De-Identification of Emails: Pseudonymizing Privacy-Sensitive Data in a German Email Corpus

Elisabeth Eder **Ulrike Krieg-Holz**
Institut für Germanistik
Alpen-Adria-Universität Klagenfurt, Klagenfurt, Austria
{elisabeth.eder | ulrike.krieg-holz}@aau.at

Udo Hahn
Jena University Language & Information Engineering (JULIE) Lab
Friedrich-Schiller-Universität Jena, Jena, Germany
udo.hahn@uni-jena.de

Abstract

We deal with the pseudonymization of those stretches of text in emails that might allow to identify real individual persons. This task is decomposed into two steps. First, named entities carrying privacy-sensitive information (e.g., names of persons, locations, phone numbers or dates) are identified, and, second, these privacy-bearing entities are replaced by synthetically generated surrogates (e.g., a person originally named *'John Doe'* is renamed as *'Bill Powers'*). We describe a system architecture for surrogate generation and evaluate our approach on CODEALLTAG, a German email corpus.

1 Introduction

With the advent and rapidly increasing adoption of electronic interaction platforms, the communication patterns of modern societies have changed fundamentally. We observe an unprecedented upsurge of digitally transmitted private communication and exploding volumes of so-called user-generated contents (UGC). As a major characteristic of these new communication habits, a sender's individual email, post, comment, tweet is distributed to an often (very) large number of addressees—the recipients of an email, other bloggers, friends or followers in social media platforms, etc.. Hence, hitherto private communication becomes intentionally public.

Responding to these changes, digital (social) media communication has become a focus of research in NLP. Yet there seems to be a lack of awareness among NLP researchers that the exploitation of natural language data from such electronic communication channels, whether for commercial, administrative or academic purposes, has to comply with binding legal regulations (Wilson et al., 2016). Dependent on each country's law system, different rules for privacy protection in raw text data are enforced (cf., e.g., two recent analyses for the US (Mulligan et al., 2019) and the EU (Hoofnagle et al., 2019)). Even privacy-breach incidents in a legal grey zone can be harmful for the actors involved (including NLP researchers).

This is evidenced dramatically in the so-called AOL search data leak.[1] In August of 2006, American Online (AOL) made a large query log collection freely accessible on the Internet for a limited time. The data were extracted over three months from their search engine to support academic research. The collection represented 650k users issuing 20 million queries *without* any significant anonymization. The result of this release, among others, was the disclosure of private information for a number of AOL users. The most troubling aspect of the data leak was the ease by which single unique individuals could be pinpointed in the logs. Even ignoring the existence of social security, drive license, and credit card numbers, the *New York Times* demonstrated the ability to determine the identity of a real user.[2] The outline of this incident and counter-measures against this privacy crash are reported by Adar (2007) from which we adopted the case description as well.

Despite this specific case, query logs from search engines might still be at the lower end of the vulnerability chain for data privacy, while UGC bundled in freely distributed corpora is clearly at its higher end, since clear names of persons, locations, etc. are dispersed all over such documents.

[1] Briefly described in https://en.wikipedia.org/wiki/AOL_search_data_leak, last accessed on July 24, 2019.

[2] https://www.nytimes.com/2006/08/09/technology/09aol.html, last accessed July 24, 2019.

Proceedings of Recent Advances in Natural Language Processing, pages 259–269,
Varna, Bulgaria, Sep 2–4, 2019.
https://doi.org/10.26615/978-954-452-056-4_030

Surprisingly, despite its high relevance for NLP operating on UGC, the topic of data privacy has long been neglected by the mainstream of NLP research. While it has always been of utmost importance for medical, i.e., clinical, NLP (Meystre, 2015), it has received almost no attention in NLP's non-medical camp for a long time (for two early exceptions, cf. Rock (2001); Medlock (2006)).

This naïve perspective is beginning to change these days with the ever-growing importance of social media documents for text analytics. However, there are currently no systematic actions taken to hide personally sensitive information from down-stream applications when dealing with chat, blog, SMS or email raw data. Since this attitude also faces legal implications, a quest for the protection of individual data privacy has been raised and, in the meantime, finds active response in the most recent work of the NLP community (Li et al., 2018; Coavoux et al., 2018).

We distinguish two basic approaches to eliminate privacy-bearing data from raw text data. The first one, *anonymization*, identifies instances of relevant privacy categories (e.g., person names or dates) and replaces sensitive strings by some artificial code (e.g., *'xxx'*). This blinding approach might be appropriate to eliminate privacy-bearing data in the medical world, but it is inappropriate for most NLP applications since crucial discriminative information and contextual clues will be erased by such a scrubbing procedure.

The second approach, *pseudonymization*, preserves such valuable information by replacing privacy-bearing text strings with randomly chosen alternative synthetic instances from the same privacy type (e.g., the person name *'Suzanne Walker'* is mapped to *'Caroline Snyder'*). As a common denominator, the term *de-identification* subsumes both, anonymization and pseudonymization.

The focus of this paper will be on pseudonymization and more precisely on the methods needed to produce realistic synthetic replacements, a process often also referred to as *surrogate generation*. We start with a discussion of related work in Section 2 and then introduce the semantic types we consider as relevant carriers of personal information in emails in Section 3. Next, we provide an overview of the email corpus our experiments are based on in Section 4, including manual annotation efforts. In Section 5, we turn to the process of surrogate generation, with focus on German language data. Since surrogate generation constitutes a highly constrained case of language generation, in Section 6 we describe the results of an evaluation study to assess the naturalness of these replacements with native speakers of German, as well as the performance of a recognizer for privacy-relevant text stretches on original and already pseudonymized data.

2 Related Work

The main thrust of work on de-identification has been performed for clinical NLP.[3] Most influential for progress in this field have been two challenge competitions within the context of the I2B2 (Informatics for Integrating Biology & the Bedside) initiative[4] which focused on 18 different types of Protected Health Information (PHI) categories as required by US legislation (HIPAA).[5] The first of these challenge tasks was launched in 2006 for 889 hospital discharge summaries, with a total of 19,498 PHI instances of person-identifying verbal expressions (Uzuner et al., 2007). The second was run in 2014 and addressed an even broader set of PHI categories (Stubbs et al., 2015a). In summary, the best system performances peaked in the high 90s (F_1 score) using classical machine learning methods, Conditional Random Fields (CRFs) in particular, and hand-written rules, or a mixture of both. As a successor to I2B2, the CEGS-NGRID Shared Tasks and Workshop on Challenges in NLP for Clinical Data created a corpus of 1,000 manually de-identified psychiatric evaluation records (Stubbs et al., 2017). Interestingly, for the automatic de-identification task performance values dropped significantly down to 79.85 F_1 for the best-performing system indicating an only modest potential for domain and text genre portability (moving from discharge summaries to psychiatric evaluation records).

Recently, the deep learning wave has also hit the (clinical) de-identification community. For this task, bidirectional Long-Short Term Memory Networks (Bi-LSTMs) became quite popular as evidenced by the work of Dernoncourt et al. (2017)

[3] Note that we have to distinguish between data protection in structured data contained in (clinical) information systems, (for which k-anonymity (Sweeney, 2002) is a well-known model to minimize a person's re-identification risk) and pseudonym-based textual variant generation for unstructured verbal data we here focus on.

[4] https://www.i2b2.org/

[5] https://www.hhs.gov/hipaa/for-professionals/privacy/index.html

who achieve an F_1 score of 97.85 on the I2B2 2014 dataset, or Liu et al. (2017) who report performance figures ranging from 95.11% over 96.98% up to 98.28% micro F_1 score under increasingly sloppier matching criteria on the same dataset.

Note that these challenges were focusing on the *recognition* of privacy-relevant text stretches textual but did not incorporate pseudonymization, a more complex task (Stubbs et al., 2015b). Carrell et al. (2013) deal with the latter challenge within the context of the 'Hiding In Plain Sight' approach where the detected privacy-bearing identifiers are replaced with realistic synthetic surrogates in order to collectively render the few 'leaked' identifiers difficult to distinguish from the synthetic surrogates—a major advantage for pseudonymization over anonymization. Targeting English medical texts SCRUB (Sweeney, 1996) is one of the first surrogate generation systems followed by work from Uzuner et al. (2007), Yeniterzi et al. (2010), Deléger et al. (2014), Stubbs et al. (2015b) and Stubbs and Uzuner (2015). Similar procedures have been proposed for Swedish (Alfalahi et al., 2012) and Danish (Pantazos et al., 2011) clinical corpora, yet not for German ones.

Work outside the clinical domain is rare. While we found no work dealing with the anonymization or even pseudonymization of emails and Twitter-style social media data,[6] anonymizing SMSes is a topic of active research. Patel et al. (2013) introduce a system capable of anonymizing SMS (Short Message Service) communication. Their study builds on 90,000 authentic French text messages and uses dictionaries as well as decision trees as machine learning technique. Their evaluation task is, however, very coarse-grained—select those SMSes from a test corpus of 23,055 messages that either have or have not to be anonymized. There is no breakdown to PHI-like categories known from the medical domain.

Treurniet et al. (2012) were taking care of privacy-relevant data for a Dutch SMS corpus (52,913 messages, in total) in much more detail. They automatically anonymized all occurrrences of dates, times, decimal amounts, and numbers with more than one digit (telephone numbers, bank accounts, etc.), e-mail addresses, URLs, and IP addresses. All sensitive information was replaced with corresponding semantic placeholder codes of the encountered semantic *type* (e.g., each specific email address was replaced by the type symbol `EMAIL`), not by an alternative semantic *token*, i.e., a pseudonym. The same strategy was also chosen by Chen and Kan (2013) for their SMS corpus that contains more than 71,000 messages, focusing on English and Mandarin. However, neither are the methods of automatic anonymization described in detail, nor are performance figures of this process reported in both papers (Chen and Kan (2013) only mention the use of regular expressions for the anonymization process).

In conclusion, pseudonymization has to the best of our knowledge only been seriously applied to medical documents, up until now. Hence, our investigation opens this study field for the first time ever to non-medical applications of pseudonymization. Such de-identified corpora can then easily be distributed via public sites and so might stimulate further NLP research.

3 Named Entities for De-Identification

Perhaps the most relevant source and starting point for determining types of personally identifying information pieces in written documents is a catalogue of *Personal Health Information* (PHI) items that has been derived from the *Health Information Privacy Act* (HIPAA) which is binding law in the US. PHI enumerates altogether 18 privacy-sensitive items organized into eight main categories (Stubbs and Uzuner, 2015):

- *Name* includes the names of patients, doctors and user names,
- *Profession* of persons mentioned,
- *Location* includes rooms, clinical departments, hospital names, names of organizations, street names, city names, state names, names of countries, ZIPs, etc.,
- *Age* of persons,
- *Date* expressions,
- *Communication* data, e.g., phone or fax numbers, email addresses, URLs, IP addresses,
- all sorts of *IDs* such as Social Security number, medical record number, health plan number, account number, license number, vehicle ID, device ID, biometric ID, etc.,
- any *Other* form of personally sensitive data.

While some types of categories from above are generally useful also for non-medical anonymization procedures, others are quite domain-specific,

[6] Lüngen et al. (2017) report on *manual* anonymization efforts for German chat data.

because they are intrinsically attached to the clinical domain (such as the names of patients, doctors or nurses, the names of hospitals and their departments, or various forms of IDs, e.g., health insurance numbers). Hence, we adapted this list for email documents while, at the same time, we tried to avoid over-fitting to this text genre.

We, finally, came up with the category set depicted in Figure 1 which we stipulate to universally account for all types of emails, irrespective of any particular natural language and email (header) encoding. The categories are organized in a concise hierarchy whose top level categories are *SocialActor (ACTOR)*, *Date (DATE)*, *FormalIdentifier (FID)*, *Location (LOC)*, and *Address (ADD)*. We anticipate that this hierarchy can be further refined and accomodated to other text genres as well.

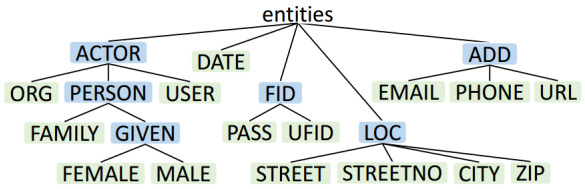

Figure 1: Hierarchy of privacy-bearing entity types (pis) relevant for emails (leaves in green)

The category of *SocialActor* can be further divided into *Organization (ORG)*, which includes all types of legal actors such as companies, brands, institutions and agencies, etc., human *Persons (PERSON)*, with subtypes *FamilyName (FAMILY)*, also including initials and credentials, and *GivenName (GIVEN)*, with another split into two subcategories, namely *FemaleName (FEMALE)* and *MaleName (MALE)*, both including nicknames and initials. Finally, *UserName (USER)* covers all kinds of invented user names for IT systems and platforms.

Date (DATE) includes all sorts of date descriptions, such as date of birth, marriage, or death, starting and ending dates of contracts, etc.

The category of *FormalIdentifier (FID)* includes *Password (PASS)* as user-provided supplementary artificial name for all kinds of technical appliances, and *UniqueFormalIdentifier (UFID)* to capture persons (students, customers, employees, members of social security systems (SSN), authors (ORCHID), etc.), computer systems (IP addresses), or other artifacts (e.g., IBANs, DOIs).

The *Location* (LOC) category subsumes *StreetName (STREET)*, *StreetNumber (STREETNO)*, *ZipCode (ZIP)*, and *CityName (CITY)* which stands for villages, towns, cities, larger metropolitan areas (e.g., *'Larger Manchester'*) and regions smaller than a state (e.g., *'Bay Area'*); it also includes derivations of these names (e g., *'Roman'*).

Finally, *Address (ADD)* encompasses *EmailAddress (EMAIL)*, *PhoneNumber (PHONE)*, including fax numbers, and *URL (URL)*, as well as other forms of domain names.

Unlike some approaches from the field of clinical NLP (Stubbs et al., 2015b), we did not take ages or professions into account, because our use case is not that sensible and ages or professions probably are mentioned far more often in clinical reports than in emails. Furthermore, unspecific dates like *'Christmas'* or *'next week'* and geographical information such as landmarks, rivers or lakes were not tagged for de-identification since their contribution to possible re-identification is fairly limited due to their generality.

4 Email Corpus and Entity Annotation

Our experiments are based on 1,390 German emails from the CODE ALLTAG$_{S+d}$ corpus (Krieg-Holz et al., 2016), which were collected on the basis of voluntary email donation. The donors have provided their explicit consent that, *after de-identification*, their emails may be made publically available. Sharing the corpus would create lots of opportunities for NLP research, since public access to private emails is generally forbidden.[7]

For the manual annotation campaign,[8] we set up a team of three annotators who tagged equally sized parts of the corpus, according to the privacy-bearing (pi) categories described in Section 3 (Figure 1). Annotation was performed on entity level. Therefore, we did not have to care about token boundaries in the surrogate generation step and, thus, no special handling for compounds and multi-token entities is required.

In order to measure the inter-annotator agreement (IAA), the annotators worked on 50 identical emails randomly selected from the corpus within the same annotation phase as the entire corpus.

[7] One of the rare exceptions is the ENRON corpus (Klimt and Yang, 2004) whose non-anonymized contents was released for open inspection by order of US judges in the course of the destruction of the Enron company as a consequence of criminal financial transactions of the Enron management. For yet another example, cf. the *Avocado Research Email Collection*, available from LDC2015T03.

[8] We used BRAT (http://brat.nlplab.org/) for annotation (Stenetorp et al., 2012).

Table 1 shows Cohen's Kappa (Cohen, 1960) as a measure for IAA for the pairs of annotators calculated on the entities represented by the BIO annotation scheme.[9] Hence, not only the token label itself but also matching starting and ending points of an entity are taken into account, as well. The agreement is quite high, especially between annotator 1 and 3.

A1 - A2	A2 - A3	A3 - A1
0.925	0.933	0.958

Table 1: Cohen's Kappa for BIO tags on CODE ALLTAG$_{S+d}$

Based on these 50 emails the annotators also examined and discussed differences of their annotations and decided on the gold standard by majority vote, which we applied for further evaluation in order to measure precision (Prec), recall (Rec) and F_1 score (F_1). Table 2 shows the outcomes regarding BIO tags per annotator and the overall result calculated over the joint annotations of the annotators. We took the averages from the outcomes of the single categories weighted by the number of true instances for each label.

	Prec	Rec	F_1
A1	98.06	95.63	96.72
A2	87.67	77.92	80.36
A3	94.14	84.79	86.82
A1+A2+A3	93.37	86.11	88.66

Table 2: Weighted average of precision, recall and F_1 score with respect to the gold standard for BIO tags of CODE ALLTAG$_{S+d}$

An error analysis revealed that, besides mostly accidental errors, a higher disagreement on tagging ORGs (due to overlap or confusion with city or product names and rather generic organizations)[10] and an uncertainty regarding DATEs could be observed. The latter problem was solved by the decision to treat all dates as *pi* regardless of their specificity. As a consequence, one annotator worked through the entire corpus to re-tag each DATE and, if necessary, also re-tag ORGs according to the findings of the error analysis. The outcome of this overhaul constituted the final gold standard annotations of CODE ALLTAG$_{S+d}$ for the de-identification task.

[9] 'B' preceding a token's tag stands for the Beginning of an entity, 'I' for its continuation (Inside), and 'O' for any stretch of text not belonging to an entity (Outside).

[10] Stubbs and Uzuner (2015) also report confusions of organizations with other subcategories from their location class (department, hospital) and Stubbs et al. (2017) witness an uncertainty for tagging quasi-generic organizations.

5 Surrogate Generation

Once *pi*-relevant named entities have been recognized they undergo a replacement process where original identifiers are substituted by synthetic, though natural, surrogates. For this step, we distinguish language-independent criteria from those which are intrinsically language-specific. Only the latter have to be re-specified for languages other than German.

5.1 Language-Independent Criteria

Personal information belonging to the categories ADD (EMAIL, PHONE or fax number, URL), FID (PASSword, UFID), USER name and ZIP code are relatively simple to replace. Each digit of the string is substituted with a randomly generated alternative digit, each alphabetic character is replaced with a randomly generated alternative letter of the same case and alphabet. Other characters like *'@'* or punctuation marks are left as is. For URLs, we also keep the subdomain *'www'* and commonly used URL schemes like *'http'*, *'https'*, *'ftp'*, *'file'* and *'mailto'*. In contrast to Stubbs et al. (2015b) we did not implement any other restrictions for the selection of characters. As a consequence, the resulting surrogates may have an unrealistic appearance.

To maintain temporal ordering in the document we generate a time shift separately for each text to make re-identification difficult, if not impossible. We shifted the dates by a random interval between 365 days forward or backward. As we try to keep the original language-dependent formats, the user has to determine the formats to be generated.

In order to maintain coreferences between *pi* entities, we replace multiple occurrences of an entity by the same surrogate. To account for different spellings of names regarding lower and upper case we treat different possibilities of combinations, the original spelling, lower case, upper case and a normalized format which is language-specific. We decided not to consider misspellings, because checking for slightly different names, e.g., employing the Levenshtein distance, could also lead to coreference breaks since quite a few names differ only in one letter (like *'Lena'* and *'Lina'*).

For resolving initials and abbreviations of GIVEN, FAMILY and CITY names, we adopt the approach by Stubbs et al. (2015b) and use letter-to-letter mappings generated for each document. This means that each name of the respective category starting with a certain character is replaced with a name with its first letter corresponding to the mapping of this character or in case of initials and abbreviations with the mapping solely. For example *'Gandalf'* and *'Gimli'* would be substituted by names starting with the same character like *'Bilbo'* and *'Boromir'*. Hence, an initial *'G.'*, will also be replaced with *'B.'*; it does not have to be disambiguated and assigned to any of the previously occurring names (a task left to a coreference resolution module we currently do not provide).

We also try to account for frequency distributions of GIVEN, FAMILY, CITY, ORG and STREET names by constraining these random letter-to-letter-mappings to map first letters to letters with a similar frequency. In this way, mappings of very common first letters, such as *'A'* in case of German first names, to rare ones, like *'X'*, can be avoided. This approach still allows rare substitutes for common names. However, we circumvent the problem of adding ambiguity to the text, if we only have few names starting with *'X'* (Stubbs et al. (2015b) also mentioned but did not implement this idea). As we map distributions in quite a rough way, we do not think that this could cause a leak of information of the original text, but distributional mappings are optional and the user may choose the granularity and distribution in his or her own module or language extension.

5.2 Language-Dependent Criteria for German

Since the categories DATE, STREET, CITY, GIVEN and FAMILY name, and ORG are affected by language-specific influences we implemented a German language module for these named entity types. In contrast to other surrogate generation systems we know of, our solution takes inflection into account (relying on the NLP tool SPACY (Honnibal and Montani, 2017)).

Dates. The formats we handle include typical combinations of day, month and year according to German formatting style (e.g., days precede months), yet also account for different spellings (e.g., *'01.06.2019'* or *'1.6.19'*). Days, months, and years occurring in isolation are processed as well. If a month is given in letter format, it is substituted with an alternative month name trying to keep differences between standard varieties (e.g., *'Januar'* (standard German) vs. *'Jänner'* (Austrian German) for *'January'*) and also preserve abbreviations (e.g., *'Jan.'*).

Street names. For names of streets, we recognize the abbreviations *'S|-str(.)'* (for *'Straße'* (street)) and *'P|-pl(.)'* (for *'Platz'* (place)) for look-up and coreference. We do not handle inflections of street names like *'Ligusterweg(e)s'* (genitive) because, in emails, street names most often are part of an address and thus lack inflection.

The list of surrogates is built from large repositories of German street names[11] jointly with Austrian ones from different provinces[12] that do not contain special characters or are composed of more than two terms (e.g., *'Albert-Einstein-Straße'*). Furthermore, we restricted them to contain only names with standard street suffixes (*'-straße'*, *'-weg'*, *'-platz'*, etc.), because we had to get rid of village names that do not have any named streets (such as with *'Wegscheid 15'*).

Given Names and Family Names. Common German proper nouns are singular and do not change number. (Duden, 2009, p. 191) Hence, our system does not process any plural forms of forenames or surnames (which rarely may occur), yet handles the genitive singular case. Therefore, our system is also capable of resolving coreferences between uninflected and inflected genitive forms.

To acquire suitable look-up dictionaries we extracted female and male names with their associated nicknames from a list of first names.[13] These lists are restricted to more or less common forenames in German-speaking countries except for names with rare first letters, where we included less frequent names, too. An alternative list of German surnames[14] is frequency-independent, thus includes also lots of uncommon names.

City names. For the CITY category, the genitive singular case is handled too. While person names and city, town or village names are mostly

[11] http://www.datendieter.de/item/Liste_von_deutschen_Strassennamen_.csv

[12] http://www.statistik.at/strasse/suchmaske.jsp

[13] ftp://ftp.heise.de/pub/ct/listings/0717-182.zip

[14] http://www.namenforschung.net/fileadmin/user_upload/dfa/Inhaltsverzeichnisse_etc/Index_Band_I-V_Gesamt_Stand_September_2016.pdf

used without determiner, a few German names for regions (e.g., *'die Steiermark'*, *'das Drautal'*) always require a determiner (Duden, 2009, pp. 299ff.). These names can be of every gender and also pluralia tantum, whereas city, town or village names are neuter and singularia tantum (Duden, 2009, pp. 160f.). Unfortunately, we currently do not dispose of repositories for gender- and number-specific CITY names large enough, so we rather tolerate possible mistakes than a potential information leakage.

In contrast to person names, we consider derivations of locations and implemented rules for signifying inhabitants (*'die Klagenfurterin'*) and adjectivized toponyms ending on *'-(i)sch/-erisch/-er'* (*'kärntnerisch'* (carinthian), *'Wiener Dialekt'* (Viennese dialect)). We check for coreferences using Levenshtein distance on previously seen CITYs. To catch lemmata occurring later we form appropriate candidates in a rule-based manner together with a lexicon look-up. For naming inhabitants, we only treat the standard forms ending on *'-er'* such as *'Wiener'* (Viennese) and do not care about non-standard names like *'Hesse'* (Hessian).

The generation of derivations from the substitute lemma is restricted to the most common cases. We produce derivatives by concatenating the lemma and *'-er'* or *'-r'*, if the lemma ends with *'-e'*. For adjectivized forms ending with *'-isch/-sch/-erisch'* that often allow a variation of these suffixes (Duden, 2016, p. 685) we decided on generating derivatives with *'-erisch'* (e.g., *'Wienerisch'* (Viennese)). Our system produces the inflectional forms for derivations by copying the original inflectional suffix to the generated form.

To maintain local national information we employed separate lists of location names for the three major German-speaking countries[15] on which we perform a dictionary look-up to determine which country the location name is from. Admittedly, this approach fails, if the place is either not mentioned or occurs in multiple countries. For substitution, we provide cleaner lists containing only villages, towns and cities for Germany,[16] Austria,[17] and Switzerland.[18]

Organizations. Similarly, we consider the genitive case for organizations. As the same company or institution in a document might be denoted by different name forms, such as its full name (*'Stadtwerke GmbH'*), with or without the corporate form (*'Stadtwerke'*) or an acronym (*'STW'*), with respect to coreference chains a more sophisticated solution is required. For now, we only check for names without a list of corporate forms. The substitution is performed with a list of German company names,[19] restricted to names not containing any GIVEN or FAMILY name. Due to gender variability and the lack of a list of institutions, we here added fictional acronyms and randomly generated letter combinations.

5.3 System Architecture

Our system for surrogate generation (see Figure 2) accepts any type of text, not only emails, but requires BRAT annotations of *pi*-relevant entities as described in Section 3. It allows for an easy adaptation to languages other than German, since the base module of the surrogate generation system can implement a language module for alternative languages, too. While language-independent categories (Section 5.1) do not need any further consideration, this language module has to provide allowed date formats and lists with language- and category-appropriate substitutes for the language-dependent classes (Section 5.2). Furthermore, frequency mappings of first letters may be specified in order to take distributions of names with respect

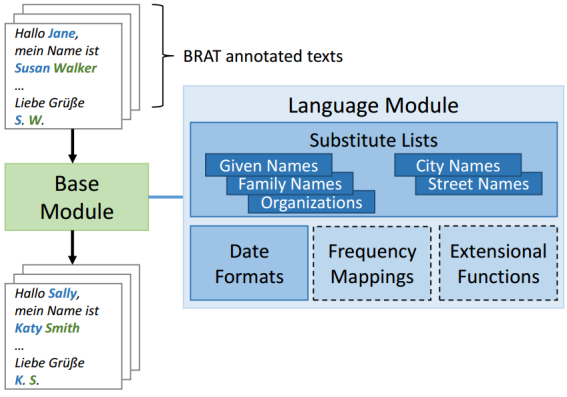

Figure 2: Schematic system architecture and surrogate generation workflow; dashed parts optional

[15] http://download.geonames.org/export/dump

[16] http://www.fa-technik.adfc.de/code/opengeodb/PLZ.tab

[17] http://www.statistik.at/strasse/suchmaske.jsp

[18] http://data.geo.admin.ch/ch.swisstopo-vd.ortschaftenverzeichnis_plz/PLZO_CSV_LV03.zip

[19] https://www.datendieter.de/item/Liste_von_deutschen_Firmennamen_.txt extracted from OpenStreetMap (http://www.openstreetmap.org/)

to their first letters into account. If a use case requires a special treatment of a category, extensional functions have to be defined, e.g., for inflection generation. Otherwise, the base system replaces the entities with default entries in the substitute lists, i.e., generally (non-inflected) lemmata.

6 Evaluation of Pseudonymization

6.1 Grammaticality and Acceptability Tests

In a first round of evaluations, we wanted to test whether the surrogates we had generated were well-formed in terms of grammaticality and semantically 'natural' in terms of acceptability. For privacy reasons, we refrained from explicitly evaluating whether coreference relations were preserved, because it is difficult to keep track of them without the original wording.

First, the pseudonymized emails were scored on both evaluation dimensions on a scale from 1 (worst) to 5 (best) in packages of about 30 emails by different annotators. Each email was annotated only once. While *grammaticality* refers to the agreement in number, gender and case between the generated surrogate and the sentence constituents the surrogate is embedded in, a surrogate is *acceptable* if it semantically fits into the surrounding context so that a reasonable semantic interpretation can be made. For example, *'We bought the car at Amici Pizza Express'* is considered as semantically inacceptable because cars normally cannot be bought at a pizzeria. In contrast common names replaced with rare ones (e.g., *'Paris'* with a small village name) are regarded as acceptable.

For the evaluation, the manually tagged 1,390 German-language emails of CODE ALLTAG$_{S+d}$ underwent the surrogate generation process twice to be sure not to reveal any original *pi* items to the annotators. The automatically generated output was checked manually, also substituting phrases not covered by our categories, such as course names (the corpus contains lots of university related emails donated by our students). After that it was fed again to the surrogate generation system.

With 4.90 for grammaticality and 4.73 for semantic acceptability the results are pretty sound. The lower outcome for semantic acceptability is mostly due to the surrogates for the ORG category. Occasionally, rather odd combination like *'Serial Knitters, IT solutions'* (for *'Institute for XX, YY-University'*) and substitutes with a different, inaccurate function were found.

6.2 Frequency Analysis

As Deléger et al. (2014) remark, large frequency imbalances between corpora may influence the performance of machine learning systems. Therefore, we also assessed frequency imbalances between different corpora: The original non-pseudonymized CODE ALLTAG$_{S+d}$ corpus with hand-annotated *pi* entities (referred to as ORIG), the pseudonymized[20] form of the original corpus (PSEUD) and pseudonymized[20] CODE ALLTAG$_{S+d}$ with automatically recognized *pi* entities (PSEUDPIR). For the latter, we retrieved the *pi* entities by training a model on $9/10$ of ORIG and applying it to the unseen part on ten different folds. We used a system for recognizing privacy-bearing information (PIR) based on NEURONER (Dernoncourt et al., 2017; Lee et al., 2016), with slight modifications of its neural network architecture from our side.

Table 3 shows that the number of tokens declines for both pseudonymized corpora compared to the original corpus. Taking a closer look, the category discrepancy almost entirely results from the category ORG and, to a much lesser extent though, also from STREET and CITY names. We conclude that our substitution dictionaries obviously contain shorter names, i.e., entities consisting of fewer tokens. Contrasting ORIG and the automatically annotated PSEUDPIR, we witness an increase of *pi* entities. For one thing, this can be explained by taking the tokens of one entity separately and thus splitting one single entity into multiple ones, which is also reflected in the smaller difference between the number of tokens. As this phenomenon especially occurs with URLs and PHONE or fax numbers (these categories are substituted randomly) it has no effect on surrogate generation. Again, ORGanizations play a major role here, because the PIR system achieves the worst results for this category.

	ORIG	PSEUD	PSEUDPIR
# token	151,229	150,166	150,425
# types[21]	21,159	22,320	22,455
# *pi* entities	8,866	8,866	9,427
# *pi* tokens	12,649	11,586	11,865

Table 3: Quantitative breakdown of the corpora used for evaluation; "#" stands for frequency count

[20] Processed by the surrogate generation system without any further reworking.

6.3 Automatic Recognition Performance

Further, we followed Yeniterzi et al. (2010) and Deléger et al. (2014) and tested the performance of the PIR system on the three corpora. Like Yeniterzi et al. (2010), we also found better results for training and testing on the PSEUD corpus than on ORIG,[22] while the performance difference decreases for PSEUDPIR and ORIG (see Table 4). Among others, this may be a consequence of the frequency imbalance, because the pseudonymized data contain fewer tokens especially those related to the hard to recognize ORGanizations.

Corpus	Prec	Rec	F_1	p-value
ORIG	83.02	82.26	82.52	
PSEUD	86.15	86.24	86.07	0.007
PSEUDPIR	82.26	85.79	83.86	0.063

Table 4: Results of the PIR system on different corpora (10-fold cross validation); significance difference (p-value) with respect to ORIG (paired t-test)

For training and testing on different corpora, we eliminated the emails found in the test set from the train set in a 10-fold cross-validation manner, even if they were pseudonymized in one of the datasets to avoid any overlap. Again, similar to Yeniterzi et al. (2010), training on ORIG and testing on PSEUD yields significantly better results than the other way round (see Table 5). Also the F_1 score plunges deeply when training on PSEUDPIR and testing on ORIG compared to ORIG and testing on PSEUDPIR.

If we subsume the language-independent categories, because they are randomized and treated similarly in surrogate generation, we get comparable outcomes for all experiments with the PIR system with an F_1 score around 90.00. In contrast, language-dependent categories, with the exception of DATEs, which accomplish equivalent results for nearly all experiments, too, consistently perform worse compared to training and testing on the same corpus, especially regarding ORGanizations, CITYs and STREETs. As the results in Table 3 reveal, the amount of word types is higher in the pseudonymized corpora; hence, they are more diverse. Further, they may contain rarer names. Both phenomena potentially lead to a decrease of performance on these data sets when trained on the ORIG corpus. Regarding the drop of the reverse experiment the fewer occurrences of 'I'-tags (from the BIO format) have an impact, too.

Train|Test	Prec	Rec	F_1	p
ORIG|PSEUD	77.97	73.12	74.93	
PSEUD|ORIG	70.25	61.32	64.57	0.0
ORIG|PSEUDPIR	85.23	75.21	78.39	
PSEUDPIR|ORIG	67.31	63.21	63.88	0.0

Table 5: Results of the PIR system trained on Train and tested on Test (10-fold cross validation without overlap); significance difference (p-value) over reverse setting (paired t-test)

In contrast to Yeniterzi et al. (2010) and our results, Deléger et al. (2014) report a smaller performance difference between training on original and testing on pseudonymized data and vice versa. This is probably due to the fact that they replaced pi entities with different entities of the same category taken from the same original corpus, thus almost retaining the original personal information. This approach bears the potential of causing a leak of personal information due to categories with limited occurrences.

7 Conclusion

In this paper, we moved the de-identification problem out of the medical domain (cf. also our work in this field described in Kolditz et al. (2019)) into the realm of user-generated content, emails in our use case. In particular, we focused on the surrogate generation step of that task, i.e., substituting named entities bearing privacy-relevant information by synthetic, yet mostly natural surrogates.

Our main contributions are the specification of a language-independent type hierarchy composed of named entities that carry privacy-relevant information, and the realization of the first German non-medical surrogate generation pipeline. It is composed of a language-dependent part for German input and a language-independent one which can readily be reused for languages other than German, without any changes.

We also ran a series of experiments on emails from the German-language CODE ALLTAG$_{S+d}$ corpus. In this evaluation of our surrogate generation system, we found high scores for the grammaticality and acceptability of the automatically

[21] Types of tokens excluding punctuation and stop words.

[22] When the k-fold splits are performed on sentences rather than on emails the results get notably better achieving 86.24 F_1 score for ORIG. But for reasons of comparability between non-pseudonymized and pseudonymized corpora we had to split on emails.

generated surrogates. A frequency analysis of different variants of CODE ALLTAG$_{S+d}$ revealed a quantitative imbalance between the original corpus, the pseudonymized one, and a de-identified variant that was built using an automatic recognizer for privacy-relevant named entities. Experiments on these three corpora further exposed differences in recognition performance already discussed in the literature.

Our future work will focus on a more adequate treatment of German derivational morphology and coreferences rooted in varying spellings. The main methodological desideratum concerns the investigation of ways to deal with organizations with different functions in order to improve semantic acceptability. Last but not least, we will have to demonstrate that the results from the small-scale corpus we currently dealt with (CODE ALLTAG$_{S+d}$) will scale up to much larger document collections (e.g., CODE ALLTAG$_{XL}$ as described in Krieg-Holz et al. (2016)).

Acknowledgments

We want to thank our RANLP reviewers for pointing out improvements of the original submission. These hints were gratefully acknowledged and incorporated in the final version of the paper.

This work was funded by the Austrian HRSM project *Kompetenznetzwerk Digitale Edition* (KONDE).

References

Eytan Adar. 2007. User 4XXXXX9: anonymizing query logs. In *Proceedings of the Workshop on Query Log Analysis: Social and Technological Challenges @ WWW 2007. Banff, Alberta, Canada, May 8, 2007*.

Alyaa Alfalahi, Sara Brissman, and Hercules Dalianis. 2012. Pseudonymisation of personal names and other PHIs in an annotated clinical Swedish corpus. In *BioTxtM 2012 — Proceedings of the 3rd Workshop on Building and Evaluating Resources for Biomedical Text Mining @ LREC 2012. Istanbul, Turkey, May 26, 2012*. pages 49–54.

David S. Carrell, Bradley A. Malin, John S. Aberdeen, Samuel Bayer, Cheryl Clark, Benjamin Wellner, and Lynette Hirschman. 2013. Hiding In Plain Sight: use of realistic surrogates to reduce exposure of protected health information in clinical text. *Journal of the American Medical Informatics Association* 20(2):342–348.

Tao Chen and Min-Yen Kan. 2013. Creating a live, public short message service corpus: the NUS SMS corpus. *Language Resources and Evaluation* 47(2):299–335.

Maximin Coavoux, Shashi Narayan, and Shay B. Cohen. 2018. Privacy-preserving neural representations of text. In *EMNLP 2018 — Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium, October 31 - November 4, 2018*. pages 1–10.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20(1):37–46.

Louise Deléger, 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, Özlem 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.

Duden. 2009. *Die Grammatik: Unentbehrlich für richtiges Deutsch*, volume 4 of *Der Duden in 12 Bänden*. Dudenverlag, Mannheim etc., 8th edition.

Duden. 2016. *Das Wörterbuch der sprachlichen Zweifelsfälle: Richtiges und gutes Deutsch*, volume 9 of *Der Duden in 12 Bänden*. Dudenverlag, Berlin, 8th edition.

Matthew Honnibal and Ines Montani. 2017. SPACY 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. https://spacy.io/.

Chris Jay Hoofnagle, Bart van der Sloot, and Frederik Zuiderveen Borgesius. 2019. The European Union general data protection regulation: what it is and what it means. *Information & Communications Technology Law* 28(1):65–98.

Bryan Klimt and Yiming Yang. 2004. The ENRON corpus: a new dataset for email classification research. In *ECML 2004 – Proceedings of the 15th European Conference on Machine Learning, Pisa, Italy, September 20-24, 2004*. Springer, pages 217–226.

Tobias Kolditz, Christina Lohr, Johannes Hellrich, Luise Modersohn, Boris Betz, Michael Kiehntopf, and Udo Hahn. 2019. Annotating German clinical documents for de-identification. In *MedInfo 2019 — Proceedings of the 17th World Congress on Medical and Health Informatics. Lyon, France, 25-30 August 2019*. IOS Press.

Ulrike Krieg-Holz, Christian Schuschnig, Franz Matthies, Benjamin Redling, and Udo Hahn. 2016. CODE ALLTAG: a German-language e-mail corpus.

In *LREC 2016 — Proceedings of the 10th International Conference on Language Resources and Evaluation. Portorož, Slovenia, 23-28 May 2016*. pages 2543–2550.

Ji Young Lee, Franck Dernoncourt, Özlem Uzuner, and Peter Szolovits. 2016. Feature-augmented neural networks for patient note de-identification. In *ClinicalNLP 2016 — Proceedings of the Clinical Natural Language Processing Workshop @ COLING 2016. Osaka, Japan, December 11, 2016*. pages 17–22.

Yitong Li, Timothy Baldwin, and Trevor Cohn. 2018. Towards robust and privacy-preserving text representations. In *ACL 2018 — Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, Victoria, Australia, July 15-20, 2018*. volume 2, pages 25–30.

Zengjian Liu, Buzhou Tang, Xiaolong Wang, and Qingcai Chen. 2017. De-identification of clinical notes via recurrent neural network and conditional random field. *Journal of Biomedical Informatics* 75(Supplement):S34–S42.

Harald Lüngen, Michael Beißwenger, Laura Herzberg, and Cathrin Pichler. 2017. Anonymisation of the Dortmund Chat Corpus 2.1. In *cmccorpora17 — Proceedings of the 5th Conference on CMC and Social Media Corpora for the Humanities. Bolzano, Italy, October 3-4, 2017*. pages 21–24.

Ben Medlock. 2006. An introduction to NLP-based textual anonymisation. In *LREC 2006 — Proceedings of the 5th International Conference on Language Resources and Evaluation. Genoa, Italy, 22-28 May, 2006*. pages 1051–1056.

Stéphane M. Meystre. 2015. De-identification of unstructured clinical data for patient privacy protection. In Aris Gkoulalas-Divanis and Grigorios Loukides, editors, *Medical Data Privacy Handbook*, Springer International Publishing, pages 697–716.

Stephen P. Mulligan, Wilson C. Freeman, and Chris D. Linebaugh. 2019. Data protection law: an overview. Technical Report CRS Report R45631, Congressional Research Service.

Kostas Pantazos, Sören Lauesen, and Sören Lippert. 2011. De-identifying an EHR database: anonymity, correctness and readability of the medical record. In *MIE 2011 — Proc. of the 23rd Conference of the European Federation of Medical Informatics. Oslo, Norway, August 28-31, 2011*. pages 862–866.

Namrata Patel, Pierre Accorsi, Diana Z. Inkpen, Cédric Lopez, and Mathieu Roche. 2013. Approaches of anonymisation of an SMS corpus. In *CICLing 2013 — Proceedings of the 14th International Conference on Computational Linguistics and Intelligent Text Processing. Karlovasi, Samos, Greece, March 24-30, 2013*. Springer, pages 77–88.

Frances Rock. 2001. Policy and practice in the anonymisation of linguistic data. *International Journal of Corpus Linguistics* 6(1):1–26.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. BRAT: a Web-based tool for NLP-assisted text annotation. In *EACL 2012 — Proc. of the 13th Conf. of the European Chapter of the Association for Computational Linguistics: Demonstrations. Avignon, France, April 25-26, 2012*. pages 102–107.

Amber Stubbs, Michele Filannino, and Özlem Uzuner. 2017. De-identification of psychiatric intake records: overview of 2016 CEGS NGRID Shared Tasks Track 1. *Journal of Biomedical Informatics* 75(Supplement):S4–S18.

Amber Stubbs, Christopher Kotfila, and Özlem Uzuner. 2015a. Automated systems for the de-identification of longitudinal clinical narratives: overview of 2014 i2b2/UTHealth Shared Task Track 1. *Journal of Biomedical Informatics* 58(Supplement):S11–S19.

Amber Stubbs and Özlem Uzuner. 2015. Annotating longitudinal clinical narratives for de-identification: the 2014 i2b2/UTHealth corpus. *Journal of Biomedical Informatics* 58(Supplement):S20–S29.

Amber Stubbs, Özlem Uzuner, Christopher Kotfila, Ira Goldstein, and Peter Szolovits. 2015b. Challenges in synthesizing surrogate PHI in narrative EMRs. In Aris Gkoulalas-Divanis and Grigorios Loukides, editors, *Medical Data Privacy Handbook*, Springer International Publishing, pages 717–735.

Latanya Sweeney. 1996. Replacing personally-identifying information in medical records, the SCRUB system. In *AMIA '96 — Proceedings of the 1996 AMIA Annual Fall Symposium. Washington, D.C., USA, October 26-30, 1996*. pages 333–337.

Latanya Sweeney. 2002. k-anonymity: a model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-based Systems* 10(5):557–570.

Maaske Treurniet, Orphée De Clercq, Henk van den Heuvel, and Nelleke Oostdijk. 2012. Collecting a corpus of Dutch SMS. In *LREC 2012 — Proceedings of the 8th International Conference on Language Resources and Evaluation. Istanbul, Turkey, May 21-27, 2012*. pages 2268–2273.

Özlem Uzuner, Yuan Luo, and Peter Szolovits. 2007. Evaluating the state-of-the-art in automatic de-identification. *Journal of the American Medical Informatics Association* 14(5):550–563.

Shomir Wilson et al. 2016. The creation and analysis of a website privacy policy corpus. In *ACL 2016 — Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany, August 7-12, 2016*. pages 1330–1340.

Reyyan Yeniterzi, John S. Aberdeen, Samuel Bayer, Benjamin Wellner, Lynette Hirschman, and Bradley A. Malin. 2010. Effects of personal identifier resynthesis on clinical text de-identification. *Journal of the American Medical Informatics Association* 17(2):159–168.