Paper ML08

Feature of SDTM Mapping and Conversion Using NLP

Eric Weising, Symbiance Inc, Princeton Junction, USA
Shiva Katepalli, Sunovion, Fort Lee, USA
Sam Tomioka, Sunovion, Fort Lee, USA

ABSTRACT
SDTM mapping process is redundant and time consuming, which always required a CDISC-SDTM Analyst intervention while creating the mapping specification in order to meet regulatory requirement. Automating the mapping process by utilizing emerging technology such as Machine Learning will effectively reduce the manual efforts and increase the quality of conversion study by study. System which uses Natural Language Processing (NLP) logics with continuous learning, eventually will help to predict not only the standard domain mapping but also will learn the creation of custom domains and will perform those mappings automatically in similar situations. Outcome of using this emerging AI technology, is gearing up the standards mapping and encouraging the effective use of SME's time. This paper will discuss one such algorithm of Machine Learning which was implemented as a tool.

INTRODUCTION
Symbiance has developed – SMART SDTM Mapper – a Tool to Automate Mapping Specification Generation. This has been done utilizing Machine Learning (ML) and Natural Language Processing (NLP) to understand the raw metadata, generation of Target SDTM Metadata from the CRF Annotation, utilizing Mapping Algorithms to Generate the Mapping Specification, generating the programming need for conversion into datasets. The programming can be run on SAS® for SDTM dataset generation.

THE CURRENT METHOD
The current method of manually mapping the SDTM specifications contains numerous drawbacks and can result with mapping mistakes that could be identified and prevented with automation. Mapping is a repetitive process and there are endless mapping possibilities and multiple mapping decisions for the same scenario. Furthermore, there is no automated way to handle mapping errors and multiple iterations. The current manual process requires highly skilled resources to execute and is very time consuming, especially when mapping non-standard raw data, identifying data types and creating code lists.

THE BENEFITS OF AUTOMATION
There are numerous benefits in considering automation. One of the major benefits of automation is the time savings. Approximately 60%-80% of manual effort can be saved using the ML/NLP tool, and these time savings can increase even further as the user feeds more study data into the tool. In addition, the ML engine is constantly trained for greater accuracy and prediction, which will also contribute to time savings. Another benefit of automation is mapping consistency across studies. Automation creates efficient use and reuse of Metadata by using the hierarchy within the tool. Study Metadata can be escalated to the Project Metadata, so subsequent studies within the same project will be mapped consistently. Study Metadata can also be escalated further to the Global Metadata, if desired. In addition, the tool utilizes efficient reuse of mapping decisions. The percentage of fallouts are very minimal to handle manually, but any user action is stored in the tool and is available for subsequent studies. Another key benefit to the tool is mapping flexibility. The raw metadata can be mapped to either CDISC Metadata or Sponsor defined standards, including the metadata itself, the IG Version and Controlled Terminologies. The next benefit of automation is the generation of standalone SAS® programs, which will save significant time, but more importantly, will reduce errors that arise from manually creating the program or copying the program from another study and forgetting to change the logic, statement, derivation, etc. These above-mentioned benefits are all factors to consider for the automation of the mapping process.

HOW TO APPLY ML/NLP
While the actual implementation of the ML/NLP into the engine itself is complex, the steps in describing this process are straightforward. First, historical mapping data/specifications need to be gathered. This is the starting point for the programmers to build the mapping repository. Next is the data cleansing and pre-processing, where the stemming and lemmatization is performed to account for spelling variations of the same text. After that, is determining the classification algorithm to use to train the model, and then training the model and testing according to the required expectations. Once this is complete, the engine is ready for use on a new study. The user loads the raw metadata into the engine, the user can adjust any missing or inappropriate mappings (which are then used to train the model even further) and then the user can generate the mapping specifications.

STEPS OF AUTOMATED MAPPING

The following steps describe the simple process of generating the SDTM mapping specs and datasets within the tool we have created. First, the user obtains the raw metadata and ingests it into our tool. Next, the tool will identify the SDTM target domains from the raw metadata and will process the intermediate mapping specs (source-target). The user then reviews the intermediate mapping specs and provides any user-required intervention, such as non-mapped terms or inappropriately mapped terms. Any user action is stored in the engine to train the model and enhance the accuracy and predictability for subsequent use. After the intermediate mapping specs are complete, the user can proceed to process final mapping with derivations (final target). Once the user reviews the final mapping, the user can download the final mapping specs and generate the programming code based on the final mapping specs. The user then submits the programs to generate the SDTM datasets. The tool also allows the creation of the trial design data/metadata, which keeps track of all the metadata in one location.

DOMAIN MAPPING

The system is trained based on CDASH and CDISC standards, yet the system can also be re-trained based on Sponsor standards. The system will automatically predict the primary domains and secondary domains based on "Topic" variables collected in the raw metadata. The system will also provide the functionality to edit the domain prediction. Once changed, the system will be re-trained by itself and it will be available for further studies within the same project.

INTERMEDIATE MAPPING

Based on the keyword, the system will also predict the target variable and will also provide the possible matches. The user can also manually map the target variable or create the custom domain. The system will get re-trained based on the user changes and the mapping will be available for subsequent studies. The system will allow the user to provide mapping and derivation comments that are used in the specification and eventually the program generator.

FINAL MAPPING

The system will determine key variables and sorting suggestions based on CDISC. The user can modify these based on the data collection and these will be used in the generated program. The system will have standard derivation logic from both global metadata and user provided derivations. The system will automatically identify and transpose "finding" type variables.

ADDITIONAL FEATURES

The system will generate separate specs for --SUPP domains. The system will perform default consistency checks, like --DTC and --DY and can be configurable by the user requirement. The system will produce one program per domain; the primary domain program will have the functionality to create its --SUPP domain if needed. The system will use SDTM domain's variable order and attributes to meet the submission guideline.

CONCLUSION

Using ML/NLP helps to solve the problem with the historical data, can learn from the user actions and can predict new mapping scenarios. The SMART SDTM Mapping Tool converts an exhaustive manual process into a significantly automated process by reducing the time and effort spent by the user, remembering/retaining all the user edits and actions, improving the accuracy of mapping over time and creating both the mapping specifications and the conversion program.

CONTACT INFORMATION

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

Eric Weising
Symbiance, Inc.
51 Everett Drive, Suite B20
Princeton Junction, NJ 08550
609-945-7501
eweising@symbiance.com

Shiva Katepalli
Sunovion Pharmaceuticals Inc.
One Bridge Plaza North, Suite 510
Fort Lee, NJ 07024
201-228-8312
shiva.katepalli@sunovion.com

Sam Tomioka
Sunovion Pharmaceuticals Inc.
One Bridge Plaza North, Suite 510
Fort Lee, NJ 07024
201-228-8039
Sam.tomioka@sunovion.com