Identification of Patient Name References within Medical Documents Using Semantic Selectional Restrictions

Ricky K. Taira, Ph.D., Alex A. T. Bui, Ph.D., and Hooshang Kangarloo, M.D.

Telemedicine Division, Department of Radiology
UCLA School of Medicine, Los Angeles, CA

***Abstract**: De-identification of a patient's personal data from medical records is a protective legal requirement imposed before medical documents can be used for research purposes or transferred to other healthcare providers (e.g., teachers, students, teleconsultations). This de-identification process is tedious if performed manually, and is known to be quite faulty in direct search and replace strategies [9]. In this paper, we report on the identification step of this process. The proposed algorithm is based on estimating the fitness of candidate patient name references to a set of semantic selectional restrictions. The semantic restrictions place tight contextual requirements upon candidate words in the report text and are determined automatically from a manually tagged corpus of training reports. Maximum entropy classifiers are used to provide a probabilistic measure of the belief of a given candidate token to a given semantic restriction. We report on the design and preliminary evaluation of the system within the domain of pediatric urology.*

INTRODUCTION

Privacy issues regarding a patient's medical record are of increasing concern given the trend toward electronic-based medical records. There are two separate issues that must be addressed: first, security (keeping information from an unauthorized person), and second, confidentiality (keeping patient-specific identifiers confidential, even from authorized users, such as researchers, teachers, etc.). The first issue is addressed by networking (*e.g.*, virtual private networks, VPNs), encryption, and user authentication methods (*e.g.*, login and passwords). The second issue, however, requires removal of patient-specific identifiers from clinical documents. Researchers (and other authorized personnel) who require access to large corpora of confidential medical documents need methods to de-identify these records, as specified by various organizations and regulatory standards set up to protect patient privacy (*e.g.*, Health Insurance Portability and Assurance Act (HIPAA), institutional review boards, Federal Policy for the Protection of Human Subjects) [5]. De-identification of medical records involves two steps: 1) the identification of personally identifying references within a medical text; and 2) the masking, coding, and/or replacing of these references with values irreversible to unauthorized personnel. We report only on the first step, using various natural language processing techniques and classifiers that employ semantic selectional restriction features. References on the task of transforming these references into another representation to assure patient anonymity can be found in [7] (a "one-way" hash) and [10] (the DataFly System).

PREVIOUS WORKS

De-identification of patient records for research has been somewhat of a lax standard. The typical approach is to perform a straightforward global search and replace strategy. In this approach, the patient's identifying information (*e.g.*, name) is given and the algorithm simply searches for all possible combinations of the patient's first, middle, and/or last name within the text report. Such an approach does not handle various false negatives such as nicknames, misspellings, and/or shortened or elaborated forms of a patient's name. Sweeney reported that on a test database of pediatric letters and physician notes, such an approach located no more than 30-60% of personally identifying information [9]. Sweeney reported on a system called *Scrub* that employed a bank of specialized detectors to locate targeted information such as phone numbers, location, age, ID number and patient name references. The detectors located these items utilizing regular expression type templates and specialized lists (*e.g.*, name lists). The Scrub system achieved a true positive detection rate of over 99%. False positive rates were not reported. Sweeney later reported on a clinical system (DataFly) that determined role-based security requirements on the fly. The DataFly system provided an additional level of assurance for patient anonymity [10]. Ruch et al. posed the problem in terms of the natural language processing tasks of word-sense disambiguation and pattern recognition [8]. Syntactic and semantic knowledge were used to classify the tokens within a report into one of about 40 semantic tags. Lexical resources included the MEDTAG lexicon, the Swiss Compendium (drug listing), and a medical devices lexicon. N-gram type rules and finite state automata were used to encode the knowledge required to disambiguate word senses. A recursive transition network (RTN) was used for extracting patient identifiers from the tagged stream of text tokens. A test set of about 800 documents was used to evaluate the algorithm. Out of 467 true instances, 452 (96.8%) were correctly removed, 3 (0.6%) partially removed, 8 (1.7%) completely removed with removal of some relevant tokens, 4 (0.9%) true instances were not removed, and no tokens removed that were not identifiers.

COLLECTION OF TRAINING DATA

The training files for our system development consists of 1350 random reports from pediatric patients generated at the UCLA Clark Urology Center. The categories of the reports included: 1) letters and reports to referring physicians, 2) discharge summaries, 3) clinical notes, and 4) operative/surgical reports.

A researcher reviewed each report within the training corpus, manually tagging all references to patient names and the local contexts in which the names were used. For example, in the sentence:

> *Johnny underwent a pyeloplasty for uretropelvic junction stenosis...*

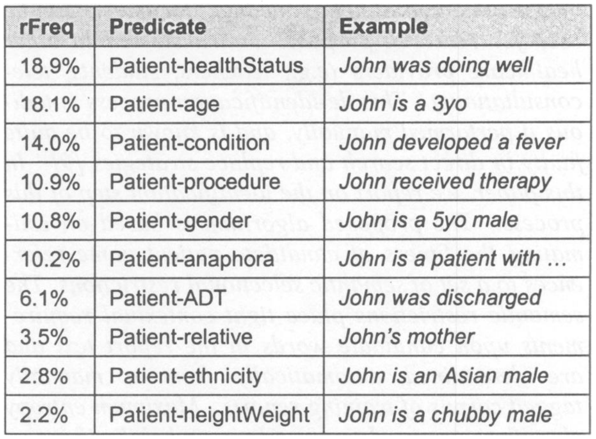
the researcher would indicate *Johnny* as the patient's name and additionally record the tokens *underwent* and *pyeloplasty* as context information. The context tokens are specified in order to learn important semantic selectional restrictions we can impose on references to patient names and their modifying concepts (*e.g.,* age, sex, procedures performed, medical condition, etc.). This tagging scheme thus captures the *logical relation*[1] instance in which the patient plays a role.

A logical relation consists of a predicate and an ordered list of one or more arguments. The predicate indicates the types of relationship between the arguments. In most cases the logical relation consists of three arguments; a head, a relation, and a value. For instance, in the example sentence above, we tag the token *Johnny* as the logical relation head, *underwent* as the relation, and *pyeloplasty* as the value. The predicate type is hasProcedure. In sentences such as:

> *Johnny is a 5 year old Caucasian male with Disease X...*

our convention was to tag the nearest token (*male*) that syntactically modifies its head (*Johnny*). The tokens (*5 year old* and *Caucasian*) modify the word *male*. Within the training set, 486 (36%) of the documents contained patient name references buried within the report text. A total of 907 patient name instances were tagged, all located within the unstructured (*i.e.,* non-header) portions of the report.

The tagged data were stored within a database that maintained for each tagged logical relation instance, the report ID, the predicate name, and the byte offsets of the head, relation and value of the logical relation arguments. Table 1 summarizes the ten most common types of logical relations tagged within the training corpus for characterizing the context in which a patient name is used. The frequencies in this table are expressed in terms of relative percentages. Note that the logical relation "Patient-ID" was also tagged, although this tends to be specific to the report header information specific to the institution.



rFreq	Predicate	Example
18.9%	Patient-healthStatus	*John was doing well*
18.1%	Patient-age	*John is a 3yo*
14.0%	Patient-condition	*John developed a fever*
10.9%	Patient-procedure	*John received therapy*
10.8%	Patient-gender	*John is a 5yo male*
10.2%	Patient-anaphora	*John is a patient with ...*
6.1%	Patient-ADT	*John was discharged*
3.5%	Patient-relative	*John's mother*
2.8%	Patient-ethnicity	*John is an Asian male*
2.2%	Patient-heightWeight	*John is a chubby male*

Table 1: The relative frequency of the ten most common logical relation types tagged within our training corpus. A total of 907 instances were tagged covering 486 positive reports. 63 percent of the reports within the corpus had no patient name references within the body of the report.

DEFINING AN INITIAL HYPOTHESIS SPACE

The first step of our algorithm is the definition of a candidate solution space for patient names within a medical document. We first locate section and sentence boundaries within an input document using the structural analyzer described in [11]. Each sentence within the report is then fed into a lexical analyzer from the same set of natural language processing (NLP) tools. The lexical analyzer step assigns each token to one of twelve syntactic tags and to one of about 200 semantic tags. No word-sense disambiguation is performed, with the first lexical entry of a particular word used. The lexical analyzer has an attached name database of over 64,000 first and last names.

Candidates are initially conservatively proposed as either tokens that match an entry in the name database or any unknown word tokens that are capitalized and that do not contain non-alphabetic characters, with the exception of a hyphen and/or apostrophe. The exception to the capitalization constraint is words from a list of name prefixes (*e.g.*, van, von, de, del, etc.) that are commonly non-capitalized parts of a patient's name. A conservative set of rule-based prefilters are then applied to eliminate obvious non patient name candidates:

- Candidates that match an entry from a 6200 entry drug name list.
- Candidates that are recognized as part of a physician name (*e.g.*, due to identity markers such as Dr., M.D., etc.).
- Candidates that are followed by words such as syndrome, disease, or procedure (*e.g.*, Potter syndrome, Gauche disease, Rashkin procedure).
- Candidates that are recognized as part of a department or institution (*e.g.*, Medical Center).
- Candidates that are recognized as part of a medical device (*e.g.*, Mersilene suture).
- Candidates with article/determiner attachments.

This initial list is conservative in proposing possible patient names and typically includes a fair number of false positives (unknown capitalized words, non-patient names, unknown chemicals, acronyms, etc.).

Hypothesis Testing with Semantic Constraints

The candidate word list provides a high recall, but perhaps unacceptable precision level for identification of patient names. This list represents the *possible* solutions, and at this point does not have attached probabilities. Our system estimates probabilities by examining how well a candidate word can take on the role of the PATIENT within the logical relations defined within the training step (see above). The rationale for this approach is that we hypothesize that patient name references are used very often within a highly focused class of communications. Again, Table 1 is an estimate of our initial exploration of these communications. Thus, the probabilities are assigned based on how well a given candidates satisfies a set of *semantic selectional restrictions*.

In brief, semantic selectional restrictions hypothesize strong associations between some classes of words (*e.g.*, *admitted*) and the semantic constraints on concepts that can fill their thematic roles (*e.g.*, patient names). Semantic selectional restriction rules have previously been used mostly with verbs and the types of words that can fulfill their argument slots. By way of illustration, the verb *underwent* strongly suggests that the head slot is filled by a patient reference. Other examples verb forms with strong associations to patients include: *vomited, administered, discharged,* and *returned*. However, as noted by Jurafsky and Martin [6], verbs are not the only types of words that can impose selectional restrictions on their arguments. Within medical documents, certain adjectives (3 year old, male, Asian) can also impose these strong associations. That is, any mention of gender or age within a medical document most likely refers exclusively to the patient. If these words can be tied grammatically to their corresponding related heads, for example, then this can provide strong contextual evidence for patient name identification.

The problem of patient name identification thus is recast as a detection problem for targeted types of logical relations (*e.g.*, see Table 1). For each of the candidate name tokens, we ask the question: how well do you (the token) fill the PATIENT role within any of the targeted logical relations in the context of the sentence under consideration? Again, the solution involves the two step process of: 1) locating candidate logical relations; and 2) estimating their probability of being a true relation.

A simple template matching technique is utilized to locate all possible candidate logical relation constructions. A logical relation template is a prototype of the pattern to be recognized. For a given sentence, the technique locates all possible combination of words that can fill the roles (*i.e.*, head, relation, and value) of a given logical relation (*e.g.*, isOfGender). This step emphasizes high recall over precision. The template matcher does not consider the spacing of words within the sentence text, but just relative word order and whether a word matches the syntactic and semantic qualifications for a given logical relation role (*i.e.*, head, relation, or value). The template itself is automatically constructed from the training data. For example, in the sentence:

John is a 5 year old male with disease X...

the candidate finder proposes the following logical relations:

isOfAge(John, is, 5 year old)
isOfGender(John, is, male)

For each candidate logical relation instance, we would like to determine the probability $p(a \mid \vec{b})$ that a is true (or a is false), given some sentence context, $\vec{b}$. To estimate this probability, we develop a maximum entropy probabilistic model for each type of logical relation. The statistical model focuses on solving a two-category classification problem: given a candidate logical relation, determine the probability that this relation represents a true instance within the context of the sentence being processed. The maximum entropy model uses the log-linear functional form shown below:

$$p(a \mid \vec{b}) = \frac{1}{Z(\vec{b})} \exp\left(\sum_{1}^{n} \lambda_i f_i(a, \vec{b}) \right)$$

where the summation is expressed over a set of indicator functions that represent the features used to characterize context for a given candidate logical relation. The weighting values, λ_i, for the features, f_i, are determined from the training data previously described. The feature functions, f_i, have the form:

$$f(x,y) = \begin{cases} 1 & \text{if } y = a \text{ and } x = \vec{b} \\ 0 & \text{otherwise} \end{cases}$$

An indicator function, f, can express either positive evidence or negative evidence. Features for the classifier can appear very much like grammar rules, *n*-gram sequences, or any possible evidence based on the semantics, syntax, and/or order of surrounding words. Example binary features may include: whether the head word of candidate logical relation precedes the value word; whether the value word is the closest value candidate to the head word; whether the value word is the object of the relational token; and whether the value word comes immediately before the head word. Features are custom defined for each logical relation classifier. The specificity of features can be very general to very specific. Features can be overlapping in their constraints and even antagonistic. The maximum entropy model is used to integrate these features into a single statistical model in a principled way. The model constrains the estimated distribution to exactly match the expected frequency of features within the training set. Beyond these constraints, the model maximizes the uncertainty so that nothing beyond what is expressed in the training data is assumed. In other words, the maximum entropy algorithm derives a probability distribution that agrees with the empirical distribution of the training data, but is maximally non-committal beyond meeting the observed evidence [1]. The model outputs a single probability value, considering the weighted aggregated evidence provided by the context, $\vec{b}$.

Using the statistical models above, a probability is calculated for each logical relation candidate. A threshold is then used to classify whether a logical relation represents a true or false instance. Instances that are classified as true are then instantiated. The final set of instantiated logical relations does not guarantee that all possible patient name references have been identified. This possibility will occur if a patient name reference is used in a context not seen in the training data. We partially remedy this problem by following the logical relation detection and identification step with one that simply employs a traditional string search strategy of all document instances that match the instantiated string of the PATIENT role slot of any of the instantiated logical relations. The search algorithm is enhanced with the integration of a modified Soundex algorithm. The identification of logical relations then can be thought of as a way of building a set of reliable guesses to patient name references. These guesses can then be used to identify all instances of these guesses within the document.

RESULTS

A preliminary evaluation of our algorithm was performed as follows:

1. 900 random test reports from the pediatric urology clinic were retrieved from our clinical database at UCLA.
2. Patient name references for each report were hand-tagged by an individual not involved in development and recorded within a database. This served as the gold standard.
3. Each of the 900 reports was processed by the system. The time to process a 5Kb report was about one-half minute on a 2GHz personal computer. The output was a vector containing the byte offset positions of all identified patient name references.
4. The system output was compared to that of the manually tagged gold standard and ROC data compiled [4]. The area under the ROC curve (A-z) is 0.9735. The best overall performance is seen at a decision threshold of .55 at which the precision score is 99.2% and recall score 93.9%.

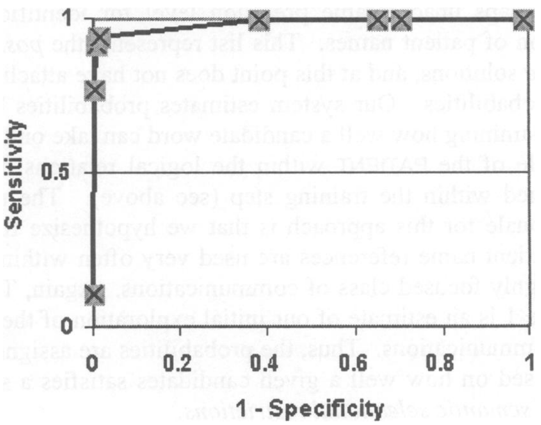


Fig 1 – ROC results for preliminary system evaluation of 900 pediatric urology reports.

DISCUSSION

We present an algorithm which estimates the probability of a patient name reference within the context of a predefined set of logical relations. The logical relations define a template for various semantic selectional restrictions placed on the descriptions of personally identifying information. Preliminary evaluation of the system shows much promise, especially since no initial clues regarding the patient's name was given. Furthermore, the attachment of patient name references to modifying words such as the patient's age, gender, and condition is a nice product of this algorithm and can be used to further de-personalize a patient's medical record.

False Positive Errors: False positive errors were of type: 1) Valid name syntax, but sematically incorrect (e.g., "*Dear Mark, Robert was in our office today*"). 2) identification of a patient's relative rather than the patient (*e.g., Johnny's sister Mary is 7 years old*). 3) Patient name and physician name the same (e.g., *Dr. Martin saw Martin today*). 4) Rare use of gender description not describing Patient name (e.g., Tanner 4 female). 5) Proper drug names that could not be ruled out (not found in drug database, e.g., Droperidol). 6) medical conditions containing a valid human name and not ruled out (*e.g.,* Costello's),

False Negative Errors: False negative errors were of type: 1) Logical relation not modeled (e.g., "*EMLA cream was applied to Johnny's right upper arm*" or "*Johnny got the equivalent of his home TPN last night*"). 2) Grammatically difficult expressions (e.g., *Dear Mark, it was indeed a pleasure for me to have in my office young Johnny*").

Future Work: Development was performed in the domain of pediatric urology. We anticipate that adaptation to new domains will require the introduction of new types of relations (*e.g.,* hasTitle Mr., Mrs.) and adult-only medical conditions (*e.g.,* pregnancy). The current system does not strictly quantify the identification of a patient name as much as it tries to identify the existence of pre-defined logical relations within the text. Identification of erroneous logical relation instances can possibly propagate false positive errors. Several enhancements to the system are planned, including ranking the reliability of a given logical relation type to exclusively identify a patient name reference and improvement of grammar features within the maximum entropy models. On-going work on implementing co-reference models into the algorithm also should give significantly improved results. Clinical utilization of the system is currently ongoing and involves integrating the patient identifier algorithm into a hospital-wide database retrieval system, called DataServer [3].

Acknowledgements

We would like to thank Kenneth M. Louie for his assistance in compiling statistics for Table 1. We also acknowledge the efforts of Carol Demise and David Lee for tagging the training and test corpora.

References

1. AL Berger, SA Della Pietra, and VJ Della Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):-39-71, 1996.
2. H Bouzelat, C Quantin, and L Dusserre. Extraction and anonymity protocol of medical file. In *Proc of AMIA Annual Fall Symposium,* pp 323-327, 1996.
3. AAT Bui, JDN Dionisio, C Morioka, U Sinha, RK Taira, and H Kangarloo. DataServer: An infrastructure to support evidence-based radiology. *Acad Radiology 9:670-678,* 2002.
4. CE Metz. Basic priniples of ROC analysis. *Seminars in Nuclear Medicine* 8:283-298, 1978.
5. Department of Health and Human Services. 45 CFR (Code of Federal Regulations), Parts 160-164. Standards for privacy of individually identifiable health information. *Federal Register,* 65(250):82461-82510, December 28, 2000.
6. D Jurafsky and JH Martin. Speech and language processing: An introduction to natural language processing. *Computational Linguistics and Speech Recognition.* Chapter 16, Prentice Hall, Upper Saddle River, NJ, 2000.
7. C Quantin, H Bouzelat, FA Allaert, AM Benhamische, J Faivre, and L Dussere. Automatic record hash coding and linkage for epidemiological follow-up data confidentiality. *Methods Inf Med* 37:271-277, 1998.
8. P Ruch, RH Baud, A-M Rassinoux, P Bouillon, and G Rober. Medical document anonymization with a semantic lexicon. In *Proc. of the AMIA Fall Symposium*, pp. 729-733, 2000.
9. L Sweeney. Replacing personally-identifying information in medical records, the SCRUB System. In *Proc of AMIA Fall Symposium*, pp.333-337, 1996.
10. L Sweeney. Guaranteeing Anonymity when Sharing Medical Data, the DataFly System, In Proc. of the AMIA Fall Symposium, pp. 51-55 1997.
11. RK Taira and S Soderland. A statistical natural language processor for medical reports. In *Proc of AMIA Fall Symposium*, 970-974, 1999.