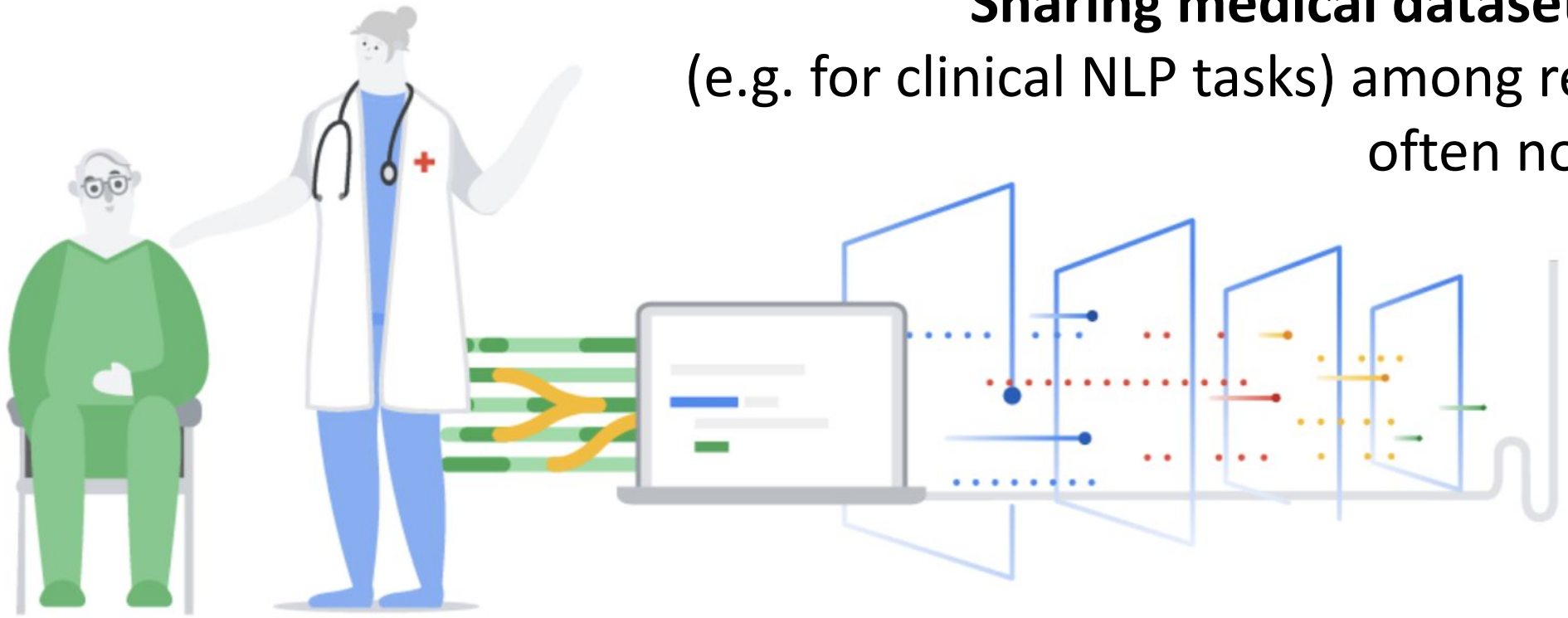
2nd Dutch meeting on Clinical NLP

Generating Artificial Electronic Health Records to Train De-Identification Models

Claudia Libbi, Jan Trienes, Dolf Trieschnigg, Christin Seifert

UNIVERSITY OF TWENTE. nedap UNIVERSITÄT DUISBURG ESSEN

Sharing medical datasets & models
(e.g. for clinical NLP tasks) among researchers is often not possible...



- **EHR contain highly privacy sensitive information**
- **It's hard to anonymize unstructured text (e.g. EHR) 100%**

Solution: Don’t use real data!

(ideally) it looks like EHR, it works like EHR, but it’s fake (in a good way)

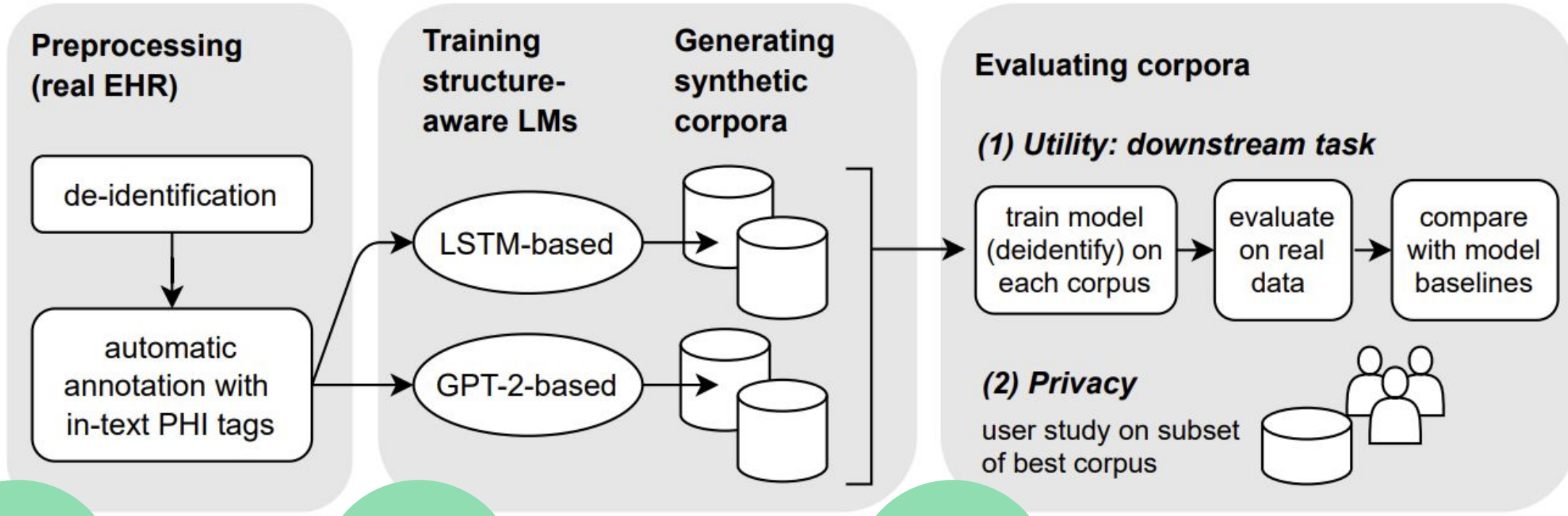
Utility + Privacy

Main contributions (spoiler)

1. Our LMs produced artificial text of **sufficient utility** to be used for training downstream ML models

2. We gained **insights into potential privacy threats** rel. generating synthetic EHR notes

Research Pipeline



1

2

3

1

Preprocessing real EHR

Data Sampling:
- ca. 1mio Dutch EHR
- from 39 customers
- ca. **52mio tokens**

- annotate PHI
- surrogate replacement (pseudonymization)

In-text annotations:
e.g. “<NameSTART> Eva <NameEND> had coffee...”

2 Training Structure-Aware LMs

Generating unstructured text (standard)

Prompt:
[Maria is meeting]

Model produces synthetic text:

Maria is meeting J.D. on January 5th.

Generating structured text (our approach)

Prompt:
[<NameSTART> Maria <NameEND>]

Model produces synthetic text with annotations:

<NameSTART> Maria <NameEND> is meeting <InitialsSTART> J.D <InitialsEND> on <DateSTART> January 5th <DateEND>.

2 Generating Artificial EHR (4 sets)

Dutch LSTM

PyTorch

→ ***Temperature Sampling***

→ ***(Nucleus) P-Sampling***

Model Architecture | *Decoding Algorithms*

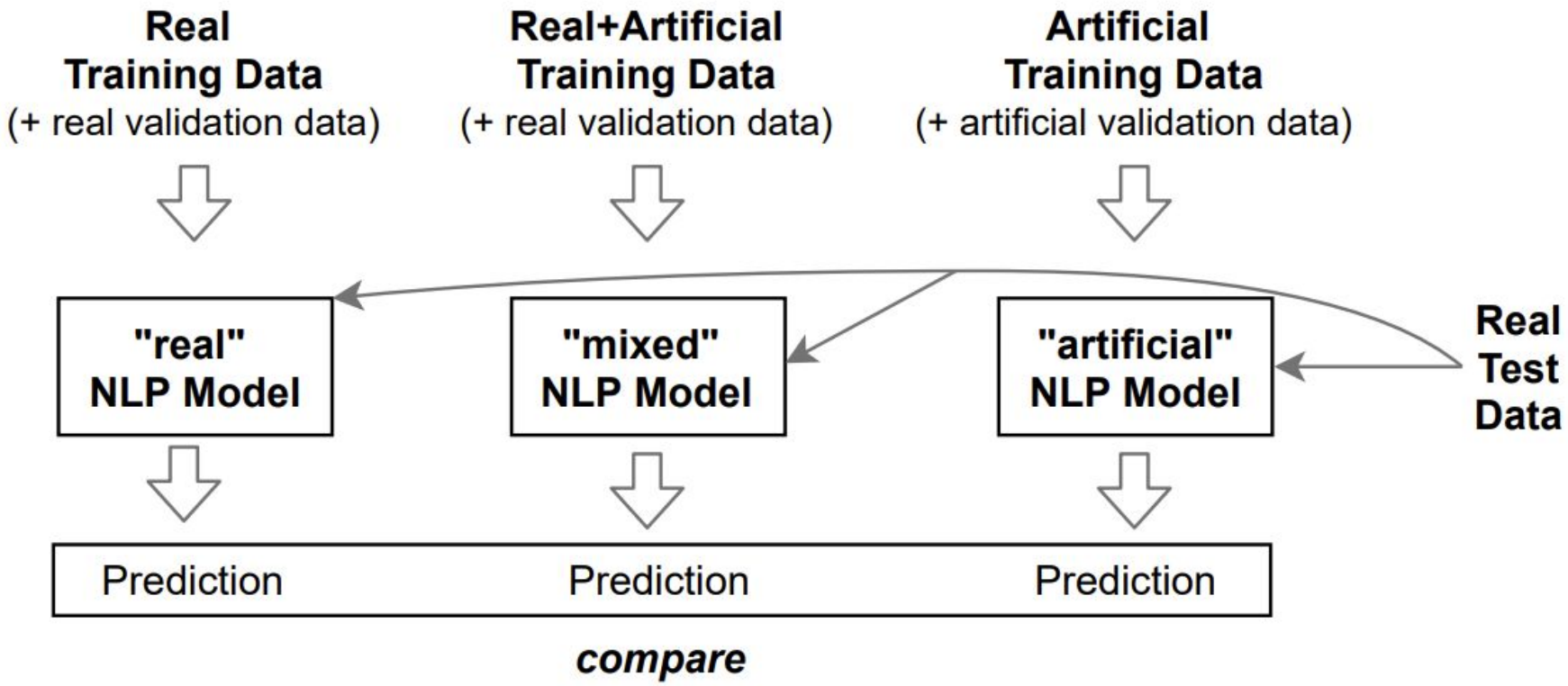
Dutch fine-tuned GPT-2

HUGGING FACE

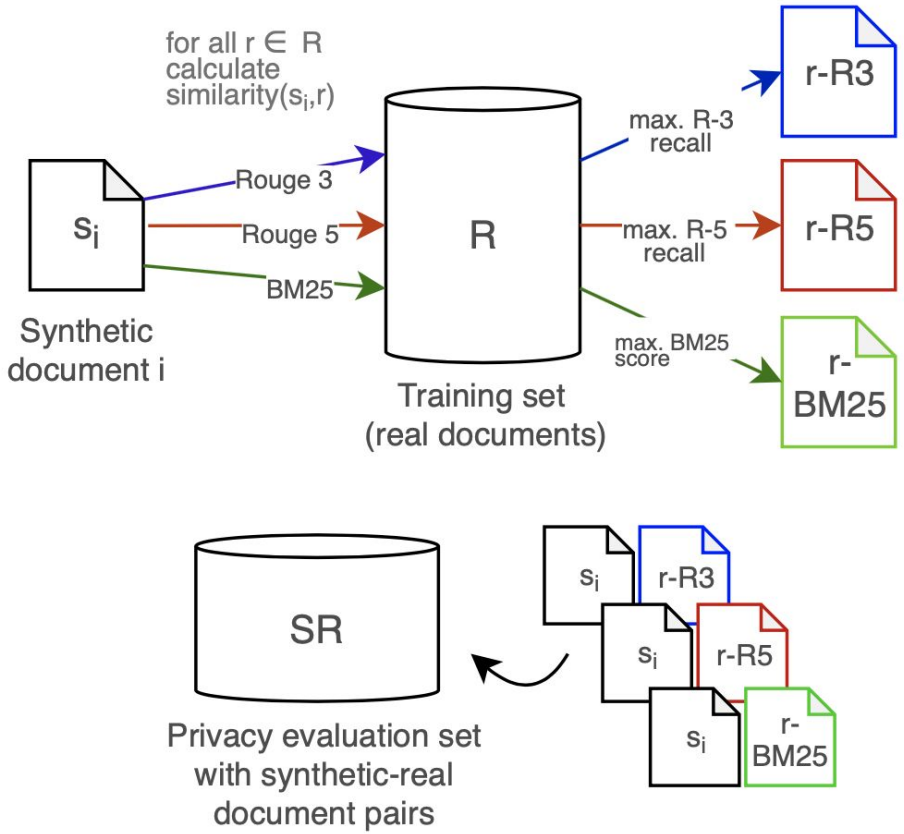
→ ***(Nucleus) P-Sampling***

→ ***Beam Search (max. PHI tags)***

3 Evaluating Utility (Downstream NER Task: De-identification)



3 Evaluating Privacy (Matching similar real-fake docs & user study)



- 122 doc-pair examples
- 12 annotators total (2 per doc-pair example)
- 5-pt. Likert-scale questions + explanation (free text)

Findings & Conclusions

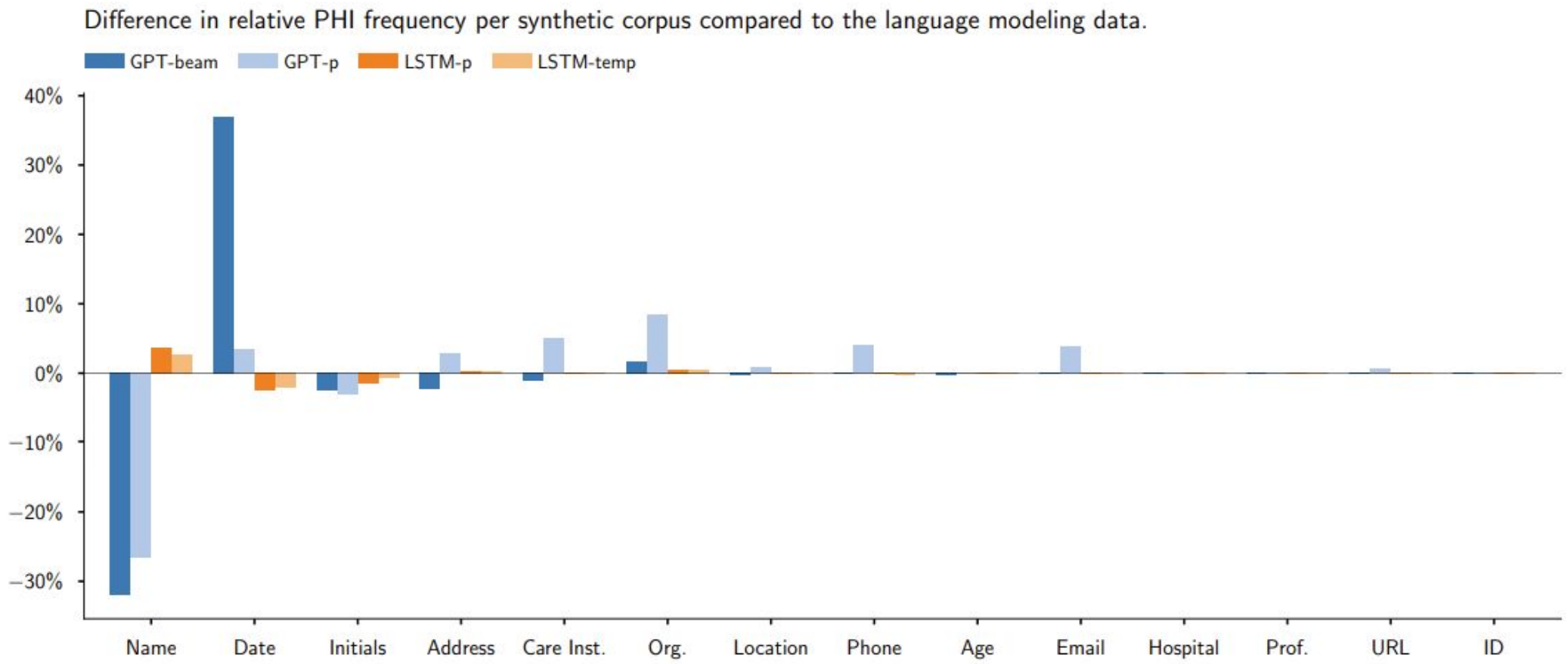
Do properties of fake EHR resemble real EHR?

	NUT [10]	LSTM-p	LSTM-Temp	GPT-p	GPT-Beam
Tokens	445,586	976,637	977,583	1,087,887	1,045,359
Vocabulary	30,252	23,052	29,485	12,149	8026
PHI instances	17,464	32,639	31,776	105,121	24,470
Sentences	43,682	70,527	72,140	128,773	83,634
Avg. tokens per sentence	10.2	13.8	13.6	8.4	12.5
Well-formed PHI tags		99.97%	99.89%	97.75%	98.84%
Malformed PHI tags		0.03%	0.11%	2.25%	1.16%

→ Not entirely...

→ We can generate **well-structured annotations**! It would be useful to control the **distribution…**

Quantitatively, the fake EHR are...



Qualitatively, the fake EHR are...

<NameSTART> J. Smith <NameEND> did a check. Dental hygiene is good and the dentures are clean. No abnormalities of the mucous membranes.

Which instruction did you give: to the nursing staff on the ward

Specifics and poss. action (s): check oral hygiene. Brush the dentures with water and soap. Please sleep without dentures and store dry. In case of no improvement, consult the nursing staff. Take care when brushing the dentures: be careful with oral care!

To whom have you instructed: (incl. names of the nurses) caregivers

Follow up action
Prevention ass. <NameSTART> A. Baker <NameEND>
Prevention ass <NameSTART> E. Williams <NameEND> oral care

Action ass. ass. from the department of the dental care <Care_InstituteSTART> The Care Home <Care_InstituteEND> for the dry mouth and the mouth of mister <NameSTART> D. Johnson <NameEND> , <Phone_faxSTART> 89-1234567 <Phone_faxEND>

→ GPT-2, beam search (good example)

→ resembles typical EHR template

But: majority of fake EHR were not coherent, especially LSTM-based.

Utility of the artificial EHR dataset is...

Split: Train/val/Test	Dataset	Precision	Recall	F1
-/-/real	NUT (rule-based) [30]	0.807	0.564	0.664
real/real/real	NUT (BiLSTM-CRF) [10]	**0.925**	0.867	0.895
Use case 1: synthetic data as a replacement for real data				
synth/synth/real	LSTM-p	0.835	0.784	0.809
synth/synth/real	LSTM-temp	0.857	0.773	0.813
synth/synth/real	GPT-p	0.776	0.700	0.736
synth/synth/real	GPT-beam	0.823	0.688	0.749
Use case 2: synthetic data as data augmentation method				
real+synth/real/real	NUT+LSTM-temp	0.919°	**0.883**▲	**0.901**°
real+synth/real/real	NUT+LSTM-p	0.916°	0.879▲	0.897°

Findings: Utility

Case 1 - replacement for real data

Better than rule-based model (case: no real training data available) on real data, but not yet practical for de-identification application.

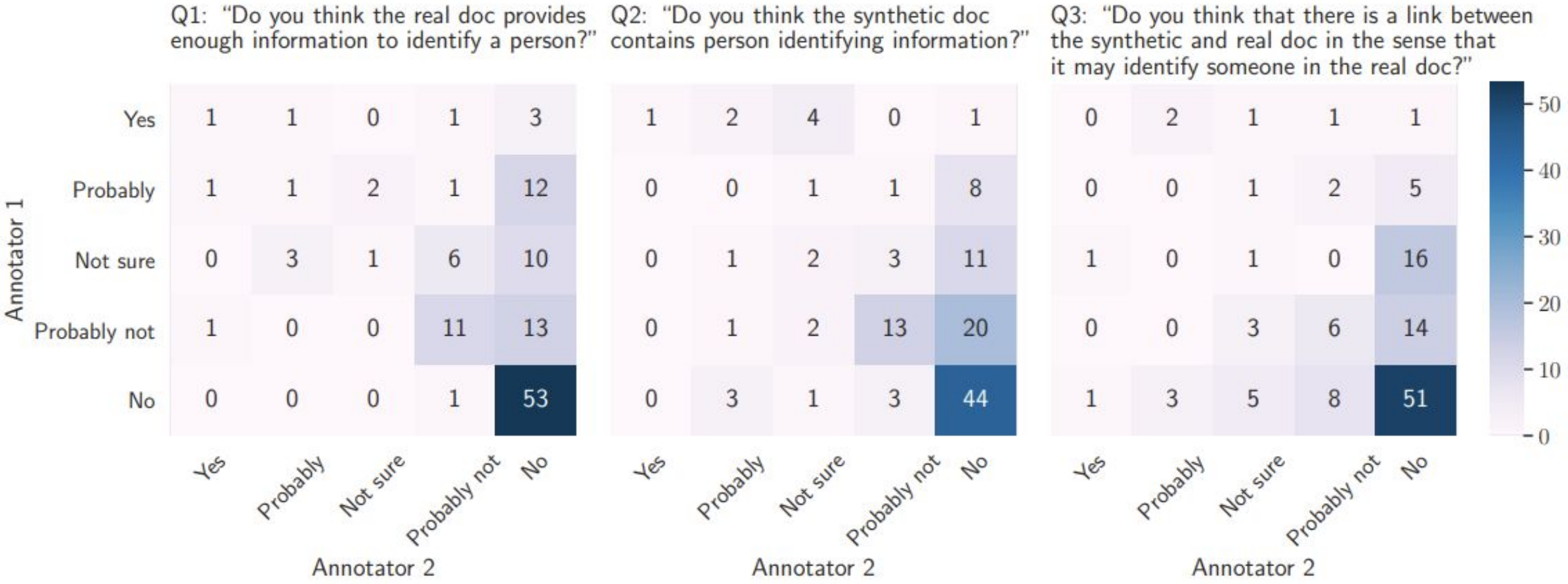
Case 2 - data augmentation to generate cheap additional training examples

Sufficient utility, benefits in case of LSTM-based data reg. recall!

Also:

1. **Greater text diversity** is beneficial for downstream task performance!
2. **Not very coherent/medically correct artificial EHR** not necessarily an issue for downstream task, syntactic correctness more important!
3. **GPT2 covers more PHI types, LSTM performs better on common PHI types**

How users judged privacy ...



Findings: Privacy

1. **Most fake EHR were not similar to original EHR they were paired with**
2. **Removal of PHI in free text not always sufficient to protect privacy** (case: specific, rare events in detail)
3. **Mediocre text-quality as protective factor** by obfuscating what is real and what is fake
4. **Larger chunks of text copied from real data** (especially in case of rare events) is concerning

Thank you! :)

ing Artificial alth Records ation Models

nes, Dolf Trieschnigg, Christin Seifert

SITY NTE.

nedap

UNIVERSITÄT DUISBURG ESSEN

future internet

MDPI

Article

Generating Synthetic Training Data for Supervised De-Identification of Electronic Health Records

Claudia Alessandra Libbi [1,2], Jan Trienes [2,3,*], Dolf Trieschnigg [2] and Christin Seifert [1,3]

1 Faculty of EEMCS, University of Twente, PO Box 217, 7500 AE Enschede, The Netherlands; alelib29@gmail.com (C.A.L.); christin.seifert@uni-due.de (C.S.)
2 Nedap Healthcare, 7141 DC Groenlo, The Netherlands; dolf.trieschnigg@nedap.com
3 Institute for Artificial Intelligence in Medicine, University of Duisburg-Essen, 45131 Essen, Germany
* Correspondence: jan.trienes@uni-due.de

Abstract: A major hurdle in the development of natural language processing (NLP) methods for Electronic Health Records (EHRs) is the lack of large, annotated datasets. Privacy concerns prevent the distribution of EHRs, and the annotation of data is known to be costly and cumbersome. Synthetic