Data Extraction Challenges

NLP4RE Tools Book Chapter

General Points

A new exclusion criterion ("Extension of an included paper") is introduced to remove duplicate tools that do not contain any significant improvement.

Marked comments

The following primary studies still had comments associated with them that were clarified and resolved according to the "solution" field.

ID	R046
Reference	Rietz, T., & Maedche, A. (2019, September). LadderBot: a requirements self-elicitation system. In 2019 IEEE 27th International Requirements Engineering Conference (RE) (pp. 357-362). IEEE.
Assigned	JUF
Comment	It is not really clear if the paper contributes an actual NLP tool or rather an interface to an existing chatbot which is outside of the papers contribution.
Clarification	The paper presents a tool called LadderBot, which simulates the "laddering interview technique for RE" (using a series of Why-questions). "As a technological foundation of LadderBot, we use the Microsoft Bot Framework on node.js. To visualize elicited ACV chains, we integrate the bot into a web application." The paper does not report any fine-tuning of the actual NLP component. The tool only seems to use it out of the box.
Solution	Include, add a comment, and discuss in the chapter

ID	R064
Reference	Sainani, A., Anish, P. R., Joshi, V., & Ghaisas, S. (2020, August). Extracting and classifying requirements from software engineering contracts. In 2020 IEEE 28th international requirements engineering conference (RE) (pp. 147-157). IEEE.
Assigned	MUN
Comment	The task type could here be either extraction or classification. The authors call it extraction, but what they do is to classify statements from contracts into obligations and not obligations. Obligations are the extracted potential requirements. They continue then to classify the extracted requirements into types relevant for contracts.
Clarification	

Solution	Label as *classification* and use it as an example for “overriding” reported categories

ID	R113
Reference	Alhoshan, W., Ferrari, A., & Zhao, L. (2023). Zero-shot learning for requirements classification: An exploratory study. Information and Software Technology, 159, 107202.
Assigned	JUF
Comment	Borderline case: the purpose of this paper is to explore a general approach (zero-shot learning), not to propose a new tool. Still, in order to explore the approach, the authors implement some prototypical solutions.
Clarification	
Solution	Include and discuss as borderline

“Other” RE Activity

The following list contains all tools where the RE activity attribute was categorized as “Other.” The “Re-class” column contains the new classification which the authors agree upon.

ID	**Tool**	**Description**	**Re-class**
R013	Temporal Requirements Classifier	Enhancement to Nikora's baseline classifier identifying temporal requirements	Analysis
R016	PReDUS	A NLP tool for the "detection of privacy content from requirements defined as User Stories"	Analysis
R020	Business rule classifier	A deep-learning tool for classifying business rules from user requirements specifications into implementation-centric classes	Modeling
R022	RecVec Converter	"This method takes a requirement text as input and produces a ReqVec for that requirement. ReqVec is a semantic vector representation for the requirement, which could be used in many requirement-related applications, such as, requirement categorization, the relatedness of two requirements detection, etc."	**Multiple**
R023	ASFR Identifier and Classifier:	A "sequence-based modelling approach to automate the identification and classification of ASFRs from SRS documents"	Modeling
R030	MWE Disambiguator	A "knowledge-based approach to disambiguate multiword expressions (MWEs) in requirements	Analysis

		document"	
R035	Trace Link Recovery and Label Prediction	"automated recovery and prediction of labels for horizontal trace links in the ITSs."	Management
R041	Refactoring recommender	A "novel approach that recommends refactorings based on the history of the previously requested features and applied refactorings."	(exclude)
R042	Interpretable Requirements Classifier	Requirements classifier (FR/NFR) using interpretable machine learning features.	Analysis
R047	Requirements demarcator	A "practical, accurate and fully automated approach for domain-independent requirements demarcation in textual [Requirements Specification]"	Analysis
R060	NoRBERT	"Non-functional and functional Requirements classification using BERT"	Analysis
R085	Co-AI	Colab-based abstraction identification	Analysis
R116	Requirements Classifier	"An end-to-end system to classify functional and non-functional requirements with minimal NLP-based preprocessing and no feature engineering."	Analysis
R128	Business rule classifier	Automatic classifier of business rules (BRs) in requirements specifications according to Ross BR classification taxonomy	Analysis
R136	PRCBERT	"Prompt learning for Requirement Classification using BERT (PRCBERT), which applies flexible prompt templates to achieve accurate classification of software requirements"	Multiple
R144	SBVR Extractor	A tool that extracts business vocabularies and business rules from UML use case diagrams using a custom-trained POS tagger.	Modeling
R151	CORG	Component-oriented synthetic textural requirements generator: a tool that "automatically generate comprehensive combinations of structurally-diverse synthesised textual requirements." from a disctionary of domain words and verb frames.	(exclude, but discuss)
R153	LTL2NL	A tool converting linear temporal logic (LTL) formula into natural language descriptions using a neural machine translation tool (OpenNMT)	Modeling
R154	Req2Spec	A tool that converts natural language requirements into formal specifications (for the Hanfor tool) using named entity recognition.	Modeling
R189	AffectedCodePredictor	A tool to predict code affected by a requirement based on semantically similar requirements and their trace link to code.	(exclude)

Unclear Tasks

The following list contains all tools where the tool task attribute contained a comment and might be subject to change. The column “Re-Class” contains the updated categorization in case the raters agree on the change.

ID	Tool	Description	Class	Re-Class
R023	ASFR Identifier and Classifier	A "sequence-based modelling approach to automate the identification and classification of ASFRs from SRS documents"	Classification	
R071	BTC EmbeddedPlatform	Requirements formalization from natural language to formal language.	Extraction	Modeling
R118	QAssist	An "AI-based QA approach aimed at providing assistance with requirements analysis. Given a question posed in NL about the requirements in an SRS, QAssist employs Natural Language Processing (NLP) to retrieve [...] relevant text passages [...] from the SRS and [...] from a domain-specific corpus."	Generation	Search & Retrieval
R125	DeepSTL	A "tool and technique for the translation of informal requirements, given as free English sentences, into Signal Temporal Logic (STL), a formal specification language for cyber-physical systems"	Extraction	Modeling
R153	LTL2NL	A tool converting linear temporal logic (LTL) formula into natural language descriptions using a neural machine translation tool (OpenNMT)	Generation	Modeling
R154	Req2Spec	A tool that converts natural language requirements into formal specifications (for the Hanfor tool) using named entity recognition.	Generation	Modeling
R157	Requirements Rewriter	A tool that classifies the quality of a natural language requirement as either good or bad and then improves a detected bad requirement using genetic algorithms	Extraction	Generation
R158	ReFeed	A tool that associates a requirement to a set of related user feedback, extracts relevant properties, and prioritizes the requirement based on this. The tool uses an ontology containing domain knowledge and NLP techniques.	Extraction	Extraction
R162	REVV-Light	A tool that identifies near-synonyms of natural language requirements using	Extraction	

		conceptual model extraction and semantic similarity		
R180	RAPID	Rational API Designer, "a novel conversational assistant that aids software developers in addressing non-functional requirements in the design of web APIs."	Generation	Search & Retrieval
R186	User Story Instability Predictor	A tool that predicts the instability (i.e., number of potential changes) of a user story	Classification	Extraction

Secondary tasks

During the resolution of the remaining comments, some tools emerged as fulfilling more than one NLP task. While we only record the primary task of each tool, the following list of tools (referenced by their paper ID) are candidates for additional tasks:

- R157: classification
- R158: tracing and relating
- R162: modeling

In case the extraction is extended in the future and the relationship between tools and their task type is updated from 1:1 to 1:n, these additional types can be considered.