Bridging Multi-disciplinary Collaboration Challenges in ML Development Workflow via Domain Knowledge Elicitation

Soya Park
MIT CSAIL
soya@mit.edu

Abstract

Building a machine learning model in a sophisticated domain is a time-consuming process, partially due to the steep learning curve of domain knowledge for data scientists. We introduce Ziva, an interface for supporting domain knowledge from domain experts to data scientists in two ways: (1) a `concept creation` interface where domain experts extract important concept of the domain and (2) five kinds of `justification elicitation` interfaces that solicit elicitation how the domain concept are expressed in data instances.

1 Introduction

In recent decades, machine learning (ML) technologies have been sought out by an increasing number of professionals to automate their work tasks or augment their decision-making (Yang et al., 2019). Broad areas of applications are benefiting from integration of ML, such as healthcare (Cai et al., 2019a,b), finance (Culkin and Das, 2017), employment (Manyika et al., 2017), and so on. However, building an ML model in a specialized domain is still expensive and time-consuming for at least two reasons. First, a common bottleneck in developing modern ML technologies is the requirement of a large quantity of labeled data. Second, many steps in an ML development pipeline, from problem definition to feature engineering to model debugging, necessitate an understanding of domain-specific knowledge and requirements (Piorkowski et al., 2021). Data scientists therefore often require input from domain experts to obtain labeled data, to understand model requirements, to inspire feature engineering, and to get feedback on model behavior. In practice, such knowledge transfer between domain experts and data scientists is very much ad-hoc, with few standardized practices or proven effective approaches, and requires significant direct interaction between data scientists and domain experts. Building a high-quality legal, medical, or financial model will inevitably require a data scientist to consult with professionals in such domains. In practice, these are often costly and frustrating iterative conversations and labeling exercises that can go on for weeks and months, which usually still do not yield output in a form readily consumable by a model development pipeline.

In this work, we set out to develop methods and interfaces that facilitate knowledge sharing from domain experts to data scientists for model development. We developed a domain-knowledge acquisition interface **Ziva** (With **Z**ero knowledge, How do **I** de**V**elop **A** machine learning model?). Instead of a data-labeling tool, Ziva intends to provide a diverse set of elicitation methods to gather knowledge from domain experts, then present the results as a repository to data scientists to serve their domain understanding needs and to build ML models for specialized domains. Ziva scaffolds the knowledge sharing in desired formats and allows asynchronous exchange between domain experts and data scientists. It also allows flexible re-use of the knowledge repository for different modeling tasks in the domain.

Specifically, Ziva focuses on eliciting key concepts in the text data of a domain (`concept creation`), and rationale justifying a label that a domain expert gives to a representative data instance (`justification elicitation`). In the current version of Ziva, we provide five different `justification elicitation` methods – bag of words, simplification, perturbation, concept bag of words, and concept annotation.

2 Ziva System

2.1 Concept creation

Creating a taxonomy is an effective way of organizing information (Laniado et al., 2007; Chilton et al., 2013). Ziva provides an interface where SMEs can

Proceedings of the 2nd Workshop on Data Science with Human in the Loop: Language Advances, pages 44–46

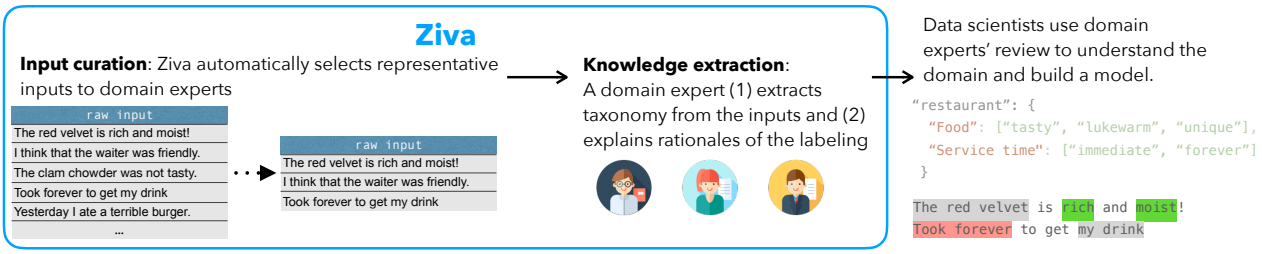

Figure 1: To facilitate domain knowledge sharing, Ziva presents representative instances and to interfaces to review the instances to domain experts, then which will be used by data scientists.

extract domain concepts. Users are asked to categorize each example instance, presented as a card, via a card-sorting activity. Users first group cards by topic (general concepts of the domain such as atmosphere, food, service, price). Cards in each topic are then further divided cards into descriptions referencing specific attributes for a topic (e.g., cool, tasty, kind, high).

2.2 Justification-elicitation interface

Once a domain expert finishes the concept extraction, they review each instance using one of elicitation interfaces, which ask the domain expert to justify an instance's label (this information is then intended for consumption by data scientists).

The `justification elicitation` interfaces were designed through an iterative process of paper prototyping, starting with initial designs inspired by our preliminary interviews. As we conducted paper prototyping, we examined if (1) the answers from different participants were consistent and (2) the information from participants' answers were useful to data scientists.

Bag of words. This base condition reflects the most common current approach. Given an instance and a label (e.g., positive, negative), the domain experts are asked to highlight the text snippets that justify the label assignment.

Instance perturbation. Inspired by one of our data scientists in the formative study, this condition asks a domain expert to *perturb* (edit) the instance such that the assigned label is no longer justifiable by the resulting text. For example, in the restaurant domain, *"our server was kind"*, can be modified to no longer convey a positive sentiment by either negating an aspect (e.g., *"our server was not kind"*) or altering it (e.g., *"our server was rude"*).

Instance simplification. This condition asks domain experts to shorten an instance as much as possible, leaving only text that justifies the assigned label of the original instance. For example, *"That's right. The red velvet cake... ohhhh.. it was rich and moist"*, can be simplified to *"The cake was rich and moist"*, as the rest of the content does not convey any sentiment, and can therefore be judged irrelevant to the sentiment analysis task.

Concept bag of words. This condition incorporates the concept extracted in the prior step. Similar to the Bag of words condition, domain experts are asked to highlight relevant text within each instance to justify the assigned label; however, each highlight must be grouped into one of the concepts. If, during `Concept creation`, the domain expert copied a card to assign multiple topics and descriptions, then the interface prompts multiple times to highlight relevant text for each one. For example, if they classified the instance, *"That's right. The red velvet cake... ohhhh.. it was rich and moist"*, into the concept *"food is tasty"*, they can select *rich*, *moist* and *cake* as being indicative words for that concept.

Concept annotation. This condition is similar to the above Concept bag of words condition. However, when annotating the instance text, domain experts are directed to distinguish between words relevant to the topic and words relevant to the description. Given the above sample instance, the domain expert would need to indicate which part of the sentence applies to *food* (e.g., *cake*) and which to *tasty* (e.g., *rich and moist*). Both this and the previous concept condition are motivated by the well-established knowledge that a variety of NLP tasks, such as relation extraction, question answering, clustering and text generation can benefit from tapping into the the conceptual relationship present in the hierarchies of human knowledge (Zhang et al., 2016). Learning taxonomies from text corpora is a significant NLP research direction, especially for long-tailed and domain-specific knowledge acquisition (Wang et al., 2017).

Details of the interface design and the evaluation can be found in Park et al. (2021).

References

Carrie Jun Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019a. "hello ai": Uncovering the onboarding needs of medical practitioners for human-ai collaborative decision-making.

Carrie Jun Cai et al. 2019b. Human-centered tools for coping with imperfect algorithms during medical decision-making.

Lydia B Chilton, Greg Little, Darren Edge, Daniel S Weld, and James A Landay. 2013. Cascade: Crowdsourcing taxonomy creation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1999–2008.

Robert Culkin and Sanjiv R Das. 2017. Machine learning in finance: The case of deep learning for option pricing. *Journal of Investment Management*, 15(4):92–100.

David Laniado, Davide Eynard, Marco Colombetti, et al. 2007. Using wordnet to turn a folksonomy into a hierarchy of concepts. In *Semantic Web Application and Perspectives-Fourth Italian Semantic Web Workshop*, pages 192–201.

James Manyika, Michael Chui, Mehdi Miremadi, et al. 2017. A future that works: Ai, automation, employment, and productivity. *McKinsey Global Institute Research, Tech. Rep*, 60.

Soya Park, April Yi Wang, Ban Kawas, Q Vera Liao, David Piorkowski, and Marina Danilevsky. 2021. Facilitating knowledge sharing from domain experts to data scientists for building nlp models. In *26th International Conference on Intelligent User Interfaces*, pages 585–596.

David Piorkowski, Soya Park, April Yi Wang, Dakuo Wang, Michael Muller, and Felix Portnoy. 2021. How ai developers overcome communication challenges in a multidisciplinary team: A case study.

Chengyu Wang, Xiaofeng He, and Aoying Zhou. 2017. A short survey on taxonomy learning from text corpora: Issues, resources and recent advances. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1190–1203. Association for Computational Linguistics.

Qian Yang, Aaron Steinfeld, and John Zimmerman. 2019. Unremarkable ai: Fitting intelligent decision support into critical, clinical decision-making processes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–11.

Hao Zhang et al. 2016. Learning concept taxonomies from multi-modal data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1791–1801. Association for Computational Linguistics.