Analysis Methods in Neural Language Processing: A Survey

Yonatan Belinkov[1,2] James Glass[1]

[1]MIT Computer Science and Artificial Intelligence Laboratory [2]Harvard John A. Paulson School of Engineering and Applied Sciences

Background

In feature-rich NLP systems, one could in theory examine how different features are used by the system, in contrast to end-to-end neural networks that are thought to be **opaque**. As neural networks replace many of their feature-rich counterparts, researchers seek to analyze and evaluate neural networks in novel and more fine-grained ways.

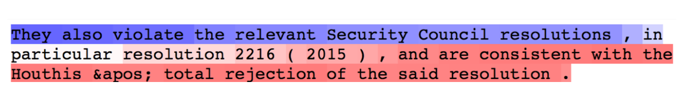
In this survey paper, we:

- Review analysis methods in neural NLP.
- Categorize methods by prominent trends.
- Highlight limitations and future directions.

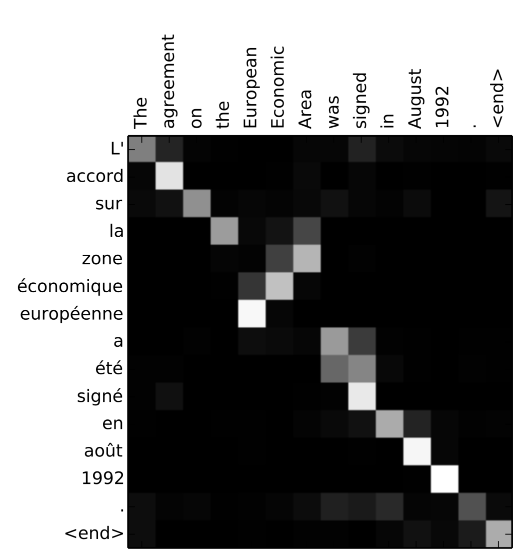
Visualization

Visualization is a valuable tool for analyzing neural networks; usually done on individual examples.

- Activations.
- Attention weights.
- Saliency of input features.
- Clusters of embeddings.
- Online tools: `LSTMVis`, `Seq2Seq-Vis`, `NeuroX`, `BertViz`, etc.



Heatmap of a position neuron.



Bahdanau et al. (2014)

Limitations: evaluation

- Evaluation is difficult and usually qualitative.
- Exceptions: human evaluation of which visualization is more accurate or credible.

Finding linguistic information in neural models

A primary goal is to determine what linguistic information is captured in neural networks when they are trained on various tasks.

- **Methods**: Probing tasks: (1) train neural model; (2) generate representations; (3) train a classifier to predict a linguistic property.
- **Linguistic phenomena**: phonology, morphology, syntax, semantics, etc.
- Different network **components**: embeddings, states, attention, etc.
- **Example**: predict POS tags from hidden states on a neural MT encoder.

Some insights

- Networks learn a substantial amount of linguistic information, especially about frequent properties, less so about rare cases.
- Hierarchical representations: lower layers capture simpler properties than higher layers. But, this may depend on architecture and task.

Limitations: methodological issues

- Correlation $\neq$ causation: Predictability of a property does not entail that the end model is using it.
- The nature of the predictor/classifier is rarely discussed.

Challenge sets

Most benchmarks evaluate performance in the average case. Challenge sets (or test suites) evaluate systems systematically on fine-grained phenomena.

- **Task**: mostly NLI/entailment and MT; also word/sentence embeddings.
- **Linguistic phenomena**: earlier work exhaustive, recent more focused
- **Languages**: Almost only English, with exceptions in MT evaluation.
- **Scale** : from small and manually constructed to large and automatic.
- **Methods**: modify benchmarks, design templates, form contrastive pairs.

Limitations

- Poor language and task coverage.
- Conflict: Should systems perform well in extreme or average cases?

Adversarial examples

Given a neural network model f and an input example x, generate an adversarial example x' that will have a minimal distance from x, while being assigned a different label by f:

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

Problems with discrete input: **measuring** and **minimizing** $||x - x'||$.

- **Adversary's knowledge**: In **white-box** attacks, word embeddings are perturbed, but the result may not be a known word. In **black-box attacks**, texts are usually edited (e.g., typos).
- **Attack specificity**: Targeted attacks are rare (being white-box).
- **Linguistic unit**: usually characters or words.
- **Task**: text classification, reading comprehension, MT. Less work on low-level tasks.

Limitations: coherence & perturbation measurement

- Need to apply constraints on few edit operations or filter replacements by semantic similarity.
- Few human evaluations of grammaticality or similarity of adversarial examples to original ones. More are needed.

Explaining predictions

Explaining specific predictions is important for increased accountability. Current solutions are limited:

- Generate explanations along with the prediction; requires manual annotations of explanations.
- Treat parts of input as explanation; ignores internal computations.

Conclusion

- Still much work to do in analysis of neural NLP.
- **Online appendix** has tables with categorizations of many studies. Contributions welcome!

Acknowledgements: This work was supported by the Qatar Computing Research Institute and the Harvard Mind, Brain, and Behavior Initiative.