

# Deep Learning for NLP 2020

## Shared Task

May 29, 2020

---

### Contents

---

<b>1 Task</b>	<b>1</b>
<b>2 Submissions</b>	<b>2</b>
2.1 Training, Development and Test Datasets . . . . .	2
2.2 Evaluation on CodaLab . . . . .	2
<b>3 Organizational Matters</b>	<b>2</b>
3.1 Group Size . . . . .	2
3.2 Phases . . . . .	3
<b>4 System Description Paper and Presentation</b>	<b>3</b>
4.1 System Description Paper . . . . .	3
4.2 Presentation . . . . .	4
<b>5 Grading</b>	<b>4</b>

---

## 1 Task

---

The DL4NLP 2020 shared task deals with semantic textual similarity under adversarial attacks. In the shared task, we define *semantic textual similarity (STS)* as a supervised regression task in which the semantic similarity of two pieces of text (typically sentences) should be determined. However, the test datasets have been perturbed with so-called *adversarial attacks*, which are modifications to the input of a model that do not change the label or score. Usually, their goal is to fool a deep learning model but not humans. Thus, attacks are often chosen in such a way that humans are robust to them.

Examples for adversarial attacks are:

- **disemvoweling:** Some of the vowels have been removed.
- **visual attacks:** Similar-looking characters have been replaced by each other.

**Your task is to design models that can deal with corrupted inputs in a similar way as humans can.**

---

## 2 Submissions

---

### 2.1 Training, Development and Test Datasets

---

The labeled datasets for this task contain entries that comprise a real-numbered similarity score between 0 and 1, a first sentence, and a second sentence. The unlabeled datasets' entries comprise only the two sentences. In our case, each line in the datasets corresponds to a single entry, and the score and the sentences are separated by tabs '\t'.

Throughout the shared task, the following datasets will be used:

- The **training dataset** contains labeled, unperturbed instances.
- The **development dataset** contains labeled, unperturbed instances.
- The **test dataset from hex06** contains labeled instances that have been perturbed with disemvoweling and visual attacks.
- The **test dataset for the scoreboard** contains unlabeled instances that have been perturbed with additional adversarial attacks.
- The **test dataset for the final score** contains unlabeled instances that have been perturbed with additional adversarial attacks.

---

### 2.2 Evaluation on CodaLab

---

CodaLab handles the evaluation of your approaches. You will submit .txt files in which every row contains the real-numbered similarity score between 0 and 1 for the same row in the corresponding unlabeled test dataset. We will use Spearman's rank to evaluate your predictions.

During the shared task, you can submit predictions for the **test dataset for the scoreboard**, which will be evaluated to determine a leaderboard that you can use to find out how well your approaches do in comparison to others. To prevent you from overfitting the test dataset, you will only be able to make two submissions per person per day.

In the final days of the shared task, we will upload the **test dataset for the final score**. This dataset has been perturbed with additional perturbations, so it is not enough to overfit the other datasets. You must then apply your models to this dataset and upload the predictions. Their evaluation will determine your final score.

---

## 3 Organizational Matters

---

### 3.1 Group Size

---

You have to enter a group in Moodle to work on the shared task. Groups can consist of **up to four people**. We recommend that there are at least two people in each group. Consider posting in the shared task discussion forum if you are looking for team members. The group selection is open until **08.06.2020 18:00**.

---

## 3.2 Phases

---

The shared task is divided into three phases:

1. During the **development phase**, you develop your approaches. Once we set up CodaLab, you may upload your predictions of the **test dataset for the scoreboard**, which determine your position in CodaLab and serve as a comparison to other participants. This phase lasts **from 01.06.2020 18:00 to 01.07.2020 18:00**.
2. During the **test phase**, we will upload the **test dataset for the final score**, and you will apply your models to it and submit its predictions. Your evaluation results will influence your final ranking in the shared task. This phase lasts **from 01.07.2020 18:00 to 03.07.2020 18:00**.
3. Research is not only about trying things out, but also about documenting and discussing the results. In shared tasks, participants are expected to submit a short paper describing their approach. **Until 12.07.2020 18:00**, you will submit your **system description papers**, **code**, and **presentation** to Moodle.

---

## 4 System Description Paper and Presentation

---

---

### 4.1 System Description Paper

---

After you submitted your predictions to CodaLab, you will be given time to prepare a system description paper with **two to four pages** in length. The paper must be submitted in PDF format and use the ACL 2020 template<sup>12</sup>.

Your paper must include the following information:

- introduction
- description of your approach
- research question, experiment descriptions and experiment setups
- results (preliminary results on the different datasets)
- discussion of your results
- conclusion

As an inspiration for how your paper could look like, have a look at this paper for the CoNLL 2017 shared task: <http://universaldependencies.org/conll17/proceedings/pdf/K17-3002.pdf>. You are encouraged to perform multiple experiments and ablation tests, for example, to find out how well different approaches perform or how the performance changes if you modify individual parts of the system.

---

<sup>1</sup><http://acl2020.org/downloads/acl2020-templates.zip>

<sup>2</sup><https://www.overleaf.com/latex/templates/acl-2020-proceedings-template/zsrkcwjtpcd>

---

## 4.2 Presentation

---

In the final lecture, some of you will give a **10 to 15 minutes long presentation** about your approach. The goal of the presentation is to communicate the central ideas behind your approach as well as the main outcomes of your experiments. Every group must hand in a presentation of **10 to 15 slides** as a PDF file.

---

## 5 Grading

---

To be eligible for the exam bonus, you must develop a model that beats our baseline on the **test dataset for the final score**. Your model must be able to do this on its own without any user intervention. However, your code does not have to run in the Docker container for the home exercises. In addition to your code, you must submit a two to four pages system description paper as well as a 10 to 15 slides presentation to Moodle and eventually give a 10 to 15 minutes presentation in the final lecture. Finally, you must include two to three sentences per person about who did what in the project.