Identification of Entity References in Hospital Discharge Letters

Dimitrios Kokkinakis
Göteborg University
Department of Swedish Language,
Språkdata
Sweden
svedk@svenska.gu.se

Anders Thurin
Clinical Physiology
Sahlgrenska University Hospital/Östra
Sweden
anders.thurin@vgregion.se

Abstract

In the era of the Electronic Health Record the release of medical narrative textual data for research, for health care statistics, for monitoring of new diagnostic tests and for tracking disease outbreak alerts imposes tough restrictions by various public authority bodies for the protection of (patient) privacy. In this paper we present a system for automatic identification of named entities in Swedish clinical free text, in the form of discharge letters, by applying generic named entity recognition technology with minor adaptations.

1 Introduction

There is a constantly growing demand for exchanging clinical and health-related information electronically. On a daily basis, hospitals store vast amounts of patient data as free text, but due to confidentiality requirements these texts remain inaccessible for research and knowledge mining. Therefore, an anonymisation or de-identification system can provide a broad spectrum of services related to the growing demands for better forms of dissemination of confidential information about individuals (Personal Health Information – PHI) found in electronic health records (EHR) and other clinical free text (e.g. discharge letters).

In this paper we present an anonymisation system for Swedish, which re-uses components of a generic named entity recognition system (NER) (Kokkinakis, 2004). Generic NER is the process of identifying and marking all single or multi-word named persons, location and organizations, including time and measure expressions, or other entities of interest in free text. NER is considered a mature technology that has numerous applications in a number of human language technologies, including information retrieval and extraction, topic categorization and machine translation. NER serves also as an important supporting technology for providing annotations for the Semantic Web.

We start by defining the notions of anonymisation and de-identification (Section 2) and give an overview of related work in Section 3. The method we use is described in Section 4. The system is based on two major components, a rule-based mechanism which makes use of classificatory criteria provided by the local context (e.g. trigger words, morphological prefixes/suffixes), and extended lists of various types of named entities. Section 5 provides a description of the medical data, while experimental results and evaluation are described in Section 6. Finally, Section 7 presents our conclusions.

2 Anonymisation *vs.* De-Identification

We define as *permanent anonymisation*, or simply *anonymisation* the process of recognizing and deliberately removing named entities and other identifying information about entities, including time expressions. Information about individuals, e.g. patients, may also include numerical, e.g. demographic or nominative information, such as age, sex, nationality and social security number, hence making the re-identification of those entities (particularly individuals) extremely difficult.

We further define as *de-identification* or *depersonalization* the process of recognizing and deliberately changing, masking, replacing or concealing the names and/or other identifying information of relevance about entities. Identified information

Joakim Nivre, Heiki-Jaan Kaalep, Kadri Muischnek and Mare Koit (Eds.)
NODALIDA 2007 Conference Proceedings, pp. 329–332

may be stored separately in an identification database. The linking between text and the identification database can be made by a unique identifier. Hence, making the re-identification or linking of individuals extremely difficult without the use of an appropriate "key".

3 Related Work

Sweeney (1996) describes the "Scrub" system, a set of detection algorithms utilizing word lists and templates that each detected a small number of name types in 275 pediatric records. Sweeney reports high rates on identified PHIs, 99-100%. Ruch et al. (2000) present comparable results with a similar system. However, it is unclear in both studies what the recall figures were. Taira et al. (2002) present a de-identification system using a variety of NLP tools. Each sentence found in a medical report was fed into a lexical analyzer (a database of 64,000 names) which assigned to each token syntactic and semantic information. Rule-based pre-filters were then applied to eliminate non-name candidates (e.g. by using drug name lists). 99,2% precision and 93,9% recall figures are reported. Thomas et al. (2002) used a method based on lists of proper names and medical terms for finding and replacing those in pathology reports. Their approach was based on identifying trigger words such as "Dr" and on the heuristic that "proper names occur in pairs". 98,7% correct identification on the narrative section and 92,7% on the entire report were reported. Sweeney (2002) describes a method, *k-anonymisation*, which de-associates sensitive attributes from the corresponding identifiers. Each value of an attribute, such as date of birth, is suppressed (i.e. replacing entries with a "*") or generalized (i.e. replacing all occurrences of for instance "070208", "070209" etc. with "0702*). Gupta et al. (2004) discusses the interplay between anonymisation and evaluation within the framework of the *De-Id system* for surgical pathology reports. Three evaluations were conducted in turn, and each time specific changes were suggested, improving the system's performance. As the authors claim, "by the end of the evaluation the system was reliably and specifically removing safe-harbor identifiers and producing highly readable de-identified text". For a description of a number of methods for making data anonymous, see Hsinchun et al. 2005. Finally, in the "Challenges in NLP for Clinical Data" workshop (Uzuner et al., 2006) one can find details of the systems that participated in a shared task dealing with the automatic de-identification (age, phone, date, hospital, location, doctor and patient) of medical summaries.

4 Method

Parts of the NER system we use for the annonymisation originate from the work conducted between 2001-03 in the Nomen-Nescio (*cf.* Bondi Johannessen et al., 2005). Five are the major components of the Swedish system:

- lists of multiword entities
- a rule-based component that uses finite-state grammars, one grammar for each type of entity recognized
- a module[1] that uses the annotations produced by the previous two components in order to make decisions regarding entities not covered by the previous two modules[2]
- lists of single names (approx. 80 000)
- a revision/refinement module which makes a final control on an annotated document with entities in order to detect and resolve possible errors and assign new annotations based on existing ones, e.g. by combining annotation fragments.

In the current work, seven types of NEs are recognized[3]: *persons*, *locations*, *organizations*, names of *drugs* and *diseases*, *time expressions* and a set of different types of *measure expressions* such as "age" and "temperature" (Table 1). The annotation uses the XML identifiers ENAMEX, TIMEX and NUMEX; for details see Kokkinakis (2004).

The lack of annotated data in the domain prohibits us from using, and thus training, a statistically

[1] The module is inspired by the *document centred approach* by Mikheev et al. (1999). This is a form of on-line learning from documents under processing which looks at unambiguous usages for assigning annotations in ambiguous words. A similar method has been also used by Aramaki et al., 2006, called *labelled consistency* for de-identification of PHIs.

[2] This module has not used in the current work, since we applied bulk annotation on a very large sample, while this module has best performance in single, coherent articles.

[3] These name categories are a subset of the original system which also covers three more entities, namely *artifacts*, *work&art* and *events* (e.g. names of conferences).

based system. Since high recall is a requirement, and due to the fragmented, partly ungrammatical nature of the data, the rule-based component of the system seemed an appropriate mechanism for the anonymisation task. Only minor parts of the generic system have been modified. These modifications dealt with: i) multiword place entities with the designators "VC", "VåC", "Vårdc" and "Vårdcentral" in attributive or predicative position, which all translate to *Health Care Center*, e.g. "Tuve VC" or "VåC Tuve" – designators frequent in the domain, which were inserted into the rule-based component of the system; ii) the designators "MAVA" *acute medical ward*, "SS", "SS/SU" and "SS/Ö", where "SS" stands as an acronym for the organization "Sahlgrenska Sjukhuset" *Sahlgrenska Hospital* and iii) the development and use of medical terminology, particularly names of pharmaceutical names (www.fass.se) and diseases, particularly eponyms (mesh.kib.ki.se), in order to cover for a variety of names that conflict with regular person names. E.g., the drug name "Lanzo" *lansoprazol* is also in the person's name list, while "Sjögrens" in the context "Sjögrens syndrom" *Sjogren's syndrome* and "Waldenström" in the context *Mb Waldenström* could also be confused with frequent Swedish last names. Therefore, the drug's and disease's modules (which were also evaluated, see Section 6) are applied before the person/location in order to prohibit erroneous readings of PHIs.

An example of annotated data, before [a] and after [b] anonymisation, is given below. The content of anonymised NEs (b) is translated as: uppercase *X* for capital letters, lowcase *x* for lower case characters, and *N* for numbers, while punctuation remains unchanged. The number of the dummy characters in each anonymised NE corresponds to the length of the original NE. However other translation schemes are under consideration. Examples of various NE types are given in table 1.

a. Pat från <*ENAMEX TYPE="LOC"*>Somalia <*/ENAMEX*> op <*TIMEX TYPE= "TME"*>-91<*/TIMEX*> med [...] får <*ENAMEX TYPE= "MDC"*>Waran<*/ENAMEX*> [...] <*ENAMEX TYPE="PRS"*>dr Steffan A. Janson <*/ENAMEX*> rekommenderar biopsi […]

b. Pat från <*COUNTRY*>Xxxxxxx <*/COUNTRY*> op <*TIME*>-NN<*/TIME*> med [...] får Waran [...] <*PERSON*>dr Xxxxxxx X. Xxxxxx<*/PERSON*> rekommenderar biopsi [...]

Entity	Examples
Person	HUM[human]: Dr Janson CLC[group]: 10 HIV-patienter
Location	LOC[country]: Somalia FNC[functional]: Åre VåC
Organization	ORG[organization]: VOLVO
Measure	PSS[pressure]: 120/80 mmHg DSG[dosage]: 10 mg 1x1
Time	TME[time]: aug. 2006
Disease	MDD[disease]: Tourettes
Drugs	MDC[drug]: Waran

Table 1.Entity examples

5 A Corpus of Clinical Data

In this study we used a large corpus (~1GB) of discharged letters extracted from the EHR system MELIOR© used by the Sahlgrenska Univ. Hospital. The corpus consists of database posts taken from tables of special interest for further research (text and data mining) such as "clinical history" and "final diagnoses". The subcorpus we used for the evaluation consists of 200 randomly extracted passages, which we believe gives a good indication of the performance of the NER system. A passage may consist of one or more sentences. The size of the evaluation material was 14,000 tokens. The only pre-processing of the texts has been the tokenization, while the anonymisation and evaluation work was conducted on a locally installed version of the system at the department of Clinical Physiology, at the Sahlgrenska/Östra Univ. Hospital in Gothenburg, behind a firewall, which, at this stage, guarantees maximum security[4].

6 Results and Evaluation

For the evaluation we manually examined the selected sample. We calculated precision, recall and f-score using the formulas: *P = (Total Corr. + Partially Corr.) / All Produced* and *R = (Total Corr. + Partially Corr.) / All Possible*. Partially correct means that an annotation is not completely correct but partial credit should be given, e.g., if the system produces an annotation for "Alzheimers sjukdom" (Alzheimer's disease) as <*ENAMEX TYPE="MDD"*>*Alzheimers*<*/ENAMEX*>*sjukdom*, instead of <*ENAMEX TYPE="MDD"*>Alzheimers

[4] We are currently investigating ways to get the appropriate clearance by the hospital's ethical committee in order to make some of the material available for research, although this might be difficult if results don't reach almost perfect scores.

sjukdom</*ENAMEX*>, then such annotations are given a half point, instead of a perfect score. F-score is calculated as: *F=2*P*R/P+R.*

Entity	P	R	F-score
Person	95,65%	95,65%	95,65%
Location	94,11%	59,25%	72,71%
Organization*	60%	85,71%	70,59%
Time	98,99%	76,03%	86%
Measure	99,19%	93,75%	96,39%
Disease	97,94%	86,81%	92,03%
Pharma/names	95,16%	92,63%	93,82%
Total	96,97%	89,35%	93%

Table 2. Evaluation Results (* only 7 occurr.)

The error analysis we conducted indicates that the performance of the generic NER system is influenced by the features of the domain. We emphasize the word "generic", since simple means can increase the P&R figures dramatically. E.g., the majority of unmarked time expressions were of the form "Number/Number –Number" (1/7 -00), characteristic of the data and not part of the Swedish standard for designating time. The analysis of the results, particularly for the cases that the system failed to produce an annotation (insufficient coverage) or when the annotation was erroneous, revealed that many errors where due to 3 types: i) spelling errors & ungrammatical constructions (e.g. 'ischaemi' – instead of 'ischemi'), ii) insufficient context/short sentences (e.g. 'ACB-op -94' – 'by-pass operation 1994') and iii) abbreviations (e.g. 'på Ger' – at the Geriatric unit – instead of 'på Geriatriken' and 'skivepitel-ca' – squamous cell cancer – instead of 'skivepitel-cancer').

7 Conclusions

We have described a system for anonymising hospital discharge letters using a generic NER system slightly modified in order to cope with some frequent characteristic features of the domain. The coverage of our approach provides a ground for accessing the content of clinical free text in a manner that enables one to draw inferences without violating the privacy of individuals, although some work still remains to be done. For the near future, we intend to: i) evaluate a larger sample and propose adjustments for increased performance; ii) integrate more NE types and iii) get the appropriate approval from the appropriate ethical committees, for releasing some of the data for further research.

Acknowledgements

This work has been partially supported by the "Semantic Interoperability and Data Mining in Biomedicine" – NoE, under EU's Framework 6.

References

Aramaki E., Imai T., Miyo K. and Ohe K. 2006. *Automatic Deidentification by using Sentence Features and Label Consistency*. Challenges in NLP for Clinical Data Workshop. Washington DC.

Bondi Johannessen J. et al. 2005. *Named Entity Recognition for the Mainland Scandinavian Languages*. Literary and Linguistic Computing, Volume 20:1.

Gupta D., Saul M. and Gilbertson J. 2004. *Evaluation of a Deidentification (De-Id) Software Engine to Share Pathology Reports and Clinical Documents for Research*. Am. J. of Clin. Pathology. 121(6): 176-186.

Hsinchun C., Fuller S.S., Friedman C. and Hersh W. 2005. *Medical Informatics – Knowledge Management and Data Mining in Biomedicine*. Pp. 109-121. Springer Series in Information Systems.

Kokkinakis D. 2004. *Reducing the Effect of Name Explosion.* LREC-Workshop: Beyond Named Entity Recognition - Semantic Labeling for NLP. Portugal.

Mikheev A., Moens M. and Grover C. 1999. *Named Entity Recognition without Gazeteers.* Proc. of the 9th European Chapter of the Assoc. of Computational Linguistics (EACL). Pp. 1-8, Bergen, Norway.

Ruch P. et al. 2000. *Medical Document Anonymisation with a Semantic Lexicon*. AMIA: Session S81 - Clinical Information Confidentiality and Security.

Sweeney L. 1996. *Replacing Personally-Identifying Information in Medical Records, the Scrub System*. J. of the American Medical Informatics Association. Washington, DC: Hanley & Belfus, Inc. Pp. 333-337.

Sweeney L. 2002. *k-anonymity: a Model for Protecting Privacy*. Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 10(5): 557-570.

Taira R.K, Bui A.A. and Kangarloo H. 2002. *Identification of Patient Name References within Medical Documents Using Semantic Selectional Restrictions*. AMIA. Pp. 757-61.

Thomas S.M., Mamlin B., Schadow G. and McDonald C. 2002. *A Successful Technique for Removing Names in Pathology Reports Using an Augmented Search and Replace Method*. AMIA. Pp. 777-81.

Uzuner O., Kohane I. and Szolovits P. 2006. *Challenges in NLP for Clinical Data*. (www.i2b2.org/NLP/)