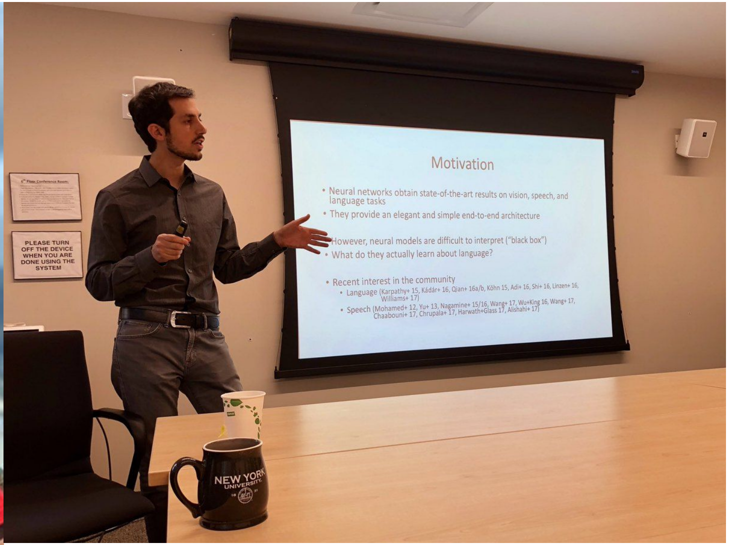
Analyzing Contextualized Representations and Individual Neurons in Deep NLP models

Nadir Durrani
Qatar Computing Research Institute

7th International Conference on Language and Technology

Core Team of NeuroX

Hassan Sajjad **Fahim Dalvi** **Nadir Durrani**



Motivation

- Neural networks obtain state-of-the-art results on vision, speech, and language tasks
- They provide an elegant and simple end-to-end architecture
- However, neural models are difficult to interpret ("black box")
- What do they actually learn about language?
- Recent interest in the community
  - Language (Karpathy+ 15, Kádár+ 16, Qian+ 16a/b, Köhn 15, Adi+ 16, Shi+ 16, Linzen+ 16, Williams+ 17)
  - Speech (Mohamed+ 12, Yu+ 13, Nagamine+ 15/16, Wang+ 17, Wu+King 16, Wang+ 17, Chaabouni+ 17, Chrupala+ 17, Harwath+Glass 17, Alishahi+ 17)

PLEASE TURN OFF THE DEVICE WHEN YOU ARE DONE USING THE SYSTEM
NEW YORK UNIVERSITY

Yonatan Belinkov
Collaborator at MIT

Motivation - Compared to Statistical MT

Word alignment

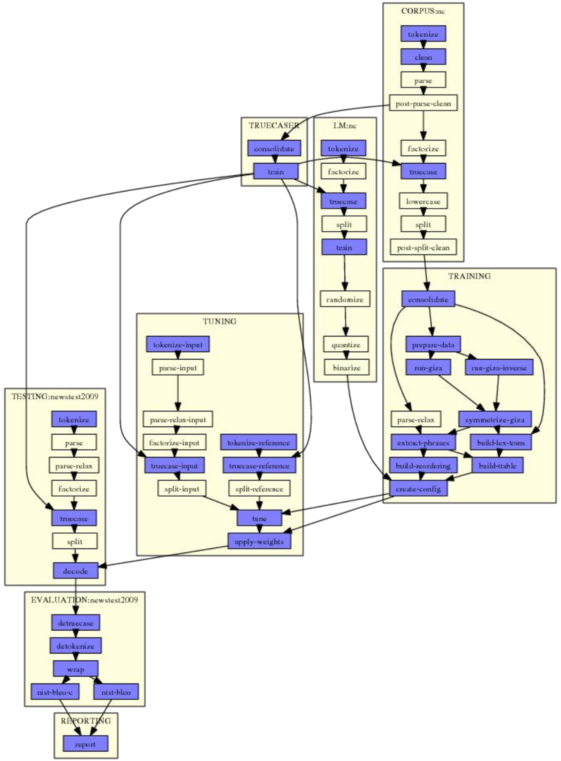
Reordering model

Lexical Model

Language Model

Tuning of features

...



Source: http://www.statmt.org/moses/?n=FactoredTraining.EMS

Motivation - Compared to Statistical MT

Word alignment

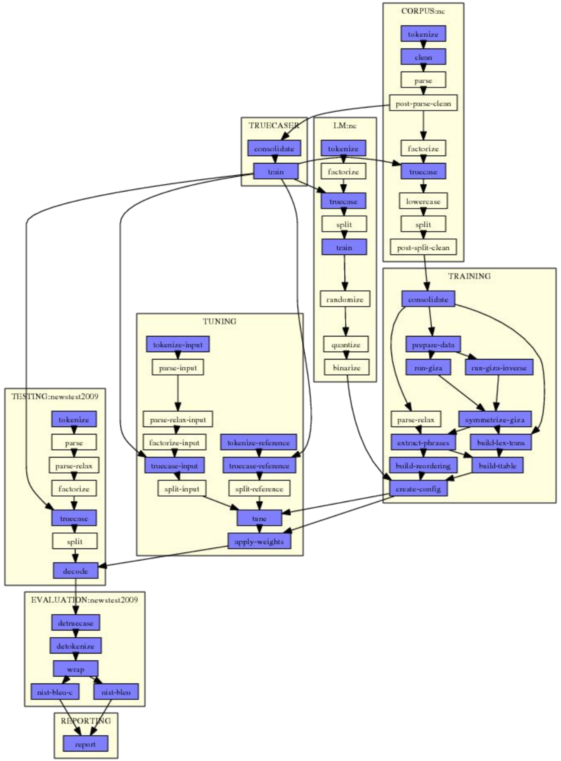
Reordering model

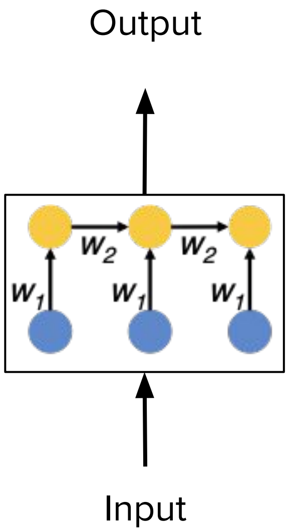
Lexical Model

Language Model

Tuning of features

...

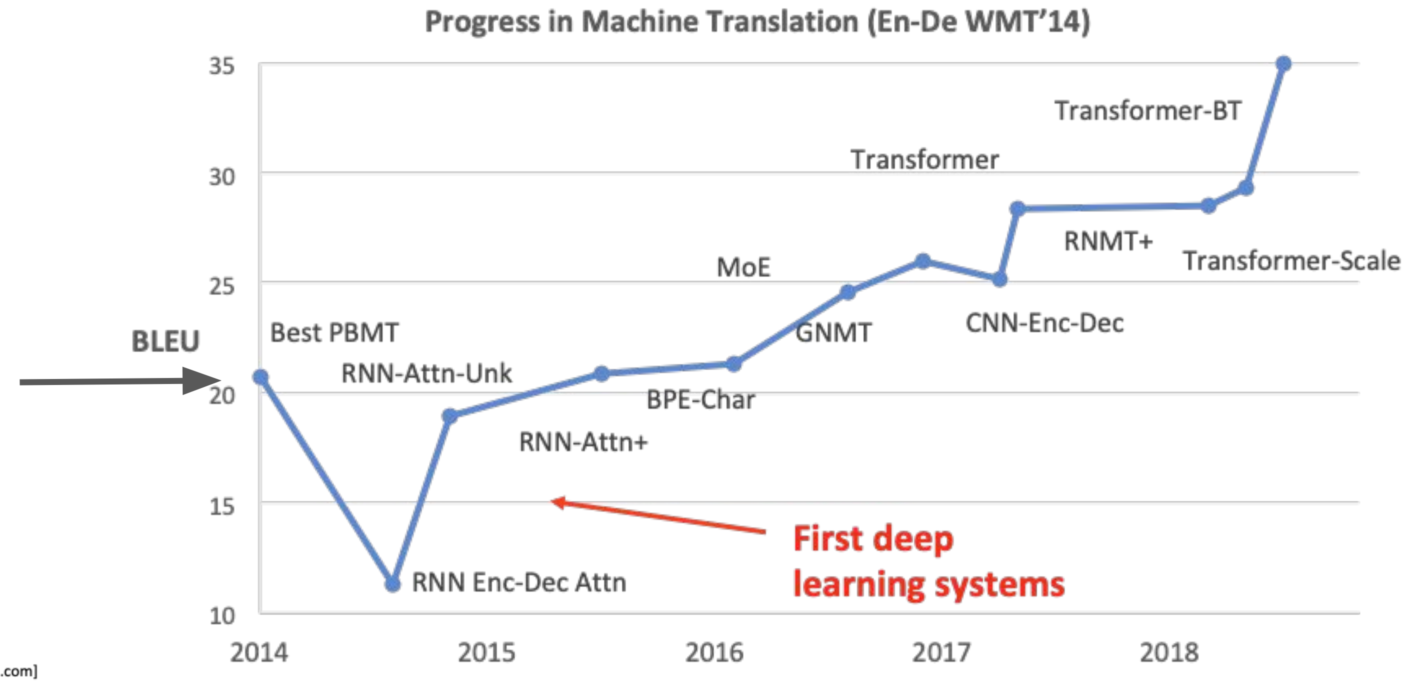




Source: http://www.statmt.org/moses/?n=FactoredTraining.EMS

Motivation

- Deep neural models: state-of-the-art for many tasks



First deep learning systems

[Sources: eff.org, nlpprogress.com]

Motivation - Compared to Statistical MT

Word alignment

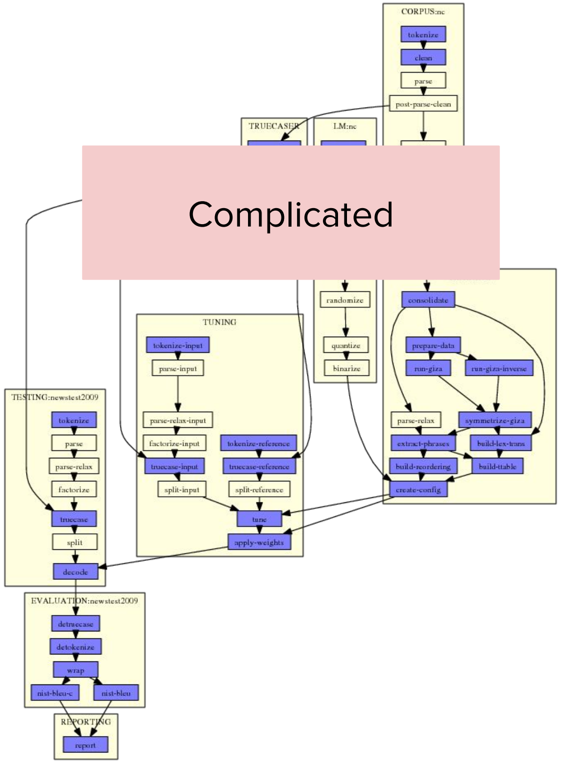
Reordering model

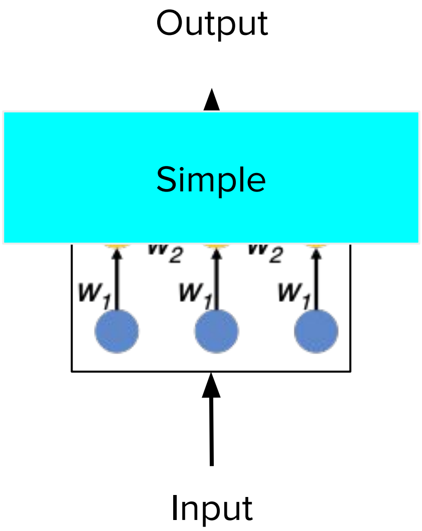
Phrase-table

Language Model

Tuning of features

...





Source: http://www.statmt.org/moses/?n=FactoredTraining.EMS

Motivation - Compared to Statistical MT

Word alignment

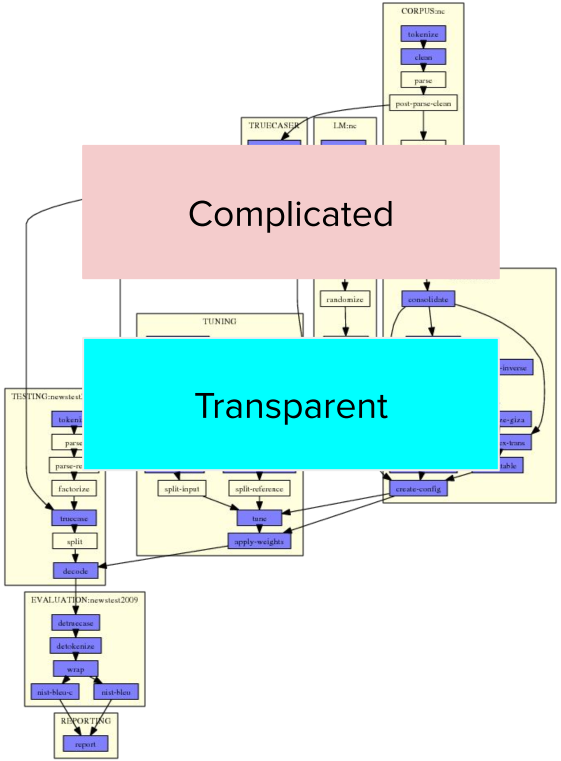
Reordering model

Phrase-table

Language Model

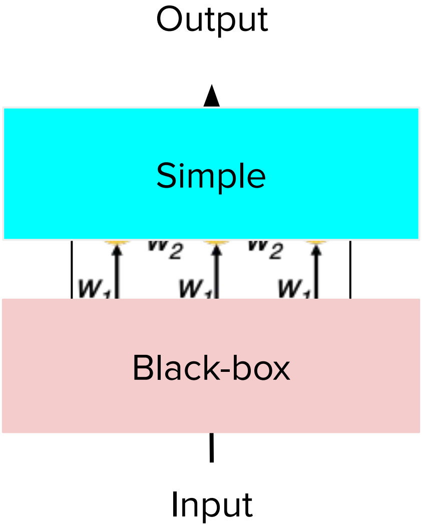
Tuning of features

...



Complicated

Transparent



Output

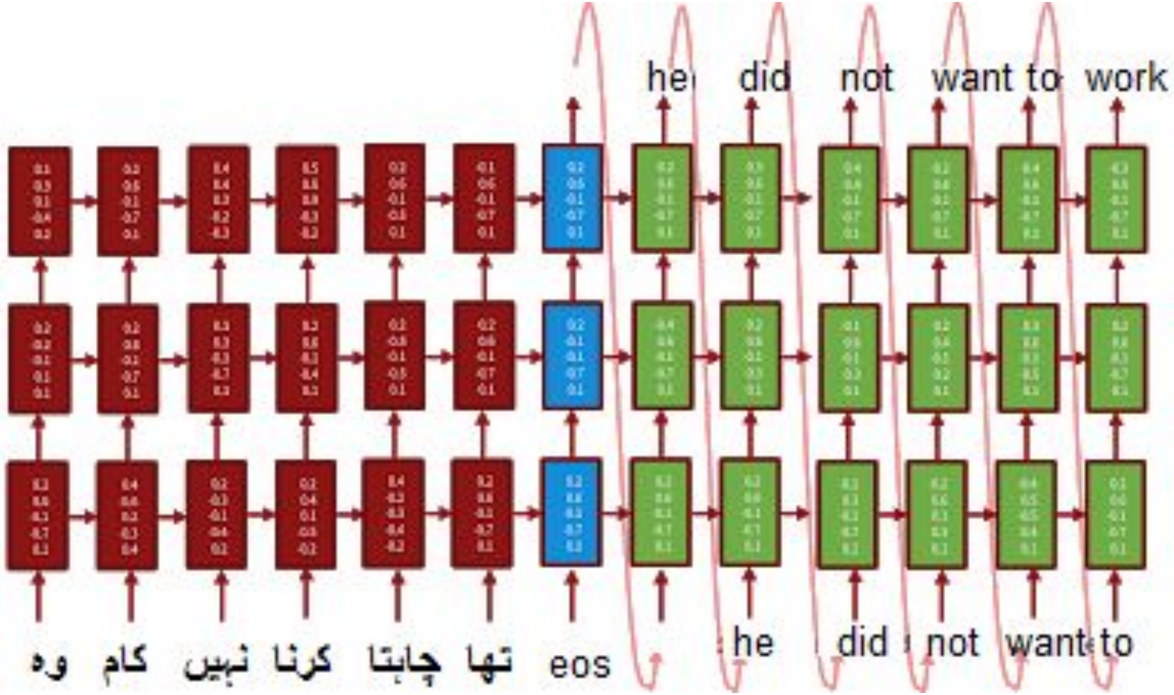
Simple

Black-box

Input

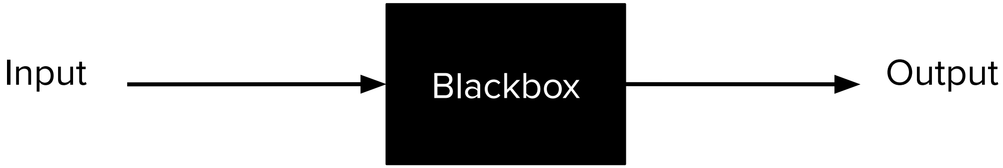
Source: http://www.statmt.org/moses/?n=FactoredTraining.EMS

Motivation


he
did
not
want to
work
وہ
کام
نہیں
کرنا
چاہتا
تھا
eos
he
did
not
want
to

Motivation

- Deep neural models: state-of-the-art for many tasks

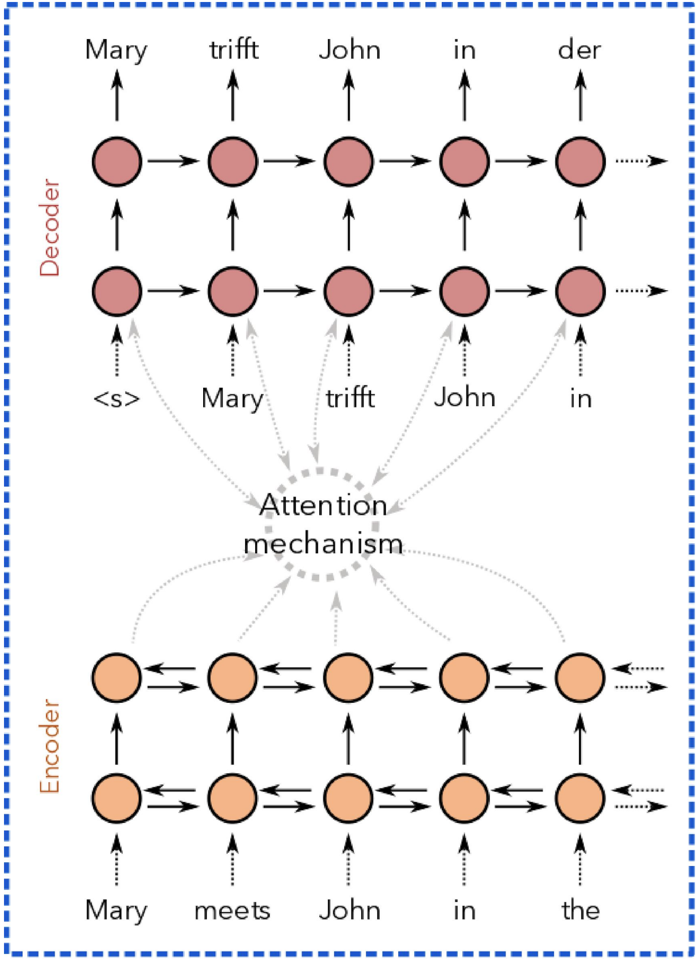


- Issue: opaqueness

- Interpretation is important
  - Debugging
  - Better understanding
  - Increasing trust in AI systems
  - Assisting ethical decision making
  - ...

In this talk ...

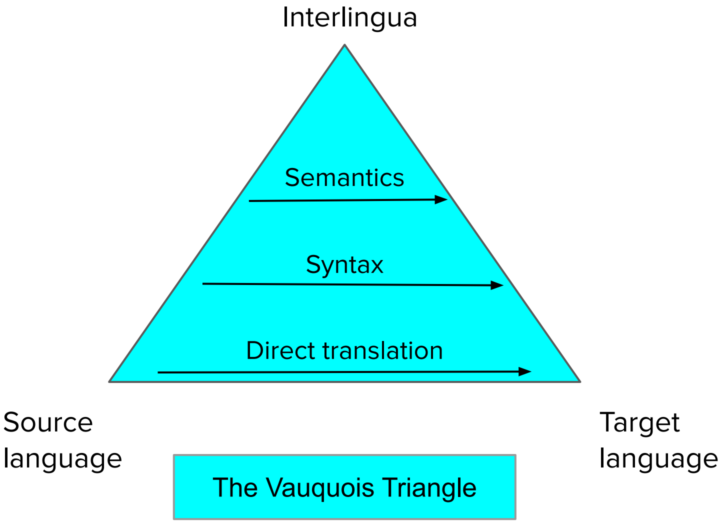
- Analyze representations
  - Model and Layer-wise
  - Neuron-level

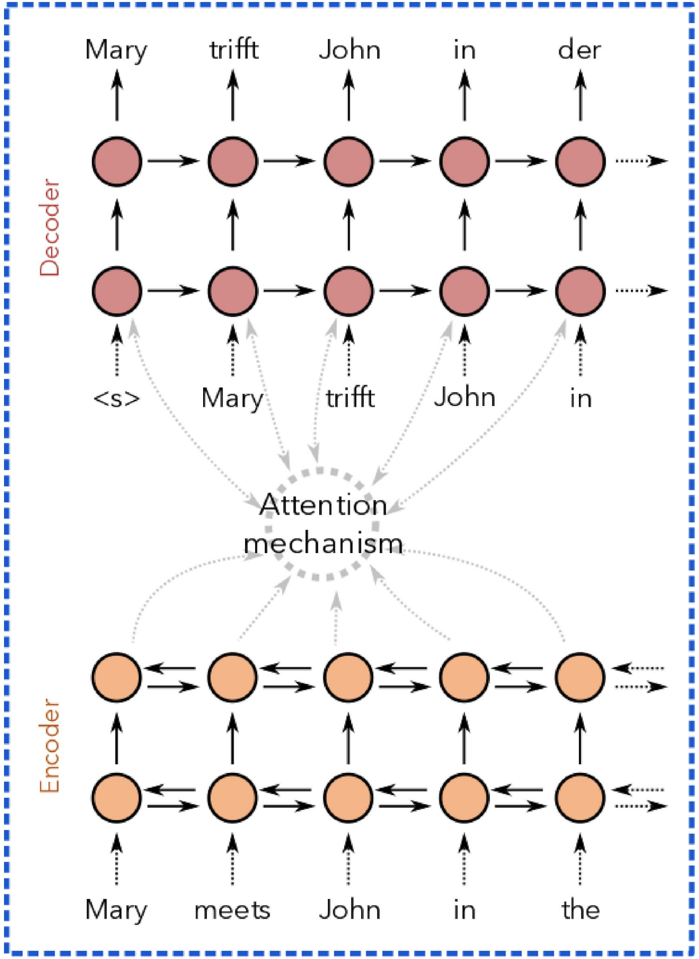


Pretrained Neural Machine Translation Model

In this talk ...

- Analyze core language properties





Pretrained Neural Machine Translation Model

Part of Speech and Morphological Tagging

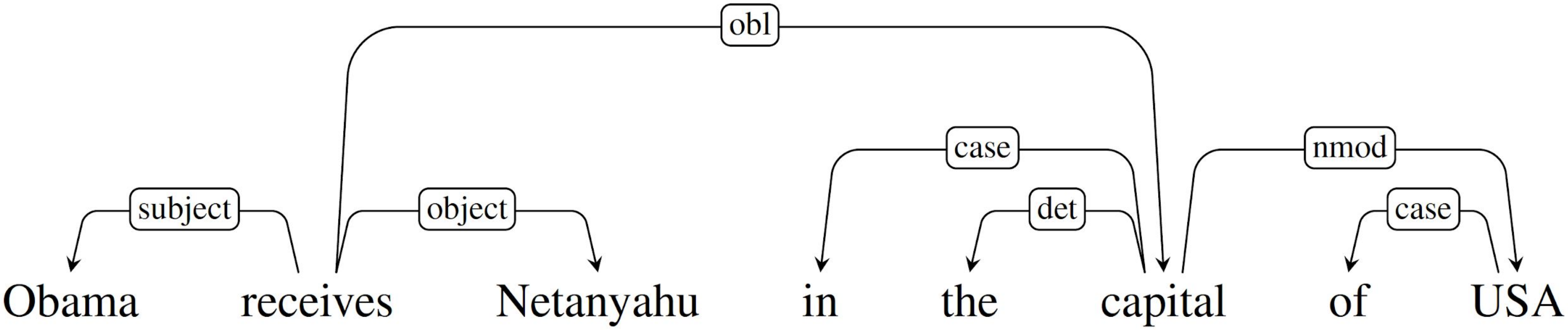
- Word-level annotations
- Rich morphological tags for a few languages

Sentence	Obama	receives	Netanyahu	in	the	capital	of	USA
POS	NP	VBZ	NP	IN	DT	NN	IN	NP

Sentence	Obama	empfäangt	Netanyahu	in	der	Hauptstadt	der	USA
Morph	nn.nom.seg.neut	vvfin.3.sg.pres.ind	ne.nom.sg.*	appr.-	art.dat.sg.fem	nn.dat.sg.fem	art.gen.pl.*	nn.gen.pl.*

Syntactic Relations

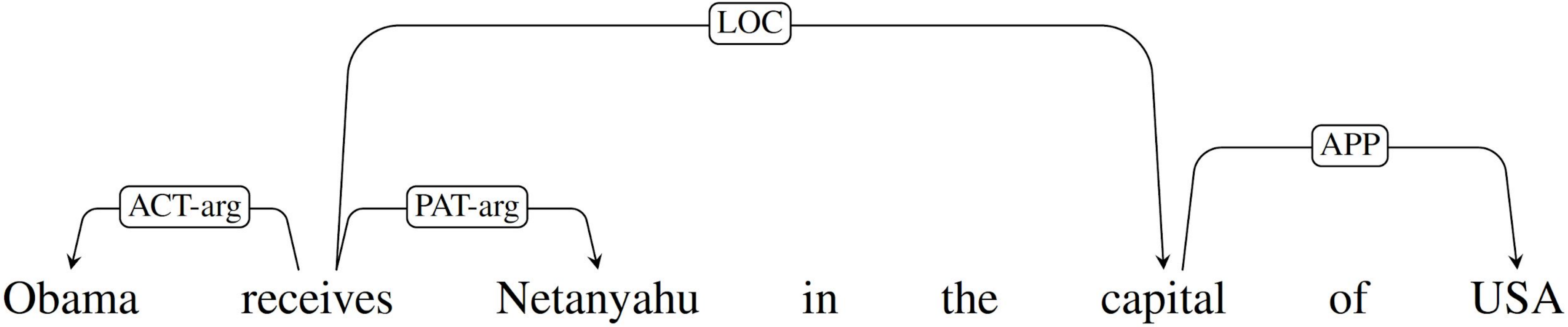
- Predict relation between two words



https://universaldependencies.org/guidelines.html

Semantic Dependency Relations

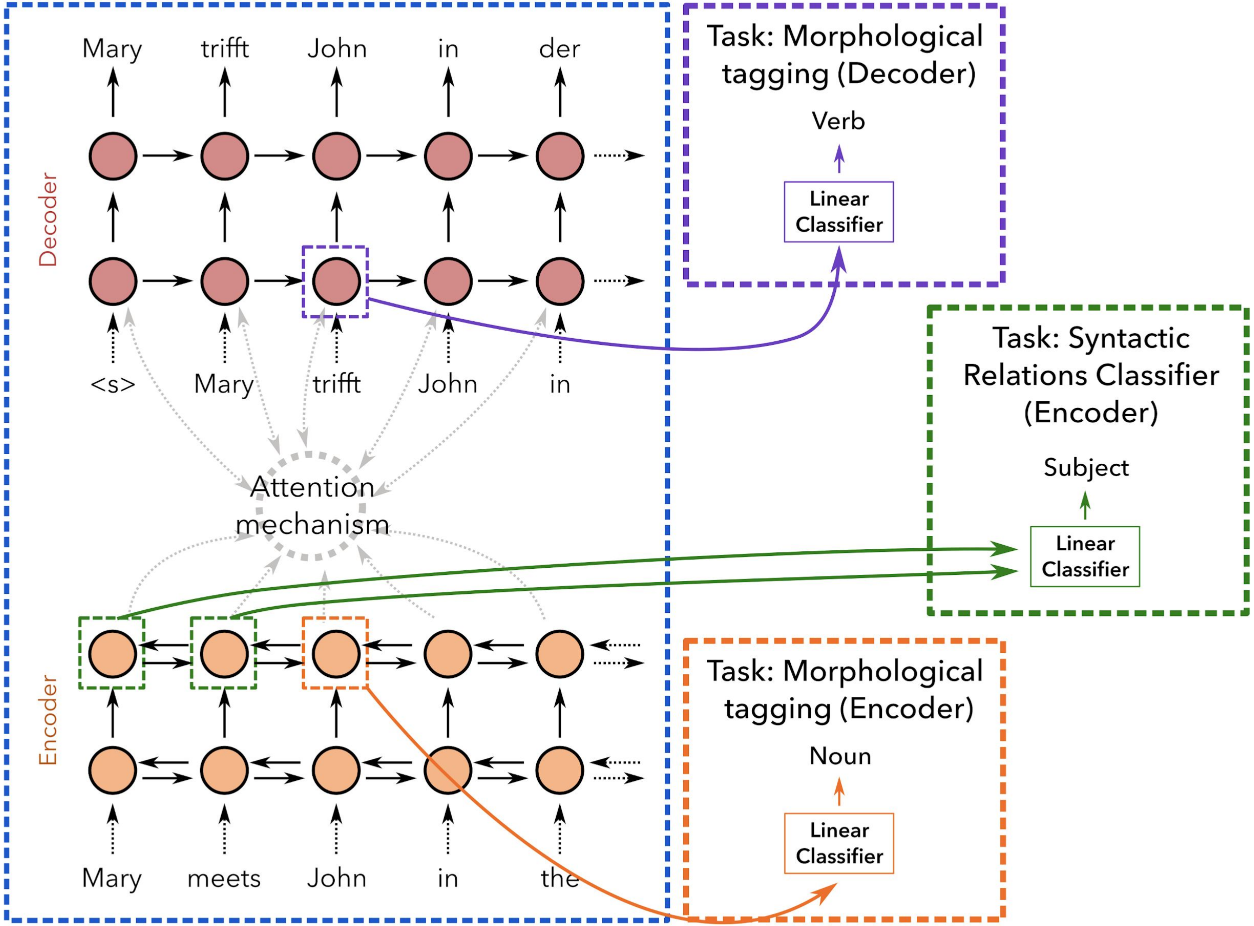
- Predict relationship between two words



For details on the tagging, see Cinková et al. 2004

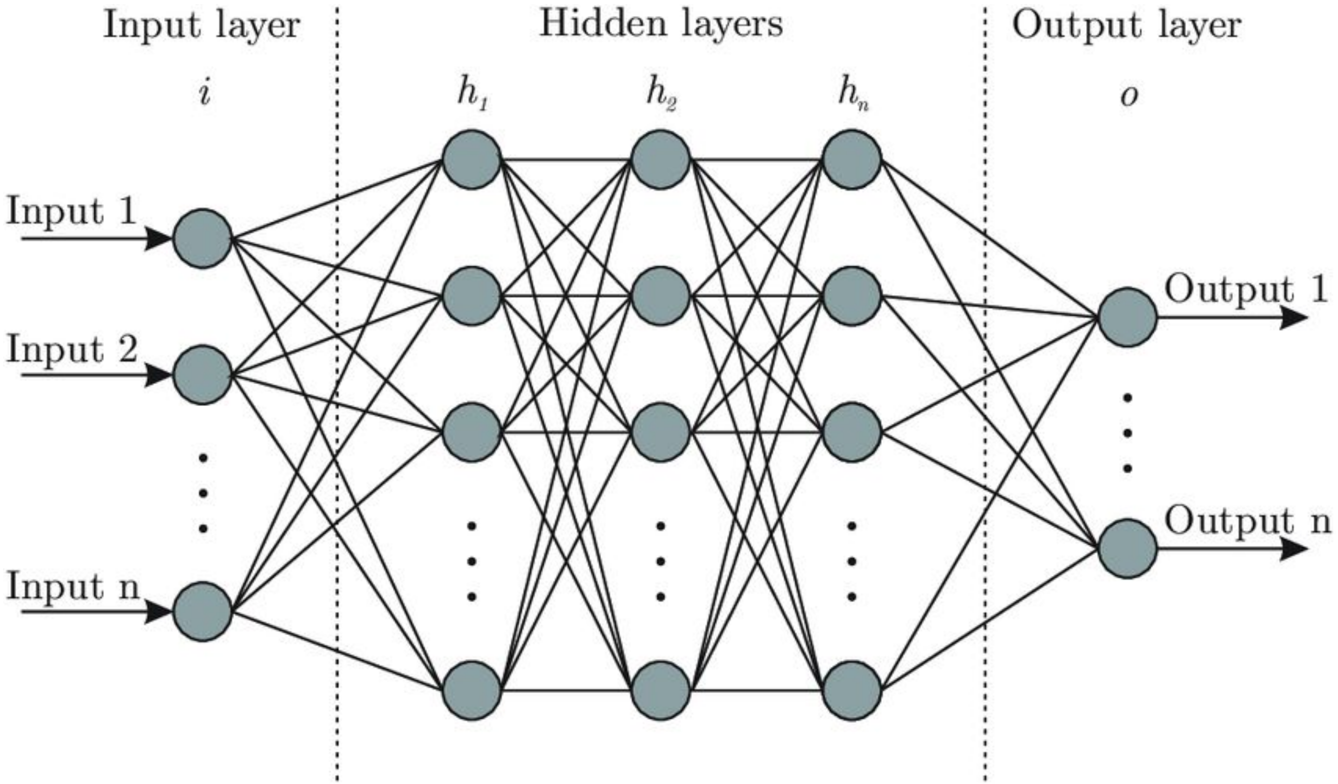
Methodology

Methodology

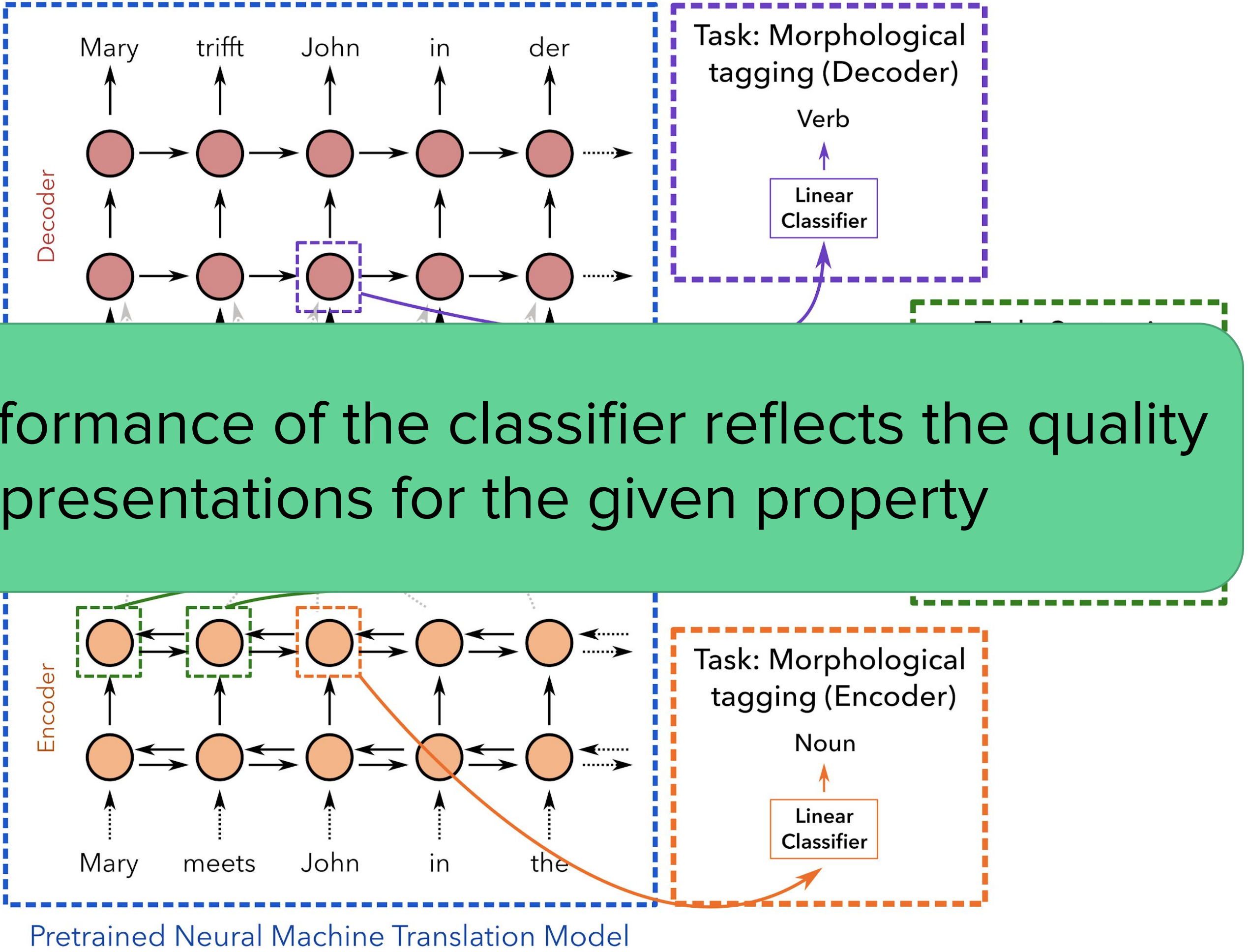


Pretrained Neural Machine Translation Model

Neural Activations



Methodology


Mary
trifft
John
in
der
Decoder
Task: Morphological tagging (Decoder)
Verb
Linear Classifier
Encoder
Mary
meets
John
in
the
Task: Morphological tagging (Encoder)
Noun
Linear Classifier
Pretrained Neural Machine Translation Model

Assumption: Performance of the classifier reflects the quality of the representations for the given property

Methodology

Formally,

- Let $x = \{x_1, x_2, \ldots, x_n\}$ denotes a sequence of input features
- $\mathbb{M}$ is a neural network model
- $\mathbb{M}$ maps input x to a sequence of latent representations $z = \{z_1, z_2, \ldots, z_n\}$

Methodology

Formally,

- Let $x = \{x_1, x_2, \ldots, x_n\}$ denotes a sequence of input features
- $\mathbb{M}$ is a neural network model
- $\mathbb{M}$ maps input x to a sequence of latent representations $z = \{z_1, z_2, \ldots, z_n\}$

- Consider a **classification task** that predicts a property $\mathbf{l}$ in a property set $\mathcal{P}$ that we believe is intrinsically learned in the model $\mathbb{M}$
- We assume that the supervision is available as $\{x_i, \mathbf{l}_i\}$ where x_i is the input word and $\mathbf{l}_i$ is its label

Methodology

- Logistic regression classifier on the $\{z_i, \mathbf{l}_i\}$ pair
  - Linear model for better explainability

- Minimize negative log likelihood of the training data

$$\mathcal{L}(\theta) = -\sum_i \log P_\theta(\mathbf{l}_i | x_i)$$

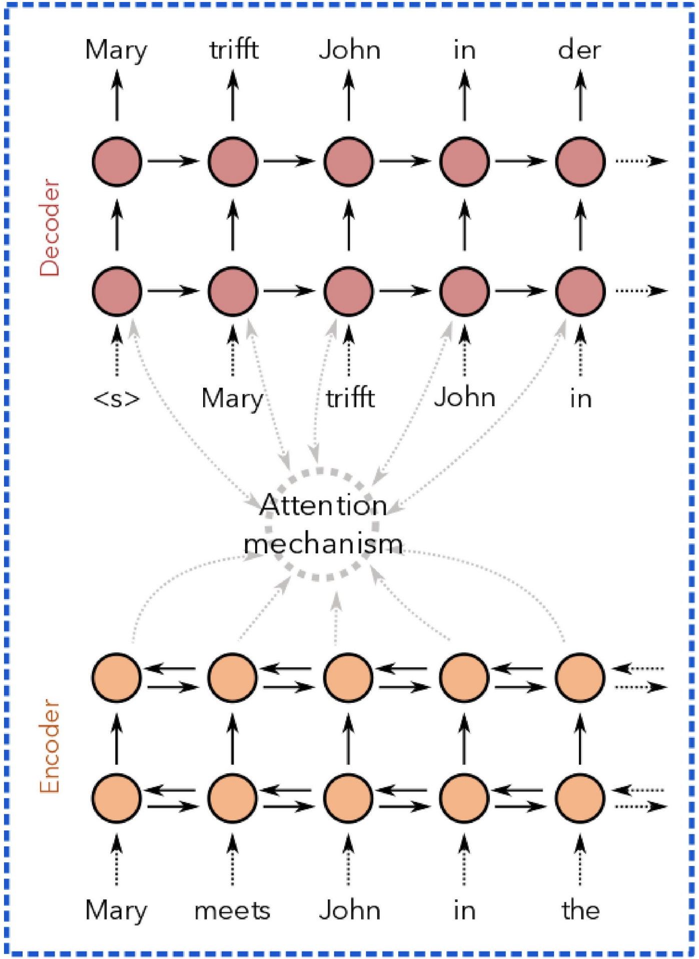
- Performance of the classifier reflects the quality of the representation with respect to the property

Experimental Setup

- Sequence to sequence with attention mechanism
    - Bi-directional LSTM
    - 2 layers and 4 layers models
- English to/from German French, Spanish, Czech, Arabic, Hebrew
- NMT Training Data - WMT, IWSLT, UN corpora
- Linguistic properties
    - Morphology
    - Semantics, Syntax (concatenate representations of words)

Questions

- Linguistic information
    - What, Where and How much
    - Effect of training choices
    - Effect of different granularities
- Role of individual neurons
- Focused vs. distributed
- ...



Pretrained Neural Machine Translation Model

Model-level Analysis

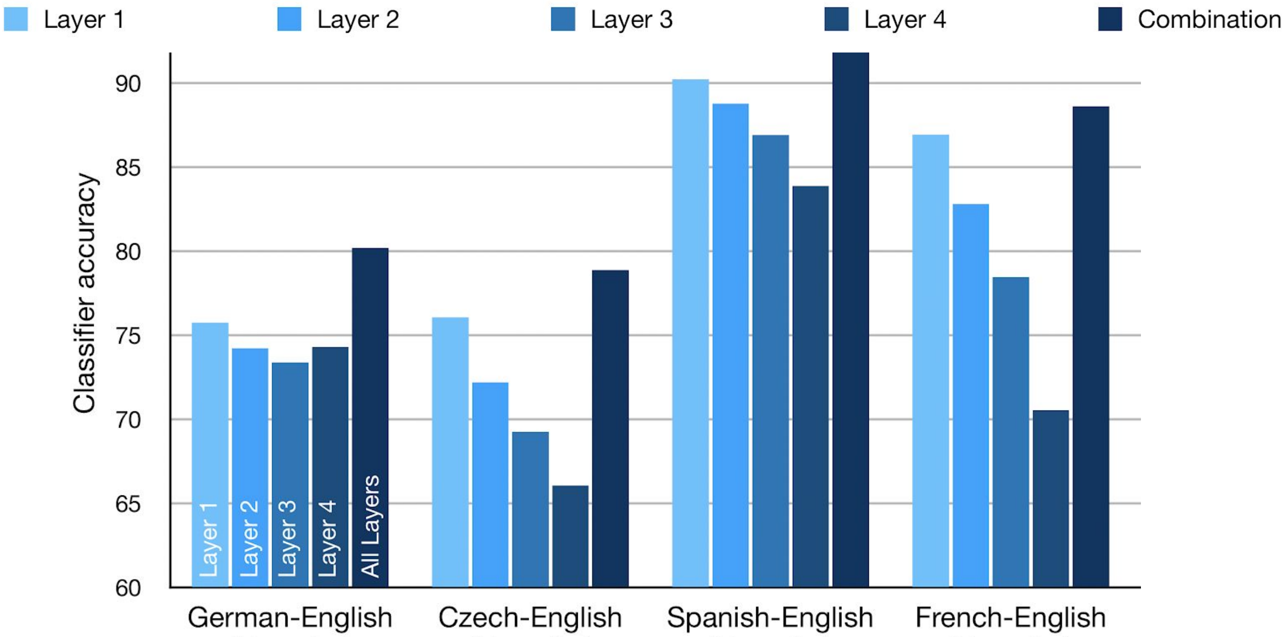
- Overall performance on the auxiliary tasks
- Average performance across several languages

	POS	Morphological	Syntax	Semantics
Majority	90.4	74.6	67.3	84.2
MT Classifier	95.4	85.4	89.2	91.4
Task-specific Classifier	96.6	91.6	-	-

Findings: Morphological Learning

- **Layer-wise** learning

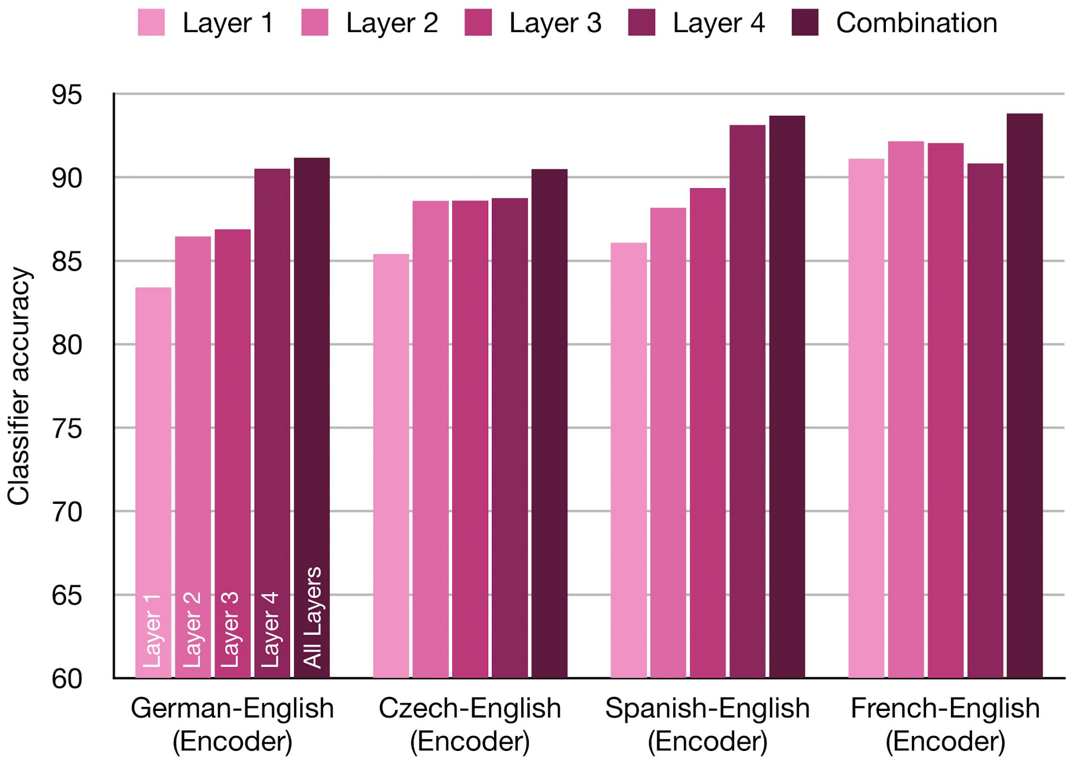
Layer 1 learns the most about the morphology but information is distributed



Findings: Syntactic Learning

- **Layer-wise** learning

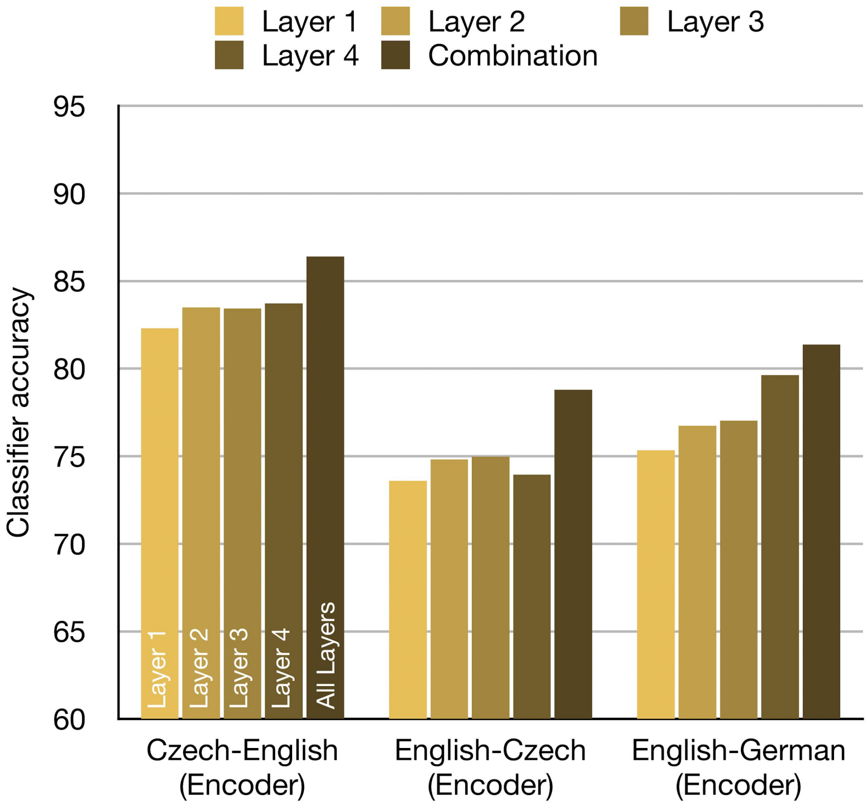
Higher layers learn more about syntax!



Findings: Semantic Learning

- **Layer-wise** learning

Higher layers learn more semantics!



Comparing systems trained using different granularities

- Various representation units because of
  - Vocabulary reduction
  - Unknown word problem
  - Morphological segmentation

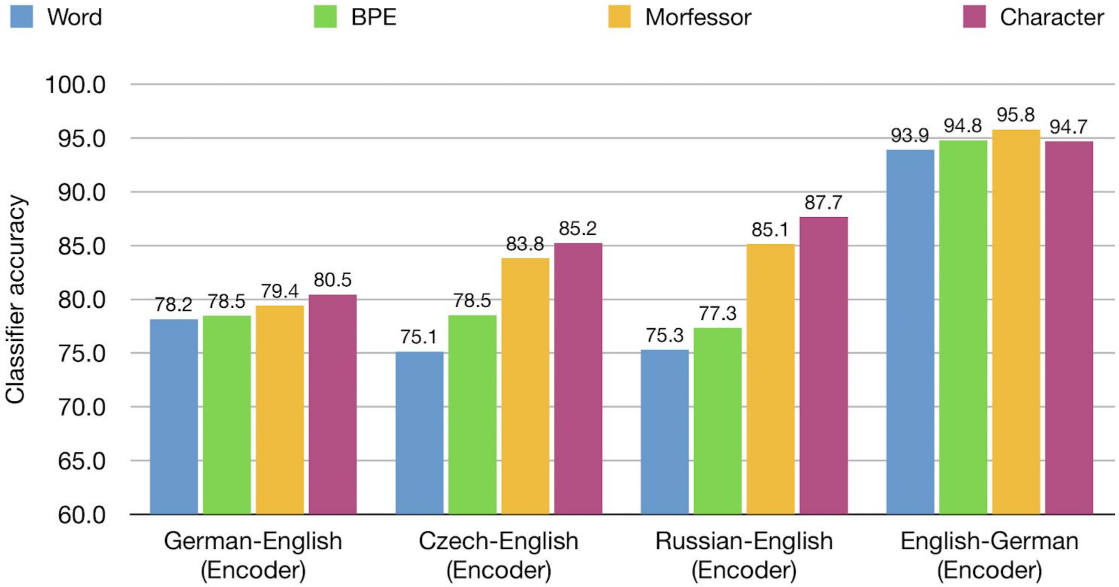
Words	Professor admits to shooting girlfriend
BPE	Professor admits to sho@@ oting gir@@ l@@ friend
Morfessor	Professor admit@@ s to shoot@@ ing girl@@ friend
Characters	P r o f e s s o r _ a d m i t s _ t o _ s h o o t i n g _ g i r l f r i e n d

Morphological Learning

- Comparing **input representations**

Character representations are **better at learning Morphology**

Character representations are **more robust towards noise**

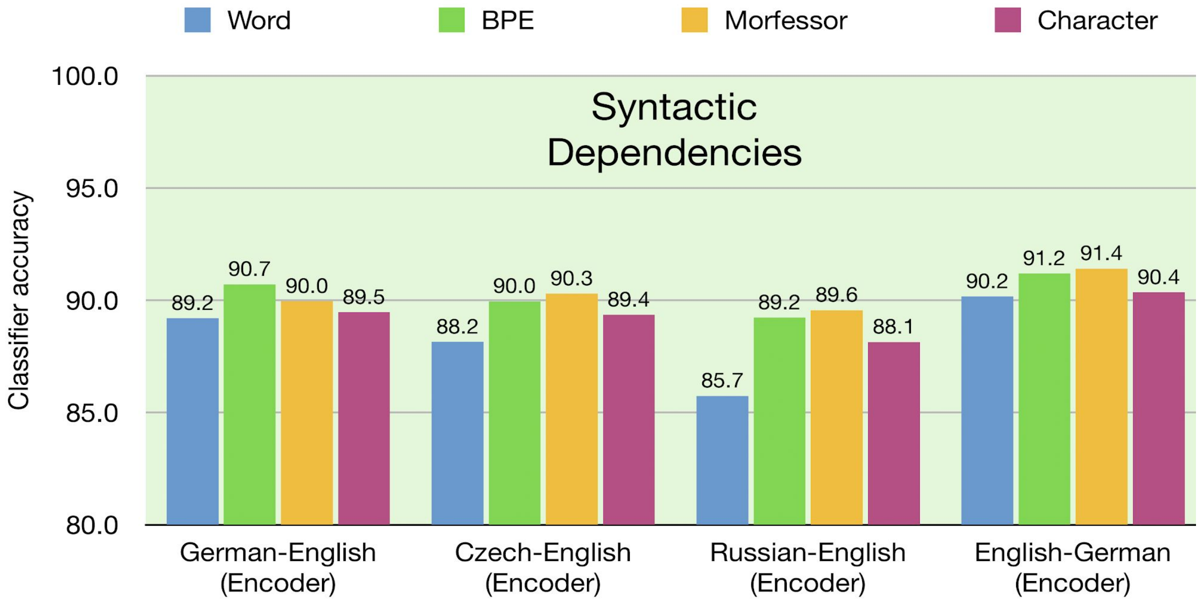


Syntactic Learning

- Comparing **input representations**

Character-based representations are **worse** in **learning syntax**

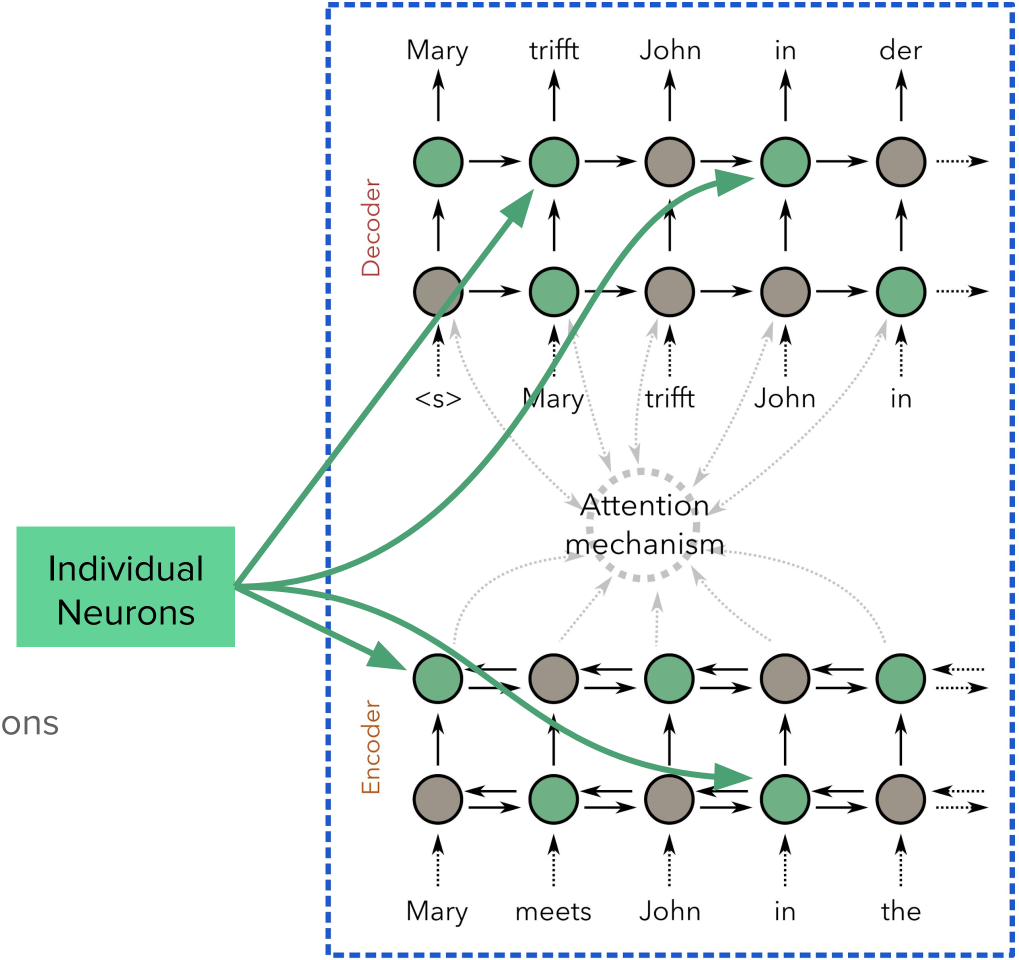
Character-based: suffer from **long-range dependency**



Neuron-level Analysis

Individual Neurons

- What is the role of individual neurons?
- Several open questions
  - Learning pattern
  - Representation of information
  - Role of individual neurons
  - Important vs. less important neurons
  - ...



Pretrained Neural Machine Translation Model

Individual Neurons

We propose two methods:

- Linguistic correlation
- Cross-model correlation

Use Case

- Controlling model behavior
- Feature selection and model distillation
    - Useful in transfer learning

Linguistic Correlation Analysis

- Goal: Identify neurons with respect to a property
  - Parts of speech properties like noun, verb, adjective
  - Semantic properties
  - Month of year, position in a sentence
  - ...

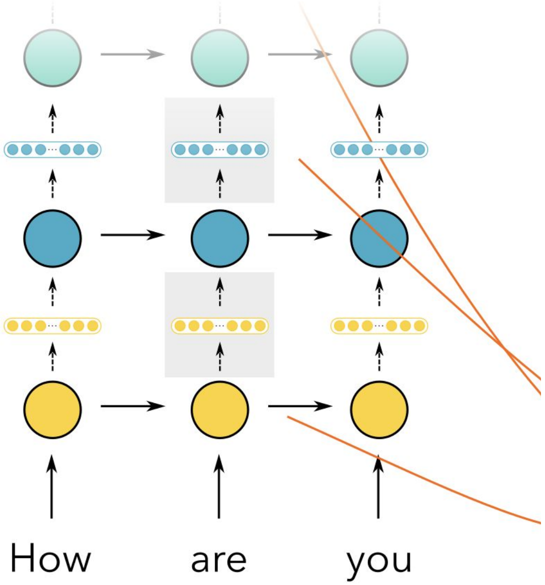
Linguistic Correlation Analysis

- Goal: Identify neurons with respect to a property
  - Parts of speech properties like noun, verb, adjective
  - Semantic properties
  - Month of year, position in a sentence
  - …

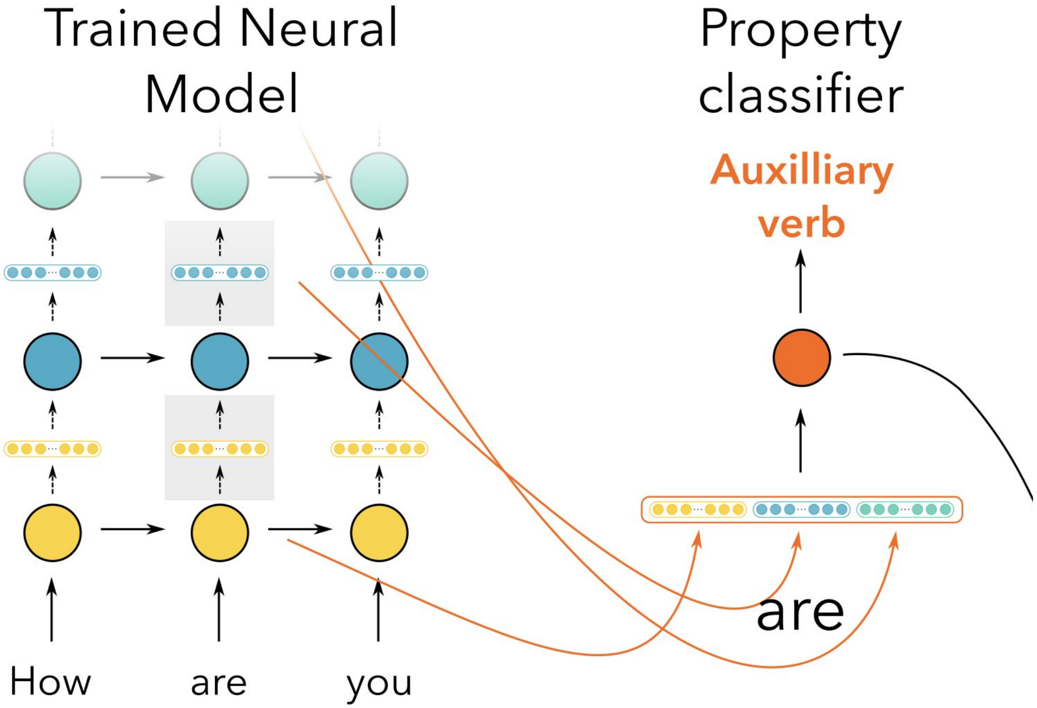
- Extrinsic supervised classification
- Extract important neurons that capture a given property

Methodology

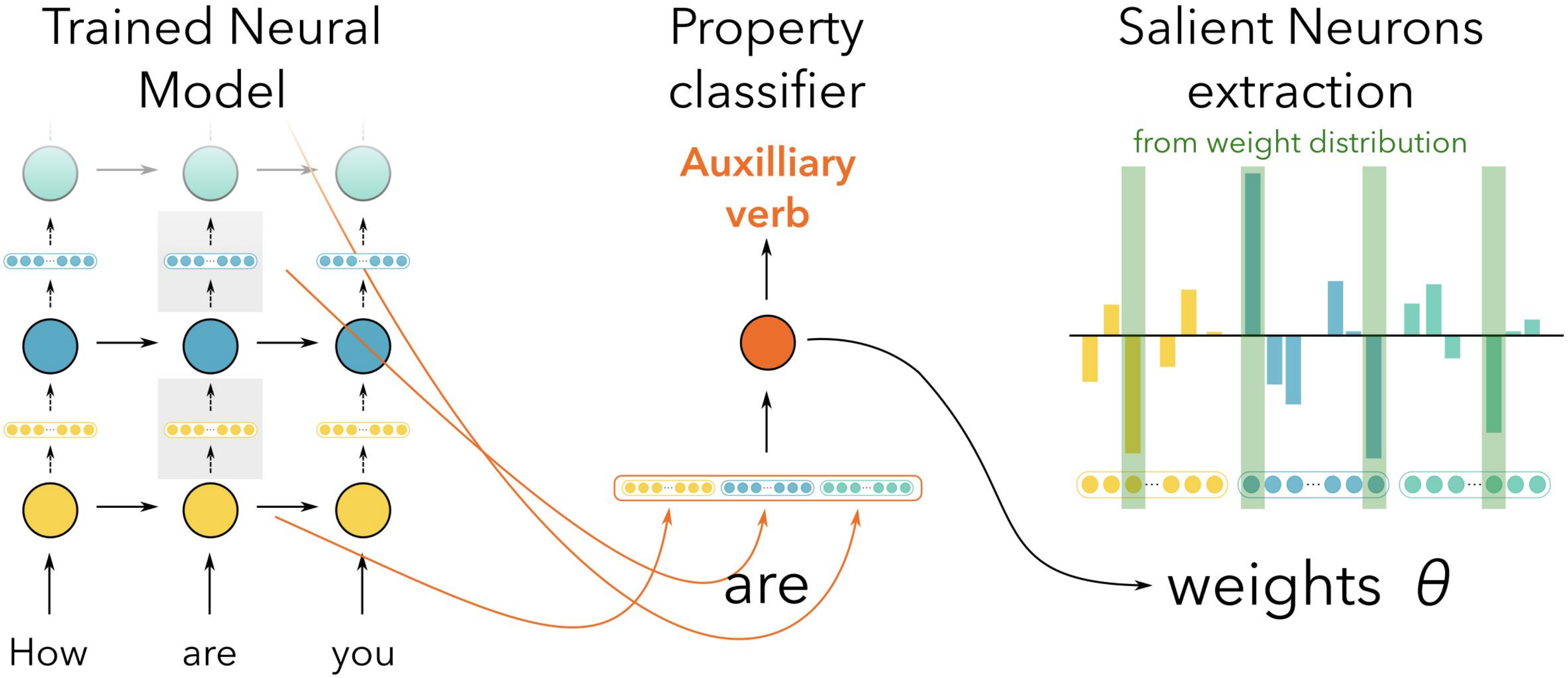
Trained Neural Model



Methodology



Methodology



Methodology

- Logistic regression classifier on the $\{z_i, \mathbf{l}_i\}$ pair
  - Linear model for better explainability

- Learned weights: Measure of the importance of each neuron z_i

- To encourage feature ranking: use **elastic net regularization**

$$\mathcal{L}(\theta) = -\sum_i \log P_\theta(\mathbf{l}_i|x_i) + \lambda_1 \|\theta\|_1 + \lambda_2 \|\theta\|_2^2$$

Methodology

- Choice of using elastic net is critical to identify both focused and distributed neurons
- The lasso regularization part of elastic net brings in focused neurons
- The ridge regularization part of elastic net brings in group of correlated neurons
- Elastic net strikes a good balance between localization and distributivity

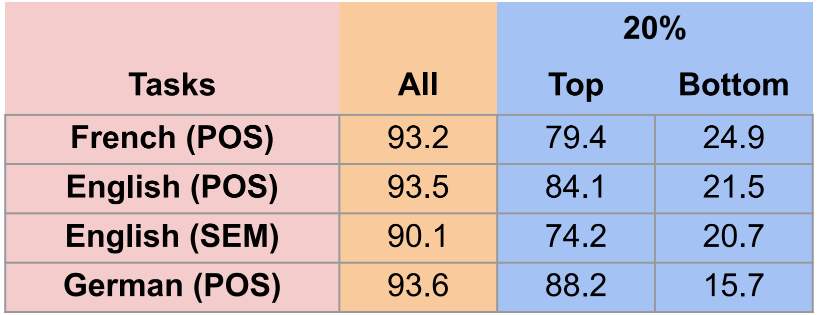
Evaluation - Ablation in Classifier

- How good are the rankings?

Evaluation - Ablation in Classifier

- How good are the rankings?

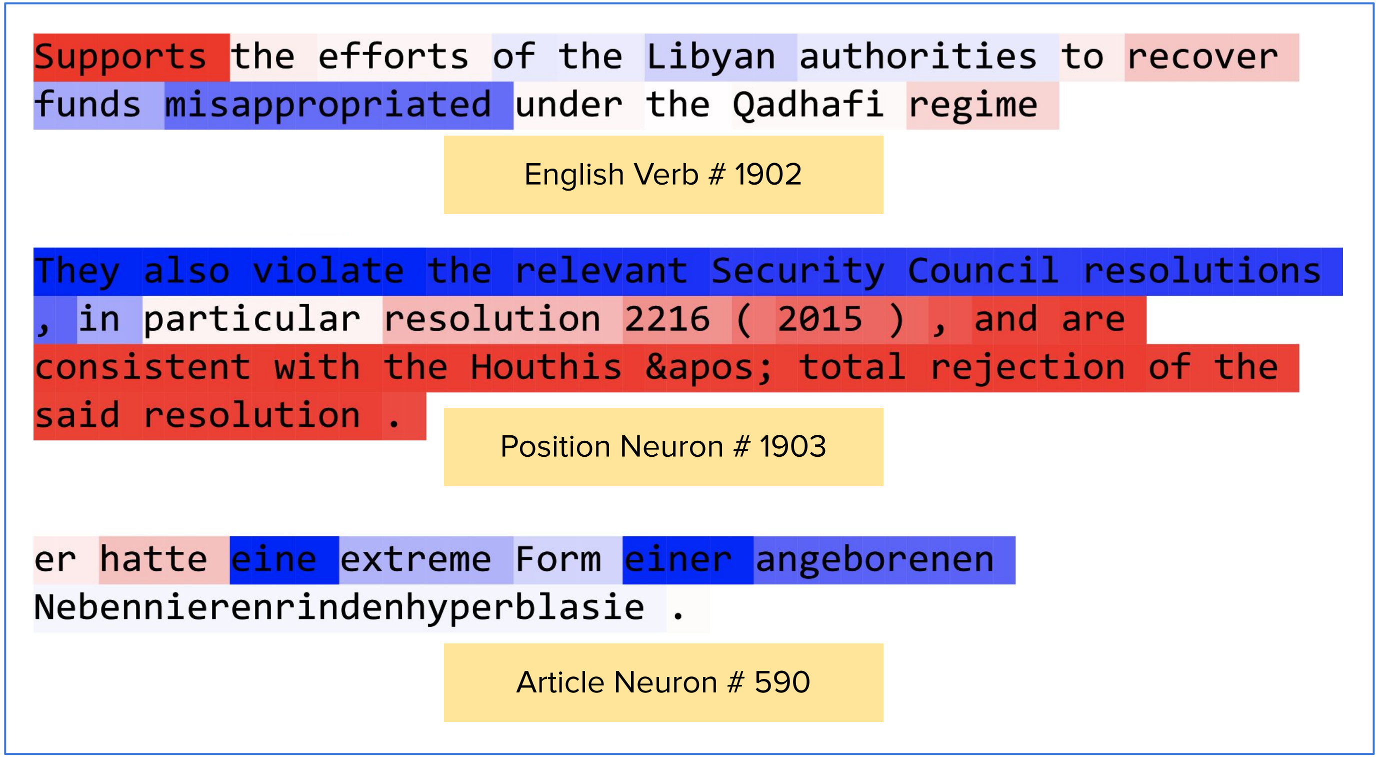
- Keep top/bottom N% neurons
- Accuracy: top N% vs. bottom N% neurons



		20%	
Tasks	**All**	**Top**	**Bottom**
French (POS)	93.2	79.4	24.9
English (POS)	93.5	84.1	21.5
English (SEM)	90.1	74.2	20.7
German (POS)	93.6	88.2	15.7

- Part-of-speech (POS) tagging and semantic (SEM) tagging

Visualization - Top Neurons



Supports the efforts of the Libyan authorities to recover funds misappropriated under the Qadhafi regime

English Verb # 1902

They also violate the relevant Security Council resolutions , in particular resolution 2216 (2015) , and are consistent with the Houthis &apos; total rejection of the said resolution .

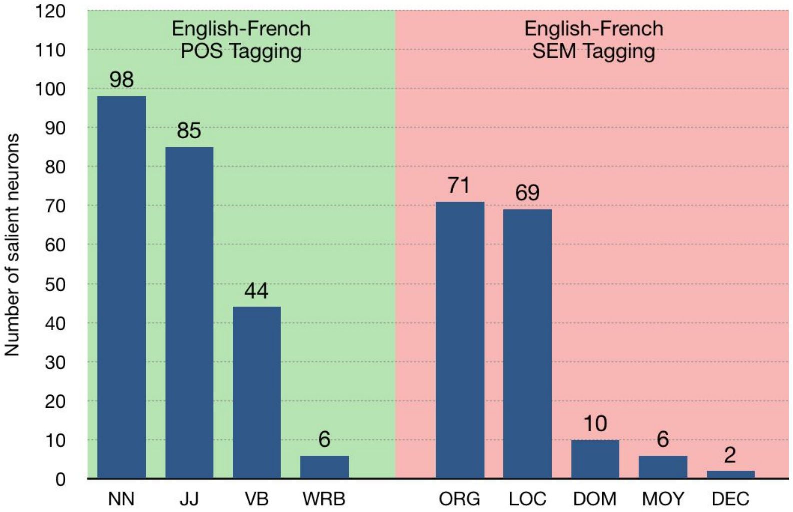
Position Neuron # 1903

er hatte eine extreme Form einer angeborenen Nebennierenrindenhyperblasie .

Article Neuron # 590

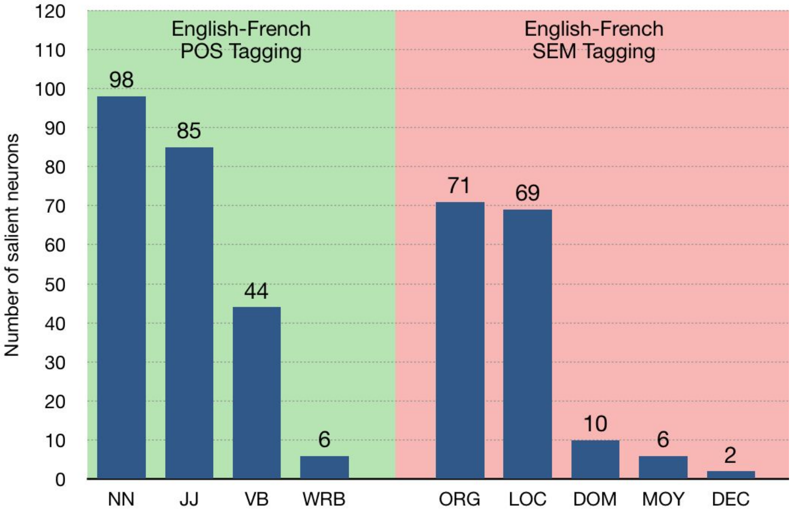
Focused vs. Distributed Neurons

- The open class categories are distributed
- The closed class categories are focused



Focused vs. Distributed Neurons

- The open class categories are distributed
- The closed class categories are focused



Neuron	Top 10 Words
#1925	August, July, January, September, October, presidential, April, May, February, December
#1960	no, No, not, nothing, nor, neither, or, none, whether, appeal
#1590	50, 10, 51, 61, 47, 37, 48, 33, 43, 49

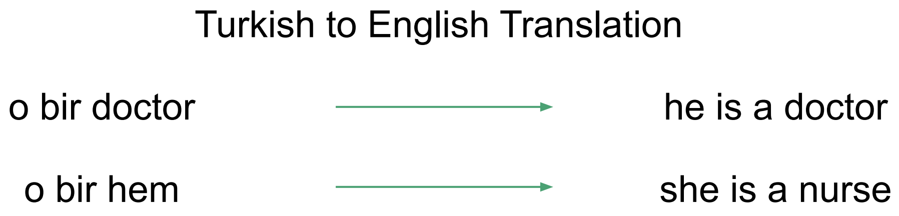
Controlling Systems’ Behavior

Controlling Systems’ Behavior

- Neurons responsible for specific properties

Can we use this information to control models?

- Benefit: Mitigating bias in models, e.g. gender bias



Turkish to English Translation

o bir doctor → he is a doctor

o bir hem → she is a nurse

mashable.com/2017/11/30/google-translate-sexism

Controlling Systems’ Behavior

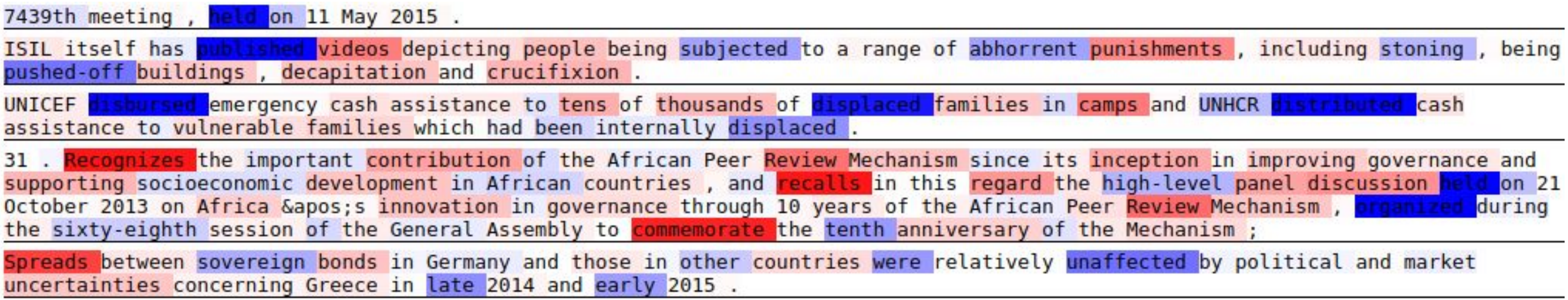
- Intervene in neuron activations at test time

Process

- Identify neuron(s) with respect to a property
- At test time, encode the source sentence as usual
- Set the activation of a particular neuron(s) in the encoder state to α
- α is a function of mean activations over a property
- Experimented with gender, number and tense

Controlling Translations

For example, consider the top neuron of verb past tense



7439th meeting , held on 11 May 2015 .

ISIL itself has published videos depicting people being subjected to a range of abhorrent punishments , including stoning , being pushed-off buildings , decapitation and crucifixion .

UNICEF disbursed emergency cash assistance to tens of thousands of displaced families in camps and UNHCR distributed cash assistance to vulnerable families which had been internally displaced .

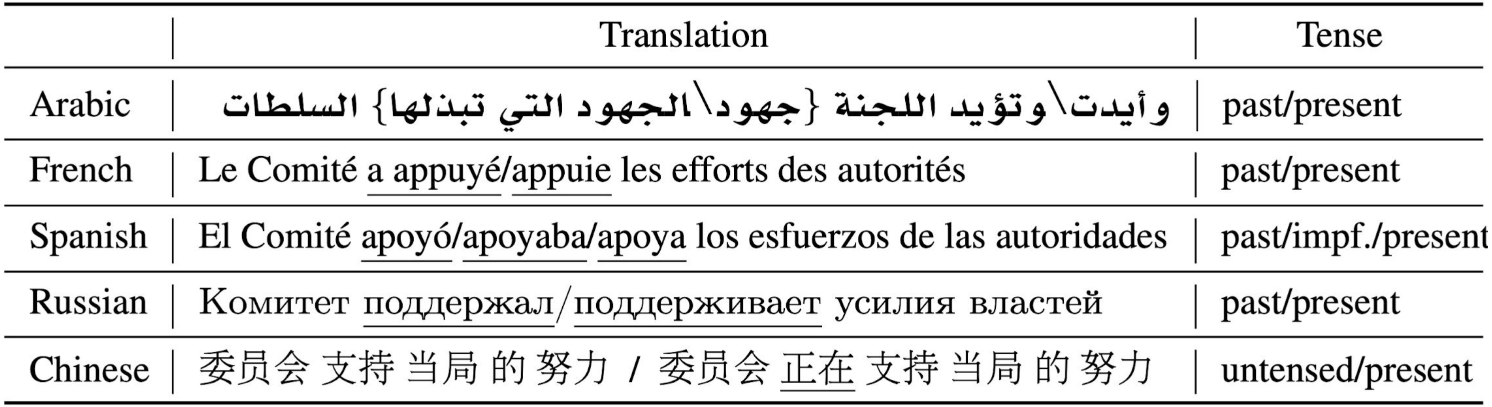
31 . Recognizes the important contribution of the African Peer Review Mechanism since its inception in improving governance and supporting socioeconomic development in African countries , and recalls in this regard the high-level panel discussion held on 21 October 2013 on Africa &apos;s innovation in governance through 10 years of the African Peer Review Mechanism , organized during the sixty-eighth session of the General Assembly to commemorate the tenth anniversary of the Mechanism ;

Spreads between sovereign bonds in Germany and those in other countries were relatively unaffected by political and market uncertainties concerning Greece in late 2014 and early 2015 .

- Fix its value to enforce tense

Controlling Translations

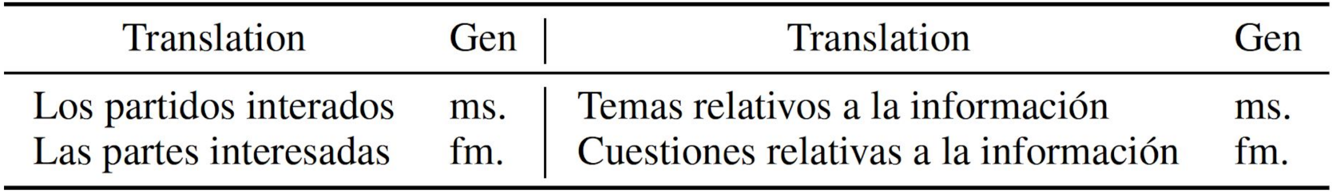
- Result of changing tense neuron



	Translation	Tense
Arabic	وأيدت\وتؤيد اللجنة {جهود\الجهود التي تبذلها} السلطات	past/present
French	Le Comité a appuyé/appuie les efforts des autorités	past/present
Spanish	El Comité apoyó/apoyaba/apoya los esfuerzos de las autoridades	past/impf./present
Russian	Комитет поддержал/поддерживает усилия властей	past/present
Chinese	委员会 支持 当局 的 努力 / 委员会 正在 支持 当局 的 努力	untensed/present

Controlling Translations

- Result of changing gender



Translation	Gen	Translation	Gen
Los partidos interados	ms.	Temas relativos a la información	ms.
Las partes interesadas	fm.	Cuestiones relativas a la información	fm.

Controlling Translations

- Open research question
- Whether all properties are manipulatable?
- Gender is the hardest in our case
- Train models with additional nobs of controlling

Cross-Model Correlation Analysis

Linguistic correlation analysis

- Requires linguistic annotations
- Assumption: properties are important

What does the model care about?

Cross-Model Correlation Analysis

Linguistic correlation analysis

- Requires linguistic annotations
- Assumption: properties are important

What does the model care about?

Cross-model correlation analysis

- Salient neurons for the model
- No annotation

Cross-Model Correlation Analysis

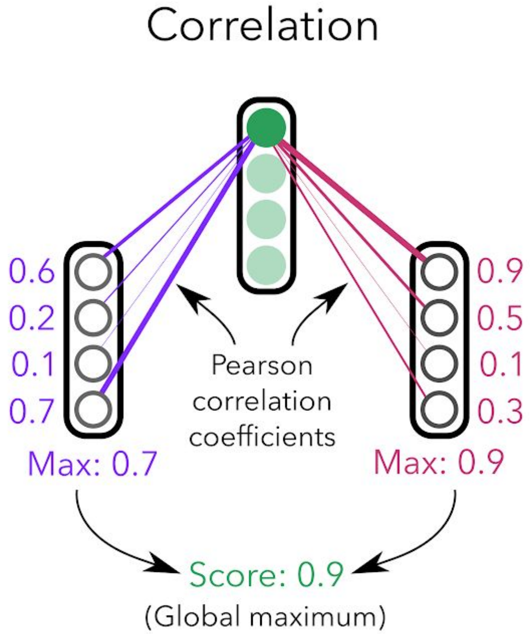
Basic Idea

- Hypothesis: Different models learn similar properties
- Search for neurons that share similar patterns in different networks
- Use correlation between neurons as a measure of their importance

Cross-Model Correlation Analysis

- Correlation of a neuron

- Models
  - Different checkpoints
  - Different random initialization
  - Different languages



Cross-Model Correlation Analysis

Consider u_i^m denotes i-th neuron activations in the m-th model

Maximum Correlation: highest correlation of u_i^m with any neuron in all other models

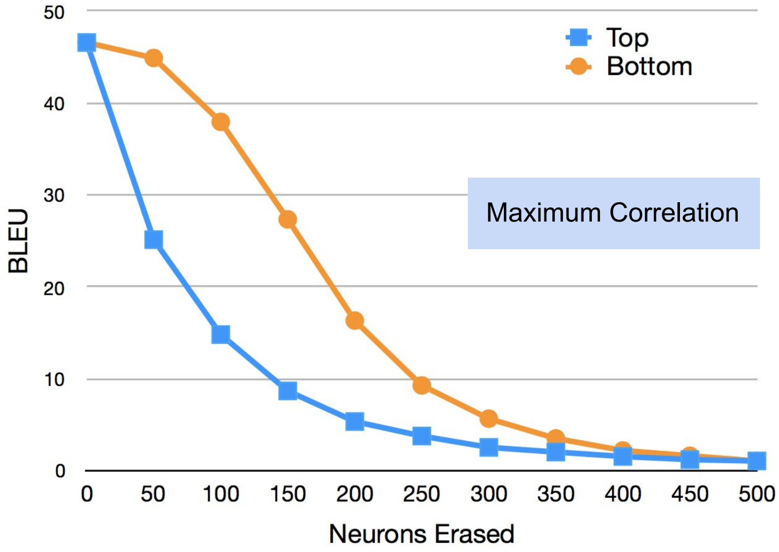
$$MaxCorr(u_i^m) = \max_{j,m' \neq m} |\rho(u_i^m, u_j^{m'})|$$

ρ is the pearson correlation

Evaluation - Ablation

Top 10%: drop by 15-20 BLEU points

Bottom 10%: drop by 2-3 BLEU points



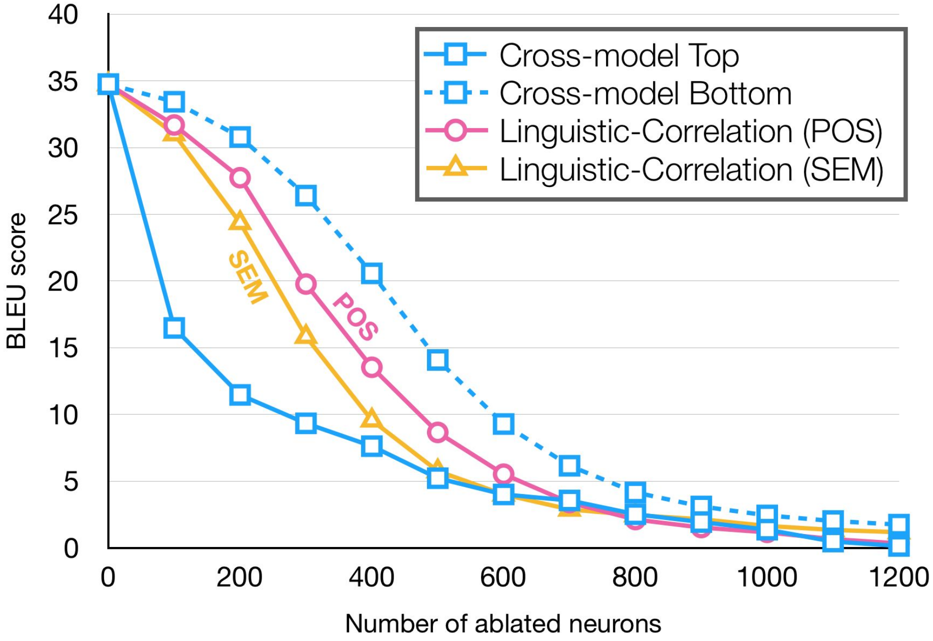
Maximum Correlation

English - Spanish Model

Evaluation - Cross-Model vs. Linguistic Rankings

- Cross-model rankings are most salient to the model

- In other words, there are properties more important than POS and SEM to generate better translations



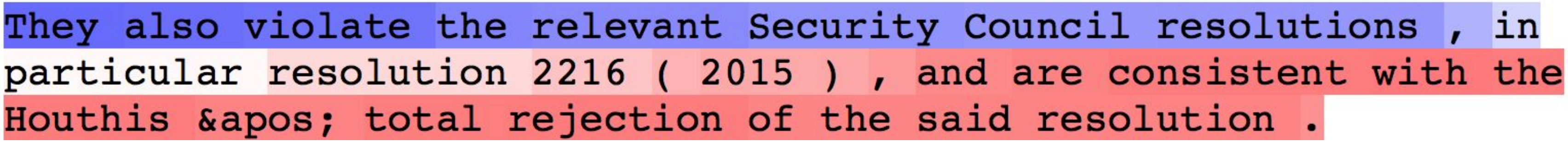
English - French

Evaluation - Visualization

- Information captured by top neurons

Evaluation - Visualization

- Many top neurons capture **position**
  - Activates negatively to positively

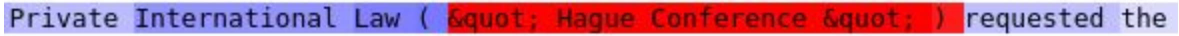


They also violate the relevant Security Council resolutions , in particular resolution 2216 (2015) , and are consistent with the Houthis &apos; total rejection of the said resolution .

- Having position neurons among the top neurons means that this phenomenon is important for the model to learn

Evaluation - Visualization

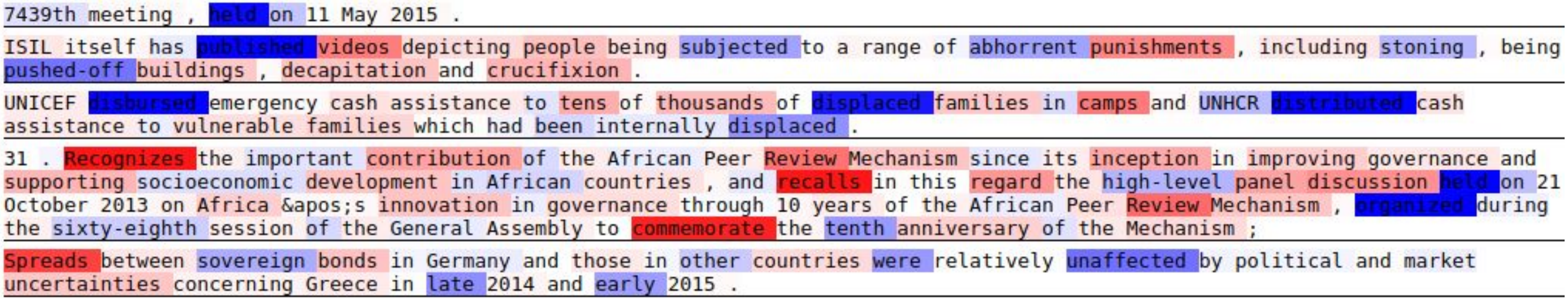
- Other top neurons: **location neurons**, tense neurons, etc.
    - Neuron activates positively for words inside parentheses and negatively for words outside parentheses



```
Private International Law ( &quot; Hague Conference &quot; ) requested the
```

Evaluation - Visualization

- Other top neurons: location neurons, **tense neurons**, etc.
  - Neuron activates positively on present tense (“recognizes, recalls, commemorate)
  - Neuron activates negatively on past tense (published, disbursed, held)



7439th meeting , held on 11 May 2015 .

ISIL itself has published videos depicting people being subjected to a range of abhorrent punishments , including stoning , being pushed-off buildings , decapitation and crucifixion .

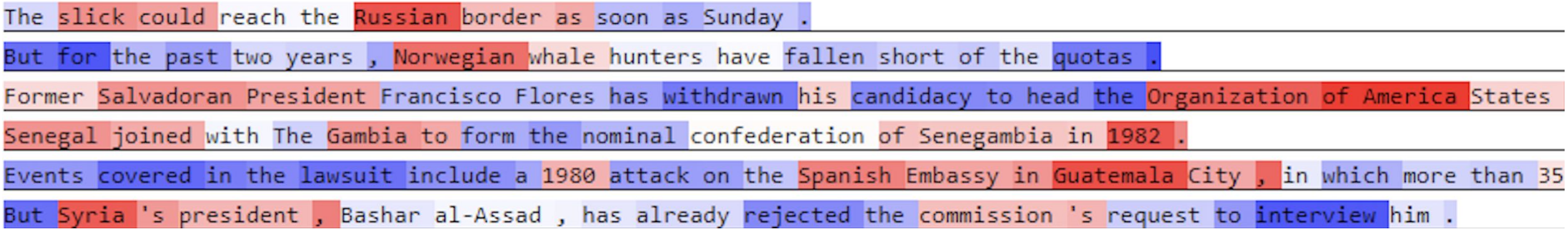
UNICEF disbursed emergency cash assistance to tens of thousands of displaced families in camps and UNHCR distributed cash assistance to vulnerable families which had been internally displaced .

31 . Recognizes the important contribution of the African Peer Review Mechanism since its inception in improving governance and supporting socioeconomic development in African countries , and recalls in this regard the high-level panel discussion held on 21 October 2013 on Africa &apos;s innovation in governance through 10 years of the African Peer Review Mechanism , organized during the sixty-eighth session of the General Assembly to commemorate the tenth anniversary of the Mechanism ;

Spreads between sovereign bonds in Germany and those in other countries were relatively unaffected by political and market uncertainties concerning Greece in late 2014 and early 2015 .

Visualization

- Country names
- Phrase-level information



The slick could reach the Russian border as soon as Sunday .

But for the past two years , Norwegian whale hunters have fallen short of the quotas .

Former Salvadoran President Francisco Flores has withdrawn his candidacy to head the Organization of America States

Senegal joined with The Gambia to form the nominal confederation of Senegambia in 1982 .

Events covered in the lawsuit include a 1980 attack on the Spanish Embassy in Guatemala City , in which more than 35

But Syria 's president , Bashar al-Assad , has already rejected the commission 's request to interview him .

Summary

- Network learns linguistic information at various level of granularity
- Lower-layers are good in learning word-level concepts while higher-layers focus more on abstract and syntactic concepts
- Information is both focused and distributed
- Position information is among the most salient property
- Neurons capture multiple related properties -- present and past tense
- Potential applications
  - Controlling the model

Thank you !!!

- On the Linguistic Representational Power of Neural Machine Translation Models. **Computational Linguistics.**
- One Size Does Not Fit All: Comparing NMT Representations of Different Granularities. **NAACL-HLT.**
- What is One Grain of Sand in the Desert? Analyzing Individual Neurons in Deep NLP. **AAAI 2019.**
- NeuroX: A Toolkit for Analyzing Individual Neurons in Neural Networks. **AAAI Demo Track 2019.**
- Identifying and Controlling Important Neurons in Neural Machine Translation. **ICLR 2019.**
- What do Neural Machine Translation Models Learn about Morphology?. **ACL 2017.**
- Evaluating Layers of Representations in Neural Machine Translation on Part-of-Speech and Semantic Tagging Tasks. **IJCNLP 2017.**
- Understanding and Improving Morphological Learning in the Neural Machine Translation Decoder. **IJCNLP 2017.**