Transforming Assessment with NLP

On-screen: [As part of our work to reassess the future and support the next generation of physicians, NBME is exploring how innovations, such as the application of natural language processing (NLP), can move us beyond traditional multiple-choice exams.]

NLP holds great potential in improving teaching and learning both within medical education and assessment and within education as a whole.

As we try to measure new constructs and we try to develop new item formats, this comes with its own set of challenges. For example, when we have items that require free-text responses, this comes with the challenge of how are we actually going to rate these free-text responses.

On-screen: [How are we trying to improve teaching and learning through NLP?]

There are many obstacles currently existing when we talk about scoring free-text responses or clinical patient notes. The first one that comes to mind is how resource intensive the whole exercise is, recruiting experts having them go through very thorough training and having them rate a very large number of patient notes, for example.

Another reason why it is challenging is because we use human raters very often as black boxes. Yes, human raters can justify why they have made a certain rating decisions, but because they have to rate so many patient notes or so many responses, we rarely ask them to identify the specific phrases in the response that correspond to a concept in the rubric. Instead, they give an overall rating, while NLP scoring systems such as the one developed in the Kaggle competition can efficiently and reliably identify the exact phrase in a text and map it to a clinical concept resulting in a very interpretable, well visualized rating of the patient note.

On-screen: [What are some of the obstacles educators face when scoring free-text responses, such as clinical patient notes?]

The field of medicine is very specific when it comes to NLP applications and that would come up with two main reasons. The first one is that clinical text contains a lot of specialized terms and concepts that are complex and have intricate relationships with each other. And so detecting these relationships automatically is a very challenging problem.

Another very big hurdle in clinical NLP is the fact that clinical data often contains personal information and so it cannot easily be shared between institutions. At the same time, NLP tools, they feed off having relevant data to train on and so that lack of data availability results in a lag in clinical NLP compared to other areas of NLP development.

On-screen: [What are some of the obstacles educators face when scoring free-text responses, such as clinical patient notes?]

Even though NLP is an efficient way to have the automated scoring of patient notes and free-text responses in the clinical domain, there are still challenges that remain. One of these challenges is that to adapt the algorithms to new cases, for example, we still need to have some level of annotation for these new cases to retrain the models on. And so obtaining that annotation is a requirement.

Also, annotators make mistakes and a model that is trained on these human annotations and benchmarked on them can only be as good as the human annotators are. There is no very long history of collaboration between NLP specialists, data scientists and the medical education assessment community. And creating that history of collaboration and shared understanding of challenges and potential solutions is an important thing we need to build to tackle these problems.

On-screen: [How can NBME better collaborate for change? Please share your thoughts and complete the survey. www.reassessthefuture.org/survey.]