2017 IEEE 19th Conference on Business Informatics

Extracting Business Process Models using Natural Language Processing (NLP) Techniques

Konstantinos Sintoris
Department of Applied Informatics
University of Macedonia
Thessaloniki, Greece
k.sintoris@uom.edu.gr

Kostas Vergidis
Department of Applied Informatics
University of Macedonia
Thessaloniki, Greece
kvergidis@uom.gr

***Abstract*— This Doctoral Consortium paper discusses how NLP can be applied in the domain of BPM in order to automatically generate business process models from existing documentation within the organization. The main idea is that from the syntactic and grammatical structure of a sentence, the components of a business process model can be derived (i.e. activities, resources, tasks, patterns). The result would be a business process model depicted using BPMN – a dedicated business process modeling technique**

Keywords— Business Process; Natural Language Processing; Information Extraction; Business Process Modelling; Business Process Management Systems; Part-Of-Speech Tagging;

I. Context and Motivation of Research

Business Process Management (BPM) can be described as a management discipline. Every business, regardless its size and its final product, follows some procedures and performs certain activities in order to reach a result. BPM aims to organize and orchestrate as much efficiently as possible [1]. The above description gives an idea of what BPM is concerned, as a universally accepted definition is yet existent. A comprehensive definition is the following.

"Supporting business processes using methods, techniques, and software to design, enact, control, and analyze operational processes involving humans, organizations, applications, documents and other sources of information." [2]

Natural Language Processing (NLP) is an interdisciplinary field which studies and develops algorithms and systems enabling computers to understand and perform tasks involving human language. NLP can also be referred to as computational linguistics, computer speech and language processing or human language technology[3]. NLP is concerned with analyzing human language both in written and oral form and consequently extract commands or useful information.

II. Research Aim and Objectives

The aim of the present research is to explore how NLP can be applied in the BPM domain in order to automatically generate business process models from existing documentation within an organization.

The main idea is that from the syntactic and grammatical structure of a sentence the components of a business process model can be derived (i.e. activities, resources, tasks, patterns). The result would be a business process model depicted using BPMN – a dedicated business process modeling technique.

The extraction can be achieved by applying well known NLP techniques such as Part of Speech tagging, Named Entity Recognition and Co-reference Resolution. The ultimate goal is the proposal of a system which is able to generate conceptual models needed for a BPM-initiative from text. Using our approach a system analyst is relieved from the time-intensive modeling task. By pursuing this research goal, we evaluated related works in the field of BPM and NLP, developed a novel transformation approach, and created and evaluated it. Therefore, the following contributions will be considered:

- Literature Review where several works will be collected, dealing with the problem of automated process model generation and evaluated their strengths and weaknesses.
- Categorization of Issues. Based on the above findings a theoretical framework will be developed which categorizes important issues which are relevant for a transformation approach.
- Novel Transformation Approach. The theoretical framework will be transferred into practice and a novel transformation approach will be developed
- Comprehensive Test Data Set. In order to evaluate the above approach a comprehensive test data set containing different process descriptions and manually created models covering various domains will be collected.

2378-1971/17 $31.00 © 2017 IEEE
DOI 10.1109/CBI.2017.41

- Evaluation Approach. To assess the accuracy of our transformation approach an evaluation methodology using graph edit distance will be created reporting the results.

The research methodology applied can be divided in the following stages:

- Literature Review. In this stage the author will study in depth the Natural Language Processing field gaining a firm grasp of the techniques and current methods used in the domain as well as technical experience in working with existing tools. The same applies for Business Process Management and Modelling which will be the second main subject of this study.
- Prototype Tool. This stage involves thorough investigation, testing and experimentation with a prototype software tool that utilizes NLP to generate business process models in BPMN diagrams.
- Analysis of Syntactic and Semantic Patterns By analyzing the linguistic patterns we will be able to map text constructions to their corresponding modelling fragments. This also reveals issues regarding the quality of the syntax parse and the text itself.
- Transformation Rules Derivation. To effectively mitigate the detected issues we will derive appropriate transformation heuristics. The rules will be implemented in our research prototype to assess the output.
- Modification, customization & improvement. Our contributions will be made in the form of upgrades, alterations and modifications to the key components of the existing tool.
- Evaluation. In this step a series of experiments will be conducted to determine the performance of the prototype tool before and after our additions.

III. Background/related research

Applications of NLP are becoming more widespread each day. Modern smartphones are shipped with some kind of intelligent personal assistant like Apple's Siri and Google's Now. These programs let you control your phone and search the Web [4] with voice commands. Then they answer you with the results and the information requested. Thus we have complete conversational agents as they accept language input (speech recognition) and language output (speech synthesis).

Another great example where NLP is used to improve human communication is machine translation. Machine translation aims in automatically translate a text from one language to another. Nowadays machine translation has reached a point where it can produce pretty accurate results, due to the large scientific research and enterprise effort of the last decade.

Finally, the Web has evolved and grown so much that the information it contains, are overwhelming. Though search engines are doing a great job at organizing and retrieving the appropriate web pages, still users have to find manually the wanted information, regardless how specific their query was. Researchers are trying to develop question answering (QA) systems utilizing many aspects of NLP. QA can be considered a whole new discipline as it relies heavily on modeling knowledge and information retrieval apart from NLP.

IV. Thesis and problem statement

Information Extraction (IE) is the task of extracting structured information from unstructured or semi-structured text. In general, IE is achieved by applying Natural Language Processing (NLP) tasks and techniques. There is an increasing amount of textual information residing in digital form, thus IE finds use in a vast range of domains. Financial professionals already use IE and NLP to seek specific pieces of information from news articles to discover mergers and aid their decision making. In the same manner the various business process documentations can be treated in an automated way using little or no input from an expert business process analyst at least in the initial generation of a business process model solely from text-based resources.

The research will unfold all the aspects of the problem in its current situation and will also describe in detail the solution that is proposed. In more detail:

- The literature review discusses the main concepts of the present study. It will be divided in two main sections: The first section will introduce the natural language process and its main concepts. The second section will be concerned with the definition of business processes and the modeling approaches.
- The solution approach will be presented to the text-to-model problem and the code of the implemented prototype tool will also be examined.
- The contributions made on the prototype tool and the reason behind of each addition and the way they were implemented will be described.
- The evaluation method will be explained.
- The results that the prototype tool generates will be demonstrated and analyze.
- The conclusions extracted from the experiments conducted and also discusses about contribution to knowledge and the limitations of the current work as well as potential research impact will be provided.

V. Transformation Challenges

Natural language is very ambiguous, thus making the creation of computer applications that completely understand human language very challenging. Friedrich [5] presents in detail a collection of challenges that have to be overcome in order to analyze natural language process descriptions successfully is presented by Friedrich [5] . Here they are discussed briefly.

A. Semantics vs. Syntax

Different syntactic structures can be used, maintaining the same meaning. So this category distinguishes three key issues.

- Active - Passive Voice

- Rewording/ Word Order
- Implicit Conditions

B. Atomicity

Atomicity describes the issue of how a sentence is transformed to activities. Sentences consist of several relative clauses, which can contain information relevant to the activity or a whole new activity. Three issues arise.

- Complex Sentences
- Action Split over Several Sentences
- Relative Clauses

C. Relevance

Inside the process description one can find sentences that are relevant to the process but do not add useful information to the final model. Also sentences often are examples, in order to help the human reader comprehend the process. Another observed issue is when the process is described on a meta level, *"After the Process starts, a Task is performed to […]"*. So we have the following issues.

- Relative Clause Importance
- Example Sentences
- Meta-Sentences

D. Referencing

One main issue of referencing is anaphora resolution, which was analyzed on the previous chapter. Moreover sentences usually refer to other sentences, namely creating textual links, that way affecting the flow of the process model. The author distinguishes three types of links **forward**, **backward** and **jumps**. Finally, a great challenge is how to determine when two branches of a process should join back together. So this category consists of the following issues.

- Anaphora
- Textual Links
- End-of-block Recognition

VI. Solution Approach

The issues stated above show the complexity of transforming text into business process models. Friedrich created a solution in theoretical as well as a practical level. The solution is implemented as a Java project and combines different computational linguistics tools. These are *Stanford Parser*, *WordNet*, *FrameNet* and an *Anaphora Resolution Technique*. It is divided in three distinct phases, the **sentence level analysis**, the **text level analysis** and the **process model generation**. The first two phases include algorithms that utilize basic NLP tasks presented in section 2. In the last stage the extracted information is mapped to BMPN elements and the final model is created. The different stages communicate with each other with an ontology called *World Model.* This ontology serves as an intermediate data structure and is inspired from field of robotics [6].

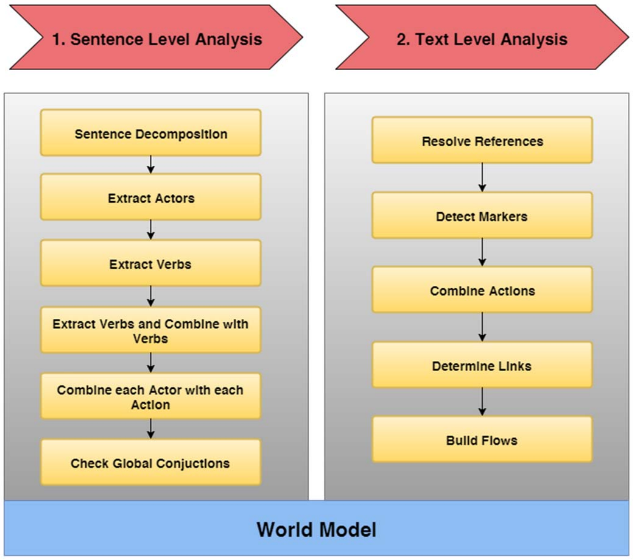

Figure 1. First two phases of the solution

The project has two main components. The first component, which will be called *Text-to-Process* from now on, is responsible for processing the text input and generating the model. The second component is responsible for visualizing the model using the BPMN language. Also it gives us the ability to edit, save and export the model. Actually this part is an external tool, called *Process Editor*, a fully equipped business process editor made in Java. This part is not discussed any further as it is not aim of this thesis.

The three distinct phases, the **sentence level analysis**, the **text level analysis** and the **process model generation** are being described below:

E. Sentence Level Analysis

First step of the sentence level analysis is the preprocessing of the text. In this step the text is loaded and split in sentences using the Stanford *sentence split* algorithm. Then the execution proceeds with *Stanford Parser* so that the syntactic and grammatical structure of the sentence is acquired. In other words, *POS tagging* [7] is performed and then *Stanford Dependencies* [8] are extracted. Important here is to note, that for the parser the English Factored model is used and all algorithms later on use the typed collapsed dependencies. The second step is to break down the sentences into phrases and will be discussed in a later stage of the research.

F. Text Level Analysis

In the Text level analysis firstly is to be defined, if a determiner is to be resolved, then a corresponding Action is searched, else if it is a pronoun then an Actor is searched.

Second step of this stage is the detection of conditional markers. These markers can either be a single word like “if”, “then”, “meanwhile” or “otherwise”, or a short phrase like “in the meantime” or “in parallel”. All of these markers have specific characteristics and can be mapped to different BPMN constructions, mainly gateways.

The third step is called action combination. Here actions that are split up over several sentences are merged into one. To consider two Actions as a candidate for a merger, a reference had to be established between them during the anaphora resolution phase.

Next step is to determine the links that exist between actions. If two actions should be combined a link is established. Last step is to determine which of the three defined types of textual reference it is, namely forward, backward or jump reference.

The last step of the text level analysis is the generation of Flows. A flow can be an activity, event or gateway BPMN element. Therefore, during the process model generation such Flows can be translated to BPMN connecting objects. In this step the previously extracted actions are matched to nodes and the gateways are determined based on the conjunctions words used in the text. Important assumption about the creation of the Flows is that a process is described sequentially.

G. Process Model Generation

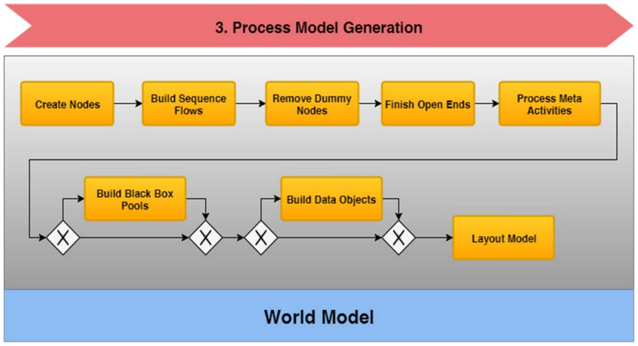

Figure 2. First two phases of the solution

The last phase of the transformation approach can be divided in two stages, consisting of nine steps in total, as depicted in figure 2. In the first stage the model is created, information contained in the World Model is transformed into its BPMN representation.

- First step is to create the nodes of the model
- The second step is to connect the nodes with the three different types of connecting objects.

VII. Results to Date and Their Validity

The research is still at an early stage.

VIII. Status of Work and Next Steps

The tool when fully developed will be capable system that aids the extraction of Business Process Models from existing specifications or completely automates it without the intervention of a dedicated business process analyst at least at the early stages of model generation.

The most important outcome of the research that needs to be highlighted is that natural language processing will be applied successfully to business processes modelling in terms of approach and methodology and independently of the results quality. Natural language processing proves to be of high importance in this new era of Mobile and Cloud computing as well as to the emerging Internet of Things. It can help to cope with this vast abundance of information that surrounds us and can promote human-computer interaction to a whole new level. Last but not least NLP already gives a competitive advantage to the products and businesses that utilize it. Consequently it is high time for the field of Business Process Management to benefit from it.

IX. Current and Expected Contributions

The prototype tool as a software project will be written in Java and will combine three major independent linguistic tools, namely the Stanford Parser, WordNet and FrameNet. Stanford Parser among these is the more substantial component of the prototype and has been under active development from Stanford. Moreover Stanford has integrated its various projects into one fully equipped suite of linguistic tools, the CoreNLP.

Therefore, as a first contribution it is decided the project to be upgraded by using CoreNLP and not to limit only to the newest parser libraries. This decision provides the opportunity not only to take advantage of any improvements made to the parser but also to incorporate new tools in our project as it is demonstrate by the next contribution.

The second contribution is driven by three facts. Reference resolution plays a very important role in every NLP application that seeks to extract information, as in this case. The second fact is that the existing tool uses a simple custom made anaphora resolution technique which performance had room for improvement. The third fact is that CoreNLP offers the state-of-the-art Stanford Coreference Resolution system. Therefore it is reasonable to test if it can offer any significant improvement to the final result.

The third and last contribution is the Semi-manual GED Similarity Framework. The Graph Edit Distance (GED) metric was selected as a similarity measure between business process models and it was implemented into a framework to facilitate evaluation of the generated process models. This contribution serves as an evaluation metric on how successful the computer generated model was compared to the one modelled from an expert.

X. Conclusion

The most important outcome of the thesis that needs to be highlighted is that natural language processing is about to be applied to business processes modeling in terms of approach and methodology and independently of the results quality. Natural language processing proves to be of utmost importance in this new era of Mobile and Cloud computing as well as to the emerging Internet of Things. It can help humans to cope with this vast abundance of information that surrounds us and can promote human-computer interaction to a whole new level. Last but not least NLP already gives a competitive advantage to the products and businesses that utilize it. Consequently it is high time for the field of Business Process Management to benefit from it.

XI. References

[1] M. Weske, *Business Process Management*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012.

[2] W. M. Van Der Aalst, A. H. Ter Hofstede, and M. Weske, “Business process management: A survey,” in *Business process management*, Springer, 2003, pp. 1–12.

[3] D. Jurafsky and J. H. Martin, *Speech and language processing*, vol. 3. Pearson, 2014.

[4] A. Franz and B. Milch, “Searching the web by voice,” in *Proceedings of the 19th international conference on Computational linguistics-Volume 2*, 2002, pp. 1–5.

[5] F. Friedrich, “Automated generation of business process models from natural language input,” Citeseer, 2010.

[6] C. B. Achour, “Writing and correcting textual scenarios for system design,” in *Database and Expert Systems Applications, 1998. Proceedings. Ninth International Workshop on*, 1998, pp. 166–170.

[7] M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini, “Building a large annotated corpus of English: The Penn Treebank,” *Computational linguistics*, vol. 19, no. 2, pp. 313–330, 1993.

[8] M.-C. De Marneffe and C. D. Manning, “Stanford typed dependencies manual,” Technical report, Stanford University, 2008.