Riordan et al. 2020

An empirical investigation of neural methods for content scoring of science explanations

NGSS science standards dimensions

- DCI (disciplinary core ideas
- CCC (cross cutting concepts)
- SEP (science and engineering practices

KI rubric:

- > “involves a process of building on and strengthening science understanding by. incorporating new ideas and sorting out alternative perspectives using evidence

- rewards conecting evidence to claims in their explanations

Data

Constructed reponse (CR) items are evaluated. The ones chosen are cases where SEPs need to be used while showing understanding of CCCs and DCIs.

CR Items:

- Musical Instruments and the Physics of Sound Waves (MI)
- Photosynthesis and Cellular Respiration (PS)
- Solar Ovens (SO)
- Thermodynamics Challenge (TC)

Two separate rubrics in parallel:

- KI rubric
  - linkage with subsets of the ideas described in the evidence statements
    - Photosynthesis (PS) listed 5 ideas related to energy and matter changes during photosynthesis
  - Scores from 1-5
- NGSS subscore rubric
  - two of three dimensions for each CR
    - Only those that are relevant given the prompt are used (e.g. a question where the answer doesn't depend upon science and engineering practices would not have a score for that dimension)
  - scores from 1-3

The thermodynamics challenge item was particularly challenging.

Sometimes there were less annotated data available for the NGSS dimension models compared to the KI models.

Models

Each item and score type were trained independently. 10-fold cross validation with train/val/test splits, evaluating on concastenated predictions across folds.

SVR

- binary word unigrams and bigrams

RNN

- pretrained word embeddings (GloVe 100) fed into a bidirecitonal GRU encoder.
- Hidden states of GRU are pooled (max)
- Encoder output aggregated in a fuly-connected feedforward layer using sigmoid act (giving scalar score).
- **Presumably the same scaling and unscaling is happening that we worked with before because sigmoid should be squishing everything to be between 0,1**
- exponential moving average across weights used during training
- 50 epochs

Pretrained transformer

- bert-base-uncased
- using `[CLS]` token output, fed through a non-linear layer to obtain the scalar score.
- exponential moving average across weights used during training
- 20 epochs
- When identifying best hyperparameters, for each fold, taking the epoch where validation performance is highest for evaluation.
- During final training, validation and training data are concatenated and then the model is retrained.
  - **I assume this is done for all the models but it's only mentioned for the PT model**

Results

KI models

The Pretrained transformer models are more robust, they're always ahead of the RNN on all metrics (sometimes not by much though).

The items that were highly skewed showed lower levels of human-machine agreement (lower thant he 0.7 threshold for QWK in real world scoring applicaitons) **Where does that threshold come from??**

Revision #6

Created Thu, Jan 21, 2021 2:11 PM by kenneth

Updated Thu, Jan 21, 2021 3:50 PM by kenneth