

Graph-Linguistic Fusion: Using Language Models for Wikidata Vandalism Detection

Authors:
Mykola Trokhymovych
Lydia Pintscher
Diego Saez-Trumper
Ricardo Baeza-Yates

We introduce a next-generation Wikidata vandalism detection system that uses Graph2Text to unify structured and textual changes for processing by a single multilingual language model.

Introduction Wikidata is a large, multilingual, open repository of structured knowledge that supports numerous web services and applications. With around 8 changes per second, the platform is constantly evolving, but not all changes are made in good faith, making vandalism detection essential. While existing models assist in identifying bad-faith changes, challenges remain in improving detection accuracy, fairness, and unified processing of both structured and textual content to safeguard Wikidata’s knowledge integrity.

Contributions

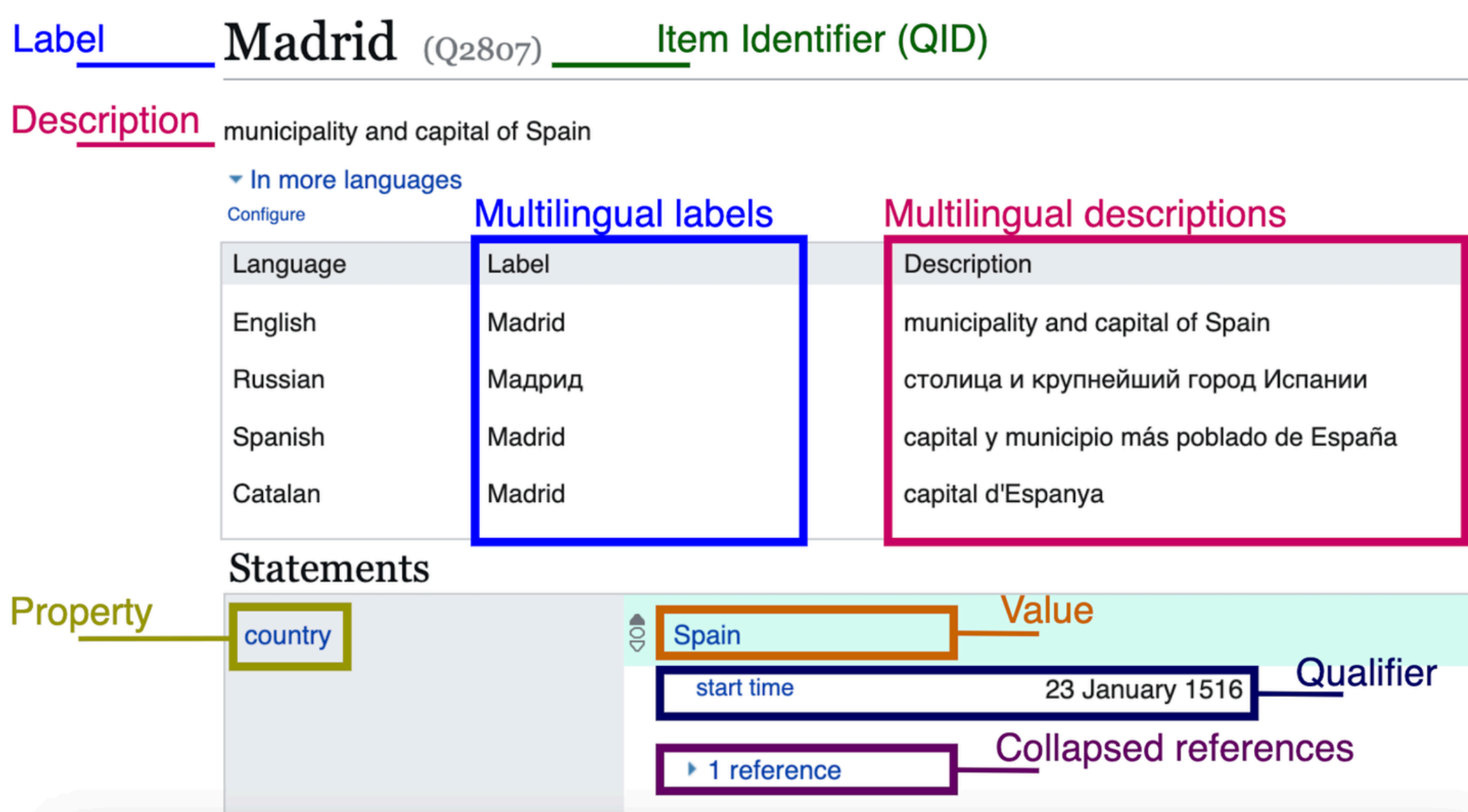
- The next-generation vandalism detection system for Wikidata, utilizing a multilingual language model to improve accuracy and fairness compared to the current production model.
- System productionalization addressing limitations imposed by resource-constrained infrastructure and product requirements.
- The publication of a new open benchmark dataset for vandalism detection in Wikidata, containing about 5M unique samples.

Objective Create a model to help editors to identify edits that require patrolling.

Approach

- Use implicit annotations (revert) to train the ML model.
- Apply Graph2Text to convert all structured knowledge into text, enabling unified content processing with a single multilingual language model.

Wikidata record elements



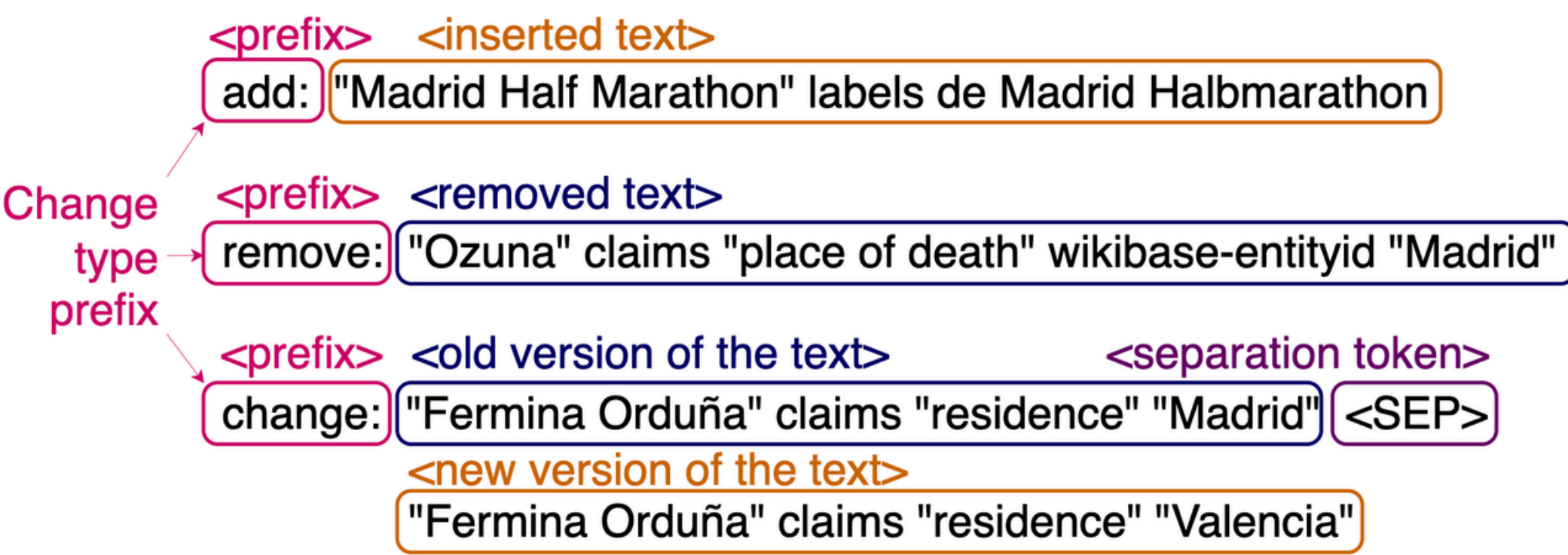
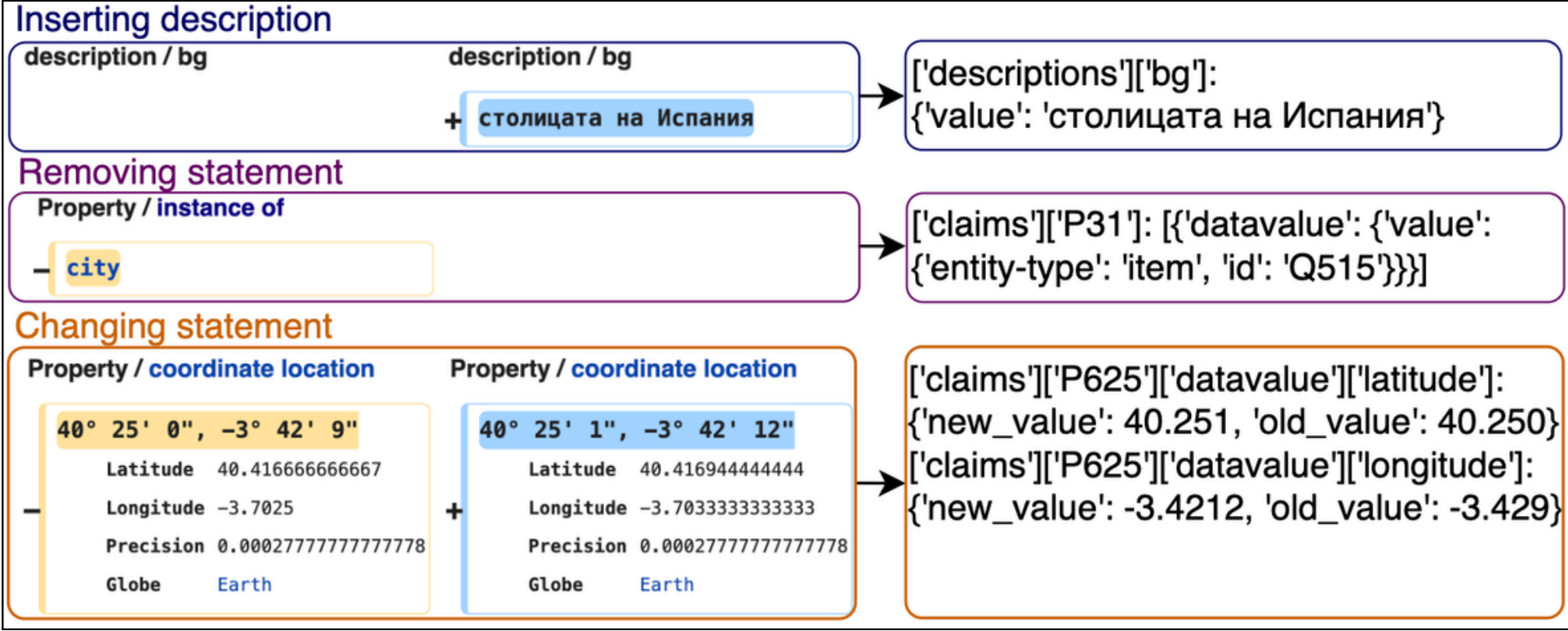
Dataset

Characteristic	Training	Hold-out
# of samples	4,197,231	645,264
Period	21 months	3 months
Anon. rate	10.7%	8.3%
Revert rate	7.9%	6.2%

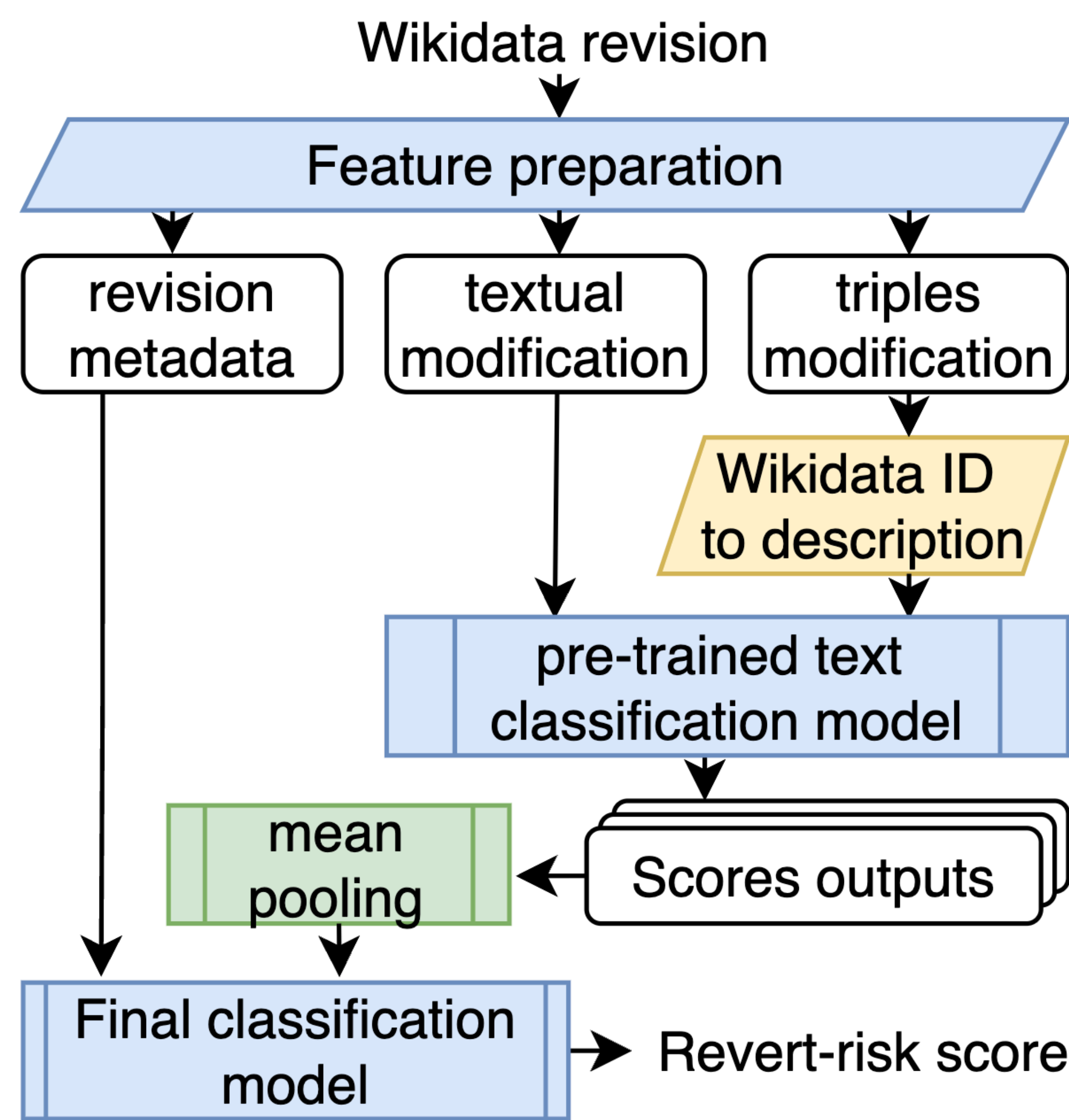
Filters applied:

- Filter for "revision-wars"
- Filter for "self-reverts"
- Filter bots revisions

Text processing



System design



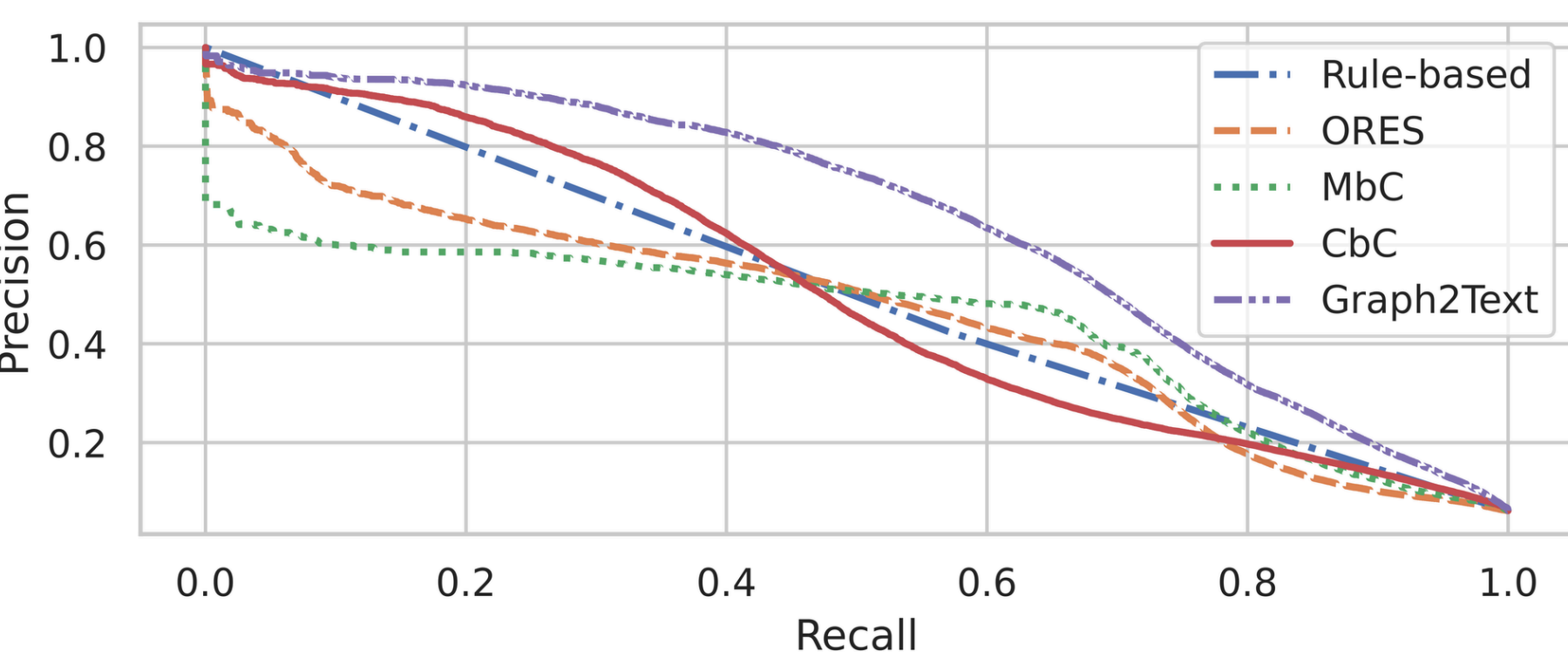
Evaluation

System performance on holdout testing set

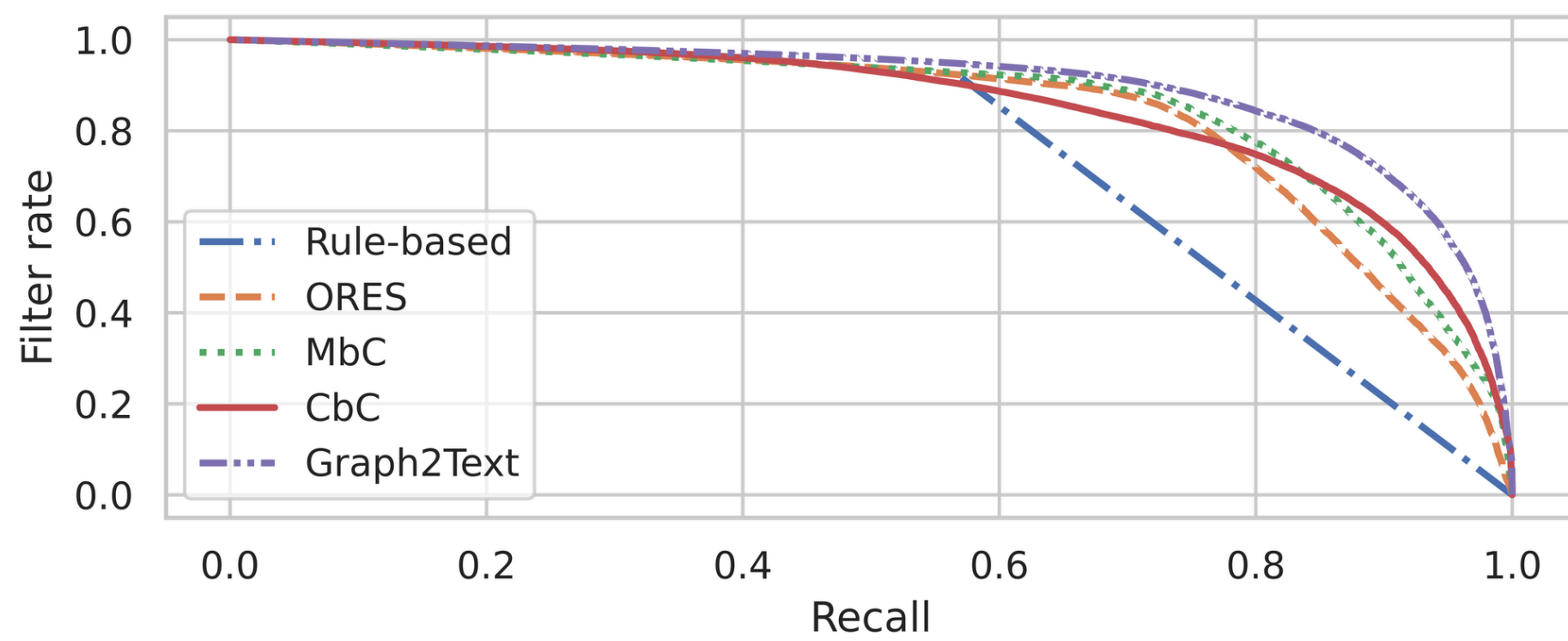
Model	AUC	CI	FR@99	FR@90	FR@70
Rule-based	0.760	[0.74, 0.78]	0.0	0.0	0.92
ORES	0.859	[0.84, 0.87]	0.45	0.88	0.94
MbC	0.880	[0.87, 0.89]	0.55	0.89	0.94
CbC	0.876	[0.86, 0.89]	0.60	0.82	0.93
Graph2Text	0.924	[0.91, 0.93]	0.71	0.91	0.96

System performance on expert-labeled data

Model	AUC	CI	FR@99	FR90	FR70
ORES	0.885	[0.879, 0.892]	0.593	0.799	0.881
Graph2Text	0.932	[0.926, 0.937]	0.698	0.846	0.918



The precision/recall curves for models.

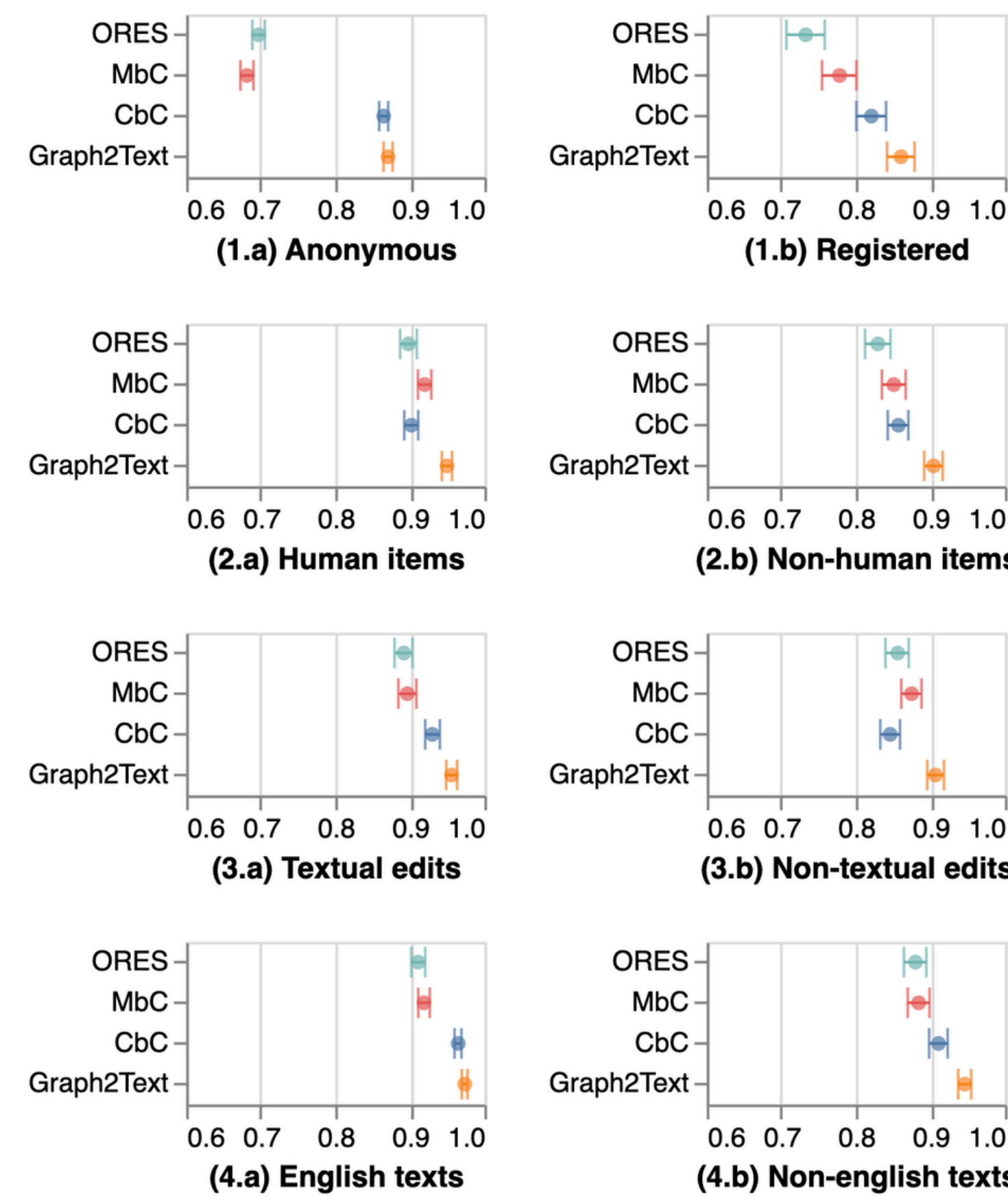


The filter rate/recall curves for models

System fairness based on holdout testing set

Model	DIR ^{anon}	DAUC ^{anon}	DIR ^{new}	DAUC ^{new}
ORES	5.69	0.035	1.37	-0.193
MbC	4.09	0.097	1.15	-0.155
CbC	2.07	-0.04	1.08	-0.027
Graph2Text	4.43	-0.01	1.24	-0.096

• **Rule-based:** anon. users are always reverted.
• **ORES:** previous SOTA model;
• **MbC:** Metadata-based Classifier (MbC)
• **CbC:** Content-based Classifier (CbC)
• **Graph2Text:** proposed model (CbC + MbC)



Model performance (AUC) comparison across various Wikidata edit characteristics.

Experiment

