Research Proposal: Automatic construction of student misconception ontology from online forum data

Candidate Number: 10140, Seminar Group 6, Word Count: 1636 words

January 15, 2019

Rationale for Study

Background

The Fourth Industrial Revolution - following the revolutions of steam, electricity and electronics - defines a present phenomenon whereby almost all aspects of the human experience are undergoing rapid disruption, and will continue to do so exponentially (Schwab, 2016). There is no doubt that such a disruption can have positive effects, since technological advances in the last 200 years have clearly played a central role in the (often overlooked) positive global developments outlined by Rosling, Rönnlund and Rosling (2018). What should be cause for concern however, is the potential for negative consequences that result from a lack of understanding and guidance of developing technologies moving forward.

One such consequence is the Fourth Industrial Revolution's effect on global inequality. McAfee and Brynjolfsson (2014) point out that, due mainly to automation and a predisposition for technical skills, the rapid change of global labour markets may aggravate worldwide inequality. What is clear is that individuals around the globe who lack basic infrastructure (nutrition, shelter, electricity, internet, literacy) will struggle to keep up as the rest of the world sprints ahead. The question therefore is how we can ensure that the developing world participates in the Fourth Industrial Revolution.

I believe that the most promising solution to this question is to address global education, precisely by capitalising on the potential that the Fourth Industrial Revolution has to disseminate decentralised learning. I believe this is absolutely crucial to addressing inter- and intra-inequality worldwide - my long term goal therefore is to explore how we can transfer valuable insights from the interaction of education and technology to developing countries in need of educational reform. To address this overarching problem, my starting point is exploring the extent to which we can retrieve information on students' misconceptions[1] online.

Research Question

Put explicitly, the research question is "To what extent can we map students' misconceptions from online data?" One avenue to address this question is to use online forum data containing students' requests for assistance on courses that they are physically or virtually enrolled in. I intend to apply exploratory **Natural Language Processing (NLP)** to a Massive Open Online Course (MOOC) that is free at the point of use[2] (and thus open to any student with an internet connection) to extract students' most common misconceptions. Furthermore, I aim to ensure that the model (1) is scalable to support continuous analysis of additional data (i.e. can handle streaming data as well as generalising to unseen datasets), and (2) **enables applications of more sophisticated NLP methods.**

Introduction to the Literature

The disruptive changes of what the literature has coined Technology Enhanced Learning (TEL) appear to only recently have been documented (Christensen *et al.*, 2011; Dellarocas and Van Alstyne, 2013; Lucas, 2014). Researchers also predicts that technology-based learning platforms and tools will continue to enhance education through adaptive learning and "recommender systems" (Brusilovsky and Nejdl, 2004; Graf *et al.*, 2012; Peña-Ayala, 2013).

Drachsler *et al.* (2015) have done an extensive overview of recommender systems to support learning and as seen in the results of Bauman and Tuzhilin (2018), remedial recommender systems improve student performance. Crucially however, what recommender systems and many other TEL tools lack is educator participation. I believe this is a problem, because as Hanushek (2010) illustrates, great teachers substantially promote student learning. The main difference between my research problem and many TEL interventions therefore, is that

[1] This could relate to the questions and problems with learning material that students are encountering, or it could be assertions that students make that are false. The latter is defintion would be a much harder challenge.

[2] Being free is nice but unnecessary for this research. Can argue this as long as it generalizes to these environments. Must include in an 'Impact of Research' section.

I want to ensure that educators are included and empowered, because I believe this adds more value than just empowering students alone[3].

Most relevant to the work is the investigation by Wise, Zhao and Hausknecht (2013) on how students contribute and attend to messages of others in online discussions[4]. They present a pedagogical model which translates findings to guide practice for students and educators in assessing online discussion participation. My research approach differs from Wise, Zhao and Hausknecht (2013) in that it is a simplified quantitative exploration of basic question data and emphasises applicability to a diverse range of datasets rather than investigating student learning with a particular, rich dataset. My goal of quantifying large online question data, with scalability for additional data and NLP methods, is a novel contribution as this has not been completed and thus represents a gap for research.

Data Collection Strategies

The data collection strategy necessitates either access to, or "scraping[5]" of a MOOC forum containing questions posted in natural language format. This dataset need only consist of questions, with answers/replies and attributional data relating to users constituting additional information useful for descriptive analysis.

The population of interest are students that actively request help on certain concept(s), in a given MOOC. It should be noted that this definition excludes students who participate in a course but don't need assistance, as well as students who need help but don't request it. Clearly the latter group of students also experience misconceptions but since we cannot measure this, they are excluded from the population - the result being that findings will only be valid for students that request assistance online.

Conveniently, the availability of cloud computing services like AWS and Google Cloud ensures that dataset size is not a restraint for analysis - this therefore allows us to choose from the largest and most popular online courses (given no data restrictions) to gather and analyse as much data as possible. In this way, it is also possible to survey the entire population of a given MOOC, promoting validity of the results for the MOOC population and eliminating many idiosyncrasies regarding sample selection. Another problem that analysis of the full population helps address is the possibility that some students are not able to communicate

[3] Need teacher in the loop reference

[4] Also need to look at Wang 2016: Towards triggering higher-order thinking behaviours in MOOCs.

[5] As in (ethically) extracting data from online web pages manually.

their misconceptions fully or clearly. The assumption here is that most of the population is able to ask coherent questions from which at least one concept can be extracted.

Data Analysis Strategies and Outcomes

Owing to the data analysis falling under the domain of NLP, there is a strong temptation to begin immediately employing sophisticated (and exciting) unsupervised NLP techniques to quantify the data. What is crucial however, is that the entire research approach hinges on accurately capturing nuanced, implicit and possibly abstract misconceptions inherent in students' online questions. Consequently, the approach chosen for this proposal is to employ trivial NLP where language isn't being processed in essence, but rather enumerated[6].

With the assumption that the majority of students are able to coherently express at least one concept in their questions, my data analysis approach is simply to merge the textual data together and to calculate the frequencies of key words. This approach is sufficient to yield the anticipated outcome of an interpretable, visual "heat-map" of the most common misconceptions among students in a specific course[7].

The end goal of the process is to ensure that the workflow is scalable and expandable to include additional (large) datasets, as well as to serve as a stepping stone for more advanced methods for text corpora classification[8] to better ascertain student misconceptions. Thus the impact and value of this exercise is not only that further research can employ more sophisticated NLP tools to gain insights from different datasets, but also that educators will be empowered to allocate resources and implement interventions more intentionally and efficiently (by addressing student misconceptions).

[6] Hence why the NLP literature has not been surveyed in this proposal.

[7] Need to thinkin about word embeddings/parsers like Princetons Wordnet

[8] For examples see Latent Dirichlet Allocation and Bag of Tricks discussed by Blei *et al.* (2003) and Joulin *et al.* (2016) respectively.

Limitations and Further Research

Limitations

As previously discussed, one limitation is that the population of interest omits students who don't seek assistance online, thus excluding information about misconceptions solved by learning processes externally. The reader is reminded though, that the my research question does not pertain to how students learn, but rather to mapping what students are struggling to learn. I have also mentioned the limitation regarding levels of coherence in student questions (i.e. how many students don't know how to ask questions about what they don't know), but I do believe that the assumption of most students being able to express at least one misconception is realistic.

In terms of ethical limitations, careful consideration must be devoted to the protection of students' data and anonymity if their identities are traceable from the data and this data must also be sourced ethically. Although risk of harm does not appear to be an issue since no interventions will be implemented in this study, where and how educators intervene from the findings in this research is of equal, if not more, importance.

Another limitation is that the quantitative tallying of concepts is a relatively simplistic analysis - but as has been mentioned, the purpose of this project is to serve as a stepping stone for further exploratory and confirmatory analysis. Regardless of the use of sophisticated methodologies however, NLP does present a number of hurdles. Since the data will be in question format from students, some challenges in the data-processing stage include misspelling, differences in American/British English, text-message vocabulary and text decorated in Latex/markdown code. These aspects would affect accurate frequency calculation of concepts, and thus much effort needs to be devoted to address this.

One last limitation is that the research question is limited to those who can access free online courses. This does appear to contradict my overarching goal of addressing educational reform in developing countries, where internet access (and other infrastructure) for many students may be non-existent. My view is that this research serves as an entry point for me to fully acquaint myself with TEL in the context of the Fourth Industrial Revolution, moreover, by exploring a small part of the evolving landscape of modern education, I hope to form the basis for PhD research based on addressing worldwide educational inequality.

Further Research

The immediate next step for further research is expanding the repertoire of NLP tools to extract more abstract and nuanced misconceptions from students' questions. Another extension could include qualitative real-time surveys that request students to rank concepts they are grappling with, providing an opportunity to cross-reference and confirm results. Furthermore, analysis of richer datasets with discussions, attributional data (age, gender, number of posts) and temporality would yield the ability to compare misconceptions across student characteristics, over time and across educational platforms. Owing to the ultimate scalability of this project, the sky really is the limit in terms of further research.

References

Bauman, K. and Tuzhilin, A. (2018) ‘Recommending Remedial Learning Material to Students by Filling Their Knowledge Gaps’, *MIS Quarterly*, 42(1), pp. 313–332. doi: 10.25300/MISQ/2018/13770.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003) ‘Latent Dirichlet Allocation’, *Journal of Machine Learning Research*, 3, pp. 993–1022.

Brusilovsky, P. and Nejdl, W. (2004) ‘Adaptive Hypermedia and Adaptive Web’, in Singh, M. P. (ed.) *Practical handbook of internet computing.* Boca Raton, FL: CRC Press.

Christensen, C. *et al.* (2011) ‘Disrupting College: How Disruptive Innovation Can Deliver Quality and Affordability to Postsecondary Education’, *Innosight Institute.*

Dellarocas, C. and Van Alstyne, M. (2013) ‘Money Models for MOOCs’, *Communications of the ACM*, 56(8), pp. 25–28.

Drachsler, H. *et al.* (2015) *Panorama of Recommender Systems to Support Learning.* Boston, MA: Springer, pp. 421–451.

Graf, S. *et al.* (2012) *Intelligent and Adaptive Learning Systems: Technology Enhanced Support for Learners and Teachers.* Hershey, PA: IGI Global.

Hanushek, E. A. (2010) ‘The Difference is Great Teachers’, in *Waiting for ‘superman’: How we can save america’s failing public schools*, pp. 81–102.

Joulin, A. *et al.* (2016) ‘Bag of tricks for efficient text classification’, *arXiv preprint arXiv:1607.01759.*

Lucas, H. (2014) ‘Disrupting and Transforming the University’, *Communications of the ACM*, 57(10), pp. 32–35.

McAfee, A. and Brynjolfsson, E. (2014) *The second machine age: Work, progress, and prosperity in a time of brilliant technologies.* WW Norton & Company.

Peña-Ayala, A. (ed.) (2013) *Intelligent and Adaptive Educational-Learning Systems: Achievements and Trends.* Berlin: Springer.

Rosling, H., Rönnlund, A. and Rosling, O. (2018) *Factfulness: Ten Reasons We’re Wrong* *about the World–and why Things are Better Than You Think*. Flatiron Books.

Schwab, K. (2016) 'The Fourth Industrial Revolution: what it means, how to respond'. Available at: https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-hov

Wise, A., Zhao, Y. and Hausknecht, S. (2013) 'Learning Analytics for Online Discussions: A Pedagogical Model for Intervention with Embedded and Extracted Analytics', *In Proceedings of the third international conference on learning analytics and knowledge*, ACM, pp. 48–56.