Anonymisation Models for Text Data: State of the Art, Challenges and Future Directions

Pierre Lison[1], Ildikó Pilán[1], David Sánchez[2], Montserrat Batet[2], and Lilja Øvrelid[3]

[1]Norwegian Computing Center, Oslo, Norway
[2]Universitat Rovira i Virgili, CYBERCAT, UNESCO Chair in Data Privacy, Spain
[3]Language Technology Group, University of Oslo, Norway

{plison,pilan}@nr.no {david.sanchez,montserrat.batet}@urv.cat liljao@ifi.uio.no

Abstract

This position paper investigates the problem of automated *text anonymisation*, which is a prerequisite for secure sharing of documents containing sensitive information about individuals. We summarise the key concepts behind text anonymisation and provide a review of current approaches. Anonymisation methods have so far been developed in two fields with little mutual interaction, namely *natural language processing* and *privacy-preserving data publishing*. Based on a case study, we outline the benefits and limitations of these approaches and discuss a number of open challenges, such as (1) how to account for multiple types of semantic inferences, (2) how to strike a balance between disclosure risk and data utility and (3) how to evaluate the quality of the resulting anonymisation. We lay out a case for moving beyond sequence labelling models and incorporate explicit measures of *disclosure risk* into the text anonymisation process.

1 Introduction

Privacy is a fundamental human right (Art. 12 of the Universal Declaration of Human Rights) and a critical component of any free society, among others to protect citizens against social control, stigmatisation, and threats to political expression. Privacy is also protected by multiple national and international legal frameworks, such as the General Data Protection Regulation (GDPR) introduced in Europe in 2018. This right to privacy imposes constraints on the usage and distribution of data including personal information, such as emails, court cases or patient records. In particular, personal data cannot be distributed to third parties (or even used for secondary purposes) without legal ground, such as the explicit and informed consent of the individuals to whom the data refers.

As informed consent is often difficult to obtain in practice, an alternative is to rely on *anonymisation techniques* that render personal data no longer personal. Access to anonymised data is a prerequisite for research advances in many scientific fields, notably in medicine and the social sciences. By facilitating open data initiatives, anonymised data can also help empower citizens and support democratic participation. For structured databases, anonymisation can be enforced through well-established privacy models such as k-anonymity (Samarati, 2001; Samarati and Sweeney, 1998) or differential privacy (Dwork et al., 2006). These privacy models and their implementations are, however, difficult to apply to unstructured data such as texts. In fact, text anonymisation has been traditionally enforced manually, a process that is costly, time-consuming and prone to errors (Bier et al., 2009). These limitations led to the development of various computational frameworks designed to extend automated or semi-automated anonymisation to the text domain (Meystre et al., 2010; Sánchez and Batet, 2016; Dernoncourt et al., 2017).

In this paper, we review the core concepts underlying text anonymisation, and survey the approaches put forward to solve this task. These can be divided into two independent research directions. On the one hand, NLP approaches rely on *sequence labelling* to detect and remove predefined categories of entities that are considered sensitive or of personal nature (such as names, phone numbers or medical conditions). On the other hand, privacy-preserving data publishing (PPDP) approaches take the notion of *disclosure risk* as starting point and anonymise text by enforcing a privacy model. Anonymisation consists of a sequence of transformations (such as removal or generalisation) on the document to ensure the requirements derived from the privacy model are fulfilled.

This position paper makes the case that none of these approaches provide a fully satisfactory account of the text anonymisation problem. We illustrate their merits and shortcomings on a case study and discuss three open challenges:

1. How to ensure that anonymisation is robust against multiple types of semantic inferences, based on background knowledge assumed to be available to an adversary ;
2. How to transform the text in order to minimise the risk of disclosing personal data, yet retain as much semantic content as possible ;
3. How to empirically evaluate the quality (in terms of disclosure risk and utility preservation) of the resulting anonymisation.

We argue in this paper that NLP and PPDP approaches should be viewed as complementary (one focusing on linguistic patterns, the other on disclosure risk) and that future anonymisation approaches for text should seek to reconcile these two views. In particular, we contend that text anonymisation models should combine a data-driven *editor model* (which selects masking operations on the document) with an *adversary* seeking to infer confidential attributes from edited documents.

2 What is Anonymisation?

The most common definition of privacy amounts to self-determination, which is the ability of individuals, groups or organisations to seclude information about themselves selectively (Westin, 1967). Information related to an identified or identifiable person is known as personal data, or more precisely *personally identifiable information* (PII). Datasets with PII cannot be released without control as this would impair the privacy of the data subjects.

2.1 Legal Requirements

Various legal frameworks regulate how PII can be collected and processed. In particular, the General Data Protection Regulation introduced in Europe (GDPR, 2016) states that data owners must have a legal basis for processing PII, the most important one being the explicit consent of the data subjects.Alternatively, data owners may choose to *anonymise* the data to ensure it can no longer be attributed to specific individuals. Anonymised data is no longer regulated by the GDPR and can therefore be freely released.

Table 1 defines some of the key terms related to data anonymisation (Elliot et al., 2016). This terminology is, however, not always applied consistently, as several authors seem to use e.g. the terms “anonymisation” and “de-identification” interchangeably (Chevrier et al., 2019).

Direct Identifier: A (set of) variable(s) unique for an individual (a name, address, phone number or bank account) that may be used to directly identify the subject.

Quasi Identifier: Information (such as gender, nationality, or city of residence) that in isolation does not enable re-identification, but may do so when combined with other quasi-identifiers and background knowledge.

Confidential Attribute: Private personal information that should not be disclosed (such as a medical condition).

Identity Disclosure: Unequivocal association of a record/document with a subject’s identity.

Attribute disclosure: Unequivocal inference of a confidential attribute about a subject.

Anonymisation: Complete and irreversible removal from a dataset of any information that, directly or indirectly, may lead to a subject’s data being identified.

De-identification: Process of removing specific, predefined direct identifiers from a dataset.

Pseudonymisation: Process of replacing direct identifiers with pseudonyms or coded values (such ”John Doe” → ”Patient 3”). The mapping between coded values and the original identifiers is then stored separately.

Table 1: Key terms related to data anonymisation.

GDPR-compliant anonymisation is the *complete* and *irreversible* process of removing personal identifiers, both direct and indirect, that may lead to an individual being identified. *Direct identifiers* correspond to values such as names or social security numbers that directly disclose the identity of the individual. However, removing direct identifiers is not sufficient to eliminate all disclosure risks, as individuals may also be re-identified by combining several pieces of information together with some background knowledge. For instance, the combination of gender, birth date and postal code can be exploited to identify between 63 and 87% of the U.S. population, due to the public availability of US Census Data (Golle, 2006). These types of personal identifiers are called *quasi-identifiers* and encompass a large variety of data types such as demographic and geospatial data. Anonymisation therefore necessitates both the removal of direct identifiers and the masking of quasi-identifiers.

Other legal frameworks have adopted a different approach. In the US, the Health Insurance Portability and Accountability Act (HIPAA) (HIPAA, 2004) lists 18 data types, such as patient's name, address or social security number, which qualify as *protected health information* (PHI) and should be removed from the data prior to release. This process of removing predefined categories of identifiers is called *de-identification*[1]. In other words, while HIPAA-based de-identification is limited to specific categories of direct identifiers, the anonymisation process defined by GDPR requires us to consider any direct or indirect information that, combined with background knowledge, may lead to re-identifying an individual. The California Consumer Privacy Act (CCPA) introduced in 2018 adopts a position relatively similar to GDPR regarding anonymisation and asserts that any data that can be linked directly or indirectly to a consumer must be considered as personal information.

We highlight these legal differences as they have important implications on how anonymisation tools should be designed and evaluated (Rothstein, 2010; Hintze, 2017). In particular, GDPR- or CCPA-compliant anonymisation cannot be restricted to the detection of predefined classes of entities but must consider how any textual element may contribute to the disclosure risk, either directly or through semantic inferences using the background knowledge assumed to be available to an adversary.

2.2 Disclosure Risks

Legal regulations for privacy and data protection (such as GDPR and HIPAA) typically focus on *identity disclosure*. However, personal information may also be disclosed without re-identification. In particular, *attribute disclosure* occurs when the value of a confidential attribute (e.g., a medical condition) can be inferred from the released data, for instance when all records sharing some characteristics (e.g. age) have the same confidential value (e.g. suffering from AIDS). Identity disclosure can be seen as a special case of attribute disclosure when the confidential attribute corresponds to the person identity. Data anonymisation should prevent identity disclosure but, in most cases, attribute disclosure, which is usually more harmful from a privacy perspective, should also be avoided.

The removal of personal information necessarily entails some data utility loss. Because the ultimate purpose behind data releases is to produce usable data, the best anonymisation methods are those that optimise the trade-off between minimising the disclosure risk and preserving the data utility.

3 NLP Approaches

3.1 De-identification

NLP research on text anonymisation has focused to a large extent on the tasks of de-identification, and, to a lesser extent, pseudonymisation. De-identification is generally modelled as a sequence labelling task, similar to Named Entity Recognition (NER) (Chiu and Nichols, 2016; Lample et al., 2016). Most work to date has been performed in the area of clinical NLP, where the goal is to detect Protected Health Information (PHI) in clinical texts (Meystre et al., 2010; Aberdeen et al., 2010). Several shared tasks have contributed to increased activity within this area, in particular through the release of datasets manually annotated with PHIs. The 2014 i2b2/UTHealth shared task (Stubbs and Uzuner, 2015) includes diabetic patient medical records annotated for an extended set of PHI categories. Another influential dataset stems from the 2016 CEGS N-GRID shared task (Stubbs et al., 2017) based on psychiatric intake records, which are particularly challenging to de-identify due to a higher density of PHIs.

Early approaches to this task were based on rule-based and machine learning-based methods, either alone or in combination (Yogarajan et al., 2018). Dernoncourt et al. (2017) and Liu et al. (2017) present the first neural models for de-identification using recurrent neural networks with character-level embeddings, achieving state-of-the-art performance on the i2b2 2014 dataset.

A central challenge in clinical de-identification is the availability of annotated data and the lack of universal annotation standards for PHI, making it difficult to transfer data across domains. Hartman et al. (2020) examine how to adapt de-identification systems across clinical sub-domains. They compare the use of labelled or unlabelled data for domain adaptation with in-domain testing and off-the-shelf de-identification tools, and show that manual labelling of even small amounts of PHI examples yields performance above existing tools.

[1] GDPR also introduces the equivalent concept of *pseudonymisation*, which is a useful privacy-enhancing measure, but it does not qualify as full anonymisation.

Further, embeddings trained on larger amounts of in-domain, unlabelled data can be employed to adapt models to a new domain (Yang et al., 2019). Finally, Friedrich et al. (2019) present an adversarial approach for learning privacy-preserving text representations, thereby allowing data to be more easily shared to train de-identification tools.

Outside of the clinical domain, Medlock (2006) presents a dataset of e-mails annotated with both direct identifiers (person names, transactional codes, etc.) and quasi-identifiers (organisations, course names, etc.). Some annotation efforts are also geared towards de-identification for languages other than English. Eder et al. (2020) present a de-identification dataset consisting of German e-mails. For Swedish, Velupillai et al. (2009); Alfalahi et al. (2012) present efforts to collect and standardise annotated clinical notes, while Megyesi et al. (2018) present a pseudonymised learner language corpus. For Spanish, a recently held shared task on clinical de-identification released a synthetic Spanish-language dataset (Marimon et al., 2019).

The problem of replacing identifiers with surrogate values is rarely addressed in NLP. Most approaches simply replace detected identifiers with dummy values such as *X*, although some models attempt to preserve the gender of person names and provide dedicated rules for e.g. dates and addresses (Sweeney, 1996; Alfalahi et al., 2012; Eder et al., 2019; Chen et al., 2019) or to a somewhat broader range of identifiers (Volodina et al., 2020).

A few studies have analysed the re-identification risk of de-identified or pseudonymised texts (Carrell et al., 2013; Meystre et al., 2014b). The data utility of de-identified texts is analysed in Meystre et al. (2014a), concluding that the impact of de-identification is small, but non-negligible.

3.2 Obfuscation Methods

Beyond de-identification, several research efforts have looked at detecting and obfuscating social media texts based on quasi-identifying categories such as gender (Reddy and Knight, 2016) or race (Blodgett et al., 2016). A number of recent approaches have sought to transform latent representations of texts to protect confidential attributes, using adversarial learning (Elazar and Goldberg, 2018), reinforcement learning (Mosallanezhad et al., 2019) or encryption (Huang et al., 2020). However, those methods operate at the level of latent vector representations and do not modify the texts themselves. One notable exception is the text rewriting approach of Xu et al. (2019) which edits the texts using back-translations.

3.3 Challenges

NLP approaches to anonymisation suffer from a number of shortcomings. Most importantly, they are limited to predefined categories of entities and ignore how less conspicuous text elements may also play a role in re-identifying the individual. For instance, the family status or physical appearance of a person may lead to re-identification but will rarely be considered as categories to detect. On the other hand, those methods may also end up removing *too much* information, as they will systematically remove all occurrences of a given category without examining their impact on the disclosure risk or on the utility of the remaining text.

4 PPDP Approaches

Privacy-preserving data publishing (PPDP) develops computational techniques for releasing data without violating privacy (Chen et al., 2009).

The PPDP approach to anonymisation is privacy-first: a *privacy model* specifying an *ex ante* privacy condition is enforced through one or several data masking methods, such as noise addition or generalisation of values (Domingo-Ferrer et al., 2016). The first widely-accepted privacy model is k-anonymity (Samarati, 2001): a dataset satisfies k-anonymity if each combination of values of quasi-identifier attributes is shared by at least k records. With $k > 1$, no unequivocal re-identifications are possible, thereby preventing identity disclosure.

Most of the attention of the PPDP community has been on structured databases. Privacy models such as k-anonymity assume that datasets consist of records, each one detailing the attributes of a single individual, and that attributes have been classified beforehand into identifiers, quasi-identifiers and confidential attributes. Moreover, most masking methods employed to enforce privacy models have been designed with numerical data in mind, and barely (and poorly) manage categorical or nominal attributes (Rodríguez-García et al., 2019).

4.1 k-anonymity and Beyond

Solutions for anonymising unstructured text are scarce and mostly theoretical. The first approaches adapted k-anonymity for collections of documents. In (Chakaravarthy et al., 2008), the authors presented the notion of K-safety. They assume a collection of entities e to be protected against disclosure, each one characterised by a set of terms $C(e)$ that represent their contexts (i.e. words co-occurring with e and that may be known to an attacker). Then, a document D containing an entity e is said to be K-safe if the terms appearing in D also belong to the contexts of, at least, $K-1$ entities other than e. Terms not fulfilling the property are redacted before release. The privacy guarantee offered by this approach is sound because the probability of disclosing the protected entity is reduced to $1/K$. However, it requires exhaustive collections of contexts for all entities to be protected, which is unfeasible. It also assumes that the detection of sensitive terms is already performed. This approach is only feasible for very constrained domains and non-dynamic sets of entities, such as collections of sensitive diseases, and documents with homogeneous contents.

Another approach built on k-anonymity is Cumby and Ghani (2011), where a multi-class classifier is trained to map input documents to (predefined) sensitive entities. This aims at reproducing the inferences that a potential attacker may perform to disclose sensitive entities. A document x referring to a sensitive entity y is then said to be K-confusable if the classifier outputs at least k classes other than y. Documents are redacted via term removal or generalisation until the property is fulfilled. To be applicable, sensitive entities should be static and the documents to be protected should match that of the corpus used for training.

Anandan et al. (2012) present a privacy model for document protection named t-plausibility. They seek to generalise terms identified as sensitive according to the t-plausibility property: a protected document is said to fulfil t-plausibility if, at least, t different *plausible* documents can be derived by specialising the generalised terms. In other words, Even though the privacy guarantee is intuitive, one can hardly predict the results for a certain t, because they depend on the document length, the number of sensitive entities and the granularity of the knowledge base employed to obtain term generalisations. Assuming that sensitive entities have already been detected also circumvents the most challenging task of document protection.

4.2 C-sanitise

Sánchez and Batet (2016, 2017) tackles the anonymisation problem from a different perspective. Instead of expressing privacy guarantees in terms of probability of disclosure, it defines risk as an information theoretic characterisation of disclosed semantics. The proposed privacy model, C-sanitise, states that given a document d, background knowledge K available to potential attackers, and a set of entities to protect C, d' is the C-sanitised version of d if d' does not contain any term t that, individually or in aggregate, unequivocally disclose the semantics encompassed by any entity in C by exploiting K. The semantic disclosure incurred by t on any entity in C is quantified as their pointwise mutual information (Anandan and Clifton, 2011) measured from their probability of (co-)occurrence in the Web, which is assumed to represent the most comprehensive knowledge source (K) available to attackers (Chow et al., 2008). This approach is able to automatically detect terms that may cause disclosure and can encompass dynamic collections of entities to protect. Obtaining accurate probabilities of co-occurrence from large corpora is, however, costly.

4.3 Differential Privacy

Differential privacy (DP) is a privacy model that defines anonymisation in terms of randomised algorithms for computing statistics from the data (Dwork et al., 2006). DP provides guarantees that the statistics cannot be used to learn anything substantial about any individual. However, the goal of DP is to produce randomised responses to controlled queries, and applying it to data publishing leads in poor data utility (Domingo-Ferrer et al., 2021). DP cannot be directly employed to edit out personal information from text while preserving the content of the rest of the document, and is thus outside the scope of this paper. However, DP can be employed for other privacy-related tasks such as in producing synthetic texts (Fernandes et al., 2018; Bommasani et al., 2019), deriving differentially-private word representations (Feyisetan et al., 2019) or learning machine learning models with privacy guarantees (McMahan et al., 2017).

4.4 Challenges

Compared to NLP approaches, proposals built around privacy models allow defining what should be protected and how. This not only allows enforcing privacy requirements, but also makes it possible to tailor the trade-off between data protection and utility preservation. On the negative side, PPDP methods are hampered by practical constraints, either because of their unfeasible assumptions, their cost or their dependency on external resources, such as large knowledge repositories, training corpora or social-scale probabilities. To the exception of C-sanitise, PPDP methods also assume that sensitive entities have already been detected in a preprocessing step. Furthermore, PPDP approaches typically reduce documents to flat collections of terms, which facilitates the formalisation of the data semantics for each document, but also ignores how terms are influenced by their context of occurrence (which is important to resolve potential ambiguities) and are interconnected through multiple layers of linguistic structures.

5 Case Study

To investigate the performance of NLP and PPDP methods, we carried out a case study where 5 annotators annotated 8 English Wikipedia page extracts. The extracts were all biographies from the "*20th century scientists*" category, with a length between 300 and 500 characters. Wikipedia articles are generic enough not to require expert domain knowledge and are commonly adopted for the evaluation of PPDP approaches (Chow et al., 2008; Sánchez and Batet, 2016). Their informativeness and density make them particularly challenging to anonymise.

The annotation task[2] consisted of tagging text spans that could re-identify a person either directly or in combination with publicly available knowledge. The annotators were instructed to prevent identity disclosure but otherwise seek to preserve as much semantic content as possible. The five annotators were researchers without previous experience in text anonymisation. The guidelines were left intentionally general to examine how annotators interpret and carry out the complex task of anonymisation – and not only de-identification – where multiple correct solutions are possible.

The task is challenging since these biographies relate to publicly known scientists for which extensive background material can be found online. Inter-rater agreement between the five annotators for the binary masking decisions was low: 0.68 average observed agreement and Krippendorff's $\alpha = 0.36$. This low agreement illustrates that, contrary to traditional sequence labelling, several solutions may exist for a given anonymisation problem. Direct identifiers were generally agreed on, while quasi-identifiers such as professions and roles (e.g. *founder*) triggered mixed decisions.

To shed further light on the anonymisation problem, we go on to compare the performance of existing tools with the manual annotations:

- A neural NER model (Honnibal and Montani, 2017) trained on the OntoNotes corpus with 18 entity types (Weischedel et al., 2011). All detected entities were masked.[3]

- Presidio[4], a data protection & anonymisation API developed by Microsoft and relying on a combination of template-based and machine learning models to detect and mask PII.

- The C-sanitise privacy model (Sánchez and Batet, 2016) described in Section 4, where the required probabilities of (co-)occurrence of terms were gathered from Google.

5.1 Metrics

To account for the multiple ways to anonymise a document, we measured the performance of the three tools above with *micro-averaged* scores over all annotators and texts. Note that, while micro-averages are typically used in NLP to aggregate measures over output classes, we are here computing an average over multiple *ground truths*.

For each annotator $q \in Q$ and document $d \in D$, let Y_d^q correspond to token indices masked by q in d, and $\hat{Y}_d$ to the token indices masked by the anonymisation tool. Precision and recall are then computed as:

$$P = \frac{\sum_{d \in D} \sum_{q \in Q} |\hat{Y}_d \cap Y_d^q|}{|Q| \sum_{d \in D} |\hat{Y}_d|} \tag{1}$$

$$R = \frac{\sum_{d \in D} \sum_{q \in Q} |\hat{Y}_d \cap Y_d^q|}{\sum_{d \in D} \sum_{q \in Q} |Y_d^q|} \tag{2}$$

An anonymisation tool will thus obtain a perfect micro-averaged recall if it detects all tokens masked by at least one annotator. The metric implicitly assigns a higher weight to tokens masked by several annotators – in other words, if all five annotators mask a given token, not detecting it will have a

[2] The guidelines and annotated data are publicly available: https://github.com/IldikoPilan/anonymisation_ACL2021

[3] Although NERs do not specifically focus on data protection, they are often used to de-identify generic texts (except clinical notes, for which domain-specific tools are available).

[4] https://github.com/microsoft/presidio

		P	R	F_1
NER	IOB-Exact	0.5	0.49	0.47
	IOB-Partial	0.61	0.48	0.54
	Binary	0.64	0.51	0.57
Presidio	IOB-Exact	0.63	0.22	0.33
	IOB-Partial	0.74	0.24	0.36
	Binary	0.76	0.25	0.38
C**-sanitise**	IOB-Exact	0.51	0.66	0.57
	IOB-Partial	0.57	0.68	0.62
	Binary	0.58	0.69	0.63

Table 2: Micro-averaged scores for NER, C-sanitise and Presidio over all texts for annotators a1, a4, a5.

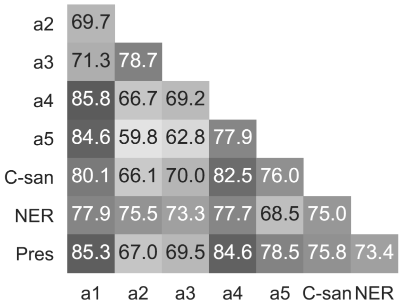
larger impact on the recall than a token masked by a single annotator. Recall expresses the level of privacy protection while precision is related to the degree of utility preservation.

The most consistent manual annotations (a1, a4, a5) were compared to system outputs at token level both as binary labels (*keep* or *mask*) and as *IOB* tags expressing annotation spans[5]. To go beyond token-level comparisons, we also computed a partial match score for IOB tags, by assigning a weight of 0.5 to partial true positives (i.e. I instead of B tags and vice versa), as in the SemEval 2013 evaluation scheme (Diab et al., 2013).

5.2 Results and Error Analysis

Table 2 presents the micro-averaged precision, recall and F_1 scores obtained for the three systems.

C-sanitise provided the best performance in terms of recall and F_1 score, while precision was higher for NER and Presidio. Figure 1 illustrates the average observed agreement for all annotators and tools on the binary, token-level masking decisions. Observed agreement with annotators was, on average, approximately the same for NER and C-sanitise, ca. 75% and ca. 77% for Presidio. We can distinguish two subgroups among the annotators in terms of mutual agreement, namely (a2, a3) and (a1, a4, a5) with 79% and 83% agreement respectively. Divergent choices in entity segmentation – e.g. splitting a consecutive mention of department and university or not – was found to play an important role in the differences among annotators, and between annotators and systems.



	a1	a2	a3	a4	a5	C-san	NER
a2	69.7						
a3	71.3	78.7					
a4	85.8	66.7	69.2				
a5	84.6	59.8	62.8	77.9			
C-san	80.1	66.1	70.0	82.5	76.0		
NER	77.9	75.5	73.3	77.7	68.5	75.0	
Pres	85.3	67.0	69.5	84.6	78.5	75.8	73.4

Figure 1: Pairwise average observed agreement. a1 to a5 correspond to the human annotators.

The proportion of masked tokens was around 50% for a1, a2 and C-sanitise, $< 30\%$ for a3, a4, a5 and NER and 11% for Presidio.

We conducted a detailed error analysis to gain a better understanding about the advantages and shortcoming of the three anonymisation tools described above. The NER tool masked generic entities such as *Second World War*, although this term was not masked by any annotator or by C-sanitise. In the phrase "*a Christian charity dedicated to helping the people of Cambodia*", most annotators did not mask any tokens, while NER masked both *Christian* and *Cambodia*, and C-sanitise *Christian charity*. On the other hand, NER ignored terms that were highly correlated with the individual and should have been masked, such as book titles authored by the person. Another interesting error can be found in the sentence "*In 1964 and 1965 he was a Visiting Professor at the University of Wisconsin–Madison on a Fulbright Program fellowship*" where the university was masked by most annotators but left untouched by C-sanitise (as the university does not frequently co-occur with this person in web documents). Presidio had the lowest recall and ignored the majority of quasi-identifiers (including organisations). Consequently, Presidio's masking should be considered a de-identification process rather than full anonymisation. See Appendix A for an annotated example document.

6 Challenges and Future Directions

The case study illustrates a number of issues facing current methods for text anonymisation. We discuss below three overarching challenges: the need to protect against several types of *semantic inferences*, the formalisation of possible *masking operations* to apply on documents, and, last but not least, the design of *evaluation metrics* to empirically assess the anonymisation performance.

[5] B(eginning) represents the first token of a span, I(nside) the subsequent tokens, and O(ut) is the label assigned to all tokens that are not part of a span.

6.1 Semantic Inferences

Most works on PPDP address anonymisation from a statistical perspective (Batet and Sánchez, 2018). Their main focus is on the statistical properties of (numerical) data and how these may allow attackers to re-identify an individual or uncover confidential data. However, the most harmful inferences in text documents are *semantic* in nature – that is, they are based on the actual meaning expressed in the texts instead of their statistical distributions.

NLP approaches do not explicitly account for semantic inferences, and simply mask all text spans belonging to predefined categories irrespective of their impact on the disclosure risk. In many PPDP approaches (Chakaravarthy et al., 2008; Cumby and Ghani, 2011; Anandan et al., 2012), the adversary is assumed to know sets of attributes associated with each entity, and semantic inferences thus correspond to combinations of attributes enabling the adversary to single out the entity to protect. However, in most practical settings, human adversaries do not have access to the original documents. They do, however, make extensive use of external background knowledge available, e.g., on the web. Such external background knowledge is captured in Sánchez and Batet (2016, 2017) using (co-)occurrence counts of terms on the web.

Other types of semantic inferences may be taken into account, such as lexical and taxonomic relations (synonyms, antonyms, hypernyms, hyponyms) between words or entities. For instance, the word “AIDS” will lead to the disclosure of the confidential attribute “immune system disease”. In Sánchez and Batet (2017), those relations are taken into account by enforcing consistency between known taxonomic relations and the information content of each term. Semantic relations can, however, extend beyond individual terms and exploit various syntactic patterns, as shown in e.g. textual entailment (Dagan et al., 2013).

Semantic inferences can also be drawn from structured data sources such as census data or medical knowledge bases. In the “*Wisconsin-Madison*” example above, the search for Fullbright recipients at that university in 1964-65 would likely allow the individual to be re-identified. Such logical inferences require specifying which background knowledge may be available to a potential intruder and would be relevant for a given text domain.

Although semantic inferences have been studied in isolation in previous work, how to integrate and chain together those inferential mechanisms into a single framework remains an open question. Formally, assuming a document d transformed into d' by an anonymisation tool in charge of protecting a set of entities C, one can design an adversary model $adv(c, d', K)$ seeking to predict, based on document d' and background knowledge K, whether the entity c was part of the original document d or not. Ideally, this adversary model should allow for multiple types of semantic inferences based on domain-relevant background knowledge (word co-occurrences in text corpora, taxonomic relations, knowledge bases, etc.).

6.2 Masking Operations

NLP approaches to text anonymisation essentially focus on *detecting* personal identifiers and rarely discuss what to do with the detected text spans, generally assuming that those should be either redacted or replaced with coded values. This approach may, however, lead to unnecessary loss of data utility, as it is often possible to replace quasi-identifiers by more generic (but still informative) entries.

How to transform a dataset to balance disclosure risk and data utility is a central research question in privacy-preserving data publishing. Various transformations have been put forward: one can *remove* values altogether, *generalise* them into less detailed categories, or *perturb* the values by adding noise or swapping them (Domingo-Ferrer et al., 2016).

In the text domain, several PPDP approaches have shown how to generalise terms using ontologies (Anandan et al., 2012; Sánchez and Batet, 2016). However, these approaches are intrinsically limited to entities present in such ontologies, and are difficult to extend to more generic text entries. Another possible transformation is to introduce noise into the text. The perturbation of data points through noise is a common type of transformation in data privacy (McSherry and Talwar, 2007). This idea of perturbation has notably been applied to word embeddings (Feyisetan et al., 2019), but it produces perturbed word distributions rather than readable documents. Semantic noise has also been defined to perturb nominal values (Rodríguez-García et al., 2017).

Formally, one can define an *editor model* $edit(d)$ taking a document d and outputting an edited document d' after applying a sequence of masking operations. This model can be e.g. expressed as a neural text editing model (Mallinson et al., 2020).

Its optimisation objective should include both minimising the risk of letting an adversary disclose at least some of the protected entities C through semantic inferences (as described in the previous section) and minimising the number of masking operations necessary to map d to d'.

6.3 Evaluation Metrics

Let D be a set of documents transformed into D' by an anonymisation tool. How can we empirically evaluate the quality of the anonymisation?

The most common method is to rely on human annotators to manually mark identifiers in each document $d \in D$, and then compare the system output with those human-annotated identifiers using IR-based metrics such as precision, recall and F_1 score. The recall can be seen as reflecting the degree of protection of the confidential information, while the precision is correlated with the remaining data utility of the documents D'.

This evaluation procedure has a number of shortcomings. As observed in our case study, there may be several equally valid solutions to a given anonymisation problem. Furthermore, IR-based metrics typically associate uniform weights to all identifiers, without taking into account the fact that some identifiers may have a much larger influence on the disclosure risk than others. For instance, failing to detect a full person name is more harmful than failing to detect a quasi-identifier.

Finally, such type of evaluation procedure is limited to the detection of direct and indirect identifiers, but ignore the subsequent step of transforming the textual content. Evaluating the quality of masking operations is tightly coupled with the problem of evaluating how data utility is preserved through the anonymisation process (Sánchez and Batet, 2016; Rodríguez-García et al., 2019). However, how to empirically measure this data utility remains an open question.

An alternative which has so far received little attention is to conduct so-called privacy attacks on the edited documents D'. This can be achieved by e.g. providing the documents D' to human experts and instruct them to re-identify those documents with the help of any information source at their disposal. Such human evaluations can help uncover weaknesses in the anonymisation model (such as semantic inferences that had been overlooked). However, they are also costly and time-consuming, as they must be repeated for each version of the anonymisation model.

7 Conclusion

This position paper discussed a number of unresolved challenges in text anonymisation. Text anonymisation is defined as the removal or masking of any information that, directly or indirectly, may lead to an individual being identified (given some assumptions about the available background knowledge). As illustrated in our case study, text anonymisation is a difficult task (also for human annotators), which goes beyond the mere detection of predefined categories of entities and may allow for several solutions. How to properly anonymise text data is a problem of great practical importance. In particular, access to high-quality data is a key ingredient for most scientific research, and the lack of good anonymisation methods for text documents (allowing data to be shared without compromising privacy) is a limiting factor in fields such as medicine, social sciences, psychology and law.

We surveyed two families of approaches with complementary strengths and weaknesses: NLP models are well-suited to capture textual patterns but lack any consideration of disclosure risk, while PPDP approaches provide principled accounts of privacy requirements, but view documents as bag-of-terms void of linguistic structure.

As outlined in the last section, a promising approach is to couple a neural editor model (applying transformations to the text) with an adversary model (capturing possible semantic inferences to uncover confidential entities). These two models can be optimised jointly using adversarial training, taking into account the necessary balance between disclosure risk and utility preservation.

Finally, we lay out a case for designing evaluation metrics that go beyond traditional IR-based measures, and account in particular for the fact that some identifiers and quasi-identifiers are more important than others in terms of their influence on the disclosure risk.

Acknowledgements

We acknowledge support from the Norwegian Research Council (CLEANUP project[6], grant nr. 308904), the Government of Catalonia (ICREA Acadèmia Prize to D. Sánchez and grant 2017 SGR 705) and the Spanish Government (project TIN2016-80250-R “Sec-MCloud”).

[6] see http://cleanup.nr.no/

References

John Aberdeen, Samuel Bayer, Reyyan Yeniterzi, Ben Wellner, Cheryl Clark, David Hanauer, Bradley Malin, and Lynette Hirschman. 2010. The MITRE identification scrubber toolkit: design, training, and assessment. *International Journal of Medical Informatics*, 79(12):849–859.

Alyaa Alfalahi, Sara Brissman, and Hercules Dalianis. 2012. Pseudonymisation of personal names and other PHIs in an annotated clinical Swedish corpus. In *Third LREC Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2012)*, pages 49–54.

Balamurugan Anandan and Chris Clifton. 2011. Significance of term relationships on anonymization. In *Proceedings of the 2011 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Workshops, WI-IAT 2011*, pages 253–256, Lyon, France.

Balamurugan Anandan, Chris Clifton, Wei Jiang, Mummoorthy Murugesan, Pedro Pastrana-Camacho, and Luo Si. 2012. t-plausibility: Generalizing words to desensitize text. *Transactions on Data Privacy*, 5(3):505–534.

Montserrat Batet and David Sánchez. 2018. Semantic disclosure control: semantics meets data privacy. *Online Information Review*, 42(3):290–303.

Eric A. Bier, Richard Chow, Philippe Golle, Tracy H. King, and J. Staddon. 2009. The rules of redaction: Identify, protect, review (and repeat). *IEEE Security and Privacy Magazine*, 7(6):46–53.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. Association for Computational Linguistics.

Rishi Bommasani, Steven Wu, Zhiwei, and Alexandra K Schofield. 2019. Towards private synthetic text generation. In *NeurIPS 2019 Workshop on Machine Learning with Guarantees*, Vancouver, Canada.

David Carrell, Bradley Malin, John Aberdeen, Samuel Bayer, Cheryl Clark, Ben 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.

Venkatesan T. Chakaravarthy, Himanshu Gupta, Prasan Roy, and Mukesh K. Mohania. 2008. Efficient techniques for document sanitization. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM 2008*, pages 843–852, Napa Valley, California, USA.

Aipeng Chen, Jitendra Jonnagaddala, Chandini Nekkantti, and Siaw-Teng Liaw. 2019. Generation of surrogates for de-identification of electronic health records. In *MEDINFO 2019: Health and Wellbeing e-Networks for All - Proceedings of the 17th World Congress on Medical and Health Informatics, Lyon, France, 25-30 August 2019*, volume 264 of *Studies in Health Technology and Informatics*, pages 70–73. IOS Press.

Bee-Chung Chen, Daniel Kifer, Kristen LeFevre, and Ashwin Machanavajjhala. 2009. *Privacy-Preserving Data Publishing*. Foundations and Trends in Databases. Now Publishers Inc.

Raphaël Chevrier, Vasiliki Foufi, Christophe Gaudet-Blavignac, Arnaud Robert, and Christian Lovis. 2019. Use and understanding of anonymization and de-identification in the biomedical literature: Scoping review. *Journal of Medical Internet Research*, 21(5):e13484.

Jason P.C. Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, 4:357–370.

Richard Chow, Philippe Golle, and Jessica Staddon. 2008. Detecting privacy leaks using corpus-based association rules. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, page 893–901, New York, NY, USA. Association for Computing Machinery.

Chad M. Cumby and Rayid Ghani. 2011. A machine learning based system for semi-automatically redacting documents. In *Proceedings of the Twenty-Third Conference on Innovative Applications of Artificial Intelligence*, pages 1628–1635, San Francisco, California, USA.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. 2013. Recognizing textual entailment: Models and applications. *Synthesis Lectures on Human Language Technologies*, 6(4):1–220.

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.

Mona Diab, Tim Baldwin, and Marco Baroni, editors. 2013. *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*. Association for Computational Linguistics, Atlanta, Georgia, USA.

Josep Domingo-Ferrer, David Sánchez, and Alberto Blanco-Justicia. 2021. The limits of differential privacy (and its misuse in data release and machine learning). *Communications of the ACM*, 64(7):34–36.

Josep Domingo-Ferrer, David Sánchez, and Jordi Soria-Comas. 2016. *Database Anonymization: Privacy Models, Data Utility, and Microaggregation-based Inter-model Connections*. Synthesis Lectures on Information Security, Privacy & Trust. Morgan & Claypool Publishers.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating Noise to Sensitivity in Private Data Analysis. In *Theory of Cryptography*, pages 265–284, Berlin, Heidelberg. Springer Berlin Heidelberg.

Elisabeth Eder, Ulrike Krieg-Holz, and Udo Hahn. 2019. De-identification of emails: Pseudonymizing privacy-sensitive data in a German email corpus. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 259–269, Varna, Bulgaria. INCOMA Ltd.

Elisabeth Eder, Ulrike Krieg-Holz, and Udo Hahn. 2020. CodE Alltag 2.0 — a pseudonymized German-language email corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4466–4477, Marseille, France. European Language Resources Association.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21, Brussels, Belgium. Association for Computational Linguistics.

Mark Elliot, Elaine Mackey, Kieron O'Hara, and Caroline Tudor. 2016. *The anonymisation decision-making framework*. UKAN Manchester.

Natasha Fernandes, Mark Dras, and Annabelle McIver. 2018. Generalised differential privacy for text document processing. *CoRR*, abs/1811.10256.

Oluwaseyi Feyisetan, Tom Diethe, and Thomas Drake. 2019. Leveraging hierarchical representations for preserving privacy and utility in text. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 210–219. IEEE.

Max Friedrich, Arne Köhn, Gregor Wiedemann, and Chris Biemann. 2019. Adversarial learning of privacy-preserving text representations for de-identification of medical records. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5829–5839, Florence, Italy. Association for Computational Linguistics.

GDPR. 2016. General Data Protection Regulation. European Union Regulation 2016/679.

Philippe Golle. 2006. Revisiting the uniqueness of simple demographics in the US population. In *Proceedings of the 5th ACM Workshop on Privacy in electronic society*, pages 77–80. ACM.

Tzvika Hartman, Michael D Howell, Jeff Dean, Shlomo Hoory, Ronit Slyper, Itay Laish, Oren Gilon, Danny Vainstein, Greg Corrado, Katherine Chou, et al. 2020. Customization scenarios for de-identification of clinical notes. *BMC Medical Informatics and Decision Making*, 20(1):1–9.

Mike Hintze. 2017. Viewing the GDPR through a de-identification lens: a tool for compliance, clarification, and consistency. *International Data Privacy Law*, 8(1):86–101.

HIPAA. 2004. *The Health Insurance Portability and Accountability Act*. U.S. Dept. of Labor, Employee Benefits Security Administration.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Yangsibo Huang, Zhao Song, Danqi Chen, Kai Li, and Sanjeev Arora. 2020. TextHide: Tackling data privacy in language understanding tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1368–1382, Online. Association for Computational Linguistics.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California.

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:S34–S42.

Jonathan Mallinson, Aliaksei Severyn, Eric Malmi, and Guillermo Garrido. 2020. Felix: Flexible text editing through tagging and insertion. *arXiv preprint arXiv:2003.10687*.

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*, pages 618–638.

H. Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. 2017. Learning Differentially Private Recurrent Language Models. *arXiv:1710.06963 [cs]*.

Frank McSherry and Kunal Talwar. 2007. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103.

Ben Medlock. 2006. An introduction to NLP-based textual anonymisation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, pages 1051–1056, Genoa, Italy. European Language Resources Association (ELRA).

Beáta Megyesi, Lena Granstedt, Sofia Johansson, Julia Prentice, Dan Rosén, Carl-Johan Schenström, Gunlög Sundberg, Mats Wirén, and Elena Volodina. 2018. Learner corpus anonymization in the age of GDPR: Insights from the creation of a learner corpus of Swedish. In *Proceedings of the 7th workshop on NLP for Computer Assisted Language Learning*, pages 47–56, Stockholm, Sweden. LiU Electronic Press.

Stéphane M Meystre, Óscar Ferrández, F Jeffrey Friedlin, Brett R South, Shuying Shen, and Matthew H Samore. 2014a. Text de-identification for privacy protection: a study of its impact on clinical text information content. *Journal of Biomedical Informatics*, 50:142–150.

Stephane M Meystre, F Jeffrey Friedlin, Brett R South, Shuying Shen, and Matthew H Samore. 2010. Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC Medical Research Methodology*, 10(1):70.

Stéphane Meystre, Shuying Shen, Deborah Hofmann, and Adi Gundlapalli. 2014b. Can physicians recognize their own patients in de-identified notes? *Studies in Health Technology and Informatics*, 205:778—782.

Ahmadreza Mosallanezhad, Ghazaleh Beigi, and Huan Liu. 2019. Deep reinforcement learning-based text anonymization against private-attribute inference. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2360–2369, Hong Kong, China. Association for Computational Linguistics.

Sravana Reddy and Kevin Knight. 2016. Obfuscating gender in social media writing. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 17–26, Austin, Texas. Association for Computational Linguistics.

Mercedes Rodríguez-García, Montserrat Batet, and David Sánchez. 2017. A semantic framework for noise addition with nominal data. *Knowledge-Based Systems*, 122(C):103–118.

Mercedes Rodríguez-García, Montserrat Batet, and David Sánchez. 2019. Utility-preserving privacy protection of nominal data sets via semantic rank swapping. *Information Fusion*, 45:282–295.

Mark A. Rothstein. 2010. Is deidentification sufficient to protect health privacy in research? *The American Journal of Bioethics*, 10(9):3–11.

Pierangela Samarati. 2001. Protecting respondents' identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering*, 13(6):1010–1027.

Pierangela Samarati and Latanya Sweeney. 1998. Protecting Privacy when Disclosing Information: k-Anonymity and its Enforcement through Generalization and Suppression. Technical report, SRI International.

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

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

Latanya Sweeney. 1996. Replacing personally-identifying information in medical records, the scrub system. In *Proceedings of the AMIA annual fall symposium*, pages 333–337. American Medical Informatics Association.

David Sánchez and Montserrat Batet. 2016. C-sanitized: A privacy model for document redaction and sanitization. *Journal of the Association for Information Science and Technology*, 67(1):148–163.

David Sánchez and Montserrat Batet. 2017. Toward sensitive document release with privacy guarantees. *Engineering Applications of Artificial Intelligence*, 59:23–34.

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):19 – 26.

Elena Volodina, Yousuf Ali Mohammed, Sandra Derbring, Arild Matsson, and Beata Megyesi. 2020. Towards privacy by design in learner corpora research: A case of on-the-fly pseudonymization of Swedish learner essays. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 357–369, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ralph Weischedel, Eduard Hovy, Marcus. Mitchell, Palmer Martha S., Robert Belvin, Sameer S. Pradhan, Lance Ramshaw, and Nianwen Xue. 2011. OntoNotes: A large training corpus for enhanced processing. In *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*. Springer.

Alan F. Westin. 1967. *Privacy and Freedom*. Atheneum, New York.

Qiongkai Xu, Lizhen Qu, Chenchen Xu, and Ran Cui. 2019. Privacy-aware text rewriting. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 247–257, Tokyo, Japan. Association for Computational Linguistics.

Xi Yang, Tianchen Lyu, Qian Li, Chih-Yin Lee, Jiang Bian, William R Hogan, and Yonghui Wu. 2019. A study of deep learning methods for de-identification of clinical notes in cross-institute settings. *BMC Medical Informatics and Decision Making*, 19(5):232.

Vithya Yogarajan, Michael Mayo, and Bernhard Pfahringer. 2018. A survey of automatic de-identification of longitudinal clinical narratives. *arXiv preprint arXiv:1810.06765*.

A Appendix

We present below the annotation of one short biography of a 20th century scientist (Alexander Frumkin) according to 5 human annotators, C-sanitize, the neural NER model and the Presidio anonymisation tool (see paper for details). The annotation task consisted of tagging text spans that could re-identify a person either directly or in combination with publicly available knowledge. The annotators were instructed to prevent identity disclosure, but otherwise seek to preserve the semantic content as much as possible. The five annotators were researchers in statistics and natural language processing.

The first five (gray) lines denotes the five human annotators, while the cyan line corresponds to C-sanitise, the blue line to the neural NER model, and the green line to the Presidio tool.

Due to page limits, we only present here one single biography, but the annotations for all 8 texts (along with the annotation guidelines and raw data) are available in the GitHub repository associated with the paper.

A.1 Alexander Frumkin

Alexander Naumovich Frumkin (Александр Наумович Фрумкин) (October 24, 1895–May 27, 1976) was a Russian/Soviet electrochemist, member of the Russian Academy of Sciences since 1932, founder of the Russian Journal of Electrochemistry Elektrokhimiya and receiver of the Hero of Socialist Labor award. The Russian Academy of Sciences' A. N. Frumkin Institute of Physical Chemistry and Electrochemistry is named after him. Frumkin was born in Kishinev, in the Bessarabia Governorate of the Russian Empire (present-day Moldova) to a Jewish family; his father was an insurance salesman. His family moved to Odessa, where he received his primary schooling; he continued his education in Strasbourg, and then at the University of Bern. Frumkin's first published articles appeared in 1914, when he was only 19; in 1915, he received his first degree, back in Odessa. Two years later, the seminal article "Electrocapillary Phenomena and Electrode Potentials" was published. Frumkin moved to Moscow in 1922 to work at the Karpov Institute, under A. N. Bakh. In 1930 Frumkin joined the faculty of Moscow University, where in 1933 he founded—and would head until his death—the department of electrochemistry. During the Second World War, Frumkin led a large team of scientists and engineers involved in defense issues. This contribution did not save him from being dismissed in 1949 as the director of the Institute of Physical Chemistry, when he was accused of "cosmopolitanism". Frumkin's most fundamental achievement was the fundamental theory of electrode reactions, which describes the influence of the structure of the interface between electrode and solution on the rate of electron transfer. This theory has been confirmed and extended within the framework of contemporary physical electron transfer models. Frumkin introduced the concept of the zero charge potential, the most important characteristic of a metal surface.

Alessandro Volta's question—a topic of discussion for over 120 years—about the nature of the EMF of electrochemical circuits was resolved using Frumkin's approach. Frumkin developed the Frumkin isotherm, an extension of the Langmuir isotherm in describing certain adsorption phenomena. Frumkin's students developed novel experimental methods that would, in time, become standard. Several applied electrochemical processes, including ones related to chemical sources of electrical power, industrial electrolysis, and anti-corrosion protection, were successfully developed under Frumkin's supervision. Frumkin was married three times, including a brief first marriage to Vera Inber.