Research Statement

John P. Lalor

Current Research

My research focuses on evaluation, both of humans and machine learning models. I have done work on applying methods from Item Response Theory (IRT) in psychometrics to building new evaluation scales for humans in the area of health literacy, and in natural language processing (NLP) for tasks such as natural language inference (NLI) and sentiment analysis (SI). These tests use latent parameters of the test questions ("items") to estimate a latent ability of the test taker (human or machine) to place the test taker on a population scale with regards to other test takers. Evaluation and analysis of ML models beyond aggregate measures such as accuracy is important to understand how models perform in specific conditions so that they can be trusted when they are put into use. This also allows for researchers and stakeholders to better understand model performance and provide a level of transparency to model performance. This research has also led to new learning methods that take advantage of the same learned item parameters, but to inform data selection for machine learning (ML) model training. My work thus far has been published in top NLP conference proceedings [1, 2]. I have also published work in peer-reviewed workshops focused on learning with uncertainty in deep neural networks [3, 4]. My applied work on a new test of health literacy has been published in top medical informatics journals [5, 6].

In my dissertation, I introduce the methods from IRT to building test sets for ML problems, specifically in NLP. I also develop new ways to use the learned item parameters, specifically difficulty, to improve ML model generalization in terms of performance on held out test sets. Incorporating IRT for ML models allows for more robust testing of the models by taking into account the difficulty of the specific test set items. The analyses resulting from this work has also shown that certain deep neural network models learn in similar patterns as humans, that is they learn easy items more quickly than they do hard items.

Evaluating machine learning models Typically machine learning (ML) models are evaluated using aggregate performance scores on a held-out test set (e.g. accuracy). However these metrics fail to account for characteristics of specific test set examples that may have an impact on performance. I have used IRT methods from psychometrics to develop new test sets for ML models in NLP tasks to show that aggregate metrics do not always represent performance accurately [1]. IRT enables us to estimate latent difficulty and discriminatory parameters of the items in a test set. For example, in the NLP task of NLI, the difficulty of the items in the test set makes a meaningful difference in how well an NLP model performs with respect to a large population of Amazon Mechanical Turk (AMT) crowdsourcing workers. Both the AMT workers and the NLP models label most easy examples correctly, and therefore the high performance of the NLP model is not as impressive. For harder examples the NLP model does worse in terms of raw accuracy, but because the AMT workers also perform worse, the model's performance with respect to the population is higher. By using IRT to construct test sets for problems in NLP, we can get a better understanding of how model

performance should be interpreted with respect to a population of test-takers instead of as a single raw aggregate score.

Knowing the latent difficulty of test set items also allows for more in-depth analysis of the performance of deep learning models as the model training setup is changed. We trained three high-performing deep learning models on two NLP tasks: NLI and SA, varying the size of the training set to simulate better- and worse-performing models. A regression to predict whether the model would label a test item correctly as a function of the item's difficulty and the training set size of the model found that not only were difficulty and training set size significant independently, but also that the interaction was significant [2]. Difficulty and training set size make sense, since harder items should be harder, and more training data should lead to higher performance. However we also found that as more data is used to train the model, the easy items get easier faster than the harder items. That is, the model is learning the easy items faster than the hard items, which is consistent with results found previously in how humans learn. It is also interesting that the difficulty parameters used for this analysis were learned from a human population of responses, so those human difficulties are informative for the deep learning models as well. This result is intuitive but had not been previously shown empirically. There is now an empirical justification for many of the heuristics behind work in areas such as curriculum learning.

Incorporating uncertainty It is typical when training an ML model in a single-class classification problem to assume that all examples of a class are equally appropriate for that class. However it may be the case that certain examples are more or less appropriate for the class due to ambiguity or uncertainty. We have used this insight to develop a new method for deep learning models that uses distributions over labels ("soft labels") to introduce uncertainty to the model training process. I used AMT to collect a large number of annotations for a subset of data in NLI and SA datasets and treated the distribution of responses over labels as the soft labels for each item. Using the soft labels for a small subset of training examples (1%) for both tasks led to significant improvement over several baselines, including a "comparable effort" baseline where the annotation budget to create the soft labels was instead used to label new training examples [3]. In addition, examining the output of the fine-tuned models shows that certain changes in model output reflect a better representation of uncertainty than in the baseline models.

Measuring and improving patient health literacy In my applied work I have focused on problems in health informatics, specifically patient health literacy. With my coauthors I developed a new test of health literacy to assess a patient's ability to understand his or her electronic health record (EHR) free-text notes [5]. The ComprehENotes test is the first test of its kind, and was built from de-identified patient notes using Sentence Verification Technique (SVT) to generate the questions and IRT to validate the question set. I used the ComprehENotes test to confirm prior self-reported results on patient EHR note comprehension that is consistent with demographic trends associated with low health literacy.

With the ComprehENotes test it is now possible to quantitatively evaluate the effectiveness of tools designed to improve patient EHR note comprehension. One such tool is NoteAid, which was developed in our lab[1]. NoteAid is a tool that automatically identifies and defines medical terms and jargon in patient EHR notes. NoteAid had previously been shown to improve patient EHR note understanding, but those scores were self-reported. I conducted a randomized experiment of Amazon Mechanical Turk crowd workers to assess whether access to NoteAid led to improved scores on the ComprehENotes test. Turkers were given the ComprehENotes test, either with no

[1] `http://www.clinicalnotesaid.org/emrreadability/emrsimplifier.uwm`

accompanying information, with a link to the MedlinePlus online medical information repository, or with NoteAid definitions embedded into the questions. The MedlinePlus group was included to simulate what resources are available to a patient interested in learning more about his or her own medical history. At present, most resources are active in that a patient has to go and search for his or her symptoms or condition to find information. NoteAid is a passive system for the patient, where definitions are immediately available to the patient without any additional effort. The patients with NoteAid ComprehENotes scores werer significantly higher than the other two groups [6].

Future Research

There are several directions for future work that I plan on pursuing, both on core machine learning problems and in applied work. First, building new test sets with IRT for machine learning to expand the types of problems in ML for which we have access to such tests. New tests are always needed in the machine learning community to keep up with the rapid development of new models. New state of the art numbers are being posted almost daily, and it is important to make sure that the tests being used are doing a good job of evaluating performance. As more models are released and used for a particular task, IRT can be used to learn latent item parameters and estimate latent ability for these models. Second, there are open research questions about the scalability of IRT to large ML datasets as opposed to the short tests typically built with IRT. In order to model latent parameters of datasets with hundreds of thousands or millions of items, we would not be able to use human response data as input to the IRT models. Instead we would have to use the responses of trained ML models for the specific task. Whether this data is appropriate as input is still an open question that I plan to address in my future research. Third, I plan on building and evaluating new ways to improve patient health literacy, specifically with regards to patients' own EHR notes.

Large-scale psychometric testing of machine learning models At present, IRT models do not scale well, because they never had to. Typically, a dataset for an IRT model would consist of several thousand students providing response patterns for a few hundred items at most. For example, building a standarized test such as the GMAT does not require hundreds of thousands of questions, but a standard machine learning benchmark, the MNIST handwritten digit dataset, contains 50,000 training examples and 10,000 test examples. In order to scale these methods to the datasets used in machine learning, particularly deep learning, new methods need to be developed. That way IRT models can be fit using the response patterns of ML models instead of humans. Learning IRT models for ML datasets allows for several new and exciting areas of research, from new learning algorithms that leverage the difficulty of specific examples during learning or optimize for latent ability to new test sets built using psychometric methodologies from the existing large-scale test sets currently available. There are two important questions that need to be answered in order to do this: (1) Are the learned IRT parameters interpretable when the input data is machine response patterns instead of human response patterns? (2) Can IRT models scale to hundreds of thousands or millions of items while maintaining the underlying model assumptions? The answer to question 1 is not obvious. It is not necessarily true that the response patterns of machine learning models will lead to similar observations as those of humans. This needs to be tested to ensure that learning IRT models with ML model data makes sense. I plan to test this assumption at a small scale, using models that we have already fit from human response patterns to compare models fit using ML model response patterns. I don't expect that the learned latent parameters will be identical, but if these methods are to be useful moving forward, I hypothesize that certain patterns will emerge that are consistent with the human-data models (e.g. positive correlation in terms of difficulty rankings).

The second question is important if these models are to be used on a large number of machine learning problems. Most datasets used to train ML models are very large and getting larger, consisting of hundreds of thousands or millions of data points. Typical IRT models are fit using marginal maximum likelihood methods which cannot scale to datasets of that size. There exist variational inference methods for fitting models that have been proposed for human-scale IRT models, but have never been tested at the scale that we would require. Using variational inference methods would allow for the learning of latent ability and item parameters using machine response patterns on entire datasets instead of data subsets. I plan to test these methods qualitatively by looking at the "easiest" and "hardest" elements in a dataset using machine response patterns. Confirming the validity of ML model data as input for IRT models and being able to learn the models with large-scale datasets opens a number of interesting research questions about how we use these models. For existing ML test sets, each new state of the art model developed can be evaluated not only on raw accuracy, but also with the latent ability estimate as determined by the model responses to the test set. Learned item parameters such as difficulty can be used to filter training sets to only include easy or hard items, and can also be used to develop new selection criteria for curriculum learning and active learning methods.

Improving patient health literacy With the ComprehENotes test we now have the ability to quantitatively meassure a patient's ability to understand his or her EHR notes. This leads to new research questions in the areas of (1) measuring the effectiveness of certain interventions on patient EHR note comprehension as measured by ComprehENotes and (2) work on the development of more personalized measures of note comprehension. Ideally, a patient could be assessed based on questions from his or her own EHR notes. Being able to generate personalized questions that are useful for assessment required new methods from NLP in terms of automated question generation, distractor identification, and test question assessment. I will conduct research into these methods as a core NLP problem with a focus on applying the methods to the task of medical question generation. With new NLP methods in place, we can use ComprehENotes as a benchmark to compare personalized test performance across patients.

References

[1] John P Lalor, Hao Wu, and Hong Yu. Building an evaluation scale using item response theory. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2016, page 648, 2016.

[2] John P Lalor, Hao Wu, Tsendsuren Munkhdalai, and Hong Yu. Understanding deep learning performance through an examination of test set difficulty: A psychometric case study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2018, 2018.

[3] John P Lalor, Hao Wu, and Hong Yu. Soft label memorization-generalization for natural language inference. In *UAI Workshop on Uncertainty in Deep Learning*, 2018.

[4] John P Lalor, Hao Wu, and Hong Yu. Scaling item response theory with stochastic variational inference. In *NIPS Bayesian Deep Learning Workshop (under review)*, 2018.

[5] John P Lalor, Hao Wu, Li Chen, Kathleen M Mazor, and Hong Yu. Comprehenotes, an instrument to assess patient reading comprehension of electronic health record notes: Development and validation. *Journal of Medical Internet Research*, 20(4), 2018.

[6] John P Lalor, Beverly Woolf, and Hong Yu. Improving ehr note comprehension with noteaid: A randomized trial of ehr note comprehension interventions with crowdsourced workers. *Journal of Medical Internet Research (in press)*, 2018.