Applied Soft Computing 71 (2018) 199–212

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

A clustering based methodology to support the translation of medical specifications to software models☆

Francesco Gargiulo*, Stefano Silvestri, Mario Ciampi

Institute for High Performance Computing and Networking of National Research Council, ICAR-CNR, Via Pietro Castellino 111, 80131 Naples, Italy

ARTICLE INFO

Article history:
Received 6 September 2017
Received in revised form 8 February 2018
Accepted 25 March 2018
Available online 4 July 2018

Keywords:
Clustering
Medical Specification Document
HL7 CDA R2
Validation
Natural Language Processing (NLP)
Schematron

ABSTRACT

In this paper we propose a methodology to reduce the complexity to realize a software validation model, starting from medical specifications written in Italian natural language text. In order to obtain an automatic validation system it is necessary to manually translate the specification documents into software models. This task is long, tedious and error prone, due to the manual effort needed. To speed up this process and to reduce the errors that can occur, an important boost can be obtained from the grouping of the conformance rules belonging to the same pattern. Clustering algorithms can accomplish this task, but there is the need to know a priori the total cluster number, and this is not possible in this kind of problem. At this aim, we propose two innovative automatic cluster selection methodologies able to evaluate the optimal number of clusters, based on an iterative internal cluster measure evaluation. These approaches consider three different Vector Space Models (VSMs), two different clustering algorithms and the impact of the using the Principal Component Analysis technique. The experimental assessment has been performed on four different datasets extracted from the HL7 CDA R2 Italian language conformance rules specification documents, demonstrating the effectiveness of the proposed methodology. Finally, in order to compare the results of all possible configurations, we realized a non-parametric statistical analysis. The obtained results demonstrated the effectiveness of the proposed methodology for automatic cluster number selection.

1. Introduction and motivations

Nowadays almost all kinds of textual documents are natively produced in a digital form, while it is even in progress the digitalization of all existing not digital texts. Despite this large availability, it is not always easy to automatically process digital text data. One of these cases is Natural Language (NL) documents. Due to complexity, richness and variability of human language, many different methodologies have been developed to process NL, ranging from Natural Language Processing (NLP) tools, to Machine Learning (ML) and Deep Learning (DL) techniques. Big steps towards the machine readability and understandability of NL texts have been made, but many issues are still open.

In many applications the automatic processing of NL text is essential, for example when the size of data is very big or there are time constraints. On the other hand, when an high level of precision is also required like in medical domains, a human supervision is needed to ensure the desired level of correctness.

Automatic text processing requires a structured or a coded language to obtain best results [1]. Thus, an easy way to overcome all problems related to NL could be the substitution of NL itself with a coded or structured language, more suitable to machine processing. This solution is not the best approach when the documents are mainly oriented to human-readability and the use of a coded language is not possible because it is not easy understandable by not expert or trained people.

Examples of the aforementioned class of texts are the implementation guides and the conformance rules. These classes of documents are respectively used to describe the contents of a digital document and how this document must be built to be conformed to a specific standard. They are written in a language understandable by all people that must deal with these documents, not only by computer scientists or domain experts. In addition, they must

☆ This paper is an extended, improved version of the paper: Francesco Gargiulo, Stefano Silvestri, Mariarosaria Fontanella, Mario Ciampi, A Methodology to Reduce the Complexity of Validation Model Creation from Medical Specification Document, presented at the Special Session on Smart Medical Devices – From Lab to Clinical Practice (SmartMedDev) 2017 and published in: Proceedings of the 10th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2017) – Volume 5: HEALTHINF, Porto, Portugal, February 21–23, 2017, pp. 497–507. SciTePress 2017, ISBN 978-989-758-213-4.

* Corresponding author.
E-mail address: francesco.gargiulo@icar.cnr.it (F. Gargiulo).

https://doi.org/10.1016/j.asoc.2018.03.057
1568-4946/

be processed to define the methodologies needed to generate an automatic validation schema. Another examples of the same class of texts are the requirements documents, where the functionalities and constraints of a system are formally described. Even in this cases there is the need to translate all the NL descriptions in a more formal language, like Unified Modeling Language (UML), to verify their consistency and to assist the implementation of the system. Requirement Engineering studies the methodologies to translate NL requirements to a more suitable automatic processing representation.

The above examples are only a few part of all the applications where there is the need to deal with documents written in complex and semantically rich natural language text and to automatically process them with an high level of precision. From these two opposite needs a trade off arises: choosing a more schematic and coded language, more easily computer understandable but less human readable, or use a more complete and rich language that can better express and describe concepts, but it is more difficult to be automatically analysed.

As described in next Section 2, in literature many different solutions to this problem have been proposed, mainly adopting NLP and ML approaches, or alternatively defining techniques able to generate requirements text in a more structured form. Unfortunately, there are some cases, like the one presented in this paper, where the only available solution is a manual translation of the NL text into a coded language [2] before process them. Being this task very time consuming, tedious and eventually error prone, the development of methodologies able to speed up and to help the accomplishment of this work is an important research area. At this aim, we propose in this paper a cluster-based methodology, which exploits a novel automatic cluster number selection approach.

Our assessment is focused on medical domain. We know that health professionals must deal with the editing of many different official clinical documents, such as Patient Summaries, Laboratory Tests Reports and Medical Prescriptions. All of them are structured or semi-structured text documents and, furthermore, they even require the presence of certain data, like, for example, a doctor name, a date or a disease code. At least, their structure and content must respect official guidelines, established by law and often described within conformance rules. To ensure the standardization and interoperability of these documents many standards have been proposed, like the ones promoted by HL7, that not only can ensure the semantic and formal correctness of the digital version of these documents, but can even support an effective and reliable automatic processing and the interoperability between different systems [3]. Due to the importance of these tasks, the definition of the conformance rules is a long and critical process, that involves many specialists from medical, clinical, legal and computer science fields and, of course, the governments and health-care agencies. The results of their work are usually documents written in natural language, containing a set of conformance requirements rules that define the format and the content of each medical documents type. The rules description must be understandable by human and simultaneously automatically processable by validation systems. This is just one of the cases above described, where a NL text requires a precise, fast and reliable automatic processing.

In Italy, Agencies and Government have produced the conformance specifications documents[1] for the digital version of the Patient Summary, the Laboratory Medicine Report, the Hospital Discharge Letter and the Medical Prescription, which are actually part of HL7 International standards in the Italian context. To implement a complete and reliable validation software model for each rule listed in the standard, computer scientists and engineers must perform a long task, analysing the NL text in the conformance specifications documents [4]. This task can be performed only by an hand-made translation of each NL rule in a software model for validation purposes, using, for example, the Standard Schematron (ISO/IEC 19757-3:2016) [5], or other rule-based validation languages. A great boost in the realization of the validation schema can be obtained decreasing the number of software model assertions that has to be manually built. The grouping of the rules following the same pattern will accomplish this goal. In this way, the same assertion function can be applied to more rules, speeding up the development of the validation model.

In this paper we present a methodology able to reduce the complexity of the realization of a validation model for a NL implementation guide by automatically grouping the NL rules, exploiting clustering techniques on different kind of Vector Space Models (VSMs). The main contribution is the definition and implementation of two different automatic cluster number selection methods, assessed on HL7 Italy CDA R2, a set of conformance rules documents.

The paper is structured as following. The next Section 2 presents actual the state of the art research in automatically validation and clustering optimization areas. Section 3 describes the details of the proposed methodologies, able to group together the rules with the same pattern belonging to the same validation schema and to determine the optimal cluster number. Section 4 is devoted to detail the proposed architecture and the materials used in its implementation. Section 5 describes the experimental assessment used to evaluate correctness of the proposed methodology. Finally, in Section 6 it will be given the conclusion and considerations about the obtained results.

2. Related works

In biomedical domain text mining and text processing are very important tasks. The authors of [6] provided an overview of the methodologies developed to solve the specific issues of this field, mainly caused by the variability of Natural Language and the use of non-standardized formats. The processing of such data is very challenging, but machine learning approaches for text mining are needed due to the increasing volumes of unstructured text to be processed. A large number of linguistic approaches for biomedical texts have been developed, known as biomedical natural language processing (BioNLP) methods. The paper focus specifically on statistical methods (Latent Semantic Analysis, Probabilistic Latent Semantic Analysis, Latent Dirichlet Allocation, Hierarchical Latent Dirichlet Allocation, Principal Component Analysis, and Support Vector Machines), applying these methods to examples extracted from the biomedical domain and comparing the obtained performances. After this analysis, some open problems and future challenges of this research area have been highlight, like the difficult of capturing the semantic content of the text.

The automatic validation of NL conformance rules is one of the specific task of NLP and involves not only the biomedical domain, but many different research areas, ranging from Clinical information models to Requirements Engineering. In [2] they considered implementation guides, a specification mechanism used to define clinical information models and to describe the contents of Electronic Health Records. These documents contain all the constraints and rules that clinical information must obey, written in NL and typically oriented to human-readability. For those reasons, to allow their computer processing, a human user must reinterpret and manual transform them into an executable language such as *Schematron* or *Object Constraint Language* (OCL): this operation is usually long, hard and error prone. To solve these problems, the

[1] Implementation Guide CDA R2, downloadable from http://www.hl7italia.it/node/34.