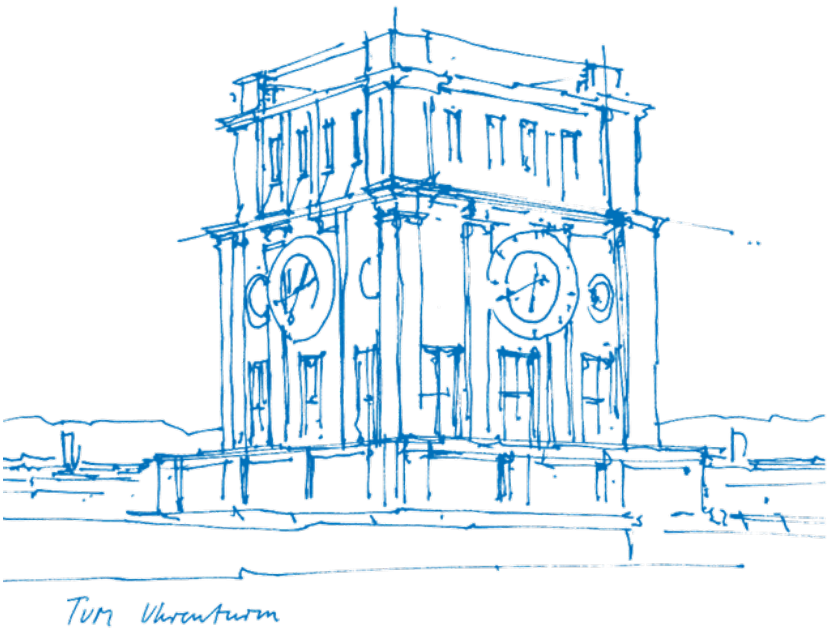
Chair of Information Systems and Business Process Management (i17)
Department of Computer Science
TUM School of Computation, Information and Technology
Technical University of Munich

Bachelor's Thesis in Information Systems

Vincent Derek Held

An enhanced automated approach for transforming natural language process descriptions to BPMN2.0 process diagrams – with an evaluation of the application to ISO-Norm process descriptions



Chair of Information Systems and Business Process Management (i17)
Department of Computer Science
TUM School of Computation, Information and Technology
Technical University of Munich

Bachelor's Thesis in Information Systems

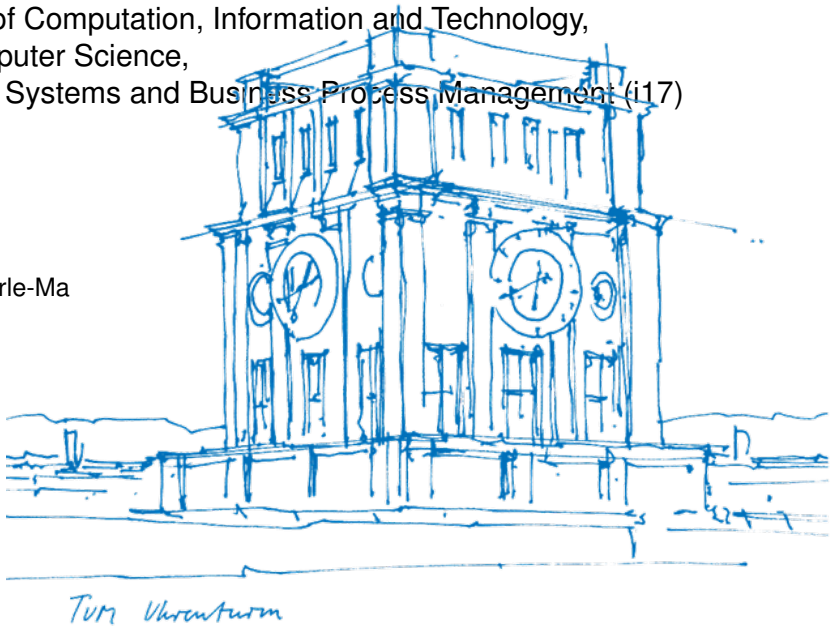
Vincent Derek Held

An enhanced automated approach for transforming natural language process descriptions to BPMN2.0 process diagrams – with an evaluation of the application to ISO-Norm process descriptions

Ein verbessertes automatisiertes Verfahren zur Umwandlung natürlichsprachlicher Prozessbeschreibungen in BPMN 2.0 Prozessdiagramme – mit einer Evaluierung der Anwendung auf ISO-Norm-Prozessbeschreibungen

Thesis for the Attainment of the Degree
Bachelor of Science

at the TUM School of Computation, Information and Technology,
Department of Computer Science,
Chair of Information Systems and Business Process Management (i17)



Examiner
Prof. Dr. Stefanie Rinderle-Ma

Supervised by
Catherine Sai

Submitted by
Vincent Derek Held

Submitted on
15.12.2023

Declaration of Academic Integrity

I confirm that this bachelor's thesis is my own work and I have documented all sources and material used.

I am aware that the thesis in digital form can be examined for the use of unauthorized aid and in order to determine whether the thesis as a whole or parts incorporated in it may be deemed as plagiarism. For the comparison of my work with existing sources I agree that it shall be entered in a database where it shall also remain after examination, to enable comparison with future theses submitted. Further rights of reproduction and usage, however, are not granted here.

This thesis was not previously presented to another examination board and has not been published.

Garching, 15.12.2023 Vincent Derek Held

Abstract

A comprehensive understanding of business processes is crucial for the digitization of these processes. The utilization of Business Process Model and Notation (BPMN) 2.0 process diagrams has emerged as a pivotal tool in both research and industry for representing and analyzing business workflows. These processes are influenced by an ever growing amount of regulatory documents and process execution data.

This thesis is a contribution to develop a state-of-the-art approach for transforming natural language process descriptions to BPMN2.0 process diagrams. The aim is to evaluate how recent developments have evolved compared to existing methods and how well the approach works for process descriptions from more complex regulatory documents (e.g. ISO standards or data protection regulations). Furthermore, it will be investigated which technologies are best suited to visualize the extracted process information in a BPMN2.0 process model.

Keywords: *Natural Language Processing, Business Process Compliance, Natural Language to Process, Business Process Model Generation*

Contents

List of Tables

List of Figures

Introduction

Motivation

Process modeling is a common technique for a better understanding and documentation of organizational processes and structures, as well as for process improvements and standardizations. [1] Processes are not only used in companies, but they are also common within organizations and in research. Usually, lots of process documentation are stored in natural language. [2] Unfortunately, experience is necessary for process modeling, which creates a need for professional knowledge. BPMN2.0 is a standardized notation used to represent business processes graphically and is widely used in various industries and science. Process modeling is an elementary part of information systems and structural design. It is time-consuming, as 60% of business process management is spent on modeling [2]. Additionally, process descriptions and diagrams must be frequently adapted due to the constant optimization of processes. Furthermore, processes need to be adapted to new regulations [3]. Natural Language Processing (NLP) is a subfield of artificial intelligence that focuses on enabling computers to understand, interpret, and produce human language (natural language). By applying NLP techniques, this project can enable the computer to understand natural language and to identify the relevant details from regulatory documents for the process [3]. Automated creation of process diagrams from textual descriptions is a promising approach to making processes more efficient and minimizing errors. This approach has already shown positive results in various application areas.

However, there are only a few papers so far that explicitly deal with the extraction of process information from legal texts. Therefore, it is essential to focus on this application since law texts are often complex and Business Process Compliance (BPC) is due to the introduction of the General Data Protection Regulation (GDPR) and its financial fines for violations a highly relevant topic for companies [4]. Mostly, lawyers must interpret the legal text correctly and help implement internal company processes. Automating the creation of process diagrams from regulatory documents can facilitate the implementation of laws and standards and minimize errors. By comparing the automatically created process diagrams from legal texts with the process diagrams of, e.g., company processes, it would be possible to make the realization in companies and organizations easier to implement. Furthermore, the costs can be reduced since the productivity in the process analysis

is high, as well as the expenditure for uniform documentation, lawyers, and conversion is more effortless.

With the combination of NLP for extracting relevant information from regulatory documents and BMPN2.0 for visualization, our approach aims to improve understanding of regulatory documents, facilitate implementation through visualization, and minimize errors due to misunderstandings. In summary, process modeling is an essential technique for understanding and improving organizational processes, but it can be time-consuming and requires specialized knowledge. Natural Language Processing (NLP) is a promising approach for automating process information extraction from textual descriptions. Automated process diagram creation has already shown positive results in various application areas. However, there is a need for more research specifically focused on the extraction of process information from legal texts. To address these gaps in the existing literature, this study aims to investigate the following research questions:

Research Questions

TODO: In the explanation also tell how you plan to prove that your potential future solution is good. About 1 page.

RQ 1: How can business process models be automatically generated from textual descriptions using NLP techniques? This research question aims to refactor the current state approach to automatically convert textual process descriptions into graphical representations using NLP techniques and to investigate the methods and techniques that can be employed.

RQ 2: How do recent developments compare to existing methods? The second research question focuses on evaluating and comparing recent advancements in the field of automatic business process modeling, specifically those utilizing NLP techniques, with existing methods. It aims to assess the improvements, innovations, and potential advantages of these recent developments in terms of ease of use, model quality, and overall performance. By conducting a comparative analysis, this research question seeks to identify the strengths and limitations of the state-of-the-art approaches compared to established methods.

RQ 3: How well does the approach work concerning process descriptions from more complex regulatory documents (e.g., ISO Norms)? This research question aims to assess the effectiveness and applicability of the proposed approach, specifically focusing on its performance when applied to more complex regulatory documents such as ISO Norms. It seeks to understand how well the NLP-based approach handles the challenges of intricate and detailed process descriptions within these regulatory documents. The research will evaluate the accuracy and ability of the approach to transform complex regulatory language into meaningful and accurate BPMN2.0 process diagrams.

RQ4: Which technologies best visualize the extracted process information in a BPMN2.0 process model? This research question aims to explore the various technologies that can effectively visualize the process information extracted from textual descriptions in a BPMN2.0 process model. Once the relevant details have been extracted using Natural Language Processing (NLP) techniques, it is crucial to represent this information in a visually understandable and standardized format. The visualization of extracted process information plays a significant role, as it enables stakeholders to comprehend the process model easily and facilitates communication and collaboration among team members involved in process analysis, improvement, and implementation. The research will

identify a best-practice technology for visualizing the extracted process information in a BPMN2.0 process model. Additionally, the research will explore any advancements or innovations in process visualization technologies that can enhance the representation and understanding of process models. Factors to consider include ease of use, flexibility in representing different process elements, support for BPMN2.0 notation, layout, and visualization quality. This understanding will contribute to the overall goal of improving the comprehension, implementation, and standardization of processes in organizations, ultimately leading to enhanced process efficiency, reduced errors, and better decision-making.

Contribution

The significance of this work lies in the effort to investigate and further develop the basic approach formulated by [TODO: Shuawei]. This basic approach provides an instrumental solution based on the conversion of textual process descriptions into BPMN diagrams. While the foundation of the approach is embedded in the work of [TODO: Friedrich], the evaluation metrics were mainly taken from academic, industrial and textbook-based process descriptions. Against this background, the main contributions of this work can be highlighted as follows:

1. **Enhanced scope of the evaluation:** this study pioneers by attempting to evaluate the baseline approach using regulatory documents, in particular ISO standards and articles from the General Data Protection Regulation (GDPR). This extends the assessment framework to cover a diverse range of textual descriptions beyond traditional academic and industry boundaries.
2. **Identification of limitations:** By applying the program to evaluation data consisting of process descriptions and legal texts, this study systematically identifies the underlying limitations of the baseline approach. This insights are crucial as it identifies opportunities for improvement and adaptation, especially when dealing with texts that contain complicated legal nuances.
3. **Differential analysis:** The research analyses the differences between regulatory documents and traditional business process descriptions. Understanding these contrasts is crucial as it sheds light on the inherent complexity and specifications of regulatory documents that may not be present in standard process descriptions and provides insight into possible contextual adaptations.
4. **Strategic adjustments**: Given the limitations and differences uncovered, this work takes the next logical step by implementing possible strategies and methods for fine-tuning the automatic generation of business processes. By addressing the characteristics of legal texts, it provides a road map for refining and improving the tool's adaptability and accuracy.

In essence, this work is a bridge that connects the fields of business process modelling and legal regulatory documentation. It strives to improve the capabilities of the basic tool, highlight the specifics of regulatory text, and ensure that it is robust and versatile enough to handle a broader and more complex range of textual descriptions.

Through comprehensive assessment and detailed analysis, this study aims to strengthen [TODO: Shuawei]'s foundational work and ensure that it is equipped to meet the challenges of an evolving text landscape.

Methodology

As an implementation project is the core of the thesis, we will follow the design science research (DSR) methodology proposed by Hevner [5], [6] [7]. This guides developing and evaluating our solutions to problems described in Chapter 1. The DSR approach is iterative and involves designing, building, evaluating, and refining an artifact until a satisfactory solution is achieved. This chapter will describe the DSR methodology we use in this study.

The DSR method consists of six stages, as follows:

1. problem identification and motivation
2. definition of the objectives for a solution
3. design and development
4. demonstration
5. evaluation
6. communication

In the following, it is explained how the six stages will be applied within the project:

Problem identification and motivation

In the DSR methodology, the initial step is to recognize the issue and establish the reason for the study. Research in the literature (as seen in the next chapter) indicates that no modern solution is currently available for automatically generating process diagrams. Specifically, there is no investigation of the possibility of automatically generating process diagrams from regulatory documents. This study aims to develop an automated approach to extract relevant data from regulatory documents and create process diagrams. This will enhance understanding and implementation of regulations while also reducing mistakes.

Definition of the objectives for a solution

The second step in the DSR methodology is to define objectives. This thesis aims to develop an enhanced proven method for automatically generating process models from a textual description using current technologies. For this purpose, given approaches will be evaluated, and the most advanced approach will be reconstructed. Here, the aim is to evaluate the accuracy, compared to the approach to be followed, as well as to minimize or eliminate errors and improve the output

quality. Additionally, this thesis focuses on analyzing the approach for creating process diagrams of regulatory documents and improving the implemented algorithm for this purpose.

Design and development

The third step in the DSR methodology is to design and develop the artifact. In our case, we develop a prototype that uses NLP techniques to extract relevant information from regulatory documents and create process diagrams (BPMN 2.0 models). The software prototype will be developed using Python programming language, Spacy library for NLP processing, and the BPMN 2.0 standard for the visualization of processes.

Demonstration

The fourth step in the DSR methodology is to demonstrate and evaluate the artifact. In our case, we demonstrated the developed Python code by using it to extract relevant information from a set of example regulatory documents and create process diagrams. Afterward, we evaluated the results of this proof-of-concept implementation based on the accuracy and completeness of the process models. The code will be published on GitHub, accessible to all stakeholders. Additionally, a set of natural process descriptions and human-modeled process diagrams, which have been used for the evaluation, will be published there as well, together with the output diagrams created by the algorithm.

Evaluation

The fifth step in the DSR methodology is to evaluate the artifact. The created approach will be evaluated by checking the completeness and the correct order of the process steps. Therefore, a set of text-based descriptions, corresponding human-modeled process diagrams, and the results of the automatically generated models will be compared to measure the accuracy with the help of an evaluation matrix.

Communication

Finally, the last step is to communicate the results. The outcomes of this project will be presented in a scientific paper (thesis). As mentioned in the demonstration part of DSR, the code will be accessible with some examples in a GitHub repository. Additionally, this thesis will be communicated and presented in a thesis defense to all relevant stakeholders.

Evaluation

The evaluation aims to assess the performance and effectiveness of the proposed approach in addressing the research objectives outlined in Chapter 1, thereby providing critical insights into its practical application and potential impact on process modeling.

This approach focuses on qualitative evaluation since quantitative decision criteria such as runtime are superior to human creation. Additionally, the runtime is not comparable to the outcomes of Friedrich et al. due to the hardware progress within the last decade.

A set of qualified input data from several sources will be collected. We will also use the same input data as Friedrich et al. to ensure comparability to the results. Afterward, this approach focuses on the evaluation of regulatory documents. Therefore, we will use descriptions of ISO norms and the General Data Protection Regulation (GDPR).

Criteria for selecting the test data are a textual process description and a corresponding BPMN model created by a human model, which will be used as the gold standard. In the following, we will call these sets “text-model pairs”.

As evaluation metrics, we will use the **Graph edit distance** (GED) to compare the similarity of the different outputs. The graph edit distance between two graphs, g_1 and g_2, is written as $GED\ (g_1, g_2)$. $GED\ (g_1, g_2)$ is minimum of editing operations to transform g_1 into g_2 and is denoted as

$$GED\ (g_1, g_2) = \min_{(e_1,\ldots,e_k)\in\mathcal{P}(g_1,g_2)} \sum_{i=1}^{k} c\ (e_i) \tag{1.1}$$

where $P\ (g_1, g_2)$ denotes the number of edits ($|e|$) multiplicated by the costs (c) per edit type to transform g_1 into g_2 and is the cost of each graph edit operation [8].

For calculating $GED\ (g_1, g_2)$the following steps will be performed:

1. **Definition of edit operations:** Operations can be adding or removing Nodes (N), Gateways (G), or Edges (E). Other activities could be changing the order of activities (OA), renaming activities (RA), or splitting activities (SA). $\{N, G, E, OA, RA, SA\} \in e$

2. **Assigning costs to the operations:** Assigning costs to each edit operation based on the complexity and significance of the process. For example, the costs of adding or removing a node are written as $C(|N|)$.
3. **Calculating the graph edit distance:** Using the graph edit distance algorithm, to calculate the minimum required cost to transform the graph g_1 to the graph g_2.

The result of the graph edit distance algorithm is the minimum required costs of transforming our automatically created process model to the human model process model. It enables us to compare the different approaches in terms of the similarity of the outcomes.

This technique is used, to determine

$$GED1\left(g_{\text{ourApproach}}, g_{\text{human-modeld}}\right) \tag{1.2}$$

and

$$GED2\left(g_{\text{Friedrich}}, g_{\text{human-modeld}}\right). \tag{1.3}$$

As the result of the GED represents the similarity of an approach to the gold standard, we will be able to compare the accuracy of GED1 and GED2. This evaluation will be done using the text-model pairs used by Friedrich et al.

Afterward, the results of our approach will be compared to the results of human-modeled diagrams based on the text-model pairs of **regulatory documents**

$$GED3\left(g_{\text{ourApproach}}, g_{\text{human-modeld}}\right). \tag{1.4}$$

Within the evaluation, the results of this approach will be represented, and the limitations of this approach will be identified and discussed. The outcomes will be compared to related work to identify similarities, differences, and potential explanations for variations. Unsolved questions and challenges that emerged during the project will be outlined, and suggestions about further research or improvements will be prepared.

Structure

TODO: Neuformulieren / Anpassen The remainder of this paper is structured as follows. Section 2 discusses related works. Section 3 introduces related definitions and problem of this work. Section 4 describes the proposed methods to improve the accuracy of the generated process diagrams and adaptions to achieve improved results, if the input are regulatory documents. Section 5 evaluates and compares the proposed approach with existing methods and evaluates our approach on a dataset. Section 6 provides a brief overview of the threats to validation. Finally, Sect. 7 shows the conclusions and future work.

Related Work

Predictive Compliance Monitoring in Process-Aware Information Systems: State of the Art, Functionalities, Research Directions

Rinderle-Ma et al. [9] **Keywords**: Predictive Compliance Monitoring, Predictive Compliance Monitoring System, Predictive Process Monitoring, Systematic Literature Review, Research Directions **//Motivation:**

- "Business process compliance is **a key area of business process management**"
- and aims at ensuring that processes obey to compliance constraints such as regulatory constraints or business rules imposed on them.
- "Process compliance can be checked during"
  - "process design time based on verification of process models"
  - "at runtime based on monitoring the compliance states of running process instances"
- "Compliance Monitoring (CM)(9, 10) is an integral part for monitoring and managing business processes in changing, complex regulatory environments such as the financial domain."
- "Yet, reactive management through auditing is still most prominent in compliance management of companies,"
- "Building a PCM system is a complex task, resulting from a multitude of compliance constraints stemming from different and constantly changing regulatory documents [12]"
- In particular, research should avoid assuming the constraints to be readily available and stated in some logic, but often have to be extracted and updated based on regulatory documents [158].
- "Life cycle handling" of process with adaption to (legal) constraints
- "Typically, constraints are stated in natural language and scattered across multiple regulatory documents (cf. Ex. 1)."
- "Moreover, this work assumes that compliance constraints are formalized u**sing some notion**." –> Darüber hinaus wird in dieser Arbeit davon ausgegangen, dass Konformitätsbeschränkungen mit Hilfe eines Begriffs formalisiert werden

//Contribution

- "This work, hence, analyzes existing literature from compliance monitoring as well as predictive process monitoring and provides an updated framework of compliance monitoring functionalities. Moreover, it raises the vision of a comprehensive predictive compliance monitoring system that integrates existing predicate prediction approaches with the idea of employing PPM with different prediction goals such as next activity or remaining time for prediction and subsequent mapping of the prediction results onto the given set of compliance constraints (PCM)."
- "By combining the respective capabilities of PPM and CM, research can offer companies a means to proactively assess and manage their business processes with respect to future outcomes, compliance status, and risks."
- "Predicate prediction – which is mainly followed by existing approaches, e.g., [16] – encodes each compliance constraint as prediction goal into prediction models."
- "Yet, especially in combination with updating compliance violations, **an open challenge remains how to define and update** the compliance degree while new events arrive throughout the event stream and to predict compliance states of single instances."

//Discussion:

- "Update 3: Continuous update of prediction results and compliance violations" –> ""Life cycle handling" of process with adaption to (legal) constraints" –> PCM (Predictive Compliance Monitoring)

Mining Process Models from Natural Language Text: A State-of-the-Art Analysis

Riefer et al. [10]

//Motivation

- "Workflow projects are time-consuming processes."
- "knowledge extraction and the creation of process models."
- "necessary information is often available as textual resources."
- "process model mining from natural language text has been a **research area of growing interest.**"
- "Organizations are constantly trying to **analyze and improve their business processes.**"

- "This is only possible if the knowledge about the processes is available **in a structured form like a business process model**."
- "85% of the knowledge and information are estimated to be available in an unstructured form, mostly as text documents (Blumberg and Atre 2003) [11]".
- "It gives people with no knowledge about process modeling the possibility to create process models, which is an important goal in view of the fact that structured data becomes more important."
- "Text Mining approaches have been developed for UML class diagrams (Bajwa and Choudhary 2011), entity relationship models (ERM) (Omar et al. 2008) or business process models (Friedrich et al. 2011). Current approaches do n**ot aim at replacing an analyst but at helping him to create better models in less time**."
- **Goal of this paper:** "general overview in terms of a state-of-the-art analysis is missing"
-

//Introduction -> Structure of Thesis:

- "The used research methodology and the identification of the relevant literature are presented in section 2, while section 3 introduces the most important methods and terms for processing natural language texts. Section 4 gives an overview of current approaches. The comparative analysis, which consists of a detailed theoretical analysis and a proposed practical analysis, is conducted in section 5. The results are discussed in Section 6, followed by a conclusion in section 7."
-
-

//Methodology "To define the current state of research and to identify different approaches for mining process models from natural language text, a systematical literature review was conducted. The three literature databases Google Scholar, SpringerLink and Scopus were used for the research. The following search keywords were derived from the title and thematic of this paper: natural language processing, process model, process modeling or process model generation, process model discovery, text mining, process mining and workflow. These were used in various combinations. As the used keywords cover broad research areas, they lead to a high number of search results. That complicated the identification of the relevant literature. The search results were checked through a title and

abstract screening to identify the relevant work. Hence, only publications which explicitly mention text-to-model transformations were considered as relevant. There turned out to be a problem with the author's way of describing their work: instead of referring to text mining or natural language processing, they often used the text type, such as use cases or group stories, to outline their work. The search in a database was aborted when a significant amount of repetitions or loss of precision was noticed. It turned out, that the keyword search provided a low degree of relevant papers. Hence the keyword search was skipped and changed to a cross reference search. Table 1 shows the literature search results."

Table 1: Results of literature research

Afterwards, further works were detected through a backwards search. The work of **(Leopold 2013)** provided a proficient starting point. The focus was set on approaches which generate a business process model. Five appropriate approaches could be identified:

BPMN model from text artefacts (Ghose et al. 2007) BPMN model from group stories (Goncalves et al. 2009) BPMN model from use cases (Sinha and Paradkar 2010 BPMN model from text (Friedrich et al. 2011) Model from text methodologies (Viorica Epure et al. 2015)

Design-time business process compliance assessment based on multi-granularity semantic information Xiaoxiao Sun [4]

- Business Process Compliance (BPC) is an essential part of BPM that measures how effectively an organization's business processes comply with all relevant laws, regulations, guidelines, and standards [2].
- Examples of critical regulatory documents are the Health Insurance Portability and Accountability Act (HIPAA), the Sarbanes-Oxley Act (SOX), and the General Data Protection Regulation (GDPR) [3].
- Companies that violate these regulatory documents risk losing the trust of investors, incurring financial fines, and facing criminal charges. As a result, adhering to rules from multiple sources has become essential for every organization to avoid huge fine losses and also to improve business process transparency [4].

- However, in company's practice, checking and ensuring the consistency of the organization's business processes with regulatory documents, i.e., BPC checking, is still largely done manually with lots of efforts.
- In addition, the costs of manual review increase significantly due to the constant changes in regulatory environment [5, 6].
- **However, formal languages often posed challenges in terms of comprehension [9].**
- In this study, we focus on parsing BPMN, which is one of the most widely used notations for business processes with the latest version of 2.0 [28].
- BPMN model comprises three types of nodes: events, activities, and gateways. Events, represented by circles, indicate occurrences within the process. Activities, depicted as rounded rectangles, represent tasks performed within the business process. Gateways, shown as diamonds, control the flow of the business process. However, the proposed approach is generic and can be applied to other modeling languages.

Solution Design

- Business Process Model and Notation

  TODO: More detail: elements and also references
- Spacy
- Problems addressed by Freidrichs
- Filtering of "proccess irrelevant information according to the PET dataset -> not directly relevent to the business process -> increases undestanbaility to the human ready but increases difficulty wehen processing the text -> which sentences are filtered Scores
-
-

Business Process Model and Notation

For the automatically generation of process diagrams it is important to use a in industry and research common standardized notation. Business Process Model and Notation 2.0 (BPMN) is an industry-standard notation specifically designed for business process modeling [12]. BPMN was developed by the Object Management Group to support business process management, for both technical users and business users, by providing a notation that is intuitive to business users yet able to represent

complex process semantics. It provides businesses with the capability of understanding their internal business procedures in a graphical notation and gives organizations the ability to communicate these procedures in a standard manner.

[13] The BPMN offers a variety of process elements. In the following the most important elements are listed and explained:

Spacy

Process Piper

Categorization of Issues

Identification of Introduction Sentences

Identification of Actors

To identify actors, we leverage the dependency labels of tokens. Specifically, for active voice sentences, the nominal subject dependency label, $token.dep_{=}$ = "$nsubj$", is utilized. For passive constructions, the agent dependency, denoted as "agent", is considered. Through this approach, the baseline method yields precise outcomes. Given the multifaceted nature of natural language an entity might be referenced using multiple terminologies.

As illustrated in Figure [TODO: Reference to Picture of Text 01 by Shuawei], several actors have been accurately identified. Due to linguistic complexities, two distinct terms reference a singular actor in the example:

- Member of the sales department
- Sales department

In the field of process diagrams, a department is always represented by a representative member. Consequently, the two expressions above refer to a single unit: the sales department.

Implementation: To address the challenge of synonymous terms representing the same actor (as elucidated in Chapter 03), we devised an algorithm. This procedure assesses the similarity between actors prior to appending a new entity to the list of valid actors. This list is subsequently employed for generating both the syntactical structure for the process diagram and the diagram itself.

Algorithm 1 Determine Actor Similarity Utilizing SpaCy's Functionality

Require: $Actor1$: string, $Actor2$: string, nlp : any
Ensure: $similarity\_score$: float compare_actors_with_similarity$Actor1, Actor2, nlp$
1: $doc1 \leftarrow nlp(Actor1)$
2: $doc2 \leftarrow nlp(Actor2)$
3: $similarity\_score \leftarrow$ ROUND($doc1$.similarity($doc2$), 2)
4: **return** $similarity\_score$ =0

Algorithm 1 necessitates two actor strings as inputs. For the calculation of similarity we use the inherent functions of SpaCy. Therefore the use of the "$en\_core\_web\_lg$" pipeline is required and disqualifies the use of the previously used pipeline "$en\_core\_web\_trf$". This is primarily due to the lack of pre-trained word vectors in the transformer pipeline, which are essential for the similarity estimation process. Each vocabulary term possesses a linked vector, a multi-dimensional construct encapsulating semantic nuances determined by contextual associations in extensive corpora. Derived from the input strings, two SpaCy document objects (Doc) are instantiated. Subsequently, SpaCy's in-built 'similarity()' function evaluates the semantic proximity of these documents, contingent on their respective vectors. This operation computes the cosine similarity, interpreting the cosine of the angle delineating two vectors. Cosine similarity values oscillate between -1 and 1. Notably, SpaCy normalizes this value, ensuring the resultant similarity scores range between 0 (indicative of orthogonal vectors, implying dissimilarity) and 1 (identical vectors).

Table 1

Similarity between Actors

	member of sales department	sales department	member of legal department
member of sales department	1	0.67	0.92
sales department	0.67	1	0.46
member of legal department	0.92	0.46	1

As the cosine similarity, is for "member of legal sales" and "member of legal department" pretty close, but as they refer to different entities, we implemented additionally approach that is token based (TODO: Alg 2). This function is designed to compute a similarity ratio between two actor names by comparing the lemmas (base forms) of their tokens, with an emphasis on significant content words.

Again, given two strings representing actors as input parameters, the function processes them through a predefined natural language processing pipeline, denoted as "nlp", to generate respective Doc objects. Prior to any comparison, the function systematically excludes tokens that are characterized as "stop words". Stop words, refer in linguistics and natural language processing to frequently

occurring words in a language that, in analytical contexts, are considered to offer limited semantic value. Subsequently, the function undertakes a pairwise comparison of the lemmas of the tokens derived from the two actors. The objective of this phase is to enumerate the quantity of matching lemmas between the two sets. Acknowledging the potential variance in token counts between different actor strings, the function uses a normalization procedure to ensure a balanced evaluation of similarity without distortion.

Algorithm 2 Compare Actor Tokens Using SpaCy

Require: $Actor1$: string, $Actor2$: string, nlp : any
Ensure: $similarity\_ratio$: float compare_actors_with_token$Actor1, Actor2, nlp$

$doc1 \leftarrow nlp(Actor1)$
$doc2 \leftarrow nlp(Actor2)$
$tokens1 \leftarrow$ FILTER_OUT_STOP_WORDS$(doc1)$
$tokens2 \leftarrow$ FILTER_OUT_STOP_WORDS$(doc2)$
$num\_tokens1 \leftarrow$ length$(tokens1)$
$num\_tokens2 \leftarrow$ length$(tokens2)$
$matching\_tokens \leftarrow 0$
if $num\_tokens1 \leq num\_tokens2$ **then**
  for each $token1$ in $tokens1$ **do**
    for each $token2$ in $tokens2$ **do**
      if $token1.lemma\_ == token2.lemma\_$ **then**
        $matching\_tokens \leftarrow matching\_tokens + 1$
        break
      end if
    end for
  end for
else
  for each $token2$ in $tokens2$ **do**
    for each $token1$ in $tokens1$ **do**
      if $token2.lemma\_ == token1.lemma\_$ **then**
        $matching\_tokens \leftarrow matching\_tokens + 1$
        break
      end if
    end for
  end for
end if
$avg\_tokens \leftarrow (num\_tokens1 + num\_tokens2)/2.0$
if $avg\_tokens > 0$ **then**
  $similarity\_ratio \leftarrow matching\_tokens/avg\_tokens$
else
  $similarity\_ratio \leftarrow 0.0$
end if
return $similarity\_ratio$ =0

member of legal department member of sales department sales department

Table 1

Textual descriptions by source type

	Amount	Frequency
academic	10	50.00%
industry	6	30.00%
textbook	4	20.00%
Total	20	100.00%

Custom Sentencizer wit LS

Including Sentences

More Actors

Regulatory Documents NLP is for unstructured texts

Implementation

Evaluation

Text Input Analysis

For the evaluation

Comparison between the Gold Standard and the Baseline System

Business Processes

Regulatory Documents

Comparison between the Gold Standard and the Proposed Approach

Business Processes

Regulatory Documents

Discussion

Conclusion

Bibliography

[1] H. Leopold, J. Mendling, and A. Polyvyanyy, "Generating natural language texts from business process models," in *Active Flow and Combustion Control 2018*, R. King, Ed., vol. 141, Series Title: Notes on Numerical Fluid Mechanics and Multidisciplinary Design, Cham: Springer International Publishing, 2012, pp. 64–79, ISBN: 978-3-319-98176-5 978-3-319-98177-2. DOI: 10.1007/978-3-642-31095-9_5. [Online]. Available: http://link.springer.com/10.1007/978-3-642-31095-9_5 (visited on 03/07/2023).

[2] F. Friedrich, J. Mendling, and F. Puhlmann, "Process model generation from natural language text," in *Active Flow and Combustion Control 2018*, R. King, Ed., vol. 141, Series Title: Notes on Numerical Fluid Mechanics and Multidisciplinary Design, Cham: Springer International Publishing, 2011, pp. 482–496, ISBN: 978-3-319-98176-5 978-3-319-98177-2. DOI: 10.1007/978-3-642-21640-4_36. [Online]. Available: http://link.springer.com/10.1007/978-3-642-21640-4_36 (visited on 03/05/2023).

[3] K. Winter, H. van der Aa, S. Rinderle-Ma, and M. Weidlich, "Assessing the compliance of business process models with regulatory documents," in *Conceptual Modeling*, G. Dobbie, U. Frank, G. Kappel, S. W. Liddle, and H. C. Mayr, Eds., vol. 12400, Series Title: Lecture Notes in Computer Science, Cham: Springer International Publishing, 2020, pp. 189–203, ISBN: 978-3-030-62521-4 978-3-030-62522-1. DOI: 10.1007/978-3-030-62522-1_14. [Online]. Available: http://link.springer.com/10.1007/978-3-030-62522-1_14 (visited on 01/16/2023).

[4] X. Sun, S. Yang, C. Zhao, and D. Yu, "Design-time business process compliance assessment based on multi-granularity semantic information," en, *The Journal of Supercomputing*, Sep. 2023, ISSN: 0920-8542, 1573-0484. DOI: 10.1007/s11227-023-05626-0. [Online]. Available: https://link.springer.com/10.1007/s11227-023-05626-0 (visited on 10/06/2023).

[5] A. Hevner, "Design science in information systems research," *Management Information Systems Quarterly*, vol. 28.1, 2008.

[6] J. vom Brocke, A. Hevner, and A. Maedche, "Introduction to design science research," in *Design Science Research. Cases*, J. vom Brocke, A. Hevner, and A. Maedche, Eds., Series

Title: Progress in IS, Cham: Springer International Publishing, 2020, pp. 1–13, ISBN: 978-3-030-46780-7 978-3-030-46781-4. DOI: 10.1007/978-3-030-46781-4_1. [Online]. Available: http://link.springer.com/10.1007/978-3-030-46781-4_1 (visited on 01/16/2023).

[7] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A design science research methodology for information systems research,"

[8] Z. Abu-Aisheh, R. Raveaux, J.-Y. Ramel, and P. Martineau, "An exact graph edit distance algorithm for solving pattern recognition problems," presented at the 4th International Conference on Pattern Recognition Applications and Methods 2015, Jan. 10, 2015. DOI: 10.5220/0005209202710278. [Online]. Available: https://hal.science/hal-01168816 (visited on 06/11/2023).

[9] S. Rinderle-Ma, K. Winter, and J.-V. Benzin, "Predictive compliance monitoring in process-aware information systems: State of the art, functionalities, research directions," *Information Systems*, vol. 115, p. 102 210, May 2023, ISSN: 03064379. DOI: 10.1016/j.is.2023.102210. arXiv: 2205.05446[cs]. [Online]. Available: http://arxiv.org/abs/2205.05446 (visited on 06/11/2023).

[10] M. Riefer, S. Ternis, and T. Thaler, *Mining Process Models from Natural Language Text: A State-of-the-Art Analysis*. Mar. 9, 2016.

[11] R. Blumberg and S. Atre, "The problem with unstructured data," *Dm Review*, vol. 13, no. 42-49, p. 62, 2003.

[12] G. Aagesen and J. Krogstie, "BPMN 2.0 for Modeling Business Processes," *Handbook on Business Process Management 1: Introduction, Methods, and Information Systems*, pp. 219–250, Apr. 2015, ISSN: 978-3-642-45099-0. DOI: 10.1007/978-3-642-45100-3_10.

[13] "Business Process Model and Notation (BPMN), Version 2.0," en,

Appendix

Table 1
Your first table

Value 1	**Value 2**	**Value 3**
α	β	γ
1	1110.1	a
2	10.1	b
3	23.113231	c

A note describing the table.

Figure 1
My Figure Caption

A note describing the figure