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 feed-backs, 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 liguistic 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 in the original process description. 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

[1] de A. R. Gonçalves, J.C., Santoro, F.M., Baião, F.A.: Business process mining from group stories. In: Borges, M.R.S., Shen, W., Pino, J.A., Barthès, J.A., Luo, J., Ochoa, S.F., Yong, J. (eds.) Proceedings of the 13th International Conference on Computers Supported Cooperative Work in Design, CSCWD 2009, April 22-24, 2009, Santiago, Chile. pp. 161–166. IEEE (2009). https://doi.org/10.1109/CSCWD.2009.4968052, https://doi.org/10.1109/CSCWD.2009.4968052

[2] de A. R. Gonçalves, J.C., Santoro, F.M., Baião, F.A.: A case study on designing business processes based on collaborative and mining approaches. In: Shen, W., Gu, N., Lu, T., Barthès, J.A., Luo, J. (eds.) Proceedings of the 2010 14th International Conference on Computer Supported Cooperative Work in Design, CSCWD 2010, April 14-16, 2010, Fudan University, Shanghai, China. pp. 611–616. IEEE (2010). https://doi.org/10.1109/CSCWD.2010.5471899, https://doi.org/10.1109/CSCWD.2010.5471899

[3] de A. R. Gonçalves, J.C., Santoro, F.M., Baião, F.A.: Let me tell you a story - on how to build process models. J. UCS **17**(2), 276–295 (2011). https://doi.org/10.3217/jucs-017-02-0276, https://doi.org/10.3217/jucs-017-02-0276

[4] van der Aa, H., Carmona, J., Leopold, H., Mendling, J., Padró, L.: Challenges and opportunities of applying natural language processing in business process management. In: Bender, E.M., Derczynski, L., Isabelle, P. (eds.) Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018. pp. 2791–2801. Association for Computational Linguistics (2018), https://www.aclweb.org/anthology/C18-1236/

[5] van der Aa, H., Ciccio, C.D., Leopold, H., Reijers, H.A.: Extracting declarative process models from natural language. In: Giorgini, P., Weber, B. (eds.) Advanced Information Systems Engineering - 31st International Conference, CAiSE 2019, Rome, Italy, June 3-7, 2019, Proceedings. Lecture Notes in Computer Science, vol. 11483, pp. 365–382. Springer (2019). https://doi.org/10.1007/978-3-030-21290-2_23, https://doi.org/10.1007/978-3-030-21290-2_23

[6] van der Aa, H., Leopold, H., Reijers, H.A.: Detecting inconsistencies between process models and textual descriptions. In: Motahari-Nezhad, H.R., Recker, J., Weidlich, M. (eds.) Business Process Management - 13th International Conference, BPM 2015, Innsbruck, Austria, August 31 - September 3, 2015, Proceedings. Lecture Notes in Computer Science, vol. 9253, pp. 90–105. Springer (2015). https://doi.org/10.1007/978-3-319-23063-4_6, https://doi.org/10.1007/978-3-319-23063-4_6

[7] van der Aa, H., Leopold, H., Reijers, H.A.: Comparing textual descriptions to process models - the automatic detection of inconsistencies. Inf. Syst. **64**, 447–460 (2017). https://doi.org/10.1016/j.is.2016.07.010, https://doi.org/10.1016/j.is.2016.07.010

[8] van der Aa, H., Leopold, H., Reijers, H.A.: Checking process compliance against natural language specifications using behavioral spaces. Inf. Syst. **78**, 83–95 (2018). https://doi.org/10.1016/j.is.2018.01.007, https://doi.org/10.1016/j.is.2018.01.007

[9] van der Aalst, W.M.P.: Process Mining - Discovery, Conformance and Enhancement of Business Processes. Springer (2011). https://doi.org/10.1007/978-3-642-19345-3, https://doi.org/10.1007/978-3-642-19345-3

[10] Ackermann, L., Volz, B.: model[nl]generation: natural language model extraction. In: Gray, J., Kelly, S., Sprinkle, J. (eds.) Proceedings of the 2013 ACM workshop on Domain-specific modeling, DSM@SPLASH 2013, Indianapolis, Indiana, USA, October 27, 2013. pp. 45–50. ACM (2013). https://doi.org/10.1145/2541928.2541937, https://doi.org/10.1145/2541928.2541937

[11] Epure, E.V., Martín-Rodilla, P., Hug, C., Deneckère, R., Salinesi, C.: Automatic process model discovery from textual methodologies. In: 9th IEEE International Conference on Research Challenges in Information Science, RCIS 2015, Athens, Greece, May 13-15, 2015. pp. 19–30. IEEE (2015). https://doi.org/10.1109/RCIS.2015.7128860, https://doi.org/10.1109/RCIS.2015.7128860

[12] Ferreira, R.C.B., Thom, L.H., Fantinato, M.: A semi-automatic approach to identify business process elements in natural language texts. In: Hammoudi, S., Smialek, M., Camp, O., Filipe, J. (eds.) ICEIS 2017 - Proceedings of the 19th International Conference on Enterprise Information Systems, Volume 3, Porto, Portugal, April 26-29, 2017. pp. 250–261. SciTePress (2017). https://doi.org/10.5220/0006305902500261, https://doi.org/10.5220/0006305902500261

[13] Friedrich, F., Mendling, J., Puhlmann, F.: Process model generation from natural language text. In: Mouratidis, H., Rolland, C. (eds.) Advanced Information Systems Engineering - 23rd International Conference, CAiSE 2011, London, UK, June 20-24, 2011. Proceedings. Lecture Notes in Computer Science, vol. 6741, pp. 482–496. Springer (2011). https://doi.org/10.1007/978-3-642-21640-4_36, https://doi.org/10.1007/978-3-642-21640-4_36

[14] Han, X., Hu, L., Dang, Y., Agarwal, S., Mei, L., Li, S., Zhou, X.: Automatic business process structure discovery using ordered neurons LSTM: A preliminary study. CoRR **abs/2001.01243** (2020), http://arxiv.org/abs/2001.01243

[15] Herbst, J.: An inductive approach to the acquisition and adaptation of workflow models. In: Proceedings of the IJCAI'99 Workshop on Intelligent Workflow and Process Management: The New Frontier for AI in Business. pp. 52–57 (1999)

[16] Honkisz, K., Kluza, K., Wisniewski, P.: A concept for generating business process models from natural language description. In: Liu, W., Giunchiglia, F., Yang, B. (eds.) Knowledge Science, Engineering and Management - 11th International Conference, KSEM 2018, Changchun, China, August 17-19, 2018, Proceedings, Part I. Lecture Notes in Computer Science, vol. 11061, pp. 91–103. Springer (2018). https://doi.org/10.1007/978-3-319-99365-2_8, https://doi.org/10.1007/978-3-319-99365-2_8

[17] Koschmider, A., Reijers, H.A.: Improving the process of process modelling by the use of domain process patterns. Enterprise IS **9**(1), 29–57 (2015). https://doi.org/10.1080/17517575.2013.857792, https://doi.org/10.1080/17517575.2013.857792

[18] Leopold, H.: Natural Language in Business Process Models - Theoretical Foundations, Techniques, and Applications, Lecture Notes in Business Information Processing, vol. 168. Springer (2013). https://doi.org/10.1007/978-3-319-04175-9, https://doi.org/10.1007/978-3-319-04175-9

[19] Leopold, H., van der Aa, H., Reijers, H.A.: Identifying candidate tasks for robotic process automation in textual process descriptions. In: Gulden, J., Reinhartz-Berger, I., Schmidt, R., Guerreiro, S., Guédria, W., Bera, P. (eds.) Enterprise, Business-Process and Information Systems Modeling - 19th International Conference, BPMDS 2018, 23rd International Conference, EMMSAD 2018, Held at CAiSE 2018, Tallinn, Estonia, June 11-12, 2018, Proceedings. Lecture Notes in Business Information Processing, vol. 318, pp. 67–81. Springer (2018). https://doi.org/10.1007/978-3-319-91704-7_5, https://doi.org/10.1007/978-3-319-91704-7_5

[20] López, H.A., Debois, S., Hildebrandt, T.T., Marquard, M.: The process highlighter: From texts to declarative processes and back. In: van der Aalst, W.M.P., Casati, F., Conforti, R., de Leoni, M., Dumas, M., Kumar, A., Mendling, J., Nepal, S., Pentland, B.T., Weber, B. (eds.) Proceedings of the Dissertation Award, Demonstration, and Industrial Track at BPM 2018 co-located with 16th International Conference on Business Process Management (BPM 2018), Sydney, Australia, September 9-14, 2018. CEUR Workshop Proceedings, vol. 2196, pp. 66–70. CEUR-WS.org (2018), http://ceur-ws.org/Vol-2196/BPM_2018_paper_14.pdf

[21] Maqbool, B., Azam, F., Anwar, M.W., Butt, W.H., Zeb, J., Zafar, I., Nazir, A.K., Umair, Z.: A comprehensive investigation of BPMN models generation from textual requirements - techniques, tools and trends. In: ICISA. Lecture Notes in Electrical Engineering, vol. 514, pp. 543–557. Springer (2018)

[22] Riefer, M., Ternis, S.F., Thaler, T.: Mining process models from natural language text: A state-of-the-art analysis. Multikonferenz Wirtschaftsinformatik (MKWI-16), March pp. 9–11 (2016)

[23] Santoro, F.M., Borges, M.R.S., Pino, J.A.: Tell us your process: A group storytelling approach to cooperative process modeling. In: Proceedings of the 12th International Conference on CSCW in Design, CSCWD 2008, April 16-18, 2008, Nanyang Hotel, Xi'an Jiaotong University, Xi'an, China. pp. 29–34. IEEE (2008). https://doi.org/10.1109/CSCWD.2008.4536950, https://doi.org/10.1109/CSCWD.2008.4536950

[24] Sawant, K.P., Roy, S., Sripathi, S., Plesse, F., Sajeev, A.S.M.: Deriving requirements model from textual use cases. In: Jalote, P., Briand, L.C., van der Hoek, A. (eds.) 36th International Conference on Software Engineering, ICSE '14, Companion Proceedings, Hyderabad, India, May 31 - June 07, 2014. pp. 235–244. ACM (2014). https://doi.org/10.1145/2591062.2591193, https://doi.org/10.1145/2591062.2591193

[25] Schumacher, P., Minor, M.: Extracting control-flow from text. In: Joshi, J., Bertino, E., Thuraisingham, B.M., Liu, L. (eds.) Proceedings of the 15th IEEE

International Conference on Information Reuse and Integration, IRI 2014, Redwood City, CA, USA, August 13-15, 2014. pp. 203–210. IEEE Computer Society (2014). https://doi.org/10.1109/IRI.2014.7051891, `https://doi.org/10.1109/IRI.2014.7051891`

[26] Schumacher, P., Minor, M., Schulte-Zurhausen, E.: Extracting and enriching workflows from text. In: IEEE 14th International Conference on Information Reuse & Integration, IRI 2013, San Francisco, CA, USA, August 14-16, 2013. pp. 285–292. IEEE Computer Society (2013). https://doi.org/10.1109/IRI.2013.6642484, `https://doi.org/10.1109/IRI.2013.6642484`

[27] Silva, T.S., Thom, L.H., Weber, A., Palazzo Moreira de Oliveira, J., Fantinato, M.: Empirical analysis of sentence templates and ambiguity issues for business process descriptions. In: Panetto, H., Debruyne, C., Proper, H.A., Ardagna, C.A., Roman, D., Meersman, R. (eds.) On the Move to Meaningful Internet Systems. OTM 2018 Conferences - Confederated International Conferences: CoopIS, C&TC, and ODBASE 2018, Valletta, Malta, October 22-26, 2018, Proceedings, Part I. Lecture Notes in Computer Science, vol. 11229, pp. 279–297. Springer (2018). https://doi.org/10.1007/978-3-030-02610-3_16, `https://doi.org/10.1007/978-3-030-02610-3\_16`

[28] Sintoris, K., Vergidis, K.: Extracting business process models using natural language processing (NLP) techniques. In: Loucopoulos, P., Manolopoulos, Y., Pastor, O., Theodoulidis, B., Zdravkovic, J. (eds.) 19th IEEE Conference on Business Informatics, CBI 2017, Thessaloniki, Greece, July 24-27, 2017, Volume 1: Conference Papers. pp. 135–139. IEEE Computer Society (2017). https://doi.org/10.1109/CBI.2017.41, `https://doi.org/10.1109/CBI.2017.41`

[29] Thom, L.H., Lau, J.M., Iochpe, C., Mendling, J.: Extending business process modeling tools with workflow pattern reuse. In: Cardoso, J.S., Cordeiro, J., Filipe, J. (eds.) ICEIS 2007 - Proceedings of the Ninth International Conference on Enterprise Information Systems, Volume EIS, Funchal, Madeira, Portugal, June 12-16, 2007. pp. 447–452 (2007)

[30] Thom, L.H., Reichert, M., Chiao, C.M., Iochpe, C.: Applying activity patterns for developing an intelligent process modeling tool. In: Cordeiro, J., Filipe, J. (eds.) ICEIS 2008 - Proceedings of the Tenth International Conference on Enterprise Information Systems, Volume ISAS-1, Barcelona, Spain, June 12-16, 2008. pp. 112–119 (2008)

[31] Thom, L.H., Reichert, M., Iochpe, C.: Activity patterns in process-aware information systems: basic concepts and empirical evidence. IJBPIM **4**, 93–110 (2009)

[32] Thom, L.H., Reichert, M., Iochpe, C.: Activity patterns in process-aware information systems: basic concepts and empirical evidence. IJBPIM **4**(2), 93–110 (2009). https://doi.org/10.1504/IJBPIM.2009.027778, `https://doi.org/10.1504/IJBPIM.2009.027778`

[33] Thom, L.H., Iochpe, C., Reichert, M.: M.: Workflow patterns for business process modeling. In: In: 8th Int. Workshop on Business Process Modeling, Development, and Support (BPMDS. pp. 349–358 (2007)