Digital Personalized Health and Medicine
L.B. Pape-Haugaard et al. (Eds.)

doi:10.3233/SHTI200165

Integration of Unstructured Data into a Clinical Data Warehouse for Kidney Transplant Screening -Challenges & Solutions

Maximilian ZUBKE[a,1] , Matthias KATZENSTEINER[a] and Oliver J. BOTT[a]
[a] *University of Applied Sciences Hannover, Hannover, Germany*

Abstract. After kidney transplantation graft rejection must be prevented. Therefore, a multitude of parameters of the patient is observed pre- and postoperatively. To support this process, the Screen Reject research project is developing a data warehouse optimized for kidney rejection diagnostics. In the course of this project it was discovered that important information are only available in form of free texts instead of structured data and can therefore not be processed by standard ETL tools, which is necessary to establish a digital expert system for rejection diagnostics. Due to this reason, data integration has been improved by a combination of methods from natural language processing and methods from image processing. Based on state-of-the-art data warehousing technologies (Microsoft SSIS), a generic data integration tool has been developed. The tool was evaluated by extracting Banff-classification from 218 pathology reports and extracting HLA mismatches from about 1700 PDF files, both written in german language.

Keywords. NLP, image processing, information extraction, data warehouse, graft rejection, kidney transplant

1. Introduction

Due to demographic change an increasing number of severe kidney disease cases and, as a result, an increasing need for kidney transplants can be expected. Early detection of graft rejection is therefore of particular importance in the therapy of kidney transplanted patients.

The joint project “Screen-Reject: A lateral flow test for rejection diagnostics” focuses on innovative diagnostics for the aforementioned purpose and is stratified into three subprojects. The subproject "Screen-Reject: Clinical data warehouse for graft rejection diagnostics" of the project network is concerned with the provision of a clinical data warehouse (CDWH) as a starting point for the development of an expert system to support rejection diagnostics based on clinical data. [1]

In the course of the medical monitoring of patients, both structured treatment-relevant data and unstructured findings texts are recorded. These unstructured findings texts pose special challenges for data extraction.

[1] Corresponding Author, Maximilian Zubke, University of Applied Sciences Hannover, Expo Plaza 12, 30539 Hannover, Germany; E-Mail: Maximilian.zubke@hs-hannover.de.

1.1. Requirements of clinical practitioners and scientists

Our clinical partners are especially interested on the Banff classification [2,3] as well as several information from examinations of human leukocyte antigens (HLA). A requirement analysis resulted in the following four requirements for the CDWH:

- Obtaining the distribution of Banff classification for all the patients included in our study.
- Selecting patients by Banff classification and coding results.
- Estimating the most common HLA mismatches.
- Selecting donor/receiver pairs having certain HLA mismatches.

However, since that information are only described in narrative instead of structured reports, it is not yet possible to select certain variables and filter them by value.

Thus methods are needed to extract the relevant information from the narrative reports.

2. Method

2.1. Data Staging

Initially, reports from the local pathology lab and the clinic for transfusion medicine were acquired. While the pathology reports could be provided as plaintext block from another clinical data warehouse [4], the findings from transfusion medicine were only available as PDF files. Thus the data staging environment have been extended by a data lake. Any information that could not be imported directly into the data warehouse was temporarily stored in the data lake.

2.2. Extracting Banff classification from staged plaintext blocks

The extraction, transformation and loading of defined information from narrative plaintext was realized by a SSIS[2] package. This package requires a source table containing plaintext resp. file paths to text data and processes the rows sequentially by a JSON-based communication with external text mining functionality. In detail, one or several external text mining tools can be connected with the SSIS package by corresponding script tasks. Finally, the data provided by the script task is written to previous designed tables of the data warehouse. Figure 1 illustrates this workaround.

[2] SQL Server Integration Services (https://docs.microsoft.com/en-us/sql/integration-services/sql-server-integration-services?view=sql-server-ver15)

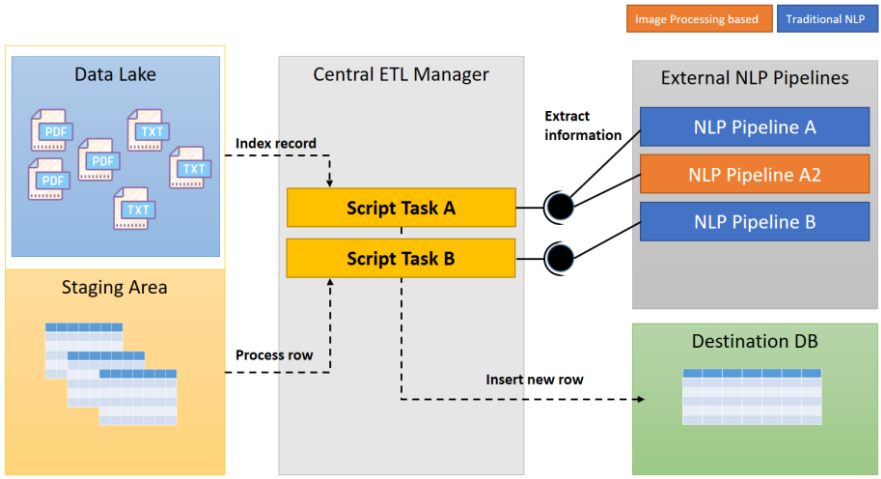

Figure 1. Workflow of the SSIS based ETL package: A central component coordinates the data flow and communicates with external NLP Pipelines via corresponding SSIS Script Tasks.

In the present case, the determined Banff categories [2,3] including optional comments as well as the Banff coding [2,3] should be extracted from pathology reports.

Considering that, two script tasks have been created, getting the same input but interacting with different NLP pipelines (one for each information to be extracted) and sending the output to two different tables in the destination database. The NLP pipelines can be any executable program which generates structured JSON output that is interpretable by the corresponding script task. In our use case, all NLP pipelines where represented by compiled python3 scripts. The information extraction functionality was realized using regular expressions.

2.3. Reverse engineering of HLA mismatch table from unstructured PDF files

In contrast to pathology findings, reports about HLA mismatches could not be made available in form of plaintext blocks, provided by another data warehouse. Instead, the only available descriptions were PDF files generated by third-party software or a scanner.

For this reason, the first step was the transformation of PDF files to text files. A preprocessing step that converts each PDF file to a text file without losing information like headlines or line breaks has been integrated. However, this approach was not able to recognize tables: Visual lines that separate rows and columns were not extracted and thus the whole table was transformed to a text block, whereby internal borders between rows and columns as well as external borders to previous and following document content. Even conversions to Word Documents or other formats failed in most cases.

However, from a visual perspective, tables are obvious and clearly delimitable parts of a document. Due to this, we decided to read the table of mismatches by image processing instead of natural language processing. Therefore, the PDF is considered as an image and the first step of this approach is the detection of horizontal and vertical lines [5]. Next, intersections of these lines are calculated. Thereafter, reading the document from left to right, each recognized intersection marks the border between columns and horizontal lines having intersection with the same vertical lines mark the rows.

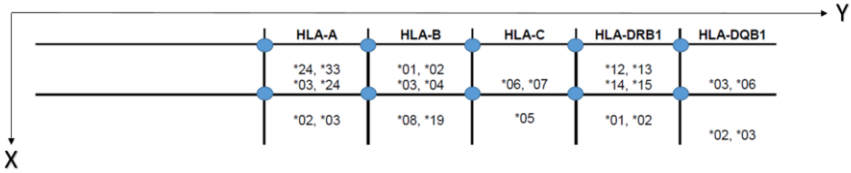

Figure 2. Sample image of a HLA table: Cell values are antigen stereotypes that extend the HLA name in the headline of the same column. Each row describes a sample from a patient.

Finally, the image regions surrounded by the intersection points become extracted and analyzed by an OCR component, OCR errors are prevented by comparing the results with the plaintext of the PDF. Geometric coordinates about the position of lines and intersection points are used to associate regions with each other and deriving the semantic information; For example, in Figure 2 the cell containing "HLA-A" and the cell containing "*02,*03" are surrounded by the same vertical lines and located in the same geometric region on Y-axis. Image regions having the same coordinates on the X-axis are interpreted as fields of the same row.

Overall, using visual features instead of linguistic or keyword based approaches, works fine to recognize the sematic structure of a mismatch table within a PDF-based report.

The extraction of the remaining content of interest, which is represented as a section or list is done by the same approach as described in section 2.2.

3. Result

As illustrated in Figure 3, the requirements mentioned in the introduction were fulfilled by extending ETL-pipelines with methods from natural language processing and image processing.

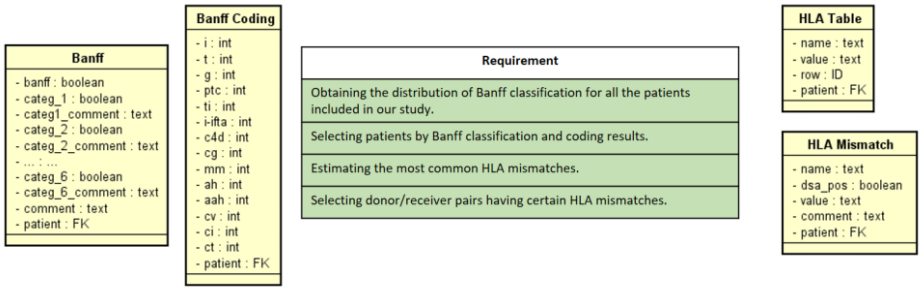

Figure 3. Requirements (green) were fulfilled by providing detailed information tables (yellow).

Technically, the described process was implemented by a generic SSIS package, which can be reused for further, similar data integration projects.

3.1. Evaluation

The described data integration approach was used for 218 pathology reports and 1700 HLA reports, both in German, given by the source dataset provided for this research project.

The correctness of our ETL approach was checked by manual verification of the mapping results. Our approach reaches a precision of 100% and a recall of 100% for pathology reports and 85% for the HLA reports. The latter recall of only 85% has been caused by unexpected document structures. After appropriate modifications, also the NLP Pipeline for the HLA reports reached a recall of 100% and finally all requirements were accomplished.

4. Conclusions

Even this work has confirmed that clinicians tend to document crucial information in form of free texts. Thus, clinical information systems shall be able to extract information of interest from such documentations and transform it into structured representations required for secondary use in research and care. The transformation of information from free text to structured formats is typically done using natural language processing. However, it has been shown, that methods from image processing can outperform traditional NLP in some specific tasks like the extraction of coherent document parts which are marked visually like e.g. tables.

5. Acknowledgement

We want to thank the center for information management of the hanover medical school for providing servers and enabling a connection to their clinical data warehouse.

The research for this paper was done within the joint project “Screen Reject”, founded by the European Regional Development Fund (ERDF) and the State of Lower Saxony, Germany.

References

[1] Screen Reject: Subproject 3: Clinical data warehouse for NTx rejection diagnostics, http://screen-reject.f3.hs-hannover.de/about/ [cited 2019 April 11].

[2] K. Solez, R.B. Colvin, L.C. Racusen, M. Haas et al., Banff 07 classification of renal allograft pathology: updates and future directions, *American journal of transplantation* **8** (2008), 753-760.

[3] C. Roufosse, N. Simmonds, M. Clahsen-van Groningen, M. Haas et al., A 2018 reference guide to the Banff classification of renal allograft pathology, *Transplantation* **102** (2018), 1795–1814.

[4] S. Gerbel, H. Laser, N.Schönfeld and T. Rassmann, The Hannover Medical School Enterprise Clinical Research Data Warehouse: 5 Years of Experience*, International Conference on Data Integration in the Life Sci*ences, *Springer* (2018), 182-194.

[5] R.O. Duda and P.E. Hart, Use of the Hough transformation to detect lines and curves in pictures, *Communications of the ACM* **15** (1971), 11-15.