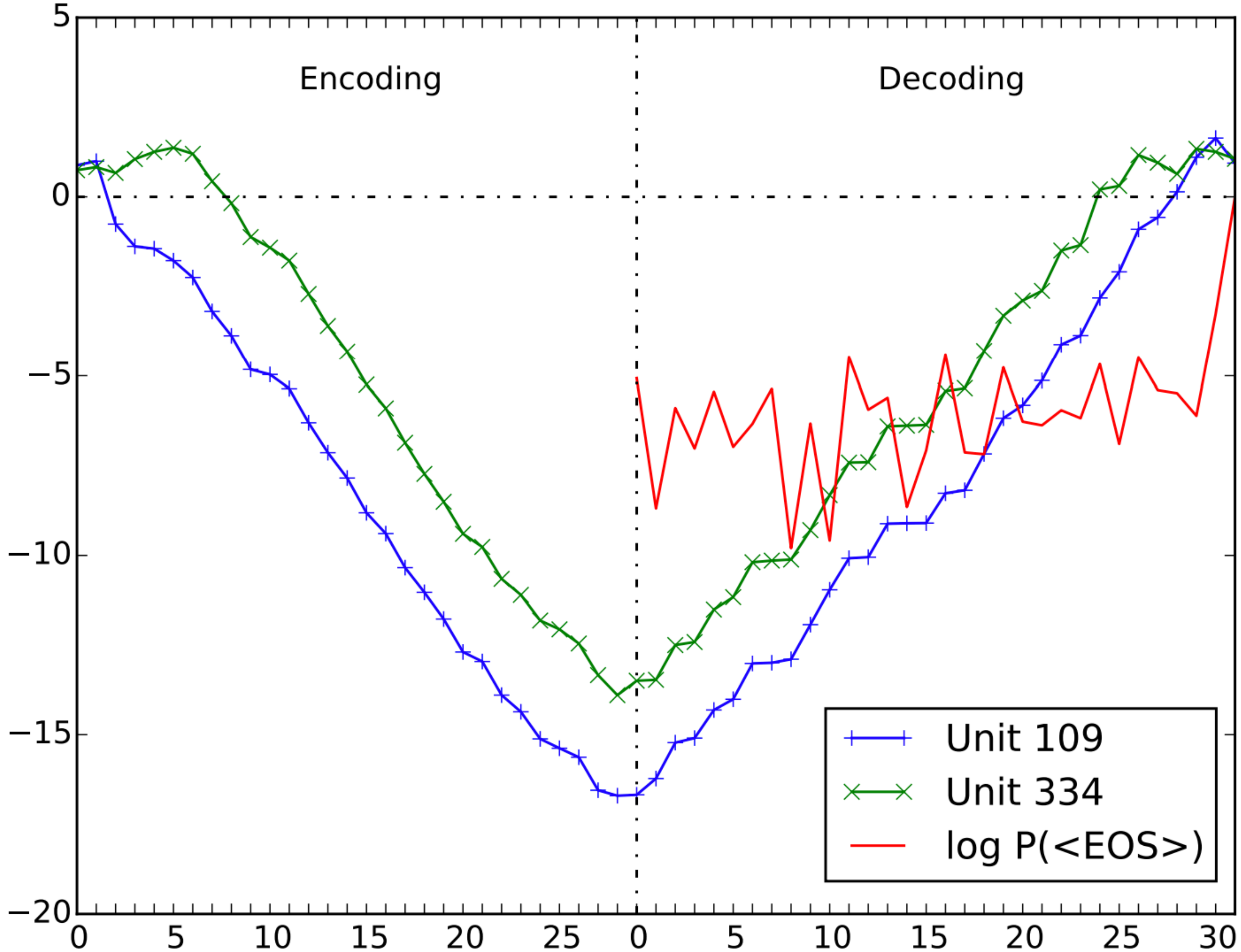
Does String-Based Neural MT Learn Source Syntax? Shi et al. EMNLP 2016

Source Syntax in NMT



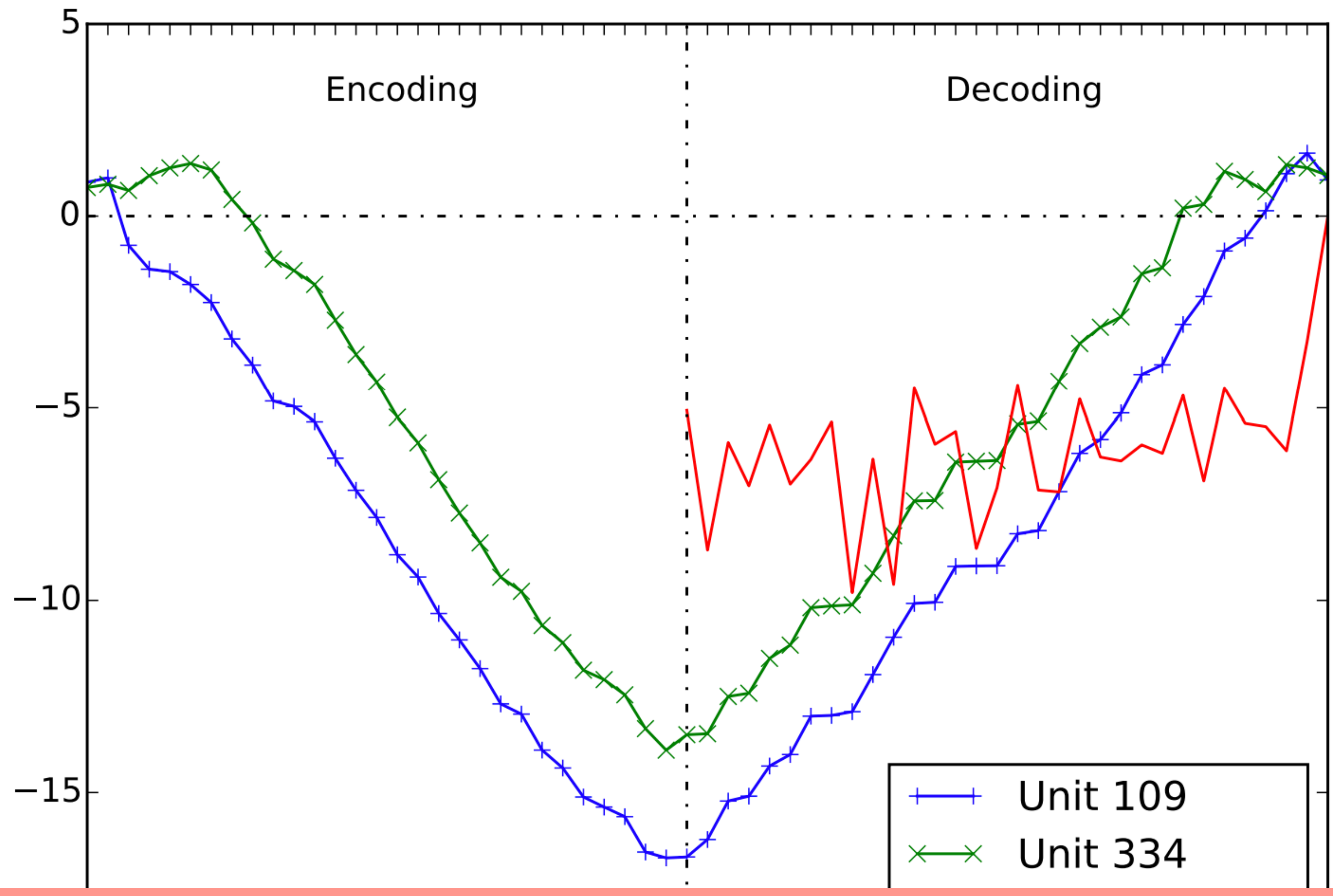
Does String-Based Neural MT Learn Source Syntax? Shi et al. EMNLP 2016

Why neural translations are the right length?



Shi et al. EMNLP 2016

Why neural translations are the right length?



Note: LSTMs can learn to count, whereas GRUs can not do unbounded counting (Weiss et al. ACL 2018)

Shi et al. EMNLP 2016

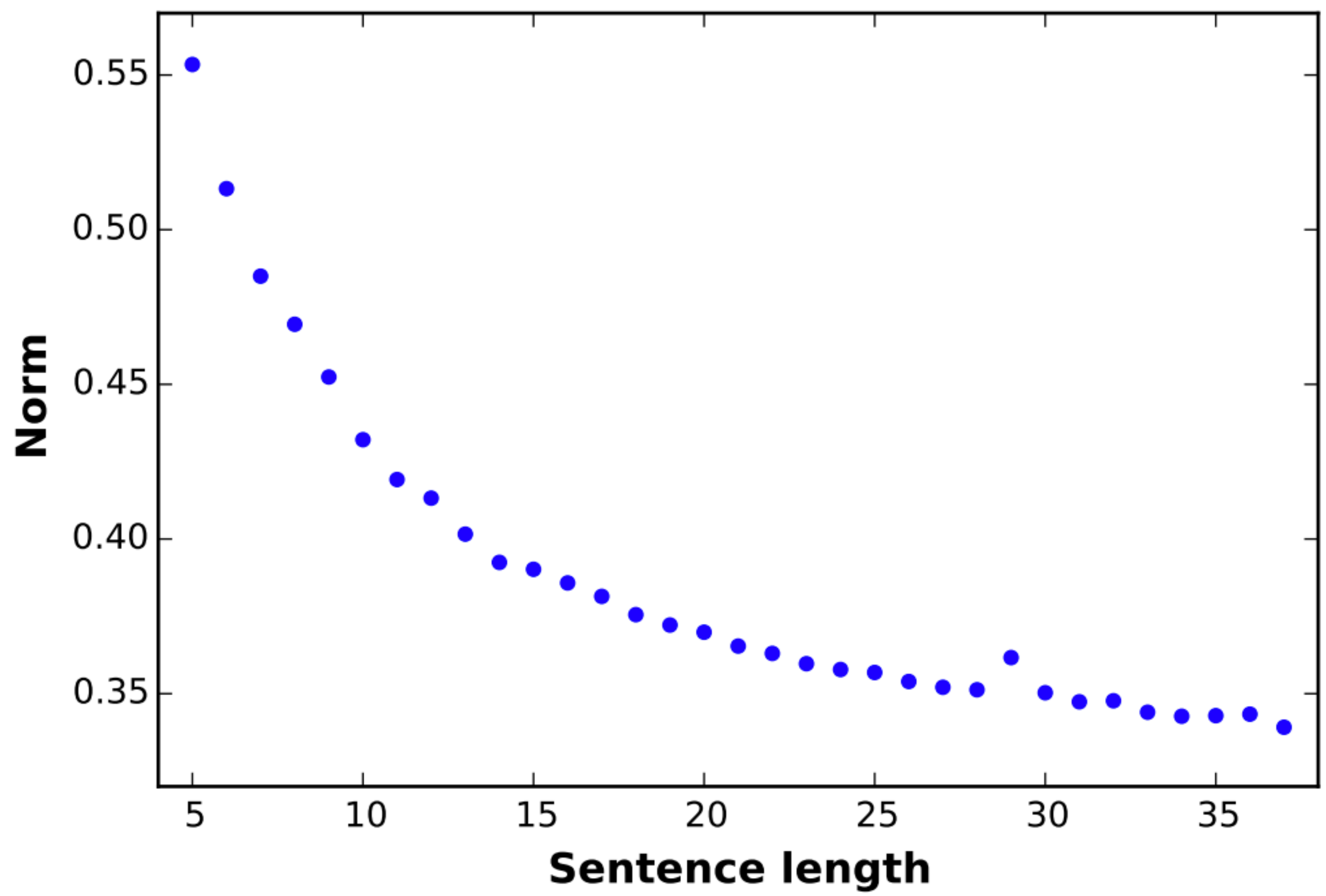
Fine grained analysis of sentence embeddings

- Sentence representations: word vector averaging, hidden states of the LSTM

- Auxiliary Tasks: predicting length, word order, content

- Findings:
  - hidden states of LSTM capture to a great deal length, word order and content
  - word vector averaging (CBOW) model captures content, length (!), word order (!!)

Fine grained analysis of sentence embeddings



(b) Average embedding norm vs. sentence length for CBOW with an embedding size of 300.

More work…

- Discuss the following two in some detail
- Fine-grained analysis of sentence embeddings using auxiliary prediction tasks
- What you can cram into a single vector: Probing sentence embeddings for linguistic properties
- Point to a survey and the table here: https://boknilev.github.io/nlp-analysis-methods/table1.html

What you can cram into a single vector: Probing sentence embeddings for linguistic properties

- "you cannot cram the meaning of a whole %&!$# sentence into a single $&!#* vector" — Ray Mooney

- Design 10 probing tasks: len, word content, bigram shift, tree depth, top constituency, tense, subject number, object number, semantically odd man out, coordination inversion

- Test BiLSTM last, BiLSTM max, Gated ConvNet encoder

Summary: What is the model learning?

https://boknilev.github.io/nlp-analysis-methods/table1.html

Explain the prediction

How to evaluate?

Training Phase

Some **x, f(x)** pairs

Some **x, f(x), E** triples

Test Phase

Input **x**
Predict **f(x)**

Input **x**
Predict **f(x)**

Automatic evaluation

Morphosyntactic Agreement

The link provided by the editor above **encourages ….**

Poerner et al, ACL 2018

Automatic evaluation

<u>Morphosyntactic Agreement</u>

The link provided by the editor above **encourages ….**

<u>Hybrid documents</u>

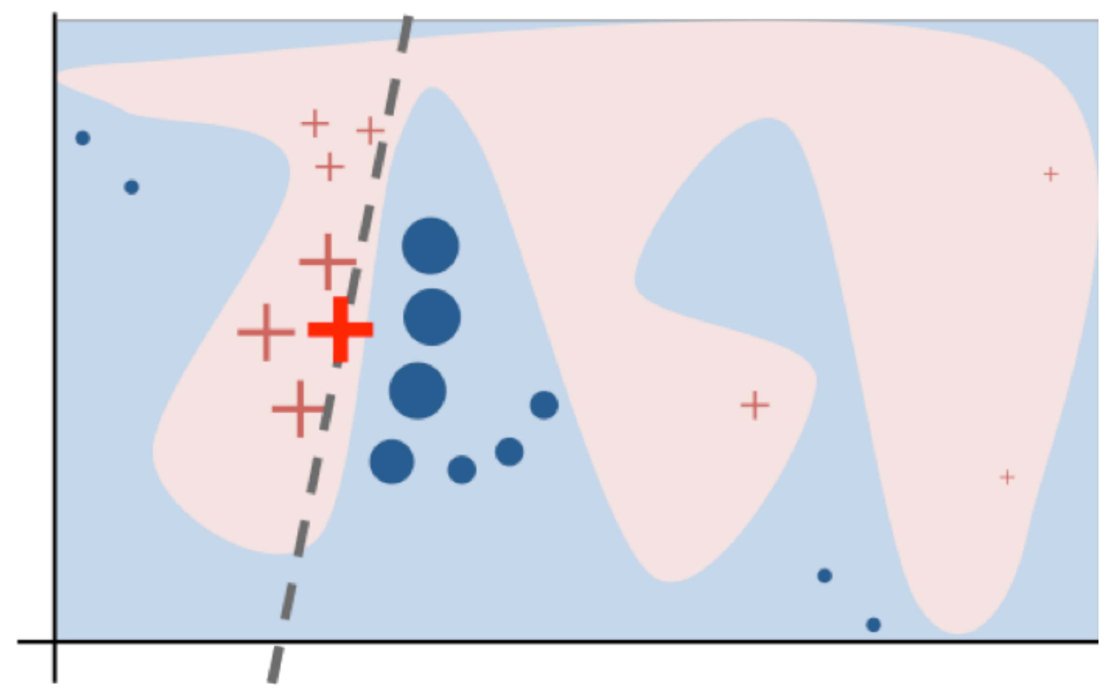
This is collected from Document 1. This text comes from Document 2. …. This text is taken from Document n.

Poerner et al, ACL 2018

Explanation Technique: LIME

Ribeiro et al, KDD 2016

Explanation Technique: LIME



Ribeiro et al, KDD 2016

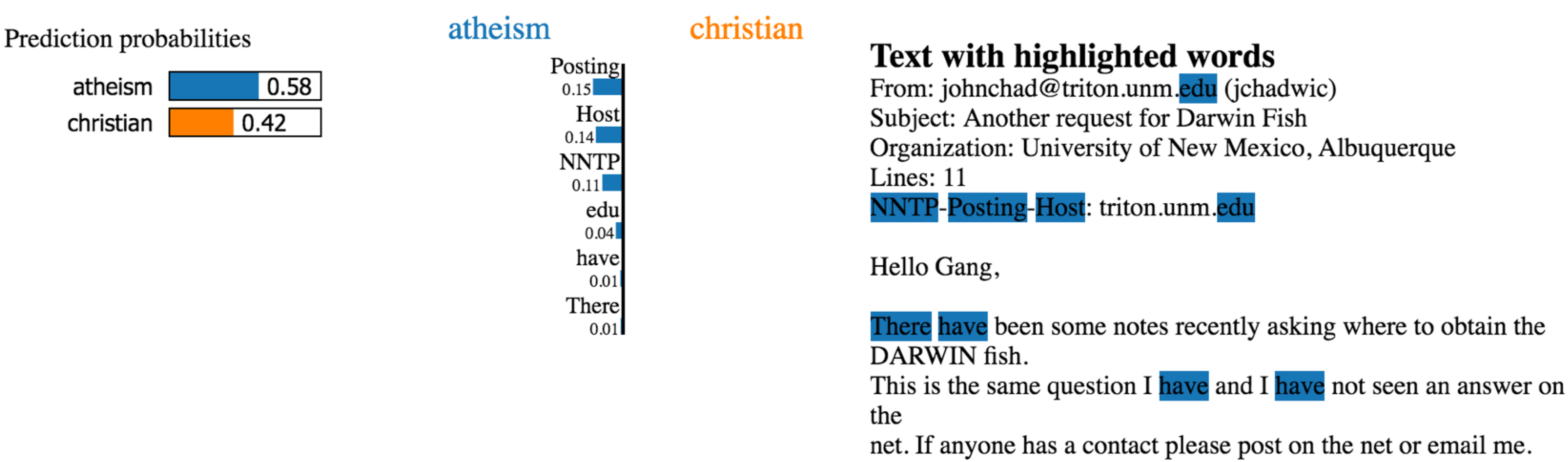
Explanation Technique: LIME

(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception network, highlighting positive pixels. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$)**, "Acoustic guitar"** ($p = 0.24$) **and "Labrador"** ($p = 0.21$)

Ribeiro et al, KDD 2016

Explanation Technique: LIME

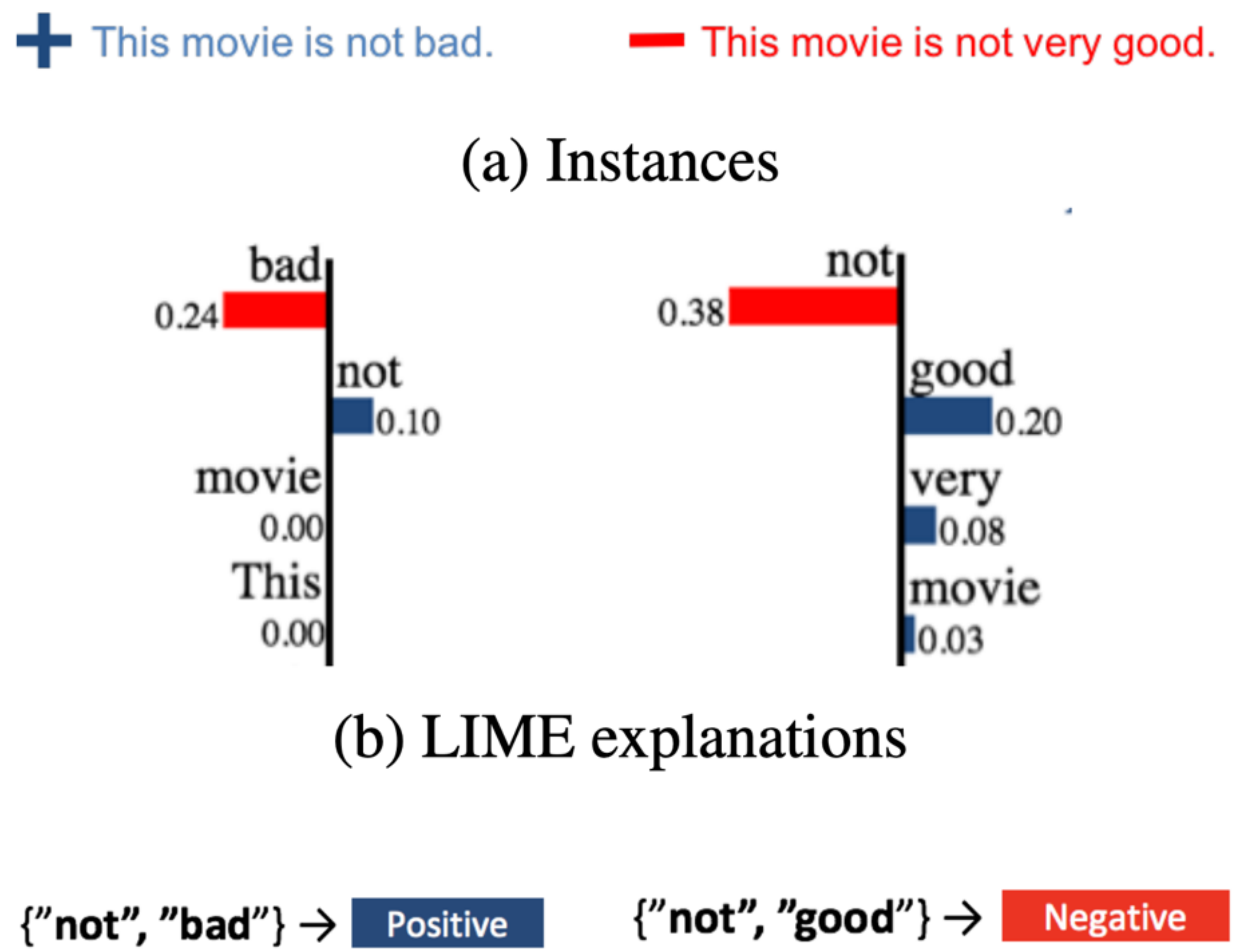


Ribeiro et al, KDD 2016

Explanation Technique: Anchors

Ribeiro et al, AAAI 2018

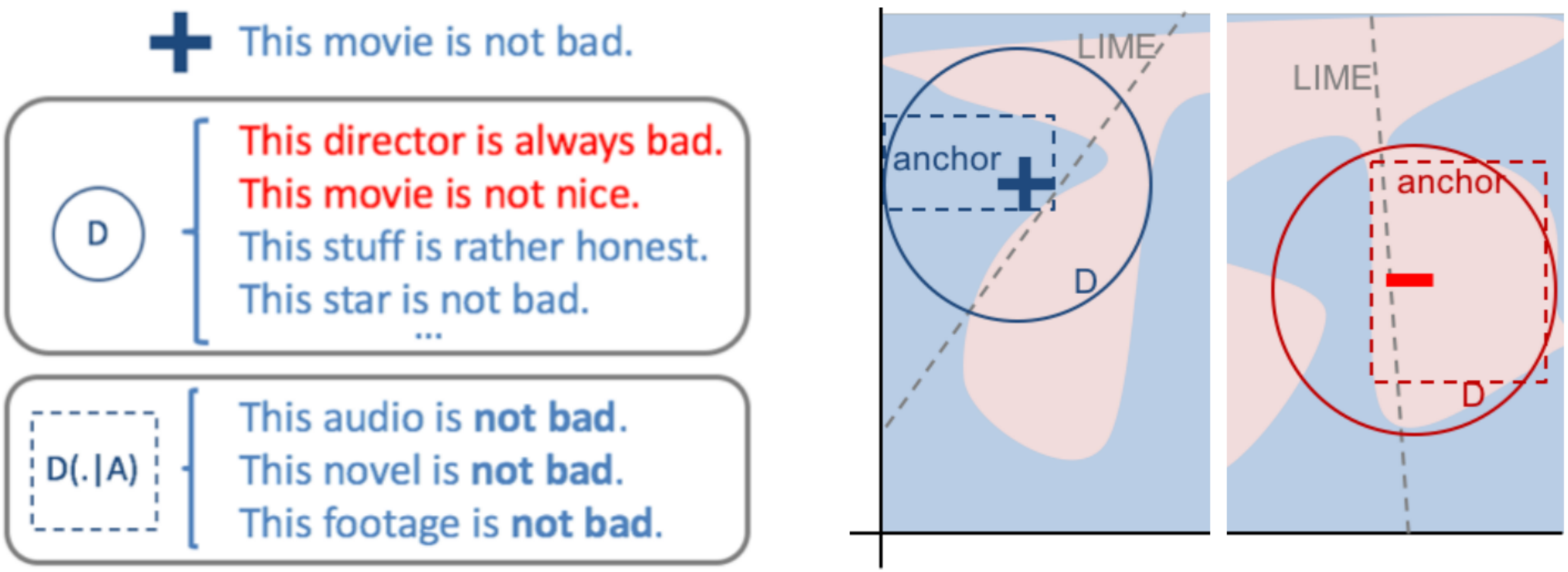
Explanation Technique: Anchors



(c) Anchor explanations

Ribeiro et al, AAAI 2018

Explanation Technique: Anchors



(a) $\mathcal{D}$ and $\mathcal{D}(.|A)$ (b) Two toy visualizations

$$\mathbb{E}_{\mathcal{D}(z|A)}[\mathbb{1}_{f(x)=f(z)}] \geq \tau,$$

Ribeiro et al, AAAI 2018

Explanation Technique: Anchors

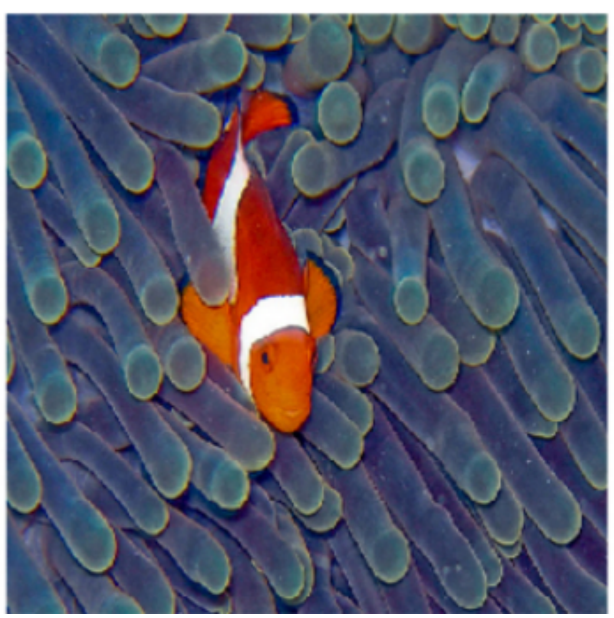
English	Portuguese
This is the **question** we must address	Esta é a questão que temos que enfrentar
This is the **problem** we must address	Este é o problema que temos que enfrentar
This is what we must address	É isso que temos de enfrentar

Table 2: Anchors (in bold) of a machine translation system for the Portuguese word for "This" (in pink).

Ribeiro et al, AAAI 2018

Explanation Technique: Influence Functions

- What would happen if a given training point didn't exist?
- Retraining the network is prohibitively slow, hence approximate the effect using influence functions.



Most influential train images

Koh & Liang, ICML 2017

Explanation Techniques: gradient based importance scores

Method	Attribution $R_i^c(x)$
Gradient * Input	$x_i \cdot \dfrac{\partial S_c(x)}{\partial x_i}$
Integrated Gradient	$(x_i - \bar{x_i}) \cdot \displaystyle\int_{\alpha=0}^{1} \dfrac{\partial S_c(\tilde{x})}{\partial(\tilde{x_i})}\Bigg\vert_{\tilde{x}=\bar{x}+\alpha(x-\bar{x})} d\alpha$
ϵ-LRP	$x_i \cdot \dfrac{\partial^g S_c(x)}{\partial x_i}, \quad g = \dfrac{f(z)}{z}$
DeepLIFT	$(x_i - \bar{x_i}) \cdot \dfrac{\partial^g S_c(x)}{\partial x_i}, \quad g = \dfrac{f(z) - f(\bar{z})}{z - \bar{z}}$

Example of attributions on MNIST

ReLU Tanh Sigmoid Softplus

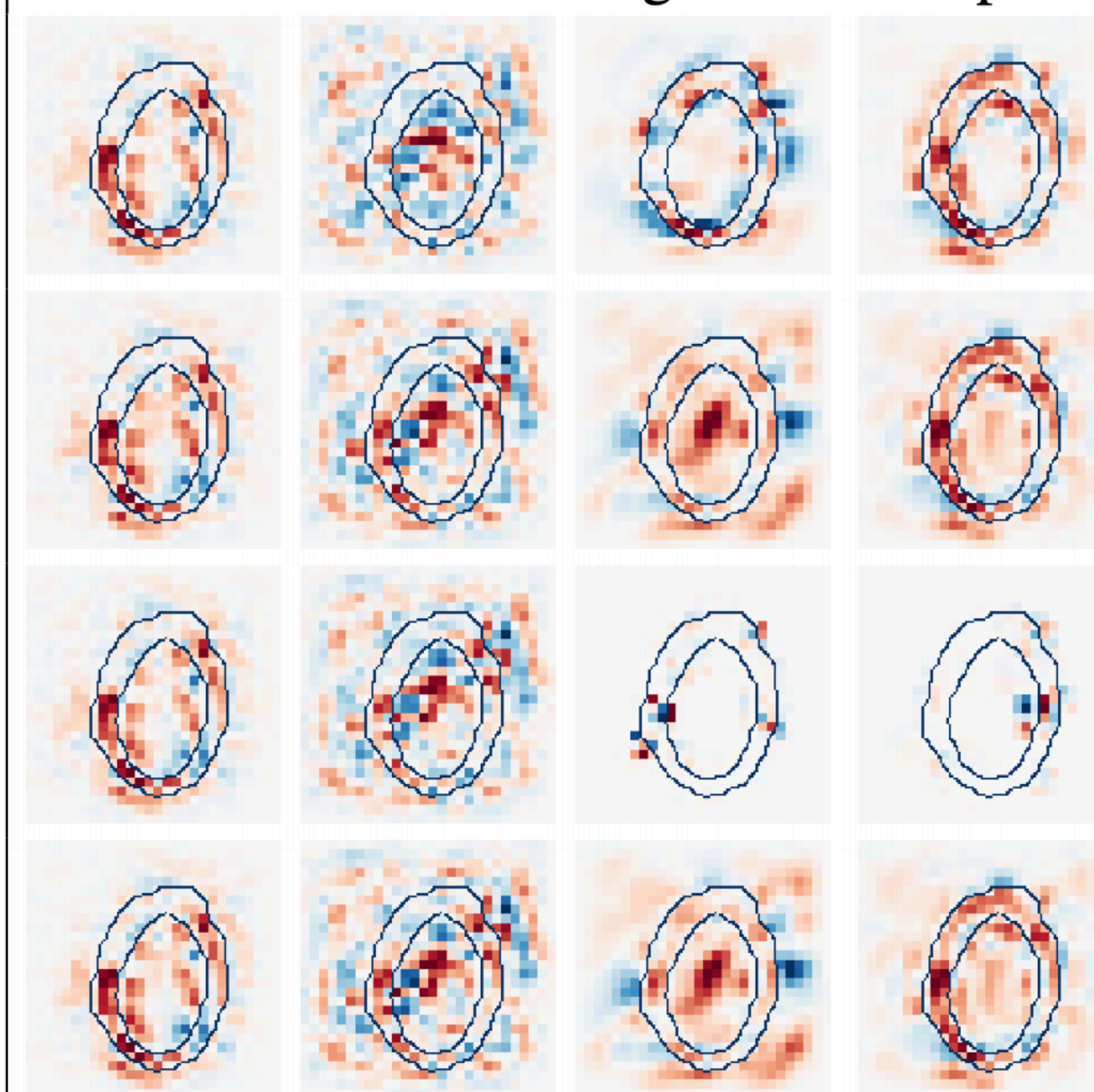


Figure from Ancona et al, ICLR 2018

Explanation Technique: Extractive Rationale Generation

Key idea: find minimal span(s) of text that can (by themselves) explain the prediction

- Generator (x) outputs a probability distribution of each word being the rational
- Encoder (x) predicts the output using the snippet of text x

- Regularization to support contiguous and minimal spans

Review

the beer was n't what i expected, and i'm not sure it's "true to style", but i thought it was delicious. **a very pleasant ruby red-amber color** with a relatively brilliant finish, but a limited amount of carbonation, from the look of it. aroma is what i think an amber ale should be - a nice blend of caramel and happiness bound together.

Ratings *Look*: 5 stars *Smell*: 4 stars

Figure 1: An example of a review with ranking in two categories. The rationale for Look prediction is shown in bold.

Future Directions

- Make the process of explanations interactive
  - Ask for details
  - What did you read (or see) to believe that
  - Contrastive explanations "Why X, why not Y"
- Complete the feedback loop: update the model based on explanations

Thank You!

Questions?