Business Process Model Information Extraction from Documents Using In-Context Learning

Patrizio Bellan,[1, 2] **Mauro Dragoni,** [1] **Chiara Ghidini** [1]

[1] Fondazione Bruno Kessler
[2] Free University of Bozen-Bolzano
pbellan@fbk.eu

Abstract

The extraction of business process model information from textual documents is a research area that still lacks the ability to scale to a variety of real-world texts.

In this paper, we explore for the first time in literature the usage of GPT-3 and in-context learning to exploit the information contained in process model descriptions. We tackle the problem in a conversational fashion that mimics a dialog with a domain expert. GPT-3 is used to simulate a domain expert who answers questions. The answers correspond to process information that allows for building the process model diagram described. In particular, we explore how to tailor GPT-3 with prompt customizations that rely on conceptual definitions and/or few examples of the task to solve.

Our results demonstrate that in-context learning in a few-shot fashion can be used to extract process information from text.

Introduction

The extraction of business process model information from textual documents is a research area that still lacks the ability to scale to a variety of real-world texts. The actual exploitation of the information contained in process model descriptions is often hampered by having to manually analyze unstructured information. The exploitation of process information is made even harder if we consider the fact that several definitions exist of many business process elements, but they often present different wordings and even conflicting characteristics (Adamo, Di Francescomarino, and Ghidini 2020) In addition, the ambiguous nature of natural language, the multiple possible writing styles, and the great variability of possible domains of application make this task extremely challenging. Indeed, recent surveys on this topic (van der Aa et al. 2018; Bellan, Dragoni, and Ghidini 2021; Maqbool et al. 2018) highlight that after more than ten years of research from the seminal work in (Friedrich, Mendling, and Puhlmann 2011), extraction of business process models from text is a task far from being resolved.

By looking at state-of-the-art works, most of the existing approaches rely on templates and rule-based approaches, which often lack the flexibility needed to fully cover the great variability of writing styles and process domains (van der Aa et al. 2019; Ackermann and Volz 2013; Friedrich, Mendling, and Puhlmann 2011; Honkisz, Kluza, and Wisniewski 2018; Quishpi, Carmona, and Padró 2020). Few recent works (Han et al. 2020; Qian et al. 2020) try to leverage modern Natural Language Processing (NLP) approaches based on deep learning, but they (somehow ironically) restrict the format of the source text to a structured text (Han et al. 2020) or to sequential lists of tasks such as recipes or assembly instructions (Qian et al. 2020), thus avoiding the challenge posed by real-world business process descriptions.

One of the problems of leveraging the potential of deep learning NLP is the **lack of the high quantities of carefully annotated data on textual descriptions** needed to make these techniques work, which newly available annotated datasets, such as PET dataset (Bellan et al.), are not yet able to address. Recent advances in NLP, and in particular the availability of large pre-trained language models and the introduction of *in-context learning strategies* (Brown and et al. 2020), are opening up new perspectives on the construction of information extraction systems that support search and question-answering (among other tasks) using multi-turn dialogs in written or spoken forms.

Approach

In (Bellan, Dragoni, and Ghidini 2022b,a) we investigate the usage of the GPT-3 (Brown and et al. 2020) language model and in-context learning to address the problem of *business process model information extraction from process description documents*. GPT-3 has been proven to be able to deal with a variety of possible writing styles and domains while relying on a few annotated data. The vast knowledge the GPT-3 model has, together with its proven reasoning skills allowed us to deal with this low-resource scenario. For the first time in literature, we explore the feasibility of our approach in a conversational fashion by mimicking a dialog with a domain expert. GPT-3 is used to simulate the expert who answers questions. An example of the dialog script is proposed in Figure 1. We develop a pipeline script to extract process information incrementally. We explore how to tailor the model with prompt customizations (the input feeds into the GPT-3 model) that rely on the usage of conceptual definitions and/or a very limited number of examples.

The starting point is a set of process elements and their relationships, which we aim at extracting. We are interested in extracting process elements that constitute the building blocks of every process model: *activities*, *actors*, and

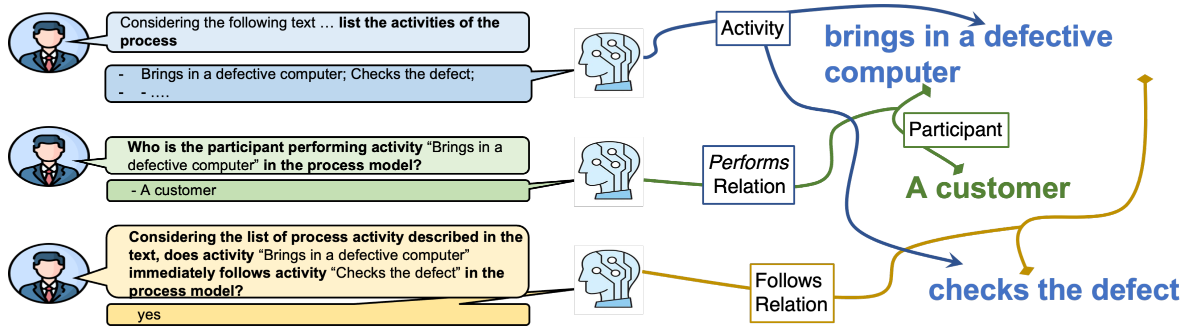

Figure 1: Example of multi-turn dialog script.

the *actor-performer* relation (that represents the relation between an activity and the actor/-s responsible for its execution). We extract also the order temporal relation between activities to reconstruct the process model logic described. We formulate a series of incremental questions posed to GPT-3 in a sequential manner that enables the extraction of those specific entities and relationships. These questions become the specific tasks that the model has to solve with the help of specific prompts. Questions do not necessarily correspond to elements in a 1:1 manner, and answers to a question *can* be used as inputs to formulate further questions. As shown in Figure 1, we start the dialog by asking which are the activities described in the text. Next, for each activity retrieved by the model, we ask the model to provide us who is the actor performing that activity. Through this question, we gather two types of information: the actors described and the performing relations. In order to reconstruct the process model logic, for each pair of activities extracted through the first question we ask the model to tell us if they stand in a follow relation or not.

We experimented with various prompt customizations using two types of information: (i) *contextual knowledge* and definitions of the process elements and relations we aim at extracting, and (ii) few shots *examples* of the task at hand. *contextual knowledge* should enable GPT-3 to identify the specific domain at hand (BPM in our case) and the elements to be extracted for the different tasks. We also include in our prompts the *task instructions* and the *process description text* upon which to perform the task. Once ready the prompts are fed into the model to generate the answer. We select two representative documents from the PET dataset to create the gold-standard examples to fill the example-component of the experimental prompts. The approach described has been empirically evaluated by performing a fine-grained analysis on a selection of representative documents from the PET dataset. Since this is a first observational study of a promising groundbreaking strategy, we decided to select documents having specific characteristics to perform an ad-hoc analysis of how the GPT-3 model worked on them. Hence, to better understand the impact of the information provided by us to enrich the language model, the most suitable way was to observe such behavior on a small but characteristic subset of documents.

Discussion

The results obtained highlight few interesting patterns. We can see that prompts that do not rely on in-context learning (without the example component) provide unsatisfactory results, thus highlighting that the native GPT-3 model is not able to address the task of extracting different process elements from text in a satisfactory manner. Prompts that rely on in-context learning approaches are more effective in providing answers than the native GPT-3 model and can lead to good results. Adding contextual BPM knowledge and definitions is useful in specific cases - thus hinting at a possible positive effect - but does not appear to be a valid general strategy, overall. The result we can report here is that adding the right contextual BPM knowledge may be a nontrivial problem that needs to be carefully investigated. An exception to the overall satisfactory performance of the in-context learning prompts without contextual knowledge is given by the Follow relation. While providing the two examples is useful to increase the performances w.r.t. the native GPT-3 model, the gain is often limited. This result may indicate that the knowledge of temporal relationships of GPT-3 is insufficient for the BPM domain and better ad-hoc training is needed. While our approach does not depend upon the particular process elements we extract, we have decided to use it for the extraction of *activities*, *participants*, the *performing* relation, and the sequence relation between activities. The incremental order of the questions is interesting because it can be used to mimic the way we often build conceptual models using follow-up questions. An important aspect in setting up the questions is their correspondence with the elements to be extracted and the specific wording to be used, that is, learning to ask the *right* question.

Conclusion

Our results demonstrate that in-context learning in a few-shot fashion can be used to extract process information from text, with the need to better investigate the extraction of the temporal relations. Finally, we observed how the performances are not related to the length of the documents. This aspect is particularly interesting since it means that the proposed strategy may be applied in different scenarios without considering the length of the document as a criticality to address.

References

Ackermann, L.; and Volz, B. 2013. model[NL]generation: natural language model extraction. In *Proc. of 2013 ACM workshop DSM@SPLASH 2013*, 45–50. ACM.

Adamo, G.; Di Francescomarino, C.; and Ghidini, C. 2020. Digging into Business Process Meta-models: A First Ontological Analysis. In *Proc. of CAiSE 2020*, volume 12127 of *LNCS*, 384–400. Springer.

Bellan, P.; Dragoni, M.; and Ghidini, C. 2021. Process Extraction from Text: state of the art and challenges for the future. *CoRR*, abs/2110.03754.

Bellan, P.; Dragoni, M.; and Ghidini, C. 2022a. Assisted Process Knowledge Graph Building Using Pre-Trained Language Models. In *Proceedings of the 19th International Conference of the Italian Association for Artificial Intelligence (AIxIA2022), Udine, November 28th, 2022*. To appear.

Bellan, P.; Dragoni, M.; and Ghidini, C. 2022b. Extracting Business Process Entities and Relations from Text Using Pre-trained Language Models and In-Context Learning. In Almeida, J. P. A.; Karastoyanova, D.; Guizzardi, G.; Montali, M.; Maggi, F. M.; and Fonseca, C. M., eds., *Enterprise Design, Operations, and Computing - 26th International Conference, EDOC 2022, Bozen-Bolzano, Italy, October 3-7, 2022, Proceedings*, volume 13585 of *Lecture Notes in Computer Science*, 182–199. Springer.

Bellan, P.; van der Aa, H.; Dragoni, M.; Ghidini, C.; and Ponzetto, S. P. ???? PET:.

Brown, T. B.; and et al. 2020. Language Models are Few-Shot Learners. In *Annual Conf. on Neural Information Processing Systems 2020, NeurIPS 2020*.

Friedrich, F.; Mendling, J.; and Puhlmann, F. 2011. Process Model Generation from Natural Language Text. In *Advanced Information Systems Engineering - 23rd International Conference, CAiSE 2011. Proceedings*, volume 6741 of *LNCS*, 482–496. Springer.

Han, X.; Hu, L.; Mei, L.; Dang, Y.; Agarwal, S.; Zhou, X.; and Hu, P. 2020. A-BPS: Automatic Business Process Discovery Service using Ordered Neurons LSTM. In *2020 IEEE International Conference on Web Services, ICWS 2020*, 428–432. IEEE.

Honkisz, K.; Kluza, K.; and Wisniewski, P. 2018. A Concept for Generating Business Process Models from Natural Language Description. In *Knowledge Science, Engineering and Management - 11th INt. Conf., KSEM 2018, Proc., Part I*, volume 11061 of *LNCS*, 91–103. Springer.

Maqbool, B.; Azam, F.; Anwar, M. W.; Butt, W. H.; Zeb, J.; Zafar, I.; Nazir, A. K.; and Umair, Z. 2018. A Comprehensive Investigation of BPMN Models Generation from Textual Requirements - Techniques, Tools and Trends. In *Information Science and Applications - ICISA 2018*, volume 514 of *LNEE*, 543–557. Springer.

Qian, C.; Wen, L.; Kumar, A.; Lin, L.; Lin, L.; Zong, Z.; Li, S.; and Wang, J. 2020. An Approach for Process Model Extraction by Multi-grained Text Classification. In *32nd Int. Conf., CAiSE 2020, Proc.*, volume 12127 of *LNCS*, 268–282. Springer.

Quishpi, L.; Carmona, J.; and Padró, L. 2020. Extracting Annotations from Textual Descriptions of Processes. In *Business Process Management - 18th Int. Conf., BPM 2020, Proc.*, volume 12168 of *LNCS*, 184–201. Springer.

van der Aa, H.; Carmona, J.; Leopold, H.; Mendling, J.; and Padró, L. 2018. Challenges and Opportunities of Applying Natural Language Processing in Business Process Management. In *Proc. of 27th Int. Conf. on Computational Linguistics, COLING 2018.*, 2791–2801. ACL.

van der Aa, H.; Di Ciccio, C.; Leopold, H.; and Reijers, H. A. 2019. Extracting Declarative Process Models from Natural Language. In *Proc. of CAiSE 2019*, volume 11483 of *LNCS*, 365–382. Springer.