Natural Language Processing Framework to Assess Clinical Conditions

HENRY WARE, CHARLES J. MULLETT, MD, PHD, V. JAGANNATHAN, PHD

Abstract
Objective: The authors developed a natural language processing (NLP) framework that could be used to extract clinical findings and diagnoses from dictated physician documentation.

Design: De-identified documentation was made available by i2b2 Bio-informatics research group as a part of their NLP challenge focusing on obesity and its co-morbidities. The authors describe their approach, which used a combination of concept detection, context validation, and the application of a variety of rules to conclude patient diagnoses.

Results: The framework was successful at correctly identifying diagnoses as judged by NLP challenge organizers when compared with a gold standard of physician annotations. The authors overall kappa values for agreement with the gold standard were 0.92 for explicit textual results and 0.91 for intuited results. The NLP framework compared favorably with those of the other entrants, placing third in textual results and fourth in intuited results in the i2b2 competition.

Conclusions: The framework and approach used to detect clinical conditions was reasonably successful at extracting 16 diagnoses related to obesity. The system and methodology merits further development, targeting clinically useful applications.


Introduction
Informatics for Integrating Biology at the Bedside (i2b2) Challenge
To stimulate further advancements in the field of healthcare data extraction from free text clinical documents, the group Informatics for Integrating Biology at the Bedside (i2b2) challenged teams to identify obesity and other co-morbidities from a corpus of dictated discharge summaries. This is their second Shared Task and Workshop Competition and the first one focused on de-identification of clinical documents and detecting smoking status. We chose to participate as it offered us a chance to compare the success of our methods and techniques with others in the field, and, the physician among the authors hopes that NLP extraction techniques will allow free-text physician notes to ‘win out’ over more restrictive templated physician notes for documentation in this era of computerized clinical decision support and automated billing and coding systems.

Background
By extracting data and information from dictated medical reports, NLP can potentially be used to facilitate and improve the process of medical care—particularly when coupled with contemporary clinical information systems. Academic investigators have largely focused on extraction of findings and problems, while the healthcare industry has developed commercial business products. Computer-assisted coding for the submission of billing claims is a thriving market. Our group has previously published an assessment of the use of commercial NLP engines for medication information extraction.

Methods
Research Corpus
The research corpus was supplied by the NLP Challenge organizers. It consisted of a training set of approximately 700 de-identified discharge summaries, each abstracted by two physician experts for the presence or absence of 16 different patient conditions (obesity, diabetes mellitus, etc). The abstractions were further broken down into “textual” and “intuitive.” The textual analysis registered that the patient condition was directly mentioned in the text of the discharge summary, and each scoring for “textual” analysis had four possible values: Yes, No, Questionable, Unknown. The intuitive analysis identified those patient conditions that may not have been directly mentioned, but whose presence or absence could be deduced or inferred by the text of the document. Each scoring for “intuitive” analysis had three possible values: Yes, No, and Unknown. Therefore, the
training dataset that accompanied the 700 discharge summary documents contained roughly $16 \times 2 \times 700$ individual data points.

The challenge organizers focused on obesity and its comorbidities.

**Overview of Procedure**

Documents were first preprocessed to remove unnecessary elements. For each principal concept, the document was textually evaluated by a three-step process for the concept itself and for any secondary concepts. First, each concept would be searched for by regular expression (explicit textual). Then, for each match found, the neighborhood of the match was examined for context. Third, the contextualized matches were brought together in an evaluation for that concept. Once all the principal and secondary concepts had been evaluated, a determination was made for the principal concept.

**Preprocessing**

Portions of the documents that were not germane to the task at hand were removed.

Many of the documents contained a family history section. While these sections did occasionally contain information about the current patient’s current condition, they were mainly focused on the conditions of other family members. In early trials, we determined that these mentions of familial conditions confused the bag-of-words-based context finder. Therefore, this section was removed in a preprocessing step.

Some of the documents contained text identifying an override of the computerized clinical decision alert along with the responsible physician. An example is shown below:

ECASA (ASPIRIN ENTERIC COATED) 325 MG PO QD

Override Notice: Override added on 11/9/01 by

FUDD, ELMER J., M.D.

on order for COUMADIN PO (ref # 00944322 )

POTENTIALLY SERIOUS INTERACTION: ASPIRIN & WARFARIN

Reason for override: md aware

These medico legal snippets sometimes included references to medications the patient was not taking. In addition, these sections confused human readers. Consequently, they were removed in the preprocessing step.

Finally, when the first line of the document contained an admitting diagnosis, this line was removed, as this often represented a working hypothesis and not a confirmed diagnosis. Admitting diagnosis sections were not removed.

**Feature Extraction**

For feature selection we used a combination of guided and manual methods.

Most features were identified by medical relevance. We attempted to first answer the question “why did the physician grade the discharge summary this way?” and then to mimic that process.

Medical concepts were linked to the original 16 whenever they were suspected of affecting an annotation. For example, when it was conjectured that a patient’s history of a myocardial infarction was used by the physician evaluators to establish a diagnosis of CAD, myocardial infarction was added as a concept, linked to CAD.

The most common class of the secondary concepts was medications. Medications were added semantically: with brand names, generic names, and common abbreviations. The concept would generally be at the level of a drug class, such as: steroids, non-steroidal anti—inflammatory drugs (NSAIDs), ACE inhibitors, loop diuretics, thiazide diuretics, β-blockers, and antidepressants. A few concepts were more specific, such as nitroglycerin or albuterol. For example, use of inhaled albuterol suggests the diagnosis of asthma.

Treatments were also used as secondary concepts. For instance, gastric bypass surgery suggested obesity, and pressure stockings indicated venous insufficiency.

Additionally, word bigrams were considered with an inverse document frequency. This brought out a few phrases such as “GI Bleed” in a GERD context.

Each concept had a synonym list. For example, the list of synonyms for albuterol was “salbutamol”, “albuterol”, “ventolin”, “proventil”, and “proair”. For medications, we used the Apelon terminology engine to provide these synonym sets. For a given drug class, this list might contain hundreds of synonyms.

Numerical features were also supported. For example, obesity could be inferred from a numeric BMI or from a numeric weight and height or from a numeric weight alone (for, e.g., >90 kg). Many other numerical features were recognized including ejection fraction, the lipid panel results, and hemoglobin A1c values.

Synonym lists were converted to regular expressions for matching against the clinical documents. The regular expressions were generally case insensitive and matched whole words, but these features could be overridden. For example, the synonym list for myocardial infarction was “MI”, “ami”, “imi”, “myocardial infarc”, “septal infarc”, and “heart attack”.

**Concept Context**

When a concept match was found, the neighborhood around the concept was examined. This approach follows NegEx and related work by Chapman et al. The examination looked for the presence of key phrases within the neighborhood in order. Our implementation differed from Chapman’s in two important ways. First, we processed at the level of characters rather than words—just regular expressions without preliminary lexing. Second we used a neighborhood bounded by any punctuation mark, rather than a five word window.

The first phrases searched for were obliterators. If an obliterator from a concept’s obliterator list was found, the match was determined to be not a match at all. For example, for the concept osteoarthritis, the phrase “arthritis” was taken as a possible match. However, the phrase “rheumatoid arthritis” was used as an obliterator as it describes a different disease from osteoarthritis.

The next phrases searched for were pseudonegators such as “no further ...”; The pseudonegators look like negations but are not. Then, we looked for hypotheses such as “evaluate for.” These indicate that the condition is possible, but not certain. Next, history markers were considered. These are
strings like “h/o”—meaning history of—that indicate the
temporality of the condition. Finally, plain negators were
sought, such as “denies” or “not signs of.”

Rules
We developed and inserted sets of rules to assist with the
intuitive determinations. These rules were framed in a
domain specific language (DSL) written in the scala pro-
gramming language.

In the development of the rules, there was always a tension
between what seemed to achieve better agreement with the
annotators in the training set and what was more medically
correct. Following the annotators too closely can lead to over
fitting, matching noise rather than signal. By contrast, a rule
which does not at first seem medically sound may proxy
some real condition of the patient or thought process of the
annotator.

An illustrative example is Congestive heart failure (CHF). When the concept is not directly mentioned in the text,
several ways of inferring the patient’s status were used.

This condition is unequivocally indicated by a low numeric
ejection fraction; we took 50% as a threshold. Most patients
with CHF will be treated with ACE inhibitors; however ACE
inhibitors are also used to treat other conditions such as
hypertension. In treating hypertension ACE inhibitors are
often prescribed with a thiazide diuretic; this combination for
CHF is rarer. So an ACE inhibitor without a thiazide diuretic
is a possible treatment for CHF; but we would also like to see a
condition associated with CHF. The findings we identified
were a history of heart transplant, the presence of pulmonary
edema, or the finding of a high numeric wedge pressure.

Sample rules described in the DSL are shown below.

```scala
val chfMed = hasAce and! hasThiazide;
Val chfSymptom = hasHeartTransplant |
hasPulmonaryEdema |
hasHighWedgePressure;
Val chfRules = hasLowEjectionFraction |
(chfMed and chfSymptom);
```

Technologies and Support Systems Used

Expert and rule-base systems have a long history in health-
care, going back to the early efforts with Mycin19,20 and
Internist.21 There are several commercial and open source
solutions that implement forward chaining, backward
chaining or even blackboard systems.22 We chose to imple-
ment our solutions from scratch as our goals were more
experimental. The difficult part of building the system was
to develop the feature set—the implementation of the infer-
ce logic was more straightforward. We relied on Apelon
terminology environment to determine relevant medications
for conditions used in the challenge.23

Evaluation Metrics Used

We used Cohen’s kappa metric to measure the progress of
our evolving solution versus the physician gold-standard
document annotators with the training set of documents.
The kappa statistic for two person inter-rater agreement
attempts to generate a measurement of agreement beyond
that expected by chance. The main reason for the selection
of this metric was that a good number of answers could simply
be guessed correctly—i.e., “the person does not suffer from
X” will be mostly true and sometimes overwhelmingly
true.24 Use of this measure allowed us to focus on optimiz-
ing a single metric, as opposed to a suite of metrics.

The NLP challenge organizers provided a utility to measure
six metrics—micro and macro versions of precision, recall
and F-measure. These measures are reported in the results
section.

Results and Discussion

Results
The contest results were graded versus the gold standard
using an F-measure score. This metric allows use of a single
number to describe both precision and recall. The metric can be
calculated by document to generate an F-Micro score and by
category to generate an F-macro score. The F-macro score
weights documents from rare classes much more heavily—in
this case, documents in the “Questionable” category were very
heavily weighted.

Table 1 and Table 2 show our results against the testing set
of documents. Overall, our total average F-measure scores
for the textual and intuitive measurements were very similar
for the test set as they were for the training set, suggesting
that either the two sets of documents were very similar, or
that our system is reasonably robust.

Disease-based kappa scores comparing our system versus
the gold standard of the annotators with adjustment for
inter-annotator agreement are provided in the Jamia “online
only” version of this article at http://www.jamia.org.

For our system, the F-Micro represents largely the behavior
of the synonym lists and rules; whereas the difference
between the F-Micro and the F-macro scores largely indicate
the success of the contextual algorithms. This is because
most mentions of a term are in a positive context. So merely
recognizing the concept by rules and synonyms results in a
good F-Micro score. The F-macro, by contrast rests largely
on correctly determining the context in which a concept
appears, because instances in a non-positive context domi-
nate the F-macro score.

Table 3 shows the computation of the overall kappa. Our
overall kappa values for agreement with the gold standard
were 0.92 for explicit textual results and 0.91 for intuited
results.

Discussion

Textual

General textual performance of the system was good; F-Micro
was over 0.98 for most diseases. For three categories, asthma,
osteoarthritis, and obstructive sleep apnea, the system out-
scored the average estimated human versus the gold standard.

One feature these conditions share is relatively little context: in
particular they are unlikely to be negated.

Our least successful textual area by F-Micro was for coro-
ary artery disease. We were unable to get a good handle on
when a mention of a stenosed artery on cardiac catheteri-
zation was scored as textual CAD, when it was scored as
intuitive CAD and when it was not read as indicative of any
CAD. For the challenge, we went with a short synonym list
with no linkages or presence of terms related to arteries. Our
F-Micro in this area was 0.903.
Intuitive

Our performance on diagnoses evaluated as “questionable” was, perhaps, our weakest overall area. Optimizing our system for this category was generally difficult, as there were relatively few questionable documents in the training set. However, we did not focus as much effort on this area as we could have. Across all sixteen intuitive categories, we correctly identified zero questionable documents in the test set. As a result, our F-macro scores were consistently higher in categories where no questionable documents were present.

Results otherwise were generally good but not as good as the textual results. Five of the 16 categories received an F-Micro of 0.98 or better. The best categories were hypertriglyceridemia and Gout, the least successful category was high cholesterol.

General Observations
We identified concepts in clinical documents using a collection of synonymous terms. If these concepts were found to be in the right context, they provided a reasonable basis for

### Table 1 • Textual Analysis of the Testing Data Set

<table>
<thead>
<tr>
<th>Disease</th>
<th>P-Micro</th>
<th>P-Macro</th>
<th>R-Micro</th>
<th>R-Macro</th>
<th>F-Micro</th>
<th>F-Macro</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obesity</td>
<td>0.9696</td>
<td>0.7373</td>
<td>0.9696</td>
<td>0.4910</td>
<td>0.9696</td>
<td>0.4891</td>
</tr>
<tr>
<td>Depression</td>
<td>0.9783</td>
<td>0.9407</td>
<td>0.9783</td>
<td>0.9757</td>
<td>0.9783</td>
<td>0.9572</td>
</tr>
<tr>
<td>Hypertriglyceridemia</td>
<td>0.9921</td>
<td>0.9960</td>
<td>0.9921</td>
<td>0.8000</td>
<td>0.9921</td>
<td>0.8730</td>
</tr>
<tr>
<td>Gallstones</td>
<td>0.9822</td>
<td>0.9896</td>
<td>0.9822</td>
<td>0.8057</td>
<td>0.9822</td>
<td>0.8095</td>
</tr>
<tr>
<td>OSA</td>
<td>0.9920</td>
<td>0.9817</td>
<td>0.9920</td>
<td>0.6651</td>
<td>0.9920</td>
<td>0.6565</td>
</tr>
<tr>
<td>Asthma</td>
<td>0.9881</td>
<td>0.8995</td>
<td>0.9881</td>
<td>0.8721</td>
<td>0.9881</td>
<td>0.8563</td>
</tr>
<tr>
<td>CAD</td>
<td>0.9034</td>
<td>0.8642</td>
<td>0.9034</td>
<td>0.8988</td>
<td>0.9034</td>
<td>0.8791</td>
</tr>
<tr>
<td>PVD</td>
<td>0.9921</td>
<td>0.9886</td>
<td>0.9921</td>
<td>0.9754</td>
<td>0.9921</td>
<td>0.9819</td>
</tr>
<tr>
<td>Gout</td>
<td>0.9861</td>
<td>0.9607</td>
<td>0.9861</td>
<td>0.9753</td>
<td>0.9861</td>
<td>0.9678</td>
</tr>
<tr>
<td>Diabetes</td>
<td>0.9682</td>
<td>0.9387</td>
<td>0.9682</td>
<td>0.7169</td>
<td>0.9682</td>
<td>0.7949</td>
</tr>
<tr>
<td>CHF</td>
<td>0.9335</td>
<td>0.7653</td>
<td>0.9335</td>
<td>0.8417</td>
<td>0.9335</td>
<td>0.7931</td>
</tr>
<tr>
<td>Venous Insufficiency</td>
<td>0.9803</td>
<td>0.7500</td>
<td>0.9803</td>
<td>0.9899</td>
<td>0.9803</td>
<td>0.8283</td>
</tr>
<tr>
<td>GERD</td>
<td>0.9861</td>
<td>0.9815</td>
<td>0.9861</td>
<td>0.7404</td>
<td>0.9861</td>
<td>0.7359</td>
</tr>
<tr>
<td>OA</td>
<td>0.9801</td>
<td>0.9541</td>
<td>0.9801</td>
<td>0.9788</td>
<td>0.9801</td>
<td>0.9659</td>
</tr>
<tr>
<td>Hypercholesterolemia</td>
<td>0.9681</td>
<td>0.9023</td>
<td>0.9681</td>
<td>0.7590</td>
<td>0.9681</td>
<td>0.8123</td>
</tr>
<tr>
<td>Hypertension</td>
<td>0.9461</td>
<td>0.8383</td>
<td>0.9461</td>
<td>0.7747</td>
<td>0.9461</td>
<td>0.8031</td>
</tr>
</tbody>
</table>

System

<table>
<thead>
<tr>
<th>Textual</th>
<th>P-Micro</th>
<th>P-Macro</th>
<th>R-Micro</th>
<th>R-Macro</th>
<th>F-Micro</th>
<th>F-Macro</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.9718</td>
<td>0.8314</td>
<td>0.9718</td>
<td>0.7542</td>
<td>0.9718</td>
<td>0.7821</td>
</tr>
</tbody>
</table>

OSA = obstructive sleep apnea; CAD = coronary artery disease; PVD = peripheral vascular disease; CHF = congestive heart failure; GERD = gastroesophageal reflux disease; OA = osteoarthritis.

### Table 2 • Intuitive Analysis of Test Data

<table>
<thead>
<tr>
<th>Disease</th>
<th>P-Micro</th>
<th>P-Macro</th>
<th>R-Micro</th>
<th>R-Macro</th>
<th>F-Micro</th>
<th>F-Macro</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obesity</td>
<td>0.9821</td>
<td>0.9812</td>
<td>0.9821</td>
<td>0.9824</td>
<td>0.9821</td>
<td>0.9818</td>
</tr>
<tr>
<td>Depression</td>
<td>0.9706</td>
<td>0.9472</td>
<td>0.9706</td>
<td>0.9709</td>
<td>0.9706</td>
<td>0.9584</td>
</tr>
<tr>
<td>Hypertriglyceridemia</td>
<td>0.9959</td>
<td>0.9789</td>
<td>0.9959</td>
<td>0.9789</td>
<td>0.9959</td>
<td>0.9789</td>
</tr>
<tr>
<td>Gallstones</td>
<td>0.9857</td>
<td>0.9916</td>
<td>0.9857</td>
<td>0.9562</td>
<td>0.9857</td>
<td>0.9729</td>
</tr>
<tr>
<td>OSA</td>
<td>0.9939</td>
<td>0.9855</td>
<td>0.9939</td>
<td>0.6639</td>
<td>0.9939</td>
<td>0.6589</td>
</tr>
<tr>
<td>Asthma</td>
<td>0.9597</td>
<td>0.8928</td>
<td>0.9597</td>
<td>0.9703</td>
<td>0.9597</td>
<td>0.9259</td>
</tr>
<tr>
<td>CAD</td>
<td>0.9585</td>
<td>0.9733</td>
<td>0.9585</td>
<td>0.6377</td>
<td>0.9585</td>
<td>0.6387</td>
</tr>
<tr>
<td>PVD</td>
<td>0.9785</td>
<td>0.9742</td>
<td>0.9785</td>
<td>0.6377</td>
<td>0.9785</td>
<td>0.6392</td>
</tr>
<tr>
<td>Gout</td>
<td>0.9920</td>
<td>0.9692</td>
<td>0.9920</td>
<td>0.9954</td>
<td>0.9920</td>
<td>0.9818</td>
</tr>
<tr>
<td>Diabetes</td>
<td>0.9749</td>
<td>0.9704</td>
<td>0.9749</td>
<td>0.9704</td>
<td>0.9749</td>
<td>0.9704</td>
</tr>
<tr>
<td>CHF</td>
<td>0.9521</td>
<td>0.9678</td>
<td>0.9521</td>
<td>0.6409</td>
<td>0.9521</td>
<td>0.6376</td>
</tr>
<tr>
<td>Venous Insufficiency</td>
<td>0.9578</td>
<td>0.8776</td>
<td>0.9578</td>
<td>0.7536</td>
<td>0.9578</td>
<td>0.8013</td>
</tr>
<tr>
<td>GERD</td>
<td>0.9249</td>
<td>0.9455</td>
<td>0.9249</td>
<td>0.5746</td>
<td>0.9249</td>
<td>0.5903</td>
</tr>
<tr>
<td>OA</td>
<td>0.9481</td>
<td>0.9508</td>
<td>0.9481</td>
<td>0.6210</td>
<td>0.9481</td>
<td>0.6192</td>
</tr>
<tr>
<td>Hypercholesterolemia</td>
<td>0.9072</td>
<td>0.9107</td>
<td>0.9072</td>
<td>0.9162</td>
<td>0.9072</td>
<td>0.9070</td>
</tr>
<tr>
<td>Hypertension</td>
<td>0.9484</td>
<td>0.9271</td>
<td>0.9484</td>
<td>0.9207</td>
<td>0.9484</td>
<td>0.9293</td>
</tr>
</tbody>
</table>

System

<table>
<thead>
<tr>
<th>Intuitive</th>
<th>P-Micro</th>
<th>P-Macro</th>
<th>R-Micro</th>
<th>R-Macro</th>
<th>F-Micro</th>
<th>F-Macro</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.9654</td>
<td>0.6410</td>
<td>0.9654</td>
<td>0.6399</td>
<td>0.9654</td>
<td>0.6404</td>
</tr>
</tbody>
</table>

OSA = obstructive sleep apnea; CAD = coronary artery disease; PVD = peripheral vascular disease; CHF = congestive heart failure; GERD = gastroesophageal reflux disease; OA = osteoarthritis.
making conclusions regarding the patient condition. Whether our methods are scalable to other clusters of diagnoses with only a reasonable amount of work is another question. We expect that analyzing large volumes of dictated data and using supervised learning techniques to guide the selection and creation of rules could allow us to successfully and semi-automatically scale this current approach.

For a variety of applications, ranging from documenting conditions present on admission to the identification of variations in care from published clinical guidelines, the approach investigated in this i2b2 Obesity Challenge appears promising.

**Conclusions**

The results show that our methods were successful as we finished fourth in the intuitive scoring and placed third in the textual scoring in the NLP challenge contest. We conclude that our method of optimizing the matching produced excellent results and provides a viable framework for future applications.

**References**