Examining Item Difficulty in NLP: To What Extent Do Examinees Affect Item Difficulty?

Yo Ehara
Tokyo Gakugei University, Koganei, Tokyo, Japan

Abstract

Recent research in Natural Language Processing (NLP) has focused on estimating the difficulty of text content, culminating in a shared task conducted in 2025. However, since many researchers in NLP are not experts in educational psychology, the item difficulty in these shared task datasets is commonly defined by the proportion of examinees who answer an item correctly, and language model performance is evaluated accordingly. This definition is inherently sensitive to changes in the set of examinees who answer correctly, thereby altering item difficulty. To overcome this issue, educational psychology employs item response theory (IRT) to separate item difficulty from the examinee population. In this study, we investigate the extent to which language model performance evaluations differ when using IRT compared to the traditional method, based on the proportion of examinees who answered items correctly.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*