Natural Language Processing with Process Models (NLP4RE Report Paper)

Jan Mendling
Wirtschaftsuniversität Wien
Vienna, Austria
jan.mendling@wu.ac.at

Henrik Leopold
Kühne Logistics University
Hamburg, Germany
henrik.leopold@the-klu.org

Lucineia Heloisa Thom
Federal University of
Rio Grande do Sul
Porto Alegre, Brazil
lucineia@inf.ufrgs.br

Han van der Aa
Humboldt-Universität zu Berlin
Berlin, Germany
han.van.der.aa@hu-berlin.de

Abstract

This paper is a report paper that focuses on research at the intersection of business process management and requirements engineering. It gives an overview of the research on natural language processing with process models organized in terms of 25 challenges. This research line is pursued in a cross-university collaboration between the authors and further colleagues. We describe the most important contributions of the authors and highlight directions for future research.

1 Team Overview

The research team has a track record of joint work in the area of natural language processing with process models of over ten years. Since several members have changed affiliation, the collaboration has evolved towards a virtual research team. The team works in the area of business process management [DRMR18] and conducts research on the analysis of business process models including process model verification, refactoring, change propagation, matching, process mining, conformance checking, guidelines, and human comprehension of process models.

Business process management according to our tradition strongly builds on research into the design of workflow systems in the 1990s and the configuration of ERP Systems in the 2000s by the help of business process models. It is in line with requirements engineering in its ambition of understanding the application domain, operational constraints, and functionality needed by stakeholders [Som05] and more specific with its focus on a special class of systems, namely systems that support an organization to execute their business processes.

2 Past Research on NLP for Requirements Engineering with Process Models

In order to organize our previous research in the area of NLP for Requirements Engineering with a specific focus on process models, we have developed a framework that includes a list of 25 challenges. These challenges are associated with integrating requirements as a process model more efficiently, validating their correctness, completeness and consistency, and extracting information to support the design and implementation of a system that supports the execution of the business process. The 25 challenges can be organized into three major categories as Figure 1 illustrates: challenges in relation to automatically processing labels (C1-C7), in relation to labels in process models (C8-C19), and in relation to overall repositories (C20-C25) [MLP14]. Various of these challenges have been addressed by our research and also by other research teams. In the following, we discuss a selection of our works in order to illustrate the spectrum of contributions that have been made in this area of research. Several of these works have been published in renowned journals including IEEE Transactions on Software Engineering, Information & Software Technology, Decision Support Systems, and Information Systems.

The initial spark for this research was laid by the observation that the textual labels of process models can be formulated in a good and bad way. This observation provided the motivation for utilizing natural language processing techniques to improve the text labels of process model. Such a technique can be understood as a specific type of refactoring of process models with the aim to make them easier to understand by humans. Towards this end, we developed a technique to identify different styles of labels automatically [LSM11] and guideline violations [LEM+13], based on which we could then refactor them [LSM12]. Recently, we developed a novel label parsing techniques, which can be used to better address the aforementioned use cases [LvdAOR19]. With these works, we addressed the Challenges C1 and C2. This foundational set of techniques was then further extended into different directions. Most notable are translation, semantic processing, and conformance checking between process model and text, as discussed next.

2.1 Translations between Process Models and Text

An important question for processing of text and models is to which extent automatic translations are feasible. We addressed this question in both directions: from text to process model and from process model to text.

Our research on the translation from text to process model [FMP11] addresses various challenges that we organize in four categories. The first category, *Syntactic Leeway*, includes problems that stem from changing active and passive voice of input text, potential rewording and changes of order and conditions that are not explicit. The second category, *Atomicity*, refers to the fact that sentences can be as complex as whole model fragments, that activities can be split across sentences and that relative clauses have to be dealt with. The third category, *Relevance*, acknowledges that relative clauses, example sentences or meta-statements should not lead to model elements. The fourth category, *Referencing*, deals with anaphora, textual links and end-of-block recognition. The proposed translation technique works from the sentence level to the text level and creates a process model automatically. Using a test set of 47 text-model pairs, we achieve an average translation accuracy of 77%. This work has been recently extended with a structural analysis of the texts and an analysis of sentence templates in order to address potential issues of ambiguity [STW+18] and is currently being integrated into a service-oriented architecture for the generation of process-oriented text.

Our complementary research on the translation from process model to text for validation purposes [LMP14] addresses various challenges that stem from parsing the formal structure of the process model. More specifically, we distinguish four categories of challenges. The first category, *Text Planning*, deals with linguistic information extraction, model linearization and text structuring. The second category, *Sentence Planning*, includes lexicalization and message refinement. The third category, *Surface Realization*, relates to interfacing with established realizers. The fourth category, *Flexibility*, addresses variations of input data and adaptation of output. The proposed translation technique starts with information extraction from process model elements to graph parsing the process model into the refined process structure tree and text structuring based on the tree fragments. This data is fed into a deep syntax tree where a technique for message refinement is applied. Finally, a realizer generates the resulting natural language text. Our evaluation demonstrates that the generated texts are highly accurate and that a back translation hardly entails any loss of information.

2.2 Semantic Processing of Process Models and Text

Each of these translation techniques takes the textual content as given. This is problematic, because terms are often ambiguous. This is the starting point of our research on the automatic detection and resolution of lexical ambiguity in process models [PLM15]. The corresponding technique covers homonym detection and resolution as much as synomym detection and resolution. The technique is evaluated using a collection of more than 2,000 process models from practice with altogether more than 20,000 text labels. The evaluation indicates that homonymous usage of terms like *application*, *case* or *incident*, as well as synonymous word pairs such as *check-control*, *create-produce*, and *customer-client* are found. Automatic resolution significantly reduces ambiguity.

A key problem of processing text labels of models in practice is that practitioners often do not use these labels in a canonical way. Examples are activity labels like *Screen delivery documents if necessary* or *update*

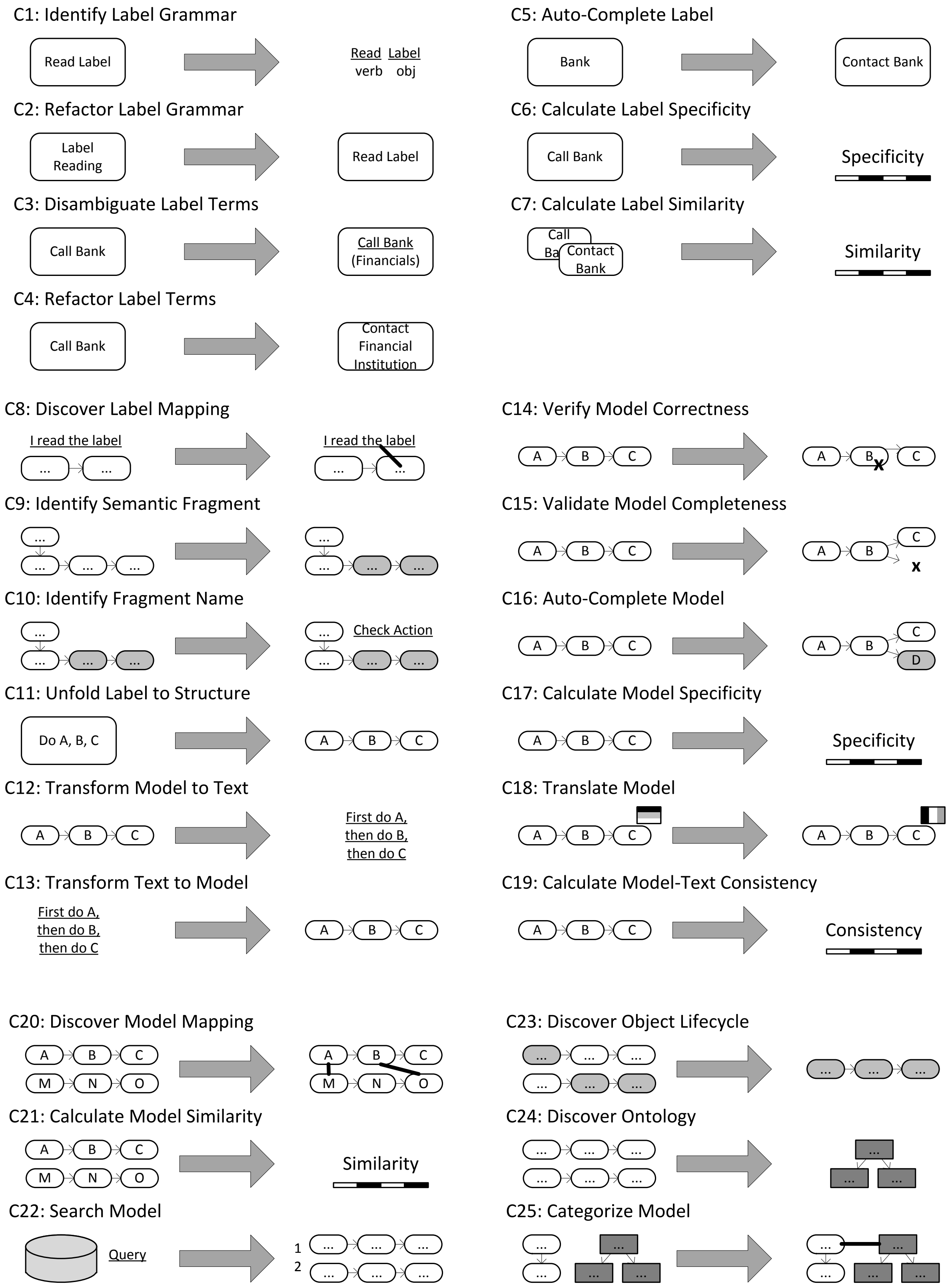
C1: Identify Label Grammar
Read Label
Read Label
verb obj
C2: Refactor Label Grammar
Label Reading
Read Label
C3: Disambiguate Label Terms
Call Bank
Call Bank (Financials)
C4: Refactor Label Terms
Call Bank
Contact Financial Institution
C5: Auto-Complete Label
Bank
Contact Bank
C6: Calculate Label Specificity
Call Bank
Specificity
C7: Calculate Label Similarity
Call Bank
Contact Bank
Similarity
C8: Discover Label Mapping
I read the label
...
...
I read the label
...
...
C9: Identify Semantic Fragment
...
C10: Identify Fragment Name
...
Check Action
C11: Unfold Label to Structure
Do A, B, C
A
B
C
C12: Transform Model to Text
A
B
C
First do A, then do B, then do C
C13: Transform Text to Model
First do A, then do B, then do C
A
B
C
C14: Verify Model Correctness
A
B
C
x
C15: Validate Model Completeness
A
B
C
x
C16: Auto-Complete Model
A
B
C
D
C17: Calculate Model Specificity
Specificity
C18: Translate Model
C19: Calculate Model-Text Consistency
Consistency
C20: Discover Model Mapping
A
B
C
M
N
O
C21: Calculate Model Similarity
Similarity
C22: Search Model
Query
1
2
...
C23: Discover Object Lifecycle
...
C24: Discover Ontology
...
C25: Categorize Model
...

Figure 1: 25 Challenges of Semantic Process Modeling [MLP14]

inventory and achieve documents. Canonicity refers to the specification of process model elements in such a way that they correspond to exactly one element [LPM17]. The paper identifies a series of patterns of such wrong usage of labels along with automatic refactorings. The transformation rules replace one model element with a non-canonical text label with a fragment of several elements. For example, the *Screen delivery documents if necessary* yields a decision block and the *update inventory and achieve documents* a sequence.

2.3 Conformance Checking between Process Models and Text

We are also able to automatically check the conformance between process models and corresponding text. A specific conformance checking technique has been developed that automatically compares recorded process executions (captured in event logs) to natural language specifications of processes [vdALR18]. A particular challenge in this regard is the inherent ambiguity of natural language, which can lead to different possible interpretations of how a process should be executed. The developed technique uses probabilistic conformance checking to take this ambiguity into account to provide reliable results.

Several works also consider that process models and textual process descriptions are often used alongside each other in organizations, given their complementary nature [vdALvdWR17]. Techniques have been developed that establish alignments between a model and a corresponding text [SvdACP18], that use such alignments to detect inconsistencies [vdALR17], and a process querying technique that can search repositories of both textual and model-based process descriptions simultaneously [LvdAP$^+$17].

Many of the proposed techniques also help to match process models. Process model matching can be defined as the task of automatically aligning the text labels of one process model with the labels of a second model [C$^+$13]. The task is rather easy if it can be assumed that there is a 1:1 match between the elements. In practice, this is hardly the case. Often aspects are represented in one model, which are not represented in the second one, and the other way around. Difficult are also matches that bridge different levels of granularity such as 1:n and n:m matches. The process matching contest promotes research in this area [C$^+$13].

3 Research Plan on NLP for Requirements Engineering with Process Models

Many of the developed techniques are important to make business process management smarter [MBBF17], though various challenges remain. Many of them can be related to the 25 Challenges illustrated above, but also beyond. In our own future research, we aim to address the following problems.

First, our current approach for process model elements identification in natural language text is based on a reduced set of BPMN elements (e.g. activity, subprocess, start, intermediate and end events). As future work we consider to extend our approach to support a larger number of elements as well as to filter natural language texts by process perspectives such as data and events. Second, we have observed that the quality of process descriptions in practice is often low. This calls for research on future techniques that are able to check quality and refactor poor text. One option is to use domain ontologies to check the consistency of process descriptions and respective ontological concepts. Benefits of ontology usage in this context has already been studies empirically in [GMB$^+$17]. Third, while existing work on the extraction of process models from natural language focuses on *imperative* process descriptions and models, we are currently working on the extraction of *declarative* process constraints from natural language [vdACLR19]. In this way, we aim to deal with rule-based descriptions of processes.

Acknowledgements

Lucineia Heloisa Thom is a CAPES scholarship holder, Program Professor Visitante no Exterior, Process Number: 88881.172071/2018-01. Han van der Aa is funded as a research fellow of the Alexander von Humboldt foundation.

References

[C$^+$13] Ugur Cayoglu et al. Report: The process model matching contest 2013. In *International Conference on Business Process Management*, pages 442–463. Springer, 2013.

[DRMR18] Marlon Dumas, Marcello La Rosa, Jan Mendling, and Hajo A. Reijers. *Fundamentals of Business Process Management, Second Edition.* Springer, 2018.

[FMP11] Fabian Friedrich, Jan Mendling, and Frank Puhlmann. Process model generation from natural language text. In *CAISE*, pages 482–496. Springer, 2011.

[GMB+17] Jonas Bulegon Gassen, Jan Mendling, Amel Bouzeghoub, Lucinéia Heloisa Thom, and Jos'e Palazzo M. de Oliveira. An experiment on an ontology-based support approach for process modeling. *Information & Software Technology*, 83:94–115, 2017.

[LEM+13] Henrik Leopold, Rami-Habib Eid-Sabbagh, Jan Mendling, Leonardo Guerreiro Azevedo, and Fernanda Araujo Baião. Detection of naming convention violations in process models for different languages. *Decision Support Systems*, 56:310–325, 2013.

[LMP14] Henrik Leopold, Jan Mendling, and Artem Polyvyanyy. Supporting process model validation through natural language generation. *IEEE Trans. Software Eng.*, 40(8):818–840, 2014.

[LPM17] Henrik Leopold, Fabian Pittke, and Jan Mendling. Ensuring the canonicity of process models. *Data Knowl. Eng.*, 111:22–38, 2017.

[LSM11] Henrik Leopold, Sergey Smirnov, and Jan Mendling. Recognising activity labeling styles in business process models. *Enterprise Modelling & Inf. Systems Architectures*, 6(1):16–29, 2011.

[LSM12] Henrik Leopold, Sergey Smirnov, and Jan Mendling. On the refactoring of activity labels in business process models. *Inf. Syst.*, 37(5):443–459, 2012.

[LvdAOR19] Henrik Leopold, Han van der Aa, Jelmer Offenberg, and Hajo A Reijers. Using hidden markov models for the accurate linguistic analysis of process model activity labels. *Information Systems (accepted for publication)*, 2019.

[LvdAP+17] Henrik Leopold, Han van der Aa, Fabian Pittke, Manuel Raffel, Jan Mendling, and Hajo A Reijers. Searching textual and model-based process descriptions based on a unified data format. *Software & Systems Modeling*, pages 1–16, 2017.

[MBBF17] Jan Mendling, Bart Baesens, Abraham Bernstein, and Michael Fellmann. Challenges of smart business process management: An introduction to the special issue. *Decision Support Systems*, 100:1–5, 2017.

[MLP14] Jan Mendling, Henrik Leopold, and Fabian Pittke. 25 challenges of semantic process modeling. *Int. J. of Inf. Systems and Software Engineering for Big Companies*, 1(1):78–94, 2014.

[PLM15] Fabian Pittke, Henrik Leopold, and Jan Mendling. Automatic detection and resolution of lexical ambiguity in process models. *IEEE Trans. Software Eng.*, 41(6):526–544, 2015.

[Som05] Ian Sommerville. Integrated requirements engineering: A tutorial. *IEEE software*, 22(1):16–23, 2005.

[STW+18] Thanner Soares Silva, Lucinéia Heloisa Thom, Aline Weber, Jos'e Palazzo Moreira de Oliveira, and Marcelo Fantinato. Empirical analysis of sentence templates and ambiguity issues for business process descriptions. In *OTM, Proceedings, Part I*, pages 279–297, 2018.

[SvdACP18] Josep Sànchez-Ferreres, Han van der Aa, Josep Carmona, and Lluís Padró. Aligning textual and model-based process descriptions. *Data Knowl. Eng.*, 118:25–40, 2018.

[vdACLR19] Han van der Aa, Claudio Di Ciccio, Henrik Leopold, and Hajo A. Reijers. Extracting declarative process models from natural language. In *CAISE (accepted for publication)*, 2019.

[vdALR17] Han van der Aa, Henrik Leopold, and Hajo A. Reijers. Comparing textual descriptions to process models - the automatic detection of inconsistencies. *Inf. Syst.*, 64:447–460, 2017.

[vdALR18] Han van der Aa, Henrik Leopold, and Hajo A. Reijers. Checking process compliance against natural language specifications using behavioral spaces. *Inf. Syst.*, 78:83–95, 2018.

[vdALvdWR17] Han van der Aa, Henrik Leopold, Inge van de Weerd, and Hajo A. Reijers. Causes and consequences of fragmented process information: Insights from a case study. In *23rd Americas Conference on Information Systems*, 2017.