Avoid playing learner and system off against each other

Ji-Ung Lee, Christian M. Meyer, Iryna Gurevych

Ubiquitous Knowledge Processing (UKP) Lab and Research Training Group AIPHES
Technische Universität Darmstadt, Germany
https://www.ukp.tu-darmstadt.de

Digital learning platforms for self-directed learning, such as Doulingo[1] gained increasing popularity in recent years. A major challenge for such platforms is providing learning material that fits the needs and proficiency of a particular learner – a task that has been done by human teachers in the traditional classroom setting and that must now be fully automatic in order to scale up the platform and enable learners to learn at their own pace whilst receiving immediate feedback on their inputs. Therefore, there has been ample research on *automated exercise generation* (Mitkov et al., 2006; Chinkina and Meurers, 2017) and *automated difficulty prediction* (Beinborn et al., 2014; Pilán et al., 2016) using machine learning (ML).

However, introducing ML and artificial intelligence in general into the learning process raises two important ethical issues: *i)* Systems may fail to recognize correctly given answers, or even worse, suggest wrong answers, and *ii)* they provide learners with unsuitable (i.e. too easy or too difficult) exercises outside their *Zone of Proximal Development* (Vygotsky, 1978). Both issues may severely harm the learning progress. As Hovy and Spruit (2016) point out, there is yet little work on mitigating such issues in our community.

Due to the scarcity of available training data, researchers increasingly rely on crowdsourcing (Heffernan et al., 2016) or active ML techniques (Zesch et al., 2015) to overcome the so-called cold-start problem of ML. These approaches are especially problematic, since they solely aim at improving the system and its underlying ML model – at the cost of the learning goals of the users. Learners are reduced to cheap labelers suffering from incorrect system feedback and varying task difficulty (cf. Settles et al., 2008). This demotivates learners, reduces their learning speed, and might even yield misconceptions.

In our work, we explore these issues for a language learning use case: *i)* We use *automatically generated C-tests* (Klein-Braley and Raatz, 1982), which have a very small solution space and thus, prevent incorrect system responses. *ii)* We integrate the learner's goal into the active ML objective of the *automated difficulty prediction*. To this end, we can jointly optimize for the learner's goal of quickly reaching their next proficiency level and the system's goal of reliably estimating exercise difficulty at high accuracy. This will contribute to learning platforms that effectively support learners without the necessity of vastly existing training data and without treating the learners as mere data labelers.

[1] https://www.duolingo.com

Acknowledgments. This work has been supported by the Hessian research excellence program "Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz" (LOEWE) as part of the *a! - automated language instruction* project under grant No. 521/17-03 and by the German Research Foundation as part of the Research Training Group "Adaptive Preparation of Information from Heterogeneous Sources" (AIPHES) under grant No. GRK 1994/1.

References

Beinborn, L., Zesch, T., and Gurevych, I. (2014). Predicting the difficulty of language proficiency tests. *Transactions of the Association of Computational Linguistics*, 2(1):517–529.

Chinkina, M. and Meurers, D. (2017). Question Generation for Language Learning: From ensuring texts are read to supporting learning. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, Copenhagen, Denmark.

Heffernan, N. T., Ostrow, K. S., Kelly, K., Selent, D., Van Inwegen, E. G., Xiong, X., and Williams, J. J. (2016). The future of adaptive learning: Does the crowd hold the key? *International Journal of Artificial Intelligence in Education*, 26(2):615–644.

Hovy, D. and Spruit, S. L. (2016). The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany.

Klein-Braley, C. and Raatz, U. (1982). Der C-Test: ein neuer Ansatz zur Messung allgemeiner Sprachbeherrschung. *AKS-Rundbrief*, 4:23–37.

Mitkov, R., An Ha, L., and Karamanis, N. (2006). A computer-aided environment for generating multiple-choice test items. *Natural Language Engineering*, 12(2):177–194.

Pilán, I., Volodina, E., and Zesch, T. (2016). Predicting proficiency levels in learner writings by transferring a linguistic complexity model from expert-written coursebooks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2101–2111, Osaka, Japan.

Settles, B., Craven, M., and Friedland, L. (2008). Active learning with real annotation costs. In *Proceedings of the NIPS workshop on cost-sensitive learning*, pages 1–10. Vancouver, Canada.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge: Harvard University Press.

Zesch, T., Heilman, M., and Cahill, A. (2015). Reducing annotation efforts in supervised short answer scoring. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 124–132, Denver, CO, USA.