Lithium-Ion Batteries

Beta Writer

Lithium-Ion Batteries

A Machine-Generated Summary of Current Research

Springer

Beta Writer
Heidelberg, Germany

ISBN 978-3-030-16799-8 ISBN 978-3-030-16800-1 (eBook)
https://doi.org/10.1007/978-3-030-16800-1

Library of Congress Control Number: 2019936280

This book was machine-generated.

Scientific Advisor: Steffen Pauly

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

1 Introduction

Henning Schoenenberger

Advances in technology around Natural Language Processing and Machine Learning have brought us to the point of being able to publish automatically generated meaningful research text.

We have seen the rise of automated text generation in popular fiction (with quite diverse and fascinating results), automated journalism such as in sports, stock market reports or auto-produced weather forecast (data-to-text), automated medical reviews and not to forget the remarkable progress in dialog systems (chat bots, smart speakers).

As far as it concerns scholarly publishing, many attempts in this area up to now have had a negative perception, and the outcome has fallen short of expectations. Often such texts have been however quite successful in demonstrating flaws in the scientific reviewing processes, clearly serving as an important corrective.

1.1 The First Machine-Generated Research Book

What you read here on your mobile device or on your computer screen or even hold in your hand as a printed book is of a different kind. In fact it is the first machine-generated research book. This book about *Lithium-Ion Batteries* has the potential to start a new era in scientific publishing. With the exception of this preface it has been created by an algorithm on the basis of a re-combined accumulation and summarization of relevant content in the area of Chemistry and Materials Science.

The book is a cross-corpora auto-summarization of current texts from Springer Nature's content platform "SpringerLink", organized by means of a similarity-based clustering routine in coherent chapters and sections. It automatically condenses a large set of papers into a reasonably short book. This method allows for readers to speed up the literature digestion process of a given field of research instead of reading through hundreds of published articles. At the same time, if needed, readers are always able to identify and click through to the underlying original source in order to dig deeper and further explore the subject. It can assist anyone who, for example, has to write a literature survey or requires a quick start into the topic. This book proposes one solution (out of many others) to the problem of managing information overload efficiently.

As it involves a number of experimental aspects, the *Beta Writer* was developed in a joint effort and in collaboration between Springer Nature and researchers from Goethe University Frankfurt, Germany. The current implementation will be subject to ongoing refinement, with the machine-generated book on Lithium-Ion Batteries providing the basis to explore strategic improvements of the technology, its integration into production and consumption workflows of scientific literature, the merits it provides and the limitations that it currently faces.

1.2 Why Lithium-Ion Batteries

More than 53,000 articles were published in the last three years, presenting research results on lithium-ion batteries. Rechargeable batteries are a crucial part of our daily life, energizing smartphones, tablets, laptops, alarm clocks, screwdrivers and many other devices. They will become even more important as energy storage systems for electric and hybrid vehicles as well as photovoltaic systems. Therefore, they are a key technology for limiting carbon-dioxide emissions and slowing down climate change. The future of mankind depends on progress in research on lithium-ion batteries, and we need to think of innovative ways to enable researchers to achieve this progress. This is where the potential of natural language processing and artificial intelligence (AI) comes in that might help researchers stay on top of the vast and growing amount of literature.

This first machine-generated book on the topic of lithium-ion batteries is a prototype which shows what is possible today if a researcher wants to get a summarized overview of the existing literature.

Next to Chemistry we are planning to publish prototypes in other subject areas as well, including the Humanities and Social Sciences, with special emphasis on an interdisciplinary approach, acknowledging how difficult it often is to keep an overview across the disciplines.

1.3 A Technological and Publishing Challenge

From the very beginning of developing this prototype we have considered our assignment equally as a technological as well as a publishing challenge. It was evident to us that numerous questions would arise from machine-generated content and from the generation process itself.

Many of these questions, related to machine-generated research content, remain open, and some of them we may not even be aware of yet. Hence we will use this preface as starting point to raise a number of questions which all stakeholders of the scientific community have to answer in a responsible and also collaborative manner. This prototype about Lithium-Ion Batteries is meant to commence an important and necessary discussion that the scholarly community will need to have much sooner than later.

We aim to explore the opportunities and limits of machine-generated research content and simultaneously suggest answers to a number of questions related to the impact of Artificial Intelligence on the scholarly publishing industry and its potential implications.

These questions focus on the crucial elements of scientific publishing:

Who is the originator of machine-generated content? Can developers of the algorithms be seen as authors? Or is it the person who starts with the initial input (such as "Lithium-Ion Batteries" as a term) and tunes the various parameters? Is there a designated originator at all? Who decides what a machine is supposed to generate in the first place? Who is accountable for machine-generated content from an ethical point of view?

We have held extensive discussions on these topics, and we have come to the preliminary conclusion that one possible answer is that there may be a joint accountability which is shared by the developers and the involved publishing editors. However, this is far from being finally decided. And there might be quite some different and equally valid answers.

1.4 Why Transparency Is Important

Full transparency is essential for us to discover both the opportunities of machine-generated content and the current limitations that technology still confronts us with.

But it was also an ethical decision that if we start this journey, we want to do so in a correct and responsible manner, in order to enable a discussion in the research communities that is as open-minded as possible.

During the entire process—from the idea to produce the first machine-generated research book to its realization—there has always been a large consensus that full transparency is one of the key elements of this project.

We also hope that the publication of this book encourages as much feedback as possible to help up learn and improve.

1.5 Continuous Improvement

We are genuinely convinced that exposing the way we work—step by step—failure as integral part of the progress—continuous feedback loop into the development—iterative approach to continuously improve—encourage criticism and learn from it—will help us turning this into a successful prototype and in the long run, shape a product that will be suitable for a large variety of use cases, increasing efficiency and allowing researchers to spend their time more effectively. That is also the reason why we decided to outline the technological side of the implementation of this book in this very preface (see below). Truly, we have succeeded in developing a first prototype which also shows that there is still a long way to go: the extractive summarization of large text corpora is still imperfect, and paraphrased texts, syntax and phrase association still seem clunky at times. However, we clearly decided not to manually polish or copy-edit any of the texts due to the fact that we want to highlight the current status and remaining boundaries of machine-generated content. We have experimented on quite a number of components, and we developed alternative implementations for most of them. Some of the more advanced modules we implemented did not find their way into the final pipeline, and we were following the preferences of the subject matter experts consulted during the development process for their selection. For example, this includes neural techniques, which will improve with additional training data and development time. While we expect them to eventually yield better results, for now they will be held in reserve as we move forward upon the solid foundation of this initial publication.

How will the publication of machine-generated content impact our role as a research publisher? As a global publisher it is our responsibility to take potential implications into consideration and therefore start providing a framework for machine-generated research content. As with many technological innovations we also acknowledge that machine-generated research text may become an entirely new kind of content with specific features not yet fully foreseeable. It would be highly presumptuous to claim we knew exactly where this journey would take us in the future.

1.6 The Role of Peer-Review

It was already pointed out that the technology is still facing a variety of shortcomings which we plan to deal with in a transparent way. We do expect that continuous improvement is necessary to constantly increase the level of quality to be delivered by machine-generated content. On the other hand, we know that the quality of machine-generated text can only be as good as the underlying sources which have been used to generate it. At Springer Nature, we publish research which stands up to scientific scrutiny. In consequence, machine-generated content makes it even more necessary to re-emphasize the crucial role of peer-review itself. Though peer-review is also in the course of being continuously re-defined (and in the future we expect to see substantial progress in machine-support also in this regard) we still think that for the foreseeable future we will need a robust human review process for machine-generated text.

Especially in the area of Deep Learning it becomes increasingly difficult to understand how a result has been actually derived. While concepts such as Explainable Artificial Intelligence (XAI) become more and more crucial, also the review process on machine-generated research content needs refinement, if not a complete re-definition. The term peer itself indicates a certain inadequacy for machine-generated research content. Who are the peers in this context? Would you as a human reader consider yourself as peer to a machine? And should an expert in a specific research field become an expert of neural networks and Natural Language Processing as well in order to be able to evaluate the quality of a text and the related research? In the field of machine-summarization of texts this might not be an issue yet, especially since the underlying sources are peer-reviewed. However, soon enough we will see machine-generated texts from unstructured knowledge bases that will lead to more complex evaluation processes. Also in this area, we have to work together to find answers and define common standards related to machine-generated content. Once more we would like to consider this book as an opportunity to initiate the discussion—as early and anticipatory as possible.

1.7 The Role of the Scientific Author

Finally, what does all this mean for the role of the scientific author? We foresee that in future there will be a wide range of options to create content—from entirely human-created content to a variety of blended man-machine text generation to entirely machine-generated text. We do not expect that authors will be replaced by algorithms. On the contrary, we expect that the role of researchers and authors will remain important, but will substantially change as more and more research content is created by algorithms. To a degree, this development is not that different from automation in manufacturing over the past centuries which has often resulted in a decrease of manufacturers and an increase of designers at the same time. Perhaps the future of scientific content creation will show a similar decrease of writers and an increase of text designers or, as Ross Goodwin puts it, writers of writers:

"When we teach computers to write, the computers don't replace us any more than pianos replace pianists—in a certain way, they become our pens, and we become more than writers. We become writers of writers."[1]

[1] https://medium.com/artists-and-machine-intelligence/adventures-in-narrated-reality-6516ff395ba3.

We do join Zackaray Thoutt's enthusiasm who indicates that "technology is finally on the cusp of breaking through the barrier between interesting toy projects and legitimate software that can dramatically increase the efficiency of humankind."[2]

We have started this exciting journey to explore this area, to find answers to the manifold questions this fascinating field offers, and to initiate a broad discussion about future challenges and limitations, together with the research communities and with technology experts. As a research publisher with a strong legacy, expertise and reputation we feel committed to push the boundaries in a pioneering and responsible way and in continuous partnership with researchers.

2 Book Generation System Pipeline

Christian Chiarcos, Niko Schenk

Automatically generating a structured book from a largely unstructured collection of scientific publications poses a great challenge to a computer which we approach with state-of-the-art *Natural Language Processing* (NLP) and Machine Learning techniques. Book generation involves numerous problems that have been addressed as separate research problems before, and solved to a great extent, but the challenge in its entirety has not found a satisfying solution thus far. The present volume aims to demonstrate what can be achieved in this regard if expertise in scientific publishing and natural language processing meet.

We aim to demonstrate both possible merits and possible limitations of the approach, and to put it to the test under real-world conditions, in order to achieve a better understanding of what techniques work and which techniques do not. In addition, we wish to better understand the demands and expectations of creators, editors, publishers and consumers towards such a product including their reactions to its limitations, and their assessment of its prospective value, both economically and scientifically.

2.1 Choosing a Methodology

As mentioned above, the development process involved both computer scientists and engineers, and editorial subject matter experts who both formulated possible topics and evaluated generated manuscripts and their shortcomings. A key insight from the development process is that different strands of science (and possibly, different personalities) formulate different constraints and preferences regarding the balance between 'creative' automated writing and a mere collation of existing publications. While it is possible to emulate the style and phrasing of prose

[2] https://blog.udacity.com/2017/08/neural-network-game-of-thrones.html.

descriptions rather accurately (e.g., from plain key words, as in Feng et al. [1], or by mean of state-of-the-art language models as reported in the recent work of Radford et al. [2]), the factual accuracy of such 'more creative' reformulations remains questionable. As creators and consumers of scientific publications tend to value correctness over style, we eventually decided for a relatively conservative approach, a workflow based on

1. document clustering and ordering,
2. extractive summarization, and
3. paraphrasing of the generated extracts.

A requirement was to produce novel content, with novelty provided by the organization of sources into a coherent work, and in the generation of chapter introductions and related work sections. We initially considered two thematic domains, chemistry and social sciences, and in both areas, subject matter experts urged us to stay as close to the original text as possible. In other areas of application, where better training and test data for developing advanced summarization workflows may exist, many technical preferences would have been different, but for these branches of research, we designed a workflow according to the premise to preserve as much as possible from the original text—while still producing readable, factually correct, compact, and, of course, novel descriptions. The interested reader may decide to what extent we achieved this goal, but more importantly, let us know where we failed, as it is human feedback—and human feedback only—that can improve the advance of artificial authoring.

2.2 System Architecture

We implement book generation as a modular pipeline architecture, where the output of one module serves as input to the next. Input to the system is a collection of publications that define the scope of the book—typically in the range of several hundred documents. For the present volume, this collection consisted of 1086 initial publications which were identified by keywords and further restricted by year of publication (cf. the next section for details). Output of the system is a manuscript in an XML format which can be rendered in HTML or further processed in the regular publishing workflow.

Main components of the pipeline are illustrated in Fig. 1 and include:

1. **Preprocessing** of input documents, i.e., conversion into the internal format, bibliography analysis, detection of chemical entities, linguistic annotation for parts of speech, lemmatization, dependency parsing, semantic roles, coreference, etc., and re-formulation of context-sensitive phrases such as pronominal anaphora, and normalization of discourse connectives.

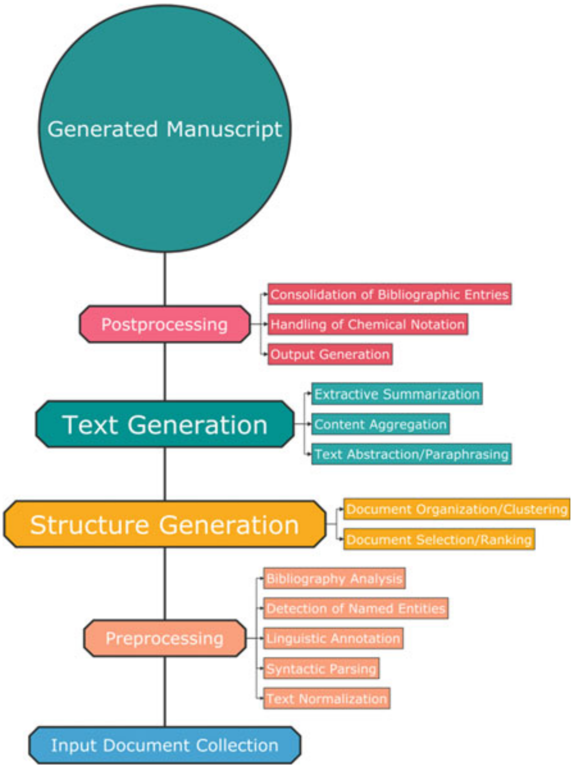

Fig. 1 Book generation system pipeline and NLP components

2. **Structure Generation**

    a. **Document organization** in order to identify the specific contribution and scope of individual input documents, to use this information to group them into chapter- and section-level clusters. As a result, we obtain a preliminary table of contents, a list of associated publications, and keywords that characterize chapters and sections.
    b. **Document selection** is a subsequent processing step during which we identify and arrange the most representative publications per section-level cluster.

3. **Text Generation**

    a. **Extractive summarization** creates excerpts of the selected documents which serve as a basis for subsections.
    b. **Content aggregation** techniques are applied to create sections with introductions and related research from multiple individual documents. Unlike document-level extractive summarization, these are composed of re-arranged

fragments of different input documents, such that information is presented in a novel fashion.

c. **Abstraction** is implemented in a conservative fashion as a postprocessing step to extraction (resp., aggregation). Here, we take single sentences into consideration and employ syntactic and semantic paraphrasing.

4. **Postprocessing** includes the consolidation of bibliographical references, chemical entities, and conversion into an output format that is suitable for generating HTML as well as a manuscript to be handed over to the publishing editor.

For every single component (resp., modules within a component), we provide alternative implementations, and eventually select among these possibilities or combine their predictions according to the preferences of the subject matter experts.

We focus on functionality, less on design. We do not provide a graphical user interface, but the feedback we obtained from subject matter experts during their qualitative evaluation of our system, resp., the generated candidate manuscripts represent invaluable input for the requirement specification of user interfaces to the book generation pipeline.

The pipeline itself is implemented as a chain of command-line tools, each configured individually according to the preferences of the subject matter experts. One premise has been to design an end-to-end system that generates manuscripts from input documents, so the scientific contribution is the overall framework and architecture, not so much the implementation of elementary components for basic machine learning or fundamental NLP tasks. For these, we build on existing open source software (e.g., Manning et al. [3], Clark and Manning [4], Cheng and Lapata [5], Barrios et al. [6]) wherever possible. It should be noted, however, that we do not depend on any specific third-party contribution, but that these are generic components for which various alternatives exist (and have been tested).

2.3 Implementational Details

In **preparation** for generating a book, we identify a seed set of source documents as a thematic data basis for the final book, which serve as input to the pipeline. These documents are obtained by searching for keywords in publication titles or by means of meta data annotations.[3] The document types can be of various kinds: complete books, single chapters, or journal articles.

For **structure generation**, we provide two alternative clustering methods operating on two alternative similarity metrics. As for the latter, we explored bibliography overlap and document-level textual similarity. As bibliography

[3] In the present volume this includes, e.g., any realization of “li-ion battery”, “lithium-ion batteries”, etc. and all occurrences containing “anode” and/or “cathode” as found in either article, chapter, book titles or document meta data.

overlap comes with a considerable bias against publications with a large number of references, we eventually settled on textual similarity as a more robust and more generic metric.

As for clustering methods, *hierarchical clustering* creates a tree structure over the entire set of documents. Clusters can be mapped to chapters and sections according to preferences with respect to size and number. However, we found that the greedy mapping algorithm we implemented for this purpose produces clusters of varying degrees of homogeneity. For the current volume, we thus performed *recursive non-hierarchical clustering* instead: (i) over the set of all documents, core thematic topics are automatically detected (*chapter generation*), and (ii) subtopics are identified within these (*section generation*). If a restriction on the number of input documents per section is defined, the *n* most representative publications per cluster (closest to the center) are chosen, and ordered within the manuscript according to their prototypicality for the cluster (i.e., distance from the cluster center). More advanced selection and ordering mechanisms are possible, but will be subject for future refinements. Figure 2 shows a graphical illustration of the cluster analysis of this book. Each color represents the membership to one of four chapters, bigger labeled dots represent chapters and sections, respectively, small dots show documents.

Interestingly, due to their proximity in the 2-D visualization the graphics shows that Chaps. 1 and 2 are thematically much more closely related (anode versus cathode materials) than Chaps. 3 and 4 (model properties and battery behaviour).

Even though the structure generation for the manuscript is fully automated, here, a number of parameter values can be set and tuned by the human expert who uses the program, such as the desired number of chapters (i.e., cluster prototypes) and

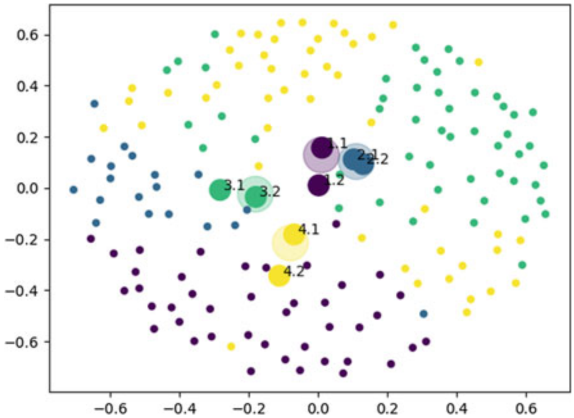

Fig. 2 Two-dimensional projection (PCA) of the cluster analysis with 4 chapters and 2 subsections, and a maximum of 25 documents per section

sections, as well as the number of document assignments per section.[4] The result of this process is a structured table of content, i.e., a manuscript skeleton in which pointers to single publications serve as placeholder for the subsequent text.

At this level, subject matter experts requested the possibility for manual refinement of the automatically generated structure. We permit publications to be moved or exchanged between chapters or sections, or even removed if necessary, for example, if they seem thematically unrelated according to the domain expertise of the editor.[5] We consider the resulting publication nevertheless to be machine-generated, as such measures to refine an existing structure are comparable to interactions between editors of collected volumes and contributing authors, e.g., during the creation of reference works.

Chapter and section headings are represented as a list of automatically generated keywords. Technically, these keywords are the most distinctive linguistic phrases (*n-gram* features) as obtained as a side-product of the clustering process and are characteristic for a particular chapter/section. Again, human intervention is possible at this stage, for instance, in order to select the most meaningful phrases for the final book. In the present volume, the keywords remained unchanged. More advanced techniques include the generation of headlines from keywords using neural sequence-to-sequence methods, and while we provide an implementation for this purpose, we found it hard to ensure a consistent level of quality, so that we eventually stayed with plain keywords for the time being.

Text generation is the task to fill the chapter stubs with descriptive content. In the present volume, this is based on *extractive, reformulated summaries*. Every chapter consists of an introduction, a predefined number of sections as determined in the previous step, a related work section, a conclusion, and a bibliography. We elaborate on each of them in the following.

Every *chapter introduction* contains a paragraph-wise concatenation of extractive summaries of all individual document introductions which have been assigned to the chapter.[6] Since all documents of a chapter have been identified to belong to the same topic, the motivation here is to combine the content of individual introductions from the publication level and merge them into a global one on the book

[4]For the present volume, the number of chapters, sections, and the maximum number of documents per section have been initially set to 4, 2, and 25, respectively. The document clusters, i.e. document to chapter assignments, are produced by k-means clustering on the term-document matrix with different weighting schemes, e.g. TF-IDF. Additional, advanced parameters to be set include the minimum/maximum document frequency of a term, the total number of features (*n-grams*) used, or whether to use stemming or other types of text normalization.

[5]For the present volume, 9 documents have been moved between chapters, and 8 documents were excluded from the final book. Overall, the generated book is based on 151 distinct publications.

[6]We elaborate on extractive summarization below.

level. The summary length (in words and as a proportion of the original text length) is parameterizable by the human editor who uses the system.[7] The conclusion of the book is built in the same way. The introduction produced in this way is conservative in that it reflects the introductions of the input documents selected for the chapter—both in order and content. As an alternative, we also implemented an approach for *reordering and combining sentences* obtained from *different* publications in single paragraphs, based on an arrangement of semantically similar sentences closer to each other and the elimination of near-duplicates. For the present volume, however, the more simplistic implementation was eventually selected, as subject matter experts found that the coherence of the resulting text suffered from the heterogeneity of the underlying documents. For future generation experiments, it would be desirable to allow an expert trying to produce a book with this technology to compare the conservative and the aggregated introductions for each chapter. For more homogeneous chapters, the latter approach may be favorable.

On the *section level*, following the introduction, publication stubs are filled with extractive summaries obtained according to different technologies:

- *Unsupervised extractive summarization*: A classical baseline for extractive summarization is the application of the page rank algorithm to the text itself, respectively, the graph of linguistic annotations obtained from it. As a result, both important phrases and relevant sentences are augmented with relevance scores, and a ranking according to these, and extractive summarization boils down to retrieving the *x* most relevant sentences until a certain length threshold is met. The method has the great advantage of being simple, mature and well-tested. It is, however, context-insensitive, in that essential information may be lost, or that sentence and keyword relevance in the context of a book project diverts from their relevance within the original publication in isolation.
- *Supervised extractive summarization*: An alternative approach to produce a ranking of sentences is to train a regressor to approximate a pre-defined score, e.g., the extraction probability of a sentence. Unfortunately, such training data is not available for our domain and cannot be created without massive investments. As a shorthand for such data, we measure the textual similarity of each body sentence with the sentences of the abstract of the same publication. We then trained a regressor to reproduce these scores, given the sentences (resp., their embeddings) in isolation. This regressor was trained for all publications from the Chemistry and Material Science, as well as Engineering domains, and it assigns every sentence a score, and thus, all sentences from a document a domain-specific rank, which then serves as a basis for extraction.
- *Extended abstracts:* An extended abstract is a reformulated and compressed version of the original abstract of a document, potentially enriched with sentences from the body of the publication as useful additional information can be retrieved

[7] The summary length has been set to either 270 words or 60% of the original text length—depending on which one was shorter. This combined metric handles the trade-off between too lengthy summaries on the one hand, and summaries which contain almost every sentence of the source, on the other.

from the document itself. For the current volume, we append sentences from the body to any sentence in the original abstract by a similarity metric (provided they exist). Similarity is measured by customizable n-gram overlap. An alternative, and slightly more aggressive implementation, is to replace sentences from the original abstract with the most similar sentences from the body.

- *Weighted combined ranking:* Each of the aforementioned methods assign sentences a score (or a classification, which can be interpreted as a binary score), from which a rank can be calculated. We provide a re-ranker that uses the weighted sum of ranks produced by different components to produce an average rank, so that more conservative approaches (extended abstracts) can be combined with context-sensitive, machine-learning based rankers (supervised extraction) and with context-insensitive, graph-based methods (unsupervised extraction) according to the relative weights the user of the system assigns to each component.

Subject matter experts from the chemistry domain found that the first two methods (and their combination) are prone to factual errors, if applied to original sections on methodology and experimental setup, so that such sections (which constitute the majority of text in this domain) must not be summarized but either dropped or quoted. This may be a characteristic of the chemistry domain, where instructions on replicating a particular experiment must be followed carefully and any omission of a step in the procedure or an ingredient is potentially harmful. For the present volume, we thus operate with extended abstracts only.

Conclusions are aggregated in the same way as introductions. It is followed by a related work section which is compiled from the citations of the input documents.

The *related work section* is typically short and organized around pivotal publications. A pivotal publication is defined as a DOI that is referred to within different publications from a chapter, and we take the number of documents referring to this publication as an indicator of its relevance. The user of the system defines a threshold *n* for the number of documents that define a pivotal publication. For instance, if *n* is set to 4, at least four different documents within a chapter need to cite the same publication DOI. From each document, we retrieve the citation context, i.e., the sentences that contain the reference, and arrange them according to their textual similarity. We thus obtain 4 sentences, at least. The frequency threshold *n* needs to be set by the user of the system.[8] It should increase in proportion with a greater number documents per chapter. Ideally, the most frequently referred to publications by distinct sources have global importance within a chapter.

The final component of text generation is **text abstraction**, i.e., here the linguistic reformulation of the original sentences, respectively. In order to create text which is not only novel with respect to its arrangement, but also with respect to its formulation, and in order to circumvent issues related to copyright of the original texts, we attempt to reformulate a majority of the sentences as part of the generated book, while trying to preserve their original meaning as best as possible. At the

[8] For the present volume, $n = 2$.

same time, subject matter experts urged us to stay as close to the original formulation as possible. Although we do perform deep parsing, this is used to inform reformulation rules only, but not as a basis for the summarization itself: The original text is preserved and reformulated, rather than being reduced to a graph and then re-generated from scratch. In this more conservative approach, we provide annotation-based reformulation components for which integrated different NLP modules have been integrated in a preprocessing step, as outlined above: identification of word boundaries,[9] part-of-speech label assignment for words (i.e. word categories, such as noun, verb, etc.), and the application of syntactic and semantic parses to each sentence in order to obtain a linguistic analysis in terms of dependency structure and semantic roles. Furthermore, coreference resolution is employed in order to detect mentions in the text and associated referential expressions (e.g., personal or possessive pronouns in subsequent sentences).

We provide the following modules:

- *Rule-based simplification*: Sentence-initial adverbials, discourse markers and conjunctions are removed as they would otherwise appear out of context after text summarization.
- *Sentence compression*: Using relevance scores such as created during keyword extraction above, and a reduction threshold of, e.g., 90%, eliminate the least relevant parts of the sentence until the reduction threshold is met. An alternative implementation shortens a sentence by removing omittable modification information, e.g., non-core, local/temporal cues, or discourse modification.
- *Sentence restructuring*: A range of syntactic transformation rules were implemented which operate on the automatically produced structure, for instance, to turn an active utterance into its passive variant.
- *Semantic reformulation*: In a final step, we substitute single words as well as longer phrases if we find synonymous expressions that exceed a predefined similarity threshold.[10] What constitutes a phrase is automatically detected by high pointwise mutual information of word co-occurrences. Note that all synonyms are automatically learned from large amounts of raw unlabeled texts using state-of-the-art methods for unsupervised learning of word representations with neural networks.

Along with these reformulation components, a module for anaphora resolution is applied to replace intersentential pronominal anaphora with the respective last nominal representation: We replace pronouns (e.g., sentence-initial *"It"*) by interpretable mentions of the same coreferential chain that are found in the prior context (e.g., full noun phrases such as *"the first study in this field"*) in order to prevent the rendering of sentences in which single pronouns appear without context after text

[9]Specifically, we developed special analyzers on the sentence level to detect, normalize and later on reinsert chemical notations, textual content in brackets (such as references and supplementary information) and other entities which need to be treated holistically and must not be parsed or split into parts.

[10]When more than one word exceeds the threshold, we select one synonym randomly.

summarization. Note, however, that this is applied during preprocessing already, in order to guarantee that extractive summarization does not create unresolvable anaphoric references.

Apart from the fully automated text generation module, the human user still has influence on the quality of the text, for example by specifying a list of prohibitive synonym replacements, or by setting the thresholds for the replacements. For compiling this volume, we selected among the aforementioned modules and adjusted their respective threshold in accordance with the feedback from subject matter experts. It is to be noted, however, that users would apparently like to scale freely between different degrees of reduction and reformulation, ranging from literal quotes to complete paraphrases. Our implementation does not provide such an interface, but developing such a tool may be a direction for future extensions.

As an example of two reformulated sentences compared to their original source sentences, involving preposing of temporal information and most of the NLP techniques described above, consider the following sentences (synonym replacements in bold, syntactic changes and coreference replacements underlined).

Source[11]:

Lithium-ion batteries have played a major role in the development of vehicle electrification since the 2000s. They are currently considered to be the most efficient technology in this market.

Automatically reformulated:

*Since the 2000s, lithium-ion batteries have played a **main** role in the development of vehicle electrification. **Lithium-ion batteries** are **presently regarded** to be the most **effective** technology in this market.*

Advanced syntactic reformulation, e.g., turning active into passive voice is illustrated in the next example.

Source:[12]

Finally, these results can develop a test methodology to determine the management of lithium batteries pack that experiences a potential heating threat.

Automatically reformulated:

*A test **approach** to **specify** the management of **Li-ion** batteries pack that experiences a potential heating threat **could be devised by these** results.*

In total, for the present volume, approximately ¾ of all sentences were syntactically reformulated, i.e. for 74% at least one transformation rule triggered. Semantic replacements (unigram, bigram, or trigrams substituted) were made to 14.7% of all tokens. More than 96% of all sentences were modified by at least one semantic substitution. Sentence compression was kept in a very conservative mode and removed only a small portion of 0.9% of the tokens. In order to acknowledge the original source, every sentence is coupled with the DOI of its source document. In addition, sentences which were not affected by reformulation, synonym replacements, or sentence compression are marked as literal quotes (1.2% of all sentences).

[11] Sabatier et al. [7] https://link.springer.com/chapter/10.1007%2F978-3-319-55011-4_3.

[12] Chen et al. [8] https://link.springer.com/article/10.1007/s10973-017-6158-y.

2.4 Challenges and Future Directions

Our book generation pipeline has been designed to not only compile extractive summaries, but also to rephrase and make creative modifications to the original text wherever possible. At the same time, however, it is forced to be conservative enough to preserve the original meaning of the sentences. Besides selecting the most important sentences in extractive summaries, this tradeoff can be seen as the most difficult challenge in the design and implementation of the system.

The system in its current version is a minimalist implementation of core components of a book generation workflow and can be refined and extended in many ways. This preliminary state is also indicated by the name of the virtual author, *Beta Writer*. Aside from creating a scalable end-to-end system for the generation of books from large bodies of scientific publications, we see our main contribution as the first successful attempt to push a machine-generated book beyond mere technical challenges through an established publication workflow up to the level of a printed book. At the same time, the name entails a commitment for future extensions and refinements, for which manifold possibilities exist, including the following:

- *Improving linguistic quality*: Current limitations of the system are mainly due to error propagation in the NLP pipeline. For instance, the very basic preprocessing steps, word and sentence identification, are both non-trivial tasks, especially for texts containing various chemical notations, numbers, or abbreviations in which punctuation symbols do not necessarily indicate a sentence or word boundary. Wrongly detected words and sentences lead to faulty linguistic annotations by the part-of-speech taggers, ultimately to wrong parses, and finally to restructured sentences which are meaningless.
- *Improving paraphrasing*: Issues regarding legibility, grammaticality, and correctness, are also partly due to the component which replaces words by synonyms: This component is not yet sensitive to aspect, or context and, thus, in some cases a substitution of a word is acceptable (*revealed* good performance -> *showed* good performance), in others not (it is *revealed* -> it is *showed*). Even more problematic in this regard is a well-known disadvantage associated to word embeddings, namely that antonyms have very similar distributions compared to synonyms. Replacing a word by its antonym, however, changes the meaning and is prohibitive in sentence reformulation. We have tried to overcome these issues as well as in any way possible using conservative similarity thresholds.
- *Headline generation*: The generation of suitable, narrative headlines (for chapters and sections) is yet another highly complex task which we did not approach in the current version of the system, but rather prompted us to stick to the keywords that we obtained as a result of text clustering. Note that the keywords themselves are not necessarily the most interpretable and meaningful phrases to a human reader, even though technically they are in fact the most distinctive n-gram features. Future research will address their combination into syntactically more appealing descriptions.

- *Improving coherence*: In this current version, we have not addressed any discourse properties of the texts. Typically, sentences do not occur in isolation. Instead, they are part of a well-formed and coherent text structure which is signaled either explicitly (e.g., using discourse markers but, next, if, etc.) or sometimes even implicitly. In fact, our extractive summaries break up and remove parts from the discourse structure of the original source documents and, in future versions of the system, special focus needs to be taken to ensure that the reformulated extractive summaries adhere to the original discourse structure and its associated global meaning. This would also entail fusing sentences and reintroducing discourse markers where applicable. We want to point out, however, that such a feature is not only non-trivial to implement but also extremely hard to evaluate.
- *Reordering*: A related challenge is the sequential order of sentences—and, similarly, the sequential ordering of sections within a chapter. Here, we have implemented different simplifications that either preserve the original order of sentences or perform re-ordering in a way that maximizes similarity between adjacent sentences. More advanced implementations could build on formal representations of discourse structure as also necessary for improving coherence.
- *Abstraction via graph representations*: A book generation pipeline based on full-fledged abstractive summarization requires the decomposition of texts and sentences into their logical parts, their representation as a graph, and the re-generation of natural language from the abstract graphs. At the moment, this is an area of intense research, and several experimental prototypes already do exist, but we estimate that a production-ready implementation will not be available for another, 3–5 years. For the academic partners in this enterprise, this is of course one of the aspects of the book generation challenge that we are particularly interested in.
- *Neural abstraction*: Another way of abstraction is the application of neural sequence-to-sequence models to translate full sentences into their paraphrases. Again, this would be a strategic goal, but we currently lack training data for our domain, and where training data is synthesized (e.g., by means of a neural noisy channel model), it is virtually impossible to guarantee a consistent level of quality in the generated output. Our own experiments show that the output that can be produced is superficially readable, but often has severe flaws when it comes to its meaning and factual correctness. For the present system, and the eventual pipeline we developed, we thus went for a conservative, extraction-based architecture. Nevertheless, this is an area of intense research.
- *Creative writing*: Another scientific challenge is the production of novel text fragments *from contextual cues* rather than from a given input sentence. While on a technical level, this is similar to neural headline generation, such apparent simulation of creative behaviour is probably the most fascinating aspect of modern-day AI. In fact, it is fairly easy to build and train a model to re-generate sentences given the previous and the next sentence. However, the quality of the generated output is even less controllable than the results we achieved by neural abstraction. Again, this remains an area of research.

- *Including structured data sources*: At the moment, the *Beta Writer* builds on three pillars: Established NLP techniques, word embeddings for the target domain, and vast amounts of scientific publications to optimize both and to create summaries from. There is another possible component that we did not take into account so far: Structured knowledge graphs can provide additional background information, e.g., about chemical entities and relations between them. In fact, such information is already available, and Springer Nature can build on the Springer Nature SciGraph in this regard. For the creation of this publication, however, we focused on core functionalities of a generic book generation pipeline, which will permit domain-specific knowledge base integration in future iterations.
- *The nasty little details*: Last but not least, we have to mention that a great deal of the errors that we are currently facing are due to specifics of the domain and the data. The interested reader will immediately spot such apparently obvious errors—with rather obvious solutions. This includes, for example, the occasional use of *us*, *ourselves*, *this paper* etc. which refers back to the original publication but is clearly misplaced in the generated book. The solution to these is a simple replacement rule, the challenge in this solution is the sheer number and the distribution of errors that require a domain-specific solution each, sometimes referred to as 'the long tail'. While we made some efforts to cover such obvious cases, continuous control and refinement of an increasingly elaborate set of repair rules is necessary, and will accompany the subsequent use and development of the *Beta Writer*.
- *Getting the human in the loop*: Error correction can potentially also be covered by a human expert—or, in a book production workflow, as part of copyediting. But even beyond this level of manual meddling with the machine-generated manuscript, a clear, and somewhat unexpected result of our internal discussions with subject matter experts on chemistry and social sciences was that editors would like to maintain a certain level of control. At the moment, the system remains a blackbox to its users, and we manually adjust parameters or (de)select modules according to the feedback we get about the generated text, then re-generate, etc. At the same time, it is impossible to optimize against a gold standard—because such data does not exist. One solution is to provide a user interface that allows a user to switch parameters on the fly and see and evaluate the modifications obtained by this and thus optimize the machine-generated text according to personal preferences, and—also depending on the feedback we elicit on this volume—developing such an interface is a priority for the immediate future.

We are well aware of experimental approaches that improve upon the current state of our implementation. With a publication that links every generated sentence with its original form in the original publication, we aim to establish a reference point for evaluation by the scientific community and a baseline for future systems to meet. Yet, at the core of this challenge is not so much scientific originality, but the balance of having an automated system performing autonomous and 'creative' operations and the degree to which the factual accuracy of the underlying text can be preserved. Guided by subject matter experts on chemistry and social sciences, we eventually went for a conservative approach to book generation, in that as much information is preserved from the original as possible. We are aware of the expectations in trustworthiness and verifiability in scientific publications which—for the time being—, a more radical, abstraction-based approach on book generation would be impossible to meet. We expect this to change in the immediate future, and we are working towards it, but at the same time as Artificial Intelligence—or, for that matter, neural Natural Language Processing—is about to reach the fringes of creativity, we still need to learn how to restrict its creativity to producing content that remains factually true to the data its predictions are generated from.

Another technical challenge that we identified during the creation of this book was that human users aim to remain in control. While an automatically generated book may be a dream come true for providers and consumers of scientific publications (and a nightmare to peer review), advanced interfaces to help users to guide the algorithm, to adjust parameters and to compare their outcomes seem to be necessary to ensure both standards of scientific quality and correctness. Advanced interfaces will also help to identify areas where it is possible to deviate from the cautious, conservative approach on text generation applied for producing the present volume, and to include more experimental aspects of AI.

References

1. Feng X, Liu M, Liu J, Qin B, Sun Y, Liu T (2018) Topic-to-essay generation with neural networks. In: Proceedings of the twenty-seventh international joint conference on artificial intelligence (IJCAI-18), pp 4078–4084
2. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I (2019) Language models are unsupervised multitask learners. https://blog.openai.com/better-language-models/
3. Manning CD, Surdeanu M, Bauer J, Finkel J, Bethard SJ, McClosky D (2014) The stanford coreNLP natural language Processing Toolkit. In: Proceedings of the 52nd annual meeting of the association for computational linguistics: system demonstrations. pp 55–60
4. Clark K, Manning CD (2016) Deep reinforcement learning for mention-ranking coreference models. In: Proceedings of EMNLP
5. Cheng J, Lapata M (2016) Neural summarization by extracting sentences and words. In: Proceedings of the 54th annual meeting of the association for computational linguistics (Volume 1: Long Papers) pp 484–494. https://doi.org/10.18653/v1/P16-1046
6. Barrios F, López F, Argerich L, Wachenchauzer R (2015) Variations of the similarity function of textRank for automated summarization. Anales de las 44JAIIO. Jornadas Argentinas de Informática. In: Argentine Symposium on Artificial Intelligence
7. Sabatier J, Guillemard F, Lavigne L, Noury A, Merveillaut M, Francico JM (2018) Fractional models of lithium-ion batteries with application to state of charge and ageing estimation. In: Madani K, Peaucelle D, Gusikhin O (eds) Informatics in control, automation and robotics. Lecture notes in electrical engineering, vol 430. Springer, Cham
8. Chen M, Yuen R, Wang JJ (2017) Therm Anal Calorim 129:181. https://doi.org/10.1007/s10973-017-6158-y

Contents