Comparing Rule-based, Feature-based and Deep Neural Methods for De-identification of Dutch Medical Records

Jan Trienes, Dolf Trieschnigg, Christin Seifert, Djoerd Hiemstra

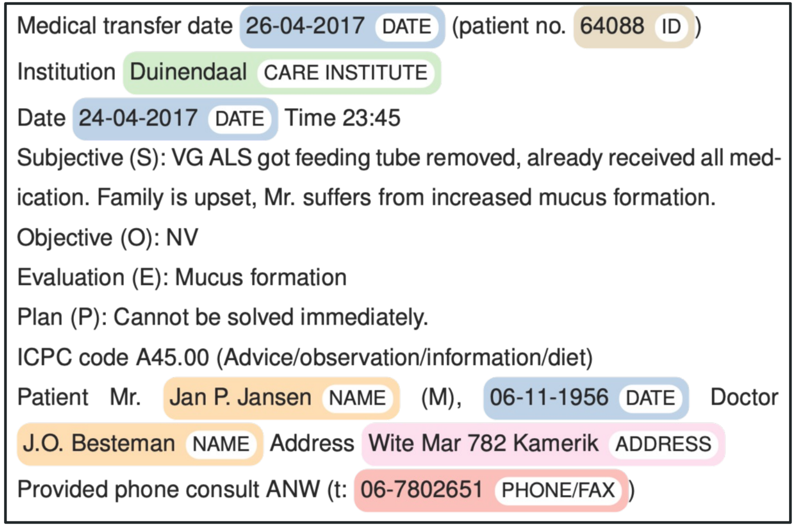
Health Search and Data Mining | Houston, February 2020

UNIVERSITY OF TWENTE.
nedap
Radboud University

Text de-identification phrased as information extraction task

Medical transfer date 26-04-2017 (patient no. 64088)
Institution Duinendaal
Date 24-04-2017 Time 23:45
Subjective (S): VG ALS got feeding tube removed, already received all medication. Family is upset, Mr. suffers from increased mucus formation.
Objective (O): NV
Evaluation (E): Mucus formation
Plan (P): Cannot be solved immediately.
ICPC code A45.00 (Advice/observation/information/diet)
Patient Mr. Jan P. Jansen (M), 06-11-1956 Doctor J.O. Besteman Address Wite Mar 782 Kamerik
Provided phone consult ANW (t: 06-7802651)

1 Detect



2 Mask, remove or replace with realistic surrogates

3 Use de-identified data for purpose

Applications

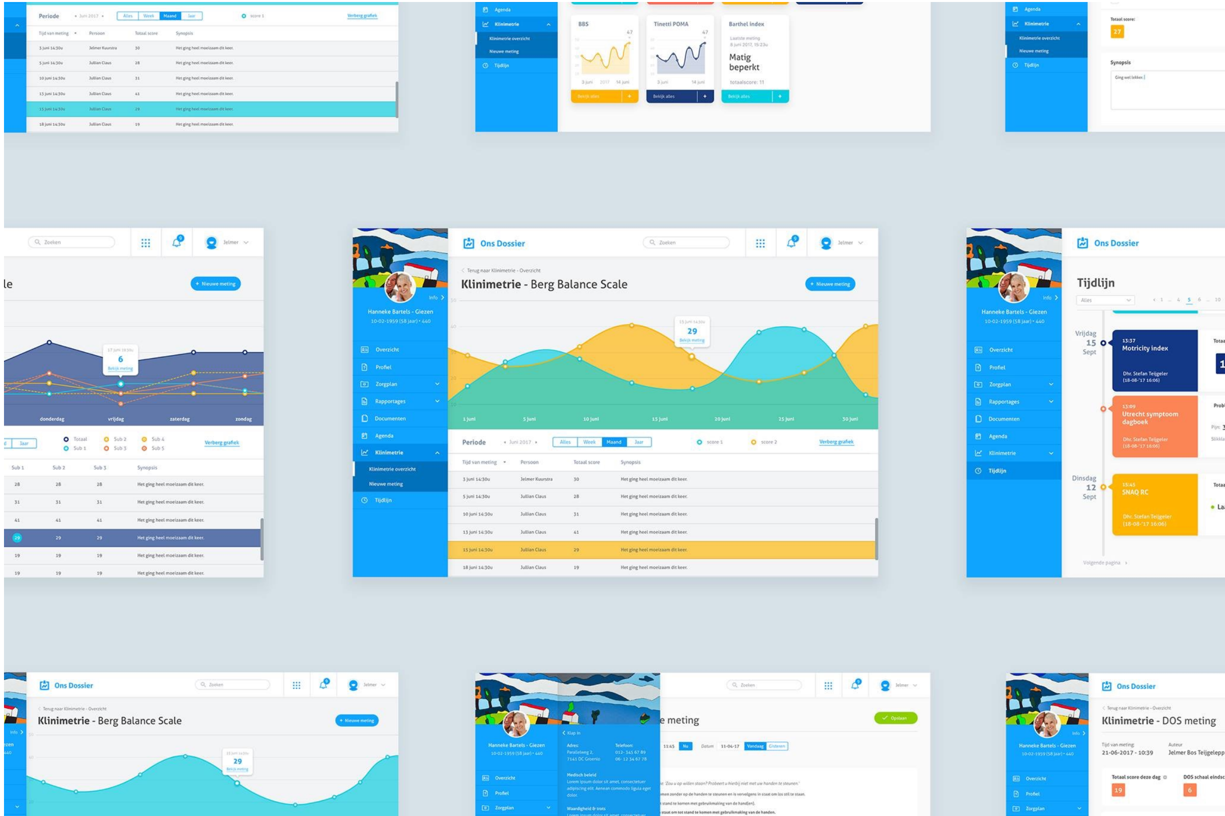
Data anlysis

Research

Customer support

Development

UX Design



Challenge 1:

Lack of openly-available de-identification resources

Challenge 2:

How do methods generalize to new domains and languages?

Comparing methods for de-identification of medical records

- Dataset and methods
- Dutch de-identification
- Generalizability

We construct a heterogeneous dataset by sampling from EHRs of multiple care domains

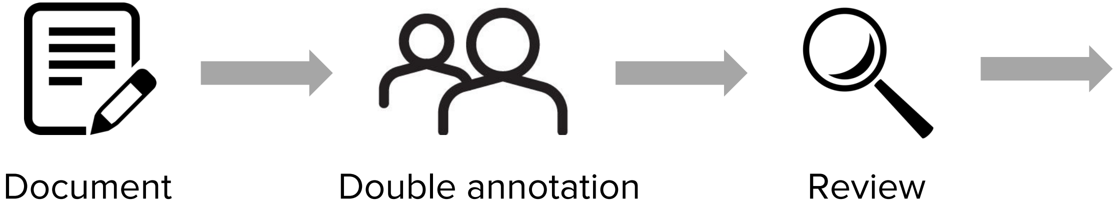
9 organizations across
different care domains
Elderly, mental, disabled

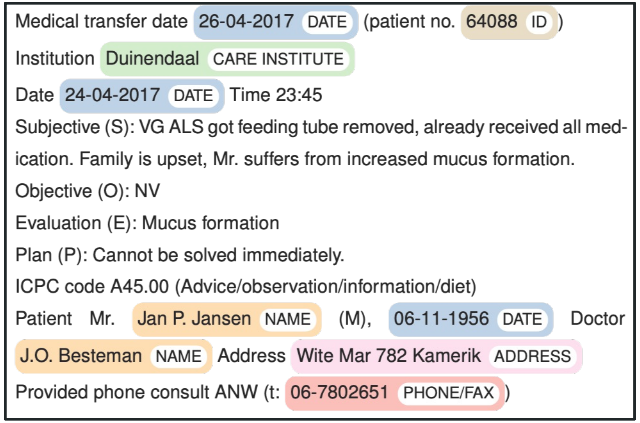
2 document types
Surveys & medical reports

Sample

1260 documents
450k words

We also need examples of protected health information




Medical transfer date 26-04-2017 DATE (patient no. 64088 ID)
Institution Duinendaal CARE INSTITUTE
Date 24-04-2017 DATE Time 23:45
Subjective (S): VG ALS got feeding tube removed, already received all medication. Family is upset, Mr. suffers from increased mucus formation.
Objective (O): NV
Evaluation (E): Mucus formation
Plan (P): Cannot be solved immediately.
ICPC code A45.00 (Advice/observation/information/diet)
Patient Mr. Jan P. Jansen NAME (M), 06-11-1956 DATE Doctor J.O. Besteman NAME Address Wite Mar 782 Kamerik ADDRESS
Provided phone consult ANW (t: 06-7802651 PHONE/FAX)

12 annotators
17,500 annotations in 1260 docs.
80h annotation + 20h review = 12.6 docs/h

We compare three recent de-identification methods

1 DEDUCE

Pattern matching & heuristics
Developed on clinical text

2 Conditional Random Field

Feature-engineering
Semantic, syntactic and orthographic features

3 BiLSTM-CRF

Generic sequence-labeling architecture
Pre-trained contextual string embeddings

[1] Menger V., et al. (2018). DEDUCE: A pattern matching method for automatic de-identification of Dutch medical text.
[2] Liu Z., et al. (2015). Automatic de-identification of electronic medical records using token-level and character-level conditional random fields.
[3] Akbik A., et al. (2018). Contextual string embeddings for sequence labeling

Comparing methods for de-identification of medical records

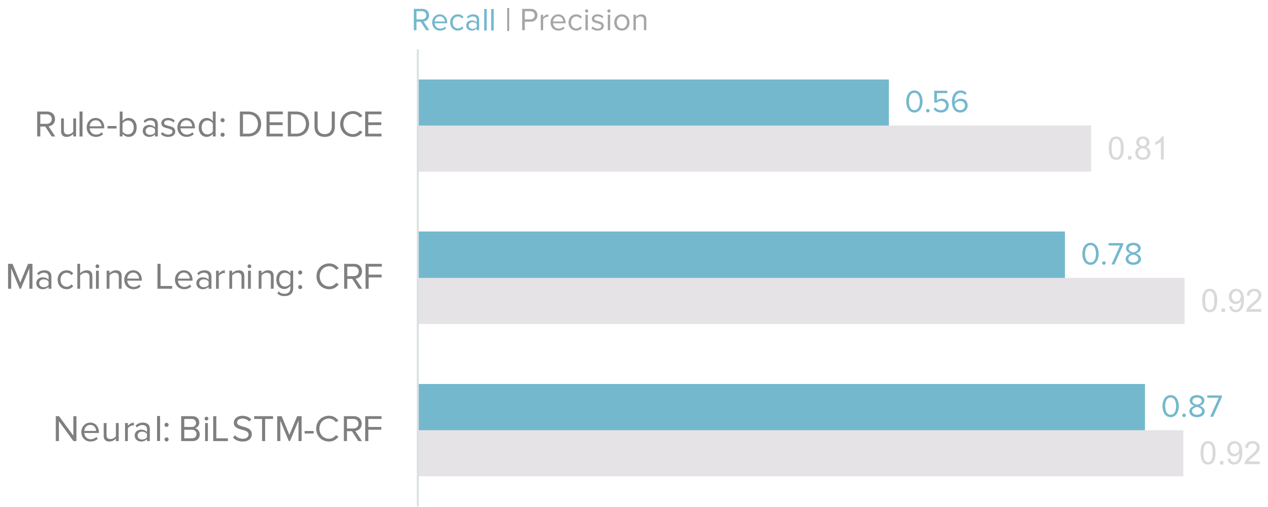
Dataset and methods

Dutch de-identification

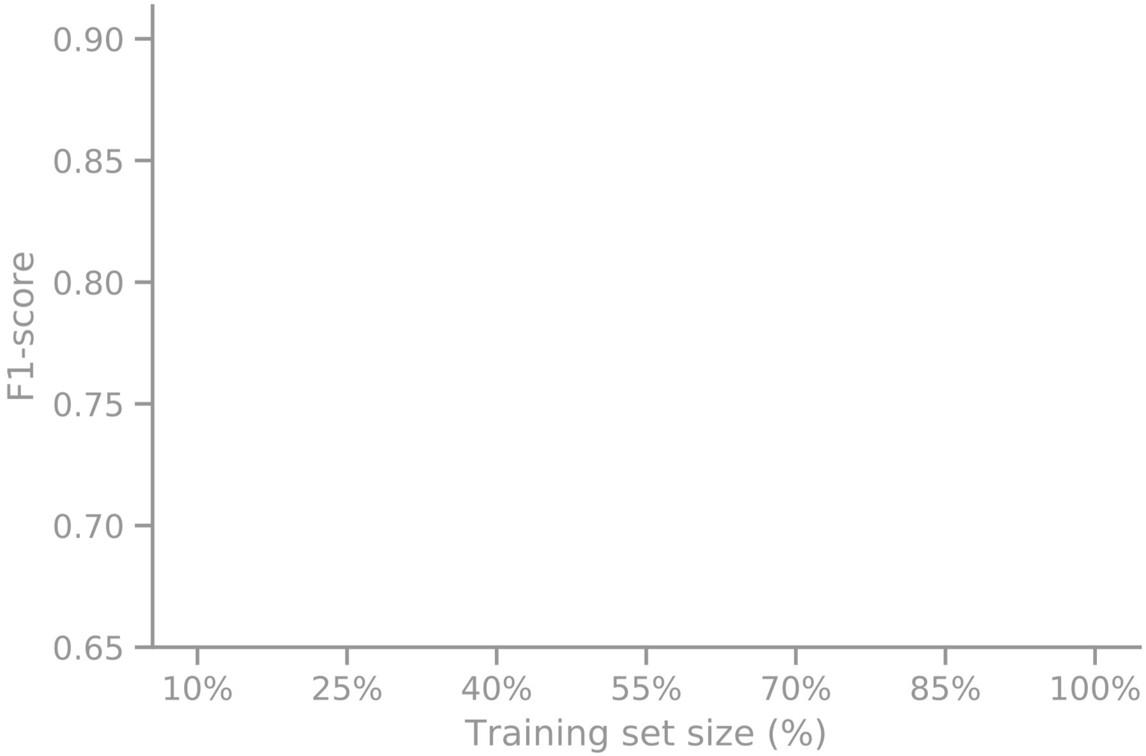
Generalizability

Neural method is most effective

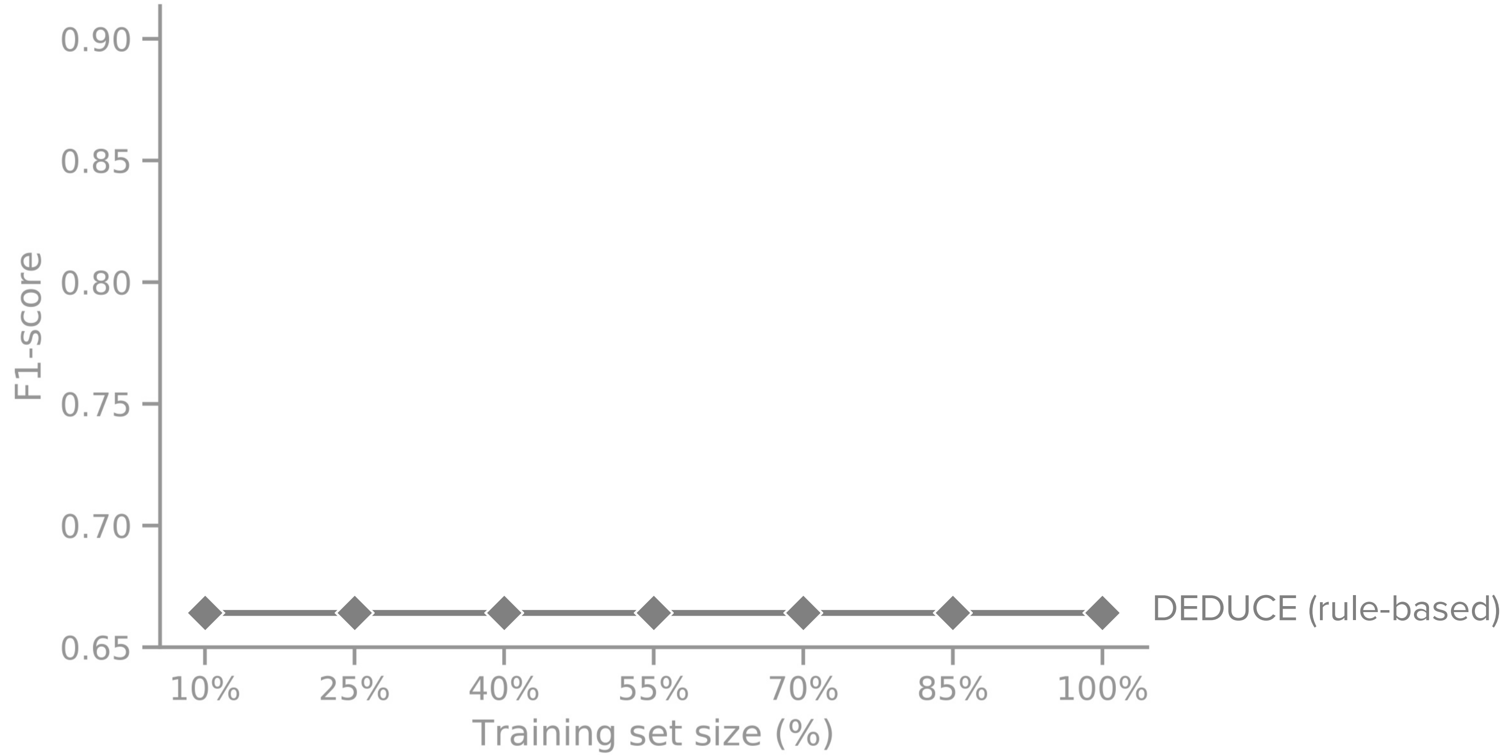
Rule-based method does not generalize to new dataset



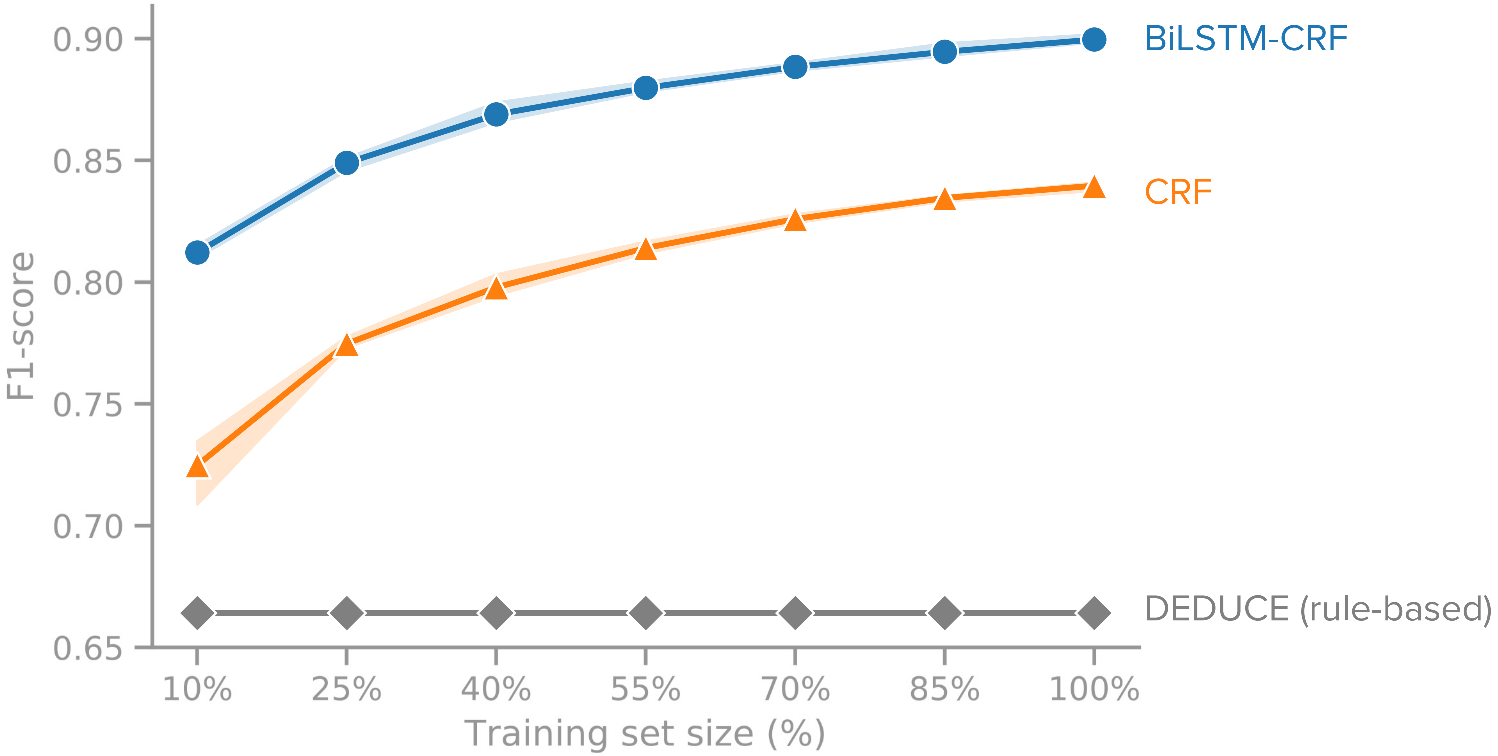
Neural method superior even with limited training data



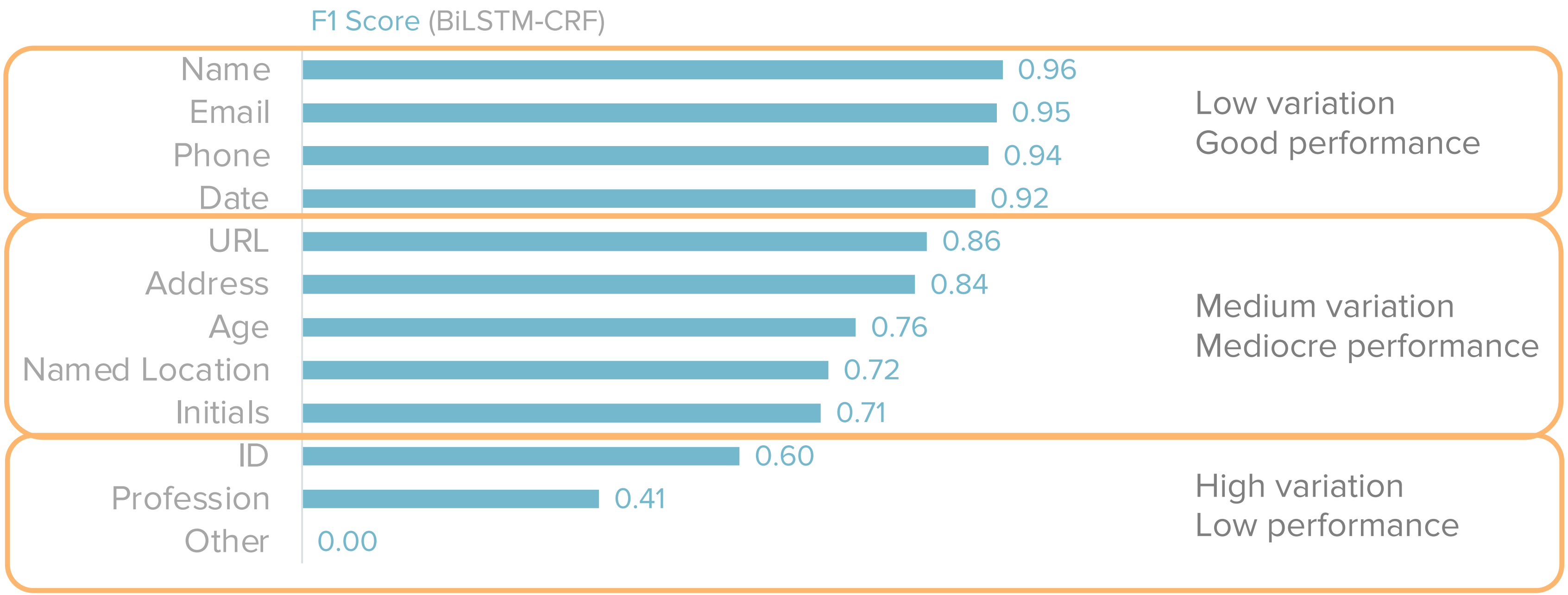
Neural method superior even with limited training data



Neural method superior even with limited training data



Sensitive information with high variation is hard to capture



Sensitive information with high variation is hard to capture

Common language

“works behind the cash register”	instead of	“cashier”
“halfway to the eighty”	instead of	“75 years”

IDs

176, 78449083, 354LO

Is this an ID, measurement or medical code?

Other category

The airing of her appearance in NBC late night makes her feel…

Comparing methods for de-identification of medical records

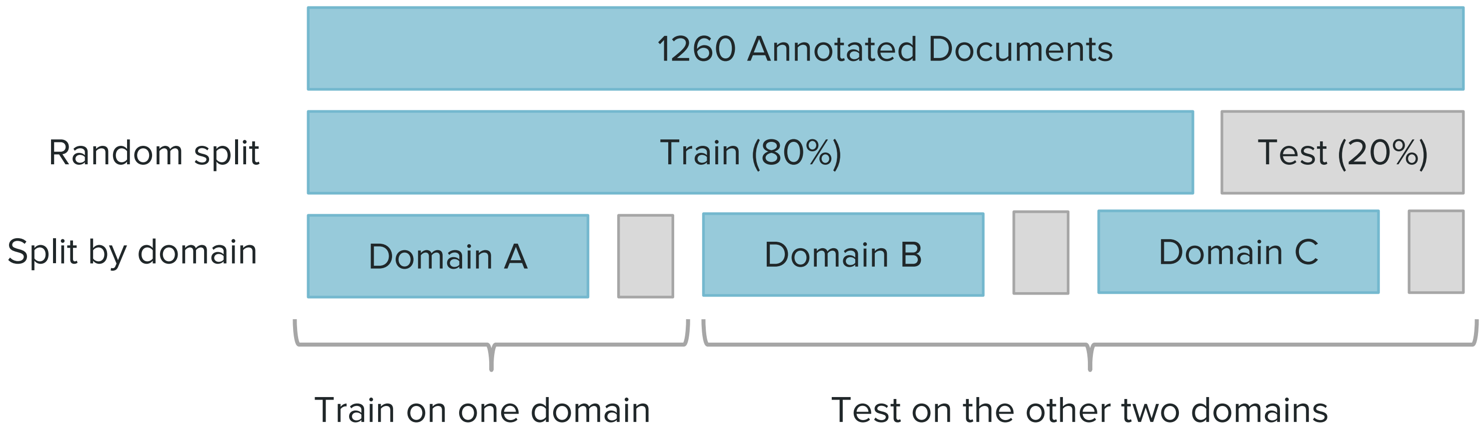
Dataset and methods

Dutch de-identification

Generalizability

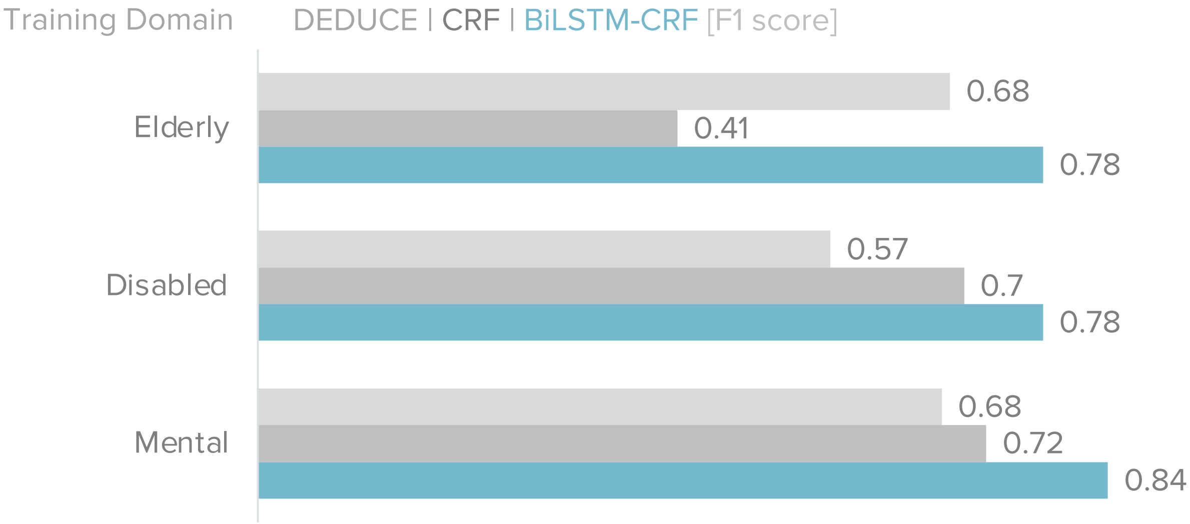
How do the methods generalize to new domains?

We split Dutch data by domains



Neural method generalizes best to new domains

Rule-based has stable performance

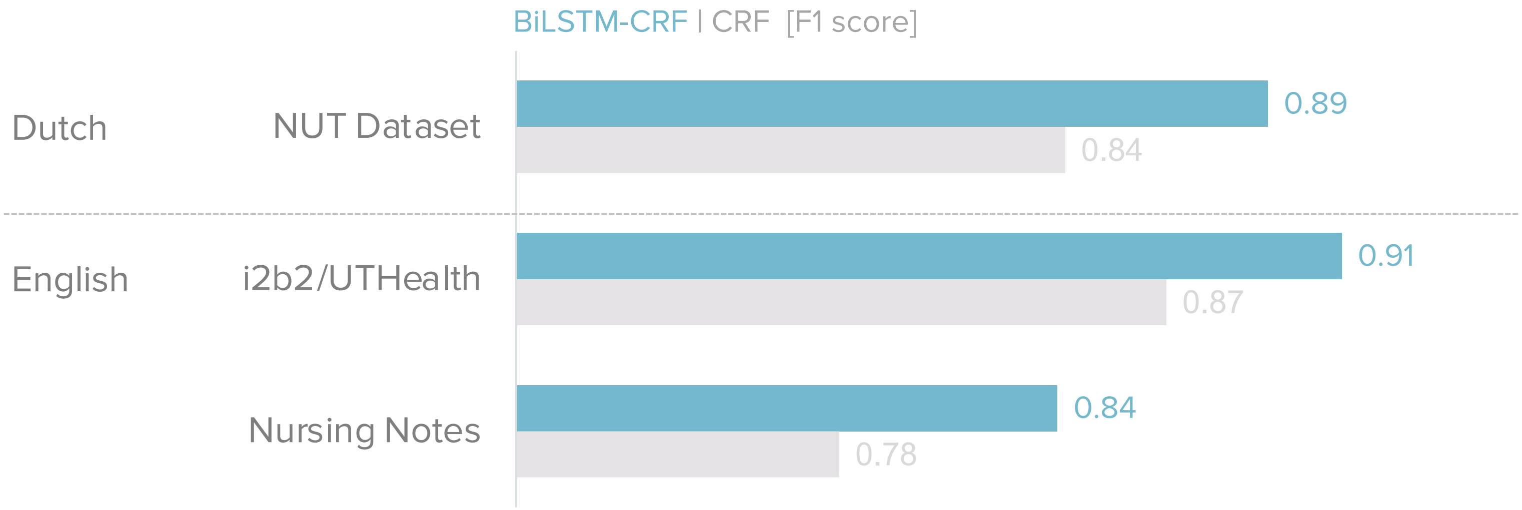


Rule-based system outperforms feature-based CRF

Neural method generalizes best to new domains

But: effectiveness is mediocre

Across datasets neural method is also most effective



Wrap up

Conclusion

- Rule-based method least effective on new data
- Neural method is a good default (even with limited data)
- Effectiveness substantially differs across domains

Future work

- Improve generalizability: transfer learning
- Combine rule-based and machine learning methods
- How to capture sensitive information with high variation?

Conclusion

We share code and pre-trained models with the community.

github.com/nedap/deidentify
jan.trienes@nedap.com

nedap
UNIVERSITY OF TWENTE.
Radboud University