Creating and Evaluating a Synthetic Norwegian Clinical Corpus for De-Identification

Synnøve Bråten
Department of Computer
and Systems Sciences
Stockholm University
Kista, Sweden
synnovebr@hotmail.com

Wilhelm Wie
Department of Computer
and Systems Sciences
Stockholm University
Kista, Sweden
w.wie@gmx.com

Hercules Dalianis
Department of Computer
and Systems Sciences
Stockholm University
Kista, Sweden
hercules@dsv.su.se

Abstract

Building tools to remove sensitive information such as personal names, addresses, and telephone numbers - so called Protected Health Information (PHI) - from clinical free text is an important task to make clinical texts available for research. These de-identification tools must be assessed regarding their quality in the form of the measurements precision and recall. To assess such tools, gold standards - annotated clinical text - must be available. Such gold standards exist for larger languages. For Norwegian, however, there are no such resources. Therefore, an already existing Norwegian synthetic clinical corpus, *NorSynthClinical*, has been extended with PHIs and annotated by two annotators, obtaining an inter-annotator agreement of 0.94 F_1-measure. In total, the corpus has 409 annotated PHI instances and is called *NorSynthClinical PHI*. A de-identification hybrid tool (machine learning and rule-based methods) for Norwegian was developed and trained with open available resources, and obtained an overall F_1-measure of 0.73 and a recall of 0.62, when evaluated using *NorSynthClinical PHI*. *NorSynthClinical PHI* is made open and available at Github to be used by the research community.

1 Introduction

The data contained within Electronic Health Records (EHRs) are of significant value to medical researchers and for administrative purposes, but privacy and patient confidentiality legislation restricts access. However, de-identification of such data – removing the Protected Health Information (PHI) within – allows it to be shared between researchers (El Emam et al., 2009). This process can be done manually; however, manual de-identification has proven to be inefficient with regards to cost, time and quality (Dernoncourt et al., 2017).

Tools for automatic de-identification of clinical data have been studied extensively. However, most of the published research is concerned with structured records and not clinical free-text, and few de-identification tools are made publicly available (Neamatullah et al., 2008). Furthermore, most research focus on English and other languages with many native speakers. Despite the fact that the Norwegian language has comparatively few native speakers[1], hospitals and organisations like the Cancer Registry of Norway are in possession of comprehensive collections of clinical data. Enabling research on this valuable and unique information could reveal new discoveries and would be of great importance for the future health care.

To ensure that de-identification applications can successfully de-identify clinical texts, they must be evaluated in a quantitative manner (Dalianis, 2018). For this purpose, verified, annotated corpora are used to test and score the applications (Pustejovsky and Stubbs, 2012). These corpora are referred to as gold standards (or reference standards), and are typically made by domain experts or linguists - following specific guidelines. A gold standard does not need to contain real PHI, and it can be developed using synthetic data. Consequently, a gold standard developed with synthetic data can be made publicly available.

This study describes the efforts of creating and evaluating the first publicly available gold standard for de-identification of Norwegian Bokmål[2] clinical text, describing and discussing the devel-

[1] Norwegian has approximately 4,320,000 native speakers, (Rehm and Uszkoreit, 2012)

[2] Norwegian Bokmål - One of the two official written variants of Norwegian.

opment and evaluation of the gold standard.

2 Related research

Marimon et al. (2019) created a gold standard corpus of Spanish synthetic clinical text. The corpus is called *Spanish Medical Document Anonymization (MEDDOCAN)* and consists of 250 clinical cases manually enriched with PHI phrases. The gold standard was applied in a community challenge track in order to evaluate the performance of de-identification tools focusing on the Spanish language. 63 systems were evaluated and 61 received an F_1-measure score above 0.70, and the highest score was 0.97. As the gold standard seems to have served its purpose, Marimon et al. (2019) provides a good example of how to solve data sparsity problems.

The lack of publicly available clinical text in Norwegian places limitations on the development of gold standards and tools for de-identification of Norwegian clinical text. Recently, there have been developments of open datasets for Named Entity Recognition (NER) of the Norwegian language, most notably *NorSynthClinical* (Rama et al., 2018) and *NorNE* (Jørgensen et al., 2020). *NorSynthClinical* is a small dataset of synthetic clinical text, focusing on family history information (further described in Section 3) (Rama et al., 2018). While the development of *NorNE* resulted in a sizeable dataset with approximately 300,000 tokens for each written variant of Norwegian and a rich entity set, most PHI entity types are missing (Jørgensen et al., 2020).

Only a few attempts aiming at developing de-identification tools focusing on the Norwegian language have previously been made. One of these was conducted by Bjurstrøm and Singh (2013). They tackled de-identification of Norwegian free text clinical notes for their master's thesis project, employing a combination of pattern recognition and simplistic statistical methods, reporting an F_1-measure of 0.72. Furthermore, they developed a reference in order to evaluate their developed tool, consisting of 225 records manually annotated and de-identified. It was, however, not evaluated further or made publicly available (Bjurstrøm and Singh, 2013).

As previously mentioned, most of the existing tools and gold standards for de-identification of clinical text are written in, and for, the English language (Dalianis, 2018). One of the most well-known gold standards is the *Multiparameter Intelligent Monitoring in Intensive Care (MIMIC II)* corpus (Saeed et al., 2002).

In Sweden, the development of both de-identification tools and gold standards has come further than in Norway. In 2008, a group of Swedish researchers developed a gold standard corpus for de-identification of Swedish clinical text (Velupillai et al., 2009). The researchers manually annotated and de-identified 100 electronic patient records (EPRs) deriving from five different clinics (*Neurology, Orthopaedia, Infection, Dental Surgery* and *Nutrition*) at Karolinska University Hospital. The gold standard consists of unstructured text (around 174,000 tokens in total) and is known as the *Stockholm EPR PHI* corpus. It has 4,700 annotated instances distributed over 8 PHI-classes. It has been further developed to *Stockholm EPR PHI Pseudo* corpus, which contains only surrogate names, addresses, phone numbers, etc., and is partly available for research (Dalianis, 2019).

3 Data

3.1 NorSynthClinical

A corpus of Norwegian synthetic clinical text, the *NorSynthClinical* corpus[3], formed the basis of the created gold standard. *NorSynthClinical* is considered the first publicly available resource of Norwegian clinical text (Rama et al., 2018). It is written by one clinician with large experience with clinical work and genetic cardiology. The corpus describes patients' family history relating to cases of cardiac disease, and according to Rama et al. (2018), it consists of 477 sentences and 6030 tokens. Only a few of these tokens can be characterised as PHI.

4 Method

The development of the gold standard involved two main steps: extension and annotation. The gold standard was evaluated by measuring the Inter-Annotator Agreement (IAA) and by testing it on a hybrid de-identification tool.

4.1 Extension

The original dataset, *NorSynthClinical*, contains very few PHIs. Therefore, it was extended with

[3] NorSynthClinical, `https://github.com/ltgoslo/NorSynthClinical`.

synthetic PHIs (see example below). Where applicable, substatements and single words, or tokens, were manually added to the corpus. Most of the tokens were randomly selected from publicly available lists, such as Statistics Norway's lists of personal names used by 200 Norwegians or more[4]. The rest of the tokens were invented. They did, however, follow specific Norwegian formats, such as for social security numbers[5] and phone numbers[6]. For more details regarding the extension, see (Bråten, 2020).

1. Original sentence in Norwegian: *Moren har visstnok noen hjerteproblemer, hun er 75 år gammel.* (The mother apparently has some heart problems, she is 75 years old.)

2. Extended sentence: *Moren har visstnok noen hjerteproblemer, hun er 75 år gammel og bor på Bakklandet Menighets Omsorgsenter.* (The mother apparently has some heart problems, she is 75 years old and lives at Bakklandet Menighets Omsorgsenter.)

4.2 Annotation

The second step of the gold standard development involved annotation. Named Entity Tagging, using the tags provided in Table 1, as proposed by (Dalianis and Velupillai, 2010), was applied in order to mark up elements of PHI. Annotation guidelines were developed[7], and the tags were assigned in the following way in the following Norwegian sentence:

3. *Moren har visstnok noen hjerteproblemer, hun er <Age>75 år</Age> gammel og bor på <Health_Care_Unit>Bakklandet Menighets Omsorgsenter</Health_Care_Unit>*. In Eng. (The mother apparently has some heart problems, she is <Age>75 years </Age>old and lives at<Health_Care_Unit> Bakklandet Menighets Omsorgsenter </Health_Care_Unit>.)

PHI tags
First_Name
Last_Name
Age
Health_Care_Unit
Phone_Number
Social_Security_Number
Date_Full
Date_Part
Location

Table 1: The Named Entity Tag set used to mark up elements of PHI.

Two annotators annotated the whole corpus separately in order to facilitate error detection and comparative evaluation. The annotators, one master of medical science student, A1, and one finance manager, A2, were both Norwegian native speakers. No specific medical knowledge was needed to carry out the annotation.

4.3 Evaluation using Inter-Annotator Agreement and a hybrid de-identification tool

As mentioned, the gold standard was evaluated by measuring the IAA. This is a common evaluation method for providing a quantitative score of how accurate an annotation task is (Pustejovsky and Stubbs, 2012). The two annotated corpora written in UTF-8 encoding format, were converted to CoNLL[8] format, using a Python3 script, to enable the measurement of IAA. During this process, a token was defined as a string of characters between two spaces or a delimiter. The symbols that were defined as a part of a token, were percentage symbols located to the right of a number as well as hyphens and full stops between two letters or numbers. Moreover, the named entity tags were assigned *IOBES* schema, indicating whether a token was *Inside, Outside*, in the *Beginning* or in the *End* of an entity, or whether the entity was represented by a *Single* token, (Collobert et al., 2011). The evaluation metrics used to measure the IAA were precision, recall and F_1-measure.

Further evaluation was conducted by executing the de-identification tool developed for Nor-

[4] Norwegian personal names, https://www.ssb.no/statbank/table/12891/ and https://www.ssb.no/statbank/table/10501//

[5] Social security numbers, https://www.skatteetaten.no/en/person/National-Registry/Birth-and-name-selection/Children-born-in-Norway/National-ID-number/

[6] Phone numbers, https://www.nkom.no/telefoni-og-telefonnummer/telefonnummer-og-den-norske-nummerplan/alle-nummerserier-for-norske-telefonnumre

[7] Annotation guidelines, https://github.com/synnobra/NorSynthClinical-PHI/raw/master/Annotation_guidelines.pdf

[8] CoNLL, Conference on Natural Language Learning

NorNE Label	PHI Tags Label
B-PER	First_Name
I-PER	Last_Name
B-ORG	S/B Health_Care_Unit
I-ORG	I/E Health_Care_Unit
B-LOC	S/B Location
I-LOC	I/E Location

Table 2: NorNE labels matched to PHI Tags labels. S = Single, B = Beginning, E = Ending, I = Inside, O = Outside

wegian pathology reports, employing the same metrics of precision, recall and F_1-measure as for the IAA. The de-identification tool is a hybrid de-identification tool utilizing a Conditional Random Fields (CRF)[9] machine learning (ML) model trained on the Bokmål half of the *NorNE* corpus and regular expressions (REGEX) rule-based pattern matching. NorNe is a corpus of Norwegian non-clinical text made publicly available (Jørgensen et al., 2020), The hybrid de-identification tool is further described in (Wie, 2020).

Some, but not all PHI entities in the developed gold standard are found in the *NorNE* training data set. Furthermore, the labels in the *NorNE* data set differ from the gold standard's PHI both in label names and annotation schema[10]. The labels are matched as seen in Table 2[11]. As the CRF machine learning model is unable to recognize entities not found in the training set, some entities are detected by ML and some by REGEX, see Table 3.

Label	Method
First_Name	CRF
Last_Name	CRF
S/B Health_Care_Unit	CRF
I/E Health_Care_Unit	CRF
Location	CRF
Age	REGEX
Date	REGEX
Phone_Number	REGEX
Social_Security_Number	REGEX

Table 3: Method for detecting labels.

The evaluation done by the de-identification application is based on the CoNLL format described earlier in this chapter. The de-identification application was not designed to distinguish between Date_Part and Date_Full, so these entities were combined for the evaluation. Furthermore, the REGEX for phone numbers, dates and social security numbers were not designed to recognize entities split into more than one token.

5 Results

5.1 Extension and Annotation

An extended and annotated version of the *NorSynthClinical* corpus has been created. It has been given the name *NorSynthClinical PHI* and made publicly available on GitHub[12]. In total, it consists of 8,270 tokens and 409 PHI instances. The distribution of the PHI categories and an overview of the number of tokens added during the extension, is provided in Table 4. Moreover, Figure 1 shows the number and distribution of annotations where the annotators agreed and not, resulting in a micro-averaged overall IAA of 0.94, see Table 5. Only annotations with exactly the same tag and span were considered matching.

5.2 De-identification

The initial evaluation test yielded the results seen in Table 6 - a micro-averaged F_1-measure of 0.553. Following the initial test, the two following modifications were implemented:

1. The Health_Care_Unit entity label and Location entity label were merged.
2. The labels for entities predicted by rule-based methods were reduced – leaving the single-token instances and the first token in multi-token instances as is, and removing the rest.

These modifications yielded a micro-averaged F-measure of 0.730 and a recall of 0.619, see Table 7, and are discussed further in the analysis and discussion chapter.

6 Analysis

6.1 The NorSynthClinical PHI corpus

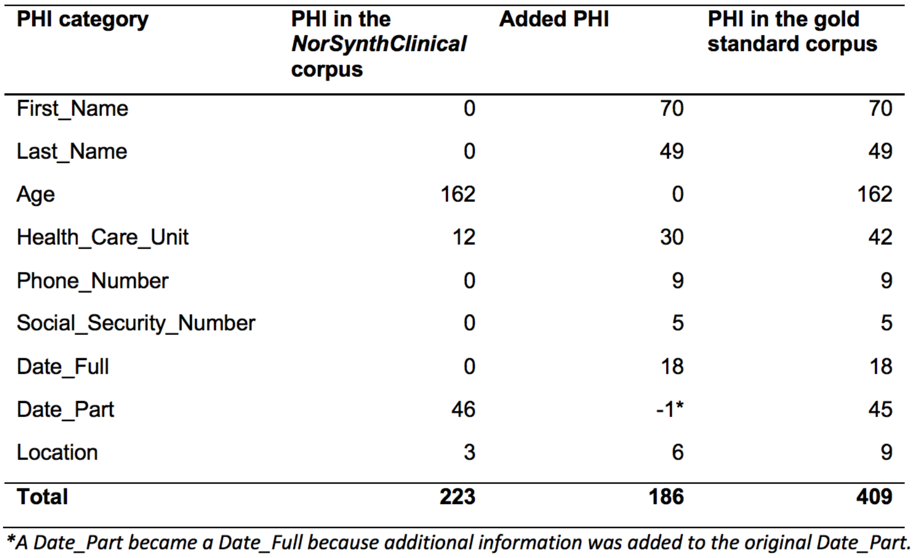
The amount of PHI in the extended corpus, *NorSynthClinical PHI*, constitutes around 5% of

[9] sklearn-crfsuite, `https://github.com/TeamHG-Memex/sklearn-crfsuite`

[10] NorNE uses the IOB2 schema (Jørgensen et al., 2020)

[11] Most notable is the matching of ORG and Health_Care_Unit

[12] NorSynthClinical PHI, `https://github.com/synnobra/NorSynthClinical-PHI`.



PHI category	PHI in the *NorSynthClinical* corpus	Added PHI	PHI in the gold standard corpus
First_Name	0	70	70
Last_Name	0	49	49
Age	162	0	162
Health_Care_Unit	12	30	42
Phone_Number	0	9	9
Social_Security_Number	0	5	5
Date_Full	0	18	18
Date_Part	46	-1*	45
Location	3	6	9
Total	**223**	**186**	**409**

**A Date_Part became a Date_Full because additional information was added to the original Date_Part.*

Table 4: The distribution of PHI categories in the original *NorSynthClinical* corpus containing 7,863 tokens) and the extended *NorSynthClinical PHI* corpus containing 8,270 tokens).

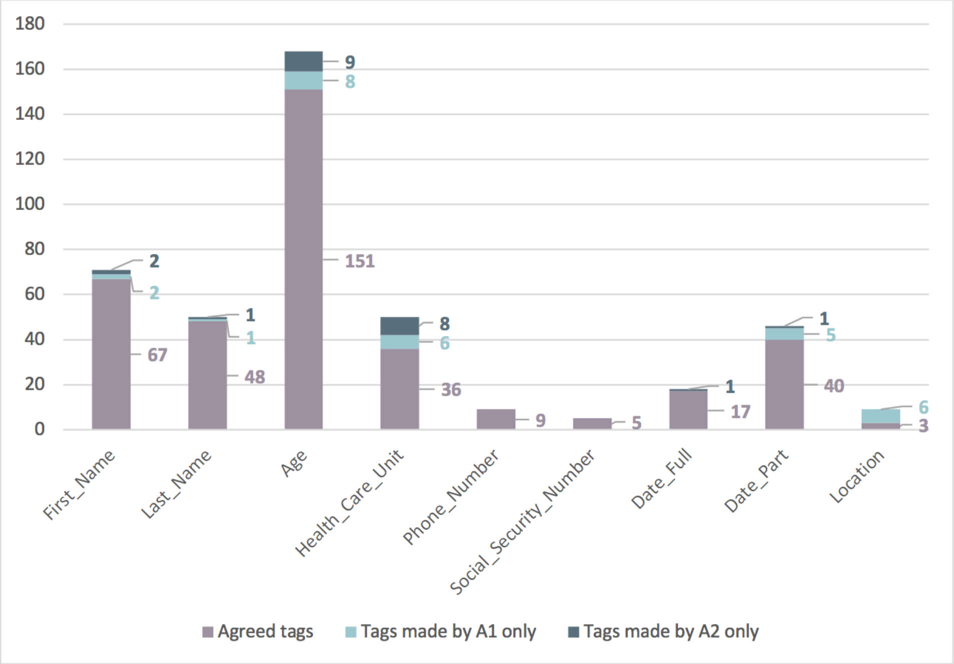
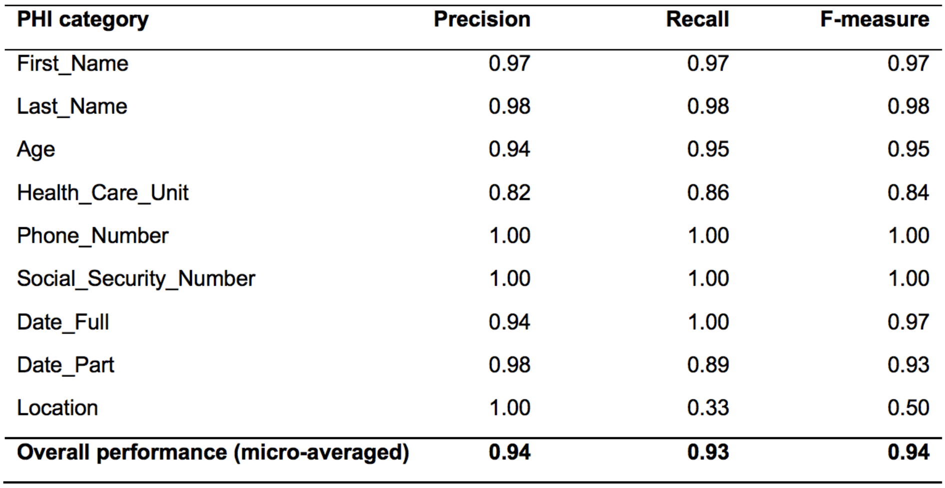


Figure 1: The distribution of agreed (n=376) and disagreed (n=50) annotation tags in each PHI category made by the two annotators A1 and A2

the content. This is above the average of 2% (Dalianis, 2018), but quite similar to the 4.3% reported by Bjurstrøm and Singh (2013). Even the distribution of the different PHI categories resembles the distribution in other clinical texts where names and dates make up the largest categories (Neamatullah et al., 2008; Dalianis and Velupillai, 2010; Deleger et al., 2014; Hanauer et al., 2013). In the extended corpus, names (including First_Name and Last_Name combined) make up almost one third of the overall PHI, and dates (including Date_Full and Date_Part) more than 15%. The most common category in the extended corpus, however, is Age. In the *NorSynthClinical PHI* corpus, Age constitutes around 39% of all PHI, while in other corpora, it constitutes no more than 1% (Neamatullah et al., 2008; Dalianis and Velupillai, 2010; Deleger et al., 2014).



PHI category	Precision	Recall	F-measure
First_Name	0.97	0.97	0.97
Last_Name	0.98	0.98	0.98
Age	0.94	0.95	0.95
Health_Care_Unit	0.82	0.86	0.84
Phone_Number	1.00	1.00	1.00
Social_Security_Number	1.00	1.00	1.00
Date_Full	0.94	1.00	0.97
Date_Part	0.98	0.89	0.93
Location	1.00	0.33	0.50
Overall performance (micro-averaged)	**0.94**	**0.93**	**0.94**

Table 5: The agreement between the two annotators that annotated the extended *NorSynthClinical* corpus.

Label	**Precision**	**Recall**	**F_1-measure**	**Support**
First_Name	0.951	0.806	0.872	72
Last_Name	0.946	0.964	0.955	55
S/B Health_Care_Unit	0.090	0.167	0.117	42
I/E Health_Care_Unit	0.833	0.192	0.346	26
Location	0.209	1.000	0.346	9
Age (REGEX)	0.985	0.259	0.410	247
Date (REGEX)	0.862	0.770	0.797	74
Social_Security_Number (REGEX)	1.000	0.286	0.444	7
Phone_Number (REGEX)	1.000	0.217	0.357	23
Micro avg.	0.675	0.468	0.553	555

Table 6: Initial evaluation test with the hybrid de-identification tool.

Label	**Precision**	**Recall**	**F_1-measure**	**Support**
First_Name	0.951	0.806	0.872	72
Last_Name	0.946	0.964	0.955	55
S/B Health_Care_Unit	0.767	0.647	0.702	51
I/E Health_Care_Unit	1.000	0.231	0.375	26
Age (REGEX)	0.985	0.395	0.564	162
Date (REGEX)	0.783	0.857	0.818	63
Social_Security_Number (REGEX)	1.000	0.400	0.571	5
Phone_Number (REGEX)	0.800	0.444	0.571	9
Micro avg.	0.893	0.619	0.731	443

Table 7: Final evaluation with the modified hybrid de-identification tool. The entities Health_Care_Unit and Location observed in Table 6 were merged into Health_Care_Unit. in this table

6.2 Inter-Annotator Agreement

The IAA score of 0.94 indicates that the agreement between the two annotators is high. This is especially true for the categories Phone_Number and Social_Security_Number, which the annotators completely agreed on, see Figure 1. However, these and most other categories contain a small number of PHI instances, questioning the reliability of the statistical analysis. The categories that the annotators disagreed on the most, were Health_Care_Unit and Location, see Figure 1. On five occasions, a PHI instance was annotated as Location by one of the annotators and as Health_Care_Unit by the other annotator. Other disagreements were caused by differences in the annotation span or in the interpretations of the provided annotation guidelines.

6.3 Evaluation with a hybrid de-identification tool

While the overall score for First_Name and Last_Name was high, the scores for Health_Care_Unit and Location were low, see Table 6. The low scores were suspected to be due to health care units often being named after locations and being syntactically similar, resulting in the CRF model frequently labelling Health_Care_Unit as Location – which was confirmed with a manual review of the incorrect predictions.

> "Of the 35 incorrect predictions where the correct label was B-ORG, 24 were labelled as B-LOC (approx. 69%).
> Of the 21 incorrect predictions where the correct label was I-ORG, 6 were labelled as I-LOC (approx. 29%)." (Wie, 2020)

For the entities processed by the rule-based part (REGEX) of the hybrid de-identification tool the initial precision was high (0.908 micro avg.). However, the recall was low for all entities except date (0.770). This was attributed to the CoNLL conversion of the *NorSynthClinical PHI* corpus splitting the pertinent entities into more tokens, which the de-identification application was not designed to handle. Another consequence of some of these entities being split is an inflation of the support for these categories. An example being the original nine instances of Phone_Number in the *NorSynthClinical PHI* corpus being counted as 23 instances – skewing the recall score, see Table 8. Applying the aforementioned modification of reduction based on prefixes resulted in the same instance support as the original.

7 Discussion and conclusion

What makes the *NorSynthClinical PHI* special and valuable is the fact that it is synthetic. As it does not contain any real personal information, the gold standard can be accessed by anyone and utilized in the development of tools for de-identification of Norwegian clinical text. Hopefully, this will facilitate more research on the content of clinical notes, and eventually a better health care.

Label	Original	CoNLL
Age	162	247
Date	63	74
Social_Security_Number	5	7
Phone_Number	9	23

Table 8: Converting from SGML format to CoNLL format support inflation.

The major weakness of the created gold standard is its small size. The English corpus *MIMIC II* consists of 412,509 clinical notes and the *Stockholm EPR PHI* corpus consists of 100 patient records (Dalianis, 2018). As mentioned in (Velupillai et al., 2009), the latter contributes 174,000 tokens. In comparison, the *NorSynthClinical PHI*, which consists of 8,270 tokens, is very small. Besides, it is very specific to the area of cardiology, written by one cardiologist, and extended by a layman. Therefore, there might be a lack of linguistic variety. Furthermore, the gold standard is written in Norwegian Bokmål and not in Nynorsk. However, it would be relatively uncomplicated to translate the gold standard from Bokmål to Nynorsk.

The de-identification tool used for evaluating *NorSynthClinical PHI* corpus was initially designed for another purpose[13] and trained on publicly available data. The effect of fundamental incompatibilities between the training set and the gold standard, like the disparity between *ORG* and *Health_Care_Unit*, is difficult to estimate. However, no other de-identification system for Norwegian is available.

The final evaluation of the modified hybrid de-identification tool for Norwegian using *NorSynthClinical PHI* gave an F_1-measure of 0.731 and a recall of 0.619.

A de-identification tool is aiming on a higher recall to remove all possible PHIs, also on the cost of lower precision.

Further improvements could be made to the de-identification tool. Implementing dictionary-based algorithms could improve the accuracy of certain entity types. Task-specific dictionaries for Norwegian health care units and/or medications are feasible implementations and would

[13] De-identification of Norwegian pathology reports.

likely improve accuracy on clinical texts. Furthermore, implementing tokenization directly in the de-identification tool would allow for de-identification of untokenized text, and minimize incompatibilities between the input and de-identification algorithm.

The gold standard has its limitations and cannot alone decide whether a specific tool provides sufficiently de-identified outcomes. Therefore, we encourage to further expansions of the gold standard corpus, in addition to more evaluation research, in order to make it more reliable and improve its quality.

Contributions of each author

SB made and evaluated the gold standard corpus, and wrote in the article. WW developed the hybrid de-identification tool and tested it on the gold standard corpus and co-authored the paper. HD supervised the study, gave comments and wrote in the article.

References

Roar Bjurstrøm and Jaspreet Singh. 2013. De-identification of Norwegian Health Record Notes: An Experimental Approach. Master's thesis, Institutt for datateknikk og informasjonsvitenskap.

Synnøve Bråten. 2020. Extending a Synthetic Norwegian Clinical Corpus for De-Identification. Master's thesis, Department of Computer and Systems Sciences, Stockholm University and Karolinska Institutet, https://daisy.dsv.su.se/fil/visa?id=230054.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12:2493–2537.

Hercules Dalianis. 2018. *Clinical text mining: Secondary use of electronic patient records*. Springer Nature, Open Access.

Hercules Dalianis. 2019. Pseudonymisation of Swedish Electronic Patient Records Using a Rule-Based Approach. In *Proceedings of the Workshop on NLP and Pseudonymisation*, pages 16–23, Turku, Finland. Linköping Electronic Press.

Hercules Dalianis and Sumithra Velupillai. 2010. De-identifying Swedish clinical text-refinement of a gold standard and experiments with Conditional random fields. *Journal of Biomedical Semantics*, 1(1):6.

Louise Deleger, Todd Lingren, Yizhao Ni, Megan Kaiser, Laura Stoutenborough, Keith Marsolo, Michal Kouril, Katalin Molnar, and Imre Solti. 2014. Preparing an annotated gold standard corpus to share with extramural investigators for de-identification research. *Journal of Biomedical Informatics*, 50:173–183.

Franck Dernoncourt, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. 2017. De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, 24(3):596–606.

Khaled El Emam, Fida Kamal Dankar, Romeo Issa, Elizabeth Jonker, Daniel Amyot, Elise Cogo, Jean-Pierre Corriveau, Mark Walker, Sadrul Chowdhury, Regis Vaillancourt, et al. 2009. A globally optimal k-anonymity method for the de-identification of health data. *Journal of the American Medical Informatics Association*, 16(5):670–682.

David Hanauer, John Aberdeen, Samuel Bayer, Benjamin Wellner, Cheryl Clark, Kai Zheng, and Lynette Hirschman. 2013. Bootstrapping a de-identification system for narrative patient records: cost-performance tradeoffs. *International Journal of Medical Informatics*, 82(9):821–831.

Fredrik Jørgensen, Tobias Aasmoe, Anne-Stine Ruud Husevåg, Lilja Øvrelid, and Erik Velldal. 2020. https://www.aclweb.org/anthology/2020.lrec-1.559 NorNE: Annotating named entities for Norwegian. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4547–4556, Marseille, France. European Language Resources Association.

Montserrat Marimon, Aitor Gonzalez-Agirre, Ander Intxaurrondo, Heidy Rodriguez, Jose Lopez Martin, Marta Villegas, and Martin Krallinger. 2019. Automatic De-identification of Medical Texts in Spanish: the MEDDOCAN Track, Corpus, Guidelines, Methods and Evaluation of Results. In *IberLEF@ SEPLN, La Sociedad Española para el Procesamiento del Lenguaje Natural*, pages 618–638.

Ishna Neamatullah, Margaret M Douglass, H Lehman Li-wei, Andrew Reisner, Mauricio Villarroel, William J Long, Peter Szolovits, George B Moody, Roger G Mark, and Gari D Clifford. 2008. Automated de-identification of free-text medical records. *BMC Medical Informatics and Decision Making*, 8(1):32.

James Pustejovsky and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications*. O'Reilly Media, Inc.

Taraka Rama, Pål Brekke, Øystein Nytrø, and Lilja Øvrelid. 2018. Iterative development of family history annotation guidelines using a synthetic corpus of clinical text. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 111–121.

Georg Rehm and Hans Uszkoreit. 2012. The Norwegian Language in the European Information Society. In *The Norwegian Language in the Digital Age*, pages 45–51. Springer.

Mohammed Saeed, Christine Lieu, Greg Raber, and Roger G Mark. 2002. MIMIC II: a massive temporal ICU patient database to support research in intelligent patient monitoring. In *Computers in cardiology*, pages 641–644. IEEE.

Sumithra Velupillai, Hercules Dalianis, Martin Hassel, and Gunnar H Nilsson. 2009. Developing a standard for de-identifying electronic patient records written in Swedish: precision, recall and F-measure in a manual and computerized annotation trial. *International Journal of Medical Informatics*, 78(12):e19–e26.

Wilhelm Wie. 2020. De-identification of Norwegian Clinical Text A Hybrid Approach Using Publicly Available Data. Master's thesis, Department of Computer and Systems Sciences, Stockholm University, https://daisy.dsv.su.se/fil/visa?id=230198.