FINAL PROJECT

Deadline: 11:59 pm, April 17, 2025

Submit via Blackboard with a VeriGuide receipt.

Please follow the course policy and the school's academic honesty policy.

1 Overview

The final group project is designed to give students a chance to collaborate and apply what they have learned in class to an NLP task, specifically focusing on Named Entity Recognition (NER). The group composition will remain the same as for the paper presentation in the mid-term presentation. This project includes two components: a presentation and a written report. Together, the final project accounts for 30% of the course's total grade.

1.1 Named Entity Recognition Task

Named entities are real-world objects such as people, organizations, and locations that can be identified with proper names. The purpose of Named Entity Recognition (NER) is to detect and categorize these named entities within text into predefined groups. Here is an example:

[ORG U.N.] official [PER Ekeus] heads for [LOC Baghdad].

In this instance, the NER system must identify the terms "U.N.," "Ekeus," and "Baghdad," and classify them into the categories ORG (organization), PER (person), and LOC (location), respectively.

1.2 Dataset

The core dataset for this project is the CoNLL-2003[1] [4] corpus, a widely recognized benchmark for English NER tasks. While the original CoNLL-2003 also includes German annotations, this project focuses exclusively on its English subset, which contains labeled named entities extracted from news articles. The dataset retains its standard partitioning into training, development, and test sets to ensure consistency with prior research.

To streamline implementation, we recommend leveraging the **datasets** library from Hugging Face, which provides direct access to preprocessed CoNLL-2003 data. This integration simplifies data loading and compatibility with modern NLP frameworks like PyTorch or TensorFlow.

For extended exploration (optional), you may evaluate your model on complementary datasets such as the English part of the WikiAnn[2] [2]. Comparing performance across distinct domains (e.g., news vs. wikipedia) could yield insights into model generalizability. While this extension is not mandatory, it would enrich the project's analytical depth.

[1] https://huggingface.co/datasets/eriktks/conll2003
[2] https://huggingface.co/datasets/unimelb-nlp/wikiann

1.3 Evaluation

The primary metric for the NER task is the F-1 score. When $\beta = 1$, the equation simplifies to the F-1 score, which is the harmonic mean of precision and recall.

$$F_\beta = \frac{(\beta^2 + 1) \times \text{precision} \times \text{recall}}{\beta^2 \times \text{precision} + \text{recall}}$$

Precision is the ratio of correctly identified named entities by the system. Recall is the ratio of named entities in the dataset that the system identifies. A named entity is correct only if it exactly matches the corresponding entity in the data file. **The F-1 score on the test set is the primary metric for this final project.** It will be the key metric for fair comparison across groups.

In addition to the F-1 score, you can introduce other metrics for system evaluation. You can assess your system in a few-shot setting (e.g., using only 32 randomly selected samples for training). You may also measure time and memory usage. Innovative methods that provide insights into your system from various perspectives are highly valued.

1.4 Method

You can use any method from a paper published in the last five years as your baseline. For instance, you might consider *Deep Contextualized Word Representations* [3] as a reference. You're encouraged to enhance your baseline with techniques like:

- Data augmentation
- Advanced training algorithms
- Complex network structures
- Pretrained language models
- Custom engineering techniques

When using methods from other papers, cite them and clearly indicate what your original contributions are in your report. This includes implementing non-open-source techniques or providing detailed comparisons of different methods not found in published work.

If you introduce new methods, explain the issues existing methods don't address and how your approach could solve them. It's okay if your methods don't work as expected, but ensure you conduct a thorough analysis and clarify why the results differ from your expectations.

1.5 Computational Resources

For this project, you may require substantial computational resources, especially if training machine learning models. We strongly recommend using Google Colab [3] as your primary development environment. Colab provides free access to GPU and TPU resources, eliminating the need for high-end local hardware. Its cloud-based Jupyter notebook interface allows seamless collaboration, automatic dependency installation, and cross-device accessibility. While the free tier has usage limits (e.g., session timeouts), it remains ideal for most coursework-level tasks. Start early to avoid last-minute resource constraints!

[3] https://colab.research.google.com/

2 Submission

2.1 Deliverables

Project Presentation (40% of the final project grade, i.e., 12% of the overall grade of this course): At the Week 13 lecture and tutorial (Apr. 15 and 17), each group will present their work.

- You have 8 minutes to describe your work and 2 minutes for Q&A.
- There is no need for all members to do the presentation.
- Detailed time slots for each team's project presentation will be determined and announced before Week 13.

Project Report (60% of the final project grade, i.e., 18% of the overall grade of this course): After the project presentation, each group will have one more week to finish their project report.

- There is no requirement on the report templates, and you are free to use anything that you feel comfortable with (e.g., LaTeX, Word, etc.) as long as you deliver your work clearly.
- The outcome should be a single PDF file that contains all contents (please see the later Section *Report Contents*).

Project Code Your final project code should also be submitted together with your report. You only need to submit code files that you write or modify and do not submit external libraries, packages, or open-sourced code bases. Although we do not grade your code, we will go through it to check the reproducibility, and it is also required in case of plagiarism.

Supplemental Material (optional) The supplemental material is not required. If you do make some efforts that are hard to incorporate in the presentation and report, you can submit them as supplemental material. We will appreciate all your efforts.

2.2 Due Dates and Submission Notes

Slides Submission The due date of slides submission is Apr. 17, 23:59 PM for submitting the project presentation slide. All teams need to submit their slides to Blackboard before this deadline. Each team only needs one team member to do the submission and the latest slide will be used for the presentation. You will use your own laptop and the projector in the classroom for the presentation. Please bring an adapter as the projector only has an HDMI port.

Report Submission The due date of report submission is Apr. 24, 23:59 PM for submitting the project report and other materials. You are given one more week to work on the project report after the project presentation because the Professor and TAs might give feedback during the project presentation, and you might want to incorporate those suggestions into your report. You should submit only one zipped file (containing your project report, code, and other materials) via the Blackboard and each team only needs to designate one member to finish the submission. Multiple submissions are accepted and only the latest version will be graded.

3 Report Contents

The first page of your report should include the report title, team name, team member information (their names and student IDs). Next is the main part of your report. It should have 6 ~ 8 pages (except References and Appendix), and you might include but do not limit to the following parts as following.

Abstract Briefly describe what you have done in this project, e.g., what problem you solved, what method you proposed, what phenomenon you observed, etc.

Introduction Explain what the purpose of this final project is, a basic introduction of the current status of the field that this project is on, what the problem this final project aims to solve, the motivation, what you have done in this final project, a summary of your experiment results, etc. You might also state what your own work is here.

Related Work Give a survey of published papers that are related to your final project, e.g., papers that serve as your baselines, papers that have a similar idea to yours but in other tasks, papers that solve your problem differently, papers that inspire you, etc.

1. There might be many related papers, and you might not have enough time to walk through them. You might only include 3 4 recently published and most related papers, though more related works are encouraged to be discussed here.
2. There is no need to discuss all the details of the papers you included here. You only need to give a brief introduction to the ideas behind these papers.

Methods Provide details about the methods this project will use, and explicitly indicate your original work. For example, if you propose a new neural network architecture, you might differentiate it from techniques that you used but were not originally proposed by you (and give citations to them). You might also include a detailed description of the baselines (like BERT[4] [1]). You are also encouraged to use diagrams to better illustrate your ideas.

Experiments You might include the following parts:

1. **Data**: Describe details about the data preprocessing steps, hyper-parameters, and etc. If you use other datasets, please also describe them in detail and give proper citations.
2. **Metric**: If you define a new metric, please give a detailed description as well as a justification of your metric.
3. **Experiment Details**: Describe other details that are crucial to reproduce your experiment, e.g., hyper-parameters of your method if you propose something new in your report. You might also give links to the open-source repositories if your work is adapted from them and specify which part is implemented yourself.
4. **Results**: You might use tables or figures to illustrate the final performance of all methods that you experiment with. You might give a detailed comparison of their pros and cons. You are encouraged to highlight interesting observations that appear in your experiments.
5. **Analysis**:

[4] https://huggingface.co/google-bert/bert-base-uncased

   a. If you observe any interesting phenomenon in your experiment, please give a detailed analysis on why this happens (and additional experiments if needed to justify your conclusion).
   b. You should give an analysis on why and when your method works, how each part of your method contributes to the final performance (a.k.a., ablation study), what are the limitations of your method, or why it does not work at all.
   c. Please provide solid experiments to support any other conclusion that you might state.

6. **Conclusion**: Briefly summarize your work here. It is similar to the abstract, but you can discuss more here, e.g., the future direction, potential pitfalls, etc.
7. **Limitation (Optional)**: You should discuss any potential risk of your proposed method and features that should be improved.
8. **References**: You have unlimited pages to put all your references here. We do not require you to use any specific reference format, but please use one of the standards like APA.
9. **Appendix (Optional)**: You have unlimited pages to put your additional experiments, proofs, or other contents that are related to your report but do not affect the understanding if not placed in the main part. This section should be placed after References.

Your report should also contain a **Contribution** section in the main part. In this section, you will state which part of your project is done by which members. This helps to ensure that every member contributes to the group project. In principle, all members in the same team will receive the same grade, but we might weigh the grade differently for each member if the workload is truly unbalanced.

4 Rubrics

Here are the scoring criteria for evaluating presentations and reports. You can use them as a reference to improve your presentations and reports. The rubrics for the presentation and report are presented in Table 1 and 2, respectively.

Table 1: Rubric for project presentation. Weight is the percentage of one criterion in the grade of the whole project presentation.

Criteria	Exemplary (5)	Needs Work (1)	Weight
Structure	• The presentation has a concise and clearly stated focus. • The presentation is well-structured with a clear storyline. • Ideas are arranged logically; they strongly support the presentation focus. • Sections are well-connected with a smooth transition.	• The presentation lacks a focus. • The presentation is ill-structured. • Ideas are presented without obvious order or logical connection. • Transitions between sections are jumpy.	20%
Content	• Materials are coherently organized, demonstrating the presenter's mastery of the subject knowledge. • All materials presented are relevant and lead naturally to the conclusion. • Ideas are supported by evidence, with appropriate use of facts, examples, statistics, and references.	• The content is fragmented; it fails to demonstrate the presenter's subject knowledge. • The materials presented are not clearly linked to the conclusion. • Ideas are stated without support or references.	40%
Communication	• The presenter is fluent and articulate; the use and variation of tone and pace are effective. • The presenter demonstrates good grammar and choice of words. • The presenter answers questions clearly and timely.	• The presenter does not speak clearly, speaks too fast or too slowly, rarely uses tone or pace variation to help the delivery. • The presenter uses very limited vocabulary and poor grammar. • The presenter could not answer questions.	20%
Visualization	• Visualizations are clear, relevant, and well-designed. • Creative effort is evident in making the presentation more captivating.	• Visualizations are irrelevant, difficult to understand, or poorly designed. • Ineffective use of media.	10%
Time management	• The presentation has a specific time limit.	• The presentation lasts less than or more than the time limit.	10%

Table 2: Rubric for project report. Weight is the percentage of one criterion in the grade of the whole project report.

Criteria	Exemplary (5)	Needs Work (1)	Weight
Organization	• Each part of the report is connected naturally by a storyline and supports each other. • The content of the report is self-contained.	• The report does not have a clear focus and it is jumpy between sections. • Readers need a strong background or extra effort to understand the content.	5%
Correctness	• References are correctly formatted according to the standards and the information is exact. • Citations are properly added to give credits for existing works. • Equations (if any) and symbols are used correctly. • Baselines/literature are correctly implemented/understood.	• References are missing or the information is incorrect. • Miss most citations for unoriginal work. • Equations and symbols are used incorrectly. • Baselines/literature are implemented/understood incorrectly.	25%
Completeness	• The report reviews a complete list of the most related work in recent years. • Methods are compared thoroughly and fairly. • Experiments are well-designed to investigate different aspects of the models or algorithms.	• The report does not review the most related work in recent years. • Methods are compared in part or under an unfair setting. • Have not performed a complete investigation of models or algorithms.	25%
Clarity	• The report contains only a few grammar errors and typos. • Sentences are fluent and paragraphs break naturally. • Descriptions are clear and easy to follow. • Equations (if any) and symbols are well-explained. • Tables and figures could be easily understood with captions only.	• The report contains many grammar errors or typos. • Sentences and paragraphs look unnatural. • Descriptions are hard to understand. • Equations, symbols, tables, and figures have no explanation or it is incomplete.	15%
Analyses	• The report clearly states the motivation. • The choice of baselines and the design of new methods are well-motivated and well-justified by experiments or references.	• The report does not provide the motivation or it is unclear. • The report does not provide evidence to support the choice of baselines or the design of new methods.	25%

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019.

[2] Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[3] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics, 2018.

[4] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147, 2003.