Data output

* 1. **What data should be de-anonymized at the end of the processing pipeline?**

The C2R pipeline

* 1. **What requirements are needed to ensure long term future use of the developed pipeline**
  2. **What model is best suited to be used as Named Entity Recognition model for the Dutch and English language?**
  3. **How can fast and asynchronous computation be accomplished?**
  4. **How can privacy be ensured throughout the process?**
  5. **What placeholders should be used after NER processing?**
  6. **What entities should stay de-identified after NER processing and de-identification has taken place?**
  7. **How should eponymous diseases be detected and handled during the de-identification steps?**

Eponymous diseases

1.5 What database should be used to gather a list of eponymous diseases?

1.6 How should eponymous diseases be handled in the NER and placeholder replacement process?

NER models

1.8 How should the models be contained? – huggingface

# Introduction

# Literature review

Terminology used:

|  |  |
| --- | --- |
| C2R | Care2Report |
| ASR | Automated Speech recognition |
| NER | Named Entity Recognition |
| IS | Information System |
| (NE-)WER | (Named Entity-) Word Error Rate |
| NLP | Natural Language Processing |
| LM | Language Model |

Care2Report aims to utilize different types of automated detection mechanisms such as speech recognition and video action recognition in order to automate the process of medical reporting for medical professionals. The processed data is not anonymous which could pose a vulnerability and introduce privacy concerns among its users and therefor endanger the adoptability of the system (<https://link.springer.com/content/pdf/10.1007%2Fs10916-013-9966-z.pdf>).

The aim of this paper is to design a time-proof, model dynamic, edge-case detecting supplementary pipeline that is able to anonymize data to a sufficient degree where it can be used for further processing and does not risk leaking personal data to external servers or entities. To fulfill earlier mentioned requirements several external ready-to-use technologies are used.

Our idea is to utilize a number of different ready-to-use pipelines and models to successfully de-identify and eventually anonymize multilingual audio transcriptions. The audio transcription process is not included into this pipeline. The current speech processing pipeline as deployed in the current C2R speech module is the following:

All steps are performed on personalized data (red) which is not desirable. The proposed pipeline following from this paper is the following.

Only audio pre-processing and transcription of audio are performed on personalized data. All further data processing is performed on de-identified data (green). After report generation, the placeholders that were implemented at the de-identification step can be replaced back to the original entities. This results in a speech processing pipeline that does the actual data processing on anonymized data which ensures possible privacy issues are omitted.

To successfully create such a pipeline a number of questions need to be answered. First of which is, how should such a pipeline be developed to ensure future, long-term use.

This pipeline can be looked upon as an information system, as in “*Information systems for managers: with cases*” [1] was stated that an information system is a formal, sociotechnical, organizational system designed to collect, process, store and distribute information. The system receives, processes and stores certain pieces of information whereafter that information is re-distributed back to the original C2R system. Information systems can be designed using several design principles. Because this system will be applied withing a medical context, prevention of failure should be the main concern, *Patterns for Fault-Tolerant Software* [2]discusses several metrics and patterns to design an error resistant system. In chapter Three of the book, Real-Time processing systems are discussed. The book sets the premise that a superset of the concepts mentioned in the book are needed to successfully develop a fault-tolerant so-called hard real-time system. The concepts mentioned in the book are:

* Coverage
* Reliability
* Availability
* Dependability
* Hardware Reliability
* Reliability Engineering and Analysis
* Performance

Current other challenges that are being looked into within the Care2Report research project that could interfere with the future use of this pipeline are:

* 1. **How should Named Entity Recognition be applied onto incoming datastreams?**

**NER and ASR together.**

**NER trained on ASR.**

**NER trained loosely from ASR.**

**Language detection requires different ASR NER models.**

**ASR NER models for Dutch are scarce.**

**ARS model and different NER models are not scarce.**

Named entity recognition is the process of identifying named entities in unstructured text through trained models. These models are often trained using supervised learning. The researcher feeds large amount of labelled data into the model which are used to train and improve the recognition process of certain types of entities such as a person, location, organization. The model then tags these entities in the text using entity types. These can be outputted and stored loosely. A lot of literature using ASR (Automated Speech Recognition) in combination with NER uses these components separately. The usual ASR NER pipeline consists of two different steps. An ASR model is trained to recognize the speech data. Then the output is used to perform NER on [3]. The NER model is trained on this initial output data or different transcription data. Most of the times this training data consists of perfect audio transcriptions which will never be the case. ASR has a certain degree of error embedded in its output. This error rate will be propagated into the NER, resulting in an even higher overall error rate since NER models are not trained to detect noise or ASR errors within text [4] but rather only detect NE in it [5]. Combining ASR and NER into a single pipeline could prove to be a solution. This was researched in [6]. [6] proposes a pipeline in which NER is performed directly on speech data. The upside to this is that the NER WER is lower. However, established pretrained models then cannot be used as transcription model. This would also mean audio has to be processed by the pipeline instead of textual data which is easier to split up in a per-sentence processing module.

* 1. **What model is best suited to be used as natural language detection model?**

e transcription de-identification and re-identification is contained in two separate steps. The first of the two removes personalized entities from the data and stores those that should be remembered in a database. This data is then used in the seconds step to make the report identifiable again.

## Anonymization

[1] – Language detection

EXPLAIN WHAT THIS PAPER SOLVES

EXPLAIN HOW IT SOVES THIS

EXPLAIN THE STEPS CONTAINED INSIDE THE SOLUTION AND PROVIDE SCIENTIFIC SUPPORT FOR EACH SOLUTION.

# Results

# Conclusion

# Discussion

Care2Report aims to utilize different types of automated detection mechanisms such as speech recognition and video action recognition in order to automate the process of medical reporting for medical professionals. The processed data is not anonymous which could pose a vulnerability and introduce privacy concerns among its users and therefor endanger the adoptability of the system (<https://link.springer.com/content/pdf/10.1007%2Fs10916-013-9966-z.pdf>).

The GDPR states that special categorical data such as “data concerning health” needs to be processed safely and securely. To minimize leaks and vulnerabilities, data should be processed anonymously. This will also allow to further process the data as anonymous data does not fall under the “special category data” which is described in the GDPR. It is actually completely outside of the scope of EU data protection laws. Which allows for far more extensive processing and use.

Therefore, data anonymization needs to be applied throughout different steps of the process. This paper will discuss different requirements and technologies of such a de-identification pipeline. The basis for implementing such a de-identification pipeline is to fit it into the current dataflow or processing pipeline.

The Care2Report processing pipeline for speech is as follows:

Personalized data is utilized in every step of this process. To minimize data exposure, de-identification can be applied. Using the method described in this paper, the amount of steps processing identifiable data will be reduced to two out of five. The result will be as follows:

Both identification steps (yellow) contain a number of steps which are performed on each of the sentences that are coming from the speech transcription and report generation step. The de-identification is performed after the transcription has been made. The re-identification is performed after the final report is generated.

|  |  |
| --- | --- |
| **De-identification** | **Re-identification** |
| 1. Detect language. 2. Detect and classify entities contained in the sentence. 3. Re-Classify detected entities. 4. Replace entities in sentence. | 1. Replace de-identified entities with actual entities. |

The aim of the pipeline is to be:

1. Multi-Lingual, to make implementation as dynamical as possible and future proof.
2. Dynamically implementable and fast, to allow for a wide variety of implementation possibilities (externally/locally).
3. Fast, data processing should be performed instantly to allow for real-time processing.
4. Model interchangeable, to allow for future improved and more efficient NER models being implemented into the solution quickly and easily.
5. Two-way, after de-identifying the data. Preset entities can be restored to re-identify the data.
6. Minimally data retaining, to ensure privacy data should only be retained during processing and be discarded when not needed.

### Multi-lingual

The current C2R system is not multi-lingual. However, the aim is to make the system multi-lingual in the future. In order to support that goal, it is vital to make sure new components will be supporting multi-langual pipelines.

Lui, Marco and Timothy Baldwin (2011) Cross-domain Feature Selection for Language Identification, In Proceedings of the Fifth International Joint Conference on Natural Language Processing (IJCNLP 2011), Chiang Mai, Thailand, pp. 553—561. Available from <http://www.aclweb.org/anthology/I11-1062>

### Dynamically implementable

### Fast

https://archive.arisuchan.jp/%CE%BB/src/1498631786854-1.pdf

### Model interchangeable

https://hal.archives-ouvertes.fr/file/index/docid/843211/filename/hatmi.pdf

<https://hal.archives-ouvertes.fr/hal-01987740/document>

<https://arxiv.org/pdf/1805.12045.pdf>

<https://www.researchgate.net/profile/Sophie-Rosset/publication/319185089_Investigating_the_Effect_of_ASR_Tuning_on_Named_Entity_Recognition/links/5ae6dab90f7e9b9793c7ddc3/Investigating-the-Effect-of-ASR-Tuning-on-Named-Entity-Recognition.pdf>

Dutch: GroNLP/bert-base-dutch-cased

English: bert-base-uncased

### Two-way

### Minimally data retaining

* Implementable into the current text pipeline with sufficient security and deployment options.
* Extreme edgecase handling?
* Lightweight.
  + Limit database hits
* Realtime processing possible
* Both way processing possible 🡪
  + First anonymize
  + Then de-anonymize
* Secure
  + All local, both runner and database.
  + Limited information storage.
  + Active removal of non-needed instances/entities
* Future proof
* Model interchangeable
* Dynamical input and output options
* Storing entities that should be de-anonymized. Rest can be deleted/should not be retained.
* DataStream?
* Classification of Entities into categories
  + Client - data
  + Unknown Entity - data
  + Eponymous disease / medical terms - data
  + Caregiver – data

Important factors:

* Placeholder output 🡪 standardized? Impact on actual summarization.
* Delivery and output methods 🡪 impact on speed.
* Handling of undetected Names Entities 🡪 can be ignored, because context is sufficiently changed?
* Eponymous disease

Why is it necessary to solve this issue?

How are we going to solve this issue?

Technologies user:

* Postgres database
* Python
* Huggingface model repository
* Snomed CT disease database
* C2R pipeline

Research question:

How should audio transcription data be anonymized before and after automated summarization to resolve privacy concerns regarding data processing?

Sessionid assigned

|

Language detection

|

NER processing using preset models (huggingface)

|

Placeholder replacement

|

Output cleaned text

|

…..

|

Input sentences from processing tasks

|

Check for placeholders in output and de-anonymize

|

Output de-anonymized text