Processing of authentic TNM classifications for automated evaluation

P. Hossner[1], I. Shumeiko*[1], C. Spreckelsen[2], E. Dahl[1] and P. Leusmann[1]
[1]Institute of Pathology [2]Institute of Medical Informatics

GEFÖRDERT VOM Bundesministerium für Bildung und Forschung

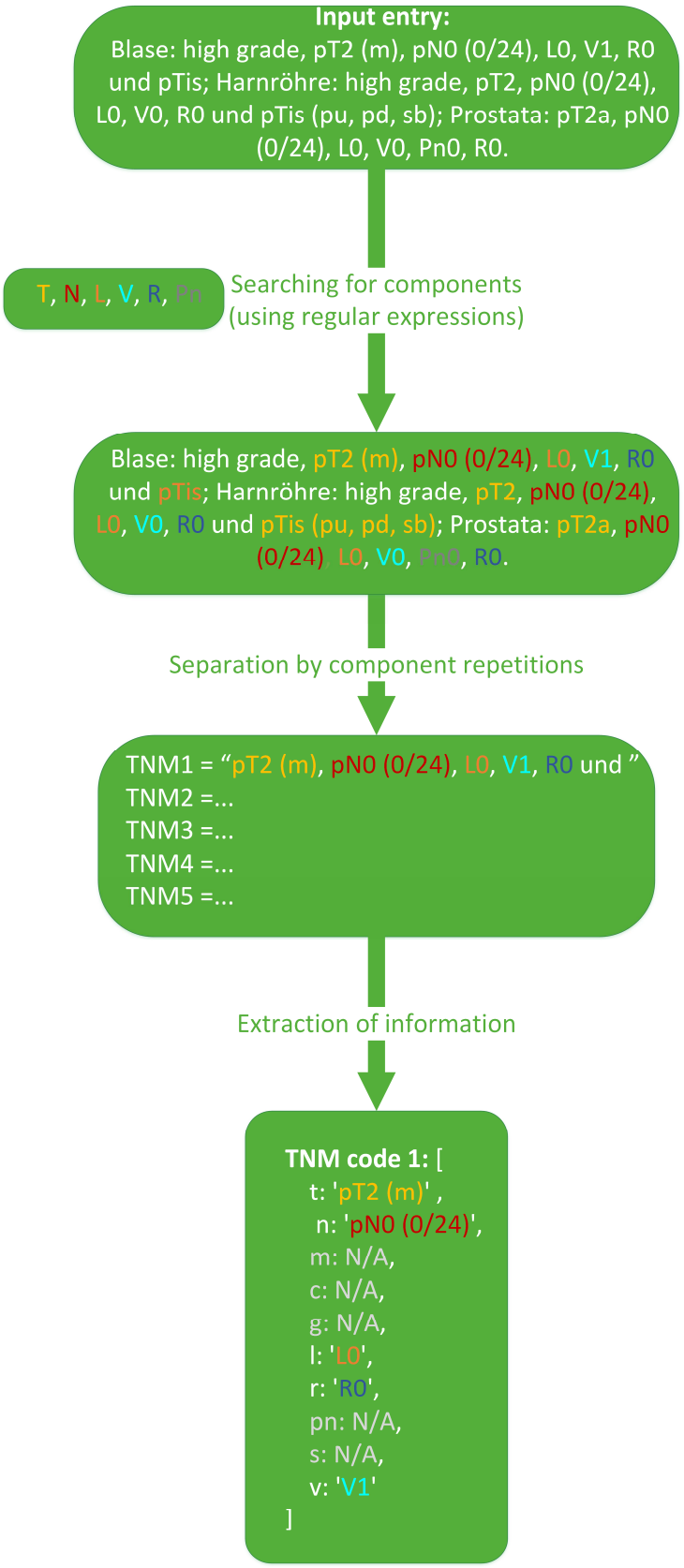

Fig. 1: Workflow of the implemented parsing procedure. Input entry processed with help of regular expression to detect all the TNM components contained in the entry. That make possible to delimit boundaries of TNM codes contained in the entry. After separation each found TNM code is stored in a TNM object. The object contains well structured information on the parsed TNM code.

BACKGROUND

In Pathology TNM codes are regularly used to encode cancer stages of cancer tumors. It is defined during microscopic examinations of tumor samples by pathologists. Pathological records are imported to clinical information systems to enable data search and analysis.
Unfortunately, this information is not standardized due to different styles and structures of pathological reports, e.g. some entries contain more than one TNM code without clear delimiters, making automatic separation difficult. Thus data must be cleaned and normalized before it can be imported to outside analytical systems.

AIM

This work aims at creating a reliable method to extract TNM codes from an input string of arbitrary length. The TNM components must be recognized correctly and grouped together into the TNM code sets without changing the TNM semantics. The output should provide the code in a machine readable format to allow further processing or queries.

METHODS

D'Avolio et. al. extracted T, N and M stages of TNM codes using regular expressions (REs) from medical reports. This good idea was improved in our work to parse not only for T, N and M stages but also the seven supplementary components described by the UICC (Union internationale contre le cancer). Apart from stage identifiers, each component provides additional properties of the tumor containing material, that were mostly ignored in D'Avolio et al. We developed an algorithm that can separate sequential TNM records for two or more samples stored in one entry. It discovers succeeding TNMs on basis of detecting TNM components, that are met in the inspected entry more than once, e.g. one entry contains two T components, indicating a presence of several TNMs.
A Java program was designed to perform parsing. As a result of the program execution, a list of TNM objects are obtained. Each object contains the extracted information about one TNM code and comprises also objects that represent information about the TNM components. This information includes additional properties of the investigated biomaterial and can be received by repeated parsing of TNM components. For that purpose several auxiliary regular expressions were constructed and applied. As mentioned, these properties were not taken into consideration in the work by D'Avolio et al.
The conceptual workflow of the program is represented in figure 1.

RESULTS

We tested our program on a set of 1600 pathological records. Each of the records could contain from 1 to 5 TNMs. The TNMs were automatically separated and their properties were parsed to be represented in an object oriented manner. A table with results in a printed form was produced and directed to a pathologist for assessment.

CONCLUSIONS

In most cases the pathologist confirmed correctness of the extracted information. Not covered entries were immoderately erroneous or ambiguous and hard for the pathologist to derive encoded information. This fact states a promising performance of the applied approach. We also prepared TNMs to import to the analytical system to provide a search on samples using attributes of TNMs as search criterion in ontology based data warehouse I2B2.

Disclosure Statement of Financial Interest: I, I. Shumeiko DO NOT have a financial interest or affiliation with one or more organizations that could be perceived as a real or apparent conflict of interest in the context of the subject of this presentation