2021 47th Euromicro Conference on Software Engineering and Advanced Applications (SEAA) | 978-1-6654-2705-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/SEAA53835.2021.00009

Using Natural Language Processing to Build Graphical Abstracts to be used in Studies Selection Activity in Secondary Studies

Vinicius dos Santos
University of São Paulo (ICMC/USP)
São Carlos, SP, Brazil
vinicius.dos.santos@usp.br

Érica Ferreira de Souza
Federal University of Technology – Paraná (UTFPR)
Cornélio Procópio, PR, Brazil
ericasouza@utfpr.edu.br

Katia Romero Felizardo, Willian Massami Watanabe
Federal University of Technology – Paraná (UTFPR)
Cornélio Procópio, PR, Brazil
katiascannavino, wwatanabe@utfpr.edu.br

Arnaldo Cândido Júnior
Federal University of Technology – Paraná (UTFPR)
Medianeira, PR, Brazil
arnaldoc@utfpr.edu.br

Sandra Maria Aluísio
University of São Paulo (ICMC/USP)
São Carlos, SP, Brazil
sandra@icmc.usp.br

Nandamudi Lankalapalli Vijaykumar
National Institute for Space Research (INPE)
São José dos Campos, SP, Brazil
vijay.nl@inpe.br

***Abstract*—Context: Secondary studies, as Systematic Literature Reviews (SLRs) and Systematic Mappings (SMs), have been providing methodological and structured processes to identify and select research evidence in Computer Science, especially in Software Engineering (SE). One of the main activities of a secondary study process is to read the abstracts to decide on including or excluding studies. This activity is considered costly and time-consuming. In order to speed up the selection activity, some alternatives such as, structured abstracts and graphical abstracts (e.g. Concept Maps – CMs), have been proposed. Objective: This study presents an approach to automatically build CMs using Natural Language Processing (NLP) to support the selection activity of secondary studies. Method: First, we proposed an approach composed by two pipelines: (1) perform the triple extraction of concept-relation-concept based on NLP; and (2) attach the extracted triples in a structure used as a template to scientific studies. Second, we evaluated both pipelines conducting experiments. Results: The preliminary evaluation revealed that CMs extracted are coherent when compared with their source text. Conclusions: NLP can assist the automatic construction of CMs. In addition, the experiment results show that the approach can be useful to support researchers in the selection of studies in the selection activity of secondary studies.**

***Index Terms*—Concept Maps, Graphical Abstract, Secondary Studies, Natural Language Processing, Open IE**

I. Introduction

The use of Evidence-Based Software Engineering (EBSE) employs appropriate research methods to build a body of knowledge about the Software Engineering (SE) practice. Within this context, secondary studies, as Systematic Literature Reviews (SLRs) and Systematic Mappings (SMs), have been providing methodological and structured processes to identify and select research evidence. One of the main steps in applying secondary studies is to read the abstracts to decide whether to include or exclude these studies. However, in Computer Science, especially in SE, there is no culture that determines the creation of structured abstracts of the studies and this increases the cost in identifying the relevance of the studies. Literature has provided evidence that unstructured and poorly written abstracts may compromise the selection activity [1]. One potential solution to minimize such problem is to promote the use of structured and graphical abstracts [1].

Concept Maps (CMs) are graphical representations of knowledge in a particular topic [2]. They provide support to organize and represent knowledge as graphs. CM has been investigated to support selecting studies during the conduction of secondary studies. In Felizardo et al. [3], was conducted a controlled experiment in order to compare performance, effectiveness (in terms of correctness of inclusion/exclusion of studies), and level of tiredness/boredom of graduate students in selecting candidate studies manually and using graphical abstracts. As one of the main results is that students widely pointed out that graphical abstracts in CMs make selection activity less boring and it is quite relevant for researchers that intend to conduct secondary studies. In Santos et al. [4], we conducted a mapping study in order to identify CMs initiatives in Computer Science. According to them, the main problems that may arise from the use of CM are cognitive overload and difficulty in finding the correct concepts and relations. In this sense, technological advances have boosted the development of new technological approaches that help the automatic construction of a CM [5].

Although, the automatic construction of CMs from texts is still an ongoing research, especially when the CM should represent a summarization of a complex text, important results can be found in the literature. In this context, Natural Language Processing (NLP) is pointed out as an important instrument to construct CMs [5].

978-1-6654-2705-0/21/$31.00 ©2021 IEEE
DOI 10.1109/SEAA53835.2021.00009

This paper presents emerging results from an approach to automatically build CMs based on NLP techniques. The approach has the objective of summarizing the CMs from abstracts of scientific articles, in order to support the selection activity of secondary studies. Our approach composed of two pipelines able to perform the triple extraction of concept-relation-concept based on NLP and attach the extracted concepts and relationships, considering a classifier, in a structure used as a template. The guidelines to construct the CMs were based on the structure proposed by [3] that describes a basic structure for representing scientific papers. In order to evaluate the approach, we grouped a set of 497 structured abstracts from Computer Science and Software Testing areas and applied evaluation techniques to measure the efficiency of the classifier. In addition, we conducted a controlled experiment with post-graduate students. The experiment evaluates the quality of the CM generated automatically.

The contribution of this work is an initiative of automatic construction of CMs based on NLP and machine learning in order to help the study selection activity in a secondary study. We present in this article new and insightful ideas, and promising results from a research project in progress that involves CMs, NLP, and secondary studies. We believe that the results achieved so far in this project could contribute to the academic community that has been making efforts to improve and automate the processes of conducting secondary studies.

This paper is organized as follows. Section II describes the Background and a some important concepts related with the approach. Section III shows how our approach is structured and the techniques are used. Section IV discusses how we validated our approach. Section V presents the discussion about the results. Finally, Section VI presents conclusions and future directions for this research.

II. Background

Secondary studies, including Systematic Literature Reviews (SLR) also known as Systematic Reviews (SR) and Systematic Mappings (SM) aim to identify and summarize research evidence on several research topics [6]. According to Kitchenham [6], the process is divided into 3 main steps: In the planning stage, the motivation and protocol for conducting an SR are identified. In the second stage of the review, in the review execution, the objective is to find primary studies capable of answering the research questions. From the studies collected in the databases, the researcher makes the selection of studies based on the inclusion and exclusion criteria. In order to optimize this process, some steps are taken for the data selection activity. First, an initial selection is made in which the inclusion and exclusion criteria are applied only to the title, abstract, and keywords. If the researcher identifies that the study does not fit within the search criteria, it is excluded. Next, another stage of selection is conducted. In this stage, the studies included in the previous activity are read again and the inclusion and exclusion criteria are applied. However, in this stage, the researcher uses the full text. At the end of the selection process, a sample of the studies is reviewed to ensure that the inclusion and exclusion criteria have been applied correctly. The last step is related to writing and disseminating the results to potential stakeholders.

Even with benefits, the execution of a secondary study can be tedious and time-consuming, especially the study selection activity. The following are the characteristics that motivated the choice of CMs as a visual representation for the creation of graphic abstracts in scientific articles.

CMs have emerged from the need to find the best way to represent the conceptual understanding of children about science [2]. CMs are graphical tools for the organization and representation of knowledge. They include concepts, usually within circles or frames, and relationships between them, which are indicated by lines. In these lines, there are words or sentences of connection, which specify the relationships between two concepts. CM can follow a hierarchical model in which the most inclusive concepts are at the top of the hierarchy, that is, at the top of the map. On the other hand, the more specific, less comprehensive concepts are at the base of the hierarchy [2].

Since CMs can help in the understanding of a study, the creation of graphical abstracts to support the selection activity in secondary studies using CMs have been studied [3], [4]. Graphical abstracts should enable reviewers to manipulate images to better understand the knowledge which is represented since analyzing data in graphical format requires less cognitive effort from the reviewer to extract information [7]. In this context, CMs can be useful tools to summarize a complex structure of textual information, contributing to identify the most relevant information in an paper. However, the construction of a CM requires time and effort in identifying and structuring knowledge in unstructured text. In order to mitigate this problem, NLP techniques have been employed and have contributed to automating the extraction of concepts and relationships from texts.

NLP is a subset of Artificial Intelligence (AI) which aims to gather knowledge on how humans understand and use language to design computer algorithms and tools that can process written and spoken language to make systems understand and manipulate natural languages to perform desired tasks [8]. The goal of NLP is to get computers to perform useful tasks involving human language, tasks like enabling human-machine communication, improving human-human communication, or simply doing useful processing of text or speech [9].

NLP aims to extract a complete representation of an idea in unstructured databases, that is, language expressed in text format. From NLP, it is possible to capture the semantics of a sequence of words (sentences, paragraphs, pages) [10]. NLP makes use of tasks such as Part-of-Speech and grammatical structure. It has to deal with the resolution of anaphoras (alternative ways to referring to entities, for example, using pronouns) and ambiguities. NLP makes use of various knowledge representations, such as: a lexicon of words and their meanings, grammatical properties, a set of grammar rules, and several other resources.

In Santos et al. (2019) [11], a SM was conducted to address

the approaches proposed to build CM using NLP. 23 relevant studies were found and despite the construction of CMs using NLP was considered a recent field, it has been proven to be effective in assisting the automatic construction of CMs. In addition, none of the approaches found were built specifically to support the creation of CMs of scientific studies to support secondary studies process.

III. An approach to build CMs based on NLP

As highlighted before, the approach presented in this work is based on the structure proposed by Felizardo et al. [3] for representing an paper by means of graphical abstracts. This structure is presented in Figure 1.

The template represented in Figure 1 is a graphical hierarchical view, containing elements of a scientific study as concepts and relationships between them. A graphical hierarchy was used because this representation suggests an optimal sequence to organize scientific material [2]. The template is organized by levels. Level 0 contains the main concept (positioned in the center of the map), around which other concepts related to the main concept are drawn and grouped in levels. The concept at the top of the map is the most inclusive. In the lower level, more details are described, as illustrated in Figure 1. The 1^{st} template level is composed of five fixed concepts representing each heading of structured abstracts: Context, Objective, Method, Result, and Conclusion. In the 2^{nd} level, each concept previously defined (i.e., the headings) can be refined into new concepts: the optional concepts. For example, the concept "*context*" (see Figure 1 – Level 1) "*is composed of*" three other concepts (see Figure 1 – Level 2): "*research area*", "*gap*", and "*motivation*". Finally, the 3^{rd} level contains variable concepts that will be attached by the user of the template.

Based on the presented template, we propose an approach to extract concepts and attach them at the 3^{rd} level of the template. We divided our approach in four steps: (1) Feeding the approach with structured abstracts; (2) Pipeline 1 – Extraction of concept and relationships (triples extraction); (3) Pipeline 2 – Classification; and (4) Summarization of CMs based on the proposed template and results presentation. These steps are represented in Figure 2 and detailed follows.

1) Input – Our approach uses two different inputs (see Figure 2): (1) a single abstract (structured or unstructured); and (2) a collection of scientific abstracts grouped by area (e.g. Computer Science, Engineering, etc.).
2) Pipeline 1 – Extraction of concept and relationships – inside this pipeline, the abstract sentences are classified according to their header (context, objective, method, result, or conclusion) using the MAZEA software [12]. Next, the classified sentences (structured abstract) are used as input to an Open IE system [13]. This system extract triples that will be used in the final CM.
3) Pipeline 2 – Classification: The purpose of this pipeline is to build a classifier capable of distinguishing whether a sentence belongs to a given class, e.g. one of the classes defined by Felizardo et al. [3] – Level 2 (see Figure 1). The preprocessing was accomplished through: (1) Part-of-Speech Tagging algorithm [14] extracts morphological function of each word, defining its grammatical class, e.g. nouns, verbs, adjective, etc; (2) Removal of stopwords, such as, pronouns, conjunctions, and interjections; (3) Removal of unwanted symbols by means of a list of symbols that we do not want in the final result; (4) Application of the reduction of words to their roots in order to reduce variations of the same term (stemming [15]); and (5) Replace numbers found by a single token (numbers parsing).
   Next, we used a Term-Frequency Inverse Document Frequency (TF-IDF) algorithm [16] to process our sentences, further used as input to train a classifier. The classifier training (Pipeline 2) is performed once, as result, it generates a decision tree, which is used to support the summarization CM creation.
4) Summarization – CM Creation: The final step is to connect the extracted triples (step 1) into a node of the template. Using the classifier trained in Pipeline 2, it is possible to select which node we can attach the triple extracted in Pipeline 1. After performing all connections, the CM is ready to be shown for user evaluation.

A. Building a CM automatically

In order to demonstrate the proposed approach, the Santos et al. [4] study was chosen. The study presents an SM on CMs in the Computer Science. The data used as input to the approach is part of structured abstract. In this abstract, the elements **context, objectives, methods, results and conclusions**, were previously identified by the author.

Is important to emphasize that for this demonstration only the "Results" element of the structured abstract was considered. However, the approach can be extended to the other elements of the structured abstract, taking into account the particularities of each element. As shown in Figure 1, the "Results" element can be subdivided in two optional concepts: Quantitative and Qualitative. The optional concepts are part of level 2 of the graphic representation structure defined by the authors. Subsequently, level 3 contains concepts called "variable concepts", which can be entered (or not) by the user while instantiating the model. The toolkit created to create the concept maps is available [1].

The process of automatic construction of a CM will be presented in 4 steps, they are: (1) Input, (2) Pipeline 1 – Triple Extraction, (3) Pipeline 2 – Classification, and (4) Summarization.

(2) Pipeline 1 – Triplet Extraction

When the structured abstract results are submitted to Pipeline 1, the sentences are separated into smaller sentences.

1) *from the mapping study, we identified 108 studies*

[1]CMtoolkit: https://github.com/csm-applications/CSM-CMtoolkit

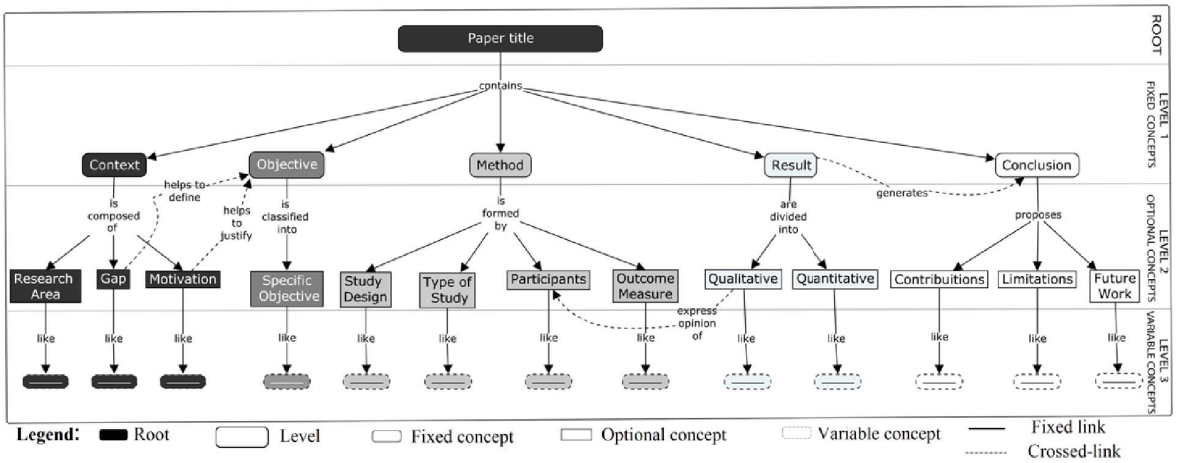


Figure 1: Structure proposed by Felizardo et al. [3]

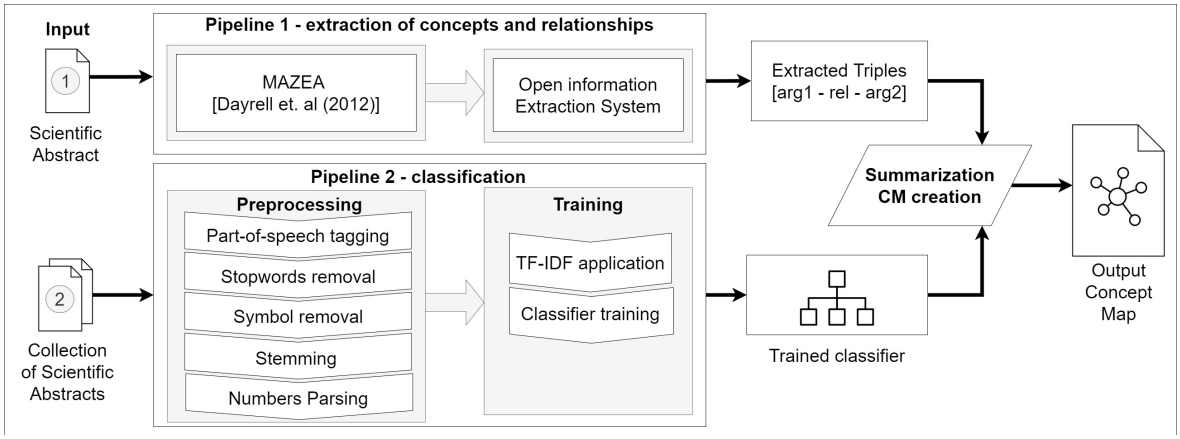


Figure 2: Approach of extracting and summarizing CMs

2) *addressing CMs initiatives in different subareas of Computer Science that were reviewed to extract relevant information to answer a set of research questions*
3) *The mapping shows an increasing interest in the topic in recent years*
4) *and it has been extensively investigated due to support in teaching and learning.*

Next, using the Part-of-Speech Tagging algorithm, each word is classified according to its function in the sentence. As a result, the Information Extractor returns possible (C) concepts, which are connected by possible (R) relationships:

- (C) *we*– (R) *identify*– (C) 108 *study*
- (C) *mapping*– (R) *show*– (C)*increase interest in recent year*
- (C) *it*– (R) *have*– (C) *have extensively investigate*
- (C) CMs *initiative* – (R) *be in*– (C) *different subarea of computer science*

(3) Pipeline 2 – Classification

In this step, we used the trained classifier to predict whether a sentence belongs to the element "Quantitative" or "Qualitative". For instance, the sentence "We identify 108 studies", has been classified as a quantitative element. On the other hand, the sentence "The mapping shows an increasing interest in the topic in recent years" belongs to the work qualitative results. Next, we detail how the output of the Pipelines 1 and 2 is summarized.

(4) Summarization

After the implementation of Pipelines 1 and 2, the CM summarization is done by attaching the extracted triples (Pipeline 1) to the model for scientific studies proposed by Felizardo et al. [3] using the predictions obtained from the classifier (Pipeline 2). The result is shown in Figure 3.

IV. Approach Validation

In order to evaluate our approach, two strategies were used: (i) evaluation of the classifier effectiveness; and (ii) CMs evaluation generated by conducting a controlled experiment considering the opinion of Master degree students comparing automatically generated CMs with manually generated CMs in terms of similarity with respect to their representativeness

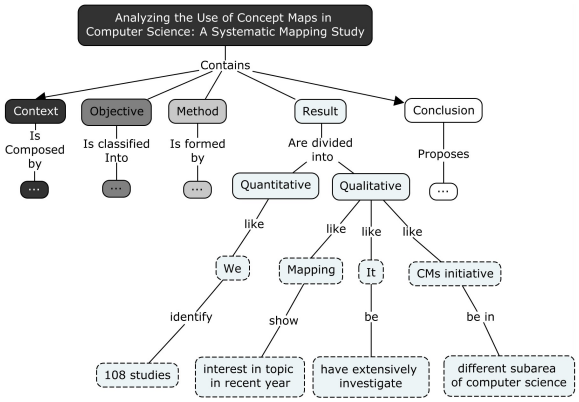


Figure 3: Example created by the proposed approach

and association to their respective textual version content in the abstracts. A replication package is available[2].

A. Classifier Evaluation

1) Dataset: In order to analyze the classifier effectiveness, 497 sentences out of 215 studies from Software Testing were selected. We used different sources to ensure a reasonable number of examples to train the classifier. The dataset contains studies published in important SE conferences and journals, such as: Conference of Evaluation and Assessment in Software Engineering (EASE) and Information and Software Technology (IST) and Journal of Systems and Software (JSS). In addition, we used the term "Software Test" in Scopus[3] as search string and collected papers which not necessarily use structured abstracts aiming to improve classification accuracy. Next, we manually annotated the selected sentences in "quantitative" or "qualitative". The complete list of abstracts used, the annotated corpus is available[4].

2) Classifiers: In order to generate a classifier we used TF-IDF technique to extract the frequency of each term in dataset and put in Weka database format (ARFF file). Next, we used the pre-processed dataset to train a supervised machine learning classification algorithm. In the experiment, we compare three classifiers recommended in literature to document classification task: (1) J48 [17]; (2) Naive Bayes [18]; (3) Random Forest [19].

3) Setup: The dataset was evaluated using a 10-fold Cross-Validation technique. We considered the Cross-Validation and Learning outcomes for fine-tuning the classification model and identifying features which improved the classification.

4) Results: Considering the results presented in Figure 4 – Table I & II it is possible to see the performance measured by precision and recall of the classifiers used. Naive Bayes classifier showed the lowest performance in precision and recall. The classifier with best results was the J48 achieving the precision of 0.838 for qualitative sentences and 0.871 for quantitative sentences.

[2]Labkit: https://github.com/CSM-Research/04-EXP-CMAutoGeneration
[3]https://www.scopus.com/
[4]Corpus: https://github.com/csm-applications/CSM-CMtoolkit

Qualitative sentences			
	Precision	Recall	F-measure
J48	0.838	0.65	0.732
Naive Bayes	0,525	0.594	0.577
Random Forest	0.875	0.441	0.586

Table I: Qualitative prediction evaluation

Quantitative sentences			
	Precision	Recall	F-measure
J48	0,871	0,949	0,908
Naive Bayes	0,827	0,783	0,805
Random Forest	0,812	0,975	0,886

Table II: Quantitative prediction evaluation

B. Validating the CMs built Automatically

With an objective to validate the CMs quality created by the proposed approach we conducted a survey to understand if the automatically built CMs were similar to CMs built by humans.

1) Hypothesis and Research Questions (RQ): CMs were evaluated considering three aspects: (1) representativeness of the concepts; (2) validity of the ideas expressed by the relationships; and (3) the coverage level. Three hypothesis were defined:

Our main RQ is: **Are the automatically generated CMs similar to the CMs generated by the study authors?**. Next, we divided the main research question in 3 specific questions:

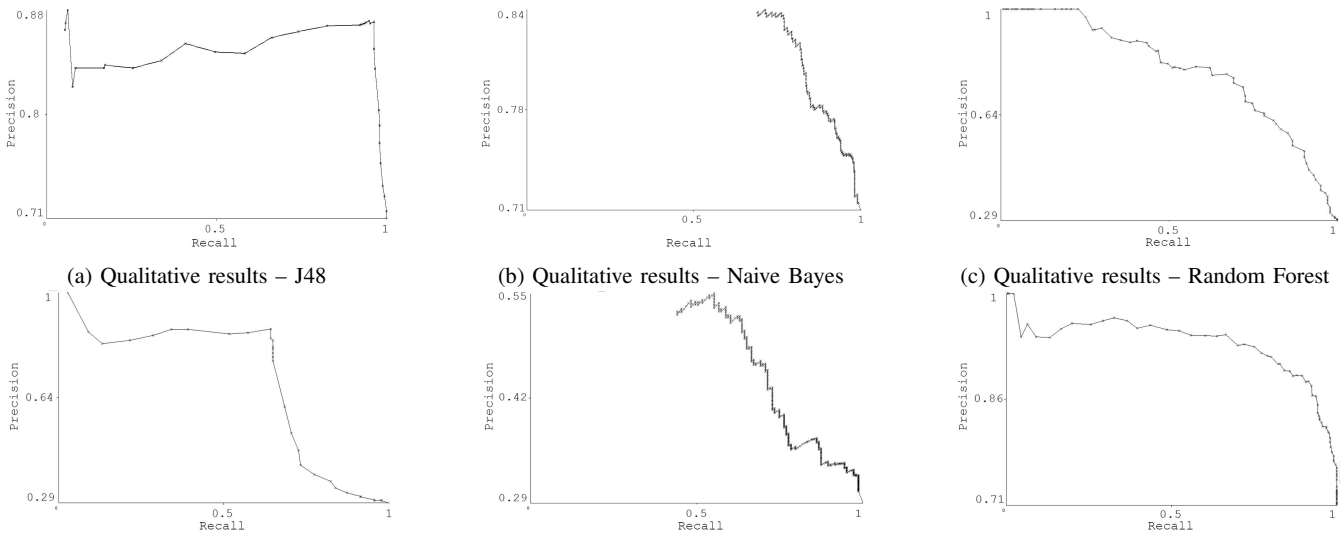
- **Research Question 1 (RQ1)**: Are the concepts presented in the automatically generated CM similar to the CMs generated manually?
- **Research Question 2 (RQ2)**: Are the relationships (links) presented in the automatically generated CM similar to the CMs generated manually?
- **Research Question 3 (RQ3)**: The automatically generated CM cover all information related to the element "Results" similarly to CM generated manually?

Associated to RQ1, RQ2 and RQ3, we defined two hypothesis:

- Null hypothesis (H0): The concepts, relationship and coverage presented in the automatically generated CM are not similar to the CMs generated manually.
- Alternative hypothesis (H1): The concepts, relationship and coverage presented in the automatically generated CM are similar to the CMs generated manually.

2) Method of the experiment: In order to build the control group, authors of six articles were asked to build a CM that represents the results obtained in their work. The target group was built using the same six studies, but using our proposed approach to generate the CMs automatically.

32 Master degree students in Computer Science were selected to answer the RQ. The automatically and author generated CMs were presented to students. Next, the students answered a questionnaire to check if they agreed with the CM concepts representativeness, validity of ideas expressed by the



(a) Qualitative results – J48 (b) Qualitative results – Naive Bayes (c) Qualitative results – Random Forest

(d) Quantitative results – J48 (e) Quantitative results – Naive Bayes (f) Quantitative results – Random Forest

Figure 4: Precision Recall curves using multiple classifiers

relationships and the coverage level. The experimental study was divided in two phases: training and execution.

(1) Training: Aiming to answer all questions of participants about the experiment. The students received a document prepared by the authors containing the CM generated from the article [4]. Next, the participants answered the three RQ.

(2) Execution: In this phase the participants received a document containing six CMs and their respective abstracts to be evaluated. The six studies were submitted to the initiative proposed in Section III. Six CMs corresponding to the data presented in the abstract of each study were automatically generated.

(3) Groups division: The authors of the six studies used in the experiment were asked to build a CM that represent the results obtained in their work, considering only what was written in the "Results" element of the paper abstract. Preparing the documents for the experiment execution and distributing to the participants, two versions were used: (1) Target group: CMs automatically generated; and (2) Control Group: CMs developed by the authors of the studies. This division was not communicated to participants to reduce the risk of interference with results.

3) Experiment results: The results presented in this section are still preliminary and are not conclusive, however they do present evidences which brings light on the RQ. The 32 students were of age ranging from 24 to 62 years. The assessment was done using the Likert scale [20]. For each question the participant selected one of the options (strongly disagree (1), disagree (2), undecided (3) , agree (4), strongly agree (5)).

According to Tables III and IV, it is possible to see that the median and mode for the automatically and author generated MCs is 4.

Table III: Automatic Generated CMs

	1 (%)	2 (%)	3 (%)	4 (%)	5 (%)	Median	Mode
RQ1	4%	19%	9%	50%	18%	4	4
RQ2	5%	21%	15%	49%	10%	4	4
RQ3	4%	18%	7%	49%	22%	4	4

Table IV: Author Generated CMs

	1 (%)	2 (%)	3 (%)	4 (%)	5 (%)	Median	Mode
RQ1	4%	19%	9%	50%	18%	4	4
RQ2	3%	19%	18%	50%	10%	4	4
RQ3	5%	19%	24%	33%	19%	4	4

C. RQ1 – Evaluation of Concepts Used

Applying the questionnaire, 68% of the participants who classified the automatically generated and author generated CMs, strongly agree or agree that CM concepts represent ideas and information related to the corresponding abstract. A t-test comparison between both groups presented no significant differences, with p-value=0.1503, df=176.6 and t=1.4447. These results support H1 associated to RQ1, in which the concepts of the automatically generated CMs represent ideas and information of the "Results" element of the abstract. Moreover, the automatically generated CMs presented an average performance with no significant difference from the author generated CMs.

D. RQ2 – Evaluation of Relationships Used

When evaluating the connections used in the CMs, it was possible to verify that on average 59% of the participants agree or strongly agree that the automatically generated CMs have relationships (links) that adequately connect the "Results" concepts of the corresponding abstract. With respect to CMs generated by the authors, on average 60% of participants strongly agreed or agreed that automatically generated CMs

have relationships that adequately connect the concepts of the "Results" element of the corresponding abstract. A t-test comparison between both: the automatically generated CMs (Target group) and author generated CMs (Control group), as a result they did not present statistical differences between the samples either, with p-value=1, df=175.18 and t=0. These results support H1 associated to RQ2, which states that the relationships properly connect the concepts related to the "Results" element of the abstract. Additionally, the automatically generated CMs also performed similarly to the author generated CMs, in this criterion.

E. RQ3 – Evaluation of CM Coverage

When questioned about CMs coverage, on average 71% of experiment participants agreed or strongly agreed that automatically generated CMs cover all information related to the "Results" element of the corresponding abstract. Regarding the scope of CMs generated by the authors, on average 52% of the participants strongly agreed or agreed that the CMs generated cover all the information related to the "Results" element of the corresponding abstract. Both values distributions showed no significant differences in t-test, with p-value=0.105, df=177.6 and t=1.6292. These results also support H1 associated to RQ3, in which the CMs cover all information related to the "Results" element of abstracts. Moreover, the comparison also supports the hypothesis that the automatically generated CMs, performed similarly to author generated ones, in this study.

V. Discussions

The main objective of this comparison was to evaluate if our approach is able to generate CMs similar to those generated by the authors, allowing the approach to be a tool to help in the construction of CMs.

In [21], a text summarization system was proposed to enhance productivity and reduce errors in the traditional data extraction process in secondary studies. We found some differences that are important: (i) the golden standard proposed and the validation was done in the medical area; (ii) the approach uses a single document to train a machine learning regression model; (iii) the approach proposes to aid the data extraction phase and not the selection phase; and (iv) the approach do not use CMs as visual representation to aid the data extraction phase. Another important approach was proposed by [22]. The goal of this study is to aid the building of CMs from text documents. We identified some similarities to our approach, such as the use of techniques as Part-of-Speech tagging. We also identified the use of a variations of some metrics. For instance, VF-ICF is a metric modified by Punuru and Chen [23], and shows that verbs occurring with only a few sets of candidate terms are more significant, while verbs occurring with too many candidate terms tend to be overly general and do not denote important semantic relations. This modification does not need a set of different documents, instead this approach considers the full study and not only the abstract. Affinity Propagation was also used to cluster related terms and also Anaphora resolution to extract complex relationships that uses pronouns to represent the nouns. Aguiar et al. [5] proposed a new method for automatic generation of CMs. An important difference found is the use of a semiautomatic approach that the user chooses the domain in the input. Also, [5] mention that the Anaphora resolution is still far from satisfactory and one of the limitations mentioned is that some relevant domain concepts were lost in the extraction. Atapattu et al. [24] present a method to transform teaching resources into integrated network models such as CM. However, the manual construction of CM from teaching materials places an additional workload on the academics involved. The researchers developed a set of NLP algorithms to support concept-relation-concept triple extraction to form CM. Structural and graph-based features are utilized to rank the triples according to their importance.

Even though the initiatives mentioned are important contributions, this work presents important characteristics that differ from the others. The most important difference of our approach is the use of the template proposed by [3]. This model helps the researchers in finding important information to include or exclude a study when performing a secondary study. Based on this template, other NLP and machine learning techniques were combined to generate the full CM.

The main contribution of this study is to present an approach to build CMs automatically using NLP techniques. Aiming to investigate the similarity between CMs generated automatically and manually by the authors, an experiment was conducted. In this experiment the element "Results" from abstracts was used. However, the techniques used in both pipelines of our approach are flexible and can be applied in other abstract elements. The classifiers can be trained using specific corpus for each abstract element. Also, according to [25], Open IE techniques do not consider the text domain to perform the extraction of triple. Consequently, the Open IE techniques used in our approach are applicable to any unstructured text.

The experimental study results indicate that the generated CMs presented valid concepts and relationships, as well as a good comprehensiveness. The preliminary results indicate that our approach can be an option to generate a representative CM. Also, the experiment results indicate that: "The concepts presented in the automatically generated CM are similar to concepts of the CMs generated manually", "The links (relationships) presented in the automatically generated CM are similar to the links of the CMs generated manually" and "The automatically generated CM covers all information related to the 'Results' element of its corresponding abstract". We believe that using the results obtained and presented so far, authors of scientific articles can automatically create a suggestion on how to structure the CMs. This can help in the creation of CMs to represent scientific studies. The automation in the creation of CMs is a way of popularizing the use of graphic summaries. In the future CMs can become a tool capable of accelerating the process of selecting primary studies in an SLR or SM.

A. Threats to validity

The main threats to the validity of this study are listed as follows: (i) only one element of structured abstract was analyzed. Despite the possibility of extending our approach for other element of structured abstract this task may not be trivial. Therefore, the proposed approach is still considered preliminary and it is not possible to generalize the results. However, we believe that our approach provides a feasible solution that can be improved in future works; (ii) the students selected to participate experiment were in graduation program at UTFPR while this research was conducted. There was no check of the participants' ability to understand the papers. To mitigate this threat we performed a detailed training session aiming to answer any questions of participants; (iv) the experiment was conducted with a small number of participants (32); (v) the low level of experience of the participants with the theme of the studies used; and (vi) English was the language used, however, participants were not native speakers.

VI. Conclusions and future work

In this study we proposed an initiative of automatic construction of CMs based on NLP and machine learning in order to contribute with the studies selection activity in a secondary study (SLR or SM). This study presented an approach to automatically build CMs employing NLP in order to support the selection activity of secondary studies. We propose an approach to extract concepts and attach them. The approach is divided 4 steps: (1) Feeding the system with structured abstracts; (2) Pipeline 1 – Extraction of concept and relationships; (3) Pipeline 2 – using a classifier; and (4) Summarization of CMs based on the proposed guidelines and results presentation. Using this technique it becomes possible to select where a specific triplet will be attached in the CM structure proposed in [3].

We started this study by investigating element "Results" of the abstract. We evaluated the classifier and the experiment reported an accuracy rate up to 86.74% in classifying the results of the studies in "Quantitative" or "Qualitative". The preliminary results of the experiment conducted shows that the approach is useful to help researchers to build graphic representations of studies. As future work, we intend to extend the classification range to cover all the concepts proposed in the model, such as context, objective, methods and conclusion.

Acknowledgment

This study was supported by São Paulo Research Foundation (FAPESP) grant: 2019/23663-1 and by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES): PROEX-11308091/D.

References

[1] B. A. Kitchenham, P. Brereton, S. Owen, J. Butcher, and C. Jefferies, "Length and readability of structured software engineering abstracts," *IET Software*, vol. 2, no. 1, p. 37, 2008.

[2] J. Novak and A. Cañas, "The theory underlying concept maps and how to construct and use them," Florida Institute for Human and Machine Cognition, Tech. Rep. Report IHMC CmapTools 2008-1, 2008.

[3] K. Felizardo, E. Souza, S. Hesae, N. Vijaykumar, and E. Nakagawa, "Analysing the use of graphical abstracts to support study selection in secondary studies," in *Conferência Iberoamericana em Software Engineering*, Buenos Aires, Argentina, 2017, pp. 1–10.

[4] S. V., E. Souza, K. Felizardo, and L. V. Nandamudi, "Analyzing the use of concept maps in computer science: A systematic mapping study," *Informatics in Education*, vol. 16, no. 2, pp. 257–288, 2017.

[5] C. Z. Aguiar and D. Cury, "Automatic construction of concept maps from texts," in 4^{th} *International Conference on Concept Mapping*, Tallinn, Estonia, 2016, pp. 1–6.

[6] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," Keele university, Durham, Tech. Rep., 2007.

[7] M. C. F. Oliveira and H. Levkowitz, "From visual data exploration to visual data mining: A survey," *IEEE Transactions on Visualization and Computer Graphics*, vol. 9, no. 3, pp. 378–394, 2003.

[8] G. Chowdhury, "Natural language processing," *Annual Review of Information Science and Technology*, vol. 37, no. 1, pp. 51–89, Jan. 2005.

[9] D. Jurafsky and J. H. Martin, *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*, Upper Saddle River, N.J., 2009.

[10] E. Cambria and B. White, "Jumping nlp curves: A review of natural language processing research," *IEEE Computational Intelligence Magazine*, vol. 9, pp. 48–57, 2014.

[11] V. Santos, E. Souza, K. Felizardo, N. Watanabe, W.M.and Vijaykumar, S. Aluísio, and A. Cândido Júnior, "Conceptual map creation from natural language processing: a systematic mapping study," *Revista Brasileira de Informática na Educação*, vol. 27, no. 03, pp. 150–176, 2019.

[12] C. Dayrell, A. Candido Junior, G. Lima, D. Machado Junior, A. A. Copestake, V. D. Feltrim, S. E. Tagnin, and S. M. Aluísio, "Rhetorical move detection in english abstracts: Multi-label sentence classifiers and their annotated corpora," in *LREC*, Istanbul, Turkey, 2012, pp. 1604–1609.

[13] G. Angeli, M. J. J. Premkumar, and C. D. Manning, "Leveraging linguistic structure for open domain information extraction," in 7^{th} *International Conference on Natural Language Processing*, Beijing, China, 2015, pp. 344–354.

[14] K. Toutanova, D. Klein, C. Manning, and Y. Singer, "Feature-rich part-of-speech tagging with a cyclic dependency network," in *North American Chapter of the Association for Computational Linguistics on Human Language Technology*, Edmonton, Canada, 2003, p. 173–180.

[15] M. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130–137, 1980.

[16] K. Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation*, 1972.

[17] R. Patil and V. M. Barkade, "Class-specific features using j48 classifier for text classification," in *International Conference on Computing Communication Control and Automation*, Pimpri Chinchwad, India, 2018, pp. 1–5.

[18] Y. H. Li and A. K. Jain, "Classification of Text Documents," *The Computer Journal*, vol. 41, no. 8, pp. 537–546, 1998.

[19] M. Klassen and N. Paturi, "Web document classification by keywords using random forests," in *Networked Digital Technologies*, Berlin, 2010, pp. 256–261.

[20] R. Likert, "A technique for the measurement of attitudes," *Archives of psychology*, vol. 1, 1932.

[21] D. D. A. Bui, G. Del Fiol, J. F. Hurdle, and S. Jonnalagadda, "Extractive text summarization system to aid data extraction from full text in systematic review development," *Journal of Biomedical Informatics*, vol. 64, pp. 265–272, 2016.

[22] I. Qasim, J. Jeong, J. Heu, and D. Lee, "Concept map construction from text documents using affinity propagation," *Journal of Information Science*, vol. 39, pp. 719–736, 2013.

[23] J. Punuru and J. Chen, "Learning non-taxonomical semantic relations from domain texts," *Journal of Intelligent Information Systems*, vol. 38, pp. 191–217, 2012.

[24] A. Atapattu, K. Falkner, and N. Falkner, "A comprehensive text analysis of lecture slides to generate concept maps," *Computers & Education*, vol. 115, no. Supplement C, pp. 96 – 113, 2017.

[25] A. Yates, M. Cafarella, M. Banko, O. Etzioni, M. Broadhead, and S. Soderland, "Textrunner: Open information extraction on the web," in *Human Language Technologies - Conference of the North American Chapter of the ACL*, Rochester, New York, USA, 2007, pp. 25–26.