Analysis Methods in Neural Language Processing: A Survey

Yonatan Belinkov and James Glass

TACL'19

Trends

- Before deep neural network era,
  - Design **human understandable features** like morphological properties, syntactic categories, semantic relations to solve our NLP task.
  - Observe the **importance** to these features assigned by a statistical NLP model to gain a better understanding of the model.
- Now,
  - Build an **end-to-end neural network model** that takes input (say, word embeddings) and generates an output (say, a sentence classification).
  - **Get good performance gains. Hard to interpret the model.**
  - **Goal of analysis work is to understand how linguistic concepts that were common as features in NLP systems are captured in neural networks.**

Tons of work published in trying to understand neural model for NLP

Goal of this survey is to organize the literature into several themes

Themes – Analysis methods

- What kind of linguistic information is captured in neural networks?
- Visualization methods
- Challenge sets or test suites
- Adversarial examples to probe weakness of neural networks
- Explaining model predictions

What kind of linguistic information is captured in neural networks?

- Three dimensions:
  - which **methods** are used for conducting the analysis?
    - e.g. predict properties from activations of the neural network
  - what kind of **linguistic** information is sought?
    - e.g. sentence length, simple word order
  - which **components** in the neural network are being investigated?
    - e.g. RNN hidden state, sentence embeddings

What kind of linguistic information is captured in neural networks?

- Probing task
  - Classification problem that focuses on simple linguistic properties of sentences.
  - e.g. Categorize sentence by the tense of their main verb.
- Example Setup:
  - Given an encoder (e.g., an LSTM) pre-trained on a certain task (e.g. MT), we use the sentence embeddings it produces to train the tense classifier (without further embedding tuning).
  - Assumption: **If the classifier succeeds, it means that the pre-trained encoder is storing readable tense information into the embeddings it creates.**

What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. Conneau et al. ACL'18.

What kind of linguistic information is captured in neural networks?

- **Surface**
  - **SentLen** – Predict length of sentence
  - **WC** – Predict which of target words appear in the sentence
- **Syntactic**
  - **TreeDepth** – Predict the maximum depth of syntactic tree underlying the sentence
  - **TopConst** – Predict the top constituents immediately below sentence node in tree
  - **BShift** - Predict whether two consecutive tokens have been inverted or not
- **Semantic**
  - **Tense** – Predict the tense of the main verb
  - **SubjNum**, **ObjNum** – Predict the number of the subject of main clause, direct object of MC
  - **SOMO** - Predict if a sentence occurs as-is in the source corpus, or whether a randomly picked noun or verb was replaced with another form with the same part of speech.
  - **CoordInv** – Predict if original sentence and sentence where the order of two coordinated clausal conjoints has been inverted purposely.

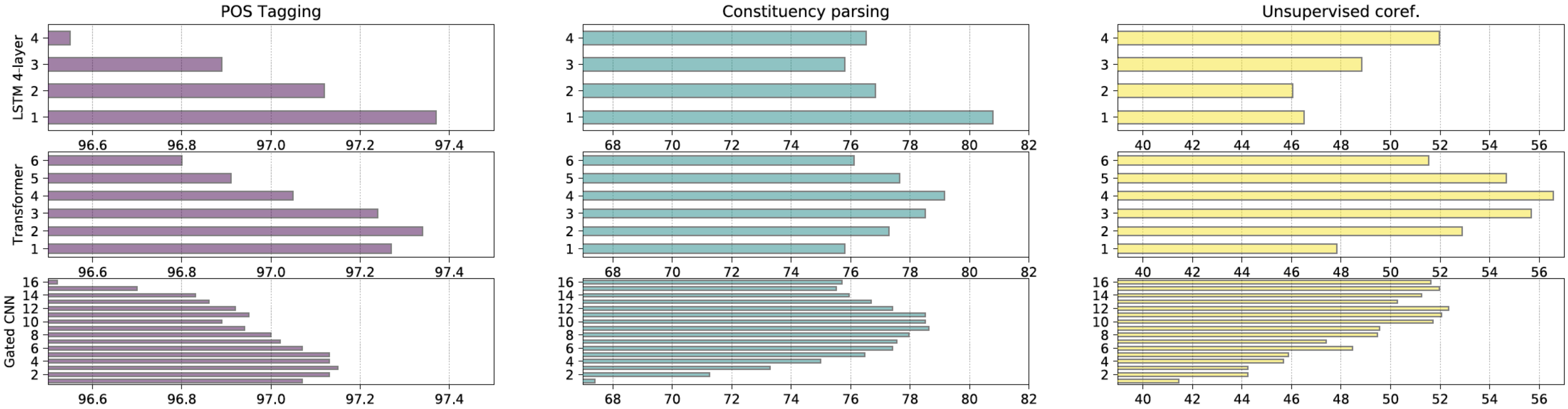
What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. Conneau et al. ACL'18.

What kind of linguistic information is captured in neural networks?

Task	SentLen	WC	TreeDepth	TopConst	BShift	Tense	SubjNum	ObjNum	SOMO	CoordInv
Baseline representations										
Majority vote	20.0	0.5	17.9	5.0	50.0	50.0	50.0	50.0	50.0	50.0
Hum. Eval.	100	100	84.0	84.0	98.0	85.0	88.0	86.5	81.2	85.0
Length	**100**	0.2	18.1	9.3	50.6	56.5	50.3	50.1	50.2	50.0
NB-uni-tfidf	22.7	**97.8**	24.1	41.9	49.5	77.7	68.9	64.0	38.0	50.5
NB-bi-tfidf	23.0	95.0	24.6	53.0	**63.8**	75.9	69.1	65.4	39.9	**55.7**
BoV-fastText	66.6	91.6	**37.1**	**68.1**	50.8	**89.1**	**82.1**	**79.8**	**54.2**	54.8
BiLSTM-last encoder										
Untrained	36.7	43.8	28.5	76.3	49.8	84.9	84.7	74.7	51.1	64.3
AutoEncoder	**99.3**	23.3	35.6	78.2	62.0	84.3	84.7	82.1	49.9	65.1
NMT En-Fr	83.5	**55.6**	42.4	81.6	62.3	88.1	89.7	89.5	52.0	71.2
NMT En-De	83.8	53.1	42.1	81.8	60.6	88.6	89.3	87.3	51.5	**71.3**
NMT En-Fi	82.4	52.6	40.8	81.3	58.8	88.4	86.8	85.3	52.1	71.0
Seq2Tree	94.0	14.0	**59.6**	**89.4**	**78.6**	**89.9**	**94.4**	**94.7**	49.6	67.8
SkipThought	68.1	35.9	33.5	75.4	60.1	89.1	80.5	77.1	**55.6**	67.7
NLI	75.9	47.3	32.7	70.5	54.5	79.7	79.3	71.3	53.3	66.5
BiLSTM-max encoder										
Untrained	73.3	**88.8**	46.2	71.8	70.6	89.2	85.8	81.9	73.3	68.3
AutoEncoder	**99.1**	17.5	45.5	74.9	71.9	86.4	87.0	83.5	73.4	71.7
NMT En-Fr	80.1	58.3	51.7	81.9	73.7	89.5	90.3	89.1	73.2	75.4
NMT En-De	79.9	56.0	52.3	82.2	72.1	90.5	90.9	89.5	73.4	**76.2**
NMT En-Fi	78.5	58.3	50.9	82.5	71.7	90.0	90.3	88.0	73.2	75.4
Seq2Tree	93.3	10.3	**63.8**	**89.6**	**82.1**	**90.9**	**95.1**	**95.1**	73.2	71.9
SkipThought	66.0	35.7	44.6	72.5	73.8	90.3	85.0	80.6	**73.6**	71.0
NLI	71.7	87.3	41.6	70.5	65.1	86.7	80.7	80.3	62.1	66.8

What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. Conneau et al. ACL'18.

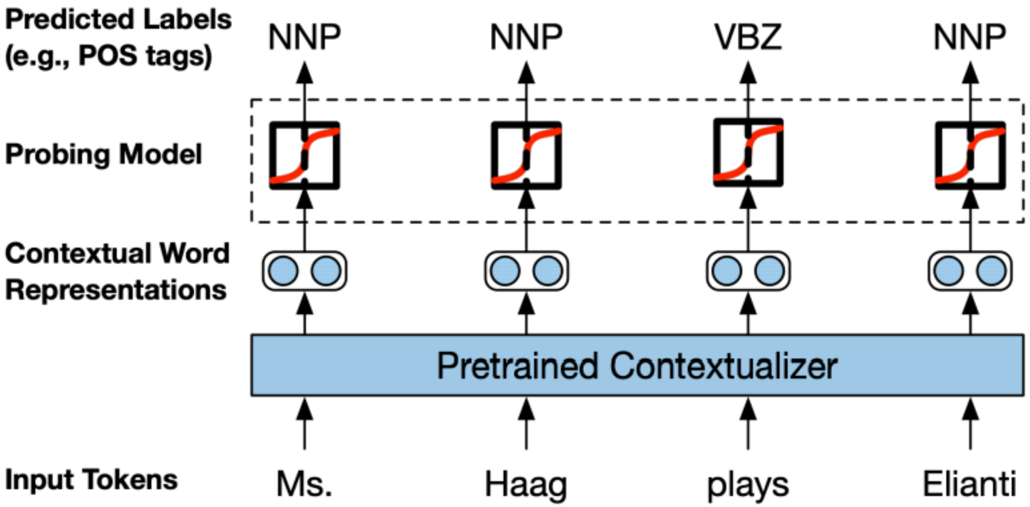
What kind of linguistic information is captured in neural networks?



- PoS, syntax in lower layers
- Coreference in higher layers
- => **biLM learn a hierarchy of contextual info.**

Dissecting Contextual Word Embeddings: Architecture and Representation. Peters et al. EMNLP'18

What kind of linguistic information is captured in neural networks?



- Focus on understanding the CWRs of individual or pairs of words.
- **Defines 16 probing tasks**
  - Token Labeling, Segmentation, Pairwise relations

Linguistic Knowledge and Transferability of Contextual Representations. Liu et al. NAACL'19

What kind of linguistic information is captured in neural networks?

Pretrained Representation	Avg.	CCG	POS		Chunk	NER	ST	GED	Supersense ID		EF
			PTB	EWT					PS-Role	PS-Fxn	
ELMo (original) best layer	81.58	93.31	97.26	95.61	90.04	82.85	93.82	29.37	75.44	84.87	73.20
ELMo (4-layer) best layer	81.58	93.81	**97.31**	95.60	89.78	82.06	**94.18**	29.24	74.78	85.96	73.03
ELMo (transformer) best layer	80.97	92.68	97.09	95.13	93.06	81.21	93.78	30.80	72.81	82.24	70.88
OpenAI transformer best layer	75.01	82.69	93.82	91.28	86.06	58.14	87.81	33.10	66.23	76.97	74.03
BERT (base, cased) best layer	84.09	93.67	96.95	95.21	92.64	82.71	93.72	43.30	**79.61**	87.94	75.11
BERT (large, cased) best layer	**85.07**	**94.28**	96.73	**95.80**	**93.64**	**84.44**	93.83	**46.46**	79.17	**90.13**	**76.25**
GloVe (840B.300d)	59.94	71.58	90.49	83.93	62.28	53.22	80.92	14.94	40.79	51.54	49.70
Previous state of the art (without pretraining)	83.44	94.7	97.96	95.82	95.77	91.38	95.15	39.83	66.89	78.29	77.10

- CWR >> Glove
- CWR is competitive with TSM. => Linear model extract good info. from CWR.
- ELMo, BERT > GPT => Bidirectionality is crucial.
- **CWR do not capture much transferable information about entities and coreference phenomena.**

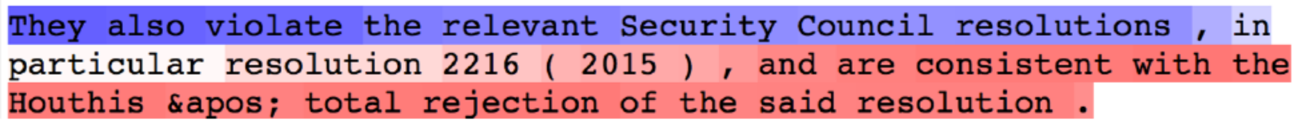
Linguistic Knowledge and Transferability of Contextual Representations. Liu et al. NAACL'19

What kind of linguistic information is captured in neural networks?

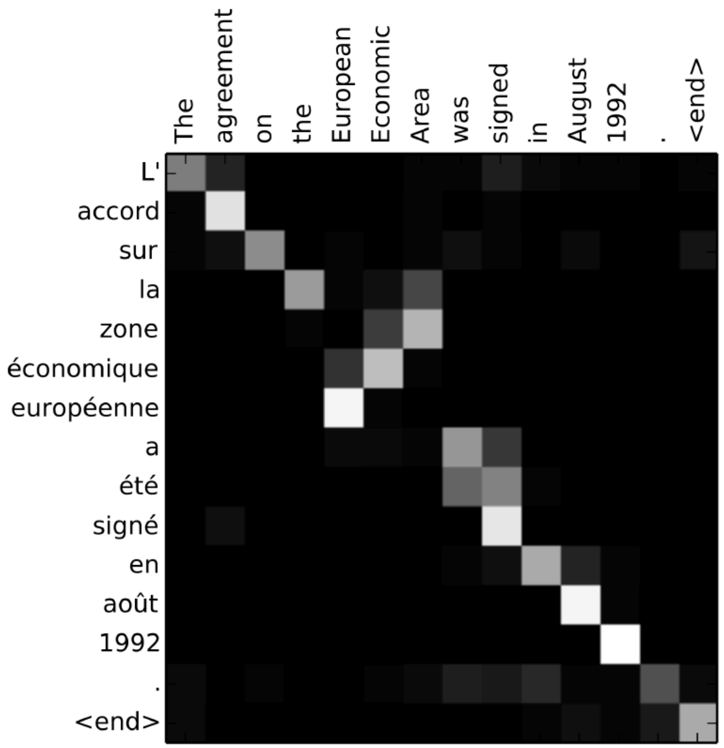
- **Limitations**
  - Probing task finds that a certain amount of linguistic info. is captured in the neural network. This does not mean the information is used by the network.
  - Most work is concerned with correlation: how **correlated** are neural network components with linguistic properties?
    - What may be lacking is a measure of **causation**: how does the encoding of linguistic properties affect the system output?
  - Better theoretical or empirical results required to understand why nuanced linguistic knowledge (e.g., tree depth) benefits from deeper probing classifier.

Visualization methods

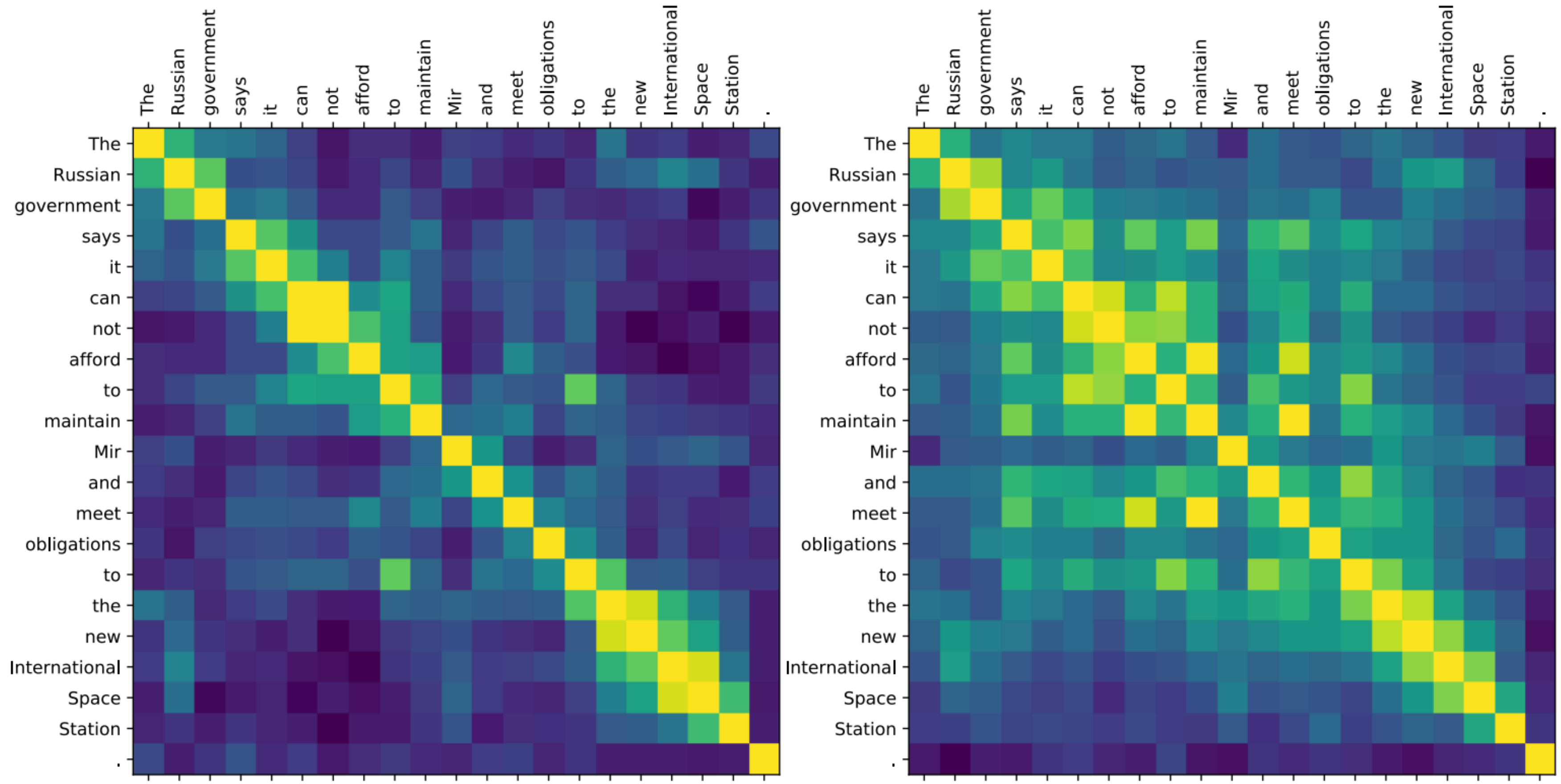
- Visualizing activations on specific examples in neural networks for language
  - e.g. the activations of a neuron that captures position in the sentence



- Attention visualization
  - Seq2seq problems like MT
  - Question Answering
  - Text Classification



Visualization methods – ELMo Contextual Similarity



- Lower layer encodes **local** info while top layer encode **longer range relationships**.
- Lower layer => words from the same syntactic constituents are in similar parts of the vector space e.g., “the new international space station”, “can not”
- Top layer => all verbs have high similarity e.g., “says”, “can”, “afford”, “meet”
- Top layer => perform co-reference resolution e.g., “it” to “government”

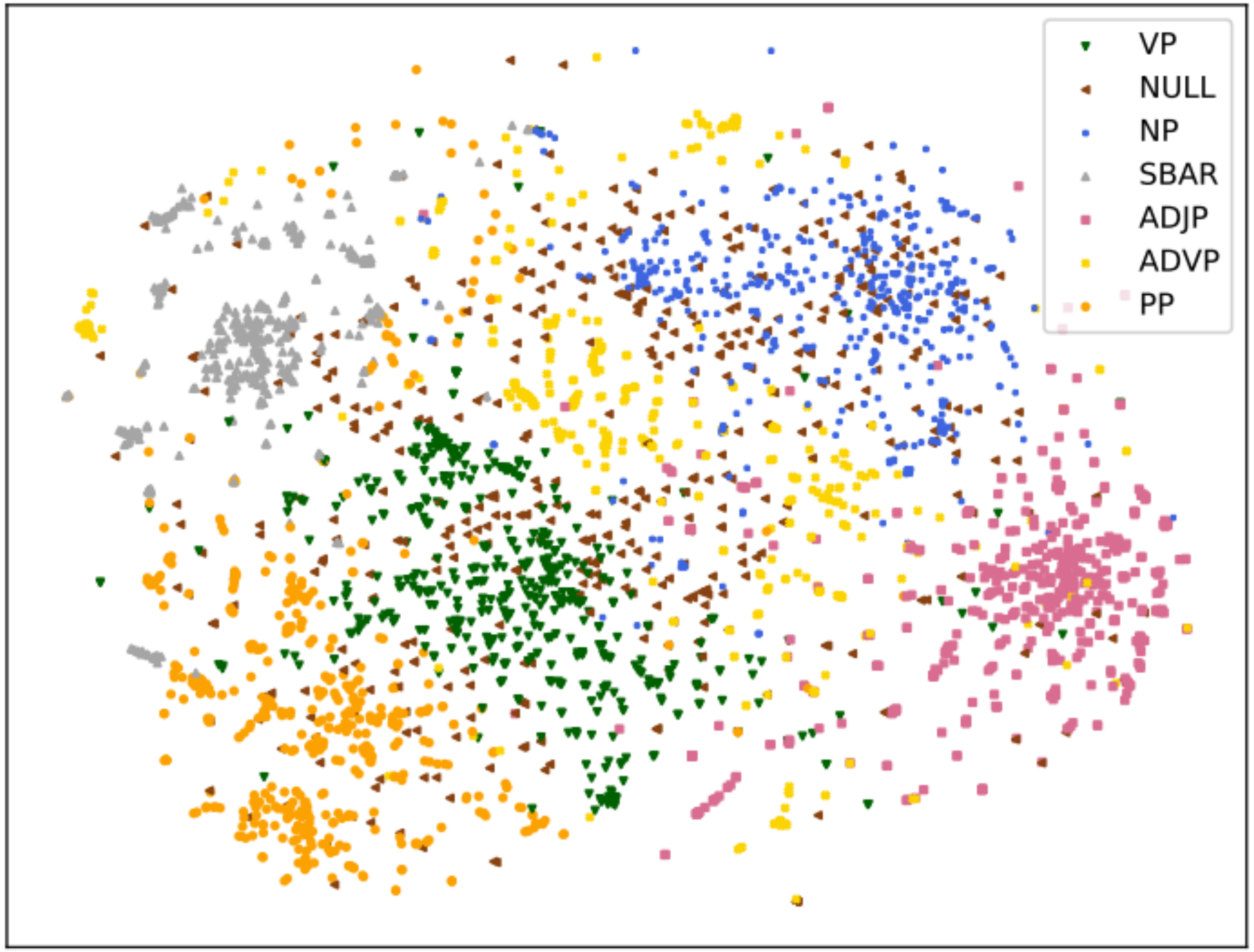
Dissecting Contextual Word Embeddings: Architecture and Representation. Peters et al. EMNLP'18

Visualization methods – t-SNE

[NP He] [VP reckons] [NP the current account deficit] [VP will narrow] [PP to] [NP only # 1.8 billion] [PP in] [NP September] .

- Compute **span representation** from first and last contextual representation of ELMo
- Get labels for some spans from CoNLL Chunking dataset
- Plot t-SNE
- Span representation capture elements of syntax.

t-SNE visualization of span representation



Dissecting Contextual Word Embeddings: Architecture and Representation. Peters et al. EMNLP'18

Visualization methods - Limitations

- **Evaluating visualization quality** is difficult and often limited to qualitative examples.
- Remains to be seen how useful visualizations turn out to be.

Challenge Sets

- Using benchmark datasets can let us evaluate system performance in the average case and may fail to capture a **wide range of phenomena**.
- Challenge datasets or test suite targets specific linguistic phenomenon.
- Criteria:
  - Task they seek to evaluate
  - Linguistic phenomena they aim to study
  - Languages they target
  - Their size
  - Their method of construction
  - How performance is evaluated

Challenge Sets – Contrastive Translation Pairs

- Analyzing translation quality (BLEU is too coarse grained)
- Idea: **NMT model has to assign a higher probability to a reference translation than to an example that introduces a specific error.**

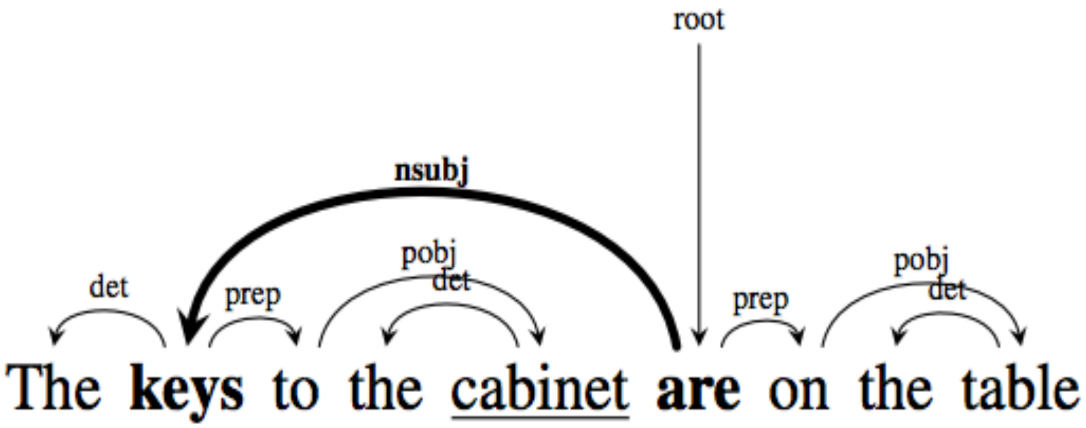
category	English	German (correct)	German (contrastive)
NP agreement	[...] **of the** American **Congress**	[...] **des** amerikanischen **Kongresses**	* [...] **der** amerikanischen **Kongresses**
subject-verb agr.	[...] that the **plan will** be approved	[...], dass der **Plan** verabschiedet **wird**	* [...], 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

- NP agreement – Change gender of singular definite determiner
- SV agreement – Change grammatical number of a verb
- SV particle – Replace separable verb particle with one not observed with the verb
- Polarity & Transliteration

How Grammatical is Character-level Neural Machine Translation? Assessing MT Quality with Contrastive Translation Pairs, Rico Sennrich, EACL'17.

Challenge Sets – Subject-verb agreement

- Create proxy task for probing syntax – subject-verb agreement task
- **Classify if the main verb is singular or plural based on its subject.**
- Run your LSTM till the word before the main verb and try to find the number of main verb based on the hidden representation.
- Alternatively, you can run a trained biLM upto 'cabinet' and try to compare the probability for the next word.
- **Success iff Prob(are|context) > Prob(is|context)**
- Conclusion: LSTM captures syntax sensitive structures really well. Needs supervision for harder cases.



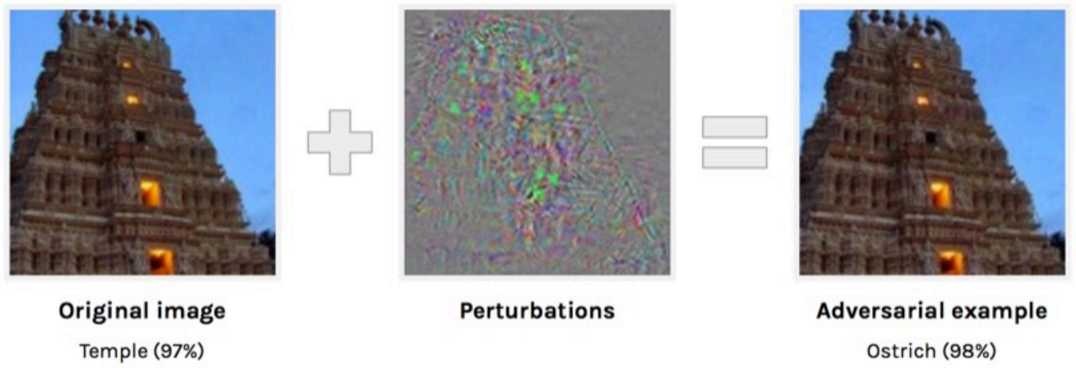
The **building** on the far right that's quite old and run down **is** the Kilgore Bank Building.

Tal Linzen, Emmanuel Dupoux & Yoav Goldberg (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. TACL.

Challenge Sets – Limitations

- Most challenge sets are in **English**.
- Some authors wish to test systems on extreme or difficult cases, beyond **normal operational capacity**. Depending on one's goal, we need to choose on specially chosen cases as opposed to the average case.
- We should compare model performance to **human performance** on the same task.

Adversarial Examples



- Understanding a model requires also an understanding of its failures
- In the vision domain, small changes to the input image can lead to **misclassification**, even if such changes are indistinguishable by humans.
- Basic setup: Given a neural network model f and an input example x, we generate an adversarial example x’ that will have a minimal distance from x, while being assigned a different label by f:

$$\min_{x'} ||x - x'||$$
$$\text{s.t.} \quad f(x) = l, f(x') = l', l \neq l'$$

Adversarial Examples

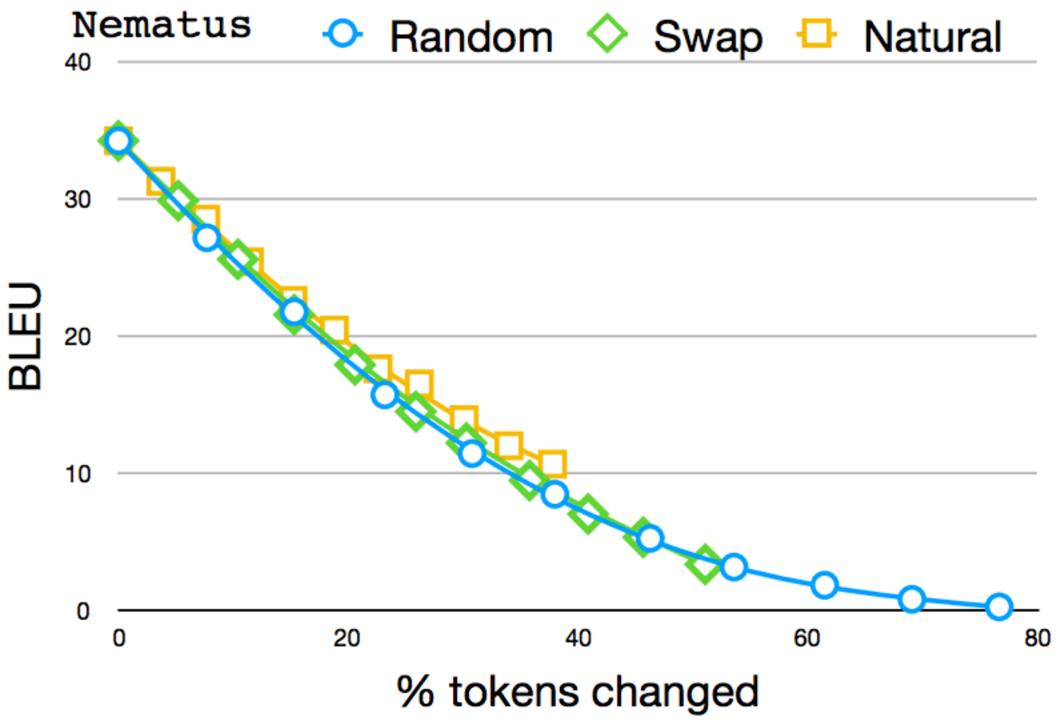
- Problems in generating adversarial text samples
  - Not clear how to measure distance between x and x' which are **discrete objects**
  - Minimizing the distance can't be formulated as an optimization problem, as this requires computing gradients with respect to a **discrete input**.
  - Difficulty in generating **imperceptible** changes in text
- Criteria:
  - Adversary's knowledge (white or black box)
  - Specificity of the attack (specific label or any label other than l)
  - Linguistic unit being modified (character or word level)
  - Task on which the attacked model was trained on (MT)

Adversarial Examples

- A **white** box attack example
  - Compute gradients with respect to the input word embeddings and perturb the embeddings.
  - Since this can result in a vector that does not correspond to any word, one could search for the **closest word embedding** in a given dictionary.
- A **black** box attack example
  - Using text edits that are thought to be natural or commonly generated by humans, such as **typos, misspellings**.

Adversarial Examples

- A **black** box attack example
  - Machine Translation
  - German to English
  - Noise on source side:
    - **Random** permutation of a word (e.g. noise -> niose)
    - **Swapping** a pair of adjacent letters (e.g. noise -> iones)
    - **Natural** human errors (source: wiki edit) (e.g. noise -> noide)
  - These examples can also be used for robust training and modeling



Synthetic and Natural Noise Both Break Neural Machine Translation. Belinkov et al. ICLR'18.

Explaining Predictions

- Explaining why a deep model makes a certain prediction is not trivial.
- Under-researched area
- Approaches
  - **Generate explanations** along with its primary prediction
    - cons: requires manual annotations of explanations
    - e.g. Explainable SNLI –Predicting explanation improves the model
  - **Use parts of the input** as explanations.

Premise: An adult dressed in black holds a stick.
Hypothesis: An adult is walking away, empty-handed.
Label: contradiction
Explanation: Holds a stick implies using hands so it is not empty-handed.

Premise: A child in a yellow plastic safety swing is laughing as a dark-haired woman in pink and coral pants stands behind her.
Hypothesis: A young mother is playing with her daughter in a swing.
Label: neutral
Explanation: Child does not imply daughter and woman does not imply mother.

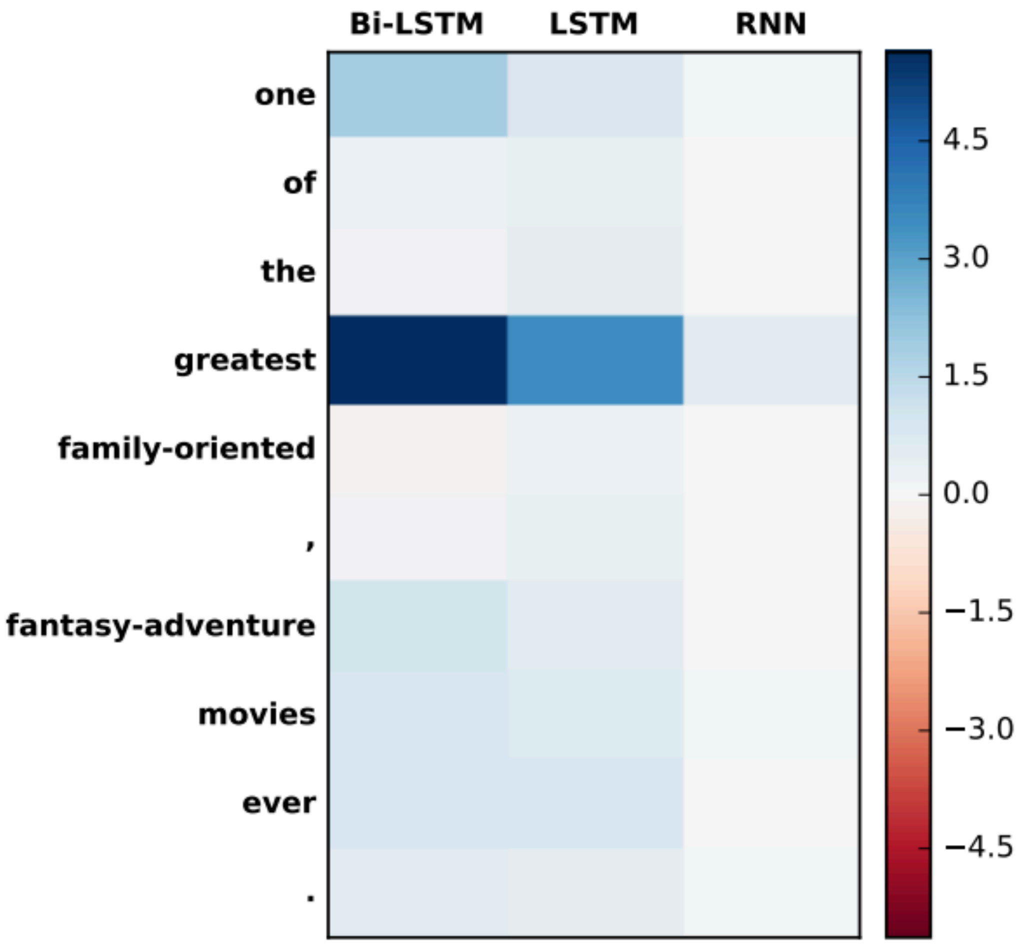
Premise: A man in an orange vest leans over a pickup truck.
Hypothesis: A man is touching a truck.
Label: entailment
Explanation: Man leans over a pickup truck implies that he is touching it.

Figure 1: Examples from e-SNLI. Annotators were given the premise, hypothesis, and label. They highlighted the words that they considered essential for the label and provided the explanations.

e-SNLI: Natural Language Inference with Natural Language Explanations. Camburu et al. NeurIPS'18.

Explaining Predictions

- **Representation Erasure**
  - **Visualize the network activations for specific examples.**
  - Importance of an input is the change in model confidence when we remove it (set the dimensions to 0).
  - e.g. sentiment analysis
    - Importance (greatest) = P(positive|input) – P(positive|input with greatest removed)



Understanding Neural Networks through Representation Erasure. Li et al. 2016

Future Work

- Probing task tells us how correlated are neural network components with linguistic properties?
  - We may need a **measure of causation**: how does the encoding of linguistic properties affect the system output.
- Challenge sets can check if model can work on difficult cases for a task
  - This might depend on **one's goals**. Hence its better to establish human performance on the sets and compare with the model performance.
  - Challenge sets needed for tasks besides NLI and MT.
- Evaluation of analysis work is often limited or qualitative
  - Newer forms of evaluation for **determining the success of different methods** are needed.
- Relatively little work on **explaining predictions** of neural network models
- Much of the analysis work is focused on the **English** language.