Bootstrapping Text Anonymization Models with Distant Supervision

Anthi Papadopoulou[1], Pierre Lison[2], Lilja Øvrelid[1], Ildikó Pilán[2]

[1]Language Technology Group, University of Oslo [2]Norwegian Computing Center
Oslo, Norway

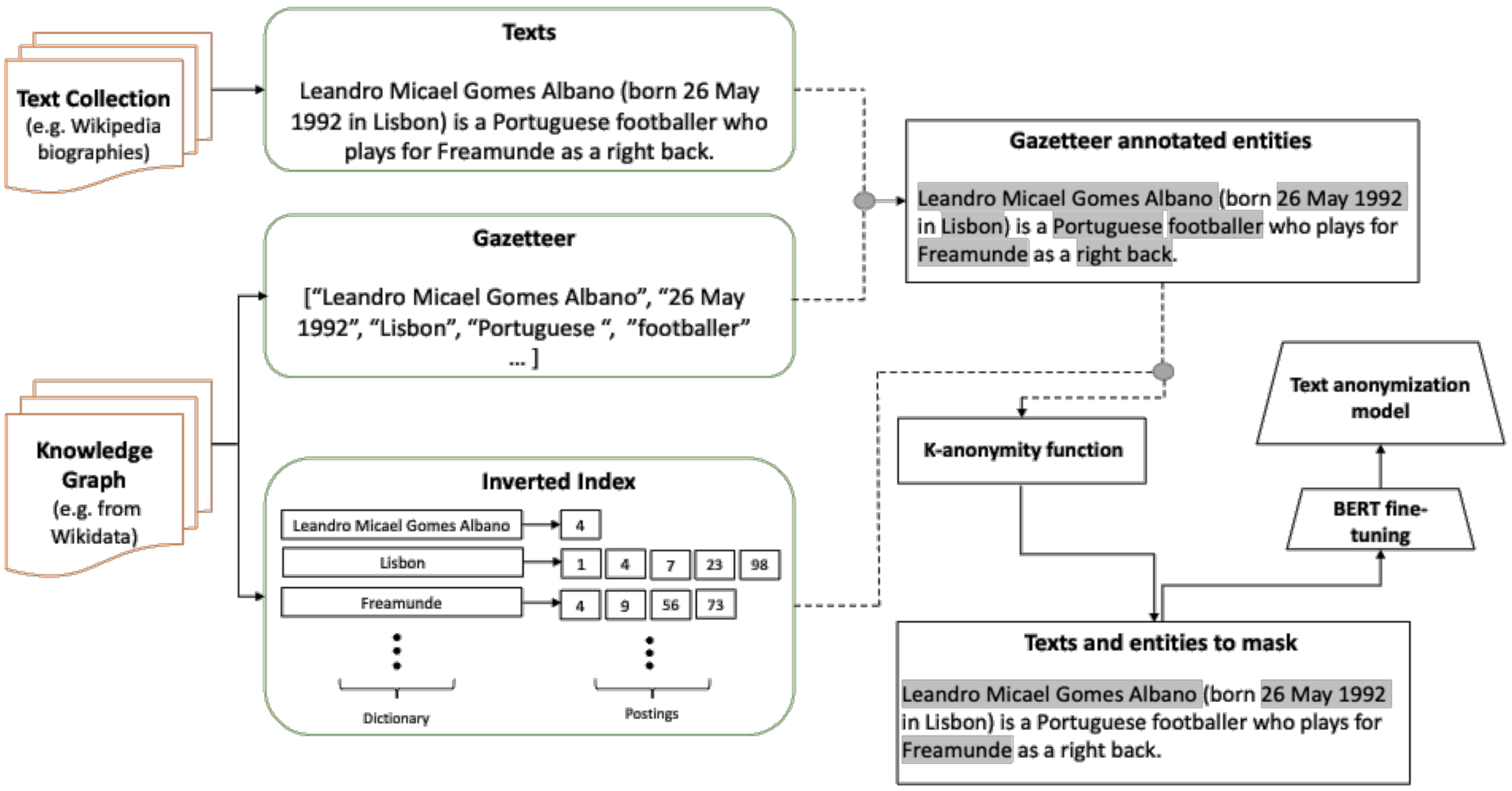
1. Motivation

- Anonymization of text data is **challenging**
- **Lack of labeled corpora** for this task due to data protection regulations and cost of annotation

2. Approach

- **Knowledge graph** as attacker's background information
- **Inverted index** as distant supervision source for automatic annotations
- ***k*-anonymity** to determine which tokens to mask based on the inverted index

→ **Automatic annotations** to fine-tune a pretrained LM



3. Evaluation Dataset

- 553 **manually annotated** Wikipedia biographies
- Two steps:
  I. Detect personal information
  II. Decide what to mask to protect the individual

Entity type	# of instances	%
DIRECT	1579	14%
QUASI	6281	56%
NO_MASK	3357	30%

4. Experimental Results - Wikipedia

System	Precision	$Recall_{all}$	$Recall_{direct}$	$Recall_{quasi}$	F1 score
$RoBERTa^{NER}$	0.770	0.845	0.810	0.801	0.805
$Greedy_{BERT}$	0.669	0.836	0.898	0.774	0.743
$Random_{BERT}$	0.650	0.832	0.895	0.770	0.730

Dataset and annotation guidelines available

https://github.com/anthipapa/textanonymization

5. Error Analysis

Original Text

Jenn Mierau is a Canadian electropop musician originally from Winnipeg who is now based in Montreal.

Human annotator

*********** *is a Canadian electropop musician originally from* ********, *who is now based in Montreal.*

Mask from supervised NER model

*********** *is a* ******** *electropop musician originally from* ********, *who is now based in* ********.

Mask from distantly supervised BERT model

**** ****** *is a Canadian* ********** *musician originally* ***********, *who is now based in* ********.

6. Experimental Results – TAB dataset

System	Precision	$Recall_{all}$	$Recall_{direct}$	$Recall_{quasi}$	F1 score
$RoBERTa^{NER}$	0.441	0.906	0.940	0.874	0.565
$TAB^{Longformer}$	0.836	0.919	1.000	0.916	0.876
$Greedy_{BERT}$	0.260	0.814	0.782	0.847	0.394
$Random_{BERT}$	0.263	0.668	0.530	0.806	0.377

7. Main takeaways

- Performance is dependant on the **quality and coverage** of the knowledge graph
- **No "gold" answer**, as long as the identity of the individual is protected
- Generic anonymization systems **over-mask** text that is domain specific

8. Future work

- Enhance quality & coverage of the knowledge graph
- Filter out information that is not sensitive
- Extend the inverted index with other background knowledge (e.g. co-occurrence estimates from raw, web-scale data)