Information Processing and Management 54 (2018) 969–984

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Automated estimation of item difficulty for multiple-choice tests: An application of word embedding techniques

Hsu Fu-Yuan[a,d,*], Lee Hahn-Ming[a,e], Chang Tao-Hsing[b], Sung Yao-Ting[c]

[a] Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, No.43, Sec.4, Keelung Rd., Da'an Dist., Taipei City 106, Taiwan
[b] Department of Computer Science and Information Engineering, National Kaohsiung University of Applied Sciences, No.415, Jiangong Rd., Sanmin Dist., Kaohsiung City 807, Taiwan
[c] Department of Educational Psychology and Counseling, National Taiwan Normal University, No.162, Sec.1, Heping E. Rd., Da'an Dist., Taipei City 106, Taiwan
[d] Research Center for Psychological and Educational Testing, National Taiwan Normal University, No.162, Sec.1, Heping E. Rd., Da'an Dist., Taipei City 106, Taiwan
[e] Institute of Information Science, Academia Sinica, Taiwan, No.128, Sec.2, Academia Rd., Nankang Dist., Taipei City 115, Taiwan

ARTICLE INFO

Keywords:
Multiple-choice item
Item difficulty estimation
Cognitive processing model
Semantic similarity
Word embedding
Machine learning

ABSTRACT

Pretesting is the most commonly used method for estimating test item difficulty because it provides highly accurate results that can be applied to assessment development activities. However, pretesting is inefficient, and it can lead to item exposure. Hence, an increasing number of studies have invested considerable effort in researching the automated estimation of item difficulty. Language proficiency tests constitute the majority of researched test topics, while comparatively less research has focused on content subjects. This paper introduces a novel method for the automated estimation of item difficulty for social studies tests. In this study, we explore the difficulty of multiple-choice items, which consist of the following item elements: a question and alternative options. We use learning materials to construct a semantic space using word embedding techniques and project an item's texts into the semantic space to obtain corresponding vectors. Semantic features are obtained by calculating the cosine similarity between the vectors of item elements. Subsequently, these semantic features are sent to a classifier for training and testing. Based on the output of the classifier, an estimation model is created and item difficulty is estimated. Our findings suggest that the semantic similarity between a stem and the options has the strongest impact on item difficulty. Furthermore, the results indicate that the proposed estimation method outperforms pretesting, and therefore, we expect that the proposed approach will complement and partially replace pretesting in future.

1. Introduction

This section is divided into two subsections. The first subsection introduces the background of the study. The second subsection describes the purpose and research questions of this study.

* Corresponding author at: Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, No.43, Keelung Rd., Sec.4, Da'an Dist., Taipei City 106, Taiwan.
E-mail addresses: D9715009@mail.ntust.edu.tw, kevinhsu@ntnu.edu.tw (F.-Y. Hsu), hmlee@mail.ntust.edu.tw, hmlee@iis.sinica.edu.tw (H.-M. Lee), changth@cc.kuas.edu.tw (T.-H. Chang), sungtc@ntnu.edu.tw (Y.-T. Sung).

https://doi.org/10.1016/j.ipm.2018.06.007
Received 7 November 2017; Received in revised form 11 June 2018; Accepted 19 June 2018

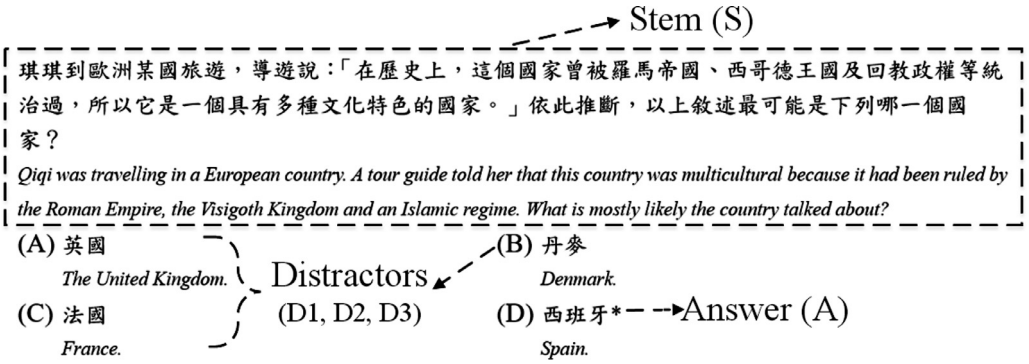

Fig. 1. Practical example of item elements.

1.1. Background

The widespread use of the Internet and the rapid development of information technology have exerted a strong influence on assessment development. Recent advancements in educational assessment and evaluation techniques reflect a trend in movement away from conventional paper-based testing towards computer-based testing (CBT). Advances in psychometrics and the development of the item response theory (IRT) (DeMars, 2010) have directed CBT's evolution into computer adaptive testing (CAT). CAT is advantageous in that it provides more efficient estimates of ability using fewer items compared to conventional testing. To ensure that CAT achieves its assessment targets, item analysis is indispensable for determining item quality, which is used to identify items that require modification or deletion before banking. Item difficulty is an index of primary importance in item analysis. Based on the preceding statement, we infer that accurately estimating item difficulty is of considerable importance in testing, as obtaining individual item difficulty values allows for the estimation of the difficulty of an entire test and the abilities of student test takers.

Item types are closely related to item difficulty, even though the implementation of item difficulty estimation varies across different item types. In this study, we explore the difficulty of multiple-choice items (MCIs), which consist of the following item elements: a question (stem) and typically three to five alternatives (or options), from which students must select. The alternatives include a correct option (answer) and a few some plausible but incorrect options known as distractors. A practical example of item elements is shown in Fig. 1. This type of item requires students to integrate stem information with their background knowledge to select the correct option. Reliable tests using MCIs are straightforward to implement, prevent the requirement for manual intervention in scoring, and are well suited for learning diagnosis and achievement evaluation. Thus, MCI-based tests are widely used in education, professional certification, and licensure. These attributes and applications have led MCI tests to be considered one of the most effective and successful forms of educational assessment (Gierl, Bulut, Guo, & Zhang, 2017).

Before drafting a test, large-scale testing organizations typically set parameters including item format, content range, and item difficulty configuration, according to test objectives. Then, they employ algorithms such as linear programming or genetic algorithms, to select the best combination of test items from an item bank that meets the criteria established by the parameters. Item difficulty is presented as categorical information, and it is generally classified into the following five levels: very easy, easy, moderate, difficult, and very difficult. At present, the primary methods for estimating item difficulty are pretesting and subjective expert judgment (Attali, Saldivia, Jackson, Schuppan, & Wanamaker, 2014). The subjective expert judgment of item difficulty does not require students. Instead, this approach obtains estimates based on experts' experience and the intuitive judgments of item difficulty. It is difficult to evaluate result stability owing to the subjectivity of expert judgment. Alternately, item difficulty can be assessed empirically by pretesting items prior to employing them in an examination. Pretesting informs the item selection process based on item difficulty, which is obtained from the analysis of the collected item responses of representative subjects that are randomly sampled from the exam population. Even though this process can achieve highly accurate estimates of item difficulty, it is relatively labor intensive and time consuming and must consider item exposure, particularly in the development of high-stakes tests (Loukina, Yoon, Sakano, Wei, & Sheehan, 2016). Thus, there is significant value in developing an automated procedure that could evaluate item difficulty and ensure sufficient psychometric quality of test items.

Even though numerous variables that affect item difficulty have already been identified by scholars, much opportunity remains for improvement in the automated estimation of item difficulty. This is largely attributed to the complex interaction among these variables and the complicated relations between item difficulty and item demands (El Masri, H., Ferrara, Foltz, & Baird, 2017; Pollitt, Ahmed, & Crisp, 2007). Ferrara, Svetina, Skucha, and Davidson (2011) define item demands as the knowledge, comprehension, and cognitive process required for examinees to correctly answer an item. Until now, research on automatic estimation of item difficulty has focused on language proficiency tests, while studies on content subjects, such as social sciences, natural sciences, medicine, and law, have received less attention. However, content subject tests are widely used in education assessment, certification, and licensure examinations. Usually, studies estimating the difficulty of items in language proficiency tests employ tools to automatically analyze linguistic features (Sheehan, 2017), or rely on external word lists and electronic lexical databases such as WordNet (Ronzano, Anke, & Saggion, 2016) to extract item difficulty characteristics to direct the automated estimation of item difficulty. However, in content subject tests, examinees must apply the knowledge and materials learned in class to a stem to select the correct answer from multiple