Leaf: Multiple-Choice Question Generation

Kristiyan Vachev[1], Momchil Hardalov[1], Georgi Karadzhov[2], Georgi Georgiev[3],
Ivan Koychev[1], and Preslav Nakov[4]

[1] FMI, Sofia University "St. Kliment Ohridski", Bulgaria
[2] University of Cambridge, UK
[3] Releva.ai, Bulgaria
[4] Qatar Computing Research Institute, HBKU, Qatar

Abstract. Testing with quiz questions has proven to be an effective way to assess and improve the educational process. However, manually creating quizzes is tedious and time-consuming. To address this challenge, we present Leaf, a system for generating multiple-choice questions from factual text. In addition to being very well suited for the classroom, Leaf could also be used in an industrial setting, e.g., to facilitate onboarding and knowledge sharing, or as a component of chatbots, question answering systems, or Massive Open Online Courses (MOOCs). The code and the demo are available on GitHub.[5]

Keywords: Multiple-choice questions, education, self-assessment, MOOCs.

1 Introduction

Massive Open Online Courses (MOOCs) have revolutionized education by offering a wide range of educational and professional training. However, an important issue in such a MOOC setup is to ensure an efficient student examination setup. Testing with quiz questions has proven to be an effective tool, which can help both learning and student retention [38]. Yet, preparing such questions is a tedious and time-consuming task, which can take up to 50% of an instructor's time [41], especially when a large number of questions are needed in order to prevent students from memorizing and/or leaking the answers.

To address this issue, we present an automated multiple-choice question generation system with focus on educational text. Taking the course text as an input, the system creates question–answer pairs together with additional incorrect options (distractors). It is very well suited for a classroom setting, and the generated questions could also be used for self-assessment and for knowledge gap detection, thus allowing instructors to adapt their course material accordingly. It can also be applied in industry, e.g., to produce questions to enhance the process of onboarding, to enrich the contents of massive open online courses (MOOCs), or to generate data to train question–answering systems [10] or chatbots [22].

---

[5] https://github.com/KristiyanVachev/Leaf-Question-Generation

2   Related Work

While Question Generation is not as popular as the related task of Question Answering, there has been a steady increase in the number of publications in this area in recent years [1,18]. Traditionally, rules and templates have been used to generate questions [29]; however, with the rise in popularity of deep neural networks, there was a shift towards using recurrent encored–decoder architectures [2,8,9,33,40,46,47] and large-scale Transformers [7,20,23,27,36].

The task is often formulated as one of generating a question given a target answer and a document as an input. Datasets such as SQuAD1.1 [37] and NewsQA [44] are most commonly used for training, and the results are typically evaluated using measures such as BLEU [32], ROUGE [25], and METEOR [21]. Note that this task formulation requires the target answer to be provided beforehand, which may not be practical for real-world situations. To get over this limitation, some systems extract all nouns and named entities from the input text as target answers, while other systems train a classifier to label all word n-grams from the text and to pick the ones with the highest probability to be answers [45]. To create context-related wrong options (i.e., distractors), typically the RACE dataset [19] has been used along with beam search [3,11,31]. Note that MOOCs pose additional challenges as they often cover specialized content that goes beyond knowledge found in Wikipedia, and can be offered in many languages; there are some open datasets that offer such kinds of questions in English [5,6,19,28,42] and in other languages [4,12,13,15,16,24,26,30].

Various practical systems have been developed for question generation. Web-Experimenter [14] generates Cloze-style questions for English proficiency testing. AnswerQuest [39] generates questions for better use in Question Answering systems, and SQUASH [17] decomposes larger articles into paragraphs and generates a text comprehension question for each one; however, both systems lack the ability to generate distractors. There are also online services tailored to teachers. For example, Quillionz [35] takes longer educational texts and generates questions according to a user-selected domain, while Questgen [34] can work with texts up to 500 words long. While these systems offer useful question recommendations, they also require paid licenses. Our Leaf system offers a similar functionality, but is free and open-source, and can generate high-quality distractors. It is trained on publicly available data, and we are releasing our training scripts, thus allowing anybody to adapt the system to their own data.

3   System

System architecture: Leaf has three main modules as shown in Figure 1. Using the *Client*, an instructor inputs a required number of questions and her educational text. The text is then passed through a REST API to the *Multiple-Choice Question (MCQ) Generator Module*, which performs pre-processing and then generates and returns the required number of question–answer pairs with distractors. To achieve higher flexibility and abstraction, the models implement an interface that allows them to be easily replaced.
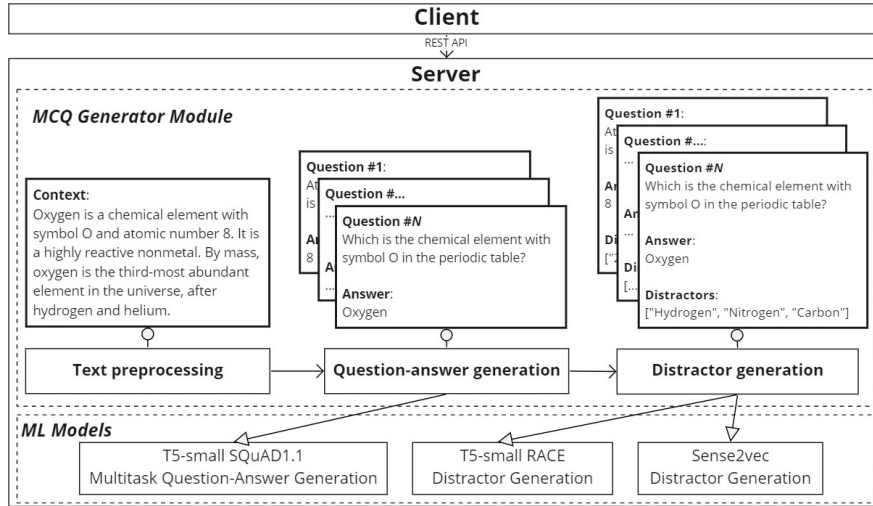
Fig. 1. The general architecture of Leaf.

Question and Answer Generation: To create the question–answer pairs, we combined the two tasks into a single multi-task model. We fine-tuned the small version of the T5 Transformer, which has 220M parameters, and we used the SQuAD1.1 dataset [37], which includes 100,000 question–answer pairs. We trained the model to output the question and the answer and to accept the passage and the answer with a 30% probability for the answer to be replaced by the `[MASK]` token. This allows us to generate an answer for the input question by providing the `[MASK]` token instead of the target answer. We trained the model for five epochs, and we achieved the best validation cross-entropy loss of 1.17 in the fourth epoch. We used a learning rate of 0.0001, a batch size of 16, and a source and a target maximum token lengths of 300 and 80, respectively. For question generation, we used the same data split and evaluation scripts as in [9]. For answer generation, we trained on the modified SQuAD1.1 Question Answering dataset as proposed in our previous work [45], achieving an Exact Match of 41.51 and an F1 score of 53.26 on the development set.

Distractor Generation: To create contextual distractors for the question–answer pairs, we used the RACE dataset [19] and the small pre-trained T5 model. We provided the question, the answer, and the context as an input, and obtained three distractors separated by a `[SEP]` token as an output. We trained the model for five epochs, achieving a validation cross-entropy loss of 2.19. We used a learning rate of 0.0001, a batch size of 16, and a source and a target maximum token lengths of 512 and 64, respectively. The first, the second, and the third distractor had BLEU1 scores of 46.37, 32.19, and 34.47, respectively. We further extended the variety of distractors with context-independent proposals, using sense2vec [43] to generate words or multi-word phrases that are semantically similar to the answer.
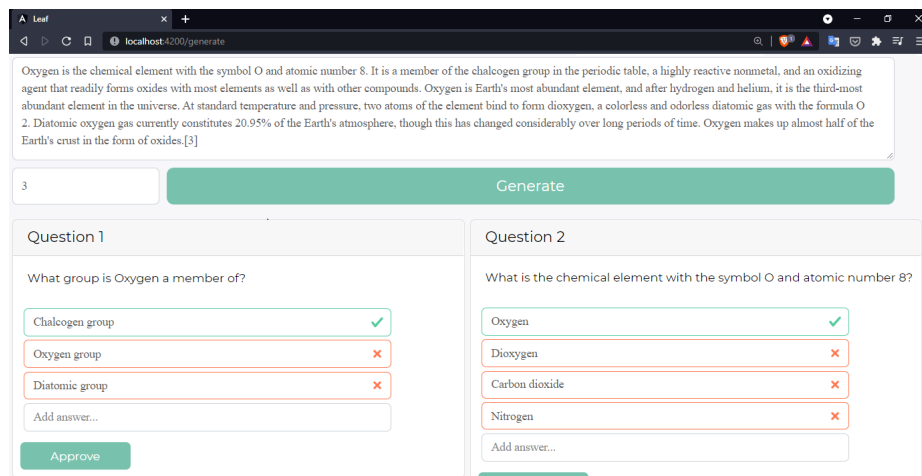
Fig. 2. Screenshot of Leaf showing the generated questions for a passage from the Wikipedia article on Oxygen. All distractors in *Question 1* are generated by the T5 model, and the last two distractors in *Question 2* are generated by the sense2vec model.

User Interface: Using the user interface shown on Figure 2, the instructor can input her educational text, together with the desired number of questions to generate. Then, she can choose some of them, and potentially edit them, before using them as part of her course.

4   Conclusion and Future Work

We presented Leaf, a system to generate multiple-choice questions from text. The system can be used both in the classroom and in an industrial setting to detect knowledge gaps or as a self-assessment tool; it could also be integrated as part of other systems. With the aim to enable a better educational process, especially in the context of MOOCs, we open-source the project, including all training scripts and documentation.

In future work, we plan to experiment with a variety of larger pre-trained Transformers as the underlying model. We further plan to train on additional data. Given the lack of datasets created specifically for the task of Question Generation, we plan to produce a new dataset by using Leaf in real university courses and then collecting and manually curating the question–answer pairs Leaf generates over time.

Acknowledgements

References

1. Jacopo Amidei, Paul Piwek, and Alistair Willis. Evaluation methodologies in automatic question generation 2013-2018. In *Proceedings of the 11th International Conference on Natural Language Generation*, INLG '20, pages 307–317, Tilburg University, The Netherlands, 2018. Association for Computational Linguistics.

2. Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Jianfeng Gao, Songhao Piao, Ming Zhou, and Hsiao-Wuen Hon. UniLMv2: Pseudo-masked language models for unified language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *ICML '20*, pages 642–652. PMLR, 2020.

3. Ho-Lam Chung, Ying-Hong Chan, and Yao-Chung Fan. A BERT-based distractor generation scheme with multi-tasking and negative answer training strategies. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4390–4400. Association for Computational Linguistics, 2020.

4. Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470, 2020.

5. Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? Try ARC, the AI2 Reasoning Challenge. *arXiv:1803.05457*, 2018.

6. Peter Clark, Oren Etzioni, Tushar Khot, Daniel Khashabi, Bhavana Dalvi Mishra, Kyle Richardson, Ashish Sabharwal, Carissa Schoenick, Oyvind Tafjord, Niket Tandon, Sumithra Bhakthavatsalam, Dirk Groeneveld, Michal Guerquin, and Michael Schmitz. From 'F' to 'A' on the N.Y. Regents Science Exams: An overview of the Aristo project. *AI Mag.*, 41(4):39–53, 2020.

7. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '19, pages 4171–4186, Minneapolis, Minnesota, USA, 2019. Association for Computational Linguistics.

8. Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019*, NeurIPS '19, pages 13042–13054, Vancouver, British Columbia, Canada, 2019.

9. Xinya Du, Junru Shao, and Claire Cardie. Learning to ask: Neural question generation for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, ACL '17, pages 1342–1352, Vancouver, Canada, 2017. Association for Computational Linguistics.

10. Nan Duan, Duyu Tang, Peng Chen, and Ming Zhou. Question generation for question answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, EMNLP '17, pages 866–874, Copenhagen, Denmark, 2017. Association for Computational Linguistics.

11. Yifan Gao, Lidong Bing, Piji Li, Irwin King, and Michael R. Lyu. Generating distractors for reading comprehension questions from real examinations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33 of *AAAI '19*, pages 6423–6430, 2019.

12. Momchil Hardalov, Ivan Koychev, and Preslav Nakov. Beyond English-only reading comprehension: Experiments in zero-shot multilingual transfer for Bulgarian. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, RANLP 19, pages 447–459, Varna, Bulgaria, 2019. INCOMA Ltd.

13. Momchil Hardalov, Todor Mihaylov, Dimitrina Zlatkova, Yoan Dinkov, Ivan Koychev, and Preslav Nakov. EXAMS: A multi-subject high school examinations dataset for cross-lingual and multilingual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, EMNLP '20, pages 5427–5444. Association for Computational Linguistics, 2020.

14. Ayako Hoshino and Hiroshi Nakagawa. WebExperimenter for multiple-choice question generation. In *Proceedings of HLT/EMNLP 2005 Interactive Demonstrations*, HLT/EMNLP '05, pages 18–19, Vancouver, British Columbia, Canada, 2005. Association for Computational Linguistics.

15. Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *ICML '20*, pages 4411–4421. PMLR, 2020.

16. Yimin Jing, Deyi Xiong, and Zhen Yan. BiPaR: A bilingual parallel dataset for multilingual and cross-lingual reading comprehension on novels. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, EMNLP-IJCNLP '19, pages 2452–2462, Hong Kong, China, 2019. Association for Computational Linguistics.

17. Kalpesh Krishna and Mohit Iyyer. Generating question-answer hierarchies. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, ACL '19, pages 2321–2334, Florence, Italy, 2019. Association for Computational Linguistics.

18. Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30, 2019.

19. Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, EMNLP '17, pages 785–794, Copenhagen, Denmark, 2017. Association for Computational Linguistics.

20. Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. In *Proceedigs of the 8th International Conference on Learning Representations*, ICLR '20, Addis Ababa, Ethiopia, 2020. OpenReview.net.

21. Alon Lavie and Abhaya Agarwal. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, WMT '07, pages 228–231, Prague, Czech Republic, 2007. Association for Computational Linguistics.

22. John Lee, Baikun Liang, and Haley Fong. Restatement and question generation for counsellor chatbot. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 1–7. Association for Computational Linguistics, 2021.

23. Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL '20, pages 7871–7880. Association for Computational Linguistics, 2020.

24. Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL '20, pages 7315–7330. Association for Computational Linguistics, 2020.

25. Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Proceedigs of the Workshop on Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, 2004. Association for Computational Linguistics.

26. Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. Few-shot learning with multilingual language models. *arXiv:2112.10668*, 2021.

27. Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv:1907.11692*, 2019.

28. Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? A new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, EMNLP '18, pages 2381–2391, Brussels, Belgium, 2018. Association for Computational Linguistics.

29. Ruslan Mitkov and Le An Ha. Computer-aided generation of multiple-choice tests. In *Proceedings of the HLT-NAACL 03 Workshop on Building Educational Applications Using Natural Language Processing*, BEA '03, pages 17–22, Edmonton, Alberta, Canada, 2003.

30. Kiet Van Nguyen, Khiem Vinh Tran, Son T. Luu, Anh Gia-Tuan Nguyen, and Ngan Luu-Thuy Nguyen. Enhancing lexical-based approach with external knowledge for Vietnamese multiple-choice machine reading comprehension. *IEEE Access*, 8:201404–201417, 2020.

31. Jeroen Offerijns, Suzan Verberne, and Tessa Verhoef. Better distractions: Transformer-based distractor generation and multiple choice question filtering. *arXiv:2010.09598*, 2020.

32. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, ACL '02, pages 311–318, Philadelphia, Pennsylvania, USA, 2002. Association for Computational Linguistics.

33. Weizhen Qi, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. ProphetNet: Predicting future n-gram for sequence-to-sequence pre-training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2401–2410. Association for Computational Linguistics, 2020.

34. Questgen. Questgen: AI powered question generator. `http://questgen.ai/`. Accessed: 2022-01-05.

35. Quillionz. Quillionz - world's first AI-powered question generator. `https://www.quillionz.com/`. Accessed: 2022-01-05.

36. Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits

of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

37. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, EMNLP '16, pages 2383–2392, Austin, Texas, USA, 2016. Association for Computational Linguistics.

38. Henry L. Roediger III, Adam L. Putnam, and Megan A. Smith. Chapter one - ten benefits of testing and their applications to educational practice. In *Psychology of Learning and Motivation*, volume 55, pages 1–36. Academic Press, 2011.

39. Melissa Roemmele, Deep Sidhpura, Steve DeNeefe, and Ling Tsou. AnswerQuest: A system for generating question-answer items from multi-paragraph documents. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, EACL '21, pages 40–52, Online, 2021. Association for Computational Linguistics.

40. Linfeng Song, Zhiguo Wang, Wael Hamza, Yue Zhang, and Daniel Gildea. Leveraging context information for natural question generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '18, pages 569–574, New Orleans, Louisiana, USA, 2018. Association for Computational Linguistics.

41. Yuni Susanti, Takenobu Tokunaga, Hitoshi Nishikawa, and Hiroyuki Obari. Evaluation of automatically generated english vocabulary questions. *Research and practice in technology enhanced learning*, 12(1):1–21, 2017.

42. Oyvind Tafjord, Peter Clark, Matt Gardner, Wen-tau Yih, and Ashish Sabharwal. Quarel: A dataset and models for answering questions about qualitative relationships. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33 of *AAAI '19*, pages 7063–7071, 2019.

43. Andrew Trask, Phil Michalak, and John Liu. sense2vec - a fast and accurate method for word sense disambiguation in neural word embeddings. *arXiv:1511.06388*, 2015.

44. Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, RepL4NLP '17, pages 191–200, Vancouver, Canada, 2017. Association for Computational Linguistics.

45. Kristiyan Vachev, Momchil Hardalov, Georgi Karadzhov, Georgi Georgiev, Ivan Koychev, and Preslav Nakov. Generating answer candidates for quizzes and answer-aware question generators. In *Proceedings of the Student Research Workshop Associated with RANLP 2021*, RANLP '21, pages 203–209. INCOMA Ltd., 2021.

46. Dongling Xiao, Han Zhang, Yu-Kun Li, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. ERNIE-GEN: an enhanced multi-flow pre-training and fine-tuning framework for natural language generation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI '20, pages 3997–4003. ijcai.org, 2020.

47. Qingyu Zhou, Nan Yang, Furu Wei, Chuanqi Tan, Hangbo Bao, and Ming Zhou. Neural question generation from text: A preliminary study. In *Natural Language Processing and Chinese Computing*, pages 662–671, Cham, 2018. Springer International Publishing.