A Deep Learning-based Virtual Assistant for better and faster Clinical Insights

Authors:

Nekzad Shroff – VP, Product Management

Aditya Gadiko – Director, Clinical Informatics

Acknowledgement - Madhup Pandey, Lead, AI Eng

nekzad.shroff@saama.com
aditya.gadiko@saama.com

www.saama.com

Table of Contents

I. The Challenge of Analytics in modern clinical trials

The cost and timeline for getting new drugs and therapies approved and available to patients has been exponentially increasing over the past decade. There are several reasons for this problem that are outlined as follows:

- Increasing complexity of science
- Increasing sophisticated protocols
- Time-consuming, manual and unscalable clinical development processes

A natural consequence of the information age is a “big bang” increase in the variety and volume of information and data points used in Clinical Development, which continue to grow at a dizzying pace. As the amount and complexity of available data has grown, it has become harder for clinical development personnel to scale existing methods of analysis to perform their job roles.

The challenges of applying traditional data analysis in modern trials include:

Scrolling, scrolling and more scrolling!

Clinical trials today collect operational and clinical data, which require development personnel to navigate and analyze in order to derive insight and make decisions. Each job role needs a different view of the same data in order to perform their job. For example, Study Managers need to aggregate study level metrics to manage the milestones of the trial while Medical Monitors need to analyze patient level data in order to ensure patient safety. The need for clinicians to have custom analytics and dynamic slices of larger, complex and diverse data sets causes traditional analytic approaches to fail for modern clinical trials.

Diversity of data sources

Clinical trials used to rely predominantly on patient data captured through EDC systems. However, modern decentralized and hybrid trials have changed that equation, with new distributed data such as e-Source, eCOA, ePRO and wearables playing a big part in data collection outside of traditional channels. Each trial protocol is different and the variety of data coming in from these new sources means that manual processing, custom visualization and ad hoc analysis of data by clinicians using traditional reporting tools cannot scale any more.

Manual data quality processes

Many labor hours are spent to clean data, that is, removing duplicate records, checking for rows and columns with missing values and ensuring clinical consistency. This is an important step in delivering accurate results but has predominantly required painstaking line by line review and checking by data managers.

As a result of the above challenges, pharma companies today are faced with a predicament of exploding timelines and resources needed to manage modern trials using outdated analysis techniques and infrastructure.

II. Can Artificial Intelligence help?

The above problems have grown over the years, but there has been historical skepticism and risk aversion of clinical development leaders to apply modern technologies such as artificial intelligence (AI) in the Clinical Domain. However, the unique challenges presented by the Covid-19 pandemic forced a change in mindset. Covid-19 increased the urgency of the pharmaceutical industry to bring therapies to market quicker despite the logistical challenges presented by the proposition of conducting trials amidst pandemic restrictions. This change in mindset turned the spotlight on AI's ability to automate and enable faster clinical development processes.

Virtual assistants are the face of AI. They are trained through Natural Language Processing (NLP) to match user text or voice inputs to executable commands. While virtual assistants such as Alexa and Siri have seen adoption in the consumer sphere to answer sophisticated questions, their inability to understand specialized clinical vocabulary has meant that they have not historically seen much application yet in this space. However, this is now changing.

At Saama, we have focused on building our Life Science Analytics Cloud (LSAC) suite of intelligent applications powered by AI to understand, interpret and automate the key processes in Clinical Development. One of the key innovations in the LSAC platform is DaLIA, our deep-learning powered virtual assistant specifically trained to understand and provide insights with clinical trial context.

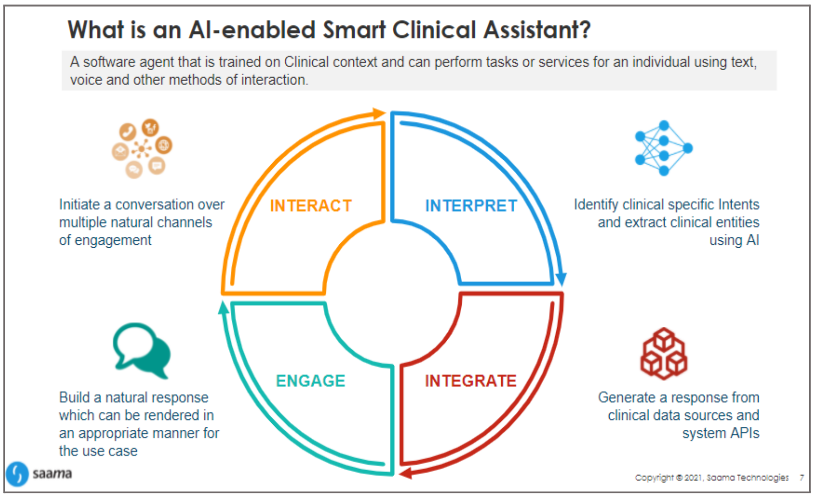


Figure 1: Process flow of an AI-based Clinical Virtual Assistant

Some key advantages of clinically aware virtual assistants over traditional analytics include:

- Accelerating data cleaning – Machine learning can be trained to eliminate duplicate and missing chunks of data without needing as much manual involvement.
- Understanding context - Assimilating past data and user context, virtual assistants can directly understand complex clinical queries vs traditional analytics which would require a lot of clicking and filtering.
- Providing "suggestions" – Based on historical trends and understanding of patterns which have been shown of highest concern or value to the user.
- Train of thought – Focusing on questions asked, pulling up similar or related queries that can seem relevant.
- Auto completion - Through frequent querying, the AI tool is able to grasp on clinical keywords to suggest relevant queries or can complete the user's query or sentence.
- Accessibility – The virtual assistant can be accessed from any smart device.

In this paper, we explore the use of an AI-enabled conversational virtual assistant for enabling and assisting clinical development personnel to automate manual tasks and speed up clinical operations and medical monitoring processes in a live clinical trial. We'll look at key areas where virtual assistants are best suited, and we'll explore the techniques by which they provide clinical insight and speed up the human workflow.

Intelligent use of virtual assistants in business processes have promise in getting better and faster insights from rich clinical data, by understanding complex clinical queries, and presenting targeted responses through powerful AI which performs automated computations in the backend.

III. Introduction to the DaLIA Smart Clinical Assistant

In this section, we will provide an overview of the process, architecture and logic behind the assistant with 'Clinical Intelligence' which understands plain English queries to perform tasks ranging from developing custom visualizations, predictive analytics, suggestions, risk identification, sending alerts and many more.

DaLIA is trained to understand and answer questions, in English, in the clinical development domain. It complements the power of packaged LSAC Analytics and Dashboards by allowing users to ask ad-hoc questions and get specific answers without browsing a specific dashboard or KPI.

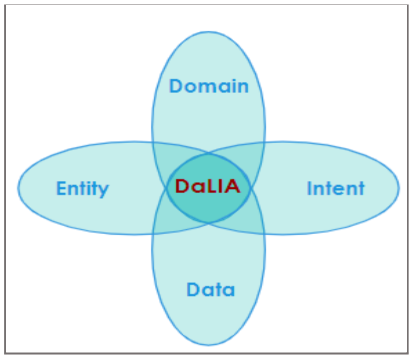
User interaction is framed in the following areas:

Domain: Focus on the study conduct domain of life sciences. Recognizing the domain, DaLIA establishes the context of your query.

Intent: Framing intent. Interpreting what you want to know [intent], that is, what, when and how many.

Entity: Identifying entities. DaLIA identifies the set of parameters [entities] needed to properly answer your question, such as, study names, site locations, time, and so on.

Data: With an understanding of the query, DaLIA uses the data resources available from your enterprise LSAC deployment.



IV. Components and Workflow of DaLIA

DaLIA works on a request-response workflow to interpret a user's plain English question and return an insight in the form of data or visualization to the user in seconds.

When user inputs are received through text or voice, DaLIA uses Natural Language Processing (NLP) to convert these **text and voice inputs** into executable commands using **interpretation** tools and algorithms from AI that identify intents and entities. Using the context and keywords established, DaLIA runs an **analysis** to gather relevant and insightful data from the vast available datasets and then delivers a **response** in the most natural way that is best understood by the user, displaying response through graphs, charts, tables or numbers. Finally, user **feedback** on the quality and usefulness of the response is used to further train the virtual assistant models.

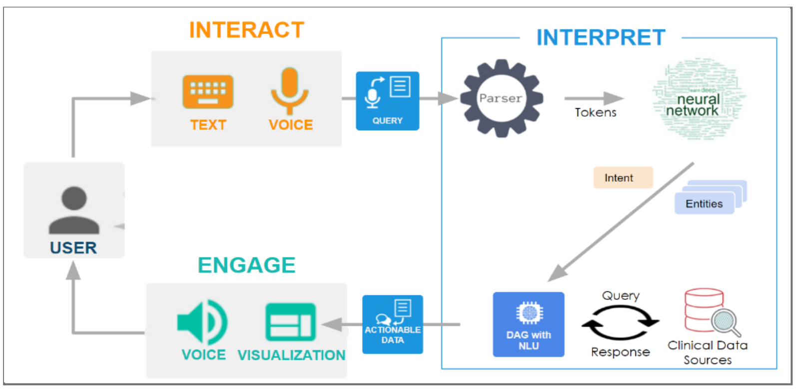


*Figure SEQ Figure * ARABIC 3: Request-Response flow of DaLIA*

A. Intent Parsing & Recognition

Intent recognition, also known as intent classification, is the task of taking a written or spoken phrase in natural language and classifying the input as an intent in the right context of the query.

Examples of Intent recognition:

- User input: *count of studies in USA*; Intent identified: **Count of Studies**
- User input: *countries for study S1000*; Intent identified: **List of Countries**

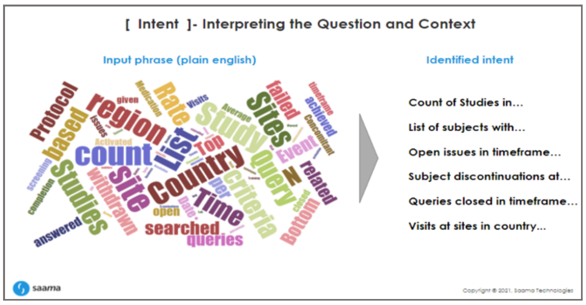


Figure SEQ Figure * ARABIC 4: Intent Recognition by DaLIA

DaLIA is programmed to recognize intents by combining the results of neural network-based models. To improve the accuracy of our neural networks models we used the Bagging technique, which is the ensemble learning method. It follows the parent-child intent relationship model, in which many child intents are recognized on the basis of the parent intent. The following three neural network algorithms are used to identify the parent intent.

a. DPCNN (Deep Pyramid Convolutional Neural Networks for Text Categorization) – https://docs.research.saama.com/docs/models/intent-extractor/
b. StarSpace - A general purpose neural embedding model that helps in text classification - https://docs.research.saama.com/docs/models/starspace/
c. ULMFIT (Universal Language Model Fine-Tuning for Text Classification) – https://arxiv.org/abs/1801.06146

To identify child intents in DaLIA, we use Python's *Difflib* library. This module provides classes and functions for comparing sequences.

The hyperparameters used in the above three models are tuned using Bayesian Optimization – *bayes_opt* and *hyperopt*. We have found that significantly higher accuracy is achieved on running data models multiple times using the Bayesian Optimization technique and each time the previous model's output is used to select the hyperparameters required to generate the new data model, thus promoting accuracy to a fine degree.

So far, the three algorithms, mentioned above, have been trained on 5000+ clinical sentences and comprise of 100+ intents.

B. Clinical Entity Extraction

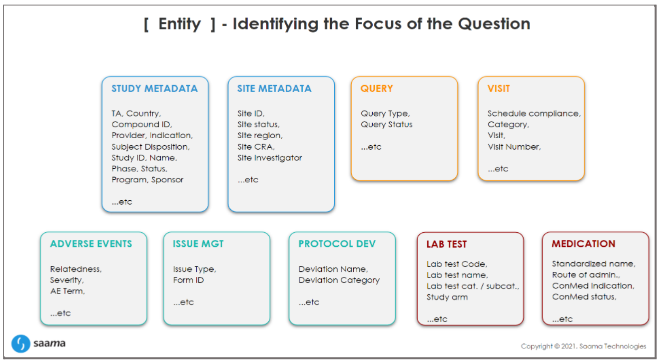
Entities are parameters that follow the intent in adding the context of the question. They are passed as parameters to form the query on the underlying data structures.

Examples of Entity Extractions:

- Phrase: "Total number of queries in oncology in the USA".
    - Entities extracted are: Therapeutic area - Oncology, country - USA.

- Phrase: “Count of enrolled subjects in study S1000”
    - Entities extracted are: Subject disposition - Enrolled, Study – S1000

Before the entity extraction, the user input undergoes a spell-check and synonym replacement. First, the sentence is split into n-grams (2 grams, 3 grams). Every string in grams is compared with the same gram entity/synonym words present. Four string-similarity methods facilitate the process of entity recognition in DaLIA. These are:



i. Jaccard – This is a token-based set similarity algorithm
ii. Ratcliff – This is a sequence-based algorithm based on the longest common substring.
iii. Jaro-Winkler – This algorithm gives high scores to two strings if, (1) they contain the same characters, but within a certain distance from each other, and (2) the order of the matching characters is the same.
iv. Difflib – This python module is used to compare the sequence of strings. The Sequence Matcher method will compare two provided strings and return the data representing the similarity between the two strings.

An Engineering Extractor technique is used to combine the above four algorithms based on confidence score, all synonyms/abbreviations/acronyms are replaced by the correct entity names. In addition, a Regular-expression technique is used to extract important keywords, patterns such as numbers, and words of comparison (more, less, equal).

C. Analysis

Using the clinical context acquired from the intent and the entities extracted, DaLIA then proceeds to apply that context to analyze the right sets of data from the standardized clinical datasets, thus forming an insight that would be clinically relevant to the user’s purpose. DaLIA uses a Directed Acyclic Graph (DAG) technique with intent interpretation and entity extraction for applying the right context and generating the right action query automatically on the fly.

D. Response

The insight formed from the action query is delivered to the user as a suitable response based on the specific use case and intent. For example, responses are rendered as tabular listings of data, graphic charts (bar chart/pie chart/ donut chart/scatter plot), a time series sequence or a sentence answering the specific question.

A data access security model is automatically applied so that the user is only allowed to see responses based on the access level granted to that specific user.

Additional analysis can be performed on the outputs such as application of filters, sorting, freezing columns etc to complete the analysis needed just like in an ad hoc analysis tool.

Adjacent “train of thought” suggestions: In addition to the response to the specific question, DaLIA also uses a K-means clustering algorithm to suggest related questions which the user could directly select to get further insight and continue the analysis train of thought.

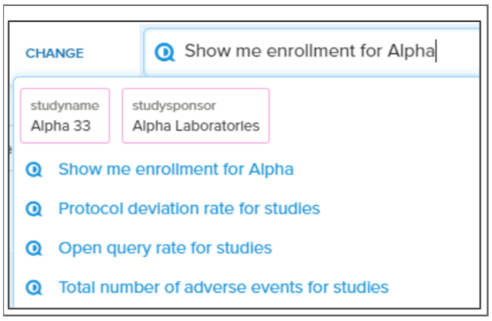
E. Feedback

Given the vast array of English sentences and responses, it's essential to monitor the performance of the virtual assistant in real world usage. DaLIA provides users with a "thumbs up" or "thumbs down" button to provide feedback on the usefulness and accuracy of the response, along with an optional comment which are used to adjust and retrain the model parameters periodically to improve accuracy.

V. DaLIA use cases and user experience

The user's DaLIA interaction experience can happen in 3 different formats:

I. Smart Search



DaLIA Smart Search is a common and intuitive interface to get quick answers by typing in a text query into a search box. The intent and entity recognition happens on the fly and are presented back as autocomplete suggestions to the user to aid in quickly finding the right question.

DaLIA's response is delivered in the form of a response card which could have a table, visualization or sentence as appropriate. As a search based interface, the context of DaLIA is reset after each search and DaLIA is ready for a fresh search.

II. Smart Clinical Bot

Figure SEQ Figure * ARABIC 7: Conversational flow with DaLIA

The Smart Clinical Bot user experience provides an ability for a clinician to "converse" with the clinical data. In this DaLIA conversational flow, the user types in a question and receives a response. DaLIA is then ready for the next question and keeps the context and keywords of the initial query, to drill in further or perform adjacent analyses. These sequential scrolling responses are well suited to "train of thought" type of deep dive analysis by users.

The conversational bot experience is implemented through a Directed Acyclic Graph (DAG) which maintains a network of processing elements for context.

<u>Example of Multi-step Conversation with DaLIA</u>

User: *Tell me enrolled subjects for study S1000 in USA*

User: *Tell me sites* (**Response:** This will show sites for study S1000 in the country USA)

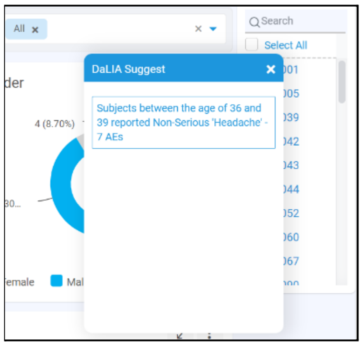
User: *How many are screened* (**Response:** This will show screened subjects for study S1000 in country USA)

User: *Tell me enrolled % and withdrawn percentage for site Rockefeller NY* (**Response:** This will show the percentage of enrolled and withdrawn for S1000 and site in country USA)

User: *Show me AEs over time* (**Response:** This will show the time series of number of adverse events recorded in study S1000, in site Rockefeller, NY, in country USA)

III. Smart Suggestions

Smart Suggestions is a prescription / recommendation based user experience for DaLIA. This is a powerful tool which reverses the workflow and makes the virtual assistant proactive instead of reacting to user requests.



Business case - Prediction of subjects at a risk of early withdrawal will help Medical Monitors/Scientists take a risk mitigative action if required.

In the Smart Suggestion concept, DaLIA uses machine learning models to detect patterns and trends of interest *in the background* and presents clinically relevant results as "smart suggestion" recommendations for the user to investigate further. This approach is particularly useful for datasets with large amounts of complex clinical data which would be hard for medical monitors to wrangle manually. This approach has the potential to significantly speed up the ability of the medical monitor to proactively identify cohorts or patients of concern (for example, we estimate based on the evidence that this speeds up the patient safety monitoring workflow analysis by 60-70%).
Methods used for suggestions

1. Pattern recognition: The objective here is to identify and capture succinct unique patterns in clinical data and to be able to detect risk factors in the early phases of study. In order to analyze and extract meaningful information from complex clinical data, a powerful computing tool is necessary. 'K-means clustering' is the technique we use to identify such patterns.

2. Statistical Correlation Analysis framework: Clinical data contains many potential clinically relevant correlations which might not be visible to the human eye manually. However, the Smart Suggestions are able to surface these signals proactively to medical monitors for quick actioning.

Sampling technique for Smart Suggestions:

To identify correlation in clinical domains, we need good quality data to feed the models. Most of the data is highly imbalanced and so a binary classifier model is used to predict all outcomes in favor of the majority class and to achieve an accuracy of 99%. In cases where class distribution is skewed, the accuracy metric is biased and not preferable.

Resampling of data in favor of minority classes needs to be done so as to not lose the actual information of both class labels. Synthetic Minority Over-Sampling Technique (SMOTE) is a solution we use for managing imbalanced data.

SMOTE is an oversampling technique where the synthetic samples are generated for the minority class. It focuses on the feature space to generate new instances with the help of interpolation between the positive instances that lie together.

First, we select the number of synthetic points to be generated and that is usually taken as 1:1 minority and majority class ratio but we tune that down further as per specific requirement. KNN is used to identify the synthetic samples. We choose k-neighbors of minor samples and create instances using any distance metric; the difference in distance between the feature vector and its neighbors is calculated.

There are cases when just oversampling is not enough to approach this problem and hence, we use a hybrid approach which first oversamples the minor samples and using Tomek reduces the majority class which does not lead to any information loss. This method is called SMOTETomek.

Once imbalanced data is handled, we have to choose a way of defining correlation which might differ as per the type of independent and dependent
features.

A. Categorical - Categorical correlation has been identified using Cramer's V algorithm
B. Continuous - Continuous has been identified using Pearson correlation coefficient
C. Categorical - Continuous relationship can be identified using binary classifier where dependent var is categorical, hence the more the predictor class is able to classify labels on test data, the better the correlation has been identified between variables.

All these target vs various domain correlation, discussed above, are identified individually and the best among them will be taken as combined input to classifier which will further identify the best input feature based on feature-ranking-by-feature importance metrics

VI. Conclusion

With recent advancements in AI/ML and Deep Learning techniques it is easier to find patterns/trends, correlation, predictions etc and provide better and faster insights into clinical data, however, a machine can never substitute human intelligence and the level of clinical context a physician or reviewer has. The goal is to leverage the strength of AI/ML and guide the end user to data of value faster so that the user can cut out the noise and focus on data to take definitive actions.

Contact Information

Saama Technologies, Inc.
900 East Hamilton Avenue Suite 200
Campbell, CA 95008
USA

Work Phone: +1 408-371-1900/+1 888-205-3500
Email: info@saama.com
Web: www.saama.com