

Creating an automatic genetic entity extraction pipeline for clinical notes in using Active Learning

Automated Extraction of Genomic Entities from Spanish and Catalan Clinical Notes using Active Learning

Motivation: There are no open-source NLP tools capable of extracting genetic entities from clinical notes in Spanish and Catalan. Integration of genomic data into the OMOP CDM is key to linking genetic biomarkers with phenotypes and adverse drug reactions.

Results: The configurations for the final iteration model achieved the following averaged Macro F1-scores over 5-fold cross-validation:

Config	LR	WD	Warmup	Class Weights	F1
C15	3e-05	0.1	0.2	(0.1, 10.0, 10.0)	0.9408
C12	3e-05	0.01	0.2	(0.5, 5.0, 5.0)	0.9331
C16	3e-05	0.1	0.2	(0.5, 5.0, 5.0)	0.9319
C13	3e-05	0.1	0.1	(0.1, 10.0, 10.0)	0.9305
C7	2e-05	0.1	0.2	(0.1, 10.0, 10.0)	0.9279
C11	3e-05	0.01	0.2	(0.1, 10.0, 10.0)	0.9276
C6	2e-05	0.1	0.1	(0.5, 5.0, 5.0)	0.9265
C8	2e-05	0.1	0.2	(0.5, 5.0, 5.0)	0.9265
C4	2e-05	0.01	0.2	(0.5, 5.0, 5.0)	0.9254
C1	2e-05	0.01	0.1	(0.1, 10.0, 10.0)	0.9249

More results:

- Prototype used for the first iteration of Active Learning labeled genes in 16% of notes.
- Higher learning rates were generally more favorable.
- 491 error records were collected across all 16 configurations and folds.
- FN account for 317 cases (64.6%), FP for 159 (32.4%), and Boundary Errors for only 15 (3.1%).

GENE 1

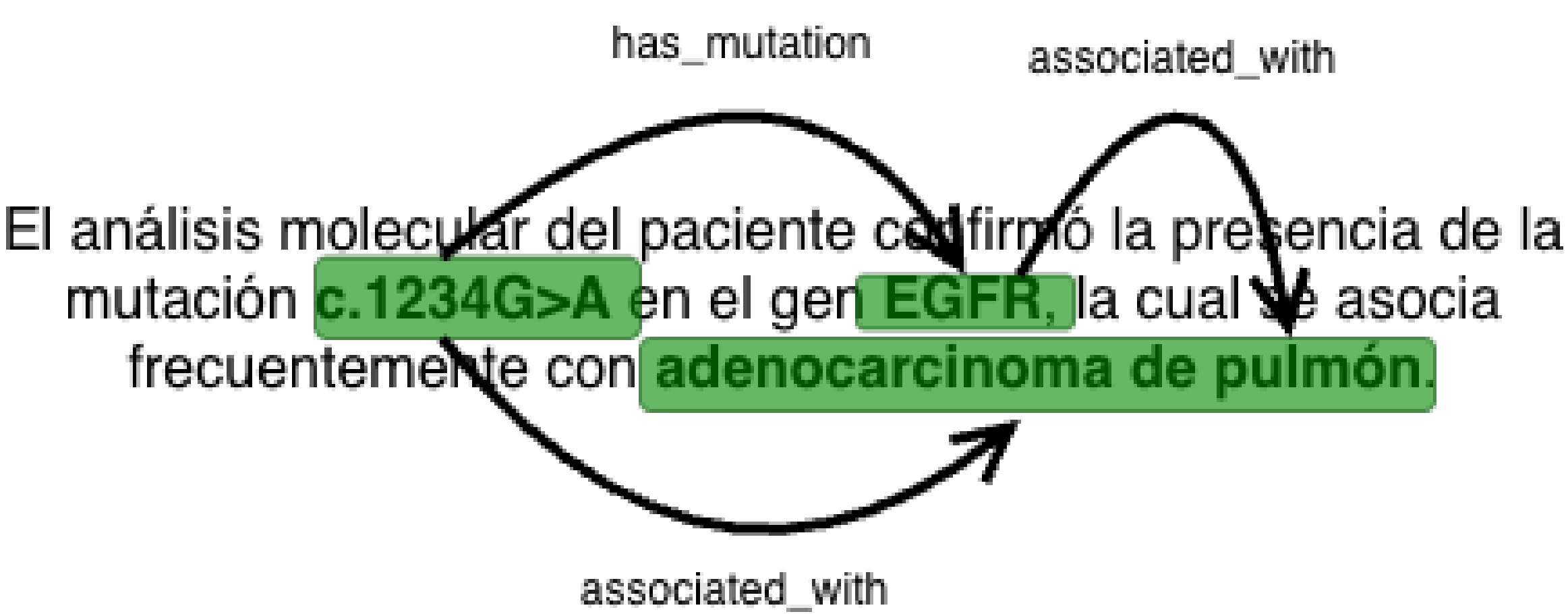
ADENDA: Comentat amb AP, NGS fet de la ultima biopsia pero per error informatic no estava penjat. KRAS MUTAT p.G12V~~KRAS: MUTAT c.35G>T, p.(Gly12Val) 36,6% ~A mes TP53 MUTAT c.743G>A, p.(Arg248Gln) 51,5% i FBXW7: MUTAT c.1394G>A, p.(Arg465His) 19,2%

Methods

- 1 **Fine-tuned transformer** model roberta-base-biomedical-clinical-es with a CRF classification head with penalty weights to mitigate class imbalance.
- 2 **Bootstrap Active Learning:** used a fine-tuned baby model trained with 50 labeled clinical notes to pre-annotate a larger dataset of 500 clinical notes. This iterative process helps with annotation's cold start.
- 3 **Hyperparameter tuning:** A grid search approach was used to test 16 different configurations regarding: learning rate, weight decay, warmup ratio and class weights.

Under development:

- Expansion to other entities such as mutation and phenotype, transforming the model from a simple gene tagger into a multi-class extractor capable of parsing the full genomic context of a clinical note.
- Adding a Relation Extraction model to link genes and mutations to phenotypes.



We present a useful method for creating genomic NLP tools in low-resource settings, successfully overcoming the obstacles related to data scarcity and extreme class imbalance inherent in clinical narratives. We are developing the first open dataset and model for genetic NERC in Spanish and Catalan.

