Revisiting Mechanical Turkers? How to Evaluate Learning Results
with Semantic Properties

Fred Morstatter and Huan Liu
School of Computing, Informatics, and Decision Systems Engineering
Arizona State University
Tempe, Arizona 85281

1. INTRODUCTION

Some machine learning algorithms offer more than just predictive power. Such algorithms provide additional insight into the underlying data. Examples of these algorithms are topic modeling algorithms such as Latent Dirichlet Allocation (LDA) [Blei et al. 2003], whose topics are often inspected as part of the analysis that many researchers perform on their data. Recently deep learning algorithms such as word embedding algorithms like Word2Vec [Mikolov et al. 2013] have produced models with semantic properties. These algorithms are immensely useful; they tell us something about the environment from which they generate their predictions.

One pressing challenge is how to evaluate the quality of the semantic information produced by these algorithms. When we employ algorithms for their semantic properties, it is important that these properties can be understood by a human. Currently, there are no established approaches to carry out this evaluation automatically. This evaluation (if done at all) is usually carried out via user studies. While this type of evaluation is sound, it is expensive both from the perspective of time and cost, and thus cannot be easily reproduced independently. These experiments have the additional drawback of being hard to scale and difficult to generalize. We pose the challenge of evaluating the information quality of these semantic properties – Can we find automatic methods for evaluating the semantic properties of algorithms that apply as easily as traditional textbook evaluation metrics for the predictive quality of models.

2. EVALUATING INFORMATION QUALITY WITH SEMANTIC PROPERTIES

Topic modeling algorithms have gained immense popularity for their ability to find underlying patterns in a dataset. LDA [Blei et al. 2003], one popular topic modeling algorithm, produces two types of information: 1) a clustering of unseen documents, and 2) the “topics” discovered within the text. The output of topic modeling algorithms has been examined by researchers across many domains in order to understand their data [Grimmer and Stewart 2013]. While the analysis is common, scientifically evaluating how well humans can understand the underlying topics is not common. [Chang et al. 2009] attempted to answer this question by proposing methods that involve crowdsourced workers to answer the questions. By showing topics to users with questionnaires about which words do not fit with the topic, the authors proposed a new measure to assess topic quality.

Word embeddings are another class of algorithms that produce additional information about the data. These algorithms learn representations of words by building vector representations for each word in the corpus. In addition to their predictive ability, these vectors are also semantically meaningful. For instance, the models generate vectors for each word in the vocabulary where words placed closer together are meant to have a higher semantic similarity. Automated methods to measure how well these embeddings match human understanding are needed. These measures will generate embeddings that are more interpretable to humans. However, measures for the semantic performance of Word2Vec have not been formalized.
3. CLOSING THE GAP

One natural approach to closing the gap is to apply well-established measures for predictive performance. After all, these measures are objective and easily reproducible. In fact, measures of predictability can be negatively correlated with those of interpretability. Among possible reasons, it is mostly due to the models’ tendency to overfit the data, yielding patterns that are too opaque for humans to interpret, as confirmed in [Chang et al. 2009]. Predictive measures will not be suitable for this challenge.

In [Morstatter et al. 2015], an attempt is made to answer how well humans can understand topics by proposing a new measure to judge Turkers’ understanding of the topics. The idea is that LDA topics represent real-world topics, such as a “sports” or “business” section in a newspaper. As shown in Figure 1, the measure judges the capability of the Turkers to judge the contents of the topic as either “sports” (blue) or “business” (red). The better the Turkers identify the correct distribution of categories by looking at the top words in the topic, the better the score. In addition to this measure that relies on crowdsourcing, the authors go further to propose offline measures that can approximate the crowdsourced measure. These offline measures work by hypothesizing how humans make judgments about topics. This shows that it is possible to find automatic indicators of semantic performance. While this is a step toward automatic semantic evaluation, further work is needed to generalize these measures.

It is an incredible advantage to researchers across many disciplines that modern algorithms can not only predict, but also provide additional information about the data. The onus is on us, researchers in data quality, to devise objective evaluation measures to ensure low-cost, scalable, and easily reproducible evaluation results so as to replace crowdsourced workers and to help advance research in the age of big data.

REFERENCES


