Natural Language Requirements Processing: from Research to Practice

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

ABSTRACT

Automated manipulation of natural language requirements, for classification, tracing, defect detection, information extraction, and other tasks, has been pursued by requirements engineering (RE) researchers for more than two decades. Recent technological advancements in natural language processing (NLP) have made it possible to apply this research more widely within industrial settings. This technical briefing targets researchers and practitioners, and aims to give an overview of what NLP can do today for RE problems, and what could do if specific research challenges, also emerging from practical experiences, are addressed. The talk will: survey current research on applications of NLP to RE problems; present representative industrially-ready techniques, with a focus on defect detection and information extraction problems; present enabling technologies in NLP that can play a role in RE research, including distributional semantics representations; discuss criteria for evaluation of NLP techniques in the RE context; outline the main challenges for a systematic application of the techniques in industry. The crosscutting topics that will permeate the talk are the need for domain adaptation, and the essential role of the *human-in-the-loop*.

CCS CONCEPTS

• **Software and its engineering** → *Requirements analysis*; • **Computing methodologies** → *Natural language processing*;

KEYWORDS

NLP, requirements engineering

ACM Reference Format:
Alessio Ferrari. 2018. Natural Language Requirements Processing: from Research to Practice. In *ICSE '18 Companion: 40th International Conference on Software Engineering , May 27-June 3, 2018, Gothenburg, Sweden.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3183440.3183467

1 INTRODUCTION

Natural language processing (NLP) techniques have been largely applied to automate several requirements engineering (RE) tasks, including model synthesis [1], requirements categorisation [4], traceability [5], detection of equivalent requirements [6], information extraction [12], defect detection [18], and, more recently, classification of online product reviews [13]. In the latest years, we have observed a novel golden age of NLP [11], with relevant advances in machine/deep learning methods. However, the current uptake of recent NLP techniques in RE is quite limited. This technical briefing aims to provide the knowledge to best to profit from the golden age of NLP also in RE, considering the work performed in the past, and the challenges that need to be addressed to transfer current technologies to industry. The talk will be based on the knowledge gained by the author in safety-critical industrial contexts, in which requirements are the cornerstone of any development activity. It will provide practical examples that require minimal programming knowledge, and pointers to available tools. Hopefully, it will be useful for professional requirements engineers, but also for researchers in RE and software engineering in general, since natural language artifacts – e.g. documentation, code comments – are pervasive within the whole software process.

2 INDUSTRIALLY-READY TECHNIQUES AND APPLICATIONS

The first works on the application of NLP techniques to RE tasks started in the nineties, with valuable experiences mainly oriented to model synthesis (e.g., [1, 15]). After a long period of incubation, in which the techniques were experimented in research settings for different tasks, in recent years we have observed an increasing number of robust case studies on applications of NLP to real-world RE problems (e.g., [2, 7, 17]). The maturity of current NLP tools contributed to this advancement. Besides the Python NLTK[1] library for text processing, one relevant example is GATE (General Architecture for Text Engineering)[2], a user friendly tool for information extraction and pattern identification that was applied in RE industrial case studies by Arora *et al.* [2] and by Rosadini *et al.* [17]. The technical briefing will use the tool to illustrate the basic steps of most NLP tasks, and will showcase an application of GATE for defect detection in requirements.

Among other promising industrially-ready techniques, it is worth illustrating approaches for information extraction that focus on the identification of multi-word terms. Requirements use a domain-specific and even project-specific jargon, in which multi-word terms such as "automatic train protection" are extremely frequent. Automatically identifying these terms can support glossary definition, and any other task that requires to extract relevant entities from requirements documents, e.g., retrieval or categorisation. Among available techniques, *contrastive analysis* – used, e.g., by Ferrari *et al.* [9] and by Nasr *et al.* [16] – will be presented in details.

[1] https://www.nltk.org
[2] https://gate.ac.uk

3 ADVANCED NLP TECHNIQUES FOR RE

Recent NLP research provided novel technologies that, if appropriately used, can further improve currently available approaches. Among these techniques, the most relevant ones are known with the term *word embeddings*. This term encompasses a series of techniques for representing the meaning of a word in a dense numerical vector. Given these representations, the semantic similarity among words can be computed by measuring the similarity between vectors. Among the various word embeddings techniques, one of the most widely known is implemented in the software package `word2vec`[3]. Given an input corpus, `word2vec` generates, for each word, a semantic laden vector representation, i.e., the word embedding. Word embeddings have the interesting characteristic that words that have similar meanings in the input corpus are represented through vectors that are similar. By enriching the words of a requirement with their semantics, the use of word embeddings in RE can improve tasks such as traceability, equivalent requirements identification and requirements categorisation.

Pre-trained word embeddings, based on domain-generic input corpora are available with `word2vec`. However, requirements use domain-specific jargon in which generic words may have a non conventional meaning. Therefore, in RE, `word2vec` needs to be trained on domain-specific corpora. During the talk, the use of Wikipedia Crawling to define domain-specific word embeddings will be shown, based on work of Ferrari *et al.* [10].

4 EVALUATING NLP TOOLS IN RE

The evaluation of NLP tools for RE tasks is subject to an ongoing debate in the RE community [3], and, in the talk, a snapshot of the debate will be given by focusing on two main issues, namely the gold standard to be used for evaluation, and the context-dependent usage of evaluation measures.

Gold Standard: when evaluating NLP tools, a common approach is to use typical information retrieval measures, such as precision and recall [14]. These are measures that can be evaluated based on a gold standard, i.e., a manually annotated dataset, in which all the correct answers that the tool is expected to produce for a specific task are established in advance. If the gold standard is not well-defined, the evaluation of the tool – and also its training process, if supervised learning is applied – will be unsound. This may happen, for example, because the correct answers are not defined by the right experts. Domain expertise, or even project expertise is required to establish correct trace links for a traceability task, or for assessing requirements defects. This and other issues related to gold standard construction, together with solutions, will be discussed in the talk.

Context: the evaluation of an NLP tool for RE depends on the context in which the tool will be used. The context is characterised by several features, which range from the task to be addressed, to the industrial process in which the tool will be included, to the expected user of the tool. Different contexts require different treatments, in terms of evaluation, and should take into account the process cost that the inaccuracy of the tool may have. The talk will show how evaluation should be performed in representative contexts, and will provide guidance to perform a sound cost-based evaluation.

[3]https://radimrehurek.com/gensim/models/word2vec.html

5 CHALLENGES AND CONCLUSION

Modern NLP techniques are statistical in nature and require large datasets to properly work [8]. Furthermore, domain expertise is required to annotate the requirements to be used for supervised machine learning algorithms. Unfortunately, requirements are often confidential assets of companies, and domain experts are rarely available to annotate datasets. Furthermore, given the company specificity of requirements languages, an algorithm trained on the data of a company may not necessarily work on a different one. A possible solution to these problems is to develop tools that can be trained *on the job*, which gather data and learn their tasks from domain experts. These tools need to be integrated in the company process, and embedded in existing tools used by practitioners. These challenges and additional ones, related to domain adaptation and the human in the loop, will be discussed as a conclusion to the technical briefing.

REFERENCES

[1] Vincenzo Ambriola and Vincenzo Gervasi. 1997. Processing natural language requirements. In *ASE'97*. IEEE, 36–45.

[2] Chetan Arora, Mehrdad Sabetzadeh, Lionel Briand, and Frank Zimmer. 2015. Automated checking of conformance to requirements templates using natural language processing. *IEEE TSE* 41, 10 (2015), 944–968.

[3] Daniel M Berry, Jane Cleland-Huang, Alessio Ferrari, Walid Maalej, John Mylopoulos, and Didar Zowghi. 2017. Panel: Context-Dependent Evaluation of Tools for NL RE Tasks: Recall vs. Precision, and Beyond. In *RE'17*. IEEE, 570–573.

[4] Agustin Casamayor, Daniela Godoy, and Marcelo Campo. 2012. Functional grouping of natural language requirements for assistance in architectural software design. *KBS* 30 (2012), 78–86.

[5] Jane Cleland-Huang, Adam Czauderna, Marek Gibiec, and John Emenecker. 2010. A machine learning approach for tracing regulatory codes to product specific requirements. In *ICSE (1)*. ACM, 155–164.

[6] Davide Falessi, Giovanni Cantone, and Gerardo Canfora. 2013. Empirical principles and an industrial case study in retrieving equivalent requirements via natural language processing techniques. *IEEE TSE* 39, 1 (2013), 18–44.

[7] Henning Femmer, Daniel Méndez Fernández, Stefan Wagner, and Sebastian Eder. 2017. Rapid quality assurance with requirements smells. *JSS* 123 (2017), 190–213.

[8] Alessio Ferrari, Felice DellâĂŹOrletta, Andrea Esuli, Vincenzo Gervasi, and Stefania Gnesi. 2017. Natural Language Requirements Processing: A 4D Vision. *IEEE Software* 34, 6 (2017), 28–35.

[9] Alessio Ferrari, Felice dellâĂŹOrletta, Giorgio Oronzo Spagnolo, and Stefania Gnesi. 2014. Measuring and improving the completeness of natural language requirements. In *REFSQ'14*. Springer, 23–38.

[10] Alessio Ferrari, Beatrice Donati, and Stefania Gnesi. 2017. Detecting Domain-specific Ambiguities: an NLP Approach based on Wikipedia Crawling and Word Embeddings. In *RE'17 Workshops*. IEEE, 393–399.

[11] Gregory Goth. 2016. Deep or shallow, NLP is breaking out. *Commun. ACM* 59, 3 (2016), 13–16.

[12] Xiaoli Lian, Mona Rahimi, Jane Cleland-Huang, Li Zhang, Remo Ferrari, and Michael Smith. 2016. Mining Requirements Knowledge from Collections of Domain Documents. In *RE'16*. IEEE, 156–165.

[13] Walid Maalej and Hadeer Nabil. 2015. Bug report, feature request, or simply praise? on automatically classifying app reviews. In *RE'15*. IEEE, 116–125.

[14] Christopher D Manning, Prabhakar Raghavan, Hinrich Schütze, et al. 2008. *Introduction to information retrieval*. Vol. 1. Cambridge University Press.

[15] Luisa Mich. 1996. NL-OOPS: from natural language to object oriented requirements using the natural language processing system LOLITA. *Natural language engineering* 2, 2 (1996), 161–187.

[16] Sana Ben Nasr, Guillaume Bécan, Mathieu Acher, João Bosco Ferreira Filho, Nicolas Sannier, Benoit Baudry, and Jean-Marc Davril. 2017. Automated extraction of product comparison matrices from informal product descriptions. *JSS* 124 (2017), 82–103.

[17] Benedetta Rosadini, Alessio Ferrari, Gloria Gori, Alessandro Fantechi, Stefania Gnesi, Iacopo Trotta, and Stefano Bacherini. 2017. Using NLP to detect requirements defects: an industrial experience in the railway domain. In *REFSQ'17*. Springer, 344–360.

[18] Sri Fatimah Tjong and Daniel M Berry. 2013. The Design of SREE: A Prototype Potential Ambiguity Finder for Requirements Specifications and Lessons Learned. In *REFSQ'13*. Springer, 80–95.