Automatic Learner Assessment on Text Understanding: Development of an Automated short answer grading system

Proposer(s) / Proposatzailea(k):

Itziar Aldabe

Oier Lopez de Lacalle

Contact / Kontaktua:

itziar.aldabe@ehu.eus

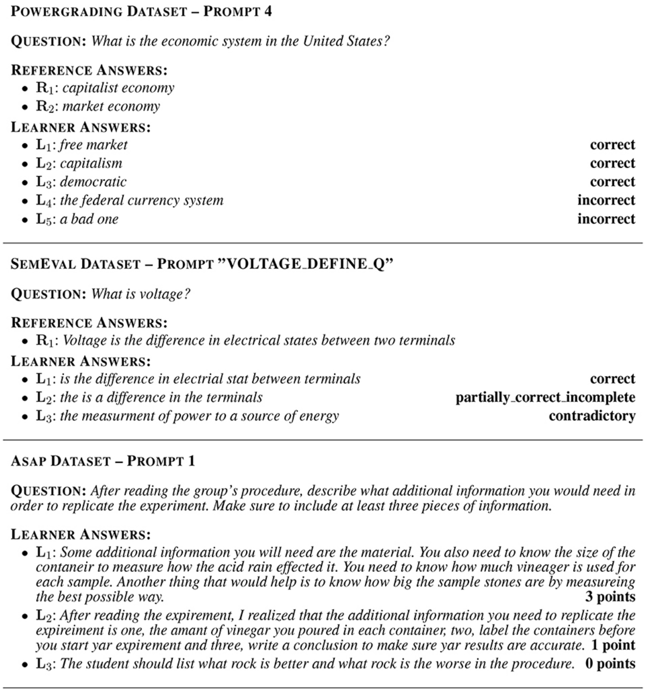
oier.lopezdelacalle@ehu.eus

Description / Deskribapena

Learner assessment for reading refers to the evaluation of individual's ability to understand the text, and it is an important application in the area of automatic educational assessment. In particular, assessment of learner's short natural language responses have been recognized as a tool to perform a deeper assessment of the student's knowledge than, for example, multiple choice questions. They require active formulation instead of just selecting the correct answer from a set of alternative, i.e., they test production instead of recognition.

In a traditional classroom-setting, answers to such exercises are manually scored by a teacher, but in recent years, their automatic scoring has received growing attention. Automated short answer grading (ASAG) is a task from the field of educational natural language processing (NLP). In this task, a free-text answer written by students should be automatically assigned a score or correctness label in the same way as a human teacher would do.

Answers to short-answer questions have a typical length between a single phrase and two to three sentences, as shown in the figure below (*source*: Horbach and Zesch (2019)). Typically, the input consist in a **question**, **learner answers**, **reference answers**, and a **scoring label** (which can be numeric or categorical).



POWERGRADING DATASET – PROMPT 4

QUESTION: *What is the economic system in the United States?*

REFERENCE ANSWERS:
- R_1: *capitalist economy*
- R_2: *market economy*

LEARNER ANSWERS:
- L_1: *free market* **correct**
- L_2: *capitalism* **correct**
- L_3: *democratic* **correct**
- L_4: *the federal currency system* **incorrect**
- L_5: *a bad one* **incorrect**

SEMEVAL DATASET – PROMPT ”VOLTAGE_DEFINE_Q”

QUESTION: *What is voltage?*

REFERENCE ANSWERS:
- R_1: *Voltage is the difference in electrical states between two terminals*

LEARNER ANSWERS:
- L_1: *is the difference in electrial stat between terminals* **correct**
- L_2: *the is a difference in the terminals* **partially_correct_incomplete**
- L_3: *the measurment of power to a source of energy* **contradictory**

ASAP DATASET – PROMPT 1

QUESTION: *After reading the group's procedure, describe what additional information you would need in order to replicate the experiment. Make sure to include at least three pieces of information.*

LEARNER ANSWERS:
- L_1: *Some additional information you will need are the material. You also need to know the size of the contaneir to measure how the acid rain effected it. You need to know how much vineager is used for each sample. Another thing that would help is to know how big the sample stones are by measureing the best possible way.* **3 points**
- L_2: *After reading the expirement, I realized that the additional information you need to replicate the expireiment is one, the amant of vinegar you poured in each container, two, label the containers before you start yar expirement and three, write a conclusion to make sure yar results are accurate.* **1 point**
- L_3: *The student should list what rock is better and what rock is the worse in the procedure.* **0 points**

Early grading work relied on patterns manually extracted from expert-provided reference answers (Mitchell et al., 2002). With advances in NLP, various approaches relied on measuring the semantic overlap between the student answer and the model answer using knowledge- and corpus-based similarity measures (Mohler et al., 2011, Burrows et al., 2015; Roy et al., 2015). In contrast, recent advances in distributional semantics and deep learning methods, a variety neural architectures have been proposed obtaining significant improvement of the state-of-the-art in ASAG (Kumar et al. ,2019) .

Goals / Helburuak

The **main objective** of the project is to analyse and implement a deep learning approaches to short answer grading. The key objectives are the following:

1. Analysis of the state of the art techniques for developing automatic grading systems
2. Design of a deep learning architecture provides scoring labels (numeric or categorical) for given short-answers
3. Implementation and evaluation of the model on publicly available datasets such as (Mohler et al. (2011), (Dzikovska et al., 2013), (Shermis, 2015)

Requirements / Betebeharrak

English. Machine learning. Good programming skills, basic math skills.

Although it is not a requirement, taking the course "**Seminar on language technologies. Deep Learning**" (see below) will allow the student to accomplish more ambitious goals. Contact us for further details.

The dissertation can be written in Basque, English or Spanish.

Framework / Esparrua

NLP applications for education

Python, pytorch/tensorflow

Tasks and plan / Atazak eta plana

- Analyze ASAG datasets.
- Analyze the state of the art in ASAG and related topics
- Design and implement an ASAG system.
- Test and evaluate the implemented algorithm on public datataset.
- Analyse the output of the system to 1) perform an error analysis and 2) purpose possible improvements.
- Write up the report.

The student can start any time, but sensible plan can be the following (month below approximate):

- Dec-Jan: Study literature
- Jan: Attend course "Seminar on language technologies. Deep Learning"
- Mar-May: Development and experiments
- June: Write down and presentation

References

Burrows S., Gurevych I., Stein B. (2015). The Eras and Trends of Automatic Short Answer Grading. International Journal of Artificial Intelligence in Education, 25(1):60-117, 2015.

Dzikovska M., Nielsen R., Brew C., Leacock C., Giampiccolo D, Bentivogli L., Clark P., Dagan I, and Hoa Trang Dang. (2013). Semeval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge. In *SEM 2013

Horbach Andrea, Zesch Torsten. The Influence of Variance in Learner Answers on Automatic Content Scoring . Frontiers in

Education, vol 4. 2019. https://doi.org/10.3389/feduc.2019.00028

Kumar V., Joshi N., Mukherjee A., Ramakrishnan G., Jyothi P. (2019). Cross-Lingual Training for Automatic Question Generation. In ACL 2019

Mitchell T., Russel T., Broomhead P., Aldridge N. (2002). Towards Robust Computerised Marking of Free-Text Responses. In Proocedings of the 6th International Computer Assisted Assessment Conference.

Mohler M., Bunescu R., Mihalcea R. (2011). Learning to Grade Short Answer Questions using Semantic Similarity Measures and Dependency Graph Aligments. In ACL 2011

Roy S., Narahari Y., Deshmukh O.D. (2015) A Perspective on Computer Assisted Assessment Techniques for Short Free-Text Answers. In Computer Assisted Assessment.

Shermis M.D. (2015) Contrasting state-of-the-art in the machine scoring of short-form constructed responses. Educational Assessment 20(1): 46-65.

RECOMMENDED COURSE: Seminar on Language Technologies. Deep learning.

Deep Learning neural network models have been successfully applied to natural language processing. These models are able to infer a continuous representation for words and sentences, instead of using hand-engineered features as in other machine learning approaches. The seminar will introduce the main deep learning models used in natural language processing, allowing the students to gain hands-on understanding and implementation of them in Tensorflow.

Topics

- Introduction to machine learning and NLP with Tensorflow Deep learning
- Word embeddings
- Language modeling and recurrent neural networks
- Convolutional neural networks
- Attention mechanisms

Prerequisite. Basic programming experience, a university-level course in computer science and experience in Python. Basic math skills (algebra or pre-calculus) are also needed.