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

Proposal to Bachelor's thesis

Improved auto-generation of business process models from natural language texts of various complexity

Shuaiwei Yu

Thesis for the Attainment of the Degree
Bachelor of Science

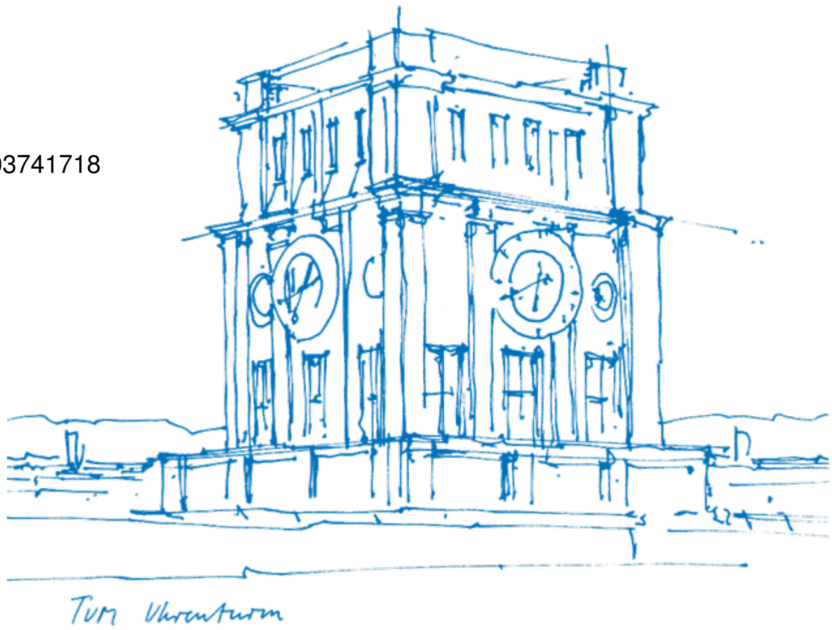
at the TUM School of Computation, Information and Technology,
Department of Computer Science,
Chair of Information Systems and Business Process Management (i17)

Examiner
Prof. Dr. Stefanie Rinderle-Ma

Supervised by
Catherine Sai, M. Sc.

Submitted by
Shuaiwei Yu
Matriculation Number: 03741718

Submitted on
15.04.2023



Contents

List of Tables

List of Figures

Motivation

Business processes are fundamental elements for companies and organizations. They aggregate all the tasks, activities, and timelines involved in companies' workflow whose aim is to provide business services or to create value [3]. Business Process Modeling Notation, also known as BPMN is a modeling language describing such workflows by using graphical notations and thus provides an easily understandable overview of the operations performed in the organization for all business users [14].

Due to the importance of Business processes, leveraging the BPMN techniques can positively affect an organization's performance. Organizations can flexibly adapt to constantly changing business conditions through business process management by providing processes standardization, improvement and quick execution of the activities. [5] However, not everyone is familiar with the BPMN designing techniques. Consequently, managers, along with other process participants prefer using natural language to define business processes. As a result, organizations usually have a large amount of information stored as text documents [3]. However, process modeling is not a simple task, but it is time-consuming, and experts with professional knowledge are required. According to [5], process modeling requires 60% of the time within a business process management project. Adopting the approach that identifies and extracts process models can minimize the time and effort of the process modeling. [12] also suggests that BPMN leads to process improvements, resulting in process cost reduction, quality increment, and higher revenue production.

Over the past years, the development of AI techniques brought solutions to many technical difficulties. Natural Language Processing (NLP), as one of the AI's branches, could possibly address the problem of the difficulties in process modeling. Natural Language Processing is an interdisciplinary discipline focusing on the study of algorithms that enable the computer to understand and process the human language[17]. During the understanding and processing of the natural language text, NLP performs three types of analysis: Firstly, morphological analysis is performed, which analyze the structure of words. The syntactic analysis then explores the grammar relationship between words in sentences, deciding which grammar category the word belongs to. Finally, semantic analysis is executed, which leverages the afore analyses to define the meaning of the text based on the knowledge of sentence structure and the relationship between words [3].

The unique features of the NLP technique make it very suitable for exploiting information from the text documents that record the firm's business process and then analyzing the data to generate the process models automatically. This paper serves as a proposal to suggest using NLP to extract the information from text written in nature language and automatically generate the corresponding business model.

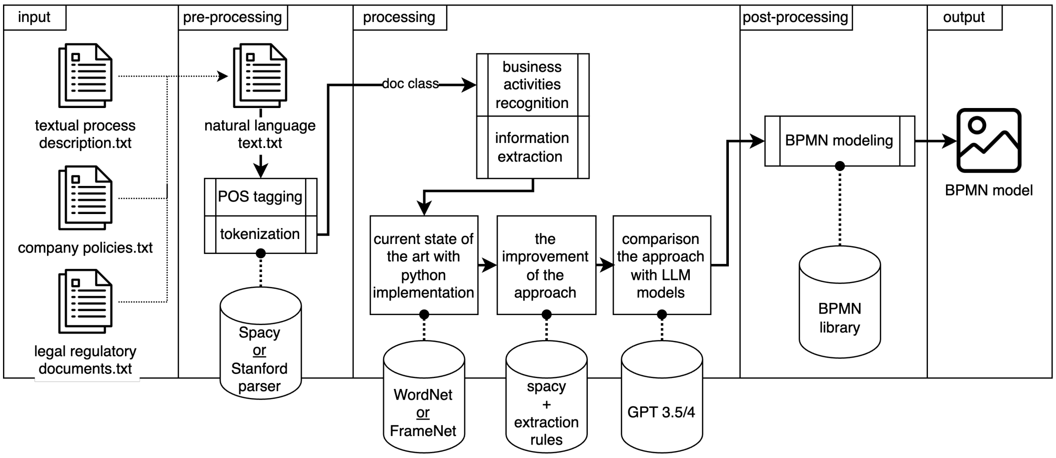
Research Questions

The main research question (**RQ**) is formulated as: "*How can business process models be automatically generated from textual descriptions using Natural language processing techniques?*". To better answer the main research question, three embedded aspects can be revealed: **RQ1**: "*Which NLP methods can be used to extract information?*"; **RQ2**: "*How can the extracted information be analyzed and composed to generate business process models?*" and finally **RQ3**: "*How does the proposed approach perform with different kinds of input documents?*"

Currently, there exist various tools, libraries, and dependencies for NLP. Therefore, the first research question **RQ1** tries to figure out which methods are the most suitable ones to use to extract information from textual descriptions. Currently, there exist several Natural language processing tools in academia and industry, and this work wants to explore their strengths and limitations. The work also aims to use the most suitable state-of-art-technique to perform the information extraction from the natural language text documents. In the next step, **RQ2** explores how can the information in the text descriptions be analyzed and extracted and then be built into the business process model. The activities recognition here is an essential but challenging task here because the identified business activities and actors serve as a basis for the BPMN generation. To address this problem, the work aims to discover the typical structure of the business activities' textual representations in natural language documents by exploring the syntactical and grammatical relationships of the words. Furthermore, reconstructing the extracted business information is also challenging. The work will focus on identifying the conditional and sequential relationships of the business activities to develop algorithms that can generate a business process model with high accuracy. The last research question **RQ3** tries to compare different complexity levels of the natural language documents. This research question aims to discover the adaptability of the proposed approach. The work will reveal the performance of our approach given natural language documents with different complexity levels, e.g., usual textual process descriptions, company policies, legal regulatory documents, and so on.

Since such regulatory documents with higher complexity levels are always formulated in a more complex manner. The work should determine whether the explored extraction strategies still apply to such documents and whether the recognition accuracy drops.

Figure 1
Overview of the approach design



Research Methodology

Design science is a paradigm of real-world problem-solving by creating innovative artifacts. Therefore, Design science research tightly connected the IT artifact with the application domain. Furthermore, the need and desire to improve the current environment and methods motivate Design science research and therefore requires innovative artifacts to address such problems [10]. The work adopted the research methodology of [15] here and followed the research process model given in their work.

Problem centered approach

Although some work in the current field was done, this work aims to develop better tools to automatically extract the business process model for the broad audience of end users, i.e., users within a business organization with little knowledge about business process modeling or underlying technologies. Such motivation provides us with an opportunity to work on creating the tool mentioned above. This problem-centered approach leads us to the first step of the research process, according to [15].

Problem Identification And Motivation

Currently no automated approach exists that can create business process visualizations from text in sufficient quality and generalizability

Objectives of a solution

Our objective is to create an easy-to-use tool that uses the Nature language processing technology to automatically extract information from organizations' texual documents and generate business process models.

Design & development

The development of the new artifact adopted the critical success chain (CSC) method, which uses literature to support and consolidate the conceptual basis of the artifact designing [15]. This work addressed the issues and the needs identified earlier, such as how to find a proper tool to extract information from texual documents or how to process such information to generate a BPMN model. The work conducted a literature review and used the helpful information from the selected papers to combine their ideas and develop our own artifact. The intended artifact is to develop a prototype that leverages the NLP technique to automatically extract business process models from texual documents.

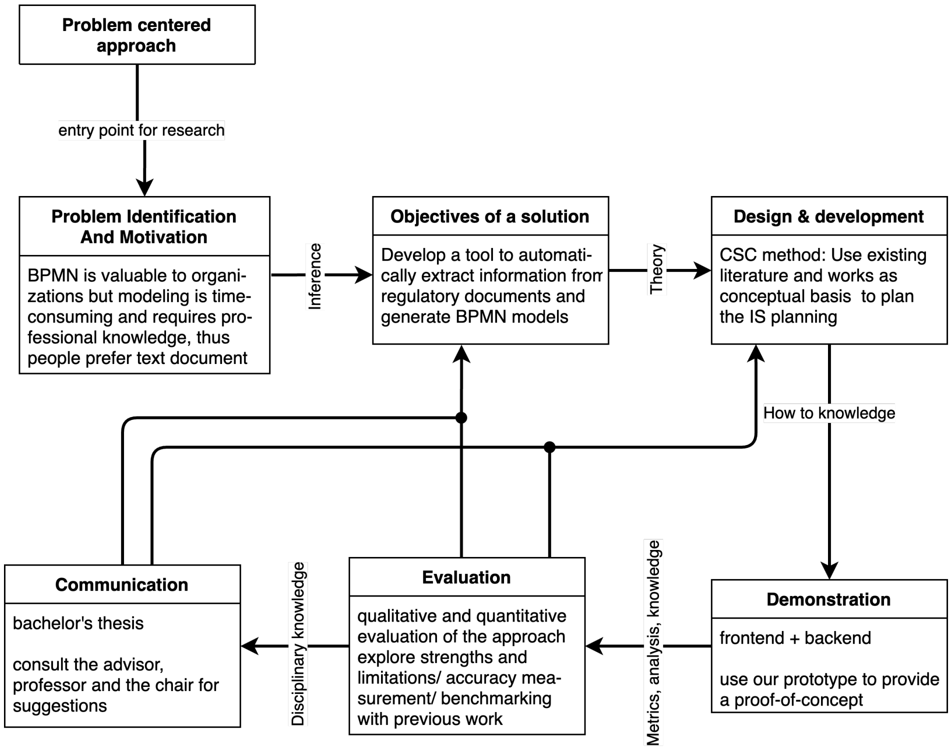
Demonstration

In the demonstration activity, The work want to illustrate how can one use the proposed new artifact to solve instances of problem [4]. The work will use the proposed artifact to convert the textual input into business process models where the users can quickly generate the BPMN models even though they have no knowledge of process modeling. We will run simulations to support our approach and to prove that our approach can minimize the workload of business process modeling.

Evaluation

The evaluation phase is vital in the design of an artifact. The evaluation examines how well the designed artifact solves the problem, which involves comparing the actual output of the problem and the output generated by the artifact [4]. The evaluation is divided into qualitative and quantitative parts. In the qualitative evaluation, The strengths and limitations of the approach are

Figure 2
Design Science Research Process



Overview of the research and artifact design steps for research methodology

explored, and we will communicate with the BPMN modeling experts to check the model compliance of our generated BPMN model. In quantitative evaluation, The accuracy of the approach will be numerically identified. The work intended to first manually generate business process models using several different kinds of documents. In the next step, the same documents will be used by the artifact to automatically generate BPMN models. Then the similarity of these two models will be computed and compared to determine the quality of the artifact. [8] suggests that the metric of *Graph Edit Distance* can be leveraged to achieve such goal. Another aspect of the quantitative aspect is to benchmark the proposed approach with the previous work. The work will look into the approach's accuracy and program execution time compared to the previous work. Based on the evaluation results, we will decide whether to iterate back the step three to modify and improve the artifact if the outcome is not very promising or to continue to the communication phase. Further improvement is also possible to be left to the subsequent projects.

Communication

To ensure the delivery of the desired artifact, every aspect of the problem and the design of the artifact will be communicated and discussed with the relevant stakeholders [4]. Since this is a bachelor's thesis, the primary contact is the author's advisor. Furthermore, the work will also seek advice and suggestions from Prof. Dr. Stefanie Rinderle-Ma and the corresponding Chair of Information Systems and Business Process Management (i17).

Related Work

In order to learn the current state-of-the-art methods of auto-generating business process models and thus answer the research question comprehensively, a systematic literature review must be performed so that what kind of efforts are made can be learned as well as what are the most preferred techniques and what open challenges exist. The literature review is conducted under the guidance of Kitchenham et al. given in their paper [13]. The work consists of several stages: Firstly, the electronic database used to run the search is chosen. Then the selection criteria are defined, and articles are filtered accordingly. After that, a horizontal search will be run to cover as many papers as possible. Finally, a list of the final literature is studied carefully, and helpful information is extracted.

To perform a comprehensive literature review, three most famous electronic databases are chosen, i.e., IEEE, Springer, and ACM. Nevertheless, only using these three databases, There is still a minor chance that some important articles will be missed. Therefore, Google scholar was also used as a complement because it covers a wide range of literature, from conference papers to degree theses. The search string used for the literature review is developed using two phrases, which are the most important ones for our research: *business process model* and *natural language processing*.

In the next step, inclusion and exclusion criteria should be defined. They describe a list of desired and undesired features for the literature selection to obtain relevant studies and support our research and future work. Inclusion criteria were developed as follows: **IC-1**: NLP should have high relevance to the research paper. **IC-2**: BPMN should have high relevance to the research paper. **IC-3**: The research paper should describe the generation of the BPMN model using NLP. Exclusion

Table 1

Overview of Systematic Literature Review Protocol

Database	hits	selected	
IEEE	56	5	
Springer	275	8	
ACM	201	2	
Google scholar	31	3	
Result horizontal search	563	18	
Vertical search		4	add papers
Overall		22	

number of hits using the search string in different databases.

criteria were: **EC-1**: the research paper is not written in English. **EC-2**: The research paper is not in the form of a proper scientific article.

During the literature selection, the first step was to identify duplicates since multiple electronic databases were used. Duplicates refer to articles that have the same title and authors. In the next step, the article's title, abstract and introduction parts were read and the inclusion and exclusion criteria were applied to shape the final result further. Finally, the whole article was read and then a vertical search was performed to identify the related papers used in our selected papers. As a final result, 17 papers were chosen.

Among the chosen papers, [14] [3] [12] [16] are literature reviews that analyzed the development and usage of process model generation methods. [12] points out that the NLP is the most widely adopted method and it can also be combined with other methods to increase accuracy. [14] and [12] give a list of tools for NLP and process model generation that have been used in previous works. [16] compares several papers using NLP to extract process models with different inputs and concludes the typical steps that have to be performed. [18] and [19] propose their findings in identifying the inconsistencies between the textual description and the generated process model. [9] however proposed a new part-of-speech tagging method that is specifically trained for business process management, which can effectively reduce the error rate in grammatical tagging.

A novel breakthrough is made in the work of [8], where a method was developed which extracts information from textual descriptions to automatically generate the business process models regardless of the structure of the input text. The paper gives an excellent overview of the steps that should be executed during the process model extraction and the potential challenge one might encounter. The Authors performed three vital steps to process the text input: (i) syntax parsing using the part-of-speech tagging method, (ii) semantic analysis using FrameNet and WordNet, and (iii) anaphora resolution. Finally, they can generate a process model based on the data. Nevertheless, their work was published in 2012, and some of the techniques could already be outdated. We see great potential here to work on an updated approach to improve the results of this work based on the paper.

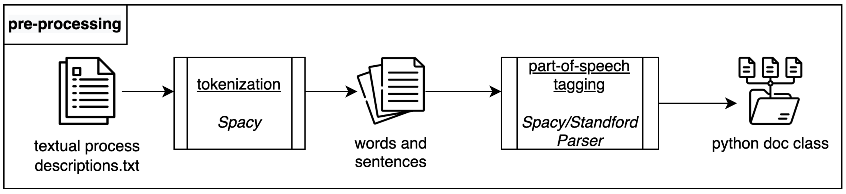
Some limitations of [8] are also addressed in [6]: The textual description must be grammatically correct, otherwise the model will produce an incorrect output. Furthermore, the process in the description must develop sequentially and cannot contain examples or questions. Another work offered in [1] focuses on the extraction of declarative process models to address the problem that many NLP models can only handle the imperative process models. This is done by introducing many grammatical constraints to analyze the relationship of words. This idea illustrates us to also expand the analysis of the semantic analysis so that our model is also able to deal with the textual documents that have a complex description of the process.

Method/Approach (theoretical)

Extracting information from documents written in text is not a simple task due to the nature of the complexity of natural language. [8] identified several obstacles to performing the information extraction: *Syntactic Leeway* describes the problem of inconsistency between the semantic and syntactic aspects of the textual representation. *Atomicity* refers to the problem of adequately mapping the phase-activities. *Relevance* checks whether some part of the text input is irrelevant to the process model, such as examples offered by authors, which helps the human reader to understand the described process but introduces noise for information extraction. *Referencing* deals with the question of how to identify the references between sentences, e.g., the pronouns "This" and "it", from the sentence "After this step, it will be delivered to ...".

Our approach will consider using textual process descriptions as input files in the format of *.txt*. In the first step, the input file will be pre-processed. The documents will be split into sentences and words using tokenization. Correctly identifying the end of sentences is crucial for further information processing. Then, the words in the sentence should be tagged with a proper

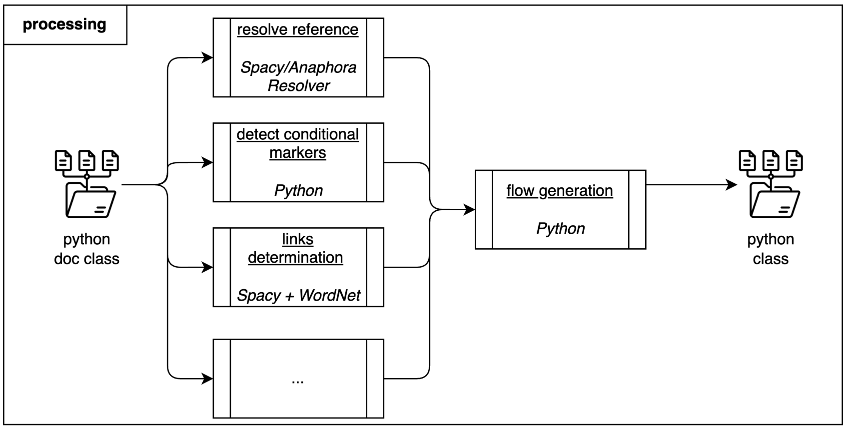
Figure 3
Overview of pre-processing



grammatical label using the part-of-speech technique so that the relationship between words can be analyzed. An essential step of this part is to identify the business activities. [5] and [7] suggest that the identification can be achieved by the pre-defined rules based on the grammatical properties of the words. Once the business activities are identified, they can be used as the fundament of the work.

This step should also tackle the problem of active and passive voice. After all tasks in pre-processing, the tagged documents will be used for the analysis of the relationship between sentences.

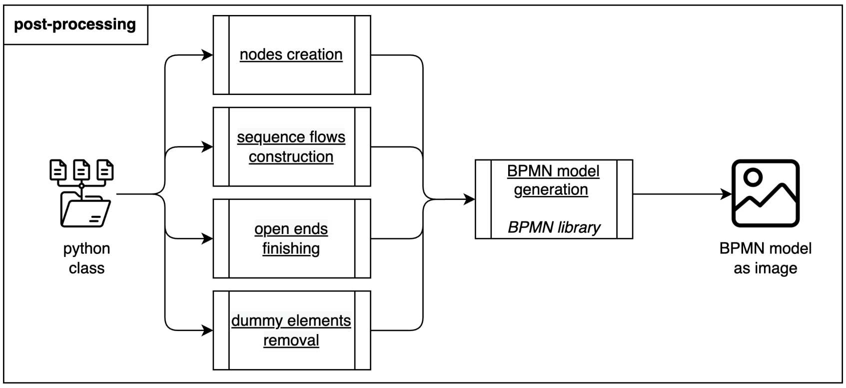
Figure 4
Overview of processing



The primary step of information extraction is text-level analysis, where sentences' sequential, conditional relationships will be exploited. In the pre-processing, the work has already tackled the problem of business activity identification. However, the identified business activities might not be complete because of the referencing problem and active/passive voice problem. In the central part of the work, the anaphora resolution problem has to be solved, which refers to the word that represents a word or a phrase that occurred beforehand [16]. Next, the approach has to solve the problem of finding conditional relationships between sentences. The conditional relationship is usually represented through a conditional word like "if", "else", "otherwise", etc. Finding these relationships is very crucial for the construction of logical conjunctions in the business model. Another essential task in text-level analysis is flow generation. A flow indicates how the business activities are related to each other and could be used to translate the processed information above into the business process model [8].

After the flows of the model are generated, the post-processing phase could now be performed. Post-processing is about generating BPMN representation using the information acquired in the last two steps. [8] suggests four steps of model generation: nodes creation, sequence flows construction, dummy elements removal, and open ends finishing. The nodes and edges will be created first to create the BPMN model. Then the dummy actions will be skipped, which are used to insert between gateways. Finally, the Start and the End events are to be created. As a result, all necessary elements of the BPMN model is created.

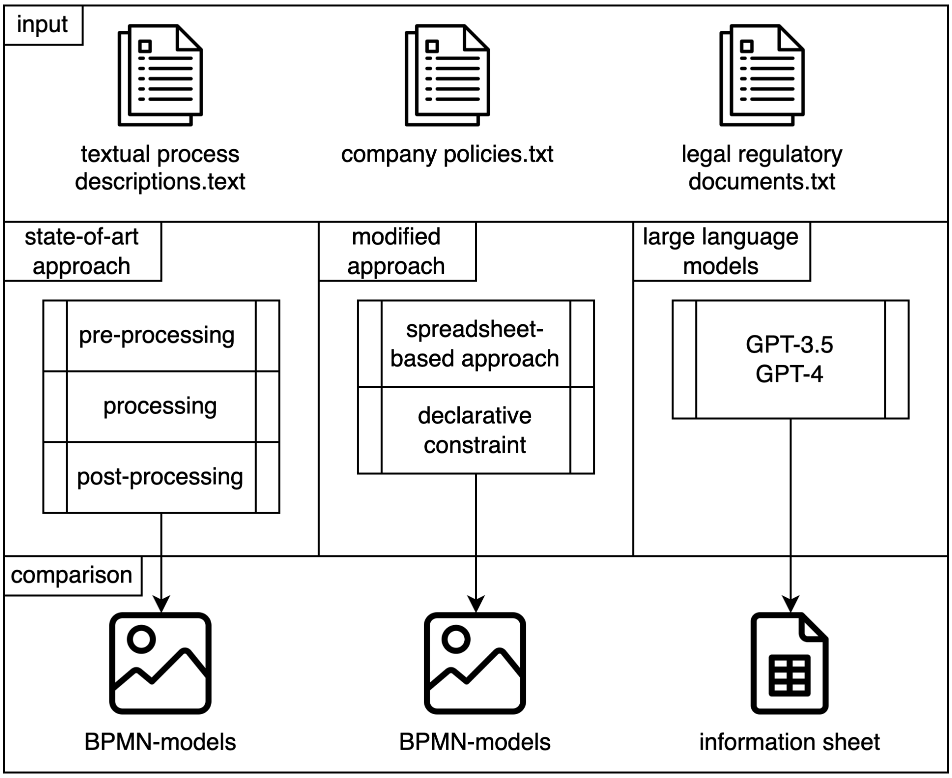
Figure 5
Overview of post-processing



Application (practical)

This part will discuss our model's practical implementation alternatives and programming tools. The work will explore three alternatives for implementing text-to-model transformation. [8] used *Java* as the programming tool in their work, yet their work was published 12 years ago, and during this period, a lot of tools in Artificial intelligence emerged. Therefore, the first step is to reconstruct and implement the current state-of-art approach with more up-to-date technologies. The practical implementation should also be divided into three parts: pre-processing, processing, and post-processing, which correspond to the theoretical design in the last chapter.

Figure 6
Logical partition of approach's functions



Overview of the practical implementation of the proposed approach

After thorough research, *Python* as the programming language with the *Spacy*[1] is decided. *Spacy* is an open-source library for performing natural language processing tasks. This library is chosen because it has a good accuracy of 90.53% on average with a relatively good execution time [7]. The approach will initially take the textual process descriptions with the format of .txt as input.

According to the theoretical design, the input file should be broken into words and tagged with correct grammatical labels. *Spacy* integrates these functions in its core and will make the document pre-processing efficient. The processed sentences are stored in a python *doc* class, which is an object that stores a series of natural language properties of the words, e.g. tokenization, part-of-speech tagging, entity recognitions [7]. Several works, [8] [1] [2], also suggest that the *Stanford parser* is a widely adopted part-of-speech tagging tool with good recognition accuracy and a wide range of *Stanford Dependencies*, which represents the grammatical relationships between words. The work will try to compare these two part-of-speech tagging tools.

[1] https://spacy.io

The information stored in the python *doc* class will be used for the main procedure of text-level analysis. [16] [8] address the anaphora resolution problem using the *WordNet* and *FrameNet*, which are a lexical database of English used to perform semantic analysis. To detect conditional markers, a list of signal words can be predefined. while [8] gives four indicators to identify the signal words: ConditionIndicators, ParallelIndicators, ExceptionIndicators, and SequenceIndicators, which accordingly represents the exclusive gateway, parallel gateway, error intermediate events and continuation of a branch of a gateway. [8] also gives a good illustration of how to generate the flows between activities, which represent their interactions.

Finally, the identified business activities connected using flows can be used to generate BPMN models. A list of rules can be created to convert the flows into the process models. In this step, the work will use the flow information to establish connections between business activities. The start and end activities will first be determined by looking into the semantic meanings of the business activities. End activities should be recognized because once an action leads to an end activity, the node should no longer be connected to other nodes. Then, the work will use the conditional relationship to create the corresponding conditional flows. Then, all the business activities can be joined together. Once such connections are established, they will be converted into images.[14] also suggests a list of BPMN modeling tools that can be leveraged to generate process models, which this work will look further into.

After these processes, A prototype replicating the current state-of-art approach in business model auto-generation is available, but implemented using *Python* and *Spacy*, which is more up-to-date and offers better performance. In the second step, relevant works will be analyzed, and constructive thoughts and ideas will be leveraged to improve the developed prototype qualitatively and quantitatively. The qualitative improvement, on the one hand, aims to extend the functions and usability of the prototype. [11] suggests a spreadsheet-based process model extraction where a spreadsheet is generated and serves as an intermediate media between textual process description and business process model. The spreadsheet provides a good overview of the entire process and allows the human user to check the correctness of the generated process models. Quantitative improvement, on the other hand, refers to the improvement whose aim is to increase the accuracy of the proposed approach.[5], and its subsequential work [7], develop the idea of using rule-based matchers to identify the business process element and summarize a list of XOR and AND gateways to identify the conditional relationships between sentences. [1] focuses on further exploring the relationship between words and designs an algorithm to generate the declarative constraint, allowing the approach to deal with more kinds of input text form.

In the second step, the prototype's performance with input text of different complexity levels will also be examined. The general textual process description will be used as input to generate process model. Yet, this work intends to find out the range of the usability of the developed prototype. Therefore, company policies and legal regulatory documents will then be used which are different structured compared to general textual process description to determine the adaptability of the prototype.

Finally, in the last step, this work will compare the performance of the developed prototype with the large language models, like *GPT-3.5* or *GPT-4* [2]. *GPT-3.5* models are described as being able to understand and generate natural language, while OpenAI describes the *GPT-4* as a large multimodal model and is able to solve complex problems with greater accuracy. Currently, *GPT* models can yet generate process models in the form of images, but we will prompt it to generate a spread containing all process model related information. By comparing these with our prototype, this work is able to gain a deeper understanding of the strengths and limitations of our work and gain insight into future works.

Evaluation

Evaluation is an essential aspect of artifact construction. It allows us to analyze the quality of the proposed artifact and based on which we can decide whether further modification and improvement should be made. The general goal of evaluation is whether the proposed approach satisfies the need of auto-generating BPMN models from textaul documents in actual practice and how well the proposed approach performs in extracting the BPMN model manually. The evaluation can be divided into to part: qualitative and quantitative evaluation.

In the qualitative evaluation, the work plana to explore the pros and cons of the proposed approach. During this phase, the basic components of the approach will be analyzed to find out their strengths and limitations in handling different kinds of text input. For example, we could use *Spacy* or *Stanford Parser* to perform part-of-speech tagging, and a qualitative evaluation is also needed

[2] https://platform.openai.com/docs/models

to see which one is more suitable for the approach. Furthermore, information extraction stage will be examined and analyzed whether the proposed information strategies are logical. Through such inspections, this work could probably gain new insights and are able to optimize the proposed approach. Moreover, we will also consult the experts from the BPMN modeling field for their suggestions on the generated process model to ensure our model complies with the BPMN modeling rules. The objective of the qualitative evaluation is to provide us with the functional information of the approach so that corresponding improvements can be made or relevant such tasks can be left as a potential challenge for future work.

As for quantitative evaluation, the work want to compare how well the automatically extracted BPMN models approximate the models created manually. In order to perform such accuracy checking, this work will, in the first step, collect text documents and other types of documents from different sources and manually model the BPMN model regarding the BPMN modeling rules. Then, the same documents will be used as input for our approach and let the program automatically generates BPMN models. In the next step, the similarity of these two models will be compared by using the metric of *Graph Edit Distance* as proposed in [8], which captures the similarity of two graphs and quantitatively measures them.

Another task this work should perform in the quantitative evaluation is benchmarking the approach with the previous works. The purpose of this task is to examine whether the proposed model offers a novel breakthrough. To achieve this, the documents used in the previous works will be gathered and this work will try to the documents generate the BPMN models. Then calculate the similarity will be calculated and this work will benchmark it with the accuracy rate given in the previous works. Furthermore, the running time for the model generation is also desired. The work would like to find out whether the newly-emerged technique applied offers a better running time compared to the techniques invented years ago.

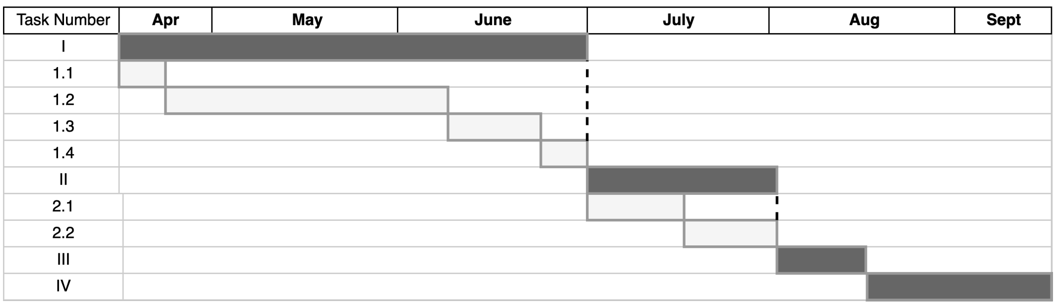
Suppose our prototype generates BPMN models that are highly similar to the models created by experts. In that case, it seems fair to regard our prototype as an effective solution for BPMN model generation and can assist with the time-consuming modeling tasks. However, if the generated model's accuracy is low, the work must head back to the prototype design steps and improve business activity recognition and model generation. If the running time of our approach is too long, some works have to be done to reduce the time complexity of the proposed approach.

Time Plan

The period for the entire project is five months. Suppose our thesis begins on the 15th. April, then the submission deadline is the 15th. September. To ensure the delivery of the project with good quality, a time plan should be developed to arrange the tasks that should be performed during these five months. The tasks should be distributed fairly within the project period so that the tasks can be accomplished with caution and precision.

The tasks for the projects can be divided into three phases and each phase contains one or several parts: **(I) Implementing the state-of-art approach**: **(1.1)** Pre-processing of the input documents, where the input documents are separated into sentences, and the part-of-speech tagging is performed. **(1.2)** Information extraction, where we will focus on extracting further information from the sentences. This part is significant work for the whole approach. During this step, the active/passive voice issue will be addressed; the anaphora problem will be solved; the conditional markers will be detected; and all other necessary steps will be performed in this step. **(1.3)** Flow generation, where each business activity's interactions are identified and connected. **(1.4)** Post-processing, where we use the flows to generate a BPMN model and convert it into images. **(II) modify the approach**: **(2.1) modify the approach**: **(2.1)** implement the spreadsheet methods **(2.2)** rule-based activities recognition, where we will develop the rules for identifying the business rules in the text documents. **(III) benchmark with the LLM** and finally **(VI)** Evaluation and redesign, where we evaluate our approach and make improvements regarding the identified limitations of our approach.

Figure 7
Gantt chart of project implementation



Bibliography

[1] Han van der Aa et al. "Extracting declarative process models from natural language". In: *Advanced Information Systems Engineering: 31st International Conference, CAiSE 2019, Rome, Italy, June 3–7, 2019, Proceedings 31*. Springer. 2019, pp. 365–382.

[2] Lars Ackermann and Bernhard Volz. "Model [nl] generation: natural language model extraction". In: *Proceedings of the 2013 ACM workshop on Domain-specific modeling*. 2013, pp. 45–50.

[3] Ana Cláudia de Almeida Bordignon et al. "Natural language processing in business process identification and modeling: a systematic literature review". In: *Proceedings of the XIV Brazilian Symposium on Information Systems*. 2018, pp. 1–8.

[4] Jan vom Brocke, Alan Hevner, and Alexander Maedche. "Introduction to design science research". In: *Design science research. Cases* (2020), pp. 1–13.

[5] Renato César Borges Ferreira, Lucinéia Heloisa Thom, and Marcelo Fantinato. "A Semi-automatic Approach to Identify Business Process Elements in Natural Language Texts." In: *ICEIS (3)*. 2017, pp. 250–261.

[6] Renato César Borges Ferreira et al. "Assisting process modeling by identifying business process elements in natural language texts". In: *Advances in Conceptual Modeling: ER 2017 Workshops AHA, MoBiD, MREBA, OntoCom, and QMMQ, Valencia, Spain, November 6–9, 2017, Proceedings 36*. Springer. 2017, pp. 154–163.

[7] Thomas Freytag et al. "NLP as a Service: An API to Convert between Process Models and Natural Language Text." In: *BPM (PhD/Demos)*. 2021, pp. 146–150.

[8] Fabian Friedrich, Jan Mendling, and Frank Puhlmann. "Process model generation from natural language text". In: *Advanced Information Systems Engineering: 23rd International Conference, CAiSE 2011, London, UK, June 20-24, 2011. Proceedings 23*. Springer. 2011, pp. 482–496.

[9] Xue Han et al. "A novel part of speech tagging framework for NLP based business process management". In: *2019 IEEE International Conference on Web Services (ICWS)*. IEEE. 2019, pp. 383–387.

[10] Alan R Hevner et al. “Design science in information systems research”. In: *Management Information Systems Quarterly* 28.1 (2008), p. 6.

[11] Krzysztof Honkisz, Krzysztof Kluza, and Piotr Wiśniewski. “A concept for generating business process models from natural language description”. In: *Knowledge Science, Engineering and Management: 11th International Conference, KSEM 2018, Changchun, China, August 17–19, 2018, Proceedings, Part I 11*. Springer. 2018, pp. 91–103.

[12] Uce Indahyanti, Arif Djunaidy, and Daniel Siahaan. “Auto-Generating Business Process Model From Heterogeneous Documents: A Comprehensive Literature Survey”. In: *2022 9th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*. IEEE. 2022, pp. 239–243.

[13] Barbara Kitchenham. “Procedures for performing systematic reviews”. In: *Keele, UK, Keele University* 33.2004 (2004), pp. 1–26.

[14] Bilal Maqbool et al. “A comprehensive investigation of BPMN models generation from textual requirements—techniques, tools and trends”. In: *Information Science and Applications 2018: ICISA 2018*. Springer. 2019, pp. 543–557.

[15] Ken Peffers et al. “A design science research methodology for information systems research”. In: *Journal of management information systems* 24.3 (2007), pp. 45–77.

[16] Maximilian Riefer, Simon Felix Ternis, and Tom Thaler. “Mining process models from natural language text: A state-of-the-art analysis”. In: *Multikonferenz Wirtschaftsinformatik (MKWI-16), March* (2016), pp. 9–11.

[17] Konstantinos Sintoris and Kostas Vergidis. “Extracting business process models using natural language processing (NLP) techniques”. In: *2017 IEEE 19th conference on business informatics (CBI)*. Vol. 1. IEEE. 2017, pp. 135–139.

[18] Han Van der Aa, Henrik Leopold, and Hajo A Reijers. “Detecting inconsistencies between process models and textual descriptions”. In: *Business Process Management: 13th International Conference, BPM 2015, Innsbruck, Austria, August 31–September 3, 2015, Proceedings 13*. Springer. 2015, pp. 90–105.

[19] Sheeza Zaheer, Khurram Shahzad, and Rao Muhammad Adeel Nawab. “Comparing manual- and auto-generated textual descriptions of business process models”. In: *2016 Sixth International Conference on Innovative Computing Technology (INTECH)*. IEEE. 2016, pp. 41–46.

Declaration of Academic Integrity

I hereby declare that the thesis submitted is my own unaided work. All direct or indirect sources used are acknowledged as references.

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, 15.04.2023 Shuaiwei Yu