BLiMP: A Benchmark of Linguistic Minimal Pairs for English

The paper and dataset can be found here.

Dataset

- 67 sub-datasets each containing 1000 minimal pairs isolating specific contrasts in syntax morphology or semantics.
- Automatically generated using grammars
- Used to evaluate several different types of state of the art language models
    - > identify morphological contrasts reliably but struggle with semantic restrictions on distribution of quantifiers negative polarity items and subtle syntactic phenomena such as extraction islands

Island phenomenon are the hardest for language models to deal with, scores on these minimal pairs are near chance. (which is funny because these are fairly robust restrictions in human grammacality judgements).

*What if you trained a language model with negative examples using the minimal pairs provided? Frame it as a classification problem and see how they compare then? Kind of avoids the issue at heart since the paper is looking at how well existing language models address these phenomenon but it would be interesting to see if these architectures **can** model this information*

The CoLA dataset features 10,000 judgements and that shows BERT and company doing well at that task. (this is mentioned later)

Related work

Language modelling

recent shifts to transformer models have resulted in reduced perplexity however, "this doesn't give insight into these models' linguistic knowledge."

> Evaluation on downstream task benchmarks (Wang et al. 2018, 2019a) is more informative, but might not prsent abroad enough challenge or represent grammatical distinctions at a sufficiently fine-grained level.

Evaluation of linguistic knowledge

There have been previous works that examined using minimal pairs to infer whether language models learn about specific linguistic phenomenon.

However, most of these works have been limited in what they investigated:

- LInzen et al. (2016) look closely at subject verb agreement
- Marvin and LInzen (2018) look at a larger set of stuff including NPI and reflexive licensing.
- Things like control, raising, ellipsis etc have not been included despite being well studied linguistic phenomenon.

There are corpora that contain grammaticality judgements for sentences. The most recent and largest is CoLA (WArstadt et al. 2019b). CoLA is included in the GLUE benchmark.

Current transformer models can be trained to give excellent results on this data.

looks like my previous idea has already been done

When Warstadt and Bowman (2019) investigaated the performance of pretrained language models including an LSTM, GPT and BERT, they found that the models did well on "sentences with marked

argument structure" and did worse on sentences with long-distance dependencies (though transformer models did better there).

> “ evaluating supervised classifiers prevents making strong conclusions about the models themselves, since biases in the training data may affect the results

Performance could be due to the occurance of similar examples in the training dataset.

When language models are evaluated on minimal pairs, this evades the problem.

The authors say the probability of a sentence (and thus the inverse perplexity) can be used as a proxy of acceptability.

Dataset

The data is automatically generated using expert crafted grammars.

- ensures that there are sufficient unacceptable answers (which are very very rare in naturally occuring text)
- allows for fully controlled dataset with isolation of each linguistic phenomenon.

Data generation procedure:

- Use a basic template
- pull from a vocab of 3,000 morphologically, syntactically, and semantically annotated words
  - These features are needed to create grammatical and felicitous sentences
- Code is available in this github repo

Sometimes implausible sentences can be generated but the authors view this as a non-issue.

The authors consider frequent inclusion of a phenomenon i na syntax/semantics textbook as an informal proxy for what is core linguistic phenomenon for English. (not especially useful when examining non-English languages as few are taught from the perspective of a different language. E.g. a minimalist syntax textbook that only discusses French)

These are the phenomenon

covered

- Anaphor agreement
- Argument structure
- Binding
- Control/Raising
- Determiner/Noun agreement
- Ellipsis
- Filler-Gap
- Irregular forms
- Island effects
- NPI licensing
- Quantifiers
- Subject-Verb agreement

Comparison to related resources

with 3000 words, this has the widest vocabulary of any related generated dataset. 11 different verb subcategorization frames.

Other works like Linzen et al (2016) that use a larger lexicon size but use data-creation methods that are limited in control or scope.

- Linzen (2016) change number marking on present tense verbs but this is a strategy that is specific to subject agreement phenomenon
- Lau et al. (2017) build a dataset but doing multiple round trip machine translations but this creates a number of grammatical violations and does not offer the granular minimal pairs that this paper's data generation method provides.

Validation

Used mechanical turk 20 annotators rated 5 pairs from each of the 67 paradigms. aggregate human agreement is estimated at 96.4%.

Only cases where the annotators agreed with Blimp on 4/5 examples from each paradigm were included. (the 67 paradigms included passed, 2 additional ones were rejected on these grounds).

Individual human agreement approximated at 88.6%

Evaluation of language models

GPT-2 achieves highest scores, n-gram the lowest, LSTM and Transformer-XL tied.

> the results seem to indicate that access to training data is the main driver of performance on Blimp for the neural models we evaluate

They point to the fact that the LSTM and Tranformer-XL models performed about the same despite wildly different architectures and GPT-2 had 100x the training data but similar architecture to Transformer-XL.

Phenomenon specific results

models perform best to human level on morphological phenomena (anaphor agreement, determiner-noun agreement, and subject-verb agreement). *Possibly because english doesn't have that much of this*

GPT2 is the only model that performs above chance on Islands but it is still 20 points behind humans. they are very hard in general.

Wilcox et al (2018) concluded that LSTMs have knowledge of some island conditions which contridicts the findings here. However, Wilcox et al. compare four related sentences with or without gaps, obtaining wh-licensing as a metric of hos strongly the language model identifies filler-gap dependency in a single spot, the lm has learned the constraint if the probability is close to 0. *this is difficult to parse, I think I need to read the original paper*

This paper finds that neural models can identify long-distance dependencies but not the domains where these dependencies are blocked.

weak performance on argument structure is somewhat strange because previous work has suggested that argument structure is a solid domain for neural models. However, these works (Warstadt and Bowman (2019)) trained the model on CoLA and didn't do direct language

modelling.

Contribution

I am unfamiliar with the creation of minimal pair datasets for evaluation of neural language models. It seems that this paper's main contribution, though, is the creation of their new dataset that approaches minimal pairs with more breadth: including examples of many more types of English linguistic phenomena. They have the widest vocabulary of any generated dataset like this, including a large number of verb subcategorization frames.

Revision #7

Created Mon, Jan 27, 2020 7:07 PM by kenneth

Updated Tue, Jan 28, 2020 4:25 AM by kenneth