De-identification of Dutch Medical Text

Erik Tjong Kim Sang[1], Ben de Vries[1], Wouter Smink[2], Bernard Veldkamp[2], Gerben Westerhof[2], Anneke Sools[2]
[1]Netherlands eScience Center, Amsterdam, The Netherlands
[2]University of Twente, Enschede, The Netherlands

April 21, 2019

Abstract

We present a method for de-identifying Dutch medical text based on an entity tagger and rules for identifying additional entity classes. A key problem for evaluating de-identification systems is that data used for evaluation cannot be shared. We propose to use public data for evaluation, which after an automatic conversion process can be made similar to our medical text. We compare our system with another system for de-identifying Dutch medical text and find that our system performs better both with respect to recall and F1.

Introduction

De-identification or anonymization of text involves the removal of information revealing the identity of persons mentioned in the text [1, 2, 3]. It is an essential preprocessing operation for sensitive documents before their release for research. A key problem in the development of systems for the automatic de-identification is obtaining data for training and testing. The sensitive parts of the data prohibit inspection by the researchers working with the data and prevent release to other parties for verification.

This paper describes our efforts to overcome this problem. We aim at developing an data collection which is suitable for evaluating de-identification software. We apply the collection to evaluate our own de-identification approach for Dutch online treatment sessions as well as assessing related software by others [4].

After this introduction, we present related work in section 2. Section 3 contains a description of the our de-identification task involving Dutch online health correspondence and our approach to collecting open data for evaluating this task[1]. In section four, we describe the results of the evaluation process. We conclude in section 5.

Related work

Uzuner et al. [1], Stubbs et al. [2] and Stubbs et al. [3] describe shared tasks related to the de-identification of English medical data. The first (2006) dealt with medical discharge records, the second (2014) with data related to heart disease of diabetic clients while the third (2016) involved psychiatric intake records. Data sets for these tasks were released after replacing private health information with surrogates (pseudonymization). Between six and ten teams participated in the tasks. Two popular techniques used were machine learning and rules. Both approaches performed well with F1 scores of the best systems exceeding 90% in the 2016 shared task.

We are aware of two studies for de-identifying texts written in the language Dutch: Menger et al. [4] built a pattern-based system for removing private information from nursing notes and treatment plans. It obtained an F1 score of 86% without taking into account duplicate entities. Scheurwegs et al. [5] trained a

[1]The software as well as the data used in this study are available at `https://github.com/e-mental-health/data-processing`

Name	Clients	Mails	Tokens
ADB	1,987	45,495	17,254,441
OVK-AS	55	1,007	378,540
OVK-ES	58	969	483,190
Wikipedia	-	-	11,301

Table 1: Dutch data used in our de-identification work. The number of tokens is equal to the number of words plus the number of punctuation signs. The ADB, OVK-AS and OVK-ES data sets involve therapeutic email correspondence which cannot be shared. The annotated Wikipedia data will used for evaluation of the de-identification process and will be shared with the community.

machine learner on manually anonymized data which was de-anonymized with fictitious values. Their best learner achieved an F1 score of 91%.

Method

We need de-identification to remove private health information from Dutch emails from online therapeutic sessions. We have data from three therapies, one related to alcohol dependence (ADB: 1987 clients) [6] and two involving therapeutic writing (OVK-AS: 55 clients and OVK-ES: 58 clients) [7]. We have both emails from clients and their counselors as well as some metadata, like diaries and questionnaires with answers. Not all clients have completed their therapy. We want to search the data for text patterns revealing cognitive developments of the clients and learn what actions of the counselor have contributed to positive developments [8]. The sessions frequently involve writing about the client's personal situation so the emails need to be de-identified before they will be released for further analysis.

The quality of the de-identification process is important. If we can convince our data providers that this process removes all sensitive information from the data, it will be easier to obtain data for future studies. We would like to have an evaluation procedure which can be verified by external researchers, including open evaluation data sets. However, we either do not have access to the original data sets or they cannot be released because they contain sensitive information. Therefore we have started looking for other similar data which can be used for evaluating de-identification software.

There are two data sources which contain texts which are similar in style to the autobiographic emails in our data: online forums and diaries. There are several relevant Dutch online forums where people write about their experiences with health, medicine and addiction, for example `medischforum.nl`. While the texts are publicly available for anybody to read it is unclear if they can be applied for scientific research. Re-issuing these text in a external data set is probably not allowed given the most recent privacy laws [9].

Dutch literature includes several publicly available autobiographically written works which could serve as evaluation data for our task. The most famous Dutch diary by Anne Frank is still under copyright but Multatuli's Max Havelaar can be obtained from Project Gutenberg [10]. However, this book was released in 1860 and the language used is quite out-of-date. The organization Nederlands Dagboekarchief (Dutch Diary Archive)[2] collects contemporary Dutch diaries of common people which could be of interest for our purpose. However, because of privacy concerns their diary collection is not available in digital form.

In order to obtain relevant contemporary text for de-identification system evaluation, we turned to the Dutch Wikipedia. It contains several biographies of famous people which could provide good evaluation material for our task. We converted the biographies to auto-biographies in an egofication process: replacing the subject name and *hij/zij* (the Dutch versions of *he* and *she* by *ik* wherever appropriate and performing similar replacement for related pronouns. We did not replace verbs (*does* → *do*) since most of the articles are written in the past tense which has the same form for the first and third form in Dutch. We used two Wikipedia articles, the one for the painter Rembrandt van Rijn and the one for the female spy Mata Hari.

For our task, de-identification involves hiding all names (persons, organizations, locations and miscellaneous [11]), dates, email addresses, urls and numbers (including phone numbers). It is not necessary to keep track of which exact entity was hidden so we will use the unnumbered strings PER, ORG, LOC, MISC,

[2] `dagboekarchief.nl`

Name	Measure	UNL	AVG	DATE	LOC	MISC	NUM	ORG	PER
Our system	Precision	84%	54%	98%	49%	14%	55%	6%	71%
	Recall	85%	57%	16%	70%	7%	95%	29%	68%
	F1	84%	55%	28%	58%	10%	70%	10%	70%
Deduce	Precision	76%	60%	54%	88%	0%	100%	0%	57%
	Recall	27%	21%	13%	21%	0%	1%	0%	54%
	F1	40%	31%	21%	33%	0%	1%	0%	55%

Table 2: Evaluation of de-identification of Dutch Wikipedia text. We rated the recognition of six entity classes (Date, Location, Miscellaneous, Number, Organization and Person, but not Mail and Url), as well as the micro average of these class scores (AVG) and unlabeled de-identification (UNL). Our system outperforms the system Deduce [4] with respect to F1 score and recall on all eight rates.

DATE, MAIL, URL and NUM as replacements. For preprocessing text, performing linguistic analysis and performing entity identification, we use the Dutch linguistic processing program Frog [12]. The analysis program was applied to the Wikipedia text and the results were manually checked and corrected. Additionally, date strings were identified and marked. The texts contained neither email addresses nor urls, so these will not be included in the evaluation.

Results

Our system for de-identification of Dutch text, relies on the syntax analysis and entity recognition of the tool Frog [12]. This tool uses machine learning techniques and name lists to identify five of the eight entity classes of interest: locations, miscellaneous entities, numbers, organizations and person names. We used rules for identifying the other three entity classes (dates, email addresses and urls). After identifying an entity, the system replaces it with a dummy token reflecting the entity class, for example PER for person names. We did not use indexes to reflect that replaced strings were identical both because we did not need the information and because this would create an extra risk of privacy breach.

We applied both our system and Deduce [4] to the Wikipedia data described in the previous section. The output of the systems was compared with the manual gold standard with the evaluation software of the CoNLL-2003 shared task on named entity recognition [11]. Entities were only regarded as correct when they contained exactly the same tokens and when they were assigned the correct class. Since the class information is not very important for our application, we also performed an evaluation without taking into account the entity classes (using unlabeled entities). The results of the evaluation can be found in Table 2.

Our system obtains a low micro averaged F1 score (55%) but the unlabeled F1 score (84%) is reasonable. Surprisingly, the Deduce system [4] does not do very well on this task. There are two reasons for this. First the system was not developed to identify all six entity classes used in this evaluation: miscellaneous entities were not included and numbers and organizations were restricted to patient ids and health institutions, respectively. The second reason for the low performance is that Deduce uses only name lists for identifying entities while our system relies on an entity recognition system which uses both name lists and machine learning, which improves the system's recall and generalizability.

An error analysis revealed that the most common mistake by our system involved confusion between dates and numbers. The Wikipedia articles contain many years which our system classified as numbers, because it lacks explicit context information like an adjacent month name. The system also sometimes failed to include uncapitalized words as parts of person names, like *van* (*from*) in *Jan van Amsterdam*, in particular when the following word was also the name of a location. Miscellaneous entities provided problems but this mixed bag of entities poses a problem for any classifier. Deduce also tagged years as numbers rather than dates. It frequently incorrectly included punctuation and the phrase *van mij* (*of mine*) in entities.

Based on this analysis, we confident that we could apply our system for de-identifying Dutch medical text. We perform one extra action in order to increase recall rates: we require the system to generate a list of all non-sentence-initial capitalized words. After processing all documents, these words are inspected manually, without looking at their context in order to avoid privacy breaches. Missed words that are suspected to be

part of names are then removed from all documents before using them for subsequent study.

Concluding remarks

We presented a study on the de-identification of Dutch medical texts. We employ a system which relies on an existing entity classifier for Dutch combined with entity detection rules. A problem in evaluating and verifying such systems is that the nature of the data prohibits sharing data and thus comparing systems. We propose to instead use public data for evaluation, like biographies from Wikipedia, which after an automatic egofication process can be turned into texts which are very similar to the egodocuments used in our mental health study. We evaluated our system against these data and compared it with the only available system for de-identification of Dutch texts: Deduce [4]. We found that our system outperformed Deduce with respect to recall and F1 on all tested classes and class sets.

We see several opportunities for future work. First, we would like to expand the evaluation data set, both in size and with respect to the covered entity classes. Second, we would like to explore methods for improving the coverage of our de-identification system. This could involve both using analysis components of the Frog entity tagger that we did not use yet, as well as combining it with other entity taggers. Finally, it would be interesting to explore evaluation of the system based on how much private information can be derived from the de-identified text. Such an analysis could provide useful information about the requirements of de-identification systems.

References

1. Uzuner Ö, Luo Y, Szolovits P. Evaluating the State-of-the-Art in Automatic De-identification. Jorunal of the American medical Informatics Association. 2007;14(5).
2. Stubbs A, Kotfila C, Uzuner O. Automated systems for the de-indentification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1. J Biomed Inform. 2015;.
3. Stubbs A, Filannino M, Uzuner Ö. De-identification of psychiatric intake records: Overview of 2016 CEGS N-GRID shared tasks Track 1. Journal of Biomedical Informatics. 2017;75:S4–S18.
4. Menger V, Scheepers F, Van Wijk L, Spruit M. DEDUCE: A pattern matching method for automatic de-identification of Dutch medical text. Telematics and Informatics. 2017;35:727–736.
5. Scheurwegs E, Luyckx K, Van der Schueren F, Van den Bulcke T. De-Identification of Clinical Free Text in Dutch with Limited Training Data: A Case Study. In: Proceedings of the Workshop on NLP for Medicine and Biology. RANLP, Hissar, Bulgaria; 2013. p. 18–23.
6. Postel MG. Well Connected - Web-based Treatment for Problem Drinkers. PhD thesis, Radboud University Nijmegen, The Netherlands; 2011.
7. Bohlmeijer E, Westerhof G. Op verhaal komen - Je autobiografie als bron van wijsheid. Boom uitgevers, Amsterdam, The Netherlands; 2012. (In Dutch).
8. Wiegersma S, Sools AM, Veldkamp BP. Exploring L̈etters from the Futureb̈y Visualizing Narrative Structure. In: Proceedings of the 7th Workshop on Computational Models of Narrative (CMN 2016). Dagstuhl Publishing, Germany; 2016. .
9. EU. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). European Union; 2016.
10. Multatuli. Max Havelaar. J. de Ruyter, Amsterdam; 1860. Available from: `gutenberg.org/ebooks/11024`.
11. Tjong Kim Sang EF, De Meulder F. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In: Daelemans W, Osborne M, editors. Proceedings of CoNLL-2003. Edmonton, Canada; 2003. p. 142–147.
12. Van den Bosch A, Busser B, Canisius S, Daelemans W. An efficient memory-based morphosyntactic tagger and parser for Dutch. LOT Occasional Series. 2007;7:191 – 206.