Final Project

# Team sign up

### In the final project, you can form a team with maximum three members. Exceptions can apply if you choose a hard level (by combining two or more medium hard level ideas). In such case, a team can have four members. Please talk to instructor for approval.

You can use the discussion board on Canvas to help team formation. Note that all members in the same team will have the same grade. So discuss your roles, responsibilities, and your team protocols when you form a team.

# Option 1

Extend the research on QTL NER task (<https://github.com/liyp0095/CuPUL>)

## Direction 1 (research oriented. Difficulty level: medium hard – hard)

Improve strict F1

If you choose this direction, you can try the following ideas

1. One reason of low strict F1 is because that token level prediction didn’t consider span level information (for example, phrase structure), causing issues such as incorrect boundary and missing predictions for long entities. You can consider a post-processing with help of constituency parsers or phrase extractions to correct such mistakes. (Difficulty level: medium hard)
2. Another reason is because of bias in training. Due to the dictionary matching method, it is harder to annotate long entities comparing to short entities. You can consider fuzzy matching to improve the distant annotation quality. (Difficulty level: medium hard)

Direction 2 (research oriented or project oriented. Difficulty level: easy – medium hard – hard)  
Consider discontinued NER

Many entities in the annotations are discontinuous entities. For example, in “milk protein and fat content”, the current annotation is “milk protein” and “fat content”, but the better annotation is “milk protein content” and “milk fat content”. For the course project, let’s focus on discontinuous entities due to conjunctions. (<https://github.com/qingwang-isu/CoRec>)

If you choose this direction, you can try the following ideas

1. Re-annotate the test data (you can design methods which can automate the process in certain degrees, but you will still need to manual check some to ensure quality) (Difficulty level: easy, but effort consuming)
2. Re-design the dictionary matching method to annotate the discontinued named entities. You can consider fuzzy matching to improve the distant annotation quality. (Difficulty level: medium hard)
3. Design a post-processing algorithm to allow discontinued named entities from current results. (Difficulty level: medium hard)

## Direction 3 (project oriented. Difficulty level: easy – medium hard)

Develop a web application for annotation visualization (<https://spacy.io/usage/visualizers>).

User query input: PMID -> Display: title, abstract, and trait annotations

1. Fetch paper through PMID (here is a tutorial <https://erilu.github.io/pubmed-abstract-compiler/>)
2. Get the NER result (you can run the dictionary matching method instead of NER model)
3. Visualize the result in a web application

# Option 2

Evaluation of Note taking and short answers from different scientific domains.

## Direction 1 (research oriented. Difficulty level: easy – medium hard)

Classification task for note taking. Data is provided in note\_classification.zip

Background: In this experiment, the participants were asked to watch a lecture recording of a specific subject and take notes. The experiment aims to study if the participants captured all key points (IdeaUnit) from the lecture.

You are given the Idea Unit of the lectures (provided by the lecturer), and the entire note of each participants, and annotation for one note per subject as training. Your task is to classify the rest of the notes under this 1-shot setting.

You will need to develop and compare three models, such as rule-based model, BERT-based model, and LLM-based model.

#### Data description

Notes.csv: Experiment,Topic,ID,Segment1\_Notes,Segment2\_Notes,Segment3\_Notes,Segment4\_Notes

Test.csv & train.csv: Experiment,Topic,ID,Segment,IdeaUnit,label

You can retrieve the note for each record in train.csv by ID and Segment. For example, the first record in train.csv has “6260226” and “1” as ID and Segment, respectively. You can find the corresponding note with ID “6260226”, Segment1\_Notes from Notes.csv to get the full note.

You will need the note, IdeaUnit, and label to train the model. Label=1 means the note covers the required IdeaUnit. Label=0 means the note does not cover the required IdeaUnit.

You will test your model performance on Test.csv

## Direction 2 (research oriented. Difficulty level: easy – medium hard)

Classification task for short answer questions. Data is provided in short\_answer\_grading.zip

In this task, participants were asked with several short answer questions. You will need to train/prompt a model to grade those short answer questions.

You will need to develop and compare three models, such as rule-based model, BERT-based model, and LLM-based model.

#### Data description

Test.csv & train.csv: Experiment,Topic,ID,Question,Response,CorrectAnswer,label

You will need the Question, Response, CorrectAnswer, and label to train the model. Label=1 means the response is correct. Label=0 means the response is similar to the correct answer but not precise. Label=-1 means the response is incorrect.

You will test your model performance on Test.csv.

# Other Options

If you are a PhD student working on related research projects, you can propose your own project. Please schedule a meeting with the instructor for approval.

# Submission (All deadlines are at 11:59PM UTC-12:00 (“anywhere on Earth”).

### Stage 1. Due March 24

Sign up group on Canvas, and submit the following information on Canvas:

1. Title (of your choice)
2. Team member: name, email, and affiliation
3. Team protocols: team meeting format and frequency; time commitment, role and responsibility of each member; if you plan to extend the course project into a paper submission/creative component, you will need to discuss authorship

### Stage 2. Due May 2

You will also need to submit the following documents on Canvas

1. your code (optional: readme)
2. a draft report (max 4 pages) in pdf using this template: <https://www.overleaf.com/latex/templates/association-for-computational-linguistics-acl-conference/jvxskxpnznfj>

There is no page limit on Limitation, References, and Appendix.

Your report will follow the principle of short paper requirement of ACL (<https://aclrollingreview.org/cfp#short-papers>), and follow the ACL policy of publication ethics (<https://www.aclweb.org/adminwiki/index.php/ACL_Policy_on_Publication_Ethics>)

Required component includes:

1. Title
2. Authors: name, email, and affiliation
3. Abstract, Introduction, Related work, Methodology, Experiments, Conclusion, Limitation, and References.

Optional component includes: Appendix.

### Stage 3. Due May 14

You will need to submit the following items on Canvas

1. A video recording of your final presentation. (max 10 minutes long)
2. Your final report (max 5 pages) in pdf