Guidelines for Annotating the Dataset of FoRC (subtask II)

1. Introduction

In this project, you will annotate **500** publications in the field of Computational Linguistics (CL) according to a hierarchical taxonomy of specific topics and sub-topics within CL.
Simply put, your job will be as follows:

- Thoroughly review the taxonomy of CL sub-topics.
- Review the current CL paper you are presented with.
- Label the paper with one or more sub-topics from the taxonomy.
- Move on to the next paper.

The documents you will annotate are all extracted from the ACL Anthology (https://aclanthology.org/), which is a big corpus of the proceedings of the most prominent conferences in the field of CL.

The core idea behind this annotation project is to develop a dataset of labeled CL papers, which can then be used for training classifiers to automatically label papers into specific and fine-grained topics and sub-topics within CL. These classifiers can then be used for downstream applications such as scientific search engines and recommender systems to better serve the needs of researchers in the community. Moreover, the classification of research topics will give a nice overview of past and current topics and their advancements and will allow the comparison of different findings.

Now that you have an idea of the project, let's delve into the specifics.

2. Using INCEpTION

Downloading and installing

In this project, we will use the annotation tool INCEpTION. For that you will need to:

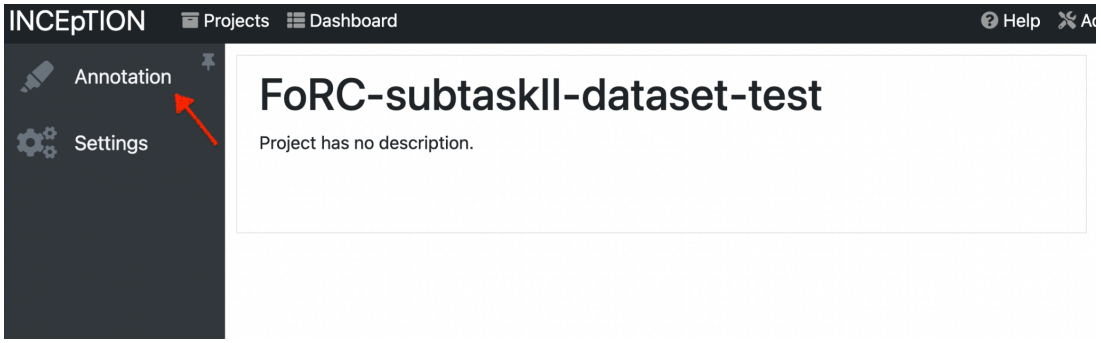
- Go to the INCEpTION website: https://inception-project.github.io/.
- Download the latest INCEpTION version by clicking on the 'Download' box on the right side of the page.
- A step-by-step guide for installation can be accessed here: Installing and Starting INCEpTION.

Accessing our project instance

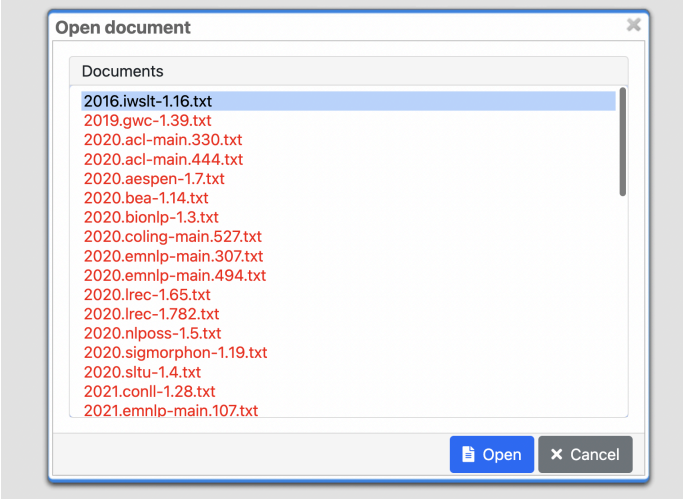
To access our INCEpTION instance, you will need to connect to the DFKI VPN first. After you are connected, you can access the instance here: http://172.16.150.205:8080/?2.
Please use the username and password provided privately to you.

3. Accessing documents to annotate

Once you have access to the project, upon opening it, you should see a screen similar to this one:



Click on the 'Annotation' option to view the list of documents you have to Annotate. You should be able to see a list similar to the following:



Highlight a specific document and click 'Open' to go to the annotation interface.

4. How to annotate

This is the important phase where you will have to review the documents and choose the correct label(s). Kindly note that your full attention should be given to this task, so make sure there are no distractions around and that you are able to focus on annotating. We recommend taking a break of at least 10 minutes after annotating a batch of 15-20 documents.

The document

The document displayed before you includes the title and abstract of a paper from ACL Anthology. Note that the name of the document includes an 'acl anthology id'. E.g.:

FoRC-subtaskII-dataset-test/2016.iwslt-1.16.txt
Project name ACL ID

To view the full PDF document of the paper before deciding on one or more labels, you can copy the ACL ID and paste it into the following URL template:

https://aclanthology.org/2016.iwslt-1.16.pdf
ACL Anthology domain ACL ID

This will give you access to the full paper so you can review it more thoroughly before making an annotation decision.

The taxonomy

For this project, we developed a hierarchical taxonomy of CL topics and sub-topics in a semi-automatic manner. It is important to familiarise yourself with all the taxonomy labels so that you are aware of all annotation options.

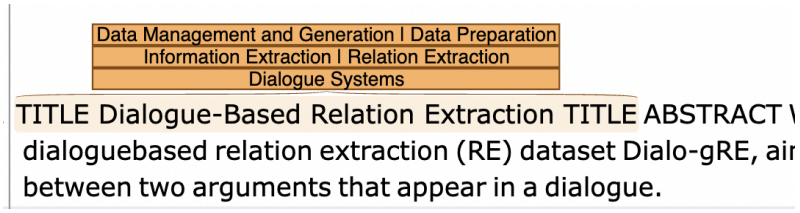
You can view the taxonomy labels in these manners:

- Tree view: Taxonomy - a Hugging Face Space by katebor
- Flat view (helpful for ctrl+F): label definitions - NLP taxonomy

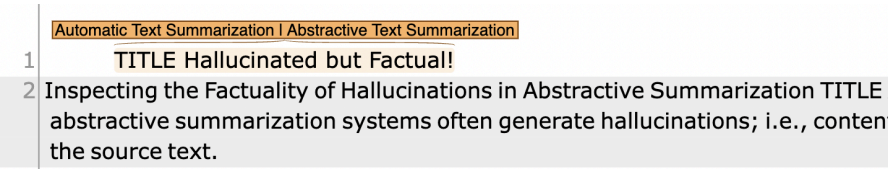
The annotation span

Once you reviewed the document and the taxonomy, choose the labels that correspond to the main ideas and contributions of the document. Once those are chosen, follow these steps:

- IMPORTANT: highlight the annotation span: the span should start from the first TITLE token and end with the second TITLE token, like so:



- If you are not able to do that because the second TITLE token is **in a different sentence**, please make sure to highlight the entire first line, like so:



Choosing the right label

After highlighting the correct span, you need to choose the label(s). In the right-side column, you will see the annotation options. The 'Layer' should be set to 'FoR-hierarchy' and the 'Text' should denote the span you highlighted.

Then, you will see three annotation options, namely Level1, Level2, and Level3, which correspond to the hierarchy levels in the taxonomy.
After choosing the Level1 label, you have the option to select more fine-grained labels within that field. Note that when you choose Level1 label, all its child nodes will be bolded and moved to the beginning of the drop-down list of Level2. The same goes for child nodes of Level2 in the Level3 drop-down list.
Only choose bolded options from Level2 and Level3!

IMPORTANT! Please choose the most specific and fine-grained label possible. E.g., if the paper is about 'Neural Machine Translation', it is not enough to choose the label 'Machine Translation' in Level1. You have to choose 'Neural Machine Translation' in Level2 as well.

IMPORTANT! INCEpTION technically gives you the option to choose Level2/Level3 labels that are not a child node of the already chosen Level1 label. Please refrain from doing that and double-check with the hierarchy tree.

Note that it is possible (and advisable!) to choose more than one label for the same document. To add a new label, simply highlight the span again and repeat the annotation process. You should be able to see all of your annotations highlighted in the text with the labels you chose.

Once you are done annotating one document, lock the document by pressing the lock icon and navigate to the next document using the arrow. Note that once locking the document you cannot go back and change it.

Annotation examples

In this section, we will give you some example annotations to demonstrate some correct and incorrect labels. Remember that only the main ideas and contributions of a document should be labelled. So, e.g., if a paper is about using different models for Sentiment Analysis, there is no need to label the document with all the models used.

Example 1

Straight-forward labels: Some documents will be straight-forward in terms of their main topics and contributions. For example, the following paper:
https://aclanthology.org/2020.nlposs-1.5.pdf

Can be labeled as:

- Biases in NLP -> Gender Bias
- Embeddings -> Word Embeddings

Example 2

Choosing a higher-level (non-specific) label: In some cases, the papers discuss some specific topics that do not exist in the taxonomy. In those cases, you should tag the higher-level topic, even if a specific one does not exist. For example: https://aclanthology.org/2020.emnlp-main.307.pdf

Can be tagged as:

- Model Architectures

Since one of its main topics is using a variational autoencoder for sentiment analysis. We label it as 'Model Architectures' since there is no 'Variational Autoencoder' label in the taxonomy.

Example 3

Implicit label: note that not all labels will appear explicitly in the abstract or the text of the paper. For example: https://aclanthology.org/2021.conll-1.28.pdf

Should be tagged, among other labels, as:

- Discourse Analysis

Even though the exact phrase does not appear in the text. However, the main topic of the document is constructing a corpus of presuppositions in English, which is a topic within discourse analysis.

Note that if, in your opinion, there is no appropriate label, you can leave the document unlabeled. If this happens, please compile a list of all unlabeled documents and a suggestion of what they could be tagged as (that does not exist in the taxonomy). This way we can have feedback about which topics need to be added to the taxonomy.

5. Remember to check

The details on dataset and model architectures are often not mentioned in the title and abstract. In order to identify a novel dataset/model architecture and the language of data, we **encourage** you to always look at the respective sections in the full papers.

6. Questions to ask yourself while annotating

a. Do the authors propose a new architecture? If yes, the paper should probably be labelled as Model Architectures.
b. Do the authors propose a new dataset? If yes, use the label Data Preparation.
c. Is the dataset based on a low-resource language? Remember that any language other than English is regarded as low-resource.
d. Is it dealing with a specific domain? Use the Domain-specific NLP label or its subtopics.
e. What is the learning paradigm?

7. Additional support

If you have any questions or concerns regarding any part of this process, please contact us! We would be happy to help!

Thank you for participating!

Raia Abu Ahmad (raia.abu_ahmad@dfki.de)
Ekaterina Borisova (ekaterina.borisova@dfki.de)