CONTENTS

1   Information overload - the challenge
2   Machine-generated books
3   Covid-19 - From Report to App
4   TL;DR Extreme Summarization
5   Natural Language Generation
INFORMATION OVERLOAD - THE CHALLENGE
OUR BRANDS PUBLISH A LOT OF RESEARCH!
Managing the Information Overload
MANAGING THE INFORMATION OVERLOAD
Our role as a publisher in AI services

**STRUCTURE EXISTING CONTENT**
- develop and offer new products like machine-generated books, reports or APPs
- auto-clustering of subject areas

**HELP RESEARCHERS**
- save time entering new field
- save time staying up to date
- help to overcome research silos

**PROVIDE SERVICE TO SN AUTHORS**
- instant literature overview
- automated ToC generation
- text generation

**ANSWER PUBLISHING QUESTIONS**
- engage with our communities
- shape the future of publishing
MACHINE-GENERATED BOOKS
MACHINE-GENERATED BOOKS

Why we as a publisher?

The challenge is 50 % technology and 50 % publishing

We also want to answer a number of important publishing-relevant questions related to auto-generated content:

- Who is the originator?
- Who is accountable?
- What about copyright and intellectual ownership?
- How does the review process look like and what does peer-review mean here?
- Who selects what a machine is supposed to generate?
- How to answer ethical and legal concerns?

Who – if not we as a publisher – should answer these questions together with the research communities and technology experts.
MACHINE-GENERATED BOOKS

The first challenge

Publish the first machine-generated scientific book

So far, only short texts, fake science and fiction exist that are written by algorithms but no scientific book yet.

We generate new scientific knowledge on the basis of re-combining existing knowledge.
MACHINE-GENERATED BOOKS
The first machine-generated academic book ever

And here it is ...

- published April 2nd, 2019
- first academic book not written by a human author
- compiled by an algorithm jointly developed by Springer Nature and Goethe University Frankfurt, in cooperation with Digital Science

Press Release
MACHINE-GENERATED BOOKS
The first machine-generated academic book ever

- on the basis of a re-combined accumulation of content.

- the book is a cross-corpora summarization of existing texts, organized into a similarity-based sequence.

- the eBook is available for free download on SpringerLink to any interested reader. https://link.springer.com/book/10.1007/978-3-030-16800-1
The first machine-generated academic book ever

Great Media Coverage: Springer Nature’s Thought Leadership

- "Advances the information industry"
- "Coolest things on earth"
- "A real breakthrough"
- "More interesting than it sounds"
The first machine-generated academic book ever

Great Media Coverage: Springer Nature’s Thought Leadership

Is this creative?
Yes and No: the algorithm produces no new results, but the unbiased summary of literally all known facts provides a new perspective, researchers confirm
MACHINE-GENERATED BOOKS
The first machine-generated academic book ever

Publishers without authors? Utopia or Dystopia?

We can create new books automatically, so ….

- Why pay any royalties?
- Do we need acquisitioning editors?
- We can also create journal articles.
- The role of researchers will change.
- The role of publishers will change.

But of course reality is more complicated…
The algorithm and how it works (in brief)

1. Editor selects topic
2. Algorithm pipeline gets access to relevant data
3. Algorithm clusters the content in multi-dimensional similarity matrix
4. Algorithm creates chapter structure based on clusters
5. Algorithm auto-summarizes the content
6. Algorithm paraphrases the text, e.g. by syntactic restructuring
7. Regular book production & publication

The process is made fully transparent

Interested? You can find more information in the introduction to the book (https://link.springer.com/book/10.1007/978-3-030-16800-1)
new internal tool, which makes summarization AI available to SN editorial staff
platform allows user to get from keywords to machine-generated texts
we always work in collaboration with communities (editors, authors or reviewers)
MACHINE-GENERATED BOOKS
The future of machine-generated books

spectrum of human-machine-interaction

100 % human written

Provide authors with machine-generated content (ToC/literature list) for their manuscripts

100 % machine-generated

Enrich human written books with auto-summarizes / combine content

Create machine-generated books edited or reviewed by subject matter experts

Auto-generated content as a service - fully automatic adhoc literature reviews
TECHNOLOGY

Machine-created new knowledge: The algorithms will combine existing knowledge to form new research hypothesis.

PUBLISHERS AND LIBRARIES

We are in the knowledge creation business and content providers and have to rethink our role.

RESEARCHERS

Researchers work changes, as they can concentrate on the creative part, and will not waste time on literature search.
Can this algorithm support the research community in times of crisis?
COVID-19 - FROM REPORT TO APP
New solutions need to tame the coronavirus literature.

“The COVID-19 literature has grown in much the same way as the disease’s transmission: exponentially. The NIH’s COVID-19 Portfolio, a website that tracks papers related to the SARS-CoV-2 coronavirus and the disease it causes, lists more than 28,000 articles — far too many for any researcher to read.”

*nature technology feature, 09 June 2020*
World Health Organisation Director-General Tedros Adhanom Ghebreyesus said, “We’re not just fighting an epidemic; we’re fighting an infodemic”. He was referring to the excessive amount of information surrounding the COVID-19 pandemic. Data dashboards, aggregators and charts of all types have formed the basis of much of what we know about the pandemic, lending a veneer of legitimacy to often contradictory or competing claims.

OECD-DEVELOPMENT-MATTERS.ORG, 17 June 2020

COVID-19 pandemic leads to flood of ‘useless’ science

Academics speak out against ‘ridiculous’ studies that are getting submitted to journals and urge ‘self-restraint’ be restored

SCIENCE | BUSINESS News, 25 June 2020
HOW CAN AI SUPPORT THE RESEARCH COMMUNITY IN TIMES OF CRISIS?
Given the scale and urgency, cutting through the Corona clutter is more important than ever.

First step: A simple overview of a range of recent Springer Nature publications using an AI-based report
- Focus on user needs and pain points: Collaboration with virulogists
- Rapid prototype in the early days of experimental and development

This minimal viable product is in its first iteration a report that groups publications meaningfully together based on their topic, provides snippets of text to grasp what they are about quickly, and then links directly to the respective Springer Nature publication.

Metrics like platform downloads, citations and social media mentions are of importance to measure the reach and potential impact.

Springer Nature's The Source, 01 April 2020
BUT...

... would this be just another solution nobody is aware of?

“So far, these sites have received modest traffic. As of late May, COVID-19 Primer has welcomed 14,000 unique visitors per month, and SciSight has seen 11,000 since it launched. COVIDScholar receives about 500 visitors each day, and COVID-19 KnetMiner has had that many in total. The Kaggle challenge has received the most attention, with 1.7 million page views since it launched in mid-March. Of the COVID-19 researchers we contacted, most had not heard of the majority of these tools. And still more such tools are being developed around the world, including Vilokana in India and CovidAsk in South Korea.”

nature technology feature, 09 June 2020
We leverage a unique combination of
i) strong brand awareness for this community
ii) deep understanding of needs and networks
iii) applicable tech for entity & topic modelling, author graphs and auto-aummarisation
vi) industry collaborations (e.g. Digital Science).

After positive feedback we experiment in various directions in this domain:

● Domain, persona and task specific content recommendations
● Content across all publishers using public CORD-19 dataset
● Reading list, automatic summaries, most important entities
● Most prolific potential collaborators to amplify wayfinders
DON’T GIVE ME THE DETAILS, JUST THE SUMMARY?

TLDR as a community acronym stands for “too long; didn’t read”.

**BENEFIT**

An alternative to abstracts, TLDRs of scientific papers leave out nonessential background or methodological details and capture the key important aspects of the paper, such as its main contributions.

**CHALLENGE**

Writing a TLDR of a scientific paper requires expert background knowledge and complex domain-specific language understanding to both identify the salient aspects of the paper while maintaining faithfulness to the source and correctness of the written summary.

Slide heading: Shashi Narayan et al., ACL 2018
After trying out the public demo:

- Positive paper author feedback on correctness and usefulness
- Next step is extending this to other domains (training corpora)

<table>
<thead>
<tr>
<th>Paper Title</th>
<th>TLDR</th>
<th>Correct</th>
<th>Useful? (5 very, 1 not at all)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimization of Hybrid Electric Drive System Components in Long-Haul Vehicles for the Evaluation of Customer Requirements</td>
<td>An evolutionary optimization algorithm to optimize the drivetrain components for a European long-haul truck concept under consideration of a highly detailed TCO model.</td>
<td>yes</td>
<td>5</td>
</tr>
<tr>
<td>Multi-Objective Optimization of a Long-Haul Truck Hybrid Operational Strategy And a Predictive Powertrain Control System</td>
<td>A rule-based control strategy for a parallel hybrid topology with fuel-saving functions and Predictive Cruise Control.</td>
<td>yes</td>
<td>5</td>
</tr>
</tbody>
</table>
NATURAL LANGUAGE GENERATION
What are language models?
Simply put, they can predict the most likely next words and their probabilities given an input primer, similar to Google query autocomplete. Natural Language Processing (NLP) is increasingly moving in the direction of creating pre-trained language models using huge (!) text corpora, which can later be fine-tuned on specific text domains and tasks.

Passing the Turing Test?
Transformers have risen admirably to the challenge of language generation with many models capable of generating impressive sequences of text that is not only grammatically correct but also cohesive and meaningful.
HOW TO AUTOMATE SCIENTIFIC CONTENT GENERATION?
Apart from technical advances, many challenges and caveats remain.

DEMO
Trained GPT-2 on ACL reference corpus (20k papers’ introductions)

DISCUSSION
A.I. = Augmented / Actionable Intelligence / Insights

We cannot delete the human in the loop.
- Blackbox: origin of the generated text
- Full transparency and ethics imperative
- Controlling many guided generations

+ Help researchers to kick-start writing
+ Use this tech for science journalism
+ Dynamically created topical pages

| Primer | Springer Nature should publish a computer-generated book about computational linguistics because... it offers a broad perspective. Recently, the computational field of Natural Language Processing has seen an exponential expansion due to the rapidly growing volume of written and spoken language data, which has led to interest in the automatic interpretation of natural language texts. (...) In order to produce such books, we will need a large number of computer-generated texts. However, such a large number of texts is not a realistic goal for the foreseeable future. We believe that it is possible to produce texts of a certain level of quality that would be of interest to a large audience.

Image on the top right: STM Tech Trends 2024
THANK YOU

Presentation prepared by
Dr. Stephanie Preuss
Senior Editor
stephanie.preuss@springernature.com

Markus Kaindl
Senior Manager
markus.kaindl@springernature.com
Q & A SESSION AND DISCUSSION

What is the role of the library in all of this?