Semantic Filtering of Textual Requirements Descriptions

Jorge J. García Flores *

LaLICC
Université de Paris 4
96, bd Raspail,
75006 Paris, France
jorge.gflores@paris4.sorbonne.fr

Abstract. This paper explores the use of semantic filtering techniques for the analysis of large textual requirements descriptions. Our approach makes use of the Contextual Exploration Method to extract, within large textual requirements descriptions, those statements considered as relevant from requirements engineering perspective: concepts, relationships, aspecto-temporal organisation, cause and control statements. We investigate to what extent filtering with these criteria can be the base of requirements analysis and validation processing, and what kind of software tools are necessary to support contextual exploration systems for this purpose.

1 Introduction

Many software development projects fail because of a deficient requirements strategy. Requirements engineering (RE) studies the processes, techniques and methods that would allow a system to meet the purpose for which it was intended [1]. Natural language plays an important role in RE. A recent study [2] shows that 73% of the documents available for requirements analysis are written in natural language. Use cases, scenarios, user stories, transcriptions of conversations for requirements elicitation [3] and even rough sketches are examples of textual requirements descriptions (TRD). Natural Language Processing (NLP) provides useful techniques to extract information from this documents, which can reach several hundred of pages in large projects.

The use of linguistic knowledge for natural language requirements processing is not new. In the past, linguistic instruments have been used to extract conceptual schema from NL requirements ([4, 5]), to analyse and validate requirements through a conceptual representation ([6-8]) or to improve the linguistic quality of TRD documents ([9, 10]). However, few of these approaches take into account the size of the input text. Large TRD documents raise interesting questions for

* PhD research funded by CONACyT (Mexican Government). Thanks to Joël Bourgeois, Marie-Josée Goulet, Jean-Luc Minel, Camille Salinesi and Marie Chagnoux for their feedback on earlier versions of this paper.
NL requirements processing techniques. What should be the processing’s scope? Is it useful to process the whole mass of documents, or is it better to limit it to some extent? If there are formal representations involved, how do they deal with such amount of input text? What are the effects of large TRD on the NL processing strategy?

This paper postulates that filtering relevant text fragments according to semantic criteria enhances large TRD processing. Its purpose is to explore the use of a linguistic technique, the Contextual Exploration (CE) Method [11], to extract semantically relevant sentences in order to support requirements analysis and validation. Four semantic viewpoints are considered as relevant for large TRD processing: 1) concepts relationships, 2) aspecto-temporal organisation, 3) control and 4) causality. The paper is organised as follows: section 2 introduces the CE method. Section 3 is devoted to the discussion of how semantic filtering would support requirements analysis and validation. It proposes an architecture for a TRD semantic filtering system. Section 4 presents the conclusions and sketches future work.

2 The Contextual Exploration Method

In the frame of the Cognitive and Applicative Grammar (GAC in French) linguistic theory [12], the CE Method [11] was originally designed to solve lexical and grammatical polysemy problems. The method rests on the principle that domain independent linguistic elements structure text meaning. Its purpose is to access the semantic content of texts in order to extract relevant information according to a certain task.

According to contextual exploration, all signs occurring in a text (the textual context) must be taken into account to determine the semantic value of a sentence. This example illustrates how indeterminacy is solved:

1. **In spite of all precautions, he was captured the day after**
2. **Without all precautions, he was captured the day after**

From the aspecto-temporal point of view, the tense of “was captured” is the linguistic marker of a semantic value that can be new-state (he was captured) or unrealized (he was not captured). However, the tense itself is not enough to decide which one of both values must be assigned, so the context has to be analysed in order to get more clues. In this case “in spite” and sans “without” are the clues that determine the semantic values of these sentences.

Linguistic markers correspond to plausible hypothesis that must be confirmed by the presence of certain clues. The heuristic mechanism is based in rules. A rule R is defined as follows:

\[
R_k = [K, \{I_p, C_p\}, D_k]
\]

1 This example is from G. Guillaume, quoted by Desclès, [11]
Where K is a class of linguistic marker, I a class of clue and, C the research context and D the decision to take [13]. Rules are organised by tasks; each task conveys a particular semantic notion. For instance, the task Static Relations [14] specifies a set of 238 rules, 6149 markers and 1777 clues in order to assign a set of 14 semantic values. This task would assign the **ingredient** semantic value to the following sentences (taken from a TRD of an insurance system):

(3) **A joint policy** includes the joiners age, the joiners gender and his smoker condition.

(4) **The agent displays** billing mode, effective date, due date, bill number and total premium amount, which are part of the policy’s billing detail.

(5) **For each** mandatory rider, the agent should specify the following values:
   - Face amount
   - Due period
   - Increased Periods

Within the “concepts relationships” viewpoint, the **ingredient** semantic value filters "part-of" relations from a text. In the above example linguistic markers are highlighted in bold. In sentence (5) the linguistic marker “For each” is not strong enough to assign the **ingredient** value, but its plausibility may be confirmed by non-lexical clues, like the typographical sign “.” or the presence of a list. Linguistic clues may not only be lexical entries. Punctuation signs and text structural elements, like titles and sections, are allowed in a clue class.

So far, the CE method has been applied to automatic summarisation [13], filtering of causality relations [15], relations between concepts [14], aspecto-temporal relations and to extract information from the web using semantic viewpoints [16]. The CE method presents specific aspects that distinguish it from other semantic tagging techniques used in the context of NL requirements processing:

1. It is oriented to large document sets, so it can handle large TRD while taking account of the linguistic context.
2. It has rule-based structure that could allow to define semantic validations beyond the limits of a sentence, and even between requirements that are far away from each other in the requirements document.
3. CE markers and clues are not restricted to linguistic elements, but also typographical and structural ones. Despite the importance of structural elements (titles, subtitles, etc.) in industrial TRD documents, little research has been done on their exploitation by NL requirements processing tools.
4. CE semantic tags are not tied to any software design method. The difference between CE and other methodology neutral approaches ([8, 7]) is that CE produce an abridged text unit (the filtered text) which could be the input to others NL requirements processing tools.
3 Semantic Filtering of TRD

Semantic filtering aims at improving large TRD processing by drawing together a small number of statements that share a certain semantic viewpoint. However, extraction is made using relevance criteria, and relevance, as it has been pointed out by Minel [13], depend for the most part on the reader’s point of view. The following viewpoints are considered as semantically relevant from a RE point of view: I) Concepts relationships, II) Aspecto-temporal organisation, III) Control and IV) Causality.

Each one of this viewpoints conveys a semantic concept that, according to the GAC linguistic model, may structure and organise meaning [12], and each one represents an important aspect in requirements analysis as well. Much effort have been devoted to build conceptual schema from TRD’s relations between concepts for requirements analysis ([4, 5, 8]). The value of extracting event’s and processes temporal organisation (the "dynamic" aspect) for NL requirements validation has been remarked by Burg [6]. Control issues, i.e. specifying if actions are "environment controlled" or "machine controlled", are of primary importance in RE [17], and a precise understanding of the causal organisation of actions is necessary to specify the rules that a system must obey [18]. The following example shows different views of one TRD paragraph according to these semantic viewpoints. Relevant sentences are marked in bold. Under-braces indicate semantic values assigned by CE rules according to linguistic markers and clues (underlined)\(^2\):

- Relations between concepts viewpoint:
  When the start button \underline{is pressed}, if there is an original in the feed slot, the photocopier makes \underline{N copies of it}, and places them in the output tray. \underline{N} is the number currently registering in the count display.

- Aspecto-temporal organisation viewpoint:
  \underline{When the start button is pressed, if there is an original in the feed slot, the photocopier makes \underline{N copies of it}, and places them in the output tray. \underline{N} is the number currently registering in the count display.}

- Control viewpoint:
  \underline{When the start button \underline{is pressed}}, if there is an original in the feed slot, \underline{the photocopier makes \underline{N copies of it}}, and

\(^2\) This example is taken from a requirements document from Kovitz [18]
places them in the output tray. N is the number currently registering in the count display. If the start button is pressed while photocopying is in progress, it has no effect. The number N in the count display updates in response to button pressed according to the state table.

- Causality viewpoint:
  When the start button is pressed, if there is an original in the feed slot, the photocopier makes N copies of it, and places them in the output tray. N is the number currently registering in the count display. If the start button is pressed while photocopying is in progress, it has no effect. The number N in the count display updates in response to button pressed according to the state table.

Every viewpoint would produce, after TRD filtering, a two-folded output: the filtered text and its associated semantic values. Our first hypothesis is that semantic values may allow to find semantic conflicts in large TRD, specially in requirements that are far away from each other in the TRD document. The following is an example taken from a in insurance system TRD:

(page 60) When the product's life-cycle is over, the system should trigger a premium-collection event.

(page 234) The system can prevent a premium-collection event but only an agent can cause it.

In the statement of page 234, filtered by the causality viewpoint, establishes that only an agent can cause a premium-collection event, while in page 60's statement (control viewpoint) there are traces of a machin-controlled situation over the premium collection event, where the system triggers the event.

The detection of conflicts needs the definition of validation rules based on semantic values. These semantic-value based rules would allow to detect inter-viewpoint conflicts and conflicts inside a viewpoint (for instance, to verify that, in incidence relations from the concepts relationships viewpoint, the relation between an element and its parts does never get reversed). They allow requirements analysis as well (for instance, a high proportion of cause and control statements may be sign of a Jackson's control problem frame [19]).

Our second hypothesis is that the filtered texts can be the input of other NL requirements processing tools that don't support large TRD processing. For instance, the conceptual schema generation tools from Circé [8] or ColorX [6] could be used to process the filtered text, regardless of its semantic values.

Figure 1 shows the proposed architecture to supports TRD semantic filtering. The rules configuration tool is intended to support a linguistic configuration phase, where the application's domain glossary is included into the linguistic resources, the CE and semantic-value based rules are fine-tuned in order to adapt them to a new application domain. A CE system would receive large TRD as an input and would filter it according to CE rules, calculating semantic values. TRD Viewpoint browser is intended to allow a user to browse between the source TRD and partial viewpoint-based views, and to allow further filtering according to application domain criteria. The Semantic filtering broker supports exploitation tools by handling semantic value and rules requests.
4 Conclusion and further work

This paper has proposed a semantic-based approach for NLRE, which extracts semantically relevant sentences from large TRD making use of linguistic-based rules. It has exposed the CE method, which organises rules, linguistic markers and clues, assigning semantic values according to four major viewpoints, considered as relevant from a RE perspective: relations between concepts, aspecto-temporal organisation, control and causality. Furthermore, this paper has outlined how semantic viewpoints could improve requirements analysis and validation in light of other NLRE approaches.

Currently, work is being done to evaluate the precision of CE rules, markers and clues (most of them issued from linguistic studies on scientific corpus) on industrial requirements documents, as well as on the implementation of a declarative language for semantic-value based rules in a way that could allow inter-operability between viewpoints. Based on other evaluations experiences [13,?] we can conclude that two kinds of evaluation will be necessary in order to know to which extent the proposed approach improve large TRD processing: a linguistic one, which will evaluate the quality of the filtered text fragments, and an empirical one, where real system analyst system analyst could use semantic filtering tools on real TRD.
References