PURE: a Dataset of Public Requirements Documents

Alessio Ferrari
ISTI-CNR
Pisa, Italy
Email: alessio.ferrari@isti.cnr.it

Giorgio O. Spagnolo
ISTI-CNR
Pisa, Italy
Email: spagnolo@isti.cnr.it

Stefania Gnesi
ISTI-CNR
Pisa, Italy
Email: stefania.gnesi@isti.cnr.it

***Abstract*—This paper presents PURE (PUblic REquirements dataset), a dataset of 79 publicly available natural language requirements documents collected from the Web. The dataset includes 34,268 sentences and can be used for natural language processing tasks that are typical in requirements engineering, such as model synthesis, abstraction identification and document structure assessment. It can be further annotated to work as a benchmark for other tasks, such as ambiguity detection, requirements categorisation and identification of equivalent requirements. In the paper, we present the dataset and we compare its language with generic English texts, showing the peculiarities of the requirements jargon, made of a restricted vocabulary of domain-specific acronyms and words, and long sentences. We also present the common XML format to which we have manually ported a subset of the documents, with the goal of facilitating replication of NLP experiments.**

I. Introduction

Requirements are normally expressed with the most flexible of the communication codes, which is natural language (NL) [1]. Several authors have applied natural language processing (NLP) techniques in requirements engineering (RE) to address multiple tasks, including model synthesis [2], classification of requirements into functional/non-functional categories [3], classification of online product reviews [4], traceability [5], [6], ambiguity detection [7]–[9], structure assessment [10], detection of equivalent requirements [11], completeness evaluation [12], and information extraction [13]–[15]. With some exceptions, most of the works use proprietary or domain-specific documents as benchmarks, and replication of the experiments as well as generalisation of the results have always been an issue [16]. This paper presents PURE (PUblic REquirements dataset), a dataset of 79 publicly available requirements documents retrieved from the Web. The dataset is oriented to replication of NLP experiments and generalisation of results. The documents cover multiple domains, have different degrees of abstraction, and range from product standards, to documents of public companies, to university projects. We also defined a general XML schema file (XSD) to represent these different documents in a uniform format. We have currently ported a subset of the documents to this format to ease rigorous comparison of NLP experiments. The paper extends a recent conference contribution [17]. With respect to this previous work, we provide statistical information on the NL content of the dataset, we present the XSD schema adopted to format the documents, and we provide recommendations on the usage of the dataset.

The remainder of the paper is structured as follows. In Sect. II, we present the documents retrieved from the Web. In Sect. III, we describe the XSD file that we defined. In Sect. IV, we discuss the usage of the dataset and its limitations. Finally, Sect. V provides final remarks.

II. The PURE Dataset

The PURE dataset is composed of public requirements documents retrieved from the Web. To retrieve the documents, we queried Google with the OR-linked keywords *Requirements Documents*, *Requirements Specification*, *System Specification*, *Software Specification*, *SRS*, and we selected those links that pointed to requirements documents (in .doc, .pdf, .html, .rtf). In addition, we navigated the source Websites in which the documents were located, to search for additional requirements documents. Our search led to the identification of 79 documents. The whole dataset – together with the XSD file, and the XML files currently ported – can be downloaded from our Web-site (http://fmt.isti.cnr.it/nlreqdataset/). We do not claim a systematic coverage of all the public requirements documents available in the Web, also given the dynamic nature of Internet searches – e.g., some of the document source links could not be retrieved already at the time of writing, other documents may be retrieved in future searches. Instead, PURE should be considered as a sample of the requirements that can be retrieved from the Web. We informally inspected each document, and labelled it according to the following fields, plus additional ones, which provide some first-stage qualitative information.

- **Doc Name:** an alphanumerical ID that identifies the document.
- **Pages:** the number of pages of the document.
- **Structure:** a letter, or combinations of letters, indicating how requirements are structurally expressed. Can be: S = *structured*: if the requirements are expressed in a structured format, as, e.g., use-cases; U = *unstructured*: if requirements are expressed as unstructured NL descriptions; O = *one statement*: if each requirement is expressed in a single NL statement. If mixed ways of expressing requirements were used – e.g., if in the same document, we found both structured requirements (S) and

2332-6441/17 $31.00 © 2017 IEEE
DOI 10.1109/RE.2017.29

unstructured ones (U) –, we combined the letters with the + operator (i.e., S + U).

- **Level:** a letter indicating the degree of abstraction of the requirements. Can be H = *high-level* requirements, or L = *low-level* requirements. The judgment was subjectively given according to the following rationale. If further refinement of the document was required before the system could be implemented, we labelled the document with H. If the content of the document was ready for implementation, we labelled it with L. Analysis on the level of single requirements is left as future work.
- **Source:** a letter indicating if the source of the requirements is a University (U), or an Public/Private Organization (I). Documents tagged with U normally include case studies, or assignments performed by university students. Documents tagged with I include industrial strength requirements, both coming from private companies and from public bodies, including standardisation groups.

A complete table that summarizes all the requirements documents, and all the fields is available from our Website. Here, we show some statistics concerning part of the fields, and aggregate statistics on the NL content automatically extracted from the documents.

a) Pages: In Fig. 1, we show the number of pages for each document. We have a maximum of 288 pages, a minimum of 7 pages, an average of 47 pages, with a quite high standard deviation of 45 pages. This indicates a strong variability of the dataset in terms of length. More accurate indicators of the documents length (e.g., number of requirements) will be provided when all the documents will be formatted in XML.

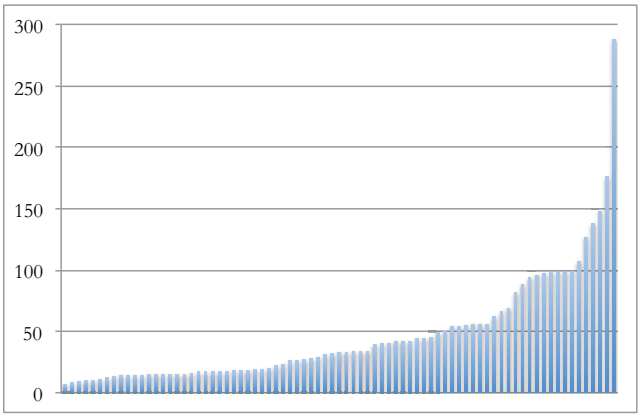

Fig. 1: Length of the different documents in terms of Pages.

b) Structure: In Fig. 2, we show the distributions of the different classes of structure. We see that the majority of the documents include a combination of unstructured content and requirements expressed in one sentence (U + O, 38%). Document with uniform formats – i.e., U, S, or O – are equally distributed, with about 15% of the documents for each class. Less represented are the other composite classes. However, the dataset appears already quite general and balanced for what concerns the structure of the requirements.

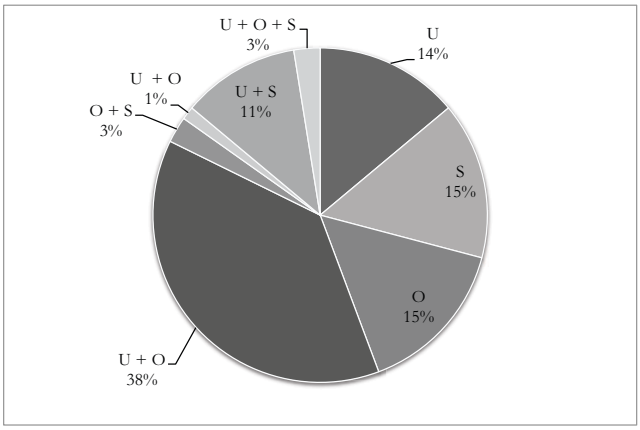

Fig. 2: Distribution of the different documents in terms of Structure.

c) Level and Source: Concerning the level of the requirements – not shown in the pictures – we have a dominance of high-level (H) requirements, with 71% of the documents classified with H, and 29% of them classified with L. Overall, more low-level requirements documents shall be added to the dataset to increase the balance. The dataset is also slightly unbalanced for what concerns the source of the requirements. Indeed, we have 62% of the requirements coming from Public/Private Organizations (I), and 38% from Universities (U). Additional industrial requirements are needed, since each company has its specific jargon [16], and, although the dataset includes documents from companies, it does not cover all the potential writing styles.

d) NL Content: We automatically extracted the text from the different documents, and in Table I we provide some aggregate statistics on the NL content of the documents. To give some meaning to our statistics, we compare our corpus with the classical Brown corpus, which includes 500 samples of English-language texts [18]. We consider the Brown corpus as a reference of generic English, since it includes text from 15 genres, which range from journalism to narrative. In the table, we differentiate between *tokens* (i.e., all separate text items including words and numbers, but excluding punctuation marks) and *lexical words* (i.e., all terms with the exception of numbers, punctuation marks, and stopwords, such as articles, pronouns, etc.).

The two corpora have a similar total number of tokens and lexical words, which implies that our analysis is performed between corpora of comparable size – also the distribution of tokens, not shown here, is comparable. On the other hand, the size of their *vocabularies* severely differ both in terms of lexical words (21,791 vs 46,018) and stems, i.e., morphological roots (16,011 vs 29,846). In particular, we can say that, in requirements documents, the vocabulary is *about a half* of the vocabulary used in generic texts. This is also indicated by the value of lexical diversity [19], computed as the number of different unique word stems and the total number of words (0.031 vs 0.054). If we compare the words used in the two

TABLE I: Aggregate statistics on the NL content of the documents.

Indicator	PURE	Brown
Number of Tokens	865,551	1,034,378
Number of Lexical Words	522,444	542,924
Vocabulary Size (Lexical Words)	21,791	46,018
Vocabulary Size (Stems)	16,011	29,846
Number of Sentences	34,268	57,340
Average Sentence Length (Tokens)	25	18
Average Sentence Length (Lexical Words)	15	10
Lexical Diversity	0.031	0.054

corpora, we see that **62%** of the lexical words used in PURE do not appear in Brown (values not reported in the table). This confirms that requirements documents use a specific vocabulary [13], with acronyms and domain specific terms, which highly differs from the common English vocabulary. Looking at the length of the sentences, we see that sentences in PURE are seven tokens longer than the sentences in Brown. Overall, we can say that requirements in the corpus have a more restricted and specific vocabulary, but longer sentences, with respect to generic texts.

In Fig. 3, we report the frequency of the most common *lexical* words in the corpus. The typical requirements-related words, e.g., *system*, *shall*, *data*, *requirements*, *user*, *software*, *specification*, etc. appear at the top of this list. We also see that some domain specific acronyms, such as *npac* (Number Portability Administration Center) and *tcs* (Train Control System, Telescope Control System, Tactical Control System) also appear in the list. The first acronym occurs in the largest document of the corpus (288 pages), while the latter is an *ambiguous* acronym, occurring in four large documents of different domains. These observations indicate that (a) requirements documents use a peculiar terminology that is common to different documents, but (b) are also characterised by domain-specific expressions, which are highly frequent in the single documents. This analysis suggests that NLP tools trained on generic English texts (e.g., statistical parsers [20]) might not be suitable for requirements analysis, and proper customisations might be needed.

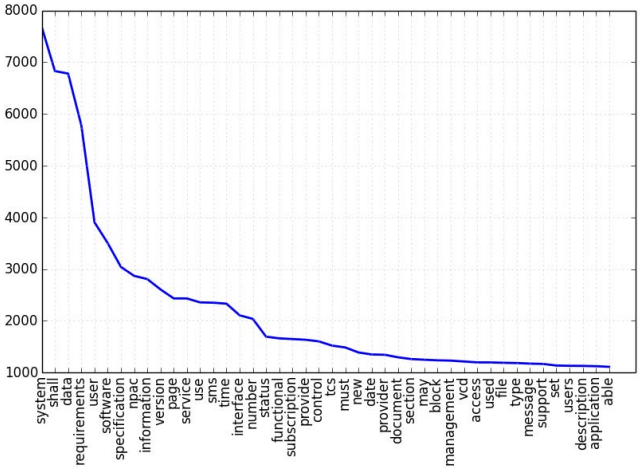

Fig. 3: Most frequent words in the overall dataset.

```
<element name="p">
 <complexType mixed="false">
  <sequence>
   <element ref="title" minOccurs="0" maxOccurs="1"/>
   <choice maxOccurs="unbounded">
    <element ref="p" minOccurs="0" maxOccurs="unbounded"/>
    <element ref="req" minOccurs="0" maxOccurs="unbounded"/>
    <element ref="b" minOccurs="0" maxOccurs="unbounded"/>
    <element ref="glossary" minOccurs="0" maxOccurs="1"/>
   </choice>
  </sequence>
  <attribute name="id" type="string"/>
 </complexType>
</element>
[...]
<element name="req">
 <complexType mixed="false">
   <sequence>
    <element ref="b"/>
    <element ref="modifier" minOccurs="0" maxOccurs="1"/>
   </sequence>
   <attribute name="id" type="string" use="required"/>
 </complexType>
</element>
[...]
<element name="req_document">
 <complexType mixed="true">
  <sequence>
   <element ref="title" minOccurs="1" maxOccurs="1"/>
   <element ref="version" minOccurs="1" maxOccurs="1"/>
   <element ref="issue_date" minOccurs="0" maxOccurs="1"/>
   <element ref="file_number" minOccurs="0" maxOccurs="1"/>
   <element ref="source" minOccurs="0" maxOccurs="1"/>
   <element ref="change_log" minOccurs="0" maxOccurs="1"/>
   <choice maxOccurs="unbounded">
    <element ref="p"/>
    <element ref="req"/>
   </choice>
  </sequence>
 </complexType>
</element>
```

Fig. 4: Excerpt of the XSD file to represent requirements documents.

III. XSD and XML Files

We defined a preliminary version of an XML schema file, i.e., an XSD file, based on a review of the gathered requirements documents. A simplified excerpt of the XSD file is shown in Fig. 4. The element `req_document` (at the bottom of the figure), is the root element. It includes several mandatory (`minOccurs` = "1") and optional (`minOccurs` = "0") elements that might appear at the beginning of the requirements document. Moreover, it includes a set of paragraphs (`p`) and requirements elements (`req`) (Fig. 4, top).

Paragraphs `p` are complex types, which can have a `title`, and can include other elements in sequence. These elements are other paragraphs, glossary and requirements elements, or text body elements – complex elements identified with `b` in the figure, but called `text_body` in the original file. Requirements `req` are sequences of text body elements `b` with modifiers (e.g., M = mandatory requirement, O = optional requirement). Both paragraphs and requirements have identifiers in string format (e.g., 1.1.a, 2.7.i, *etc.*). All the other elements that are not defined here, but appear in Fig. 4, are reported in our original file, together with the definition of other useful elements, as, e.g., lists, cross-references, and glossary items.

The XSD was defined based on an informal review of the gathered documents. It is designed to be simple, clear and

sufficiently comprehensive, but, on the other hand, we cannot claim that all the peculiarities of the different documents are taken into account. Modifications on the XSD file will be performed along with the porting of the documents. We have manually represented 12 documents according to the XSD format. We have been faithful to the source, i.e., we did not correct formatting or other errors in the documents. For representative examples, we recommend to look at `2007-ertms.xml` and `2007-eirene_fun_7-2.xml`, two requirements documents of the railway domain available from our Website.

IV. Research Topics Enabled and Limitations

In this section, we discuss the NLP tasks that can be performed on the current documents (i.e., the research topics enabled by PURE), as well as those tasks that require additional work on PURE before it can be used as a benchmark.

The current documents can be used as they are for research on abstraction identification, document structure assessment, and model synthesis. Research on abstraction identification, and information extraction in general, can leverage the glossaries available for part of the documents, and compare automatically extracted abstractions with the terms of the glossaries, as performed, e.g., by Gacitua *et al.* [13]. Research on structure assessment can compare automated structure extraction techniques with the XML structure of the documents, as in Ferrari *et al.* [10]. Model synthesis research can use the different documents to produce graphical summary models to be further evaluated by human assessors, as performed, among others, by Robeer *et al.* [2].

Other tasks, such as requirements categorisation [21], ambiguity detection [7], and equivalent requirements identification [11], require further annotation to be performed by the RE community interested in NLP. For each specific RE task, manual annotations have to be provided for the requirements, in order to use the documents as training, test and validation sets, for supervised machine-learning algorithms, or as *gold standards* for unsupervised algorithms [20]. In practice, categories of functional/non-functional requirements shall be provided, as well as annotation of terms and phrases perceived as ambiguous and requirements considered as equivalent.

For the traceability task, the dataset is currently not suitable. Indeed, to identify requirements traces between requirements at different levels of abstraction, as performed, e.g., by Gervasi and Zowghi [22], we need high-level and low-level requirements belonging to the same project, with traceability links. Hence, for the traceability task, we suggest to refer to the benchmarks used by other authors (e.g., NASA CM1 [5], [22]).

V. Conclusion

This paper presents PURE, a dataset for natural language requirements processing. It is oriented to enable replication of NLP experiments and generalisation of results in RE. The dataset is currently composed of 79 documents in various formats, and 12 documents that have been ported to a common XML format. As future work, we are committed to port the whole dataset to XML, to enrich the dataset with other public documents, and to further share it in public hosting sites. We also expect the RE researchers interested in NLP to annotate the requirements for specific tasks, and to share the annotation schemes adopted, so that they can be reused. As highlighted in our previous work [17], APIs for manipulating the XML files are under development, an the RE community is also encouraged to contribute to PURE with additional documents.

References

[1] M. Kassab, C. Neill, and P. Laplante, "State of practice in requirements engineering: contemporary data," *Innovations in Systems and Software Engineering*, vol. 10, no. 4, pp. 235–241, 2014.

[2] M. Robeer, G. Lucassen, J. M. E. van der Werf, F. Dalpiaz, and S. Brinkkemper, "Automated extraction of conceptual models from user stories via nlp," in *RE'16*. IEEE, 2016, pp. 196–205.

[3] A. Casamayor, D. Godoy, and M. Campo, "Functional grouping of natural language requirements for assistance in architectural software design," *KBS*, vol. 30, pp. 78–86, 2012.

[4] W. Maalej and H. Nabil, "Bug report, feature request, or simply praise? on automatically classifying app reviews," in *RE'15*. IEEE, 2015, pp. 116–125.

[5] H. Sultanov and J. H. Hayes, "Application of reinforcement learning to requirements engineering: requirements tracing," in *RE'13*. IEEE, 2013, pp. 52–61.

[6] J. Cleland-Huang, A. Czauderna, M. Gibiec, and J. Emenecker, "A machine learning approach for tracing regulatory codes to product specific requirements," in *ICSE (1)*. ACM, 2010, pp. 155–164.

[7] S. F. Tjong and D. M. Berry, "The design of SREE: a prototype potential ambiguity finder for requirements specifications and lessons learned," in *REFSQ'13*. Springer, 2013, pp. 80–95.

[8] H. Femmer, D. M. Fernández, S. Wagner, and S. Eder, "Rapid quality assurance with requirements smells," *JSS*, vol. 123, pp. 190–213, 2017.

[9] C. Arora, M. Sabetzadeh, L. Briand, and F. Zimmer, "Automated checking of conformance to requirements templates using natural language processing," *IEEE TSE*, vol. 41, no. 10, pp. 944–968, 2015.

[10] A. Ferrari, S. Gnesi, and G. Tolomei, "Using clustering to improve the structure of natural language requirements documents," in *REFSQ'13*. Springer, 2013, pp. 34–49.

[11] D. Falessi, G. Cantone, and G. Canfora, "Empirical principles and an industrial case study in retrieving equivalent requirements via natural language processing techniques," *IEEE TSE*, vol. 39, no. 1, pp. 18–44, 2013.

[12] A. Ferrari, F. dellOrletta, G. O. Spagnolo, and S. Gnesi, "Measuring and improving the completeness of natural language requirements," in *REFSQ'14*. Springer, 2014, pp. 23–38.

[13] R. Gacitua, P. Sawyer, and V. Gervasi, "On the effectiveness of abstraction identification in requirements engineering," in *RE'10*. IEEE, 2010, pp. 5–14.

[14] T. Quirchmayr, B. Paech, R. Kohl, and H. Karey, "Semi-automatic software feature-relevant information extraction from natural language user manuals," in *REFS'17*. Springer, 2017, pp. 255–272.

[15] X. Lian, M. Rahimi, J. Cleland-Huang, L. Zhang, R. Ferrari, and M. Smith, "Mining requirements knowledge from collections of domain documents," in *RE'16*. IEEE, 2016, pp. 156–165.

[16] A. Ferrari, F. Dell'Orletta, A. Esuli, V. Gervasi, and S. Gnesi, "Natural Language Requirements Processing: a 4D Vision," *IEEE Software (to appear)*, 2017.

[17] A. Ferrari, G. O. Spagnolo, and S. Gnesi, "Towards a dataset for natural language requirements processing," in *REFSQ'17 Workshops*, 2017.

[18] W. N. Francis and H. Kucera, "Brown corpus manual," *Brown University*, vol. 15, 1979.

[19] D. Malvern, B. Richards, N. Chipere, and P. Durán, "Lexical diversity and language development," *Quantification and Assessment. Houndmills, Basingstoke, Hampshire: Palgrave Macmillan*, 2004.

[20] C. D. Manning and H. Schütze, *Foundations of statistical natural language processing*. MIT Press, 2003.

[21] A. Casamayor, D. Godoy, and M. Campo, "Identification of non-functional requirements in textual specifications: A semi-supervised learning approach," *IST*, vol. 52, no. 4, pp. 436–445, 2010.

[22] V. Gervasi and D. Zowghi, "Supporting traceability through affinity mining," in *RE'14*. IEEE, 2014, pp. 143–152.