Theses and Dissertations

May 2014

Adverse Drug Event Detection, Causality Inference, Patient Communication and Translational Research

Balaji Polepalli Ramesh
University of Wisconsin-Milwaukee

ADVERSE DRUG EVENT DETECTION, CAUSALITY INFERENCE, PATIENT

COMMUNICATION AND TRANSLATIONAL RESEARCH

by

Balaji Polepalli Ramesh

A Dissertation Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

in Biomedical and Health Informatics

at

The University of Wisconsin-Milwaukee

May 2014

ABSTRACT

ADVERSE DRUG EVENT DETECTION, CAUSALITY INFERENCE, PATIENT

COMMUNICATION AND TRANSLATIONAL RESEARCH

by

Balaji Polepalli Ramesh

The University of Wisconsin-Milwaukee, 2014
Under the supervision of Professor Susan McRoy and Professor Hong Yu

Adverse drug events (ADEs) are injuries resulting from a medical intervention related to a drug. ADEs are responsible for nearly 20% of all the adverse events that occur in hospitalized patients. ADEs have shown to increase the cost of health care and the length of stay in hospital. Therefore, detecting and preventing ADEs for pharmacovigilance is an important task that can improve the quality of health care and reduce the cost in a hospital setting. In this dissertation, we focus on the development of ADEtector, a system that identifies ADEs and medication information from electronic medical records and the FDA Adverse Event Reporting System reports. The ADEtector system employs novel natural language processing approaches for ADE detection and provides a user interface to display ADE information. The ADEtector employs machine learning techniques to automatically processes the narrative text and identify the adverse event (AE) and medication entities that appear in that narrative text. The system analyzes the entities recognized to infer the causal relation that exists between AEs and medications by automating the elements of Naranjo causality assessment scale using knowledge and rule based approaches. The Naranjo causality assessment scale is a validated tool for finding

the causality of a drug induced adverse event or ADE. The scale calculates the likelihood of an adverse event related to drugs based on a list of weighted questions. The ADEtector also presents the user with evidence for ADEs detected by extracting figures that contain ADE related information from biomedical literature. A brief summary is generated for each of the figures that are extracted to help users better comprehend the figure. This will further enhance the user experience in understanding the ADE information better. The ADEtector also helps patients better understand the narrative text by recognizing complex medical jargon and abbreviations that appear in the text by providing definitions and explanations for them from external knowledge resources. This system could help clinicians and researchers in discovering novel ADEs and also hypothesize new research questions within the ADE domain.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

As with any endeavor in life, completing a Ph.D. dissertation takes perseverance to overcome the intellectual and emotional challenges that come over the years. This research could not have been completed without the help of many people, all of who deserve more than just a mention here.

I cannot express enough gratitude to Prof. Hong Yu for the support and guidance she provided through every step of this project. She has played a significant role in my development as a researcher and as a person. I am inspired by her passion for her work, her dedication to the students, and the value she places on a balanced life. I have always been able to turn to her for advice, feedback, motivation, and help. Hong is a great role model, and I look forward to working with her in the future.

I would like to thank Dr. Susan McRoy. As my advisor and co chair of the program, she has been extremely helpful and an excellent source of support and ideas. I am extremely fortunate to have a very knowledgeable and helpful thesis committee and would like to thank each and every one of them for providing excellent suggestions throughout this project.

I would like to thank the faculty and staff at the University of Wisconsin-Milwaukee and University of Massachusetts Medical School for providing an environment and a support system where I could come up with ideas and work on them.

In addition, I would like to thank my friends over the years who have made my time here as a student so enjoyable and I also would like to thank them for all of the fun times which gave me a break from work and school.

I am truly blessed to have such wonderful parents and family who are always there for me holding me up in love and prayer. I particularly thank my wife, Shourya, for her amazing love, support, encouragement, and patience. Their unfailing and endless support for of my academic pursuits from the start of my education to the completion of my doctorate has always strengthened me.

Chapter 1: Introduction

An *adverse event* (AE) is an injury or untoward medical occurrence to a patient or clinical investigation subject who was administered a pharmaceutical product that does not necessarily have a causal relationship with the treatment received by the patient [1,2]. An *adverse drug event* (ADE) is an injury resulting from a medical intervention related to a drug including the harm caused by a drug (adverse drug reactions and overdoses) and harm from the use of the drug (including dose reductions and discontinuations of drug therapy) [3,4]. Studies have reported that ADEs account for nearly 20% of all adverse events that occur in hospitalized patients [5-7]. In the United States alone, ADEs account for more than 770,000 injuries and deaths annually [8-10], and an increased average length of stay in hospitals at a cost of $1.56-$5.60 billion annually [3,11]. Improved methods for ADE detection and analysis may identify novel drug safety signals and lead to improved methods for avoiding ADEs, with their inherent burden of morbidity, mortality, and cost.

Some ADEs are identified during clinical trails but these trials are often conducted on a small cohort of patients and do not represent the entire population. The majority of ADEs come to light only after the drug is widely used in the market by patients and hospitals. Electronic Medical Records (EMRs) in hospitals contain vast amounts of data regarding patient health, drug usage and unprecedented events that occurred during the hospital stay. A study by Cullen et al. [10] found that information regarding ADEs appear in the unstructured or narrative text. Studies have also shown that patient discharge summaries

incorporate ADE information [11,12]. However, currently, ADE information is mainly extracted manually [2,13]. Figure 1 below shows a sample discharge summary from the University of Pennsylvania Medical Center (UPMC). We can see that medication information and other information related to ADE are in the form of narrative text.

As part of a major effort to support post-marketing drug safety surveillance, the US Food and Drug Administration (FDA) receives mandatory reports on ADEs from manufacturers through the FDA Adverse Event Reporting System (FAERS). The FAERS is a database that captures information concerning adverse events and medication errors associated with FDA approved prescription drugs. Currently, FAERS contains over four million reports of adverse events dating from 1969 to present [12]. It serves as a rich resource for pharmacovigilance – the study of drug-related injuries for the purpose of making warning or withdrawal recommendations for pharmaceutical products [4]. A typical FAERS report incorporates both structured data and unstructured free text, as shown in Figure 2. The structured data entries incorporate each patient's personal and demographic information, a list of prescribed drugs, and the class of drug reaction (in this example, "anaphylactic reaction") (Figure 2). The event/problem narrative contains additional information relevant to describing the event, assessing causality, and grading severity (Figure 2). In the example, the narrative text contains the phrase (highlighted in yellow) that indicates the causality between paclitaxel and anaphylactic reaction and "experienced a life threatening anaphylactic reaction" shows the severity of the event, which is not coded in the structured data.

Although EMRs and FAERS reports are an excellent resource to study drug effects [13],

```
<type>DS</type>
<chief_complaint>SYNCOPE</chief_complaint>
<admit_diagnosis>780.2</admit_diagnosis>
<discharge_diagnosis>458.0,581.1,920,E888.9,182.0,294.8,250.00,V58.67,401.9,227.0,782.3,573.9,574.20
,</ discharge_diagnosis>
<year>2007</year>
<report_text>[Report de-identified (Safe-harbor compliant) by De-ID v.6.22.07.0]
```

DISCHARGE SUMMARY

DATE OF ADMISSION: **DATE[Jun 24 2007]

DATE OF DISCHARGE:
DISCHARGE DIAGNOSES
1. Frequent falls.
2. Orthostatic hypotension.
3. Proteinuria.
4. History of hypertension.
...
CONSULTANTS
1. Dr. **NAME[UUU] for gynecologic malignancy.
2. Dr. **NAME[TTT] for proteinuria and lower leg edema. 3. Dr. **NAME[SSS] Work for placement.
4. PT/OT for rehabilitation.

HOSPITAL COURSE
This patient is an **AGE[in 80s]-year-old female with multiple medical problems as mentioned above. In the
last past one month she had three hospital
admissions including this one and is the first of two that is about her
progressive lower extremity edema and extremity cellulitis ... carcinoma
the patient had Megace which had been discontinued due to the patient being
ALLERGIC TO MEGACE. We reconsulted Dr. **NAME[UUU] about this ... meetings will need to be called
according to her son.

DISPOSITION
At this point of time I discharged Mrs. **NAME[AAA] to **INSTITUTION ... scan should be performed.

DISCHARGE MEDICATIONS
1. Tylenol 650 mg by mouth every six as needed for pain or fever.
2. Vicodin 500/5 mg one tablet by mouth every four as needed for pain. ...

DISCHARGE INSTRUCTIONS
1. This patient has frequent falls and also she is very demented, and fall precaution must be performed at all
times.
...

```
</report_text>
</report>
```

Figure 1: A sample electronic medical record (EMR). The adverse events are highlighted
in red and the medications in yellow.

the structured data do not incorporate confounding factors, including concomitant

medications and patient medical histories, which limits effectiveness of these data for

pharmacovigilance. In contrast, such confounding factors are frequently described in the unstructured FAERS narratives. Making these data computationally available is critical for pharmacovigilance.

```
FDA Adverse Event Reporting System (FAERS)
ISR Report for report # ISR 4901614-2
Reporter Name:                                        Reporter Phone:
Reporter Org:                                         Reporter City:
Reporter Street:                                      Reporter State:
Reporter Zip:                                         Reporter Country:
Product(s)                      Route                    Indication(s)
TAXOL                           INTRAVENOUS              BREASTCANCER
KYTRIL                          INTRAVENOUS              PREMEDICATION
DECADRON PHOSPHATE              INTRAVENOUS              PREMEDICATION
GLUCOPHAGE
ACIPHEX
LEVAQUIN
PYRIDIUM
Reaction(s)
ANAPHYLACTIC REACTION
Disease/Surgical Procedure
FOOD ALLERGY
DRUG HYPERSENSITIVITY
DIABETES MELLITUS
BREAST CANCER
DEPRESSION

Event/ Problem Narrative
    A nurse reported that a 60- year old female consumer experienced a delayed allergic reaction
    while on paclitaxel therapy. The patient received her first dose of paclitaxel 280 mg on
    [DATE]. On [DATE], hours later, the consumer presented at the emergency room with pain in
    her hands and feet, feeling of throat closing and a rash (generalized). Epinephrine was
    administered and her symptoms resolved. She was discharged to home on the same day. The
    patient's medical history was a significant for diabetes, left breast cancer, depression
    and allergies of shell fish and sulfa. Supplemental information was received from the
    reporting nurse on [DATE]; The patient experienced a life threatening anaphylactic reaction
    after receiving one dose of paclitaxel, on [DATE] for treatment of breast cancer. The
    Anaphylactic reaction was characterized by symptoms of pain in her hands and feet, feeling
    of throat closing and rash (generalized).
```

Figure 2: A sample FAERS report. (A) structured data. (B) narrative free text

Currently, manual abstraction is required for identification of relevant data in EMR and FAERS narratives. Given the enormous number of EMR reports generated at the hospitals and FAERS reports received by the FDA, manual abstraction is impractical and expensive. Therefore, it is important to develop computational approaches to automatically extract information from these narratives. In this dissertation, our goal is to

develop natural language processing (NLP) approaches to automatically extract ADE information from the narratives of both EMRs and FAERS reports. This is an important step toward enriching the existing capacity of narrative text for pharmacovigilance.

The task is, however, very challenging. The detection of ADE involves identification of not only medication and AE entities but also inferring the causal relation that exists between them. We focus on the development of a system, ADEtector, which employs novel NLP approaches for ADE detection. After recognizing the ADE, we find evidence from the biomedical literature in the form of figures using a simple keyword search to further support the finding. This helps the users to visually see the ADE information in literature and corroborate it. The EMR and FAERS reports often contain domain-specific terms that may be hard for users to comprehend. To help users better understand the report, we build an NLP system that identifies complex medical terms that appear in the report and provide definitions and explanations to them from external knowledge resources. We also develop a new user interface integrating all these systems that can aid in viewing the ADE information and facilitate discovery of novel ADEs from the narrative text.

We hypothesize such a tool can be useful to various group of users. Patients can use this tool to better understand the content of the reports and examine if any ADEs were reported in their records. Clinicians and researchers can use it to find the presence of ADE in a report and use it as a tool to help in translational research by further investigating the biomedical literature for ADE information using the figure evidence component. The tool could also be used by regulatory agencies such as the FDA, for

pharmacovigilance to monitor ADEs in the reports and investigate them further by looking at the figure evidence and definitions for the terms that appear in the report.

We envision that ADEtector will not only be a tool for pharmacovigilance but also an application that can improve communication between physicians and patients in comprehending adverse drug events. As such we need to simplify the EHR text to make it readable by patients who have little clinical knowledge.

Figure 3 below shows the schematic of the proposed system architecture. The system consists of five components.

1. Adverse Event and Medication Named Entity Tagger

2. Causality Inference Engine

3. Figure Evidence Generator

4. NoteAid

5. ADE View

Given an AERS report or an EMR narrative, the first component, *Adverse Event and Medication Named Entity Tagger*, automatically processes the narrative and identifies the adverse event and medication-related entities that appear in that report. The second component, *Causality Inference Engine*, is an inference engine that derives a relation between the AE and the medication entities. The component analyzes the ADEs using rule- and knowledge-based approaches to derive the causality between drugs and adverse events. We automate the elements of Naranjo Causality Assessment Scale. We also develop a discourse connective identifier with sense detector that can be used to aid the automation of elements of the Naranjo score. Discourse connectives are words or phrases that connect or relate two coherent sentences or phrases and indicate the presence of

discourse relations. The discourse connective recognizer identifies the presence of explicit discourse cue that appear in the narrative text. The Naranjo Adverse Drug Reaction Probability Scale [14] is a validated tool for finding the causality of a drug-induced adverse event or ADE. The scale calculates the likelihood of an AE related to drugs based on a list of weighted questions or elements. The scale also examines factors such as the temporal association of drug administration and event occurrence, alternative causes for the event, drug levels, dose–response relationships, and previous patient experience with the medication. The response to each of these questions is assigned points, and the ADE is assigned to a probability category from the total score, as follows: *definite* if the overall score is 9 or greater, *probable* for a score of 5-8, *possible* for 1-4, and *doubtful* if the score is 0.

Table 1: The Naranjo Adverse Drug Reaction Probability Scale to assess the causality of an ADE

	Yes	No	Do not know
1. Are there previous *conclusive* reports on this reaction?	+1	0	0
2. Did the adverse event occur after the suspected drug was administered?	+2	-1	0
3. Did the adverse reaction improve when the drug was discontinued or a specific antagonist was administered?	+1	0	0
4. Did the adverse reaction reappear when the drug was readministered?	+2	-1	0
5. Are there alternative causes (other than the drug) that could have on their own caused the reaction?	-1	+2	0
6. Did the reaction reappear when a placebo was given?	-1	+1	0
7. Was the drug detected in the blood (or other fluids) in concentrations known to be toxic?	+1	0	0
8. Was the reaction more severe when the dose was increased or less severe when the dose was decreased?	+1	0	0
9. Did the patient have a similar reaction to the same or similar drugs in *any* previous exposure?	+1	0	0
10. Was the adverse event confirmed by any objective evidence?	+1	0	0

Table 1 shows the Naranjo elements that are used to assess the causality of an ADE by a drug. Previous work has shown that physicians are able to assess the causality with the Naranjo scores [15-18]. In this dissertation, we automate three elements of the Naranjo scores due to limitations of the data availability and infer the causality using the AE and medication entities recognized by rule- and knowledge-based approaches.

The third component, *Figure Evidence Generator*, extracts figures related to the ADE from biomedical literature as evidence [19] to support the ADE information extracted from the narrative text using a simple keyword-based search. It also generates a concise summary for each figure that is extracted to help in comprehending the figure better. The fourth component, *NoteAid,* processes the narrative text and identifies clinical concepts such as medical jargon, abbreviations, and complex disease and medication names. It then fetches definitions and explanations for the identified concepts from external knowledge resources to help understand the narrative text. The fifth and the final component, *ADE View* is a user interface that displays the ADE inferred along with the figure evidence, a summary for each of the figures and NoteAid simplified text to aid clinicians and researchers in discovering novel ADEs from the narrative text.

Although several studies have identified ADEs, most employ the simple approach of co-occurrence between the mention of a drug and an AE, which does not necessarily reflect the causal relation between the two entities. In this dissertation, we employ machine-learning approaches to identify AE and medication named entities and infer the causal relation that exists between AE and medication entities.
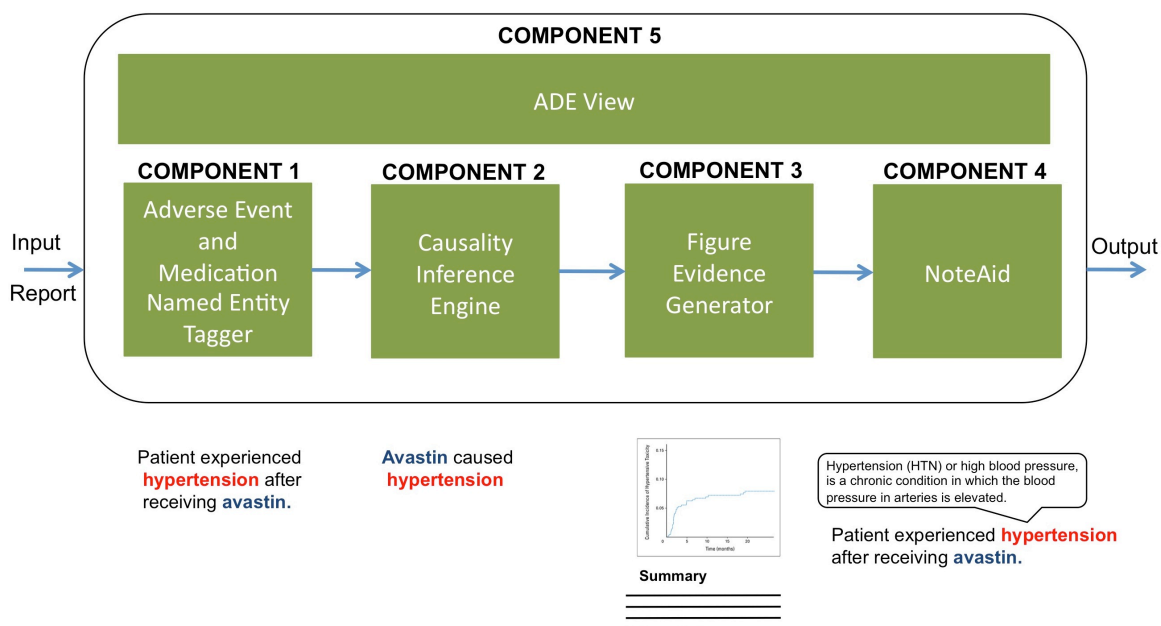
Figure 3: Schematic of the proposed architecture of the system. The first component is Medication and Adverse Event Named Entity Tagger, the second component is Causality Inference Engine, the third component is Figure Evidence Generator, the fourth component is NoteAid and the final one is ADE View, which displays the output of the all the four components

The remaining parts of the dissertation are organized as follows: Chapter 2 discusses the supervised machine-learning methods utilized for the identification of adverse events and medication-related named entities. Chapter 3 and Chapter 4 discuss the automatic recognition discourse connective and the automation of the Naranjo Adverse Drug Reaction Probability Scale elements. Chapter 5 talks about the methods used to generate summaries for the figure evidence that is extracted from biomedical literature and its evaluation. Chapter 6 describes the NoteAid system developed to help users understand the content of the narrative text better and its evaluation. The development of the user interface to display the ADE information is discussed in Chapter 7.

Chapter 2: Named Entity Recognition from FDA AERS Narratives and EMR Reports

In this chapter, we describe recognizing adverse events, medications and other named entities in narrative text, the first component of the ADEtector system. The entities recognized by the named entity recognizer are used to infer the causality between medication and adverse event. We use only adverse event and medication entities in this dissertation since we automate only three Naranjo elements. Although the other entities are not utilized, they are required to automate the other Naranjo elements. The narrative text in EMR and FAERS reports is very rich and contains a lot of information such as adverse events, indications, laboratory findings, medications, dosage, and others. Identification of the named entities is an important step toward recognizing ADE. We develop an annotation guideline and annotate medication information and adverse event-related entities on 122 FAERS reports and 150 EMR narratives comprising of ~23K and ~103K word tokens respectively. A named entity tagger using supervised machine-learning approaches is built for detecting medication information and adverse event entities using various categories of features. The annotated corpus had an agreement of over 0.9 Cohen's κ for medication and adverse event entities. The best performing tagger achieves an overall performance of over 74% for detection of medication, adverse event and other named entities.

Related Work

There is extensive research related to AE and ADE detection and analysis from a variety of data sources. Earlier work has examined patients' paper medical records to determine whether AE and ADE can be reliably abstracted based on the information conveyed in those records. For example, Hiatt et al. [20] was among one of the early studies that defined AE as an injury caused at least in part by medical mismanagement (negligence). They then manually abstracted ADE from patients' paper-based clinical medical records. Similarly, other early studies (e.g., [3,7,21]) defined AEs and ADEs and then manually abstracted them from clinical records. These studies indicate the feasibility and value of clinical records for ADE surveillance and prevention.

When electronic medical records (EMRs) became available, computational approaches were developed to automatically identify AE and ADE information from EMRs. Studies used rule-based approaches for detecting ADEs from EMR data [22-24]. Tinoco et al. [25] compared a rule-based computer surveillance system called Health Evaluation through Logical Processing (HELP) [26] with manual chart reviews of 2,137 patient admissions. They reported that HELP detected as many ADEs as were found by manual chart review, suggesting that NLP systems could improve ADE detection from EMR narrative data.

Many studies applied NLP [27] to detect AEs and then inferred a causality relationship between a drug and an AE – called an ADE – using logical rules, statistical analyses, and supervised machine-learning approaches. Hazlehurst et al. [28] developed MediClass, a knowledge-based system that deploys a set of domain-specific logical rules to medical

concepts that are automatically identified from EMR narratives (e.g., progress notes) or precoded data elements (e.g., medication orders). The system achieved a precision of 64% for detecting vaccine-related AEs [29]. A number of studies [30-32] applied the NLP system, MedLEE [33] to detect AEs from discharge summaries and hospitalization records. For example, Wang et al. [30] applied MedLEE to detect terms and mapped them to the UMLS semantic types. Subsequently, they detected medication and AEs when the terms were mapped to the UMLS concepts with the semantic types of Clinical Drug (T200) and Disease or Symptom (T047), respectively. The causality relationship between a medication and an AE was extracted from 25K discharge summaries based on a χ^2-statistical analysis of medication and AE. Evaluation of seven drugs for known ADEs led to a recall and precision of 75% and 31% respectively. Aramaki et al. [34] manually annotated 435 discharge summaries for drugs and ADEs and then applied supervised machine learning to detect these named entities. They identified the causality between drugs and AEs using pattern matching and SVM techniques. They reported a recall and precision of 0.81 and 0.87 for drug and 0.80 and 0.86 for AE detection respectively. For inferring causality they achieved recall and precision of 0.92 and 0.41 using pattern matching and 0.62 and 0.58 using SVM technique respectively.

In addition to EMRs, studies have explored other data sources for ADE information, including biomedical literature [35,36], product labels [37], social media and the Internet [38,39]. Shetty and Dalal [40] mined ADEs from the PubMed citations. They first built a document classifier to identify relevant documents that incorporate ADE relationships using MeSH terms. For example, if an article is assigned with "chemically induced" or "adverse effects," then the article is likely to incorporate an ADE. They then identified

ADE signals using disproportionality analysis in which the rate at which a particular AE of interest co-occurs with a given drug is compared to the rate an AE occurs without the drug in the collection. Their evaluation of a predefined set of 38 drugs and 55 AEs showed that their literature-based approach could uncover 54% of ADEs prior to FDA warnings.

A number of studies have explored approaches for extracting medication-related entities from clinical text [41-45]. In 2009 i2b2 organized a medication information extraction competition, in which 20 teams from around the world participated [46]. From all the teams that participated, the medication entities achieved F1 scores over 0.75, while the duration and indication entities achieved the best performance of 0.52 and 0.46 respectively.

There is also a rich store of literature for ADE detection on Spontaneous Reporting Systems (SRS) such as the FAERS reports and WHO VigiBase [47]. Studies have explored several statistical data mining and machine-learning techniques on SRS for the detection of ADE signals [13,48-73]. However, all aforementioned approaches for ADE detection from FAERS are based on its structured data. In this study, we report the development and evaluation of supervised machine-learning approaches for automatically detecting medication information and adverse events from the FAERS narratives. We speculate that such information can be a useful addition to the FAERS structured data for ADE detection.

Materials and Methods

Annotation Data and Procedure

The annotation was carried out using Knowtator (http://knowtator.sourceforge.net), a plugin for Protégé (http://protege.stanford.edu). The Knowtator interface allows users to define entities that need to be annotated and configure the relationships between them. The annotated narratives were used as both training and testing data for machine learning approaches and evaluated using cross-validation. We also report Cohen's κ, a well-known statistic used to assess the degree of Inter-Annotator Agreement (IAA) between annotators [74].

FAERS Reports

Through our collaboration at Northwestern University [75], we obtained a collection of 150 de-identified FAERS narratives; a sample is shown in Figure 2. The data collection originally came as a scanned PDF image file. We manually transcribed the PDF file into a computer-readable text file.

We randomly selected a set of 28 narratives for developing the annotation guideline. Our annotation guideline was based on the i2b2 challenges in NLP for Clinical Data Medication Extraction [46,76]. A balanced interdisciplinary team consisting of a linguist (NF), a physician (SB), two informaticians (BPR and HY) and a physician informatician (ZFL) developed the annotation guideline through an iterative process. Following the final annotation guideline, two annotators (ZFL, designated as AnnPhy, and NF, designated as AnnLing), both of whom were the primary annotators for the i2b2

medication event detection challenge [76] in which we participated, independently annotated the remaining 122 AERS narratives. A physician (SB) served as a tiebreaker and resolved annotation disagreements. This collection of 122 narratives is comprised ~23K word tokens and the average number of words per narrative is 190.2±130.3.

The annotated data were grouped into four collections each containing 122 narratives: *AnnPhy* and *AnnLing* – data annotated by annotators AnnPhy (ZFL) and AnnLing (NF), respectively; *Comb* – a joint set of annotations agreed upon by both AnnPhy and AnnLing; and *Tie* – a joint set of AnnPhy and AnnLing annotations where disagreements were resolved by the tiebreaker SB. We use these four sets of data to build robust supervised machine-learning classifiers to identify entities.

EMR reports from Pittsburgh repository

We randomly selected 150 de-identified discharge summary reports from Pittsburgh NLP repository, as shown in Figure 1. We incorporated the guideline from FAERS annotation and further refined the guideline iteratively to represent various attributes of entities. For example, for the dosage entity we included fields to represent the strength, form and type. The adverse event, indication and OSSD entities incorporated attributes, present and history to indicate their status. Following the final annotation guideline, AnnLing independently annotated all the 150 reports. AnnPhy independently annotated a subset of 25 reports in the set of 150 articles to measure the IAA. This collection of 150 EMR narratives comprised ~103K word tokens and the average number of words per narrative is 685.3 ± 346.3.

Supervised Machine Learning

Three supervised machine-learning approaches were explored for automatically identifying medication information and adverse events: Naïve Bayes (NB), Support Vector Machines (SVMs), and Conditional Random Fields (CRFs) [77]. We built NB and SVM classifiers using Weka [78] and the CRF model using ABNER toolkit [79]. NB is a simple model that assumes that all attributes of the examples are independent of each other given the context of the class. SVMs are a well-known statistical machine-learning algorithm and have shown very good performance in many classification tasks [80,81]. CRFs have shown success in named entity recognition in biomedical domain [79,82].

Learning Features

We explored a variety of features, such as syntactic features, semantic features based on the external knowledge resource (UMLS), morphological and contextual features, presence of negation, hedging and discourse connectives as a feature in addition to ABNER default features which include bag of words and orthogonal features. We describe each of these in detail below.

The syntactic features include the part-of-speech (POS), the phrasal class of each token, and the POS of the token immediately to the left of the token under consideration. The syntactic features were extracted from the constituency parse tree generated by Charniak–Johnson parser [83] trained in the biomedical domain. The parser was evaluated to have the best performance when tested on the GENIA corpus [84]. Figure 4 shows a sample constituency parse tree. In this example, the POS features DT (determiner), JJ (Adjective), NN (Noun) are the POS of tokens "A," "female," and "patient" respectively.

Further, the phrasal class for all the three tokens is NP. The left sibling POS value of "A" is NONE assuming it is the start of the sentence. The left sibling POS of "female" and "patient" tokens are DT and JJ respectively.

We applied the UMLS Metamap [85] (http://metamap.nlm.nih.gov/) to extract semantic features, which are concepts and semantic types represented in the UMLS Metathesaurus. The morphological features were obtained by considering various characteristics of the word. We took attributes of the word such as whether it was a digit, was capitalized, its alphanumeric order (if the token started with letters and was followed by numerals or vice versa), and the presence of punctuation such as commas and hyphens. These features were extracted using a simple pattern-matching technique. The first (prefix) and last (suffix) three and four characters of the token were added as affix features.

NP – noun phrase   VP – verb phrase   JJ - Adjective
S- simple declarative clause   DT – determiner   NN -Noun
SBAR – subordinate clause   VBD - Verb, past tense
IN - Preposition or subordinating conjunction
VBG - Verb, gerund or present participle
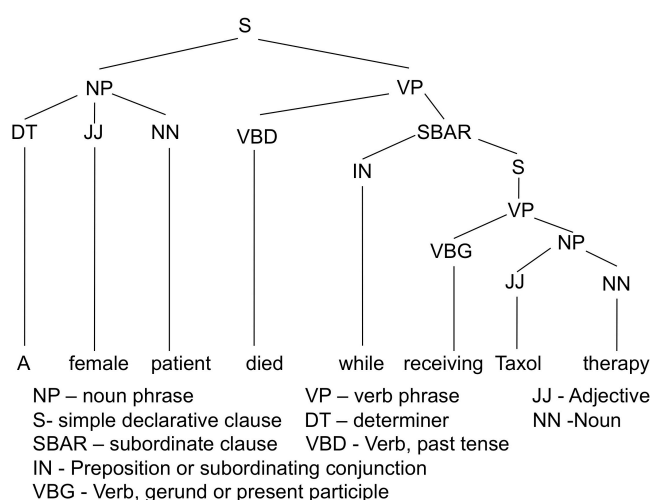
Figure 4: The sample constituency parse tree

Our manual examination of the data showed that named entities appeared around negation, hedging and discourse connectives. We further investigated our observation by conducting a small distributional study of the presence of discourse connectives around

entities and of negation and hedging scopes incorporating entities. Table 2 shows the distribution of negation and hedging scopes incorporating entities and the presence of discourse connective around named entities on 20 randomly selected FAERS reports.

Table 2: Distribution of negation and hedging scopes and discourse connectives on randomly selected 20 FAERS reports

Number of Reports	20
Number of Connectives	41
Number of Connectives around entities	26
Number of negation scopes	41
Number of negation scopes incorporating entities	20
Number of hedging scopes	29
Number of hedging scopes incorporating entities	25

Table 2 shows that 63.4% of the time connectives appeared within a two-word window of entities. The negation and the hedging scopes incorporate entities 49% and 86% of the times. In Table 3 below, example 1 shows that the adverse event "cerebral hemorrhage" appears after the connective "due to" and example 2 shows connective "while" appears in between the "hypersensitive reaction" and "Taxol." Examples 3 and 4 show instances where the named entities are incorporated within the scope of negation. Similarly examples 5 and 6 show instances where the hedging scope incorporates the named entities. Therefore, the negation and hedging cues with their scope that were detected automatically by the systems [86] and [87] and the presence of discourse connectives that were automatically detected by discourse parser [88] were added as features.

Table 3: Examples from FAERS reports with entities around discourse connectives and negation and hedging scope incorporating named entities. Connectives are shown in **bold.** The scope of negation and hedging is shown in *italics* and the cue is shown in ***bold italics***

Discourse connectives around entities	1) The patient died **due to** cerebral hemorrhage. 2) Patient experienced a hypersensitive reaction **while** on Taxol
Entities within negation scope	3) *…death **not** related to Taxol.* 4) She experienced *pleural suffusion **without** pueumothorax*
Entities within hedging scope	5) *…**possibly** related to vinorelbin and underlying disease* 6) *… death **possibly** related to paclitaxel and radiotherapy.*

Systems

We developed several taggers to evaluate (a) the complexity of the task for identifying medication information and adverse events and (b) the impact of features.

Systems to Evaluate Task Complexity

In this experiment, we built two baseline systems to compare the performance of ML algorithms. First, *BaseDict*, a system based on dictionary matching. A lexicon of AEs and medication is compiled from UMLS Metathesaurus using the semantic types as defined in [30]. The baseline system *BaseDict* tags all the instances of the lexicon that match the text. Second, *MetaMapTagger*, a system based on UMLS Metamap. We apply Metamap to tag AEs and medications using UMLS semantic types similar to *BaseDict*.

The baseline systems were compared with taggers built using bag of words as default feature – *NBTagger*, an NB-based tagger; *SVMTagger*, an SVM-based tagger and *SimpleTagger*, a CRF-based tagger built using ABNER default features. We then evaluate the taggers by adding all the features defined in learning features, which we call

NBTagger⁺, *SVMTagger*⁺ and *CombinedTagger* for NB, SVM and CRF based taggers respectively. We also built and ensemble classifier *EnsembleTagger*, an SVM classifier that combines the output of the SVM and CRF taggers.

Systems to Evaluate Impact of Features

We evaluate the impact of various features on the performance of the tagger. We used the ML technique that achieved the best performance in our previous experiment. In addition to the default features trained as *SimpleTagger*, we individually added syntactic features (*SyntacticTagger),* semantic features (*SemanticTagger*), morphological features (*MorphologicalTagger*), affix features (*AffixTagger)*, negation and hedging features (*NegHedgeTagger*), discourse connective features (*ConnectiveTagger*), and a tagger incorporating all the features (*CombinedTagger*) were trained to identify the named entities.

Machine Learning Evaluation Metrics

All the AE taggers trained were evaluated using ten-fold cross validation. We reported recall, precision, and the F1 score. Recall is the ratio of the number of entities of a certain class correctly identified by the system and the number of entities of that class in gold standard. Precision is the ratio of the number of entities of a certain class correctly identified by the system and the number of entities of that class predicted by the system. F1 score is the harmonic mean of precision and recall.

Results

Corpus Characteristics and Annotation Agreement

In this section, we discuss the corpus characteristics and the annotation agreement between the annotators on FAERS reports and EMR narratives. The annotation agreement is calculated based on two criteria: *strict* in which the two annotations have an exact match, and *relaxed* in which there exists an overlap of at least one word between the two annotations. We measured the agreement using *relaxed* criteria to estimate the agreement when the boundary of the entity is ignored.

FAERS Reports

Table 4 below shows the definitions of adverse event and medication-related named entities, the number of annotated instances, and Cohen's κ value. The table also shows the number of instances annotated in all four datasets.

As shown in Table 4, *adverse event* (*AE*) was the most frequently annotated entity followed by *medication* entity in FAERS reports. *Duration* had the least number of annotated instances and lowest kappa value (0.34) for *strict* criteria. *Indication* had the second highest kappa value for *relaxed* criteria (0.93) after *medication* (0.95), since most of the *indication* entities were followed by explicit and unambiguous patterns, such as "for the treatment of", "diagnosed with", "due to", "enrolled in breast cancer study" and so on.

Table 4: Named entity definition, number of instances annotated, inter annotator agreement measured by Cohen's κ for both strict and relaxed criterion of named entities on FAERS reports

Named Entity	Definition	Number of Instances Annotated				κ (strict)	κ (relaxed)
		AnnPhy	AnnLing	Comb	Tie		
Medication	Name of the drug administered to patient including drug class name or medications referred to with pronouns.	1231	1278	1152	1286	0.92	0.95
Dosage	Amount of a single medication used in each administration.	143	315	137	205	0.59	0.82
Route	Method for administering the medication.	115	244	107	132	0.59	0.64
Frequency	How often each dose of the medication should be taken.	25	56	21	42	0.58	0.74
Duration	How long the medication is to be administered.	34	153	24	51	0.34	0.87
Indication	Medical conditions for which the medication is given.	175	148	126	175	0.76	0.93
Adverse event (AE)	Harm directly caused including the pronouns referring to it by the drug at normal doses and during normal use.	1689	2083	1646	1842	0.83	0.93
Other Signs, Symptoms and Diseases (OSSD)	Other symptoms associated with the disease.	234	140	90	147	0.50	0.71
Treatment	Treatment the patient received for the disease.	77	216	62	153	0.39	0.77
Total		3723	4633	3365	4033		

EMR Reports

Table 5 below shows the adverse event and medication-related named entities, the number of annotated instances, and Cohen's κ value on EMR reports. As shown in Table 5, *Other Signs Symptoms and Diseases* (*OSSD*) was the most frequently annotated entity followed by *medication* entity in EMR reports. *Adverse event* had the least number of annotated instances followed by *Duration. Duration had the* lowest kappa value (0.06) for *strict* criteria. *Frequency*, *dosage* and *medication* had the kappa values greater than 0.9 for *relaxed* criteria.

Table 5: Number of entities annotated and their inter annotator agreement measured by Cohen's κ for both strict and relaxed criterion of named entities on EMR reports

Named Entity	Number of Instances Annotated	κ (strict)	κ (relaxed)
Medication	1290	0.92	0.93
Dosage	599	0.89	0.94
Route	484	0.88	0.90
Frequency	964	0.94	0.97
Duration	49	0.06	0.37
Indication	196	0.37	0.75
Adverse event (AE)	26	0.39	0.81
Other Signs, Symptoms and Diseases (OSSD)	2123	0.61	0.88
Total	5731		

Results of Supervised Learning

FAERS Reports

Table 6 below reports recall, precision, and the F1 score of the AETaggers for identifying the AE and other medication-related named entities on each of the four datasets as described in the annotation and data procedure on FAERS reports.

The baseline system *BaseDict* and *MetamapTagger* that matches only *AE* and *medication* achieved an F1 score of 0.45, 0.41, 0.46, 0.42 and 0.41, 0.41, 0.42, 0.40 on *AnnPhy*, *AnnLing*, *Comb*, and *Tie* datasets respectively. Among the taggers using a bag of words as features, CRF-based *SimpleTagger* had the best performance. The addition of features improved the performance of ML classifiers. The *CombinedTagger* achieved best performance of 0.69, 0.74 and 0.73 F1 scores on *AnnPhy*, *AnnLing* and *Comb* datasets respectively. The *SVMTagger$^+$* had the best performance of 0.66 F1 score on *Tie* dataset. The difference in performance between *CombinedTagger* and *SVMTagger$^+$* taggers was statistically significant only on *AnnLing* dataset (t-test, $p < 0.05$). The ML-based taggers clearly outperform the baseline method. The CRF-based tagger had the best overall performance and was therefore chosen as the system to be adapted for subsequent experiments measuring impact of features. *EnsembleTagger* improved the performance on all the four datasets, but the improvement in performance was statistically significant only on the *Tie* dataset (t-test, $p < 0.05$).

We trained CRF-based AETagger's using different features as described in learning features section. The results show that *CombinedTagger* achieved the highest performance on all datasets. Our results also show that the *AnnLing* dataset has the highest performance while *Tie* performs the lowest. *Comb* outperforms both *Tie* and *AnnPhy*.

Since the *Comb* dataset's performance (0.73 F1 score) is close to the highest (0.74 F1 score) and contains annotations agreed by both annotators, we further report feature

Table 6: The performance of the taggers (Mean ± std dev) on each of the four FAERS datasets * (t-test, $p < 0.05$)

	Dataset / ML	AnnPhy F1 score (Precision, Recall)	AnnLing F1 score (Precision, Recall)	Comb F1 score (Precision, Recall)	Tie F1 score (Precision, Recall)
Task Complexity	**BaseDict**	0.45 ± 0.10 (0.86 ± 0.08, 0.31 ± 0.09)	0.41 ± 0.09 (0.91 ± 0.07, 0.27 ± 0.08)	0.46 ± 0.12 (0.82 ± 0.06, 0.32 ± 0.11)	0.42 ± 0.10 (0.86 ± 0.13, 0.28 ± 0.08)
	MetaMapTagger	0.41 ± 0.17 (0.41 ± 0.16, 0.42 ± 0.18)	0.41 ± 0.10 (0.47 ± 0.20, 0.37 ± 0.15)	0.42 ± 0.18 (0.41 ± 0.17, 0.43 ± 0.19)	0.40 ± 0.16 (0.46 ± 0.19, 0.36 ± 0.14)
	NBTagger	0.22 ± 0.08 (0.39 ± 0.17, 0.15 ± 0.05)	0.23 ± 0.08 (0.45 ± 0.14, 0.16 ± 0.06)	0.24 ± 0.08 (0.40 ± 0.17, 0.17 ± 0.05)	0.20 ± 0.06 (0.47 ± 0.19, 0.13 ± 0.04)
	SVMTagger	0.55 ± 0.05 (0.77 ± 0.10, 0.44 ± 0.04)	0.55 ± 0.05 (0.78 ± 0.07, 0.43 ± 0.05)	0.58 ± 0.04 (0.78 ± 0.10, 0.46 ± 0.04)	0.59 ± 0.04 (0.80 ± 0.05, 0.46 ± 0.05)
	SimpleTagger	0.67 ± 0.09 (0.77 ± 0.09, 0.60 ± 0.09)	0.72 ± 0.08 (0.81 ± 0.06, 0.66 ± 0.10)	0.71 ± 0.08 (0.81 ± 0.09, 0.63 ± 0.08)	0.63 ± 0.09 (0.69 ± 0.08, 0.55 ± 0.10)
	NBTagger[+]	0.45 ± 0.09 (0.38 ± 0.10, 0.56 ± 0.06)	0.44 ± 0.06 (0.39 ± 0.07, 0.50 ± 0.06)	0.46 ± 0.09 (0.37 ± 0.11, 0.60 ± 0.04)	0.43 ± 0.07 (0.38 ± 0.08, 0.51 ± 0.07)
	SVMTagger[+]	0.66 ± 0.07 (0.78 ± 0.10, 0.58 ± 0.06)	0.67 ± 0.07 (0.78 ± 0.07, 0.59 ± 0.07)	0.70 ± 0.06 (0.80 ± 0.11, 0.63 ± 0.05)	0.66 ± 0.07 (0.78 ± 0.06, 0.57 ± 0.08)
	CombinedTagger	0.69 ± 0.09 (0.77 ± 0.10, 0.62 ± 0.09)	0.74 ± 0.08 (0.81 ± 0.07, 0.68 ± 0.09)	0.73 ± 0.08 (0.81 ± 0.10, 0.66 ± 0.07)	0.65 ± 0.08 (0.71 ± 0.08, 0.60 ± 0.09)
	EnsembleTagger	0.72 ± 0.09 (0.82 ± 0.10, 0.64 ± 0.09)	0.77 ± 0.08 (0.83 ± 0.07, 0.71 ± 0.09)	0.76 ± 0.07 (0.86 ± 0.10, 0.68 ± 0.06)	0.71 ± 0.08 (0.78 ± 0.06, 0.65 ± 0.10)
Impact of Features	**SimpleTagger**	0.67 ± 0.09 (0.77 ± 0.09, 0.60 ± 0.09)	0.72 ± 0.08 (0.81 ± 0.06, 0.66 ± 0.10)	0.71 ± 0.08 (0.81 ± 0.09, 0.63 ± 0.08)	0.63 ± 0.09 (0.69 ± 0.08, 0.55 ± 0.10)
	AffixTagger	0.67 ± 0.09 (0.78 ± 0.09, 0.60 ± 0.09)	0.73 ± 0.09 (0.81 ± 0.06, 0.66 ± 0.10)	0.70 ± 0.08 (0.81 ± 0.09, 0.63 ± 0.08)	0.61 ± 0.09 (0.70 ± 0.08, 0.52 ± 0.10)
	ConnectiveTagger	0.67 ± 0.09 (0.77 ± 0.09, 0.60 ± 0.09)	0.73 ± 0.08 (0.81 ± 0.06, 0.66 ± 0.10)	0.71 ± 0.08 (0.81 ± 0.09, 0.63 ± 0.08)	0.63 ± 0.09 (0.70 ± 0.07, 0.57 ± 0.10)
	MorphologicalTagger	0.68 ± 0.09 (0.77 ± 0.08, 0.60 ± 0.10)	0.73 ± 0.08 (0.81 ± 0.06, 0.66 ± 0.09)	0.71 ± 0.08 (0.80 ± 0.09, 0.63 ± 0.08)	0.64 ± 0.08 (0.71 ± 0.07, 0.59 ± 0.09)

NegHedgeTagger	0.66 ± 0.09 (0.77 ± 0.09, 0.59 ± 0.10)	0.72 ± 0.08 (0.81 ± 0.06, 0.65 ± 0.10)	0.71 ± 0.08 (0.81 ± 0.09, 0.63 ± 0.08)	0.61 ± 0.09 (0.69 ± 0.08, 0.54 ± 0.10)
SemanticTagger	0.68 ± 0.09 (0.77 ± 0.10, 0.61 ± 0.09)	0.70 ± 0.09 (0.78 ± 0.07, 0.64 ± 0.10)	0.72 ± 0.09 (0.80 ± 0.11, 0.65 ± 0.08)	0.63 ± 0.09 (0.69 ± 0.10, 0.58 ± 0.09)
SyntacticTagger	0.68 ± 0.09 (0.78 ± 0.09, 0.61 ± 0.10)	0.72 ± 0.08 (0.80 ± 0.06, 0.65 ± 0.09)	0.71 ± 0.08 (0.80 ± 0.09, 0.64 ± 0.08)	0.63 ± 0.08 (0.70 ± 0.08, 0.58 ± 0.09)
CombinedTagger	0.69 ± 0.09 (0.77 ± 0.10, 0.62 ± 0.09)	0.74 ± 0.08 (0.81 ± 0.07, 0.68 ± 0.09)	0.73 ± 0.08 (0.81 ± 0.10, 0.66 ± 0.07)	0.65 ± 0.08 (0.71 ± 0.08, 0.60 ± 0.09)

analyses using the *Comb* dataset. Table 7 below shows how different learning features affect AETagger's performance on FAERS data. The results show that adding a single feature added little to overall performance, although the performance of different entities varied. Affix features improved *route* and *duration* but decreased *AE, medication,* and *dosage*. This could be because affix features capture the explicit patterns of *route* such as "venous" in "intravenous." The decrease in performance of affix features could be due to the sparsity of such common patterns. Connective features increased the performance of *dosage*, *route*, and *indication*; however, the performance of *medication* decreased. This could be again due to the presence of connectives around *dosage*, *route*, and *indication* entities such as "…surged from an endometrial carcinoma, she ***then*** had chemo-radiotherapy…" The connective "then" appears around the *indication* "endometrial carcinoma." But the decrease in performance for medication could be due to the sparsity of such patterns. Other features (morphological, negation, hedge, semantic, and syntactic) showed similar patterns. On the other hand, when all features were added, the overall performance increased to 0.73 F1 score (default 0.71), although the increase was not statistically significant (t-test, $p < 0.05$).

Table 7: The F1 score performance of different named entities with different features of *Comb* dataset on FAERS dataset.

Feature group	AE	Medication	Dosage	Frequency	Route	Duration	Indication	OSSD	Treatment	Overall
Default	0.70 ± 0.10	0.82 ± 0.10	0.59 ± 0.35	0.57 ± 0.46	0.36 ± 0.33	0.20 ± 0.42	0.57 ± 0.12	0.44 ± 0.45	0.60 ± 0.52	0.71 ± 0.08
Affix	0.69 ± 0.11	0.81 ± 0.12	0.58 ± 0.37	0.59 ± 0.45	0.55 ± 0.37	0.40 ± 0.52	0.57 ± 0.09	0.51 ± 0.44	0.60 ± 0.52	0.70 ± 0.08
Connective	0.70 ± 0.10	0.81 ± 0.10	0.69 ± 0.31	0.57 ± 0.46	0.44 ± 0.36	0.20 ± 0.42	0.60 ± 0.15	0.44 ± 0.45	0.60 ± 0.52	0.71 ± 0.08
Morphological	0.70 ± 0.10	0.82 ± 0.10	0.57 ± 0.35	0.59 ± 0.45	0.32 ± 0.32	0.20 ± 0.42	0.62 ± 0.12	0.47 ± 0.43	0.60 ± 0.52	0.71 ± 0.08
NegHedge	0.69 ± 0.10	0.82 ± 0.10	0.56 ± 0.36	0.59 ± 0.45	0.36 ± 0.33	0.20 ± 0.42	0.59 ± 0.11	0.50 ± 0.43	0.60 ± 0.52	0.71 ± 0.08
Semantic	0.71 ± 0.11	0.82 ± 0.11	0.56 ± 0.35	0.65 ± 0.40	0.34 ± 0.33	0.30 ± 0.48	0.64 ± 0.13	0.43 ± 0.39	0.60 ± 0.52	0.72 ± 0.09
Syntactic	0.70 ± 0.10	0.81 ± 0.11	0.61 ± 0.35	0.59 ± 0.45	0.32 ± 0.31	0.34 ± 0.47	0.58 ± 0.11	0.44 ± 0.45	0.60 ± 0.52	0.71 ± 0.08
All	0.72 ± 0.10	0.83 ± 0.11	0.61 ± 0.37	0.59 ± 0.44	0.32 ± 0.31	0.34 ± 0.47	0.65 ± 0.11	0.55 ± 0.39	0.60 ± 0.52	0.73 ± 0.08

EMR Reports

Table 8 below reports recall, precision, and F1 score of the AETaggers for identifying the AE and other medication-related named entities on EMR reports.

The baseline system *BaseDict* and *MetamapTagger* that matches only *AE* and *medication* achieved an F1 score of 0.45 and 0.42 respectively, similar to the performance on FAERS reports. Among the taggers using bag of words as features, CRF-based *SimpleTagger* had the best performance. The *SVMTagger*[+] achieved an F1 score of 0.62. Addition of features did not improve the performance in the case of EMR reports. The difference in

performance between *CombinedTagger* and *SVMTagger*[+] taggers was statistically significant (t-test, $p < 0.05$). The ML-based taggers clearly outperform the baseline method. The CRF-based tagger had the best overall performance and was therefore chosen as the system to be adapted for subsequent experiments measuring the impact of features.

We trained CRF-based AETagger's using features described in learning features section. The results show that all the taggers performed equally well and addition of features did not affect the performance of the *SimpleTagger*.

Table 8: Performance of the taggers (Mean ± std dev) on EMR reports.

	System	Precision	Recall	F1 score
Task Complexity	**BaseDict**	0.76 ± 0.14	0.33 ± 0.12	0.45 ± 0.13
	MetaMapTagger	0.41 ± 0.17	0.41 ± 0.10	0.42 ± 0.18
	NBTagger	0.64 ± 0.29	0.10 ± 0.02	0.17 ± 0.02
	SVMTagger	0.82 ± 0.03	0.55 ± 0.04	0.66 ± 0.04
	SimpleTagger	0.81 ± 0.03	0.72 ± 0.04	0.77 ± 0.03
	NBTagger[+]	0.40 ± 0.05	0.45 ± 0.03	0.42 ± 0.04
	SVMTagger[+]	0.75 ± 0.17	0.53 ± 0.16	0.62 ± 0.18
	CombinedTagger	0.81 ± 0.03	0.72 ± 0.04	0.76 ± 0.03
	EnsembleTagger	0.83 ± 0.03	0.75 ± 0.04	0.79 ± 0.03
Impact of Features	**SimpleTagger**	0.81 ± 0.03	0.72 ± 0.04	0.77 ± 0.03
	AffixTagger	0.82 ± 0.03	0.72 ± 0.05	0.77 ± 0.04
	ConnectiveTagger	0.82 ± 0.03	0.72 ± 0.05	0.77 ± 0.04
	MorphologicalTagger	0.82 ± 0.03	0.73 ± 0.04	0.77 ± 0.04
	NegHedgeTagger	0.82 ± 0.03	0.72 ± 0.05	0.77 ± 0.04
	SyntacticTagger	0.82 ± 0.03	0.72 ± 0.04	0.76 ± 0.04
	CombinedTagger	0.81 ± 0.03	0.72 ± 0.04	0.76 ± 0.03

Later we trained another tagger by combining the output of *CombinedTagger* and *SVMTagger*[+] taggers to produce an *EnsembleTagger*. The *EnsembleTagger* had the best F1 score of 0.79. The difference in improvement of *EnsembleTagger* against *CombinedTagger* and *SVMTagger*[+] taggers was statistically significant (t-test, $p < 0.05$).

Table 9 below shows how different learning features affect AETagger's performance on EMR reports. It was interesting to see that adding features had little or no influence on the overall performance of the taggers and their performance difference was not statistically significant (t-test, $p < 0.05$). From this we could infer that the features are really very sparse and do not have a strong influence on the performance of the ML model. The CRF model incorporated in the ABNER with its default features works the best.

Table 9: The F1 score performance of the different named entity categories with different features on EMR reports.

Feature group	AE	Medication	Dosage	Frequency	Route	Duration	Indication	OSSD	Overall
Default	0.29 ± 0.10	0.83 ± 0.04	0.85 ± 0.07	0.92 ± 0.03	0.89 ± 0.03	0.52 ± 0.13	0.30 ± 0.10	0.64 ± 0.04	0.77 ± 0.03
Affix	0.26 ± 0.15	0.84 ± 0.03	0.87 ± 0.05	0.92 ± 0.05	0.88 ± 0.06	0.53 ± 0.13	0.32 ± 0.11	0.64 ± 0.05	0.77 ± 0.04
Connective	0.29 ± 0.19	0.84 ± 0.04	0.87 ± 0.06	0.91 ± 0.04	0.88 ± 0.04	0.53 ± 0.14	0.30 ± 0.12	0.64 ± 0.05	0.77 ± 0.04
Morphological	0.26 ± 0.20	0.85 ± 0.03	0.87 ± 0.05	0.91 ± 0.04	0.88 ± 0.05	0.52 ± 0.14	0.30 ± 0.13	0.63 ± 0.04	0.77 ± 0.04
NegHedge	0.29 ± 0.19	0.83 ± 0.04	0.86 ± 0.06	0.91 ± 0.05	0.88 ± 0.04	0.51 ± 0.13	0.32 ± 0.13	0.64 ± 0.05	0.77 ± 0.04
Syntactic	0.30 ± 0.17	0.84 ± 0.03	0.87 ± 0.06	0.90 ± 0.05	0.88 ± 0.04	0.51 ± 0.12	0.33 ± 0.12	0.64 ± 0.04	0.76 ± 0.04
All	0.27 ± 0.11	0.84 ± 0.03	0.87 ± 0.05	0.91 ± 0.04	0.87 ± 0.06	0.52 ± 0.13	0.33 ± 0.10	0.66 ± 0.05	0.76 ± 0.03

Annotation Disagreements

We manually analyzed the annotation disagreements and found they can be organized into three main categories:

(1) Boundary inconsistencies – disagreement due to inconsistent boundary annotation.

(2) Missed named entity annotations – disagreement where one annotator annotated an entity and the other annotator completely failed to annotate it.

(3) Inconsistent named entity annotations – disagreement due to inconsistent categorization of entities.

There were a total of 2,955 disagreed token instances in FAERS reports, of which 1,591 (~54%) were related to *AE* and *medication* named entities. Similarly, there were a total of 2,227 disagreed token instances in 25 EMR reports, of which 1,625 (~73%) were related to OSSD.

Boundary Inconsistencies

We found that inconsistencies related to boundary accounted for nearly 13.9% (412 of 2,955) of disagreement in FAERS reports and 8.6% (191 of 2,227) disagreement in EMR reports. In all the examples in the article, the named entity instance is shown in **bold** and the named entity type is shown within the "[]."

Example 1: She received approximately less than two minutes of therapy with intravenous Taxol (paclitaxel), 280 mg in a **three hour** [duration]

```
infusion [route] for phase IIID ovarian cancer, when the symptoms
occurred.
```

In example 1, AnnLing annotated "three hour" as *duration* and "infusion" as *route*, AnnPhy annotated "three hour infusion" as *duration* only. This inconsistency exemplifies differences between the linguist and the physician. While the linguist can separate the linguistic differences between different named entities, we found that physicians (both ZFL and SB) frequently overlook the differences, which leads to inconsistent annotations.

Missed Named Entity Annotations

Missed named entity annotation was the major cause for disagreement. Among 2,955 disagreed token instances, 2,355 (~79.7%) belong to this category in FAERS report and among 2,227 disagreed token instances, 1872 (~84%) belong to this category in EMR reports. Table 10 shows instances of *medication* that were annotated by one annotator and missed by other on FAERS reports. Examples 1-5 were annotated by AnnPhy but missed by AnnLing; examples 6-10 were annotated by AnnLing but missed by AnnPhy.

AnnLing explained that "blood transfusion," "fluids," and "red packed cells" shown in examples 1, 2 and 5, were not *medication*, but referred to a kind of treatment or medical procedure. In example 3, AnnLing missed annotating "normal saline" as *medication*. In example 4, "oxygen" was not annotated because AnnLing felt it did not represent *medication.* Annotators did not reach any consensus on whether to annotate "oxygen" as *medication* or not. The differences here exemplify the strength of the physician as a

domain expert who may interpret the semantics of EMR notes more accurately than the linguist.

In contrast, in examples 7, 8, and 10, AnnPhy did not annotate "treatment," "Re-exposure," and "chemotherapy" as these entities were anaphoric references; AnnLing, being a linguist, annotated these anaphoric references as *medication*. In example 6, AnnLing annotated "drug" as *medication* but AnnPhy did not annotate the entity because the text did not refer to any *medication*. Later, AnnLing agreed that while there is mention of entities, they do not refer to specific entities such as "drug" in example 6 and therefore should not be annotated. Example 9 was a special case where "concomitant drug" refers to the role or function of the drug "Solupred" rather than referring to a drug. AnnPhy did not annotate such instances. These examples demonstrated that annotating medical texts is a complex and cumbersome task. Further refinement of guidelines in such instances may improve the consistency of annotations.

Table 10: Disagreement in medication annotation (medication is shown in bold) on FAERS reports

Annotated by AnnPhy but not annotated by AnnLing	1. Given multiple **blood transfusions** (hemoglobin: 4.8). 2. Pressors continued with **fluids**. 3. He was admitted to the hospital and hydrated with **normal saline**. 4. The event was treated with steroids and **oxygen**. 5. Pancytopenia, treated with G-CSF, erythropoetin, and **red packed cells.**
Annotated by AnnLing but not annotated by AnnPhy	6. Causality assessment **drug** relationship is unable to determine for Taxol 7. The 4th previous courses of **treatment** were well tolerated. 8. During the first infusion of paclitaxel, the patient experienced a decrease in blood pressure and was unconscious for a short while. **Re-exposure** elicited the same symptoms. 9. The **concomitant drug** prescribed was oral Solupred instead of Solumedrol 10. A female patient possibly received non-therapeutic dosages of intravenous Taxol (paclitaxel), Paraplatin (carboplatin), and/or Platinol (cisplatin) for the treatment of ovarian cancer and subsequently expired. It was reported that the pharmacist possibly diluted **the chemotherapy** improperly.

Inconsistent Named Entity Category Annotations

We have annotated a total of eight different named entity types, as shown in Table 4 and Table 5. The third type of inconsistency was caused by inconsistent named entity assignments. Among 2,955 disagreed token instances, 188 (~6.4%) belong to inconsistent named entity category in FAERS report and among 2,227 disagreed token instances, 164 (~7.4%) belong to inconsistent named entity category. We manually examined few instances and examples 2-6 below show the annotated sentences where inconsistency occurred in FAERS reports. Examples 7 and 8 show the annotated sentences where annotators disagreed in EMR reports. Example 2 is an instance where both annotators agreed on the AE annotation.

Example 2: *The patient then became **lightheaded** [AE], **collapsed** [AE], and was **unconscious** [AE].*

Example 3, however, shows an instance where AnnPhy and the tiebreaker agreed on "haematologic toxicity" as an AE whereas AnnLing did not initially annotate the entity. This inconsistency suggests that domain knowledge is required for annotation. After discussion with two other annotators, AnnLing agreed that "haematologic toxicity" should be annotated an AE.

Example 3: *Investigator considers that **haematologic toxicity** [AE] of methotrexate could be increased by interaction with apranax (naproxene) and sintrom (acenocoumarol).*

Example 4 shows an instance where AnnLing and tiebreaker agreed on "cardiogenic shock" as an AE but AnnPhy annotated it as *OSSD*. AnnPhy argued that "cardiogenic

shock" caused "death;" therefore "death" should be an *AE* and "cardiogenic shock" is the reason for death and therefore was annotated as *OSSD*. This example shows the complexity of clinical cause.

Example 4: *On [words marked], the patient died, presumed to be a result of **cardiogenic shock** [AE]. Prior to death, on [words marked], the patient was noted for having an increase in troponin T level, and found to be more unresponsive.*

In example 5, the tiebreaker annotated "allergy" as an AE, whereas AnnPhy annotated it as OSSD and AnnLing did not annotate it as an AE because it refers to the patient's history of "allergy" and does not represent a current instance of AE. We will need to refine our annotation guideline to add current or past status in addition to the named entity annotation.

Example 5: *Moderate anaphylactoid symptom appeared after administration of docetaxel and recovered later. After the end of administration, convulsion appeared. Anti-convulsion agent could not be administered due to **allergy** [AE].*

Example 6 shows an instance of boundary inconsistency. AnnPhy and AnnLing both annotated "NCI/CTC grade 4 neutropenia without fever" as an AE whereas the tiebreaker annotated "NCI/CTC grade 4 neutropenia" as an AE and "fever" as OSSD. This is a case in which annotators interpret clinical texts differently. Such an inconsistency is difficult to address due to the nature of ambiguity in clinical texts.

Example 6: *...days after the last Vinorelbine intake patient was hospitalized due to **NCI/CTC grade 4 neutropenia** [AE] without **fever** [OSSD] ...*

Example 7 shows an instance of inconsistent entity categorization in EMR reports. AnnPhy annotated the entire span "baby aspirin" as medication, whereas AnnLing annotated "baby" as dosage and "aspirin" as medication. This shows the complexity on medical narratives and how annotators interpret text differently as we saw before.

Example 7: *... I have also advised her to change her aspirin from 325 to a **baby** [dosage] **aspirin** [medication] especially in view of now the concomitant Coumadin use.*

In Example 8, AnnPhy annotated the "minimally invasive esophagectomy" as OSSD, but AnnLing failed to annotate it showing the advantage of having domain knowledge.

Example 8: *...gentleman with a complicated medical history secondary to a **minimally invasive esophagectomy** [OSSD] with complications of an esophageal leak requiring an initial takedown...*

Error Analyses

FAERS Reports

For error analyses on FAERS reports, we focused on *CombinedTagger* because it yielded the highest performance (as shown in Table 6) and the *Comb* dataset because it contained annotations agreed on by both annotators. We randomly selected 100 named entities predicted wrongly by *CombinedTagger* and manually analyzed them. As shown in Figure 5, we group the errors into a total of five types of errors and give an illustrative example for each. In all examples, annotated named entities are shown in **bold**, the tagger output in {*italicized*} and the named entity type is shown within "[]". The leading type of error was data sparseness (35%). Data sparseness is a common problem and the major cause of

poor performance. For instance, the gold standard consisted a number of singleton instances (instances that appear only once) like "cytolysis," "sodium chloride solution 0.9% 100ml," and "neoplasm of unspecified nature of respiratory system" that created sparseness in the data.

The second cause of error was inconsistent inclusion of punctuation (21%). The gold standard had an inconsistency in inclusion of punctuation (eg., a period [.] in "neutropenia.") as a part of the named entity. This boundary inconsistency reduced the overall performance. Figure 5 shows an instance where the gold standard included a period as part of named entity ("neutropenia.") but the tagger failed to include it ("neutropenia"). This was followed by an error caused by ambiguous named entities (15%). The instances in gold standard that were assigned to multiple named entity categories resulted in ambiguous entities. For example, "death" was annotated as either *AE* or *OSSD*. This could have confused the ML algorithm and yielded a lower performance. In Figure 5, the instance "death" was not annotated as *AE* in the *Comb* dataset due to disagreements between annotators, but tagger identified it as an *AE*. The missed pronoun annotations such as "the event," contributed to 8% of the errors. The final category was for the other type of errors (21%), for which the exact cause of error could not be determined. In Figure 5, "seizure" was annotated as an *AE* but the tagger failed to identify it. The exact cause for miscategorization could not be determined.

For error analyses on EMR reports, we focused on *EnsembleTagger* because it yielded the highest performance (as shown in Table 8). We construct an error matrix of word tokens as shown in Table 11 below to investigate the errors caused by the

EnsembleTagger. We can see that a huge portion of errors belonged to *OSSD* and *Indication* categories.

Sampled Error

Category	Bar	Example
Pronoun Error		… and asthenia. **The event** [AE] was considered medically serious.
Ambiguity Error		Cause and date of death {*death* [AE] } were unspecified.
Punctuation Error		…was admitted to hospital to febrile **neutropenia.** [AE] {*neutropenia* [AE] }…
DataSparseness Error		… Liver function tests showed **cytolysis** [AE] with transaminases-6000…
Other Errors		… event which occurred was probably a **seizure** [AE] {*seizure*}.

Num of Errors (0, 5, 10, 15, 20, 25, 30, 35)

Figure 5: Error categories, their frequency and an illustrative example of error category on 100 randomly sampled instances in FAERS reports. The annotated entities are shown in **bold**, the entity type is shown in "[]," and tagger output is *{italicized}*.

Table 11: Error matrix showing the distribution of tokens and their categories on EMR reports by *EnsembleTagger*

		Predicted Category								
		O	**ad**	**m**	**route**	**indication**	**f**	**OSSD**	**du**	**do**
Gold Standard	**O**	67849	11	231	108	69	89	2435	26	76
	ad	69	24	1	1	1	1	97	0	2
	m	339	1	3309	19	6	7	68	0	45
	route	131	0	12	1392	0	17	10	0	35
	indication	308	0	18	3	111	10	555	0	13
	f	281	0	5	10	2	3236	10	5	8
	OSSD	4308	5	70	47	130	0	10603	0	1
	du	175	0	0	3	0	9	0	164	6
	do	356	0	53	33	1	18	12	6	3316

Similar to the FAERS report, we randomly selected 100 named entities predicted wrongly by *EnsembleTagger* and manually analyzed them. We group the errors into a total of five types of errors and give an illustrative example for each as shown in Figure 6. The leading type of error was ambiguous named entities in EMRs (59%). For example,

"prophylaxis" was annotated as either *indication* or *OSSD*. This could have confused the ML algorithm and yielded a lower performance as in the FAERS reports. In Figure 6, the instance "prophylaxis" was not annotated as *indication* in EMR reports, but the tagger identified it as an *OSSD*. Data sparseness was the second major reason of error (18%). For instance, the gold standard consisted of a number of singleton instances (instances that appear only once) like "slurred speech," "pulse is in the low 100s," and "popliteal veins" that created sparseness in the data.

This was followed by errors due to inconsistent inclusion of punctuation, which also caused 10% of the error. The gold standard had an inconsistency in inclusion of punctuation (eg., a period [.] in "insulin.") as a part of a named entity. This boundary inconsistency reduced the overall performance. Figure 6 shows an instance where the gold standard included a period as part of a named entity ("insulin.") but the tagger failed to include it ("insulin"). This was followed by error caused by ambiguous named entities (15%). The missed pronoun annotations such as "medications," contributed to 2% of the errors. The final category was other type of errors (11%), for which the exact cause of error could not be determined. In Figure 6, "hemodialysis" was not annotated, but the tagger identified it as an *OSSD*. The exact cause for miscategorization could not be determined.

Annotation Inconsistencies

As predicted, annotation inconsistency played an important role on AETaggers' performance as our Pearson correlation results (coefficient of 0.73 on FAERS and 0.83 on EMR reports) show that the IAA value (Cohen's κ) is positively correlated with

machine-learning performance of named entity recognition. This is not surprising because inconsistent annotations confuse the machine-learning systems.
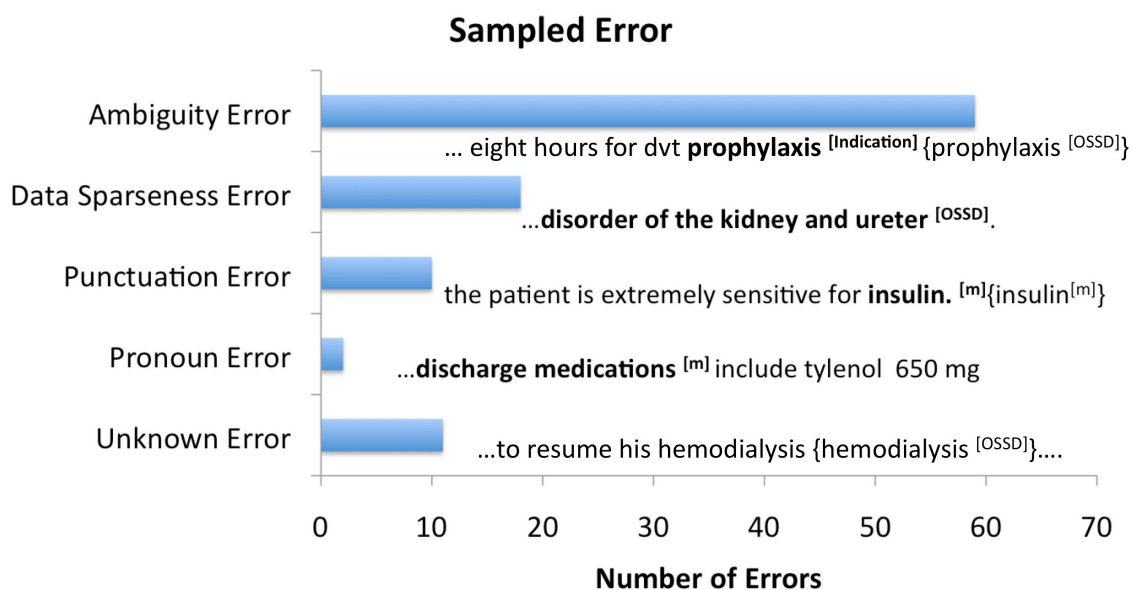
Figure 6: Error categories, their frequency and an illustrative example of error category on 100 randomly sampled instances on EMR reports. The annotated entities are shown in **bold**, the entity type is shown in "[]," and tagger output is *{italicized}*.

Our manual analysis of inconsistency revealed that nearly 20% of errors were due to inconsistent inclusion of punctuation in the annotation in FAERS reports. When we removed the inconsistency in punctuation, the F1 score of *CombinedTagger* increased from 0.73 to 0.79, which was statistically significant (t-test, $p < 0.05$). Unlike FAERS reports, manual analysis of EMR reports showed only ~6% of errors were related to punctuation inconsistencies. The missed pronoun annotations of *AE* and *medication*,

although they can be fixed readily, also contributed to the lower performance of the tagger.

Data Sparseness

Data sparseness is a common problem and the major cause of poor performance. The performance of AETagger was positively correlated with the size of annotated data for each named entity (a Pearson correlation coefficient of 0.64 in FAERS reports and 0.49 in EMR reports). In the cases of *frequency*, *duration, OSSD,* and *treatment* entities, data were very sparse (Table 4) and taggers showed low performance on these named entities. In addition to low performance, data sparseness also contributed to a higher standard deviation (Table 7). When the training data incorporate instances of a named entity but the testing data do not, the precision decreases. When the training data miss instances of a named entity but the testing data do not, then recall suffers.

Learning Features

To further understand the contribution of learning features on the performance of AETagger in FAERS data, we first trained the tagger with all the features and used it as a baseline system (*CombinedTagger*). We then removed each feature category one at a time. Table 12 below shows the performance of taggers when a feature category was removed iteratively from the *CombinedTagger*. Consistent with Table 7, the results show that each feature contributed to the performance differently. But in EMR reports, the removal of features did not have any effect on the performance of the taggers.

Discussion

Our results show that medication and adverse events can be reliably annotated (Cohen's κ value of 0.64~0.95 IAA as shown in Table 4) in the FAERS narratives and (Cohen's κ value of around 0.9 IAA for most of the categories as shown in Table 5) in EMR reports. Many named entities (e.g., *indication*) that had shown low annotation agreements in the i2b2 challenge [76] had good annotation agreements on FAERS dataset.

Table 12: The precision, recall, and F1 score (mean ± std dev) of taggers with feature categories removed one at a time on each of the four annotated datasets.

Dataset / Tagger	AnnPhy F1 score (Precision, Recall)	AnnLing F1 score (Precision, Recall)	Combined F1 score (Precision, Recall)	Tie F1 score (Precision, Recall)
All Features	0.67 ± 0.09 (0.77 ± 0.10, 0.62 ± 0.09)	0.74 ± 0.08 (0.81 ± 0.07, 0.68 ± 0.09)	0.73 ± 0.08 (0.81 ± 0.10, 0.66 ± 0.07)	0.65 ± 0.08 (0.71 ± 0.08, 0.60 ± 0.09)
No Affix Features	0.68 ± 0.09 (0.76 ± 0.10, 0.62 ± 0.09)	0.71 ± 0.10 (0.78 ± 0.07, 0.65 ± 0.11)	0.71 ± 0.09 (0.79 ± 0.11, 0.64 ± 0.09)	0.64 ± 0.08 (0.70 ± 0.08, 0.60 ± 0.08)
No Connective Features	0.69 ± 0.09 (0.77 ± 0.10, 0.62 ± 0.09)	0.74 ± 0.08 (0.81 ± 0.06, 0.69 ± 0.09)	0.73 ± 0.08 (0.81 ± 0.10, 0.66 ± 0.07)	0.65 ± 0.08 (0.71 ± 0.08, 0.60 ± 0.09)
No Morphological Features	0.69 ± 0.09 (0.78 ± 0.10, 0.62 ± 0.09)	0.73 ± 0.08 (0.81 ± 0.06, 0.66 ± 0.09)	0.73 ± 0.08 (0.82 ± 0.10, 0.66 ± 0.07)	0.65 ± 0.08 (0.72 ± 0.08, 0.60 ± 0.08)
No Negation and Hedge Features	0.68 ± 0.09 (0.77 ± 0.10, 0.62 ± 0.09)	0.74 ± 0.08 (0.81 ± 0.07, 0.68 ± 0.09)	0.72 ± 0.08 (0.81 ± 0.10, 0.65 ± 0.07)	0.64 ± 0.09 (0.71 ± 0.09, 0.59 ± 0.09)
No Semantic Features	0.67 ± 0.08 (0.77 ± 0.08, 0.60 ± 0.09)	0.74 ± 0.08 (0.82 ± 0.05, 0.68 ± 0.10)	0.71 ± 0.08 (0.80 ± 0.09, 0.64 ± 0.08)	0.64 ± 0.08 (0.71 ± 0.07, 0.59 ± 0.08)
No Syntactical Features	0.68 ± 0.09 (0.77 ± 0.10, 0.61 ± 0.09)	0.73 ± 0.08 (0.80 ± 0.07, 0.68 ± 0.09)	0.71 ± 0.09 (0.80 ± 0.11, 0.64 ± 0.08)	0.64 ± 0.08 (0.70 ± 0.09, 0.58 ± 0.09)

The improvements were attributed to improved annotation guidelines and the quality and domain specificity of the FAERS narratives.

With a good IAA, we still found room to further improve the annotation guideline. For example, our error analyses (Figure 5) show that inconsistencies were introduced by annotation boundary; therefore it can be further refined. Although *medication* had the highest IAA (0.95 on FAERS reports and 0.93 on EMR reports), our analysis (Table 10) found that the inconsistency in *medication* was introduced by whether instances like "fluids" could be considered as medication or not. In the future, we may separate *medication* into two classes: *strict medication* and *relaxed medication*. The names and mentions of all drugs appearing in the United States Pharmacopeia will belong to *strict medication;* any substances or chemicals — including oxygen, fluids, drinks, and others – given to patients during the treatment will be classified as *relaxed medication*. Refining the guideline to annotate previous and potential AEs like "allergy" (example 5) may further reduce the inconsistency.

We explored various ML methods and compared them with a baseline string matching system to assess the complexity of the task. Our model achieved comparable performance of 0.74 F1 score on FAERS data and 0.76 on EMR data to our previous work in i2b2 that had an F1 score of 0.76 [76]. The CRF-based tagger had the best performance. Further analyses of the CRF tagger found that data sparseness affected the taggers' performance (Figure 5 and Figure 6). For example, the standard deviation of *treatment* is high because we found that the testing data did not incorporate *treatment* instances in FAERS report. Similar behavior was also observed for other sparse entities (Table 7).

Using the best performing ML technique, we explored a variety of features (Table 6 and Table 7). The features had a mixed effect on the performance of the taggers and the

combination of all the features improved overall performance slightly. This suggests the robustness of the default features for CRFs. Since most of the features were extracted automatically, for example, negation, hedge cues, and discourse connectives were extracted using the taggers [86,87] and parser [88] we developed. The accuracy of the extracted features played an important role in overall performance of the tagger. To avoid the noise introduced by automatic feature extraction, one may explore the features manually annotated such as POS in PennTree Bank [89]. This is, however, expensive. An alternative is to further improve the performance of the BioNLP systems for feature extraction.

Throughout the study, we found that additional features may be further included. For example, we observed that *OSSD* most often appeared in the patient's medical history. We therefore, added a feature representing patient history and found that the performance of the *CombinedTagger* on *OSSD* increased 1.2% absolute (results not reported in the Result section), although the increase was not statistically significant (t-test, $p < 0.05$).

Our study has limitations. The AETaggers were trained on the FAERS and EMR report corpus we constructed. Like any other NLP system, the performance of the tagger on other types of records can vary based on the structure and content of the narrative text. On the other hand, since our selection of the FAERS and EMR reports corpus was through a random process, we speculate that the data are representative. Although the taggers performed well, the training and evaluation were based on a relatively small training data. We speculate that increasing the size of training can further improve the performance.

Conclusion

In this study, we developed an annotation guideline for medication and adverse event information from the FAERS and EMR narratives. Our annotation of 122 FAERS narratives (a total of ~23K tokens) and 150 EMR reports (a total of ~103K tokens) showed a reliable inter-rater annotation agreement. We then developed machine-learning based models for automatically extracting medication and adverse event information from the FAERS narratives. We explored different learning features. The results show that features such as syntactic, semantic, morphological, and affix improved the performance and the best performing system had an overall F1 score of 0.73 on FAERS reports and 0.77 on EMR reports. In the future, we would like to refine further the annotation guideline, explore features and increase the annotation size to improve system performance. We will also explore approaches for normalizing the entities by mapping them to standard terminologies like MedDRA and identify the causal relation between a medication and an adverse event.

Chapter 3: Automatic Discourse Connective Detection in Biomedical Text

This chapter discusses the identification of discourse connectives in biomedical text. Discourse connectives are used as features in the named entity recognition task as they appeared 63.4% of the time around entities (Table 2). The identification of the discourse connectives can also aid in the automation of the elements of the Naranjo Score that calculates the likelihood of adverse events caused due to drugs.

Relation extraction in biomedical text mining systems has largely focused on identifying clause-level relations, but increasing sophistication demands the recognition of relations at discourse level. A first step in identifying discourse relations involves the detection of discourse connectives: words or phrases used in text to express discourse relations. We describe the development and evaluation of supervised machine-learning approaches for automatically identifying discourse connectives in biomedical text. Two supervised machine-learning models (support vector machines and conditional random fields) for identifying discourse connectives in biomedical literature were explored. We trained in-domain supervised machine-learning classifiers on the Biomedical Discourse Relation Bank (BioDRB), an annotated corpus of discourse relations over 24 full-text biomedical articles (~112,000 word tokens), subset of the GENIA corpus. We also explored novel domain adaptation techniques to leverage the larger open-domain Penn Discourse Treebank (PDTB) (~1 million word tokens). We evaluated the models using the standard evaluation metrics of precision, recall, and F1 scores. Supervised machine-learning approaches can automatically identify discourse connectives in biomedical text, and the

novel domain adaptation techniques yielded the best performance: 0.761 F1 score. A demonstration version of the fully implemented classifier BioConn is available at: http://bioconn.askhermes.org. Further experiments for identifying the sense of explicit discourse connectives show the connective itself as a highly reliable indicator for coarse sense classification achieving a performance of 0.9 F1 score.

We apply this model developed in biomedical journal articles to narrative text written by health-care providers. While, the model performs well in the biomedical journal domain, we expect there would be a drop in performance when applied to medical text. Future work will focus on developing an annotated gold standard with discourse relations and identifying discourse connectives on medical text.

Introduction

The desire for knowledge discovery through text mining of biomedical literature has led to a great deal of research on the extraction and retrieval of valuable and useful information from biomedical text, through natural language processing (NLP) methods developed for recognizing entities (e.g., proteins, genes, drugs, diseases, etc.), facts, hypotheses, events, and relations between entities. However, with the exception of some recent work on coreference resolution [90], much of this processing has been restricted to the level of the clause, focusing on identifying entities and relations within a clause, and has ignored the importance of identifying relations expressed at the level of discourse, i.e., relations expressed across clauses or sentences. In example 1, for instance, queries regarding the inhibitory effect of IL-10 could be answered more accurately when the "concession" relation between the two sentences is identified, signaled by the word

However. Taking the first sentence alone would otherwise lead to the false inference that the IL-10 mediated inhibitory effect is unrestricted.

> ***Example 1:*** *IL-10-mediated inhibition of CD4+ T-cell cytokine production is principally dependent on its inhibition of macrophage antigen-presenting cell function [1].* **However**, <u>this indirect inhibitory effect is thought to be restricted at the site of T-cell activation in RA</u>… (Concession: Contra-expectation)

Knowledge of such relations, called discourse relations, can be very useful in extracting various kinds of biomedical information. In this chapter, we present the first investigations into identifying discourse relations in biomedical literature. We focus on identifying "discourse connectives," which are words or phrases used to indicate the presence of discourse relations, such as the word *However* in example 1. Following the terms and definitions of the Penn Discourse Treebank (PDTB) [91], discourse relations hold between abstract objects (AOs), such as eventualities and proposition, which serve as the arguments to the relation. Each discourse relation is assumed to hold between precisely two arguments (named Arg1 and Arg2). Discourse relations are characterized in terms of several semantic (or sense) classes, including "contrast," "conjunction," "cause," "condition," and "instantiation," among others. In example 2, the word *but* is a discourse connective that indicates the presence of a "contrast" relation between the eventualities expressed by the two sentences. In all the examples in this paper, Arg2, the argument syntactically associated with or bound by the connective is underlined, while Arg1 is shown in italics. The discourse connective is in bold. The semantics (or sense) of the connective is shown in parentheses at the end of the examples.

Example 2: *The phosphorylation of signal transducer and activator of transcription 3 was sustained in both blood and synovial tissue CD4+ T cells of RA,* **but** <u>it was not augmented by the presence of 1 ng/ml IL-10</u>. (Contrast)

Identifying the presence of discourse relations can help in the extraction of valuable information from natural language text and also benefit many natural language processing applications [92-98]. For example, identifying causal discourse relations will make it possible to generate repositories of "why" questions from biomedical text [99]. In general, question-generation systems [96], as well as question-answering systems, stand to benefit greatly from recognizing discourse relations because it will allow for the generation and answering of complex questions about biomedical events and situations.

Discourse relations can also be used to benefit information extraction from clinical narratives. Unique adverse drug event (ADE) information often appears in narratives of electronic health records. While most BioNLP (Biomedical Natural Language Processing) algorithms for ADE extraction are based on co-occurrence of an adverse event and a drug, problems exist with such an approach, as illustrated by examples 3 and 4. In example 3, the connective *after* has a temporal function describing the administration of atenolol; in Example 4, the connective *after* has a causal interpretation. The "bradycardia" is caused by atenolol. We can utilize the presence of such connectives to automate the elements of Naranjo elements that require identifying the presence of temporal and causal relations that exist between the drug and the adverse event.

Example 3: *…atenolol should be continued while he is at hospital and* **after** <u>he is</u>

discharged. (Temporal: Succession)

Example 4: *Patient was noted to bradycardia as heart rate fell to low 50s* **after** taking atenolol, (Cause: Result)

The examples below show instances where identifying discourse relations are very helpful in information retrieval and extraction tasks. The connective *also* in example 5 suggests that its sentence, taken in isolation, does not provide the complete information about where the IgMhi cells are found. In order to complete this information, the previous sentence must be taken into account as well.

Example 5: In control B6 mice, IgMhi cells (in red) were present in the MZ, outside of the MOMA-1+ cells. IgMhi MZ B cells were also found outside of the MOMA-1 ring in p50-/- mice, but in reduced numbers, as expected from the drastic reductions in MZ B cells in these mice. (Conjunction)

Discourse relations can also be useful for categorizing citations and the relations between the citations to enhance information retrieval: the connective *In contrast* in example 6 signals a contrast relation between two cited articles, 48 and 49, mentioned in two different sentences.

Example 6: *The importance of PU.1 for Btk gene regulation is underlined by the fact that the absence of PU.1 leads to a two- to 3-fold reduction of Btk expression (48).* **In contrast**, the deficiency of Sp1 that also stimulates Btk promoter activity together with PU.1 (49) had no influence on Btk expression (49). (Contrast)

For summarization tasks, it is useful to identify summary sentences, as well as the larger text segments that such sentences summarize. Connectives like *In conclusion* in example 7 are important indicators of such relations.

> ***Example 7***: *Consistent with our binding studies, we observed that BOB.1/OBF.1 together with Oct2 was able to activate the murine Btk promoter ~150-fold in a dose-dependent manner (Figure 5A, B and data not shown). Transfection experiments using NIH/3T3 cells revealed that BOB.1/OBF.1 together with PU.1 only marginally enhanced PU.1-mediated Btk promoter activity. In contrast, co-transfection of Oct2 together with PU.1 stimulated PU.1-mediated Btk promoter activity significantly (from 6- to 75-fold). Moreover, co-transfection of PU.1 together with Oct2 and BOB.1/OBF.1 led to an even stronger and synergistic activation (325-fold) of the murine Btk promoter (Figure 5C).* **In conclusion**, these findings indicate that the transcriptional coactivator BOB.1/OBF.1 regulates the Btk promoter activity in B cells in vitro as well as in vivo, in concert with Oct and PU.1 proteins. (Restatement: Generalization)

Causal and justification relations also constitute a very important part of the knowledge dealt within information extraction: for example, the connective *since* in example 8 signals a causal relation between the two clauses. In other words, the fact that "HeLa cells do not express Oct2" is the reason (or reason for believing) that "the addition of an anti-Oct2 antibody did not interfere with complex formation."

> ***Example 8***: *The addition of an anti-Oct2 antibody did not interfere with complex formation,* **since** HeLa cells do not express Oct2. (Cause: Reason)

Example 9 illustrates the importance of accurately disambiguating ADE causal relations. Here, with a co-occurrence approach, both "Solu-Medro" and "cyclosporin" present themselves as the causes of the "acute renal failure." On the other hand, by recognizing the connective *so* and its arguments, we can accurately select "cyclosporin" as the drug causing the renal failure.

> ***Example 9:*** *In the emergency department, he was given one dose of Solu-Medro 500 mg, however, he was found to have elevated cyclosporin levels at 679,* **_so_** *this was thought to be the likely cause of his acute renal failure and his cyclosporin was temporarily held. Since that time, on his hospital day #1, his cyclosporin levels trended down to the point at which there were just slightly over 100 on hospital day #3 and cyclosporin was reinitiated at lower doses. He was dialyzed on admission with removal of 4 liters of fluid, CVM, BK, and LDH were sent from dialysis. His creatinine improved, so further dialysis and biopsy were deemed unnecessary. (A narrative excerpt released by the i2b2 organizer [44].)*

Therefore, the identification of discourse relations would enable text-mining engines to discover not only entities and events but also relations between biological or medical events, such as the temporal and causal relations, relations between facts, and relations between experimental evidence and their conclusions.

Words that function as discourse connectives in some instances may have non-discourse related functions in others. Therefore, one cannot identify discourse connectives by simply using a list of connective expressions and applying pattern matching over the texts. For instance, the word *so* functions as a connective in example 10(a), expressing a result relation, while acting as an intensifier in example 10(b) with no discourse function at all. A similar example of such functional ambiguity is given for "briefly" is shown in example 11. In example 11a, the word "Briefly" is used to express the elaboration or specification relation in discourse, whereas the same word in example 11b functions as a temporal adverbial modifier for an action verb.

*Example 10(a)***:** *however, CsA also inhibits activation of the JNK pathway following TcR/CD3 and CD28 stimulation [29,30]*, and **so** CsA pretreatment may act to prevent early T cell activation of these pathways, thus blocking cytokine production and protecting mice from the effects of subsequent SEB exposure. (Cause: Result)

Example 10(b): It is striking that ductal growth is **so** exquisitely focused in the end buds.

Example 11(a): *CD4+ T cells were isolated from ST samples, as previously described [27]*. **Briefly,** fresh ST samples were fragmented and digested with collagenase and DNase for 1 hour at 37°C. (Restatement:Specification)

Example 11(b): 2.5×10^6 cells were lysed in lysis buffer [100 mM N-2-hydroxyethylpiperazine-N'-2-ethanesulfonic acid (HEPES), pH 7.9, 10 mM KCl,

0.1 mM EDTA, 1.5 mM MgCl2, 0.2% Nonidet P-40, 1 mM dithiothreitol (DTT), and 0.5 mM PMSF], **briefly** vortexed at a moderate speed, then incubated on ice for 5 minutes.

Automatic discourse parsing comprises several subtasks, including discourse connective detection, argument detection, discourse connective sense categorization, and discourse structure composition. The first step toward a full-fledged discourse relation detection system and parser is the detection of discourse connectives. In this study, we explore supervised machine-learning approaches to automatically identify discourse connectives in biomedical literature and compare them with simple lexical pattern matching-based approaches. Later we predict the class-wise sense of the discourse connectives. The main contributions are 1) we are the first group to identify discourse connectives in the biomedical domain; 2) we explore the use of domain-specific features in addition to the normal syntactic features used in machine learning and 3) we use domain adaptation techniques to leverage larger open-domain data sets and further improve the performance of the discourse connective identification.

Related Work

A great deal of work has been performed to explore methods for discourse parsing [100-102] and discourse identification in the open domain [103,104]. Pitler and Nenkova [105] explored supervised machine-learning approaches to identify explicit discourse connectives and disambiguate their sense in the PDTB.

In contrast, work on discourse parsing in the biomedical domain has been limited. BioNLP tasks have traditionally focused on sentence-level analysis and information extraction. Studies [106-108] have explored approaches to segment biomedical text into sections and topics. Szarvas et al. [109] created BioScope, a corpus annotated with negative and speculative keywords and their linguistic scope in biomedical text. Agarwal and Yu [110,111] subsequently developed a system to automatically identify negation and hedging cues and their scope in biomedical text.

The most closely related work is the development of an annotated corpus of discourse relations called the BioDRB [112,113], and studies on the sense disambiguation of discourse connectives [112]. Studies have also examined certainty [114] and future research direction in biomedical literature [115] using discourse structure. Other discourse aspects have been researched in the biomedical domain, such as the annotation of co-reference relations [90,116-118] and anaphora resolution [119].

We developed a preliminary CRF-based classifier to identify discourse connectives using the PDTB and BioDRB corpora. Then we expanded the classifier by exploring new features including syntactic and domain specific semantic features and novel domain adaptation techniques. We also explored supervised machine-learning techniques to identify the sense of connective and classified it into one of four categories: comparison, expansion, contingency and temporal.

Materials and Methods

Discourse Relations Corpora

The two annotated corpora we used in this study are the Penn Discourse TreeBank [91] (PDTB 2.0, http://www.seas.upenn.edu/~pdtb) and the Biomedical Discourse Relation Bank [112] (BioDRB, http://biodiscourserelation.org/). The PDTB annotations are done over 2,159 texts (over 1 million word tokens) from the *Wall Street Journal* (WSJ) articles collection of the Penn Treebank [120]. The Penn Treebank is an open domain large-scale annotated corpus of syntactic phrase structure that has been very widely used by researchers for data-driven parser development. The source *WSJ* articles have also been annotated for other kinds of linguistic information, including semantic roles [121] and coreference [122], among others. The PDTB was developed to further enrich the WSJ annotations at the level of discourse and provides annotations of explicit and implicit discourse relations their arguments, their senses, and the attributions of discourse relations and each of their two arguments.

The BioDRB is a corpus of discourse relations annotated over 24 full-text articles (~112,000 word tokens) taken from the GENIA corpus [123]. The GENIA articles were selected by querying the PubMed for "blood cells" and "transcription factors" and were considered representative of scientific articles in this domain by the GENIA research group [124]. Discourse relation annotations of the BioDRB largely follow the PDTB guidelines and, like the PDTB, include annotations of explicit and implicit discourse relations, their arguments and their semantics. Unlike the PDTB, however, the BioDRB

does not currently annotate attribution. An overall agreement of 85% was reported among annotators of BioDRB [113].

The PDTB and BioDRB contain annotations for 18,459 and 2,637 total explicit connectives, or 18.5 and 26.4 discourse connectives per 1,000 tokens, respectively. After connective stemming (e.g., "three days after" stemmed to "after") there are 100 unique explicit discourse connectives in the PDTB and 123 in the BioDRB.

Our analysis shows that 56% of the explicit discourse connectives in the BioDRB occur in the PDTB, including common connectives like *and*, *also*, *so*, and *however*. Thirty-three percent of the connectives in BioDRB comprise the class of "subordinators" like *followed by*, *in order to*, and *due to,* which are not annotated as connectives in the PDTB corpus (connectives in the PDTB are defined as belonging to three grammatical classes: subordinating conjunctions, coordinating conjunctions, and discourse adverbials). The final 11% of the connectives in the BioDRB consist of lexical items that do not occur in the PDTB texts and were therefore not classified as connectives. Examples of these include: *In outline, As a consequence,* and *In summary*.

Figure 7 below shows the frequency of the tokens in the BioDRB corpus and their frequency as connectives. From our analysis of the BioDRB data we found that 76% of the connectives had both discourse and non-discourse usage and 43.5% of the connectives occur only once in the entire corpus as connectives.
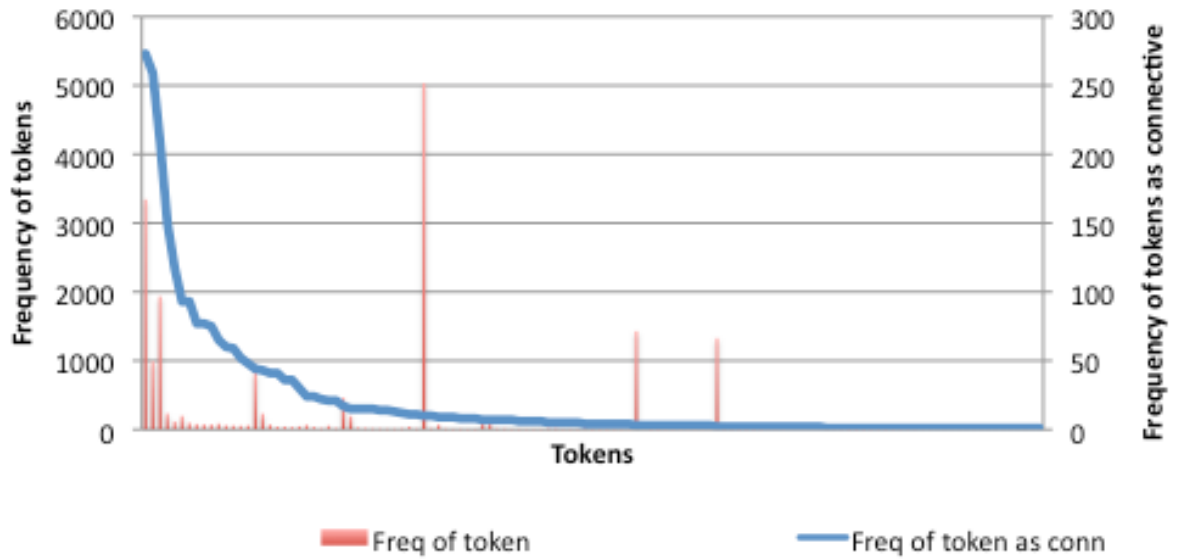
Figure 7: Frequency of the tokens in the BioDRB corpus and their frequency as connectives

Domain Adaptation Approaches

In order to compensate for the relatively small size of the BioDRB (~112K tokens) and to leverage the much larger open-domain PDTB (~1 million tokens), we explored domain adaptation approaches to build models trained on both corpora. In domain adaptation, the larger corpus is referred to as source domain (PDTB in this case) and the smaller one as the target domain (BioDRB in this case). In this study, we explored three supervised domain adaptation techniques:

Instance weighting combines the data from both corpora but assigns different weights to them during the training phase. The weights are usually inversely proportional to the size of the corpus to compensate for the larger number of training examples and to avoid

overfitting to the source domain. The classifier was then trained using this weighted training dataset.

Instance pruning actively removes misleading training instances. For example, if for training example *d*, we find different labels for *d* in the source and target domains, then we remove all such instances of *d* from the source domain training data. To apply instance pruning, we first trained a classifier on the target domain data (BioDRB) and then applied this classifier to the source domain data (PDTB). All the instances in the source domain that were incorrectly classified are pruned from the source training set (~ 1% of data was pruned). The final classifier was trained using this pruned source domain dataset.

Feature augmentation is a method in which additional metafeatures are added to indicate whether a specific feature came from the source or target dataset. For each training example, the feature vector is expanded to contain not only the original features, but also indicators representing the domain from which each feature was taken. This makes it possible for us to represent the effect of individual features in the source and target domain respectively, and for the machine-learning algorithms to distinguish between features important to the respective domains. The classifier is then trained on the combined dataset with the additional features. Consider the example, *"...industry is regulated by commodity futures ..."* in the source domain and *"...resulted in a small overlap in regulated mRNAs at 4 ..."* in the target domain. The word "regulated" is used as a verb in source domain whereas it is used as an adjective in target domain. In the feature vector for the word "regulated," the source-specific indicator linked to "verb" and

the target-specific indicator linked to "adjective" is set.

Supervised Machine Learning

The two supervised machine-learning approaches we explored were Conditional Random Fields (CRFs) and Support Vector Machines (SVMs). Our aim in using these two approaches was to explore whether it was more beneficial to cast the problem of identifying discourse connectives as a sequence-labeling task (with CRFs) or as a classification task (with SVMs).

CRFs are a probabilistic modeling framework [125] commonly used for sequence labeling problems. In our experiments, we treated documents as a sequence of words, and the classifier determined whether or not each word in the sequence was part of a connective. We built the CRF classifiers using the ABNER toolkit [126].

To test connective identification as a classification task, we built an SVM classifier using Weka. SVMs are a well-known statistical machine-learning algorithm and have shown very good performance in many classification tasks [80,81]. We used the SVM to classify each word in a sentence as either a discourse connective token or a non-discourse token. We also trained SVM and logistic regression-based classifier to predict the class-wise discourse sense of the connective and classify it into one of the four expansion, contingency, comparison or temporal categories.
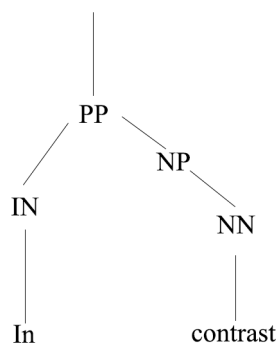
Figure 8: Sample parse tree

In addition to the default ABNER features, we evaluated syntactic and domain-specific learning features. We explored the syntactic features that have been shown to be important in previous studies [100,103,105], namely part-of-speech (POS) of the token, the label of the immediate parent of the token's POS in a parse tree, and the POS of the left sibling (the token to the left of the current word inside the innermost constituent). Figure 8 shows a sample parse tree. The tags IN and NN are used as the POS features; PP and NP are used as the label of the immediate parent of POS for the word tokens "in" and "contrast." The left sibling value is NONE assuming it is the start of the sentence. The syntactic features were obtained using the Charniak–Johnson parser trained in the biomedical domain. The parser was evaluated to have the best performance when tested on the GENIA corpus [127]. We also explored domain-specific features by using Metamap, the BANNER [128] gene tagger, and the LINNAEUS [129] species tagger to map text elements to the Unified Medical Language System (UMLS) [130] semantic types, and to identify named entities including gene and species.

Experiments and Systems

We developed several systems to evaluate (a) the complexity of connective identification in open domain, (a) the complexity of connective identification in biomedical domain, (c) the impact of different syntactic and domain-specific features for connective identification, (d) the impact of different domain adaptation models for connective identification, and (e) discourse connective sense identification. Since there are a total of 24 articles in the BioDRB, to simplify the task, we used 12-fold cross-validation rather than the common 10-fold so that an article (not a segment of it) was assigned as either a training or a testing article.

Complexity of discourse connective identification in open domain

In this experiment we built systems to measure the complexity by gradually increasing the training size from 0.24, 0.48, 0.7 to 1 million tokens on the PDTB corpus. We compare the performance of the systems using syntactic features generated from open domain Stanford parser and compare it with the features annotated by humans using the Penn Tree Bank [120]. Experiments were performed to ascertain the value of syntactic features with the open domain classifier. Starting with just the default ABNER feature set, each syntactic feature was considered independently, and then various combinations of features were evaluated.

Complexity of discourse connective identification in biomedical domain

In this experiment, we develop two heuristic baseline systems and compare their performance with our in-domain CRFs and SVM-based classifiers.

Baseline Systems

The first baseline system, *BaseLex*, uses a lexical heuristic, creating a lexicon by extracting the connectives annotated in the BioDRB corpus and then tagging all instances of these words in the text as connectives.

The second baseline system, *BaseLexPunct*, is a combination of the lexical heuristic from *BaseLex* and additional heuristics related to observed punctuation patterns associated with connectives. In particular, we observed that connectives were often either preceded or followed by a comma or appeared as the first word in the sentence. The system first identifies all connective terms from the lexicon in the text and then filters out the instances that do not match with the manually created punctuation heuristic.

Supervised Machine Learning Systems

The two baseline systems were compared against our supervised machine-learning systems: *In-domainSVM*, the SVM classifier, and the *In-domainCRF,* the CRFs-based classifier. Both the classifiers were trained and tested on BioDRB, using syntactic features discussed above.

Measuring the Impact of Semantic Features for Discourse Connective Identification

In this experiment, we evaluated the impact of different types of features; in particular we wished to determine the relative performance of syntactic versus domain-specific features. For this reason we built variants of the best performing classifier from the first experiment using different features, as follows: The *UMLS* classifier exclusively uses UMLS features extracted using Metamap; the *GeneSpecies* classifier exclusively uses the

gene and species categories extracted with BANNER and LINNAEUS as features. We then evaluate both of these classifiers after adding the features used in previous experiment, which we call *UMLS$^+$* and *GeneSpecies$^+$* respectively. Finally, we combined all of the features into a classifier, which we will call *Semantics$^+$*.

Systems to Measure Impact of Domain Adaptation for Discourse Connective Identification

In this experiment, we evaluated the impact of the domain-adaptation approaches described in the domain adaptation section, for which we compared several classifiers with and without domain adaptation. We used the classifier type and feature sets found to have the best performance in our previous experiments.

Baseline Systems

The following systems did not incorporate domain adaptation and were used as the baseline: the *In-domain* classifier, trained exclusively on the target domain; the *Cross-domain* classifier, trained on the source domain; and the *Unweighted* classifier, trained on the merged source and target domains.

Domain Adaptation Systems

To test the various domain adaptation techniques, we developed three classifiers: the *InstanceWeighting* classifier, where source domain data were given a weight 0.1 times that of target domain data[1]; the *InstancePruning* classifier and the *FeatAugment*

---

[1] The value of 0.1 was used as an approximation of the relative sizes of the datasets.

classifier, which were trained using the instance weighting, instance pruning and feature augmentation approaches respectively.

Combined Domain Adaptation Systems

The following systems incorporated combinations of the domain adaptation techniques: the *Weighted-Pruning* classifier, trained using a combination of instance weighting and instance pruning approaches; the *Weighted-FeatAugment* classifier, trained using a combination of instance weighting and feature augmentation approaches; the *Hybrid* classifier, trained using a combination of instance pruning and feature augmentation approaches; and finally, the *Weighted-Hybrid* classifier, trained using the combination of all three approaches. For the combined methods using instance weighting, the source weight was changed from 0.1 to 0.5 to reflect the effects of the other two adaptation methods.

Discourse Connective Sense Identification

In this experiment, we built two classifiers to identify the class-wise sense of the discourse connective as either expansion, contingency, comparison or temporal. We extract syntactic features of the connective as described previously and build a SVM and logistic regression-based classifier *SVMSense* and *LogisticSense* respectively to identify the sense of the connective.

Evaluation Metrics

All the classifiers (including the baseline classifiers) were run at the token level (i.e., the word level, marking each token in the evaluation corpus as either connective or not). The

classifiers trained in the open domain (PDTB) for connective identification and the classifiers trained for connective sense identification were evaluated using ten-fold cross-validation. Other classifiers trained using biomedical domain data were evaluated with twelve-fold cross-validation, except for the *Cross-Domain* classifier, which was trained on the source domain and evaluated on the target domain. For systems using the combination of the BioDRB and the PDTB, the training for each fold was always done on the entire PDTB with eleven-twelfths of the BioDRB and the evaluation done on the remaining BioDRB data. The standard evaluation metrics of recall, precision, and F1 score were used to measure the performance of all systems.

Results

Table 13 shows the precision, recall, and F1 score of the classifier trained on the 0.24, 0.48, 0.7, and 1 million tokens of the PDTB corpus. Automatically generated syntactic features reduce the performance as compared to human annotated syntactic features. As shown in Table 13, in the performance of 0.24 million token dataset, the F1 score decreases from 0.918 to 0.829, and this difference is statistically significant ($p < 0.05$, t-test, two tails). Table 14 shows the precision, recall, and F1 score of the ten-fold cross-validation results of this experiment. The Parent Category feature is the single most effective feature, resulting in an F1 score of 0.922, while the Left Sibling feature also improves performance. Alone, the POS tag feature has a very small effect on the performance. In experiments with combined syntactic features, the POS tag feature still seems to be the least effective. The performance of all features combined is the same as the performance of all features, except the POS tag. The Left Sibling feature improves

performance in combination with the Parent category feature (from 0.922 F1 score to 0.930), but the Parent category feature appears from these experiments to be the most valuable by far.

Table 13: The performance (average±Std) of open domain classifier for identifying discourse connectives on different data sizes

	Stanford parser - 0.24 million tokens	Gold syntax - 0.24 million tokens	Gold syntax - 0.48 million tokens	Gold syntax - 0.7 million tokens	Gold syntax - 1 million tokens
Precision	0.875 ± 0.018	0.935 ± 0.021	0.938 ± 0.012	0.944 ± 0.007	0.935 ± 0.016
Recall	0.789 ± 0.034	0.902 ± 0.026	0.923 ± 0.010	0.931 ± 0.008	0.925 ± 0.017
F1 score	0.829 ± 0.021	0.918 ± 0.015	0.931 ± 0.008	0.937 ± 0.004	0.930 ± 0.012

Table 14: Performance (average±Std) of open domain classifier with combinations of syntactic features

	Precision	Recall	F1 score
Default	0.876±0.021	0.809±0.012	0.841±0.013
Default+POS	0.878±0.018	0.810±0.024	0.842±0.016
Default+Parent	0.932±0.016	0.914±0.018	0.922±0.013
Default+LeftSib	0.908±0.020	0.875±0.018	0.891±0.012
Default+POS+LeftSib	0.897±0.032	0.875±0.020	0.890±0.014
Default+POS+Parent	0.934±0.016	0.917±0.020	0.926±0.013
Defaul+LeftSib+Parent	0.936±0.017	0.925±0.019	0.930±0.012
All Features	0.935±0.016	0.925±0.017	0.930±0.012

Default+POS: Default features of ABNER + POS feature

Default+Parent: Default features of ABNER + Parent Category feature

Default+LeftSib: Default features of ABNER + Left Sibling feature

Default+POS+LeftSib: Default features of ABNER + POS + Left Sibling

Default+POS+Parent: Default features of ABNER + POS + Parent Category

Default+LeftSib+Parent: Default features of ABNER + Left Sibling + Parent Category

Table 15 shows the performance evaluation of the in domain classifiers relative to the baseline systems, as described in previous sections. The heuristic baseline systems *BaseLex* and *BaseLexPunct* had an F1 score of 0.33 and 0.272, respectively. The supervised machine-learning classifiers *In-domainSVM* and *In-domainCRF* had an F1 score of 0.657 and 0.757, respectively. The supervised machine-learning methods clearly outperform the baseline methods. The CRF-based system had the best performance overall and was therefore chosen as the system to be adapted for subsequent experiments.

It is clear from the data in Table 15 that the addition of domain specific semantic features did not help improve classifier performance. The *In-domain* classifier, trained using only the syntactic features, had the best performance, F1 score 0.757, followed by the *Gene-Species* classifier, F1 0.753. While the difference between the two scores is not statistically significant, the additional features were clearly not providing any benefit. Therefore, in subsequent experiments with domain adaptation, we used only syntactic features to train classifiers. We can also see from the table that identification of non-discourse connectives had good performance in all systems.

Table 15: Task Complexity: performance (average±Std) of different classifiers for the task complexity measurement. Effect of learning features: performance (average±Std) of in-domain CRF classifiers trained with different learning features

	Classifier Type	Overall Performance (F1 score) (Precision, Recall) Discourse Connectives	Overall Performance (F1 score) (Precision, Recall) Non Discourse Connectives
TASK COMPLEXITY	**BaseLex**	0.330 ± 0.044 (0.198 ± 0.032, 1.000 ± 0.000)	0.948 ± 0.005 (1.000 ± 0.000, 0.901 ± 0.010)
	BaseLexPunct	0.272 ± 0.058 (0.165 ± 0.041, 0.790 ± 0.072)	0.946 ± 0.006 (0.994 ± 0.001, 0.901 ± 0.010)
	In-domainSVM	0.657± 0.061 (0.773 ± 0.066, 0.575 ± 0.07)	0.945 ± 0.002 (0.998 ± 0.001, 0.897 ± 0.004)
	In-domainCRF	**0.757± 0.059** (0.817 ± 0.058, 0.711± 0.086)	**0.994 ± 0.001** (0.992 ± 0.002, 0.996 ± 0.001)
EFFECT OF LEARNING FEATURES	**UMLS (UMLS Semantic features)**	0.681 ± 0.063 (0.786 ± 0.050, 0.606 ± 0.086)	0.993 ± 0.001 (0.990 ± 0.003, 0.996 ± 0.001)
	Gene-Species (Gene + Species features)	0.686 ± 0.058 (0.797 ± 0.050, 0.608 ± 0.082)	0.993 ± 0.001 (0.990 ± 0.002, 0.996 ± 0.001)
	UMLS[+] (Syntactic + UMLS Semantic features)	0.744 ± 0.061 (0.806 ± 0.051, 0.696 ± 0.087)	0.992 ± 0.001 (0.986 ± 0.003, 0.997 ± 0.001)
	Gene-Species[+] (Syntactic + Gene + Species features)	0.753± 0.052 (0.814 ± 0.045, 0.703 ± 0.075)	**0.994 ± 0.001** (0.992 ± 0.002, 0.996 ± 0.001)
	In-domain (Syntactic features)	**0.757± 0.059** (0.817 ± 0.058, 0.711± 0.086)	**0.994 ± 0.001** (0.992 ± 0.002, 0.996 ± 0.001)

	Semantics[+] (All features)	0.747 ± 0.059	**0.994 ± 0.001**
		(0.810 ± 0.048, 0.698 ± 0.086)	(0.992 ± 0.002, 0.996 ± 0.001)

Table 16 shows the performance of all CRF classifiers with the impact of different domain adaptation models, as described in the previous section. Among the simple domain adaptation techniques, the *InstanceWeighting* classifier had the best performance; with F1 score 0.730, compared to other individual domain adaptation-based classifiers *InstancePruning* and *FeatAugment*, for which F1 scores were 0.637 and 0.677, respectively.

None of the methods, however, performed better than the baseline *In-domainCRF* classifier. Some classifiers increased recall (*InstanceWeighting*) while others increased precision (*InstancePruning*). This indicates that systems combining multiple domain adaptation techniques may be more robust, and therefore produce better F1 scores. Results of these combinations are shown in the last four rows. The *Hybrid* classifier had the best performance among all classifiers, with an F1 score of 0.761. All the classifiers shown in Table 16 were statistically significant (t-test, $p < 0.05$) when compared with the *Cross-domain* classifier. The performance of classifiers trained using simple domain adaptation methods were statistically significant (t-test, $p < 0.05$) when compared with the classifiers trained using combined domain adaptation methods. In contrast, the classifiers trained using combined domain adaptation techniques did not produce statistically significant differences in their results.

The performance of the classifier for identifying the class-wise sense of the discourse connective is shown below in Table 17. The *SVMSense* classifier achieved the highest F1 score of 0.9.

Table 16: Performance (average±std) of different classifiers based on CRFs for identifying the discourse connectives using domain adaptation techniques for various categories

Classifier Type	Overall Performance (F1 score) (Precision, Recall) Discourse Connectives	Overall Performance (F1 score) (Precision, Recall) Non Discourse Conn
Cross-domain	0.592 ± 0.066 (0.834 ± 0.061, 0.461 ± 0.065)	0.992 ± 0.001 (0.986 ± 0.002, 0.998 ± 0.001)
UnWeighted	0.677 ± 0.071 (0.810 ± 0.061, 0.585 ± 0.085)	0.993 ± 0.001 (0.989 ± 0.002, 0.997 ± 0.001)
In-domain	0.757 ± 0.059 (0.816 ± 0.058, 0.711 ± 0.086)	**0.994 ± 0.001** (0.992 ± 0.002, 0.996 ± 0.001)
Weighted	0.730 ± 0.053 (0.805 ± 0.052, 0.671 ± 0.075)	0.993 ± 0.001 (0.991 ± 0.002, 0.996 ± 0.001)
Pruning	0.637 ± 0.076 (0.844 ± 0.070, 0.514±0.079)	0.993 ± 0.001 (0.987 ± 0.002, 0.998 ± 0.001)
FeatAugment	0.695 ± 0.056 (0.760 ± 0.048, 0.647 ± 0.090)	0.993 ± 0.001 (0.990 ± 0.002, 0.996 ± 0.001)
Weighted-Pruning	0.753 ± 0.057 (0.816 ± 0.051, 0.703 ± 0.083)	**0.994 ± 0.001** (0.992 ± 0.002, 0.996 ± 0.001)
Weighted-FeatAugment	0.757 ± 0.045 (0.809 ± 0.050, 0.716 ± 0.068)	**0.994 ± 0.001** (0.992 ± 0.002, 0.996 ± 0.001)
Hybrid	**0.761 ± 0.051** (0.813 ± 0.041, 0.719 ± 0.079)	**0.994 ± 0.001** (0.993 ± 0.002, 0.996 ± 0.001)
Weighted-Hybrid	0.757 ± 0.050 (0.807 ± 0.047, 0.717 ± 0.076)	**0.994 ± 0.001** (0.992 ± 0.002, 0.996 ± 0.001)

Table 17: Performance (average±std) of the classifiers for indentifying class-wise sense of the discourse connectives.

	Class	Precision	Recall	F1 Score
SVM Sense	**Comparison**	0.92 ± 0.05	0.86 ± 0.09	0.88 ± 0.05
	Contingency	0.90 ± 0.05	0.94 ± 0.03	0.92 ± 0.03
	Expansion	0.84 ± 0.07	0.89 ± 0.06	0.86 ± 0.05
	Temporal	0.94 ± 0.04	0.86 ± 0.06	0.89 ± 0.03
	Overall	0.90 ± 0.03	0.90 ± 0.03	0.90 ± 0.03
Logistic Sense	**Comparison**	0.91 ± 0.08	0.86 ± 0.09	0.88 ± 0.07
	Contingency	0.90 ± 0.06	0.94 ± 0.02	0.92 ± 0.03
	Expansion	0.85 ± 0.07	0.88 ± 0.07	0.86 ± 0.05
	Temporal	0.88 ± 0.09	0.86 ± 0.06	0.87 ± 0.04
	Overall	0.89 ± 0.04	0.89 ± 0.04	0.89 ± 0.04

Error Analysis

For error analysis, we focused on analyzing the CRF classifiers trained on syntactic features, since they showed the best performance. Error analysis revealed that most of the errors were due to the common problem of data sparseness. Specifically, most of the false negatives did not appear in the training set or appeared only once as a connective in the entire corpus. Therefore, we assessed classifier performance while taking these distributions into account. We first categorized the connectives based on their occurrence distributions in the PDTB and BioDRB corpora. There were three categories: connectives that were present and annotated in both corpora (BioDRB ∩ PDTB), present in both but annotated only in BioDRB (BioDRB ∉ PDTB), and finally, present and annotated only in BioDRB (BioDRB ∅ PDTB). We then investigated the performance of each domain-adapted classifier on each of the categories for tokens that appear at least once as connectives in the corpus. Table 18 shows the percentage of connectives identified by the

classifier in each category, the classifier's performance in that category for identifying the token as a discourse connective and non-discourse connective. We can observe that the weighting technique improved the performance across all three categories.

The impact of the frequency of the connectives on the performance of the classifier was analyzed. Figure 9 shows the graph of the number of connectives and the performance of the top-performing *Hybrid* classifier against the frequency of connectives in the BioDRB.

Table 18: Performance (F1 Score) of the classifiers for identifying the discourse connectives by their distribution in BioDRB

	BioDRB ∩ PDTB			BioDRB ∉ PDTB			BioDRB ∅ PDTB		
	% Of conns	Performance as DCONN	Performance as Non DCONN	% Of conns	Performance as DCONN	Performance as Non DCONN	% Of conns	Performance as DCONN	Performance as Non DCONN
Cross-domain	96.7%	0.62	0.92	3.3%	0.03	0.97	0%	0	0.86
Unweighted	84.3%	0.70	0.93	10.5%	0.21	0.98	5.2%	0.55	0.91
In-domain	74%	0.78	0.94	19.8%	0.65	0.98	6.2%	0.63	0.91
Weighted	75.7%	0.75	0.94	17%	0.51	0.98	7.3%	0.7	0.92
Pruning	93.4%	0.67	0.93	3.3%	0.08	0.97	3.3%	0.14	0.87
FeatAugment	72.8%	0.70	0.93	21%	0.58	0.98	6.2%	0.5	0.9
Weighted-Pruning	75.3%	0.77	0.94	18.5%	0.60	0.98	6.2%	0.63	0.91
Weighted-FeatAugment	72.6%	0.77	0.94	20.2%	0.67	0.98	7.2%	0.7	0.92
Hybrid	74.4%	0.78	0.94	19.5%	0.66	0.98	6.1%	0.67	0.92
Weighted-Hybrid	73.8%	0.78	0.94	20.2%	0.65	0.98	6%	0.67	0.92

In general, as the frequency of the connectives increased, the performance of the classifier for identifying those connectives increased. This is to be expected, as increased

training data resulted in improved classification. The decrease in performance for very frequent connectives can be explained by a small number of very frequent but very ambiguous connectives.
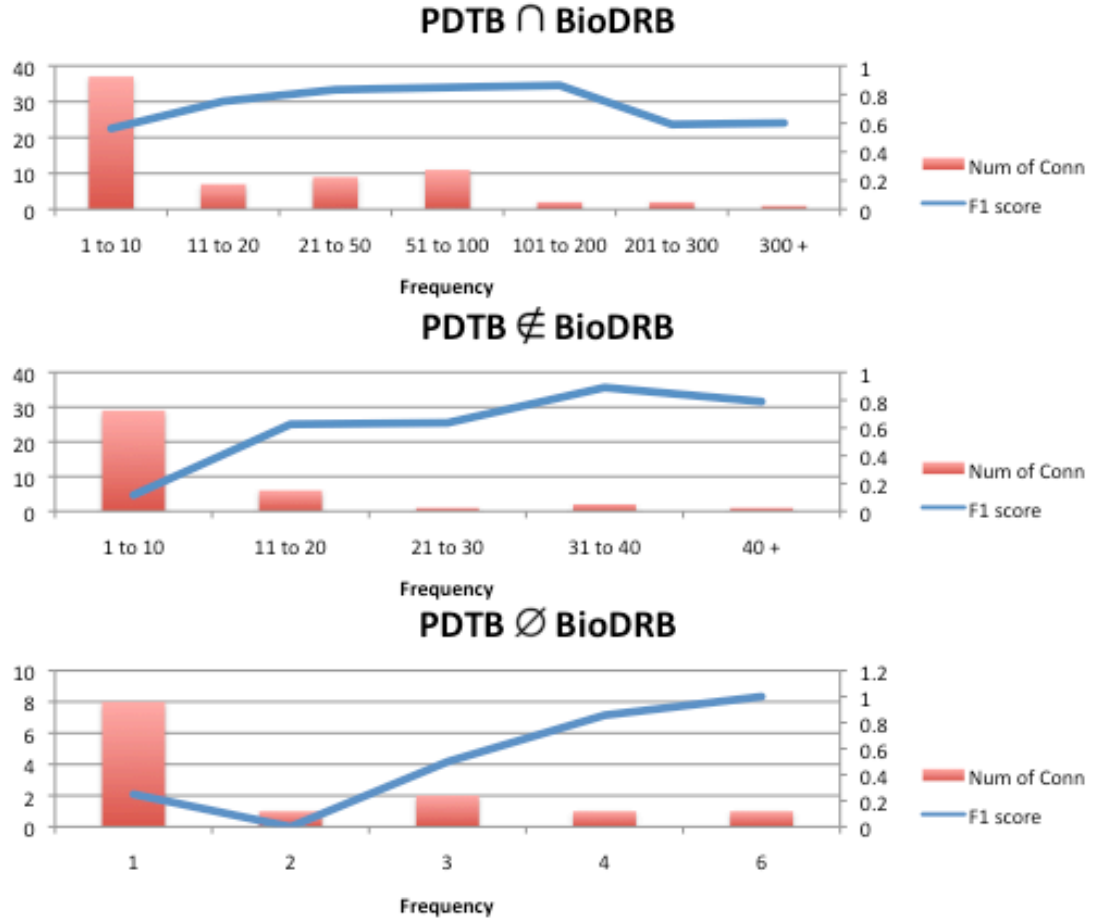
Figure 9: The graph of performance of Hybrid classifier over different distributions of the connectives

Table 19 below shows the five most common connective forms, the likelihood of each form occurring as a connective, and the F1 scores of the classifiers on these connectives. The *Hybrid* and *In-domain* classifiers performed better for frequent connectives (>100 occurrences as connectives). The connective *and* had an F1 score of approximately $0.7 \pm$ 0.04 for all the classifiers except for the *FeatAugment* classifier. For the connectives *by*,

to, and *after,* the table shows that as domain adaptation techniques were applied, the performance increased over *Cross-domain and Unweighted* classifiers.

Table 19: The top 5 connectives in BioDRB and their F1 scores on the classifiers

Classifiers	and (8.1%)	by (26.1%)	to (10.8%)	after (52.7%)	however (100%)
Cross-domain	0.72	0.03	0	0.06	0.98
Unweighted	0.73	0.3	0.07	0.67	**1**
In-domain	0.7	**0.64**	0.66	**0.74**	**1**
Weighted	0.7	0.52	0.53	0.72	**1**
Pruning	**0.74**	0.04	0	0.5	0.99
FeatAugment	0.26	0.55	0.58	0.65	0.99
Weighted-Pruning	**0.74**	0.57	0.6	0.73	**1**
Weighted-FeatAugment	0.67	0.59	**0.67**	0.72	**1**
Hybrid	0.67	**0.64**	**0.67**	0.72	**1**
Weighted-Hybrid	0.67	0.62	0.66	0.71	**1**

Our results show that a significant percentage of errors were introduced by two of the most frequent connectives, *by* and *to*, which were annotated in the BioDRB. We assessed the performance of all the classifiers discussed earlier after removing the connectives *by, to* and singleton connectives. The connective *by* sometimes appears as Noun Phrase (NP) and as Clause introduced as a subordinating conjunction (SBAR) few times. In either case it may or may not be a connective; therefore, the connectives *by* and *to* were removed. Experiments were then performed on this modified set of data. The results of the experiments are shown in Table 20. The overall performance of all the classifiers increased significantly except for *Weighted-FeatAugment*.

The performance of the *Cross-domain* classifier increased significantly to 0.673. This increase is due to the removal of connectives *by* and *to*, which are highly ambiguous and not annotated in the PDTB corpus. The *Unweighted* classifier had a performance of 0.766.

The *In-domain* classifier had an F1 score of 0.791. The performance of all classifiers using simple domain adaptation techniques increased with the *FeatAugment* classifier performing as well as *In-domain* with an F1 score of 0.791. *Weighted* and *Pruning* classifiers had an F1 score of 0.770 and 0.718, respectively.

The performance of the combined domain adaptation techniques also improved except for *Weighted-FeatAugment*, which performed only as well as the *Cross-domain*. The performance of *Weighted-Pruning*, *Weighted-FeatAugment*, *Hybrid,* and *Weighted-Hybrid* are 0.788, 0.690, 0.792, and 0.789 respectively. The *Hybrid* classifier still had the best performance.

Table 20: Performance (average±Std) of various classifiers for identifying the discourse connectives without singleton connectives and connectives *by* and *to*.

Classifier Type	Precision	Recall	F1 Score
Cross-domain	0.824 ± 0.057	0.570 ± 0.064	0.673 ± 0.058
Unweighted	0.826 ± 0.063	0.715 ± 0.068	0.766 ± 0.059
In-domain	0.846 ± 0.060	0.746 ± 0.074	0.791 ± 0.056
Weighted	0.825 ± 0.068	0.725 ± 0.062	0.770 ± 0.053
Pruning	**0.847 ± 0.060**	0.625 ± 0.072	0.718 ± 0.062
FeatAugment	0.835 ± 0.056	0.755 ± 0.072	0.791 ± 0.054
Weighted-Pruning	0.835 ± 0.061	0.750 ± 0.079	0.788 ± 0.060
Weighted-FeatAugment	0.824 ± 0.067	0.596 ± 0.073	0.690 ± 0.068
Hybrid	0.836 ± 0.058	**0.757 ± 0.074**	**0.792 ± 0.053**
Weighted-Hybrid	0.839 ± 0.055	0.749 ± 0.073	0.789 ± 0.053

Examples

In this section we manually examined the set of classified instances to evaluate the classifier that had the poorest performance (*Cross-domain*) and the classifier that had the best performance (*Hybrid*).

> ***Example 12*: One day after** <u>injection</u>, *the swelling of the ears was determined with a gauge (Hahn & Kolb, Stuttgart, Germany).* (Temporal: Succession)

> ***Example 13:* In view of the fact that** <u>NF-κB was also activated by anti-CD3/anti-CD28, IL-15 or mitogens in our experiments</u>, *it is most likely that the NF-κB pathway is also actively involved in the induction of IL-17 in RA PBMC.* (Cause: Justification)

Examples 12 and 13 show instances in which both the *Cross-domain* and *Hybrid* classifiers failed to identify the connectives. The connectives *One day after* and *In view of the fact that* appear only once in the entire BioDRB corpus and do not occur at all in the PDTB corpus. Since the classifiers encounter these connectives for the first time during testing, they fail to recognize them as discourse connectives. Example 6 suggests that collecting an exhaustive list of discourse connectives will not be feasible because any number could be inserted into the expression *One day after*.

> ***Example 14:* In order to** <u>explain this differential efficacy</u>, *several parameters were analyzed.* (Purpose: Goal)

> ***Example 15*: Due to** <u>the high level of sensitivity of nested RT-PCR</u>, *even low*

levels of illegitimate transcription in PBMNCs can cause false-positive results [2-5]. (Cause: Reason)

Examples 14 and 15 show instances that were correctly identified by the *Hybrid* classifier but were incorrectly classified by the *Cross-domain* classifier. Both *in order to* and *due to* are subordinators that were not annotated as connectives in the PDTB corpus but were annotated as connectives in the BioDRB corpus. Since the *Hybrid* classifier is trained for the biomedical domain using BioDRB, it identified them as connectives; however, the *Cross-domain* classifier failed to identify them as connectives as its training set did not contain such instances. In fact, the only connective in the BioDRB $\notin$ PDTB class that *Cross-domain* correctly classifies is *as an example*, which shares words with common connectives in PDTB.

> **Example 16**: *We considered this to be an appropriate positive control, as any cell that is detected using the immunobeads should express the EpCAM gene.* Tests of the single tumor cell and 100 PBMNC aliquots with EpCAM showed that it was **also** expressed to a sufficient level to enable detection of the tumor cell in 31/35 (89%) cases after 45 cycles of PCR amplification. (Conjunction)

> **Example 17:** *The accelerating effect of the mAb RIB5/2 was reproduced in two additional treatment experiments,* **and** this effect was observed despite a variable onset of AA in the PBS-treated animals (day 9 to 11). (Conjunction)

Examples 16 and 17 show instances that were correctly identified by the *Cross-domain* classifier but incorrectly classified by the *Hybrid* classifier. The connectives *also* and *and*

occur in both the PDTB and BioDRB corpora. Table 4 shows that the connective *and* had a better F1 score for the *Cross-domain* and *In-domain* classifiers compared with the *Hybrid* classifier. In addition, the *Hybrid* classifier incorporates feature augmentation, whose difficulty classifying *and* is clearly illustrated in Table 4.

Discussion

Automatic identification of discourse connectives is a challenging task. We performed various syntactic feature selection experiments (Table 14). The experiments show that the parent category feature had the highest single impact on the performance of the classifier compared with the other two syntactic features. The POS category had the least impact on the performance of the classifier, suggesting that POS information is largely redundant with the information about the word itself and not very useful. On the other hand, POS features may be valuable for new target domains, as they may help identify previously unseen connectives. The unadapted source-domain data may thus hurt adaptation performance by reducing the weight of POS features. Future work should focus on exploring POS features and related syntactic features and their benefit to the biomedical domain.

On BioDRB corpus, we found 76% of connectives to be ambiguous. As such, it is not surprising that using simple lexical features based on connective-matching system did not perform well (0.33 F1 score as shown in Table 15). Our results show that the supervised machine-learning approaches significantly outperformed the simpler pattern-matching approaches, yielding a maximum 0.757 F1 score.

We explored two different machine-learning models: SVM and CRF. We found that the

CRF model outperformed the SVM model, yielding 0.757 F1 score, 10% higher than that of the SVM model. Note that the performance of both systems was much lower than in the open domain (0.94 F1 score). For comparison, we trained and tested CRF models on the PDTB with the published feature set [105]. The classifier yielded similar results (0.937 F1 score), which demonstrated that our models are state-of-the-art.

Our results have shown that in-domain classifiers out-performed cross-domain classifiers. While the CRF-based in-domain classifier achieved the highest performance of 0.757 F1 score, the best cross-domain classifier yielded only 0.592 F1 score. The results demonstrate that the biomedical domain needs domain-specific models for discourse connective identification.

We explored different learning features. Similar to previous open-domain work [105], we found that syntactic features are important. In contrast, adding domain-specific semantic features (e.g., features based on UMLS) did not improve the performance. We speculate that the additional features may have introduced noise that is responsible for decreased performance.

Previous work has demonstrated that domain-adaption approaches can significantly improve the performance of tasks such as semantic role labeling [131]. In contrast, our experiments show that different domain adaptation methods have complementary effects on performance and can be combined for further improvement. Our new domain adaptation model *Hybrid*, which is a CRF model trained with a combination of instance pruning and feature augmentation domain adaptation techniques, outperformed all other models achieving an F1 score of 0.761. The *Hybrid* classifier used the advantages of both

the instance pruning (improved precision) and feature augmentation (improved recall) approaches thus increasing the overall performance.

Data sparseness is a very common problem in statistical NLP. In our study, 43.5% of the connective types appeared only once in the entire corpus. However, our results show that removal of these singleton connectives did not drastically affect system performance. This may be explained by the fact that the singleton connectives accounted for only a small portion (3%) of all discourse connective instances. This suggests that future work should focus on identifying improved features for disambiguating commonly occurring and highly ambiguous (such as *by* and *to*) connectives.

Predicting the sense of discourse relations is an important subtask of discourse parsing and previous studies have shown [132] that the task is fairly easy and have achieved very good performance (F1 score of 0.91). Similarly, our approach yielded a very high F1 score of 0.9 for a classifier trained on SVM to identify the class-wise sense of discourse connectives.

As we observed in our results, the *Crossdomain* model had a reduced performance due to different characteristics of the data. The model trained on BioDRB corpus might have a similar effect when applied on the medical text written by health-care professionals, as they will have different characteristics compared to that of the text in biomedical domain. Therefore, future work should focus on developing a gold standard annotated with discourse relations and building models specific to the medical text.

Conclusion and Future work

We have presented a method to automatically identify discourse connectives in biomedical text. This task is difficult and poses many challenges. The *Hybrid* classifier based on CRF with a combination of instance pruning and feature augmentation domain adaptation techniques had the best performance (F1 score 0.761) in the biomedical domain, while performance in open domain is still better (F1 score 0.93). We explored various supervised machine-learning based algorithms for automatically identifying explicit discourse connectives and evaluated different domain adaptation techniques to adapt models trained on the PDTB to the biomedical domain with various novel features. Although performance of *Hybrid* classifier is not statistically significant than *In-domain* classifier, leveraging the large corpus from another domain makes the classifier trained for biomedical domain more robust when the data are sparse. Future work will explore features to disambiguate the commonly occurring and confounding connectives like *by* and *to*. The SVM classifier to identify the sense of connective had a performance of 0.9 F1 score. Later we will extend this work to identify the arguments of explicit discourse connective, the next step toward developing a discourse parser. We will also explore techniques to identify the presence of implicit discourse relations in the text. Then we plan to develop a corpus annotated with discourse relations on medical text and apply similar techniques to identify discourse connectives and relations.

Chapter 4: Automation of Naranjo Scale Elements

A causal relation exists between the drug and adverse event in an adverse drug event. In this chapter, we discuss the Causality Inference Engine, which automates the Naranjo Causality Assessment Probability Scale that infers the causality between the drug and adverse event as shown in Table 1. We use a knowledge-based approach to automate the elements of the scale. In this research we automate 3 out of 10 elements of the Naranjo Scale due to constraints of data availability. The remaining elements of the Naranjo required longitudinal patient data and objective evidence from laboratory charts that were not available. We obtain information from various knowledge resources and apply knowledge-based and rule-based methods to automate the elements of the Naranjo Causality Assessment Probability Scale. A score is assigned to each of the element automated based on the Naranjo Scale.

There are many challenges in building such a system. As stated earlier, one of the challenges is the availability of the structured and longitudinal data of the patient. The second challenge is the construction of data to evaluate the tool. At present there is no gold standard that is available to conduct an evaluation to measure the effectiveness of the system. We have to build a gold standard of reports that are assessed and scored for the Naranjo elements by experts.

Therefore, the system at its current state automates 3 out of the 10 questions of the Naranjo Causality Assessment Probability Scale and the remaining elements can be automated by utilizing the structured data from the patients health records, discourse analysis as shown in examples 3 and 4 in chapter 3 and using temporal analyses systems.

We use *AE* and *medication* entities from the named entity recognizer module to automate these three elements and the other entities recognized could be utilized to automate the other elements of Naranjo. We also evaluate the tool on a small set of three randomly selected discharge summaries that contain adverse events related to "aspirin."

Related Work

Causality assessment is the evaluation of the likelihood that a particular treatment is the cause of an observed adverse event [133]. It assesses the relationship between a drug treatment and occurrence of an adverse event. It is an important component of pharmacovigilance, contributing to better evaluation of the risk-benefit profiles of medicines [134] and is an essential part of evaluating ADE reports in early warning systems and for regulatory purposes [135].

Causality assessment in pharmacovigilance may involve making a decision based on the information on the relationship between a drug exposure and suspected ADE from a single adverse event or suspected ADE report (or a series of reports). Early solutions to this problem were frequently in the form of standardized decision aids (SDAs) or algorithms [136]. Standardized decision aids are a group of causality assessment methods that have several features in common: they pose a series of predetermined questions; the question are usually answered "yes," "no," or sometimes "unknown/not applicable"; the answers to each of the questions has a preset weight, and they are combined in an explicit manner. The final weight score then is converted to a probabilistic statement regarding the strength of the causal relationship such as "possibly" or "probably" drug related.

These methods are explicit, fast, and increase reproducibility. Standardized causality assessment is now a routine procedure at pharmacovigilance centers around the world; it is aimed at decreasing ambiguity of the data and also plays a key role in data exchange and limits the drawing of erroneous conclusions [137].

Agbabiaka et al. [138] performed a systematic review of the causality assessment scales. They classified the articles into three main broad categories: expert judgment/global introspection, algorithms and probabilistic methods. They concluded that no method can be universally used to assess the causality of the ADEs.

Karch and Lasagna [139] initially designed a decision table algorithm to identify ADEs, poisonings, noncompliance, and excessive self-administration. Their algorithm estimates the certainty of an adverse event and suspected drug. Kramer et al. [140] proposed an algorithm that provides detailed criteria for ranking the probability of causation when ADE is suspected between a drug and a clinical manifestation. The algorithm provides a scoring system of six axes of decision strategy: previous general experience with the drug, alternative etiologic candidates, timing of events, drug levels and evidence of overdose, dechallenge, and rechallenge. The sum of the scores is ordinally partitioned to rate the candidate ADE as definite, probable, possible, or unlikely. Later Naranjo et al. [141] developed an ADE probability scale consisting of ten questions that are answered as "yes," "no," "un-known," or "inapplicable" to assess the causality of a drug in a variety of clinical situations using the conventional categories and definitions of "definite," "probable," "possible" and "doubtful." Scores assigned to each question ranged from −1 to +2. The event is assigned to a probability category based on the total

score. A total score of ≥9 is "definite," "probable" is 5–8, "possible" 1–4 and "doubtful" ≤0. This scale assesses the likelihood of an ADE associated with only one drug, not for adverse drug events resulting from interactions between two drugs [142]. The Naranjo Scale does not address the main points that are necessary in causality evaluation of potential drug interactions. Nevertheless, the adverse reaction scores obtained by using the Naranjo Scale correlate well with those derived with Kramer's algorithm, which does address those points [103].

Hutchinson and Lane [143] propose that the metrics of validity and reliability are not the optimal measures by which to assess causality assessment methods. Then they propose a set of six criteria by which ADE causality assessment methods should be assessed. They evaluated a number of algorithms, all of which failed to satisfy the criteria they proposed. They suggested that the problem is at a fundamental level: the lack of a clear understanding of the real nature of the causality assessment problem and, therefore, the failure to develop and use a coherent theoretical framework for its solution. They formulate a Bayesian equation that satisfies all of the requirements. Danan and Benichou [144] described the development of a new algorithm for scoring the confidence to assess the causal connection between a drug and outcome called Roussel Uclaf Causality Assessment Method (RUCAM). The method is relatively simple like Naranjo's but adds some sub-scores in each question so that the system is less rigid. They evaluated and validated their new method on drug-induced liver injuries. Four experts evaluated 400 cases of liver injury and had an agreement rate of 37% among four experts for identifying

drug-induced liver injury. Koh and Li [145] developed another algorithm by simplifying the algorithm in [140]. They found that their algorithm is congruent with Naranjo.

Studies have also explored a number of Bayesian causality assessment scores [146-148]. Lane et al. [149] developed a new Bayesian approach to identify adverse reaction called Bayesian Adverse Reaction Diagnostic Instrument (BARDI). BARDI is explicit in the information that is used and how each piece of information is weighted. It uses Bayesian statistics to combine factors coherently. The Bayesian method can include any relevant information and can consider multiple possible causes. The tool was tested in many studies and was shown to be working well and valid[150,151]. Lanctot and Naranjo [146] developed MacBARDI-Q&A by computerizing BARDI to facilitate clinical use. The tool consists of three components: a UI, a spread sheet with database access and a knowledge base organized by clinical manifestations with a set of finite drugs. Later Lanctot and Narnajo [147] compared BARDI with another algorithm that calculates the probability of an adverse event induced by drug called Adverse drug reaction Probability Scale (APS). They report a high correlation (rs = 0.45, p <0.001) between BARDI and APS to calculate the probability of causality between drug and the adverse event. However, BARDI better distinguished cases that were highly probable (n = 83; PsP > or = 0.75) or highly improbable (n = 30; PsP < or = 0.25), whereas the APS rated the majority of these cases in the midrange (n = 128; range of APS, 1 to 8.9). They concluded that although APS is an effective screening tool, BARDI can better discriminate drug from nondrug-induced cases better and may be more appropriate for serious cases of adverse drug reactions.

Other studies have explored techniques based on regression [152] and genetic algorithm [153] to assess the adverse event causality. In the current study, we automate the elements of Naranjo Causality Assessment Probability Scale using knowledge- and rule-based techniques to assess the causality of drug-induced adverse event.

Methods

We use knowledge- and rule- based approaches to automate the Naranjo elements shown in Table 1. We automate three elements of the Naranjo Causality Assessment Probability Scale as shown in Table 21 below.

Table 21: The elements of the Naranjo elements automated by the tool

Q1: Are there previous conclusive reports on this reaction?
Q2: Did the adverse events appear after the suspected drug was given?
Q9: Did the patient have a similar reaction to the same or similar drugs in any previous exposure?"

Given a report with the named entities recognized, we process them at sentence level and consider all possible drug-AE pairs that appear in the sentence to automate the elements of the Naranjo Scale.

Automation of Naranjo Elements

As described above, we automate three elements of Naranjo causality assessment scale.

Element One: "Are there previous conclusive reports on this reaction"

The first Naranjo element we automate checks if there are any reports of this reaction. For each adverse event and drug pair, we examine if that drug and adverse event pair

occur in any of the four sources, namely: FAERS report, package inserts, SIDER database and Ask a patient website.

The first source, the FAERS, are reports from the FDA AERS database containing information of adverse events and medication errors that are reported to the FDA by health-care providers and consumers. The second source, package inserts, provides additional information about the drug, such as the generic name of the drug, pharmacology, indications, usage, contraindication, warnings, adverse reactions, how it is supplied, and others. The third source, SIDER database, contains information on marketed drugs and their recorded adverse drug reactions. The information is extracted from public documents and package inserts. They also contain information regarding side effect frequency, drug and side effect classifications, as well as links to further information, such as drug–target relations. The fourth source, AskaPatient database, contains more than 4,000 chemically prepared prescription drugs approved by the FDA's Center for Drug Evaluation and Research. It also includes some popular biological drugs such as vaccines, blood and blood components, allergies, somatic cells, and gene therapy products.

Element Two: Did the adverse events appear after the suspected drug was given?

For the second element we automate checks if the AE appear after the drug was given. To accomplish this task, we examine if the reports contain keywords that indicate the presence of AE after drug was given such as "in association with" and "experienced" and examine if the AE and drug appear in a given word token window around the keyword.

We compiled a list of seven keywords empirically by examining the reports. We use these keywords as cues and find the named entities around it.

Element Nine: Did the patient have a similar reaction to the same or similar drugs in any previous exposure?

This element examines if patient experienced any reaction previously when exposed to drugs belonging to the same drug class. We use the classification of drugs defined in the drugbank to automatically find drugs that belong to the same class. Given a drug, we find all the drugs belonging to its class and examine the report to find a mention of those drugs. If we find a mention of a drug belonging to the same class, we use the rules developed to automate element two to see if the patient had similar reaction to the drugs belonging to the same class.

Evaluation

We apply the Naranjo automator tool that was developed on a set of discharge summaries and obtain the score for each of the summaries to measure the performance of the tool. The elements of the Naranjo are assessed at sentence level. For every sentence, we assess the causality by considering all possible combinations of drug-AE pairs that appear in the sentence.

Evaluation Data

We selected discharge summaries containing known drug and the adverse events caused by the drug, due to the lack of a gold standard annotated with the Naranjo elements. We randomly selected a very small set of three de-identified discharge summary notes that

had mention of "aspirin" in them from the Pittsburgh NLP repository [154], which contains a variety of de-identified clinical reports, including discharge summaries and progress notes. Two of the three discharge summaries contained known adverse events related to aspirin. There were a total of 189 sentences and 2,547 words tokens in the three reports.

Results

The Naranjo automator automated three questions and assigned scores on three EHR reports. The tool obtained a score of 0 for the report that did not have any aspirin related adverse event. The tool assigned a score of 2 for one report and 3 for the other report. Example 1 below shows the instance from the report that resulted in a score of 2 and Example 2 shows the instance from the report that resulted in a score of 3. In all the examples below, the adverse events are shown in **bold** and medication is underlined.

In example 1, the tool assigned a score of two as it identified the keywords "hypersensitive to" along with the drug and the adverse event around it. Hence, the tool assigned a score of 2 from question 2.

In the case of example 2, the tool identified "GI bleed" as an AE due to "aspirin" from package inserts and gave a score of 1 for question 1. For question 2, it assigned a score of 2 as the tool identified the keywords "related to" and the drug and the medication around it.

> ***Example 1***: `Apparently the patient was hypersensitive to` aspirin
> `and started` **`oozing blood`** `through his trach insertion site and`
> `pacemaker insertion site.`

>*Example 2*: `Doctor's assessment was that the patient had a small`
>`volume upper ` **`GI bleed`** ` probably related to her alcohol intake and`
>`aspirin` ` and Lovenox.`

Limitations, Conclusion and Future Work

We automated 3 out of the 10 elements of the Naranjo Causality Assessment Probability Scale. Since our tool was still in its initial stage and only three elements were automated, we applied our tool on a very small set of three discharge summaries. Although the tool performed well on assigning a score, the evaluation data were very small. The second limitation is that we automated only 3 elements out of 10 due to limitations of data availability. In the future, we would like to access the structural and longitudinal patient data and automate all the elements of the Naranjo Scale. We will further evaluate the tool on a larger set of data to measure its efficiency in finding the causality between the adverse event and the drug.

Chapter 5: Figure Evidence through Figure Associated Text Summarization

The third component of the ADEtector is the Figure Evidence Generator. This component retrieves figures related to ADE from the biomedical literature as evidence. To help users comprehend the figure a concise summary is generated for each figure retrieved by the Figure Evidence Generator component. We describe the generation of automatic summary in this chapter. We hypothesize that this tool will help researchers to identify the causes of ADE for translational research.

Millions of figures have been published in biomedical literature; these figures are a rich and important knowledge resource for scientists. Researchers need access to the figures and the knowledge they represent in order to validate research findings, as well as to generate new hypotheses. By themselves, these figures are sometimes incomprehensible to both humans and machines, and their associated texts are therefore essential for full comprehension. The associated text of a figure, however, is scattered throughout its full-text article and contains redundant information content.

In this chapter, we report the development and evaluation of unsupervised figure summarization systems, *FigSum+*, that automatically identify associated text, remove redundant information, and generate a text summary for every figure in an article. We created and published a dataset on figshare.com[2] that consists of 94 figures from 19

---

[2] http://figshare.com/articles/Figure_Associated_Text_Summarization_and_Evaluation/858903

biomedical articles. We evaluated our systems on two gold standards, the *FigSumGS1* dataset and the *FigSumGS2* dataset, that were derived from the published articles using different techniques in order to highlight the robustness and efficacy of our system.

We hypothesize that such figure summarization reduces information overload while maintaining information content and allows users to navigate content more efficiently. We conducted an intrinsic evaluation of the summarization systems, *FigSum+*, and measured its performance against both baseline and state-of-the-art supervised and unsupervised systems using precision, recall, F1, and ROUGE scores. The *FigSum+* systems achieved the best F1 score of 0.66 and ROUGE-1 score of 0.97. We found that the *FigSum+* systems surpassed both the baseline and state-of-the-art supervised and unsupervised systems in all measures except recall on the *FigSumGS2* dataset, where its score was comparable to the state-of-the-art supervised system. The overall superior performance of our approach suggests that these systems can be used to efficiently summarize figure content. We also designed a two-tiered extrinsic evaluation approach: (1) a self-reported scoring of the efficiency and usefulness of a figure summarization system and (2) a task-driven, randomized, controlled cognitive evaluation of full article comprehension. We evaluated two figure summarization systems *FigSum* and *FigSum+* within a novel user interface. *FigSum* generates figure summary by identifying sentences that are most similar to the figure caption and by classifying sentences to the structured categories of background, methods, results, and conclusion. *FigSum+* incorporates figure-referring paragraphs as figure summary. Our results show that the novel user

interface incorporating figure summary reduces the time spent on comprehending the figure content and improves the quality of responses for users to answer a biological question. Our results also show that users favor *FigSum+*, the system that associates a figure with paragraphs that describe them in the full-text article. The evaluations show that users like the figure summaries incorporated in novel user interfaces. The users are able to answer questions regarding the main content of the article using a subset of information contained in the full text article thereby reducing the users' information overload.

Introduction

Figures in biomedical publications are an essential part of biomedical knowledge. They help researchers by providing evidence to support their finding, report their discovery and generate new research hypotheses. Futrelle [155] indicated that nearly 50% of the article content in the biological domain is figure related. Therefore, we are developing an intelligent figure search engine (http://figuresearch.askhermes.org) to assist biological researchers. Currently our figure search engine is available as a SciVerse API and has indexed over 4 million full-text biomedical journal articles published by Elsevier.

Given the enormous number of figures in biomedical literature, a key aspect in building an effective figure search engine is the ability to automatically interpret figure content. A number of studies have examined various approaches for analysis and retrieval of

relevant figures from literature [156-167]. The ImageCLEF[3] competition for automatic annotation and retrieval of images from literature has been held annually for the last 10 years. But there is very limited research on extracting information related to figures from the full paper text in the biomedical domain [168].

Demner-Fushman [169] emphasized the importance of analyzing the text associated with the figure for its comprehension. Our evaluation study [170] showed that for a figure to be comprehended, it must be interpreted in conjunction with the text that refers to it in the article. We evaluated figure comprehension when a figure was presented (1) with its caption only, (2) with its caption along with the article title and abstract, and (3) with the article full text. The study found that presentation of the figure to biomedical researchers with just title and abstract failed to convey 30% of the information, compared to comprehension of the figure with the full text article. For example, Figure 10 shows a graph along with its caption. The caption information alone is not sufficient for complete comprehension of the figure. Hence, the associated text from the full-text of the article is required to completely understand the figures [171]. However, the associated text can be scattered throughout the full-text article and, moreover, can be redundant [170].

We therefore developed a figure summarization system called *FigSum* [168] that automatically generates a summary for every figure by extracting summary sentences from a full-text article based on word-level similarities between sentences and figure caption. A pilot evaluation showed biologists like the summaries generated by our pilot

---

[3] http://www.imageclef.org/

FigSum system [172]. Such figure summarization systems can provide users with a succinct figure summary who would otherwise have to spend time navigating through the full-text article. Figure 11 shows the summary generated by our summarization system for the figure shown in Figure 10. The summary helps to better understand the figure. The summarization system also has the potential of improving figure retrieval and mining knowledge from figures.

BRCA1 by terciles	N	Median (95% CI)	P
Low tercile	14	84 (---)	0.009
Intermediate tercile	14	14 (9.31-18.69)	
High tercile	13	36 (17.66-54.34)	
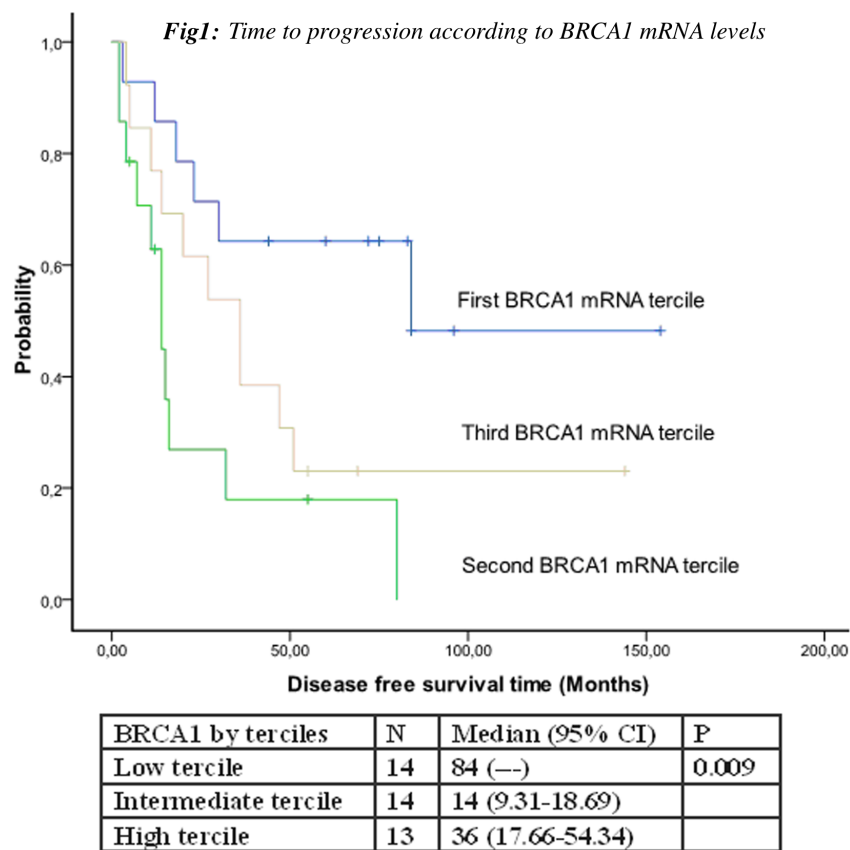
Figure 10: A sample figure with its caption. Figure1 appearing in article [173]

There were no differences in patient characteristics according to BRCA1 mRNA levels. Only one patient attained a pathological complete response (in tumor but not in axilla); it is thus impossible to correlate response with BRCA1 mRNA levels. Low levels of BRCA1 mRNA were associated with better TTP (84 months versus 14 and 36 months; p=0.009) (Fig 1). Median OS was not reached in patients with low levels of BRCA1 mRNA, while it was 21 months for those with intermediate and 50 months for those with high levels (p=0.03).

Figure 11: The summary generated by our system for figure shown in Figure 10

In the current study, we explore and evaluate a number of different summarization approaches, which we refer to as *FigSum+* systems. Specifically, in the first approach, we aggregate the sentences associated with a figure and remove any redundant sentences. In the second approach, we generate a figure summary by identifying the most relevant paragraph associated with the figure. In our third approach, we rank sentences based on content centroid. We perform intrinsic evaluation of these summarization approaches and report their performance.

We also build a user interfaces incorporating the best performing *FigSum+* system and *FigSum* system and conduct two extrinsic evaluations to measure the effectiveness of the summary for a specific task [174]. The first extrinsic evaluation measures the efficiency of the generated summary in summarizing figure content. The second extrinsic evaluation is a cognitive evaluation to measure the effectiveness of the summaries in allowing biomedical researchers to comprehend and answer questions related to the main content of the article. Our study suggests that summarization of figures using NLP techniques can help in comprehension.

Related Work

Summarization is one of the most extensively studied fields in natural language processing (NLP). The summarization approaches can be broadly classified as extractive or abstractive [175,176]. Extractive approaches extract and concatenate sentences from the text corpus to construct a summary, whereas abstractive summarization relies on natural language generation approaches that build new sentences representing the content of a text corpus to be summarized. In this work, we focus on the task of extractive summarization based on the text associated with a biomedical figure. The following sections review related work in open domain text summarization, text summarization in the biomedical domain, and figure summarization.

Open Domain Summarization

Extractive summarization identifies sentences that subsume the key points of a whole document (or collection of documents). One straightforward method is to select sentences based on word frequencies in the document. Very early work on summarization by Luhn, [177] proposed a simple idea based on the intuition that words occurring frequently in a document tend to describe the main topic. The approach selected sentences that contained many high-frequency words. Later studies improved this strategy by adding weight to words, using techniques such as log-likelihood and others [178-182]. For example, Brunn et al. [178] used syntactic parsing to identify important words for summarization. Approaches that identify summary sentences based on location were also developed. For

example, Nakov et al. [183] used citance (text that surrounds a citation reference) to summarize a document.

In other early work for summarization, Edmundson [184] applied a linear function that combines different factors, including resemblance to the title, indicative context cues (e.g., *in summary*), keywords, and sentence location. Myaeng and Jang [185] extended this work by adding centrality of the sentence to the document to select summary sentences.

Later studies explored various information retrieval (IR) techniques, such as the TF $\times$ IDF weighting scheme, [179,186-188] which alleviates the negative impact of overweighting of some common words, and latent semantic analysis, [189,190] which derives an implicit representation of text semantics based on observed word co-occurrences for summarization. For instance, Hovy and Lin [179] developed SUMMARIST, which integrates IR approaches, topic signatures (words that are highly descriptive of a document), dictionaries, and semantic knowledge derived from WordNet [191] to generate summary. Inspired by link analyses and page rank algorithms for Web document retrieval, Mihalcea et al. [192] and Erkan et al. [193] applied a graph-based ranking method to select important sentences based on the graph derived from words and sentences. Radev et al. [194] developed a MEAD summarizer that generates summaries based on a cluster centroid calculated by TF $\times$ IDF word similarity.

Studies also explored supervised machine-learning approaches for summarization [180, 195-198]. Kupiec et al. [180] developed a Naïve Bayes classifier incorporating contextual

and document features to select summary sentences. Wang et al. [195] and Hirao et al. [196] ranked sentences using a support vector machine classifier to generate summaries. Leskovec et al. [198] built semantic graphs to extract subject–object–predicate triplets from sentences and then trained a support vector machine classifier to extract salient sentence triplets for summarization.

Biomedical Summarization

Open domain summarization approaches are based on similarity and term occurrence approaches and have not shown to be the optimal choice for biomedical text due to domain-specific characteristics. For example, biomedical named entities (e.g., gene, protein, and chemical names) are frequently multiple words. Therefore, summarization systems are built upon biomedical knowledge resources, including the Medical Subject Headings (MeSH), the Unified Medical Language System (UMLS), and the Gene Ontology (GO) project.

Chiang et al. [199] developed GeneLibrarian, which generates a viewgraph of genes related to the input query based on GO similarity. The system also generates a summary of a gene by selecting sentences based on term occurrences. Ling et al. [200] developed approaches to automatically generate a structured gene summary by first retrieving gene-related documents and then extracting sentences containing factual information about the target gene. Jin et al. [201] developed a query-based gene summarization system that integrates the page rank algorithm, sentence similarity, and the function of the gene represented by GO.

Many studies have focused on summarizing content in biomedical text using semantic resources. Bhattacharya et al. [202] proposed a method to compute similarities between the MeSH terms assigned to an article and every sentence in the article and then return the top N-ranked sentences as summary sentences. Plaza [203] generated summaries based on sentence location. Reeve et al. [204] developed the BioChain system using the concept chaining technique. The technique links semantically related concepts in text using UMLS [205] and sentences with strong chains (where strength is based on the number of concepts) are used to form the summary. Fiszman et al. [206] applied handcrafted transformation rules to the output of SemRep[4] to summarize content. SemRep extracts biomedical concepts and relations relevant to a given query from MEDLINE records using MetaMap[5], which maps free text to UMLS concepts for concept extraction and the UMLS semantic network for relations between concepts. This work was modified by Workman et al. [207] to generate summaries relevant to the genetic etiology of a disease from biomedical literature to support the genetic database curation. Workman and Hurdle [208] applied SemRep to citations obtained from PubMed. They analyzed the output using statistical methods to automatically identify salient data in bibliographic text and generate summaries for bibliographic-based data. Shang et al. [209] extended the work of Fiszman et al. [206] to develop a multi-document summarizer for a given biomedical concept. Concepts and relations in sentences are extracted using SemRep. The sentences that contain high-frequency relations are then extracted as a

[4] http://semrep.nlm.nih.gov/

[5] http://metamap.nlm.nih.gov/

summary. Other studies [210,211] have explored knowledge from the UMLS to construct a graph and then select summary sentences based on node clustering.

Biomedical Figure Summarization and User Interfacing

Futrelle [212] proposed the idea of diagram summarization. He described the challenges related to summarizing figures and emphasized the importance of captions and referring text. Bhatia and Mitra [213] applied a supervised approach to summarize document objects such as figures, tables, and algorithms on a set of 290 document elements. Wu and Carberry [214] identified relevant paragraphs for images in news domain articles.

We developed a pilot summarization system, *FigSum* [168] for the biomedical domain. *FigSum* first classifies sentences into the introduction, methods, results, and discussion categories using a supervised machine-learning classifier [215]. Each sentence is then scored based on its TF × IDF weighted cosine similarity with the figure caption and the article's central theme. The top-scoring sentence in each category is included in the summary. The *FigSum* system is integrated into our larger figure search system (http://figuresearch.askhermes.org). An online survey revealed that 65.2% of participants found that *FigSum* summaries improved figure comprehension [172]. The current study explores additional figure summarization methods and performs an intrinsic and extrinsic evaluation of these approaches. In this dissertation we compare the performance of our approaches with supervised and unsupervised summarization approaches.

We performed [216] an evaluation of the novel user interface built to allow bioscience to efficiently access all the figures in an article and emphasize the most important figure.

All the figures that appear in the article were shown as thumbnails, and the most important figure is enlarged. The evaluation showed that 92% of the bioscience researchers preferred the user interface. Figure 12 shows a snapshot of this user interface that incorporates the *FigSum* summary. The interface is divided into two panes. The left pane presents the article title, authors, journal information, abstract, and thumbnails of all the figures. When the user selects a figure thumbnail, the right pane displays the enlarged figure along with its caption and summary.
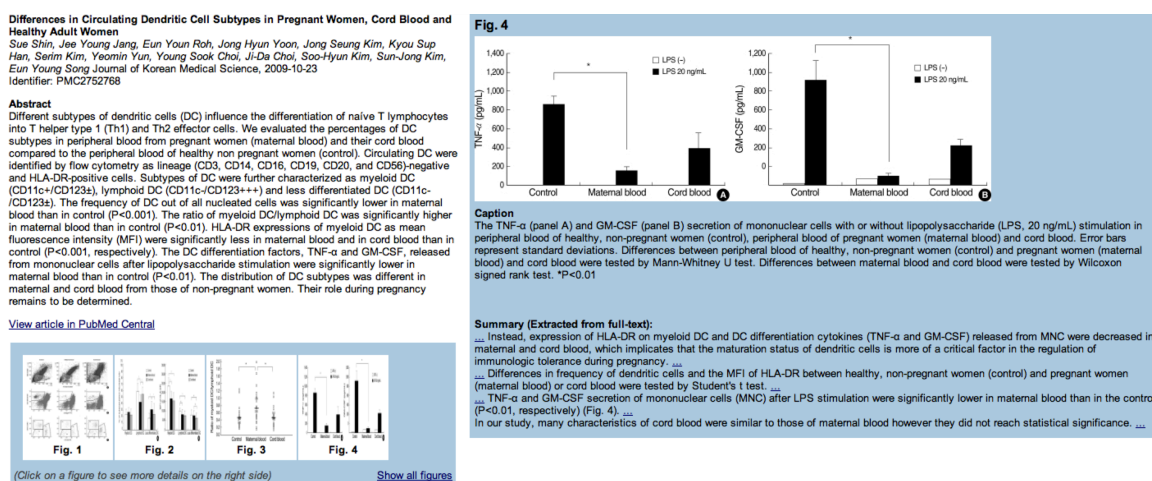
Figure 12: Snapshot of the interface incorporating *FigSum* system.

Evaluation

Evaluation is important for all NLP tasks. Mani [217] discussed various summarization evaluation criteria, including coherence, informativeness, relative utility, and relevance of the summary. Other studies have used evaluation methods that include word similarity measures (cosine similarity) [218], overlap of a sequence of words that includes n-grams (sequences of n-word tokens) and longest common subsequence [219,220], and the Bleu

[221] machine translation evaluation measure [222] for summarization. The Document Understanding Conference adopted the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) package for content-based evaluation [223]. Among different summarization evaluation metrics [224,225], the ROUGE score is widely used and is calculated based on co-occurrence between the gold standard and the summary generated. All aforementioned evaluation approaches can be classified as intrinsic evaluation as the summary generated by the system is evaluated against a gold standard summary generated by human.

A number of studies [226,227] have discussed extrinsic task-driven evaluation of summarization that measure the usefulness of summaries for specific tasks such as question answering and comprehension [228,229], information retrieval [230-232] and document categorization [233,234]. Studies have [235,236] evaluated summarization systems on the user's ability to find answers quickly while satisfying information needs. Murray et al. [237] performed an extrinsic evaluation of speech summarization for understanding how decisions were made in a decision audit task.

In biomedical domain, Fizman et al. [238] evaluated the summaries and determined its usefulness in helping clinicians to provide quality patient care. Yang et al. [239] performed a hybrid evaluation, partway between intrinsic and extrinsic system evaluation, to suggest the usefulness of a gene summarization system for a micro array analysis.

In the present study, we perform an extrinsic evaluation of the best performing version of *FigSum+*, the system that extracts paragraphs that contain figure-referring sentences. The evaluation focuses on measuring the effectiveness of the systems to summarize figures and improve article comprehension while reducing user information seeking overload.

Methods

Intrinsic Evaluation

Features used for Summarization

We explored a number of features to build figure summarization systems.

1) Similarity and IR based features

    a) Caption similarity feature – the cosine similarity value between each of the candidate sentences in the full text and the figure caption.

    b) Title similarity feature – the cosine similarity between each of the candidate sentences in the full text and the article title.

    c) Reference sentence similarity feature – the cosine similarity between each of the candidate sentences and the sentences referring to the figure.

    d) TFIDF feature – the text association between each of the candidate sentences in the full text and the figure caption is computed by calculating the TF × IDF vector for every candidate sentence and figure caption. A score is calculated as the cosine similarity of the TF × IDF vectors of candidate sentences and the figure caption.

2) Reference Features

a) Figure reference sentence feature – this feature represents if the sentence contains a figure mention (i.e., a sentence that incorporates figure reference cues such as *Fig. X*).

b) Figure reference paragraph feature – this feature represents if the sentence belongs to a paragraph that has a figure mention.

3) Hybrid feature – we first identify paragraphs in the full text article that contains figure referring sentences. We apply MEAD [194], a centroid-based text summarizer as described earlier on these sentences that are a part of the figure-referring paragraphs. The *n* top scoring sentences are selected as summary sentences.

4) Position

a) Distance from start feature - the position of the sentence from the start of the article.

b) Distance from end feature - the position of the sentence from the end of the article.

c) Distance from reference sentence feature – this is a binary feature that indicates if the candidate sentence is within 10 sentences of the reference sentence.

5) Sentence length feature – the length of the sentence.

6) Cue words and phrase feature – authors of articles use certain cue words and phrases to describe document elements such as figures, as discussed in [213]. We use the list of 140

cue words and phrases listed in [213] and add the presence or absence of these cue words in the sentence as a binary feature.

Figure Summarization Systems

In this section we describe our unsupervised *FigSum+* methods and other unsupervised and supervised systems we built for comparison with our system, *FigSum+*.

FigSum+ Systems

Figure 13 shows the general pipeline of the *FigSum+* systems. Given a full text article, the Text Extractor module extracts individual sentences from the article. If the article is in XML file format, an XML parser module will process the text to extract sentences from the XML file. If the article is in PDF format, the PDF-to-text converter (PDFTextStream[6]) tool extracts the text from the PDF document and then we split the text into individual sentences using an in-house sentence splitter that splits sentences by determining sentence boundaries such as periods. The figure summarization module utilizes various unsupervised techniques described below to summarize figures in the article and generate a summary for each figure.

We describe five different implementations of *FigSum+* systems based on the technique used in the figure summarization module. Each implementation of the *FigSum+* system differs by including one of the following five figure summarization modules: 1a, 1b, 2a, 2b, or 3.

---

[6] http://snowtide.com

(1) IR-based approaches: we explore two IR-based approaches for summarization.

    (a) *Similarity* – we compute the value of the caption similarity feature and present the top scoring sentences as the summary for the figure.

    (b) *TFIDF* – we compute the value of the TFIDF feature and present the top-scoring sentences as the summary for the figure..

(2) Surface-cue approaches: We identify summary content using surface cues.

    (a) *SurfaceCue* – extracts all figure referring sentences in the full text and presents them as the summary for the figure.

    (b) *Paragraph* – extracts all paragraphs containing figure referring sentences and they are presented as a summary of the figure.

(3) *Hybrid* – we compute the value of the hybrid feature and present the top scoring sentences as the figure summary.
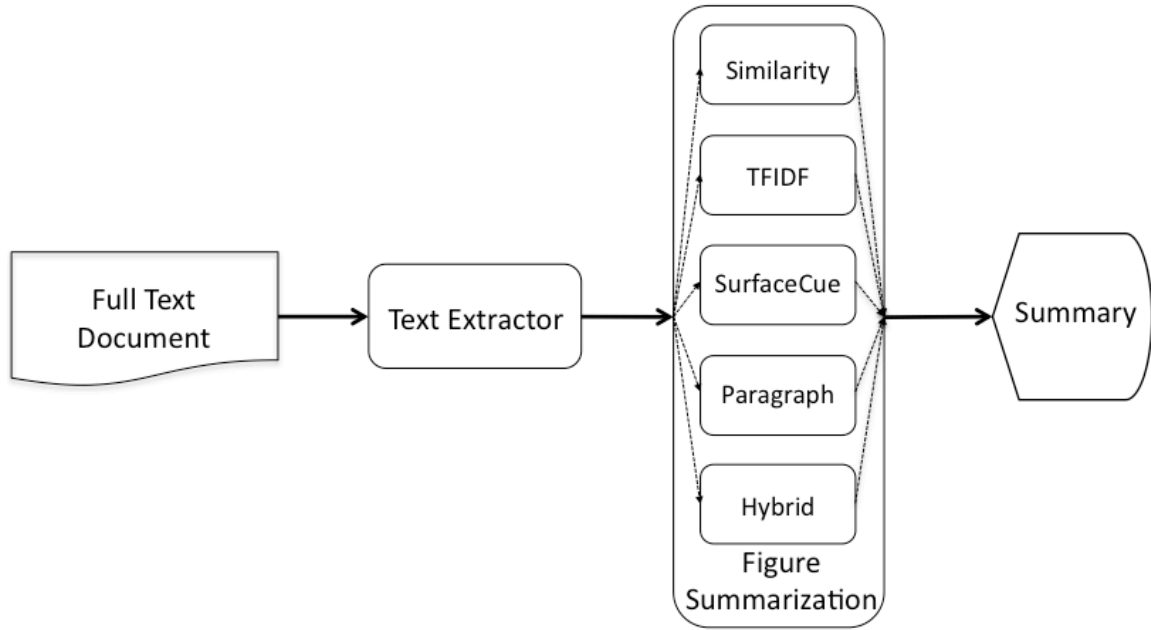
Figure 13: The general pipeline of the figure summarization systems. Each implementation of the *FigSum+* system differs by including only one of the five modules shown in the Figure Summarization component above: Similarity, TFIDF, SurfaceCue, Paragraph, or Hybrid.

Unsupervised Baseline Systems

We built three baseline unsupervised systems to compare the performance of *FigSum+* systems: *RandomSent*, *RandomPara,* and *MEAD*. The *RandomSent* system randomly selects *n* sentences from the article as the summary for the figure. The *RandomPara* system randomly selects *n* paragraphs and extracts the first sentence of all the randomly selected paragraphs as the summary for the figure. For the last baseline system, *MEAD,* we applied the centroid summarizer MEAD to the entire full text article and select *n* top scoring sentences as the summary for each figure.

Supervised Baseline Systems

In *FigSum+*, we use five features, namely: caption similarity feature, TFIDF feature, figure reference sentence feature, figure reference paragraph feature and hybrid feature.

We explored each of these individually and trained baseline supervised machine-learning models to generate figure summaries using each of these features. Each individual feature was used with both a naive bayes (NB) and Support Vector Machine (SVM) classifier, thus resulting in 10 baseline supervised systems: *NBSimilarity*, *NBTFIDF*, *NBSurfaceCue*, *NBParagraph*, *NBHybrid*, *SVMSimilarity*, *SVMTFIDF*, *SVMSurfaceCue*, *SVMParagraph* and *SVMHybrid*.

Unsupervised State-of-the-Art System

We also implemented the state-of-the-art unsupervised system, *FigSum*, which summarizes the figure as described earlier, for performance comparison with *FigSum+*.

Supervised State-of-the-Art System

We implemented the state-of-the-art system described in [213] by building two systems, *NBSOTA* and *SVMSOTA*, using the NB and SVM techniques respectively, with the features described in [213]. The features used are: figure reference sentence, figure reference paragraph, caption similarity, reference sentence similarity, distance from reference sentence and cue words.

We then extended the state-of-the-art system and build two more systems, *NBSOTA+* and *SVMSOTA+* using NB and SVM respectively, that incorporate all the features described above.

Evaluation Metrics

We calculate the microaveraged (as datasets are of different sizes) recall (R), precision (P), and F1 (F) scores to evaluate the summaries generated by each of the *FigSum+* implementations described in the figure summarization systems section. Recall is defined as the ratio of the number of sentences correctly identified by the system to the total number of sentences in the gold standard; precision is defined as the ratio of the number of sentences correctly identified by the system to the total number of sentences identified by the system; and the F1 score is the harmonic mean of recall and precision:

$$Recall = \frac{\# \ of \ sentences \ correctly \ identif \ ie \ dby \ the \ system}{Total \ \# \ of \ sentences \ in \ the \ gold \ standard} \quad (1)$$

$$Precision = \frac{\# \ of \ sentences \ correctly \ identif \ ie \ dby \ the \ system}{Total \ \# \ of \ sentences \ identif \ ie \ dby \ the \ system} \quad (2)$$

$$F1 \ Score = \frac{(2 \times Precision \times Recall)}{(Precision + Recall)} \quad (3)$$

We also compute the ROUGE score using the parameters established by DUC 2007 [240]. Eq (4) gives the formula to calculate ROUGE-N, where *n* stands for the length of the n-gram, *gram$_n$,* and *Count$_{match}$(gram$_n$)* is the maximum number of n-grams co-occurring in a candidate summary and a set of reference summaries. For every sentence in the summary generated by the *FigSum+* implementation, we calculate the ROUGE score against every sentence in the gold standard using the formula in Eq (4) and retain the best scores. Then we calculate the average of the best ROUGE score sentences for

every figure: ROUGE-1 (R1) compares summaries based on the co-occurrence of unigrams (single words), ROUGE-2 (R2) compares summaries based on the co-occurrence of bigrams (two consecutive words), and ROUGE-SU4 (RSU4) compares summaries based on the co-occurrence of skip bigrams with a maximum gap length of four [223].

$$ROUGE - N = \frac{\sum\limits_{S \in \{Ref\,erenceSummaries\}} \sum\limits_{gram_n \in S} Count_{match}(gram_n)}{\sum\limits_{S \in \{Ref\,erenceSummaries\}} \sum\limits_{gram_n \in S} Count(gram_n)} \quad (4)$$

Evaluation Data

We evaluate our *FigSum+* the approaches on a set of 19 full-text biomedical articles. Nine articles were randomly selected from our BioDRB corpus, a collection of 24 GENIA full-text articles fully annotated by us for discourse connectives and relations [132]. Four biologists with expertise in the biology domain each selected either two or three additional articles from various biomedical journals, for a total of 10 additional articles. The combined dataset of 19 articles comprises 94 figures and is made publicly available on figshare.com. The five *FigSum+* implementations are evaluated against the following two gold standards developed on these full-text articles; we selected two gold standards built using different approaches to show the robustness and efficacy of the five different techniques to figure summarization:

a) *FigSumGS1* – a gold standard of 94 figures from 19 articles from various biomedical journals was created as follows: four biologists ($B_1 - B_4$) read two papers each, for a

sub-total of 8 articles, and then selected sentences within each article that summarized figure content. In addition, two ($B_1 - B_2$) of the four biologists, read and selected sentences from 11 additional articles, thus yielding a total of 19 articles in the gold standard. The two biologists (B_1 and B_2) identified 303 and 383 sentences, respectively. They had an inter-annotator agreement (IAA) of 0.68 Cohen's κ value on the subset of 11 articles, which indicates a fair agreement between the annotators. The gold standard consists of a total of 678 sentences from 19 articles with a microaverage of 7.21 sentences per figure and a macroaverage of 7.73 sentences per figure.

b) *FigSumGS2* – a second gold standard consisting a subset of 17 articles from the 19 articles collected in (a) was created using the guideline that was developed to evaluate the *FigSum* system [168]. Seven annotators with advanced degrees (MS and above) selected three to four sentences that best described the background of the figure, the methods used to generate the figure, the outcome of the figure, and the conclusion inferred from the figure on the subset of 17 articles consisting of 84 figures; this subset was chosen from the 19 articles due to constraints of manual annotation. Hence, for each figure, a summary consisting of 12 to 16 sentences was obtained. All seven annotators together identified 869 unique sentences from the 17 articles with a microaverage of 10.34 unique sentences per figure and a macroaverage of 10.44 unique sentences per figure.

Table 22 shows the number of sentences and figures that appear in each article, the average number of unique sentences selected per figure and the total number of sentences annotated for both gold standards.

Table 22: Statistics of FigSumGS1 and FigSumGS2 gold standards

Article	# of sents	# of figs	FigSumGS1		FigSumGS2	
			Avg # of unique sents per fig	# of sents annotated	Avg # of unique sents per fig	# of sents annotated
1	190	3	5.0	15	11.7	35
2	144	3	18.0	54	11.7	35
3	173	7	5.0	35	8.0	56
4	160	5	8.6	43	10.2	51
5	172	4	12.8	51	10.5	42
6	140	5	8.4	42	10.8	54
7	281	9	7.8	70	11.8	106
8	137	9	4.7	42	6.3	57
9	142	5	6.2	31	11.2	56
10	87	5	6.4	32	8.4	42
11	162	6	6.0	36	9.7	58
12	34	2	7.5	15	6.0	12
13	50	3	8.0	24	11.0	33
14	138	3	5.0	15	12.7	38
15	119	3	12.3	37	11.0	33
16	120	5	9.2	46	12.4	62
17	152	7	5.1	36	14.1	99
18	157	4	6.2	25	-	-
19	184	6	4.8	29	-	-

Extrinsic Evaluation

Novel User Interface

To facilitate both evaluations, we built a user interface using standard Java servlets, which presents information about the article as described earlier. In this study, we implement four different user interfaces:

1. *FigSumInt* – interface with title, abstract, and figures linked to summaries generated from *FigSum* system

2. *FigSum+Int* – interface with title, abstract, and figures linked to summaries generated from *FigSum+* system

3. *SimpleInt* – interface with title, abstract, figures and captions, but no summarization system

4. *FullTextInt* – interface with the full-text pdf of the article only

Evaluation Study 1: Comparative Evaluation of FigSum and FigSum+ Interfaces

We evaluate the interface incorporating *FigSum+* (*FigSum+Int*) alone, and then in comparison to the interface using *FigSum* (*FigSumInt*), measuring the efficiency and usefulness of the figure summary for comprehending figures. For this evaluation, we recruited authors who evaluated how well each figure summarization system performed. The following evaluations were performed:

(i) *FigSum+Int* evaluation – article authors use the interface and rank it and the summary generated by *FigSum+* on a Likert-type scale for the criteria defined in Table 23.

(ii) *FigSum+Int* vs. *FigSumInt* evaluation – article authors use the *FigSumInt* and *FigSum+Int* interfaces and, for the evaluation criteria defined in Table 23, either rate one of them as superior to the other or both as being equally good, and the authors provide comments.

Study Conditions, Subjects, and Implementation

To estimate whether the initial figure enlarged in the right pane of the interface has any effect on content comprehension, we evaluate the summaries under two different conditions.

    (a) Enlarge a random figure from the list of figures published in the article.

    (b) Enlarge the most important figure in the article.

For condition (b), the most important figure in the article is determined by a figure ranker [216]. The figure ranker ranks the figure on the notion of its degree of centrality to the article. The centrality of the figure is calculated via the similarity between the full-text article and the figure. For each of evaluations (i) and (ii) described above, we evaluate conditions (a) and (b), resulting in four evaluations:

1. *FigSum+Int* evaluation with random figure enlarged initially – (i)(a)

2. *FigSum+Int* evaluation with most important figure enlarged initially – (i)(b)

3. *FigSum+Int* vs. *FigSumInt* evaluation with random figure enlarged initially – (ii)(a)

4. *FigSum+Int* vs. *FigSumInt* evaluation with most important figure enlarged initially – (ii)(b)

We apply *FigSum* and *FigSum+* on the PubMed Central article collection. As authors of the articles best understand the article, requests were sent to 3000 randomly selected corresponding authors, i.e., 750 requests for each of the four evaluations.

Table 23: Evaluation criteria for comparative interface evaluation

a) Overall quality of figure summary
b) Helpfulness of summary to understand the figure with caption
c) Relevance of summary to figure
d) Conciseness and coverage of information related to figure by summary

Task-driven Cognitive Evaluation

In this section, we describe the task-driven, randomized, controlled cognitive evaluation used to compare the efficacy and accuracy for comprehension of the article's main content using the four interfaces described in previous section: *FigSumInt, FigSum+Int, SimpleInt*, and *FullTextInt*. For this evaluation, our study recruited 16 users and used 16 articles. Each of the article authors prepared a question that assessed the important aspect of the article. All 16 users then read each of the 16 articles and responded to author questions for that article.

Gold Standard

We emailed authors of articles published in *The Journal of Biological Chemistry* and *Proceedings of the National Academy of Sciences*. We asked authors to respond to the following questions "What is the most important research question that you are trying to address in your article?" and "What is the answer for this question?" The responses from 16 randomly selected authors were used for the evaluation task.

Study Design, Subjects and Procedure

We designed a 16 × 16 Latin square evaluation to counterbalance the interfaces. We recruited 16 graduate students and researchers (9 women and 7 men, aged 25 to 35; 14 Asians and 2 Caucasians) from the Department of Biology at the University of

Wisconsin – Milwaukee to participate in the study. There were nine early stage doctoral students, four advanced doctoral students, two postdoctoral fellows, and one lab research manager. All individuals had experience using information systems like PubMed to access biomedical articles. Each user was given a $20 Amazon gift card for participating in the study.

Each of the 16 users evaluated 16 articles, resulting in a 16 × 16 Latin square as shown in Table 29. The users were presented with a question page containing all 16 questions, a link to an interface, and a text box to provide answer for each question. Each question was assigned to one of the four interfaces as determined by the Latin square design. All evaluation studies were conducted in December 2011.

Users were under no time constraints and were allowed to copy and paste text from the interface into their answer. They could skip to the next question and come back to it later. If users were not able to find the answer, they were to respond that the information provided was not sufficient. There were a total of 256 responses (16 users × 16 questions); users took 76.75 ± 27.65 min on average to complete the task.

We recorded user interaction and activity, including user comments, using Morae usability software. Morae software has the ability to monitor and record a wide range of events and activities performed while the user interacts with the system, including mouse clicks, text entry, browsing behavior, change of Web page, and moving between various windows. Morae provides analysts the ability to code, timestamp, and categorize events.

Analysis

To evaluate the quality of answers provided by the participants, we requested the first authors of the articles to rate the answers on a scale of 1 to 4, with 4 being "very good" and 1 being "very poor." Eight out of 16 authors responded with evaluation scores. The remaining responses were evaluated similarly by a bioinformatician. Later, we performed cognitive evaluation as defined in [235,241,242] by identifying goals and actions common to all the interfaces. Table 24 shows a list of tasks performed by subjects. We measured time spent, number of actions, quality of answers, and analyzed participant comments thematically. The novel user interface design required users to make a number of clicks to browse through the figures. *FullTextInt* required users to scroll to browse information. Scrolling and clicks were counted as distinct actions. The number of actions and time spent included browsing through the interface, answering the question, adjusting the browser, and copying and pasting the answer. An office assistant coded the recordings as defined in Table 24. It took five to seven hours to code each subject's recording.

Table 24: Description of the task performed by the subjects to answer questions

Task	Definition
Reading Question	The task begins when user starts a question and ends when user completes responding to the question and moves on to next question.
User Comment	The task begins when user is reading out loud or commenting on usability of interface.
Reading Article	The task begins when user starts to examine the interface to find answers and ends when user leaves the interface.
Adjust browser	The task begins when the user is adjusting the browser with actions such as expanding, moving, repositioning, or closing browser and ends when user returns to other task.
View Title	The task begins when user starts viewing the title and ends when user moves on to a different part of the article or away from the article to answer the question.
View Abstract	The task begins when user starts viewing the article abstract and ends when user moves on to a different part of the article or away from the article to answer the question
View Full text	The task begins when user starts viewing the article full text except for title and abstract and ends when user moves to title, abstract, or away from the article to answer the question.
View Figure	The task begins when the user starts viewing a figure, including its captions and summary.
Scroll	The task begins when the user is navigating the document too quickly and unlikely be performing any of the previously defined actions.
Copy/paste	The task begins when user starts to copy information from the interface and ends after the information is pasted.
Return to Question Page	The task begins when user returns to question page that contains list of questions and box for answers and ends when user navigates to another page.
Answer the Question	The task begins when user starts entering text in answer box and ends when user navigates away from answer box.
Searching for text	The task begins when user starts searching for text in the interface using Cntl + F option and ends when user finished searching.
Encountered User error	The task begins when user encounters a browser-related error due to problems with internet connection or takes significant time (>1 second) to navigate the interface or question page.

Results

Intrinsic Evaluation

The intrinsic evaluation compares the performance of all five *FigSum+* implementations against baseline and state-of-the-art unsupervised and supervised systems. Table 25 and Table 26 show the average performance of the various systems we built for summarization on the *FigSumGS1* and *FigSumGS2* datasets respectively. We chose the value of top *n* to be equal to the average number of sentences per figure in the gold standard. Hence, *n* is equal to 8 and 11 sentences per figure for *FigSumGS1* and *FigSumGS2* datasets respectively.

Baseline Systems Result

For unsupervised baseline case, the *RandomSent* system had an F1 score performance of 0.06 and 0.08 and R1 scores of 0.28 and 0.32 on *FigSumGS1* and *FigSumGS2* datasets. The *RandomPara* system had an F1 score performance of 0.01 on both gold standards and R1 scores of 0.22 and 0.32 on *FigSumGS1* and *FigSumGS2* datasets respectively. The *MEAD* system achieved an F1 score performance of 0.05 and 0.07 and R1 scores of 0.30 and 0.36 on *FigSumGS1* and *FigSumGS2* datasets respectively. Whereas the state-of-the-art unsupervised method *FigSum* system had an F1 score performance of 0.22 and 0.18 and R1 score of 0.51 and 0.55 on *FigSumGS1* and *FigSumGS2* datasets respectively.

For supervised baseline case, all the implementations of the baseline SVM systems, except for the system using the hybrid feature, failed to generate summaries. Both the

NB- and SVM-based systems using the hybrid feature, *NBHybrid* and *SVMHybrid*, performed similarly and had the best baseline F1 score performance of 0.49 and 0.26 and R1 performance of 0.95 and 0.75 on the *FigSumGS1* and *FigSumGS2* datasets respectively.

State-of-the-Art Systems Result

We compared our *FigSum+* systems with the unsupervised *FigSum* system. The *FigSum* system had an F1 score performance of 0.22 and 0.18 and R1 score of 0.51 and 0.55 on *FigSumGS1* and *FigSumGS2* datasets respectively. For supervised state-of-the-art case, the NB-based supervised systems performed well compared to the SVM-based model similar to performance in article [213]. On *FigSumGS1* dataset, the NB-based state-of-the-art systems *NBSOTA* and *NBSOTA+* had an F1 score performance of 0.53, but *SVMSOTA+* achieved the second best R1 score of 0.95. Similarly, on *FigSumGS2* dataset, *NBSOTA* and *NBSOTA+* had the best F1 score performance of 0.38, and *SVMSOTA+* achieved the best R1 score of 0.76. The results of the *FigSum* and Bhatia and Mitra systems are shown in *italics* in Table 25 and Table 26.

Our FigSum+ Systems Result

The *SurfaceCue* implementation of *FigSum+* achieves the highest precision on both gold standards (0.96 and 0.63 on *FigSumGS1* and *FigSumGS2* datasets respectively) and the *Paragraph* implementation results in the highest recall (0.82 and 0.42 on *FigSumGS1* and *FigSumGS2* datasets respectively) and the highest F1 score (0.66 and 0.41 on *FigSumGS1* *FigSumGS2* datasets respectively). The *Hybrid* implementation performs second best,

yielding F1 scores of 0.62 and 0.39, respectively, on *FigSumGS1 and FigSumGS2*

datasets.

Table 25: Average performance and ROUGE scores of (average ± standard deviation) of figure summarization techniques on *FigSumGS1* dataset. Bold indicates the best performance.

		System	Precision	Recall	F1 score	R1	R2	RSU4
Baseline	**Unsupervised**	RandomSent	0.06±0.09	0.06±0.12	0.06±0.09	0.28±0.09	0.11±0.10	0.13±0.09
		RandomPara	0.04±0.18	0.01±0.05	0.01±0.05	0.22±0.16	0.07±0.18	0.08±0.17
		MEAD	0.05±0.09	0.06±0.11	0.05±0.08	0.30±0.08	0.12±0.09	0.14±0.09
	Supervised	NBSimilarity	0.48±0.18	0.15±0.12	0.20±0.12	0.50±0.32	0.40±0.31	0.40±0.31
		SVMSimilarity	-	-	-	-	-	-
		NBTFIDF	-	-	-	-	-	-
		SVMTFIDF	-	-	-	-	-	-
		NBSurfaceCue	0.44±0.11	0.17±0.20	0.18±0.15	0.57±0.19	0.45±0.24	0.46±0.24
		SVMSurfaceCue	-	-	-	-	-	-
		NBParagraph	0.54±0.20	0.74±0.19	0.59±0.14	0.73±0.20	0.66±0.25	0.66±0.25
		SVMParagraph	-	-	-	-	-	-
		NBHybrid	0.80±0.19	0.37±0.15	0.49±0.15	0.95±0.13	0.94±0.17	0.94±0.17
		SVMHybrid	0.80±0.19	0.37±0.15	0.49±0.15	0.95±0.13	0.94±0.17	0.94±0.17
State-of-the-art	**Unsupervised**	FigSum	*0.28±0.24*	*0.19±0.19*	*0.22±0.19*	*0.51±0.18*	*0.36±0.22*	*0.37±0.21*
	Supervised	NBSOTA	*0.44±0.15*	*0.74±0.17*	*0.53±0.12*	*0.63±0.12*	*0.53±0.15*	*0.53±0.14*
		SVMSOTA	*0.58±0.15*	*0.17±0.20*	*0.23±0.22*	*0.41±0.47*	*0.39±0.47*	*0.39±0.47*
		NBSOTA+	0.47±0.16	0.70±0.19	0.53±0.13	0.67±0.16	0.57±0.20	0.57±0.20
		SVMSOTA+	0.78±0.17	0.34±0.14	0.47±0.14	0.95±0.14	0.93±0.18	0.93±0.18
Our System	**FigSum+**	Similarity	0.28±0.20	0.38±0.28	0.30±0.20	0.52±0.17	0.38±0.20	0.38±0.20
		TFIDF	0.30±0.25	0.34±0.24	0.30±0.22	0.51±0.21	0.38±0.25	0.38±0.24
		SurfaceCue	**0.96±0.13**	0.41±0.22	0.54±0.21	**0.97±0.07**	**0.97±0.10**	**0.97±0.10**
		Paragraph	0.64±0.27	**0.82±0.23**	**0.66±0.20**	0.74±0.20	0.67±0.25	0.68±0.24
		Hybrid	0.67±0.28	0.64±0.27	0.62±0.24	0.77±0.19	0.71±0.25	0.71±0.24

The ROUGE score evaluation of *SurfaceCue* resulted in the highest R1, R2, and RSU4 scores, all above 0.97, on *FigSumGS1* dataset. Similarly, *SurfaceCue* resulted in the highest R1 score of 0.76 on *FigSumGS2* dataset.

Table 26: Average performance and ROUGE scores (average ± standard deviation) of figure summarization techniques on *FigSumGS2* dataset. Bold indicates the best performance.

		System	Precision	Recall	F1 score	R1	R2	RSU4
Baseline	Unsupervised	RandomSent	0.08±0.08	0.09±0.11	0.08±0.09	0.32±0.08	0.15±0.09	0.16±0.08
		RandomPara	0.04±0.16	0.01±0.04	0.01±0.05	0.32±0.08	0.14±0.10	0.16±0.09
		MEAD	0.08±0.10	0.07±0.09	0.07±0.09	0.36±0.08	0.17±0.10	0.19±0.10
	Supervised	NBSimilarity	0.42±0.14	0.10±0.08	0.14±0.08	0.48±0.28	0.36±0.25	0.37±0.26
		SVMSimilarity	-	-	-	-	-	-
		NBTFIDF	-	-	-	-	-	-
		SVMTFIDF	-	-	-	-	-	-
		NBSurfaceCue	0.49±0.06	0.05±0.04	0.08±0.05	0.05±0.15	0.03±0.11	0.03±0.11
		SVMSurfaceCue	-	-	-	-	-	-
		NBParagraph	0.43±0.16	0.41±0.18	0.40±0.13	0.66±0.18	0.55±0.23	0.56±0.23
		SVMParagraph	-	-	-	-	-	-
		NBHybrid	0.55±0.17	0.18±0.08	0.26±0.11	0.75±0.25	0.66±0.33	0.66±0.33
		SVMHybrid	0.55±0.17	0.18±0.08	0.26±0.11	0.75±0.25	0.66±0.33	0.66±0.33
State-of-the-art	Unsupervised	FigSum	*0.31±0.20*	*0.13±0.10*	*0.18±0.13*	*0.55±0.14*	*0.40±0.18*	*0.41±0.18*
	Supervised	NBSOTA	*0.37±0.14*	***0.43±0.19***	*0.38±0.13*	*0.59±0.11*	*0.46±0.13*	*0.47±0.13*
		SVMSOTA	*0.54±0.12*	*0.10±0.11*	*0.15±0.15*	*0.41±0.42*	*0.37±0.41*	*0.37±0.41*
		NBSOTA+	0.37±0.13	**0.43±0.20**	0.38±0.13	0.60±0.15	0.47±0.18	0.47±0.18
		SVMSOTA+	0.54±0.16	0.18±0.12	0.26±0.11	**0.76±0.25**	0.67±0.33	0.67±0.33
Our System	FigSum+	Similarity	0.31±0.16	0.28±0.16	0.29±0.15	0.55±0.13	0.40±0.16	0.41±0.15
		TFIDF	0.27±0.22	0.20±0.14	0.22±0.16	0.51±0.18	0.36±0.22	0.36±0.21
		SurfaceCue	**0.63±0.36**	0.16±0.13	0.24±0.17	**0.76±0.24**	**0.68±0.32**	**0.68±0.31**
		Paragraph	0.51±0.24	0.42±0.22	**0.41±0.17**	0.66±0.18	0.56±0.22	0.56±0.22
		Hybrid	0.54±0.24	0.33±0.19	0.39±0.18	0.70±0.16	0.60±0.21	0.60±0.21

Extrinsic Evaluation

Comparative Evaluation Results

Table 27 presents the results of the comparative evaluation. As described in comparative evaluation section, we conducted four evaluations. Some of the values were missing, as the authors did not provide response to all the criteria in Table 23.

The first two columns of Table 27 show the result of the *FigSum+Int* evaluation. The vast majority of authors either moderately or strongly agreed that *FigSum+* was better for all four criteria, and we find no statistically significant difference between which figure was enlarged initially.

The last two columns in Table 27 show the result of the *FigSum+Int* vs. *FigSumInt* evaluation, again comparing a random figure and most important figure enlarged initially. We performed chi-square analysis and obtained chi-square value (χ^2) and degrees of freedom (df) to test statistical significance of the responses received. For both initial figure conditions, many authors indicated that *FigSum+Int* was either better or equal to *FigSumInt* on all four criteria. The chi-square analysis revealed that author responses were statistically significant ($p < 0.05$) for comparing the interfaces on all criteria for condition (a) random figure enlarged, whereas for condition (b), the most important figure enlarged, all criteria except overall quality of summary are statistically significant ($p < 0.05$). A chi-square analysis of both conditions combined (random figure and most important figure enlarged initially) showed that *FigSum+Int* is significantly better than *FigSumInt* ($p < 0.05$). The authors generally spoke positively about the summaries and the interface during the evaluation. A few of the comments are shown in Table 28.

Table 27: Results of the comparative summary evaluation

System		FigSum+Int			FigSumInt vs. FigSum+Int	
		Random Fig	*Ranked Fig*		*Random Fig*	*Ranked Fig*
# of requests sent		750	750		750	750
# of responses		60	71		53	53
Overall Quality Good	*Strongly agree*	60.0% (36)	50.7% (36)	*FigSum better*	17.0% (9)	36.5% (19)
	Moderately agree	30.0% (18)	43.7% (31)	*Both are same*	26.4% (14)	23.1% (12)
	Neither agree nor disagree	3.3% (2)	1.4% (1)	*FigSum+ better*	56.6% (30)	40.4% (21)
	Moderately disagree	3.3% (2)	4.2% (3)	(χ^2, df)	(13.61,2)	(2.57,2)
	Strongly disagree	3.3% (2)	0.0% (0)			
Summary Helpful	*Strongly agree*	51.7% (31)	43.7% (31)	*FigSum better*	22.0%(11)	36.5% (19)
	Moderately agree	33.3% (20)	42.2 (30)	*Both are same*	24.0% (12)	13.5% (7)
	Neither agree nor disagree	3.3% (2)	7.0% (5)	*FigSum+ better*	54.0% (27)	50.0% (26)
	Moderately disagree	8.3% (5)	2.8% (2)	(χ^2, df)	(9.63,2)	(10.64,2)
	Strongly disagree	3.3% (2)	1.4% (1)			
Summary Relevant	*Strongly agree*	51.7% (31)	35.2% (25)	*FigSum better*	13.7% (7)	37.3% (19)
	Moderately agree	30.0% (18)	40.8% (29)	*Both are same*	39.2% (20)	13.7% (7)
	Neither agree nor disagree	8.3% (5)	7.0% (5)	*FigSum+ better*	47.1% (24)	49.0% (25)
	Moderately disagree	10.0% (6)	12.7% (9)	(χ^2, df)	(9.29,2)	(9.88,2)
	Strongly disagree	1.7% (1)	1.4% (1)			
Summary Concise	*Strongly agree*	50.0% (30)	33.8% (24)	*FigSum better*	19.2% (10)	38.0% (19)
	Moderately agree	28.3% (17)	45.1% (32)	*Both are same*	30.8% (16)	8.0% (4)
	Neither agree nor disagree	13.3% (8)	11.3% (8)	*FigSum+ better*	50.0% (26)	54.0% (27)
	Moderately disagree	5.0% (3)	7.0% (5)	(χ^2, df)	(7.53,2)	(16.34,2)

	Strongly disagree	1.7% (1)	1.4% (1)			

Table 28: Comments received from evaluators who were first authors of articles

Positive Comments:
Indeed, this is a great tool and I found it useful!
It is really very convenient for the readers. I feel it very helpful for the researchers.
It is excellent way of representing the summary of the events related to the figures in the article as well as brings the best out for the reader to comprehend. It is very useful and will give a new dimension to writing and publishing one's scientific article. The summary is concise and informative.
Negative Comments:
Summary needs to be little more informative regarding the figure by adding the observation & result of the figure.

Task-Driven Cognitive Evaluation Results

The excerpts below illustrate the coding process followed to characterize participant actions. Later, we report time spent, number of actions, and quality of answer for all four interfaces evaluated, as described in the task-driven cognitive evaluation section. Table 29 below shows the Latin square design used in the study. Each cell contains the interface the user was presented and the score assigned by the author/bioinformatician for the user response.

Illustration of Coding the Task

We recorded the participant's interaction for all 16 questions. The following excerpts show the process of coding the Morae recordings for each question based on the interface they received.

Table 29: Latin square design. Users were assigned 16 articles. Each user was presented with one of the four interfaces. Each cell shows the interface presented to user and the score assigned to the user response on a scale of 1 to 4, with 4 being "very good" and 1 being "very poor." A – *FullTextInt,* B – *FigSumInt,* C – *SimpleInt,* D – *FigSum+Int*

Q\U	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	A 3	B 2	C 4	D 2	C 2	D 4	A 3	B 2	A 3	B 2	C 4	D 2	C 1	D 1	A 2	B 4
2	B 3	C 3	D 3	A 1	D 2	A 3	B 4	C 3	B 2	C 2	D 3	A 1	D 4	A 1	B 3	C 4
3	C 4	D 4	A 4	B 3	A 2	B 3	C 1	D 3	C 1	D 1	A 4	B 3	A 4	B 3	C 4	D 4
4	D 2	A 1	B 4	C 2	B 2	C 2	D 4	A 2	D 1	A 2	B 2	C 3	B 1	C 1	D 2	A 4
5	D 3	A 1	B 1	C 3	C 3	D 1	A 2	B 3	D 1	A 3	B 2	C 4	A 4	B 4	C 2	D 1
6	A 2	B 3	C 3	D 1	D 2	A 4	B 3	C 2	A 2	B 3	C 2	D 1	B 1	C 2	D 2	A 4
7	B 4	C 4	D 3	A 1	A 1	B 3	C 4	D 1	B 3	C 2	D 1	A 1	C 1	D 2	A 4	B 4
8	C 4	D 4	A 4	B 1	B 2	C 3	D 2	A 2	C 1	D 3	A 2	B 4	D 4	A 1	B 2	C 4
9	B 1	C 4	D 4	A 1	C 2	D 4	A 2	B 3	C 1	D 1	A 1	B 2	C 1	D 1	A 1	B 4
10	C 1	D 4	A 2	B 1	D 4	A 2	B 4	C 1	D 1	A 1	B 4	C 4	D 4	A 3	B 3	C 4
11	D 1	A 1	B 3	C 1	A 1	B 3	C 3	D 2	A 1	B 1	C 2	D 4	A 4	B 3	C 2	D 1
12	A 2	B 1	C 3	D 1	B 1	C 3	D 1	A 1	B 1	C 2	D 1	A 3	B 1	C 1	D 2	A 1
13	D 3	A 3	B 3	C 1	C 1	D 2	A 2	B 1	B 2	C 2	D 1	A 1	A 2	B 1	C 3	D 4
14	A 1	B 1	C 4	D 1	D 1	A 3	B 1	C 2	C 1	D 1	A 2	B 4	B 1	C 1	D 4	A 1
15	B 4	C 4	D 4	A 1	A 1	B 2	C 4	D 2	D 3	A 4	B 2	C 3	C 1	D 1	A 4	B 4
16	C 3	D 4	A 4	B 3	B 2	C 3	D 3	A 1	A 3	B 1	C 3	D 2	D 4	A 1	B 3	C 4

Excerpt 1 – *FullTextInt*: The subject had completed five questions and was over an hour into the session. The subject was provided the article [243] and was asked the question, "Does the cancer stroma vary in between cancers?" The subject was presented with *FullTextInt* to answer the question. The participant took 9.37 min to answer this question.

1:08:31   START Question

1:08:40   END ACTION Return to Question interface

1:08:41   START ACTION Article: Click on the link for article

1:08:45   START ACTION View Abstract

1:09:01   START ACTION Adjust browser

1:09:04   END ACTION Adjust browser

1:09:22   START ACTION Adjust browser

1:09:27   END ACTION Adjust browser

1:09:28   END ACTION View Abstract

1:09:29   START ACTION Examine Full text

1:11:55   COMMENT: The figures and figure summaries are better. The
summary describes the figure. The caption should specify the complete
explanation of type of figure. For example, if you are looking at image
of electro micrograph of cancerous cells, the summary should specify
that. The figure summary does it in a nice and concise way. Going
through the full-text article is very hard.

1:13:51   END ACTION Examine Full text

1:13:53   START ACTION View Abstract

1:13:56   END ACTION View Abstract

1:13:57   START ACTION Return to Question interface

1:14:03   END ACTION Return to Question interface

1:14:04   START ACTION Examine Full text

1:14:56   END ACTION Examine Full text

```
1:14:58     START ACTION View Abstract

1:15:43     END ACTION View Abstract

1:15:43     START ACTION Return to Question interface

1:15:49     END ACTION Return to Question interface

1:15:53     COMMENT: I'm trying to scroll through all the figures.

1:15:54     START ACTION Return to Question interface

1:15:55     END ACTION Return to Question interface

1:15:59     START ACTION Return to Question interface

1:15:59     START ACTION Answer Question

1:16:08     END ACTION Answer Question

1:16:09     END ACTION Return to Question interface

1:16:09     START ACTION Examine Full text

1:16:27     END ACTION Examine Full text

1:16:27     END ACTION Article

1:16:27     START ACTION Return to Question interface

1:16:28     COMMENT: I think I found answers in figures.

1:16:28     START ACTION Answer Question
```

```
1:18:08     END ACTION Answer Question
```

```
1:18:09     END Question
```

Excerpt 2 – *FigSum+Int*: The subject had completed three questions and was nearly eight minutes into the evaluation. The subject was provided with *FigSum+Int* and asked, "What is the mechanism of substrate transport and binding at the binding site?" for article [244]. The participant took 5.37 min to answer this question.

```
07:45.1     START Question

07:45.8     START ACTION Article: Click on the link for article

07:47.7     START ACTION Adjust browser

08:09.2     END ACTION Adjust browser

08:09.6     START ACTION View Abstract

10:02.8     START ACTION Copy/Paste

10:19.5     END ACTION View Abstract

10:21.3     START ACTION Return to Question interface

10:22.0     START ACTION Answer Question

10:24.7     END ACTION Copy/Paste

10:26.0     END ACTION Answer Question

10:26.6     END ACTION Return to Question interface

10:27.4     START ACTION Figure 1

10:52.7     COMMENT: Interface is good

11:03.5     END ACTION Figure 1

11:05.0     START ACTION Figure 2

11:09.6     END ACTION Figure 2
```

```
11:10.7     START ACTION Figure 3

11:11.2     COMMENT: The figure summary really helps. If the interface
            had a one-line heading in bold describing the figure, it
            would be useful.

11:45.8     END ACTION Figure 3

11:46.6     START ACTION Figure 4

11:54.1     END ACTION Figure 4

11:55.7     START ACTION Figure 5

12:23.1     COMMENT: Found the answer should I copy/paste the entire
section?

12:41.0     START ACTION Copy/Paste

12:43.9     END ACTION Figure 5

12:44.1     END ACTION Article

12:44.8     START ACTION Return to Question interface

12:45.3     START ACTION Answer Question

13:21.0     END ACTION Copy/Paste

13:21.4     END ACTION Answer Question

13:21.8     END Question
```

The above excerpts represent the nature of interaction of the user and the features each system offers. The actions performed by the participants on *SimpleInt* and *FigSumInt* were similar to *FigSum+Int*, as shown in Excerpt 2. The users browse through the figures, captions and summary, if available. As described in Table 24, we consider viewing a figure, its caption and summary as one action; hence no separate action is described for viewing summary and caption.

One participant could not find the answer to the question with *FullTextInt*. Similarly, there were 4, 4, and 3 number of participants who could not answer the questions with *SimpleInt*, *FigSumInt*, and *FigSum+Int* interfaces respectively. Participants spent an average of 5.01 min with the interface before indicating that they could not answer the question.

Table 30 reports descriptive statistics for the cognitive evaluation criteria: time spent per question, number of actions per question, and quality of answer. *FullTextInt* took the longest time (ave. 5.02 min), as users had to browse through the entire text. *SimpleInt* took the least time (ave. 3.35 min), as the system had minimum information (abstract + figures + caption). *FigSumInt* and *FigSum+Int* took 4.27 min and 4.42 min on average, respectively. The average number of actions and quality of answers for all the interfaces was 16 and 2.4, respectively. There were no statistically significant difference between interfaces for number of actions and quality of answers, except for time spent on *SimpleInt* (t-test, $p < 0.05$).

Table 30: Descriptive statistics (avg ± stddev) of task-based cognitive evaluation

System	*FullTextInt*	*SimpleInt*	*FigSumInt*	*FigSum+Int*
Time Spent	5.02 ± 3.05 min	3.35 ± 1.58 min	4.27 ± 3.14 min	4.42 ± 2.58 min
Number of Actions	15.92 ± 9.7	15.09 ± 5.4	15.8 ± 7.84	16.73 ± 7.48
Quality of Answer (4 = very good; 1=very poor)	2.17 ± 1.16	2.5 ± 1.15	2.45 ± 1.09	2.34 ± 1.23

Discussion

Intrinsic Evaluation

In this study, we developed and investigated five implementations of *FigSum+* to automatically summarize figure-associated text from the article. These approaches remove redundant information by extracting sentences associated with the figure and decrease user information overload. We evaluated the performance of these approaches against two sets of gold standards. The first gold standard was comprised of 94 figures from 19 *PMC* articles (*FigSumGS1* dataset) and the second, a subset of 84 figures from 17 articles in the *FigSumGS1* dataset *(FigSumGS2* dataset). The *FigSumGS1* dataset showed a good IAA of 0.68 Cohen's κ for a subset of 11 articles.

We first compared the performance of the five *FigSum+* systems against unsupervised baseline (*RandomSent, RandomPara*, and *MEAD*) and unsupervised state-of-the-art (*FigSum*) systems. The improvement in both F1 score and ROUGE performance of *SurfaceCue*, *Paragraph, Hybrid* compared to all unsupervised systems was statistically significant (t-test, $p < 0.05$) on the *FigSumGS1* dataset. Whereas, for the *FigSumGS2* dataset comparison of unsupervised baseline systems, the improvement in the ROUGE score performance of *SurfaceCue*, *Paragraph*, and *Hybrid* was statistically significant (t-test, $p < 0.05$), but the F1 score performance of only *Paragraph* and *Hybrid* was statistically significant (t-test, $p < 0.05$).

Supervised baseline systems using the same individual features as in the *FigSum+* systems were built using the NB and SVM machine-learning techniques. All baseline SVM systems except for the system using the hybrid feature failed to generate figure

summaries on both datasets. Among the supervised baseline systems based on NB, the system using the reference paragraph feature achieved an F1 score performance of 0.59 and 0.40 on *FigSumGS1* and *FigSumGS2* datasets respectively. The NB system using the hybrid feature had the highest R1 performance of 0.95 and 0.76 on *FigSumGS1* and *FigSumGS2* datasets respectively. The difference in F1 and ROUGE score performance of NB-based systems was statistically significant over the *Paragraph* and *Hybrid* (t-test, $p < 0.05$).

We also compared the performance of the *FigSum+* systems against state-of-the-art supervised systems (*NBSOTA* and *SVMSOTA*). We extended these systems (*NBSOTA+* and *SVMSOTA+*) further by adding additional features. The F1 score performance of the *Paragraph and Hybrid* systems were statistically significantly better than all state-of-the-art supervised systems (t-test, $p < 0.05$). In addition, the F1 score performance of *SurfaceCue* was statistically significantly better than the *SVMSOTA* system (t-test, $p < 0.05$) on *FigSumGS1* dataset. In terms of the ROUGE score performance, *SVMSOTA+* achieved the best scores using supervised approaches, and the difference in performance against the best performing *SurfaceCue* was not statistically significant. The systems performed similarly on *FigSumGS2* dataset, but the improvement of state-of-the-art supervised systems on *Paragraph and Hybrid* systems were not statistically significant.

The better performance of our unsupervised *FigSum+* systems over the state-of-the-art supervised systems [213] (*NBSOTA* and *SVMSOTA*) could be attributed to a number of reasons. First, our systems were limited to the biomedical domain. Hence, these features could be better tuned to outperform in our domain. Second, although we used the same set of features as described in [213], the implementation of the similarity feature between

our systems and [213] was different, as we used the cosine similarity instead of the Okapi BM25 similarity. Third, the evaluation data used in [213] were different from the data used in our experiments.

We also found that features such as TFIDF and caption similarity could be adding noise to the supervised machine learning, as Table 25 and Table 26 showed that the performance of the baseline supervised systems using these features yielded low performance. The superior performance of our unsupervised *FigSum+* approaches demonstrates their ability to generate comprehensive figure summaries to enhance user figure comprehension.

The *FigSum+* approaches *SurfaceCue, Paragraph,* and *Hybrid* had average F1 scores ranging between 0.79 and 0.26, 0.84 and 0.27, and 0.82 and 0.21, respectively, for *FigSumGS1* dataset and between 0.62 and 0.10, 0.62 and 0.24, and 0.64 and 0.21, respectively, for *FigSumGS2* dataset. Human-generated summaries often show such variations as well [245,246]. The difference in performance of the various *FigSum+* techniques can be attributed to variations in the quality of the gold standard generated by the annotators.

Further analysis of the *FigSum+* performance on *FigSumGS1* dataset using Spearman Rank Correlation showed that there was no correlation between the F1 score and the length of the article or the number of figures. However, the F1 score of *SurfaceCue* showed moderate negative correlation (rho = -0.51, $p < 0.05$) with the average number of sentences per figure. For *FigSumGS2* dataset, the length of the article had a moderate negative correlation with the performance of *Paragraph* (rho = -0.52, $p < 0.05$) and *Hybrid* (rho = -0.50, $p < 0.05$) implementations and the average number of sentences per

figure and had a negative correlation with the performance of *Paragraph* (rho = -0.71, *p* < 0.05) and the *Hybrid* (rho = -0.74, *p* < 0.05) implementations. This finding suggests that longer summaries tend to have lower quality.

The *SurfaceCue* system had a near perfect ROUGE score for *FigSumGS1* dataset, since the annotators picked figure-referring sentences as part of the gold standard. Although the *SurfaceCue* approach had a very high ROUGE score, it also had a very low recall (0.41 for *FigSumGS1* and 0.16 for *FigSumGS2* datasets) compared to the *Paragraph* and *Hybrid* approaches. There was no correlation between the ROUGE score performance and the length of the article, the number of figures, or the average number of sentences per figure for the *FigSumGS1* dataset. Similarly, there was no correlation between the number of figures or the average number of sentences per figure except length of the article, which had a negative correlation with *SurfaceCue* (rho = -0.72, *p* < 0.05) for *FigSumGS2* dataset.

The *FigSum+* approaches performed well against two different gold standards constructed using different criteria, demonstrating the robustness of the approaches and their efficacy in rendering comprehensive figure summaries. It was also interesting that one article in *FigSumGS1* dataset had an F1 score of 0.44 for the *Hybrid* approach but achieved an R1 score of 0.85, indicating that the quality of the summaries extracted by the *FigSum+* implementations were as good as human-generated summaries.

One of the inherent problems of extractive summaries is that they lack coherence and certain sentences do not make sense when taken out of context (e.g., as in the *SurfaceCue* implementation). For example, Figure 14 shows a figure along with its caption and the sentence extracted by the *SurfaceCue* method. The sentence "*The summary risk*

difference was 0.27% (−0.10% to 0.63%, P=0.15, I2=0%; fig 2) with no indication of publication bias in the funnel plot," provides very little context for the figure. To overcome this problem, we extracted whole paragraphs where figure-referring sentences appeared, as in the *Paragraph* approach. Figure 15 shows the summary extracted by the *Paragraph* method for the figure shown in Figure 14. The summary provides more information and context to help understand the figure better. We believe this method provides users with the sentence context and improves the overall comprehension of the figure while reducing user information overload.
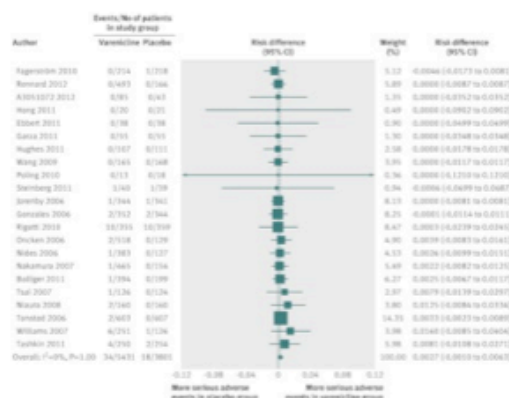
Fig 2: *Difference in risk of treatment emergent, cardiovascular serious adverse events associated with varenicline use in 22 double blinded, placebo controlled, randomised trials.*

Sentence from SurfaceCue: The summary risk difference was 0.27% (−0.10% to 0.63%, P = 0.15, I2=0%; fig 2) with no indication of publication bias in the funnel plot.

Figure 14: A sample figure with its caption and the summary generated by *SurfaceCue*. Figure 2 appearing in article [247].

> Across the 22 studies, the crude rates of treatment emergent, cardiovascular serious adverse events were 0.63% (34/5431) for the varenicline group and 0.47% (18/3801) for the placebo group. No events occurred in eight trials, including three trials with more than 100 participants per arm. The summary risk difference was 0.27% (−0.10% to 0.63%, P=0.15, I2=0%; fig 2) with no indication of publication bias in the funnel plot. For comparison, based on 14 studies with at least one event, the relative risk was 1.40 (0.82 to 2.39, P=0.22, I2=0%; table 2), the Mantel-Haenszel odds ratio was 1.41 (0.82 to 2.42, P=0.22, I2=0%), and the Peto odds ratio was 1.58 (0.90 to 2.76, P=0.11, I2=0%).

Figure 15: The summary generated by *Paragraph* method for the figure in Figure 14

There are, however, certain limitations to the study. The current results are based on only 94 figures from 19 biomedical articles. Although this number of figures is small, it is on a par with other studies that also require extensive manual annotation [213]. The results indicate that the *FigSum+* approaches − specifically *Paragraph* and *Hybrid* − can generate summaries that are closely related to the information deemed important by experts to explain the content of figures. We also found that the *FigSumGS1* had a fair IAA of 0.68 Cohen's κ value. In the future, we will further refine the guideline to improve the IAA. Another limitation is that *FigSum+* systems do not consider the semantics of the sentence when generating summaries. Hence, we intend to expand our study by annotating more articles and to use semantic information to potentially improve system performance.

Extrinsic Evaluation

We designed and developed figure summarization systems and incorporated them within a user interface around findings from our previous work [170] that text associated with figures other than the figure caption is helpful in understanding the meaning of a figure in full-text biomedical articles. We sought to minimize information overload and scrolling through the full-text article while providing maximum information in limited space using NLP techniques. The interface provides access to all figures published along with their captions and summaries. Study users indicated that the interface helped them understand the content easily.

The comparative evaluation (Table 27) to evaluate the efficiency of figure summaries indicate that authors find the figure summary useful for comprehending figures, and provided positive comments (Table 28). One limitation of this evaluation is that we had no control over the author's interaction with the interface. Hence, we do not know whether the authors examined the entire interface before evaluating the summaries.

Further, we compared the performance of the two summarization systems in comparative evaluation and their effectiveness against *SimpleInt* and *FullTextInt* in a task-based cognitive evaluation using a 16 × 16 Latin square design (Table 29). We recorded the interaction of all the 16 users using Morae software and analyzed their interaction. The analysis and coding of these recordings was time consuming, manually intensive and expensive process that took five to seven hours to analyze each evaluation.

The users liked the figure summarization systems with interface and its ease of use. Although all the interfaces except *FullTextInt* present the user with a subset of the overall article information, users could provide answers using the interface incorporating

summarization systems (Table 30). The difference in quality of answers from the various systems is not statistically significant. Hence, users are able to answer the questions using the figure summaries incorporated in the interfaces as well as if they had access to the full text article. All users mentioned that the summarization systems were helpful and reduced their effort in going through the entire article.

Table 30 shows that users spent less time and had fewer actions on *SimpleInt*. This is because *SimpleInt* provides title, abstract, and figures with captions, without providing any additional information to find evidence for supporting answers. When more information is provided, users spend time further investigating the content provided via the interface to find evidence and corroborate their answers. Hence, users were more confident with their responses using *FigSumInt* and *FigSum+Int* although it increased the number of actions and the time spent.

We also found that few full-text articles contain an author summary (i.e., a summary of the article described by the author other than the abstract). Users mentioned that the author summary, when available, was helpful in answering the questions. However, an author summary was only included in the *FullTextInt*. As one would expect, the article and the question also influenced the quality of response and time taken by subjects. For example, for article [248] we asked, "How do the components of Mitotic Checkpoint Complex (MCC) inhibit the Anaphase Promoting Complex (APC) in fission yeast?" Participants with *FullTextInt* took 11.00 ± 3.05 min to determine the answer. Whereas, article [249] "Is DETC an effective drug in vivo in a murine model of *Leishmania* infection?" - needed only 1.75 ± 1.05 min to answer with *FullTextInt*. Also, participants

had expertise in different areas within the biological domain. If participants encountered a question in their area of expertise, they so indicated and found answers quickly.

In addition, the task-based cognitive evaluation measures the comprehension of the article main content but did not measure individual figure comprehension. This is because we hypothesize that biomedical researchers are interested in determining the main research question addressed in the article rather than simply comprehending the content of individual figures.

Conclusion

This study explored a number of supervised and unsupervised approaches to summarize figures in biomedical articles by aggregating sentences associated with a figure and removing redundant sentences. We developed the unsupervised *FigSum+* systems and performed an intrinsic evaluation against two different gold standards constructed on 19 PMC articles consisting of 94 figures and reported the ROUGE and F1 scores. The *FigSum+* systems achieved the best F1 score of 0.66 and ROUGE-1 score of 0.97. We performed two extrinsic evaluations of summarization systems incorporated within the user interface for presenting article content The comparative evaluation showed that *FigSum+* summaries accessed through the interface were superior to its predecessor, *FigSum*, in efficiency and usefulness for comprehending figures. A task-driven cognitive evaluation of figure summaries within the user interface showed that users could provide answers to questions addressing the main content of the article with summarization systems without the full-text articles. Therefore, figure summaries provide an approach to

reducing information overload while improving user information seeking behavior and maintaining information content.

Chapter 6: Improving comprehension on EMR Notes with NoteAid

In this chapter we discuss the NLP approaches we explored to aid patient comprehension of EHR notes so that the ADEtector tool can improve not only pharmacovigilance but also the patient engagement and physician and patient communication. Allowing patients direct access to their electronic medical record (EMR) notes has been shown to enhance medical understanding and may improve healthcare management and outcomes. The EMR notes contain a lot of domain specific jargon, complex disease names and abbreviations that make them difficult for patients to fathom. Therefore, we built NoteAid, a NLP system that processes text and provides consumer-oriented, simplified explanations and definitions to complex medical concepts using external knowledge resources. We hypothesize such a system will help patients better comprehend the text and the ADE related information present in the narrative text. We conducted a pilot evaluation for linking EHR notes — through NoteAid — to three external knowledge resources: MedlinePlus, the Unified Medical Language System (UMLS), and Wikipedia. Our results show that Wikipedia significantly improves EHR note readability. Preliminary analyses show that MedlinePlus and the UMLS need to improve both content readability and content coverage for consumer health information. A demonstration version of the fully functional NoteAid system is available at http://clinicalnotesaid.org

Introduction

Allowing patients direct access to their electronic health record notes (EHNs) has been shown to enhance medical understanding and improve medication adherence [250-256]. However, the average American reads at or below an 8th grade level [257], and over 90 million Americans have limited health literacy [258]. Studies have shown that lower health literacy leads to poor health-related knowledge [259, 260], lower use of preventive health services [261,262], increased risk of hospitalization [263,264], decreased medication adherence [265,266], greater probability of depressive symptoms [267,268], poorer health status [269,270], higher healthcare costs [271,272], poorer self-management [273,274], and increased mortality [275,276]. Patients who have limited health literacy may have difficulty understanding written medical information, communicating with healthcare providers, and navigating complex EMR systems.

A recent study [277] suggested that lack of health literacy has the potential to reach unacceptable degrees of negative consequences on patients. Lober et al. [278] found that overall health literacy, manifest by questions about conditions, medications, terminology, and more, presented a barrier to almost of third of their subjects. Another study concluded that lower health literacy had a direct impact on information seeking of patients [279].

We are therefore developing NoteAid, a biomedical natural language processing (BioNLP) system to help patients comprehend EMRs by providing comprehensible terms and concepts tailored to the patient, and by linking the EMR to external patient education materials. Studies have shown that patient education is effective in improving health literacy, decreasing disease severity, improving self-management behaviors, and reducing

hospitalizations [253,254,280]. We therefore hypothesize that NoteAid will improve patient comprehension of their EMRs and therefore increase the quality of patient care.

Related Work

Studies have shown that patients commonly have difficulty understanding at least part of their EMRs [250,252,281,282]. Chapman et al. [283] found that a substantial proportion of the lay public does not understand phrases often used in cancer consultations. Another study showed that patients understand less than 30% of commonly used medical terms in the emergency department [284]. Since medical notes were complex to comprehend, Keselman and Smith [285] developed a classification scheme of comprehension errors and categorized the errors based on lay individuals' re-tellings of two documents containing clinical text.

Zeng-Treitler et al. [286] designed and implemented a prototype text translator to make reports more comprehensible to consumers. The translator identified difficult terms and replaced them with easier synonyms, generated and inserted explanatory texts for them. Their prototype did not show significant better comprehension. Kandula et al. [287] developed a multimedia computer-based program for diabetic patients. They focused on communicating the learning objectives succinctly and eliminated unnecessary information. They showed their system to 190 patients and found that literacy levels had significant increases in knowledge scores after viewing the system ($p < 0.05$).

Hong et al. [288] identified 340 unique diagnosis term/patient-friendly term pairs from UMLS. They found that use of patient-friendly terms could help to bridge the language gap between providers and consumers but not always. Zeng-Treitler et al. [289]

investigated whether multilingual machine translation could help make medical record content more comprehensible to patients who lack proficiency in the English. They translated 213 medical record sentences from English into Spanish, Chinese, Russian and Korean. Evaluation of comprehensibility and accuracy of the translation found that majority of the translations were incomprehensible and/or incorrect and suggested that machine translation tool can potentially be improved.

Smith et al. [290] evaluated the comprehension of clinical document by increasing the coherence of the text. They defined coherence as the connectedness of ideas in a text, which affects comprehension. They improved the coherence by adding background information about patients and providing more information of the disease without increasing the readability level. The study found that coherence has a small effect on consumer comprehension of clinical text, but the task is extremely labor intensive and not scalable.

A substantial amount of work has been done to compile a consumer health vocabulary [291, 292] by analyzing user queries to Web sites at the National Library of Medicine [293,294]; consumer texts [295,296]; social media, including email content [297], and online support groups (e.g., PatientsLikeMe [298]). Approaches have been developed to predict term familiarity with linguistic/stylistic features [299], term frequency [300], as well as machine-learning approaches [301]. Tools have also been developed to simplify EHR note content using both syntactic and semantic approaches (e.g., [302,303]).

The Patient Clinical Information System (PatCIS) [256] was created to serve as a test bed for exploring issues related to patient access of EHR records. It provides patients with online information resources and educational material, and evaluations by patients have

been positive [304]. However, researchers mainly compiled the education material in the PatCIS system manually after reading the EHRs. In this chapter we discuss the development of NoteAid, a system that automatically links EHR notes to patient education materials to assist their EHR note comprehension.

Materials and Methods

NoteAid has two main components: A knowledge resource comprised of patient education materials and BioNLP approaches that link EHR notes to the knowledge resource. In the following, we first describe three knowledge resources. We then describe BioNLP approaches and conclude with an evaluation design.

External Knowledge Resources

The Unified Medical Language System (UMLS) [305] is a rich biomedical knowledge resource; Metathesaurus (MT) is a large, multi-purpose, and multi-lingual thesaurus that contains millions of biomedical and health related concepts, their synonym names, and their relations, from over 150 vocabularies. UMLS makes available the lexical tool MetaMap [85], which maps text to UMLS concepts and semantic types. We use UMLS MT version 2011AB in our system.

MedlinePlus [306] is a National Institutes of Health's Web site for patients and their families and friends. Medline Plus provides current and reliable information about over 900 diseases, conditions and treatment to users in simple language. The links to various health topics are added daily and the content is reviewed once every six months.

Wikipedia (Wiki) is a collaborative, community developed web-based encyclopedia that has evolved to be an important medical resource for a wide spectrum of audiences including healthcare professionals [307]. Among online health information resources, Wiki has shown to be a prominent source, ranking among the first 10 results in 71-85% of search engines and keywords tested [308].

The NoteAid System

Our goal was to assist patients to understand the content of their EHR notes. For this purpose, we decided to link the complex medical concepts that appear in the text to simple consumer oriented definitions and explanations from external sources of information as described earlier. These definitions describe the complex medical concepts and jargon that appear in these EHR notes.
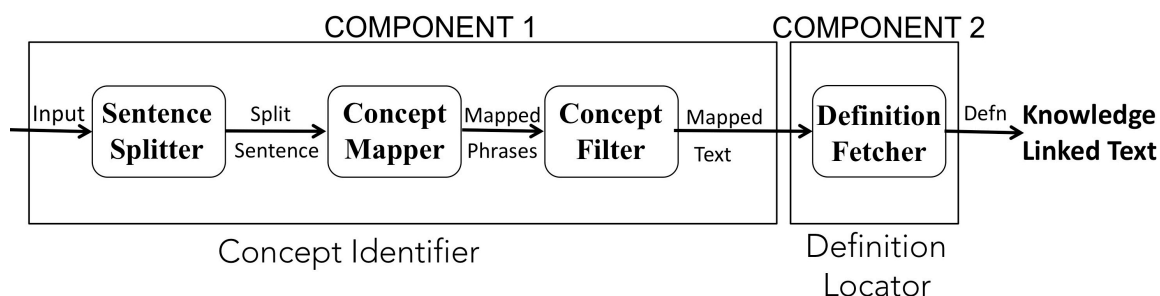
Figure 16: Schematic representation of NoteAid system

Figure 16 above shows the schematic representation of the NoteAid system. The system is comprised of two components. The first component is Concept Identifier (CI). CI processes input text and maps terms to the corresponding UMLS concepts. The second component is Definition Locator (DL). DL fetches definitions from UMLS, MedlinePlus and Wikipedia.

CI consists of three modules: Sentence Splitter, Concept Mapper, and Concept Filter. Sentence Splitter splits input text into individual sentences. Concept Mapper is built upon the Metamap tool [309] which identifies concepts and their UMLS semantic types. Concept Filter identifies clinical concepts by selecting the concepts that belong to the semantic types: Acquired Abnormality, Antibiotic, Cell or molecular Dysfunction, Clinical Attribute, Diagnostic Procedure, Disease or syndrome, Experimental model of disease, Finding, Laboratory procedure, Laboratory or Test result, Organ or Tissue function, Pathologic function, Physiologic function, Pharmacologic substance, Sign or symptom and Therapeutic or preventive procedure.

After concepts are identified, DL retrieves definitions from UMLS, Medline Plus and Wikipedia using Definition Fetcher module. The UMLS MRDEF file contains definitions of 107,604 unique concepts. We parsed the Medline Plus content and extracted over 900 health related topics and their summaries. We automatically extracted definitions from the summaries by using handcrafted rules. For Wiki, we made use of the web service WikiAPI to return a Wiki page given a query topic (concept). We filter the definition fetched by Wiki by adding a filter, which fetches a definition if the article is health-related. Wiki assigns each article a set of categories, which are organized into a direct acyclic graph. We recognize an article as health-related if any of the assigned categories or the corresponding hierarchical categories belong to the following two terms: clinical and health. When an article page is returned, DL extracts the first three lines of the Wikipedia content. We found such a simple method works very well for extracting definitions from Wikipedia.

Evaluation Procedure and Metrics

To evaluate whether NoteAid improves EHR note comprehension, we evaluated four NoteAid implementations, namely: Medline Plus (linking EHR concepts to definitions in Medline Plus), UMLS (linking EHR concepts to their synonyms and definitions in the Unified Medical Language System), Wikipedia (Wiki, linking EHR concepts to health related articles in Wikipedia) and the hybrid system that integrates the three aforementioned implementations using de-identified EHR notes. We conducted two sets of evaluations described below.

Subjects

With IRB approval, we recruited subjects from Amazon Mechanic Turk (AMT). We used AMT because the subjects have various background and qualifications, and therefore are representative in terms of health literacy. Many research studies use AMT for data collection and survey and have proven to be a reliable resource [303].

Evaluation Data and Readability Score

We randomly selected 20 de-identified progress note reports (PGN) and 20 de-identified discharge summary reports (DS) from the Pittsburgh NLP repository [154], which contains a variety of de-identified clinical reports including discharge summaries and progress notes. We used the Flesch-Kincaid ease score and Flesch-Kincaid grade level [310] to score readability; the higher the Flesch Readability ease scores, the higher the readability. In contrast, a lower Flesch-Kincaid grade level indicates higher readability.

We conduct two different evaluations. The first evaluation used both DS and PGN and for second evaluation we used only PGNs.

Evaluation Process

As mentioned above, we conducted two different evaluations. The first evaluation examines the self-reported comprehension score before and after applying the NoteAid system and the second evaluation examines the self reported comprehension of each implementation individually. For quality control, we gave each subject a question related to his/her evaluation data. The evaluation was hosted and stored on a local server. At the end of the evaluation, subjects received a code to confirm their participation in the study and receive payment for the task. Each subject spent 30-40 minutes to complete the entire evaluation.

Evaluation One

For each NoteAid implementation, we asked each subject to read each assigned EHR note before and after the NoteAid system and score his/her level of comprehension (on a scale of 1 to 5, with 1 the poorest and 5 the best comprehension). Each subject was asked to complete the evaluation of either 20 PGNs or 20 DSs. Each subject was given a link to a welcome page describing the study, followed by demographic information page, qualifying question page, pages containing EHR notes to evaluate, and finally the thank you page along with the validation code.

We recruited 64 subjects: 8 subjects for each of the 8 evaluation tasks (4 systems, 2 types of EHR notes). A total of 3 subjects did not complete the evaluations and 2 subjects withdrew from the study. Our results were based on the analyses of the evaluation of the remaining 59 subjects who completed their tasks.

Evaluation Two

We evaluate a total of five systems: the baseline system in which a clinical note is presented without a NoteAid implementation and four NoteAid implementations where a note is presented with different NoteAid implementations. We recruited 25 subjects, 5 subjects for each of the 5 systems. Each subject was asked to evaluate 20 PGN notes. For each note (either the note alone or with a NoteAid implementation), we asked the subject to read and score his/her level of comprehension on a scale of 1 to 5. Each subject was given a link to a welcome page describing the study, followed by demographic information page, qualifying question page, pages containing EHR notes to evaluate and finally a thank-you page along with the validation code.

Four subjects failed to complete the evaluations. Our results were based on the analyses of the evaluation of the remaining 21 subjects who completed their tasks.

Evaluation Criteria

We report the average comprehension scores before and after each of the NoteAid implementations: MedlinePlus, UMLS, Wiki, and Hybrid. The non-parametric Wilcoxon signed-rank test was used to compare subjects' scores on PGNs or DSs before and after each implementation.

In order to evaluate whether the comprehension scores represent readability, we report both Flesch readability ease score and Flesch-Kincaid grade level and calculate the non-parametric Spearman correlation coefficient. We also show the scatter-plot of the comprehension scores before and after the NoteAid systems, between the two readability scores, and between the comprehension and the readability scores.

Demographic Information of Subjects

Evaluation One

Of the 59 subjects (23 female and 36 male) completed the evaluation. The number of Asian, White, African American and Alaskan Native was 34, 23, 1, and 1, respectively. Nearly 24% of all subjects reported having Hispanic or Latin ethnicity. The subjects of the study had a wide range of educational backgrounds. Twenty three (39%) of them had Bachelors degree, 15 (25.4%) of them had a Masters degree, 12 (20.3%) of them had an Associate degree and the remaining 9 (15.3%) had a high school diploma.

Evaluation Two

Twenty one subjects (9 female and 12 male) completed the evaluation. The number of White American (White), Asian, and Black American (Black) were 15, 4, and 2 respectively. The subjects in the study had a wide range of educational (Edu) backgrounds. Six (28.57%) subjects had a Masters degree, 6 (28.57%) had Bachelors degree, 2 (9.52%) of them had an Associate degree and the remaining 7 (33.34%) subjects had a high school diploma.

Results

Table 31 shows the characteristics of the EHR note data used in the evaluation. The DS

and PGN have an average Flesch Readability ease score of 38.5 and 43.9 and an average

Flesch-Kincaid Grade Level of 8.8 and 9.76, respectively.

Table 31: Statistics of NoteAid Evaluation Data

Type	Discharge Summaries	Progress Notes
No. of Reports	20	20
Total (Avg) # of sentences	355 (17.8)	473 (23.7)
Total (Avg) # of Words	2362 (118)	4862 (243)
Avg Flesch ease score	38.5	43.9
Avg Flesch-Kincaid Grade Level	8.8	9.8

Evaluation One

Table 32 shows the average comprehension scores of the four NoteAid implementations

(before and after each implementation). As shown in the table, all three NoteAid

implementations except for MedlinePlus improve the comprehension in both DSs and

PGNs. None of the improvement is statistically significant except for the Wiki

implementation on PGNs. The Hybrid implementation has a p value of 0.06 for

improvement on PGNs.

Table 32: Average standard deviation of comprehension values of four NoteAid implementations

System	Discharge Summaries		Progress Notes	
	Before	**After**	**Before**	**After**
MedlinePlus	3.52±0.73	3.49±0.87	3.18±0.38	2.86±0.55
UMLS	3.80±0.16	3.81±0.48	3.75±0.55	4.01±0.86
Wiki	3.57±0.68	4.14±0.49	3.45±0.55	4.53±0.71*
Hybrid	3.86±0.69	4.02±0.73	3.40±0.55	4.54±0.53
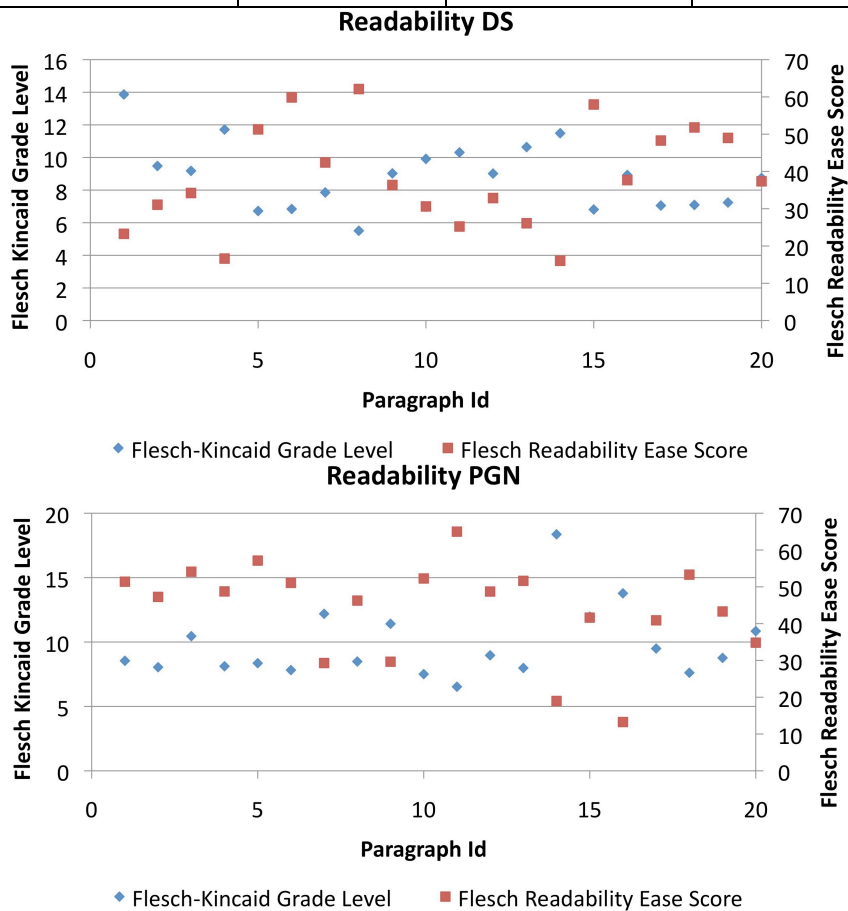
Figure 17: Readability of Evaluation Data

Figure 17 shows the scatter plot of the Flesch-Kincaid Grade Level and the Flesch Readability ease score calculated from the 20 DSs and 20 PGNs we used for the evaluation. The Spearman rank correlation on Flesch-Kincaid Grade Level and the Flesch Readability ease score demonstrated the consistency of data (for PGN: rho = -0.807 $p <$ 0.05, for DS: rho =-0.970, $p < 0.05$).
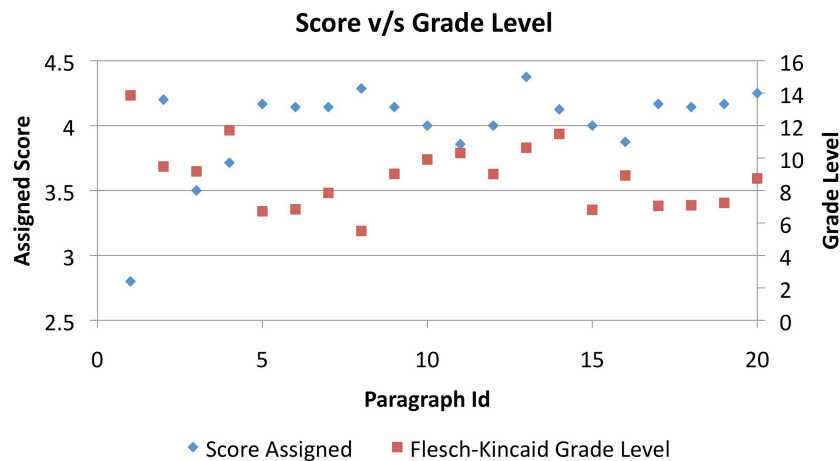
Figure 18: Scatter Plot of the assigned score and Flesch-Kincaid Grade Level in the evaluation EHR notes

Figure 18 shows the scatter plot of the Flesch-Kincaid Grade Level and text comprehension score after NoteAid system on DS reports. The data shows they had a Spearman rank correlation of -0.47, $p < 0.05$. This indicates a fair correlation between the score assigned score and the readability of the reports.

Table 33: Number of concepts that were linked to different knowledge resources by the NoteAid system

System	Discharge Summaries	Progress Notes
MedlinePlus	37	53
UMLS	171	362
Wiki	190	427

Table 33 shows the total number of concepts that were recognized by three different NoteAid implementations on the 20 DSs and 20 PGNs.

Evaluation Two

This evaluation was carried out only on the set of 20 PGNs. Table 34 below shows the average comprehension scores of PGNs without any NoteAid implementation and with each of the four NoteAid implementations. The average comprehension score of subjects and Flesch-Kincaid grade level had a spearman ranked correlation coefficient of rho =-0.77 ($p < 0.05$).

As shown in Table 34, all NoteAid implementations improved self-rated PGN note comprehension and the improvements were statistically significant ($p < 0.05$, the Mann-Whitney-Wilcoxon test). The difference in comprehension scores between different NoteAid implementation was not statistically significant except for the difference between the MedlinePlus and the UMLS implementations ($p < 0.05$, the Mann-Whitney-Wilcoxon test). Table 34 also shows the number of concepts identified by each of the NoteAid implementations.

Table 34: The average self-rated comprehension values (average ± std dev) and number of concepts identified by NoteAid implementations. (*$p < 0.05$)

System	Notes Alone	Medline Plus	UMLS	Wiki	Hybrid
Score	2.95 ± 0.67	4.12* ±0.33	3.63* ± 0.57	3.85* ± 0.47	3.92* ± 0.40
# conc	NA	52	352	436	476
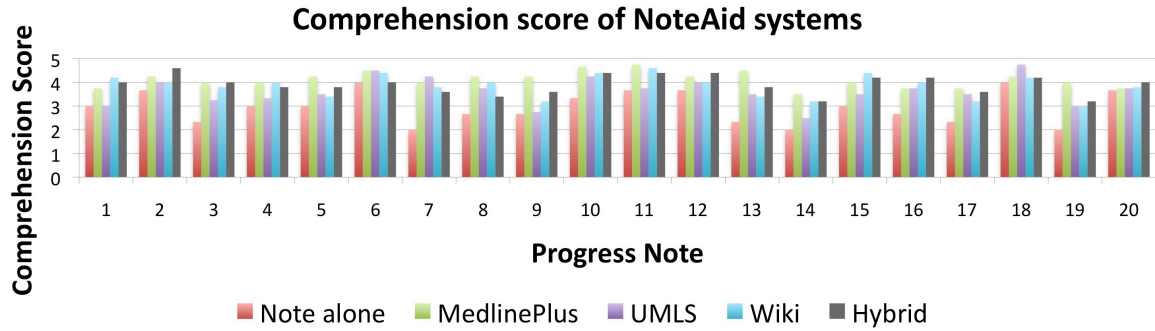
Figure 19: The average self-rated comprehension score for each note with different NoteAid implementations

Figure 19 shows the average self-rated comprehension scores of all NoteAid implementations for every PGN note, and Figure 20 shows a scatter plot of the average self-rated comprehension scores with notes alone and notes with the MedlinePlus implementation. The results as shown in both figures demonstrate a strong and consistent improvement of self-rated comprehension scores with NoteAid implementations for every PGN note.
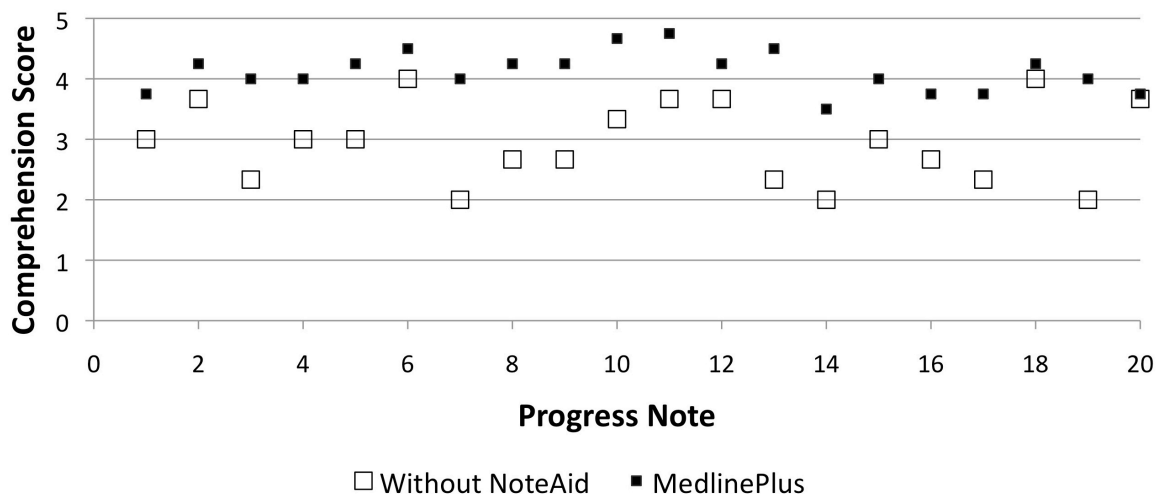
Figure 20: Scatter plot of average self-rated comprehension scores with notes alone and with the Medline Plus NoteAid implementation

The Pearson coefficient values between the subject education level and comprehension scores are: Note Alone: 0.98, MedlinePlus: 0.31, UMLS: 0.71, Wiki: -0.47 and Hybrid: 0.04.

Discussion

According to the average Flesch-Kincaid Grade Level shown in Table 31, DSs are easier to comprehend than the PGNs, corresponding to a 8th and 9th grade education, respectively. Our results on evaluation one show that subject self-reported EHR note comprehension scores fall between 3 and 4 on a five-point scale. In contrast, all 59 of our subjects have a high school education and higher. The results suggest a gap between education level, readability and health literacy. The observation of such a literacy gap is consistent with other evaluation studies in health literacy [274]. As shown in Figure 18, our results show that text readability scores positively correlate with the comprehension

scores, suggesting that our subjects' assignment of self-comprehension scoring is consistent with the readability assessment.

Our results show that overall the NoteAid systems improve comprehension. Of all four systems, the Wiki implementation on PGNs has the highest performance and statistical significance in improving EHN comprehension. In contrast, the consumer-driven authoritative resources of the UMLS and the MedlinePlus implementations yield relatively less improvement in evaluation one. The non-significant improvement in the comprehension of DS could be due to the fact that DSs are easier to comprehend than the PGNs. The self-comprehension scores are higher in DSs than in PGNs and therefore the difference in improvement is smaller.

Content coverage may partially explain performance differences among the three external resources. As shown in Table 33, EHR notes link to more Wikipedia definitions than to UMLS. MedlinePlus has the least number of definitions available. While Wikipedia incorporates over 4 million topics and articles written in English, the content of Medline Plus and UMLS are limited. For example, we found only 900 health topics in Medline Plus. As a result, the NoteAid system that links EHR notes to Wikipedia yields the best performance.

An illustrative example is shown in the following EHR note:

Example 1: Her cardiac index is 3.6. She is off of drips. We will start on baseline Coreg. history of diabetes on 80 of Lantus a day. Would try to wean her off of the insulin infusion to a low level of Lantus with a

```
sliding scale. No evidence of bleeding. Keep the chest tubes in place.
We have started her Synthroid. From a respiratory standpoint, continue
incentive spirometry, mobilization, and oral narcotics.
```

In this EHR note, Wikipedia covers 6 concepts—"cardiac index," "Coreg," "diabetes," "lantus," "bleeding," and "synthroid" and received an average comprehension score of 4.3. In contrast, the UMLS covers three concepts—"bleeding," "Synthroid," and "oral narcotics" and received an average comprehension score of 4. Medline Plus covers only two concepts "diabetes" and "bleeding" and received the lowest average comprehension score of 2.5.

Furthermore, we found that the Wikipedia content is easier to read than the UMLS or the Medline Plus content. An example is shown below.

Example 2: `The patient's bilirubin is 1.6. He is not coagulopathic.`

The definition of "coagulopathic" is complex in the UMLS: "Hemorrhagic and thrombotic disorders that occur as a consequence of abnormalities in blood coagulation due to a variety of factors such as COAGULATION PROTEIN DISORDERS; BLOOD PLATELET DISORDERS; BLOOD PROTEIN DISORDERS or nutritional conditions" which has a Flesch-Kincaid grade level of 24.

In contrast, its Wikipedia definition - "Coagulopathy is a condition in which the blood's ability to clot is impaired. This condition can cause prolonged or excessive bleeding, which may occur spontaneously or following an injury or medical and dental procedures. The normal clotting process depends on the interplay of various proteins in the blood," —

has a Flesch-Kincaid grade level of 13 and is easier to comprehend than the UMLS definition.

The evaluation one results show that the NoteAid system that integrates all three resources did not perform as well as the Wikipedia system, although the integration outperformed both the UMLS and the Medline Plus systems. This may be explained by the fact that the addition of less readable content from UMLS and MedlinePlus hurts performance.

The results in evaluation two show that, when the clinical notes are presented alone, the self-rated comprehension scores are highly correlated (0.98 Pearson coefficient) with the education levels of the subjects. The results support the validity of self-rated comprehension scores. In contrast, the correlation results are mixed with different NoteAid implementations. While the UMLS has a correlation value of 0.71, the Medline Plus and Hybrid implementations decrease to 0.31 and 0.04. The Wiki implementation has a negative correlation: -0.47. Several factors may contribute to the results. First, the definition quality of the UMLS, Wiki, Medline Plus and Hybrid resources are not yet evaluated and it is unclear whether the definitions correctly represent the semantic meanings of the notes. Secondly, although providing definitions may help comprehension, providing too much or unnecessary information (such as Wiki) may hurt those who have a better education level.

In future work, we need to conduct a comprehensive "think aloud" evaluation study to understand the behavior of users. We will also need to evaluate the quality of definitions

of different NoteAid implementations and patient comprehension by replacing complex medical jargon with its equivalent lexical lay term variants [311,312] in EHNs.

The significant improvement of Medline Plus over the UMLS implementation in evaluation two may be attributed to the lower readability of content in UMLS. Although the improvement in comprehension of Wiki over Medline Plus implementation was not statistically significant, Wiki content may not be accurate as discussed earlier.

The evaluation two results also show that all four NoteAid implementations improved EHR note self-rated comprehension significantly over Notes alone. The results are largely consistent with our evaluation one in which NoteAid implementations was evaluated in a before-and-after fashion but there are differences between the both the evaluation results. From evaluation one, we found that the Wikipedia implementation had the largest improvement and that the Medline Plus implementation decreased the self-rated comprehension scores. Such discrepancy can be explained by the limitations of our study.

Limitations, Conclusion and Future Work

First, we report subjects' self-rated note comprehension but did not evaluate to what extent they accurately comprehended the note content. The evaluation one design may be a better model as we force a subject to read the EHR note prior to her/his exposure to the improved note (note+NoteAid). A randomized design, as we have done in evaluation two, may provide an evaluation subject little incentive for comprehending the note

content. In future work, we will test subjects' comprehension based on content analyses of every clinical note. Furthermore, we will evaluate subjects' health literacy [313].

Secondly, the number of subjects in these studies is small. As a result, we can't evaluate the impact of moderators. For example, the data size is not well rounded to conclude that the subjects' education levels impact self-rated comprehension scores.

Third, our NoteAid implementations link EHR notes to definitions only, not to other education materials that MedlinePlus additionally provides. Other limitations of the study include that lay people performed our evaluation but not patients who comprehend their own EHR notes.

We have shown in this study the development of the NoteAid systems. We have evaluated four NoteAid implementations: linking EHR notes to MedlinePlus, the UMLS, Wikipedia, and Hybrid (all combined). Evaluation one results show that the system that links EHR notes to Wikipedia and the Hybrid system that links EHR notes to all three knowledge resources yield the best performance. Although Medline Plus and the UMLS are designed to facilitate consumer-oriented health information, they both need to improve their content coverage as well as readability. In contrast, Wikipedia has a broader coverage of health information as well as easy-to-read content. Evaluation two shows that Medline Plus implementation demonstrated the highest improvement.

In the future, we plan to access and improve the effectiveness of the concept filtering and coverage to improve the performance of the system. In addition, we hope to evaluate the quality of the definition provided by various educational resources and evaluate the system in a real health care setting, the next step towards building a clinical application.

Chapter 7: Integrating Components into a Unified System

In previous chapters, we discussed the development of systems for adverse event and medication-related named entity recognition (component 1), inferring causality by automating Naranjo Causality Assessment Probability Scale (component 2), generation of figure evidence by extracting relevant figures from biomedical literature and summarizing them with a figure summarization system (component 3), and processing narrative text to identify complex medical jargon and provide definitions and explanations to better comprehend the text (component 4). The final goal of the study was the development of a user interface integrating all four components. The interface helps users to see the named entities recognized, connectives identified and the Naranjo Score calculated automatically. ADEView also allows the user to view figures related to the ADE and view the text processed by NoteAid to better comprehend the EMR notes.

System Implemented

The final objective of this study is to combine these systems to develop a comprehensive application that would be made freely available online so that the general public can access and make use of this application. In this section, we describe the system developed for this dissertation – ADEtector.

ADEtector

The ADEtector application integrates all the system components and provides a common way to see the output of all the systems implemented in this dissertation through *ADEView*. As described earlier, *ADEView* displays the output of four components:

Named Entity Recognizer, Causality Inference Engine, Figure Evidence Generator and NoteAid system. Given an input report, the application processes the report and shows the output of each of component in individual tabs as shown in Figure 21 below.

The Named Entity Recognizer identifies all the entities and the UI displays all the entities that appear in the text using techniques as discussed in Chapter 2. Each entity is highlighted in a different color to distinguish between various entities as shown in Figure 21.

Figure 21: Screen shot of the ADEtector system showing the output of the named entity recognizer component. Each of the entities recognized is highlighted in different colors.

We also display the output of a discourse connective identifier with sense detector, which is used as features for NER task and can be used for automation of remaining elements of Naranjo Scale. Figure 22 shows the output of the discourse connective identifier. The interface shows the connective as hyperlinked text and when the user hovers the mouse over it, the class-wise sense of the connective is shown. The second component Causality Inference Engine consists of automated Naranjo Causality Assessment Probability Scale, which calculates the score of an adverse event related to a drug. Figure 23 shows the

output of the Naranjo Scale. The tool identified "aspirin" caused "GI bleed" and assigned

a score of 3 to it.
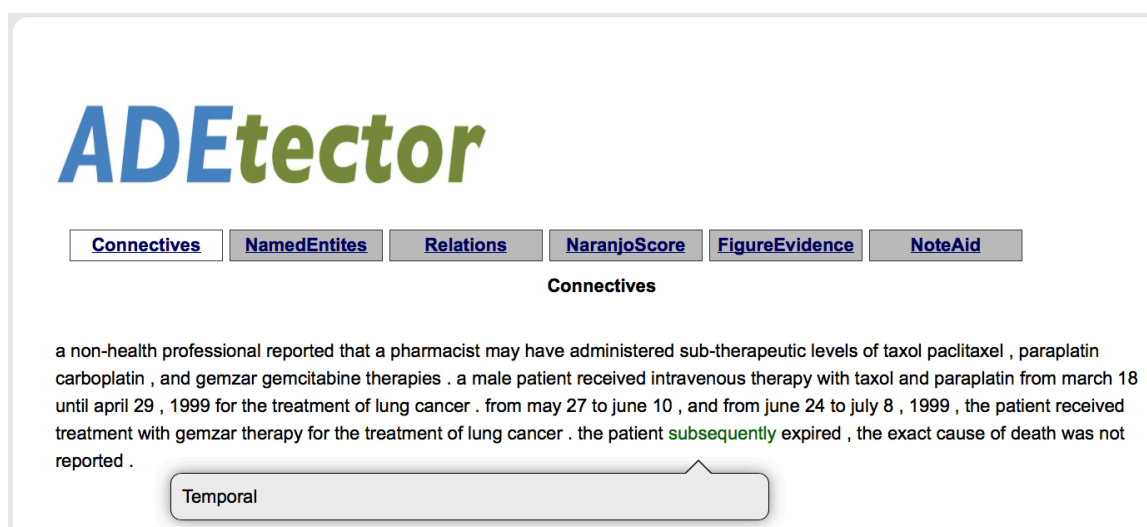
Figure 22: Screen shot of the ADEtector showing the output of the discourse connective identifier module. The interface shows the connective identified as hyperlinked text and when the user hovers the mouse over the text, a pop-up box shows the class-wise sense of the connective identified.
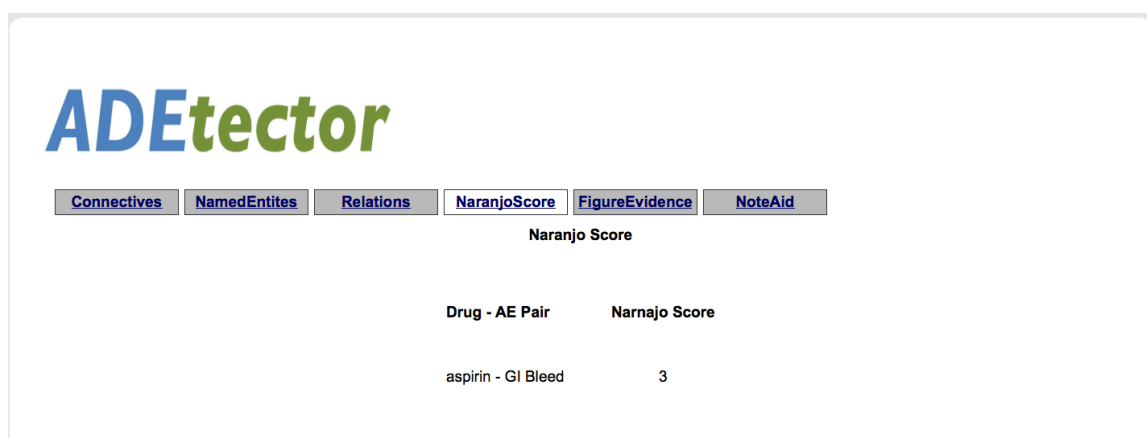
Figure 23: Screen shot of ADEtector showing the output of the Naranjo Causality Assessment Probability Scale. The tool identified GI bleed is related to aspirin and assigned a score of 3.

The third component, Figure Evidence Generator, searches evidence for ADE detected,

from the biomedical literature and then presents the user with figures along with their

summaries generated by the figure summarizer. Figure 24 below shows the screen shot of the output from the Figure Evidence Generator. It shows a grid of all the figures related to the ADE extracted from the biomedical literature. When a user clicks a figure, it shows the figure with its caption and summary along with other article-related information such as title, author information and abstract of the article as shown in Figure 25.

Figure 24: Screen shot of the ADEtector interface showing the figure evidence of the ADE that was detected by the previous components. The interface shows all the figures related to the ADE. When the user clicks on a figure, then the system shows the figure along with its summary.

Figure 25: Screen shot of the interface showing the figure along with its caption, summary and other article information.

The fourth and last component is NoteAid. This component identifies complex medical jargon in the text and provides explanations to them. Figure 26 below shows the interface with medical concepts identified as hyperlinked text. When the user hovers the mouse over the concept, the explanation of the concept appears in a pop-up box. The figure below shows the explanation of the concept "Paraplatin" in the pop-up box when the user hovers the mouse.
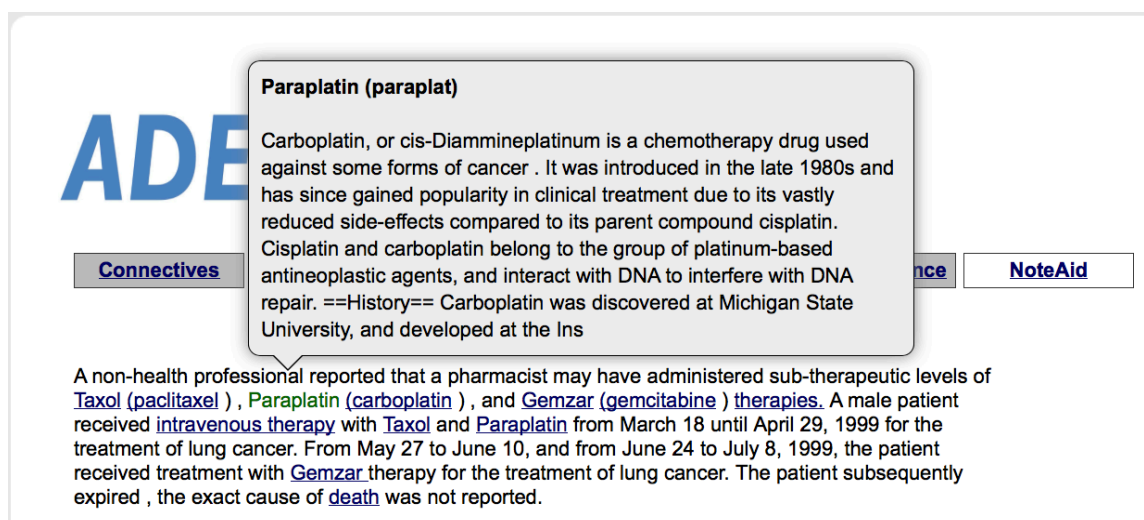
Figure 26: Screen shot of the ADEtector showing the output of the NoteAid component. The interface shows the medical concepts identified as hyperlinked text and when the user hovers the mouse over a concept, its explanation is shown.

Conclusion and Possible Improvements

The ADEtector application integrates various components such as the named entity recognition, causality inference, figure evidence through summarization and noteaid components. We hypothesize that such a system will help researchers and regulatory agencies discover adverse events quickly and easily. It also helps physicians to explain ADEs to patients and improve patient-physician communication.

There are several improvements that can be made so that these applications are more useful and attractive to researchers. For the Named Entity Recognizer, we could explore semi-supervised machine-learning approaches to further improve the performance of the system. Then we can normalize the identified entities to ontology. After normalizing, we can provide more information about the entity such as its synonyms and its position in the ontology to help researchers and regulatory agencies understand the ADE.

The Naranjo Score is currently shown as a table, and we can enhance the user experience by graphically linking the drug and the adverse event and showing the information about the ADE and the Naranjo Score on the link.

The figure evidence component can be further improved by extracting words appearing in the figure itself as keywords. This requires the use of sophisticated optical character recognition (OCR) software because many figures have poor resolution and text can often appear mixed with the biological sample image. Future work should also focus on retrieving more relevant figure to the ADE.

Currently, the NoteAid system only provides users with the definition, but we can further improve user experience by providing users with more useful information regarding the clinical concepts depending on the task, such as providing information about pharmaceutical properties of drugs and causes of adverse events to help researchers make better decisions regarding the ADE.

References

[1]   L. T. Kohn, J. M. Corrigan, M. S. Donaldson, and others, *To err is human: building a safer health system. A report of the Committee on Quality of Health Care in America, Institute of Medicine*. Washington, DC: National Academy Press, 2000.

[2]   "E 2 A Clinical Safety Data Management: Definitions and Standards for Expedited Reporting - WC500002749.pdf." .

[3]   D. W. Bates, D. J. Cullen, N. Laird, L. A. Petersen, S. D. Small, D. Servi, G. Laffel, B. J. Sweitzer, B. F. Shea, R. Hallisey, and others, "Incidence of adverse drug events and potential adverse drug events," *JAMA: the journal of the American Medical Association*, vol. 274, no. 1, p. 29, 1995.

[4]   J. R. Nebeker, P. Barach, and M. H. Samore, "Clarifying adverse drug events: a clinician's guide to terminology, documentation, and reporting," *Ann. Intern. Med.*, vol. 140, no. 10, pp. 795–801, May 2004.

[5]   A. J. Forster, R. B. Halil, and M. G. Tierney, "Pharmacist surveillance of adverse drug events," *American journal of health-system pharmacy*, vol. 61, no. 14, p. 1466, 2004.

[6]   E. J. Thomas, D. M. Studdert, H. R. Burstin, E. J. Orav, T. Zeena, E. J. Williams, K. M. Howard, P. C. Weiler, and T. A. Brennan, "Incidence and types of adverse events and negligent care in Utah and Colorado," *Medical Care*, vol. 38, no. 3, p. 261, 2000.

[7]   L. L. Leape, T. A. Brennan, N. Laird, A. G. Lawthers, A. R. Localio, B. A. Barnes, L. Hebert, J. P. Newhouse, P. C. Weiler, and H. Hiatt, "The nature of adverse events in hospitalized patients," *New England Journal of Medicine*, vol. 324, no. 6, pp. 377–384, 1991.

[8]   D. C. Classen, S. L. Pestotnik, R. S. Evans, J. F. Lloyd, and J. P. Burke, "Adverse drug events in hospitalized patients," *JAMA: the journal of the American Medical Association*, vol. 277, no. 4, p. 301, 1997.

[9]   D. J. Cullen, B. J. Sweitzer, D. W. Bates, E. Burdick, A. Edmondson, and L. L. Leape, "Preventable adverse drug events in hospitalized patients: a comparative study of intensive care and general care units," *Critical care medicine*, vol. 25, no. 8, p. 1289, 1997.

[10]   D. J. Cullen, D. W. Bates, S. D. Small, J. B. Cooper, A. R. Nemeskal, and L. L. Leape, "The incident reporting system does not detect adverse drug events: a problem for quality improvement.," *The Joint Commission journal on quality improvement*, vol. 21, no. 10, p. 541, 1995.

[11]   D. W. Bates, N. Spell, D. J. Cullen, E. Burdick, N. Laird, L. A. Petersen, S. D. Small, B. J. Sweitzer, and L. L. Leape, "The costs of adverse drug events in hospitalized patients," *JAMA: the journal of the American Medical Association*, vol. 277, no. 4, p. 307, 1997.

[12]   "FDA Adverse Events Reporting System (FAERS) > FDA Adverse Event Reporting System (FAERS) Statistics." [Online]. Available: http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/ucm070093.htm. [Accessed: 18-Apr-2013].

[13]   N. P. Tatonetti, P. P. Ye, R. Daneshjou, and R. B. Altman, "Data-driven prediction of drug effects and interactions," *Sci Transl Med*, vol. 4, no. 125, p. 125ra31, Mar. 2012.

[14]   C. Naranjo, U. Busto, E. Sellers, P. Sandor, I. Ruiz, E. Roberts, E. Janecek, C. Domecq, and D. Greenblatt, "A method for estimating the probability of adverse drug reactions," *Clinical Pharmacology & Therapeutics*, vol. 30, no. 2, pp. 239–245, 1981.

[15]   S. Belknap, H. Moore, S. Lanzotti, P. Yarnold, M. Getz, D. Deitrick, A. Peterson, J. Akeson, T. Maurer, R. Soltysik, and others, "Application of software design principles and debugging methods to an analgesia prescription reduces risk of severe injury from medical use of opioids," *Clinical Pharmacology & Therapeutics*, vol. 84, no. 3, pp. 385–392, 2008.

[16]   S. Belknap, C. Georgopoulos, D. West, P. Yarnold, and W. Kelly, "Quality of methods for assessing and reporting serious adverse events in clinical trials of cancer drugs," *Clinical Pharmacology & Therapeutics*, vol. 88, no. 2, pp. 231–236, 2010.

[17]   D. C. Classen, S. L. Pestotnik, R. Evans, and J. Burke, "Description of a computerized adverse drug event monitor using a hospital information system," *Hospital Pharmacy*, vol. 27, pp. 774–774, 1992.

[18]   R. S. Evans, S. L. Pestotnik, D. C. Classen, S. D. Horn, S. B. Bass, and J. P. Burke, "Preventing adverse drug events in hospitalized patients," *Annals of Pharmacotherapy*, vol. 28, no. 4, pp. 523–527, 1994.

[19]   H. Yu and M. Lee, "Accessing bioscience images from abstract sentences," *Bioinformatics*, vol. 22, no. 14, pp. e547–556, Jul. 2006.

[20]   H. H. Hiatt, B. A. Barnes, T. A. Brennan, N. M. Laird, A. G. Lawthers, L. L. Leape, A. R. Localio, J. P. Newhouse, L. M. Peterson, K. E. Thorpe, and others, "A study of medical injury and medical malpractice," *New England Journal of Medicine*, vol. 321, no. 7, pp. 480–484, 1989.

[21]   T. A. Brennan, L. L. Leape, N. M. Laird, L. Hebert, A. R. Localio, A. G. Lawthers, J. P. Newhouse, P. C. Weiler, and H. H. Hiatt, "Incidence of adverse events and negligence in hospitalized patients," *New England journal of medicine*, vol. 324, no. 6, pp. 370–376, 1991.

[22]   A. K. Jha, G. J. Kuperman, J. M. Teich, L. Leape, B. Shea, E. Rittenberg, E. Burdick, D. L. Seger, M. V. Vliet, and D. W. Bates, "Identifying Adverse Drug Events," *Journal of the American Medical Informatics Association*, vol. 5, no. 3, pp. 305 –314, May 1998.

[23]   H. J. Murff, A. J. Forster, J. F. Peterson, J. M. Fiskio, H. L. Heiman, and D. W. Bates, "Electronically screening discharge summaries for adverse medical events," *Journal of the American Medical Informatics Association*, vol. 10, no. 4, pp. 339–350, 2003.

[24]   S. H. Hwang, S. Lee, H. K. Koo, and Y. Kim, "Evaluation of a computer-based adverse-drug-event monitor," *Am J Health Syst Pharm*, vol. 65, no. 23, pp. 2265–2272, 2008.

[25]   A. Tinoco, R. S. Evans, C. J. Staes, J. F. Lloyd, J. M. Rothschild, and P. J. Haug, "Comparison of computerized surveillance and manual chart review for adverse events," *J Am Med Inform Assoc*, vol. 18, no. 4, pp. 491–497, Aug. 2011.

[26] R. M. Gardner, T. A. Pryor, and H. R. Warner, "The HELP hospital information system: update 1998," *Int J Med Inform*, vol. 54, no. 3, pp. 169–182, Jun. 1999.

[27] L. Ohno-Machado, P. Nadkarni, and K. Johnson, "Natural language processing: algorithms and tools to extract computable information from EHRs and from the biomedical literature," *J Am Med Inform Assoc*, vol. 20, no. 5, pp. 805–805, Sep. 2013.

[28] B. Hazlehurst, H. R. Frost, D. F. Sittig, and V. J. Stevens, "MediClass: A system for detecting and classifying encounter-based clinical events in any electronic medical record," *J Am Med Inform Assoc*, vol. 12, no. 5, pp. 517–529, Oct. 2005.

[29] B. Hazlehurst, A. Naleway, and J. Mullooly, "Detecting possible vaccine adverse events in clinical notes of the electronic medical record," *Vaccine*, vol. 27, no. 14, pp. 2077–2083, Mar. 2009.

[30] X. Wang, G. Hripcsak, M. Markatou, and C. Friedman, "Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study," *J Am Med Inform Assoc*, vol. 16, no. 3, pp. 328–337, Jun. 2009.

[31] K. Haerian, D. Varn, S. Vaidya, L. Ena, H. S. Chase, and C. Friedman, "Detection of pharmacovigilance-related adverse events using electronic health records and automated methods," *Clinical Pharmacology & Therapeutics*, 2012.

[32] G. B. Melton and G. Hripcsak, "Automated detection of adverse events using natural language processing of discharge summaries," *J Am Med Inform Assoc*, vol. 12, no. 4, pp. 448–457, Aug. 2005.

[33] C. Friedman, P. O. Alderson, J. H. Austin, J. J. Cimino, and S. B. Johnson, "A general natural-language text processor for clinical radiology.," *J Am Med Inform Assoc*, vol. 1, no. 2, pp. 161–174, 1994.

[34] E. Aramaki, Y. Miura, M. Tonoike, T. Ohkuma, H. Masuichi, K. Waki, and K. Ohe, "Extraction of Adverse Drug Effects from Clinical Records," *Studies in health technology and informatics*, vol. 160, no. Pt 1, pp. 739–743, 2010.

[35] C. Yang, P. Srinivasan, and P. M. Polgreen, "Automatic adverse drug events detection using letters to the editor," *AMIA Annu Symp Proc*, vol. 2012, pp. 1030–1039, 2012.

[36] H. Gurulingappa, A. Mateen-Rajput, and L. Toldo, "Extraction of potential adverse drug events from medical case reports," *J Biomed Semantics*, vol. 3, no. 1, p. 15, 2012.

[37] K. W. Fung, C. S. Jao, and D. Demner-Fushman, "Extracting drug indication information from structured product labels using natural language processing," *J Am Med Inform Assoc*, vol. 20, no. 3, pp. 482–488, May 2013.

[38] B. W. Chee, R. Berlin, and B. Schatz, "Predicting adverse drug events from personal health messages," *AMIA, page 10pp, Washington, DC*, 2011.

[39] R. W. White, N. P. Tatonetti, N. H. Shah, R. B. Altman, and E. Horvitz, "Web-scale pharmacovigilance: listening to signals from the crowd," *J Am Med Inform Assoc*, Mar. 2013.

[40] K. D. Shetty and S. R. Dalal, "Using information mining of the medical literature to improve drug safety," *Journal of the American Medical Informatics Association*, 2011.

[41] S. M. Meystre, J. Thibault, S. Shen, J. F. Hurdle, and B. R. South, "Textractor: a hybrid system for medications and reason for their prescription extraction from clinical text documents," *J Am Med Inform Assoc*, vol. 17, no. 5, pp. 559–562, Oct. 2010.

[42] I. Spasic, F. Sarafraz, J. A. Keane, and G. Nenadic, "Medication information extraction with linguistic pattern matching and semantic rules," *J Am Med Inform Assoc*, vol. 17, no. 5, pp. 532–535, Oct. 2010.

[43] H. Xu, S. P. Stenner, S. Doan, K. B. Johnson, L. R. Waitman, and J. C. Denny, "MedEx: a medication information extraction system for clinical narratives," *J Am Med Inform Assoc*, vol. 17, no. 1, pp. 19–24, Feb. 2010.

[44] S. Doan, L. Bastarache, S. Klimkowski, J. C. Denny, and H. Xu, "Integrating existing natural language processing tools for medication extraction from discharge summaries," *J Am Med Inform Assoc*, vol. 17, no. 5, pp. 528–531, Oct. 2010.

[45] S. M. Meystre, J. Thibault, S. Shen, J. F. Hurdle, and B. R. South, "Automatically detecting medications and the reason for their prescription in clinical narrative text documents," *Stud Health Technol Inform*, vol. 160, no. Pt 2, pp. 944–948, 2010.

[46] Ö. Uzuner, I. Solti, and E. Cadag, "Extracting medication information from clinical text," *Journal of the American Medical Informatics Association*, vol. 17, no. 5, p. 514, 2010.

[47] M. Lindquist, "VigiBase, the WHO Global ICSR Database System: Basic Facts," *Drug Information Journal*, vol. 42, no. 5, pp. 409–419, Sep. 2008.

[48] M. Hauben, D. Madigan, C. M. Gerrits, L. Walsh, and E. P. Van Puijenbroek, "The role of data mining in pharmacovigilance," *Expert Opin Drug Saf*, vol. 4, no. 5, pp. 929–948, Sep. 2005.

[49] M. Hauben and A. Bate, "Decision support methods for the detection of adverse events in post-marketing data," *Drug Discovery Today*, vol. 14, no. 7–8, pp. 343–357, Apr. 2009.

[50] R. Harpaz, W. DuMouchel, N. H. Shah, D. Madigan, P. Ryan, and C. Friedman, "Novel Data-Mining Methodologies for Adverse Drug Event Discovery and Analysis," *Clin Pharmacol Ther*, vol. 91, no. 6, pp. 1010–1021, Jun. 2012.

[51] A. Bate and S. J. W. Evans, "Quantitative signal detection using spontaneous ADR reporting," *Pharmacoepidemiol Drug Saf*, vol. 18, no. 6, pp. 427–436, Jun. 2009.

[52] J. S. Almenoff, E. N. Pattishall, T. G. Gibbs, W. DuMouchel, S. J. W. Evans, and N. Yuen, "Novel statistical tools for monitoring the safety of marketed drugs," *Clin. Pharmacol. Ther.*, vol. 82, no. 2, pp. 157–166, Aug. 2007.

[53] S. J. Evans, P. C. Waller, and S. Davis, "Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports," *Pharmacoepidemiol Drug Saf*, vol. 10, no. 6, pp. 483–486, Nov. 2001.

[54] E. P. van Puijenbroek, A. Bate, H. G. M. Leufkens, M. Lindquist, R. Orre, and A. C. G. Egberts, "A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions," *Pharmacoepidemiol Drug Saf*, vol. 11, no. 1, pp. 3–10, Feb. 2002.

[55] G. N. Norén, R. Sundberg, A. Bate, and I. R. Edwards, "A statistical methodology for drug-drug interaction surveillance," *Stat Med*, vol. 27, no. 16, pp. 3057–3070, Jul. 2008.

[56] R. Harpaz, S. Vilar, W. DuMouchel, H. Salmasian, K. Haerian, N. H. Shah, H. S. Chase, and C. Friedman, "Combing signals from spontaneous reports and electronic health records for detection of adverse drug reactions," *J Am Med Inform Assoc*, vol. 20, no. 3, pp. 413–419, May 2013.

[57] T. Sakaeda, K. Kadoyama, and Y. Okuno, "Adverse event profiles of platinum agents: Data mining of the public version of the FDA adverse event reporting system, AERS, and reproducibility of clinical observations," *International Journal of Medical Sciences*, vol. 8, no. 6, p. 487, 2011.

[58] S. Vilar, R. Harpaz, H. S. Chase, S. Costanzi, R. Rabadan, and C. Friedman, "Facilitating adverse drug event detection in pharmacovigilance databases using molecular structure similarity: application to rhabdomyolysis," *Journal of the American Medical Informatics Association*, 2011.

[59] G. N. Norén, A. Bate, R. Orre, and I. R. Edwards, "Extending the methods used to screen the WHO drug safety database towards analysis of complex associations and improved accuracy for rare events," *Stat Med*, vol. 25, no. 21, pp. 3740–3757, Nov. 2006.

[60] A. M. Hochberg, M. Hauben, R. K. Pearson, D. J. O'Hara, S. J. Reisinger, D. I. Goldsmith, A. L. Gould, and D. Madigan, "An evaluation of three signal-detection algorithms using a highly inclusive reference event database," *Drug Saf*, vol. 32, no. 6, pp. 509–525, 2009.

[61] A. Bate, M. Lindquist, I. R. Edwards, and R. Orre, "A data mining approach for signal detection and analysis," *Drug Saf,* vol. 25, no. 6, pp. 393–397, 2002.

[62] A. Bate, "Bayesian confidence propagation neural network," *Drug Saf,* vol. 30, no. 7, pp. 623–625, 2007.

[63] W. DuMouchel and D. Pregibon, "Empirical bayes screening for multi-item associations," in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining,* New York, NY, USA, 2001, pp. 67–76.

[64] A. Bate, M. Lindquist, I. R. Edwards, S. Olsson, R. Orre, A. Lansner, and R. M. De Freitas, "A Bayesian neural network method for adverse drug reaction signal generation," *Eur. J. Clin. Pharmacol.*, vol. 54, no. 4, pp. 315–321, Jun. 1998.

[65] A. Genkin, D. D. Lewis, and D. Madigan, "Large-Scale Bayesian Logistic Regression for Text Categorization," *Technometrics*, vol. 49, no. 3, pp. 291–304, Aug. 2007.

[66] O. Caster, G. N. Norén, D. Madigan, and A. Bate, "Large-scale regression-based pattern discovery: The example of screening the WHO global drug safety database," *Stat. Anal. Data Min.*, vol. 3, no. 4, pp. 197–208, Aug. 2010.

[67] R. Solomon and W. Dumouchel, "Contrast media and nephropathy: findings from systematic analysis and Food and Drug Administration reports of adverse effects," *Invest Radiol*, vol. 41, no. 8, pp. 651–660, Aug. 2006.

[68] M. Rouane-Hacene, Y. Toussaint, and P. Valtchev, "Mining Safety Signals in Spontaneous Reports Database Using Concept Analysis," in *Artificial Intelligence in Medicine*, vol. 5651, C. Combi, Y. Shahar, and A. Abu-Hanna, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 285–294.

[69] R. Harpaz, H. S. Chase, and C. Friedman, "Mining multi-item drug adverse effect associations in spontaneous reporting systems," *BMC Bioinformatics*, vol. 11 Suppl 9, p. S7, 2010.

[70] R. Harpaz, H. Perez, H. S. Chase, R. Rabadan, G. Hripcsak, and C. Friedman, "Biclustering of adverse drug events in the FDA's spontaneous reporting system," *Clin. Pharmacol. Ther.*, vol. 89, no. 2, pp. 243–250, Feb. 2011.

[71] R. Ball and T. Botsis, "Can network analysis improve pattern recognition among adverse events following immunization reported to VAERS?," *Clin. Pharmacol. Ther.*, vol. 90, no. 2, pp. 271–278, Aug. 2011.

[72] R. Harpaz, W. DuMouchel, P. LePendu, A. Bauer-Mehren, P. Ryan, and N. H. Shah, "Performance of Pharmacovigilance Signal-Detection Algorithms for the FDA Adverse Event Reporting System," *Clin Pharmacol Ther*, Apr. 2013.

[73] E. Yom-Tov and E. Gabrilovich, "Postmarket Drug Surveillance Without Trial Costs: Discovery of Adverse Drug Reactions Through Large-Scale Analysis of Web Search Queries," *Journal of Medical Internet Research*, vol. 15, no. 6, p. e124, Jun. 2013.

[74] J. Cohen, "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, Apr. 1960.

[75] "RADAR (Research on Adverse Drug events And Reports)." [Online]. Available: http://cancer.northwestern.edu/Research/research_programs/radar/index.cfm. [Accessed: 11-Apr-2013].

[76] Z. Li, F. Liu, L. Antieau, Y. Cao, and H. Yu, "Lancet: a high precision medication event extraction system for clinical text," *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 563–567, 2010.

[77] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-*, 2001, pp. 282–289.

[78] "Weka 3 - Data Mining with Open Source Machine Learning Software in Java." [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/. [Accessed: 29-Aug-2013].

[79]   B. Settles, "ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text," *Bioinformatics*, vol. 21, no. 14, p. 3191, 2005.

[80]   S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *The Journal of Machine Learning Research*, vol. 2, pp. 45–66, 2002.

[81]   T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," *Machine Learning: ECML-98*, pp. 137–142, 1998.

[82]   R. Leaman and G. Gonzalez, "BANNER: an executable survey of advances in biomedical named entity recognition," *Pac Symp Biocomput*, pp. 652–663, 2008.

[83]   E. Charniak and M. Johnson, "Coarse-to-fine n-best parsing and MaxEnt discriminative reranking," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, 2005, pp. 173–180.

[84]   D. McClosky, "Any domain parsing: Automatic domain adaptation for natural language parsing," Ph. D. thesis, Department of Computer Science, Brown University, 2009.

[85]   A. R. Aronson, "Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program.," in *Proceedings of the AMIA Symposium*, 2001, p. 17.

[86]   S. Agarwal and H. Yu, "Biomedical negation scope detection with conditional random fields," *Journal of the American Medical Informatics Association*, vol. 17, no. 6, p. 696, 2010.

[87]   S. Agarwal and H. Yu, "Detecting hedge cues and their scope in biomedical literature with conditional random fields," *Journal of biomedical informatics*, 2010.

[88]   B. Polepalli Ramesh, R. Prasad, T. Miller, B. Harrington, and H. Yu, "Automatic discourse connective detection in biomedical text," *J Am Med Inform Assoc*, vol. 19, no. 5, pp. 800–808, Sep. 2012.

[89] M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini, "Building a large annotated corpus of English: The Penn Treebank," *Computational linguistics*, vol. 19, no. 2, pp. 313–330, 1993.

[90] J. Zheng, W. W. Chapman, R. S. Crowley, and G. K. Savova, "Coreference Resolution: A Review of General Methodologies and Applications in the Clinical Domain," *Journal of Biomedical Informatics*, 2011.

[91] R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and B. Webber, "The penn discourse treebank 2.0," in *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, 2008, pp. 2961–2968.

[92] D. Marcu, "Improving summarization through rhetorical parsing tuning," in *The 6th Workshop on Very Large Corpora*, 1998, pp. 206–215.

[93] E. H. Hovy, "Automated discourse generation using discourse structure relations," *Artificial intelligence*, vol. 63, no. 1–2, pp. 341–385, 1993.

[94] H. Hernault, P. Piwek, H. Prendinger, and M. Ishizuka, "Generating dialogues for virtual agents using nested textual coherence relations," in *Intelligent Virtual Agents*, 2008, pp. 139–145.

[95] R. McDonald, K. Hannan, T. Neylon, M. Wells, and J. Reynar, "Structured models for fine-to-coarse sentiment analysis," in *ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS*, 2007, vol. 45, p. 432.

[96] P. Mannem, R. Prasad, and A. Joshi, "Question generation from paragraphs at UPenn: QGSTEC system description," in *Proceedings of QG2010: The Third Workshop on Question Generation*, 2010.

[97] B. MacCartney and C. D. Manning, "Natural logic for textual inference," in *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, 2007, pp. 193–200.

[98] I. Mani, M. Verhagen, B. Wellner, C. M. Lee, and J. Pustejovsky, "Machine learning of temporal relations," in *Proceedings of the 21st International*

Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, 2006, pp. 753–760.

[99] R. Prasad and A. Joshi, "A discourse-based approach to generating why-questions from texts," in *Proceedings of the Workshop on the Question Generation Shared Task and Evaluation Challenge Arlington, VA*, 2008.

[100] R. Soricut and D. Marcu, "Sentence level discourse parsing using syntactic and lexical information," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, 2003, pp. 149–156.

[101] D. A. DduVerle and H. Prendinger, "A novel discourse parser based on support vector machine classification," in *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*, 2009, pp. 665–673.

[102] R. Subba and B. Di Eugenio, "An effective discourse parser that uses rich linguistic information," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2009, pp. 566–574.

[103] B. Wellner and J. Pustejovsky, "Automatically identifying the arguments of discourse connectives," in *Proceedings of EMNLP-CoNLL*, 2007, pp. 92–101.

[104] R. Elwell and J. Baldridge, "Discourse connective argument identification with connective specific rankers," in *The IEEE International Conference on Semantic Computing*, 2008, pp. 198–205.

[105] E. Pitler and A. Nenkova, "Using syntax to disambiguate explicit discourse connectives in text," in *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, 2009, pp. 13–16.

[106] M. Light, X. Y. Qiu, and P. Srinivasan, "The language of bioscience: Facts, speculations, and statements in between," in *Proceedings of BioLink 2004 workshop on linking biological literature, ontologies and databases: tools for users*, 2004, pp. 17–24.

[107] T. Mullen, Y. Mizuta, and N. Collier, "A baseline feature set for learning rhetorical zones using full articles in the biomedical domain," *ACM SIGKDD Explorations Newsletter*, vol. 7, no. 1, pp. 52–58, 2005.

[108] D. Biber and J. K. Jones, "Merging corpus linguistic and discourse analytic research goals: Discourse units in biology research articles," *Corpus Linguistics and Linguistic Theory*, vol. 1, no. 2, pp. 151–182, 2005.

[109] G. Szarvas, V. Vincze, R. Farkas, and J. Csirik, "The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts," in *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, 2008, pp. 38–45.

[110] S. Agarwal and H. Yu, "Biomedical negation scope detection with conditional random fields," *Journal of the American Medical Informatics Association*, vol. 17, no. 6, p. 696, 2010.

[111] S. Agarwal and H. Yu, "Detecting hedge cues and their scope in biomedical literature with conditional random fields," *Journal of biomedical informatics*, 2010.

[112] R. Prasad, S. McRoy, N. Frid, A. Joshi, and H. Yu, "The biomedical discourse relation bank," *BMC bioinformatics*, vol. 12, no. 1, p. 188, 2011.

[113] H. Yu, N. Frid, S. McRoy, R. Prasad, A. Lee, and A. Joshi, "A pilot annotation to investigate discourse connectivity in biomedical text," in *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, 2008, pp. 92–93.

[114] V. Rizomilioti, "Exploring epistemic modality in academic discourse using corpora," *Information technology in languages for specific purposes*, pp. 53–71, 2006.

[115] F. Lisacek, C. Chichester, A. Kaplan, and Á. Sandor, "Discovering paradigm shift patterns in biomedical abstracts: application to neurodegenerative diseases," in *first international symposium on semantic mining in biomedicine*, 2005, pp. 11–13.

[116] G. K. Savova, W. W. Chapman, J. Zheng, and R. S. Crowley, "Anaphoric relations in the clinical narrative: corpus creation," *Journal of the American Medical Informatics Association*, vol. 18, no. 4, p. 459, 2011.

[117] A. Coden, G. Savova, I. Sominsky, M. Tanenblatt, J. Masanz, K. Schuler, J. Cooper, W. Guan, and P. C. de Groen, "Automatically extracting cancer disease characteristics from pathology reports into a Disease Knowledge Representation Model," *Journal of biomedical informatics*, vol. 42, no. 5, pp. 937–949, 2009.

[118] A. Roberts, R. Gaizauskas, M. Hepple, G. Demetriou, Y. Guo, I. Roberts, and A. Setzer, "Building a semantically annotated corpus of clinical texts," *Journal of biomedical informatics*, vol. 42, no. 5, pp. 950–966, 2009.

[119] J. Castano, J. Zhang, and J. Pustejovsky, "Anaphora resolution in biomedical literature," 2002.

[120] M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini, "Building a large annotated corpus of English: The Penn Treebank," *Computational linguistics*, vol. 19, no. 2, pp. 313–330, 1993.

[121] M. Palmer, D. Gildea, and P. Kingsbury, "The proposition bank: An annotated corpus of semantic roles," *Computational Linguistics*, vol. 31, no. 1, pp. 71–106, 2005.

[122] R. Weischedel, M. Palmer, M. Marcus, E. Hovy, S. Pradhan, L. Ramshaw, N. Xue, A. Taylor, J. Kaufman, M. Franchini, M. El-Bachouti, R. Belvin, and A. Houston, "Ontonotes Release 4.0," Linguistic Data Consortium, Philadelphia, Technical Report.

[123] J. D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii, "GENIA corpus—a semantically annotated corpus for bio-textmining," *Bioinformatics*, vol. 19, no. suppl 1, p. i180, 2003.

[124] K. Verspoor, K. B. Cohen, and L. Hunter, "The textual characteristics of traditional and Open Access scientific journals are similar," *BMC bioinformatics*, vol. 10, no. 1, p. 183, 2009.

[125] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-*, 2001, pp. 282–289.

[126] B. Settles, "ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text," *Bioinformatics*, vol. 21, no. 14, p. 3191, 2005.

[127] D. McClosky, "Any domain parsing: Automatic domain adaptation for natural language parsing," Ph. D. thesis, Department of Computer Science, Brown University, 2009.

[128] R. Leaman and G. Gonzalez, "BANNER: an executable survey of advances in biomedical named entity recognition," in *Pacific Symposium on Biocomputing*, 2008, vol. 13, pp. 652–663.

[129] M. Gerner, G. Nenadic, and C. Bergman, "LINNAEUS: a species name identification system for biomedical literature," *BMC bioinformatics*, vol. 11, no. 1, p. 85, 2010.

[130] O. Bodenreider, "The unified medical language system (UMLS): integrating biomedical terminology," *Nucleic acids research*, vol. 32, no. suppl 1, p. D267, 2004.

[131] D. Dahlmeier and H. T. Ng, "Domain adaptation for semantic role labeling in the biomedical domain," *Bioinformatics*, vol. 26, no. 8, p. 1098, 2010.

[132] R. Prasad, S. McRoy, N. Frid, A. Joshi, and H. Yu, "The biomedical discourse relation bank," *BMC Bioinformatics*, vol. 12, no. 1, p. 188, May 2011.

[133] "The use of the WHO-UMC system for standardised case causality assessment." .

[134] A. F. Macedo, F. B. Marques, C. F. Ribeiro, and F. Teixeira, "Causality assessment of adverse drug reactions: comparison of the results obtained from published decisional algorithms and from the evaluations of an expert panel," *Pharmacoepidemiol Drug Saf*, vol. 14, no. 12, pp. 885–890, Dec. 2005.

[135] Y. Arimone, B. Bégaud, G. Miremont-Salamé, A. Fourrier-Réglat, N. Moore, M. Molimard, and F. Haramburu, "Agreement of expert judgment in causality assessment of adverse drug reactions," *Eur. J. Clin. Pharmacol.*, vol. 61, no. 3, pp. 169–173, May 2005.

[136] M. D. Stephens, "The diagnosis of adverse medical events associated with drug treatment," *Adverse Drug React Acute Poisoning Rev*, vol. 6, no. 1, pp. 1–35, 1987.

[137] R. H. Meyboom, Y. A. Hekster, A. C. Egberts, F. W. Gribnau, and I. R. Edwards, "Causal or casual? The role of causality assessment in pharmacovigilance," *Drug Saf*, vol. 17, no. 6, pp. 374–389, Dec. 1997.

[138] T. B. Agbabiaka, J. Savović, and E. Ernst, "Methods for causality assessment of adverse drug reactions: a systematic review," *Drug Saf*, vol. 31, no. 1, pp. 21–37, 2008.

[139] F. E. Karch and L. Lasagna, "Toward the operational identification of adverse drug reactions," *Clin. Pharmacol. Ther.*, vol. 21, no. 3, pp. 247–254, Mar. 1977.

[140] M. S. Kramer, J. M. Leventhal, T. A. Hutchinson, and A. R. Feinstein, "An algorithm for the operational assessment of adverse drug reactions. I. Background, description, and instructions for use," *JAMA*, vol. 242, no. 7, pp. 623–632, Aug. 1979.

[141] C. A. Naranjo, U. Busto, E. M. Sellers, P. Sandor, I. Ruiz, E. A. Roberts, E. Janecek, C. Domecq, and D. J. Greenblatt, "A method for estimating the probability of adverse drug reactions," *Clin. Pharmacol. Ther.*, vol. 30, no. 2, pp. 239–245, Aug. 1981.

[142] S. L. Kane-Gill, L. Kirisci, and D. S. Pathak, "Are the Naranjo criteria reliable and valid for determination of adverse drug reactions in the intensive care unit?," *Ann Pharmacother*, vol. 39, no. 11, pp. 1823–1827, Nov. 2005.

[143] T. A. Hutchinson and D. A. Lane, "Assessing methods for causality assessment of suspected adverse drug reactions," *J Clin Epidemiol*, vol. 42, no. 1, pp. 5–16, 1989.

[144] G. Danan and C. Benichou, "Causality assessment of adverse reactions to drugs--I. A novel method based on the conclusions of international consensus meetings: application to drug-induced liver injuries," *J Clin Epidemiol*, vol. 46, no. 11, pp. 1323–1330, Nov. 1993.

[145] Y. Koh and S. C. Li, "A new algorithm to identify the causality of adverse drug reactions," *Drug Saf*, vol. 28, no. 12, pp. 1159–1161, 2005.

[146] K. L. Lanctôt and C. A. Naranjo, "Computer-assisted evaluation of adverse events using a Bayesian approach," *J Clin Pharmacol*, vol. 34, no. 2, pp. 142–147, Feb. 1994.

[147] K. L. Lanctôt and C. A. Naranjo, "Comparison of the Bayesian approach and a simple algorithm for assessment of adverse drug events," *Clin. Pharmacol. Ther.*, vol. 58, no. 6, pp. 692–698, Dec. 1995.

[148] P. Zapater, J. Such, M. Pérez-Mateo, and J. F. Horga, "A new Poisson and Bayesian-based method to assign risk and causality in patients with suspected hepatic adverse drug reactions: a report of two new cases of ticlopidine-induced hepatotoxicity," *Drug Saf*, vol. 25, no. 10, pp. 735–750, 2002.

[149] D. Lane, M. Kramer, T. Hutchinson, J. Jones, and C. Naranjo, "The causality assessment of adverse drug reactions using a Bayesian approach.," *Pharmaceut Med*, vol. 2, pp. 265–283, 1987.

[150] K. L. Lanctôt, B. M. Ghajar, N. H. Shear, and C. A. Naranjo, "Improving the diagnosis of hypersensitivity reactions associated with sulfonamides," *J Clin Pharmacol*, vol. 34, no. 12, pp. 1228–1233, Dec. 1994.

[151] C. Naranjo, M. Kwok, and K. Lanctôt, "Enhanced differential diagnosis of anticonvulsant hypersensitivity reactions by an integrated Bayesian and biochemical approach.," *Clin Pharmacol Ther*, vol. 56, no. 5, pp. 564–75, Nov. 1994.

[152] Y. Arimone, B. Bégaud, G. Miremont-Salamé, A. Fourrier-Réglat, M. Molimard, N. Moore, and F. Haramburu, "A new method for assessing drug causation provided agreement with experts' judgment," *J Clin Epidemiol*, vol. 59, no. 3, pp. 308–314, Mar. 2006.

[153] Y. Koh, C. W. Yap, and S. C. Li, "A quantitative approach of using genetic algorithm in designing a probability scoring system of an adverse drug reaction assessment system," *Int J Med Inform*, vol. 77, no. 6, pp. 421–430, Jun. 2008.

[154] W. Chapman, *University of Pittsburgh NLP Repository (http://www.dbmi.pitt.edu/nlpfront)*. .

[155] R. P. Futrelle, "Handling figures in document summarization," in *Proc. of the ACL-04 Workshop: Text Summarization Branches Out*, 2004, pp. 61–65.

[156] N. C. Rowe, "Efficient caption-based retrieval of multimedia information," Monterey, California. Naval Postgraduate School, 1993.

[157] E. J. Guglielmo and N. C. Rowe, "Natural-language retrieval of images based on descriptive captions," *ACM Transactions on Information Systems (TOIS)*, vol. 14, no. 3, pp. 237–267, 1996.

[158] J. R. Smith and S.-F. Chang, "VisualSEEk: a fully automated content-based image query system," in *Proceedings of the fourth ACM international conference on Multimedia*, 1997, pp. 87–98.

[159] N. C. Rowe, "Precise and efficient retrieval of captioned images: The MARIE project," *Library Trends*, vol. 48, no. 2, pp. 475–495, 1999.

[160] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra, "Relevance feedback: a power tool for interactive content-based image retrieval," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 8, no. 5, pp. 644–655, 1998.

[161] J. Jeon, V. Lavrenko, and R. Manmatha, "Automatic image annotation and retrieval using cross-media relevance models," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, 2003, pp. 119–126.

[162] D. L. Swets and J. J. Weng, "Using discriminant eigenfeatures for image retrieval," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 18, no. 8, pp. 831–836, 1996.

[163] H. B. Kekre and S. D. Thepade, "Image Retrieval using Color-Texture Features Extracted from Walshlet Pyramid," *ICGST International Journal on Graphics, Vision and Image Processing (GVIP)*, vol. 10, pp. 9–18, 2010.

[164] H. Müller, N. Michoux, D. Bandon, and A. Geissbuhler, "A review of content-based image retrieval systems in medical applications—clinical benefits and future directions," *International journal of medical informatics*, vol. 73, no. 1, pp. 1–23, 2004.

[165] E. R. Tufte, "Envisioning information," *Optometry & Vision Science*, vol. 68, no. 4, pp. 322–324, 1991.

[166] T. Hammond, B. Eoff, B. Paulson, A. Wolin, K. Dahmen, J. Johnston, and P. Rajan, "Free-sketch recognition: putting the chi in sketching," in *CHI'08 extended abstracts on Human factors in computing systems*, 2008, pp. 3027–3032.

[167] B. Paulson and T. Hammond, "PaleoSketch: accurate primitive sketch recognition and beautification," in *Proceedings of the 13th international conference on Intelligent user interfaces*, 2008, pp. 1–10.

[168] S. Agarwal and H. Yu, "FigSum: automatically generating structured text summaries for figures in biomedical literature," in *AMIA Annual Symposium Proceedings*, 2009, vol. 2009, p. 6.

[169] D. Demner-Fushman, S. Antani, M. Simpson, and G. R. Thoma, "Annotation and retrieval of clinically relevant images," *Int J Med Inform*, vol. 78, no. 12, pp. e59–e67, Dec. 2009.

[170] H. Yu, S. Agarwal, M. Johnston, and A. Cohen, "Are figure legends sufficient? Evaluating the contribution of associated text to biomedical figure comprehension," *Journal of biomedical discovery and collaboration*, vol. 4, no. 1, p. 1, 2009.

[171] H. Yu, "Towards answering biological questions with experimental evidence: automatically identifying text that summarize image content in full-text articles," in *AMIA Annual Symposium Proceedings*, 2006, vol. 2006, p. 834.

[172] S. Agarwal and H. Yu, "Figure summarizer browser extensions for PubMed Central," *Bioinformatics*, vol. 27, no. 12, pp. 1723–1724, Jun. 2011.

[173] M. Margeli, B. Cirauqui, E. Castella, G. Tapia, C. Costa, A. Gimenez-Capitan, A. Barnadas, M. S. Ronco, S. Benlloch, M. Taron, and R. Rosell, "The Prognostic Value of BRCA1 mRNA Expression Levels Following Neoadjuvant Chemotherapy in Breast Cancer," *PLoS One*, vol. 5, no. 3, Mar. 2010.

[174] *Evaluating Natural Language Processing Systems: An Analysis and Review*. Springer, 1995.

[175] I. Mani, *Automatic summarization*, vol. 3. John Benjamins Publishing Company, 2001.

[176] A. Nenkova, S. Maskey, and Y. Liu, "Automatic Summarization," in *Tutorial Abstracts of ACL 2011*, Portland, Oregon, USA, 2011, p. 3.

[177] H. P. Luhn, "The automatic creation of literature abstracts," *IBM Journal of research and development*, vol. 2, no. 2, pp. 159–165, 1958.

[178] M. Brunn, Y. Chali, and C. Pinchak, "Text Summarization Using Lexical Chains," in *in Document Understanding Conference (DUC*, 2001, pp. 135–140.

[179] E. Hovy and C. Y. Lin, "Automated text summarization in SUMMARIST," *Advances in Automatic Text Summarization*, pp. 81–94, 1999.

[180] J. Kupiec, J. Pedersen, and F. Chen, "A trainable document summarizer," in *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, 1995, pp. 68–73.

[181] J. M. Conroy, J. D. Schlesinger, and D. P. O'Leary, "Topic-focused multi-document summarization using an approximate oracle score," in *Proceedings of*

the COLING/ACL on Main conference poster sessions, Stroudsburg, PA, USA, 2006, pp. 152–159.

[182] S. Gupta, A. Nenkova, and D. Jurafsky, "Measuring importance and query relevance in topic-focused multi-document summarization," in *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, Stroudsburg, PA, USA, 2007, pp. 193–196.

[183] P. I. Nakov, A. S. Schwartz, and M. A. Hearst, "Citances: Citation Sentences for Semantic Analysis of Bioscience Text," in *In Proceedings of the SIGIR'04 workshop on Search and Discovery in Bioinformatics*, 2004.

[184] H. P. Edmundson, "New methods in automatic extracting," *Journal of the ACM (JACM)*, vol. 16, no. 2, pp. 264–285, 1969.

[185] S. H. Myaeng and D. H. Jang, "Development and evaluation of a statistically-based document summarization system," *Advances in automatic text summarization*, pp. 61–70, 1999.

[186] C. Aone, M. Okurowski, J. Gorlinsky, and B. Larsen, "A trainable summarizer with knowledge acquired from robust nlp techniques.," in *Advances in Automatic Text Summarization*, I. Mani and M. Maybury, Eds. MIT Press, 1999, pp. 71–80.

[187] E. Filatova and V. Hatzivassiloglou, "A formal model for information selection in multi-sentence text extraction," in *Proceedings of the 20th international conference on Computational Linguistics*, Stroudsburg, PA, USA, 2004.

[188] T. Kikuchi, S. Furui, and C. Hori, "Automatic speech summarization based on sentence extraction and compaction," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03)*, 2003, vol. 1, pp. I–384–I–387 vol.1.

[189] Y. Gong, "Generic text summarization using relevance measure and latent semantic analysis," in *in Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2001.

[190] J. Y. Yeh, H. R. Ke, W. P. Yang, I. Meng, and others, "Text summarization using a trainable summarizer and latent semantic analysis* 1," *Information processing & management*, vol. 41, no. 1, pp. 75–95, 2005.

[191] G. A. Miller, "WordNet: a lexical database for English," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[192] R. Mihalcea, "Graph-based ranking algorithms for sentence extraction, applied to text summarization," in *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, Stroudsburg, PA, USA, 2004.

[193] G. Erkan and D. R. Radev, "LexRank: graph-based lexical centrality as salience in text summarization," *J. Artif. Int. Res.*, vol. 22, no. 1, pp. 457–479, Dec. 2004.

[194] D. R. Radev, H. Jing, M. Sty, and D. Tam, "Centroid-based summarization of multiple documents," *Information Processing & Management*, vol. 40, no. 6, pp. 919–938, 2004.

[195] C. Wang, F. Jing, L. Zhang, and H.-J. Zhang, "Learning query-biased web page summarization," in *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, New York, NY, USA, 2007, pp. 555–562.

[196] T. Hirao, H. Isozaki, E. Maeda, and Y. Matsumoto, "Extracting important sentences with support vector machines," in *Proceedings of the 19th international conference on Computational linguistics - Volume 1*, Stroudsburg, PA, USA, 2002, pp. 1–7.

[197] J. M. Conroy and D. P. O'leary, "Text summarization via hidden markov models," in *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, 2001, pp. 406–407.

[198] J. Leskovec, M. Grobelnik, and N. Milic-Frayling, "Learning sub-structures of document semantic graphs for document summarization," in *LinkKDD Workshop*, 2004, pp. 133–138.

[199] J. H. Chiang, J. W. Shin, H. H. Liu, and C. L. Chin, "GeneLibrarian: an effective gene-information summarization and visualization system," *BMC bioinformatics*, vol. 7, no. 1, p. 392, 2006.

[200] X. Ling, J. Jiang, X. He, Q. Mei, C. Zhai, and B. Schatz, "Generating gene summaries from biomedical literature: A study of semi-structured summarization," *Information Processing & Management*, vol. 43, no. 6, pp. 1777–1791, 2007.

[201] F. Jin, M. Huang, Z. Lu, and X. Zhu, "Towards automatic generation of gene summary," in *Proceedings of the Workshop on BioNLP*, 2009, pp. 97–105.

[202] S. Bhattacharya and others, "MeSH: a window into full text for document summarization," *Bioinformatics*, vol. 27, no. 13, p. i120, 2011.

[203] L. Plaza and J. Carrillo-de-Albornoz, "Evaluating the use of different positional strategies for sentence selection in biomedical literature summarization," *BMC Bioinformatics*, vol. 14, no. 1, p. 71, Feb. 2013.

[204] L. Reeve, H. Han, and A. D. Brooks, "BioChain: lexical chaining methods for biomedical text summarization," in *Proceedings of the 2006 ACM symposium on Applied computing*, 2006, pp. 180–184.

[205] B. Humphrey, D. A. B. Lindberg, H. M. Schoolman, and G. O. Barnett, "The Unified Medical Language System: An Informatics Research Collaboration.," *Journal of the American Medical Association*, vol. 5, pp. 1–11, 1998.

[206] M. Fiszman, T. C. Rindflesch, and H. Kilicoglu, "Abstraction summarization for managing the biomedical research literature," in *Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics*, Stroudsburg, PA, USA, 2004, pp. 76–83.

[207] T. E. Workman, M. Fiszman, J. F. Hurdle, and T. C. Rindflesch, "Biomedical text summarization to support genetic database curation: using Semantic MEDLINE to create a secondary database of genetic information," *J Med Libr Assoc*, vol. 98, no. 4, pp. 273–281, Oct. 2010.

[208] T. E. Workman and J. F. Hurdle, "Dynamic summarization of bibliographic-based data," *BMC Med Inform Decis Mak*, vol. 11, p. 6, 2011.

[209] Y. Shang, Y. Li, H. Lin, and Z. Yang, "Enhancing Biomedical Text Summarization Using Semantic Relation Extraction," *PLoS ONE*, vol. 6, no. 8, p. e23862, Aug. 2011.

[210] L. Plaza, A. Díaz, and P. Gervás, "A semantic graph-based approach to biomedical summarisation," *Artif Intell Med*, vol. 53, no. 1, pp. 1–14, Sep. 2011.

[211] I. Yoo, X. Hu, and I.-Y. Song, "A coherent graph-based semantic clustering and summarization approach for biomedical literature and a new summarization evaluation method," *BMC Bioinformatics*, vol. 8 Suppl 9, p. S4, 2007.

[212] R. P. Futrelle, "Summarization of diagrams in documents," *Advances in Automated Text Summarization*, pp. 403–421, 1999.

[213] S. Bhatia and P. Mitra, "Summarizing figures, tables, and algorithms in scientific publications to augment search results," *ACM Transactions on Information Systems (TOIS)*, vol. 30, no. 1, p. 3, 2012.

[214] P. Wu and S. Carberry, "Toward extractive summarization of multimodal documents," in *Proceedings of the Workshop on Text Summarization at the Canadian Conference on Artificial Intelligence*, 2011, pp. 53–61.

[215] S. Agarwal and H. Yu, "Automatically classifying sentences in full-text biomedical articles into Introduction, Methods, Results and Discussion," *Bioinformatics*, vol. 25, no. 23, p. 3174, 2009.

[216] H. Yu, F. Liu, and B. P. Ramesh, "Automatic Figure Ranking and User Interfacing for Intelligent Figure Search," *PLoS ONE*, vol. 5, no. 10, p. e12983, Oct. 2010.

[217] I Mani, "Summarization Evaluation: An Overview," in *Proceedings of the NTCIR Workshop*, 2001, vol. 2.

[218] R. L. Donaway, K. W. Drummey, and L. A. Mather, "A comparison of rankings produced by summarization evaluation measures," in *Proceedings of the 2000 NAACL-ANLPWorkshop on Automatic summarization-Volume 4*, 2000, pp. 69–78.

[219] H. Saggion, S. Teufel, D. Radev, and W. Lam, "Meta-evaluation of summaries in a cross-lingual environment using content-based metrics," in *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, 2002, pp. 1–7.

[220] D. R. Radev, S. Teufel, H. Saggion, W. Lam, J. Blitzer, H. Qi, A. \cCelebi, D. Liu, and E. Drabek, "Evaluation challenges in large-scale document summarization," in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, Stroudsburg, PA, USA, 2003, pp. 375–382.

[221] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics*, 2002, pp. 311–318.

[222] K. Pastra and H. Saggion, "Colouring summaries BLEU," in *Proceedings of the EACL 2003 Workshop on Evaluation Initiatives in Natural Language Processing: are evaluation methods, metrics and resources reusable?*, 2003, pp. 35–42.

[223] C. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proceedings of the ACL Workshop: Text Summarization Braches Out 2004*, 2004, pp. 74–81.

[224] A. Nenkova, R. Passonneau, and K. McKeown, "The pyramid method: Incorporating human content selection variation in summarization evaluation," *ACM Transactions on Speech and Language Processing (TSLP)*, vol. 4, no. 2, p. 4, 2007.

[225] D. R. Radev and D. Tam, "Summarization evaluation using relative utility," in *Proceedings of the twelfth international conference on Information and knowledge management*, 2003, pp. 508–511.

[226] B. J. Dorr, "Text summarization evaluation: correlating human performance on an extrinsic task with automatic intrinsic metrics," DTIC Document, 2006.

[227] J. Steinberger and K. Je\vzek, "Evaluation measures for text summarization," *Computing and Informatics*, vol. 28, no. 2, pp. 251–275, 2012.

[228] A. H. Morris, G. M. Kasper, and D. A. Adams, "The effects and limitations of automated text condensing on reading comprehension performance," *Information Systems Research*, vol. 3, no. 1, pp. 17–35, 1992.

[229] S. Teufel, "Task-based evaluation of summary quality: Describing relationships between scientific papers," in *In Workshop Automatic Summarization, NAACL*, 2001.

[230] R. W. White, J. M. Jose, and I. Ruthven, "A task-oriented study on the influencing effects of query-biased summarisation in web searching," *Information Processing & Management*, vol. 39, no. 5, pp. 707–733, 2003.

[231] A. Tombros and M. Sanderson, "Advantages of query biased summaries in information retrieval," in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, 1998, pp. 2–10.

[232] I. Mani and E. Bloedorn, "Multi-document Summarization by Graph Search and Matching," *arXiv:cmp-lg/9712004*, Dec. 1997.

[233] T. F. Hand, *A Proposal for Task-based Evaluation of Text Summarization System*. 1997.

[234] I. Mani, D. House, G. Klein, L. Hirschman, T. Firmin, and B. Sundheim, "The TIPSTER SUMMAC text summarization evaluation," in *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, 1999, pp. 77–85.

[235] N. Elhadad, K. McKeown, D. Kaufman, and D. Jordan, "Facilitating Physicians' Access to Information via Tailored Text Summarization," *AMIA Annu Symp Proc*, vol. 2005, pp. 226–230, 2005.

[236] K. Mckeown, R. J. Passonneau, D. K. Elson, A. Nenkova, and J. Hirschberg, *Do Summaries Help? A Task-Based Evaluation of Multi-Document Summarization*. 2005.

[237] G. Murray, T. Kleinbauer, P. Poller, T. Becker, S. Renals, and J. Kilgour, "Extrinsic summarization evaluation: A decision audit task," *ACM Transactions on Speech and Language Processing (TSLP)*, vol. 6, no. 2, p. 2, 2009.

[238] M. Fiszman, D. Demner-Fushman, H. Kilicoglu, and T. C. Rindflesch, "Automatic summarization of MEDLINE citations for evidence-based medical treatment: a topic-oriented evaluation," *J Biomed Inform*, vol. 42, no. 5, pp. 801–813, Oct. 2009.

[239] J. Yang, A. Cohen, and W. Hersh, "Evaluation of a gene information summarization system by users during the analysis process of microarray datasets," *BMC Bioinformatics*, vol. 10, no. Suppl 2, p. S5, Feb. 2009.

[240] "DUC 2007 Task, Documents for Summarization, and Measures." [Online]. Available: http://duc.nist.gov/duc2007/tasks.html. [Accessed: 19-Apr-2013].

[241] D. R. Kaufman, V. L. Patel, C. Hilliman, P. C. Morin, J. Pevzner, R. S. Weinstock, R. Goland, S. Shea, and J. Starren, "Usability in the real world: assessing medical information technologies in patients' homes," *J Biomed Inform*, vol. 36, no. 1–2, pp. 45–60, Apr. 2003.

[242] A. W. Kushniruk and V. L. Patel, "Cognitive and usability engineering methods for the evaluation of clinical information systems," *J Biomed Inform*, vol. 37, no. 1, pp. 56–76, Feb. 2004.

[243] R. B. West, D. S. A. Nuyten, S. Subramanian, T. O. Nielsen, C. L. Corless, B. P. Rubin, K. Montgomery, S. Zhu, R. Patel, T. Hernandez-Boussard, J. R. Goldblum, P. O. Brown, M. van de Vijver, and M. van de Rijn, "Determination of Stromal Signatures in Breast Carcinoma," *PLoS Biol*, vol. 3, no. 6, Jun. 2005.

[244] I. H. Shrivastava, J. Jiang, S. G. Amara, and I. Bahar, "Time-resolved Mechanism of Extracellular Gate Opening and Substrate Binding in a Glutamate Transporter," *J Biol Chem*, vol. 283, no. 42, pp. 28680–28690, Oct. 2008.

[245] G. Salton, A. Singhal, M. Mitra, and C. Buckley, "Automatic text structuring and summarization," *Information Processing & Management*, vol. 33, no. 2, pp. 193–207, Mar. 1997.

[246] T. Nomoto and Y. Matsumoto, "Data Reliability and Its Effects on Automatic Abstracting," *In Proceedings of the Fifth Workshop on Very Large Corpora, Beijing/Hong Kong*, 1997.

[247] J. J. Prochaska and J. F. Hilton, "Risk of cardiovascular serious adverse events associated with varenicline use for tobacco cessation: systematic review and meta-analysis," *BMJ*, vol. 344, p. e2856, 2012.

[248] M. Sczaniecka, A. Feoktistova, K. M. May, J.-S. Chen, J. Blyth, K. L. Gould, and K. G. Hardwick, "The Spindle Checkpoint Functions of Mad3 and Mad2 Depend on a Mad3 KEN Box-mediated Interaction with Cdc20-Anaphase-promoting Complex (APC/C)," *J Biol Chem*, vol. 283, no. 34, pp. 23039–23047, Aug. 2008.

[249] R. Khouri, F. Novais, G. Santana, C. I. de Oliveira, M. A. Vannier dos Santos, A. Barral, M. Barral-Netto, and J. Van Weyenbergh, "DETC Induces Leishmania Parasite Killing in Human In Vitro and Murine In Vivo Models: A Promising Therapeutic Alternative in Leishmaniasis," *PLoS One*, vol. 5, no. 12, Dec. 2010.

[250] M. Baldry, C. Cheal, B. Fisher, M. Gillett, and V. Huet, "Giving patients their own records in general practice: experience of patients and staff.," *Br Med J (Clin Res Ed)*, vol. 292, no. 6520, pp. 596–598, Mar. 1986.

[251] W. J. Winkelman, K. J. Leonard, and P. G. Rossos, "Patient-Perceived Usefulness of Online Electronic Medical Records: Employing Grounded Theory in the Development of Information and Communication Technologies for Use by Patients Living with Chronic Illness," *J Am Med Inform Assoc*, vol. 12, no. 3, pp. 306–314, May 2005.

[252] J. F. Seitz, A. Ward, and W. H. Dobbs, "Granting patients access to records: the impact of the Privacy Act at a federal hospital," *Hosp Community Psychiatry*, vol. 29, no. 5, pp. 288–289, May 1978.

[253] D. A. DeWalt, R. M. Malone, M. E. Bryant, M. C. Kosnar, K. E. Corr, R. L. Rothman, C. A. Sueta, and M. P. Pignone, "A heart failure self-management program for patients of all literacy levels: a randomized, controlled trial [ISRCTN11535170]," *BMC Health Serv Res*, vol. 6, p. 30, 2006.

[254] D. Schillinger, M. Handley, F. Wang, and H. Hammer, "Effects of self-management support on structure, process, and outcomes among vulnerable patients with diabetes: a three-arm practical clinical trial," *Diabetes Care*, vol. 32, no. 4, pp. 559–566, Apr. 2009.

[255] M. Meltsner, "A Patient's View of OpenNotes," *Ann Intern Med*, vol. 157, no. 7, pp. 523–524, Oct. 2012.

[256] J. J. Cimino, V. L. Patel, and A. W. Kushniruk, "The patient clinical information system (PatCIS): technical solutions for and experience with giving patients access to their electronic medical records," *Int J Med Inform*, vol. 68, no. 1–3, pp. 113–127, Dec. 2002.

[257] L. M. Stossel, N. Segar, P. Gliatto, R. Fallar, and R. Karani, "Readability of patient education materials available at the point of care," *J Gen Intern Med*, vol. 27, no. 9, pp. 1165–1170, Sep. 2012.

[258] "The Health Literacy of America's Adults: Results from the 2003 National Assessment of Adult Literacy," 06-Sep-2006. [Online]. Available: http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2006483. [Accessed: 15-Oct-2012].

[259] M. V. Williams, D. W. Baker, R. M. Parker, and J. R. Nurss, "Relationship of functional health literacy to patients' knowledge of their chronic disease. A study of patients with hypertension and diabetes," *Arch. Intern. Med.*, vol. 158, no. 2, pp. 166–172, Jan. 1998.

[260] J. A. Gazmararian, M. V. Williams, J. Peel, and D. W. Baker, "Health literacy and knowledge of chronic disease," *Patient Educ Couns*, vol. 51, no. 3, pp. 267–275, Nov. 2003.

[261] S. White, J. Chen, and R. Atchison, "Relationship of preventive health practices and health literacy: a national study," *Am J Health Behav*, vol. 32, no. 3, pp. 227–242, Jun. 2008.

[262] T. L. Scott, J. A. Gazmararian, M. V. Williams, and D. W. Baker, "Health literacy and preventive health care use among Medicare enrollees in a managed care organization," *Med Care*, vol. 40, no. 5, pp. 395–404, May 2002.

[263] D. W. Baker, J. A. Gazmararian, M. V. Williams, T. Scott, R. M. Parker, D. Green, J. Ren, and J. Peel, "Functional health literacy and the risk of hospital admission among Medicare managed care enrollees," *Am J Public Health*, vol. 92, no. 8, pp. 1278–1283, Aug. 2002.

[264] D. W. Baker, R. M. Parker, M. V. Williams, and W. S. Clark, "Health literacy and the risk of hospital admission," *J Gen Intern Med*, vol. 13, no. 12, pp. 791–798, Dec. 1998.

[265] S. C. Kalichman, H. Pope, D. White, C. Cherry, C. M. Amaral, C. Swetzes, J. Flanagan, and M. O. Kalichman, "The Association between Health Literacy and HIV Treatment Adherence: Further Evidence from Objectively Measured Medication Adherence," *J Int Assoc Physicians AIDS Care (Chic)*, vol. 7, no. 6, pp. 317–323, 2008.

[266] S. Kripalani, M. E. Gatti, and T. A. Jacobson, "Association of age, health literacy, and medication management strategies with cardiovascular medication adherence," *Patient Educ Couns*, vol. 81, no. 2, pp. 177–181, Nov. 2010.

[267] A. Lincoln, M. K. Paasche-Orlow, D. M. Cheng, C. Lloyd-Travaglini, C. Caruso, R. Saitz, and J. H. Samet, "Impact of Health Literacy on Depressive Symptoms and Mental Health-related: Quality of Life Among Adults with Addiction," *J Gen Intern Med*, vol. 21, no. 8, pp. 818–822, Aug. 2006.

[268] I. M. Bennett, J. F. Culhane, K. F. McCollum, L. Mathew, and I. T. Elo, "Literacy and depressive symptomatology among pregnant Latinas with limited English proficiency," *Am J Orthopsychiatry*, vol. 77, no. 2, pp. 243–248, Apr. 2007.

[269] M. S. Wolf, J. A. Gazmararian, and D. W. Baker, "Health literacy and functional health status among older adults," *Arch. Intern. Med.*, vol. 165, no. 17, pp. 1946–1952, Sep. 2005.

[270] D. W. Baker, R. M. Parker, M. V. Williams, W. S. Clark, and J. Nurss, "The relationship of patient reading ability to self-reported health and use of health services.," *Am J Public Health*, vol. 87, no. 6, pp. 1027–1030, Jun. 1997.

[271] D. H. Howard, J. Gazmararian, and R. M. Parker, "The impact of low health literacy on the medical costs of Medicare managed care enrollees," *Am. J. Med.*, vol. 118, no. 4, pp. 371–377, Apr. 2005.

[272] B. D. Weiss and R. Palmer, "Relationship between health care costs and very low literacy skills in a medically needy and indigent Medicaid population," *J Am Board Fam Pract*, vol. 17, no. 1, pp. 44–47, Feb. 2004.

[273] M. C. Carolan, G. Gill, and C. Steele, "Women's experiences of factors that facilitate or inhibit gestational diabetes self-management," *BMC Pregnancy Childbirth*, vol. 12, no. 1, p. 99, Sep. 2012.

[274] M. M. Schapira, K. E. Fletcher, A. Hayes, D. Eastwood, L. Patterson, K. Ertl, and J. Whittle, "The development and validation of the hypertension evaluation of lifestyle and management knowledge scale," *J Clin Hypertens (Greenwich)*, vol. 14, no. 7, pp. 461–466, Jul. 2012.

[275] D. W. Baker, M. S. Wolf, J. Feinglass, J. A. Thompson, J. A. Gazmararian, and J. Huang, "Health literacy and mortality among elderly persons," *Arch. Intern. Med.*, vol. 167, no. 14, pp. 1503–1509, Jul. 2007.

[276] R. L. Sudore, K. Yaffe, S. Satterfield, T. B. Harris, K. M. Mehta, E. M. Simonsick, A. B. Newman, C. Rosano, R. Rooks, S. M. Rubin, H. N. Ayonayon, and D. Schillinger, "Limited literacy and mortality in the elderly: the health, aging, and body composition study," *J Gen Intern Med*, vol. 21, no. 8, pp. 806–812, Aug. 2006.

[277] J. E. Fincham, "The Public Health Importance of Improving Health Literacy," *Am J Pharm Educ*, vol. 77, no. 3, Apr. 2013.

[278] W. Lober, B. Zierler, A. Herbaugh, S. Shinstrom, A. Stolyar, E. Kim, and Y. Kim, "Barriers to the use of a Personal Health Record by an Elderly Population," *AMIA Annu Symp Proc*, vol. 2006, pp. 514–518, 2006.

[279] C. von Wagner, C. Semmler, A. Good, and J. Wardle, "Health literacy and self-efficacy for participating in colorectal cancer screening: The role of information processing," *Patient Educ Couns*, vol. 75, no. 3, pp. 352–357, Jun. 2009.

[280] R. L. Rothman, D. A. DeWalt, R. Malone, B. Bryant, A. Shintani, B. Crigler, M. Weinberger, and M. Pignone, "Influence of patient literacy on the effectiveness of a primary care-based diabetes disease management program," *JAMA*, vol. 292, no. 14, pp. 1711–1716, Oct. 2004.

[281] E. J. Stein, R. L. Furedy, M. J. Simonton, and C. H. Neuffer, "Patient access to medical records on a psychiatric inpatient unit," *Am J Psychiatry*, vol. 136, no. 3, pp. 327–329, Mar. 1979.

[282] S. E. Ross and C.-T. Lin, "The Effects of Promoting Patient Access to Medical Records: A Review," *J Am Med Inform Assoc*, vol. 10, no. 2, pp. 129–138, 2003.

[283] K. Chapman, C. Abraham, V. Jenkins, and L. Fallowfield, "Lay understanding of terms used in cancer consultations," *Psychooncology*, vol. 12, no. 6, pp. 557–566, Sep. 2003.

[284] E. B. Lerner, D. V. Jehle, D. M. Janicke, and R. M. Moscati, "Medical communication: do our patients understand?," *Am J Emerg Med*, vol. 18, no. 7, pp. 764–766, Nov. 2000.

[285] A. Keselman and C. A. Smith, "A classification of errors in lay comprehension of medical documents," *J Biomed Inform*, vol. 45, no. 6, pp. 1151–1163, Dec. 2012.

[286] Q. Zeng-Treitler, S. Goryachev, H. Kim, A. Keselman, and D. Rosendale, "Making texts in electronic health records comprehensible to consumers: a prototype translator," *AMIA Annu Symp Proc*, pp. 846–850, 2007.

[287] N. R. Kandula, P. A. Nsiah-Kumi, G. Makoul, J. Sager, C. P. Zei, S. Glass, Q. Stephens, and D. W. Baker, "The relationship between health literacy and knowledge improvement after a multimedia type 2 diabetes education program," *Patient Educ Couns*, vol. 75, no. 3, pp. 321–327, Jun. 2009.

[288] Y. Hong, K. Ehlers, R. Gillis, T. Patrick, and J. Zhang, "A usability study of patient-friendly terminology in an EMR system," *Stud Health Technol Inform*, vol. 160, no. Pt 1, pp. 136–140, 2010.

[289] Q. Zeng-Treitler, H. Kim, G. Rosemblat, and A. Keselman, "Can multilingual machine translation help make medical record content more comprehensible to patients?," *Stud Health Technol Inform*, vol. 160, no. Pt 1, pp. 73–77, 2010.

[290] C. A. Smith, S. Hetzel, P. Dalrymple, and A. Keselman, "Beyond Readability: Investigating Coherence of Clinical Text for Consumers," *J Med Internet Res*, vol. 13, no. 4, Dec. 2011.

[291] B. Smith and C. Fellbaum, "Medical WordNet: a new methodology for the construction and validation of information resources for consumer health," in *Proceedings of the 20th international conference on Computational Linguistics*, 2004, p. 371.

[292] *Consumer Health Vocabulary http://www.consumerhealthvocab.org/.* .

[293] A. T. McCray, R. F. Loane, A. C. Browne, and A. K. Bangalore, "Terminology issues in user access to Web-based medical information.," *Proc AMIA Symp*, pp. 107–111, 1999.

[294] Q. T. Zeng, T. Tse, J. Crowell, G. Divita, L. Roth, and A. C. Browne, "Identifying Consumer-Friendly Display (CFD) Names for Health Concepts," *AMIA Annu Symp Proc*, vol. 2005, pp. 859–863, 2005.

[295] T. B. Patrick, H. K. Monga, M. C. Sievert, J. H. Hall, and D. R. Longo, "Evaluation of Controlled Vocabulary Resources for Development of a Consumer Entry Vocabulary for Diabetes," *Journal of Medical Internet Research*, vol. 3, no. 3, p. e24, Aug. 2001.

[296] L. A. Slaughter, D. Soergel, and T. C. Rindflesch, "Semantic representation of consumer questions and physician answers," *Int J Med Inform*, vol. 75, no. 7, pp. 513–529, Jul. 2006.

[297] C. A. Smith, P. Z. Stavri, and W. W. Chapman, "In their own words? A terminological analysis of e-mail to a cancer information service.," *Proc AMIA Symp*, pp. 697–701, 2002.

[298] C. A. Smith and P. J. Wicks, "PatientsLikeMe: Consumer health vocabulary as a folksonomy," *AMIA Annu Symp Proc*, pp. 682–686, 2008.

[299] G. Rosemblat, R. Logan, T. Tse, and L. Graham, "Text Features and Readability: Expert Evaluation of Consumer Health Text," *MEDNET*.

[300] N. Elhadad, "Comprehending technical texts: predicting and defining unfamiliar terms," *AMIA Annu Symp Proc*, pp. 239–243, 2006.

[301] Q. Zeng, E. Kim, J. Crowell, and T. Tse, "A text corpora-based estimation of the familiarity of health terminology," *Biological and Medical Data Analysis*, pp. 184–192, 2005.

[302] S. Kandula, D. Curtis, and Q. Zeng-Treitler, "A semantic and syntactic text simplification tool for health content," in *AMIA Annual Symposium Proceedings*, 2010, vol. 2010, p. 366.

[303] G. Leroy, J. Endicott, O. Mouradi, D. Kauchak, and M. Just, "Improving perceived and actual text difficulty for health information consumers using semi-automated methods," in *AMIA Annu Symp Proc*, 2012, pp. 522–31.

[304] J. J. Cimino, V. L. Patel, and A. W. Kushniruk, "What do patients do with access to their medical records?," *Stud Health Technol Inform*, vol. 84, no. Pt 2, pp. 1440–1444, 2001.

[305] B. Humphrey, D. A. B. Lindberg, H. M. Schoolman, and G. O. Barnett, "The Unified Medical Language System: An Informatics Research Collaboration.," *Journal of the American Medical Association*, vol. 5, pp. 1–11, 1998.

[306] N. L. of M. (U.S.), "Fact SheetMedlinePlus®." [Online]. Available: http://www.nlm.nih.gov/pubs/factsheets/medlineplus.html. [Accessed: 19-Nov-2012].

[307] J. M. Heilman, E. Kemmann, M. Bonert, A. Chatterjee, B. Ragar, G. M. Beards, D. J. Iberri, M. Harvey, B. Thomas, W. Stomp, M. F. Martone, D. J. Lodge, A. Vondracek, J. F. de Wolff, C. Liber, S. C. Grover, T. J. Vickers, B. Meskó, and M.

R. Laurent, "Wikipedia: A Key Tool for Global Public Health Promotion," *Journal of Medical Internet Research*, vol. 13, no. 1, p. e14, Jan. 2011.

[308] M. Laurent and T. J. Vickers, "Seeking Health Information Online: Does Wikipedia Matter?," *J Am Med Inform Assoc*, vol. 16, no. 4, pp. 471–479, Jul. 2009.

[309] A. R. Aronson, "Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program," *Proc AMIA Symp*, pp. 17–21, 2001.

[310] L. Si and J. Callan, "A statistical model for scientific readability," in *Proceedings of the tenth international conference on Information and knowledge management*, 2001, pp. 574–576.

[311] M. Laurent and T. J. Vickers, "Seeking Health Information Online: Does Wikipedia Matter?," *J Am Med Inform Assoc*, vol. 16, no. 4, pp. 471–479, Jul. 2009.

[312] J. Proulx, S. Kandula, B. Hill, and Q. Zeng-Treitler, "Creating Consumer Friendly Health Content: Implementing and Testing a Readability Diagnosis and Enhancement Tool," in *2014 47th Hawaii International Conference on System Sciences*, Los Alamitos, CA, USA, 2013, vol. 0, pp. 2445–2453.

[313] R. M. Parker, D. W. Baker, M. V. Williams, and J. R. Nurss, "The test of functional health literacy in adults: a new instrument for measuring patients' literacy skills," *J Gen Intern Med*, vol. 10, no. 10, pp. 537–541, Oct. 1995.

Curriculum Vitae

BALAJI POLEPALLI RAMESH

brpjr@uwm.edu
http://www.linkedin.com/in/balajipolepalliramesh/

EDUCATION

Doctor of Philosophy, Biomedical and Health Informatics　　　**Sep 2009 to May 2014**
UNIVERSITY OF WISCONSIN, Milwaukee, WI
Dissertation Title: Adverse Drug Event Detection, Causality Inference and Summarization

Master of Science, Computer Science　　　**Sep 2007 to Aug 2009**
UNIVERSITY OF WISCONSIN, Milwaukee, WI
Thesis Title: ANTSENS: an Ant based routing protocol for large-scale sensor networks

Bachelor of Engineering, Computer Science　　　**Sep 2003 to June 2007**
VISVESVARAYA TECHNOLOGICAL UNIVERSITY, Belgaum, India

INTERNSHIPS & EXPERIENCE

Research Assistant　　　**June 2013 to Present**
UNIVERSITY OF MASSACHUSETTS MEDICAL SCHOOL, Worcester, MA

Research Assistant　　　**Aug 2009 to May 2013**
UNIVERSITY OF WISCONSIN, Milwaukee, WI

Software Intern　　　**Jan 2012 to Jun 2012**
NEXTBIO INC., Santa Clara, CA

Software Intern　　　**May 2008 to Aug 2008**
 METAVANTE CORP., Milwaukee, WI

Application Developer　　　**Sep 2008 to Dec 2008**
UNIVERSITY OF WISCONSIN, R2D2 Center, Milwaukee, WI

Teaching Assistant　　　**Sep 2007 to May 2009**
UNIVERSITY OF WISCONSIN, Milwaukee, WI

PUBLICATIONS

- **Polepalli B**, Xie W, Thangaraja D, Goyal M, Hosseini H, Bashir Y,*"Impact of IEEE 802.11n Operation on IEEE 802.15.4 Operation"*, In Proceedings of 5th IEEE International Workshop on Heterogeneous Wireless Networks, part of Advanced Information Networking and Applications (AINA), May 2009
- Xie W, **Polepalli B**, Goyal M, Hosseini H,*"Enhancing Simulation Mode Operation of ospfd"*, In Proceedings of Network-Based Information System (NBiS), Aug 2009
- Rohm D, Goyal M, Xie W, **Polepalli B**, Hosseini H, Divjak A, Bashir Y,*"Dynamic Reconfiguration in Beaconless IEEE 802.15.4 Networks Under Varying Traffic Loads"*, In Proceedings of IEEE Globecom 2009 and IEEE eXplore, Nov 2009
- Yu H, Liu F, **Polepalli Ramesh B**, "*Automatic Figure Ranking and User Interfacing for Intelligent Figure Search*". PLoS ONE 5(10): e12983, Oct 2010

- **Polepalli Ramesh B**, Yu H. "*Identifying Discourse Connectives in Biomedical Text*". In Proceeding of American Medical Informatics Symposium, Nov 2010
- Kim D, **Polepalli Ramesh B**, Yu H. "*Automatic Figure Classification in Bioscience Literature*", Journal of Biomedical Informatics;44(5):848-58, Oct 2011
- **Polepalli Ramesh B,** Prasad R, Miller T, Yu H. "*Automatic Discourse Connective Detection in Biomedical Text*", JAMIA; 19(5), Sept 2012
- **Polepalli Ramesh B**, Houston T, Brandt C, Fang J, Yu H. "*Improving Pateints' Electronic health Record Comprehension with NoteAid*", MedInfo 2013, Copenhagen, Denmark
- **Polepalli Ramesh B**, Yu H, "*Systems for Improving Electronic Health Record Note Comprehension*", HSD 2013 in ACM SIGIR 2013
- **Polepalli Ramesh B**, Belknap S, Li Z, Frid N, West D, Yu H. "*Automatically recognizing medication and adverse event information from the FDA AERS narratives*", JMIR Medical Informatics. (In Press)
- **Polepalli Ramesh B**, Yu H. "*Figure Summarization and Evaluation*" (Conditionally accepted in PLoSONE)
- **Polepalli Ramesh B,** Cai S, Chiriboga G, Knight K, Yu H. "*Translating HER Notes from English to Spanish with NoteAid $_{Spanish}$*" , Submitted to AMIA 2014
- Zheng J, Yarzebski J, **Polepalli Ramesh B,** Goldberg R, Yu H. "*Automatically Detecting Acute Myocardial Infarction Events from EHR Text*" , Submitted to AMIA 2014
- **Polepalli Ramesh B**, Yu H. "*Extrinsic Task Driven Evaluation of Figure Summarization and Article Comprehension*" (In preparation)

HONORS, ACCOMPLISHMENTS AND LEADERSHIP EXPERIENCE

Distinguished Dissertation Fellowship · University of Wisconsin, Milwaukee, WI

Graduate Research Student Award · Biomedical and Health Informatics Research Institute

Best Student Paper Award · MedInfo 2013, 14th World Congress in Medical Informatics

Featured UWM Graduate Student · University of Wisconsin, Milwaukee, WI

Graduate School Travel Award · University of Wisconsin, Milwaukee, WI

Asian Faculty and Staff Scholarship Award · AFSA University of Wisconsin, Milwaukee, WI

Chancellors Award · University of Wisconsin, Milwaukee, WI

Best Academic Performer Award · College of Engineering, Visvesvaraya Technological University. Belguam, India

REVIEWER SERVICES

Reviewed paper and posters for the MedInfo, AMIA, Journal of Biomedical informatics and Applied Clinical Informatics.

SYSTEMS DEVELOPED

ADEtector (http://autumn.ims.uwm.edu:8080/adetector)
- A system that detects the presence of adverse drug events in electronic health record notes by combining various NLP and knowledge based techniques

NoteAid (clinicalnotesaid.org)
- A system that helps patients comprehend their electronic health record notes

ADE Repository (aderepository.org)
- An online repository to of adverse drug events from various resources

AERS Tagger (http://23.23.239.90/aderepository/aerstagger.uwm)
- A named entity tagger to identify adverse events and medication related information in FAERS reports

EMR Tagger (http://23.23.239.90/aderepository/emrtagger.uwm)
- A named entity tagger to identify adverse events and medication related information in EMR reports

BioConn (http://www.askhermes.org/BioConn/)
- A system that automatically identifies discourse connectives in biomedical domain

Biomedical Figure Classifier (http://autumn.ims.uwm.edu:8080/FigureClassifier/)
- A system that automatically classifies biomedical figure on its modality

TECHNICAL SKILLS

NLP/Machine Learning:	Named Entity Recognition, Discourse Parsing, Syntactic Parsing, Biomedical Ontologies, Semantic Analysis, Co-reference Resolution, Relation Extraction, Ranking, Indexing, Topic Modeling, Semi-Supervised Learning
Languages:	Java, C, C++, Python, Perl, JSP, JavaScript, Unix Scripting
Software/Applications:	Weka, MATLAB, SQL, SAS, STATA, R, SPSS, Drupal, Joomla, Hadoop, Photoshop
Servers:	WebSphere, Tomcat, JBoss, Apache Solr

Major Professor   Date