APPROACHING INFORMATION SYSTEM CHALLENGES WITH NATURAL LANGUAGE PROCESSING

Topic B (rule based)

Rule-based Extraction of Constraints from Regulatory Texts

Leo Holzhauer | Feb 08 2024

TUM WS 2023/24

MOTIVATION

Ensuring efficient and robust business process compliance

Input

This methodology is applicable under the following conditions:
(a) The land subject to the project activity does not fall in wetland category;
(b) Soil disturbance attributable to the project activity does not cover more than 10 per cent of area in each of the following types of land, when these lands are included within the project boundary:
(i) Land containing organic soils;
(ii) Land which, in the baseline, is subjected to land-use and management practices and receives inputs listed in appendices 2 and 3 to this methodology.

After pre-processing and constraint search

methodology be applicable under follow condition : (a) land subject to project activity not fall wetland category ; (b) soil disturbance attributable to project activity not cover more than 10 per cent of area each of follow type of land , when land be include project boundary : (i) land contain organic soil ; (ii) land , baseline , be subject to land - use and management practice and receive input list appendix 2 and 3 to methodology .

Constraint as defined in GS

({check type of land}, {check compliance with applicability conditions of contained tools}, {directly follows}, {land_project_activity != wetland AND (containing_organic_soils == False OR (containing_organic_soils == True AND soil_disturbance_area <= 10)) AND (land_subject_to_land_use_and_management == False OR (land_subject_to_land_use_and_management == True AND receives_listed_inputs == True AND soil_disturbance_area <= 10))})

DATA

Table 1: Data sources (RW: real-world, S: synthetic)

	Symbol	Name	Constraints	Type	Source
1	AKTG	German Stock Corporation Act	8	RW	(6)
2	CDM1	Afforestation and reforestation of degraded mangrove habitats (UNFCCC CDM)	5	RW	(7)
3	CDM2	Afforestation and reforestation project activities implemented on lands other than wetlands (UNFCCC CDM)	5	RW	(8)
4	CDM3	Cable cars for mass rapid transit system (UNFCCC CDM)	6	RW	(9)
5	CDM4	Energy efficiency and/or energy supply projects in commercial buildings (UNFCCC CDM)	10	RW	(10)
6	CDM5	Electricity generation by the user (UNFCCC CDM)	14	RW	(11)
7	COFFEE	Coffee Roasting Process	26	S	(12)
8	PATG	German Patent Act	4	RW	(13)

TYPES OF CONSTRAINT ITEMS

Table 2: Types of constraint items

Type	Operators	Example
Inequality	<, <=, >, >=	`x less than y`
Equality	==, !=	`x must be y`
Meta	AND, OR	`if x, then y`

PIPELINE

Pre-processing

- split to chunks
- stop words
- lemmatize
- enumeration details
- linebreaks

Modelling

- search constraint items
- prune constraints
- build constraints

Evaluation

- similarity scores
- precision and recall

MODELLING

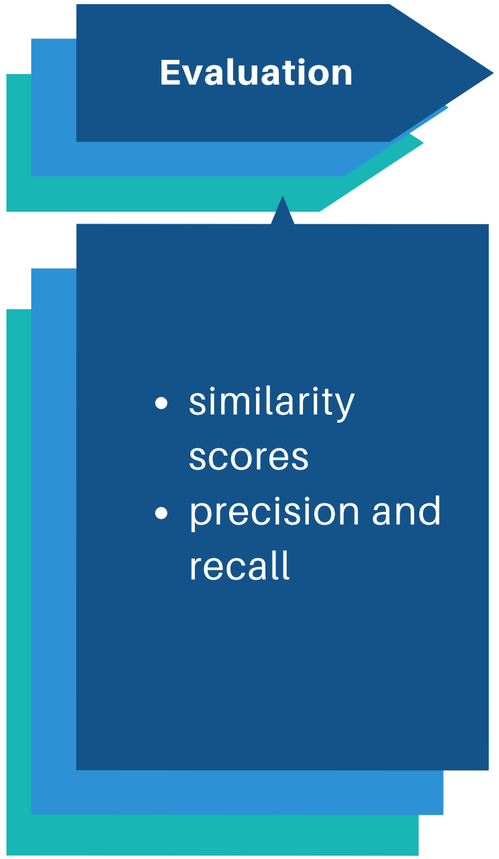
Modelling

- search constraint items
- prune constraints
- build constraints

Algorithm 1 Get Constraint

Require: nlp, $builder$, $chunk$, $eqParams$, $ineqParams$, $metaParams$, $enumSummary$, $linebreaks$
Ensure: $constraint$
Initialize $inequality$, $equality$, $meta$ searchers with respective parameters
Search for inequality CI in $chunk$
Search for equality CI in $chunk$
Search for meta enumeration CI in $chunk$ with $enumSummary$
Search for meta if and for clauses in $chunk$
Combine findings into $constraint$
if $constraint$ found **then**
    $constraint \leftarrow$ Determine context
    $constraint \leftarrow$ Insert connections
    $constraint \leftarrow$ Insert boolean CI
    $constraint \leftarrow$ Sort and prune
    if $constraint$ left **then**
        Build $formattedConstraint$
    end if
end if
Add $formattedConstraint$ to $constraint$
return $constraint$

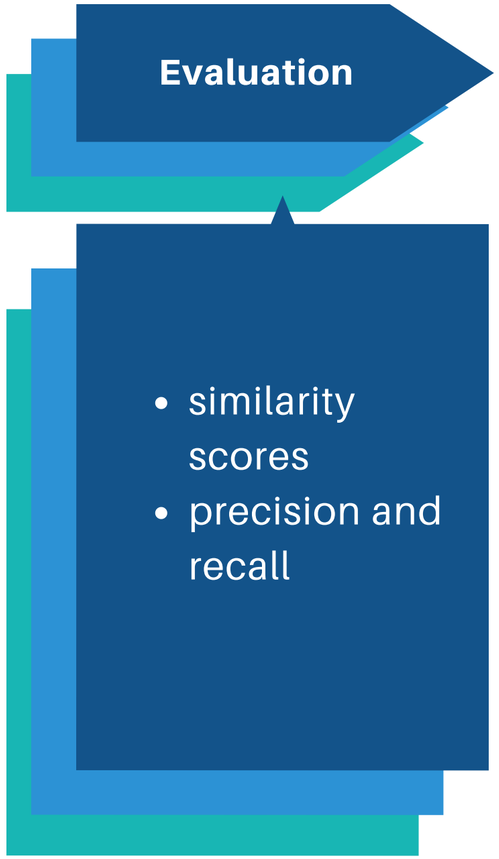
EVALUATION



$$\text{Precision} = \frac{|c_e \in C_E \mid \exists c_{gs} \in C_{GS} : S(c_e, c_{gs}) > \theta|}{|C_E|} \tag{1}$$

$$\text{Recall} = \frac{|c_e \in C_E \mid \exists c_{gs} \in C_{GS} : S(c_e, c_{gs}) > \theta|}{|C_{GS}|} \tag{2}$$

EVALUATION



Input

This methodology is applicable under the following conditions:
(a) The land subject to the project activity does not fall in wetland category;
(b) Soil disturbance attributable to the project activity does not cover more than 10 per cent of area in each of the following types of land, when these lands are included within the project boundary:
(i) Land containing organic soils;
(ii) Land which, in the baseline, is subjected to land-use and management practices and receives inputs listed in appendices 2 and 3 to this methodology.

Extracted constraint

```
methodology == follow_condition(( ) AND
(attributable_project_activity_cover <= 10 AND
((land_contain_organic_soil == True) OR
(receive_input_list_methodology == True))))
```

Constraint as defined in GS

```
land_project_activity != wetland AND
(containing_organic_soils == False OR
(containing_organic_soils == True AND
soil_disturbance_area <= 10)) AND
(land_subject_to_land_use_and_management == False OR
(land_subject_to_land_use_and_management == True
AND  receives_listed_inputs == True AND
soil_disturbance_area <= 10))
```

CONCLUSION

Annotated dataset	Modular processing pipeline
Reliability of rule-based detection	Flexibility of LLM-based construction
Theoretical framework for costraint categorisation	Precision and recall for constraint extraction
Process step - constraint matching	Tackling non-deterministic evaluation

OUTLOOK

TUM WS 2023/24

THANK YOU

Rule-based Extraction of Constraints from Regulatory Texts

Leo Holzhauer

Feb 08 2024