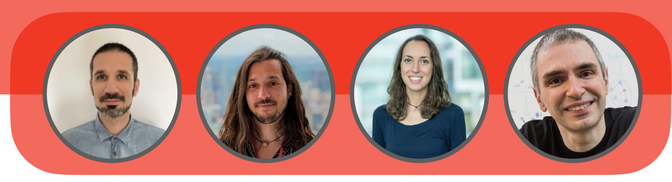
Dissecting Biases in Relation Extraction

A Cross-Dataset Analysis on People's Gender and Origin

Marco Antonio Stranisci, Pere-Lluís Huguet Cabot, Elisa Bassignana, Roberto Navigli

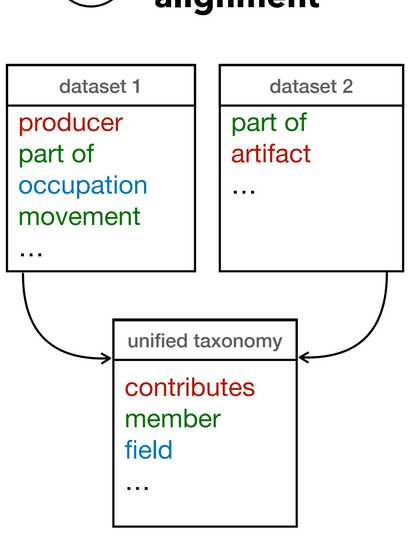


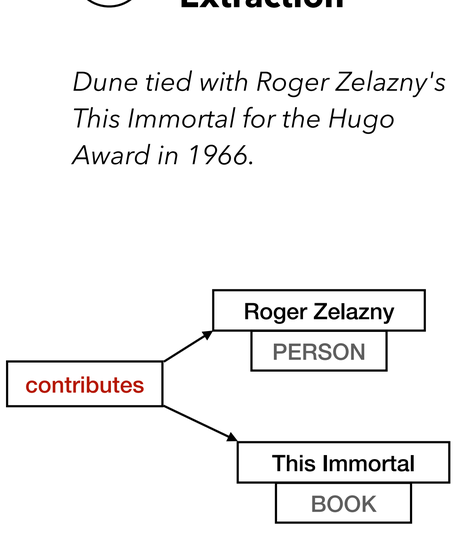
What?

We propose a methodology for the analysis of socio-demographic biases in the Relation Extraction pipeline, which is completely transparent in terms of interpretability.

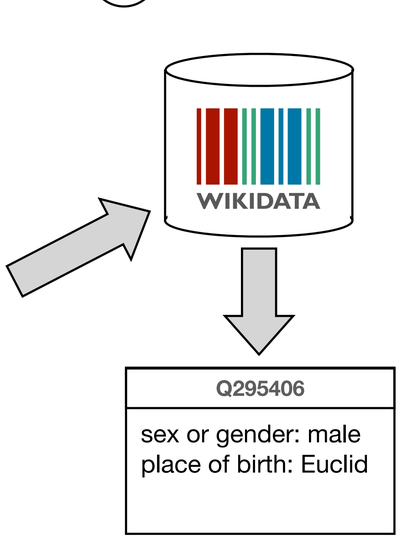
How?

1 **Label alignment** 2 **Relation Extraction** 3 **Entity Linking** 4 **Bias analysis**

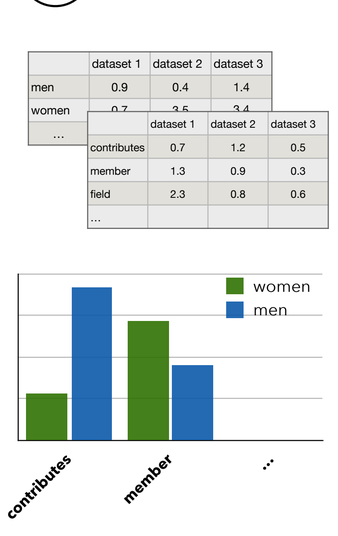




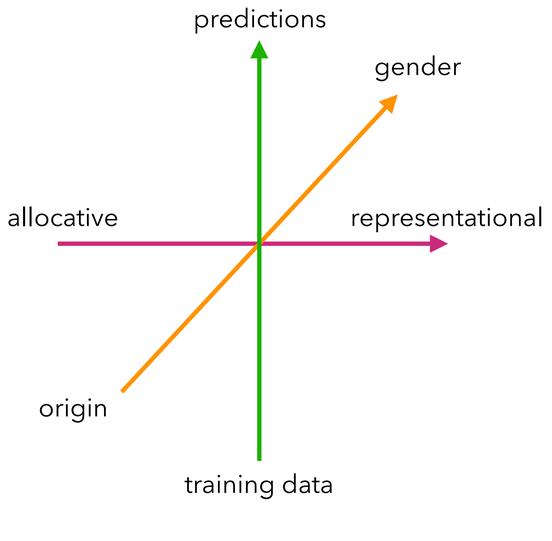
ReLiK (Orlando et al., 2024)



EntQA (Zhang et al., 2022)



predictions



Experiments and Results

Datasets:

- SRED^{FM} (Huguet Cabot et al., 2023)
- CrossRE (Bassignana and Plank, 2022)
- NYT (Riedel et al., 2010)

	SRED^{FM}	CrossRE	NYT
Women	20.0%	11.8%	17.3%
Global South	18.9%	10.0%	12.2%

Allocative bias in training data (% of entities)

	SRED^{FM}		CrossRE		NYT	
	M	W	M	W	M	W
contributes	0.28	0.475	**0.407**	0.291		
date	**1.038**	0.926				
field	**0.388**	0.291				
geographical	**0.469**	0.368	0.218	0.218	**3.251**	2.164
language	0.013	0.006				
member	0.21	0.164	0.229	0.218	**0.739**	0.283
participated	0.088	0.049	**0.278**	0.145		
position held	0.091	0.038	0.745	0.727	**0.085**	0.012
relationship	0.124	**0.215**	**0.098**	0.036	0.078	**0.211**
topic	0.001	0.001	0.018	0.018		

	SRED^{FM}		CrossRE		NYT	
	N	S	N	S	N	S
contributes	**0.758**	0.162	**0.447**	0.333		
date	**1.07**	0.993				
field	0.394	0.451			0.002	0.0
geographical	0.501	**0.64**	0.198	**0.644**	0.965	**1.019**
language	0.025	0.024				
member	0.252	0.201	0.3	0.222	**0.169**	0.121
participated	0.052	0.08	**0.218**	0.133		
position held	0.144	**0.196**	0.742	**1.200**	**0.036**	0.009
relationship	0.132	0.119	0.093	0.111	**0.077**	0.025
topic	0.002	0.002	**0.013**	0.0		

Representational bias in training data (t-test statistics)

	SRED^{FM}	CrossRE	NYT
Women	-3.5%	-2.2%	+5.6%
+SRED^{FM}		-5.8%	+0.6%
+ gen. balanced	-2.9%	-4.4%	0.0%
Global South	-1.7%	-8.3%	-2.1%
+SRED^{FM}		-6.7%	-1.6%
+ gen. balanced	-0.3%	-9.9%	-5.9%

Allocative bias in predictions
(False Positive Balance score)

The 5th Workshop on Gender Bias in Natural Language Processing

PIONEER CENTRE FOR AI
UNIVERSITÀ DI TORINO
SAPIENZA UNIVERSITÀ DI ROMA
IT-UNIVERSITETET I KØBENHAVN