NLP4RE ID Card

The NLP4RE ID-Card is a means of capturing a descriptive overview of an NLP4RE paper by highlighting key information that can be useful in diverse scenarios such as replication, teaching support, or secondary research.

The ID-Card consists of seven dimensions concerning (I) the problem tackled by the paper (RE Task), (II) the solution proposed (NLP task(s)), (III) the input and output of the solution, (IV) the raw data and the annotated dataset, (V) the annotators and annotation process, (VI) the implementation of the solution proposed in the paper (Tool), and finally (VII) how the solution is evaluated (Evaluation).

To get an informative summary, the ID-Card can be applied on one RE task at a time. Multiple Cards can be used to describe a paper tackling multiple RE tasks (e.g., domain model generation and incompleteness detection). However, the ID-Card allows multiple answers in most of the questions, since the same RE task (e.g., requirements classification) can be solved using multiple NLP tasks (e.g., information extraction and classification).

In the following, we present some hints to facilitate generating an ID-Card for a particular paper:

- We included the field "Other/Comments" as a possible answer in all questions. This field can be used to add remarks when the question is not applicable, the answer is not specified in the paper, introduce another alternative that is not among the possible answers, or further comments about the question.

- In Dimension III (NLP Task Details), we are interested in the "initial input" and the "final output" of the proposed solution. Different questions about the output are adapted to suit the NLP task.

- In case a dimension or specific questions are not considered in the paper (e.g., the paper uses a dataset from the existing literature and so no details are given about the annotation process), then the dimension/questions can be skipped.

- If a question with a free-text field has multiple answers (the paper is using multiple datasets in IV), then the respective answers should be provided in a comma-separated format (e.g., "dataset1, dataset2, dataset3").

Title and authors of the paper

An NLP approach for cross-domain ambiguity detection in requirements engineering

Alessio Ferrari, Andrea Esuli

I. RE Task

I.1. What RE Task is your study addressing?
Please choose one option only, if multiple options apply fill multiple sheets

- ○ Requirements Classification
- ○ Requirements tracing
- ○ Requirements defect detection
- ○ Model generation
- ○ Test generation
- ○ Requirements retrieval
- ○ Information extraction from legal documents
- ○ App review analysis
- ○ Dependency and relation extraction
- ○ Information extraction (e.g., features, terms) from requirements
- ◉ Other/Comments: modeling language ambiguity

II. NLP task(s)

II.1. What types of NLP task is your study tackling? *Select all options that apply*

- ☐ Classification (choose among classes)
- ☐ Translation (models, tests, etc.)
- ☐ Information Extraction (glossary, terms)
- ☑ Information Retrieval (search / rank)
- ☑ Other/Comments: language modeling

III. NLP Task Details

Input Granularity

III.1. What is the input of your NLP task?

- [x] Document
- [ ] Paragraphs
- [ ] Sentences
- [ ] Phrases
- [x] Words
- [ ] Structured/tabular text: e.g., use case specification, feature table
- [ ] Other (e.g., models, trace links, diagrams, code comments)/Comments

Output Type (NLP Task: Classification)

Fill in only you answered Classification in Section II

III.2. What type of classification is the study about?

- [ ] Binary-Single label (e.g., ambiguous vs. unambiguous)
- [ ] Binary-Multi label (e.g., functional vs non-functional)
- [ ] Multi class-Single label (e.g., predicting number of stars, given a review. any scoring task)
- [ ] Multi class- Multi label (e.g., type of quality in non-functional requirements, feature requests vs. bug reports vs. other)
- [ ] Other/Comments

III.3. What are the labels that can be assigned?
Please provide the labels, or point to the specific part of the original paper where the labels are specified

Output Type (NLP Task: Information Extraction)

Fill in only you answered Information Extraction in Section II

III.4. What is the level of granularity of the extracted elements?

- [ ] Sentences
- [ ] Phrases
- [ ] Words
- [ ] Other/Comments

III.5. What is the type of the extracted elements?

Output Type (NLP Task: Translation)

Fill in only you answered Translation in Section II

III.6. What is the type of output?

- [ ] Text (e.g., from a diagram generate description)
- [ ] Table (e.g, from requirements generate a table of features)
- [ ] Graphical diagram (e.g., generate UML diagram from text)
- [ ] Executable model (e.g., from requirements to Simulink or executable sequence/activity/state diagrams)
- [ ] Test Cases
- [ ] Other/Comments

III.7. What is the translation mapping cardinality between initial input and final output?

- [ ] 1 to 1
- [ ] 1 to many (e.g., 1 use case description to 3 diagrams)
- [ ] Many to 1 (e.g., many requirements translated into a single model, many feature table translated into a single feature diagram)
- [ ] Many to many (e.g., from many use case descriptions to one class diagram and multiple other diagrams)
- [ ] Other/Comments

Retrieved Elements (NLP Task: Information Retrieval)

Fill in only you answered Information Retrieval in Section II

III.8. What are the types of elements retrieved with the query?

- [ ] Document
- [ ] Paragraphs
- [ ] Sentences
- [ ] Phrases
- [x] Words
- [ ] Structured/tabular text: e.g., use case specification, feature table
- [x] Other (e.g., models, trace links, diagrams, code comments)/Comments: ranked list of works by cross-domain ambiguity

Output Type (NLP Task: Other)

Fill in only you answered Other in Section II

III.9. What is the type of output?

language models representing stakeholders use of the language, to be used for disambiguation

IV. Data and Dataset

IV.1. How many data items do you process?
Please report the numerical information and details about all the data that is used in your evaluation

47500 wikipedia pages from 5 domains

IV.2. In which year or interval of year were the data produced?

IV.3. What is the source of the data?

- [ ] Industrial Project, proprietary data
- [ ] Industrial Project, publicly available data
- [ ] Community-based Open Source Projects (i.e., the community is organic and there is no top-down control from a company)
- [ ] Textbook examples or cases
- [ ] Student Projects (i.e., requirements developed by students in a realistic settings, where the project lasts months)
- [ ] Toy Requirements (e.g., used for courses, developed by students during exercises)
- [ ] Legal/regulatory documents
- [ ] User generated content (e.g., app stores, user forums, ...)
- [x] Other/Comments: wikipedia

IV.4. What is the level of abstraction of the data (not limited to requirements)?

- [ ] User-level (e.g., "Uploading pictures with the app is so annoying!")
- [ ] Normative-level (e.g., "The processor shall not engage another processor without prior consent of the controller.")
- [ ] Business-level (e.g., "Provide customer with a pleasant online shopping experience")
- [ ] System-level (e.g., "When a GSI component constraint is violated, STS shall deliver a warning message to the system operator")
- [ ] Module-level (e.g., module design description, sub-system design description)
- [ ] Code-level (e.g., code or comments)
- [x] Other/Comments: encyclopedia

Type of data and formulation

IV.5. What is the format of the data?
Please consider all the data that are used in the paper

- [x] "Shall" requirements
- [ ] User stories
- [ ] Use cases
- [ ] User reviews
- [ ] Social media posts (e.g., Tweets)
- [ ] Bug/defect reports
- [ ] Messages in User forums
- [ ] Graphical Diagrams
- [ ] Scenarios (other than use cases)
- [ ] Legal text
- [ ] Other artefacts (e.g., slides)/Comments

IV.6. How rigorous is the format of the data?

- [x] Unconstrained natural language
- [ ] Template-based controlled natural language (i.e. the language follows a template or at least complies to some basic rules)
- [ ] Restricted grammar based controlled natural language (i.e., the language follows a restricted grammar)
- [ ] Semantically-augmented natural language (i.e., tags identifying semantic functions of part of the text, normally export from tools)
- [ ] Other/Comments

IV.7. What is the natural language of the data (if applicable) ?

English

Heterogeneity

IV.8. Please list which domains your data belongs to (e.g., automotive, satellite, entertainment, information systems) [comma separated text]

Computer Science, Electronic Engineering, Mechanical Engineering, Medicine, Sport

IV.9. From how many different sources your data comes from?

1

Licensing

IV.10. Is the dataset publicly available (also from other authors)?

- (•) Fully
- ( ) Partially
- ( ) Upon Request
- ( ) No
- ( ) Other/Comments

IV.11. What license has been used?

- (•) No license
- ( ) License: reuse only for non-commercial purposes
- ( ) License: reuse for any purposes
- ( ) License: modification only for non-commercial purposes
- ( ) License: modification for any purposes
- ( ) Other/Comments

IV.12. Where is the dataset stored?

- [ ] On a private/corporate website
- [x] In a repository (GitHub, GitLab)
- [ ] In a persistent platform with DOI (Zenodo, FigShare)
- [ ] Other/Comments

IV.13. Provide a URL to the dataset, if available, or to the original paper that proposed the dataset.

https://zenodo.org/record/1476902

V. Annotators and Annotation process

If no annotation has been done, go to Section VI.

Refer to the annotations for the NLP task listed in Section II.
If you have multiple NLP tasks, separate your answers below with a comma.

V.1. How many annotators have been involved?

2

V.2. How are the entries annotated?

- [x] Multiple annotators per entry
- [ ] One annotator per entry (no quality control)
- [ ] One annotator per entry (quality control, possibly on a sample, by a supervisor)
- [ ] Partly multiple annotators per entry, partly one annotator per entry
- [ ] Other/Comments

V.3. What is the average level of application domain experience of the annotators?

- [x] None or unknown
- [ ] Informed outsider
- [ ] Domain expert
- [ ] Other/Comments

V.4. Who are the annotators?

- [x] The designers of the technique/tool
- [ ] People who have direct contact with the designers
- [x] Independent annotators (e.g., crowd workers, pre-annotated data)
- [ ] Other/Comments

Annotation scheme

V.5. How was the annotation scheme established among the annotators?

- [ ] Only via class labels
- [ ] Oral agreement among the taggers
- [ ] Written guidelines with label definitions
- [x] Written guidelines with definitions and examples
- [ ] Other/Comments

V.6. Did you make the written guidelines public?

- [ ] No
- [ ] No, but are made available upon request
- [x] Yes, via a non-persistent URL
- [ ] Yes, via a persistent URL
- [ ] Other/Comments

Annotation process

V.7. Did you share other information that could support the annotators other than the elements to annotate?

- [x] No
- [ ] Surrounding context (e.g., the entire paragraph where the sentence to annotate appears)
- [ ] Entire document
- [ ] Other/Comments

V.8. Did you employ techniques to mitigate fatigue effects during the annotation sessions?

- ( ) Yes
- (•) No
- ( ) Other/Comments

Agreement

V.9. What are the metrics used to measure intercoder reliability?

- [x] Cohen's K
- [ ] Krippendorf's Alpha
- [ ] Fleiss' K
- [ ] Other/Comments

V.10. How were conflicts resolved?

- [x] Majority voting
- [ ] Discussion among taggers
- [ ] Resolution by authors
- [ ] Resolution by independent expert (not a tagger)
- [ ] Disagreements were disregarded
- [ ] Not resolved
- [ ] Other/Comments

V.11. What is the measured agreement?

varied on the tested cases

VI. Tool

By tool, we intend any implementation of the approach(es) proposed in the paper. This includes scripts, executable programs, APIs, etc.

VI.1. What is the type of proposed solution?

- [ ] Supervised ML
- [ ] Unsupervised ML
- [ ] Supervised DL
- [x] Unsupervised DL
- [ ] Rule-based
- [ ] Other/Comments

VI.2. What algorithms are used in the tool?

word2vec

VI.3. What has been released?

- [ ] Tool (e.g., a binary file) -- standalone
- [ ] Service on the web (e.g., REST)
- [ ] Library / API
- [x] Source code
- [ ] Executable notebook (e.g., Jupyter)
- [ ] Pre-trained model (ML-based trained model)
- [ ] No tool has been released
- [ ] Other/Comments

If no tool has been released go to Section VII.

VI.4. What needs to be done for running the tool?

- [x] No installation is needed (e.g., standalone tool)
- [ ] Import and integrate into your own code (e.g., via MAVEN)
- [ ] Compile and run
- [ ] Virtual machine / Docker containers
- [ ] Reproduce the tool from the explanation in the paper
- [ ] Other/Comments

VI.5. What type of documentation has been provided alongside the tool?

- [x] README file
- [ ] Pseudocode / illustration in the paper
- [ ] Wiki or dedicated website
- [ ] Tutorial
- [ ] Ready-to-use examples
- [ ] An academic paper
- [ ] No documentation
- [ ] Other/Comments

VI.6. What type of dependencies does the tool have?

- [ ] None
- [x] Open source libraries / software (e.g., python libraries, NLTK)
- [ ] Proprietary libraries / software (e.g., Google Cloud NLP)
- [ ] Specific OS
- [ ] Specific hardware (e.g., hardware GPU acceleration required)
- [ ] External knowledge bases (e.g., Wikipedia)
- [ ] Other/Comments

VI.7. How is the tool released?

- [ ] On a private/corporate website
- [x] In a repository (GitHub, GitLab)
- [ ] In a persistent platform with DOI (Zenodo, FigShare)
- [ ] Other/Comments

VI.8. What license has been used?

- (•) No license
- ( ) License: reuse only for non-commercial purposes
- ( ) License: reuse for any purposes
- ( ) License: modification only for non-commercial purposes
- ( ) License: modification for any purposes
- ( ) Other/Comments

VI.9. Where is the tool released? [url]

https://zenodo.org/record/1476902

VII. Evaluation

VII.1. What metrics are used to evaluate the approach(es)?

- [ ] Precision/Recall
- [ ] F-Score
- [ ] Accuracy
- [ ] AUC
- [ ] MAP
- [ ] LAG
- [ ] WER (word error rate)
- [ ] NIST - METEOR - ROUGE - BLEU
- [x] Other/Comments kendall's tau

VII.2. What is the validation procedure?

- [ ] Train-test split
- [ ] Cross-validation
- [ ] Cross-project validation
- [x] Entire dataset (e.g., no training)
- [ ] Other/Comments

VII.3. What baseline do you compare against?

- [ ] Existing tool or algorithm (e.g., from another paper for the same NLP task)
- [ ] Reconstructed tool from other research
- [ ] Automated, but self-defined (e.g., using ML algorithm not reported elsewhere)
- [ ] Theoretical / conceptual (e.g., majority class)
- [ ] Human baseline (i.e., comparison against human performance)
- [x] None
- [ ] Other/Comments

VII.4. Please provide more details about the baseline you compare against, if any.

Questionnaire for the participants

Time Spent

How many minutes did you spend filling out the NLP4RE ID Card? 40

Ease of Use

	Difficult	Somewhat Difficult	Neutral	Somewhat Easy	Easy
How easy is it to fill out the NLP4RE ID Card?	○	○	◉	○	○

Appropriateness
Indicate how appropriate the NLP4RE ID Card is for each of the intended use.

	Inappropriate	Somewhat Inappropriate	Neutral	Somewhat Appropriate	Appropriate
To facilitate data collection for literature surveys	○	○	○	○	◉
An educational tool on reporting research	○	○	○	○	◉
To facilitate reuse and replication	○	○	○	○	◉

Level of detail
Indicate the level of detail of the information captured by the NLP4RE ID Card for each of the intended use.

	Too generic	Somewhat too generic	Appropriate level of detail	Somewhat too detailed	Too detailed
To facilitate data collection for literature surveys	○	○	◉	○	○
An educational tool on reporting research	○	○	◉	○	○
To facilitate reuse and replication	○	◉	○	○	○

Intention to use
Indicate how likely is it that you or your team will use the NLP4RE ID Card for each of the intended use.

	Very Unlikely	Unlikely	Neutral	Likely	Very Likely
To facilitate data collection for literature surveys	○	○	○	◉	○
An educational tool on reporting research	○	○	◉	○	○
To facilitate reuse and replication	○	○	○	◉	○

Do you have any other remarks for the NLP4RE ID Card?

Maybe this paper is borderline for the format of this card, specially section IV, as it does not use a requirement dataset and it focused on language modeling for ambiguity detection which is an accessory aspect with respect to RE.

Do you have any other intended uses for the NLP4RE ID Card in addition to those we proposed?

No