

Projects - Neural methods for NLP

Master LiTL 2023-2024

chloe.braud@irit.fr

Each project will be realized in groups of 2 or 3 students. The evaluation will correspond to a report, the code and a defense:

- **Part 1:** assignment due on **15/01/24**, send by email:
 - first part of the report (pdf file) as described below
- **Part 2:** assignment due on **12/02/24**, send by email:
 - the entire report (pdf file),
 - the code (notebook or python files),
 - if needed data (e.g. drive link).
- Each file must contain “project X” + the **names of all students of the groupe**
- **Defense** on **13/02/24**: each group will present its project (about 15mn).

The groups of **2 students** will work on a task corresponding to **document classification**.

The group of **3 students** will work on a task corresponding to **sequence labeling**.

For each project, I propose a specific task, combined with a specific setting and some relevant datasets. Each group can propose some changes, either in the task or datasets, the main point is to keep the setting (i.e. low-resource, bias, sequence labeling). For the sake of simplicity, we will use datasets already available on the HuggingFace hub: <https://huggingface.co/datasets>.

Topics:

- Project 1: Sentiment analysis on a low-resource language (2 students) → <https://fr.overleaf.com/9873667126msjxzgqwxfgw#a011a2>
- Project 2: Classifying biographies under gender bias (2 students) → <https://fr.overleaf.com/6391416295ndccbwtpskrh#04728b>
- Project 3: Cross-domain Named Entity Recognition (3 students) → <https://fr.overleaf.com/6119332344jtqsgtpdbjnr#71d715>

Project 1: Sentiment analysis on a low-resource language (2 students)

The goal of the project is to build a system for sentiment analysis for a low-resource language. We will simulate this setting by using: a dataset in English considered as the **source**, and a dataset in any other language than English corresponding to our **target**. A. Our goal is to transfer our knowledge of the task from source to target, using multilingual embeddings.

The analysis of the results should investigate performance drop when using multilingual embeddings compared to monolingual ones (baseline system on the source language), the performance obtained when no data is available for the target language (transfer learning) and the number of source examples needed to improve these performance (semi-supervised setting).

Task: sentiment analysis

Setting: transfer learning for low-resource language

Datasets:

- Source language (English): https://huggingface.co/datasets/rotten_tomatoes
- Target language (e.g. Czech):
https://huggingface.co/datasets/fewshot-goes-multilingual/cs_csfd-movie-reviews

Part 1: study the data and read a research paper

Prepare the first part of your report that should contain:

- a precise description of the data (both source and target), give the statistics and all relevant information. Be careful: the label sets could be different, you are allowed to map the labels to homogenize the corpora (mandatory for the proposed transfer learning strategy).
- the description of a research paper on a similar task: ideally, find a paper presenting results on transfer learning for the task (if you find a paper with results in the same language as the target language, report the results for comparison).

Part 2: experiments, results and analysis

Write the code to perform the following experiments:

- Performance when fine-tuning on the source language using monolingual embeddings: report results when varying the type of embeddings used (at least DistilBERT and BERT).
- Performance when fine-tuning on the source language using multilingual embeddings: report results using multilingual BERT → allow to evaluate the possible loss in performance when using multilingual vs monolingual embeddings.

- Performance on transfer learning: use the system fine-tuned on the source language with multilingual embeddings to make predictions on the target language → give performance one can expect when no target data is available.
- Performance when fine-tuning on the target language using multilingual embeddings: report results using subsets of varying sizes of the target language → give the performance one can expect if a ‘few’ examples are annotated for the target language. How many examples do you need to outperform the transfer learning approach?
- Optional: Look at the most important words for the prediction using interpretability tools.

Write the second part of the report that should contain:

- a description of your approach / methodology, the research questions you seek to answer
- a description of your system, the architecture, the hyper-parameters used
- a description and analysis of your results
- a conclusion proposing ways of improvement, and where you can discuss the difficulties encountered.

Project 2: classifying biographies under gender bias (2 students)

The goal of the project is to build a system for biographies / occupation classification while investigating gender bias for the task. Gender can be easily identified from biographies and it thus could bias our model. We will try to better understand its behavior through experiments and analysis.

The analysis of the results should investigate performance on the target task with original data, performance when information about gender are removed from the original text (debiasing), and the more informative lexical elements for the task, possibly informing new debiasing strategies.

Task: occupation classification

Setting: classification under bias

Datasets:

- Dataset: https://huggingface.co/datasets/LabHC/bias_in_bios

Part 1: study the data and read a research paper

Prepare the first part of your report that should contain:

- a precise description of the data, give the statistics and all relevant information.
- the description of a research paper on a similar task or a similar problem (bias in models in general)
 - Recommended reading: *Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting*. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)

Part 2: experiments, results and analysis

Write the code to perform the following experiments:

- Performance when fine-tuning on occupation classification using monolingual embeddings: report results when varying the type of embeddings used (at least DistilBERT and BERT).
- Performance when fine-tuning on occupation classification using monolingual embeddings when lexical indicators of gender are removed (you might propose different strategies) → allow to evaluate the possible loss in performance when 'debiasing' the data.
- Evaluating the bias in data: try to perform gender classification (with and without lexical gender indicators) and use an interpretability tool to reveal strong lexical cues for the task: are there still indicators of genre, indicators of other kinds of bias?

Write the second part of the report that should contain:

- a description of your approach / methodology, the research questions you seek to answer
- a description of your system, the architecture, the hyper-parameters used
- a description and analysis of your results
- a conclusion proposing ways of improvement, and where you can discuss the difficulties encountered.

Project 3: Cross-domain NER

The goal of the project is to build a system for Named Entity Recognition and to investigate transfer across domains. We will simulate this setting by using: a dataset in one domain considered as the **source**, and datasets in other domains corresponding to our **targets**. Our goal is to explore the transfer ability when varying the domains.

The analysis of the results should investigate performance when training and evaluating on the same domain (baseline system on the source domain), the performance drop when evaluating on a different domain (transfer learning), the correlation between domain similarity and performance (source selection / domain adaptation) and the improvement when adding a few target data at training time (semi-supervised setting).

Task: named entity recognition

Setting: domain adaptation

Datasets:

- CrossNER: https://huggingface.co/datasets/DFKI-SLT/cross_ner
- Source domain: Reuters
- Target domains: Politics, Natural Science, Music, Literature, and Artificial Intelligence
- (also a corresponding github: <https://github.com/zliucr/CrossNER>)

Part 1: study the data and read a research paper

Prepare the first part of your report that should contain:

- a precise description of the data, give the statistics and all relevant information. Be careful: the label sets could be different, you are allowed to map the labels to homogenize the corpora: discuss the possibilities for transfer experiments.
- the description of a research paper on a similar task and setting.
 - Recommended reading: Liu, Z., Xu, Y., Yu, T., Dai, W., Ji, Z., Cahyawijaya, S., Madotto, A., Fung, P. (2021, May). CrossNER: Evaluating cross-domain named entity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 35, No. 15, pp. 13452-13460).

Part 2: experiments, results and analysis

Write the code to perform the following experiments:

- Performance when fine-tuning on the source domain using contextual embeddings (BERT).
- Performance on transfer learning: use the system fine-tuned on the source domain with contextual embeddings to make predictions on each target domain → give performance one can expect when no target data is available. Does it work better for specific domains?

- Performance when fine-tuning on the source domain combined with target data (training set) → give the performance one can expect if a 'few' examples are annotated for the target language. Does it improve performance?
- Optional: Look at the most important words for the prediction using interpretability tools.

Write the second part of the report that should contain:

- a description of your approach / methodology, the research questions you seek to answer
- a description of your system, the architecture, the hyper-parameters used
- a description and analysis of your results
- a conclusion proposing ways of improvement, and where you can discuss the difficulties encountered.