Short Answer Scoring

Edgar Andrés Santamaría
Mohamed Yassin Akhayat

Short Answer Scoring

Introduction

Article: Abstract, Introduction & Conclusion

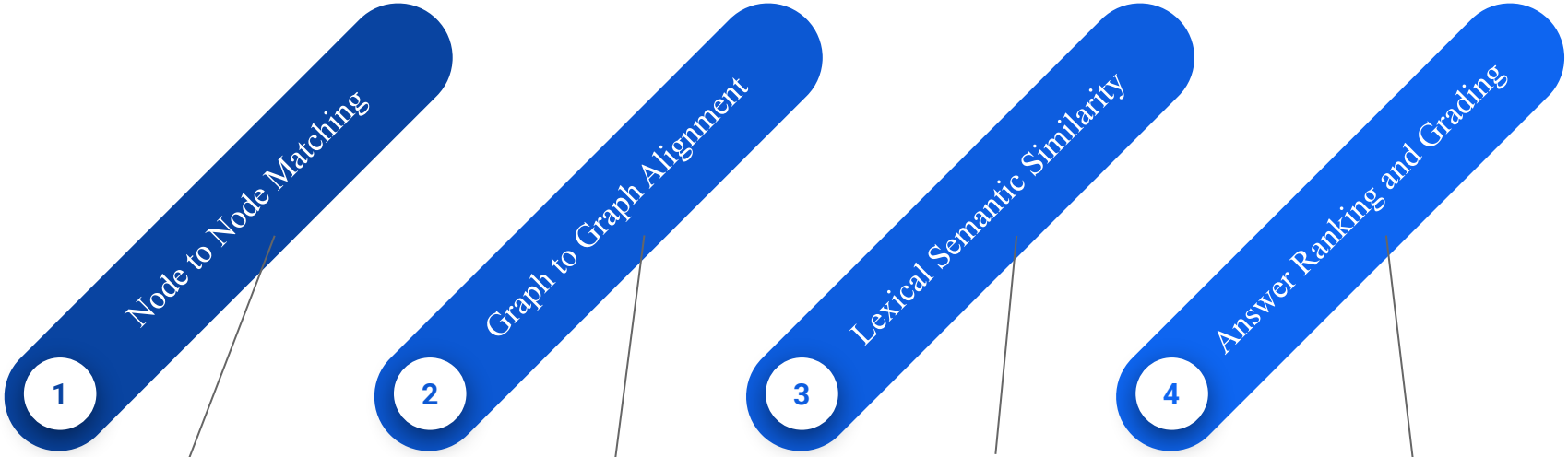
“ An automatic short answer grading system is one that automatically assigns a grade to an answer provided by a student, usually by comparing it to one or more correct answers. ”

“ we seek answers to the following questions. First, to what extent can machine learning be leveraged to improve upon existing approaches to short answer grading. Second, does the dependency parse structure of a text provide clues that can be exploited to improve upon existing BOW methodologies? ”

the model is based on a **supervised learning** method, with a labeled data set for training (handwritten as golden standard examples),

the results are based on how much better the **alignment** is done, and also the **probabilistics measures** for assigning grades.

Article : Main Content & Tools



in this process we try to implement a model for chunking, we can assign scores for the pairs of chunks for the two sentences.

in this part, as input we take the chunking scoring scoring and we try to align with a similarity algorithm

this part is for taking into account the semantic weights of our chunks in the alignment process

finally we can assign a final score to the answer based on a probabilistic method and assembling all the the scores obtained in previous steps

Related Content & Group Proposition

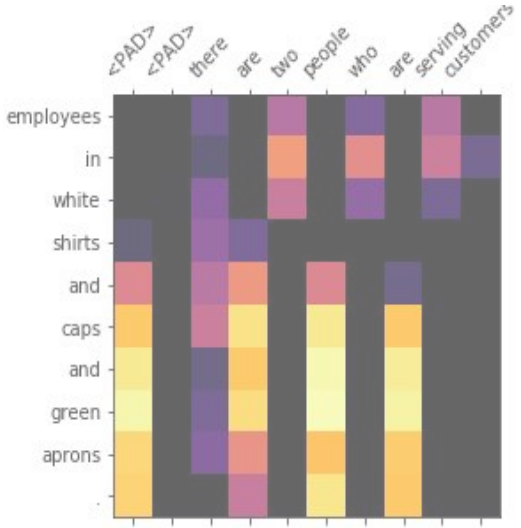
1. **Using NLP Enhancing Second Language Acquisition (SMILLE)**

We propose the implementation of this model to enhance SLA proficiency. SMILLE is a system that helps the user to notice linguistic content of the target language in any text from the web by highlighting (i.e. enhancing) language structures in context, while also offering the possibility of looking up meaning and word class.

Related Content & Group Proposition

1. **Using NLP Enhancing Second Language Acquisition (SMILLE)**

This graph shows alignments between words in each sentence. We plan on trying to highlight the pieces of information that are wrong in the students sentence as compared to the sentence provided by the teacher by using a similar approach as the one presented in the graph.



Related Content & Group Proposition

2. Students Performance: **"Intelligent language tutoring system"**

- To fulfill this part in our project we can trace back the mistakes detected on the students answers (recording generated information as historical) and assign some related exercices.

TABLE III. THE FOUR TUTORING TACTICS FOR CIRCSIM-TUTOR.

Plan	Tactics
Tutor	Ask the student a set of questions
Give answer	Ask the student to demonstrate his/her answer.
Hint	Remind ("Remember that. . . ")
Acknowledge	4 possible cases (see below)

Related Content & Group Proposition

2. **Students Performance:** **“Intelligent language tutoring system”**

- As the article proposes, it is necessary to provide tutor knowledge about students in order to clearly define next learning steps as new exercise or insight weakness on the student development

Related Content & Group Proposition

2. **Students Performance:** **“Intelligent language tutoring system”**

- This system seems consistent since it can learn once deployed, and store historical feedbacks. Those could be used for developing score dashboards for teachers even insight students performance.

REFERENCES (1)

Al, E. M., Shaalan, K., & 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI). (September 01, 2014). A Survey of Intelligent Language Tutoring Systems. 393-399.

Mohler, M., Bunescu, R., & Mihalcea, R. (January 01, 2011). Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. *Annual Meeting of the Association for Computational Linguistics and ... Conference of the European Chapter of the Association for Computational Linguistics : Proceedings of the Conference, 49,* 752-762.

Zilio, L., Wilkens, R., Fairon, C., & 11th International Conference on Recent Advances in Natural Language Processing, RANLP 2017. (January 01, 2017). Using NLP for enhancing second language acquisition. *International Conference Recent Advances in Natural Language Processing, Ranlp,* 839-846.

Short Answer Scoring

Design

Useful Information: Understanding Technology

Deep Learning: This technology aims to generalize previous tasks (and others also), by using embedded matrix systems to achieve the desired result over passing forward and backward the data, and tuning the matrix systems on each step.

- multi process to achieve better alignment.
- multi process to assign grades and scores.

Useful Information: Understanding Tasks

Chunking: This task consist on defining desired parts for given textual features.

Classification: This task consists on assigning certain category or probability for given textual features.

Regression: This task consists on assigning certain category or probability for given numerical features.

Useful Information: Understanding Representation

Word Embeddings: dense vector representation of words based on large amounts of word corpora, this aims to represent numerically the word, constitutes language model (LM).

RNN: dense vector representation of sequences based on the provided corpora, this aims to represent numerically the word.

Feedback Students SoA update

1. This task consists on automatic extracting the chunks that compose both teacher and student answers.

2. Then align those extracted chunks.

3. Finally, rank those alignments to provide feedback to student remarking those wrong chunks, those scores will also be needed to calculate the final grading for the answer.

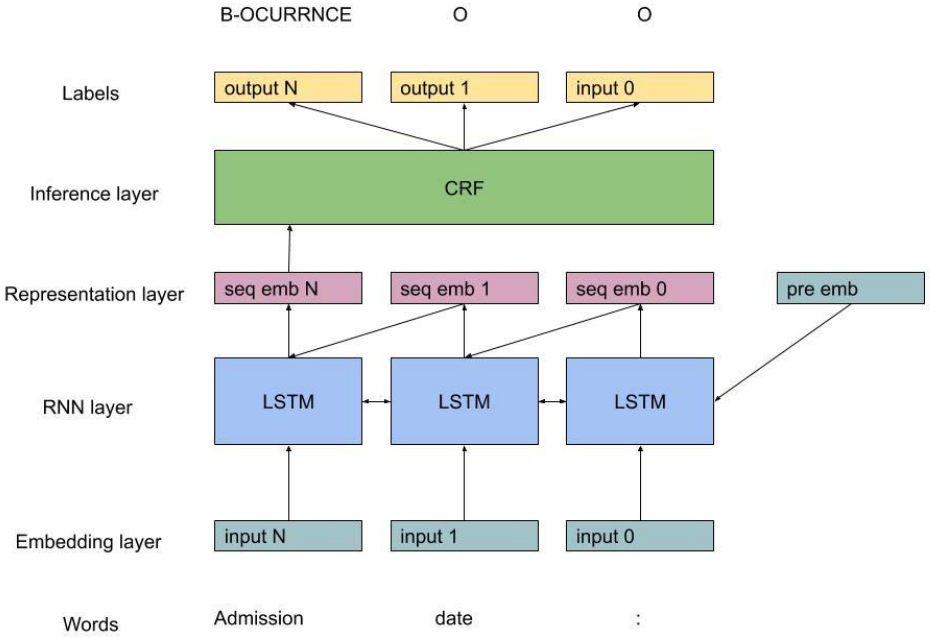
Chunking & FLAIR

Bi-LSTM + CRF (tagger)

1. The task consist on representing certain input sequence and guess its corresponding output sequence.

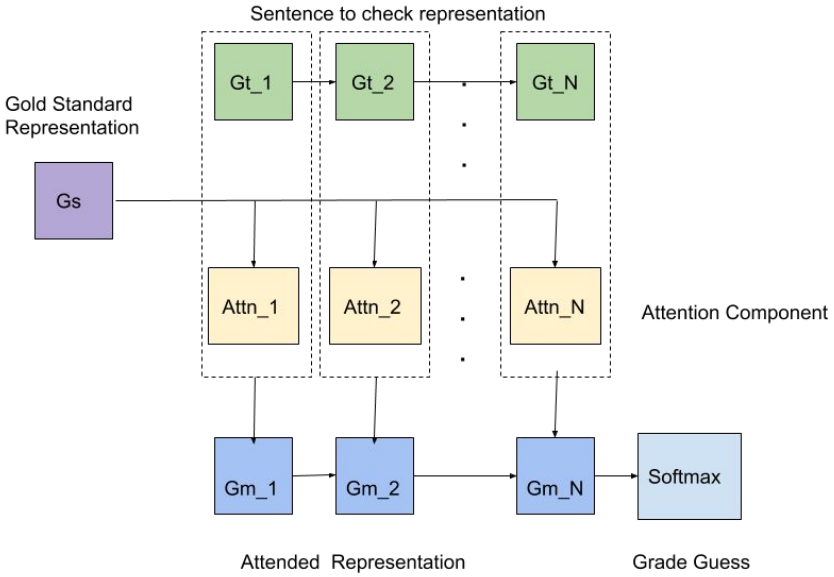
2. Representation based on bidirectional LSTM with pre-trained word embeddings.

3. Sequence guessing based on CRF, acts as stacked SVM to recognize chunks.



Alignment & SNLI model (Tags and Degrees)

3GRU + Attn + Softmax (Classifier)



1. 3 RNNs for representing the conjunct problem, first Gold standard, then sentence to check, both will be assembled for attended representation.

2. Attention component calculated at each stage (concat each Gt_N with Attn_N-1), that linked with Gs.

3. Calculate attended representation using Attn_N and Attn_N-1.(here we have word by word linked).

4. Finally, classify certain grade with the attended representation.

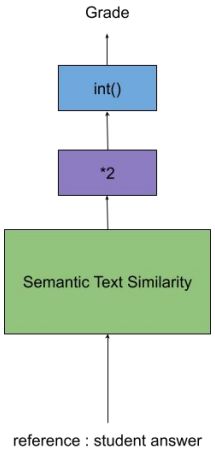
Grading Students SoA update

1. This task consists on automatic extracting certain grade for the student answer given the alignment scores.

2. Semantic similarity between both answers should be analyzed for this task.

Scoring & Semantic text similarity

external resource



1. The fine tuned Bert LM based model aids us normalizing the similarity between two sentences in range (0..5).

2. Now we calculate the expected grade multiplying by 2 and reporting as integer.

REFERENCES (2)

Edgar Andrés, IDENTIFICADOR AUTOMÁTICO DE RELACIONES TEMPORALES EN TEXTOS CLÍNICOS BASADO EN REDES NEURONALES, 2019 ***https://addi.ehu.es/handle/10810/37091***

Wang and Jiang. *Learning Natural Language Inference with LSTM, 2016*

Luong et al. *Effective Approaches to Attention-based Neural Machine Translation, 2015*

Rocktäschel *REASONING ABOUT ENTAILMENT WITH NEURAL ATTENTION, 2016*

Siddharth Narayanan, 2019, Semantic Similarity in Sentences and BERT *You can find me here !*

Find the model in the following link https://github.com/EdgarAndresSantamaria/similarity_deep_model

Short Answer Scoring

Process

Dataset (semeval - 2013)

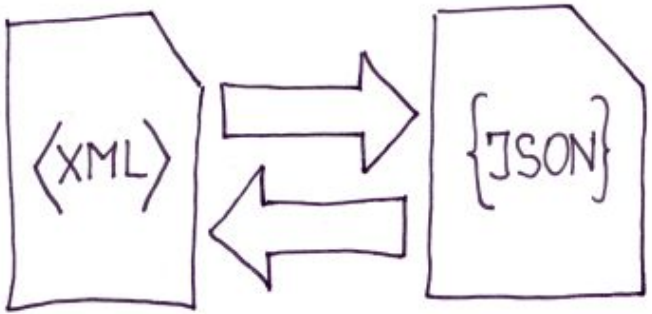
A set of Xml files, a file for each question, contains reference answers as the correct answer, and the students answers.

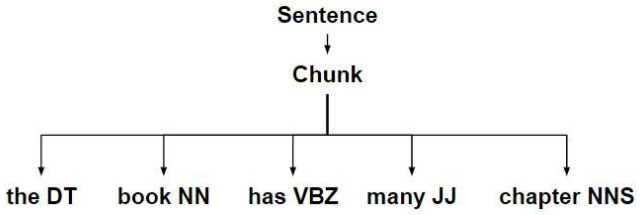
The dataset is a real experiment, so we could detect many orthograph **errors** due to the hand written samples.

Preprocessing

The most **important** process. in this part, we tried to feed our model a clean data as much as possible.

Applying NLP tools to get the proper **chunks** from each sentence, to not lose important information that can be significant (in answers) was our main goal in this part.





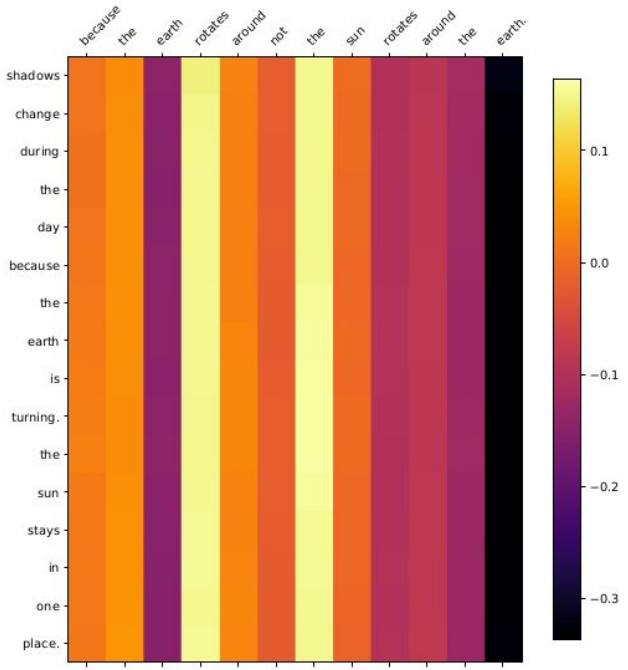
Understandability- Deep learning approach

The similarity detection between two sentences can be reached with several techniques as: euclidean distance, cosine similarity, etc. In this approach we used **pearson autocorrelation** coefficient between two sentences [reference answer] and [student answer], we also calculate the deviation and variance between both to detect structure gaps.

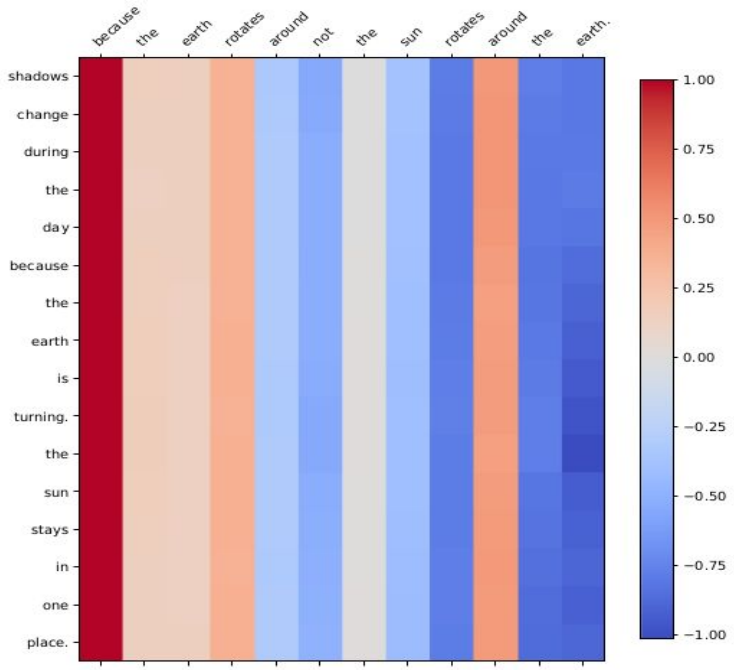
In this section we present the visual content retrieved from the model, and some insights to understand it.

text Correctness- Understandability

- The **Attention** model implemented in this project allow us to see the alignment taken into account.
- In this way we can understand the structure used for the decision.
- Based on pre-trained **word embedding** ‘Glove’ (little version).



text Similarity - Understandability

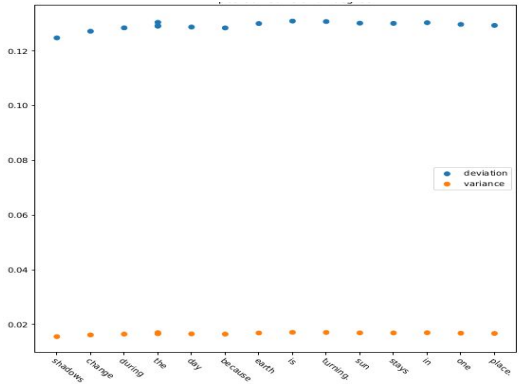


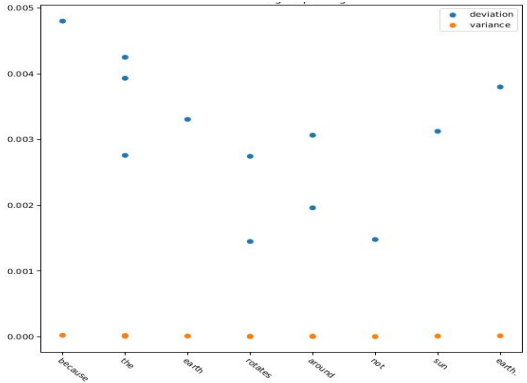
- The **Attention** model implemented in this project allow us to calculate the autocorrelation over the alignment between answer and reference.
- In this way we can understand the intern semantic structure used in the decision.

Variance and Deviation- Understandability

We use the Variation and standard deviation to have an idea about the outlier chunks contained in both ‘reference’ and ‘answer’ respect to each other. It can address better understandability for the obtained results, sometimes provided ‘answers’ aren’t well formed, that could be seen in those graphics.

reference answer



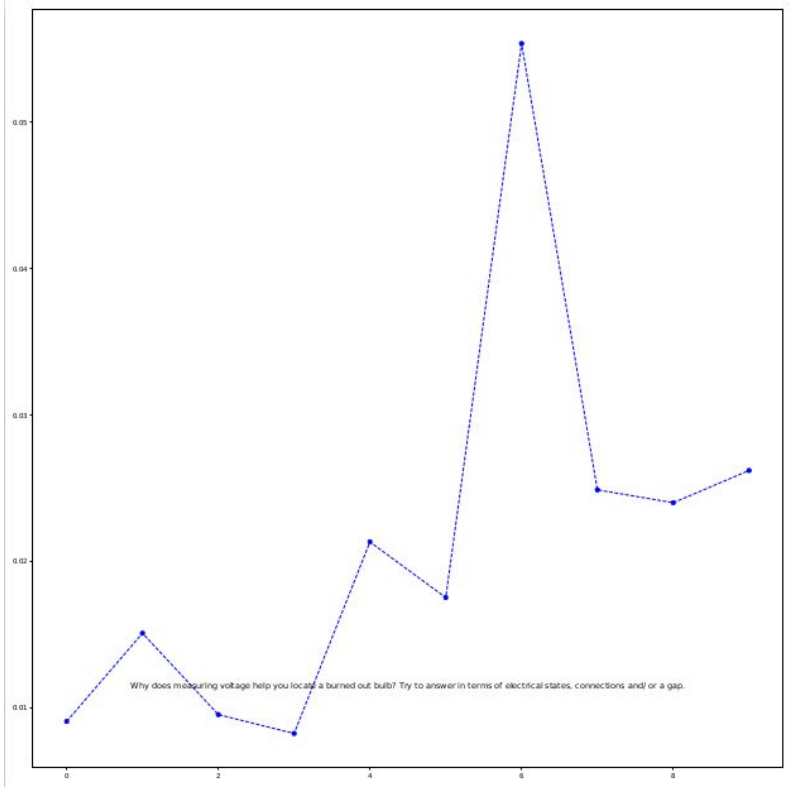


student answer

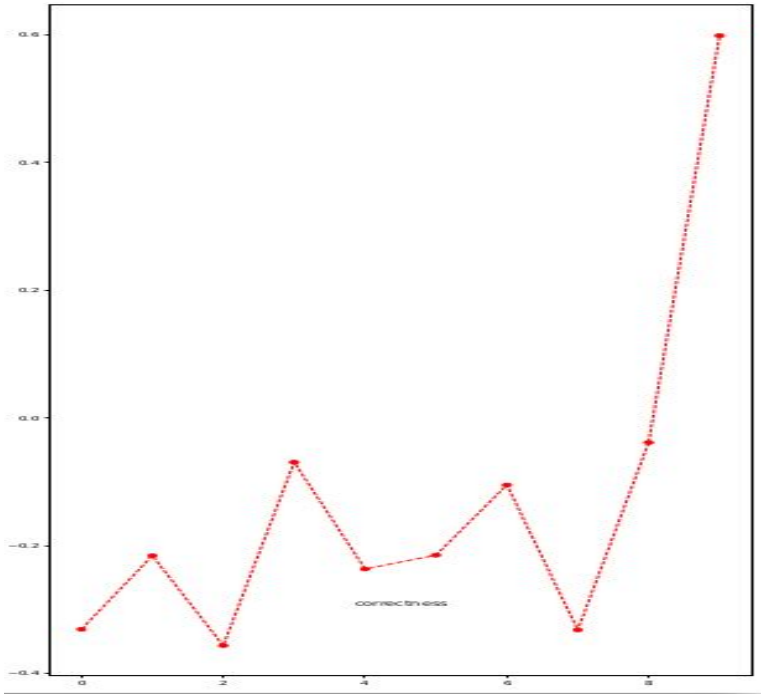
ensembled Similarity (Avg) - Understandability

We can show the **Average similarity degree** of a group of answers to the reference.

The Similarity Average is normalized as (-1 .. 1). This is taken into account with ensembled correctness to determine answers as ‘correct’ or ‘incorrect’.



ensembled Correctness (Avg) - Understandability



We can show the **Average correctness degree** of a group of answers to the reference.

The Correctness Average isn't normalized in this case, but with the Similarity Average insights that's how we decide 'correct' or incorrect' for the given answers.

Unsupervised Grading - Understandability

The normalization (0 ..5) for grading purposes is generated using external resources. ‘semantic_text_similarity’ package.

Here we use another external pre-trained model due the provided data doesn’t have any grading information involved.

Report - Understandability

While analyzing all kinds of output, you will find some mistakes, and some contradictory score assigning, that maybe due the dataset state, lack of preprocessing, miss alignment, quality of copus or even pre-trained embeddings.

Prepared Data access for those purposes was difficult, so we gathered the accessible on our own.

LOG - Understandability

The grade decisions can be seen in the log after finishing all the pipeline, you can check the ranges prefixed of answers and questions you want to test.

The data preprocessing have a big impact on the process, in real cases we can not exclude certain terms from students answers as stopwords, etc. In this experiment we processed all data.

LOG - UNDERSTANDABILITY

```
question: Elena has a male lizard that has lived for several years in the habitat she provided.
She knows that some lizards are territorial. Besides additional food and water, what should she be sure
to include in the habitat before she adds another male lizard?

 given information:

 reference: Elena should include a separate shelter for each lizard.        answer: She should add another shelter.
 score correct

 results:

 estimated similarity: [2.5420654]      estimated score: [0]
 grade: 5
```

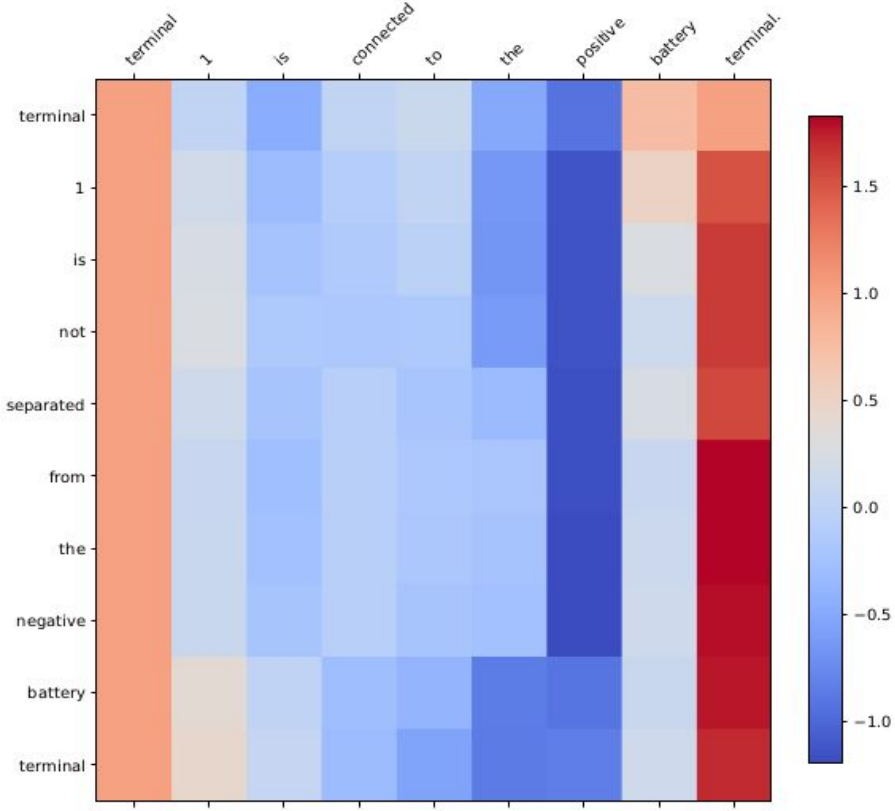
Output Example 1

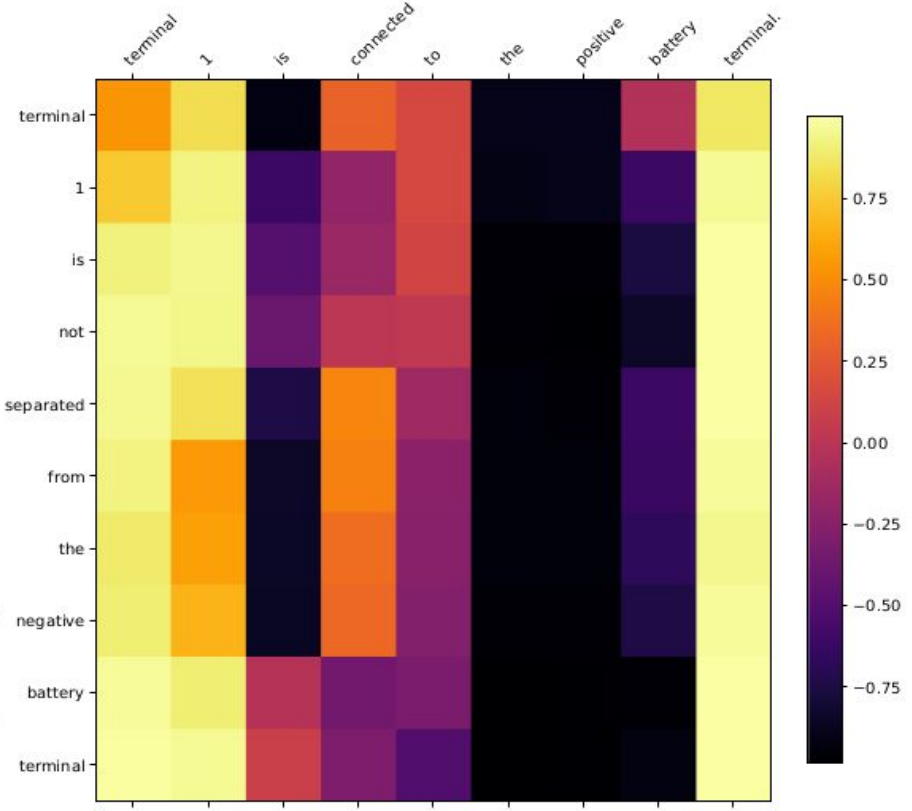
Reference answer: terminal 1 is not separated from the negative battery terminal

Student answer: Terminal 1 is connected to the positive battery terminal.

- **incorrect**
- **similarity:[3.8376303]**
- **score: 7/10**

Example 1


terminal
1
is
connected
to
the
positive
battery
terminal.
terminal
1
is
not
separated
from
the
negative
battery
terminal
1.5
1.0
0.5
0.0
−0.5
−1.0


terminal
1
is
connected
to
the
positive
battery
terminal.
terminal
1
is
not
separated
from
the
negative
battery
terminal
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

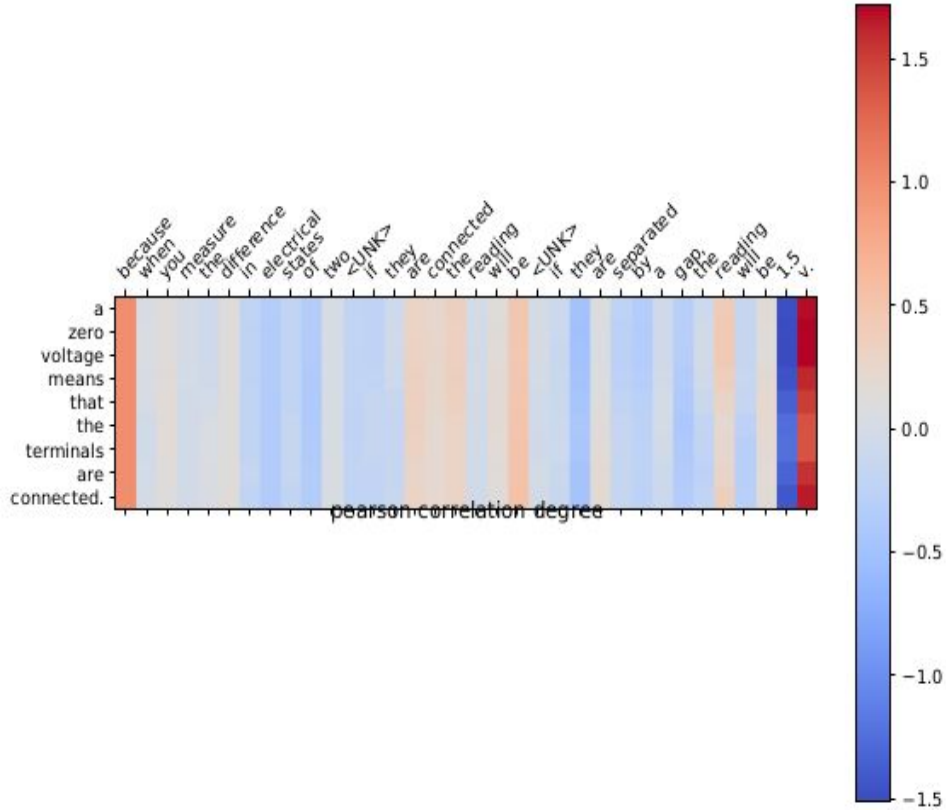
Output Example 2

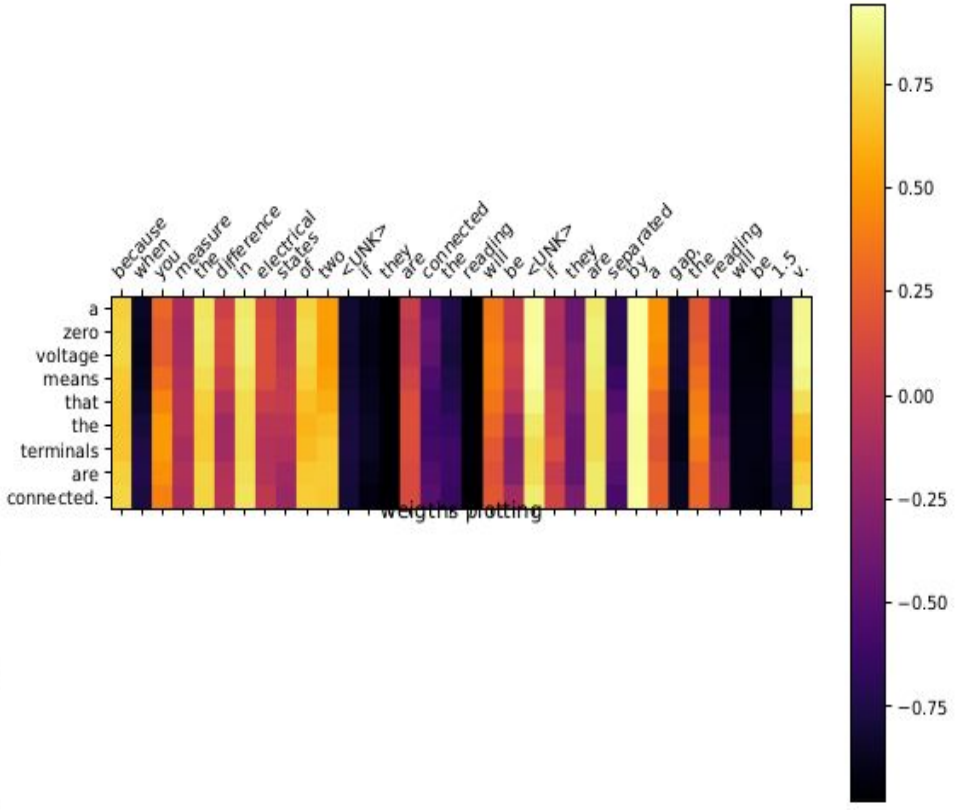
Reference answer: A zero voltage means that the terminals are connected.

Student answer: Because when you measure the difference in electrical states of two terminals, if they are connected the reading will be 0. If they are separated by a gap, the reading will be 1.5 V.

- **correct**
- **similarity:[2.5791867]**
- **score: 5/10**

Example 2


because when you measure the difference in electrical states of two <UNK> if they are connected the reading will be <UNK> if they are separated by a gap, the reading will be 1.5 v.
a zero voltage means that the terminals are connected.
pearson correlation degree
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5


because when you measure the difference in electrical states of two <UNK> if they are connected the reading will be <UNK> if they are separated by a gap, the reading will be 1.5 v.
a zero voltage means that the terminals are connected.
weights plotting
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

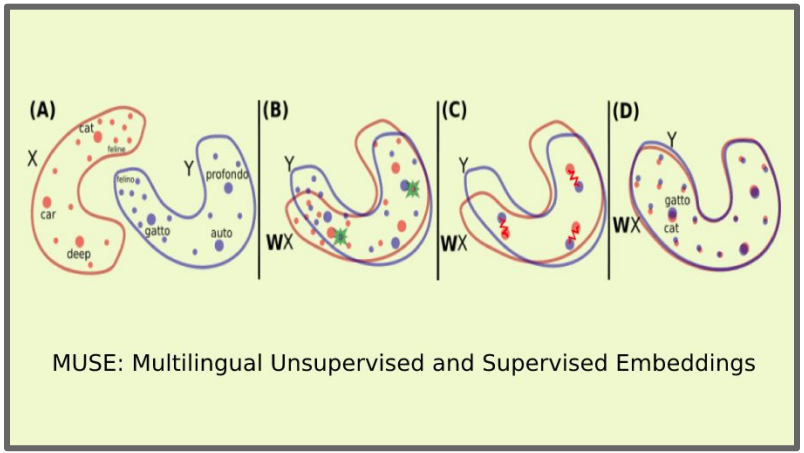
Student Performance | tutoring

Our model have the capability to integrate easily a sort of tracking, based on the **neural network** implemented, while crossing the data to detect similarity, we can assign **feedback** on **answers**, it requires proper **data related** to **questions** or tasks.

We can go deeper in students mistakes from the orthograph to the form of the answer, and also the semantics presented in the answer (it depends on the subject treated). We could use the answers provided from the students to expand our corpora, etc.

SMILE : Learning a new language

The Advantage of using an approach that is based on embeddings, is that we can apply the same process, just by adding a cross-lingual embedding which introduce semantic dimension to the chosen languages into our model.



MUSE: Multilingual Unsupervised and Supervised Embeddings

REFERENCES (3)

Vitalii Zhelezniak, Aleksandar Savkov, April Shen & Nils Y. Hammerla Babylon Health, Correlation Coefficients and Semantic Textual Similarity, *Proceedings of NAACL-HLT 2019, pages 951–962 Minneapolis, Minnesota, 2019 Association for Computational Linguistics*

Mohler, M., Bunescu, R., & Mihalcea, R. (January 01, 2011). Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. *Annual Meeting of the Association for Computational Linguistics and ... Conference of the European Chapter of the Association for Computational Linguistics : Proceedings of the Conference, 49,* 752-762.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, Janyce Wiebe SemEval-2016 Task 1: *Semantic Textual Similarity, Monolingual and Cross-Lingual Evaluation.* *Proceedings of SemEval-2016, pages 497–511,San Diego, California, June 16-17, 2016. ©2016 Association for Computational Linguistics*

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, First Joint Conference on Lexical and Computational Semantics *(*SEM), pages 385–393,Montr´eal, Canada, June 7-8, 2012.c©2012 Association for Computational Linguistics SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity.*

Short Answer Scoring

Results Conclusion

Group Exercise

In this part, you can test your capability to apply Semantic Text Similarity manually while we cross the examples and then we can compare to the (100 epoch) model's result with :

train acc = 0.8 , dev acc = 0.74 , test acc = 0.7

Note that this kind of experience require the interaction of human as a golden standard model, to feed the machine for better and reasonable results.

Question

Elena has a male lizard that has lived for several years in the habitat she provided.

She knows that some lizards are territorial. Besides additional food and water, what should she be sure to include in the habitat before she adds another male lizard?

Example 1

Reference answer: Elena should include a separate shelter for each lizard.

Student answer: She should add another shelter.

Can you make prediction ?

- incorrect / correct ?
- similarity:[0 -> 5] ?
- score: ??/10

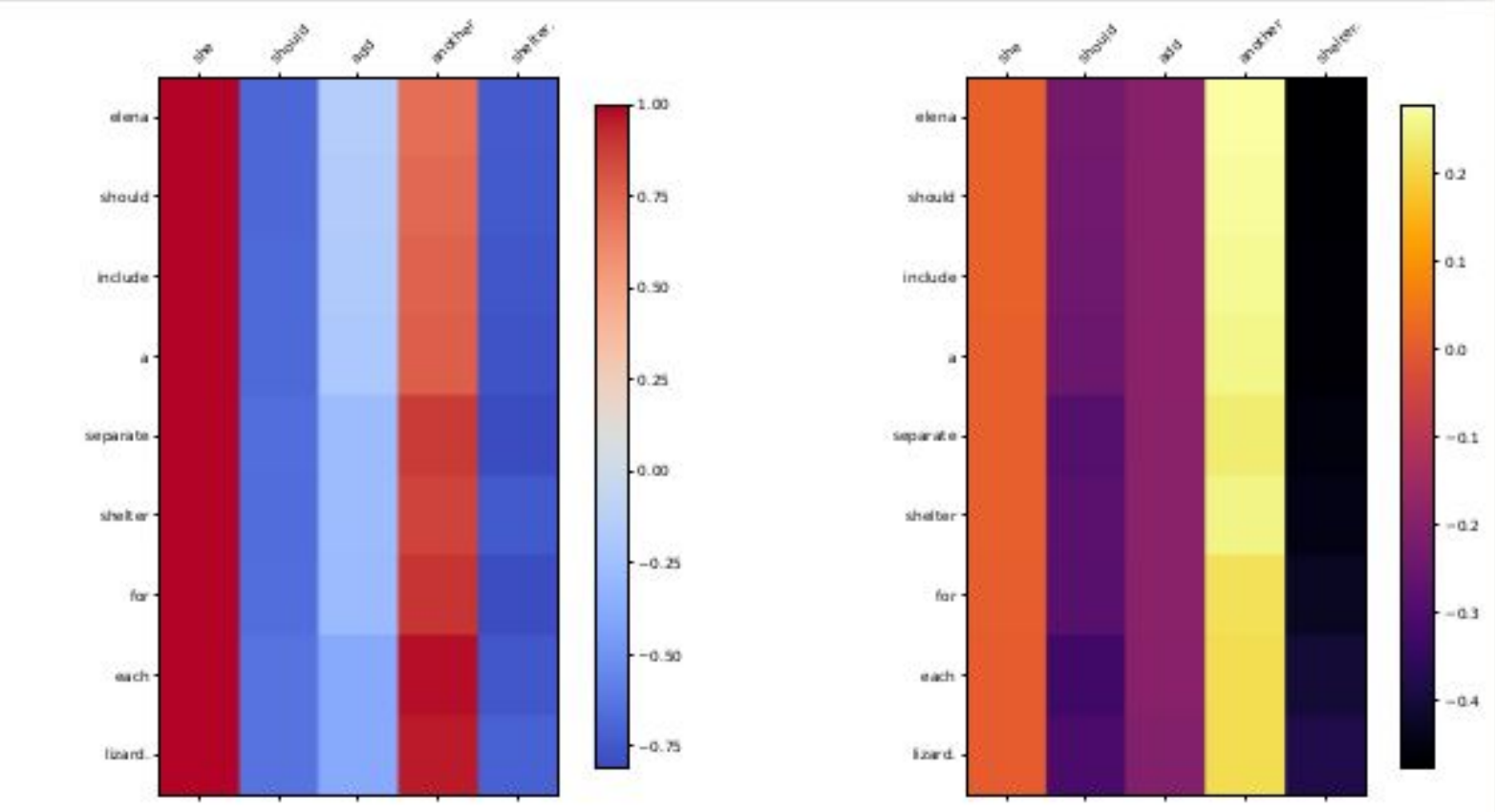
Example 1

Reference answer: Elena should include a separate shelter for each lizard.

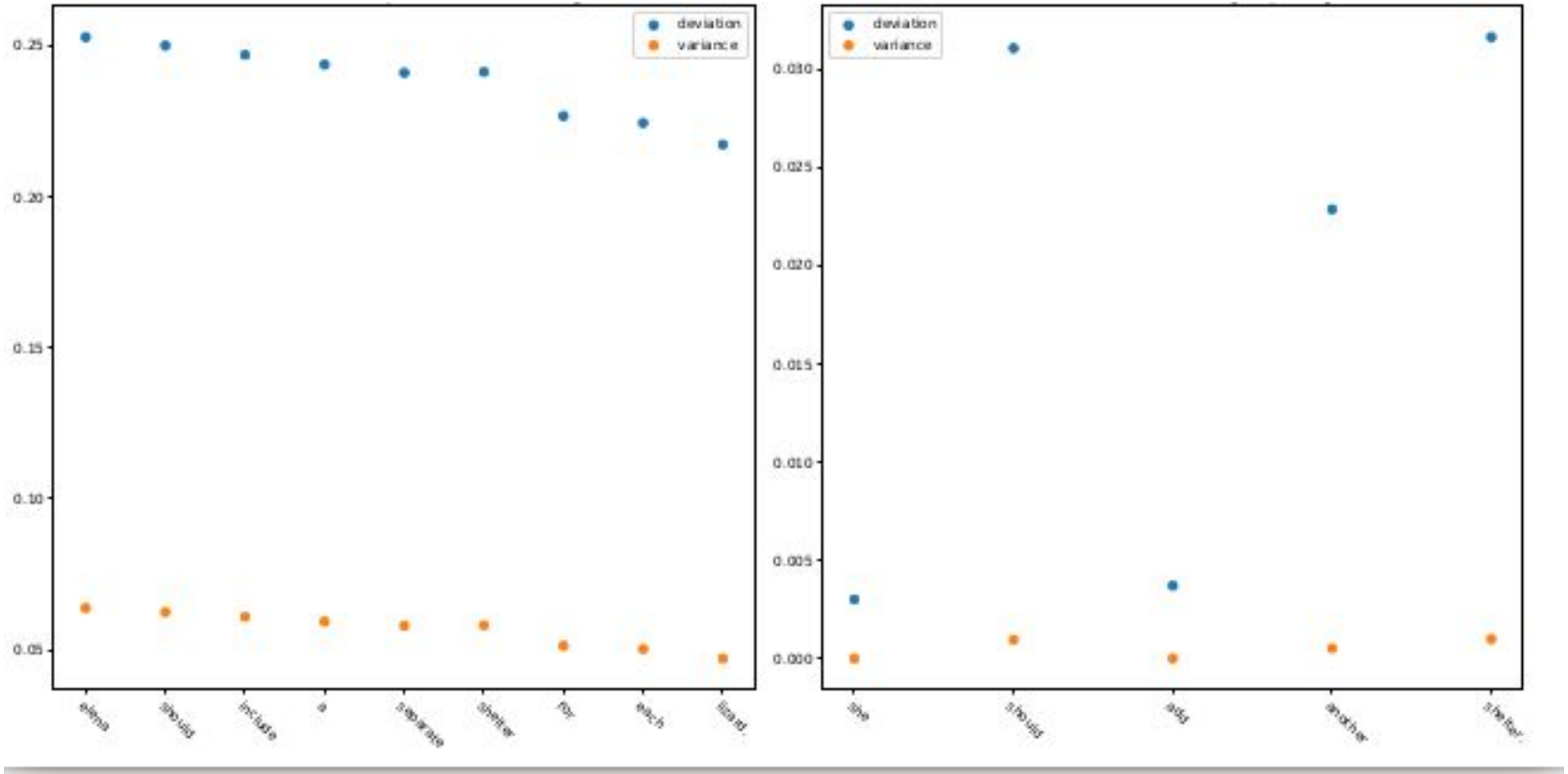
Student answer: She should add another shelter.

- **incorrect**
- **similarity:[2.5420654]**
- **score: 5/10**

Example 1


she
should
add
another
shelter.
elena
should
include
a
separate
shelter
for
each
lizard.
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4

Example 1


deviation
variance
0.25
0.20
0.15
0.10
0.05
she
should
include
a
separate
shelter
for
each
listed
deviation
variance
0.030
0.025
0.020
0.015
0.010
0.005
0.000
she
should
add
another
shelter

Example 2

Reference answer: Elena should include a separate shelter for each lizard.

Student answer: Water and food.

Can you make prediction ?

- incorrect / correct ?
- similarity:[0 -> 5] ?
- score: ??/10

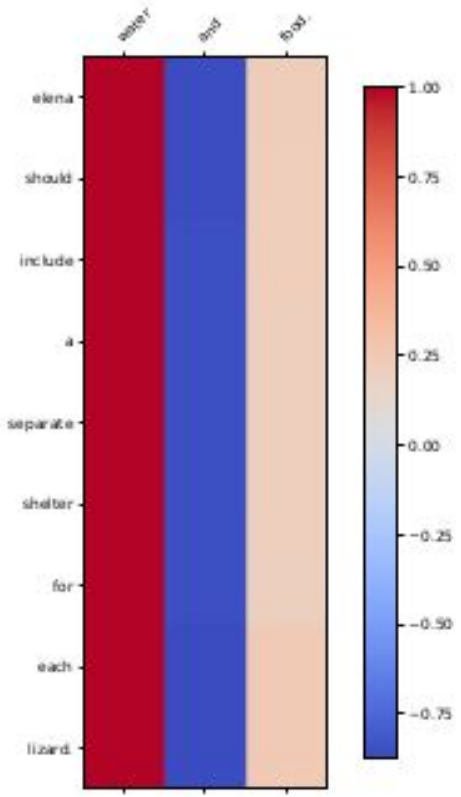
Example 2

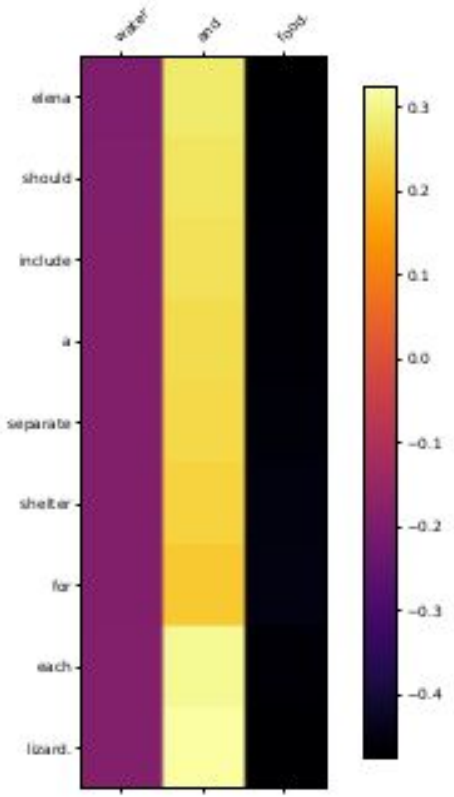
Reference answer: Elena should include a separate shelter for each lizard.

Student answer: Water and food.

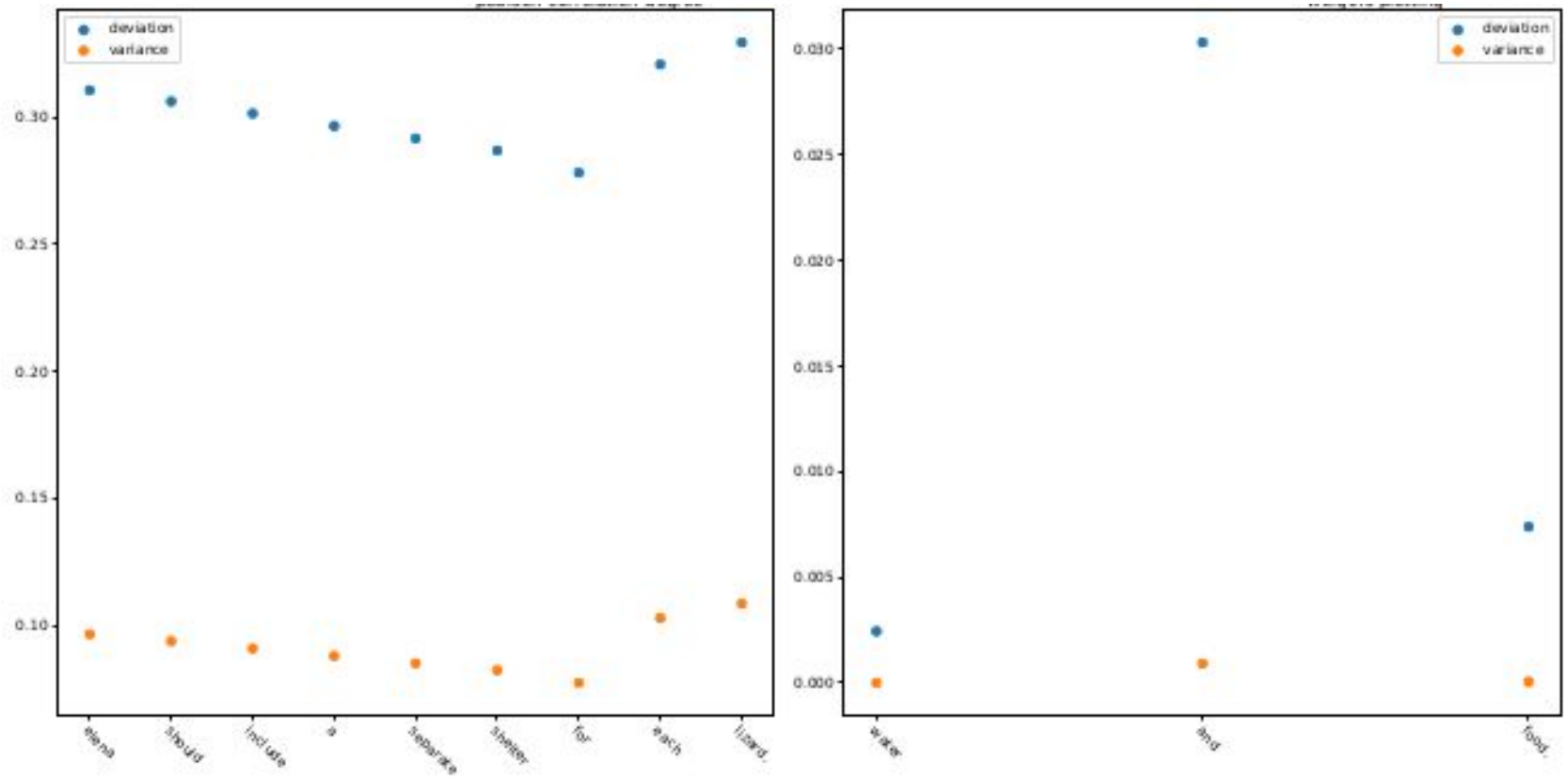
- **incorrect**
- **similarity:[0.7210027]**
- **score: 1/10**

Example 2


water
and
food
elena
should
include
a
separate
shelter
for
each
lizard
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75


water
and
food
elena
should
include
a
separate
shelter
for
each
lizard
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4

Example 2


deviation
variance
0.30
0.25
0.20
0.15
0.10
elena
should
include
a
separate
shelter
for
each
lizard
deviation
variance
0.030
0.025
0.020
0.015
0.010
0.005
0.000
water
and
food

Example 3

Reference answer: Elena should include a separate shelter for each lizard.

Student answer: Grass.

Can you make prediction ?

- incorrect / correct ?
- similarity:[0 -> 5] ?
- score: ??/10

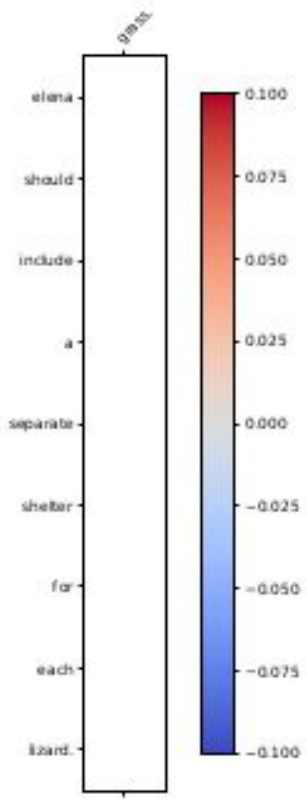
Example 3

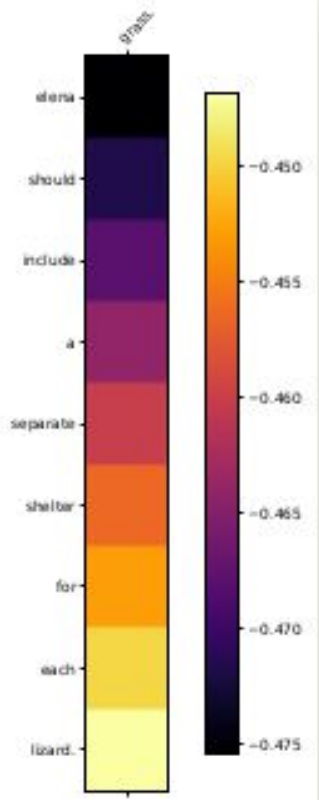
Reference answer: Elena should include a separate shelter for each lizard.

Student answer: Grass.

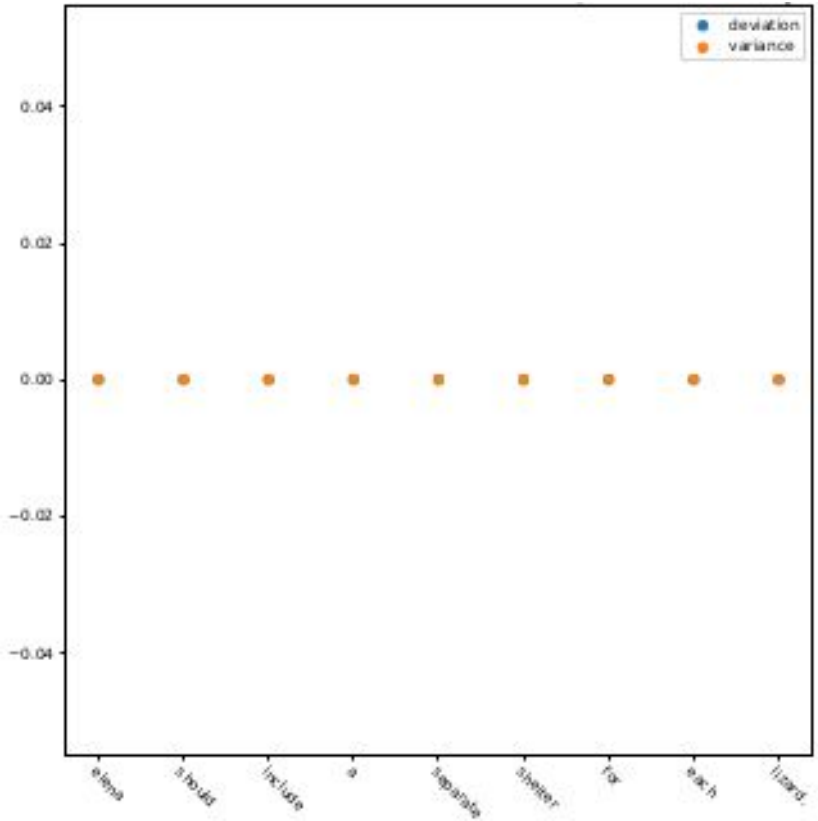
- **incorrect**
- **similarity:[0.6501440]**
- **score: 1/10**

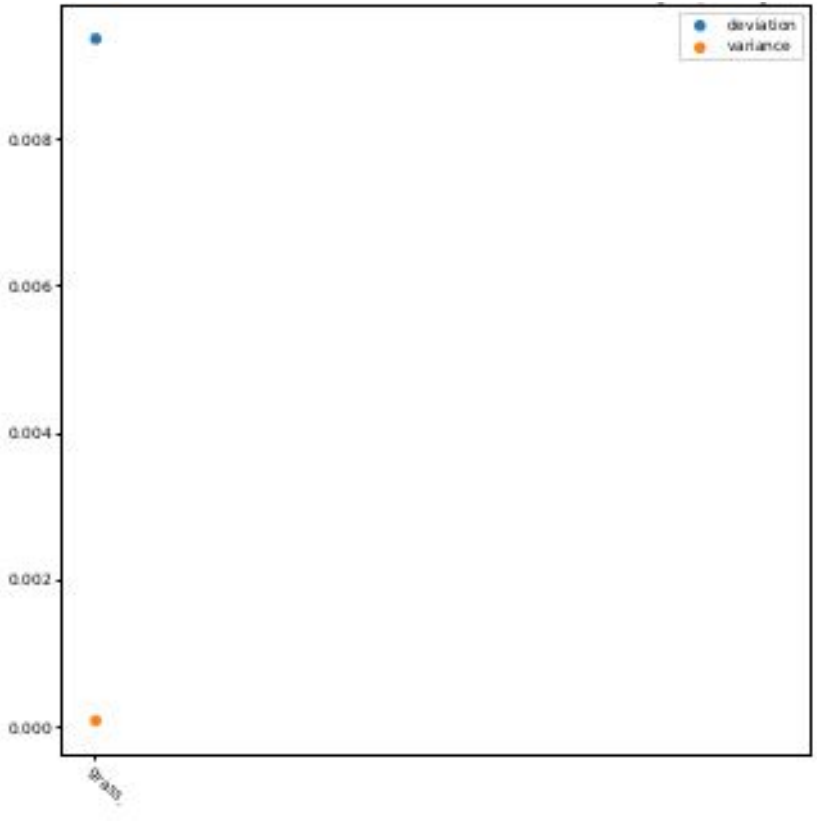
Example 3


grass
elena
should
include
a
separate
shelter
for
each
lizard.
0.100
0.075
0.050
0.025
0.000
−0.025
−0.050
−0.075
−0.100


grass
elena
should
include
a
separate
shelter
for
each
lizard.
−0.450
−0.455
−0.460
−0.465
−0.470
−0.475

Example 3

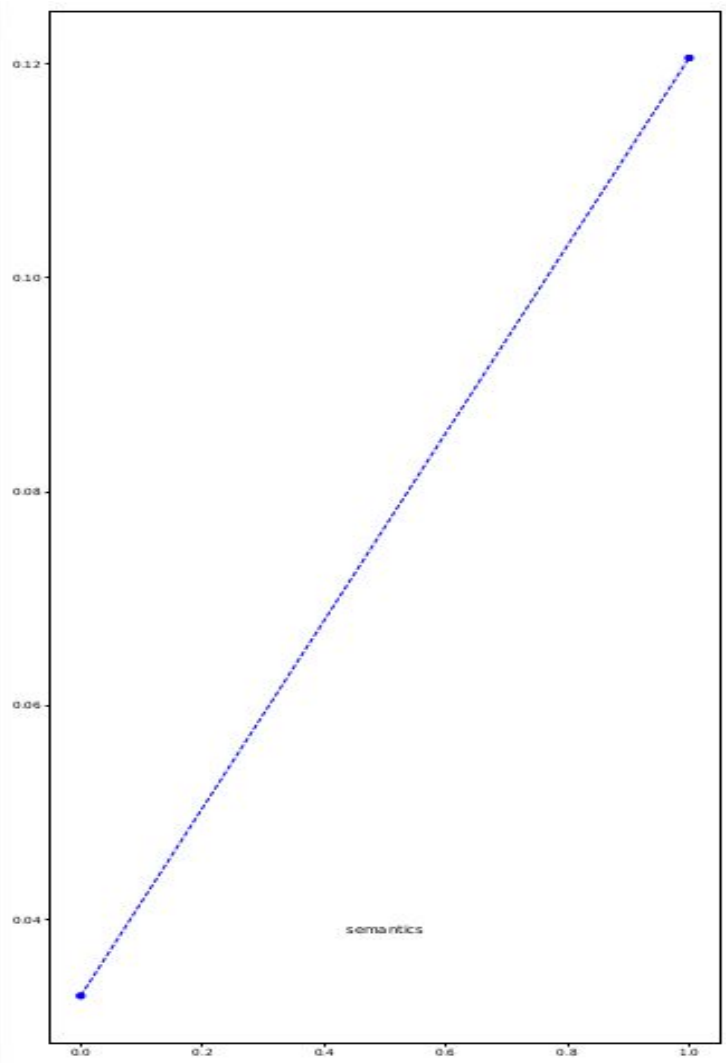

deviation
variance
0.04
0.02
0.00
−0.02
−0.04
elena
should
include
a
separate
shelter
for
each
lizard.

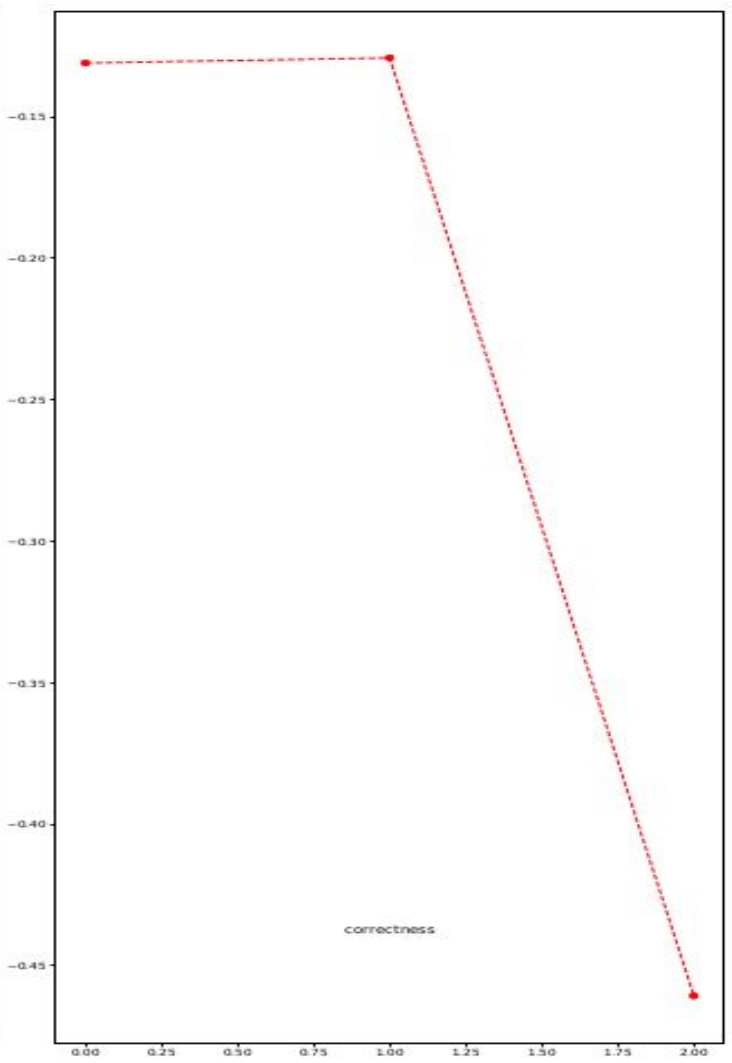

deviation
variance
0.008
0.006
0.004
0.002
0.000
grass.

AVG LINES

Notice that:

- correctness is negative.
- semantics near 0, but the first one very high.





Group conclusion

Could you feedback those students taking into account this knowledge?

You think that those models could be machine friendly and user friendly?

Conclusions - Model

In this challenging experiment we discovered a lot of techniques to achieve our objectives, the process can have many different designs, and may require a lot of powerful resources and time.

Assign a score automatically to an answer may seem simple but when you discover how complex can semantics be, will change your mind.

With all this ressources, we can create powerful educational tools that serve all the actors for a functional educational system.

Conclusions - Information / Knowledge

The results may improve using prepared Data on the process, even applying some suitable preprocess.

The knowledge access over this area is forbidden in most cases, that shall stop further conclusions or comparisons.

The access to proper Data and Papers should be strategy over those tasks were the high difficulty can address to wrong results, if researchers aim the profit we won't advance as desired, they should profit over finished models and processes.

Questions