Machine-in-the-Loop with Documentary and Descriptive Linguists

Sarah Moeller
University of Florida
smoeller@ufl.edu

Antti Arppe
University of Alberta
arppe@ualberta.ca

Abstract

This paper describes a curriculum for teaching linguists how to apply machine-in-the-loop (MitL) approach to documentary and descriptive tasks. It also shares observations about the learning participants, who are primarily non-computational linguists, and how they interact with the MitL approach. We found that they prefer cleaning over increasing the training data and then proceed to reanalyze their analytical decisions, before finally undertaking small actions that emphasize analytical strategies. Overall, participants display an understanding of the curriculum which covers fundamental concepts of machine learning and statistical modeling.

1 Introduction

This paper outlines the curriculum for "Machine-in-the-Loop for Language Documentation" workshops and shares from the experience we gained while teaching the material in different settings.[1] The workshop material gives instruction in foundational machine learning concepts, natural language processing (NLP) techniques, and statistical modeling. The full version of the material is designed to create an opportunity where documentary and descriptive linguists learn the concepts and apply NLP models to assist them in their documentary and descriptive tasks.

In hopes of increasing the effectiveness of instructors who teach NLP to linguists or who wish to improve inter-disciplinary collaboration and communication, this paper describes patterns that we observed while teaching this material in different settings. We aim the discussion towards instructors who come from a computational angle and will be teaching participants whose degrees or training is primarily in linguistics or related social sciences (We use the term "linguist" as a shorthand for any of these learners.) After teaching the material in whole or part, we noticed similar patterns of interaction with the NLP models emerged. These patterns make sense in light of typical linguistic training that promotes analytical skills and in-depth investigation of language data.

2 Motivation and Background

Unfortunately, current documentary and descriptive methods cannot scale up to match the pace of the language endangerment or provide automated methods to assist documentation (Seifart et al., 2018) because annotating large corpora of potential training data is still done mostly by hand (Duong, 2017; Palmer et al., 2010). This situation can be addressed by providing non-computational linguists with a basic understanding of statistical NLP so they are better equipped to integrate machine learning assistance into their work.

The potential for NLP to increase and improve documentary tasks has been clearly demonstrated (Felt, 2012; Moeller, 2021; Palmer, 2009; Xia et al., 2016). However, NLP research with under-documented languages often does not consider realistic factors sufficiently. For example, research tends to take a linear approach, where initial annotation of training data is the only input from human experts that the NLP system receives. In contrast, the workshop promotes human-in-the-loop approaches in realistic settings. The goal of human-in-the-loop techniques is to make optimal use and also reduce expensive human annotation while improving model performance (Bridgwater, 2016). We use the term "machine-in-the-loop (MitL)" to emphasize our conviction that technology's role is to assist humans (Zhang et al., 2022), not *vice versa*.

Although Lin et al. (2016) indicate that "reactive" learning that uses simple uncertainty sampling for denoising a corpus is not ideal, the workshop code implements this basic method, we choose this simple method because, unfortunately,

[1] https://github.com/sarahrmoeller/AI_Workshop

Proceedings of the The Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 27–32
March 21-22, 2024 ©2024 Association for Computational Linguistics

a short pedagogical event is not ideal for active learning experimentation or complex strategies. It may be worth noting that Lin et al. assume that the examples sampled by the algorithm as most "uncertain" are the only examples that will be relabeled, whereas we allow the linguists to choose what examples they will work with. Humans can analyze why a data point might have been selected as most uncertain by the algorithm and whether they agree that relabeling it for the next training cycle is likely to be impactful. Also, humans can generalize from the algorithm's simple calculations and then find and re-label multiple examples that they feel are similar in nature. These abilities may counteract some of the noted drawbacks of simple uncertainty sampling.

3 Overview of Curriculum and Workshops

The goal of the curriculum is to help linguists better understand concepts that will, hopefully, make them comfortable integrating NLP systems in their work. The material is aimed at linguists or others engaged in language-related work who do not have a background in advanced mathematics or computer science. The curriculum progresses through the learning objective listed below.

1. Know terminology related to artificial intelligence and NLP (e.g. What is NLP and where does it fit in the field of AI?)

2. Understand the foundational statistical concepts underpinning machine learning

3. Distinguish between classification (supervised) and clustering (unsupervised)

4. Understand the role of features and the importance of feature selection in classical machine learning, and the importance of data representation or selection in deep learning

5. Grasp the differences between, and the reasons for using, precision, recall, and F1 measure versus accuracy

6. Learn what a classification report and confusion matrix are and how to read them from a model trained for a task related to basic linguistic analysis

7. Interpret the model's predictions on previously unannotated data

8. Improve the model's output with any of three steps:
   - correcting noisy training data
   - increasing training data by creating new examples or by correcting the model's annotations on previously unannotated data
   - customizing the machine learning model architecture

The material uses lectures and activities to convey concepts and guide participants through a MitL approach to language documentation and description task. It includes Python code that we used to preprocess data and train the NLP models. Other instructors may find the code helpful but will probably wish to adjust it to suit their context. The activities are of two types. The first are short activities usually in the form of games or worksheets that are intended to break up lectures and reinforce concepts. These can be done in groups or individually. The final activity is essentially the same as learning objective 8. It assumes participants can be grouped into teams of 2-6 members. For the final activity, the one-day version of the curriculum uses the same data for all teams. Details about this data is provided in the released material. The three-day version assumes teams will be formed before the workshop begins and each team will work on their own field data during the final activity. With this in mind, team leaders are encouraged to invite members who are knowledgeable about the language and the task.

In a workshop setting with more than two or three teams, we found it necessary to recruit "tech coaches" for each team. These are participants who have skills to download and run the workshop code. Tech coaches do not need to run the code for preprocessing data if the instructor can do this more complicated task beforehand. The tech coaches do not necessarily need to be familiar with the language.

The material was developed from stand-alone lectures and assignments in university classrooms which were gradually combined and redesigned for non-credit workshop settings aimed at faculty, graduate students, or others working with endangered languages. The material released with this paper are designed for all-day workshops either one or three days in length. The three-day version of the workshop aims to coach participants to apply NLP systems to their own data. Below we described how we covered the material in a university classroom in some detail because we did release curriculum for this setting and then outline more briefly how a one-day and three-day workshop can be conducted.

University Classroom. The earliest version of the combined material was piloted in a cross-listed graduate/undergraduate language documentation course. The course used the Nyagbo language (ISO: nyb) as a case study and the instructors spent one week teaching Nyagbo morphology. In another week, the instructors briefly introduced NLP (objectives 1-4 and 7). At the end of the week, they showed output of a morphological parser (segmentation and glossing) that had been trained on available Nyagbo documentary field data. The students formed groups and were assigned to improve the parser with the first two steps described in objective 8 above. Students could also fulfill the assignment by writing an error analysis on the test data. Each group worked independently. They were allowed to decide what steps to take but were encouraged to assign a group member to each of these three tasks. After two weeks, students submitted their new training data containing all corrections to the original annotations and all additional annotations. The instructor trained and tested the same architecture on the new training sets. The test data was not changed. In class, the instructor explained classification metrics (objectives 5-6) and showed the classification reports from each team's new data next to the original classification report. The reports were discussed together (objective 7).

1-day Workshop. The one-day version covers all eight objectives but in less depth. The material is designed to cover objectives 1–4 in the first 3-4 hours through lectures and short activities. Then after introducing the final activity (English POS tagging with MitL approach) and allowing participants to form groups, Objectives 6-8 are are best covered by walking the groups through each step of a first iteration of the MitL activity. This will take about 1 hour. After that, groups should work independently through as many training+re-annotation iterations as they can in the remaining time. The final 45-60 minutes can be reserved for discussion and reflective learning.

3-day Workshop. The three-day version of the workshop is designed for teams to spend half of each day covering objectives 1-5 and the other half working through objectives 6-8 by applying a MiTL approach to their field data. Instructors may allow the teams to choose the NLP task for the MiTL activity, but it is recommended to limit the choice of tasks so that they suit the instructor's workload and are suitable to the participants' computing resources. We limited choices to POS tagging or morphological parsing.

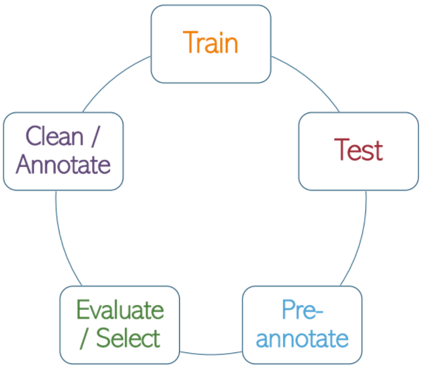

Figure 1: Machine-in-the-Loop approach to language documentation and description. A machine learning model is trained and tested for a NLP task on documentary data. The model is used to "pre-annotate" unannotated data. Linguists evaluate the pre-annotated data which is presented to them ranked according to some active learning selection strategy. The linguists then clean the initial training data or correct and add the pre-annotated data to the training data.

4 The Machine-in-the-Loop Approach

The workshop curriculum centers around a Machine-in-the-Loop (MitL) approach comprising active learning cycles of training and annotation. The goal is to guide humans to new annotation that will have most impact towards gradually improving NLP performance so that the improved model offers useful assistance to the humans. We implemented re-active labeling (Lin et al., 2016) using least confidence uncertainty sampling (Lewis and Catlett, 1994; Culotta and McCallum, 2005). Other sampling strategies can be used by instructors, but uncertainty sampling is relatively easy to explain.

As implemented for the workshop, least confidence uncertainty sampling ranks the model's predicted annotations by the differences between the model's confidences for each predicted sequence. When using Conditional Random Fields (CRF), the model's confidence C is defined as its calculated probability P that unit x should be annotated with a particular class $\hat{y}$. Confidence for a predicted sequence is calculated by normalizing each unit probability by the number of units in the sequence. For POS tagging, units are words and sequences are sentences; for morphological parsing, units are morphemes and/or morpheme glosses and sequences are characters in a word. The same calculation of confidence can be used with other models, such as the Transformer, but probability can be substituted for the absolute value of the models' negative log likelihood for a unit.

$$C = \frac{P(\hat{y}|x)}{L}$$

The sequences and the predicted annotations are ranked from lowest to highest confidence scores and written to a plain text file without the scores. Since participants naturally begin at the top of a file, this ranking nudges them to start correcting the new annotations with examples that the model found most confusing.

The goal of the MitL activity in the three-day workshops is to provide AI assistance for basic language documentation tasks, specifically annotation (interlinerization), and thereby increase the amount and the quality annotated data in endangered languages. Where possible, the activity should be prepared before the workshop begins. Teams can choose their task, and if ethically sound, share their field data with the instructor two to six weeks before the workshop begins. The instructor can run the preprocessing code to separate the annotated and unannotated units (units causing problem for the code are removed to the unannotated corpus) and divide the annotated corpus into a 9/1 training/test split. An initial model can then be trained, tested, and then used to annotate unlabeled data. Ideally, participants bringing field data should first create gold standard labels by correcting the portion of the data withheld to serve as the test set. If this is all done before the workshop begins, participants can start the MitL activity (objective 8) on the first day by examining the initial model's classification report, confusion matrix, and predicted annotations. Then they can attempt to clean manually labeled training data, correct the model's annotations, or customize the model architecture of feature selection. At the end of each day or whenever teams feel ready, new models can be trained on the version of the training data.

5 Observations

We taught the material in all settings described in section 3, primarily to participants with higher education degrees primarily in linguistics or related social science fields. In all three settings, similar patterns of interaction with the MitL approach emerged. We feel that a description of these patterns may assist instructors who come from a computational background. We have no data to confirm these patterns quantitatively, but they match our broader experience as computational and quantitative linguists teaching in linguistic departments.[2]

We observed that linguists' interaction with the MitL approach tend to reflect their analytical training rather than statistical or engineering solutions. Given free choice of the sub-steps under objective 8, the linguists followed three stages. First, they remove noise in the manually annotated data by correcting mistakes and inconsistencies. Second, they revise their previous analytical decisions and change manual annotations accordingly. Third, they take strategic actions aimed to improve the representation of the training data. The first two stages held true across all settings, but the third one was primarily observed in the in our one three-day workshop. We include it because, based on our broader teaching experience, it seems likely to hold true generally for linguists newly introduced to NLP and MitL concepts, and so may be helpful for other instructors to look for.

First, clean data. Field data is inherently noisy, either due to the dynamic nature of linguistic analysis, or as a by-product of manual work. Given a choice between the tasks under objective 8, linguists showed a preference to clean training data. This held true even though the instructors emphasized the impact of data size on statistical models. Interestingly, research (Chen et al., 2022; Lin et al., 2016) suggests, given limited time, cleaning rather than adding more data is a wise choice.

Cleaning meant correcting typos and making glosses consistent (e.g. '1.sg' and 'I' → 1.SG). We noted that teams tended not to clean all data at once, but asked to retrain the model 2-5 times and leaned on the model's output to find mistakes for the next round of corrections. It seems they found that although language data can be cleaned without NLP assistance, the MitL approach helped them

[2] One reviewer noted these observations match their experience as well.

more quickly identify issues.

Second, reanalyze. As obvious mistakes in the training data were corrected, we observed the linguists started to lean more on the MitL approach, but not in the way we expected. They looked more closely at the model's predictions for the unlabeled data. Noting that some of the model's "mistakes" were due to isomorphism or other ambiguity, they asked questions about outliers and weighting in statistical modeling. We assumed they would focus on adding new annotated sequences to the training data. Instead, they shifted their focus to the analytical decisions which had governed their original annotation. Several mentioned how the model highlighted a tension between lumping versus splitting choices they had felt while analyzing and annotating the original data.

The linguists began reanalyzing their previous annotation choices. They used the highest listed sequences (ones with lowest confidence scores) in the computer-annotated file to guide their reanalysis. This indicates that they had grasped the concept of least confidence sampling. It seems they hypothesized that if they changed their entire annotation schema, either by adding new classes or lumping others and then bulk-edited the training data, this would solve with the model's "confusions". After one or two rounds, reanalysis resulted improvement and sometimes decreased performance. This led to conversations about how an MitL approach encourages iterative work just solving the next low-hanging fruit may be more effective than linear work that provides a comprehensive analysis.

Finally, strategize additional annotation. Once the original training data had been cleaned and the teams finished their reanalysis we observed the linguists really began to integrate the principles of statistical modeling that were covered during the lectures. Their questions and discussions turned to strategies for increasing annotated training data as much as possible. A notable pattern at this stage was they either did not fully grasp the impact of data size or were daunted by the task of providing enough new annotated data. Instead of bulk-editing, they tended to correct the model's annotations in small strategic efforts that seemed guided as much by their knowledge of the language as by the uncertainty sampling strategy. Then they would request for the model to be retrained so they could see the effect. For example, they might correct the machine annotations only on the first six (least confident) sequences or they might search for sentences with a rare part of speech. Sometimes the latter approach improved the model's performance on one class.

One group who had a member with Python programming skills and worked with a CRF for POS-tagging decided to experiment with the model architecture which was optimized for English. For example, because their language is more morphologically complex than English, they programmed the model to use the first and last six letters of a word as features, instead of the first and last four letters. This was the only example of participants attempting to customize the model, and they also seemed to prefer leveraging strategic knowledge of the language, rather than a statistical strategy (i.e. lots of annotation).

What's next? Participants progressed through these three stages independently. In general, they demonstrated understanding how noise and ambiguity present significant issues for statistical models with limited data. In the closing discussions, a repeated theme was the amount of annotation needed in order to make a significant impact would exceed their time limitations. This presented a chance to introduce "engineering" solutions not covered in the workshop materials, such as data hallucination and synthetic augmentation or cross-lingual transfer learning.

6 Conclusion

While teaching this material we observed that documentary and descriptive linguists bring their analytical strengths to the MitL approach. The material does not include formal assessments but we consider it successful because participants who brought their own data left with a better quality documentary corpus. A recent grant submitted by a workshop attendee to fund this workshop in another location indicates that the participants also assessed the material positively.

We conclude with specific recommenations. First, despite many successful remote collaborations in the post-pandemic age, we found that in-person events, removed from regular work, promote focused work and encourages social interaction that counteracts the intense schedule. Second, we recommend adding one day to the longer workshop schedule just to deal with anticipated and unanticipated issues (remote server not set up in time, flight delays, code bugs, forming teams, etc.).

Acknowledgements

The curriculum development was supported by the University of Florida and the “21st Century Tools for Indigenous Languages” Partnership (funded by grant #895-2019-1012 from the Social Sciences and Humanities Research Council of Canada [SSHRC] and the University of Alberta). We are very grateful to workshop participants and students who gave feedback on the curriculum and whose abilities and work exceeded our expectations!

References

Adrian Bridgwater. 2016. Machine Learning Needs A Human-In-The-Loop. *Forbes*.

Derek Chen, Samuel R. Bowman, and Zhou Yu. 2022. Clean or Annotate: How to Spend a Limited Data Collection Budget. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 152–168, Hybrid. Association for Computational Linguistics.

Aron Culotta and Andrew McCallum. 2005. Reducing Labeling Effort for Structured Prediction Tasks:. In *Proceedings, The Twentieth National Conference on Artificial Intelligence and the Seventeenth Innovative Applications of Artificial Intelligence Conference*, Pittsburgh, Pennsylvania, USA. American Association for Artificial Intelligence.

Long Duong. 2017. *Natural language processing for resource-poor languages*. Phd thesis, University of Melbourne, Melbourne, Australia.

Paul Felt. 2012. *Improving the Effectiveness of Machine-Assisted Annotation*. Ma thesis, Brigham Young University.

David D. Lewis and Jason Catlett. 1994. Heterogeneous Uncertainty Sampling for Supervised Learning. In William W. Cohen and Haym Hirsh, editors, *Machine Learning Proceedings 1994*, pages 148–156. Morgan Kaufmann, San Francisco (CA). Cite for uncertainty sampling.

Christopher Lin, M Mausam, and Daniel Weld. 2016. Re-Active Learning: Active Learning with Relabeling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1).

Sarah Moeller. 2021. *Integrating Machine Learning into Language Documentation and Description*. Ph.d., University of Colorado at Boulder, United States – Colorado.

Alexis Palmer, Taesun Moon, Jason Baldridge, Katrin Erk, Eric Campbell, and Telma Can. 2010. Computational strategies for reducing annotation effort in language documentation. *Linguistic Issues in Language Technology*, 3(4):1–42.

Alexis Mary Palmer. 2009. *Semi-automated annotation and active learning for language documentation*. Phd thesis, University of Texas at Austin.

Frank Seifart, Nicholas Evans, Harald Hammarström, and Stephen C. Levinson. 2018. Language documentation twenty-five years on. *Language*, 94(4):e324–e345.

Fei Xia, William D. Lewis, Michael Wayne Goodman, Glenn Slayden, Ryan Georgi, Joshua Crowgey, and Emily M. Bender. 2016. Enriching a massively multilingual database of interlinear glossed text. *Language Resources and Evaluation*, 50(2):321–349.

Shiyue Zhang, Ben Frey, and Mohit Bansal. 2022. How can NLP Help Revitalize Endangered Languages? A Case Study and Roadmap for the Cherokee Language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1529–1541, Dublin, Ireland. Association for Computational Linguistics.