Supporting Dynamic Construction of Datasets with Annotator Suggestions

Jeongeon Park[1], Eunyoung Ko[1], Donghoon Han[2], Jinyeong Yim[3], Juho Kim[1]
[1] School of Computing, KAIST
[2] Superb AI
[3] Clova AI, NAVER Corporation
{jeongeon.park, eunyeongko, juhokim}@kaist.ac.kr, dhhan@superb-ai.com, jinyeong.yim@navercorp.com

Abstract

The construction process for machine learning datasets is costly for experts as it requires going through multiple iterations to build a label set and communicating and resolving issues with annotators. To make the process more efficient for experts, we propose DynamicLabels, which allows experts to refine the dataset structure with label set suggestions collected by the annotators.

Introduction

Constructing a dataset is one of the most important steps in building a machine learning model. While there are publicly accessible datasets (Deng et al. 2009; Maas et al. 2011) available for use, a lot of times model experts need to manually construct datasets from scratch for custom applications or improved performance in a particular domain.

While for some tasks the label set—a set of labels to annotate the data instances—can be quickly built with existing datasets or theories, for others building a label set requires experts to have a global understanding of the dataset. To construct a label set that is of high coverage and easily comprehensible for both experts and annotators, the expert has to go through many instances of the dataset and take multiple iterations prior to annotation. Even with multiple iterations to refine the label set, communicating with annotators on unclear labels or edge cases requires extra time of the experts.

There have been attempts to reduce the expert cost in constructing label sets, such as Cascade (Chilton et al. 2013) and Deluge (Bragg, Weld et al. 2013), which generate a taxonomy without expert involvement. However, these only target the label set building process, when more issues can be raised during the annotation phase. Other work tries to refine label sets during the annotation process through crowd-generated labels or structured labeling (Chang, Amershi, and Kamar 2017; Kulesza et al. 2014), but they mainly target binary classification.

To reduce the burden of experts, we propose DynamicLabels, a workflow that allows experts to dynamically improve the dataset with annotator suggestions. DynamicLabels consists of three stages, (1) initial label set construction (by expert), (2) annotation with suggestions (by annotators), and (3) suggestion review and dataset finalization through a dashboard (by expert).

We present a case study to build post-OCR parsing datasets to compare DynamicLabels with the baseline workflow. Results show that DynamicLabels enabled a more complete, a detailed label set construction with a similar time compared to baseline. Furthermore, annotator suggestions were approved as new labels and helped experts better understand communication issues and consider them in decision making.

Proposed Workflow: DynamicLabels

DynamicLabels is a workflow where the label set and corresponding annotations can be refined after annotation with annotator suggestions. The overall pipeline of DynamicLabels is shown in Figure 1.

Stage 1: Initial Label Set Construction The expert constructs a label set after looking at a small portion of the dataset, while this likely results in a low-coverage label set.

Stage 2: Annotation with Suggestions The annotators are asked to annotate the raw data using the label set from Stage 1. When the annotator labels the raw data, instead of labeling with the existing label set, they can optionally provide suggestions for improving the label set.

We provide two simple ways to provide label suggestions: the **close to button** and the **N/A button**. The annotator clicks the *close to button* when the label fits the existing label set, but does not perfectly describe the data to be annotated. It indicates a possible extension of the current label set (e.g., Suggesting the label *'menu - size'* for receipts when *'menu - name'* can be used). The *N/A button* is when a data instance cannot be annotated with the given label set, which can help detect edge cases the expert did not spot in Stage 1. For both suggestion types, the annotator is asked to (1) input a better label for the selected data instance by either typing in a new label or selecting an existing label that other annotators have suggested, and (2) provide a reason for the suggestion.

After completing the assigned annotations, the annotator is asked to review their suggestions. For each suggestion they provided, they need to identify if other workers' suggestions are similar to theirs. The system uses this information to group the similar suggestions and minimize the burden of experts in reviewing and resolving the suggestions.

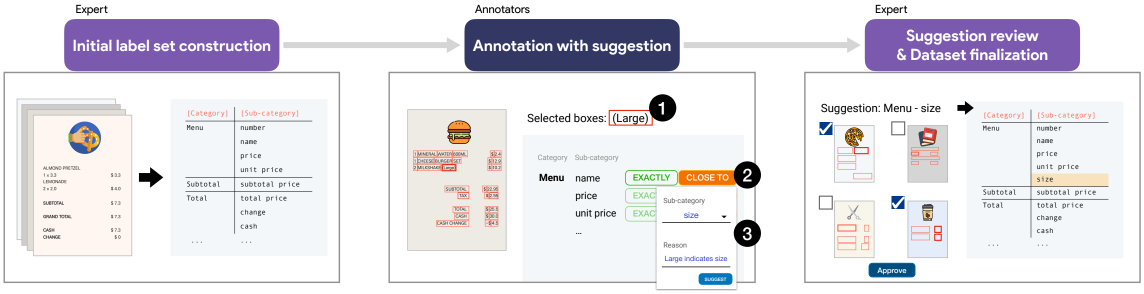

Figure 1: Three stages of DynamicLabels: (1) Initial label set construction stage (by expert), (2) Annotation with suggestion stage (by annotators), and (3) Suggestion review and dataset finalization stage (by expert). The figure illustrates how the label *'menu - size'* could be added with annotator suggestion.

Stage 3: Suggestion Review and Dataset Finalization After annotation, the expert can resolve suggestions and finalize the dataset through a **suggestion review dashboard** that presents the collected annotation and the suggestions. In the dashboard, the expert can either (1) add annotator-suggested labels to the label set, or (2) rename, merge, or change the structure of the label set if it is in a hierarchy.

To resolve grouped suggestions and corresponding annotations, the expert can (1) *approve* as the suggested label, (2) *add as new* label by renaming the suggestion, (3) *add to existing* labels, or (4) *ignore* the suggested label. Additional changes such as renaming, merging, or changing the structure can be performed directly on the label set. The changes in the label set and the corresponding annotations are reflected instantaneously as reviews take place, to help check the current status of the dataset. Annotations not marked as suggestions are merged with majority voting.

Case Study: Post-OCR Parsing

Study Design We recruited four experts to build datasets for a post-OCR parsing model with either DynamicLabels or the baseline workflow, using either receipt or event flyer documents. Each document consisted of 800 images [1].

For the baseline workflow, the expert first constructed a label set by reviewing 500 images. Then, annotators used the label set to annotate and raise issues when they detect edge cases. After annotation, the expert reviewed issues from annotators and finalized the dataset.

For DynamicLabels, the expert constructed an initial label set with 100 images, which is passed on to Stage 2. Then, the expert reviewed the grouped suggestions and refined the dataset structure using the review dashboard.

For each condition, we recruited 200 crowd workers from Amazon Mechanical Turk [2] as annotators. We assigned five workers to a single image. Each worker annotated 20 images and was paid approximately $8.0 for one hour of work.

Study Results The results are shown in Table 1 The experts in DynamicLabels had more labels in their final label set than experts in the baseline workflow. The experts in DynamicLabels were more open to add and revise the labels based on annotators' suggestions while experts in the baseline added or revised labels only for the content that they couldn't see in Stage 1. For example, with DynamicLabels the expert created labels such as *'menu - code'* or *'payment - gopay'* (Receipt dataset), which was not present in the baseline label set. The constructed dataset had a similar accuracy for both document types (Receipt: 96.2% and 95.7%, event flyer: 85.6% and 88.4%, baseline and DynamicLabels respectively).

Table 1: Results on how many issues/suggestions were raised by the annotators and how experts incorporated them into refining the label set.

	Receipt		Event flyer	
	Baseline	Proposed	Baseline	Proposed
# Annotations with issue/suggestion	218	478	229	516
# Suggestions (Grouped suggestions)	-	311 (233)	-	160 (96)
Initial label set	25	32	39	37
Final label set	26	37	39	45
# change in label set	+1	+5	0	+8

We also identified three patterns in how experts utilized the grouped suggestions provided by DynamicLabels.

Reduced the overall communication cost. In the baseline workflow, there were annotations with issues where experts had difficulties understanding why they were raised. In DynamicLabels, annotators' suggestions with matching annotations helped understand why particular suggestions were provided, which aided the annotator-expert communication.

Provided an overview of the dataset. As mentioned in Cascade (Chilton et al. 2013), having a global view of the entire dataset helps experts in label set construction. Experts who used DynamicLabels indicated the grouped annotations worked as supporting evidence for deciding whether to add labels. For event flyers, the label *'participating entity'* was added after the expert checked multiple suggestions with people's name as the annotation.

Helped renaming the labels. While the suggestions were used to improve the dataset structure, experts also used them to come up with a more general or inclusive label name.

[1] We used https://github.com/clovaai/cord for receipts and manually collected event flyers.

[2] https://www.mturk.com/

References

Bragg, J.; Weld, D. S.; et al. 2013. Crowdsourcing multi-label classification for taxonomy creation. In *First AAAI conference on human computation and crowdsourcing*.

Chang, J. C.; Amershi, S.; and Kamar, E. 2017. *Revolt: Collaborative Crowdsourcing for Labeling Machine Learning Datasets*, 2334–2346. New York, NY, USA: Association for Computing Machinery. ISBN 9781450346559. URL https://doi.org/10.1145/3025453.3026044.

Chilton, L. B.; Little, G.; Edge, D.; Weld, D. S.; and Landay, J. A. 2013. *Cascade: Crowdsourcing Taxonomy Creation*, 1999–2008. New York, NY, USA: Association for Computing Machinery. ISBN 9781450318990. URL https://doi.org/10.1145/2470654.2466265.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Kulesza, T.; Amershi, S.; Caruana, R.; Fisher, D.; and Charles, D. 2014. Structured Labeling for Facilitating Concept Evolution in Machine Learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, 3075–3084. New York, NY, USA: Association for Computing Machinery. ISBN 9781450324731. doi:10.1145/2556288.2557238. URL https://doi.org/10.1145/2556288.2557238.

Maas, A. L.; Daly, R. E.; Pham, P. T.; Huang, D.; Ng, A. Y.; and Potts, C. 2011. Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 142–150. Portland, Oregon, USA: Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P11-1015.