

## i2b2 Workshop on Natural Language Processing Challenges for Clinical Records

Ozlem Uzuner, Ph.D., Chair  
SUNY, Albany  
Albany, NY 12222  
ouzuner@albany.edu

Peter Szolovits, Ph.D.  
MIT CSAIL  
Cambridge, MA 02139  
psz@mit.edu

Isaac Kohane, M.D., Ph.D.  
Children's Hospital Boston  
Boston, MA 02115  
Isaac\_kohane@harvard.edu

**Abstract:** This workshop aims to bring together computational linguists and medical informaticians interested in automatic linguistic processing of clinical records such as medical discharge summaries and radiology reports. Lack of a publicly available and standardized data set has been one of the biggest barriers to systematic progress of Natural Language Processing techniques for clinical data.

Within the framework of the i2b2 project, we have generated and released a set of fully de-identified medical discharge summaries to the research community. We prepared two grand challenge questions around this data: What are some methods for automatic de-identification of clinical records and how well do they perform? Can we automatically identify the smoking status of patients based on their clinical records?

We prepared the gold standard for both of these grand challenge questions. We made the data set available to interested researchers and invited them to participate in the grand challenge. At the time of writing, 18 teams had committed to participate.

This workshop serves as a venue for the participants of the grand challenge to demonstrate their systems, present their papers, and discuss future directions in Natural Language Processing and Medical Informatics using Clinical Data.

### WORKSHOP DESCRIPTION

Clinical records, e.g., medical discharge summaries, radiology reports, etc., contain significant medical information which can complement laboratory data in many ways. These records provide evidence for hypothesized situations and reveal the similarities and correlations among different health problems, medications, treatments, etc. However, the information included in these documents is in the form of unstructured, ungrammatical, fragmented English text. This makes the linguistic processing, search, and retrieval of these records very challenging; currently, there are very few tools for automatic linguistic processing of these records. Existing technologies for processing structured information such as databases, and grammatical documents such as journal or news

articles, have limited utility for processing clinical records.

One barrier to the development of natural language processing technologies specific to clinical records is the difficulty of obtaining these records. In the absence of a standardized, publicly-available gold standard, efforts to build appropriate technologies have been limited and fragmented. The lack of standardized, publicly-available gold standard limits the progress of the state-of-the-art in automatic linguistic processing of clinical records, limits the development of technologies available for search and retrieval of clinical text, and as a result limits our ability to make use of the information contained in these records.

As a part of the i2b2 (Informatics for Integrating Biology to the Bedside) project funded by the National Library of Medicine, we have organized a workshop which will bring together medical informaticians, natural language processing researchers, i.e., technology developers, and medical and clinical researchers, i.e., data owners. Our ultimate goal is to foster the symbiotic relationship between these research communities so that through their interactions, they can gain a deeper understanding of possible collaborations that can push the state-of-the-art forward.

To address the problem of limited availability of data that lies at the root of uncoordinated efforts to develop technologies for automatic linguistic processing of clinical data, we have released a set of de-identified clinical records. We have designed this data to evaluate two particular grand challenge questions:

1. What is the state-of-the-art in automatic de-identification of clinical data?
2. How accurately can automatic methods evaluate the smoking status of patients based on their medical records?

We seek to evaluate various approaches to solving these problems on our standardized, public data set. Therefore, we have created a training set that could be used for the development of systems. We will

compare the performance of submitted systems on a held-out test set.

The workshop will provide a venue for the grand challenge participants to demonstrate their systems and to present a short paper or poster discussing their scientific contributions. The best performing system will also be announced during the workshop.

We believe that the availability of clinical records for research and the grand challenge related to processing of these records will be a major resource for both the Natural Language Processing community and the Medical Informatics community. These resources will give the Natural Language Processing researchers access to data that is otherwise very difficult to obtain. The products of the natural language research community will, in turn, enable use of clinical records for answering a wide variety of research questions in medical informatics.

#### **DATA PREPARATION**

The data release was approved by the Institutional Review Boards of the Partners Healthcare System, SUNY, and MIT. The records released were all fully de-identified. This process was conducted in two stages.

In the first stage, an automatic de-identification system was used.<sup>1</sup> Most approaches to de-identification rely heavily on dictionaries and heuristic rules; these approaches fail to remove most personal health information (PHI) that cannot be found in dictionaries. They also can fail to remove PHI that is ambiguous between PHI and non-PHI. Our approach showed that we can de-identify medical discharge summaries using support vector machines that rely on a statistical representation of local context. Comparing our approach with three different systems, we showed that a statistical representation of local context contributes more to de-identification than dictionaries and hand-tailored heuristics; we also showed that when the language of documents is fragmented, local context contributes more to de-identification than sentential context.

In the second stage, the output of the automatic de-identifier was validated manually. Three manual passes were made over each record. Finally, the identified personal health information was replaced with realistic surrogates.

Data for the smoking status evaluation challenge was hand annotated by pulmonologists.

#### **GENERATION OF SURROGATE PHI**

We replaced PHI in the clinical records with realistic surrogates in order to preserve the challenges present

for automatic de-identification systems. Authentic data contains some uncustomary entries for various PHI, e.g., “J Street” can be a hospital, “011406” can be a date. While generating surrogate PHI, we tried to keep such peculiar cases in the data as much as possible.

Our surrogate generation approach permutes the syllables of existing names obtained from the U.S. Census bureau but conforms to the exact format of each authentic PHI. This approach to generating surrogate PHI usually produces entries such as “Valtawnprinceel Community Memorial Hospital” or “GIRRESNET, DIEDREO A”; note that each of these entries follow the exact format of the PHI they were generated to replace. However, this approach can sometimes generate entries such as “Black” and “John” which themselves can be found in the U.S. Census bureau name lists. We make no effort to eliminate such surrogates from the corpus.

Throughout the surrogate generation process, we’ve tried to protect the integrity of the data as much as possible. For example, dates in the same record have the same offset and the proper noun references to the same named entity are replaced with the same surrogate (in the appropriate format). However, the methods employed are fairly basic; therefore, the results are not perfect.

#### **Ambiguity of PHI**

To make the de-identification task challenging, during the surrogate generation process, we introduced some ambiguity in PHI. In other words, we generated surrogate PHI that coincide with medical terms such as diseases, treatments, and medical test names.

We introduced two types of ambiguity:

1. A term can be ambiguous between PHI and non-PHI within the same record.
2. A term can be ambiguous between PHI and non-PHI in the same corpus.

Roughly ~30-40% of all surrogate PHI in our corpus are ambiguous with some disease, treatment, or test name.

#### **Implications**

Because of the randomly generated surrogate names of people, institutions, and places, dictionary-based approaches will be less successful with this data set than with real data. Note that, in reality, many foreign names (with no dictionary matches) exist and are used in discharge summaries.

Because dictionary-based lookup is often used as one source of information in even much more sophisticated

approaches than simple dictionary matching, this dataset is particularly challenging.

Most PHI appearing in the corpus that are in the dictionary are also disease names or other medical terms that were introduced to enhance ambiguity.

#### ANNOTATION FOR THE SMOKING CHALLENGE

The data for the smoking challenge were annotated by pulmonologists. Two pulmonologists annotated each record; in the case of disagreements, judgments from two other pulmonologists were obtained.

For this task, the pulmonologists were asked to classify patient records into five possible categories based on the information contained in the records and based on their medical intuitions.

The categories of smokers were:

- **PAST SMOKER:** A patient whose discharge summary asserts either that they are a past smoker or that they were a smoker a year or more ago but who have not smoked for at least one year.
- **CURRENT SMOKER:** A patient whose discharge summary asserts that they are a current smoker (or that they smoked without indicating that they stopped more than a year ago) or that they were a smoker within the past year.
- **SMOKER:** A patient who is either a CURRENT or a PAST smoker but, whose medical record does not provide enough information to classify the patient as either a CURRENT or a PAST smoker.
- **NON-SMOKER:** A patient whose discharge summary indicates that they have never smoked.
- **UNKNOWN:** The patient's discharge summary does not mention anything about smoking.

Second hand smokers are considered NON-SMOKERS for the purposes of this study, unless there is evidence in their record that they actively smoked. Similarly, as we are only concerned with tobacco smoking, marijuana smoking should not affect the patients' smoking status.

The doctors annotated a total of 1000 records were annotated. Agreement between them on the complete set of records was around 60% (as measured by Kappa). The records that the pulmonologists did not agree on were omitted from the challenge (unless a majority vote could identify a clear label for these records).

#### Implications

The low agreement among the doctors indicates that identification of the smoking status of patients based on the clinical records is challenging even for the educated, human annotators. Note that evaluation of

smoking status based on the medical intuitions of the doctors is even harder; not only because the judgments do not directly rely on the records but also the agreement between the doctors in this case is even lower (Kappa $\approx$ 0.5).

#### AGENDA FOR THE WORKSHOP

In addition to allocating enough time to talks to be delivered by the grand challenge participants, the organizers of this workshop will allow enough time to open brainstorming sessions among all participants about future directions of the field and the data necessary for the necessary progress.

#### ACKNOWLEDGEMENTS

The grand challenge and the workshop were funded by the grant number U54-LM008748 on Informatics for Integrating Biology to the Bedside from National Library of Medicine.

In addition, the organizers are grateful to the following colleagues for their contributions:

- Tawanda Sibanda, M.Eng.
- Tian He, B.S.
- Yuan Luo, B.S.
- Matthew Goldstein
- Steve Kannan, B.S.
- Marcos Athanasoulis, Dr. PH, MPH
- John Brehm, M.D.
- Mike Cho, M.D.
- Henry C. Chueh, M.D.
- Susanne Churchill, Ph.D.
- Marilyn Foreman, M.D.
- Anne Fuhlbrigge, M.D., M.S.
- Vivian S. Gainer
- John Glaser, Ph.D.
- Bree Huning
- Gary Hunninghake, M.D.
- Diane Keogh
- Ross Lazarus, M.D., MPH
- Michael Mendis
- Shawn Murphy, M.D., Ph.D.
- William A. Small, Jr.
- Margarita Sordo, Ph.D.
- Scott Weiss, M.D., MPH
- Qing Zeng, Ph.D. and last but not least,
- American Medical Informatics Association.

#### REFERENCES

- [1] Sibanda, T., and Uzuner, O. Role of Local Context in Automatic Deidentification of Ungrammatical, Fragmented Text. In Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL), 2006.