Beyond human-in-the-loop: scaling occupation taxonomy at Indeed

Suyi Tu, Olivia Cannon

Abstract

A hierarchical occupation taxonomy helps Indeed better match job seekers and jobs. Historically, scaling this hierarchical taxonomy in international markets was a time intensive and highly manual process. Leveraging the strengths of both machine learning models and subject matter experts led to the creation of an improved human-in-the-loop system that met the needs of a growing business without sacrificing quality. This paper describes this system and discusses the challenges and insights from its implementation. Specifically, it highlights the value of involving subject matter expertise beyond just labeling and therefore offers the application of an expert-in-the-loop framework for scaling taxonomy.

Keywords

Human-in-the-loop, Expert-in-the-loop, taxonomy, subject matter expert, natural language processing, scaling

1. Introduction

As the #1 job site in the world, Indeed hosts a massive volume of unstructured data in the form of millions of job descriptions and resumes. Classifying a job or work experience as an occupation is one way of extracting structured data from these documents. This improves matching between jobs and job seekers and enables personalization.

The Indeed Taxonomy team has spent years thoroughly researching and developing a hierarchical occupation taxonomy and a hand-curated rule system to classify jobs for core strategic markets. The occupation taxonomy is granular, precise, and customizable by market. Consequently, it required significant time and resource investment for international scaling and ongoing maintenance. As demand for Indeed's presence to expand into more international markets rapidly grew, the team needed new ways to scale this work faster, with fewer resources, and without sacrificing data quality.

This paper reflects on the challenges and successes encountered while introducing a human-in-the-loop (HITL) approach to scaling Indeed's occupation taxonomy development and classification to support its rapidly expanding international business strategy. In particular, it discusses a core learning: that targeted combination of domain expertise and broad application of machine learning technology has been the key to success. The scope of this combination exceeds the level of human-augmentation generally implied by HITL, and we therefore describe it more precisely as *expert-in-the-loop* (EITL).

[Hierarchy visualization of part of food & beverage occupation taxonomy]

Figure 1: Example of an occupation taxonomy branch

2. Motivation

The occupation taxonomy at Indeed is a hierarchy that leverages broader and narrower concept relationships to enrich understanding of occupation types. For example, *Bartenders* is narrower than its broader taxonomy concept, *Food & Beverage Servers*, and *Food & Beverage Servers* is narrower than its broader concept, *Food & Beverage Occupations*. It is therefore understood that *Bartenders* is a type of *Food & Beverage Occupation*. This broadest grouping is referred to as a Top Level concept. Given a job or resume document, our classification system assigns the most specific occupation concepts possible.

Indeed is a global company with a presence in over 60 countries. Occupation taxonomy design in different markets may vary due to factors such as employment landscape, economy, and the volume of jobs hosted on Indeed in that location. However, the Indeed Taxonomy team has found that occupation taxonomy design is typically more similar than it is different across markets. Compared with occupation taxonomy concepts that exist in the US, the most similar International market with complete taxonomy coverage has over 90% of occupation concepts in common, while even the least similar International market has roughly 50% in common. Having a global occupation taxonomy that demonstrates both commonality and distinction allows for sharing knowledge

RecSys in HR'22: The 2nd Workshop on Recommender Systems for Human Resources, in conjunction with the 16th ACM Conference on Recommender Systems, September 18–23, 2022, Seattle, USA.

suyitu@indeed.com (S. Tu); ocannon@indeed.com (O. Cannon)

among markets, designing market-specific features, and avoiding excessive granularity where it is not relevant or required.

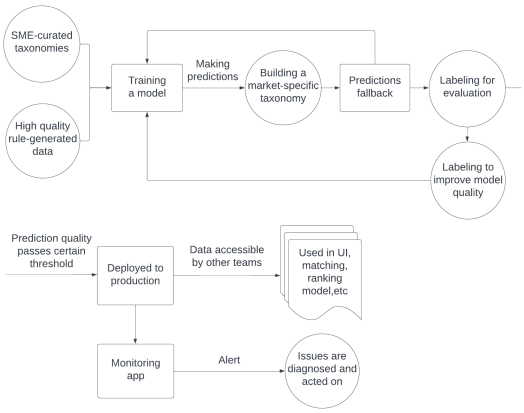
The Taxonomy team at Indeed is led by taxonomists who are highly skilled professional researchers and possess graduate degrees in Information Science, Philosophy, Linguistics, or related disciplines. Teams of Taxonomy Analysts are hired specifically to support strategically significant markets, and these individuals typically possess work experience in data analysis or professional research. They must have strong expertise in a market's employment landscape as well as spoken and written fluency in the official languages. They either possess a formal education in Information Science with hands-on experience with Taxonomy, or receive extensive on-the-job training in Information Science concepts and Taxonomy best practices. Their work is informed by consulting relevant quantitative and qualitative data, competitive analyses, and applying many of the same user-centered design principles seen in UX Research or Content Design.

Indeed's occupation taxonomy expansion for the first few international markets was based on a rule-populated manual approach and took years to achieve minimal viable state in each new market. As international business exponentially expanded, the human-intensive process of crafting separate rule systems for each new market and keeping them up-to-date became insurmountable.

There is plenty of research on applications of machine learning techniques to large-scale taxonomy development and document classification in diverse fields with minimal human intervention [1, 2, 3, 4]. However, in order to enable a consistent and high-quality user experience around the globe, it is important to identify shared occupations between markets and develop new components of the taxonomy to reflect market-specific occupation types or relationships, which involves research from domain experts. In addition, the system in production requires close monitoring to ensure data quality and freshness [5]. These challenges made adopting a purely automated system less desirable. There is also emerging research on the value of incorporating domain expertise in facilitating effective automation in highly technical fields [6, 7, 8]. We propose that occupation taxonomy development and design similarly benefits from this expert augmented approach that leverages the strength of both machine learning models and subject matter experts (SMEs). In the following sections, "experts" in our EITL system are referred to as "SMEs".

3. Expert-in-the-loop Taxonomy Scaling

We represent a simplified overview of our EITL system in Figure 2. Note that the "model" specified in the figure



[End-to-end workflow of the expert-in-the-loop system]

Figure 2: High-level overview of the expert-in-the-loop system. Circled steps are heavily SME involved; Squared steps are fully automated.

can refer to different model architectures in the various development phases and markets. This section discusses model selection and evolution in detail.

When entering a new market, instead of manually creating a rule system from scratch for occupation classifications, we employ machine learning. We start by selecting a model for training that makes the best use of the existing data. This existing data may include SME-curated occupation hierarchies and high-quality, rule-generated labels in existing markets. Linguistic and cultural resemblance between markets are two main factors guiding the model selection decision at this initial phase, because labels of the same language are often good training data sources and cultural resemblance is often reflected in the overlapping occupation concepts.

There are several model options that each come with their own benefits and tradeoffs at different phases. Model options include a multilingual BERT (M-BERT) model [9], and convolutional neural networks (CNN) [10, 11] with different training data. For new markets that share a language and concepts with established markets, a CNN model is trained to make initial predictions. If there is no resemblance in either aspect, we default to a global M-BERT model that is trained with high quality data from all the existing markets, regardless of linguistic or cultural similarity. Since the generalization power of M-BERT comes with tradeoffs of large model size and high inference latency, we have to limit the input length and therefore limit the prediction accuracy.

After the selected model generates predictions, SMEs conduct research and build out the market-specific occupation taxonomy iteratively. For example, SMEs identify novel market-specific occupations like *Judicial Scriveners*

in Japan, *Pizza Chefs* in Italy, *Hostel Wardens* in India, or removing occupations that do not apply to the specific market. The ongoing research is reflected by a dynamic layer at the inference time, which is used to process predictions into the market-specific taxonomy. This step is referred to as "Predictions fallback" in Figure 2.

To determine deployment eligibility, Taxonomy SMEs label the model predictions for evaluation on a concept-by-concept basis and push concepts to production based on a predetermined quality threshold. Since the top level taxonomy concepts are typically shared among all markets, it takes a short amount of time for a new market to deploy top level concept predictions in production.

For novel, market-specific concepts, or for concepts where prediction quality is below threshold, SMEs provide training labels for retraining the model. For deployed models and concepts, we follow a monitoring process to discover and act on issues.

After a market-specific taxonomy is built out, our research shows that a market-specific CNN model utilizing this knowledge with translated training data often outperforms the global M-BERT model. Therefore, the "model" in Figure 2 will be replaced by a market-specific model for subsequent loops.

This EITL approach leveraging machine learning models eliminates the most expensive task in the process: the need to curate and maintain a rule-based system in pursuit of capturing 100% of all possible classifiable syntaxes. It allows the Indeed Taxonomy team to divert resources to more SME-critical tasks, such as researching and creating market-specific occupation hierarchies and SME-annotated datasets, expediting deployment, and other metadata projects. We also benefit greatly by involving SME knowledge beyond just labeling like most traditional HITL systems, which we'll discuss in detail in the next section. As a result of this approach, we observed a six times faster deployment with similar SME resources was achieved in markets adopting this system (compared to the fully manual process), and this system is now in production in dozens of markets.

4. Challenges and Insights

We learned a lot from implementing this approach. In this section, we will discuss our success in leveraging subject matter expertise in the modeling process, adding transparency to a blackbox system, and establishing a monitoring process. We will also share insights from exploring different data sources and identifying the unequal nature of different errors.

4.1. Transfer subject matter expertise to understanding in data

Shadowing taxonomy SMEs and listening to their insights and challenges offered valuable information that would not otherwise have been obvious. Incorporating this information into the modeling process resulted in significant impact.

For example, we learned that although taxonomy hierarchies may vary in structure from market-to-market, there are occupation concepts across these markets that are conceptually identical but have different IDs due to those structural discrepancies. This posed an issue for cross-market training and prediction. By incorporating the mapping into model training, we saw gains in both precision and recall in all the locales, with an average increase of 5% in precision and 12% in recall.

Learning about differences in taxonomy concepts across markets also prompted us to explore and apply market-specific models after the initial set of market-specific hierarchies are built out. Experiments in the initial adopted market showed an average increase of 12% in precision when the threshold is set to keep recall the same.

4.2. Quality training and evaluation data at scale

Quality data is key for training and evaluating any automated systems. With limited resources, we explored outsourcing labeling and using user feedback data to directly approach the problem, and also revisited prioritization of the tasks based on our learning.

4.2.1. Outsourcing labeling

Crowd-sourcing labels for tasks where SME knowledge is needed has been known to be challenging [1, 12, 13]. After working with external labelers on multiple tasks with mixed outcomes in the past years, we learned that significant work needed to be put in upfront in order to gain high quality outcomes–underscoring the difference between EITL and HITL. Certain tasks need strong domain specific subject matter expertise which is hard to gain with short-term training sessions, particularly when much of the communication is asynchronous. Those tasks should not be outsourced in the first place. Guiding labelers to evaluate predictions required ongoing support in applying consistent classification heuristic to their labeling tasks. For example, coaching labelers to understand the difference between occupation and work environment was a unique challenge. Frequently, jobs for *School Custodians* would be labeled as a type of *Education & Instruction Occupation* due to their workplace environment. Occupationally, however, this should be a

Cleaning & Grounds Maintenance Occupation. Moreover, if a task involves distinguishing between specific licenses or specialties in different industries, significant research and domain knowledge are needed in order to produce high quality labels.

After identifying tasks suitable to outsource, the next steps are to work with SMEs to scope the task, provide clear guidelines, and set up monitoring-QA-calibration loops to help ramp up the quality. In a labeling task where the goal is to generate evaluation data for top level occupation predictions in a new market, we found it to be more efficient to scope the task in the binary fashion: During the labeling task, labelers are asked to label the predictions such as "Job A is a type of *Education Instruction Occupations* job" as *correct* or *incorrect*. The incorrect predictions are then sent to SMEs to assign correct labels. This second step is essential to improve model quality but extremely hard for non-SMEs since it requires familiarity with definitions of a few dozen occupation concepts.

4.2.2. User input data

User input data comes in large volume, but tends to be noisy [4, 14]. After reviewing user input data against SME labels, we learned that there could be discrepancies associated with understandings of job descriptions and concept definitions, often resulting in different or missing labels. In a model-assisted user input task, we found 20-30% of user overrides disagreed with SME labels. In addition, user input was easily influenced by UI design. A recent UI change making the editing option less obvious resulted in a 4% decrease in users overriding the predictions. These challenges reflect the quality tradeoff associated with this free data source.

4.2.3. Prioritization of labeling tasks

We learned that, in terms of labeling, quality usually comes with the tradeoff of cost and speed, especially when the dimension of label space is in the thousands. With that learning, we decided to prioritize SME resources for evaluation datasets, while using alternative data sources for training, such as taxonomy data from other markets and user input. When certain classes proved to perform poorly with the alternative training data source, we then prioritized labeling better training data for those specific classes. With this decision, we were able to significantly decrease time to the first model deployment and evaluation.

4.3. Understand what error matters

In the classic machine learning model measurements, error is defined to be a binary concept. However, with our close collaborations with SMEs on labeling tasks, we learned that there is usually a certain level of subjectivity associated with decisions, which naturally brought us this question: How bad are the errors? For example, if the model makes the same error as a non-SME user, it likely does not hurt user experience as much as a completely out-of-place error. After examining different definitions, considering trade-offs between information granularity and interpretability, we ended up defining three levels of error severity with easy interpretations: 1) *the mild errors* are within 2 hops of the correct labels, typical examples include siblings, parents and children nodes; 2) *the moderate errors* are at the same top levels with the correct labels, usually due to granularity issues; and 3) *the severe errors* are in different top levels, which will result in bad user experience if not fixed.

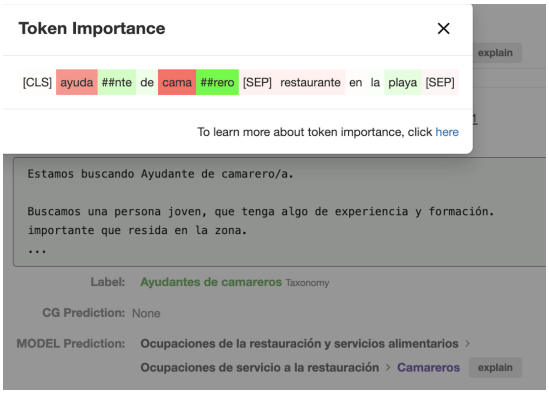
Having this insight allows the SME team to conduct more nuanced error analysis such that they can better prioritize severe errors for quality improvement tasks. It also allows the data science team to measure quality in a more practical manner, and utilize data from multiple sources which would be considered too noisy otherwise.

4.4. Establish process of monitoring, diagnosis, and action on issues

With machine learning models in production supporting dozens of locales and thousands of taxonomy concepts in total, we needed to monitor performance of the models and act on issues in a timely manner. Therefore we are exploring a workflow with SMEs, which involves a monitoring dashboard with alerts on the individual concept level and a process to diagnose and act on issues. The alerts are designed to be triggered when a drastic drop in performance is seen in a node with sufficient labels. Those alerts will then be sent to SMEs in each locale, marked with an urgency level based on the scale of the drop. The SMEs will follow a triaging decision tree, leading to either ignoring the alert or creating a ticket with their observations and judgments for data scientists or engineers to resolve.

4.5. Strive for transparency into black-box systems

Deep learning systems are often treated as black boxes, but this technical opacity can lead to inefficient cross-functional collaboration. For example, during the initial collaboration period, SMEs were tasked to provide labels for a model in a market with around a thousand concepts. After a significant amount of work was put in, we realized the evaluation metrics were inaccurate due to nonrandom selection of labeled documents. In addition, the initial instruction of providing a large fixed amount of labels per concept could be improved since concepts that perform



[Example of the token importance function applied to a job in the internal tool]

Figure 3: Visualization of the token importance function in the internal tool. Green / red highlights indicate positive / negative contributions to the model prediction.

well do not need additional training data, and the ones with fewer jobs struggle to find enough documents to be labeled and have little effect on the overall performance.

We started with knowledge sharing, Q&A, and discussion sessions to connect common operation actions with implications on the models, and gradually built out resources and documentation for FAQs. In addition, we built a function to visualize token importances in a given training sample with the *Integrated Gradients* [15] technique, which is used by SMEs to promote transparency into why a document might be predicted as a given occupation.

These insights and successes could not have been achieved without close collaboration and communication. It enabled us to effectively distribute SME resources to the highest priority task, which resulted in the research, development, and publishing of over 700 existing or new occupation taxonomy concepts in strategically significant new markets within six months.

5. Conclusion and future work

In this paper, we presented a human-in-the-loop system at Indeed that scales occupation taxonomy development and application to international markets with high quality outcomes. Over the course of developing the system, we found tremendous value in leveraging subject matter expertise throughout the process, ranging from data collection and processing, model training and evaluation, to performance monitoring and diagnosis. Therefore we propose calling this system expert-in-the-loop.

In ongoing work, we are looking to partner more closely among functions to allow for combined SMEs, explore further improvements to the current machine learning models, and build better functionality for overriding training and production data.

Acknowledgments

We would like to thank everyone on the Metadata & Taxonomy team that contributed to this system for their diligent work and thoughtful collaboration. We are grateful to the reviewers of this paper for providing valuable feedback and comments.

References

[1] S. P. Ponzetto, M. Strube, Deriving a large-scale taxonomy from wikipedia, in: AAAI, 2007.

[2] W. Wu, H. Li, H. Wang, K. Q. Zhu, Probase: a probabilistic taxonomy for text understanding, Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data (2012).

[3] F. Javed, Q. Luo, M. McNair, F. Jacob, M. Zhao, T. S. Kang, Carotene: A job title classification system for the online recruitment domain, 2015 IEEE First International Conference on Big Data Computing Service and Applications (2015) 286–293.

[4] P. Das, Y. Xia, A. Levine, G. D. Fabbrizio, A. Datta, Large-scale taxonomy categorization for noisy product listings, 2016 IEEE International Conference on Big Data (Big Data) (2016) 3885–3894.

[5] N. Noy, Y. Gao, A. Jain, A. Narayanan, A. Patterson, J. Taylor, Industry-scale knowledge graphs: Lessons and challenges: Five diverse technology companies show how it's done, Queue 17 (2019).

[6] E. D. Gennatas, J. H. Friedman, L. H. Ungar, R. Pirracchio, E. Eaton, L. G. Reichmann, Y. Interian, J. M. Luna, C. B. Simone, A. Auerbach, E. Delgado, M. J. van der Laan, T. D. Solberg, G. Valdes, Expert-augmented machine learning, Proceedings of the National Academy of Sciences 117 (2020) 4571–4577. doi:10.1073/pnas.1906831117.

[7] B. C. Wallace, K. Small, C. E. Brodley, J. Lau, T. A. Trikalinos, Deploying an interactive machine learning system in an evidence-based practice center: Abstrackr, in: Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium, IHI '12, Association for Computing Machinery, New York, NY, USA, 2012, p. 819–824. doi:10.1145/2110363.2110464.

[8] R. Ghani, M. Kumar, Interactive learning for efficiently detecting errors in insurance claims, in: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '11, Association for Computing

Machinery, New York, NY, USA, 2011, p. 325–333. doi:10.1145/2020408.2020463.

[9] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[10] Y. LeCun, Y. Bengio, Convolutional networks for images, speech, and time series, 1998.

[11] Y. Kim, Convolutional neural networks for sentence classification, in: EMNLP, 2014.

[12] C. Sun, N. Rampalli, F. Yang, A. Doan, Chimera: Large-scale classification using machine learning, rules, and crowdsourcing, Proc. VLDB Endow. 7 (2014) 1529–1540.

[13] R. Bekkerman, M. Gavish, High-precision phrase-based document classification on a modern scale, in: KDD, 2011.

[14] D. Shen, J.-D. Ruvini, B. M. Sarwar, Large-scale item categorization for e-commerce, Proceedings of the 21st ACM international conference on Information and knowledge management (2012).

[15] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, ArXiv abs/1703.01365 (2017).