Extractive and Abstractive Summarization of Organic Chemistry Patents

Darby Brown
University of California, Berkeley
Dec 2023

Abstract

Patent literature contains a wealth of technological information, with 71% of new technology published in patent literature not being published elsewhere *(Kim, 2020)*. Efficient literature review is vital for research and development, necessitating thorough patent review processes. Patent summarization poses challenges due to domain-specific vocabulary, legal nuances, and long documents. This study explores domain-specific hybrid summarization of organic chemistry patents. We employ SumBasic for extractive summarization paired with Google's Pegasus-large for abstractive summarization, with a focus on organic chemistry patents. Utilizing the Harvard USPTO Patent Dataset, we analyze the performance of different patent sections and present a proof of concept for further optimization in materials and chemical industries.

1. Introduction

Seventy one percent of new technology published in patent literature is not published elsewhere *(Kim, 2020).* Accurately evaluating this treasure-trove of information to discern latest innovations is critical to efficient research and development for any organization. To this end, lengthy patent review processes are a mainstay for companies and research institutes alike. Screening the patent landscape is also a critical activity for attorneys who must prior art when evaluating or submitting patents.

In the field of text summarization, patent summarization has proven particularly challenging. Touching nearly every field of technological innovation, patents are long-form documents detailing novel ideas and approaches. They often contain brand new words or processes and a mix of domain-specific technical and legal jargon that protects intellectual property while making interpretation by human or machine difficult. Summarizing them requires the proper interpretation of uncommon, domain-specific vocabulary and correct identification of the novelty described in the patent.

Driven by a massive increase in patent filings, researchers have worked to overcome these challenges through an array of techniques. Much of the research in this space focuses on replicating the abstract or claims of published patents in an effort to speed workflows for patent attorneys. In this work, we follow the same approach of summarizing patent abstracts from publicly available data with a specific focus on organic chemistry patents. We focus on hybrid summarization, first using SumBasic for extractive summaries, and domain-specific fine tuning of Google's *Pegasus-large.*

Optimization in this domain space would be of particular interest to materials and chemical corporations who invest heavily in patent landscaping each year and hold proprietary summaries of patents related to their product portfolio. We hope this work serves as a proof of concept for further fine-tuning on product-specific datasets.

2. Background

The application of transformers for patent summarization has blossomed in recent years with the release of more capable models such as Pegasus. This model is particularly well-suited for the task as it is specifically pre-trained for abstractive summarization. It uses full-sentence masking (GSG) as well as token masking (MLM), and certain versions of the Pegasus model have also been trained on patent summarization (BIGPATENT dataset) *(Zhang, 2020*).
Pegasus has been fine-tuned on the complete HUPD dataset to predict the first claim with strong results on PegasusBigBird *(Moreno, 2023)*, finding that the summary and abstract served as the best input for this purpose. Owing to limited compute resources and a different target summary, we opt for Pegasus-large and screen all possible sections for summarization for this research.

In 2020, Pilault et al. showed the effectiveness of adding previously extracted sentences to train language models with long dependencies *(Pilault, 2020).* We follow a similar approach, working with extractive summaries before abstractive summarization through transformers.

Outside of transformers and single-document summarization, work has been done for multi-document summarization using generative adversarial networks *(Kim, 2022).* Researchers have speculated the use of PRIMERA *(Xiao, 2022)* for multi-document summarization. The work in multi-document summarization would lend itself well as a next step for the goal of patent landscaping.

3. Dataset

We leverage the Harvard USPTO Patent Dataset (HUPD) dataset, which is a large-scale corpus of patents published with the US Patent and Trademark Office between 2004-2018. *(Suzguna, 2022)*. We use the international patent classification key to filter for only patents filed under the Organic Chemistry classification (C07). This subset accounts for only 2% of all patents filed.

4. Methods

Given the relatively small representation of organic chemistry patents, we focus on optimizing summarization performance for this domain specifically. For all experiments detailed below, SumBasic and Pegasus-large were used for extractive and abstractive summarization, respectively.

4.1 Section Selection

Patents are comprised of multiple sections:

- **Abstract** is a short summary of the invention with the most pertinent facts.
- **Claims** specify the extent of legal protection. Multiple claims can be listed in order of importance.
- **Description** provides detail for a skilled person to understand the invention.

- **Background** provides context and prior art.
- **Summary** presents all critical information of the document in shorter format, emphasizing it's nature and purpose.

Each section is treated differently per individual patent. As seen in Figure 1, the length of any given section can vary from a few sentences to many thousands of words. The first step of our research was to screen the quality of summary generated from any single section, using the Abstract section as a label. The top performing section using Pegasus-large out of the box will serve as our baseline.

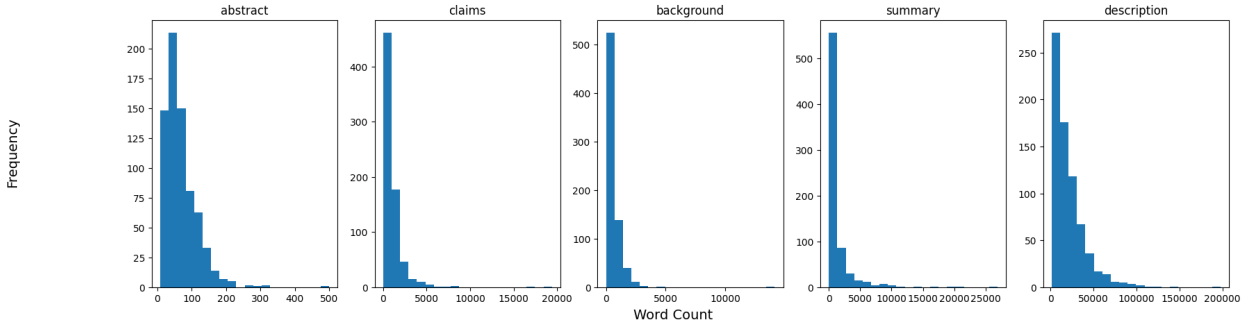

Figure 1. Distribution of section lengths for a subset of organic chemistry patents in the HUPD dataset.

4.2 Extractive Summarization

The average length of a patent in our subset was 24,000 words. We leverage extractive summarization through SumBasic to improve the capture of key ideas in these long-form documents. The summaries generated through SumBasic ensure that the sentences with highest-frequency words are passed into Pegasus-large, which has a max input length of 1024 tokens. We try extractive summarization of the full patent as well as of the most informative sections, as determined in the section selection step.

4.3 Fine Tuning

Finally, we fine-tune Pegasus-large on the best performing section and the extractive summaries to try to improve domain-specific performance. Hyperparameters were kept consistent after some testing. *Fixed Hyperparameters: learning_rate = 10e-5, weight_decay = 0.01, epochs = 5*

5. Evaluation

Despite some criticism of its utility for abstractive summarization, the standard for scoring patent summarization quantitatively is through ROUGE scoring. One important consideration for the Organic Chemistry space is that long chemical names *(for example, bis-(trichloromethyl)- benzene)* are often broken up into many tokens. The appropriate representation of chemical names is a critical component to proper summarization, and we

therefore place extra consideration on the ROUGE-L score, as the longest common subsequence will often be the relevant chemical name.

To account for the shortcomings of ROUGE scoring, we confirm scores using qualitative comparisons and observations (Appendix A).

6. Results and Discussion

6.1 Section Selection

As represented in Figure 2, the description section outperformed across all ROUGE metrics, followed by the claims section. This is a sensible outcome based on Pegasus-large's training set. The BIGPATENT dataset on which it was trained only consists of patent title, abstract, claims, and description. Additionally, the Summary and Background sections had zero text for some patents, likely contributing to their poor scores.

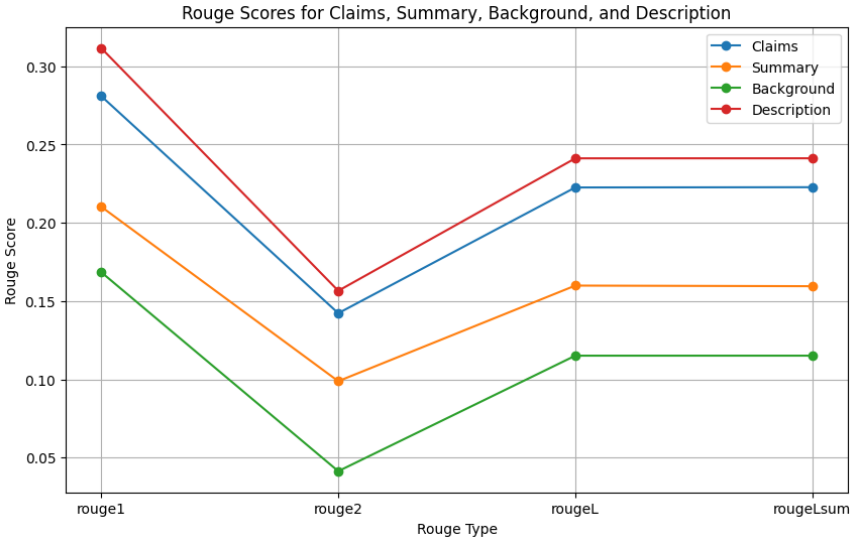

Figure 2. ROUGE scores for claims, summary, background, and description sections using the abstract as the reference summary.

6.2 Extractive Summarization and Fine Tuning

Extractive summarization did not perform as well as expected quantitatively, both on the standard Pegasus-large model and during fine-tuning. One likely cause of this, which we discuss in the limitations section, is the possibility of data leakage between HUPD and BIGPATENT datasets. The scores for the 'Desc' dataset may be artificially inflated due to this.

Despite sub-par ROUGE scores, the hybrid summaries do seem to be representative of the patents from a qualitative perspective. Further work would need to be done to screen a sufficient sample size for conclusive qualitative results, however.

When considering how to improve the performance of the hybrid summarization, SumBasic has not historically been the best extractive summarization technique for patents.

Graph-based algorithms such as TextRank may perform better in a task where capturing relationships between sentences is more important than the frequency of words.

		Dataset	
Model	ROUGE	Desc	Ext-Sum-CD
Pegasus -large	R1	31.16	22.51
	R2	15.66	8.57
	RL	24.11	15.81
	RLsum	24.15	15.8
Fine tuned Pegasus -large	R1	44.68	29.66
	R2	27.94	13.34
	RL	37.52	23.42
	RLsum	40.42	25.09

Table 1. ROUGE scores for baseline, hybrid, and fine-tuned models. **Desc**: Description section only as input_ids. **Ext-Sum-CD:** extractive summarization through SumBasic performed on the claims and description sections.

6.3 Limitations

Memory: A number of decisions had to be made in the interest of space and time limitations. The subset of HUPD Organic Chemistry patents was shrunk from 80,000 total available patents to 720 examples to allow for extractive summarization and fine tuning in Google Colab.

Data leakage: Importantly, when reviewing the results qualitatively, it became clear that the model was cheating on some patents-returning the exact abstract verbatim. This makes it clear there is overlap in the HUPD and BIGPATENT datasets, which undermines the validity of our results for summarization on wholly-unseen patent datasets. (Appendix A)

7. Conclusion

Our study delves into the complex realm of patent summarization, with emphasis on distilling information from organic chemistry patents. While Pegasus-large, fine-tuned on the description section, showed good ROUGE scored, future work should investigate the extent of data leakage and discern actual test performance relative to extractive summarization. Additionally, future work could explore alternative extractive summarization techniques and assess performance on other domain-specific datasets to enhance the utility of patent summarization for industrial stakeholders.

References:

1. Mirac Suzguna, Luke Melas-Kyriazib, Suproteem K. Sarkarc, Scott Duke Kominersc, Stuart M. Shieberc. 2022. The Harvard USPTO Patent Dataset: A Large-Scale, Well-Structured, and Multi-Purpose Corpus of Patent Applications. https://arxiv.org/pdf/2207.04043.pdf
2. Jingqing Zhang, Yao Zhao, Mohammad Saleh, Peter J. Liu. 2020. *PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. https://arxiv.org/abs/1912.08777*
3. Eva Sharma, Chen Li and Lu Wang. 2019. **BIGPATENT: A Large-Scale Dataset for Abstractive and Coherent Summarization**. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. https://evasharma.github.io/bigpatent/
4. Sunhye Kim, Byunun Yoon. 2022. **Multi-document summarization for patent documents based on generative adversarial network**. Expert Systems with Applications Vol. 207. https://www.sciencedirect.com/science/article/pii/S0957417422012118?casa_token=gN4tuSNa1JUAAAAA:mTcRIQRK5_dtpJ6DLD9jEu2U40adxswKWCb_shVi0FoU9HBeqercMxwunS1iIL_0c9tFShRnSeo
5. Jonathan Pilault, Raymond Li, Sandeep Subramanian, Christopher Pal. **On Extractive and Abstractive Neural Document Summarization with Transformer Language Models**. 2020. Proceedings of the 2020 Conference on Emperical Methods in Natural Language Processing. https://aclanthology.org/2020.emnlp-main.748/
6. Sarah Moreno. 2023. **Transformers-based Abstractive Summarization for the Generation of Patent Claims**. Politecnico di Torino Masters' Thesis https://webthesis.biblio.polito.it/26720/
7. Silvia Casola, Alberto Lavelli. 2022. **Summarization, Simplification, and Generation: The Case of Patents.** Università di Padova, Human Inspired Technology Research Centre. https://arxiv.org/abs/2104.14860
8. World Intellectual Property Organization. 2009. IPC Publication. https://ipcpub.wipo.int/?notion=scheme&version=20230101&symbol=none&menulang=en&lang=en&viewmode=f&fipcpc=no&showdeleted=yes&indexes=no&headings=yes&notes=yes&direction=o2n&initial=A&cwid=none&tree=no&searchmode=smart
9. Wen Xiao, Iz Beltagy, Giuseppe Carenini, Arman Cohan. 2022. PRIMERA: Pyramid-based Masked Sentence Pre-training for Multi-document Summarization. https://arxiv.org/pdf/2110.08499.pdf

Appendix A: Qualitative Results

	Selected example: Section Selection Step
Abstract	Paraffin compositions including mainly even carbon number paraffins, and a method for manufacturing the same, is disclosed herein. In one embodiment, the method involves contacting naturally occurring fatty acid/glycerides with hydrogen in a slurry bubble column reactor containing bimetallic catalysts with equivalent particle diameters from about 10 to about 400 micron. The even carbon number compositions are particularly useful as phase change material.
Claims Summary	<pad>A method for preparing a phase change material comprising: feeding naturally occurring fatty acids and esters to a reactor containing a liquid phase, a gas phase, and a catalyst comprising nickel, molybdenum, cobalt, tungsten, or a combination of any two or more thereof; at least partially hydrogenating/hydrogenolyzing the naturally occurring fatty acids and esters in the reactor; controlling an intracatalyst temperature gradient within the reactor to reduce thermal decarboxylation relative to oxygen hydrogenolysis; and withdrawing a</s>
Summary Summary	<pad>2 is a schematic diagram of an embodiment of an operation of a series-reactor process in accordance with the present invention. 3 is a schematic diagram of an alternative embodiment of an operation of a series-reactor process in accordance with the present invention.</s><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad><pad>
Background Summary	<pad>These factors make the latent heat for even carbon number 14 n-hexadecane (carbon number 18) and even carbon number 20 n-hexadecane (carbon number 18) more expensive than the latent heat for even carbon number 14 n-hexadecane (carbon number 18) n-hexadecane (carbon number 18) n-hexadecane (carbon number 18) n-hexadecane (carbon number 18) n-hexadecane (carbon number 18) n</s>
Description Summary	Field of the Invention The present invention relates to producing specialty materials and chemical intermediates from bio-renewable feedstocks such as animal fats, plant oils, algal oils, bio-derived greases, and tall oil fatty acid (hereafter referred to as biological feedstocks, or alternatively, fatty acids and/or glycerides depending upon the composition of the feedstock). Specifically, the present invention relates to predominantly even carbon number 231 paraffin compositions in the C12-C24 range, and the catalytic</s>

- Also qualitatively, it is apparent that Claims and Description sections as input ids do best to generate summaries from Pegasus-large.

	Selected example: Fine-tuning on different datasets
Abstract	A process for producing methyl methacrylate, the process comprising contacting reactants comprising methacrolein, methanol and an oxygen-containing gas, under reaction conditions in the presence of a solid catalyst comprising palladium, bismuth and at least one third element X selected from the group consisting of Fe, Zn, Ge, and Pb, wherein the solid catalyst further comprises a support selected from at least one member of the group consisting of silica and alumina.
Fine tuned Pegasus-large - Description	[<pad>, A, method, for, producing, meta, -, x, yl, ylene, di, iso, cyan, ate, of, the, present, invention, includes, a, reaction, step, in, which, mono, hal, ogen, ated, , benzene, ,, formaldehyde, s, ,, and, an, , amide, compound, represented, by, general, formula, (1), below, are, allowed, to, react, in, the, presence, of, an, acidic, liquid, to, produce, a, bis, amide, compound, ,, a, de, hal, ogen, ation, step, in, which, in, the, bis, amide, compound, ,, the, halogen, atom, derived, from, the, mono, hal, ogen, ated, , benzene, is, replaced, with, a, hydrogen, atom, ,, and, a, hydrogen, </s>, ...]
Fine tuned Pegasus-large - Ext-Sum-CD	[<pad>, The, present, invention, relates, to, a, process, for, producing, methyl, meth, acrylate, (, MMA,), from, me, tha, cro, lein, (, MAC,),, methanol, ,, and, oxygen, ., The, process, is, capable, of, producing, und, es, ir, ably, high, amounts, of, methyl, for, mate, as, a, by, -, product, ., The, present, invention, relates, to, a, process, for, producing, methyl, meth, acrylate, (, MMA,), from, me, tha, cro, lein, (, MAC,),, methanol, ,, and, oxygen, ., The, process, is, capable, of, producing, und, es, ir, ably, high, amounts, of, methyl, for, mate, as, a, by, -, </s>, ...]

- In the above example, despite ROUGE scores it seems that the summary produced through the hybrid approach is a strong contender, whereas the summary produced by the description alone claims we are producing the incorrect molecule.

	Selected example: Data Leakage
Abstract	The present invention relates to an antibody construct comprising a first human binding domain specific for the extracellular part of the influenza envelope protein M2 (M2e) and a second domain specific for CD3. Moreover, the invention provides a nucleic acid molecule encoding the antibody construct, a vector comprising said nucleic acid molecule and a host cell transformed or transfected with said nucleic acid molecule or vector. Furthermore, the invention provides a process for the production of the antibody construct of the invention, a pharmaceutical composition comprising said antibody construct, a medical use/method of treatment relating to said antibody construct, and a kit comprising said antibody construct.
Description	<pad>FIELD OF THE INVENTION The present invention relates to an

Summary	antibody construct comprising a first human binding domain specific for the extracellular part of the influenza envelope protein M2 and a second domain specific for CD3. Moreover, the invention provides a nucleic acid molecule encoding the antibody construct, a vector comprising said nucleic acid molecule and a host cell transformed or transfected with said nucleic acid molecule or vector. Furthermore, the invention provides a process for the production of the antibody construct of the</s>

- The above makes it clear that the model has seen this exact abstract before.