A Novel System for Extractive Clinical Note Summarization using EHR Data

Jennifer Liang and Ching-Huei Tsou
IBM Research, Yorktown Heights, NY 10598
{jjliang, ctsou}@us.ibm.com

Abstract

While much data within a patient's electronic health record (EHR) is coded, crucial information concerning the patient's care and management remain buried in unstructured clinical notes, making it difficult and time-consuming for physicians to review during their usual clinical workflow. In this paper, we present our clinical note processing pipeline, which extends beyond basic medical natural language processing (NLP) with concept recognition and relation detection to also include components specific to EHR data, such as structured data associated with the encounter, sentence-level clinical aspects, and structures of the clinical notes. We report on the use of this pipeline in a disease-specific extractive text summarization task on clinical notes, focusing primarily on progress notes by physicians and nurse practitioners. We show how the addition of EHR-specific components to the pipeline resulted in an improvement in our overall system performance and discuss the potential impact of EHR-specific components on other higher-level clinical NLP tasks.

1 Introduction

EHRs are a longitudinal record of the patient's health information consisting of structured (e.g. vitals, medications, labs, procedures) and unstructured (e.g. progress notes, discharge summaries, diagnostic test reports) information. Clinical notes within EHRs are traditionally a rich source of data where detailed information about the patient's medical history and clinical care process is documented. However, physicians at the point of care are mostly unable to review much of this unstructured information due to the abundance of notes within a patient EHR and the time constraint inherent in the clinical setting. Also, the move from paper records to EHRs have unintentionally resulted in issues of note bloat, where use of templates and copy-paste have introduced unnecessary or redundant data into clinical notes, worsening the problem of information overload and making it more difficult for physicians to identify key clinical data with potentially negative consequences (Shoolin et al., 2013; Vogel, 2013).

In the clinical care process, what is considered key clinical data within a clinical document depends greatly on the user and their task; what is important for a physician to know while diagnosing a patient is different from what is important for a social worker to know when arranging post-discharge home care. Building off the idea of a problem-oriented medical record introduced by Dr. Lawrence Weed (1968), we decided to approach this problem of information overload within EHRs from a disease-specific perspective. We propose an automated summarization system that produces an extractive summary for each note containing only the most important information relevant for managing a patient's hypertension or diabetes mellitus at the point of care.

There are multiple challenges in generating a disease-specific extractive summary on clinical text. First of all, the abundance of domain-specific terminology and presence of non-standard abbreviations and misspellings make machine comprehension of clinical text a much more complex task (Demner-Fushman et al., 2009). Secondly, the use of temporal narratives with reference to multiple diseases and the inherent interrelatedness of different diseases and other clinical concepts makes it difficult to determine what is "disease-specific" in the context of our summary. Moreover, the heavy use of templates, copy-paste, and imported data within clinical notes (Shoolin et al., 2013; Vogel, 2013) suggests that medical NLP at the concept level is insufficient for differentiation between "important" and "unimportant" information. Last of all, due to the regulations surround-

Proceedings of the 2nd Clinical Natural Language Processing Workshop, pages 46–54
Minneapolis, Minnesota, June 7, 2019.

ing use and sharing of protected health information, and the need for expert annotation, clinical NLP systems typically only have access to a limited amount of labeled data. To address these concerns, we leverage individual components, trained on separate labeled datasets, that target lower level clinical NLP tasks such as identifying note structure and specific clinical events of interest, and chain these individual components together into a pipeline that automatically generates disease-specific summaries from clinical notes.

2 Related Work

EHR summarization efforts have mostly focused on extraction of clinical variables or visualization of structured and unstructured elements in the EHR as a longitudinal data display (Pivovarov and Elhadad, 2015), with the objective being to present an overview of the entire longitudinal patient record. Savova et al. (2010) built cTAKES, an open-source NLP system for information extraction from unstructured clinical text. Rogers et al. (2006) developed the CLEF chronicle, which uses a semantic network of concepts and interrelations to represent events in a patient's medical history, that could serve as a building block for future summarization efforts. CLAMP (Soysal et al., 2017) allows users to more efficiently build customized NLP pipelines and reported good performance on named entity recognition and concept encoding in their evaluation.

Some researchers approached EHR note summarization from a problem identification perspective. Cao et al. (2004) summarized discharge summaries as problem lists. Van Vleck and Elhadad (2010) identified a list of problems relevant to a physician seeing a new patient for a given set of clinical notes. Our work differs from previous published research in that our summarization system is (1) targeted toward a single clinical encounter represented by a note, (2) specific to the management of a given disease: hypertension and diabetes mellitus, (3) generates a human readable textual summary as opposed to a list of clinical variables, and (4) extends beyond basic medical NLP with concept recognition and relation detection to also include components specific to EHR data.

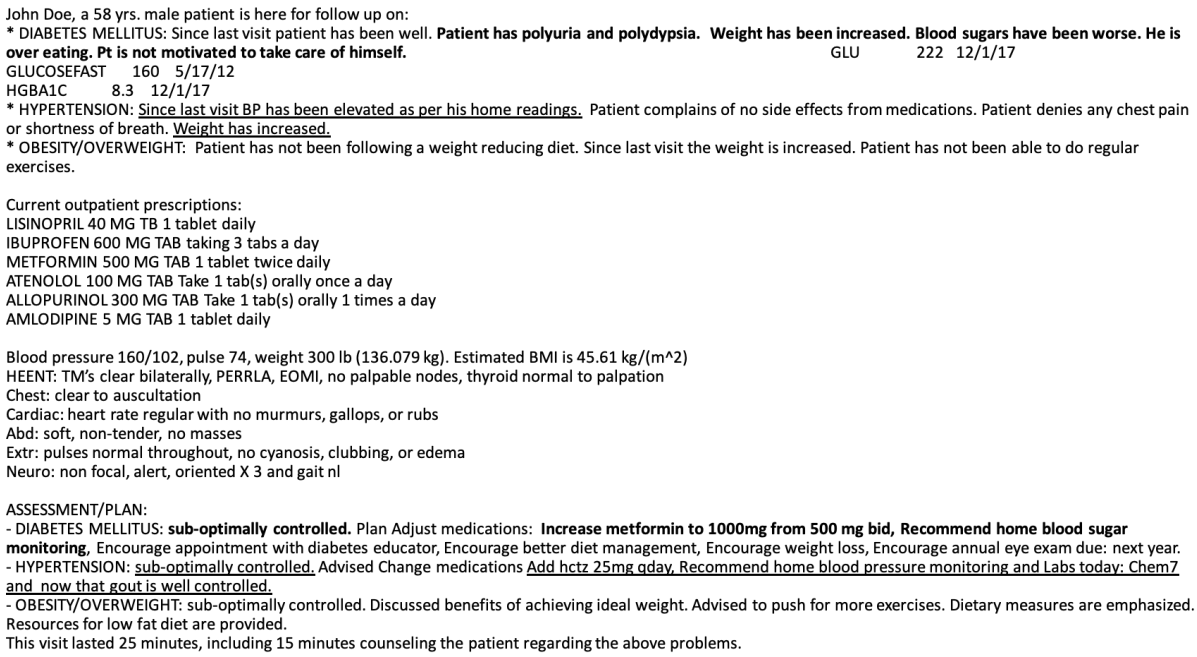
3 Method

The ultimate goal of the system is to generate a cohesive summary of a patient, similar to a summary written by an attending physician after reviewing the patient's chart. Such a system requires text summary from individual notes, reconciliation between structured data and unstructured narratives, temporal alignment of the clinical events, and natural language generation to produce the final abstractive summary. This paper focuses entirely on extracting informative sentences rather than cohesive sentences from a single clinical note. The output of the work presented here can be used as the input for downstream components to generate cohesive summaries across the longitudinal patient record.

3.1 Dataset

Our dataset consists of patient EHRs within a large ambulatory multi-specialty medical group in the US that contain a known diagnosis of hypertension and/or diabetes mellitus based on their structured encounter diagnosis list. We selected notes within these patient EHRs authored by physicians or nurse practitioners and manually reviewed approximately half of the selected notes to ensure that at least one of our diseases of interest, hypertension or diabetes mellitus, was addressed at the visit documented in that note. We made the decision to focus on physician and nurse practitioner notes because those providers are the primary decision-maker in the patients' clinical care management. Manual review was performed on approximately half of the notes to ensure a sufficient number of positive examples from the ground truth generation effort. The resulting corpus consisted of 3,453 outpatient clinical notes over 762 patients, with an average length of 138 sentences per note.

The corpus was annotated by 12 internal medicine or family medicine physicians over the course of 6 months. Physicians were asked to review each note and annotate information relevant to the physicians' decision-making for management of the patient's hypertension or diabetes mellitus, with the understanding that the annotated information would be presented together as a disease-focused summary of the note. Examples of relevant information included in the summary are statements about the current problem status, any signs or symptoms experienced by the patient, desirable and undesirable effects of current treat-


John Doe, a 58 yrs. male patient is here for follow up on:
* DIABETES MELLITUS: Since last visit patient has been well. **Patient has polyuria and polydypsia. Weight has been increased. Blood sugars have been worse. He is over eating. Pt is not motivated to take care of himself.** GLU 222 12/1/17
GLUCOSEFAST 160 5/17/12
HGBA1C 8.3 12/1/17
* HYPERTENSION: Since last visit BP has been elevated as per his home readings. Patient complains of no side effects from medications. Patient denies any chest pain or shortness of breath. Weight has increased.
* OBESITY/OVERWEIGHT: Patient has not been following a weight reducing diet. Since last visit the weight is increased. Patient has not been able to do regular exercises.

Current outpatient prescriptions:
LISINOPRIL 40 MG TB 1 tablet daily
IBUPROFEN 600 MG TAB taking 3 tabs a day
METFORMIN 500 MG TAB 1 tablet twice daily
ATENOLOL 100 MG TAB Take 1 tab(s) orally once a day
ALLOPURINOL 300 MG TAB Take 1 tab(s) orally 1 times a day
AMLODIPINE 5 MG TAB 1 tablet daily

Blood pressure 160/102, pulse 74, weight 300 lb (136.079 kg). Estimated BMI is 45.61 kg/(m^2)
HEENT: TM's clear bilaterally, PERRLA, EOMI, no palpable nodes, thyroid normal to palpation
Chest: clear to auscultation
Cardiac: heart rate regular with no murmurs, gallops, or rubs
Abd: soft, non-tender, no masses
Extr: pulses normal throughout, no cyanosis, clubbing, or edema
Neuro: non focal, alert, oriented X 3 and gait nl

ASSESSMENT/PLAN:
- DIABETES MELLITUS: **sub-optimally controlled.** Plan Adjust medications: **Increase metformin to 1000mg from 500 mg bid, Recommend home blood sugar monitoring**, Encourage appointment with diabetes educator, Encourage better diet management, Encourage weight loss, Encourage annual eye exam due: next year.
- HYPERTENSION: sub-optimally controlled. Advised Change medications Add hctz 25mg qday, Recommend home blood pressure monitoring and Labs today: Chem7 and now that gout is well controlled.
- OBESITY/OVERWEIGHT: sub-optimally controlled. Discussed benefits of achieving ideal weight. Advised to push for more exercises. Dietary measures are emphasized. Resources for low fat diet are provided.
This visit lasted 25 minutes, including 15 minutes counseling the patient regarding the above problems.

Figure 1: Sample clinical note with extractive summaries for hypertension and diabetes mellitus. Underlined sentences together form the extractive summary for hypertension; sentences in bold together form the extractive summary for diabetes mellitus.

ment, and any changes to current treatment plan. Each note was independently reviewed and annotated by two physicians, and then adjudicated by a third MD. The inter-annotator agreement is reported in the "Results" section. Figure 1 shows an example of a clinical note and its extractive summaries for hypertension and diabetes mellitus.

3.2 Extractive Summarization

Current approaches of document summarization in the clinical domain have largely been extractive rather than abstractive, so the original text as written by the physicians are preserved.

Given a clinical note consisting of a sequence of sentences, $N = \{s_1, s_2, ...s_n\}$ a sentence level, single document, extractive summarization task can be defined as to create a summary NS by selecting m sentences ($m \leq n$) from N. One of the simplest approaches is to model the task as a supervised binary classification problem, where we find a model with model parameters θ that maximize the likelihood

$$p(Y|N,\theta) = \prod_{(i=1)}^{n} p(y_i|s_i,\theta) \qquad (1)$$

where $Y = \{y_1, y_2, ...y_n\}$ and $y_i \in \{0, 1\}$.

It is obvious that in (1) each sentence is classified solely on the information contained in the sentence itself. This rudimentary approach works reasonably well for many sentence classification tasks, provided the majority of the sentences are self-contained. For clinical notes, sentences are often short and the meaning depend heavily on the context. One way to address this is to model the problem as a sequence labeling task, where additional information at the document level is also considered,

$$p(Y|N,\theta) = \prod_{(i=1)}^{n} p(y_i|s_i,N,\theta) \qquad (2)$$

A simple example of sequence model is a linear-chain CRF (Lafferty et al., 2001), which takes the previously labeled sentence(s) into consideration when predicting the label of the current sentence. Recent advances in deep neural network based approaches have shown great promises in analyzing several types of EHR data (Shickel et al., 2018), but their application in extractive note summarization is largely unexplored (Alsentzer and Kim, 2018).

In this paper, we start by creating the baseline using the 3 approaches discussed above, i.e., a linear SVM for sentence classification, a linear chain CRF in which each note is modeled as a sequence, and a simple CNN-rand (Kim, 2014) for our summarization task. All 3 models used only the note

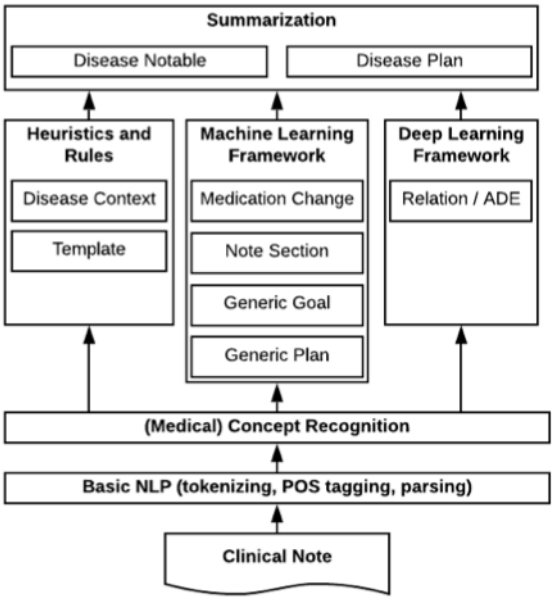


Figure 2: Single clinical document extractive summarization pipeline.

text as features: bag of n-grams for the SVM and the CRF, and randomly initialized word embeddings (updated during training) for CNN-rand. We report the F-scores for each of these models in the "Results" section.

3.3 Clinical Note Processing Pipeline

Each clinical note is ingested by a basic NLP processing layer that performs the standard NLP tasks including tokenization, lemmatization, sentence segmentation, POS tagging, and parsing, followed by a medical concept recognition component, where key medical concepts such as labs, procedures, medications, signs and symptoms, and diseases are identified. Our system used an English Slot Grammar (ESG) parser (McCord, 1990) followed by a proprietary medical concept annotator that maps terms into unified medical language system (UMLS) concepts (Bodenreider, 2004). This entity linking pipeline is similar to MetaMap (Aronson and Lang, 2010) but was optimized to process clinical notes, where sentences are not always well structured and abbreviation expansion and disambiguation plays an important role. These foundational NLP analytics, although crucial to the success of downstream components and remain an active research area, have become commodity in the recent years, for instance, CLAMP (Soysal et al., 2017), Amazon Comprehend Medical, (Amazon, 2019), and IBM Natural Language Understanding (IBM, 2019) and is not of interest in this paper.

We separate the components in the next layer into 3 categories based on how the analytics are developed, namely, heuristics and rules, assertions framework, and deep-learning framework (Figure 2). When labeled data is easier to obtain, data-driven approaches, such as deep neural network architectures, often outperform other methods due to their ability to effectively learn representations as well as model parameters. For instance, for adverse drug events (ADEs), we use an existing labeled dataset from the MADE1.0[1] NLP challenge to train a BiLSTM-CRF model with attention (Dandala et al., 2018). On the other hand, ground truth for extractive summarization requires human experts to read the entire note and is harder to acquire. With limited ground truth, we have learned that a hybrid system combining heuristics, less expressive models (e.g. linear SVM), and outputs from deep-learning based components as features generates better results than trying to train an end-to-end pure neural network based system. In the "Results" section, we will give one example from each category that has significant contribution to the overall system.

3.4 Evaluation

Evaluating a text summary is challenging. Generally, the ways of evaluating the performance of automatically generated summarizations can be categorized into intrinsic and extrinsic evaluation methods (Steinberger and Jezek, 2009). Intrinsic evaluation directly compares the generated summary to the ground truth summary. For example, co-selection measures calculate the precision, recall, and F-score at the sentence level; and content-based measures such as ROUGE (Lin, 2004) compares at word level using n-gram and/or longest common subsequence. Intrinsic evaluation can also be done qualitatively, by domain experts using a Likert-type scale. Extrinsic evaluation measures the quality of the automatic summaries indirectly, for a given task. For example, how much time a physician can save in their daily practice with or without the help of such summarization. In this work we present our results using intrinsic evaluation with co-selection measures. Studies using qualitative intrinsic measurements and quantitative extrinsic evaluation are being planned

[1] bio-nlp.org/index.php/announcements/39-nlp-challenges/

	Hypertension	Diabetes
Precision	0.723	0.726
Recall	0.646	0.671
F-score	0.682	0.697

Table 1: Inter-annotator agreement.

	Hypertension	Diabetes
SVM, linear	0.524	0.516
CRF, linear-chain	0.579	0.598
CNN	0.584	0.593

Table 2: F-scores for hypertension and diabetes mellitus summarization using different models.

for the future; those results will be reported at a later time and are beyond the scope of this paper. In this work, we evaluated the system using the co-selection measures, i.e., calculating sentence level F-score between system-identified span and a physician-annotated span, using 10-fold cross validation.

4 Results

4.1 Inter-Annotator Agreement

Agreement was calculated at the sentence level, meaning that if two annotators each marked a different span of text within the same sentence, it was considered a match. On average, annotators marked 4 to 5 sentences per note to be included in the disease-specific summary. Because the number of sentences not included in a summary is significantly larger than the number of sentences marked by physician annotators, we use recall, precision, and F-score as surrogates for the typical Cohen's Kappa in reporting inter-annotator agreement. Using one annotator's annotations as reference, we calculate the recall, precision, and F-score of the second annotator as a measure of the inter-annotator agreement, as shown in Table 1. In the "Discussion' section, we will discuss reasons for the observed differences between annotators.

4.2 Summarization Models

Table 2 shows the results of 3 different approaches for our summarization task: classification with SVM, sequence labelling with CRF, and a simple CNN with randomly initialize word embeddings. Note that in our training corpus, notes on average have 138 sentences and only 3% of the sentences are annotated as summary. Although the F-scores are at the fifties, the accuracies (including true-negatives) are well above high nineties.

	Hypertension	Diabetes
Unigram	0.555	0.581
+N-gram	0.579	0.598
+Concept	0.590	0.608
+Section	0.630	0.637
+Context	0.642	0.655
+Plan	0.646	0.662
All	0.657	0.679

Table 3: Ablation study - F-scores of disease-specific insights.

We can see that considering the document level information (CRF & CNN) is important for the task, and CNN is performing reasonably well even with limited labeled data. In this paper, our goal is to identify useful information from the entire EHR, in the context of producing extractive clinical narrative summarization. Those document level and patient level features can be used in both neural network based and non-neural network based architectures. Our current system uses CRF at the top level to label the sentences, and several deep learning based models at the component level.

4.3 Impact of EHR Components

Table 3 shows the results of an ablation study on selected components. Here we only report the numbers using the CRF model. We started by using only bag-of-words (unigram) as features and introduce a new type of feature in the next row. The last row shows the final result of using all features, including all components shown in Figure 2, such as ADEs, goals, and medication changes.

4.4 Heuristics and Rules - Disease Context

As our summarization is disease-specific, it is intuitive that knowing the disease context of each sentence can be a useful feature. For example, in the sample clinical note in Figure 1, the phrase "sub-optimally controlled" appears 3 times, each under a different disease context: "*DIABETES MELLITUS*", "*HYPERTENSION*", and "*OBESITY/OVERWEIGHT*". Depending on the disease context, each specific instance of "sub-optimally controlled" may or may not be an insight we want to extract for a given target disease.

In practice, this disease context can be an ex-

plicit section header (as in Figure 1), or conveyed more implicitly, such as a disease mentioned in the previous sentence or an encounter diagnosis code in the structured data associated with that note. A common example of implicit disease context is in specialist notes; for example, an Endocrinology note consulting on a patient's newly diagnosed diabetes mellitus. Here, the specialty in the note metadata ("*Endocrinology*") and the reason for consult ("*newly-diagnosed DM2*") both serve as the context for the entire note.

Because the disease context can be far away from the current sentence, especially for cases of implicit context, heuristics are used until we have enough labeled data to train a model with long term memory, such as recurrent neural network with attention mechanism. These heuristics were developed with input from subject matter experts familiar with how to read and interpret clinical text. We can see from Table 3 that modeling disease context explicitly improved the overall performance.

4.5 Machine Learning Framework - Note Section

Although not required, healthcare providers often follow some common structures, for example, SOAP (Lew and Ghassemzadeh, 2018), when documenting a patient encounter in a clinical note. Knowing where a sentence resides with respect to these structures or sections, will undoubtfully help the system extract important insights from the note.

As there are no set rules for indicating sections, and headers and formats are not strictly enforced, pattern matching rules using regular expression yield mediocre results. To improve the accuracy, we model note section classification as a supervised sequence labeling task using linear-chain CRF - each sentence belongs to 1 of 14 predefined note sections, namely, chief complaint, history of present illness, past medical history, past surgical history, medications, allergies, social history, family history, review of systems, vital signs, physical exam, diagnostic test results, assessment and plan, and other, and both the results from regular expression (matching predefined format and headers) as well as the words in the sentence are used as features in the CRF model. Table 3 shows that note section is another useful global feature in extractive summarization.

5 Discussion

5.1 Reasons for Annotator Differences

Although we placed our extractive summarization task in the setting of a specific user (physician), disease (hypertension and diabetes mellitus), and task (disease management), there is still a subjective component in the ground truth generation process. Some practitioners prefer a very concise summary limited to only the disease of focus, while others prefer a more informative summary that includes not only information directly related to the disease of focus, but also to related co-morbidities. Also, redundancy in clinical notes means that the same information is often presented in different ways in different parts of the same note; for example, the patient's presentation is described in detail in the history of present illness (HPI) section at the top of the note, while the same information is summarized in a more concise way in the assessment and plan section (AP) at the end of the note. Some practitioners prefer the additional detail contained in the HPI section as part of their summary, while others prefer only seeing the more concise version in the AP section.

During the ground truth generation process, we aligned physicians' perspectives and preferences the best we could through multiple discussions of what types of information should (e.g. problem status, home monitoring results, changes in disease management) and should not (e.g. direct imports from structured data, routine labs and follow-up instructions) be included as part of the summary. The discussions helped ensure better consistency of the ground truth among different physician annotators, but still resulted in an observed IAA of 0.682 to 0.697. This reflects the inherent subjective nature of the task, and demonstrates the need for a third MD to adjudicate any disagreements between annotators to produce consistent ground truth to be used by our system.

5.2 Addressing Issues of Data Scarcity

One of the challenges common in developing analytics on clinical text is the limited labeled data available for training and testing due to health information privacy concerns and the expensive cost of expert annotations. Our automatic summarization system works around this limitation by using individual components that can be trained using separate ground truth. Some components, such as note section classification, do not require annota-

tors with the same level of domain expertise and can more easily be done in-house by appropriately trained non-physician annotators. Other components, such as adverse drug events, make use of existing labeled datasets available through various clinical NLP challenges such as MADE1.0[2] and TAC[3]. Using separately trained components allows our system to make the most of the limited amount of available expert-annotated data.

5.3 Limitations in Evaluation

In this study, we use an intrinsic evaluation of our generated summary using precision, recall, and F-score to compare against ground truth created by physicians. However, these metrics do not fully capture the nuances of what should or should not be included in a clinical summary. The wide spectrum of what could be considered "important" to a physician means that not all false negatives are equivalent; some information is critical to patient management and should never be missed, while the importance and relevance of some other pieces of information are debatable amongst different physicians. Similarly, not all false positives are equivalent; some false positives are completely wrong and unrelated to the disease at hand, while others comprise of sentences that were not included in the ground truth but still provide relevant and useful information.

Adverse drug events are an example of important information that physicians are particularly sensitive to. ADEs have great impact on patient safety and is considered an important insight to extract per our annotation guidelines. These are rare yet important events for physicians to be aware of when managing a patient's care. We have been actively participating in recent ADE detection related challenges and developed our component using BiLSTM-CRF (Dandala et al., 2018). A major task for this component is to distinguish adverse drug events (e.g. "*His cough improved off lisinopril*") from indication for a drug (e.g. "*His hypertension improved on lisinopril*"), i.e., to identify the type of relation between a drug and a sign or symptom. This is often impossible without medical knowledge, and is an example of why a generic summarization algorithm from other domains will not work on clinical narratives out of the box, as the importance of a sentence depends on medical knowledge from outside of the document.

Because of the importance of ADEs to clinical care, a missed ADE by the system (false negative) or an incorrectly identified ADE (false positive) both negatively impact the overall quality of the summary significantly more compared to other types of information. As ADEs are rare, adding this component does not have significant impact to the overall summarization accuracy measure. However, we choose to discuss this component in this paper to demonstrate the need for a qualitative intrinsic evaluation that weighs each sentence based on its importance in order to capture the value that rare yet important events, such as ADEs, bring to the overall system.

Redundant information in clinical notes pose another challenge to the evaluation of our system. This redundancy led to observed cases where the ground truth has one sentence annotated while the system has annotated a different sentence containing essentially the same content. This is judged as a false negative (because the system missed the physician-annotated sentence) and a false positive (because the system-annotated sentence was not in the physician-annotated ground truth), reflecting negatively in the overall system evaluation without giving the system credit for the fact that the relevant information is still present in the generated summary. This demonstrates the need for a separate extrinsic evaluation of our generated summaries based on its usefulness in the clinical setting, which we have planned for the future.

We are planning future evaluations of our system using qualitative intrinsic measures to capture the importance of different information within the clinical summary, and quantitative extrinsic measures to evaluate the usefulness of the system-generated summaries for practicing physicians at the point of care.

6 Conclusions

We propose an automated system for disease-specific extractive summarization on a single clinical note. We describe our clinical note processing pipeline that includes a basic NLP processing layer as well as additional EHR-specific components such as note section classification, disease context identification, and adverse drug events detection. We show incremental improvement in overall system performance with addition of each component, from F-scores of 0.555 and 0.581 for

[2] bio-nlp.org/index.php/announcements/39-nlp-challenges/

[3] bionlp.nlm.nih.gov/tac2017adversereactions/

hypertension and diabetes mellitus, respectively, when using only unigrams, to 0.657 and 0.679 when all components are included in the pipeline. Our work demonstrates how analytics beyond concept recognition is necessary for a complex and higher-level clinical NLP task such as summarization. Also, until abundant labeled data in clinical narratives becomes available, generic summarization algorithms developed using non-EHR data will benefit from using EHR-specific components discussed here as global features.

References

Emily Alsentzer and Anne Kim. 2018. Extractive Summarization of EHR Discharge Notes. *arXiv:1810.12085*.

Amazon. 2019. Amazon Comprehend Medical.

Alan R Aronson and François-Michel Lang. 2010. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32:D267–D270.

Hui Cao, Michael F. Chiang, James J. Cimino, Carol Friedman, and George Hripcsak. 2004. Automatic Summarization of Patient Discharge Summaries to Create Problem Lists using Medical Language Processing. *Stud Health Technol Inform*, 107(2):1540.

Bharath Dandala, Venkata Joopudi, and Murthy Devarakonda. 2018. IBM Research System at MADE 2018: Detecting Adverse Drug Events from Electronic Health Records. In *Proceedings of the 1st International Workshop on Medication and Adverse Drug Event Detection*, volume 90 of *Proceedings of Machine Learning Research*, pages 39–47. PMLR.

Dina Demner-Fushman, Wendy W. Chapman, and Clement J. McDonald. 2009. What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics*, 42(5):760–772. Biomedical Natural Language Processing.

IBM. 2019. IBM Natural Language Understanding.

Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Valerie Lew and Sassan Ghassemzadeh. 2018. *SOAP Notes. [Updated 2019 Jan 19]. In: StatPearls [Internet]*. Treasure Island (FL): StatPearls Publishing.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Michael C. McCord. 1990. Slot grammar. *Natural language and logic*, pages 118–145.

Rimma Pivovarov and Noémie Elhadad. 2015. Automated methods for the summarization of electronic health records. *Journal of the American Medical Informatics Association*, 22(5):938–947.

Jeremy Rogers, Colin Puleston, and Alan Rector. 2006. The CLEF Chronicle: Patient Histories Derived from Electronic Health Records. In *22nd International Conference on Data Engineering Workshops (ICDEW'06)*, pages x109–x109.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

Benjamin Shickel, Patrick Tighe, Azra Bihorac, and Parisa Rashidi. 2018. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. *IEEE Journal of Biomedical and Health Informatics*, 22(5):1589–1604.

J. Shoolin, L. Ozeran, C. Hamann, and W. Bria. 2013. Association of Medical Directors of Information Systems Consensus on Inpatient Electronic Health Record Documentation. *Appl Clin Inform*, 4(2):293–303.

Ergin Soysal, Jingqi Wang, Min Jiang, Yonghui Wu, Hua Xu, Serguei Pakhomov, and Hongfang Liu. 2017. CLAMP a toolkit for efficiently building customized clinical natural language processing pipelines. *Journal of the American Medical Informatics Association*, 25(3):331–336.

Josef Steinberger and Karel Jezek. 2009. Evaluation Measures for Text Summarization. *Computing and Informatics*, 28:1001–1026.

Tielman T. Van Vleck and Noémie Elhadad. 2010. Corpus-Based Problem Selection for EHR Note Summarization. *AMIA Annu Symp Proc*, pages 817–821.

Lauren Vogel. 2013. Cut-and-paste clinical notes confuse care, say US internists. *CMAJ*, 185(18):E826–E826.

Lawrence L. Weed. 1968. Medical records that guide and teach. *New England Journal of Medicine*, 278:652657.