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

Auto-generation of business process models using natural language processing

Shuaiwei Yu

Thesis for the Attainment of the Degree
Bachelor of Science

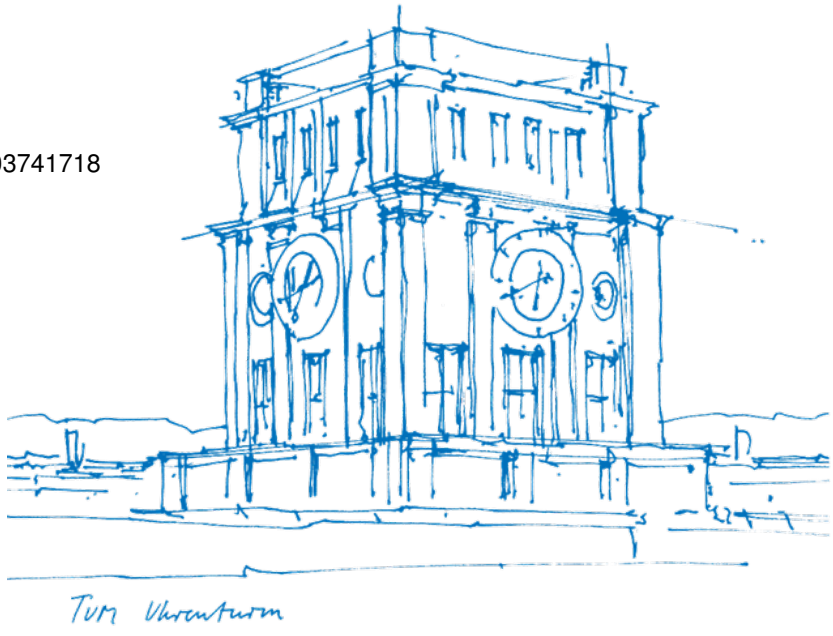
at the TUM School of Computation, Information and Technology,
Department of Computer Science,
Chair of Information Systems and Business Process Management (i17)

Examiner
Prof. Dr. Stefanie Rinderle-Ma

Supervised by
Catherine Sai, M. Sc.

Submitted by
Shuaiwei Yu
Matriculation Number: 03741718

Submitted on
15.04.2023



Abstract

Business process models visualize the abstract business process into an intuitive graphical representation. By adopting them in business logic, an organization can potentially increase its productivity. Yet, modeling business process models is a complicated work for people who receive little modeling training, and thus they tend to document the business process using text descriptions. There is an urgent need to develop a tool for organizations to translate text documents into business process models.

In this paper, we tackle this problem by using the natural language processing (NLP) technology to extract information from text descriptions and automatically generate the BPMN models. We analyzed the prior works of many scientists and developed a tool with an easy-to-use web interface in an innovative way.

Keywords: *Nature language processing, business process models, Artificial intelligence.*

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 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 [11].

Due to the importance of Business processes, leveraging the BPMN techniques can positively affect an organization's performance and thus increase its productivity. 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]. There is a need to translate text documents into the process model regarding such a situation. However, process modeling is not a simple task, but is time-consuming and experts with professional knowledge are required.

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[14]. 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 generated from regulatory documents automatically using the Nature language processing technique?*". 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 regulatory documents. The methods should be able to separate sentences, label each word in a sentence with corresponding syntactical tags, and analyze the grammatical relationships between words. By doing so, we are able to explore the information hidden behind the natural language and thus use them for further operation. In the next step, **RQ2** explores how to use syntactical and grammatical information to determine events of business processes, identify the conditional restraints ("and" or "or"), and the sequential orders of business processes. Once such information is acquired, an algorithm should be developed to combine all the business processes in a logical order. In the end, the composed process model should be well visualized. The last research question **RQ3** tries to discover the adaptability of the proposed model: How well does the method perform with the document other than the regulatory document? Does the accuracy of the outcome decrease with other kinds of documents? A sets of different input documents will be prepared and a corresponding benchmark will be performed.

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 [8]. We adopted the research methodology of [12] here and followed the research process model given in their work.

Problem centered approach

Although some work in the current field was done, we wanted 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 [12].

Problem Identification And Motivation

Modeling business processes requires experts with relevant knowledge and can be exhausting and time-consuming. Thus, small companies usually cannot afford such experts and business managers usually prefer to describe processes using natural language. As a result, an organization usually process a large amount of text documents [3]. However, the business process provides an intuitional overview of the business process and can potentially increase a company's productivity.

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' regulatory 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 [12]. We addressed the issues and the needs identified earlier, such as how to find a proper tool to extract information from regulatory documents or how to process such information to generate a BPMN model. We 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 regulatory documents.

Demonstration

In the demonstration activity, we want to illustrate how we can use our new artifact to solve instances of problem [4]. We plan to implement a web-based front-end so that customers can use our artifact rather easily, even if they have no knowledge of programming. The customer can enter the textual description which is a regulatory document into an input field on the website and click a button to send the request to convert the text to the BPMN model. Then the regulatory document will be sent to the backend and processed there. Finally, the customer will be able to see a BPMN model presented in an image on the website.

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 the actual output of the problem and the output generated by the artifact [4].

The evaluation measures how well the artifact supports a solution to the problem. This activity involves comparing the objectives of a solution to actual observed results from use of the artifact in context. Depending on the nature of the problem venue and the artifact, evaluation could take many forms. At the end of this activity the researchers can decide whether to iterate back to step three to try to improve the effectiveness of the artifact or to continue on to communication and leave further improvement to 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, we 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

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		2	add papers
Overall		20	

be performed so that we can learn what kind of efforts are made 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 [10]. 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, we chose three most famous electronic databases, 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, we also used Google scholar 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 keywords, 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 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, we read the article's title, abstract and introduction parts and applied the inclusion and exclusion criteria to shape the final result further. Finally, we read the whole article and then performed a vertical search to identify the related papers used in our selected papers. As a final result, 17 papers were chosen. Among the chosen papers, [11] [3] [9] [13] are literature reviews that analyzed the development and usage of process model generation methods. [9] points out that the NLP is the most widely adopted method and it can also be combined with other methods to increase accuracy. [11] and [9] give a list of tools for NLP and process model generation that have been used in previous works. [13] compares several papers using NLP to extract process models with different inputs and concludes the typical steps that have to be performed. [15] and [16] propose their findings in identifying the inconsistencies between the textual description and the generated process model.

A novel breakthrough is made in the work of [7], where they developed a method which extracts information from textual descriptions to automatically generate the business process models regardless of the structure of the input text. The Authors performed three vital steps to process the text input: (i) the syntax parsing using the Stanford Parser, (ii) semantic analysis using FrameNet and WordNet, and (iii) anaphora resolution. Finally, they can generate a process model based on the data. Some limitations of this work are addressed in [5]: 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. Among all works, with little focus on the visualization of the process model, [6] introduces a web-based NLP model extraction service. However, this work leverages the extraction model of others, and thus the output accuracy cannot be well guaranteed. A possible research gap is to leverage the findings in [5] and [1] to overcome the weakness of the model in [7]. Moreover, [6] provides a good idea of visualization, where a front-end interface is implemented.

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. [7] 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 model will use regulatory documents as input files. In the first step, the input file will be pre-processed. The documents will be split into sentences 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 grammatical label so that we can analyze the relationship between words. In the pre-processing, it is also very important to identify the business process elements, such as the actors and actions. 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.

The major step of information extraction is text-level analysis, where the sequential, conditional relationships of sentences will be exploited. In the central part of our work, we have to solve the anaphora resolution problem, which refers to the word that represents a word or a phrase that occurred beforehand [13]. Next, we have 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 the text-level analysis is the 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 [7].

After the flows of the model are generated, we could now perform the post-processing phase. Post-processing is about generating BPMN representation using the information acquired in the last two steps. [7] 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. [6] illustrate us to additionally implement a web interface that eases to use of the regulatory document to BPMN model transmission service. The web interface should take the text description as input and then will represent the BPMN model in the webpage after the text is processed with our model in the backend.

Application (practical)

This part will discuss our model's practical implementation and programming tools. 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. [7] 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. Thus, we decided to use *Python* as the programming language with the *Spacy*[1], which is an open-source library especially for performing natural language processing tasks. The reason why we choose this library is that it has a good accuracy of 90.53% on average. Besides, it also has a relatively good execution time [6].

First, the input file should be the regulatory documents stored in raw text form, e.g., company policies, laws, etc. According to the theoretical design, the input file should be broken into sentences 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 contains every word in the sentence stored in an object. Such an object possesses all the features of the word, such as the word's tokenization, part of speech tagging, sentence recognition, and so on [6]. Several works [7] [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.

[1] https://spacy.io

The information stored in the python *doc* class will be used for the main procedure of text-level analysis. [13] [7] 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. [6] summarized the list of XOR and AND gateways, while [7] gives four lists: ConditionIndicators, ParallelIndicators, ExceptionIndicators, and SequenceIndicators, which accordingly represents the exclusive gateway, parallel gateway, error intermediate events and continuation of a branch of a gateway. [7] 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 for the generation of BPMN models. We can create a list of rules to convert the flows into the process models. [11] also suggests a list of BPMN modeling tools that can be leveraged to generate process models, which we will look further into. As mentioned, a web-based frontend should be created to increase the usability of our conversion model. The plan is to use *Vue* as the framework to create such a website. The website should be able to have a text input area where the user can enter the regulatory text. Then the text will be delivered to our *Python* backend and processed. The final results will be sent back to the website and displayed in the form of an image.

Evaluation

Time Plan

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 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.

[6] 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.

[7] 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.

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

[9] 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.

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

Appendix

Table 2
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

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