Review of Process fragments discovery from emails: Functional, data and behavioral perspectives discovery

Reviewed paper

M. Elleuch, O. A. Ismaili, N. Laga, et W. Gaaloul, "Process fragments discovery from emails: Functional, data and behavioral perspectives discovery", Information Systems, vol. 118, p. 102229, sept. 2023, doi: 10.1016/j.is.2023.102229.

Summary

The paper presents a new method to extract business process (BP) data from emails. The extracted data consists of activities, speech acts, BP artifacts and BP fragments. These attributes are first formalized and organized in a meta-model. Base attributes are extracted from emails using classic NLP methods and grouped using overlapping clustering. Additional data such as sequential constraints are also mined in the data. The approach is evaluated quantitatively on 5000 emails from the Enron dataset, annotated by the authors.

Reasons to accept the paper

- The major strength of this submission is the ability to extract multiple perspectives of business processes, which is a real innovation in BP mining from emails. The data collected is both diverse and well connected.
- The proposed approach only uses classic NLP techniques and seems to be applicable with low compute cost on a variety of datasets.
- The ontologies are well formalized and could be reused with a different method, enabling comparison between results.

Reasons to reject the paper

- The annotation strategy could be more detailed in terms of scientific methodology and its robustness could be improved.
- Furthermore, the evaluation section could be clearer, especially the presentation of the results.
- Finally, threats to validity could be analyzed more deeply.

Comments

Considering the previous points and especially the innovative dimension of the paper, my recommendation is acceptance of the paper. The issue regarding the scientific validity of the annotation strategy should be addressed to dissipate concerns over validity of the results.

Major Comments

- Title: Process fragments could be replaced by Business process fragments in order to situate the paper in the Business process field.

- Abstract: The abstract is long (more than 400 words), due to a very detailed problem statement. The problem statement could be summarized and parts of it could be added to the introduction. Quantitative results could also be added to assess the performance of the method.
- Keywords: The keywords "Data perspective", "Behavioral perspective" and "Functional perspective" are redundant and not optimized for SEO. Other keywords could better represent the paper subjects and contribution.
- Section 3: Great overview of the whole approach. References to the example emails in Figure 4 really help to ground the approach and is helpful when diving into formalizations later.
- Figures should be safe contained: figures 17 to 22 lack a legend to understand the colors used. This is especially relevant at the end of the paper since the related text can be far away due to the high number of figures (see below).
- There is a high number of figures (i.e., 11) in the last 6 pages of the paper. While these figures present interesting results, their number and large size reduce the readability of the evaluation section and can lose the reader. This is in part due to the 2 different types of figures: the one displaying evaluation results (12, 13, 19) and the one displaying a visualization of the extracted data (14, 15, 16, 17, 18, 20, 21, 22). Creating a real split between these groups could provide a better structure. Moreover, some of the visualization figures could be removed (i.e., 16, 18, 21, 22b) as they are redundant with other figures and are already accessible on the results website.
- Section 6: The evaluation results are presented in various figures (12, 13, 14) and are always dependant on parameters. It is therefore hard to quickly understand the performance of the approach. Individual results, such as best F1 score, could be highlited and/or regrouped in a single table at the end of the section. These results could also be included in the abstract and conclusion.
- Conclusion: Explicit results of the evaluation are not present. Takeaway numbers could be included to summarize the performance of the method.
- Threats to validity: validity of the annotation methodology should be discussed. As stated in the beginning of this review, annotation strategy should include the number of annotators and the steps taken to mitigate biases. Without these precisions the validity of the strategy could be questionned considering the arbitrary nature of the tasks splits.
- Results website: The website is most welcome and provides insightful vizualisations of the results of the paper. However the reusability of the data could be improved by regrouping all the files into an archive instead of splitting it into 300 different parts. The specified format of the data does not correspond to the actual data files (.txt instead of .json).

Minor comments

- Figure 1: The figure should be in higher resolution and labels could be clearer, using color as in Figure 2 for example.
- Figure 2: Remove bolding of subtitle to be consistent with other figures.
- Figure 3: Numbering of steps could be improved by replacing numbers to the same numbering used in section 3.
- Background lines should be removed in Figures 5a, fig 6, fig 7, fig 11 for consistency and better readability.
- Figure 6: Some of the text is overlapping with lines, reducing readability.
- Section 5.5.2: "Rules based method" should not be capitalized.
- Section 6.1.4: Fix typo, replace "Besed" by "Based".
- Figures 21 and 22: Fix the trimming to remove unnessecary lines.
- Figure 22b: The figure should be reorganized to remove the rotation, as it is not easilty readable in this orientation.