NLP course 2022

Homework 3

Coreference Resolution

Prof. Roberto Navigli

Teaching assistants:
Stefan Bejgu,
Abelardo Carlos Martinez Lorenzo,
Riccardo Orlando,
Alessandro Scirè, Simone Tedeschi

SAPIENZA
UNIVERSITÀ DI ROMA

SAPIENZA NLP

Coreference Resolution

An introduction

Coreference Resolution

- Coreference is the fact that two or more expressions in a text – like pronouns or nouns – link to the same person or thing
- It is an important Natural language processing task, and it is also considered one of the key building blocks to building conversational Artificial intelligences
- Example:

__John__ went home because __he__ was tired

Coreference Resolution

This task can be solved by applying a two-steps pipeline consisting of:

1. Mention proposal
2. Mention linking

Coreference Resolution: Mention Proposal

In the mention proposal step the aim is to identify all the candidate entities in a text.

I voted for Obama because he was most aligned with my values", she said.

Coreference Resolution: Mention Linking

In the mention linking step, instead, the goal is to understand which pairs of mentions correspond to the same entity.

I voted for Obama because he was most aligned with my values", she said.

The homework

The goal of this homework

The core goal of this homework is to develop a system that performs **mention proposal** and **mention linking**

Alice Perrers is the protagonist of Emma Campion's novel, The King's Mistress. She appears in Anya Seton's novel, Katherine.

She	79	
Alice Perrers	0	TRUE
Emma Campion	36	FALSE

In the simplest formulation you have the ambiguous pronoun pre-identified **(She)** as well as its candidate entities **(Alice Perrers and Emma Campion)**, and you have select the right coreference **(Alice Perrers)**

The goal of this homework

This task can be addressed at three different complexity levels:

1. End-to-end (E2E) coreference resolution
2. Entity identification and resolution
3. Entity resolution

E2E Coreference Resolution

In end-to-end coreference resolution, you receive as input only sentences.

Alice Perrers is the protagonist of Emma Campion's novel, The King's Mistress.
She appears in Anya Seton's novel, Katherine.

And the goal is to:

E2E Coreference Resolution

In end-to-end coreference resolution, you receive as input only sentences.

Alice Perrers is the protagonist of Emma Campion's novel, The King's Mistress.
She appears in Anya Seton's novel, Katherine.

And the goal is to:

1. Identify the ambiguous pronoun (**note**: there can be more than one pronoun, but only one is ambiguous)

E2E Coreference Resolution

In end-to-end coreference resolution, you receive as input only sentences.

Alice Perrers is the protagonist of Emma Campion's novel, The King's Mistress. She appears in Anya Seton's novel, Katherine.

And the goal is to:

1. Identify the ambiguous pronoun (**note**: there can be more than one pronoun, but only one is ambiguous)
2. Recognize the entities

E2E Coreference Resolution

In end-to-end coreference resolution, you receive as input only sentences.

Alice Perrers is the protagonist of Emma Campion's novel, The King's Mistress. She appears in Anya Seton's novel, Katherine.

And the goal is to:

1. Identify the ambiguous pronoun (**note**: there can be more than one pronoun, but only one is ambiguous)
2. Recognize the entities
3. Select the entity that corresponds to the identified pronoun

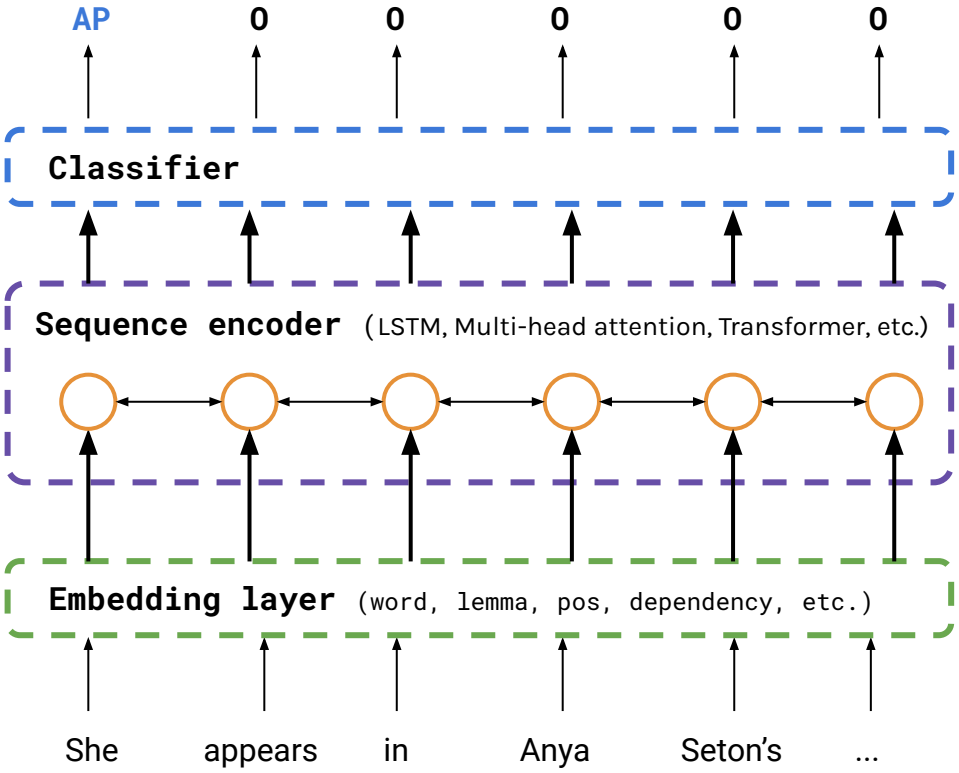
Entity identification and resolution

- The **entity identification and resolution** subtask corresponds to steps 2 and 3 of the described pipeline

- Similarly, the **entity resolution** subtask corresponds to the step 3 only
  - Note that if you do only this step, you have to select the coreferenced entity between just two candidate entities (binary classification)

Ambiguous Pronoun Identification

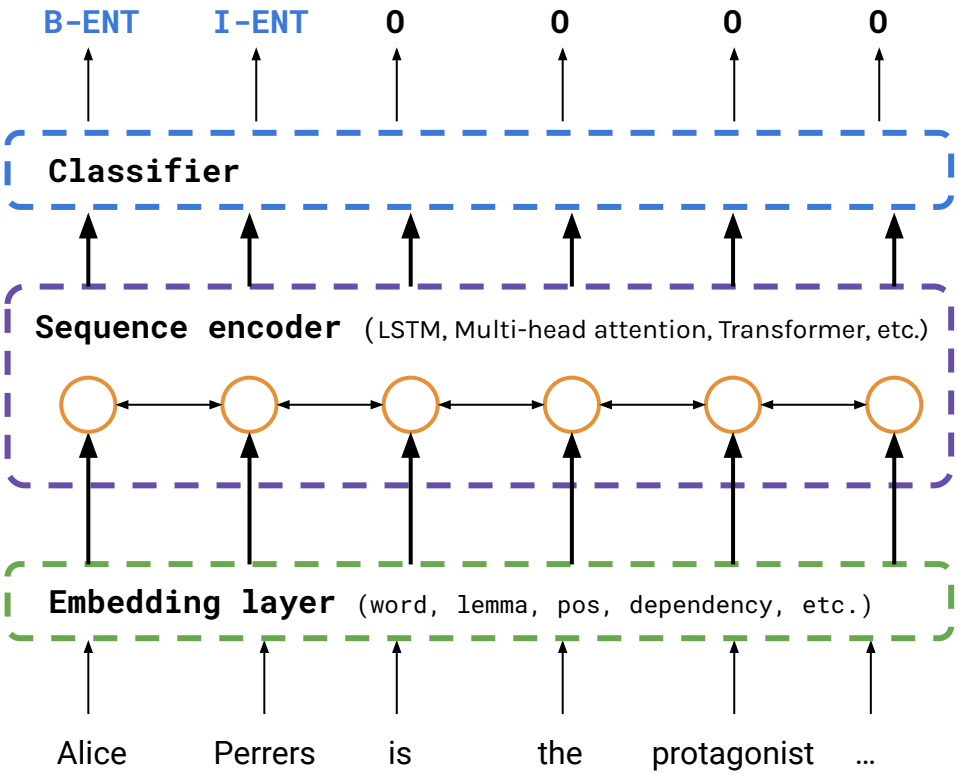
Possible approaches



Starting from the structure of the dataset entries showed at slide 8, you can construct a training set for your ambiguous pronoun classifier.

Alternatively, you can try to apply a POS tagger and, if there are multiple pronouns, apply some heuristics.

Entity Identification
Possible approaches



You can either use the system you developed for HW1 or use an improved version, e.g. including Transformers.

Note: you may not need to predict the semantic type of an entity (e.g. B-PER could not add any useful information)

Entity Resolution
Possible approaches

- For the entity resolution step we expect from you to explore **original directions**
- A good starting point could be to produce a **contextualized representation** for both the identified pronoun and entities, and select the most related entity

Coreference Resolution Evaluation

The performance will be measured by means of the accuracy score as follows:

$$\begin{cases} \textbf{1 if both the pronoun and the coreferenced entity are correctly identified} \\ \textbf{0, otherwise} \end{cases}$$

If you will work only on steps 2 and 3, the pronoun will be always considered as correct

Coreference Resolution Evaluation

The performance will be measured by means of the accuracy score as follows:

Alice Perrers is the protagonist of Emma Campion's novel, The King's Mistress.
She appears in Anya Seton's novel, Katherine.

Your Answer:
She 79
Alice Perrers 0

True Labels:
She 79
Alice Perrers 0

Accuracy = 1.0

Coreference Resolution Evaluation

The performance will be measured by means of the accuracy score as follows:

Alice Perrers is the protagonist of Emma Campion's novel, The King's Mistress. She appears in Anya Seton's novel, Katherine.

Your Answer:
The 57
Alice Perrers 0

True Labels:
She 79
Alice Perrers 0

Accuracy = 0.0

Coreference Resolution Evaluation

The performance will be measured by means of the accuracy score as follows:

Alice Perrers is the protagonist of Emma Campion's novel, The King's Mistress. She appears in Anya Seton's novel, Katherine.

Your Answer:
The 57
Emma Campion 35

Accuracy = 0.0

True Labels:
She 79
Alice Perrers 0

Coreference Resolution Evaluation

The performance will be measured by means of the accuracy score as follows:

Alice Perrers is the protagonist of Emma Campion's novel, The King's Mistress. She appears in Anya Seton's novel, Katherine.

Your Answer:
She 79
Emma Campion 35

Accuracy = 0.0

True Labels:
She 79
Alice Perrers 0

Coreference Resolution Evaluation

The performance will be measured by means of the accuracy score as follows:

Mayhem Miller brings in MMA fighter Eddie Alvarez with a record of 22 wins and 2 losses to teach them a lesson.

Your Answer:
Them 98
-

Accuracy = 1.0

True Labels:
Them 98
-

Homework extras

Increase the complexity of your model!

Take inspiration from recent papers:

- **ProBERT**, Gendered Ambiguous Pronouns Shared Task: Boosting Model Confidence by Evidence Pooling (Attree 2019)
- **Full Ensemble**, Gendered Pronoun Resolution using BERT and an extractive question answering formulation (Chada 2019).
- **PeTra**: A Sparsely Supervised Memory Model for People Tracking (Toshniwal et al. 2020)
- Or try something new!

Implement other parts of the pipeline

- Entity Identification and Resolution
  - In this setting, the entities are not given but you have to identify them by your own
- E2E Coreference Resolution
  - In this setting, neither the entities nor the pronoun are given

If you decide to implement other steps of the pipeline, you can both use a **pipeline** or adopt a **joint approach**.

Submission

What you will receive

- We will provide you with a folder organized as follows (some files are omitted):

```
- nlp2022-hw3/
  - data/
  - hw3/
    - model.py
    - stud/
  - model/
  - requirements.txt
  - test.sh
```

- You are allowed to edit **only** the items in bold!

What you will receive

- We will evaluate your work using Docker
  - You should be fine even if you don't know anything about it
- If **`test.sh`** runs on your side, it will run on ours as well
  - Just keep in mind: do not change any file but those we marked in bold as editable in the previous slide
- Additionally, we wrote a **`README.md`** to get you everything up and running
- You can find the code repository here!

What we expect from you

- The zip folder we gave you (but populated :))
- Put your training code (if you used Colab, download the notebook `.ipynb` and place it) in **`hw3/stud/`**
- If you use any additional library, modify the **`requirements.txt`** file as needed (click here for info)
- Use the data (train and dev) in the data folder
  - use each file as defined in the **standard ML conventions** (*train for training*, *dev for model selection*)

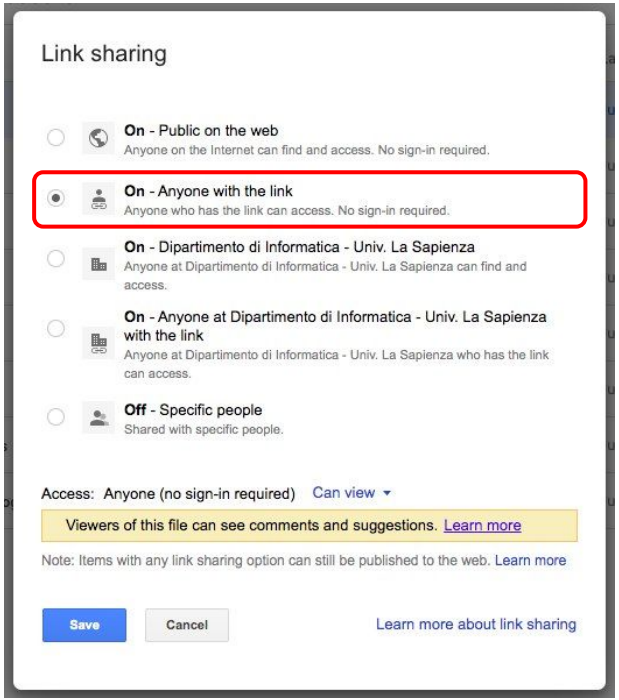
What we expect from you

- Put **everything** your model needs (vocabulary, weights, ...) inside the **`model/`** folder, and **be sure to properly load them** in your model
- In **`hw3/stud/implementation.py`** implement the **`StudentModel`** class
  - Load your model and use it in the **`predict`** method
  - You **must respect** the signature of the predict method!
  - You can add other methods (i.e. the constructor)
- In **`hw3/stud/implementation.py`** implement the **`build_model`** function
  - It should initialize your **`StudentModel`** class.

What we expect from you

- Use **`test.sh`** to check that everything works
- Add your **`report.pdf`** to the folder (yes, export it in PDF even if you are using Word!)
- Name the zip folder **`lastname_studentid_hw3.zip`**:
  - Ex: Luigi D'Andrea will submit a file named **`dandrea_1234567_hw3.zip`**
  - If you are unsure which name to put, use the one in your institutional email account

Submission Instructions



- Upload the zip on your **institutional** Drive and make it **link-shareable** and **public** to anyone (an automatic script will download it).
- Make sure it is accessible via an incognito page of your browser!
- Do **NOT modify** the folder structure
- You have to submit the homework through the submission form on Google Classroom. You will be asked to fill a form with the requested information and the **link** to the zip you uploaded on Drive.

SAPIENZA NLP

Dataset

The dataset

- GAP Coreference Dataset[1]
- Languages: English

	Train	Dev	Test
English	2000	454	2000

Number of sentences for each split.

[1]: Webster, Kellie and Recasens, Marta and Axelrod, Vera and Baldridge, Jasoni: "Mind the GAP: A Balanced Corpus of Gendered Ambiguous Pronouns" *Transactions of the ACL* 2018.

The dataset

The data is available in a tsv format

Column	Header	Description
1	ID	Unique identifer for an example (two pairs)
2	Text	Text containing the ambiguous pronoun and two candidate names. About a paragraph in length
3	Pronoun	The pronoun, text
4	Pronoun-offset	Character offset of Pronoun in Column 2 (Text)
5	A ^	The first name, text
6	A-offset	Character offset of A in Column 2 (Text)
7	A-coref	Whether A corefers with the pronoun, TRUE or FALSE
8	B ^	The second name, text
9	B-offset	Character offset of B in Column 2 (Text)
10	B-coref	Whether B corefers with the pronoun, TRUE or FALSE
11	URL ^^	The URL of the source Wikipedia page

The dataset

The dataset is a tsv file where each row contains the following fields:

- development-965
- Merrill De Maris is a writer who worked on Disney Comic Strips for King Features Syndicate. De Maris helped Floyd Gottfredson with many of his early Mickey Mouse comic strips --
- his 139
- De Maris 92 FALSE
- Floyd Gottfredson 108 TRUE
- http://en.wikipedia.org/wiki/Merrill_De_Maris

The dataset

The dataset is a tsv file where each row contains the following fields:

- development-965
- Merrill De Maris is a writer who worked on Disney Comic Strips for King Features Syndicate. De Maris helped Floyd Gottfredson with many of his early Mickey Mouse comic strips --
- his 139
- De Maris 92 FALSE
- Floyd Gottfredson 108 TRUE
- http://en.wikipedia.org/wiki/Merrill_De_Maris

The dataset

The dataset is a tsv file where each row contains the following fields:

- development-965
- Merrill De Maris is a writer who worked on Disney Comic Strips for King Features Syndicate. De Maris helped Floyd Gottfredson with many of **his** early Mickey Mouse comic strips
  --
- **his** 139
- De Maris 92 FALSE
- Floyd Gottfredson 108 TRUE
- http://en.wikipedia.org/wiki/Merrill_De_Maris

The pronoun

The dataset

The dataset is a tsv file where each row contains the following fields:

- development-965
- Merrill De Maris is a writer who worked on Disney Comic Strips for King Features Syndicate. **De Maris** helped Floyd Gottfredson with many of his early Mickey Mouse comic strips
  --
- his 139
- **De Maris** 92 FALSE
- Floyd Gottfredson 108 TRUE
- http://en.wikipedia.org/wiki/Merrill_De_Maris

The first candidate

The dataset

The dataset is a tsv file where each row contains the following fields:

- development-965
- Merrill De Maris is a writer who worked on Disney Comic Strips for King Features Syndicate. De Maris helped **Floyd Gottfredson** with many of his early Mickey Mouse comic strips --
- his 139
- De Maris 92 FALSE
- **Floyd Gottfredson** 108 TRUE
- http://en.wikipedia.org/wiki/Merrill_De_Maris

The second candidate

The dataset

The dataset is a tsv file where each row contains the following fields:

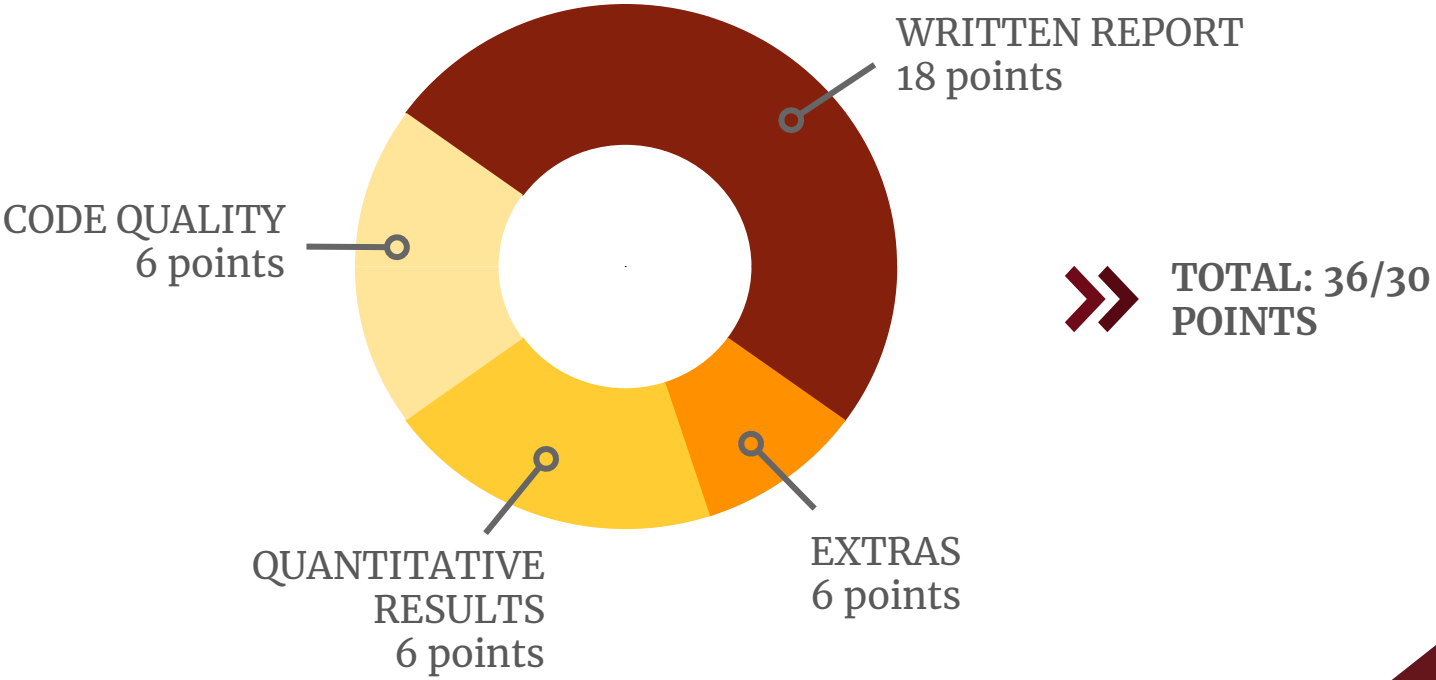
- development-965
- **Merrill De Maris** is a writer who worked on **Disney Comic Strips** for **King Features Syndicate**. **De Maris** helped **Floyd Gottfredson** with many of his early **Mickey Mouse** comic strips --
- his 139
- De Maris 92 FALSE
- Floyd Gottfredson 108 TRUE
- http://en.wikipedia.org/wiki/Merrill_De_Maris

THERE ARE MORE CANDIDATES IN THE SENTENCE

Evaluation

Evaluation

We will take into account the following criteria:



TOTAL: 36/30 POINTS

SAPIENZA NLP

Report: dos and don’ts

- **ACL 2021 paper template**
  - Available here (Word and LaTeX direct download) or here (Overleaf LaTeX template)
  - You can use either the LaTeX or the Word template, your choice
  - **DO NOT MODIFY** the template (margins, spacing, font size)
  - Use the non-anonymous flag, so you can enter your name
- **Max 3 pages**
  - For the report, including title, subtitles, etc.
  - This is a **STRICT RULE!**
- **Unlimited extra pages for images, tables and references**
  - Be sure to **include** and properly **comment** a confusion matrix, visualized as heat map
  - Every image and table must have a caption (don’t abuse them please :))
  - Tables and images must be referenced in the report

Report: what you are expected to do

We expect a good report to be:

- **Readable** and **understandable**
  - We will not give penalties for English errors, but we expect the report to follow a clear flow. We don't want to read just a sequence of statements on what you did without showing the reasoning behind your choices
- **Well-structured** and **organized**
  - Take inspiration from the many papers available online and organize your report in well-defined sections (e.g. method, setup, experiments, results...)

Report: what you are not expected to do

We expect a good report **NOT** to include:

- Unnecessary **task** or **dataset descriptions**
  - just focus on your solution to the problem.
- **Code** copy-paste
  - Your code should be self-explanatory, so no need to show it in the report. You can add **pseudocode** to show some particular algorithm, but **no code or screenshots,** please!

Report: what you are not expected to do

We expect a good report **NOT** to include:

- Unnecessary **low-level implementation details**
  - Avoid any **low-level implementation/technical details** like “I used a dictionary to store these values”, “I had to use configuration X to solve this exception”, “I could not use Y because there was a dependency issue with Z”, etc.
  - Instead, **we are interested in high-level abstractions/strategies** you decide to use to tackle the homework, as well as the **intuitions behind your choices**.

    E.g. use and description of a particular model, explanation of how and why an architecture works, etc.

Application: what you are expected to do

Your project should conform to the following rules:

- You **MUST** use PyTorch.
  - TensorFlow and other deep learning frameworks are **NOT** allowed.
  - PyTorch Lightning **is allowed**
- **Frameworks** that use PyTorch (e.g. AllenNLP, torchtext…) are **NOT** allowed.
- Libraries (such as tqdm, sklearn, NLTK) are fine, but since the line between a framework and a library is sometimes blurred, please ask in the Google Classroom group before using any external library: **any other** library **MUST be agreed with the TAs**.

Application: what you are expected to do

Your project should conform to the following rules:

- **You are now allowed** to use any architectures that have been explained in the course, in particular:
  - word embeddings (Word2Vec, GloVe, etc.) **are allowed**,
  - contextualized word embeddings (ELMo, etc.) **are allowed**,
  - Transformer-based models (BERT, BART, RoBERTa, etc.) **are allowed**.
- For any doubt, please ask the TAs on Google Classroom.
- **Comment** your code, please!

Quantitative Results

We will evaluate the **performance of your model** on a SECRET test set.

Also in this homework, your model must pass a **baseline score** for each (optional and mandatory) subtask for us to consider it.

The homeworks that surpass the baseline will be graded according to different thresholds, that will be defined based on an internal reference model and the normalized distribution of **YOUR** scores.

Extras

You can achieve **up to 6 points with some extra work!**

See Homework Extras section for some suggestions about what we consider an extra.

Don’t forget to **explain your choices** in the report! Extras that are not explained in the report will not be considered for evaluation.

Please note that you are not allowed to use tools that have not been explained **yet** in the course. For any doubt, please contact the TAs.

Evaluation

- `test.sh` is identical to what we will be using
- **If it does not run on your side, we will not correct your homework**
- Note that, if you use **any kind of hard-coded paths**, this script **won't work**
- Use **paths relative to the project root folder**, e.g.:
  - **NO:** `/home/pincopallino/my_folder/model/weights.pt`
  - **OK:** `model/weights.pt`

Warnings

Things you should be aware of

Please be aware that

This is an **individual exercise!** Collaboration among the students is **not** allowed.

We will check for **plagiarism** both manually and automatically.

It is **not allowed** to:

- Copy from other students
- Share your code with other students
- Copy from online resources (StackOverflow, GitHub, Medium, Kaggle and so on).

While we release the homework on GitHub, **DO NOT FORK THE PROJECT.**

However, you are allowed to use material from **external sources** as long as it is **not central** to the homework.

You are also allowed to use the **SOME** parts of the presented class notebooks. However, you **MUST** explicitly specify these parts in your code comments.

Use of external data

- For your experiments, **use the provided data** (train and dev) in the data folder; use each file as defined in the standard ML conventions (train for training, dev for model selection).
- If you train it using external data, include it in the data folder.
- If you train it on dev set, it will be a **FAIL**.

Tips

A few tips to organize your work:

- Test the testing infrastructure as soon as possible so if you have any issue you can contact us and try to solve the issue
- Start **as soon as possible**!
  - Training a neural network requires time, possibly hours, depending on your hardware
- Start **small**!
  - If you don't get decent results with a very simple neural network, there is a good chance that adding other things won't make your model perform better
- Leave some time for **hyperparameter** tuning!
  - Sometimes good hyperparameter combinations can do wonders for your neural network
- Use **Google** Colab (free GPUs!)

Deadline

When to deliver what

Deadlines

The students **who passed the first homework** may deliver the second one in one of the four available deadlines (2022):

1. Early submission: June 23rd (23:59 CEST) → only this date allows **late submission!**
   Late submission: June 26th (23:59 CEST)
   Presentation: 30th June

2. Submission: July 14th (23:59 CEST)
   Presentation: July 21st

3. Submission: July 19th (23:59 CEST)
   Presentation: July 26th

4. Submission: September 12th(23:59 CEST)
   Presentation: September 19th

Late submission policy

- For the first deadline **ONLY**, we will allow **late submissions** until the 26th of June.
- Also in this case, we will apply 1 point of penalty for each day of delay.
- Example: A student delivers their homework on June 26rd (17:35) ⟶ max possible grade 36 - 3 = 33.

Final Exam

Presentation

To complete the exam, you will have an **oral presentation** of all the homeworks.

Presentation will be done in classroom/Google meet **with slides.**

Homework 1+2 presentation:

- 12 minutes presentation. This is a **very strict requirement.**
  A split of 5 minutes for the first homework and 7 for the second should work.
- 10 minutes for questions

Homework 1+2+3 presentation:

- 15 minutes presentation. This is a **very strict requirement.**
- 10 minutes for questions

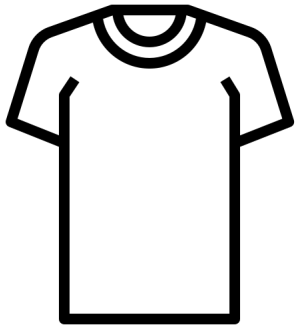
Awards 

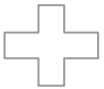
Get a **Sapienza NLP™** t-shirt

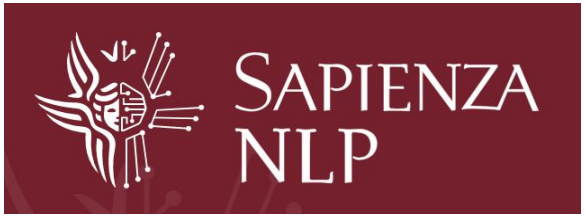
Win a Sapienza NLP t-shirt!

We will hand out amazing Sapienza NLP t-shirts to the **overall top-5** students!

The final ranking will be computed according to the scores on our **secret** test set.







That’s not all

If your work is novel, interesting and original, we will gladly invite you to work together with us to extended on a fully-fledged paper for **TOP-TIER INTERNATIONAL CONFERENCE**!

Just over the last 12 months, the Sapienza NLP group published more than a dozen of papers!

Questions?

If you have a question that may interest your colleagues, **please ask it on Google Classroom.**

Otherwise, for personal or other questions, email **ALL** of us (but please, only reach for things that can't be asked on the Google Classroom).

Our emails are:

{bejgu, martinez, orlando, scire, tedeschi}@diag.uniroma1.it

Good Luck!!

Dataset

UniteD-SRL (Tripodi et al., EMNLP 2021)

The data is available in a format that is very similar to the [CoNLL-U] format

```
# document_id = 1997/crc/c/69
# sentence_id = 216:2
# text = The Committee notes that this reservation might impede the full impl
0   The the _   _   _
1   Committee   committee   _   agent   _
2   notes   note    SPEAK   B-V _
3   that    that    _   topic   _
4   this    this    _   _   _
5   reservation reservation _   _   agent
6   might   might   _   _   _
7   impede  impede  STOP    _   B-V
8   the the _   _   _
9   full    full    _   _   _
10  implementation  implementation  _   _   theme
11  of  of  _   _   _
12  the the _   _   _
13  Convention  Convention  _   _   _
14  .   .   _   _   _
```

- Column 1: token index in the sentence.
- Column 2: tokens in the sentence.
- Column 3: lemma of each token in the sentence.
- Column 4: VerbAtlas frame of a predicate. If the token is not a predicate, then __.
- Column `4 + i`: Roles of predicate i. E.g. Column 5 lists the roles for the first predicate, i.e., `insists`.

Dataset

UniteD-SRL (Tripodi et al., EMNLP 2021)

The data is available in a format that is very similar to the [CoNLL-U](#) format

```
# document_id = 1997/crc/c/69
# sentence_id = 216:2
# text = The Committee notes that this reservation might impede the full impl
0   The the _   _   _
1   Committee   committee   _   agent   _
2   notes   note    SPEAK   B-V _
3   that    that    _   topic   _
4   this    this    _   _   _
5   reservation reservation _   _   agent
6   might   might   _   _   _
7   impede  impede  STOP    _   B-V
8   the the _   _   _
9   full    full    _   _   _
10  implementation  implementation  _   _   theme
11  of  of  _   _   _
12  the the _   _   _
13  Convention  Convention  _   _   _
14  .   .   _   _   _
```

- Column 1: token index in the sentence.
- Column 2: tokens in the sentence.
- Column 3: lemma of each token in the sentence.
- Column 4: VerbAtlas frame of a predicate. If the token is not a predicate, then __.
- Column `4 + i`: Roles of predicate i. E.g. Column 5 lists the roles for the first predicate, i.e., `insists`.

VERY IMPORTANT!

Dataset

UniteD-SRL (Tripodi et al., EMNLP 2021)

The data is available in a format that is very similar to the CoNLL-U format

```
# document_id = 1997/crc/c/69
# sentence_id = 216:2
# text = The Committee notes that this reservation might impede the full impl
0   The the _   _   _
1   Committee   committee   _   agent   _
2   notes   note    SPEAK   B-V   _
3   that    that    _   topic   _
4   this    this    _   _   _
5   reservation reservation _   _   agent
6   might   might   _   _   _
7   impede  impede  STOP    _   B-V
8   the the _   _   _
9   full    full    _   _   _
10  implementation  implementation  _   _   theme
11  of  of  _   _   _
12  the the _   _   _
13  Convention  Convention  _   _   _
14  .   .   _   _   _
```

- Column 1: token index in the sentence.
- Column 2: tokens in the sentence.
- Column 3: lemma of each token in the sentence.
- Column 4: VerbAtlas frame of a predicate. If the token is not a predicate, then __.
- Column `4 + i`: Roles of predicate i. E.g. Column 5 lists the roles for the first predicate, i.e., `insists`.

VERY IMPORTANT!

You can safely ignore the B-V tags (they indicate the position of the predicate).