Using BERT and XLNET for the Automatic Short Answer Grading Task

Hadi Abdi Ghavidel, Amal Zouaq and Michel C. Desmarais
Department of Computer Engineering and Software Engineering, Polytechnique Montréal, Montreal, Canada

Keywords: Automatic Short Answer Grading, SciEntBank, BERT, XLNET.

Abstract: Over the last decade, there has been a considerable amount of research in automatic short answer grading (ASAG). The majority of previous experiments were based on a feature engineering approach and used manually-engineered statistical, lexical, grammatical and semantic features for ASAG. In this study, we aim for an approach that is free from manually-engineered features and propose an architecture for deep learning based on the newly-introduced BERT (Bidirectional Encoder Representations from Transformers) and XLNET (Extra Long Network) classifiers. We report the results achieved over one of the most popular dataset for ASAG, SciEntBank. Compared to past works for the SemEval-2013 2-way, 3-way and 5-way tasks, we obtained better or competitive performance with BERT Base (cased and uncased) and XLNET Base (cased) using a reference-based approach (considering students and model answers) and without any type of hand-crafted features.

1 INTRODUCTION

Automatic grading of natural language answers is a highly desired goal in education. Advances in machine learning bring this goal closer to reality. Large classes and the success of Massive Open Online Courses (MOOCS) in education contribute to making this goal even more attractive. Open-ended answers provide teachers with a more accurate and detailed understanding of how a student comprehends domain-specific knowledge (Badger and Thomas, 1992). This is compared to traditional types of answers like multiple-choice questions or fill-in-the-gap items in which the student's understanding is restricted to the choices that are presented and thus not examined deeply. (Riordan et al., 2017).

For automatic grading, natural language answers can be divided into essays or short answers. According to Burrows et al. (2015), short answers have the following characteristics: The answer should *not be guessed* from the words in the question (*external knowledge*); the answer should be given in *natural language*; the *length* of the answer should be about one phrase to one paragraph; the *content* of the answer is domain-related; and the answer should be *close-ended*.

In both short answers and essays, each student answer is evaluated based on a nominal, ordinal or ratio scale (Roy et al., 2018). In the nominal scale, grades are in the format of labels like correct, incorrect, etc. Ordinal grades are in the letter format like A+, A-, etc. and ratio grades are in the numerical format like 1, 1.5, etc.

Besides grades, the student answer is usually associated with the question and a model (also called reference) answer(s). Sakaguchi et al. (2015) defined two general types of grading approaches. When automatic grading is done based only on the student answer and label, this is called a response-based approach. Otherwise, the grading is done using a reference-based approach in which the whole context is considered (model answer or question, or both along the student answer and the label). In this case, the system compares the student's answer with the model answer using several types of similarity metrics. The other approaches are hybrid, in which both response- based and reference-based techniques are taken into account simultaneously.

In a majority of natural language processing(NLP) tasks such as ASAG, language model (LM)s have proven to be successful. In essence, these models help determine the probability of a sequence of words and can predict words given previous words within a sequence (Goldberg, 2017). Traditional language models such as n-gram language models use count methods. Vector-space models based on counting n-grams have often been used for the ASAG task. For example, Mantecon et al. (2018) compared bag of n-grams

Ghavidel, H., Zouaq, A. and Desmarais, M.
Using BERT and XLNET for the Automatic Short Answer Grading Task.
DOI: 10.5220/0009422400580067
In *Proceedings of the 12th International Conference on Computer Supported Education (CSEDU 2020) - Volume 1*, pages 58-67
ISBN: 978-989-758-417-6

representations with bags of semantic annotations for ASAG.

In modern approaches, LMs are trained using neural networks. Neural language models alleviate the problems of traditional approaches in a number of ways (Goldberg, 2017). Firstly, they expand the context taken into account. Secondly, these models exhibit a generalization capability across different contexts. Initial neural models were based on recurrent neural networks (RNN) like long short term memory networks (LSTMs and BiLSTm). The most recently introduced neural models for language modeling like BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) and XLNET (Extra Long Network) (Yang et al., 2019) are based on the transformer architecture. They are reported by Devlin et al. (2018) to have robustly increased the performance of several NLP tasks like GLUE (General Language Understanding Evaluation) or question answering on SQuAD (Stanford Question Answering Dataset) (Rajpurkar et al., 2016). We aim to expand the application of these transformer models to the current ASAG task.

In this paper, we adopt a reference-based approach to train a machine learning model on one of the most challenging datasets provided in SemEval[1]-2013, which is SciEntBank (Nielsen et al., 2008). We define two grading models based on BERT and XLNET and show that both either match or improve the performance of the state of the art (SOTA) on this dataset. The main advantage of the approach presented here is that grading is performed without manually extracting features, contrary to what has been generally done in the state of the art.

The structure of the paper is as follows: In section 2, we review previous approaches to grading applied on the SciEntBank dataset. Then, we explain BERT and XLNET in the context of ASAG. In section 4, we specify the experimental setup and the evaluation measures. The next section analyzes the results obtained in all our experiments. We discuss how BERT and XLNET boost the ASAG on SciEntBank dataset in section 6. The last section describes our conclusion, limitations and future work.

2 PREVIOUS WORKS

In 2013, SemEval (Dzikovska et al., 2013) established a competition for ASAG, framed as a textual entailment task where grading the answers requires semantic inference. Semantic inference helps to identify if the content of student answers and model answers are similar or dissimilar and goes beyond a mere word overlap. Two datasets namely BEETLE (Dzikovska et al., 2010) and SciEntBank were provided for this challenge. In this paper, we trained a model on the SciEntBank dataset. According to the results of the SOTA (Dzikovska et al., 2010), this dataset is the most challenging of the two.

The SciEntBank dataset is gathered from 3rd-6th grade students (Nielsen et al., 2008). The general topics for this dataset include Life Science, Physical Science and Technology, Earth and Space Science, and Scientific Reasoning and Technology. Based on the classification from SemEval-2013, there are three ways of labeling students' responses:

- *2-way task:* correct and incorrect
- *3-way task:* correct, contradictory or incorrect
- *5-way task:* correct, partially correct, contradictory, irrelevant or not in the domain

For evaluation purposes, SemEval-2013 provided three test sets: *test of unseen-answers (TUA), test of unseen-questions (TUQ), and test of unseen-domains (TUD) scenarios*. The TUA is used for the evaluation of the system based on questions already seen in the training set (Dzikovska et al., 2013). By contrast, the questions in the TUQ are totally new and not observed in training set. Finally, questions related to domains unseen in the training set are proposed in the TUD scenario. The size of the test data is 540, 733 and 4562 answers respectively. Overall, there are 10,804 responses to 197 questions in the whole train and test datasets.

The first results on this dataset were published in SemEval-2013 (Dzikovska et al., 2013). In this competition, nine teams participated with various models. Almost all the teams benefited from using similarity metrics such as BLEU (Bilingual Evaluation Understanding) (Papineni et al., 2002) between questions, student answers and model answers. These metrics were sometimes applied to dense vectors obtained using LSA (Latent Semantic Analysis) (Deerwester et al., 1990) or similar dimension reduction techniques. Considering all the tasks in all the test sets, the overall best performance was obtained with the following systems:

- ETS (Heilman and Madnani, 2013): In this system, n-gram features and similarity features are used together with a domain adaptation technique called Daume (III, 2009) to adapt features to the context of the answers across domains.
- CoMet (Ott et al., 2013): Syntactic features like parts of speech, dependency relations and constituent structures are used in this system. It

[1]Semantic Evaluation Competition

also combined several other features and built a stacked classifier. The meta-classifier used was logistic regression.

- SOFTCARDINALITY (Jimenez et al., 2013): This system modeled the overlap between student and model answer through soft cardinality, a method proven by Jimenez et al. (2010) to boost the accuracy in measuring textual similarity.

Since the creation of the SemEval-2013 challenge, several other researchers also evaluated their systems on the SciEntBank dataset. For example, Ramachandran et al. Ramachandran and Foltz (2015) augmented the ASAG dataset by generating model answers. To accomplish this task, the authors summarized the answers of the top students. The features used were mostly similarity measures between their generated model answer, the student answer and the question. These features include word overlap, cosine similarity and Lesk similarity. When evaluated on the TUA test, their approach outperformed past works in the 3-way and 5-way tasks, but not in the 2-way task.

Sultan et al. (2016) applied a feature ensemble approach in which the authors combined text alignment, semantic similarity, question demoting, term weighting, and length ratios. They were able to achieve slightly better results than the SOTA on the 2-way task for the TUD test set.

In one of the most recent works, Saha et al. (2018) suggested a new set of features in which they partitioned the similarities into histogram bins instead of one single overall similarity. These partial similarities are based on tokens and part-of-speech tags. In addition to these similarities, the authors considered question types (like where, when, how, etc.) as another feature in their evaluation process. They combined these token features (TF) with sentence embeddings features (SF) based on InferSent (Conneau et al., 2017)) and achieved better or competitive results on three datasets, including SciEntBank. Roughly at the same time, Marvaniya et al. (2018) created a scoring rubric for ASAG. Instead of considering only the model answers as a perfect example for the comparison, the authors defined the scoring rubric as the ranked clusters of student answers for each grade. In their research, the scoring rubric, in conjunction with a student answer and question, is converted to core features such as lexical overlap and sentence embeddings extracted by InferSent (Conneau et al., 2017)).

Finally, the most similar work to ours is the approach of Sung et al. (2019) in which they conducted an experiment on only the 3-way SciEntBank dataset using BERT and improved the results over the SOTA. Since there is no access to the input setup and all the hyperparameters of the paper, their results are hardly reproducible. In contrast to this work, we tested BERT and XLNET on all the SemEval-2013 tasks (2-way, 3-way and 5-way).

In this work, we adopt the reference-based approach of using student and model answers. We propose to learn the entailment between the student answer and the model answer using BERT (Devlin et al., 2018) and XLNET (Yang et al., 2019). Our approach differs from top ASAG systems in that we carry out the task of grading without using any type of handcrafted features. We also do not use questions as input in our experiments.

3 ASAG WITH BERT AND XLNET

As the grades in the SciEntBank dataset are based on a nominal scale, we consider grading the answers as a supervised classification task. There has been a vast body of methods for representing input (tokens within student answers, model answers or questions). These representations, such as n-grams, (Heilman and Madnani, 2013; Mantecon et al., 2018) are usually learned or extracted from training data. However, deep learning networks require a large amount of data to be trained adequately. The fact that the size of SciEntBank dataset is relatively small hinders a robust training process. To relieve this issue, modern language model architectures make use of transfer learning (Goodfellow et al., 2016). In fact, one of the strengths of models such as BERT and XLNet is that they are pre-trained on very large corpora. They can then be fine-tuned on small corpora such as SciEntBank. These recent language models are built on the transformer architecture (Vaswani et al., 2017).

BERT is a bidirectional model which is trained to build a language model using a Transformer encoder (Vaswani et al., 2017; Devlin et al., 2018). The corpora used for pretraining are BooksCorpus (800M words) and Wikipedia (2,500M words). Overall, BERT beats the other similar transformer models like OpenAI (Radford et al., 2018) in that it considers both left and right contexts given a target word. BERT is pretrained using a masked language model (MLM) task and next sentence prediction (NSP) task. MLM (Taylor, 1953) is used to randomly mask some of the tokens and then predict them. In NSP, two consecutive sentences are used to gain discourse knowledge.

Another transformer model which is even more recent than BERT is XLNET. XLNET is inspired from BERT and Transformer-XL (Dai et al., 2019) but differs in the following ways:

- **XLNET and BERT:** XLNET is based on a bidirectional AR (autoregressive) language model,

while BERT is built upon bidirectional AE (autoencoder) model. In this regard, XLNET is pretrained based on the potential permutations of context words surrounding a target word. Also, XLNET takes into account dependencies between words.

- **XLNET and Transformer-XL:** Two fundamental features of Transformer-XL are integrated into XLNET permutation language modeling: *relative positional embeddings* and *the recurrence mechanism*.

Overall, ASAG using transformer-based architectures can be seen as learning the textual entailment between a student answer and a model answer. In what follows, we briefly present the model architecture and input representation for both BERT and XLNET.

3.1 Model Architecture

Both BERT and XLNET are available with two model sizes:

- BERT Base and XLNET Base: There are 12 layers, 12 attention heads and 768 neurons for these two models. The number of total parameters is 110 millions. For BERT, cased and uncased versions are trained and released. However, only the cased version is released with XLNET.
- BERT Large and XLNET Large: In contrast to BERT Base, these models include 24 layers, 16 attention heads and 1024 neuron. Similar to BERT Base, there are cased and uncased versions while only the cased version exist for the XLNET large model.

Since we ran our experiments on Google Colab and due to our limited computing power, we chose the lighter BERT and XLNET *base models*.

3.2 ASAG Input Representation and Architecture

In our models (called graders), the input is the student answer, model answer and the grades (labels) associated with each student answer. Our goal is to train a model in which the relationship between student answers and model answers is learned in association with the grade assigned to each student answer.

Similar to SemEval-2013 as discussed in Section 2, we consider ASAG as a textual entailment task. Inspired by the textual entailment literature (Dagan et al., 2005), we hypothesize that a correct student answer (text) entails the model answer (hypothesis) and we note it as follows:

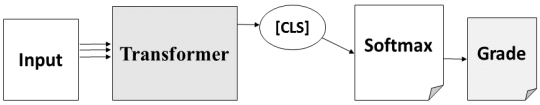

Figure 1: Overall Grading Process by Transformers.

$$\text{student answer S} \rightarrow \text{model answer M}$$

In other words, the correct grade is assigned to an answer if the answer entails the model answer. We also explored the results of the reverse experiment in which the model answer entails the student answer. The results of these reverse experiments are not reported in this paper because they are not extraordinarily different from the results of the S→M experiments.

The overall process is summarized in Figure 1. No pre-processing is applied to the raw answer texts. Overall, the input sequence is fed to the Transformer model. Then, the output of the classification model token ([CLS[2]] in Figure 1) is passed to the softmax function and a grade is assigned to the answer.

Both BERT and XLNET have a maximum number of tokens in an input sequence (max sequence lengths). In fact, each model includes different tokenizers that output different tokens. What makes these lengths dissimilar is visible in the tokenization output of the following example from model answers in SciEntBank train set:

Rub the minerals together and see which one scratches the other.

- **BERT Base Uncased:** ['rub', 'the', 'minerals', 'together', 'and', 'see', 'which', 'one', 'scratches', 'the', 'other', '.']
- **BERT Base Cased:** ['R', '##ub', 'the', 'minerals', 'together', 'and', 'see', 'which', 'one', 'scratch', '##es', 'the', 'other', '.']
- **XLNET Base Cased:** [' ', 'Rub', 'the', 'minerals', 'together', 'and', 'see', 'which', 'one', 'scratches', 'the', 'other', '.']

In the above examples, the lengths are 12, 14 and 13. We observe that the words *Rub* and *scratches* are tokenized in a different way, which produced different lengths.

4 EXPERIMENTS

In this section, we describe the experimental setup.

[2]classification

4.1 Experimental Setup

There are 135 questions in the SciEntBank train dataset. Each question is a tuple composed of: question, model answer, student answer, and label. We divided this dataset into train and validation set. As question-blind division of the dataset could bias the final trained model towards one (or a number of) specific question, we randomly selected 20% of the tuples from each question for the validation set and 80% for the train set. The test sets were provided separately by SemEval-2013. BERT and XLNet were fine-tuned on the SemEval train datasets, and tested on the test sets.

We conducted our experiments using the max sequence lengths of 165, 185 and 175 (considering the [CLS] and [SEP[3]] tokens) for BERT Base uncased, BERT Base cased and XLNET cased respectively. These parameter values were chosen based on the longest tokenized answer in the dataset.

4.2 Training the Grader with BERT and XLNET

Our hyperparameters for all the 2-way, 3-way and 5-way grading were experimentally selected and are as follows:

- Epochs = 10
- Dropout probability for all the layers = 0.1
- Warmup Proportion = 0.1
- Mini Batch size = 16
- Learning rate = 5e-6 for BERT and 5e-5 for XLNET

To avoid overfitting/underfitting, we employed one of the most common regularization techniques, which is to stop the training process on a set of initial epochs before full convergence. This is done in combination with the dropout technique. We set the number of epochs to 10 experimentally and trained our model for the full number of epochs. Then, we analyzed the flow of loss change per epoch to control the overfitting and underfitting. In the end, we evaluated our models on the validation sets and stopped the training process before the occurrence of any of these two problems. We stopped the training at the 2-4 epoch for all the models.

4.3 Evaluation Measures

We evaluate our ASAG system using the SemEval-2013 challenge measures:

- **Accuracy (ACC):** Proportion of correctly graded answers.
- **Macro Average F1 score (M-F1):** Precision, recall and F1 scores are calculated independently for each grade label and then averaged over all grade labels.
- **Weighted Average F1 score (W-F1):** Weighted average of M-F1. It takes into account the size of the classes related to each grade label.

5 RESULTS

In this section, we report the results of BERT Base uncased, BERT Base cased and XLNET Base cased. Then, we compare the performance of BERT and XLNET with the best graders in the SOTA. We do not consider the question in our ASAG task because BERT and XLNET classifiers only accepts two input sequences (student answer and model answer in our case).

5.1 Proposed Graders

The results from both experiments in all our proposed models are provided in Figure 2.

As shown in Figure 2, the results of the 2-way experiments show that BERT Base uncased is slightly better than the other two models in the S→M configuration. As expected, all evaluation measures indicate a stronger performance for TUA than TUQ and TUD.

XLNET Base cased for the 3-way task dominates almost in all the datasets in terms of F1-Macro and F1-weighted. For Acc, BERT Base uncased and XLNET Base cased behave almost in an identical manner. Overall, the results obtained with the TUD dataset are comparable with that of TUA.

Finally for the 5-way task (S→M), the performance of grading using BERT Base cased model is always better than other models. BERT Base cased and XLNET Base cased behave similarly to a certain extent in all the datasets. Overall, the results for the 5-way task are not as high as for the other tasks. However, they are still competitive with the SOTA (we return to these results in section 5.2).

Overall, all the proposed models seem to have better performance for TUA. For TUQ and TUD, the models behave roughly in the same way.

[3] separator

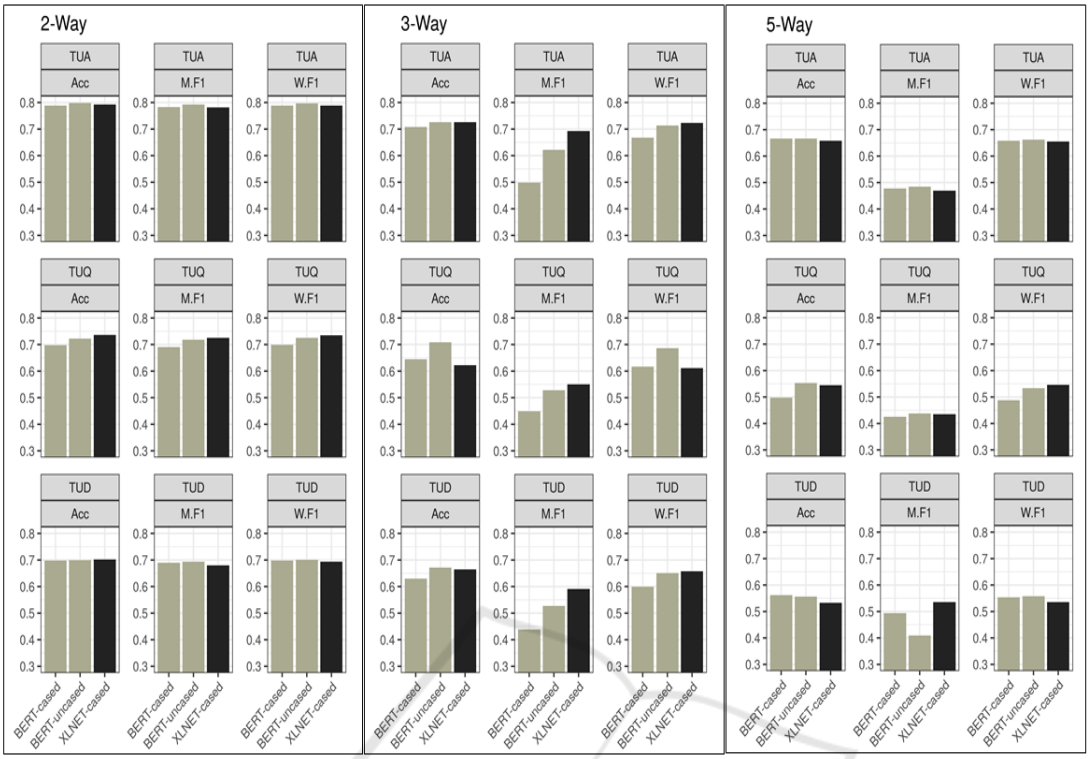

Figure 2: Comparing BERT and XLNET in the context of S→M for 2way, 3way and 5way task.

5.2 Comparison of Proposed Graders with SOTA

In this subsection, we compare the performance of all our proposed models to the other systems in SOTA. Tables 1, 2 and 3 report the results for 2-way, 3-way and 5-way tasks respectively. Regarding the SemEval-2013 (Dzikovska et al., 2013) systems, it should be noted that we include in the Tables 1, 2 and 3 only the best performing systems. In what follows, we explain the differences and similarities of our results for all the tasks. Given that (Saha et al., 2018) is the best performing system, most of our comparisons involve with its two main configurations (with or without questions (+/- Q)):

- 2-way
  - TUA: BERT Base uncased is slightly better than (Saha et al., 2018) (+Q) in terms of Acc, M-F1 and W-F1 scores. XLNET Base cased and (Saha et al., 2018) obtain a similar accuracy. If we compare our models with that of (Saha et al., 2018) (-Q) (as we do not consider the question in our model), all our proposed models outperform the SOTA.
  - TUQ: (Saha et al., 2018) (-Q) is better than all our proposed models in terms of ACC, M-F1 and W-F1 scores. Only XLNET Base cased comes close.
  - TUD: Again (Saha et al., 2018) (-Q) achieved the best results for all our metrics. Like for TUQ, the results obtained with XLNET Base cased are similar.
- 3-way
  - TUA: For all the evaluation measures, (Sung et al., 2019) achieved the best results in the SOTA. Except (Sung et al., 2019), BERT Base uncased and XLNET Base cased are the best in terms of accuracy. When not considering (Sung et al., 2019), XLNET Base cased performed better than all the other systems in the SOTA for the other two measures.
  - TUQ: BERT Base uncased performed the best based on all the measures, except when compared to (Sung et al., 2019) for M-F1.
  - TUD: For this test set, we achieved the best results in the SOTA with XLNET Base cased by all the measures. Followed by this model, BERT Base uncased is slightly better than XLNET Base cased only for the accuracy. In terms of W-F1, BERT Base uncased can be ranked second after XLNET Base cased.

Table 1: Comparison of the proposed system with SOTA on 2-way SciEntBank dataset: The highlighted numbers are the best in the SOTA.

	TUA			TUQ			TUD		
	Acc	**M-F1**	**W-F1**	**Acc**	**M-F1**	**W-F1**	**Acc**	**M-F1**	**W-F1**
COMET (Ott et al., 2013)	0.774	0.768	0.773	0.603	0.579	0.597	0.676	0.67	0.677
ETS (Heilman and Madnani, 2013)	0.776	0.762	0.77	0.633	0.602	0.622	0.627	0.543	0.574
SOFTCARDINALITY (Jimenez et al., 2013)	0.724	0.715	0.722	0.745	0.737	0.745	0.711	0.705	0.712
Sultan et. al. (Sultan et al., 2016)	0.708	0.676	0.69	0.705	0.678	0.695	0.712	0.703	0.712
Graph (Ramachandran and Foltz, 2015)	-	0.644	0.658	-	-	-	-	-	-
MEAD (Ramachandran and Foltz, 2015)	-	0.631	0.645	-	-	-	-	-	-
TF+SF [-question] (Saha et al., 2018)	0.779	0.771	0.777	**0.749**	**0.738**	**0.747**	0.708	0.690	0.702
TF+SF [+question] (Saha et al., 2018)	0.792	0.785	0.791	0.702	0.685	0.698	**0.719**	**0.708**	**0.717**
Mavarniya (Marvaniya et al., 2018)	-	0.773	0.781	-	-	-	-	-	-
BERT Base uncased	**0.798**	**0.792**	**0.797**	0.723	0.718	0.724	0.699	0.693	0.7
BERT Base cased	0.79	0.783	0.788	0.697	0.690	0.698	0.698	0.689	0.697
XLNET Base cased	0.792	0.781	0.788	0.736	0.724	0.734	0.702	0.679	0.693

Table 2: Comparison of the proposed system with SOTA on 3-way SciEntBank dataset: The highlighted numbers are the best in the SOTA.

	TUA			TUQ			TUD		
	Acc	**M-F1**	**W-F1**	**Acc**	**M-F1**	**W-F1**	**Acc**	**M-F1**	**W-F1**
COMET (Ott et al., 2013)	0.713	0.64	0.707	0.546	0.38	0.522	0.579	0.404	0.55
ETS (Heilman and Madnani, 2013)	0.72	0.647	0.708	0.583	0.393	0.537	0.543	0.333	0.461
SOFTCARDINALITY (Jimenez et al., 2013)	0.659	0.555	0.647	0.652	0.469	0.634	0.637	0.486	0.62
Sultan et. al. (Sultan et al., 2016)	0.604	0.443	0.569	0.642	0.455	0.615	0.626	0.451	0.603
Graph (Ramachandran and Foltz, 2015)	-	0.438	0.567	-	-	-	-	-	-
MEAD (Ramachandran and Foltz, 2015)	-	0.429	0.554	-	-	-	-	-	-
TF+SF [-question] (Saha et al., 2018)	0.718	0.666	0.714	0.613	0.491	0.628	0.632	0.479	0.611
TF+SF [+question] (Saha et al., 2018)	0.718	0.657	0.711	0.653	0.489	0.636	0.640	0.452	0.61
Marvaniya (Marvaniya et al., 2018)	-	0.636	0.719	-	-	-	-	-	-
Sung et. al. (Sung et al., 2019)	**0.759**	**0.72**	**0.758**	0.653	**0.575**	0.648	0.638	0.579	0.634
BERT Base uncased	0.726	0.622	0.714	**0.708**	0.528	**0.686**	**0.672**	0.528	0.6514
BERT Base cased	0.707	0.5	0.667	0.645	0.45	0.616	0.63	0.438	0.6
XLNET Base cased	0.726	0.7	0.723	0.622	0.55	0.61	0.665	**0.6**	**0.657**

- 5-way
  - TUA: BERT Base uncased performed better than all the other systems in the SOTA in terms of all the evaluation measures. Also, BERT Base cased and XLNET Base cased performed well in terms of accuracy and W-F1 when compared to the other systems.
  - TUQ: Similar to TUA, BERT Base uncased dominated the SOTA followed by XLNET Base cased. It should be noted the W-F1 for XLNET Base cased is slightly better than that of BERT Base uncased. However, BERT Base cased performs better in terms of M-F1.
  - TUD: The results from all our proposed models are significantly better than those of the SOTA. Among our models, BERT Base cased is the top model in terms of accuracy and W-F1. XLNET Base cased is better in terms of M-F1.

6 DISCUSSION

According to the results described in Section 5, our findings suggest that BERT and XLNET graders achieve better or competitive results compared to the SOTA. This is true especially on the TUA test set for all the tasks, except for (Sung et al., 2019) on the 3-way task. For all the other test sets in all the tasks, our proposed models perform better in terms of all the measures.

For the 2-way task on TUQ, XLNET Base cased compete with the best system (Saha et al., 2018) in the SOTA. The difference is around 0.1. The same is true for TUD, except the M-F1. Overall, the value of all the evaluation metrics are almost equal or above 0.7 using our proposed models.

For the 3-way grading task, XLNET Base cased grade well generally. This is highlighted in TUA and

Table 3: Comparison of the proposed system with SOTA on 5-way SciEntBank dataset: The highlighted numbers are the best in the SOTA.

	TUA			TUQ			TUD		
	Acc	**M-F1**	**W-F1**	**Acc**	**M-F1**	**W-F1**	**Acc**	**M-F1**	**W-F1**
COMET (Ott et al., 2013)	0.6	0.441	0.598	0.437	0.161	0.299	0.421	0.121	0.252
ETS (Heilman and Madnani, 2013)	0.643	0.478	0.64	0.432	0.263	0.411	0.441	0.38	0.414
SOFTCARDINALITY (Jimenez et al., 2013)	0.544	0.38	0.537	0.525	0.307	0.492	0.512	0.3	0.471
Sultan et. al. (Sultan et al., 2016)	0.489	0.3298	0.487	0.480	0.302	0.467	0.506	0.344	0.484
Graph (Ramachandran and Foltz, 2015)	-	0.372	0.458	-	-	-	-	-	-
MEAD (Ramachandran and Foltz, 2015)	-	0.379	0.461	-	-	-	-	-	-
TF+SF [-question] (Saha et al., 2018)	0.644	0.480	0.642	0.5	0.316	0.488	0.508	0.357	0.492
TF+SF [+question] (Saha et al., 2018)	0.629	0.472	0.630	0.506	0.376	0.471	0.51	0.342	0.486
Mavarniya (Marvaniya et al., 2018)	-	0.579	0.61	-	-	-	-	-	-
BERT Base uncased	**0.66**	**0.484**	**0.662**	**0.552**	**0.437**	0.533	0.557	0.41	**0.558**
BERT Base cased	**0.66**	0.478	0.658	0.5	0.424	0.487	**0.562**	0.5	0.552
XLNET Base cased	0.658	0.47	0.655	0.544	0.435	**0.545**	0.532	**0.535**	0.535

TUQ for all the measures. For TUQ, BERT Base uncased dominates the SOTA. It should be noted that the accuracy for BERT Base uncased is always the highest. Overall, the proposed models seem to improve the SOTA.

In the 5-way task, all our models work robustly compared to the SOTA. Among them, BERT Base uncased achieved the top values for all the evaluation measures. Close to BERT Base uncased, it is XLNET and BERT Base cased. XLNET Base cased even performs slightly better than our other models for TUD.

As we explore grading (or classification) from 2-way to 5-way, the performance becomes weaker. Although we improve the performance in 3-way and 5-way tasks, the value for all the evaluation measures (Acc, M-F1 and W-F1) does not exceed 0.75, especially in the 5-way task. Besides, most of the measures in the 5-way task show that the models are not robust in the SOTA (including the current study) and predicts the answers randomly to certain extent.

We explored the dataset to try to understand its challenges. One possible reason for lower classification results is that we found it difficult to differentiate classes like *correct* from *partially correct*, and *irrelevant* from *not in the domain*. In fact, we found some ambiguity in the labels associated to some answers, which adds to the complexity of grading. For instance, in the following example extracted from SciEntBank (Nielsen et al., 2008), it is a complicated to assign a *partially correct* label based only on the "scratch" concept:

- *Question:* "Georgia found one brown mineral and one black mineral. How will she know which one is harder?"
- *Model Answer:* "The harder mineral will leave a scratch on the less hard mineral. If the black mineral is harder, the brown mineral will have a scratch."
- *Student Answer:* "The one with a scratch."

Consequently, it seems likely that the performance of machine scoring systems would improve if the grade labels were more clearly defined and annotated.

Finally, the type of the test sets adds to the complexity of the task. The SOTA and our evaluation results show that TUA is the least difficult dataset, followed by TUQ and TUD. In fact, the challenge seems to increase with unknown questions or when there is a need for domain adaptation.

7 CONCLUSION AND FUTURE WORK

In this paper, we proposed approaches based on the BERT and XLNET classifiers for the ASAG task. Overall, we showed the approaches can be considered as better or comparable to SOTA satisfactory graders on the SemEval-2013 SciEntBank dataset. Our findings further suggest that XLNET and BERT seem to be strong baselines for ASAG for the 3way and 5way tasks respectively, with the exception of the 2-way TUQ and TUD conditions.

These overall good results of language models such as BERT and XLNet on ASAG seem to indicate that modern language models succeed in building a semantic representation of student answers and model answers and classifying them correctly. BERT and XLNet seem to either equal or outperform the results obtained with human engineered features. These features, however, help to explain classification results, a task that is more difficult with transformer-based models. For pedagogical purposes, the ability

to explain grading results is a must and further work should be directed towards the combination of modern language models with explainable capabilities.

The current work has a number of limitations. One of the most important is that our experiment was carried out on only one dataset. Other datasets like BEETLE or DT-Grade (Banjade et al., 2016) could also be used to confirm the promising characteristics of BERT and XLNET for ASAG. Another limitation is that that we did not use the largest BERT model due to limited computing power. We also note that we could not go beyond 10 epochs, and as a result adjusted our early stopping based on the observation on this 10 epoch experiment.

We plan to address the above-mentioned limitations in future work. We also intend to explore ensembling BERT with other classifiers to boost the grading performance, especially by considering features that were successful in the SOTA.

ACKNOWLEDGMENTS

The current research is supported by an INSIGHT grant from the Social Sciences and Humanities Research Council of Canada (SSHRC).

REFERENCES

Badger, E. and Thomas, B. (1992). Open-ended questions in reading. *Practical assessment, research & evaluation*, 3(4):03.

Banjade, R., Maharjan, N., Niraula, N. B., Gautam, D., Samei, B., and Rus, V. (2016). Evaluation dataset (dt-grade) and word weighting approach towards constructed short answers assessment in tutorial dialogue context. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 182–187.

Burrows, S., Gurevych, I., and Stein, B. (2015). The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, 25(1):60–117.

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., and Bordes, A. (2017). Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*.

Dagan, I., Glickman, O., and Magnini, B. (2005). The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop*, pages 177–190. Springer.

Dai, Z., Yang, Z., Yang, Y., Cohen, W. W., Carbonell, J., Le, Q. V., and Salakhutdinov, R. (2019). Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.

Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dzikovska, M. O., Moore, J. D., Steinhauser, N., Campbell, G., Farrow, E., and Callaway, C. B. (2010). Beetle ii: a system for tutoring and computational linguistics experimentation. In *Proceedings of the ACL 2010 System Demonstrations*, pages 13–18. Association for Computational Linguistics.

Dzikovska, M. O., Nielsen, R. D., Brew, C., Leacock, C., Giampiccolo, D., Bentivogli, L., Clark, P., Dagan, I., and Dang, H. T. (2013). Semeval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge. Technical report, NORTH TEXAS STATE UNIV DENTON.

Goldberg, Y. (2017). Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies*, 10(1):1–309.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

Heilman, M. and Madnani, N. (2013). Ets: Domain adaptation and stacking for short answer scoring. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, volume 2, pages 275–279.

III, H. D. (2009). Frustratingly easy domain adaptation. *arXiv preprint arXiv:0907.1815*.

Jimenez, S., Becerra, C., and Gelbukh, A. (2013). Softcardinality: Hierarchical text overlap for student response analysis. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, volume 2, pages 280–284.

Jimenez, S., Gonzalez, F., and Gelbukh, A. (2010). Text comparison using soft cardinality. In *International symposium on string processing and information retrieval*, pages 297–302. Springer.

Mantecon, J. G. A., Ghavidel, H. A., Zouaq, A., Jovanovic, J., and McDonald, J. (2018). A comparison of features for the automatic labeling of student answers to open-ended questions. In *TEMPLATE'06, 1st International Conference on Template Production*. International Educational Data Mining Conference.

Marvaniya, S., Saha, S., Dhamecha, T. I., Foltz, P., Sindhgatta, R., and Sengupta, B. (2018). Creating scoring rubric from representative student answers for improved short answer grading. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 993–1002. ACM.

Nielsen, R. D., Ward, W. H., Martin, J. H., and Palmer, M. (2008). Annotating students' understanding of science concepts. In *LREC*.

Ott, N., Ziai, R., Hahn, M., and Meurers, D. (2013). Comet: Integrating different levels of linguistic modeling for meaning assessment. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, volume 2, pages 608–616.

Papineni, K., Roukos, S., Ward, T., and Zhu, W. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Ramachandran, L. and Foltz, P. (2015). Generating reference texts for short answer scoring using graph-based summarization. pages 207–212.

Riordan, B., Horbach, A., Cahill, A., Zesch, T., and Lee, C. M. (2017). Investigating neural architectures for short answer scoring. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 159–168.

Roy, S., Rajkumar, A., and Narahari, Y. (2018). Selection of automatic short answer grading techniques using contextual bandits for different evaluation measures. *International Journal of Advances in Engineering Sciences and Applied Mathematics*, 10(1):105–113.

Saha, S., Dhamecha, T. I., Marvaniya, S., Sindhgatta, R., and Sengupta, B. (2018). Sentence level or token level features for automatic short answer grading? use both. In *International Conference on Artificial Intelligence in Education*, pages 503–517. Springer.

Sakaguchi, K., Heilman, M., and Madnani, N. (2015). Effective feature integration for automated short answer scoring. In *Proceedings of the 2015 conference of the North American Chapter of the association for computational linguistics: Human language technologies*, pages 1049–1054.

Sultan, M. A., Salazar, C., and Sumner, T. (2016). Fast and easy short answer grading with high accuracy. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1070–1075.

Sung, C., Dhamecha, T. I., and Mukhi, N. (2019). Improving short answer grading using transformer-based pretraining. In *International Conference on Artificial Intelligence in Education*, pages 469–481. Springer.

Taylor, W. L. (1953). "cloze procedure": A new tool for measuring readability. *Journalism Bulletin*, 30(4):415–433.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., and Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.