

Appendix A Rule-Based System Examples

The rule-based pipeline was developed through SAS[®] Visual Text Analytics 8.3 software [1], on a server accessible only within the Gemelli hospital's internal network. Consequently, we are unable to share it externally. However, Expression 1 reports an example of rules for the identification of nodule (a synonym of metastasis), where all the possible declinations of the lemma are included.

Listing 1 Rules for the Nodule Concept

```
CONCEPT: nodulo@
CONCEPT: nodulazione@
CONCEPT: formazione@ nodulare@
REGEX: m[ai]cronodul[io]
CLASSIFIER: linfadenopatia
```

Furthermore, Expression 2 presents a set of rules defining the concept of presence, listing all the potential Italian expressions.

Listing 2 Rules for the Presence Concept

```
CLASSIFIER: esito positivo
CONCEPT: sede di concept_nodule
CONCEPT: coerente@ con concept_nodule
CONCEPT: come per concept_nodule
CONCEPT: relativa@ a
CONCEPT: positivo@ per
CONCEPT_RULE:(ORDDIST_2, "_c{compatibile con}", "concept_nodule")
NO_BREAK: _c{non compatibile con}
```

Finally, as the last example, Expression 3 shows the rules used for labelling all the EHR sentences with the presence of nodules. Then when rules in 2 and 1 (that in the example 3 are indicated with the names of `concept_presence` and `concept_nodule`) are found in this order, the input text is tagged as positive to nodule.

Listing 3 Rules for the Presence of Nodule Concept

```
PREDICATE_RULE(presence, nodule):
  (SENT,(ORD,"_presence{concept_presence}",
    "_nodule{concept_nodule}"))
```

The same logic of concept rules shown above has been repeated for the semantic categories of lesion, high metabolic activity, staging and metastasis.

Appendix B Grid Search Implementation Details

Table B1 reports the hyperparameter ranges used in the grid search and the optimal values found for the 550-sample training setup. These same values were used for the 1100-sample setup without further tuning.

Table B1 Hyperparameter grid search results, performed during the training phase.

Hyperparameter	Grid	Optimal Value
body learning rate	[1e-6, 1e-3]	.2 e-4
num epochs	2, 3, 4	2
batch size	16, 32, 64	16
max iter	20, 30, 40	20
solver	newton-cg, lbfgs, liblinear	newton-cg

Appendix C Annotation Dashboard

The annotation dashboard was developed by using the Python Django framework [2] to allow the expert physicians to easily annotate each EHR with the corresponding outcome information (presence or absence of metastasis) and with any additional comments. The same interface could also be used to validate already classified data, by comparing model classifications against physicians' annotations.

The dashboard was a valuable instrument for making the labeling phase easier for clinicians who, through a simple interface, spent about 3 minutes per report.

The dashboard URL is accessible only from the hospital's internal network, then it's not sharable. However Figure C provides a screen of the annotation page.

The screenshot shows a web application interface for medical report annotation. At the top, there is a dark navigation bar with links: 'Archivi', 'NLP', 'Lista Referti', and 'Validazione'. A 'Sign Out' button is in the top right corner. Below the navigation bar, a progress indicator shows '16.0%'. A central dark bar displays 'Referti da Validare: 262' with navigation buttons 'INDIETRO', 'REFERTI', and 'AVANTI'. The main content area is divided into three columns. The left column, titled 'Dettagli', contains metadata for 'ID Referto: 287', 'Classe: metastasi', 'Validazione: metastasi', 'Metadati: Data Validazione: 26/01/24', 'Utente Validatore: admin', and 'Commenti:'. The middle column, titled 'Referto 287', contains a large text input field with the placeholder '< Input EHR Text >'. The right column, titled 'Azioni', contains a list of checkboxes for organ types: 'cute', 'ossa', 'linfonodi', 'polmone', 'fegato', 'cervello', and 'locale'. Below these are two buttons: 'NO METASTASI' and 'METASTASI', and a 'COMMENTI' button at the bottom.

Fig. C1 Annotation Dashboard

References

- [1] Anandarajan, M., Hill, C., Nolan, T., Anandarajan, M., Hill, C., Nolan, T.: Sas visual text analytics. Practical Text Analytics: Maximizing the Value of Text Data, 263–282 (2019)
- [2] Django Software Foundation. 2019: Django. <https://djangoproject.com>