Automatic Student Scoring - Module I

Applications on Language Technologies

Edgar Andrés Santamaría and Mohamed Yassin Akhayat

Basque Country University UPV-EHU
{eandres011,makhayat001}@ikasle.ehu.eus

Abstract. In the current work we present a module for a tutoring system, the automatic scoring system. The general system consists in grading and assessment modules, in this way we aim to support the knowledge acquisition process.The system decide whether certain answer is correct or incorrect given a reference answer. Using Semantic Similarity the grade is assigned to each answer.

Keywords: Tutoring system · RNN · Attention.

References

1. Al Emran, M., Shaalan, K.: A survey of intelligent language tutoring systems. In: 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI). pp. 393–399. IEEE (2014)
2. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
3. Mohler, M., Bunescu, R., Mihalcea, R.: Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. pp. 752–762. Association for Computational Linguistics (2011)
4. Rocktäschel, T., Grefenstette, E., Hermann, K.M., Kočiskỳ, T., Blunsom, P.: Reasoning about entailment with neural attention. arXiv preprint arXiv:1509.06664 (2015)
5. Wang, S., Jiang, J.: Learning natural language inference with lstm. arXiv preprint arXiv:1512.08849 (2015)
6. Zilio, L., Wilkens, R., Fairon, C.: Using NLP for enhancing second language acquisition. In: Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017. pp. 839–846. INCOMA Ltd., Varna, Bulgaria (Sep 2017). https://doi.org/10.26615/978-954-452-049-6$_1$07, https://doi.org/10.26615/978-954-452-049-6_107

1 Introduction

In the current work we aim to propose a suitable work for an automatic scoring system. Therefore our main objective consists on create a system able to decide whether a response is correct or incorrect given a reference answer. Finally the answer will be graded according to the semantic similarity between the answer and the reference.

The system will be able to perform automatic evaluation of the students, this crucial part of tutoring systems [1], student grades constitute the main metric to define the knowledge acquisition performance. Finally we aim as secondary objective to provide suitable framework for the students assessment, tutoring systems must provide feedback to enhance the knowledge acquisition [6].

The current work is a decision system that aids evaluators managing the students grading and feedback process. The system align answer and reference at sentence level , similar idea to Graph based approach [3], in that case they proposed graphs to handle the semantic similarity between words, in this case we substitute with a Language Model (LM), the LM proposed is based on Recurrent Neural Networks (RNNs) [5], this approach allow the system modeling word sequences, the direct alignment between Graphs in the current approach is handled with the attention mechanism, this consists on direct linking the representation steps of source and target sentences by trainable weights [4]. Finally the system learns to determine if the answer is correct or incorrect, the grade of each answer is obtained from a semantic similarity test based on pre-trained BERT system [2].

2 The System

In this case the proposed approach is based on Deep Learning, and follows the Architecture presented in figure 1, there exists two RNNs (Gt and Gs) that represent the student answer and the reference respectively. The reference is multiplied with an attention component (Attn), the result is stacked with the student answer representation. The stack is feed as input of the third RNN (Gm) that will learn the attention between both sentence representation. Finally we apply a fully connected layer Softmax to perform the binary decision. To achieve the numerical grade a pre-trained BERT model is used, the result [0..5] is scaled to [0..10] depending on the correctness result.

Input Output of the System The input for the system consists on a JSON format with a reference answer and a student answer. The results of the scoring system will be returned, in one hand as a plain text report with correctness results. In the other hand The system returns full report over the attention mechanism, the correlation and deviation statistics to insight the similarity process.

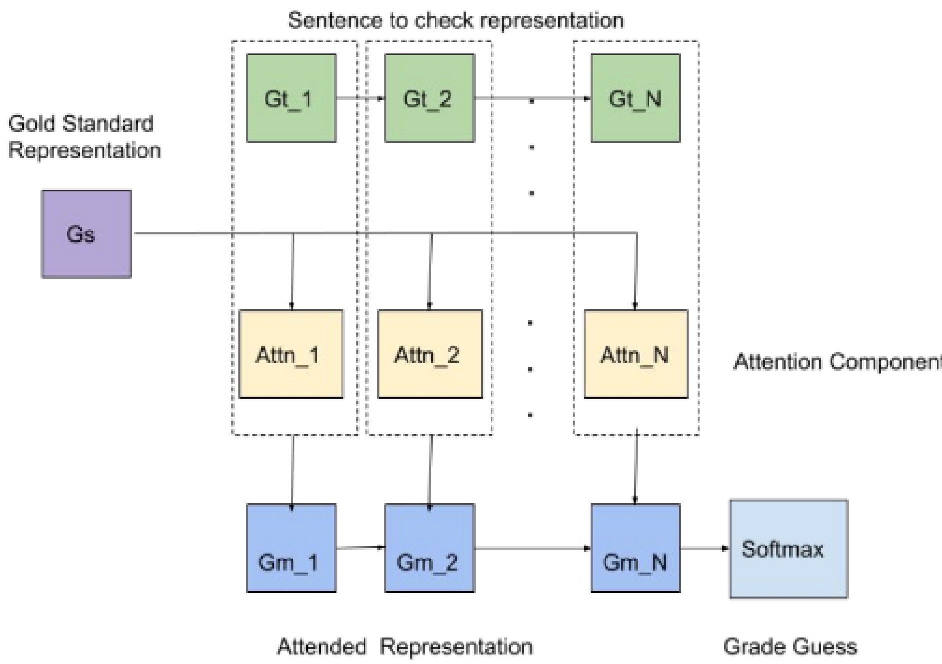

Fig. 1. Architecture for binary scoring.

Training settings For the current experiment we used the data provided by semeval-2013 Student Response Analysis, in it's 2-way task scenario, we were provided reference answer, student answer, grade (binary) and question for each example. For training purposes we processed the given (.XML) format into the required JSON format, and divided the data into 7448 train, 2482 development and 2484 test samples.

The training was performed in 25 epochs applying early stopping technique patience 5. The proposed RNNs were three GRU layers with 20 units, and the word embedding used to perform the Lexical LM was Glove_6B_50D. The stochastic gradient descent was applied using L2 regularizer with 0.001 lambda and 16 batch size. For performance purposes the sentence length was padded into 20 tokens.

3 Results

In the following figure 2 we can see the training performance, in this case we can see that the system avoids overfitting and tend to good score on development. The evaluation of the system was performed into the unseen test set and reached 0.8 accuracy, so we perform good decision in the student answer classification task.

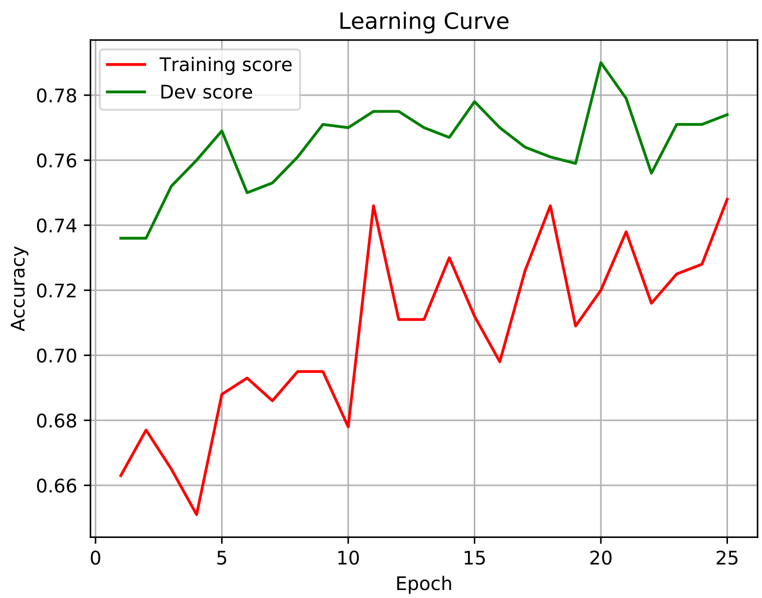

Fig. 2. Accuracy curve.

4 Conclusions

In the current approach we generated a module able to automatically classify students answer into correct / incorrect, also grade the student performed score based on semantic similarity. The proposed framework enables building new modules on top to reach our further objectives.

Future work First of all we plan to improve the evaluation report, and fine tune all the system. Finally we will propose another module II to handle the students assessment and feedback.

Using the model scoop, we can fusion the work of two models by adding more RNN layers on top, the rule of the added RNN layer will be to identify the way answers are written, the embedding layers will generate a sort of vocabulary to generate feedback on students answers,