Influence of Module Order on Rule-Based De-identification of Personal Names in Electronic Patient Records Written in Swedish

Elin Carlsson, Hercules Dalianis

Department of Computer and Systems Sciences (DSV)
Stockholm University
Forum 100, 164 40 Kista, Sweden
E-mail: {eli-carl, hercules} @dsv.su.se

Abstract

Electronic patient records (EPRs) are a valuable resource for research but for confidentiality reasons they cannot be used freely. In order to make EPRs available to a wider group of researchers, sensitive information such as personal names has to be removed. De-identification is a process that makes this possible. Both rule-based as well as statistical and machine learning based methods exist to perform de-identification, but the second method requires annotated training material which exists only very sparsely for patient names. It is therefore necessary to use rule-based methods for de-identification of EPRs. Not much is known, however, about the order in which the various rules should be applied and how the different rules influence precision and recall. This paper aims to answer this research question by implementing and evaluating four common rules for de-identification of personal names in EPRs written in Swedish: (1) dictionary name matching, (2) title matching, (3) common words filtering and (4) learning from previous modules. The results show that to obtain the highest recall and precision, the rules should be applied in the following order: title matching, common words filtering and dictionary name matching.

1. Introduction

Today, a huge amount of data are produced in electronic patient records (EPRs), however they are very seldom reused for purposes other than as notes for an individual patient. EPRs contain both structured information such as gender, age, clinic, diagnosis, but also unstructured information in free text (Dalianis et al., 2009). All this information is a valuable source for both research and educational purposes, but also for testing performance on EPR systems. However, EPRs often contain protected health information (PHI), which is information that can reveal the identity of the individual patient (HIPAA, 2003), and can therefore not be used freely. Therefore it would be valuable to de-identify patient records so that they can be used by a wider group of researchers. De-identification is the process of identifying, annotating and finally removing or replacing PHI. A high recall in de-identification systems is preferred over high precision, since the aim is to remove all possible instances of PHI.

Many methods for de-identifying EPRs have been developed, both rule-based and dictionary-based as well as statistical and machine learning-based (Uzuner et al., 2007). For patient names, there exists only very sparse annotated training material and a rule-based system for de-identification would be necessary. One problem with rule-based approaches is that the effect of each rule[1] alone or combined is not well known. This paper aims to find out what effect different rules have on precision and recall, in a rule- and dictionary-based approach for de-identifying personal names in EPRs written in Swedish[2]. Four different modules that implement one rule each are described and they are executed in all possible combinations in order to deduce their effects.

2. Related research

Named Entity Recognition (NER) is a well established research area and represents the process of identifying named entities such as names of persons, locations and organizations in free text (Chinchor, 1997). NER is related to de-identification of EPRs (Uzuner et al., 2007) and is mainly divided into two types: rule- and dictionary-based algorithms or statistical algorithms (Uzuner et al., 2008).

Mikheev et al. (1999) have carried out an experiment to investigate the usefulness of dictionaries in NER for news articles in English. They combined rule-based and statistical methods and the results indicate that dictionaries do not improve the precision and recall significantly (approximately 6 percent).

For Swedish, Kokkinakis and Thurin (2007) have developed a rule-based system for de-identifying hospital discharge letters. The system utilizes dictionaries and yields 95.65 percent precision and recall for personal name identification.

In Dalianis and Velupillai (2010), a Gold Standard developed for de-identification of EPRs written in Swedish is described. The Gold Standard is named Stockholm EPR PHI Corpus and contains, among other annotated classes, 953 first names and 932 last names. The Stockholm EPR PHI Corpus includes 100 EPRs encompassing in total 380,000 tokens. One finding was that patient names are 20 times rarer than clinician names and occur in less than 0.02 percent of the total number of tokens, and are therefore difficult to use as training material for machine learning based methods.

[1] In this paper, the word ”rule” refers to a collection of instructions which together work towards a common goal.

[2] This research has been carried out after approval from the Regional Ethical Review Board in Stockholm (Etikprövningsnämnden i Stockholm), permission number 2007/1625-31/5.

Dalianis and Velupillai (2010) applied an machine learning approach based on Conditional Random Fields using the Stockholm EPR PHI Corpus in de-identification of first names and last names. They obtained 95 percent precision and 77 percent recall on first names and 95 percent and 84 percent in precision and recall, respectively, on last names. For the very specific class patient first names they did, however, obtain only a precision of 43 percent and a recall of 8 percent. For the very specific class patient last names the system failed, probably due to the very scarce training material.

Most of the systems for de-identification of clinical texts are written for the English language. Sweeney (1996) has developed the ”Scrub” system for de-identifying medical records. The system utilizes detection algorithms that employ dictionaries and templates in order to identify one entity each. Scrub reports good results, 99-100 percent of all personal names were de-identified.

Douglass et al. (2005) describe a rule-based system for de-identifying free text nursing notes. In order to identify patient names, the system utilizes dictionaries with names, common English words and medical terms as well as titles. The system reports 44 percent and 98 percent precision and recall respectively for the process of de-identifying names.

Thomas et al. (2002) investigate whether a substitution method for de-identifying proper names, not requiring specialized natural language processing resources, would meet a recall of 90 percent. The basis of the substitution method is the assumption that proper names mostly appear in pairs. Thomas et al. utilize both dictionaries (proper names, clinical and common words) as well as title matching. The reported results are 11.8 percent and 98.7 percent in precision and recall respectively.

Neamatullah et al. (2008) describe the Deid system which applies lexical matching with dictionaries, regular expressions and simple heuristics to identify PHI in clinical text written in American English. Deid reports a precision and recall of 72.5 and 99.5 percent respectively for persons. Velupillai et al. (2009) ported Deid to Swedish, unfortunately it was not obvious to find out how the different modules influenced each other and they obtained very low precision and recall.

3. Method

Our aim is to analyze how different rules in a rule-based and dictionary-based approach, for de-identifying personal names in EPRs written in Swedish, influence precision and recall. To do this, we have identified four basic rules, each implemented in a module:

DM The dictionary module utilizes dictionaries containing female first names, male first names and last names in order to annotate names. The dictionaries contain 2,930, 2,742 and 34,894 words respectively. They are applied in the indicated order: female first names, male first names and finally last names. This means that if a word is contained in the dictionary of female first names, the word will be annotated and the two other dictionaries will not be examined. If, however, the word is not contained in this dictionary, the dictionary of male first names will be examined, etc.

TM The titles module uses title match to find names. Words not marked as a common word with a capital first letter, that is next to a title (such as ”dr” and ”ssk”, abbreviations for ”physician” and ”nurse”, respectively, in Swedish) will be annotated. The words will be annotated differently depending on different conditions: Firstly, if there are two words meeting the above criteria next to a title, they will be annotated in the order first name then last name. Secondly, if – on the other hand – there is just one word meeting these criteria, it will be annotated as a last name if the title is a doctor’s title (such as ”dr”), otherwise it will be annotated as a first name.

CM The common words module utilizes a dictionary of common words containing 5,000 words. Every word in the EPR that exists in the dictionary is marked as ”common” and will be used by the subsequent modules.

LM The learning module, inspired by Neamatullah et al. (2008) and Dalianis and Åström (2001), remembers those words that have been annotated as a name at least twice by the previous modules. It will annotate all the remembered words that have not not already been annotated. The learning module is applied on the whole set of EPRs consisting of 100 documents.

The common words module does not produce any of its own annotations, rather it just serves the other modules. Hence it will not influence the results when it is executed as the last module. The learning module depends on the annotations of previous modules and will therefore not influence the results when executed as the first module.

4. Evaluation and results

In order to find the best combination of modules, we evaluated the output of all combinations using the annotated Stockholm EPR PHI Corpus, described in Dalianis and Velupillai (2010), as a Gold Standard. We used precision, recall and F-score as evaluation metrics on first and last names, respectively, as well as combined (total). The results for the dictionary module and the title module evaluated separately, as well as the combinations that yielded the best and least favourable results, are provided in Figure 1-3. The aim of this work is not primarily to obtain a high precision and recall for each module, but to investigate the influence of the modules.

The dictionary module (DM) yields high recall for first names (approximately 89 percent) and a relatively high recall for last names (approximately 64 percent), the precision is however very low (approximately 6-8 percent for first and last names). The titles module (TM) does not yield such a high recall (approximately 19 percent for first names and 34 percent for last names), but the precision is much better (approximately 50 and 78 percent, respectively).

The highest total F-score (52.62 percent) is reported with the combination titles, common words and dictionary module (TCD-M), with a precision and recall of 45.43 and 62.53 percent, respectively. For first names, this combination also reports the highest F-score (61.01 percent), with

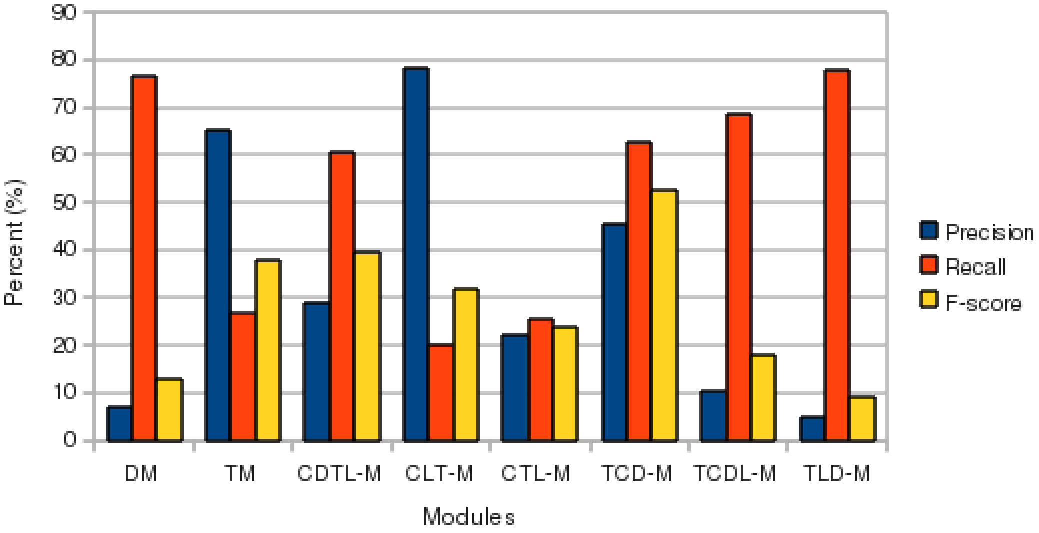

Figure 1: Results from different ordering of modules for first and last names combined (total).

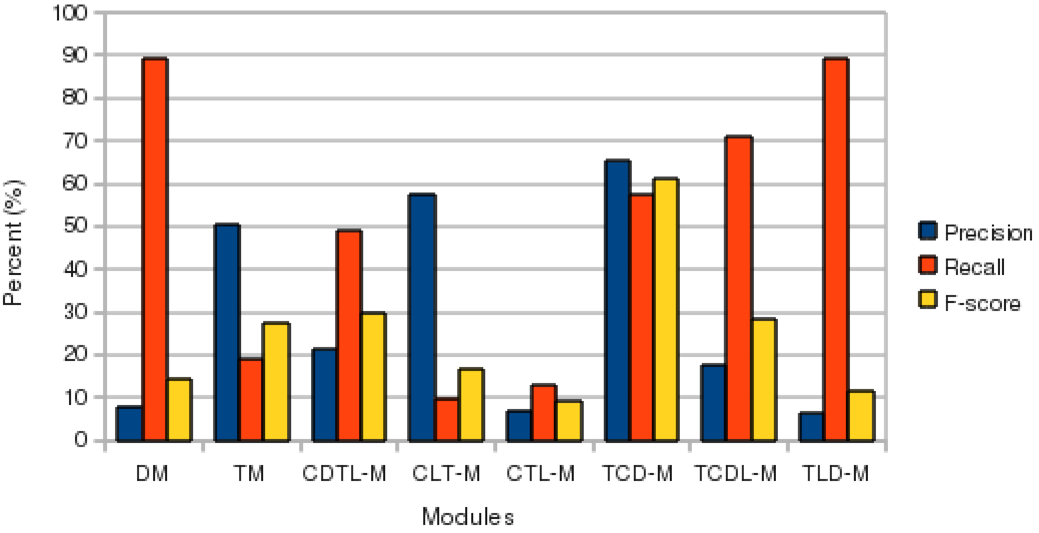

Figure 2: Results from different ordering of modules for first names.

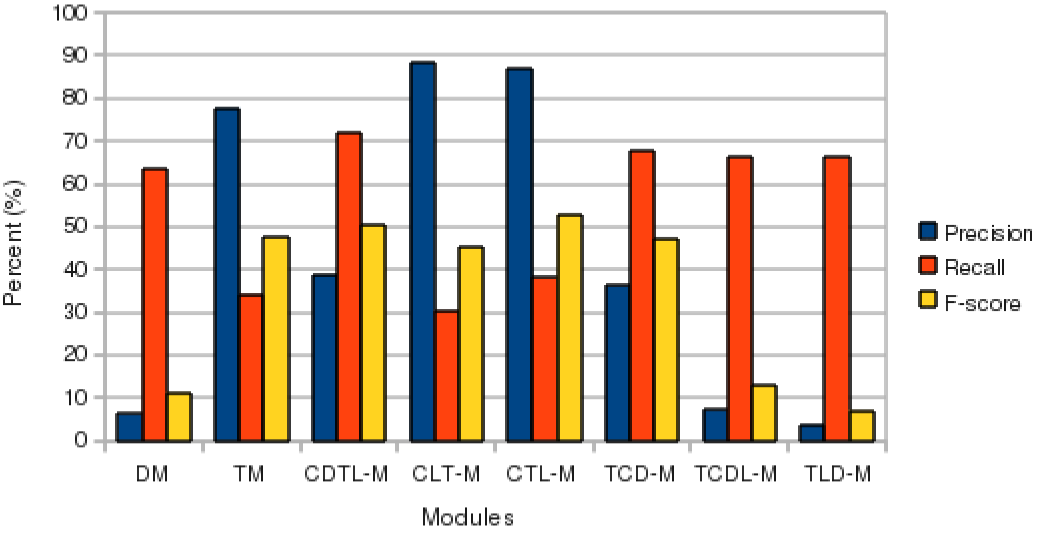

Figure 3: Results from different ordering of modules for last names.

a precision of 65.23 percent and recall of 57.31 percent. By adding the learning module to the end of execution (TCDL-M), the recall increases with approximately 14 percent but the precision decreases with approximately 48 percent. For last names, the combination of common words, titles and learning module (CTL-M) reports the highest F-score (52.88 percent), with a recall that is lower than the precision. The combination that, for last names, reports the highest recall, with an F-score which is still quite good, is the common words, dictionary, titles and learning module (CDTL-M) which yields a precision of 38.76 percent and a recall of 72.01 percent.

The lowest total F-score (9.11 percent) is reported with the combination titles, learning and dictionary module (TLD-M), with a precision and recall of 4.84 and 77.75 percent, respectively. For first names, the lowest F-score (8.96 percent) is reported with the combination common words, titles and learning module (CTL-M), with a precision of 6.90 percent and a recall of 12.78 percent. The lowest recall (9.86 percent) for first names is, however, reported with the combination common words, learning and titles module (CLT-M), which also reports the lowest recall for last names (30.36 percent). The lowest F-score (7.03 percent) for last names is reported with the combination titles, learning and dictionary module (TLD-M), with a precision of 3.71 percent and a recall of 66.20 percent.

The results of each module are also individually analyzed for their influence on precision and recall. The effect of each module is calculated by subtracting the precision and recall of the combination in which the module is represented, from precision and recall of the combination in which the module is not represented. In this way we may deduce how the precision and recall increase or decrease when one module is represented in one combination, compared to when it is not represented.

The dictionary module increases recall and overall decreases precision. The precision for last names mainly decreases much more than for first names. This means that the number of false positives increases for last names when the dictionary module is present, probably because the dictionaries for last names is about six times larger than the dictionaries for first names.

The titles module overall increases recall and decreases precision. The module reports better improvement for last names in precision and recall, than for first names.

The common words module overall improves precision and decreases recall. This means that the dictionary of common words contains words that are actually names in the EPRs. In addition, the recall decreases more for first names than for last names, which indicates that most of the names in the dictionary for common words are first names.

The learning module does not influence the results at all when executed immediately after the dictionary module but before the titles module (one third of cases). This is because all occurrences of a name have already been annotated by the dictionary module. In the other cases the learning module reports quite different results. The precision for first names decreases (by up to 53 percent) and the precision for last names improves with a maximum of one percent but otherwise decreases. This fact indicates that the module produces a large number of false positives.

5. Conclusions

We have analyzed the effects that four different rules, each implemented in a module, have on the results of the de-identification of personal names in Swedish EPRs. The analysis shows that dictionaries increase recall but decrease precision. The titles module increases recall as well, but not to the same extent as the dictionary module. The common words module mainly increases precision but decreases recall. The learning module has a negative effect, if any at all, on the results. In most of the cases, when the recall improves, the precision decreases significantly. This is in line with what Douglass et al. (2005) found, their module for identifying names that previous modules have missed out also reports reduced precision.

Name dictionaries proved to be an important resource to achieve high recall in the de-identification of personal names. Dictionaries, however, imply low precision and to improve the precision a module of common words is necessary, even though the common words module implies lower recall.

The combination titles, common words and dictionary module reports high recall for de-identifying the general entity names, and is therefore to be preferred in de-identification systems. In order to optimize the performance of de-identification it may, however, be necessary to combine the modules in a different manner for first names and last names, respectively.

The dictionary module reports better recall for first names than for last names (approximately 26 percent). The analysis of the dictionary module's influence on the results also shows that recall is better for first names when the module is present. This means that first names are more common than last names, due to the fact that the dictionaries contain the most common names in Sweden and the dictionary of last names is larger than the dictionaries of first names.

The analysis shows that the title module increases recall more for last names than for first names, which proves that the title module identifies more last names than first names.

This study should not be considered complete, in future research it would be interesting to investigate how other rules influence the results, e.g. if identification of ambiguous words would increase the precision. We plan to use the GTA – Granska Text Analyzer (Knutsson et al., 2003) – to disambiguate ambiguous words in order to improve our system. This study can also be used for investigating how the size of name dictionaries, as well as the ordering of dictionaries, influence results.

We believe that our results have broaden the field of rule-based de-identification and can be applicable on other languages than Swedish as well.

6. Acknowledgements

Thanks to Martin Hassel for providing dictionaries with common words in patient records and thanks to Sumithra Velupillai for setting up the gold standard. Finally we wish to thank Andreas Amsenius for the evaluation program.

7. References

N. Chinchor. 1997. MUC-7 Named Entity Task Definition. [Online], Published 17 september. Available at: http://acl.ldc.upenn.edu/muc7/ne_task.html [Accessed 5 March 2010].

H. Dalianis and E. Åström. 2001. SweNam-A Swedish Named Entity recognizer. Its construction, training and evaluation. Technical report, TRITA-NA-P0113, IPLab-189, NADA, KTH.

H. Dalianis and S. Velupillai. 2010. De-identifying Swedish Clinical Text – Refinement of a Gold Standard and Experiments with Conditional Random Fields. *To be published in Journal of Biomedical Semantics*.

H. Dalianis, M. Hassel, and S. Velupillai. 2009. The Stockholm EPR Corpus - Characteristics and Some Initial Findings. In *Proceedings of the 14th International Symposium for Health Information Management Research – ISHIMR 2009. Evaluation and implementation of e-health and health information initiatives: international perspectives*, pages 243–249, Kalmar, Sweden, 14-16 October, 2009.

M. Douglass, G.D. Clifford, A. Reisner, W.J. Long, G.B. Moody, and R.G. Mark. 2005. De-identification algorithm for free-text nursing notes. *Computers in Cardiology 2005*, 32:331–334.

HIPAA. 2003. HIPAA Privacy Rule and Public Health: Guidance from CDC and the U.S. Department of Health and Human Services. *Centers for Disease Control and Prevention. MMWR*, 52, April. Available at: http://www.cdc.gov/mmwr/pdf/other/m2e411.pdf (accessed March 22, 2010).

O. Knutsson, J. Bigert, and V. Kann. 2003. A Robust Shallow Parser for Swedish. In *Proc. 14th Nordic Conf. on Comp. Ling. NODALIDA-2003*.

D. Kokkinakis and A. Thurin. 2007. Identification of entity references in hospital discharge letters. In *Proceedings of the 16th Nordic Conference of Computational Linguistics NODALIDA-2007*, pages 329–332, University of Tartu.

A. Mikheev, M. Moens, and C. Grover. 1999. Named entity recognition without gazetteers. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, pages 1–8, Morristown, NJ, USA. Association for Computational Linguistics.

I. Neamatullah, M.M. Douglass, L.H. Lehman, A. Reisner, M. Villarroel, W.J. Long, P. Szolovits, G.B. Moody, R.G. Mark, and G.D. Clifford. 2008. Automated de-identification of free-text medical records. *BMC Medical Informatics and Decision Making*, 8:32.

L. Sweeney. 1996. Replacing Personally-Identifying Information in Medical Records, the Scrub System. In *Proceedings of The AMIA Annual. Fall Symposium 1996*, pages 333–337.

S.M. Thomas, B. Mamlin, G. Schadow, and C. McDonald. 2002. A Successful Technique for Removing Names in Pathology Reports Using an Augmented Search and Replace Method. In *Proceedings of the AMIA Annual Symposium 2002*, pages 777–781.

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

Ö. Uzuner, T.C. Sibanda, Y. Luo, and P. Szolovits. 2008. A de-identifier for medical discharge summaries. *Artificial Intelligence in Medicine*, 42(1):13–35.

S. Velupillai, H. Dalianis, M. Hassel, and G.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 special issue on Mining of Clinical and Biomedical Text and Data*.