i2b2 FIRST SHARED-TASK ON NATURAL LANGUAGE CHALLENGES FOR CLINICAL DATA

Ozlem Uzuner[1], Peter Szolovits[2], Isaac Kohane[3]

1. Department of Information Studies, University at Albany, SUNY; 2. MIT CSAIL; 3. Partners Healthcare System

ABSTRACT

This shared-task and workshop aim 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 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 have 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.

The workshop on Challenges in Natural Language Processing for Clinical Data will meet with the Fall Symposium of the American Medical Informatics Association in November. This workshop will provide a venue for the participants of the grand challenge to present their papers and demonstrate their systems.

This project is supported by grant U54LM008748.

Overview and Goals

Clinical records 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 the 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, **we have organized a shared-task and workshop** which will bring together medical informaticians, natural language processing researchers, and medical and clinical researchers. 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 and developed the gold-standard 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 answering these problems on our standardized, public data set. **We have created a training set that could be used for developing systems. We will compare the performance of submitted systems on a held-out test set.**

We are in the process of organizing a **workshop** that will provide a venue for the grand challenge participants to demonstrate their systems and **to present a short paper or poster discussing the scientific contributions behind their systems**. The best performing systems will also be announced during the workshop.

Natural Language Processing for Data Preparation

The data for these challenges were fully de-identified. This process was conducted in two stages.

In the first stage, an automatic de-identification system was used.[1] 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; that when the language of documents is fragmented, local context (captured by the words immediately surrounding the target word) contributes more to de-identification than global context (captured by the information contained in the complete sentence).

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 (PHI) was replaced with realistic surrogates.

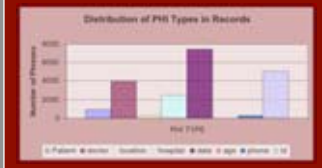
Data for the smoking status evaluation challenge was hand annotated by pulmonologists.

[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.

De-identification Challenge Data

We replaced PHI in the clinical records with realistic surrogates in order to preserve the real 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 kept 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 the 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 references to the same named entity are replaced with the same surrogate. 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. 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 identifying words appearing in the corpus that are in the dictionary are also disease names or other medical terms that were introduced to enhance ambiguity.

Smoking Challenge Data

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

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. 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=~0.5).