Answer Scoring

- Riordan et al., 2019
- Horbach et al., 2019
- Riordan et al. 2020

Riordan et al., 2019

How to account for mispellings:

Quantifying the benefit of character representations in neural content scoring models

---

Takehome:

"Models with character representations outperformed their word-only counterparts...lower MSE and higher QWK" p. 121

Datasets

- ASAP-SAS: 10 questions with large number of responses for each question, sentence or two in length
- Formative-SAS: dataset collected by ETS (relatively short answers)
- Summative-LAS: 20 questions, mean number of words is 230

Methods

Word only model

pretrained word embeddings into bidirectional GRU. Hidden states of GRUs are either pooled or go through an MLP attention mechanism Output of the encoder goes through sigmoid fully connected layer which produces a score

Character + word models

Each word is represented with a sequence of 25-dimensional character embeddings. "Character embeddings are concatenated with the word embeddings prior to the word-level encoder" (p. 119)

Results

ASAP-SAS

While adding character representations performed better than just spelling correction, the effect of adding character representations was not statistically significant in the GLMM model and using spelling corrections was not significant either.

No evidence for interaction between character representations and spelling correction in the

GLMM.

Formative K12-SAS

Same general trend as ASAP-SAS

- character and word representations outperform word representations
- spelling corrected models outperformed non-spelling corrected models

Statistical significance between the different representations and the different methods of spelling correction but no interaction observed between mispelling bins and the representation used.

"The difference between feature sets and between mispellings bins was significant even when controlling for score and number of words" (p. 123)

Large majority of responses had no spelling errors. 3 spelling bins used (0, 1, 2+)

Q: Is spelling not what the character representations are able to capture? Is it instead morphological variation?

- What if you ran a stemmer over the input? Would the difference between word+character embeddings and plain word embeddings go away? Surely someone has done this.

Q: I thought that the addition of character representations was helpful for two of the datasets but not the last one. The conclusion reached was that character representations were not as helpful as spelling correction but I think this was only significant for the 2nd dataset.

Q: Are the character representations alone enough? (what if you dropped words)

Horbach et al., 2019

The influence of variance in learner answers on automatic content scoring

Andrea Horbach and Torsten Zesch

Variance

Sources of variance

- Conceptual variance:
  - when there are multiple separate right answers to a question.
  - bigger issue is number of variants of incorrect answers. *why not focus on modelling correct answers? Could you use an approach that allows you to rely more on how close this answer is to the correct answers I saw in training (if generative, I'm not sure how this would work for discriminative) could you model correct/wrong questions as anomaly detection?*
- Variance in realization
  - different ways of forming the same conceptual answer
  - Linguistic variation
    - language provides lots of possibilities to express the same meaning *what if you*

did reparsing or something to map variant forms to roughly the same meaning

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??**