Process Extraction from Natural Language Text

Patrizio Bellan 1,2 [0000-0002-2971-1872]

1 Process and Data Intelligence group, Fondazione Bruno Kessler, Povo (Tn), Italy
2 Free University of Bozen-Bolzano, Bolzano (Bz), Italy
pbellan@fbk.eu

1 Introduction and Problem Definition

Public and private organizations always seek to achieve high standardization and improve performance of their business processes. Having control over business times, costs, errors and redundancy is vital to survive from continuous business revolutions [18, 31]. Business Process Management is a discipline that aims to discover, analyze, and optimize business processes, typically represented in model diagrams. Unfortunately, the initial elicitation of a process model from documents is a time consuming and cost intensive operation, as argued in [15, 21]. Therefore, companies and the scientific community are interested in discovering novel algorithmic procedures to alleviate the initial creation of process models from documents.

The extraction of a process model from documents is a complex task since the analysis of the natural language description of a process may produce multiple interpretation. This task is made up of three main activities. Filtering uninformative sentences of the process description out, because not all the sentences represent a process element. Then, the extraction of the process elements described in the text takes place. Finally, process elements discovered have to be logically organized following the semantic conveyed in the process description. So, defining the logical succession of process model elements is the last challenge to tackle. However, not only each sentence can describe multiple process elements, but also each word can have multiple meanings. To determine the correct intended meaning and to map it into the corresponding process element implies considering these two aspect at once. Also, there is the need to take care simultaneously of the multiple linguistics levels (syntactical, semantics and pragmatics) as well as ambiguities due to natural language.

The common solution found in the literature to solve the problem of process extraction from natural language text relies on a two-steps transformation approach with intermediate representation. Here, the model is considered as a compound function in which the first function \( f_a \) extracts process elements from a text and populates the structured intermediate representation, while the second function \( f_b \) builds the process model from the intermediate representation. However, most contributions proposed in this area date back to several years ago and, hypothetically, they may be considered outdated, given the advances of Natural Language Processing (NLP) techniques in the last few years [22]. Maqbool et al. argue in [21] that current approaches may be not able to scale up to real world scenarios, highlighting the need of research in this direction. The limited data publicly available and the heterogeneous approaches proposed in the literature highlighted a lack of a fair comparative analysis among them by making hard
the evaluation of their strengths and weaknesses [22]. While the maturity of Event-log Process Mining is well established in literature [9], this research direction is still in an early stage of development [4, 21].

Among all the contributions, the work of Friedrich et al. [13] is still the state-of-the-art, as emphasized in recent surveys and overviews on this topic such as the ones presented in [4, 21]. These two papers also highlight that after almost ten years of research, process extraction from natural language text is a task far from being resolved. Therefore further research in this direction is needed with the purpose of improving the quality of the process model generation.

2 Literature Analysis

Since the beginning of this line of research, the most common approach to mine processes from documents rely on the use of lists of signal words (also called trigger words) together with rules in the form of either patterns or templates. Important works, selected on the basis of their impact in terms of citations, that follow this approach are the ones in [1, 2, 3, 5, 6, 7, 8, 10, 11, 12, 13, 14, 16, 17, 19, 20, 23, 24, 25, 26, 27, 28, 29, 30, 32, 33].

There are three different approaches adopted in the literature. The one proposed by Han et al. in [14] that aims to build a direct mapping between a process description and its corresponding formal representation via neural transformation. A second solution, proposed in [10], is grounded on the philosophy of learning by doing, that is, learning a mapping of process model elements from a textual process description through user’s feedback, incrementally. The third solutions, adopted in [3, 5, 8, 11, 12, 13, 16, 19], uses a two-steps transformation with intermediate representation approach in which in the first step, the process elements are identified in the text and memorized in a structured representation, then in the second step the corresponding process model is generated from the structured representation.

Regarding the type of process model generated, these contributions can be divided in two groups. The first group includes the solutions that generate an imperative process model, in which each process element has a fixed relation with the other process elements [6, 7, 8, 10, 11, 12, 13, 14, 16, 19, 24, 27]. The second group includes the only two contributions found in the literature that aim to generate a declarative process model in which the behavior of a process is expressed with constraints on the relation between its process elements [5, 20].

The comparison of all these cited contributions on the experimental evaluation level reveals that the proposed solutions were tested on different aspects of the discovery task, using different data sets and evaluated according to different metrics. This highlights a lack of uniformity in the evaluation step that makes a comparison of the proposed contributions rather difficult. An important cause of this can be found in the absence of a common benchmark for the evaluation and in the absence of a common testing data set to compare the works. If we focus on the metrics we can notice that the works analyzed exploit different measures. The works proposed in [3, 5, 7, 10, 11, 12, 19, 24] adopt information retrieval metrics to quantitatively evaluate the performance of the proposed systems on the quality of the elements extracted from the process textual description.
In [13, 14] a graph-based measure quantitatively evaluate the quality of the process model created by the proposed systems from the textual description of a process.

3 Limitations

The analysis of the literature reveals that this area of research is still in an early stage of development, with many challenges still unanswered. Focusing on the research gaps found the analysis of these contributions reveals that they present three main limitations.

L1 Limitations with the techniques adopted, because current contributions are highly tailored to data input considered. In fact, when the state-of-the-arts (imperative and declarative) where tested on a private data set of standard operative procedures (SOP) they were not able to produce any diagrams.

L2 Limitations with the data. The works presented in [5, 6, 7, 8, 10, 12, 13, 19, 27] all adopt the data set (or a subset of it) proposed by [13]. The problem with this data set was not validated and thus it is not a representative sample of the variety of real scenarios. A comparison of it with some SOP documents adopted in a factory revealed that they differ greatly in: (i) the number of uninformative sentences/relative clauses, (ii) an extensive use of abbreviations, (iii) technical words, (iv) rare words; (v) writing styles, and (vi) formatting style.

L3 Limitations with the metrics adopted to judge the quality of the proposed systems, because there is a lack of metrics that consider a wide range of possible errors, assigning different weight for each type of errors, at once.

4 Research Questions

Starting from these limitations I formulate the following research questions.

RQ1 How far can the adoption of a statistical machine learning approach be superior (in terms of error reduction) to the rule-based solutions?

In particular, RQ1.1 Can the adoption of statistical machine learning classifiers enhance the performance in discriminating process elements described in a text?

RQ1.2 Which are the most predictive linguistics features to extract from a text that allow to enhance the detection of process elements?

RQ2 Is it possible to use NLP models trained in other domains into this context without re-training these models on process descriptions?

RQ3 Can I propose new bench-marking procedures with new data and new metrics (that consider a wide range of possible errors, assigning different weight for each type of errors) to judge the quality of process extraction from natural language text correctly?

5 Initial Research Plan

A promising possibility to investigate RQ1 in the task of process extraction from natural language text could be the adoption of the two steps transformation approach with intermediate representation. This approach enables to tackle the linguistics challenges linked to process extraction incrementally and independently. In RQ1.1, process
elements detection task is considered as a classification problem. The integration of statistical classifiers in this framework could increase precision and recall of process elements extracted, rejecting false positive. A strong limitation of rule-based approaches is the impossibility of taking advantage of semantic embeddings vectors to represent the meaning conveyed by a process description. Semantic embeddings together with a statistical classifier should allow taking a possibly correct decision in those cases out of any rules or word lists. The resulting framework should have the necessary generalization abilities to deal with multiple possible scenarios and different writing styles. In RQ1.2, I would investigate the adoption of semantics embeddings as well as discover which are the other possible important linguistic features that can increase classification performance. The effectiveness of many linguistic features are still unexplored. Indeed, there is a gap in the literature regarding the most predictive linguistics features to extract to better detect (in a statistical setting) the process elements described. An example of unconsidered linguistic features are: Multi-Words-Expressions (MWE), verb classes, temporal expressions, and preposition super sense.

The costly problem of creating a good resource of labeled data to train statistical models (able to scale up in real scenarios) on, data augmentation techniques have to be investigated in order to expand data availability. The difficulties related to this type of data generation concern generating an artificial valid textual process description with different words, different phrase structures, and possibly a different writing style, without changing the semantic of the process model described. The study of this type of data augmentation would partially address L2.

In RQ2, I would investigate the possibility of handling linguistics phenomena also by leveraging models trained to solve similar tasks but in different contexts. Moreover, their integration allows to consider a broader sets of features in process elements extraction tasks. But, because target classes of pre-trained models (such as event classes of an event detection system) could differ from the ones needed in these tasks, transfer learning and domain adaptation techniques must to be considered. Together these point would address L1.

These research questions are all intended to solve the first sub-problem ($f_a$). They have to be addressed before the investigations of better solutions to the sub-problems of determine the logical succession of process model elements ($f_b$) can take place. This is so, because it is vital that process elements are correctly extracted in the input text; although the final diagrams will always be wrong.

In RQ3, I would tackle the lack of a real comparison of different approaches on a single real-world benchmark, because it should shed light on the real weaknesses and strengths of the different research contributions. The metrics proposed in the literature do not make a distinction between the possible kind of errors that can be generated. For example, focusing on activities and participants only, there are no metrics that try to quantify the amount of error if an activity is attributed to the wrong participant. In a real-world context this kind of error can have significant negative consequences. Thus, to fill the gap L3 and to also finally give uniformity to this research field, I would investigate more realistic metrics able to weight for importance the different possible errors related to process extraction from natural language text.
Bibliography


