https://qdata.github.io/deep2Read/

Defending Against Neural Fake News

Zellers et al.

2019

Presented by Eli Lifland, 11/15/2019

Fake news: background

- Fake news: News designed to intentionally deceive
- Most current fake news is manually written
- Progress in language generation makes it easier to artificially generate ‘neural’ fake news
- Focus on text-only documents with two goals
  - Monetization: Get clicks to increase ad revenue
  - Propaganda: Communicate targeted information
- Current detection efforts:
  - Manual fact-checking e.g. Politifact
  - Automated detection of stylistic biases
  - However, fact checking not a panacea due to cognitive biases

Threat Model

- Adversarial game with 2 players
  - Adversary: Generates fake stories which are either viral or persuasive, which must read realistically to humans and verifier
  - Verifier: Classifies stories as real or fake. Has access to unlimited real news stories and a few fake news stories from a specific adversary

Text generation

- Current SOTA: unconditional generation

$$p(\boldsymbol{x}) = \prod_{i=1}^{N} p(x_i | x_1 \ldots x_{i-1}).$$

- Including metadata leads to joint distribution

$$p(\text{domain}, \text{date}, \text{authors}, \text{headline}, \text{body}).$$

Text generation

- At inference time, sort context fields in standard order, append field-specific start token, then sample from model

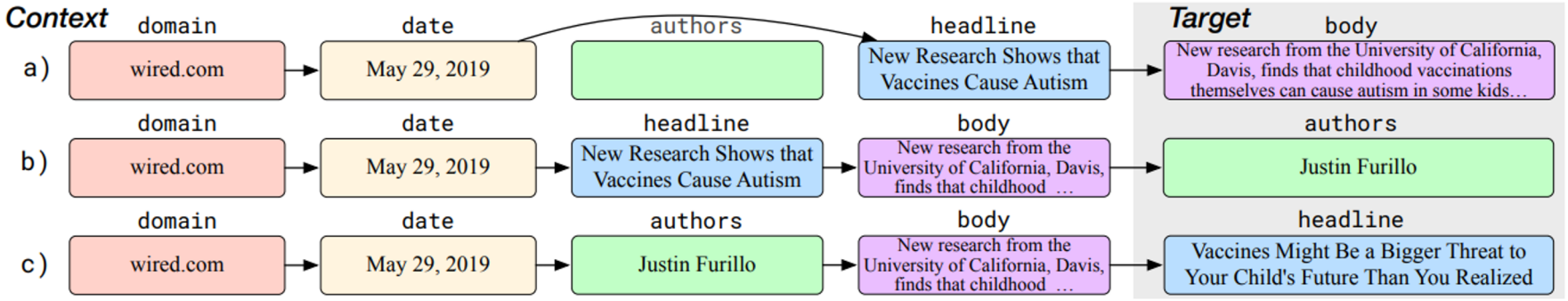

Figure 2: A diagram of three GROVER examples for article generation. In row a), the body is generated from partial context (the authors field is missing). In b), the model generates the authors. In c), the model uses the new generations to regenerate the provided headline to one that is more realistic.

Model training

- Simulate inference by partitioning article's fields into disjoint sets, randomly drop out fields, then predict tokens in first set followed by tokens in second set
- Architecture: Same as GPT2
  - GROVER-Base: same size as GPT/BERT-Base
  - GROVER-Large: same size as BERT-Large
  - GROVER-Mega: same size as GPT2
- Dataset: RealNews scraped from Common Crawl
- Learning: 2 weeks of training on 256 TPU cores

Variance of generations

- Choice of decoding algorithm is important
- Likelihood-maximization works better for close-ended generation (e.g. translation) than open-ended
- Use Nucleus Sampling (top-p)
  - For a given threshold p, at each timestep sample from most probable words with cumulative probability representing top-p% of vocabulary

LM Results

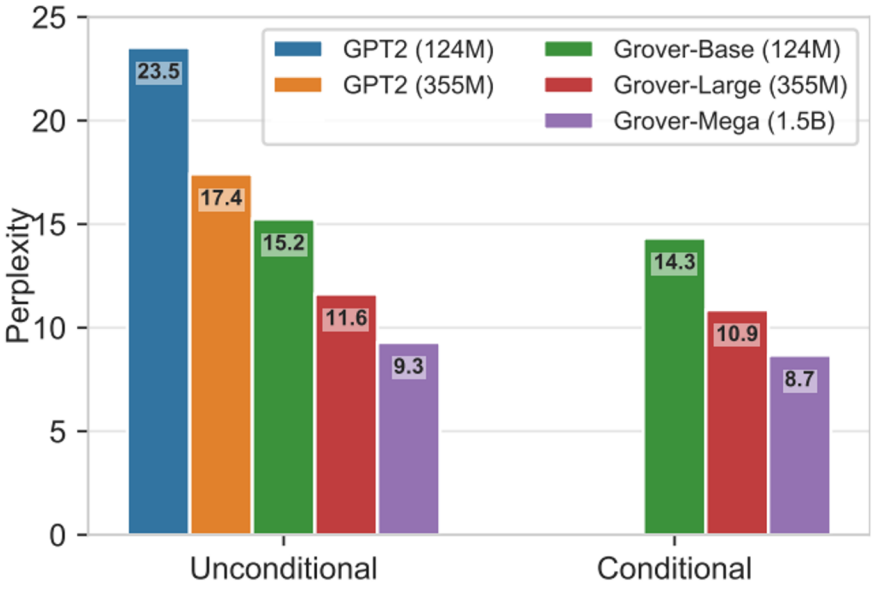

Figure 3: Language Modeling results on the body field of April 2019 articles. We evaluate in the *Unconditional* setting (without provided metadata) as well as in the *Conditional* setting (with all metadata). GROVER sees over a 0.6 point drop in perplexity when given metadata.

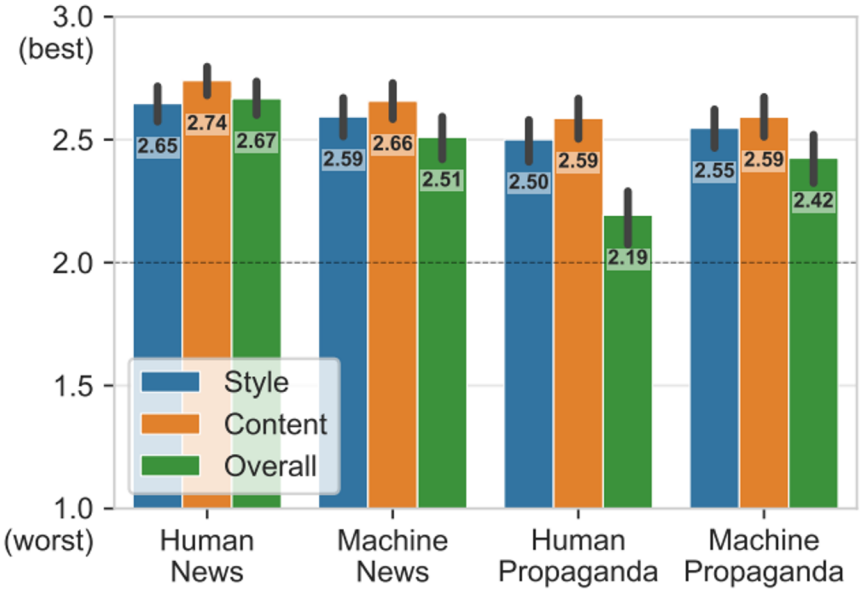

Figure 4: Human evaluation. For each article, three annotators evaluated style, content, and the overall trustworthiness; 100 articles of each category were used. The results show that propaganda generated by GROVER is rated more plausible than the original human-written propaganda.

Neural Fake News Detection

- Use models for role of verifier: classifying articles as human or machine written
  - GROVER
  - GPT2
  - BERT
  - Fast Text
- Semi-supervised setting
  - Access to entire RealNews training set
  - Limited access to generated articles; 10k using metadata from recent articles
- Unpaired vs. paired evaluation

Neural Fake News Detection: Results

Table 1: Results of discriminators versus generators, in both the paired and unpaired settings and across architecture sizes. We also vary the generation hyperparameters for each generator-discriminator pair, reporting the discrimination test accuracy for the hyperparameters with the *lowest* validation accuracy. Compared with other models such as BERT, GROVER is the best at detecting its own generations as neural fake news.

Discriminator size		Unpaired Accuracy, Generator size: 1.5B	355M	124M	Paired Accuracy, Generator size: 1.5B	355M	124M
	Chance		50.0			50.0	
1.5B	GROVER-Mega	**92.0**	**98.5**	**99.8**	**97.4**	**100.0**	**100.0**
355M	GROVER-Large	**80.8**	**91.2**	**98.4**	**89.0**	**96.9**	**100.0**
	BERT-Large	73.1	75.9	97.5	84.1	91.5	99.9
	GPT2	70.1	78.0	90.3	78.8	87.0	96.8
124M	GROVER-Base	**70.1**	**80.0**	**89.2**	77.5	88.2	95.7
	BERT-Base	67.2	76.6	84.1	**80.0**	**89.5**	**96.2**
	GPT2	66.2	71.9	83.5	72.5	79.6	89.6
11M	FastText	63.8	65.6	69.7	65.9	69.0	74.4

https://qdata.github.io/deep2Read/

Neural Fake News Detection: Weak Supervision

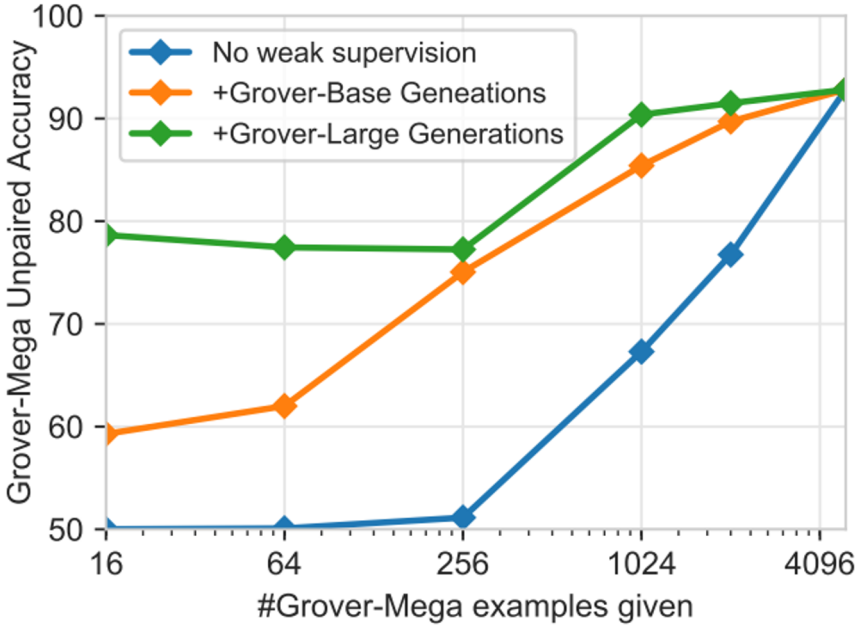

Figure 5: Exploring weak supervision for discriminating GROVER-Mega generations. With no weak supervision, the discriminator sees x machine-written articles (from GROVER Mega). For +GROVER-Base and +GROVER-Mega, the discriminator sees $5000-x$ machine-written articles given by the weaker generator in question. Seeing weaker generations improves performance when few human examples are given.

https://github.com/rowanz/grover

How does model distinguish?

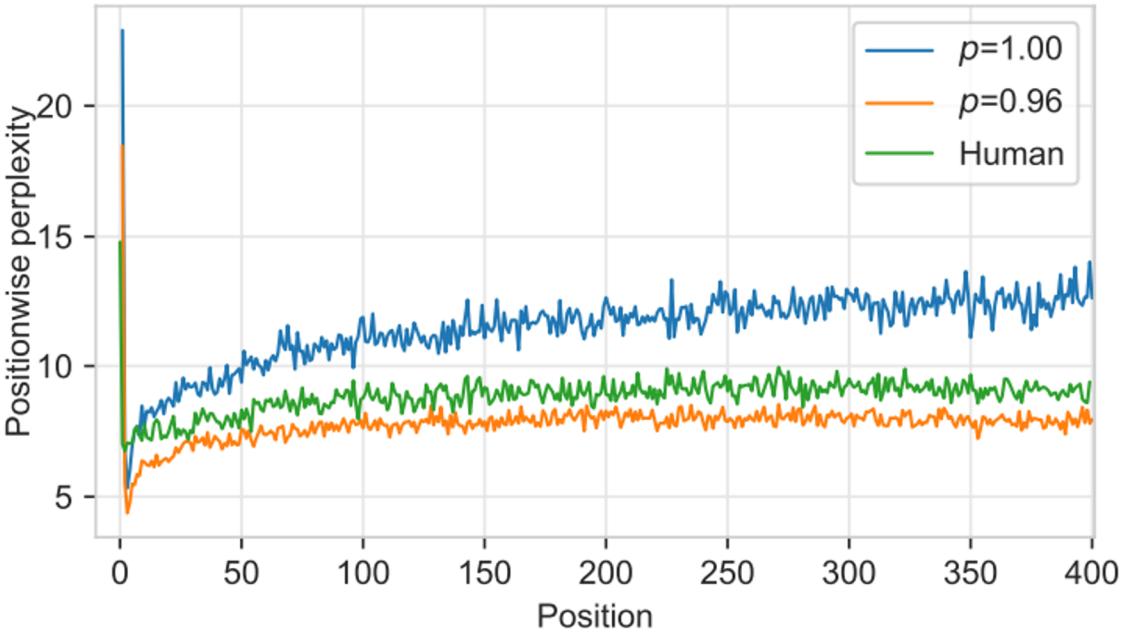

Figure 6: Perplexities of GROVER-Mega, averaged over each position in the body (after conditioning on meta-data). We consider human-written with GROVER-Mega generated text at $p=1$ (random sampling) and $p=.96$. The perplexity of randomly sampled text is higher than human-written text, and the gap increases with position. This suggests that sampling without variance reduction increasingly falls out-of-distribution.

How does model distinguish?

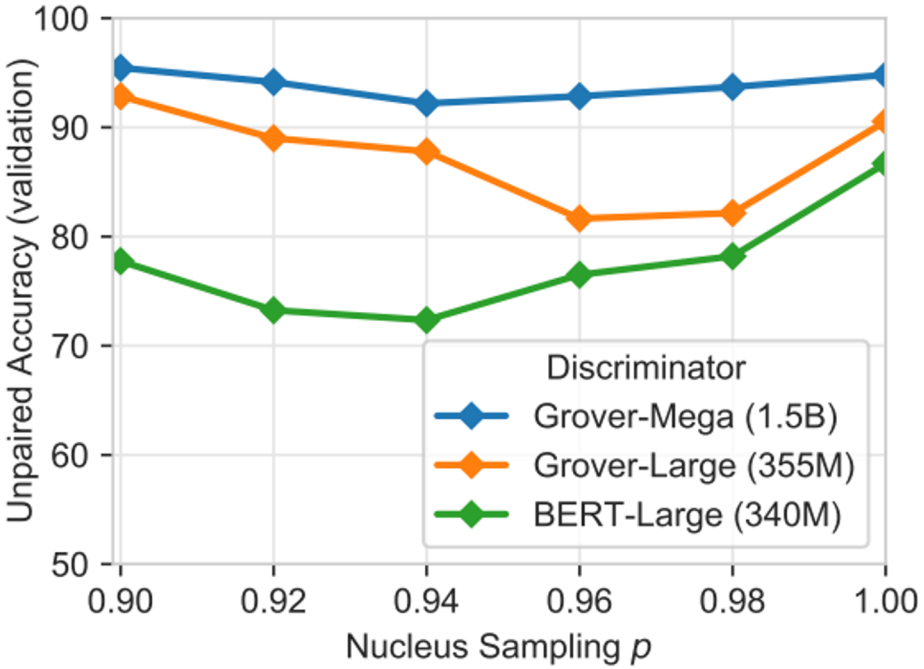

Figure 7: Unpaired validation accuracy, telling apart generated news articles (from GROVER Mega) from real articles, at different variance reduction thresholds p (for Nucleus Sampling). Results varying p show a sweet spot ($p = 0.92 - 0.96$) wherein discrimination is hardest.

Conclusion/Future Work

- Era of neural disinformation
  - Training GROVER-Mega cost “only” $25k
- Release of generators is critical
  - GROVER is an effective detector of neural fake news
- Additional threat models
- Machine-generated real news
- Integrating real-world knowledge into discriminator
- Platforms should use NNs to flag news articles as they are published