Using AI to solve business problems in scholarly publishing

Artificial intelligence (AI) tools are widely used today in many areas, and are now being introduced into scholarly publishing. This article provides a brief overview of present-day AI and machine learning as used for text-based resources such as journal articles and book chapters, and provides an example of its application to identify suitable peer reviewers for manuscript submissions. It describes how one company, UNSILO, has created a tool for this purpose, and the underlying technology used to deliver it. The article also offers a glimpse into a future where AI will profoundly change the way that academic publishing will work.

Keywords
AI; artificial intelligence; machine learning; case study; start-up; peer review

MICHAEL UPSHALL

Head of Business Development
UNSILO

Introduction

News and media articles are today full of references to artificial intelligence (AI), and indeed AI is already more pervasive than we might notice. It is in daily use around us; this is not the future, this is the present. When we open our e-mail, AI is used to remove spam from our in-box. When Amazon or Spotify suggests 'more like this', it is using AI tools to identify likely suggestions. And when we complete a form using handwritten characters and numbers, an AI tool is typically converting our handwriting to machine-readable symbols. All these capabilities come under the heading of artificial intelligence, but this broad term 'AI' actually covers a wide range of different tools and approaches.

Early approaches to AI

The tools used in AI today are something of a return to ideas that had been created many years ago, notably Bayesian reasoning (explained later). While the principle of AI has remained constant – using a machine to solve problems that humans currently solve – the approach has varied. In the last 50 years or more there have been several shifts in direction for what constitutes AI orthodoxy. For many years, AI was based around what has been named as a symbolist approach (see, for example, Wikipedia, 'Symbolic artificial intelligence'),[1] trying to use pre-existing knowledge and combine it in the form of rules to solve problems. (See, for example, Domingos.)[2] After all, this appears to be how humans reason: I know what I am looking at is a dog because it has four legs, and it barks. Yet, other animals have four legs, and wolves and foxes bark. In practice, the attempt to distil rules and procedures from observation turns out to be far more complex than anticipated. Imagine how many rules you would need to create to differentiate dogs from any other animal and achieve 100% accuracy. To create a perfect rule requires a complete knowledge of a domain, and for many tasks, including academic publishing, there is no complete knowledge available; we have to work from partial data.

'Imagine how many rules you would need to create to differentiate dogs from any other animal and achieve 100% accuracy'

Academic publishing is based on language, and researchers have struggled for years to make natural language, the language humans write and speak, intelligible to machines, by codifying it into rules. AI researchers tried to replicate the workings of the human brain in order to determine what they presumed were the universal rules of grammar. But the results were poor; however large the grammar, there always remain the exceptions of natural language, which appear to be infinite. Trying to formulate the rules by which the universe is run is as complicated as running the universe.

‘Trying to formulate the rules by which the universe is run is as complicated as running the universe’

Present-day approaches to AI

Modern AI is to a large extent based on a change in direction on the use of AI and machine-learning tools that became widespread during the last 20 years or so. It is based on a fundamental change to problem-solving. Instead of attempting to create rules based on a full knowledge of the problem, modern AI, which covers a vast range of solutions such as self-driving cars, forecasting the weather, online dating and image recognition, starts from a position of imperfect knowledge. Using what knowledge is available, it employs inference engines, which use existing data to predict new results, and combines those results with such techniques as neural networks, which ‘learn’ by assessing examples. These examples may be tagged (as when the computer is trained to identify handwritten numbers) to create a training set (as when a computer is provided in advance with 1,000 images of dogs and cats) or even, in the specific use described here, the machine trains itself: this is what is described as automatic or unsupervised concept extraction. Neural networks are frequently combined with Bayesian inference.

Using Bayesian reasoning in academic publishing

Instead of attempting to collect all the possible relevant data, which would take significant time and may in any case be impossible to achieve, Bayesians start from a position of imperfect data and estimate a probable result based on the available information. It is not surprising, therefore, that Bayesian reasoning is used to try to determine likely odds for sporting events and in gambling[3] – both cases where we are trying to predict an outcome using the available information. Bayesian reasoning provides a way of determining the probability of something when full background information is not available – such as predicting the future. It provides a remarkably wide-ranging set of solutions to many business problems. What is described here is the application of Bayesian reasoning to solve some of the common problems encountered as part of the academic researcher user journey. This involves working with academic text to provide various tools that enable the text to be ‘understood’.

‘Bayesians … estimate a probable result based on the available information’

Supervised and unsupervised

The AI discussed in this article is very different from traditional symbolist methods. Not only is it based on Bayesian reasoning, but it also makes use of unsupervised machine learning. Instead of starting with rules, unsupervised machine learning starts with nothing more than a corpus, which is just a large collection of textual content, for example book chapters or academic articles. Every word, or phrase, in a document is logged by the system, from its position within a sentence and within the document, so that words that are positioned nearby in the same sentence and within a few words can be identified. Simply by looking at the text in the context of all the other texts, a system can determine semantic information about words. For example, the phrase ‘cardiac arrest’ occurs in English language sentences in similar contexts to the phrase ‘heart attack’, and this is a powerful indicator of synonymy. Given a sufficiently large corpus of natural language, the system identifies these related meanings without any prior training in medicine, or indeed any other subject. What are the benefits of such an approach?

‘unsupervised machine learning starts with nothing more than a corpus’

One implication of this technique is that no prior subject tagging of the corpus is required. This technique does necessarily need a training set – a set of articles that have previously been coded to identify the required result. Nor is it necessary to begin by building a taxonomy, which would be the traditional approach to 'understanding' an article for identification purposes. Clearly, this is a dramatic change to more traditional approaches to coding content.

'no prior subject tagging of the corpus is required'

UNSILO and its role

UNSILO was founded in 2012 to bring AI-based solutions to the world of academic publishing, using the technology of 'unsupervised concept extraction'. Co-founder Mads Rydahl had been product director of Siri, the San Francisco-based pioneer voice-recognition company that was subsequently bought by Apple. Similar technology is in use with all the leading software companies, Amazon, Microsoft and Google, although these companies have of course a much wider range of use cases than academic publishing. General AI tools tend to have lower quality results than software that is created expressly for one sector. In fact, academic publishing is an ideal area for this new concept extraction technology. Firstly, it comprises a very precise and well-structured subset of natural language, particularly in the case of journal articles and academic monographs; academic articles typically have an abstract, a discussion, a set of assertions and a set of references, for example. Although the language type is consistent, the quantity of content is vast, in fact so great that no human researcher can keep up with the continuing flow of content; around three million new articles are published every year.[4] As academic publishing expanded dramatically during the post-war era, it required a scaling of the publishing model that many smaller publishers struggled to keep up with. Moreover, although some areas of academic publishing have existing standard classification systems (for example MeSH, the most widely used system for classifying medical articles), there is no standard classification system in use across all subjects.

'academic publishing is an ideal area for this new concept extraction technology'

UNSILO's first customer, Springer (subsequently Springer Nature), used UNSILO to identify links across all eleven million journal articles they published. (The number of articles is considerably higher today.) That use demonstrated two of the key advantages of this technology. Firstly, it is scalable, since it is able to employ both cloud storage for collection and manipulation of concepts. It takes advantage of the vast increase in computing power during the last 20 years to make it possible to analyse a corpus of some ten billion words and to capture information on every one of those words. Secondly, the subject-agnostic nature of the technology meant that a publisher such as Springer could use one tool to index all their content in a consistent way, ranging from molecular biology to business studies and economics. Finding related articles provides a quick win for the publisher: it can be delivered via an application programming interface (API) on the publisher or hosting company website, and so requires minimal work at the publisher end. It does not require any knowledge of AI to implement, and so provides an immediate benefit to users. Even though some subject areas at Springer had a taxonomy, there was no taxonomy that covered all the 25 or more subject areas in which Springer publish, hence the value of a subject-agnostic tool.

'Finding related articles provides a quick win for the publisher'

AI: just another new technology?

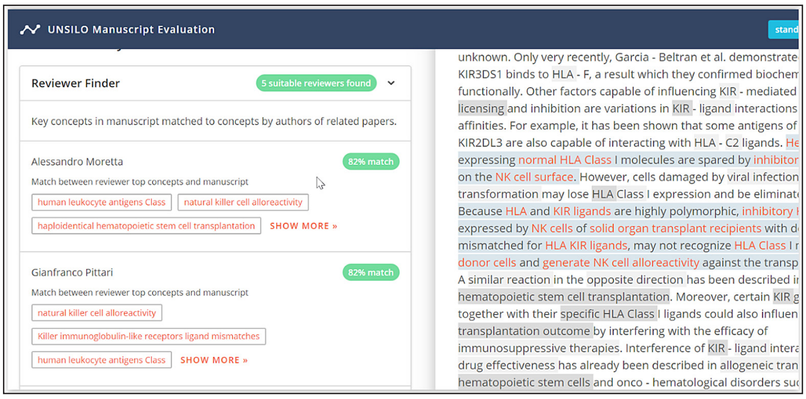
Undoubtedly, AI is a disruptive technology; it has the potential to transform many existing business processes because of its fundamentally different approach. Consider the invention of the typewriter: although it made the process of capturing text much more rapid and more efficient, it did not introduce any fundamental change in the way that an author creates content. By contrast, the AI tools described here enable some fundamental changes in the academic publishing workflow. For example, the concept extraction process can be used not just for document discovery and linking of content, but also for tasks involved in the manuscript submission process for scholarly articles. UNSILO presented at the 2018 Frankfurt Book Fair a set of automated checks (being tested with Clarivate and their Scholar One product) that help a human editor or author assess and evaluate a new submission in real time. Currently, the average time for peer review of a new article is in the region of three months.[5] In addition, it is a very labour-intensive process. Around 75% of manuscript submissions are rejected before peer review, and typically three peer reviewers will be contacted for every published article, as documented by the recent Publons report on the state of peer review.[6] While the number of peer review invitations is growing by around 10% per year, the number of accepted invitations is only growing by some 5% per year – clearly, it is a challenge finding good peer reviewers.

'AI tools … enable some fundamental changes in the academic publishing workflow'

Libraries, from the earliest times, have been based on systems for classifying content so that it can be found by users. One task that has traditionally been assumed by academic publishers is essentially an attempt to classify the content they publish in a systematic way. But the use of unsupervised concept extraction described here provides a very different way to link related content. Does this mean that all taxonomies are now irrelevant? By no means, but the choice of tool for making content discoverable may depend on what the classification system is being used for. Progress in this area of AI has been remarkable. Just a few years ago, software appeared on the market that identified subject groups, but this software required a substantial training set to be built and a taxonomy created before the tool could be used to link content. Today, the need for a taxonomy is increasingly questioned:

> 'What has moved on is the assumption that a taxonomy is required in these processes. The more recent content analysis approaches (in semantic enrichment and AI) use more statistical and grammatical analysis, rather than analysis against a taxonomy or ontology. This makes them more flexible and potentially more fine-grained in their output. It also removes the need for the upkeep of such taxonomies and ontologies. There are cases where the use of a taxonomy or ontology are still appropriate, but this should no longer be the assumed starting point.'[7]

'the need for a taxonomy is increasingly questioned'



Key concepts in manuscript matched to concepts by authors of related papers.

Alessandro Moretta

Match between reviewer top concepts and manuscript

human leukocyte antigens Class | natural killer cell alloreactivity | haploidentical hematopoietic stem cell transplantation

Gianfranco Pittari

Match between reviewer top concepts and manuscript

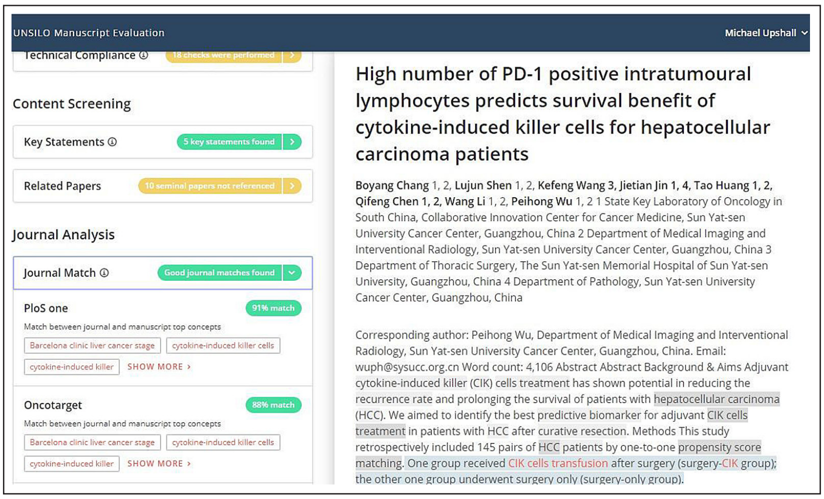
natural killer cell alloreactivity | Killer immunoglobulin-like receptors ligand mismatches | human leukocyte antigens Class

unknown. Only very recently, Garcia - Beltran et al. demonstrate
KIR3DS1 binds to HLA - F, a result which they confirmed biochem
functionally. Other factors capable of influencing KIR - mediated
licensing and inhibition are variations in KIR - ligand interactions
affinities. For example, it has been shown that some antigens of
KIR2DL3 are also capable of interacting with HLA - C2 ligands. He
expressing normal HLA Class I molecules are spared by inhibitor
on the NK cell surface. However, cells damaged by viral infection
transformation may lose HLA Class I expression and be eliminat
Because HLA and KIR ligands are highly polymorphic, inhibitory
expressed by NK cells of solid organ transplant recipients with d
mismatched for HLA KIR ligands, may not recognize HLA Class I r
donor cells and generate NK cell alloreactivity against the transp
A similar reaction in the opposite direction has been described i
hematopoietic stem cell transplantation. Moreover, certain KIR g
together with their specific HLA Class I ligands could also influen
transplantation outcome by interfering with the efficacy of
immunosuppressive therapies. Interference of KIR - ligand intera
drug effectiveness has already been described in allogeneic tran
hematopoietic stem cells and onco - hematological disorders suc

Figure 1. Example of automated peer reviewer identification

Using unsupervised concept extraction, UNSILO provides an automated peer reviewer finder (Figure 1). Using the automated concept extraction process, the system builds a profile of a submitted article and compares it with the profile of the tens of thousands of published authors – in this case, the collection of open access full-text articles and abstracts available in PubMed, over 29 million articles.[8] It then identifies the closest five matches of potential peer reviewer by creating a profile for each author based on the articles he or she has written and identifying the most relevant concepts for those articles. The human editor now has a list of suggested reviewers to contact, but there is no obligation for the editor to make use of all or any of the recommendations. In other words, the machine facilitates a human process, but in no way replaces the human input. Using the same technology, any submitted article can be matched to the most relevant of the 27,000 journals that comprise the PubMed collection (Figure 2). These are examples of how AI transforms the type of human engagement required for academic publishing. Further use cases are emerging as publishers engage more with the tools.

> 'the machine facilitates a human process, but in no way replaces the human input'



Technical Compliance ⓘ — 18 checks were performed

Content Screening

Key Statements ⓘ — 5 key statements found

Related Papers — 10 seminal papers not referenced

Journal Analysis

Journal Match ⓘ — Good journal matches found

PloS one — 91% match
Match between journal and manuscript top concepts
Barcelona clinic liver cancer stage | cytokine-induced killer cells | cytokine-induced killer

Oncotarget — 88% match
Match between journal and manuscript top concepts
Barcelona clinic liver cancer stage | cytokine-induced killer cells | cytokine-induced killer

High number of PD-1 positive intratumoural lymphocytes predicts survival benefit of cytokine-induced killer cells for hepatocellular carcinoma patients

Boyang Chang 1, 2, **Lujun Shen** 1, 2, **Kefeng Wang 3, Jietian Jin 1, 4, Tao Huang 1, 2, Qifeng Chen 1, 2, Wang Li** 1, 2, **Peihong Wu** 1, 2 1 State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Sun Yat-sen University Cancer Center, Guangzhou, China 2 Department of Medical Imaging and Interventional Radiology, Sun Yat-sen University Cancer Center, Guangzhou, China 3 Department of Thoracic Surgery, The Sun Yat-sen Memorial Hospital of Sun Yat-sen University, Guangzhou, China 4 Department of Pathology, Sun Yat-sen University Cancer Center, Guangzhou, China

Corresponding author: Peihong Wu, Department of Medical Imaging and Interventional Radiology, Sun Yat-sen University Cancer Center, Guangzhou, China. Email: wuph@sysucc.org.cn Word count: 4,106 Abstract Abstract Background & Aims Adjuvant cytokine-induced killer (CIK) cells treatment has shown potential in reducing the recurrence rate and prolonging the survival of patients with hepatocellular carcinoma (HCC). We aimed to identify the best predictive biomarker for adjuvant CIK cells treatment in patients with HCC after curative resection. Methods This study retrospectively included 145 pairs of HCC patients by one-to-one propensity score matching. One group received CIK cells transfusion after surgery (surgery-CIK group); the other one group underwent surgery only (surgery-only group).

Figure 2. Automated journal match functionality

Publisher and institutional adoption of AI tools

Until now, publishers and institutions have responded very slowly to the use of AI. There may be several reasons for this. Publishing is an industry that has been slow to innovate (even today, the take-up of XML workflow is by no means universal across the industry, and the adoption of a standard XML 'flavour' is relatively recent, whereas the use of XML in other sectors, such as the transmission of financial transactions in banking, has been a standard for many years). At the same time, the industry has been cautious about the idea of editors losing control of any aspect of the submission process. As for libraries, they also have not engaged very systematically with AI tools as described here despite the great opportunities provided by them, for example in linking an institutional repository or preprint collection to the catalogue of published materials available via the university library. As a genuinely interdisciplinary tool, unsupervised concept extraction works across subject disciplines and covers any kind of textual content, including preprints and internal documents.

Moving from technology to solutions

Clearly, technology alone does not create innovation; it requires humans who see the possibilities and who can identify business cases from that technology. That combination of technology and human knowledge of specific publishing situations requires a working knowledge of both the use and deployment of these tools. In academic libraries, just as in publishing, some shared knowledge is required for a business case to be identified and accurately assessed. One way to encourage technical innovation is by the use of simple APIs that can be integrated to existing platforms, for example to indicate related content or to find a relevant journal. These APIs require minimal technical knowledge to install but can provide valuable information on user interaction with AI tools, while being at the same time fully justified as an investment in their own right. In this way, the publisher engages with the technology without having to have a detailed knowledge of the principles of AI and machine learning.

> 'One way to encourage technical innovation is by the use of simple APIs'

Staffing implications and the role of the editor

Following the implementation of AI tools, can we therefore expect to see editorial departments reduced to zero staff in the future? Absolutely not! One of the first discoveries UNSILO made when implementing this technology is that those in control of the process want, quite rightly, to retain a level of human configurability, apart from some simple low-level tasks that can be entirely automated. An effective strategy for the implementation of machine-learning tools should respond to this request and provide a level of configurability, so a human editor or author can verify the process is delivering good results before leaving the algorithm to automate a process.

> 'machine-based systems incorporate human bias, and using only human-based tools will not remove it'

But there is another fundamental reason for including humans. They are essential because AI is not value-free; there is bias in every algorithm. Humans, however, can identify and can counteract bias. Bias is ubiquitous in any human decision-making. To give one trivial example, a peer-reviewed academic study found that in the presentation of choices to humans, there is a bias towards choosing the first option in any survey.[9] So it is not surprising that in any algorithm there is likely to be some bias. Of course, machine-based systems incorporate human bias, and using only human-based tools will not remove it – bias will still be found in the existing non-AI workflow tools currently in operation. The only way to deal with bias is to recognize, in any effective AI-based strategy, that bias from human decisions exists, and apply tools where possible to counteract it.

Summary

In this brief overview I have suggested some of the ways in which AI tools could profitably be employed by publishers and institutions. The recommendation is for editors and information specialists to get hands-on experience at the earliest opportunity, enabling them to make future decisions on the rollout of the technology based on experience rather than guesswork. The peer-review solution presented here represents just one way in which AI tools could be deployed to deliver the goal of a faster and higher-quality academic publishing workflow.

In the longer term, this AI-based technology will profoundly change the way that academic publishing works. For example, there is no reason why the submission tools discussed here could not be used direct by authors to try out their manuscripts for submission readiness. Publishers and institutions should be evaluating and implementing solutions using these new tools. The most effective implementation will come from experienced users able to identify the most effective points in the academic workflow where these tools can be introduced.

Abbreviations and Acronyms

A list of the abbreviations and acronyms used in this and other *Insight*s articles can be accessed here – click on the URL below and then select the 'full list of industry A&As' link: **http://www.uksg.org/publications#aa**

Competing interests

The author has declared no competing interests.

References

1. “Symbolic Artificial Intelligence,” *Wikipedia*, February 14, 2019: https://en.wikipedia.org/w/index.php?title=Symbolic_artificial_intelligence&oldid=883217410 (accessed March 5, 2019).
2. Pedro Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World* (London: Allen Lane, 2015).
3. Nate Silver, *The Signal and the Noise: The Art and Science of Predictio* (London: Allen Lane, 2013).
4. Rob Johnson, Anthony Watkinson and Michael Mabe, *The STM Report 2018: An Overview of Scientific and Scholarly Journal Publishing* (2018): https://www.stm-assoc.org/2018_10_04_STM_Report_2018.pdf (accessed March 5, 2019).
5. *STM Report 2018*, p5.
6. “Publons Global State Of Peer Review-2018”: https://publons.com/static/Publons-Global-State-Of-Peer-Review-2018.pdf (accessed March 5, 2019).
7. *STM Report 2018*, p164.
8. “PubMed,” *Wikipedia*: https://en.wikipedia.org/w/index.php?title=PubMed&oldid=877177825 (accessed March 5, 2019).
9. Dana R. Carney and Mahzarin R. Banaji, “First Is Best”, *PLOS ONE* 7, no. 6 (27 June 2012): e35088: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0035088 (accessed March 5, 2019).

Michael Upshall
Head of Business Development
UNSILO, DK
E-mail: michael.upshall@unsilo.com

ORCID ID: http://orcid.org/0000-0003-1115-6847

To cite this article:
Upshall M, ‘Using AI to solve business problems in scholarly publishing,’ *Insights*, 2019, 32: 13, 1–7; DOI: https://doi.org/10.1629/uksg.460

Submitted on 23 December 2018 Accepted on 11 February 2019 Published on 17 April 2019

Published by UKSG in association with Ubiquity Press.