Extracting medication information from clinical text

Özlem Uzuner, Imre Solti, Eithon Cadag

ABSTRACT
The Third i2b2 Workshop on Natural Language Processing Challenges for Clinical Records focused on the identification of medications, their dosages, modes (routes) of administration, frequencies, durations, and reasons for administration in discharge summaries. This challenge is referred to as the medication challenge. For the medication challenge, i2b2 released detailed annotation guidelines along with a set of annotated discharge summaries. Twenty teams representing 23 organizations and nine countries participated in the medication challenge. The teams produced rule-based, machine learning, and hybrid systems targeted to the task. Although rule-based systems dominated the top 10, the best performing system was a hybrid. Of all medication-related fields, durations and reasons were the most difficult for all systems to detect. While medications themselves were identified with better than 0.75 F-measure by all of the top 10 systems, the best F-measure for durations and reasons were 0.525 and 0.459, respectively. State-of-the-art natural language processing systems go a long way toward extracting medication names, dosages, modes, and frequencies. However, they are limited in recognizing duration and reason fields and would benefit from future research.

INTRODUCTION
Clinical records contain information that can be invaluable, for example, for pharmacovigilance, for comparative effectiveness studies, and for detecting adverse events. The structured and narrative components of clinical records collectively provide a comprehensive account of the medications of patients. The Third i2b2 Workshop on Natural Language Processing Challenges for Clinical Records focused on the extraction of medications and medication-related information from discharge summaries for the purpose of making this information readily available for use. We refer to this challenge as the medication challenge. This article provides an overview of the medication challenge, describes the top 10 systems, evaluates and analyzes their outputs, and provides directions for the continued improvement of the state of the art.

The medication challenge was designed as an information extraction task. The goal for each discharge summary, was to extract the following information (called ‘fields’) on medications experienced by the patient:
- Medications (m): including names, brand names, generics, and collective names of prescription substances, over the counter medications, and other biological substances for which the patient is the experiencer.
- Dosages (do): indicating the amount of a medication used in each administration.
- Modes (mo): indicating the route for administering the medication.
- Frequencies (f): indicating how often each dose of the medication should be taken.
- Durations (du): indicating how long the medication is to be administered.
- Reasons (r): stating the medical reason for which the medication is given.
- List/narrative (ln): indicating whether the medication information appears in a list structure or in narrative running text in the discharge summary.

The medication challenge asked that systems extract the text corresponding to each of the fields for each of the mentions of the medications that were experienced by the patients. The text corresponding to each field was specified by its line and token offsets in the discharge summary so that repeated mentions of a medication could be distinguished from each other. The values for the set of fields related to a medication mention, if presented within a two-line window of the mention, were linked in order to create what we defined as an ‘entry’. If the value of a field for a mention was not specified within a two-line window, then the value ‘nm’ for ‘not mentioned’ was entered and the offsets were left unspecified. Medication and list/narrative fields could never be ‘nm’. The list/narrative field required no offset as no text was extracted for it.

Figure 1 and figure 2 (available online at http://jamia.bmj.com) show two excerpts along with the ‘gold standard’ entries corresponding to the italicized portions of these excerpts. In both figures, there is one entry per mention of each medication. Each entry lists the fields that relate to the same medication mention. The fields are separated by double pipes (||).

i2b2 provided annotation guidelines for the medication challenge, released a small set of discharge summaries with their gold standard annotations, and invited the development of systems that could address the challenge. Twenty teams from 23 organizations and nine countries participated in the challenge. The findings of the medication challenge were presented in a workshop that i2b2 organized, in co-sponsorship with the American Medical Informatics Association (AMIA), at the Fall Symposium of AMIA in 2009.

RELATED WORK
The medication challenge continued i2b2’s efforts to release clinical records to the natural language processing (NLP) community for the advancement of state of the art. This challenge built on past i2b2 challenges, as well as past shared task efforts outside of the clinical domain.

The Genomics Track of Hersh et al. in the Text Retrieval Evaluation Conferences encouraged the...
development of systems that can retrieve the documents relevant to a query from a set. The BioCreAtIvE of Hirschman et al.\(^6\) evaluated systems that can extract biomedical entities and their relations from bioscience text. The ICD-9-CM coding challenge of Pestian et al.\(^2\) ranked practical systems that can automatically classify radiology reports.

Past i2b2 challenges complemented these shared-task efforts. The first i2b2 challenge included an information extraction task focused on deidentification\(^5\) and a document classification task focused on the smoking status of patients.\(^7\) The second i2b2 challenge was a multi-document classification task on obesity and 15 co-morbidities; this challenge required information extraction for disease-specific details, negation and uncertainty extraction on diseases, and classification of patient records.\(^8\)

The third i2b2 challenge—that is, the medication challenge—extends information extraction to relation extraction; it requires extraction of medications and medication-related information followed by determination of which medication belongs to which medication-related details. For an overview of studies focused on medication extraction, please see online supplements at http://jamia.bmj.com.\(^12\)–\(^16\)

**DATA**

The medication challenge was run on deidentified discharge summaries from Partners Healthcare. The summaries were released under a data use agreement that allows their use for research beyond the challenge. All relevant institutional review boards approved the challenge and the use of the discharge summaries.

A total of 1243 deidentified discharge summaries were used for the medication challenge; 696 of these summaries were released during the development period. The i2b2 team generated ‘gold standard’ annotations for 17 of the 696. A total of 547 discharge summaries were held out for testing. Challenge participants (after turning in their system submissions) collectively developed the gold standard annotations for 251 summaries out of the 547. Details of this annotation process can be found in Uzuner et al.\(^3\) Table 1 (available online at http://jamia.bmj.com) shows the number of examples of each field that were found in the gold standard annotations and the average per discharge summary.

**METHODS**

We defined two sets of evaluation metrics, referred to as the horizontal and vertical metrics, for evaluating the system outputs submitted to the challenge. We used horizontal metrics for measuring system performance on ‘entries’; we used vertical metrics for measuring performance on individual ‘fields’. We computed both metrics at phrase and token level, using precision, recall, and F-measure. For our purposes, phrases refer to the complete text of field values; tokens are delimited by spaces and punctuation. We defined phrase-level and token-level precision, recall, and F-measure on the basis of the definitions provided for evaluating the ‘list questions’ of the Question Answering track of the Text Retrieval Evaluation Conferences (see equations 1–6).\(^17\)

**Phrase-level precision (PP) =**

\[ \frac{\#Correctly \ returned \ phrases \ by \ system}{\#Phrases \ returned \ by \ the \ system} \]

**Phrase-level recall (PR) =**

\[ \frac{\#Correctly \ returned \ phrases \ by \ system}{\#Phrases \ in \ gold \ standard} \]

**Phrase-level F-measure =**

\[ \frac{(\beta^2 + 1) \times PP \times PR}{(\beta^2 \times FP) + PR} \] where \( \beta = 1 \)

**Token-level precision (TP) =**

\[ \frac{\#Correctly \ returned \ tokens \ from \ each \ phrase \ in \ system \ output}{\#tokens \ in \ system \ output} \]

**Token-level recall (TR) =**

\[ \frac{\#Correctly \ returned \ tokens \ from \ each \ phrase \ in \ system \ output}{\#tokens \ in \ gold \ standard} \]

**Token-level F-measure =**

\[ \frac{(\beta^2 + 1) \times TP \times TR}{(\beta^2 \times TP) + TR} \] where \( \beta = 1 \)

Entries that are ‘nm’ in the gold standard are omitted from the evaluation. We align entries in the system output with entries in the gold standard using the medication name and offset. In the case of multiple entries for a medication at a given offset, for each entry of the gold standard, the system entry that gives the best F-measure is selected as the alignment match; each gold standard entry can match only one entry in the system output. The list/narrative field is omitted from this process as it requires no text to be extracted from the discharge summaries. Pseudocode for this alignment is presented in Figure 3 (available online at http://jamia.bmj.com).

**Horizontal metrics**

The primary evaluation metric for the challenge computed phrase-level and token-level F-measures on entries. We refer to this metric as the horizontal metric.

**Vertical metrics**

Vertical metrics measure phrase-level and token-level performance on individual fields. For example, vertical metrics for medications measure performance only on medication names without any consideration of the entries of these medications.

**Significance tests**

We used approximate randomization for testing significance.\(^18\) For a pair of outputs \( A \) and \( B \) from two different systems, we
computed the horizontal phrase-level F-measure of each output and noted the difference in performance ($f_{j} = A - B$). Let $j$ be the number of entries in $A$, and let $k$ be the number of entries in $B$; $A$ and $B$ are combined to a superset of entries, $C$, of size $j + k$. For each iteration $i$ up to $n$ iterations, we draw $j$ entries randomly and without resampling from $C$ and create the pseudoset of entries $A_i$. The remainder of $C$ then forms the pseudoset of entries $B_i$. We compute the phrase-level horizontal F-measures for $A_i$ and $B_i$, and note the difference between them ($f_{j} = A_i - B_i$). We then count the number of times that $f_{j} = 0$, for all $i$, and divide this count by $n$ to calculate the p value between $A$ and $B$. We ran significance tests on the outputs of the top 10 teams with $n = 1000$.

**Systems**

Twenty teams participated in the medication challenge (see table 2 available online at http://jamia.bmj.com). The teams received the annotated discharge summaries and the annotation guidelines in June 2009. They were then given approximately 2.5 months to develop their systems and to (optionally) annotate their own training data. At the end of the development period, i2b2 released test discharge summaries and collected system outputs on the test summaries 3 days after their release. Each team was permitted to submit up to three system outputs for evaluation. At the time of submission, the teams classified their system outputs along three dimensions.

1. **External resources** (marked as ‘Yes’ or ‘No’): system outputs that used proprietary systems, data, and resources that were not available to other teams were considered to have been generated with the aid of external resources and were marked as ‘Yes’.
2. **Medical experts** (marked as ‘Yes’ or ‘No’): system outputs that were the product of a team that involved medical experts—for example, medical doctors, nurses, etc.—were marked as ‘Yes’.
3. **Methods** (marked as ‘Supervised’, ‘Rule-based’, or ‘Hybrid’): system outputs were marked on the basis of the approaches that the teams utilized.

Each team was evaluated on the basis of their best performing system output. Of the 20 evaluated outputs, four were declared to have utilized external resources, five were declared to have benefitted from medical experts, 10 were described by their authors as rule-based, four as supervised, and six as hybrids.

Seven out of the top 10 teams’ best performing submissions were rule-based. In general, these teams applied text filtering to eliminate the content that was not related to the medications of the patient. They built vocabularies from publicly available knowledge sources such as the Unified Medical Language System (UMLS), enriched these vocabularies with examples from the training data and the annotation guidelines, and bootstrapped examples from unlabeled i2b2 discharge summaries as well as the web. Given their vocabularies, the teams employed pattern matching and regular expressions to determine most of the medication information. However, they differed in their core tools: BME-Humboldt used GNU software for regular expressions and the Unstructured Information Management Architecture (UIMA) as their base; NLM used MetaMap for marking reasons; Open University (OpenU) employed the Genia Tagger for part of speech tagging; Vanderbilt University used the MedEx system for tagging medication names, frequencies, etc and a context free grammar for converting the text into a structured format.

The University of Utah (UoUtah), the University of Wisconsin-Milwaukee (UWisconsinM), and the University of Sydney (USyd) submitted hybrid systems. These teams also obtained, compiled, and used vocabularies; however, they utilized these resources with various machine learning approaches. University of Utah compiled a knowledge base for medical reasons and extended it with software components and databases that included OpenNLP, MetaMap, and UMLS. University of Wisconsin-Milwaukee trained conditional random fields (CRFs) and combined its output with output from a rule-based system for determining medication names. They used Adaboost trained on annotated data for pairing medication names with their doses, modes, frequencies, durations, etc. They used support vector machines (SVMs) to distinguish narrative and list entries. University of Sydney combined CRFs with SVMs and rules. They used a CRF to identify entities. They relied on SVMs to determine whether two entities were related. They used pattern matching rules to supplement the detection of fields, to assign the scope of negation, and to declare entries as list or narrative.

### RESULTS AND DISCUSSION

We ranked the system outputs on the basis of horizontal metrics.

#### Horizontal results and discussion

Tables 3 and 4 (and tables 5 and 6 available online at http://jamia.bmj.com) show phrase-level and token-level horizontal performances of complete system outputs, outputs that represent only

<table>
<thead>
<tr>
<th>Rank</th>
<th>Group (external resources, medical experts)</th>
<th>Phrase-level horizontal evaluation</th>
<th>Narrative evaluation</th>
<th>List evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Overall</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>1</td>
<td>USyd (N,Y)</td>
<td>0.896</td>
<td>0.820</td>
<td>0.857</td>
</tr>
<tr>
<td>2</td>
<td>Vanderbilt (Y,Y)</td>
<td>0.840</td>
<td>0.803</td>
<td>0.821</td>
</tr>
<tr>
<td>3</td>
<td>Manchester (N,N)</td>
<td>0.864</td>
<td>0.766</td>
<td>0.812</td>
</tr>
<tr>
<td>4</td>
<td>NLM (N,N)</td>
<td>0.784</td>
<td>0.823</td>
<td>0.803</td>
</tr>
<tr>
<td>5</td>
<td>BME-Humboldt (N,N)</td>
<td>0.841</td>
<td>0.758</td>
<td>0.797</td>
</tr>
<tr>
<td>6</td>
<td>OpenU (N,N)</td>
<td>0.850</td>
<td>0.748</td>
<td>0.796</td>
</tr>
<tr>
<td>7</td>
<td>UParis (N,N)</td>
<td>0.799</td>
<td>0.781</td>
<td>0.780</td>
</tr>
<tr>
<td>8</td>
<td>ULMIS (N,N)</td>
<td>0.827</td>
<td>0.725</td>
<td>0.773</td>
</tr>
<tr>
<td>9</td>
<td>UofUtah (N,Y)</td>
<td>0.832</td>
<td>0.715</td>
<td>0.769</td>
</tr>
<tr>
<td>10</td>
<td>UWisconsinM (N,N)</td>
<td>0.904</td>
<td>0.661</td>
<td>0.764</td>
</tr>
</tbody>
</table>

N and Y in parentheses indicate whether the system was declared to have used external resources or benefitted from medical experts. N, No; Y, Yes.
list entries, and outputs that represent only narrative entries. The narrative entries are obtained from the complete system outputs by filtering out the entries that are marked ‘list’ for list/narrative field. For both phrase-level and token-level evaluation, performance on list entries was higher than performance on narrative entries.

Table 3 (and table 5 available online at http://jamia.bmj.com) show that the University of Sydney’s system ranked first in both phrase-level and token-level horizontal evaluation. Table 4 (and table 6 available online at http://jamia.bmj.com) shows that their system performed significantly differently from the other nine systems. Vanderbilt University’s system ranked second in both phrase-level and token-level horizontal evaluation. However, it was not significantly different from the University of Manchester (Manchester), BME-Humboldt, the Open University, and the University of Wisconsin-Milwaukee systems.

Tables 3 and 4 show that systems whose performances vary greatly in terms of the horizontal phrase-level F-measures may not be significantly different from each other. For instance, despite having a lower phrase-level horizontal F-measure than all other top 10 systems, University of Wisconsin-Milwaukee’s system is statistically indiscernible from all but two systems, including one of the top three. In terms of the phrase-level horizontal F-measures, the only systems to perform significantly differently from all systems that scored below them came from the University of Sydney and the University of Manchester.

Figures 4 and 5 (available online at http://jamia.bmj.com) plot the recall and precision of the top 10 systems. University of Wisconsin-Milwaukee’s system shows the highest precision and NLM’s system gives the highest recall in both phrase-level and token-level horizontal evaluation.

Vertical results and discussion

Tables 7–10 (available online at http://jamia.bmj.com) show the vertical evaluation of the top 10 systems. In both phrase-level and token-level vertical evaluation, University of Sydney’s system had the best performance in medications, dosages, modes, and frequencies; University of Manchester’s system had the best performance in durations and reasons. In general, the top 10 teams achieved more than 0.75 for F-measures in medications, dosages, modes, and frequencies. On the other hand, the teams performed poorly on durations and reasons, with F-measures ranging from 0.03 to 0.525 in these two fields.

Table 7 and 9 (available online at http://jamia.bmj.com) show that duration and reason were the hardest fields to extract for all teams. Our analysis showed that durations and reasons showed much variation in their content, were greater in their length (as measured by number of tokens) than the rest of the fields, and also showed greater variability in length (see expanded Section 5.2 in online supplements), which made them difficult to capture.\(^{(31)}\)

CONCLUSION

The Third i2b2 Workshop on Natural Language Processing Challenges for Clinical Records attracted 20 international teams and tackled a complex set of information extraction problems. The results of this challenge show that the state-of-the-art NLP systems perform well in extracting medication names, dosages, modes, and frequencies. However, detecting duration and the reason for medication events remains a challenge that would benefit from future research.

Acknowledgments

This work was supported in part by the NIH Roadmap for Medical Research, Grant U54LM008746 in addition to grants 2 T15 LM007442-06, 5 U54 LM008748, and 1K99LM010227-0110; and grants T15 LM07442, HHSG 272200700057 C, and N00244-09-1-0081. IRB approval was granted for the studies presented in this manuscript. We thank all participants teams for their contributions to the challenge, University of Washington and Harvard Medical School for their technical support, and AMIA for co-sponsoring the workshop.

Funding

NSF, NIH. The project described was supported in part by the i2b2 initiative, Award Number U54LM008748 from the National Library of Medicine. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Library of Medicine or the National Institutes of Health.

Ethics approval

This study was conducted with the approval of the Harvard Partners Health Care, MIT, SUNY, University of Washington.

Provenance and peer review

Not commissioned; externally peer reviewed.

REFERENCES