Paper ET02

Content Authoring Automation for Productivity Gains, Improved Quality, and Lesser Submission Delays Using Advanced NLP Techniques

Kumail Fazal, ZS Associates, India Pvt. Ltd.
Debiprasad Roy, Allogene Therapeutics
Ishant Arora, ZS Associates, India Pvt. Ltd.
Nafisa Ali, ZS Associates, India Pvt. Ltd.

ABSTRACT

Pharma organizations are increasingly investing in drug registration and regulatory authoring processes. The 2020 Biomedtracker report highlights that 64% of the total regulatory spend, amounting to $7B, is allocated to drug registration & regulatory authoring. In the regulatory domain, new content and data are seldom generated afresh and are often imported from the safety and clinical functions. However, manual sourcing of information is time-consuming task, besides can potentially lead to quality gaps. This necessitates an automated authoring solution.

The paper focuses on a tool created by ZS Associates in partnership with Allogene Therapeutics that automates 75% of the time-consuming process of manually authoring an Investigator's Brochure (IB) document. This document usually takes three to four weeks by conventional process.

ZS team explored the advantages of using the tool to produce natural language textual content, summarize text and tabular data, extract entities, and incorporate user feedback to create a document version. The paper also examines the tool's methodologies and the user-driven configuration file, which enhances the tool's adaptability from the user's perspective.

The study concludes with benefits of this automated solution--reduced time and cost and minimized chances of quality gaps. Furthermore, the tool's user-friendly configuration file makes it easier to use and incorporate feedback from users thereby enhancing its overall effectiveness. Thus, this paper provides valuable insights into the development and application of automated authoring solutions, which can help improve regulatory compliance and reduce overall cost in drug development.

INTRODUCTION

This paper delves into the life of a medical writer and the four-step process for authoring a report. The four-step process of authoring new content includes picking the appropriate template, identifying what information needs to be populated in each section, mapping each section to source document(s) & assessing complexity of required information (as-is, transformed, new), and, finally, creating the formatted report.

For emerging pharmaceutical organizations, there are no standard templates available, making it more challenging to pick the right one.

Post the identification of information that needs to be populated, for each section, a mapping of source document(s) is done. The complexity of each section of the report is also identified during this mapping process. The complexity is defined based on how much identical the content is to the source document (eg, completely identical, requires some additional processing). These documents could be text documents or contain tables, listings, and figures.

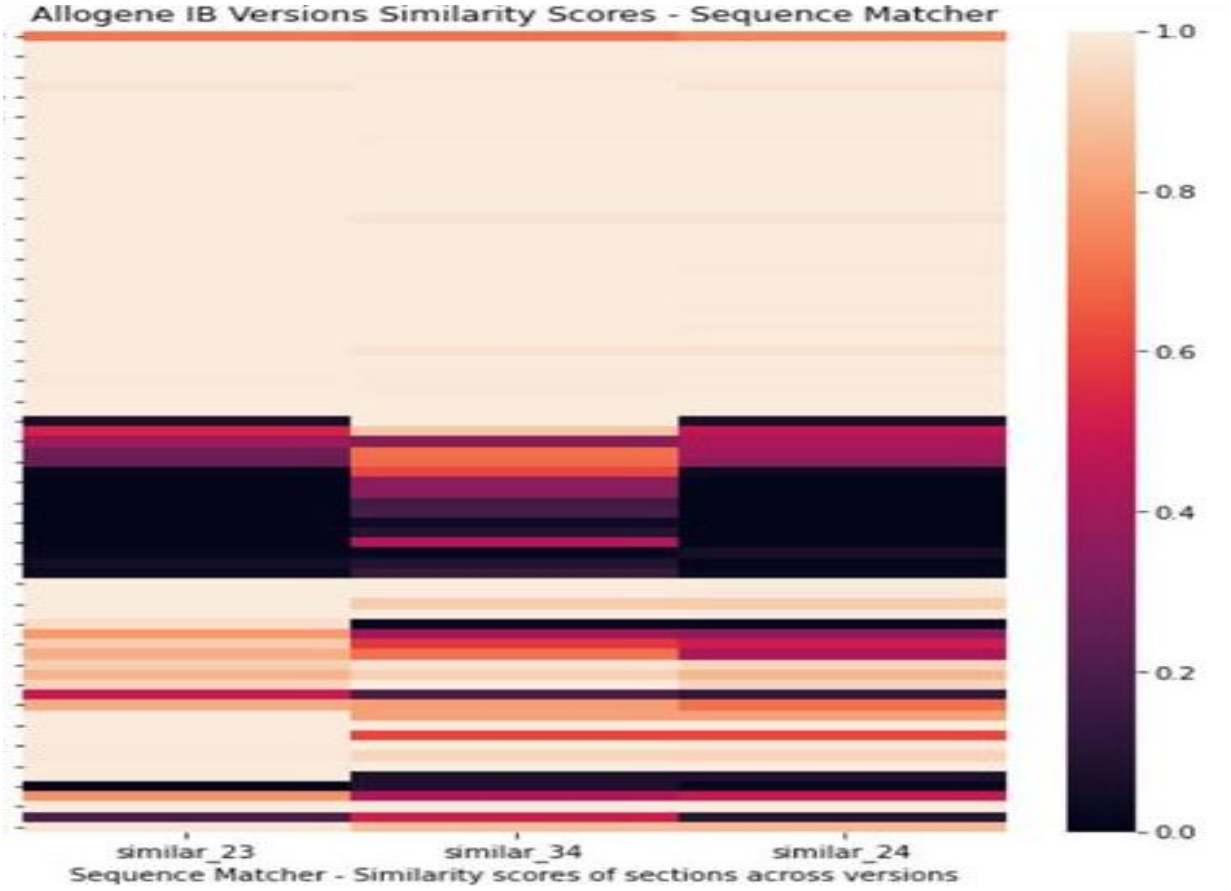
DISCOVERY PHASE

We focused on analyzing variations in data sources and consistency across Investigator's Brochure (IB) documents. We prioritized comparing textual content to calculate similarity scores across sectional content from different IB versions, excluding tabular and figure data due to inconsistency in numerical data. To feed data for similarity checks, the IB content was processed (removing extra white spaces, lowercasing, removing new line characters, tokenizing data), vectorized (Doc2vec from the gensim library for creating embeddings), & calculating cosine similarity.

The result is shown below:

*Fig 1: *Darker color - Lesser overlap*
Lighter color - More overlap

Legend:
similar_23 (comparison of version 2 & version 3)
similar_34 (comparison of version 3 & version 4)
similar_24 (comparison of version 2 & version 4)



Overview of results

Sections with high content overlap (similarity scores are higher than 90%):

- IB Sections: Summary, Introduction, Formulation and Nonclinical Studies

Sections with moderate content overlap (similarity scores between 60%-90%):

- IB sections: Exposure, Adverse Events

Sections with low content overlap (similarity scores are lower than 60%):

- IB Sections: Effects in Humans (AEs, Deaths, Lab Evaluations, Exposure, etc.), Summary of Data and Guidance for Investigator

The results redirected us to utilize benchmark templates and analyze existing standard content to identify sections that are either out of scope or particularly beneficial for medical writers. To create the benchmark template, the authors used 4-5 seed Investigator's Brochures (IB) documents with variations. The results provide a starting point for managing these sections and optimizing the authoring process for medical writers. This approach enabled a more streamlined and efficient process for authoring IBs, as it helps to identify sections that require additional attention and resources, as well as those that can be completed more efficiently.

IMPLEMENTATION PHASE

The four-step approach to automate the authoring of Investigator's Brochures (IBs).

1. Creating a benchmark document, which is especially relevant for emerging pharmaceutical companies without standard templates.
2. Identifying what information needs to be populated in each section
3. Mapping each section to source document(s) & assessing complexity of required information (as-is, transformed, new)
4. Finally, creating the formatted report.

Validation of the generated report was done. The common issues that arose (like missing information, irrelevant content, and discrepancies in statistics or milestones) were highlighted & resolved.

Configuration file

Based on the benchmark document, a user-configurable file is created that consists of all sections & corresponding information required during content automation process.

An example of a configuration file is shown below.

Benchmark Section ID	Benchmark Section Name	Data source	Level	Standardized text	Heading	Dependent sections
2.	SUMMARY	['Non-Clinical Overview', 'INTEGRATED OVERVIEW A	L1	No	No	No
2.	SUMMARY	['Protocol', 'Study Overview']	L3	No	No	No
2.	SUMMARY	['DISP', 't_disp']	L2- Fill in the blanks (from table)	As of the data cutoff, {TOTA	No	No
3.	INTRODUCTION	No	No	No	No content	No
4.1	Molecular Structure and Chemi	No	L2- Fill in the blanks (text)	{DRUG_NAME} Product Description:	No	No
6.2.2.1	Overview	No	L3	No	No	['6.2.2.2 ', '6.2.2.4 ', '6.2.2.6 ', '6.2.4 ', '6.2.2.3 ']

It consists of

1. Benchmark Section ID – The section number to be populated in the IB
2. Benchmark Section Name – The section name to be populated in IB
3. Data source – Data source and its corresponding section to refer for the content to be populated in IB
4. Level – Complexity levels (L1, L2 – Fill in the blanks (from table), L2 – Fill in the blanks (text), L3)
5. Standardized text – If a standard text has to be populated as a content
6. Heading – If the section is a heading or not
7. Dependent sections – If the section is dependent on some other sections within IB.

Alongwith this, the configuration file also includes source documents' path used during ingestion process. An example has been shown below.

Asset Code	Data Source	File name	Path
ALLO-316	Non-Clinical Overview	NA	316\NonClinical Overview\IND 026942 nonclinical-overview.pdf
ALLO-316	Protocol	NA	316\Protocol\ALLO-316-101_Monotherapy_Original_protocol_30Sep2020.pdf
ALLO-316	M3	NA	316\M3\ALLO-316 M3 Sections.pdf
ALLO-316	TLFs	t_demog	316\TLFs\2021\tlfs\t_demog.pdf
ALLO-316	TLFs	t_disp	316\TLFs\2021\tlfs\t_disp.pdf
ALLO-316	TLFs	t_teae_pt	316\TLFs\2021\tlfs\t_teae_pt.pdf
ALLO-316	TLFs	t_exp	316\TLFs\2021\tlfs\t_demog.pdf
ALLO-316	TLFs	t_teae_pt_rel316	316\TLFs\2022\tlfs\tlfs\t_teae_pt_rel316.pdf
ALLO-316	TLFs	t_sae_pt	316\TLFs\2022\tlfs\tlfs\t_sae_pt.pdf
ALLO-316	TLFs	t_lb_abnorm	316\TLFs\2021\tlfs\t_lb_abnorm.pdf
ALLO-316	TLFs	t_teae_pt_ge3	NA

Level of automation:

The degree of automation determines the level of difficulty required for mapping the content of each section from the data source. Each section that will be automated using the configuration file has an

established level of complexity (L1, L2 – Fill in the blanks (from the table), L2 – Fill in the blanks (text), L3), which the system uses to determine the approach it needs to take to automate the content.

Level 1 (L1): This complexity level lets the system copy the entire data source section content (exactly same way)

L2- Fill in the blanks (text): This complexity level lets the system use a standardized text (mentioned in the configuration file) and fill in the blanks with indicated values. This helps in maintaining the structure and consistency of the content across all the IB documents and their corresponding versions.

The values (keywords, entities, etc.) are extracted from the corresponding source documents using regex, pretrained models (Biomedical model NER trained: Stanza jnlpba), and localization methods.

For example: For the 'Molecular structure and chemical name' section, the following information was extracted:

▪ Product Description (extracted from the M3 document)
▪ Active Substance Definition (extracted from the M3 document)

L2- Fill in the blanks (from the table): This complexity level lets the system process tabular data to bring it into a text format using Camelot and Tabula Python libraries. The text data is then processed to retrieve the indicated values (indicated in the configuration file). This level also leverages certain Allogene-specific business rules and assumptions.

L3- Summarization: This complexity level lets the system perform extractive summarization on the entire content using transformer model. It summarizes the source document's corresponding section's content if there are no dependent sections within IB. If there are dependent sections within IB, then this section is performed at the end when all the dependent section(s) within IB are generated.

Validation/Cross-Verification

Content automated by the tool was thoroughly validated and examined by the ZS' & Allogene's subject matter expert (SME). Here are a few of the validation aspects that were looked into:

- Missing information on the cover page (Confidentiality Information)
- Missing section or content (Summary section)
- Content not relevant (Low similarity for content under sub-section)
- Discrepancy in content (stats, milestones) across interdependent sections

Version updates

At times, users may wish to change content in the auto-generated report and want it to be reflected into the future version. So, this tool has the flexibility to upload the one version of document and create a new version with modified changes (only for specified section(s)) incorporated.

Logs

The user can also keep a track of which section is automated (or, modified from the previous version) by using the tool's automation summary and stats feature.

Below is an example:

AUTOMATION SUMMARY

Number of sections	Status
filled	41
empty	2
no content	9
not automated	12

AUTOMATION STATS

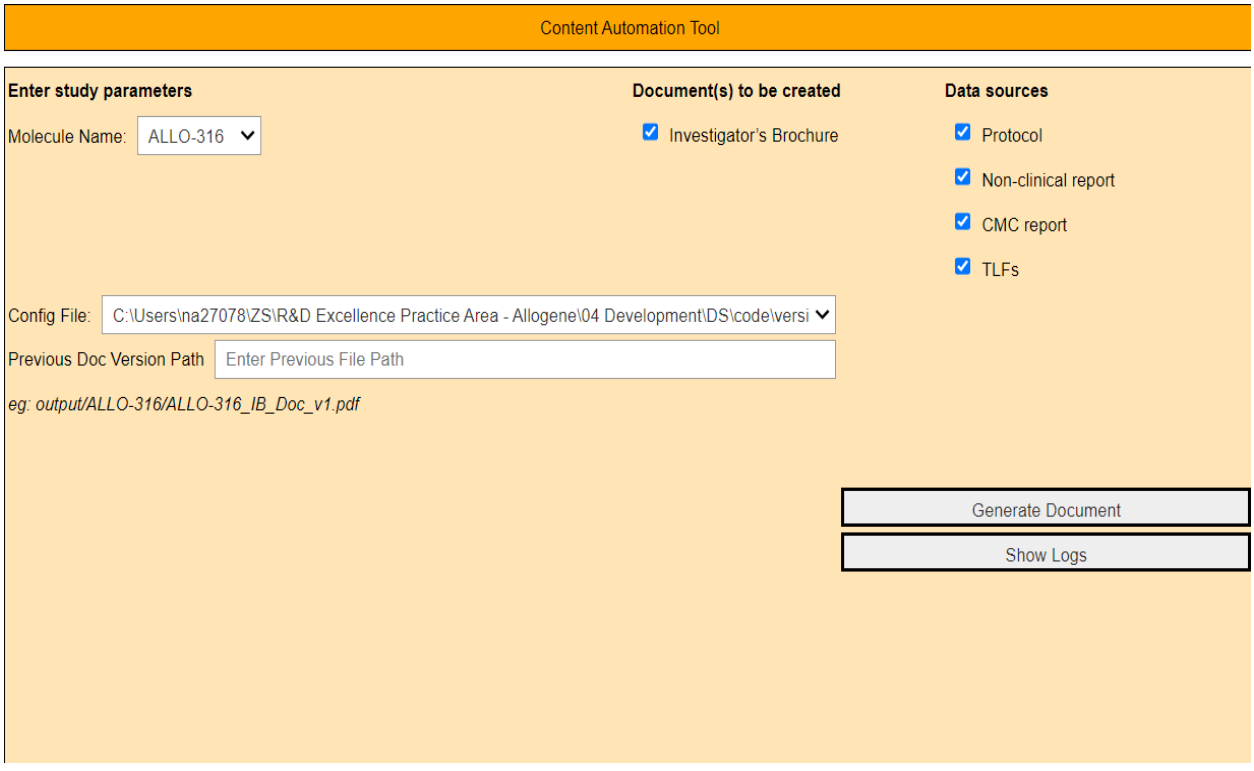
Section name	Status
2. SUMMARY	filled
3.1 Product description	filled
3.3 Clinical Plan	filled
4.1 Molecular Structure and Chemical Name	filled
4.2 Drug Product description	filled
4.3 Formulation	filled
4.4 Storage and Preparation	filled
5.1.1 Brief Summary	filled
5.1.1.1 Primary Pharmacology	filled
5.1.1.2 Secondary Pharmacology	filled
5.1.1.4 Pharmacodynamic Drug Interactions	filled
5.2.1 Pharmacokinetics	filled
5.3.1 Brief Toxicology Summary	filled
6.1 Overview of ALLO-316-101	filled
6.1.1 Disposition	filled
6.1.2 Demographics and Baseline Disease Characteristics	filled
6.2 Safety in Ongoing Study	filled
6.2.2.1 Overview	filled
6.2.2.2 Most Common TEAEs	filled
6.2.2.4 TEAEs Related to ALLO-316	filled
6.2.2.5 Safety Events of Interest	filled
6.2.2.6 Serious TEAEs	filled
6.2.4 Laboratory Evaluations	filled

[yellow] - Status change
[green] - Section addition

Output format

The tool generates a formatted word document (IB) as the output.

Snippet of the Content automation tool



CONCLUSION

The content automation tool not only can create content but also facilitate users to create new versions by updating only specific sections of previously generated report. This tool uses natural language processing techniques to process text, extract entities that are necessary for content creation, extract information from table, & summarize text. The tool's functionality is entirely driven by configuration file, making it highly adaptable to different use cases. It can provide valuable insights from tabular data and is completely customizable by users based on their specific report requirements.

By utilizing the aforementioned strategy, an advanced content automation tool can be customized to create various documents, including IBs and CSRs in an automated fashion, utilizing NLP techniques. By applying this technique, ZS and Allogene diminished the effort of manual authoring the IBs to the extent that approximately 60% of sections were automated. The tool is furnished with various functionalities such as data ingestion, data processing, content generation, and automated validation, all of which are established using the microservices architecture that allows adaptability & extendibility to other reports.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Author Name: Kumail Fazal
Company: ZS Associates
Email: kumail.fazal@zs.com

Author Name: Debiprasad Roy
Company: Allogene Therapeutics
Email: Debiprasad.Roy@allogene.com

Author Name: Ishant Arora
Company: ZS Associates
Email: ishant.arora@zs.com

Author Name: Nafisa Ali
Company: ZS Associates
Email: nafisa.ali@zs.com