Institution: University of Wolverhampton

Unit of Assessment: 26 Modern Languages and Linguistics

Title of case study: Improving High-stakes Medical Examinations through Natural Language Processing

Period when the underpinning research was undertaken: 2003 – 2020

Details of staff conducting the underpinning research from the submitting unit:

Name(s):	Role(s) (e.g. job title):	Period(s) employed by submitting HEI:
Dr Le An Ha	Senior Lecturer	2005 – Present
Dr Richard Evans	Lecturer	1998 – Present
Dr Nikiforos Karamanis	Research Associate	2005 – 2005
Dr Victoria Yaneva	Lecturer	2016 – Present
Professor Ruslan Mitkov	Professor of Computational Linguistics and Language Engineering	1995 – Present

Period when the claimed impact occurred: 2014 – 2020

Is this case study continued from a case study submitted in 2014? N

1. Summary of the impact

The National Board of Medical Examiners (NBME) is a not-for-profit organisation which develops and administers the United States Medical Licensure Examination (USMLE), a compulsory exam for roughly 90,000 US medical students per year. The students who undertake this exam do so in order to obtain a medical license, without which they would not be allowed to practice medicine within the USA and Canada. NBME is, therefore, safeguarding the American public by ensuring that all practising physicians have the necessary knowledge and skills to provide high-quality healthcare.

The Research Group in Computational Linguistics (RGCL) and NBME have collaborated since 2005, which has impacted on both public services [I1] through the adoption of new and improved technologies in the service that NBME provides (medical licensing exams) and on practitioners and the delivery of professional services [I2] through changes to workforce planning and assessment practice.

2. Underpinning research

Computational Linguistics / Natural Language Processing (NLP) is a branch of artificial intelligence that focuses on the computational processing of natural languages. NLP research areas where the RGCL has prominent contributions include machine translation, question answering, automated question generation, text simplification, readability assessment, and many others.

The collaboration with NBME began in 2005, as a result of RGCL's pioneering work on the automatic generation of multiple-choice questions [R1, R2, R3], in which NBME saw the potential to significantly improve the quality and speed of developing medical examinations. These examinations include the United States Medical Licensing Examination (USMLE), which is taken by roughly 90,000 students annually in the USA and Canada (e.g. 90,086 students for all four USMLE components in 2018) and is also widely used outside the USA, including in Brazil and Italy (e.g. 54,917 students from non-US/Canadian schools in 2018). Additionally, the organisation maintains and improves the quality and efficiency of the examinations, and ensures fairness to guarantee that medical practitioners are competent in the knowledge and skills required by

medical services in the USA and beyond. To fulfil this role, NBME is in constant need of developing new, high-quality multiple-choice questions, which was the initial reason for their interest in exploring NLP-based approaches for automating parts of their test development process.

This collaboration has resulted in the two research findings, which led to the development of application categories that have benefited NBME:

F1. The benefit of using NLP to automatically generate incorrect answer options or new multiple-choice questions

This first finding relates to the idea that test questions along with their distractors (the set of a question, correct answer, and distractors is called an *item*) can be generated automatically. Original technology developed by RGCL utilised the automated generation of test items in two different ways. The first of these [R1] helps item writers by suggesting plausible incorrect answer options for their questions, which has been shown to help train novice item-writers and reduce the time needed for item development. Another application is the fully automatic development of items from existing text passages [R2, R3]. Instead of having item-writers write the questions, their efforts are only needed to post-edit automatically generated questions, which are extracted and adapted from relevant text passages. These applications have the potential to improve the efficacy, quality, and cost of exam development by alleviating the burden of writing items from scratch.

F2. The benefit of predicting item characteristics

The second finding relates to the idea that we can extract relevant information from the linguistic content of test items. For example, we can predict how long it would take on average for test-takers to answer an item [R4] (*time intensiveness*) or what proportion of the test-takers would be able to answer it correctly [R5, R6] (*item difficulty*). This technology, described in R4, R5, and R6, is a novel application of NLP to educational testing needs in that these topics had not been investigated within the NLP field until that point. Undertaking this research has a clear benefit for improving exam fairness since it enables test development companies to ensure that exam forms seen by different groups of examinees (for example, male vs female examinees, examinees with various socio-economic backgrounds, etc.) do not differ significantly in terms of their difficulty or the time required to answer the items correctly. Failing to ensure this can potentially cause significant disadvantages for some students and diminish the fairness and reliability of the exam.

In addition to the two research findings discussed above, further NLP research conducted by RGCL staff includes applications for automatic scoring of the examinees' writing, which required developing a methodology for mining clinical text and mapping the extracted information to a score for each student. This technology provides an objective addition to the highly subjective process of scoring by human raters. Another NLP application developed by us was the automatic conversion of thousands of questions written in American English into British English, allowing their export to British institutions without requiring extensive manual conversion and thus saving significant resources.

Researchers within the UOA worked with representatives from NBME (Peter Baldwin, Janet Mee, Polina Harik, Brian Clauser), with Yaneva being jointly employed by both, and based at NBME. Ha, Evans, Karamanis, and three other members of staff also received continuous full-time funding for their research from NBME.

3. References to the research

All research outputs in this section have been peer-reviewed. R2 and R3 have been extensively cited (217 and 232 citations, respectively, as of January 2021) and R6 received the Ambassador Paper Award at the *NLP for Building Educational Applications workshop* (BEA 2019).

R1. Ha, L. A. and Yaneva, V. (2018) Automatic Distractor Suggestion for Multiple-Choice Tests Using Concept Embeddings and Information Retrieval. *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)* held in conjunction with NAACL, New Orleans, USA, 5 June 2018 (https://www.aclweb.org/anthology/W18-0548/).

R2. Mitkov, R., Ha, L. A., and Karamanis, N. (2006) A Computer-aided Environment for Generating Multiple-choice Test Items. *Journal of Natural Language Engineering*, 12(**2**), pp.177-194 (http://doi.org/10.1017/S1351324906004177).

R3. Mitkov, R. and Ha, L. A. (2003) Computer-aided Generation of Multiple-choice Tests. *Proceedings of the HLT-NAACL 03 workshop on Building Educational Applications Using Natural Language Processing,* Volume 2, pp. 17–22. Edmonton, Canada, 2003 (https://www.aclweb.org/anthology/W03-0203/).

R4. Baldwin, P., Yaneva, V., Ha, L.A., Mee, J., and Clauser, B. (2020) Using Natural Language Processing to Predict Item Response Times and Improve Test Construction. *Journal of Educational Measurement* (http://doi.org/10.1111/jedm.12264) (REF 2 Output).

R5. Yaneva, V., Ha, L. A., Baldwin, P., and Mee, J. (2020) Predicting Item Survival for Multiple Choice Questions in a High-stakes Medical Exam. *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 6812-6818. Marseille, France, 13-15 May, 2020 (https://www.aclweb.org/anthology/2020.lrec-1.841/).

R6. Ha, L. A., Yaneva, V., Baldwin, P., and Mee, J. (2019) Predicting the Difficulty of Multiple-Choice Questions in a High-stakes Medical Exam. *Proceedings of the 14th Workshop on Innovative Use of NLP for Building Educational Applications (BEA),* held in conjunction with ACL 2019, Florence, Italy, 2 August, 2019 (https://www.aclweb.org/anthology/W19-4402/).

4. Details of the impact

The principal beneficiaries of this research are NBME, predominantly through the improvement of the service that they provide [F1, F2], and the improvement of their workforce planning and research agenda [I2]. Through its adoption of NLP technology in strategic planning for exam development and scoring, NBME influences how well the knowledge of medical students is measured to protect the public and ensure a high standard of medical care in the USA and Canada. Therefore, research undertaken by the UOA has impacted a) public services [I1], and b) practitioners and the delivery of professional service [I2].

The principal benefit of the research is the improvement in fairness of the examination. This is achieved in several ways. For example, a well-known source of bias in educational measurement is the bias of human raters, who are prone to subjectivity, fatigue, misinterpretation of guidelines, and other factors. Introducing an automated way to score examinee writing is one way to reduce this bias by providing a consistent measure against which human ratings could be compared. In cases where there is high reliability, the system can be used to provide the score, thus saving resources. Another example includes the benefit of being able to predict how long certain test items would take to answer [F2]. This way, when different groups of examinees see two different sets of test items, one group would not be disadvantaged compared to the other simply because they were assigned more time-consuming items. Given the high-stakes nature of the exam, the outcome of such potential unfairness can have serious consequences for the career development of the test-takers, as well as the safety of the public.

The pathway for achieving this impact is the close working relationship between the NBME and the UOA. In addition to Yaneva being a staff member in both institutions (employed by NBME after completing a PhD and post-doc at RGCL), an annual meeting between members of the two groups sets the research agenda for the following year. Additionally, RGCL sends a monthly research report to the NBME. So far three jointly authored outputs have been published (on both F1 and F2), with several others in preparation.

I1. Impacts on public services

NBME has adopted several new technologies developed by RGCL (Section 2), which have improved the quality and efficiency of their public service [C1, C2, C3]:
Overall, the *"use of NLP enables NBME to apply the best available technology to ensure accuracy and validity for all our examination programs."* [C2]. Examples of this include the automated scoring of student writing, where human bias is reduced through the introduction of an objective way to mark the text written by examinees during NBME's assessment of Clinical Skills (CS) exam, making it a fairer assessment of their skills: "*The use of computer-assisted scoring in Step 2 CS is expected to enhance the quality and efficiency of assessment."* Other applications such as the automated generation of test items are regarded as ways to *"increase the efficiency and quality of the item writing process."* [C2]. NBME acknowledges that *"These are all long-term projects pursued by NBME as a direct result from the pilot studies developed by RGCL."* [C1]

I2. Impacts on practitioners and delivery of professional services

Further to the improvement of the examinations developed by NBME [F1, F2], the collaboration between the two institutions has had a strong impact on NBME's workforce planning and research agenda. In particular, using NLP has enabled NBME to establish itself as an NLP leader among educational testing organisations [C1], hiring NLP professionals to develop its own in-house projects.

- Workforce planning has been influenced by research: While the typical background of NBME researchers is in psychometrics and educational measurement, in 2018, NBME hired its first staff member with a background in NLP (Yaneva). Yaneva is currently jointly employed at both NBME (Philadelphia, USA) and the University of Wolverhampton.
- Funding of NBME research interns on the topic of NLP in educational measurement: Since 2018, NBME has funded annual summer internships in NLP-related projects. For example, the 2018 research intern (Kang Xue – see C4) was mentored by Yaneva and focused his research on predicting multiple-choice question difficulty [C4].
- Improvements in practice**:** As a result of the collaboration with us, NBME has made improvements to its development as a leading brand within the field of educational testing, as stated in [C1]:

"The research contributions of the team prompted an evidence-based strategic change for the organization towards embracing NLP-related technological advances, learning to conduct NLP research in-house, and publishing research findings at NLP venues. Not only has this shift helped improve the quality of our assessment and its competitiveness, as discussed above, but it has also shaped an ambition for NBME to become a leader in NLP-based solutions within the assessment community."

In addition to the above, NBME's recent strategic change to a stronger focus on NLP is also evidenced by the planning of a professional conference on the topic of NLP in Assessment (to be held on 31/10–01/11, 2021 in Philadelphia, USA) [C1], which intends to bring together practitioners and researchers from the fields of NLP and educational measurements and establish NBME as a leader in the field of NLP for educational testing. The event has a planned capacity of 120 attendees; 18 renowned international speakers are scheduled to deliver presentations. The speakers include both international academic staff and experts in resolving practical issues for the implementation of NLP systems in real-life assessment practice.

5. Sources to corroborate the impact (indicative maximum of 10 references)

C1. A Letter of Support provided by senior NBME members corroborating the RGCL's research being implemented at NBME.

C2. NBME Annual Report for 2019, Section 6, p.12. (Available at:

https://www.nbme.org/sites/default/files/2020-03/NBME_Annual_Report_2019.pdf)

C3. NBME Annual Report for 2018, p.44. (Available at: https://nbme.org/sites/default/files/2020-07/NBME_Annual_Report_2018.pdf)

C4. Xue, K., Yaneva, V., Runyon, C., & Baldwin, P. (2020). Predicting the Difficulty and Response Time of Multiple-Choice Questions Using Transfer Learning. In Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications (pp. 193-197). (Available at: https://www.aclweb.org/anthology/2020.bea-1.pdf#page=207)