Data Cleaning Tools for Token Classification Tasks

Karthik Muthuraman[2], **Frederick Reiss**[1,2], **Hong Xu**[2],
Bryan Cutler[2] and **Zachary Eichenberger**[1,3]
[1]IBM Research – Almaden, San Jose, CA 95120, USA
[2]IBM Center for Open Source Data and AI Technologies (CODAIT),
San Francisco, CA 94105, USA
[3]University of Michigan, Ann Arbor, MI 48109, USA
karthik.muthuraman@ibm.com, frreiss@us.ibm.com, hongx@ibm.com
bjcutler@us.ibm.com, zachary.eichen@gmail.com

Abstract

Human-in-the-loop systems for cleaning NLP training data rely on automated sieves to isolate potentially-incorrect labels for manual review. We have developed a novel technique for flagging potentially-incorrect labels with high sensitivity in named entity recognition corpora. We incorporated our sieve into an end-to-end system for cleaning NLP corpora, implemented as a modular collection of Jupyter notebooks built on extensions to the Pandas DataFrame library. We used this system to identify incorrect labels in the CoNLL-2003 corpus for English-language named entity recognition (NER), one of the most influential corpora for NER model research.

Unlike previous work that only looked at a subset of the corpus's validation fold, our automated sieve enabled us to examine the entire corpus in depth. Across the entire CoNLL-2003 corpus, we identified over 1300 incorrect labels (out of 35089 in the corpus).

We have published our corrections, along with the code we used in our experiments. We are developing a repeatable version of the process we used on the CoNLL-2003 corpus as an open-source library.

1 Introduction

Human-in-the-loop systems for cleaning NLP training data rely on automated sieves to isolate potentially-incorrect labels for manual review. In this work, a full version of which has been presented in (Reiss et al., 2020), we describe how we developed a novel technique for flagging potentially-incorrect labels with high sensitivity in named entity recognition corpora.

We implemented our sieve in the context of a set of extensions to the *Pandas*[1] DataFrame library. In addition to flagging errors, our extensions provide facilities for comparing NLP model results and visualizing model outputs and training data in context.

Because we built these facilities into the primary DataFrame library of the Python data analysis stack, we were able to construct an end-to-end system for NLP data cleaning as a series of Jupyter[2] notebooks. This design gives sophisticated users a view of the internals of the data cleaning process and allows for easy customization.

Our Jupyter notebooks comprises a pipeline that starts with training ensembles of models. Next, the system analyzes the outputs of the ensembles to identify potentially incorrect labels. Additional notebooks provide human annotators with a view of the suspicious labels in context. Later stages of the pipeline merge and analyze the results of manual annotation; then construct a corrected dataset and reports on the nature of the corrections.

We used this system to identify errors in the CoNLL-2003 NER corpus. The English-language portion of the CoNLL-2003 shared task (Tjong Kim Sang and De Meulder, 2003) (henceforth CoNLL-2003) is one of the most widely-used benchmarks for named entity recognition (NER) models. It consists of news articles from the Reuters RCV1 corpus (Lewis et al., 2004). Since its debut, CoNLL-2003 has played a central role in NLP research and continues to do so with more than 2300 citations. While researchers have relied heavily on the CoNLL-2003 corpus as a source of ground truth, few have paid attention to the corpus itself. Errors in the corpus could potentially mislead and even divert the course of future research.

Unlike previous analyses of this dataset that only examined small fractions of the CoNLL-2003 corpus, our work leveraged a high level of automation to analyze the entire corpus. We found over 1300 errors.

[1] https://pandas.pydata.org/

[2] https://jupyter.org

Proceedings of the 2nd Workshop on Data Science with Human in the Loop: Language Advances, pages 59–61
June 11, 2021. ©2021 Association for Computational Linguistics

2 Process

Our approach builds on previous work in semi-supervised labeling, with some key differences. Because we were looking for errors in a corpus that already had many high-quality labels, we needed a sieve with especially high sensitivity. We used ensembles of NER models trained on the corpus, and we focused on cases where the models agreed strongly on a particular label, but that label does not appear in the corpus. One of these ensembles was the outputs of the original 16 entries in the 2003 competition. We also trained two other 17-model ensembles ourselves by applying Gaussian random projections to the BERT embeddings space.

We developed extensions to the *Pandas* DataFrame library that enabled us to represent spans within documents as cells within a DataFrame. This facility allowed us to use DataFrames to track the spans of the entities that each of our models produced and to aggregate together the results across models. Using these capabilities, we developed Jupyter notebooks that analyzed our ensembles' outputs to identify labels that appeared in the outputs of multiple models but were not in the corpus.

We used our Pandas extension types' ability to render spans to HTML to view these spans in the context of the original document from within the same Jupyter notebooks.We started with labels that had a strong agreement among models and we progressed to labels with less agreement among models, the fraction of flagged labels that was actually incorrect decreased. When this fraction dropped below 20 percent, we stopped going through the ordered list of flagged labels. We had an inter annotation agreement and audit cycle for each correction made. In total, we made 12 passes (3 ensembles × 2 sets of labels × 2 human reviewers) of manual review over the `train` and `test` folds of the corpus and 8 passes over the `test` fold.

When we found that a label was incorrect, we coded the type of error and the required correction so that the error could be corrected automatically later on. We divided errors into several categories as explained in detail in the full version of this paper at (Reiss et al., 2020).

3 Corrections

In total, we examined 3182 labels our ensembles had flagged in the three folds of the corpus. We considered any label where fewer than 7 models agreed with the corpus label to be "flagged". Of these labels, 1274 came from the `test` fold, 854 came from the `dev` fold, and 1054 came from the `train` fold; accounting for 22.6%, 14.3%, and 4.5% of their folds, respectively. Figure 1 shows the split of final errors identified by ensemble and source.

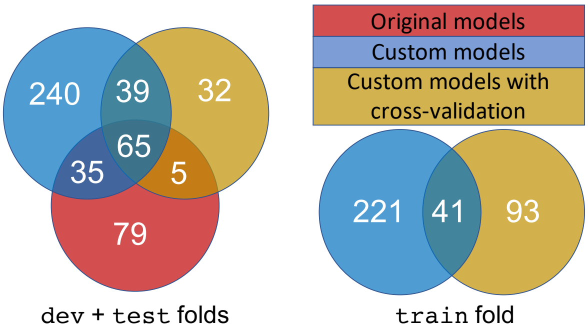

Figure 1: Number of errors flagged by different combinations of ensembles after filtering by human labelers.

Manual inspection determined that 850 of these 3182 entities (27%) were incorrect. We also found 475 additional incorrect entities in close proximity to the entities that our techniques flagged, for a total of 1320 incorrect labels across the corpus.

After identifying incorrect tags, spans and sentence boundaries, we created a corrected version of the original CoNLL-2003 dataset, which we refer to as the corrected CoNLL-2003 dataset.

4 Ongoing Work

While preparing our dataset of corrections for release, we identified additional improvements to the corrections. We have released a second version of the dataset containing these improvements plus some additional corrections pointed out by members of the open source NLP community.

We have released the code that we used in our experiments so far[3]. To facilitate the reuse of this code on other datasets, we are developing a more refined version of this code. Key changes that we are working on are reducing the number of passes of manual review required, simplifying the creation of ensembles of models, and extending the approach from NER to other token classification tasks like semantic role labeling. We plan to release these improvements.

[3] https://github.com/CODAIT/text-extensions-for-pandas

References

David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. 2004. Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397.

Frederick Reiss, Hong Xu, Bryan Cutler, Karthik Muthuraman, and Zachary Eichenberger. 2020. Identifying incorrect labels in the CoNLL-2003 corpus. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 215–226. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the SIGNLL Conference on Computational Natural Language Learning*, pages 142–147, USA. Association for Computational Linguistics.