OSIP NLP STUDY

Executive summary
Deliverable id : EXEC

Written by	***Responsibility*** + handwritten signature if no electronic workflow tool
Gaëtan Pruvost (Thales SIX)	
Florent Maye	
Gérald Garcia	
Verified by	
Approved by	

Approval evidence is kept within the document management system

This work has been performed in the frame of ESA contract 4000133471/20/NL/GLC/kk

Change Records

ISSUE	DATE	§ CHANGE RECORDS	AUTHOR
1.0	10/01/2022	Initial release	

This work has been performed in the frame of ESA contract 4000133471/20/NL/GLC/kk

Table of contents

1. High level view of the results (executive summary)

Model based approaches have proved to be efficient in supporting engineering activities, models replacing traditional document-based approaches. Nevertheless even in most advanced deployments, a lot of engineering artifacts are textual either because the return on investment of introducing models is too low on this particular case or because the information is more efficiently expressed in natural language even if consistency and correctness issues appears.

Huge progress has been made recently in AI-based Natural language processing (NLP), mainly driven by chat bots and vocal home assistant usages. The proposed idea consists in spinning in these technologies into space engineering process, studying how natural language processing can help the space engineer in daily activities.

Many engineering domains can take advantages of these technologies. The most evident one are the requirement management domains as most of the requirements are textual and are generally not formally modeled even if they have a certain level of structure and rules.

Using NLP technology semantic information can be extracted from textual requirements, which may have several advantages that will be evaluated: find related requirements to a given one, enhance search in requirement database, check consistency of traceability, smart comparison of specification contents, identification of suspicious requirements, identification of overlapping requirements, …

Natural language processing is also promising in linking related engineering textual artifacts (like design reports, justification reports, …) even with models contents.

The top-level goals for this deployment as stated in the proposal are to:

- Capturing use-cases relevant for AI and NLP technics.
- Define an overall framework for implementing solutions
- Implement search capability with business knowledge
- Implement search in structured data
- Explore similarity in a particular set of artefacts
- Explore graph based representations and merging with NLP

After a selection of use-cases, 5 use cases has been demonstrated in this study :

- “Smart searching” taking into account space domain : improving the capabilities of search engines by injecting a knowledge linked to our domain (for example Ariane 6 is a launcher).
- Searching extended to non-textual content : providing an unified search engine based on textual query that is capable to search both in textual content (requirements, documents, …) but also in model artefacts (functions, components, mass budgets, …)
- Similarities into a particular set of artefacts : improve search by relying on semantic distance between two artefacts using models trained on space domain corpus.
- Similarities into heterogeneous set of artefact : implementation of a trace assistant capable to propose to the user the most probable trace links between two set of artefacts (for example two specification for satisfaction links, between functions and related requirements, …)
- Integrating structured data (models) and unstructured data (text) into a single representation to improve textual similarities by model information.

This work has been performed in the frame of ESA contract 4000133471/20/NL/GLC/kk

These use-case has been implemented as proof of concept into this study. For each use-case an investigation phase has been done where multiple implementation option has been studied and the retained principle has been integrated into a micro-service single search architecture (with an unified web interface) capable to support almost all use-cases and sourcing its data (the engineering artefacts) into a knowledge graph structured by a ontology very similar to Osmose one.

This study has validated the interest and the feasibility of the different use-cases. The proposed use-cases for implementation has increasing level of difficulty with regards to current state of the art. Some use cases are very close to operational usage (search engine), some use-case has very encouraging results (for example the transfer learning on top of a Space BERT model to perform trace link recommendations) that has to be confirmed in other data set. Some experiment (in particular the fusion between information coming from a model and textual content, ie merging NLP transformers approach with graph ML) has to be investigated further and are still at very low TRL.