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

4. Experimental Results - Wikipedia

<table>
<thead>
<tr>
<th>Entity type</th>
<th># of instances</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIRECT</td>
<td>1579</td>
<td>14%</td>
</tr>
<tr>
<td>QUASI</td>
<td>6281</td>
<td>56%</td>
</tr>
<tr>
<td>NO_MASK</td>
<td>3357</td>
<td>30%</td>
</tr>
</tbody>
</table>

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 ******** musican originally *********, who is now based in *********.

6. Experimental Results – TAB dataset

<table>
<thead>
<tr>
<th>System</th>
<th>Precision</th>
<th>Recall_all</th>
<th>Recall_direct</th>
<th>Recall_quasi</th>
<th>F1 score</th>
</tr>
</thead>
<tbody>
<tr>
<td>RoBERTaNER</td>
<td>0.441</td>
<td>0.906</td>
<td>0.940</td>
<td>0.874</td>
<td>0.565</td>
</tr>
<tr>
<td>TABLongformer</td>
<td>0.836</td>
<td>0.919</td>
<td>1.000</td>
<td>0.916</td>
<td>0.876</td>
</tr>
<tr>
<td>GreedyBERT</td>
<td>0.260</td>
<td>0.814</td>
<td>0.782</td>
<td>0.847</td>
<td>0.394</td>
</tr>
<tr>
<td>RandomBERT</td>
<td>0.263</td>
<td>0.668</td>
<td>0.530</td>
<td>0.806</td>
<td>0.377</td>
</tr>
</tbody>
</table>

7. Main takeaways

- Performance is dependent 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)