



Automatically Locating Results to Support Systematic Reviews in Software Engineering

**José Alberto S. Torres, Daniela S. Cruzes, Laís do
Nascimento Salvador**

ESELAW – Apr/2013

ACKNOWLEDGEMENTS

- Brazilian Federal Highway Police Department (DPRF)
- National Institute of Science and Technology for Software Engineering (INES)



Schedule

- Introduction
- Related Work
- Sentence Classification Methods
- An Approach For Locating Results
- Discussion and Conclusion

Introduction

- For decades people have known the gaps between research evidence and clinical practice, and the consequences in terms of ineffective or even harmful decision making.
- Evidence-Based Software Engineering (EBSE) aims to improve decision-making on the development and maintenance of software through the integration of current best research evidence with practical experience.

Introduction

- In this context, Systematic Review (SR) has provided mechanisms to identify and aggregate research evidence providing a full and fair assessment of the state of evidence related to a particular topic of interest .
- In the last years we have noticed an improvement from the use of Systematic Reviews by researchers in Software Engineering.
- One of the challenges in conducting this Systematic Reviews of Literature is to maintain a balance between methodological rigor and the required effort.

Introduction

- ❑ Some authors have applied the techniques of text mining and machine learning in order to reduce the effort and time required for construction of Systematic Review of Literature, acting, respectively, in steps of selecting papers and automatic identification of contextual information.
- ❑ Thus, we propose in this paper a way to perform the automation of an important step in performing synthesis in Systematic Reviews, the results extraction activity.

Related Work

- ❑ Some authors have published, especially in recent years, papers about tools to automate parts of the Systematic Review as a way to reduce time and cost required for its realization:
- ❑ Felizardo created a tool to support the primary study selection activity using visual text mining techniques
- ❑ Malheiros proposed an automated tool to help researchers in the initial selection of items to be used in the SRL
- ❑ Silva Rocha developed a tool for automatic information extraction from the scientific papers context

Sentence Classification Methods

- Three methods were found
 - One pattern-recognition method
 - Two Machine-learning methods
- All methods applied to Health area papers

Sentence Classification Methods

- The first method implements a pattern recognition process that does previously a lexical and syntactic analysis of the sentences to be evaluated.

Information Type	Pattern
Objective	In this_{article paper study research work} }... DET_{motivation: aim goal objective}...
New Things	Here, we propose a novel (...) approach... This analysis reveals...
Future Works	{Further Future more}_{work investigati on observation}_{<V>}...

Sentence Classification Methods

- The second technique uses a set of 2504 words derived from the set of sentences used in the experiment conducted by the author.
- The standards were defined for the classification by the combination of the presence or absence of elements of this set of words.
- Each word of the training set was defined as an attribute of classification, thus, each sentence was defined with 2504 attributes.

Sentence Classification Methods

- The last method aimed to promote improvement in a method to automatically create summaries of papers.
- The study proposes the creation of an attribute set that characterizes various aspects of the sentences and uses Naive Bayes classification algorithm to automatically define the correct category to each element in analysis.

Sentence Classification Methods

Category	Description
BACKGROUND (<i>BASIC</i>)	Sentences describing some (generally accepted) background knowledge
OTHER (<i>BASIC</i>)	Sentences describing aspects of some specific other research in a neutral way
OWN (<i>BASIC</i>)	Sentences describing any aspect of the own work presented in this paper – except what is covered by AIM or TEXTUAL
AIM	Sentences best portraying the particular (main) research goal of the paper
TEXTUAL	Explicit statements about textual section structure of the paper
CONTRAST	Sentences contrasting own work to other work; or pointing out weakness in other research
BASIS	Statements that the own work uses or gets support from other work as its basis or starting point.

An Approach For Locating Results

□ The main objective is to define a method to Automate the task of locating result sentences in unstructured Software Engineering papers. Papers that do not follow the model IMRaD

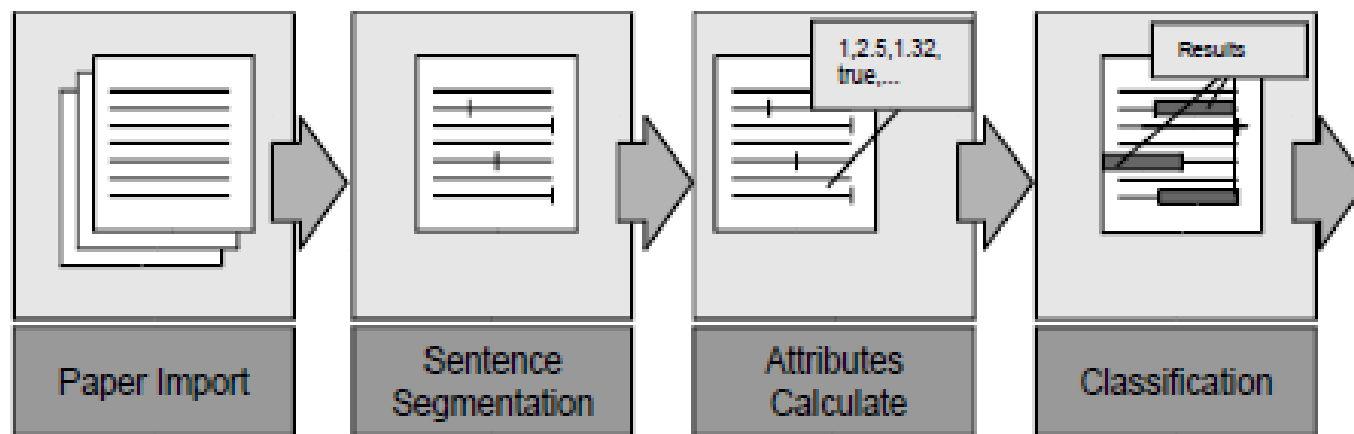


Fig. 1. - Textum Method Schema

Paper Import and Sentence Segmentation

- ❑ The first step is to convert the papers into plain text format documents. Tags and undesirable characters are removed in this stage.
- ❑ The next step is to segment sentences, pieces of text with full meaning
- ❑ Pattern recognition algorithm

Attributes Calculation

- ❑ Text classification strategies use a number of predefined attributes to automatically infer the category and classify the pieces of text.
- ❑ We Selected many attributes from several works and defined a initial list.
- ❑ To select the best attributes from the whole set, we used a genetic algorithm based technique. Classification tests were run from sets of attributes mounted from the initial group.
- ❑ The sets of attributes with the best classification results were selected and mixed to form new subgroups, used as the basis for new classification process, and so on, until define the best performing attributes to compose the final set.

Attributes Calculation

Keyword Frequency

Cue Method

Paragraph Sector

TF-IDF

TF-ISF

Sentence length

Gist Sentence

Lexical Connectivity

Section Position

Segment Position

Verbal Tense

Citations

Header

Self-indicative phrases

Comparison

Number Presence

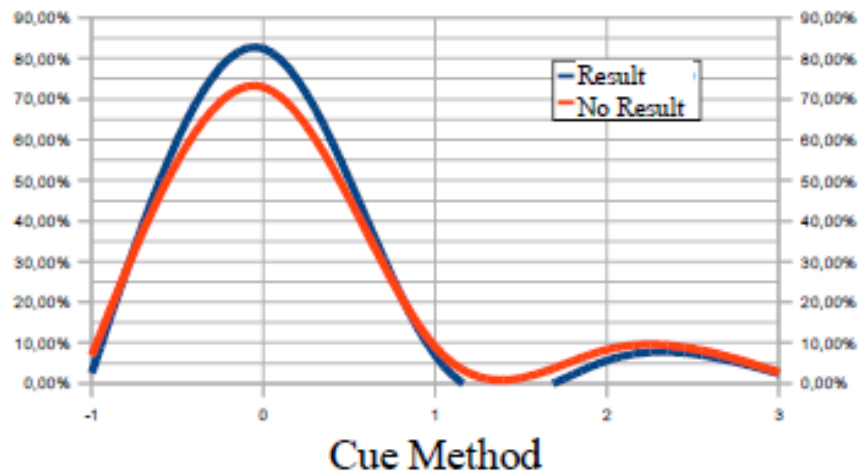
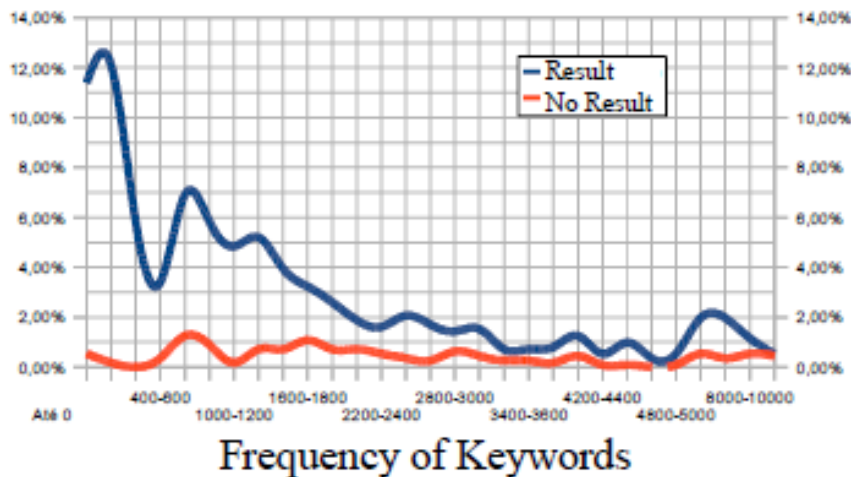
Percentage Indicator

Classification

- ❑ The classification task is performed in two steps.
- ❑ The first step uses a rule-based algorithm and the second a machine learning algorithm

Rule-based Algorithm

- ❑ The process of choosing the classification rules set is based on the analysis of the distribution of each sentence attribute value.



Rule-based Algorithm

Table 1. - Rule Set Example

- If the sentence has no verb, must be set to "no result";
- If the sentence, excluding "stop words", having size less than or equal to 3, must be defined as "no result"
- The sentences that have value less than or equal to 600 for the attribute "frequency of key words" must be defined as "no result";

Machine Learning Algorithm

- ❑ The second classification stage is based on machine learning algorithm.
- ❑ Six algorithms were tested to select the heuristic to be utilized in our method: Naive Bayes (NB); Tree J 4.8 (TJ48), Decision Table (DT), Support Vector Machines (SVM), Nearest Neighbor (IBK) and Multilayer Perceptron algorithm (MP).

Machine Learning Algorithm

Table 2. - Sentence Classification Result

Algorithm		N. Bayes	T. J.48	TRD	SVM	MP	IBK
Result Sentences	Precision	26,2%	55,6%	0,0%	31,6%	32,7%	28,1%
	Recall	17,5%	7,9%	0,0%	9,5%	28,6%	39,7%
No Result Sentences	Precision	93,0%	92,5%	91,9%	31,6%	93,8%	94,5%
	Recall	95,7%	99,4%	100,0%	9,5%	94,8%	91,1%
Average (Results and no results)	Precision	87,6%	89,5%	84,5%	87,6%	88,9%	89,1%
	Recall	89,4%	92,1%	91,9%	91,0%	89,5%	86,9%

Machine Learning Algorithm

- ❑ In our study, we tried to select the “result sentences” in order to facilitate the researcher work. In this case, if the algorithm classifies "no re-sults" sentences incorrectly is better than the algorithm does not select the "result" sentences.
- ❑ Because of this, the recall was prioritized in the selection of machine learning classification algorithm and because this Nearest Neighbor heuristic was defined as the core of the second stage of classification process

Discussion and Conclusion

- ❑ The study showed a precision rate up to 56.5% and recall rate up to 60%, depending on the rule-based classification scheme used.
- ❑ These results showed precision rates close to those described in an experiment conducted with students at the University of Maryland, in which students achieved rates close to 50%.
- ❑ The main difference between manual and automatic tests realized was the time spent to perform these procedures.

Discussion and Conclusion

- ❑ However, in practice, we have noted that a researcher needs to read the whole paragraph to understand the context of a sentence highlighted, because sentences that are in the same paragraph help to explain the results.
- ❑ When these paragraphs, where the sentences highlighted were found, were completely analyzed, it was discovered that there were several other results that have not been automatically highlighted.

Discussion and Conclusion

- ❑ The observation showed that if a tool is used to highlight the paragraphs instead of the sentences the recall rate reaches 72%, an increase of 12% compared with the original test.
- ❑ With this new paradigm, the new method achieved a precision rate of 74% in classifying paragraphs in which there are results.

Discussion and Conclusion

- ❑ Even with the change, the main goal that was to reduce the effort needed to create the review has been maintained.
- ❑ Selecting only the paragraphs, the Textum method reduced the text to be analyzed by researchers to 20%, which, in theory, would probably reduce in 80% the time spent on traditional analysis of the paper.



Thank you!

Contacts

alberto@torres.eti.br

www.torres.eti.br