Universidad Nacional de Colombia

NER MODEL IN JUDGMENTS OF THE COLOMBIAN CONSTITUTIONAL COURT

BY JORGE LEONARDO RABA

VIDEO LINK

Introduction

Named entity recognition is a fundamental task in natural language processing that serves as a support in **legal systems**, similar to machine translation, the **sequence models** are used to identify entities.

Dataset

The dataset was generated with **judgments** of the **Constitutional Court of Colombia** that were exported to **IOB format** with **47** legal **named entities**

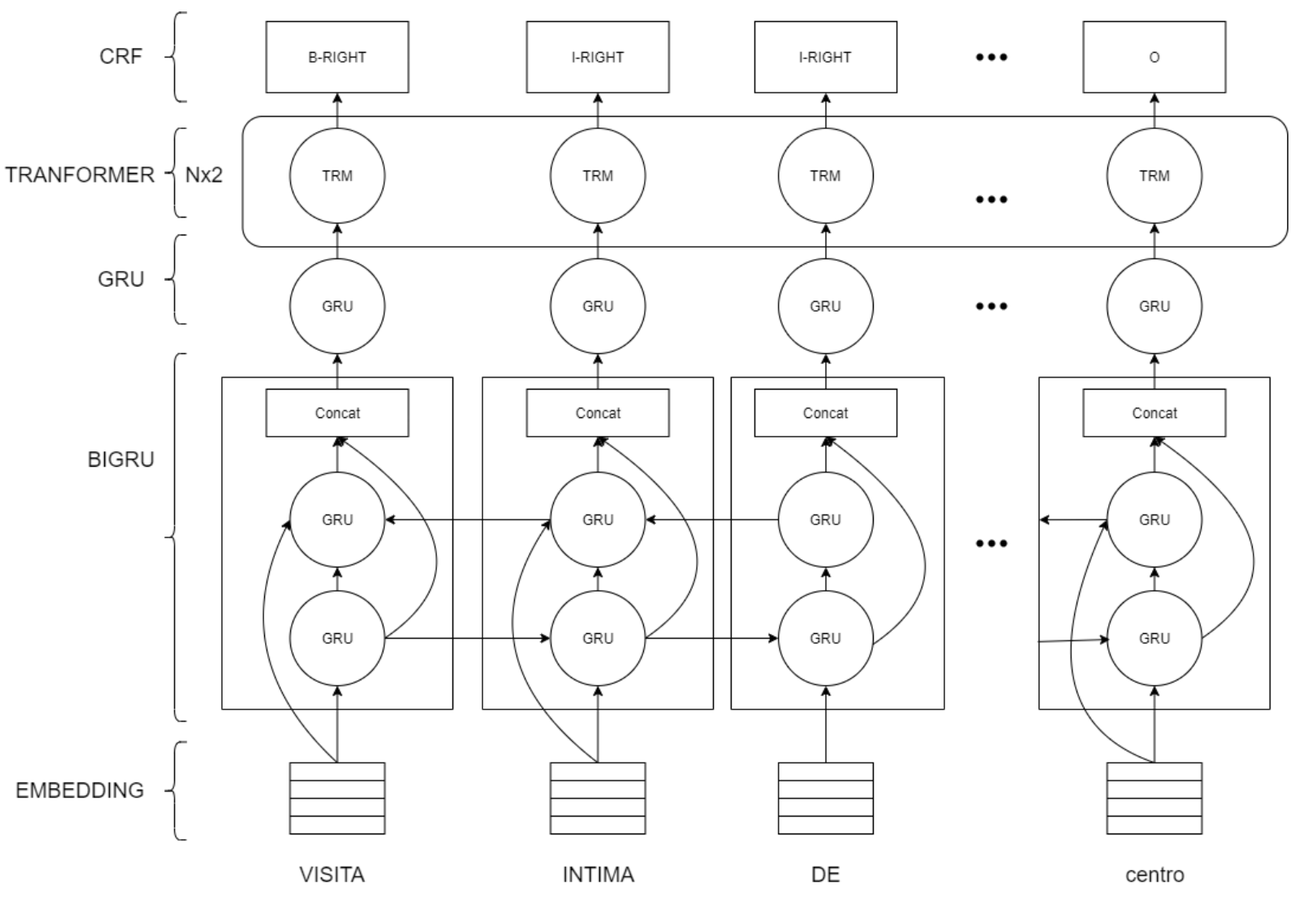
COUNTRY SENT

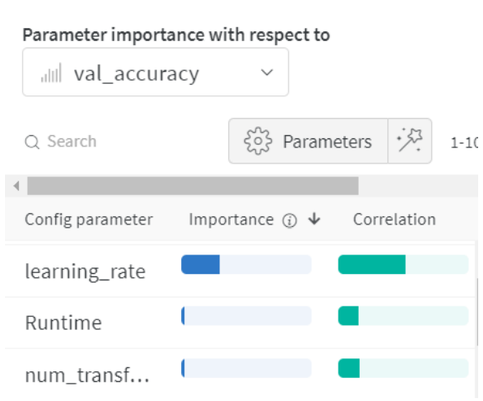
REPÚBLICA DE COLOMBIA Sentencia T-011/18 LEGITIMACION POR ACTIVA EN TUTEL
indígenas LEGITIMACION POR PASIVA EN TUTELA-Autoridad pública ACCION DE TUTE

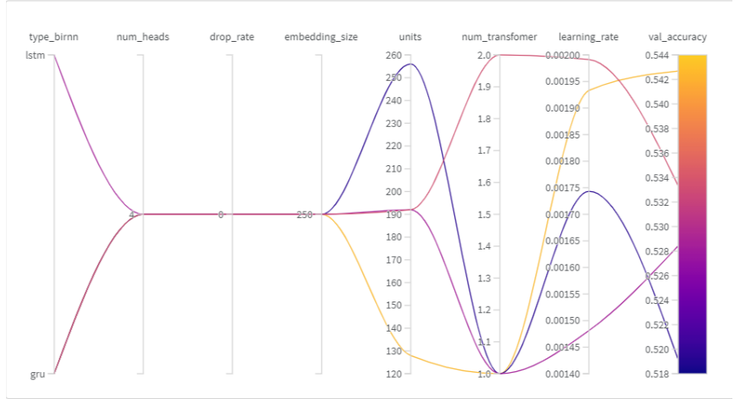
RIGHT

LA CONSULTA PREVIA-Único mecanismo judicial eficaz para garantizar que los pueblo

Model







Results

	precision	recall	f1-score	support
CHB	0.70	0.94	0.80	98
CLAIM	0.00	0.00	0.00	28
COMMUNICATION	0.00	0.00	0.00	1
COMPANY	1.00	0.23	0.38	13
COUNTRY	0.74	0.54	0.62	37
COURT	0.81	0.56	0.66	341
CRIME	0.57	0.38	0.46	21
CRT	0.00	0.00	0.00	8
CTY	0.86	0.15	0.26	39
DATE	0.51	0.83	0.63	619
DCT	0.59	0.82	0.69	181
DFD	0.61	0.73	0.66	679
DISEASE	0.00	0.00	0.00	4
DOC	0.69	0.82	0.75	567
EDUINS	0.78	0.66	0.71	38
GOV	0.39	0.53	0.45	503
GSTATE	0.00	0.00	0.00	29
HCI	1.00	0.29	0.44	7
ICOURT	0.65	0.48	0.55	23
ILAW	0.70	0.58	0.63	126
JDG	0.65	0.57	0.61	139
LAW	0.56	0.69	0.62	406
LOC	0.56	0.32	0.41	91
MAG	0.65	0.76	0.70	88
MUN	0.75	0.08	0.14	38
NORM	0.72	0.51	0.60	342
NPO	0.50	0.76	0.60	76
ORG	0.48	0.28	0.35	103
PDOCID	0.00	0.00	0.00	1
PER	0.38	0.32	0.35	81
PNT	0.65	0.46	0.54	240
POLPARTY	0.00	0.00	0.00	2
PRINCIPLE	0.32	0.39	0.35	49
PROCCESS	0.00	0.00	0.00	9
PUBPER	0.00	0.00	0.00	58
RES	0.00	0.00	0.00	22
RIGHT	0.61	0.43	0.50	699
SENT	0.74	0.64	0.69	412
SERVICE	0.85	0.38	0.53	60
SSI	0.00	0.00	0.00	2
SYSTEM	0.00	0.00	0.00	7
TIME	0.39	0.36	0.38	127
TOC	0.00	0.00	0.00	1
URL	0.00	0.00	0.00	5
VTM	0.59	0.53	0.56	103
micro avg	0.59	0.60	0.59	6523
macro avg	0.44	0.36	0.37	6523
weighted avg	0.59	0.60	0.58	6523

Conclusion

The CRF layer helps to reduce the error at the time of classification, however, as the model is sequential, it does not allow parallel processing.

To improve results, it is important to use a strategy that allows **generating artificial data** by making variations of the tagged entities and the surrounding words **keeping semantics of the paragraphs**.