Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications

Guergana K Savova,1 James J Masanz,1 Philip V Ogren,2 Jiaping Zheng,1 Sunghwan Sohn,1 Karin C Kipper-Schuler,1 Christopher G Chute1

ABSTRACT
We aim to build and evaluate an open-source natural language processing system for information extraction from electronic medical record clinical free-text. We describe and evaluate our system, the clinical Text Analysis and Knowledge Extraction System (cTAKES), released open-source at http://www.ohnlp.org. The cTAKES builds on existing open-source technologies—the Unstructured Information Management Architecture framework and OpenNLP natural language processing toolkit. Its components, specifically trained for the clinical domain, create rich linguistic and semantic annotations. Performance of individual components: sentence boundary detector accuracy = 0.949; tokenizer accuracy = 0.949; part-of-speech tagger accuracy = 0.936; shallow parser F-score = 0.924; named entity recognizer and system-level evaluation F-score = 0.715 for exact and 0.824 for overlapping spans, and accuracy for concept mapping, negation, and status attributes for exact and overlapping spans of 0.957, 0.943, 0.859, and 0.580, 0.939, and 0.839, respectively. Overall performance is discussed against five applications. The cTAKES annotations are the foundation for methods and modules for higher-level semantic processing of clinical free-text.

INTRODUCTION
The electronic medical record (EMR) is a rich source of clinical information. It has been advocated that EMR adoption is a key to solving problems related to quality of care, clinical decision support, and reliable information flow among individuals and departments participating in patient care.1 The abundance of unstructured textual data in the EMR presents many challenges to realizing the potential of EMRs. For example, most EMRs record a narrative describing the history of the current illness for an episode of care. Clinical researchers leverage this information by employing a number of domain experts to manually curate such narratives. This process can be both error-prone and labor-intensive. Automating it is essential to developing an effective EMR infrastructure. Natural language processing (NLP) techniques have demonstrated successes within the clinical domain (for an overview see Meystre et al2). However, their widespread adoption rests on developing comprehensive clinical NLP solutions based on open standards and software.

Our goal is the development of a large-scale, comprehensive, modular, extensible, robust, open-source NLP system designed to process and extract semantically viable information to support the heterogeneous clinical research domain and to be sufficiently scalable and robust to meet the rigors of a clinical research production environment. This paper describes and evaluates our system—the clinical Text Analysis and Knowledge Extraction System (cTAKES).

BACKGROUND
The clinical narrative has unique characteristics that differentiate it from scientific biomedical literature and the general domain, requiring a focused effort around methodologies within the clinical NLP field.2 Columbia University’s proprietary Medical Language Extraction and Encoding System (MedLEE)3 was designed to process radiology reports, later extended to other domains,4 and tested for transferability to another institution.5 MedLEE discovers clinical concepts along with a set of modifiers. Health Information Text Extraction (HITEx)6 7 is an open-source clinical NLP system from Brigham and Women’s Hospital and Harvard Medical School incorporated within the Informatics for Integrating Biology and the Bedside (i2b2) toolset.5 IBM’s BioTeKS9 and MediKAT10 were developed as biomedical-domain NLP systems. SymText and MPLUS11 12 have been applied to interpret the interpretations of lung scans13 to detect pneumonia14 and central venous catheters mentions.15 Other tools developed primarily for processing biomedical scholarly articles include the National Library of Medicine MetaMap,16 providing mappings to the Unified Medical Language System (UMLS) Metathesaurus concepts.17 18 Those from the National Center for Text Mining (NaCTeM),19 JULIE lab,20 and U-compare21 with some applications to the clinical domain.22 Within the Cancer Biomedical Informatics Grid23 initiative, the University of Pittsburgh’s Cancer Tissue Information Extraction System24 25 aims at extracting information from surgical pathology reports using the National Cancer Institute Enterprise Vocabulary System26 and MetaMap.

In the general domain, a number of open-source NLP toolsets exist. The OpenNLP suite27 implements a maximum entropy (ME) machine learning (ML) classifier28–30 with probabilities that maximize information entropy and that are derived from a dataset to include a diverse feature set.31 32 Buyko and colleagues33 adapted the OpenNLP to the biomedical scientific literature and demonstrated...
that the performance of OpenNLP’s components on biotexts represented by GENIA\textsuperscript{34} and PennBioIE\textsuperscript{35} is comparable to that on the newswire text.

Our unique contribution is an NLP system specifically tailored to the clinical narrative that is large-scale, comprehensive, modular, extensible, robust, open-source and tested at component and system levels.

**DESIGN AND SYSTEM DESCRIPTION**

The cTAKES is a modular system of pipelined components combining rule-based and machine learning techniques aiming at information extraction from the clinical narrative. The gold standard datasets for the linguistic labels and clinical concepts are created on content that is a subset of clinical notes from the Mayo Clinic EMR. Standard evaluation metrics are used to measure the quality of the gold standards and cTAKES performance.

**The cTAKES system**

The cTAKES consists of components executed in sequence to process the clinical narrative with each component incrementally contributing to the cumulative annotation dataset (see figure 1 for a running example). This provides the foundation for future cTAKES modules for higher-level semantic processing of the clinical free-text.

The cTAKES accepts either plain text or clinical document architecture-compliant\textsuperscript{36} XML documents. The current open-source release consists of the following components/annotators:

- Sentence boundary detector
- Tokenizer
- Normalizer
- Part-of-speech (POS) tagger
- Shallow parser
- Named entity recognition (NER) annotator, including status and negation annotators.

**An example of a sentence discovered by the sentence boundary detector:**

Fx of obesity but no fx of coronary artery diseases.

**Tokenizer output – 11 tokens found:**

Fx of obesity but no fx of coronary artery diseases.

**Normalizer output:**

Fx of obesity but no fx of coronary artery diseases.

**Part-of-speech tagger output:**

Fx of obesity but no fx of coronary artery diseases.

**Shallow parser output:**

Fx of obesity but no fx of coronary artery diseases.

**Named Entity Recognition – 5 Named Entities found:**

Fx of obesity but no fx of coronary artery diseases.

- obesity (type=diseases/disorders, UMLS CUI=C0028754, SNOMED-CT codes=308124008 and 5476005)
- coronary artery diseases (type=diseases/disorders, CUI=C0010054, SNOMED-CT=8957000)
- coronary artery (type=anatomy, CUI(s) and SNOMED-CT codes assigned)
- artery (type=anatomy, CUI(s) and SNOMED-CT codes assigned)
- diseases (type=diseases/disorders, CUI=C0010054)

**Status and Negation attributes assigned to Named Entities:**

Fx of obesity but no fx of coronary artery diseases.

- obesity (status = family_history_of; negation = not_negated)
- coronary artery diseases (status = family_history_of, negation = is_negated)

**Figure 1** Example sentence processed through cTAKES components ‘family history of obesity but no family history of coronary artery diseases.’

Fx, family history.
drugs, the latter includes terms from the Orange Book\textsuperscript{43} that have an RxNORM code. This dictionary was enriched with synonyms from UMLS and a Mayo-maintained list of terms. On the basis of the output from the shallow parser, the algorithm finds all noun phrases, which become the look-up window. The dictionary is interrogated for permutations of variations of the head and modifiers within the noun phrases to account for non-lexical variations. The NER component does not resolve ambiguities that result from identifying multiple terms in the same text span.

The negation annotator implements the NegEx algorithm,\textsuperscript{44} which is a pattern-based approach for finding words and phrases indicating negation near named entity mentions. The status annotator uses a similar approach for finding relevant words and phrases that indicate the status of a named entity.

Each discovered named entity belongs to one of the dictionary semantic types and has attributes for (1) the text span associated with the named entity (‘span’ attribute), (2) the terminology/ontology code the named entity maps to (‘concept’ attribute), (3) whether the named entity is negated (‘negation’ attribute), and (4) the status associated with the named entity with a value of current, history of, family history of, possible (‘status’ attribute). These semantic types and their attributes were selected in consultation with clinical researchers and practitioners and supported by an analysis of clinical questions and retrieval requests where the most frequent UMLS types and groups were disorders, clinical drug, sign and symptom, and procedures. Any future event is considered hypothetical; hence, the status value will be set to ‘possible’. Allergies to a given medication are handled by setting the negation attribute of that medication to ‘is negated’. Non-patient experiences are flagged as ‘family history of’ if applicable.

The cTAKES is distributed with the best-performing modules and machine learning models for the results reported below. It is released open-source under an Apache License, Version 2.0, as part of the Open Health Natural Language Processing\textsuperscript{45} Consortium. The cTAKES runs on Apache Unstructured Information Management Architecture (UIMA)\textsuperscript{46,47} and Java 1.5. It has been tested for scalability in a cloud computing environment.\textsuperscript{48}

### Corpus

Absent shared community annotated resources for the clinical domain,\textsuperscript{49} we built our own gold standard datasets for named entity\textsuperscript{50} and linguistic annotations as described in the Gold Standard Corpus and Inter-annotator Agreement subsection. Both datasets are created on content derived from the Mayo Clinic EMR. In addition, we used the gold standard linguistic annotations of the Penn TreeBank (PTB)\textsuperscript{50} and GENIA\textsuperscript{34} corpora.

To measure the quality of the in-house developed annotations, we report inter-annotator agreement (IAA) as the positive specific agreement (PSA)\textsuperscript{51} and \( \kappa \)\textsuperscript{52,53} defined as:

\[
\kappa = \frac{P(a) - P(e)}{1 - P(e)}
\]

where \( P(a) \) is the relative observed agreement among the annotators and \( P(e) \) is the hypothetical probability of chance agreement. \( P(e) \) is computed from a contingency matrix representing agreements and disagreements. \( P(e) \) is the sum of the proportion of annotator A1 assigned tags to the positive row over all assignments and the proportion of A2 assigned tags to the positive column over all assignments, which is added to the proportion of annotator A1 assigned tags to the negative row and the proportion of annotator A2 assigned tags to the negative column.\textsuperscript{53} \( P(e) \) for the shallow parse task follows.\textsuperscript{52}

### Evaluation metrics

We used standard metrics:

\[
\text{recall} = \frac{\text{TruePositives}}{\text{TruePositives} + \text{FalseNegatives}}
\]

### Table 1

(A) Inter-annotator agreement (IAA) as positive specific agreement on all annotations in compared sets (A1, A2, A3, A4 are the annotators; C1 is the gold standard created by A1 and A2; C2 is the gold standard created by A3 and A4). The overall agreement between C1 and C2 when the match criteria require that the spans overlap and the concept, negation, and status are the same is 0.746. (B) F-score and accuracy for system output compared to disorder mention gold standard. Results for attributes should be read ‘given that the spans are exact, the accuracy for attribute X is Y’ or ‘given that the spans overlap, the accuracy for attribute X is Y.’
Application of information technology

\[
\text{precision} = \frac{\text{TruePositives}}{\text{TruePositives} + \text{FalsePositives}}
\]

(5)

\[
\text{F-score} = \frac{2 \times (\text{precision} \times \text{recall})}{\text{precision} + \text{recall}}
\]

(4)

\[
\text{accuracy} = \frac{\text{TruePositives} + \text{TrueNegatives}}{\text{TruePositives} + \text{FalsePositives} + \text{FalseNegatives} + \text{TrueNegatives}}
\]

(5)

Because the NER module allows multiple annotations per span (exact and overlapping), we also computed precision, recall, and F-score for the partial matches, which we report as ‘spans overlap’ in table 1B. This measure has been referred to as soft F1-score in.54

The two OpenNLP parameters for number of iterations (the number of times the training procedure should iterate when finding the model’s parameters) and a cut-off (the number of times a feature must have been seen in order to be considered in the model) are selected by a 10-fold cross validation. The results for the cTAKES ML components are reported on the basis of a 10-fold cross-evaluation by averaging the results from each fold. The tests for statistical significance use a t test for paired-two sample means with no-difference null hypothesis.

STATUS AND REPORT
Gold standard corpus and inter-annotator agreement
Two types of gold standard annotations were created. The linguistic annotations were used to train and evaluate cTAKES machine learning models. The named entity gold standard annotations were used to evaluate the NER component.

Linguistic gold standard
In addition to the widely tested PTB and GENIA corpora, our linguistic gold standard included annotations generated on a clinical corpus sampled from the Mayo Clinic EMR. The Mayo-derived linguistically annotated corpus (Mayo) was developed in-house and consisted of 275 clinical notes (100 650 tokens; 7299 sentences; 61 consult; 1 discharge summary; 4 educational visit; 4 general medical examination; 48 limited exam; 19 multi-system evaluation; 43 miscellaneous; 1 preoperative medical evaluation; 3 report; 3 specialty evaluation; 5 dismissal summary; 73 subsequent visit; 5 therapy; 3 test-oriented miscellaneous).

For the POS tags and shallow parses, we extended the PTB annotation guidelines55 56 to the clinical domain, with specifics for the annotations of numbering, roman numerals, drug names, abbreviations, drug-related attributes, and symbols. Guidelines for annotating the shallow parses (or chunks) included specifics for the annotations of medications, key clinical phrases, negation, and shared modifiers. Examples are in tables A1 and A2 of the online data supplement at http://jamia.bmj.com. The phrasal category with the lowest agreement is the VBG (verb, gerund) and NN (noun, singular or mass) pairing (eg, ‘Her husband is disturbed by her sleeping’).

The shallow parses IAA is very strong (\(\kappa > 0.8\)) (details in tables A4 and A5 of the online data supplement at http://jamia.bmj.com). The phrasal category with the lowest agreement is PRT (particle phrase) (IAA range is 0.500–0.727). For example, in the sentence: ‘She is up for short periods’, LA1 annotated ‘up’ as a PRT, while LA2 as an adverbial phrase (ADVP). This ambiguity affects the annotations of the ADVP category, which has the second lowest agreement (IAA range is 0.841–0.899).

The disagreements for the adjectival phrase (ADJP) category were mainly in spans that included a modifier or a post-modification (eg, ‘Bowel movements otherwise regular’—one annotator chose the span for ‘otherwise regular’ while the other selected ‘regular’ as the ADJP).

The IAA results for the named entity corpus are summarized in table 1A with details described by Ogren et al.59

Sentence boundary detector
Experiments of various corpus combinations were conducted (table 2A). We experimented with iteration values ranging from 100–350 and cut-off values from 1–7. The values that resulted in the maximal accuracies in table 2A are 350 iterations and a cut-off of 6 for GENIA, 350 iterations and a cut-off of 4 for PTB, 100 iterations and a cut-off of 1 for Mayo, and 100 and 4 for the combined data.

There are several major sources of errors: (1) sentences starting with titles, especially ‘Dr’ and ‘Mr’; (2) sentences together, 1466 gold standard Disorder UMLS Semantic group named entity annotations were discovered in it (Ogren et al53; for IAA summary see table 1A). The most important datum that best represents the overall consistency of the gold standard on the entire annotation task is the agreement between C1 and C2 (0.746) for the match criteria that require the spans to overlap and the concept code, negation, and status values to match. These Disorder gold standard named entity annotations are used to evaluate the NER component.

Inter-annotator agreement
The three linguistic annotators (LA) are represented as LA1, LA2, and LA3. The POS tags IAA is excellent, with the same result of 0.993 for positive specific agreement and \(\kappa\) because F(e) is extremely small due to the large number of possible values (details in table A3 of the online data supplement at http://jamia.bmj.com). The main source of disagreement is associated with RP (particle) and RB (adverb) part-of-speech tags (eg, ‘She needs to lie down,’ where the bolded word was assigned an RB tag by LA1 and RP tag by LA2). Another source of disagreement is the VBG (verb, gerund) and NN (noun, singular or mass) pairing (eg, ‘Her husband is disturbed by her sleeping’).

The shallow parses IAA is very strong (\(\kappa > 0.8\)) (details in tables A4 and A5 of the online data supplement at http://jamia.bmj.com). The phrasal category with the lowest agreement is PRT (particle phrase) (IAA range is 0.500–0.727). For example, in the sentence: ‘She is up for short periods’, LA1 annotated ‘up’ as a PRT, while LA2 as an adverbial phrase (ADVP). This ambiguity affects the annotations of the ADVP category, which has the second lowest agreement (IAA range is 0.841–0.899).

The disagreements for the adjectival phrase (ADJP) category were mainly in spans that included a modifier or a post-modification (eg, ‘Bowel movements otherwise regular’—one annotator chose the span for ‘otherwise regular’ while the other selected ‘regular’ as the ADJP).

The IAA results for the named entity corpus are summarized in table 1A with details described by Ogren et al.59

Table 2 Accuracy results from the OpenNLP ME classifier. Rows represent training data; columns are test data. 10-fold cross validation with 80/20 data split for each fold

<table>
<thead>
<tr>
<th></th>
<th>GENIA</th>
<th>PTB</th>
<th>Mayo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genia</td>
<td>0.986</td>
<td>0.646</td>
<td>0.821</td>
</tr>
<tr>
<td>PTB</td>
<td>0.967</td>
<td>0.944</td>
<td>0.940</td>
</tr>
<tr>
<td>Mayo</td>
<td>0.959</td>
<td>0.852</td>
<td>0.947</td>
</tr>
<tr>
<td>GENIA + PTB + Mayo</td>
<td>0.986</td>
<td>0.942</td>
<td>0.949</td>
</tr>
</tbody>
</table>

(a) Accuracy for the openNLP sentence boundary detector

<table>
<thead>
<tr>
<th></th>
<th>GENIA</th>
<th>PTB</th>
<th>Mayo</th>
</tr>
</thead>
<tbody>
<tr>
<td>GENIA</td>
<td>0.986</td>
<td>0.764</td>
<td>0.804</td>
</tr>
<tr>
<td>PTB</td>
<td>0.851</td>
<td>0.969</td>
<td>0.878</td>
</tr>
<tr>
<td>Mayo</td>
<td>0.812</td>
<td>0.844</td>
<td>0.940</td>
</tr>
<tr>
<td>GENIA + PTB + Mayo</td>
<td>0.984</td>
<td>0.969</td>
<td>0.936</td>
</tr>
</tbody>
</table>

(b) Accuracy results for the openNLP POS TAGGER

ME, maximum entropy; POS, part-of-speech.
starting with abbreviations (eg, ‘US’) and month names; (3) numbers and numbered lists; and (4) initials of person names. The first case constitutes roughly 73% of all the false negatives in the PTB corpus and nearly 30% in the Mayo corpus, but it is not present in GENIA because it consists of biomedical articles. The second, third, and fourth cases are frequent in the Mayo corpus because the clinical reports contain more shorthands that are not consistently followed by periods, which creates a challenge for the machine learning models. The clinical data also includes more unconventional abbreviations and short sentences (1–3 words), which offers an explanation for the low cut-off value.

Our results are similar to those reported by Buyko et al.33

**Tokenizer**
The rule-based tokenizer achieves an accuracy of 0.938 without the context-dependent tokenizer (CDT) and 0.949 with it. The remaining errors are generally related to punctuation (the tokens ‘140–150/60–80’ and ‘S/P’ are tokenized into three tokens) or due to lack of relevant rules (the tokens ‘p.o.’ and ‘a.m.’ are tokenized into four tokens). A baseline space-delimited tokenizer achieves an accuracy of 0.716.

**Part-of-speech tagger**
We determined an optimal cut-off of 4 with an iteration cycle of 100. Using these values, different combinations of training and testing corpora were used to build and evaluate the models (table 2B). All results were obtained by using gold standard tokens.

The accuracy of 0.986 on GENIA is comparable to the score of 0.989 on the same corpus reported by Buyko et al.33 The best results for each training corpus were obtained when the model was tested on the same corpus. One could consider this result the ceiling of the tagger performance: 0.969 for PTB, 0.986 for GENIA, and 0.940 for Mayo dataset. This supports results previously reported in the literature33 57 58 and the explanation that each corpus has its specific sublanguage. The differences between the accuracy on PTB+GENIA+Mayo and ceiling accuracy are statistically significant (p<0.01). A conclusion could be drawn that the POS module built using data from PTB, GENIA, and Mayo dataset could be successfully ported across these three domains. We conclude that the optimal parameters for a POS tagger using the OpenNLP technology are a corpus of combined data, a frequency cut-off of 4, and iterations of 100. The per-tag results are high for the adjectival tags (JJ) (accuracy=0.887), noun tags (NN) (accuracy=0.938), and verb tags (VB) (accuracy=0.925), which are prerequisites for a high-performance shallow parser. Error analysis shows that the most frequent failure sources are incorrect NN (5%) or VBN (2%) tags for JJ, incorrect JJ tags (2.5%) for NN and incorrect NN (4.6%), and NNP (1%) tags for VB (details in table A6 of the online data supplement at http://jamia.bmj.com).

**Shallow parser**
The OpenNLP ME model is trained on IOB-formatted data, where B indicates the beginning of the phrase, I the elements within the phrase, and O the elements outside of the phrase. We determined an optimal cut-off of 4 with an iteration cycle of 100, which were used for training the models represented in table 3. The accuracy and F-score were computed using the methodology and evaluation script from the Conference on Computational Natural Language Learning (CoNLL) 2000 shared task. All results were obtained by using gold standard POS tags.

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Accuracy and F-score results for the OpenNLP shallow parser. Rows represent training data; columns are test data. 10-fold cross validation with 80/20 data split for each fold</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy and F-score (from CoNLL script)</td>
<td>PTB</td>
</tr>
<tr>
<td>PTB</td>
<td>0.969/0.954</td>
</tr>
<tr>
<td>GENIA</td>
<td>0.888/0.826</td>
</tr>
<tr>
<td>Mayo</td>
<td>0.880/0.834</td>
</tr>
<tr>
<td>PTB + GENIA + Mayo</td>
<td>0.969/0.953</td>
</tr>
</tbody>
</table>

The F-score of 0.93462 on GENIA is comparable to the score of 0.9560 on the same corpus reported by Buyko et al.—a state-of-the-art figure compared to the performance figures from CoNLL 2000. The best performance on the Mayo dataset is achieved through training on the combined corpus, with an overall F-score of 0.924. However, domain-specific data is critical to achieving best performance. As expected, corpus size is critical to explain the result from training Mayo/testing Mayo experiment. There is significant difference between the results for training PTB/testing PTB and training PTB+GENIA+Mayo/testing PTB, and training Mayo/testing Mayo and training PTB +GENIA+Mayo/testing Mayo (p<0.0001). Comparing the results from training GENIA/testing GENIA and training PTB +GENIA+Mayo/testing GENIA yields a p value of 0.04, which suggests that GENIA has distinct patterns from those in PTB and Mayo.

The categories essential for deep parsing (NP, PP, VP) and NER (NP and PP) have high recognition rates (0.908, 0.954 and 0.956; details in table A7 of the online data supplement at http://jamia.bmj.com).

Error analysis shows that about 7% of the NP errors, 2.6% of the PP errors, and 2.5% of the VP errors are due to incorrect B, I, and O assignments.

**Named entity recognition (NER)**
We reported preliminary results in the paper by Kipper-Schuler et al.33 and table 1B summarizes our most recent evaluation. These results represent an overall cTAKES system evaluation, as the NER module produces the annotations of greatest interest to applications that use cTAKES and because it relies on the output of all the other components. The NER performance is reported in terms of F-score, while the assignment of attribute values for each discovered named entity in terms of accuracy because each attribute is assigned a label. The best results achieve an F-score of 0.715 for exact matches and 0.824 for overlapping matches.

Consistent with our preliminary results,39 mapping to the UMLS concept unique identifier (CUI) accuracy is high for exact span matches. However, it drops when span matching is relaxed to require only that they overlap. This follows from the nature of dictionary-based approaches in which the span identified in the text is used to look up the concept. If the span identified is different than what is in the gold standard, it is no surprise that the phrases map to different terms in the dictionary. For the gold standard annotation of ‘degenerative joint disease’ with CUI C0029408, the algorithm produces the right-boundary match ‘joint disease’ with CUI C0022408 as well as the exact match ‘degenerative joint disease’ with CUI C0029408. As the boundary match is a broader term than the manual annotation, it is considered a CUI non-match. Negation detection achieves an accuracy of around 0.94 for both exact and overlapping spans. Status annotation has an accuracy of 0.859, which is comparable to the IAA of 0.860. We categorize the error sources as algorithmic errors, dictionary problems, and conceptual problems.39
Application of information technology

DISCUSSION

Remaining challenges
Although fast, our permutational dictionary look-up NER approach requires maintaining a lexically variant-rich dictionary and fails at recognizing complex levels of synonymy. In our paper, Li et al.,\(^{60}\) we investigated an alternative ML approach through conditional random fields (CRFs)\(^{61,62}\) and support vector machines\(^{63}\) that showed that CRFs with multiple features outperform a single feature of dictionary look-up (of note, the ML NER module is not part of the current cTAKES release). Alternatively, a linguistic approach could be considered in which variants of the NP candidate are generated, mapped to a dictionary, and results ranked to produce the final mapping.\(^{16}\) These approaches are a trade-off between rich dictionary inclusion and computational cost. Although we report positive specific agreement values (identical to F-scores\(^{51}\)) of 0.598 and 0.457 for exact and overlapping spans with the default MetaMap parameters in our paper,\(^{39}\) these results should not be considered as directly comparable to our cTAKES NER component as we did not attempt to optimize MetaMap parameters. We plan to conduct a detailed comparative study of a variety of approaches to eventually combine their strengths.

One of the most frequent error sources in the NER component is the selection of one unique meaning to an named entity, which potentially maps to several concepts.\(^{64}\) We investigated supervised ML Word Sense Disambiguation and concluded that it is a combination of features unique for each ambiguity that generates the best results\(^{65}\) suggesting potential scalability issues. The cTAKES currently does not resolve ambiguities. Another cTAKES limitation lies in coordination structure interpretations—for example, the phrase ‘bladder and bowel habits’ should be parsed into two concepts (‘bladder habits’ and ‘bowel habits’), which cTAKES currently processes incorrectly as ‘bladder and bowel habits’ and ‘bowel habits’.

The cTAKES named entity attributes are similar to MedLEE’s. Unlike MedLEE, the current cTAKES release does not include a module for asserting the relation between a disease/disorder, sign/symptom or procedure, and an anatomical site (MedLEE’s BodyLoc modifier). We view this as a post-NER relation assertion task, which we will address in the future in the broader context of UMLS relation discovery from the clinical narrative. Another future challenge is expanding the values of the status named entity attribute with levels of granularity to express uncertainty and relatedness to patient to reflect non-patient experiences other than ‘family history of’.

Global evaluation and applications
We have conducted global evaluations of cTAKES for two large-scale phenotype-extraction studies: (1) ascertaining cardiovascular risk factors for a case-control study of peripheral arterial disease using the EMR within the eMERGE\(^{66}\) and (2) treatment characteristics from pathology reports into a cancer disease knowledge model.\(^{9,8}\) We are actively working on cTAKES modules for coreference,\(^{74}\) temporal relation discovery,\(^{75}\) and certainty assertion, to be implemented as downstream cTAKES components following NER. These components are expected to contribute to a refined set of named entity attributes. The cTAKES is being integrated within the Ontologies Development and Information Extraction tool.\(^{76}\) We recently ported the HITEx tokenizer into cTAKES, which highlights cTAKES’s modularity. In collaboration with University of Colorado investigators, we are extending cTAKES for semantic processing. The cTAKES will be further enhanced for data normalization as part of a Strategic Health IT Advanced Research Project (SHARP) focusing on secondary use of the EMR. By making cTAKES available open-source, we present the community with the opportunity to collaboratively develop the next generation clinical NLP systems for large-scale intelligent information extraction from the clinical narrative.

Acknowledgments
We would like to acknowledge our annotators, Barbara Abbott, Debra Albrecht, Pauline Funk, and Donna Ihrie, for their excellent work; Dingcheng Li’s work towards the linguistic annotations; Serguei Pahkhomov for conceptualization; Patrick Duffy for brilliant programming; and Jim Buntrock for being an integral part of this project. The work was partially supported by a 2007 IBM UIMA grant. We would like to thank our alpha testers, Dr Leonard D’Avelio, VA Boston, and Dr Cheryl Clark, MITRE Corporation, who endured the pains of the initial code. Our appreciation goes to the three anonymous JAMIA reviewers whose detailed comments greatly improved the manuscript.

Funding
The work was partially supported by a 2007 IBM UIMA grant.

Provenance and peer review
Not commissioned; externally peer reviewed.

REFERENCES
Application of information technology


56. eMERGE. https://www.mc.vanderbilt.edu/victc/dcc/projects/acc/index.php/

57. Page Main Page.


67. ODIE. Ontology development and information extraction (ODIE) toolset. http://www.bioontology.org/ODIE.