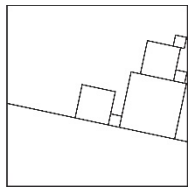
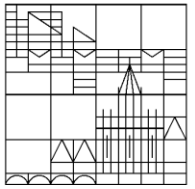
LREC-COLING 2024

Cluster of Excellence
The Politics of Inequality



Universität
Konstanz



GRIT: A Dataset of Group Reference Recognition in Italian

Sergio E. Zanotto[1], Qi Yu[1], Miriam Butt[1], Diego Frassinelli[1,2]

[1] Department of Linguistics & Cluster of Excellence 'The Politics of Inequality', University of Konstanz

[2] Center for Information and Language Processing, LMU Munich

In a Nutshell..

- **Introducing a New Task:**
  - Facing the need of a reliable identification of group references, we introduce the new task of Group Reference Recognition

- **Dataset Release:**
  - We build GRIT, a language resource manually annotated for Group References in Italian

- **Proof of concept:**
  - We verify the feasibility of the task by fine-tuning a BERT model and provide the results

Introduction

The reliable identification of group references is useful for the analysis of political discourses.

- **Group References:** Linguistic expressions that literally or figuratively refer to individuals or groups of people.

- **Group Reference Recognition (GRR):** new task for the automatic and reliable detection of group references.

Example of Group References

- (1) **Proper nouns:**
  - [...] *the Zapatists were unarmed.*
- (2) **Common nouns:**
  - *The teachers and the students of* [...]
- (3) **Relative clauses:**
  - *One of the reasons why the people who have no trust anymore* [...]

GRR vs. NER

- Proper nouns as in Example (1) *The Zapatists* might be identified using existing named entity recognition (NER) tools.
- **Common Nouns** as in Example (2) *The teachers, the students* and **Relative Clauses** as in Example (3) *The people who have no trust anymore* are out of the scope of traditional NER.
- **Figurative uses of group references** such as *Brazil won the world cup* need disambiguation from the entity *Brazil* as a country.

Related Work & Research Gap

The reliable identification of group references is useful for the analysis of political discourses:

- **Linguistics studies on group references:** sociolinguistic perspective on identity formation (e.g., Eckert, 1989; Trudgill, 2000; Labov, 2006); „Us vs Them“ dychotomy (Wodak, 2014; Zotzmann and O’Regan, 2016); semantic of group terms and their characteristics (Barker, 1992; Schwarzschild,1992; Carlson and Pelletier, 1995).

- **Social Sciences on group references:** Social Identity Theory (Tajfel and Turner, 2004); role of group appeals in political communication (Baker et al., 2008; Petrogiannis and Freidenvall, 2022).

Only very few computational studies have attempted to tackle the task of GRR:

- **Existing studies in automated GRR:**
  - manual recogniton (Russmann, 2020), dictionary approach (Haselmayer and Jenny, 2017; Decadri and Boussalis, 2020)
    → requiring extensive manual effort; not generalizable
  - machine learning approach (Licht and Sczepanski, 2023)
    → restricted to the genre of party manifestos; focusing only on English and German

Dataset Creation and Annotation

- **1000 texts from two sources:**
  - The PAISÀ Corpus (Lyding et al., 2013)
  - The Italian section of the corpus ParlaMint 2.1 (Erjavec et al., 2021).

- **Manual annotation of *group reference* labelled as *REFERENCE (REF):***
  - Mark all the tokens in the sentences that refer to people's identity (or to persons in general).
    - E.g., *the teachers*, *a student*
  - Minimal meaningful syntactic phrase: all the tokens that are fundamental to distinguish the group in the real world.
    - E.g., *the son of Luigi of Denmark*

- **2 annotators + reviewers:** Cohen's kappa of 0.82

GRIT: Summary & Access

Source	#Documents	#Sentences	#Tokens	#Tokens Labeled as REF
PAISÀ	900	3,074	85,623	14,128
ParlaMint	100	2,588	83,943	8,727
Total	1,000	5,662	169,566	22,855

Table 1: Overview of GRIT.

Open Access: GitHub - Sergio-E-Zanotto/grit

Experiment: Method

- We model the task of automated GRR as a binary token-level classification:

  Maybe the Sardinian senators should [...]

  0 1 1 1 0 0

- Training, validation and test sets:

	#Sentences	#Tokens	#Tokens Labeled as REF (Long Version / Short Version)
Train	2,968	106,482	18,043 / 13,285
Validation	371	13,495	2,379 / 1,766
Test	372	13,821	2,433 / 1,792
Total	3,711	133,798	22,855 / 16,843

Table 2: Summary of the training, validation and test set (80, 10, 10).

- Fine-tuned the `bert-base-italian-cased` model (**dbmdz/bert-base-italian-cased · Hugging Face**):
  - Long Version: minimal miningful syntactic phrase (e.g. *the teachers*)
  - Short Version: minimal miningful syntactic phrase without initial function words (e.g. *teachers*)

Experiment: Results

- **Classification results:**

	Accuracy	Precision	Recall	F1
Long	0.96	0.90	0.92	0.91
Short	0.96	0.86	0.90	0.88

Table 3: Classification results of fine-tuned BERT.

- Despite the **sparsity of group references**, the fine-tuned BERT model can **identify most of the instances**.
- Performances of short version are slightly inferior, reflecting the well-known **role of determiners in referential expressions** (Carlson and Pelletier, 1995).

Conclusion

- We **establish** the task of **GRR.**
- To this end, we introduced **GRIT, the first multi-domain and multi-genre large-scale dataset for GRR in Italian**.
- We show the **feasibility of automatizing it** in a way of being **robust to unseen data**.

LREC-COLING 2024

Cluster of Excellence
The Politics of Inequality

Universität
Konstanz

Thank you!

sergio.zanotto@uni-konstanz.de
qi.yu@uni-konstanz.de

https://github.com/Sergio-E-Zanotto/grit

SCAN ME