De-identification in Natural Language Processing

Veronika Vincze^{1,2} ¹MTA-SZTE Research Group on Artificial Intelligence ²Department of Informatics, University of Szeged Szeged, Hungary Email: vinczev@inf.u-szeged.hu

Abstract—Natural language processing (NLP) systems usually require a huge amount of textual data but the publication of such datasets is often hindered by privacy and data protection issues. Here, we discuss the questions of de-identification related to three NLP areas, namely, clinical NLP, NLP for social media and information extraction from resumes. We also illustrate how de-identification is related to named entity recognition and we argue that de-identification tools can be successfully built on named entity recognizers.

I. INTRODUCTION

Natural language processing (NLP) is an interdisciplinary field that lies between artificial intelligence and linguistics: it seeks to understand human language with automatic methods [1]. One of the most widely studied areas of NLP is information extraction, which aims at collecting relevant information from free (unstructured) texts. Doing that, new knowledge can be quickly gathered from a large amount of texts.

NLP tasks are usually handled by either rule-based or statistical systems and in the latter, one usually makes use of machine learning. One of the most widespread paradigms within machine learning is called supervised learning: in this setting, an annotated training database is required from which machine learning algorithms may collect and systematize patterns typical of the problem, which patterns are later applied to previously unseen and unannotated databases. Thus, large annotated datasets are necessary to solve NLP problems with supervised methods and generally, there are some public benchmark datasets (some of them are mentioned in Sections II and III, related to named entity recognition and medical NLP) on which the outputs of different systems handling the same task may be compared, in order to ensure the comparability of scientific results.

One issue that may arise when publishing such annotated NLP datasets concerns privacy and data protection. There are certain types of texts that contain a big amount of personal / sensitive data and thus their proper and confidential treatment is essential. However, since supervised machine learning algorithms can only learn from data, it is necessary to get access to texts which do not contain any personal data, thus questions of de-identification have already been solved, yet the dataset can be appropriately employed in the learning process. In this paper, we examine de-identification issues in three areas, namely, in clinical NLP, NLP for social media and information extraction from resumes (CVs). We will also show how de-identification is related to named entity recognition, i.e. identifying person names, locations and so on in texts.

Richárd Farkas Department of Informatics University of Szeged Szeged, Hungary Email: rfarkas@inf.u-szeged.hu

II. NAMED ENTITY RECOGNITION

Named entities are names or identifiers that uniquely refer to one entity in the world. They are essential for information extraction systems as they seek to detect relations among different entities. Named entities belong to several semantic categories, the most widely used ones being:

- person names;
- organizations;
- locations;
- product names;
- identifiers etc.

However, these categories could be further divided into subcategories, depending on the exact domain. For instance, in clinical documents it is relevant who the doctor and the patient are, a simple "person name" category will not suffice. In other cases, not all of these are relevant or new categories might be introduced like phone numbers, e-mails and taxonomy names.

Named entity recognition has a long tradition in natural language processing. It was a task assigned within the framework of the Message Understanding Conference MUC-7 [2]. Participants had to identify personal names, geographical names, organisations, and other names related to time, quantity, and descriptive terms. In 2003, The Conference on Computational Natural Language Learning (CoNNL) [3] was announced by the open tournaments. The aim was to construct a named entity recognition model that could handle English and German texts. One of the most successful and most widely used approaches for named entity recognition is sequence labeling, in particular the Conditional Random Fields (CRF) approach. The best performing systems on the CoNLL task applied a CRF approach and gave an F-score of 0.85-0.89 for English.

Named entity recognition is closely related to deidentification and anonymization since most of the categories to be de-identified are some kind of named entities and hence, techniques for named entity recognition may be fruitfully applied as a first step to remove personal data from texts.

In the following, we present three areas of natural language processing in which de-identification is an essential issue and we also show how named entity recognition may help in the task of de-identifying texts.

III. DE-IDENTIFICATION IN MEDICAL NLP

The de-identification of clinical records is crucial in the human life sciences because a de-identified text can be made publicly available in order to facilitate research on human diseases. However, clinical documents on the patients include explicit personal health information (PHI), which prevents the release of many useful data sets because this would jeopardize individual patient rights. According to the guidelines of Health Information Portability and Accountability Act (HIPAA), all the medical discharge summaries released must be free of the following categories of textual PHI:

- first and last names of patients, their health proxies, and family members;
- doctors' first and last names;
- identification numbers;
- telephone, fax, and pager numbers;
- hospital names;
- geographic locations;
- dates.

Removing these kinds of PHI is the main goal of the deidentification process. Anonymization goes one step beyond the deletion of personal information and attempts to identify and classify personal information in the text to one of the above-mentioned HIPAA-defined categories. This categorization enables the replacement of personal data instead of simple deletion, and it has several advantages. First, the replacement of PHIs with automatically generated look-alike substitutes preserves the readability of text and, second, even if deidentification is not totally perfect, that is, some data items remain unchanged, the artificial substitutes disguise those very few pieces of personal information that still remain in the document (the reader will not know whether a label was the original or a substitute).

In the literature many de-identification approaches have been introduced.¹ Some approaches target the recognition (and removal) of particular types of PHI like [5], which focuses on patient names, or [6], which seeks to identify person names (both patients and doctors). There are several approaches that perform the full de-identification of medical texts. These are based either on a pattern-matching algorithm that uses a thesaurus [7], [8]; a combination of rule-based systems and pattern matching using dictionaries [9] and the Unified Medical Language System [10] or on a statistical model [11]. The importance of this problem was also reflected in the organization of the I2B2 shared task in 2007, which focused on the de-identification of personal data in clinical documents. For a detailed description of the shared task challenge, the corpus, results and lessons learned, see [12].

In [13], some Named Entity Recognition (NER) techniques are used for the task of the de-identification of clinical records. The NER system described in [14] was adapted to the clinical domain and the main categories like patient and doctor names, hospital names, locations were identified by this tool. With other categories such as dates or phone numbers, regular expressions were applied. After the identification of the entities, they were changed to another one from the same category, e.g. *Dr. Jones* was changed to *Dr. Jackson*. The system achieved an overall F-score of 0.967 at the instance level, which yielded the best result for the challenge, proving that the de-identification of clinical documents can be successfully carried out with techniques based on named entity recognition.

IV. DE-IDENTIFICATION IN NLP FOR SOCIAL MEDIA

The role of the internet, more precisely, the role of social media in people's lives is becoming more and more significant, especially due to its importance in modern communication. The billions of tweets, wall posts and likes reveal a lot of user preferences, for instance, what type of products they choose, what type of music, books, cars or food they prefer, what destinations they travel to for holiday, what political parties they vote for, besides, their demographic data (age, location, marital status) can be also easily gathered and so on. A large amount of data is generated by the users who are communicating through this medium, and this could be exploited in a number of natural language processing applications such as information extraction and sentiment analysis. Be it Facebook, Twitter, Linkedin or Pinterest - to an enterprise, government, or a social scientist, this endless stream of big data allows the crafting of user profiles with a level of detail that was unimaginable a few years ago. There is thus a strong need for advanced analytics, to complement the abundant computer power and cheap storage, in order to allow enterprises to offer not just generalized but truly personalised interaction with their customers.

However, all of these possible applications require the intensive use of user-generated content, which may include a considerable amount of personal data. For instance, Facebook wall posts or tweets are usually paired with the name or ID of the user, the geolocation where they were created, and so on. Personal data can also be easily gathered from user profiles, where each user typically provides their name, age, gender, city, phone number, e-mail address and other types of sensitive data. Thus, data protection and de-identification issues are indispensable when working with social media texts. For instance, there was recently a shared task on computational personality recognition [15]. They released two datasets essays and a subset of the myPersonality dataset -, which include gold standard personality labels and texts (essays and Facebook status updates) written by the persons themselves. All these data items had to be anonymized before publishing them and users of the dataset are required to sign contracts to ensure privacy rights.

As a real-life example, let us take personalization, where the goal is to provide each user with a highly personalized content/offer, be it a selection of snippets of news or a smartphone that perfectly fits the user's needs. In such cases, the profiling system makes the appropriate selection for the user on the basis of the user's earlier preferences or the preferences generalized from the specific choices of other users with similar characteristics to the user in question. Quite often, people share their opinion on some products through social media tools (e.g. wall posts or comments to the site of the producer on Facebook) and hence, trends can be generalized

¹For a detailed overview of the field, please refer to [4].

from social media content like which age group likes which types of phones or which functionalities of the device are preferred by which group of users etc.

Although there is a lot of personal data in social media content, to achieve the above-mentioned goals, there are some data types that are required for personalization purposes. If all the sensitive data were removed from the texts, no useful information would remain since no valid tendencies could be deduced from overly general templates like *someone did not like the screen of some smartphone*. What would be more useful instead is to have templates filled with specific information like:

a boy between 16 and 18 did not like the screen of iPhone 4

25 women between 35 and 40 who had children had a look at Jamie Oliver's website

This kind of information can be exploited in creating personalizing and profiling systems: the user's personal characteristics and habit can be compared to those stored in the database of the system and then a personalized offer can be readily produced.

Looking at the above points from the perspective of deidentification, we may argue that not all data types should be removed or replaced from social media texts. We propose that there should be three categories for sensitive data. These are:

- **Data to be removed / replaced**: this category involves person names, addresses, phone numbers, e-mail addresses and other identifiers that uniquely refer to the person (Skype, Twitter etc. IDs and the like);
- Data to be merged into groups: this category includes quantitative data like the age of the person (for the sake of simplicity, it is normally sufficient to handle age groups rather than the individual ages) or the city where the user lives (in a similar way, it is sufficient to know whether the user lives in a big city, a town or a village, no matter what its precise name is);
- Data to be preserved: in the profiling case, it is important to leave the names of products or organizations as they are, in order to get a clear view on which products the users prefer.

We should add here that the above classification scheme is not absolute. Depending on the exact task, other categorization of sensitive data might be relevant.

V. DE-IDENTIFICATION IN INFORMATION EXTRACTION FROM CVs

With many companies, it is a constant headache to find new employees. In the CV filtering process, human resource managers do not have the capacity of having a thorough look at each CV submitted to a given position, which hinders the selection of the best candidate. Hence, a typical solution for this problem is to randomly select some tens of CVs from the database, which are then manually checked and the best candidates from these dozens of CVs are then interviewed. Information extraction systems that are able to parse CVs and then to return a ranking of the candidates for a given position are badly needed in the market. However, the main difficulties of the task lie in the following. The format of CVs change from document to document since there is no unique template that all of the candidates apply. Therefore, the first task is to collect into a database all the necessary information from the CVs which are relevant for the given job. The most common types of information that are listed in CVs are the following:

- Personal data: name, date of birth, address, phone number, e-mail, marital status;
- Education: name of schools, degrees;
- Employment: names of companies or employers the candidate previously worked for, earlier jobs or positions;
- Language skills;
- Competencies;
- Hobbies and interests.

In order to collect these pieces of data from the CVs via automatic methods, a training database of CVs is needed. However, the publication of such databases is a delicate question, especially due to data protection and anonymization issues. For these reasons, we are not aware of any publicly available CV databases as it would require the anonymization of CVs. However, not all of the above categories should be treated in the same way, unlike clinical documents, where most of the data should be unequivocally removed from the text and replaced by another of the same category (e.g. a name of a hospital should be replaced by the name of another hospital) (see Section II). Here we argue that the following methods should be applied for the different categories of sensitive data in CVs:

- **Data to be replaced:** this method, for instance, could be used for names, addresses, phone numbers and e-mails;
- Data to be preserved: data items concerning education generally contain relevant information: the degrees of the candidate have a strong effect on the selection process and sometimes it is also relevant which college / university the candidate attended and graduated from since a degree from a higher ranked university is usually considered more valuable than one from a lower ranked university. The same is true for the names of previous positions held at different companies since it reveals whether the candidate has worked in a similar position for which s/he is applying now. The name of earlier employers usually contains important information on the work experience of the candidate, so they should probably be preserved. However, this might be a delicate issue;
- **Data to be deleted:** in order to ensure equal rights for men and women, young and old, information on age, gender and marital status should be deleted from the CV.

With the help of the above-sketched anonymization procedures, databases of CVs can be made publicly available, which will surely enhance the implementation of information extraction from CVs for human resource management.

VI. CONCLUSIONS

In this paper, we showed how de-identification issues can influence certain areas of natural language processing. We argued that in order to publish datasets, they should undergo de-identification and/or anonymization by applying named entity recognition tools, which have been consistently used for that purpose, in the clinical domain. As for texts from the social media domain or CVs, the publication of large databases has been impeded by the absence of proper de-identification tools developed for the specific domain. We hope that our work will enhance the development of such tools in the near future.

ACKNOWLEDGMENT

Veronika Vincze was funded in part by the project Telemedicine-focused research activities in the fields of mathematics, informatics and medical sciences (grant no.: TÁMOP-4.2.2.A-11/1/KONV-2012-0073). Richárd Farkas was funded in part by the project FuturICT.hu (grant no.: TÁMOP-4.2.2.C-11/1/KONV-2012-0013). Both projects are financed by the European Union and the European Social Fund.

References

- D. Jurafsky and J. H. Martin, Speech and Language Processing (2nd Edition) (Prentice Hall Series in Artificial Intelligence), 2nd ed. Prentice Hall, 2008.
- [2] N. A. Chinchor, "Overview of muc-7/met-2." in Proceedings of the Seventh Message Understanding Conference (MUC-7), 1998.
- [3] E. F. Tjong Kim Sang and F. De Meulder, "Introduction to the conll-2003 shared task: Language-independent named entity recognition," in *Proceedings of CoNLL-2003*, W. Daelemans and M. Osborne, Eds. Edmonton, Canada, 2003, pp. 142–147.
- [4] S. M. Meystre, F. J. Friedlin, B. R. South, S. Shen, and M. H. Samore, "Automatic de-identification of textual documents in the electronic health record: a review of recent research." *BMC medical research methodology*, vol. 10, no. 1, pp. 70+, Aug. 2010. [Online]. Available: http://dx.doi.org/10.1186/1471-2288-10-70
- [5] R. Taira, A. Bui, and H. Kangarloo, "Identification of patient name references within medical documents using semantic selectional restrictions," in *Proceedings of the American Medical Informatics Association* 2002, I. Kohane, Ed. Philadelphia, PA, USA: Hanley & Belfus Inc., 2002, pp. 757–761.
- [6] S. M. Thomas, B. Mamlin, G. Schadow, and C. McDonald, "A successful technique for removing names in pathology reports using an augmented search and replace method," in *Proceedings of the American Medical Informatics Association 2002*, I. Kohane, Ed. Philadelphia, PA, USA: Hanley & Belfus Inc., 2002.
- [7] L. Sweeney, "Replacing personally-identifying information in medical records, the scrub system," in *Proceedings of the American Medical Informatics Association 2002*, J. Cimino, Ed. Philadelphia, PA, USA: Hanley & Belfus Inc., 1996, pp. 333–337.
- [8] P. Ruch, R. H. Baud, A.-M. Rassinoux, P. Bouillon, and G. Robert, "Medical document anonymization with a semantic lexicon," in *Proceedings of the American Medical Informatics Association 2000*, 2000, pp. 729–733.
- [9] M. Douglass, G. D. Clifford, A. Reisner, G. B. Moody, and R. G. Mark, "Computer-Assisted De-Identification of Free Text in the MIMIC II Database," *Computers in Cardiology*, vol. 32, pp. 331–334, 2005.

- [10] J. G. D. Gupta, M. Saul, "Evaluation of a Deidentification (De-Id) Software Engine to Share Pathology Reports and Clinical Documents for Research," *American Journal of Clinical Pathology*, vol. 121, no. 6, 2004.
- [11] T. Sibanda and O. Uzuner, "Role of local context in automatic deidentification of ungrammatical, fragmented text," in *Proceedings of* the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics. Morristown, NJ, USA: Association for Computational Linguistics, 2006, pp. 65–73.
- [12] O. Uzuner, Y. Luo, and P. Szolovits, "Evaluating the Stateof-the-Art in Automatic De-identification," J Am Med Inform Assoc, vol. 14, no. 5, pp. 550–563, 2007. [Online]. Available: http://www.jamia.org/cgi/content/abstract/14/5/550
- [13] Gy. Szarvas, R. Farkas, and R. Busa-Fekete, "State-of-theart anonymisation of medical records using an iterative machine learning framework," *J Am Med Inform Assoc*, vol. 14, no. 5, pp. 574–580, 2007. [Online]. Available: http://www.jamia.org/cgi/content/abstract/M2441v1
- [14] Gy. Szarvas, R. Farkas, and A. Kocsor, "A Multilingual Named Entity Recognition System Using Boosting and C4.5 Decision Tree Learning Algorithms," in *Discovery Science*, 2006, pp. 267–278.
- [15] F. Celli, F. Pianesi, D. Stillwell, and M. Kosinski, "Workshop on computational personality recognition (shared task)," in *Proceedings of* WCPR13, in conjunction with ICWSM-1, 2013.