7-2-2020

Automatic Learning of Document Section Structure for Ontology-based Semantic Search

Deya Banisakher
Florida International University, dbani001@fiu.edu

FLORIDA INTERNATIONAL UNIVERSITY

Miami, Florida

AUTOMATIC LEARNING OF DOCUMENT SECTION STRUCTURE FOR ONTOLOGY-BASED SEMANTIC SEARCH

A dissertation submitted in partial fulfillment of the

requirements for the degree of

DOCTOR OF PHILOSOPHY

in

COMPUTER SCIENCE

by

Deya Banisakher

2020

To: Dean John L. Volakis
College of Engineering and Computing

This dissertation, written by Deya Banisakher, and entitled Automatic Learning of Document Section Structure for Ontology-Based Semantic Search, having been approved in respect to style and intellectual content, is referred to you for judgment.

We have read this dissertation and recommend that it be approved.

Armando Barreto

Ning Xie

Shu-Ching Chen

Naphtali Rishe, Co-Major Professor

Mark Finlayson, Major Professor

Date of Defense: July 2, 2020

The dissertation of Deya Banisakher is approved.

Dean John L. Volakis
College of Engineering and Computing

Dean Andrés G. Gil
Vice President for Research and Economic Development
Dean of the University Graduate School

Florida International University, 2020

DEDICATION

To my mom, Fatima Attiyat, and Mubarak Banisakher. To my brother, David Banisakher, and my stepmom Bella Khamitsaeva. I would not be half the person I am (and I'm not just speaking genetically) if it wasn't for the love, support, and guidance of my parents.

ACKNOWLEDGMENTS

Thanks, everyone!

ABSTRACT OF THE DISSERTATION

AUTOMATIC LEARNING OF DOCUMENT SECTION STRUCTURE FOR ONTOLOGY-BASED SEMANTIC SEARCH

by

Deya Banisakher

Florida International University, 2020

Miami, Florida

Professor Mark Finlayson, Major Professor

Modeling natural human behaviour in understanding written language is crucial for developing true artificial intelligence. For people, words convey certain semantic concepts. While documents represent an abstract concept—they are collections of text organized in some logical structure, that is, sentences, paragraphs, sections, and so on. Similar to words, these document structures, are used to convey a logical flow of semantic concepts. Machines however, only view words as spans of characters and documents as mere collections of free-text, missing any underlying meanings behind words and the logical structure of those documents.

Automatic semantic concept detection is the process by which the underlying meanings of words are identified and retrieved. My thesis aims at bridging the semantic gap between automatic concept detection and logical document structure understanding. In my dissertation, I demonstrate an analysis and development of a framework for using logical document structure knowledge (that is, *section structure*) in detecting semantic concepts within documents in various domains. In that, I developed my research around six different document classes from four domains: medical, legal,scientific, and news reporting.The document classes are as follows: psychiatric report evaluations, hospital discharge summaries, and radiology reports in the medical domain; Patent documents in

the legal domain; environmental journal articles in the scientific domain; Finally, business and politics news articles.

I demonstrate section structure identification and discovery models over five different document classes from three domains: psychiatric evaluations, radiology reports, and discharge summaries in the clinical domain; patent documents in the intellectual property (IP) domain, and environmental scientific articles from the scientific domain. I also demonstrate various approaches for supervised ontology-based semantic concept detection. Finally, I discuss the use of section structure in scientific articles for semantic concept detection and demonstrate its efficacy and efficiency in achieving a significant performance.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Motivation

Modeling natural human behaviour in understanding language in all its forms is critical for the development of true artificial intelligence. Written text is the most effective communication medium humans use daily [Kellogg, 1999]. With the explosion of digital communication, worldwide inter-connectivity, and the abundance of storage capacity, written documents are ever-growing in every domain. Free-text documents in the medical and the scientific domains are examples of this exponential growth of data [Fang et al., 2016]. Psychiatric reports and scientific research articles, for example, are being produced and digitized at record rates as more people are treated for mental illness by medical professionals and more problems are being solved by scientists. As disjoint as they may seem, these examples of documents share common characteristics as they are highly structured and contain free-form narrative that typically communicates a single idea presented through various sub-ideas and substructures (e.g. sections of a scientific article).

My work is motivated by the need to model natural human behaviour in understanding written documents. Specifically, my work aims to model two characteristics of this behaviour: first, understanding the logical structure of documents, and second, understanding the underlying meaning behind the written word, that is semantic understanding. Documents, regardless of the domain, are written in a structured format: words make up sentences, then paragraphs, sections and subsections, which in turn make up chapters and so on. These substructures are often headed with labels starting from a document title to section headings [Power et al., 2003].

In reading research, strategic reading has been studied as part of understanding human behaviour [Paris et al., 2016]. Readers are not passive sponges, soaking up information as it is fed to them line by line. In fact, the most effective readers are aware of their objectives, monitor the relevance of each part of a document to those objectives, and select the most relevant parts to attend to [Waller, 2011]. Additionally, cognitive psychologists have established that the human understanding of verbal information draws heavily on pre-existing knowledge or frameworks, sometimes referred to as schemata. These are conceptual structures, sets of expectations, or mental scripts that we can use to make inferences that may not be explicit in the text itself. Scripts are sets of knowledge about what we expect certain situations to be like and what might normally happen in them, based on experiences we have gathered over time. This falls under the concept of schemata [Kintsch, 1974, McNamara et al., 1991, McVee et al., 2005]. For documents, this can be thought of as the logical organization of sections and other substructures.

Document structure understanding as a subfield of natural language understanding (NLU) is interested in developing models that capture the substructures aforementioned within documents in various domains. Although there has been efforts toward developing general models that encompass multiple domains [Power et al., 2003, Nordström, 2008], these models struggle when faced with different and unseen domains and therefore this problem is far from being solved. Although desired, it is difficult to have have models that understand document structure across wildly different domains - this is a task that even humans have trouble with [Allahyari et al., 2017]. Take a legal document (e.g. an patent document), a medical document (e.g. a psychiatric report), a scientific article and a book. Sure enough the article and the book share similar characteristics, but they are in no way similar to the former examples except, again, in the fact that they are merely structured in some typical way.

Thus we must direct our attention to developing domain-specific document understanding models. However, even within a single domain, documents can be written differently (and sometimes wildly differently). For example, scientific articles can be headed in a generic way (i.e. abstract, introduction, methods, etc . . .), and psychiatric evaluation reports have no strict format as to how sections should be headed or ordered. In general, document structure understanding consists of two subproblems: first, a section label understanding or text-segmentation problem, and second, a section type understanding problem (i.e. what sections share a common type or goal given the domain in question). Solutions to these problems can be helped by semantic concept understanding and extraction from text and vice versa. That is, following intuition, to learn and understand the type of a block of text, one must first analyze and understand the meaning behind it.

Semantic concept detection is the process by which an implicit meaning of text is made explicit [Drumond and Girardi, 2008, Dou et al., 2015]. That is, understanding the underlying meaning of text. For example if a computer science scientific article is referring to a neural network, a model capable of semantic concept detection will infer that it is also referring to machine learning, artificial intelligence, and other related concepts. Semantics has always been at the heart of natural language processing (NLP). There have been many strides at developing such models using various methods. One of such methods is the use of ontologies–a formal description (typically, tree-shaped) of concepts and their interrelationships [Navigli and Velardi, 2004]. When dealing with domain-specific documents, domain-specific ontologies are especially useful when compared to their domain-independent counterparts as they are more compact and descriptive of their respective domains [Dou et al., 2015].

1.2 Problem Statement and Research Components

My research problem, in large, concerns the analysis and development of a framework for using logical document structure knowledge (that is, *section structure*) in detecting semantic concepts within documents in various domains. This entails four abstract conceptual components which have driven my research. Following, I list and discuss these components

Component 1. Corpora collection and annotation. The first step in any framework development or empirical design is data collection and processing. Additionally, a common task in natural language processing for preparing a given corpus for further processing is annotation. For this, I developed and evaluated the models I discuss in this dissertation using documents from six different datasets spanning four domains: medical, legal,scientific, and news reporting.The document classes are as follows: psychiatric report evaluations, hospital discharge summaries, and radiology reports in the medical domain; Patent documents in the legal domain; environmental journal articles in the scientific domain; Finally, business and politics news articles. Each of these corpora underwent an annotation study concerning the section structure of their respective documents. Additionally, the scientific articles corpus was also annotated with environmental semantic concepts using a domain-specific ontology.

Component 2. Modeling the logical document structure. In my research, the logical document structure refers to the sections in which a document is logically organized. As I have discussed earlier, documents are listed in sections that follow a logical order (e.g. this dissertation document) that express various conceptual sub-ideas which in turn make up the central idea of the document itself. Thus modeling the section structure of documents can give us insights into and lead to better understanding of the conceptual ideas communicated in those documents. I used the annotated data for each corpus from

the first component to develop automatic section type identification models that were successful in dissecting those documents into a standard set of sections.

The first task in section structure extraction is the identification of sections, their positions and boundaries in various documents. I performed three studies for section structure identification. The first (§3.1, [Banisakher et al., 2018a]), uses an Hierarchical Hidden Markov Model (HHMM) that was developed using the psychiatric evaluation reports (Corpus 2.1). The second (§3.2), extends the HHMM approach by using Conditional Random Fields (CRFs) which I developed using three corpora: psychiatric evaluation reports, radiology reports, and discharge summaries (Corpora 2.1-2.3). In the third (§3.3), I present an extended application of the CRF approach to improving the detection of paragraph functions in news article paragraphs (Corpus 2.6).

Additionally, I developed an approach to automatically discovering section type knowledge for a document class in a data-driven fashion using a modified Bayesian model merging algorithm. I tested my approach on five different document classes from three domains: psychiatric evaluations, radiology reports, and discharge summaries (Corpora 2.1-2.3) in the clinical domain; patent documents (Corpus 2.4) in the intellectual property (IP) domain, and environmental scientific articles (Corpus 2.5) from the scientific domain.

Component 3. Detecting semantic concepts through the use of domain-specific ontologies. Semantic concept detection refers to the identification of semantic entities (i.e. concepts) that represent an underlying meaning of lexical entities (i.e. words and sentences). Ontologies serve as thesauri for such tasks, as they provide a structured format that defines conceptual entities through hierarchical relationships. Thus, using these entity structures can be greatly beneficial—and in many cases necessary—to identify domain-specific concepts. For this, I developed automatic ontology-based supervised concept learning approaches for the biogeochemical scientific literature that use a domain-specific ontology and the annotated data from the first component.

Component 4. Incorporating section structure in the detection of semantic concepts. Following intuition, one can gauge the conceptual knowledge communicated through various sections within documents given: (1) knowledge of the domain concepts the document falls under, and (2) knowledge of the logical structure of the documents within that domain. For automated computational processes, the first can be achieved through the use of ontologies per domain, while the second can be solved through learning statistical features of the documents within a specific domain or corpus. Thus, here I combined components 2 and 3 and applied a set of experiments to test and analyze the models over the scientific domain.

1.3 Dissertation Contributions

There are three key contributions of the research I present here:

1. As outlined above, I developed models for (1) the identification of section structure given a standardized ontology of sections and (2) the discovery of section types which allows to automatically detect such standard ontologies for large collections of documents in a given document class. The second was the first attempt at unsupervised section type discovery of documents in various domains. I demonstrated the efficacy of these approaches using the six corpora outlined earlier through the extraction of unique sets of features that aid those models in distinguishing the various sections.
2. Additionally, I developed models for learning to identify domain-specific ontology concepts in the academic literature, specifically for the biogeochemical domain. More importantly, I demonstrated the efficacy and efficiency of incorporating the document structure of scientific articles in the detection of semantic concepts. In that vein, I automatically extracted the section structure of scientific articles and

encoded it as a feature in an ontology-based supervised model for learning semantic concepts. Additionally, I extracted a set of unique features that are able to capture the relationships between semantic concepts found a domain-specific ontology and the free-text language contained within scientific articles.

3. Finally, for the development of the models that I detail in this dissertation, I developed gold-standard corpora for section structure identification where documents from five different datasets (spanning four domains) were annotated using a unified section structure ontology. These corpora are as follows: psychiatric report evaluations, hospital discharge summaries, and radiology reports in the medical domain; Patent documents in the legal domain; and environmental journal articles in the scientific domain.

1.4 Outline

The remainder of this dissertation is organized as follows: First, in chapter 2, I discuss the six corpora I used to develop the automatic approaches and models for section identification and concept learning. In that, I introduce the corpora in detail, the relevant statistics and ontologies, as well as their respective annotation studies and annotation results. Second, in chapter 3 I present three studies for section structure identification. The first (§3.1), uses a Hierarchical Hidden Markov Model (HHMM) that was developed using the psychiatric evaluation reports (Corpus 2.1). The second (§3.2), extends the HHMM approach by using Conditional Random Fields (CRFs) which I developed using three corpora: psychiatric evaluation reports, radiology reports, and discharge summaries (Corpora 2.1-2.3). Finally, in the third (§3.3), I present an extended application of the CRF approach to improving the detection of paragraph functions in news article paragraphs (Corpus 2.6). Then, in chapter 4, I describe an approach to automatically discovering

section type knowledge for a document class in a data-driven fashion using a modified Bayesian model merging algorithm. I tested my approach on five different document classes from three domains: psychiatric evaluations, radiology reports, and discharge summaries (Corpora 2.1-2.3) in the clinical domain; patent documents (Corpus 2.4) in the intellectual property (IP) domain, and environmental scientific articles (Corpus 2.5) from the scientific domain. Next, in chapter 5, I first present a survey of academic literature search (§5.1) where I describe various approaches in semantic search in detail, I then present an ontology-based supervised concept learning approach for the biogeochemical scientific literature that uses random decision forrest as a supervised classifier learning scientific semantic concepts from a domain-specific ontology (§5.2), I then follow with an extended approach to ontology-based supervised concept learning for the biogeochemical scientific literature that uses the section structure of scientific articles (§5.3). I end with a conclusion that revisits the results and contributions of each research component (chapter 6), Finally, I include the discussion or related work relavent to each component in each respective section.

CHAPTER 2

CORPORA AND ANNOTATION

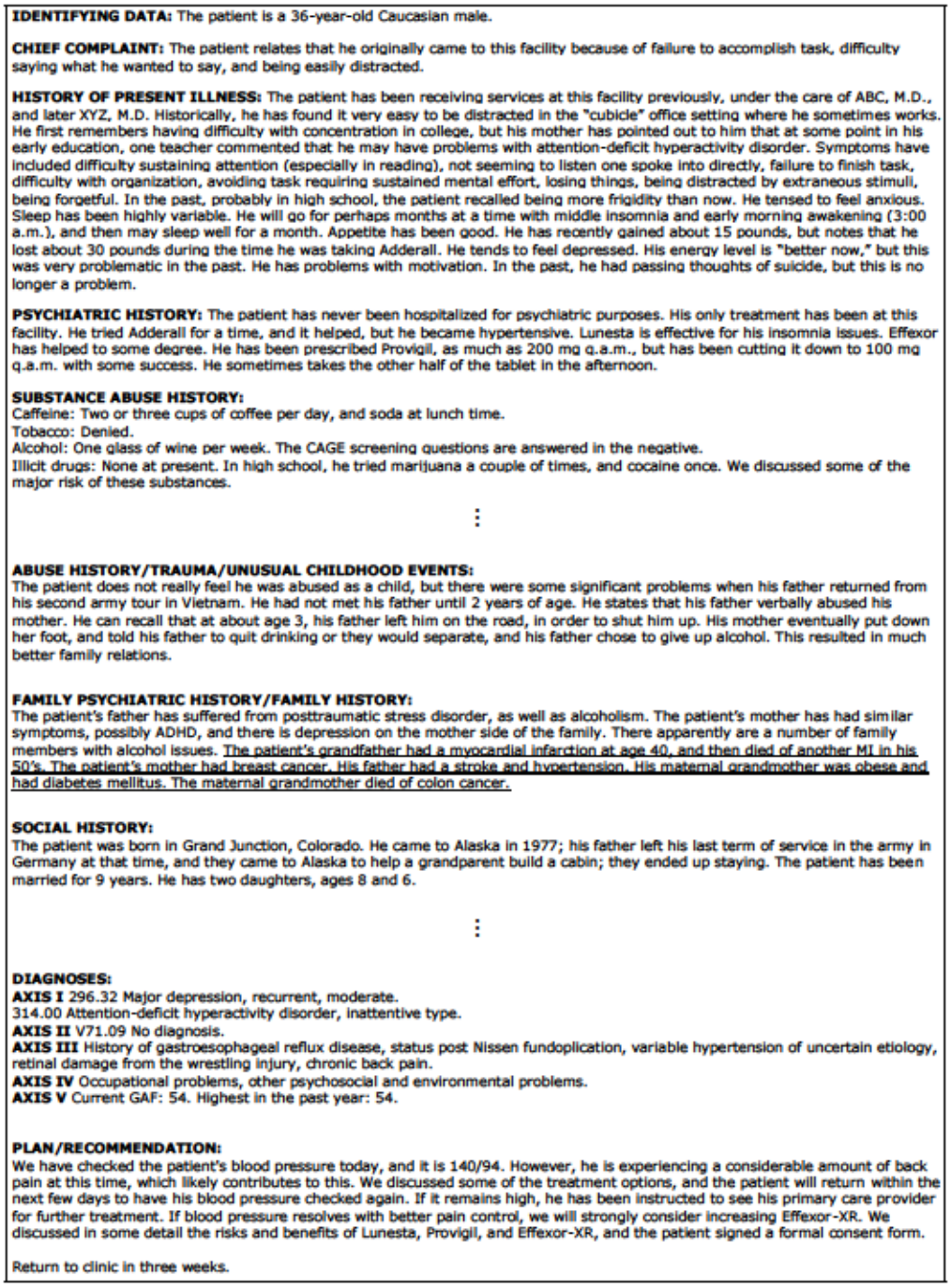
I developed and evaluated the models I discuss in this thesis using documents from six different datasets spanning four domains: medical, legal,scientific, and news reporting.The document classes are as follows: psychiatric report evaluations, hospital discharge summaries, and radiology reports in the medical domain; Patent documents in the legal domain; environmental journal articles in the scientific domain; Finally, business and politics news articles.

In this chapter, I discuss each of these document classes, and the specific corpora I used or collected. I also discuss the corpora ontologies and report detailed statistics both in this chapter and in the following relevant chapters for legibility and ease of reference. Finally, I discuss the annotation process, agreement metrics, and annotation results for each corpus.

2.1 Corpus 1: Psychiatric Evaluation Reports

A mental health assessment is the process through which a psychiatrist or a psychologist obtains and organizes necessary information about mental health patients. This process usually involves a series of psychological and medical tests (clinical and non-clinical), examinations, and interviews [Reeves and Rosner, 2016]. These procedures serve the purpose of making a diagnosis that then guides a treatment or a treatment plan [Association, 2018].

The output of a mental health assessment is a mental health report. Psychiatric reports are simpler subtype of this document type, and mainly consist of long-form unstructured text. They are the end product of psychiatric assessments in which psychiatrists summarize the information they gathered, as well as integrate the patient history, their eval-



IDENTIFYING DATA: The patient is a 36-year-old Caucasian male.

CHIEF COMPLAINT: The patient relates that he originally came to this facility because of failure to accomplish task, difficulty saying what he wanted to say, and being easily distracted.

HISTORY OF PRESENT ILLNESS: The patient has been receiving services at this facility previously, under the care of ABC, M.D., and later XYZ, M.D. Historically, he has found it very easy to be distracted in the "cubicle" office setting where he sometimes works. He first remembers having difficulty with concentration in college, but his mother has pointed out to him that at some point in his early education, one teacher commented that he may have problems with attention-deficit hyperactivity disorder. Symptoms have included difficulty sustaining attention (especially in reading), not seeming to listen one spoke into directly, failure to finish task, difficulty with organization, avoiding task requiring sustained mental effort, losing things, being distracted by extraneous stimuli, being forgetful. In the past, probably in high school, the patient recalled being more frigidity than now. He tensed to feel anxious. Sleep has been highly variable. He will go for perhaps months at a time with middle insomnia and early morning awakening (3:00 a.m.), and then may sleep well for a month. Appetite has been good. He has recently gained about 15 pounds, but notes that he lost about 30 pounds during the time he was taking Adderall. He tends to feel depressed. His energy level is "better now," but this was very problematic in the past. He has problems with motivation. In the past, he had passing thoughts of suicide, but this is no longer a problem.

PSYCHIATRIC HISTORY: The patient has never been hospitalized for psychiatric purposes. His only treatment has been at this facility. He tried Adderall for a time, and it helped, but he became hypertensive. Lunesta is effective for his insomnia issues. Effexor has helped to some degree. He has been prescribed Provigil, as much as 200 mg q.a.m., but has been cutting it down to 100 mg q.a.m. with some success. He sometimes takes the other half of the tablet in the afternoon.

SUBSTANCE ABUSE HISTORY:
Caffeine: Two or three cups of coffee per day, and soda at lunch time.
Tobacco: Denied.
Alcohol: One glass of wine per week. The CAGE screening questions are answered in the negative.
Illicit drugs: None at present. In high school, he tried marijuana a couple of times, and cocaine once. We discussed some of the major risk of these substances.

⋮

ABUSE HISTORY/TRAUMA/UNUSUAL CHILDHOOD EVENTS:
The patient does not really feel he was abused as a child, but there were some significant problems when his father returned from his second army tour in Vietnam. He had not met his father until 2 years of age. He states that his father verbally abused his mother. He can recall that at about age 3, his father left him on the road, in order to shut him up. His mother eventually put down her foot, and told his father to quit drinking or they would separate, and his father chose to give up alcohol. This resulted in much better family relations.

FAMILY PSYCHIATRIC HISTORY/FAMILY HISTORY:
The patient's father has suffered from posttraumatic stress disorder, as well as alcoholism. The patient's mother has had similar symptoms, possibly ADHD, and there is depression on the mother side of the family. There apparently are a number of family members with alcohol issues. The patient's grandfather had a myocardial infarction at age 40, and then died of another MI in his 50's. The patient's mother had breast cancer. His father had a stroke and hypertension. His maternal grandmother was obese and had diabetes mellitus. The maternal grandmother died of colon cancer.

SOCIAL HISTORY:
The patient was born in Grand Junction, Colorado. He came to Alaska in 1977; his father left his last term of service in the army in Germany at that time, and they came to Alaska to help a grandparent build a cabin; they ended up staying. The patient has been married for 9 years. He has two daughters, ages 8 and 6.

⋮

DIAGNOSES:
AXIS I 296.32 Major depression, recurrent, moderate.
314.00 Attention-deficit hyperactivity disorder, inattentive type.
AXIS II V71.09 No diagnosis.
AXIS III History of gastroesophageal reflux disease, status post Nissen fundoplication, variable hypertension of uncertain etiology, retinal damage from the wrestling injury, chronic back pain.
AXIS IV Occupational problems, other psychosocial and environmental problems.
AXIS V Current GAF: 54. Highest in the past year: 54.

PLAN/RECOMMENDATION:
We have checked the patient's blood pressure today, and it is 140/94. However, he is experiencing a considerable amount of back pain at this time, which likely contributes to this. We discussed some of the treatment options, and the patient will return within the next few days to have his blood pressure checked again. If it remains high, he has been instructed to see his primary care provider for further treatment. If blood pressure resolves with better pain control, we will strongly consider increasing Effexor-XR. We discussed in some detail the risks and benefits of Lunesta, Provigil, and Effexor-XR, and the patient signed a formal consent form.

Return to clinic in three weeks.

Figure 2.1: Excerpt from a psychiatric report showing section headings and content. Some sections are omitted for clarity. Additionally, an example of an implicit section is shown in the underlined text which represents a *FAMILY MEDICAL HISTORY* section that has no explicit heading or was included under a different heading (i.e., *FAMILY PSYCHIATRIC HISTORY* in this case).

uation, patient diagnosis, and suggested treatments or future steps [Groth-Marnat, 2009, Goldfinger and Pomerantz, 2013]. There are several types of psychiatric reports that vary depending on the type and purpose of assessment: Psychiatric evaluation reports, crisis evaluation reports, daily SOAP reports (Subjective, Objective, Assessment, Plan), mental status exam reports, and mini mental status exam reports, to name a few [Association, 2006]. Here I focus on psychiatric evaluation reports. Figure 2.1 shows a snippet example of a typical psychiatric evaluation report, while other reports also follow a similar structure. Although there is no one strict format, there are general guidelines that psychiatrists follow when writing psychiatric evaluation reports. Drawing from the general psychiatric evaluation domains, these reports start with the patient's identifying information, followed by the patient's chief complaints, presenting illness and its history, personal and family's medical history, mental status examination, and ending with the psychiatric medical diagnosis and treatment plan. This information is typically structured into an ordered list of headed sections [Association, 2006]. Table 2.1 contains a detailed list of the main sections of a psychiatric evaluation report in general order of appearance. Not all listed sections appear in all psychiatric evaluation reports, and they also do not necessarily appear in the same order, although there is usually a general pattern to the order.

To the best of my knowledge there was no corpus of psychiatric reports annotated with section labels, so I created my own. I collected 150 publicly available psychiatric evaluation report samples by crawling the web through custom search engines (Google Custom Search Engine for Medical Transcriptions[1] and GoogleMT[2]) and other sources [3].

[1] https://cse.google.com/cse/publicurl?cx=010964806533120826279:kyuedntb2fy

[2] https://www.googlemt.com/#gsc.tab=0

[3] http://www.medicaltranscriptionsamples.com/
http://mtsamples.com/
https://medword.com/psychiatry5.html
http://www.medicaltranscriptionsamplereports.com/
http://onwe.bioinnovate.co/psychological-assessment-example/

The reports I selected were complete and adhere to the general guidelines for psychiatric report writing discussed previously. Some of the reports were anonymized samples of real reports, while others were mock reports written for educational purposes.

#	Section	# Words	# Sent.	Sent. Length	% Present	% Implicit	κ
	GENERAL PATIENT INFO						
1	IDENTIFYING DATA	12	2	6	100	0	0.97
2	CHIEF COMPLAINT	27	3	9	100	0	0.96
	MEDICAL HISTORY						
3	HIST. OF PRSNT. ILLNSS.	232	29	8	95	10	0.92
4	PSYCHIATRIC HISTORY	85	8	11	82	36	0.91
5	SUBSTANCE ABUSE HIST.	98	10	10	88	44	0.90
6	REVIEW OF SYMPTOMS	150	19	8	96	51	0.89
7	SURGERIES	28	3	7	33	0	0.96
8	ALLERGIES	4	2	2	98	0	0.96
9	CURRENT MEDICATIONS	40	9	4	100	0	0.97
	FAMILY HISTORY						
10	BIRTH AND DEVELOP. HIST.	59	5	10	31	51	0.81
11	ABUSE HIST./TRAUMA	110	9	12	79	34	0.79
12	FAMILY PSYCHIATRIC HIST.	44	5	9	73	80	0.75
13	FAMILY MEDICAL HISTORY	48	7	7	92	38	0.79
14	SOCIAL HISTORY	80	7	11	76	45	0.84
15	PREGNANCY	29	3	8	47	64	0.81
16	SPIRITUAL BELIEFS	12	2	5	24	0	0.92
17	EDUCATION	32	3	8	68	0	0.93
18	EMPLOYMENT	31	3	9	79	0	0.91
19	LEGAL	10	1	5	20	0	0.95
	MENTAL STATUS						
20	MENTAL STATUS EXAM	155	18	9	95	11	0.78
21	STRENGTHS AND SUPPORTS	8	1	8	71	43	0.81
	TREATMENT						
22	FORMULATION	35	4	8	62	0	0.96
23	DIAGNOSES	63	12	5	100	0	0.97
24	PROGNOSIS	8	2	3	74	0	0.94
25	TREATMENT PLAN	121	12	10	100	0	0.96
	Max	232	29	12	100	80	0.97
	Average	61	7	7	75	20	0.90
	Min	4	1	2	20	0	0.75

Table 2.1: Section ontology and relevant statistics for Corpus 1: Psychiatric Evaluation Reports. All columns represent averages. The last three rows are the max, average, and min of averages.

#	Section	# Words	# Sent.	Sent. Length	% Present	% Implicit	κ
	CLINICAL INFORMATION						
1	CLINICAL HISTORY	80	8	10	100	0	0.95
	EXAM DETAILS						
2	EXAM	16	2	8	100	0	0.96
3	COMPARISON	16	2	8	86	10	0.85
4	CONTRAST	14	2	7	14	53	0.81
5	PROCEDURE	12	2	6	100	60	0.80
	FINDINGS						
6	FINDINGS	192	24	8	100	0	0.92
	IMPRESSION						
7	IMPRESSION	133	19	7	100	0	0.91
8	ATTENDING STATEMENT	-	-	-	0	-	-
	Max	192	24	10	100	60	0.96
	Average	66	8	8	75	18	0.89
	Min	12	2	6	0	0	0.80

Table 2.2: Section ontology and relevant statistics for Corpus 2: Radiology Reports. All columns represent averages. The last three rows are the max, average, and min of averages.

2.2 Corpus 2: Radiology Reports

A radiology report is a summary of a radiology scan such as an X-Ray or an MRI, where a radiologist communicates findings and an analysis of the output of the scans [of Radiology, 2019]. Similar to the previous two clinical report types, radiologists are typically trained to follow a general report guideline. Similar to psychiatric evaluations, this is not a strict format, as reports vary in their section structure and content based on the procedure performed, the patient's specific case, and the radiologist's and medical institution writing styles. Figure 2.2 shows a snippet of a radiology report.

I randomly extracted 423 radiology reports from the MIMIC-III database that were complete and adhered to the general radiology writing guidelines outlined by [of Radiology, 2019]. These reports covered a variety of procedures and scan types, including X-Rays, MRIs, and ultrasound. I used the ontology of section headers presented in [Tepper et al., 2012]. Table 2.1 shows this ontology along with detailed list of the main sections

of a radiology reports in usual order of appearance as well as relevant corpus statistics including: the average number of words per section, average number of sentences, and average sentence length. Additionally, I present the percentage of present sections per section type as well as the percentages of these sections appearing implicitly within the corpus.

EXAM: Noncontrast CT scan of the lumbar spine

REASON FOR EXAM: Left lower extremity muscle spasm.

COMPARISONS: None.

FINDINGS: Transaxial thin slice CT images of the lumbar spine were obtained with sagittal and coronal reconstructions on emergency basis, as requested.

No abnormal paraspinal masses are identified.

There are sclerotic changes with anterior effusion of the sacroiliac joints bilaterally.

There is marked intervertebral disk space narrowing at the L5-S1 level with intervertebral disk vacuum phenomenon and advanced endplate degenerative changes. Posterior disk osteophyte complex is present, most marked in the left paracentral to lateral region extending into the lateral recess on the left. This most likely will affect the S1 nerve root on the left. There are posterior hypertrophic changes extending into the neural foramina bilaterally inferiorly. There is mild neural foraminal stenosis present. Small amount of extruded disk vacuum phenomenon is present on the left in the region of the exiting nerve root. There is facet sclerosis bilaterally. Mild lateral recess stenosis just on the right, there is prominent anterior spondylosis.

At the L4-5 level, mild bilateral facet arthrosis is present. There is broad based posterior annular disk bulging or protrusion, which mildly effaces the anterior aspect of the thecal sac and extends into the inferior aspect of the neural foramina bilaterally. No moderate or high-grade central canal or neural foraminal stenosis is identified.

At the L3-4 level anterior spondylosis is present. There are endplate degenerative changes with mild posterior annular disk bulging, but no evidence of moderate or high-grade central canal or neural foraminal stenosis.

At the L2-3 level, there is mild bilateral ligamentum flavum hypertrophy. Mild posterior annular disk bulging is present without evidence of moderate or high-grade central canal or neural foraminal stenosis.

At the T12-L1 and L1-2 levels, there is no evidence of herniated disk protrusion, central canal, or neural foraminal stenosis.

There is arteriosclerotic vascular calcification of the abdominal aorta and iliac arteries without evidence of aneurysm or dilatation. No bony destructive changes or acute fractures are identified.

IMPRESSION:
1. Advanced degenerative disk disease at the L5-S1 level.
2. Probable chronic asymmetric herniated disk protrusion with peripheral calcification at the L5-S1 level, laterally in the left paracentral region extending into the lateral recess causing lateral recess stenosis.
3. Mild bilateral neural foraminal stenosis at the L5-S1 level.
4. Posterior disk bulging at the L2-3, L3-4, and L4-5 levels without evidence of moderate or high-grade central canal stenosis.
5. Facet arthrosis to the lower lumbar spine.
6. Arteriosclerotic vascular disease.

Figure 2.2: Excerpt from a radiology report showing section headings and content. Some sections are omitted for clarity.

HISTORY OF PRESENT ILLNESS: Mr. ABC is a 60-year-old white male veteran with multiple comorbidities, who has a history of bladder cancer diagnosed approximately two years ago by the VA Hospital. He underwent a resection there. He was to be admitted to the Day Hospital for cystectomy. He was seen in Urology Clinic and Radiology Clinic on MM/DD/YYYY.

HOSPITAL COURSE: Mr. ABC presented to the Day Hospital in anticipation for Urology surgery. On evaluation, EKG, echocardiogram was abnormal, a Cardiology consult was obtained. A cardiac adenosine stress MRI was then proceeded, same was positive for inducible ischemia, mild-to-moderate inferolateral subendocardial infarction with peri-infarct ischemia. In addition, inducible ischemia seen in the inferior lateral septum. Mr. ABC underwent a left heart catheterization, which revealed two vessel coronary artery disease. The RCA, proximal was 95% stenosed and the distal 80% stenosed. The mid LAD was 85% stenosed and the distal LAD was 85% stenosed. There was four Multi-Link Vision bare metal stents placed to decrease all four lesions to 0%. Following intervention, Mr. ABC was admitted to 7 Ardmore Tower under Cardiology Service under the direction of Dr. XYZ. Mr. ABC had a noncomplicated post-intervention hospital course. He was stable for discharge home on MM/DD/YYYY with instructions to take Plavix daily for one month and Urology is aware of the same.

DISCHARGE EXAM:
VITAL SIGNS: Temperature 97.4, heart rate 68, respirations 18, blood pressure 133/70.
HEART: Regular rate and rhythm.
LUNGS: Clear to auscultation.
ABDOMEN: Obese, soft, nontender. Lower abdomen tender when touched due to bladder cancer.
RIGHT GROIN: Dry and intact, no bruit, no ecchymosis, no hematoma. Distal pulses are intact.

PROCEDURES:
1. On MM/DD/YYYY, cardiac MRI adenosine stress.
2. On MM/DD/YYYY, left heart catheterization, coronary angiogram, left ventriculogram, coronary angioplasty with four Multi-Link Vision bare metal stents, two placed to the LAD in two placed to the RCA.

DISCHARGE INSTRUCTIONS: Mr. ABC is discharged home. He should follow a low-fat, low-salt, low-cholesterol, and heart healthy diabetic diet. He should follow post-coronary artery intervention restrictions. He should not lift greater than 10 pounds for seven days. He should not drive for two days. He should not immerse in water for two weeks. Groin site care reviewed with patient prior to being discharged home. He should check groin for bleeding, edema, and signs of infection. Mr. ABC is to see his primary care physician within one to two weeks, return to Dr. XYZ's clinic in four to six weeks, appointment card to be mailed him. He is to follow up with Urology in their clinic on MM/DD/YYYY at 10 o'clock and then to scheduled CT scan at that time.

DISCHARGE DIAGNOSES:
1. Coronary artery disease status post percutaneous coronary artery intervention to the right coronary artery and to the LAD.
2. Bladder cancer.
3. Diabetes.
4. Dyslipidemia.
5. Hypertension.
6. Carotid artery stenosis, status post right carotid endarterectomy in 2004.
7. Multiple resections of the bladder tumor.
8. Distant history of appendectomy.
9. Distant history of ankle surgery.

Figure 2.3: Excerpt from a hospital discharge summary showing section headings and content. Some sections are omitted for clarity.

2.3 Corpus 3: Hostpital Discharge Summaries

A discharge summary is the final documentation of a terminated hospital stay. Physicians supervising hospitalized patients usually do not follow patients outside the hospital. This creates a discontinuity of care that is addressed through a discharge summary. Discharge summaries are the means for this communication between inpatient and outpatient physicians. These reports are intended to summarize the course of hospital treatment by listing

the various events during hospitalization, thus preparing the outpatient physician to resume care of the patient [Horwitz et al., 2013]. Like the two previous corpora, discharge summaries are governed by general writing guidelines that suggest the information that should be included. In practice, different hospital networks and even different medical professionals within the same hospital often write these reports differently, tailoring them to specific patient cases. Figure 2.3 shows a snippet of a discharge summary document.

Similar to radiology reports, I randomly extracted 150 discharge summaries from the Medical Information Mart for Intensive Care III (MIMIC-III) database [Johnson et al., 2016]. I selected discharge summaries that were complete and that adhere to the general guidelines of medical note writing. As is the case with all MIMIC-III data, the summaries are anonymized. I used the ontology of section headers presented in [Tepper et al., 2012]. Table 2.3 shows this ontology along with a detailed list of the main sections of a discharge summary reports in usual order of appearance as well as relevant corpus statistics including: the average number of words per section, average number of sentences, and average sentence length. Additionally, I present the percentage of present sections per section type as well as the percentages of these sections appearing implicitly within the corpus.

2.4 Corpus 4: Patent Documents

Patent documents are the result of a successful patent application. Many of a patent's sections are mandatory, e.g., the claims section [WIPO, 2007]. Similarly, the description section in these documents is further composed of subsections, some of which are mandatory, while others are optional and depend on the authors' preferences as well as the patent's technical topics. In their work on patent section segmentation, [Brügmann et al., 2015] outlined the structure of the description section in a patent document into five mandatory and two optional segments.

#	Section	# Words	# Sent.	Sent. Length	% Present	% Implicit	κ
	GENERAL PATIENT INFO						
1	ADMIT DATE	3	1	3	100	0	0.97
2	DISCHARGE DATE	3	1	3	100	0	0.97
3	SERVICE	4	2	2	100	0	0.96
	PROVIDER INFO						
4	ATTENDING	2	1	2	82	0	0.96
5	ADMIT PHYSICIAN	2	1	2	100	0	0.95
6	DISCHARGE PHYSICIAN	2	1	2	100	0	0.95
	CONDITION BEFORE ADMISSION						
7	ADMISSION DIAGNOSES	96	12	8	100	0	0.92
8	HISTORY	135	15	9	76	58	0.85
9	MEDICATIONS	55	11	5	100	0	0.96
10	REASON FOR ADMISSION	162	18	9	100	0	0.95
	CONDITION AT DISCHARGE						
11	CONDITION	4	2	2	100	0	0.96
12	DISPOSITION	2	1	2	34	10	0.92
13	DISCHARGE DIAGNOSES	144	18	8	89	37	0.89
14	OTHER DIAGNOSES	-	-	-	0	-	
15	PHYSICAL EXAM ON DISCH.	45	9	5	40	38	0.90
	MEDICAL HISTORY						
16	ALLERGIES	12	3	4	100	0	0.96
17	FAMILY HISTORY	81	9	9	43	20	0.89
18	GYNECOLOGICAL HISTORY	-	-	-	0	-	
19	PAST MEDICAL HISTORY	144	16	9	100	41	0.82
20	PAST SURGICAL HISTORY	32	4	8	100	58	0.83
21	SOCIAL HISTORY	84	7	12	37	66	0.88
	HOSPITAL COURSE						
22	CONSULTATION	88	11	8	6	0	0.96
23	HOSPITAL COURSE	168	14	12	85	0	0.95
24	PHYSICAL	66	11	6	28	13	0.89
25	PROCEDURES	15	5	3	65	10	0.91
26	STUDIES	-	-	-	0	0	
	DISCHARGE INSTRUCTIONS						
27	FOLLOW UP	-	-	-	0	-	
28	DIAGNOSTIC STUDIES REC'D	-	-	-	0	-	
29	DISCHARGE INSTRUCTIONS	408	34	12	100	0	0.96
30	DISCHARGE MEDICATIONS	72	12	6	100	0	0.97
	Max	408	34	12	100	66	0.97
	Average	73	9	6	62	13	0.94
	Min	2	1	2	0	0	0.82

Table 2.3: Section ontology and relevant statistics for Corpus 3: Hospital Discharge Summaries. All columns represent averages. The last three rows are the max, average, and min of averages.

In my work, I focus on the description section of patent documents and refer to those as patent documents in my discussion throughout this paper. I randomly collected 464

#	Section	# Words	# Sent.	Sent. Length	% Present	κ
1	TECHNICAL FIELD	85	3	8	100	0.95
2	BACKGROUND ART	267	57	11	100	0.96
3	SUMMARY OF THE INVENTION	1,286	89	10	100	0.94
4	DESCRIPTION OF DRAWINGS	975	19	8	100	0.92
5	PREFERRED EMBODIMENTS	4,106	208	7	100	0.95
6	INDUSTRIAL APPLICABILITY	2,731	96	2	31	0.87
7	EXAMPLES	1,258	82	4	14	0.89
	Max	4,106	208	2	14	0.87
	Average	1,530	79	7	78	0.92
	Min	85	3	11	100	0.96

Table 2.4: Section ontology and relevant statistics for Corpus 4: U.S. Patent Documents. All columns represent averages. The last three rows are the max, average, and min of averages.

U.S. patent documents using the PATENTSCOPE database [WIPO, 2019] provided by the World Intellectual Property Organization (WIPO). The documents spanned the period between 1954 and 2010. We then extracted the description sections from the original patent documents to construct our corpus. Finally, we used the ontology of section types presented in [Brügmann et al., 2015]. Table 2.4 lists the main section types in their usual order of appearance and how often they occur in our corpus and their relevant statistics.

2.5 Corpus 5: Environmental Scientific Articles

The environmental scientific corpus was the result of an interdisciplinary collaborative project between computer scientists (including myself and other colleagues at the School of Computing and Information Sciences) and environmental scientists at Florida International University's Earth and Environment department. The project was supported by the CREST Center for Aquatic Chemistry and Environment at FIU, and was aimed at developing a domain-specific ontology-based semantic search engine for the environmental scientific literature. To the best of our knowledge there was no corpus of scientific articles annotated with ENVO concepts, so we created our own. We collected a total of 19

Query	Title	Citation	Tokens	Sentences	Unique Concepts	κ
Methyl-Mercury concentrations in Everglades water and sediment	Mercury in the Aquatic Environment ...	[Ullrich et al., 2001]	5,081	162	26	n/a
	Sulfide Controls on Mercury Speciation ...	[Benoit et al., 1999]	4,133	168	13	n/a
	Sulfate Stimulation of Mercury Methylation ...	[Gilmour et al., 1992]	3,642	160	18	n/a
	Effect of Salinity on Mercury Activity ...	[Compeau and Bartha, 1987]	3,421	150	22	n/a
Sulfate reduction occurring in Everglades pore waters and sediments	Anaerobic Microflora of Everglades Sediments ...	[Drake et al., 1996]	4,651	179	35	0.64
	Constants for mercury binding ...	[Benoit et al., 2001]	4,629	173	17	0.62
	Mercury methylation in periphyton ...	[Cleckner et al., 1999]	3,839	159	18	0.75
	Methylmercury Concentrations ...	[Gilmour et al., 1998]	4,295	183	26	0.30
	Bacterial Methylmercury Degradation ...	[Marvin-DiPasquale and Oremland, 1998]	3,696	199	27	0.44
Sulfur reduction affecting South Florida Everglades soils	Groundwater's significance to changing ...	[Harvey and McCormick, 2009]	9,650	300	73	0.63
	Variation in Soil Phosphorus ...	[Chambers and Pederson, 2006]	3,032	103	39	0.71
	Sulfur in the South Florida ecosystem ...	[Orem et al., 2011]	3,485	149	37	0.69
	Sulfur in peat-forming systems ...	[Casagrande et al., 1977]	3,998	165	35	0.71
	Effects of sulfate amendments ...	[Dierberg et al., 2011]	4,463	160	42	0.62
Everglades groundwater surface water interaction	Coastal groundwater discharge – an additional ...	[Price et al., 2006]	4,445	198	32	0.85
	The Influence of Hydrologic Restoration ...	[Sullivan et al., 2014]	5,860	220	28	0.81
	Ground Water Recharge and Discharge ...	[Harvey et al., 2004]	6,257	223	36	0.88
	Estimates of groundwater discharge ...	[Zapata-Rios and Price, 2012]	6,480	307	48	0.83
	Quantifying time-varying ground-water ...	[Choi and Harvey, 2000]	4,747	186	46	0.81
		Max	9,650	307	73	0.88
		Average	4,741	186	33	0.69
		Min	3,032	103	13	0.30
		Standard Deviation	1,604	43	15	0.15

Table 2.5: Articles used in Corpus 5: Environmental Scientific Articles. Listed are the number of tokens in each article, the number of sentences overall, the number of unique concepts, and the annotator agreement expressed as Cohen's κ.

articles (90,074 total words) using four search queries that were created by three domain experts (two PhD students and a professor of Hydrology). Our domain experts ran the queries through Google Scholar and examined from the several hundred results returned, identifying the top four or five most relevant articles for each query. Importantly, several of the articles were not ranked near the top of Google's results, and were rather found many pages deep.

We then manually annotated articles at the sentence level in two separate annotation studies: The first pertained ontological semantic concepts (i.e. scientific concepts found in the articles), and the second involved the section structure of the articles similar to the previous corpora. Regarding the semantic concept annotation, and in an pilot study of ours [Eisenberg et al., 2017], we determined that the most useful ontology for our purposes was the Environment Ontology (ENVO), a community-led, open ontology for various life science disciplines [Buttigieg et al., 2013]. According to its creators, ENVO is an attempt at establishing a standard annotation scheme for several co-dependent or

related disciplines, including, but not limited to, ecology, hydrology, environmental biology, and the geospatial sciences. ENVO contains concepts corresponding to a wide range of natural environments and environmental conditions. It is encoded in the Open Biomedical Ontologies (OBO) syntax, which is a subset of the Web Ontology Language (OWL). ENVO can be populated, managed, and maintained using the OBO-Edit ontology development tool. We annotated the articles at the sentence level using concepts from ENVO (the following sections discuss the annotation study in detail).Table 2.5 lists the queries, the corresponding articles returned from the search results, as well as article-specific statistics. The articles have an average of 4,741 tokens, 186 sentences, 261 unique ENVO concepts.

ENVO, like many ontologies, is hierarchical in design. Three of its top-level, most developed branches are *environmental system*, *environmental feature*, and *environmental material*. It's hierarchical structure allows for it to include not only entities, but also higher-level relationships between various concepts, including many standard ontological relationships such as `is-a`, `part-of`, `contained-in`, `connects`, and `has-condition`. ENVO also contains scientific and domain-specific relationships such as `derives-from`, `input-of`, `output-of`, `has-habitat`, and `biomechanically-related-to`. Furthermore, the ontology boasts a well-connected graph of synonymy relationships, encoded using different granularities including `broad`, `exact`, and `narrow`.

Additionally, Table 2.6 shows the ontology chosen for scientific article section structure along with a detailed list of the main sections in usual order of appearance as well as relevant corpus statistics including: the average number of words per section, average number of sentences, and average sentence length. I also present the percentage of present sections per section type as well as the percentages of these sections appearing implicitly within the corpus. the ontology of sections and the as well as the relevant statistics.

#	Section	# Words	# Sent.	Sent. Length	% Present	% Implicit	κ
	INTRODUCTION						
1	BACKGROUND	800	35	23	100	0	0.96
2	PROBLEM	400	19	21	100	61	0.88
3	METHOD	1,413	53	27	100	0	0.95
4	RESULT	1,925	84	23	100	0	0.96
	RELATED WORK						
5	CONNECTION	356	21	17	100	100	0.85
6	DIFFERENCE	281	14	20	100	100	0.79
7	FUTURE WORK	350	20	18	40	36	0.83
8	CONCLUSION	205	10	21	100	0	0.96
	Max	1,925	84	27	100	100	0.96
	Average	716	32	21	93	37	0.90
	Min	205	10	17	40	0	0.79

Table 2.6: Section ontology and relevant statics for Corpus 6: Environmental Scientific Articles. All columns represent averages. The last three rows are the max, average, and min of averages.

2.6 Corpus 6: News Articles

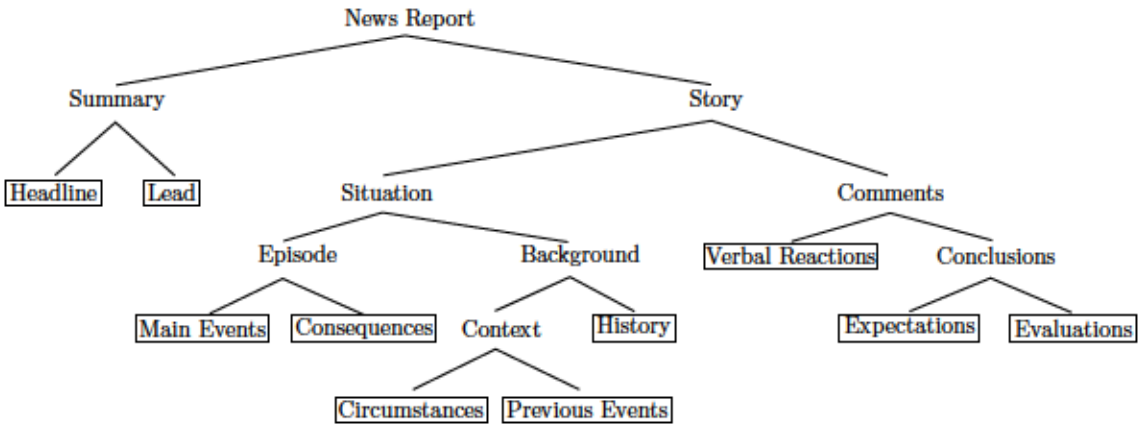


Figure 2.4: The hierarchical discourse structure of news proposed by van Dijk [van Dijk, 1988]. Boxes indicate labels that were directly annotated on the documents; other labels can be inferred. From Yarlott et al. [2018], Figure 1.

For the news domain, I used a gold-standard corpus previously developed by Yarlott et al. [2018] of van Dijk's [1988] labels (Figure 2.4) applied to a subset of the Automated Content Extraction (ACE) Phase 2 corpus [NIST, 2002]. The ACE Phase 2 corpus is a

major standard corpora of news articles that boasts three advantages: it is widely-used, has relevance to other tasks, and was readily available to researchers. This dataset comprises 50 documents containing 28,236 words divided in 644 paragraphs.Table 2.7 shows the corpus-wide statistics for the number of words and paragraphs, where each paragraph is given a single type in accordance to van Dijk's theory.

	Words	Paragraphs
Total	28,236	644
Average	564.7	12.9
Std. Dev.	322.1	4.9

Table 2.7: Corpus-wide statistics for the annotated data for Corpus 6: News Articles. Adapted from Yarlott et al. [2018], Table 1.

2.7 Annotation Processes

There was a total of seven annotation studies conducted for this research: Six studies involved the annotation of documents for their section structure (one for each corpus), five of which I developed and served as an annotator and (or) adjudicator, while I only served as an annotator for the news articles corpus study. Additionally, I co-developed and managed an extensive seventh study that involved the annotation of scientific articles for their semantic ontological concepts. The remainder of this section discusses the annotation process for each of these corpora in detail.

2.7.1 Corpora 1-4

I prepared the psychiatric evatluation corpus in two stages. First, I standardized the labels' names, selecting a single uniform name for each section type and mapping corresponding section labels found in the corpus to those names. For example, some reports contained

the section *SCHOOL* while others listed it as *EDUCATION*. Here I selected *EDUCATION* as the uniform section label across all reports.

Second, I created a hierarchy for the section names which reflected implicit embedded sections types that I found in the corpus. There were only three section types that included implicit subsections in this corpus, namely, *MEDICAL HISTORY*, *FAMILY HISTORY*, and *MENTAL STATUS*. For example, some reports containing the section *MENTAL STATUS* might in turn include information in that section about both *MENTAL STATUS EXAM* and *STRENGTHS AND SUPPORTS*. In this case I identified these implicit subsection boundaries (that is, the boundaries were not identified with a section header) and labeled those subsections with both the parent and child label. Table 2.1 lists the the parent sections that sometimes included other sections implicitly (emphasized in bold), the unified list of section types found in the collected reports (numbered sections), word and sentence level statistics, and percentage of reports containing those sections in the corpus. For both of these stages I used all 150 reports.

As for the next three corpora, I used section ontologies developed in previous studies as discussed in the previous sections. Annotation was done (over the course of a month, approximately, for each corpus) in a double-blind manner by three annotators: myself, a computer science undergraduate student with an uncompleted medical degree, and a medical doctor who also acted as an adjudicator for the medical corpora (Corpora 1-3), while I acted as the adjudicator for the patent documents corpus (Corpus 4). The annotators that took part in this project were given minimal training outside of their individual experience with annotation studies. They were provided with the psychiatric report writing guide [Association, 2006], the American College of Radiology standards handbook [of Radiology et al., 2018], as well as other templates and guidelines for discharge summary writing, to use as a reference in each corpus annotation. Additionally, an annotation guide that included the ontology of labels for the documents in each corpus was provided.

The annotators were instructed to annotate continuous blocks of text that correspond to each label in the ontology with no repetition of labels. That is, a label can not occur more than once in a report and a label block can not be segmented.

2.7.2 Corpus 5: Environmental Scientific Articles

As discussed previously, there were two annotation studies for the scientific articles corpus. The first pertained ontological semantic concepts (i.e. scientific concepts found in the articles), and the second involved the section structure of the articles similar to the previous corpora.

Semantic Concept Annotation

The purpose of manually annotating semantic concepts from the ontology was twofold: first, to show that the ontological concepts appear in the target texts and, second, to show that it is possible to automatically learn domain-specific concepts from a relevant ontology. Because developing concept detectors is a non-trivial task, in prior work my co-authors and I tested the utility of the ontology, as well as verified that it is feasible to automatically rank articles using detected ontological concepts [Eisenberg et al., 2017]. We then expanded that effort by creating a larger gold-standard corpus and demonstrating that we can identify the concepts in the articles automatically.

As discussed above, we collected a corpus of 19 articles from the biogeochemical and hydrological domains, aligned with three search queries. Our team of domain trained annotators then annotated the queries and the articles for concepts from ENVO. For each article, annotations were collected at the sentence level. Our annotation team was composed of one PhD student in hydrology and four Earth and Environment undergraduate students at FIU.

More than 20 years ago, Andren & Harriss (1973) measured relatively high % MeHg (MeHg as a percent of total Hg) in Everglades sediments, noting that samples from the Everglades were comparable to Hg-contaminated Mobile Bay sediments. [Gilmour et al., 1998, p. 328]

Text Span	Concept	ID
Everglades sediments	sediment	2007
Everglades	peat swamp	189
Mobile Bay sediments	sediment	2007

Figure 2.5: Example sentence from article [Gilmour et al., 1998, p.328]. Underlined portions of the text indicate spans that were associated with an ENVO concept; the table shows the associated ENVO concept ID.

Annotators used Protégé [Musen, 2015] to search and explore ENVO when deciding what concepts should marked for each sentence of each article. Annotators recorded their annotations in a spreadsheet, where each row represented a sentence, followed by columns representing the span of text containing the concept and the ID of the identified concept.

Figure 2.5 gives an example sentence from one of the test articles, along with the text spans which were associated with an ENVO concept.

The process of annotation involved several rounds of training, annotating, and revision of the annotation guidelines. Even for a relatively simple sentence as shown in Figure 2.5, numerous annotation decisions were needed. Below, I walk through this process phrase by phrase:

More than 20 years—This phrase does not need to be annotated, as it is a temporal expression referring to time period of the events mentioned later in the sentence.

... Andren & Harris (1973)—This phrase also does not need to be annotated, because it is a reference to a relevant article, and referring to the scientific literature isn't a concept in ENVO.

... measured relatively high %—This does not need to be annotated, as ENVO does not contain concepts related to specific chemical concentration levels.

... *MeHg*—This is the chemical formula for *methylmercury*, an environmental contaminant. The concepts of *contaminant* and *contamination* are not in ENVO. However, because this concept is relevant to the domain of interest, I did record these text spans and their related ideas so as to begin to build a set of concepts to expand ENVO in future work.

... *(MeHg as a percent of total Hg)*—Again, we identified the spans *MeHg* and *Hg* as the missing concept *contaminant*.

... *in Everglades sediments*—This phrase is tricky, because *Everglades* and *sediment* appear as individual concepts in ENVO, but when they appear in succession they form a multiword expression. *Everglades sediment* does not appear directly in ENVO. However, as it is presumably a subclass (or multiple subclasses) of sediments generally, we queried ENVO for the entity *sediment* (ENVO ID 2007), and examined its children for potential matches. *Sediment* has multiple children, namely, specific subtypes such as *lake sediment* or *contaminated sediment.* However, because there is no concept corresponding to the specific collection of different types of sediments that comprise the Everglades, we tagged this with the more general entity *sediment.*

... *noting that samples from the Everglades*—For this span, we first looked through ENVO to find a concept for *Everglades.* The closest concept is *peat swamp* (ENVO ID 189), which has no children, and so we tag this span using this concept.

... *were comparable to Hg-contaminated Mobile Bay sediments.*—For this span, we again tagged *Hg* as the missing concept *contaminant*. In the same way as above for *Everglades sediment*, the phrase *Mobile bay sediments* was tagged with the general concept *sediment*.

The first four articles were the result of a previous pilot annotation study [Eisenberg et al., 2017]. The first three co-authors in that study and a domain expert served as the annotators for those articles, and were annotated as follows: we annotated the first 50

sentences of one of the articles [Ullrich et al., 2001] cooperatively to develop the annotation guidelines, while each annotator annotated the remaining 130 sentences individually so as to allow us to calculate inter-rater reliability. After these first articles was finished, we then assigned each of the annotators one of the four remaining articles for annotation [Gilmour et al., 1998, Benoit et al., 1999, Gilmour et al., 1992, Compeau and Bartha, 1987]. The remaining ten articles were doubly annotated by a new team of trained annotators and domain experts following the developed annotation guidelines.

Section Structure Annotation

As for the annotation of section structure, I led a seperate smaller annoation study with the help of another CS PhD student. We annotated the 19 articles according to the AZ scheme that describes the rhetorical progression of scientific text. The scheme was originally introduced by Teufel and Moens [2002], who applied it to computational linguistics articles. I used the version that Mizuta et al. [2006] adapted for biology articles, with minor modifications concerning zone names as it was done by Guo et al. [2013] in their study on identifying information zones in scientific articles. The annotation of these articles was done over the span of two weeks in a double-blinded manner, were we annotated contigiguous blocks of text (at the sentence level) for their respective containing sections from the ontology shown in Table 2.6. The sections *CONNECTION* and *DIFFERENCE* were used for contiguous spans of text (sentences) that discussed related work. The other sections are self explanatory.

2.7.3 Corpus 6: News Articles

As stated before, the news articles corpus and annotation study was produced by [Yarlott et al., 2018], where I served as one of three annotators. I include the discussion of the

corpus, annotation process and results directly from [Yarlott et al., 2018] for legibility and ease of reference.

Annotation was performed over the course of a month, as time allowed. The adjudicator performed annotation of all ten sets of documents, while the other two annotators performed annotation of six sets each. Figure 2.6 illustrates this division of work. Annotation of each set took approximately 45 minutes to an hour, resulting in roughly ten hours of annotation work for the adjudicator and six hours for the other two annotators. The annotations were performed using Microsoft Word's built-in comment feature, to eliminate the need for any tool-based annotator training. When confronted with multiple labels that seemed to fit, annotators were instructed to choose the label that seemed the most applicable.

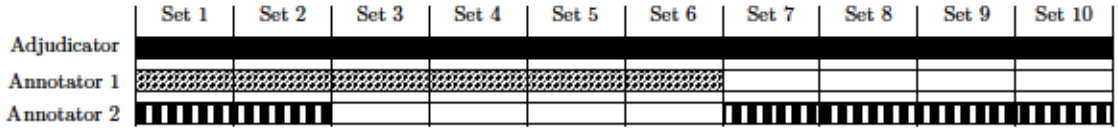


Figure 2.6: Division of work for annotation study of the news articles corpus. From Figure 3 in [Yarlott et al., 2018].

The adjudication procedure took a further hour for each set of documents, resulting in another ten hours of work for the adjudicator and another two hours for the other two annotators, who were only required to participate in adjudication of the first two sets of documents. The purpose of this group adjudication meeting was to resolve any outstanding questions or confusions regarding the annotation procedure. The annotation resulted in triple annotation for the first ten documents, and double annotation for the remaining forty documents. The multiple annotations were merged into a gold standard for every document. Additionally, although annotators were instructed to annotate the headline for each document, these labels are not included as part of the gold standard because within the ACE Phase 2 dataset, the headlines themselves are clearly annotated.

Annotation was done in a double-blind manner by three annotators, one of whom also acted as the adjudicator. All three annotators are Ph.D. students in computer science with a focus on natural language processing, with experience in both annotating and running annotation studies.

The annotators that took part in this project were given minimal training outside of their individual experience with annotation studies. Annotators were provided with a guide describing van Dijk's theory. A single adjudication meeting was held after annotation for the first two sets of documents was completed. The primary purpose of this adjudication meeting was to resolve any questions the annotators had, discover any uncertainty in the annotation guide, and revise the annotation guide to address these questions. The annotation guide contains descriptions of each discourse label in addition to an example of a fully-annotated news article, shown in Figure 2.7.

2.8 Agreement Metrics

To measure the reliability of the annotation data, I calculated the Cohen Kappa measure for inter-annotator agreement. A value for Cohen's Kappa [Landis and Koch, 1977] is calculated for each section. The ideal value of Cohen's Kappa is 1, which denotes perfect agreement. The range for Cohen's Kappa is between -1 and 1. Any value of Kappa below 0 is considered to have no agreement, between 0 and 0.2 "slight" agreement, between 0.2 and 0.4 "fair" agreement, between 0.4 and 0.6 "moderate" agreement, between 0.6 and 0.8 "substantial" agreement, and 0.8 and 1 is almost "perfect" agreement [Landis and Koch, 1977]. To calculate Cohen's Kappa, I populate a confusion matrix (Table 2.8) for each section (label).

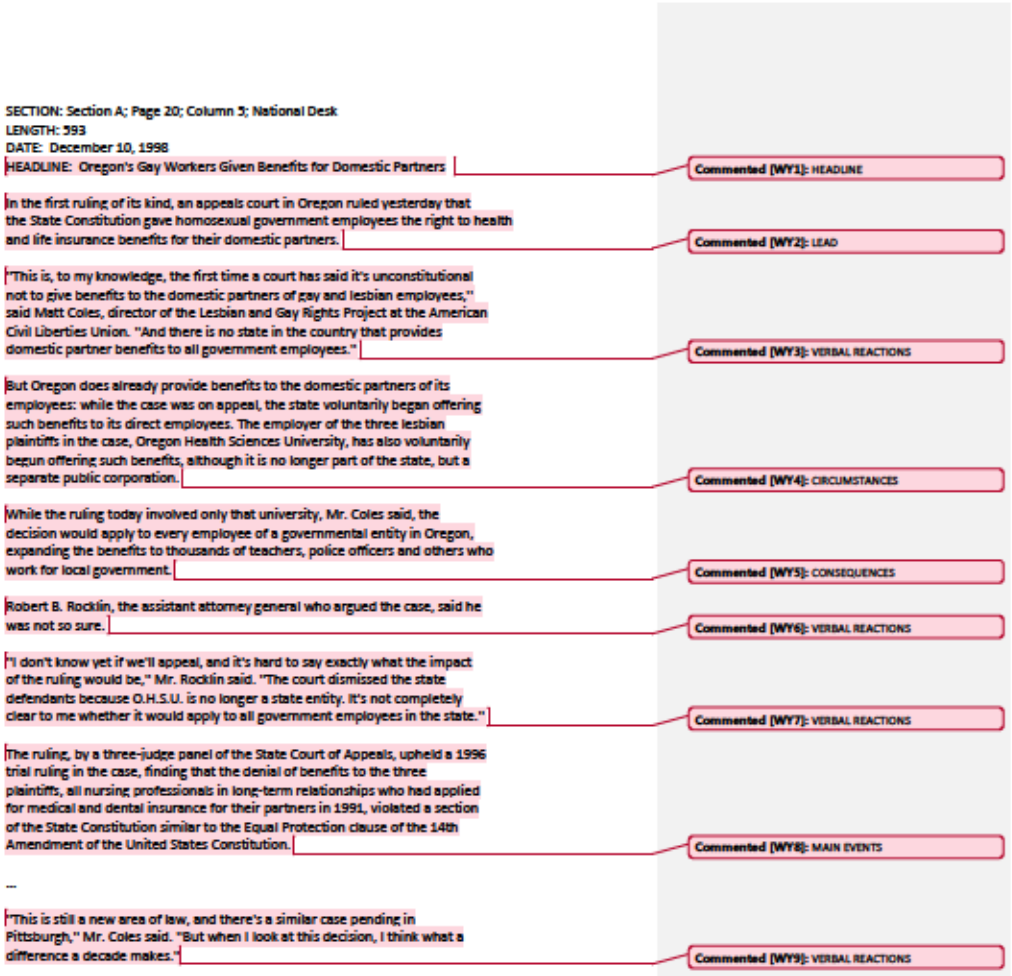


Figure 2.7: Example annotation included in the annotation guide. Some parts of the annotation have been omitted for brevity. From Figure 2 in [Yarlott et al., 2018]

		Annotator 1	
Annotator 2		true positives	false positives
		true negatives	false negatives

Table 2.8: Confusion matrix for Cohen's Kappa calculation.

2.9 Annotation Results

In this section I discuss the annotation results for each of the corpora and their respective annotation studies discussed earlier in order.

2.9.1 Coprpora 1-4

The annotation results for each sof the first four corpora are shown in the last column in each of their respective above tables (Tables 2.1, 2.1, 2.3, and 2.4). The annotation studies resulted in the highest agreement for discharge summaries with a Cohen's κ of 0.94 for discharge summaries, followed by patent documents with a 0.92, 0.90 for the psychiatric evaluation reports, and 0.89 for the radiology reports, all of which are considered "perfect" agreement. Discharge summaries and patent documents contained more sections with clearly distinct language as opposed to the psychiatric evaluation and radiology reports which contained a lot of sections that shared similar language and concepts such as the personal and family medical and psychological history sections.

The tables above also show the inter-annotator agreement for each section in each corpus. Sections that never appear implicitly were high in agreement, while the annotators disagreed more on sections that often appeared implicitly. Section with high prevalence and are implicit saw the least agreement.

2.9.2 Corpus 5: Environmental Scientific Articles

For the semantic concept annotation study, and as discussed above, the first four articles (Table 2.5) were the result of a previous pilot annotation study [Eisenberg et al., 2017]. The first three co-authors (including myself) and a domain expert served as the annotators for those articles. This produced a Cohen's κ of 0.57, which is "moderate to substantial" agreement [Artstein and Poesio, 2008]. The remaining 15 articles were annotated by larger group of domain experts as discussed previously. The resulting micro-averaged inter-annotator measure agreement over all annotator groups using Cohen's κ is 0.61 which is "substantial" agreement [Artstein and Poesio, 2008]. I also report per-document κ measures. I report a κ with zeroes columns and rows removed. This refers to

Comparison	# Docs	P	R	F_1	p_0	p_e	κ
A1 vs. A2	10	0.76	0.79	0.77	0.63	0.18	0.55
Adj. vs. A1	30	0.81	0.85	0.83	0.71	0.19	0.64
Adj. vs. A2	30	0.80	0.83	0.82	0.69	0.18	0.62
A1 vs. Gold	30	0.93	0.92	0.92	0.86	0.19	0.83
A2 vs. Gold	30	0.92	0.90	0.91	0.83	0.19	0.80
Adj. vs. Gold	50	0.93	0.87	0.90	0.81	0.18	0.77

Table 2.9: Microaveraged agreement measures between the annotators (A1, A2), adjudicator (Adj.), and the merged gold standard (Gold)—including precision (P), recall (R), balanced F-measure (F_1), relative observed agreement among raters (p_0), probability of chance agreement (p_e), and Cohen's kappa (κ, derived from p_0 and p_e). From Table 2 in [Yarlott et al., 2018]

the following situation: when analyzing the confusion matrix for a given concept, if there was a row or column that only contained the number 0, I removed it from the calculation of the average κ. I justify this because situations where there is a row or column consisting of only zeroes means that the annotators consistently marked a certain concept as two different things. An example of this is an annotator consistently marking a set of spans as the concept *watercourse*, and the other annotator consistently marking the same span as *watershed*, which are two similar concepts. They were marking the same span as different concepts, and each annotator always made the same decisions, but the problem was with what they called the concept. They were consistent, which is qualitatively represented by the fact that there is a column or row in the confusion matrix of all 0's. Due to the consistency of the mislabeling, I can justify removing their κ's from the calculation of the average κ.

As for the section structure annotation study, the resulting average Cohen κ was 0.90 which is considered "perfect" agreement. The last column in Table 2.6 shows the resulting agreement per section. Similar to the previous four corpora sections (e.g. *BACKGROUND*, *RESULTS*) that never appear implicitly were high in agreement, while the annotators disagreed more on sections that often appeared implicitly (e.g. *CONNECTION*,

FUTURE WORK). Section with high prevalence and are implicit saw the least agreement (e.g. *DIFFERENCE*).

2.9.3 Corpus 6: News Articles

This section is repeated from Yarlott et al.'s article 2018 for ease of reference. The annotation study had two goals: first, to produce a benchmark dataset of document-level discourse annotations to evaluate the impact of document-level discourse on information extraction. Second, to evaluate whether or not humans can reliably apply van Dijk's theory to actual documents. That is, the annotators have a high degree of agreement with respect to each other. To measure agreement, I use the standard F_1 score [van Rijsbergen, 1979], treating one of the annotators as the correct labels, as well as Cohen's kappa coefficient for inter-rater agreement [Cohen, 1968].

Label	Count
Lead	42
Main	60
Consequences	19
Circumstances	103
Previous Events	64
History	27
Verbal Reactions	252
Expectations	21
Evaluations	56
Total	644

Table 2.10: Distribution of the labels within Corpus 5: News Articles. The majority of paragraphs fall under the categories of verbal reactions or circumstances.

The results of the annotation study are shown in Table 2.9. Inter-annotator agreement between annotators A1 and A2 was measured over ten documents; inter-annotator agreement between the annotators and the adjudicator, as well as the annotators and the gold

standard, was measured over 30 documents. The comparison between the adjudicator and the gold standard was measured over the entire collection of 50 documents.

Finally, Table 2.10 provides the distribution of van Dijk's labels (sans headlines, of which there are 50: one for each document, were annotated within the ACE Phase 2 corpus). Verbal reactions and circumstances dominate the labels.

CHAPTER 3

AUTOMATIC SECTION STRUCTURE IDENTIFICATION

The first task in section structure extraction is the identification of sections, their positions and boundaries in various documents. As I described in detail in chapter 2, some document classes have a strict section structure, most however, do not . In this chapter, I present three studies for section structure identification. The first (§3.1), uses an Hierarchical Hidden Markov Model (HHMM) that was developed using the psychiatric evaluation reports (Corpus 2.1). The second (§3.2), extends the HHMM approach by using Conditional Random Fields (CRFs) which I developed using three corpora: psychiatric evaluation reports, radiology reports, and discharge summaries (Corpora 2.1-2.3). Finally, in the third (§3.3), I present an extended application of the CRF approach to improving the detection of paragraph functions in news article paragraphs (Corpus 2.6).

3.1 Using Hierarchical Hidden Markov Models to Automatically Identify Sections in Psychiatric Reports

Psychiatric evaluation reports represent a rich and still mostly-untapped source of information for developing systems for automatic diagnosis and treatment of mental health problems. As discussed previously in §2.1, these reports contain free-text structured within sections using a convention of headings. In this section, I present a model for automatically detecting the position and type of different psychiatric evaluation report sections.

3.1.1 Motivation

With the exponential growth of free text in electronic health records (EHRs)—which includes mental health documents—it is ever more important to develop natural language

processing (NLP) models that automatically understand and parse such text. When incorporated in other systems, these models may aid (1) clinical decision support, (2) the extraction of key population information and trends, and (3) precision medicine efforts where personalized information and trends are extracted and used in the treatment process [Demner-Fushman et al., 2009, Hripcsak et al., 2003].

The majority of clinical NLP work has focused on semantic parsing of clinical notes found in EHRs. There are several challenges in automatic understanding of unstructured text in EHRs, encompassing many levels of linguistic processing: identifying document layouts, their discourse organization, mapping lexical information to semantic concepts found in biomedical ontologies, as well as understanding inter-concept co-reference and temporal relations [Li et al., 2010]. These challenges are also present for mental health NLP applications.

I present an approach to automatically model the discourse structure of psychiatric reports as well as segment these reports into various sections. This model learns the section types, positions, and sequence and can automatically segment unlabeled text in a psychiatric report into the corresponding sections. I hypothesize that knowledge of the ordering of the sections can improve the performance of a section classifier and a text segmenter. To test this hypothesis, I trained a Hierarchical Hidden Markov Model (HHMM) that categorizes sections in psychiatric reports into one of 25 pre-defined section labels.

The remainder of this section is organized as follows: I first reintroduce psychiatric reports and their various types and conventions (§3.1.2) as done in chapter 2 for ease of reference. Next, I discuss the task definition in detail (§3.1.3). I then describe my approach including the corpus used, and the two main components of my model (§3.1.4). Additionally, I present and discuss the baselines and experiments performed as well as the results obtained from those experiments (§3.1.4). I finally conclude this section with a review of related work on document section identification and text segmentation (§3.1.6).

3.1.2 Psychiatric Evaluation Reports

As discussed in chapter 2, psychiatric reports mainly consist of long-form unstructured text. They are the end product of psychiatric assessments in which psychiatrists summarize the information they gathered, as well as integrate the patient history, their evaluation, patient diagnosis, and suggested treatments or future steps [Groth-Marnat, 2009, Goldfinger and Pomerantz, 2013]. There are several types of psychiatric reports that vary depending on the type and purpose of assessment: Psychiatric evaluation reports, crisis evaluation reports, daily SOAP reports (Subjective, Objective, Assessment, Plan), mental status exam reports, and mini mental status exam reports, to name a few [Association, 2006].

Although there is no one strict format for these reports, there are general guidelines that psychiatrists follow when writing psychiatric evaluation reports. Drawing from the general psychiatric evaluation domains, these reports start with the patient's identifying information, followed by the patient's chief complaints, presenting illness and its history, personal and family's medical history, mental status examination, and ending with the psychiatric medical diagnosis and treatment plan. This information is typically structured into an ordered list of headed sections [Association, 2006]. As listed previously in chapter 2, Table 3.1.2 contains a detailed list of the main sections of a psychiatric evaluation report in general order of appearance. I repeat the information from Table 2.1 for ease of reference. Not all listed sections appear in all psychiatric evaluation reports, and they also do not necessarily appear in the same order, although there is usually a general pattern to the order.

#	Section	# Words	# Sent.	Sent. Length	% Present	% Implicit
	GENERAL PATIENT INFO					
1	IDENTIFYING DATA	12	2	6	100	0
2	CHIEF COMPLAINT	27	3	9	100	0
	MEDICAL HISTORY					
3	HIST. OF PRSNT. ILLNSS.	232	29	8	95	10
4	PSYCHIATRIC HISTORY	85	8	11	82	36
5	SUBSTANCE ABUSE HIST.	98	10	10	88	44
6	REVIEW OF SYMPTOMS	150	19	8	96	51
7	SURGERIES	28	3	7	33	0
8	ALLERGIES	4	2	2	98	0
9	CURRENT MEDICATIONS	40	9	4	100	0
	FAMILY HISTORY					
10	BIRTH AND DEVELOP. HIST.	59	5	10	31	51
11	ABUSE HIST./TRAUMA	110	9	12	79	34
12	FAMILY PSYCHIATRIC HIST.	44	5	9	73	80
13	FAMILY MEDICAL HISTORY	48	7	7	92	38
14	SOCIAL HISTORY	80	7	11	76	45
15	PREGNANCY	29	3	8	47	64
16	SPIRITUAL BELIEFS	12	2	5	24	0
17	EDUCATION	32	3	8	68	0
18	EMPLOYMENT	31	3	9	79	0
19	LEGAL	10	1	5	20	0
	METAL STATUS					
20	MENTAL STATUS EXAM	155	18	9	95	11
21	STRENGTHS AND SUPPORTS	8	1	8	71	43
	TREATMENT					
22	FORMULATION	35	4	8	62	0
23	DIAGNOSES	63	12	5	100	0
24	PROGNOSIS	8	2	3	74	0
25	Treatment Plan	121	12	10	100	0
	Total				75	

Table 3.1: Section ontology for psychiatric evaluation reports and corpus statistics.

3.1.3 Task Definition

My goal was to build models that learn the section structure of an evaluation psychiatric report. As discussed earlier, a psychiatric evaluation report consists of several sections, often ordered in a usual way. Therefore the task I tackle here is to segment and classify blocks of unstructured text (at the sentence level) drawn from psychiatric evaluation reports into their appropriate section types. I assume that the reports follow the general

***Family History**: Her mother was depressed and was treated. Her mother is currently age 55 ... There is no family history of bipolar disorder, anxiety ... Medical history in the family is significant for her son, age 4, who is having seizures ... and several paternal great aunts had breast cancer.*

Figure 3.1: Excerpt from a psychiatric report showing an example of implicitly including two different sections within another (namely, *FAMILY PSYCHIATRIC HISTORY* in the first underlined portion, and *FAMILY MEDICAL HISTORY* in the second underlined portion within *FAMILY HISTORY*).

guidelines of psychiatric evaluation report writing discussed in (§3.1.2). There are four main challenges in section classification of clinical notes and mental health reports. First, labels that psychiatrists use to designate sections are ambiguous and various [Li et al., 2010], for example, a section titled *IDENTIFICATION OF PATIENT* by one psychiatrist might be named *REFERRAL DATA* or *IDENTIFYING INFORMATION* by another. Second, psychiatrists often omit some sections entirely or include them implicitly within other sections or under other labels, for example, the section *CHILDHOOD EVENTS* can be included in a larger section such as *FAMILY HISTORY* while *STRENGTHS AND SUPPORTS* can be listed within *Mental Status*. Figures 2.1 and 3.1 show snippets of psychiatric reports that demonstrate implicit sections. Third, the sections' order can be different between different psychiatric reports. Fourth, some section labels are omitted or skipped, especially if the information that would be placed in that section is not relevant to the patient being evaluated.

Additionally, with the section labels removed from the reports, the segmentation task was to find the section boundaries using sentences as the processing unit. This task is similar to topic shift detection in meeting minute, newscasts, and doctor-patient counseling conversations (both, written and spoken). Psychiatric reports are highly structured, with specific types of information (e.g., prescribed medications) found in particular sections (e.g.,*TREATMEN PLAN*), and with various general conventions for what informa-

tion should appear in which sections, and in what order. However, the segmentation task is not trivial as it faces the same aforementioned challenges. Additionally, one must find highly distinctive features to distinguish individual sentences (and thus, boundaries) in various sections as some of these sections can contain similar linguistic and structural features and may even contain similar topic keywords (e.g. language in *FAMILY PSYCHIATRIC HISTORY* and *SOCIAL HISTORY*.

I identify the subtasks of this problem as (1) learning and building a model for the sections' order and presence in a report, (2) learning and building models that describe the distinctive features of the various section types, and (3) applying a combination of these two model to simultaneously identifying section boundaries and label section types.

3.1.4 Approach

Given the sequential nature of the reports' sections, I treat this ordering task as a sequence labeling task. That is, given a psychiatric report with n sections $S = (S_1, \ldots, S_n)$, determine the optimal sequence of section labels $O^* = (O_1^*, \ldots, O_n^*)$ among all possible section sequences. Hidden Markov Models (HMMs) have been used successfully for sequence labeling in a wide variety of applications, including specifically natural language processing and medical informatics. In my problem formulation and approach, I follow and combine work presented by Sherman and Liu [2008] and Li et al. [2010]. Both of these approaches used HMM-based models coupled with section or topic-specific n-gram models to segment text. Sherman and Liu [2008] focused on segmenting sentences within meeting minutes into a set of predefined topics, while Li et al. [2010] focused on identifying sections within clinical note documents. I take a supervised learning approach where I learn the HMM parameters using a labeled corpus. My implementation was generally guided by the work described in Barzilay and Lee [2004] and [Rabiner, 1989].

To overcome the challenges outlined in (§3.1.3), I first created a unified hierarchy of standardize section labels types, based on observations in a 150 report corpus that I assembled. Second, while Li et al. [2010] focused on the section level when building their n-gram language models, I focus on the sentence level, similar to Sherman and Liu Sherman and Liu [2008]. Additionally, to model the inclusion of some sections within others as discussed in (§3.1.3) I built a two-level Hierarchical HMM (HHMM) [Bui et al., 2004] in which some states contain HMM models for their implicit subsections. This is in contrast to the approach presented by Li et al. [2010], who used a flat HMM, disregarding any hierarchy within the clinical notes' sections. The HHMM model was first proposed by Fine et al. [1998] as a strict tree structure where each state in the HHMM is an HHMM itself. This approach was extended and tailored by researchers for various tasks such as the approach proposed by Bui et al. [2004] who relaxed the original model to fit general HMM structures and implementations.

In summary, to tackle the first subtask from (§3.1.3) I built a two-level HHMM that models the positions and order of the reports' sections. To tackle the second subtask, I built language models (namely, n-gram models) per section type that describe distinctive lexical information for each of those sections. I then couple the HHMM with the n-gram models where the HHMM and HMM states represent the known section labels, while the states' observations are the n-grams contained within each of the individual sections. Finally, to tackle the the third subtask, that is identifying section boundaries, I follow a decoding scheme using the Viterbi algorithm (discussed briefly in §3.1.4).

In the remainder of this section I describe the corpus preperation. Next, I present the two components of the HHMM model, that is, the states (modeling the section order) and the observations (modeling the section language). Finally I briefly discuss the process by which I use the model to identify section boundaries.

Corpus

As outlined in chapter 2, I prepared the corpus in two stages. First, I standardized the labels' names, selecting a single uniform name for each section type and mapping corresponding section labels found in the corpus to those names. For example, some reports contained the section *SCHOOL* while others listed it as *EDUCATION*. Here I selected *EDUCATION* as the uniform section label across all reports.

Second, I created a hierarchy for the section names which reflected implicit embedded sections types that I found in the corpus. There were only three section types that included implicit subsections in the corpus, namely, *MEDICAL HISTORY*, *FAMILY HISTORY*, and *MENTAL STATUS*. For example, some reports containing the section *MENTAL STATUS* might in turn include information in that section about both *MENTAL STATUS EXAM* and *STRENGTHS AND SUPPORTS*. In this case I identified these implicit subsection boundaries (that is, the boundaries were not identified with a section header) and labeled those subsections with both the parent and child label. Table 3.1.2 lists the the parent sections that sometimes included other sections implicitly (emphasized in bold), the unified list of section types found in the collected reports (numbered sections), word and sentence level statistics, and percentage of reports containing those sections in the corpus. For both of these stages I used all 150 reports.

Following standard procedure for supervised machine learning, I split the corpus under a cross-validation paradigm into two sets for training and testing, where 80% of the reports were used in training and 20% for testing. This amounted to 120 and 30 reports for training and testing respectively.

Modeling the Section Orders

As discussed before, I built an HHMM where each state corresponds to a distinct section label. I introduce the terms *state* and *parent state* when discussing the HHMM. A *state*

is simply an HMM state corresponding to a distinct section. A *parent state* is an HHMM state corresponding to a collection of ordered sections. To account for sections listed implicitly, I created a two-level HHMM where *parent states* contained *states* representing the ordered subsections found in the *parent state* section. Thus the model contained 25 *states* and three *parent states* corresponding to information in Table 3.1.2. The first HHMM layer contained both *states* and *parent states*, while the second layer contained a total of 12 *states* corresponding to the potential implicit subsections for the three *parent states*. In this HHMM, each *parent state* is simply an HMM itself. Thus my discussion of HMM parameter calculation applies to both *states* and *parent states*.

The model learned transition probabilities from the labeled corpus. The state transition probabilities capture constraints on section orderings. I estimated the probabilities between each state s using Equation 3.1. Additionally, to account for sparsity (that is, unseen section orders) I smoothed the probabilities by the total number of section labels t_S following Laplace smoothing.

$$P(s_j|s_i) = \frac{count(s_i, s_j) + 1}{count(s_i) + t_S} \tag{3.1}$$

The second level HMM models contained within the *parent states* follow the same scheme in probability estimation, but differ in the smoothing parameter (t_S). Here, the total number of section labels t_S depends on the number of subsections in each of the *parent states*. For example, the *parent state MEDICAL HISTORY* contains a total of four subsections or *states*, and thus its HMM model is smoothed by $t_S = 4$. Finally, all of the model's states are linked with empty transitions in addition to self-looping ones to account for missing sections as well as a section continuation, respectively (i.e. indicating a section shift or a continuation).

Modeling Section Language

To tackle the second subtask identified in (§3.1.3), I built n-gram language models [Jain et al., 2015] that captured distinctive lexical information contained within the individual sections. This, in turn, helped classify unknown blocks of text (that is, text unseen previously by the trained models) within a report into their respective sections. I opted to use bigrams as opposed to higher n-gram models as the training corpus because was extremely sparse, and higher n-gram models had poor performance. This is consistent with significant research showing that in most applications bigrams work well and better than others [Reynar, 1998].

I built independent bigram models for each section type in the reports, using only text from that section type. Additionally, for each of the three section types represented by the *parent states* (discussed above), I built bigram models using text found in all of the contained subsections. A common problem that arises with n-gram models is sparsity of phrases or words. This is especially the case when training on a small corpus. Given the relatively small corpus, my models were quite sparse at first, however, I used Laplace Smoothing as a solution.

Similar to transition probabilities, the HHMM learned observation probabilities from the labeled corpus. I trained a bigram model for each state s of the HHMM. Equation 3.2 shows the computation for the likelihood of a sentence sequence w_0^k (i.e., a long sequence of words) to be generated by a state s. Equation 3.3 shows the computation for estimating the specific state bigram probability along with Laplace smoothing counts for the corresponding section S (V_S represents the vocabulary size for that section state).

$$P(w_0^k|s) = \prod_{0}^{k-1} P_s(w_{i+1}|w_i) \tag{3.2}$$

$$P_s(w_{i+1}|w_i) = \frac{count_S(w_i^{i+1}) + 1}{count_S(w_i) + |V_S|} \tag{3.3}$$

I used a rule-based approach to detect uniformly structured sections containing only standard medical terms such as medications and additional key terms. The sections mapped with hard-coded rules are the *CURRENT MEDICATIONS* and the standard *DSM-IV* multiaxal assessment contained within the *DIAGNOSIS* section, one of which is illustrated in Figure 3.2 below. I recognize that this standard has been dropped with the introduction of *DSM-5* in 2013, however, the dataset I used follows the older standard as most psychiatric reports in existence do since the new standard is relatively new.

Axis I	296.32	Major depressive disorder, recurrent, moderate
	305.00	Alcohol use disorder, mild
Axis II	V71.09	No diagnosis
Axis III		Hypertension
Axis IV		Problems with primary support group
Axis V	GAF = 48 (Current)	

Figure 3.2: Example of *DSM-IV* multiaxal diagnosis assessment.

For the *MEDICATIONS* section, I used publicly available datasets containing lists of medications [eMedicineHealth, 2018], and the U.S. National Library of Medicine's RxNorm dataset [Liu et al., 2005]. String-matching was additionally used to locate the *DIAGNOSIS* sections as my algorithm would search for the key headers "Axis I, II, III, IV, V".

Therefore I generated 26 bigram models, one for each section type (except for the two rule-based types), plus three parent section types.

Decoding

I integrated the bigram models with the HHMM and then used this bigram-HHMM model in a decoding framework to infer the most likely section boundaries and section types for documents with their section labels removed. I used the Viterbi algorithm and applied

the following equation to obtain the most likely labeling of sections O^*, where n is the section index, and k_n is the word index for section n:

$$O^* = \arg\max_{s} P(s)P(w_0^{k_n}|s) = \arg\max_{s_1 s_2 \ldots s_n} P(s_1)P(w_0^{k_n}|s_1) \times \prod_{i=0}^{n} P(s_i|s_{i-1})P(w_1^{k_n}|s_i) \tag{3.4}$$

3.1.5 Results and Discussion

As discussed above, I randomly split the corpus into training and testing sets in a cross-validation setup, using five folds, resulting in 120 reports for training and 30 for testing in each fold. The models were trained to learn a total of 25 distinct sections. Here I present the evaluation methods and results, describing our baseline approaches, as well as the performance of both the baselines and our method averaged across the test sets.

Evaluation Methods

There are two problems that this system solves: 1) the section labeling problem —applying the correct section type to each section—and 2) the section segmentation problem—identifying the correct section boundaries. I evaluated the system's performance on these two problems separately.

For the section ordering, I evaluated the performance of the model on each section using the F_1 measure averaged across all folds. As for the boundary detection problem, we use the WindowDiff (W_d) (Equation 3.5) [Pevzner and Hearst, 2002] and P_k (Equation 3.6) [Beeferman et al., 1999] metrics. These metrics compare the number of segmentation boundaries between a system's output and a gold standard by observing a scrolling window of text in the document, and run from 0 to 1, with scores closer to 0 being better. W_d increases (gets worse) when the boundaries are different. Similarly P_k increases when

a section type transition (i.e., a section type for this study) is different. The W_d score represents the probability that the number of boundaries found by the system is different from that in the gold standard, while the P_k score represents the probability that any two sentences are incorrectly listed as being in the same section.

$$W_d(ref, hyp) = \frac{1}{N-k} \sum_{i=1}^{N-k} (|ref - hyp| \neq 0) \tag{3.5}$$

$$P_k(ref, hyp) = \sum_{1 \leq i \leq j \leq n} D(i,j)(\delta_{ref}(i,j) \ \ XNOR \ \ \delta_{hyp}(i,j)) \tag{3.6}$$

Baseline Methods

I compared the system's performance in finding the correct labels of sections in a report to two baseline methods. The first method was introduced as a baseline by Li et al. [2010]. This method uses bigrams to independently classify each section, disregarding any section order information. For the second baseline, I followed the primary approach proposed by Li et al. [2010] which is a flat HMM model built similarly to my model as described previously (§3.1.4), but operates on a section level rather than a sentence level. Li's method ignores hierarchical information where some report sections are implicitly included within other sections. My implementation of this model included 25 states corresponding to each section within the reports. Both of these methods assume that the section boundaries are given, and as such they only generate a sequence labeling for section types.

I compared the system's performance in identifying section boundaries to two other baseline methods. The first is LCSeg—a popular text segmentation baseline [Galley et al., 2003]. LCSeg assumes that a topic change in written text occurs when chains of frequent repetitions of words begin and end. It rewards shorter chains over longer ones and further rewards chains with more repeated terms. Finally, the lexical cohesion between two

chains is evaluated using a cosine similarity. The second method is TopicTiling—an augmentation of the well-known TextTiling algorithm [Hearst, 1994]. TopicTiling [Riedl and Biemann, 2012] is LDA-based and represents segments as dense vectors of terms contained in dominant topics (as opposed to sparse term vectors).

Results

For the section labeling problem, the HHMM equaled or outperformed both baselines in all the sections. Table 3.3 shows the precision, recall, and F_1 scores for the two baselines and my model. The *DIAGNOSIS* section saw the best performance due to a rule-based approach. Similarly, *CURRENT MEDICATIONS* achieved high scores due to the use of dictionaries. All three models performed the worst in identifying the *LEGAL* section. I suspect that this is due to the low prevalence of this section and its content in the dataset. Similarly, sections with lower prevalence saw lower performance than others. Both baselines performed well in identifying the *IDENTIFYING DATA* and *DIAGNOSIS* sections due to their highly distinctive language. The HHMM model performed better for all implicit subsections, and significantly better for two (i.e., *PREGNANCY* and *BIRTH AND DEVELOPMENTAL HISTORY*. Finally, the HHMM model performed exactly the same as the Flat HMM baseline for the three parent types, as my model reduces to the Flat HMM in these cases and because the flat HMM model assumes a fixed general ordering of the sections.

Since the report sections vary in size, I computed both macro- and micro-averaged precision, recall, and F-measure (last two rows in Table 3.3). My model's micro-averaged F-measure is above 90% which is significantly higher than both the Flat-HMM and the independent bigram baselines performing at 85% and 62% respectively. Similar to Li et al. [2010], both the HHMM and the Flat-HMM baseline seemed to neither overfit nor

#	Section	Independent Bigram			Flat HMH			HHMM		
		P	R	F_1	P	R	F_1	P	R	F_1
	GENERAL PATIENT INFO	-	-	-	-	-	-	-	-	
1	IDENTIFYING DATA	0.83	0.81	0.82	0.96	0.94	0.95	0.95	0.95	**0.97**
2	CHIEF COMPLAINT	0.68	0.65	0.67	0.88	0.74	0.80	0.89	0.89	**0.91**
	MEDICAL HISTORY	0.66	0.66	0.65	0.93	0.88	0.90	0.88	0.88	**0.90**
3	HIST. OF PRSNT. ILLNSS.	0.69	0.67	0.68	0.91	0.86	0.88	0.94	0.86	**0.90**
4	PSYCHIATRIC HISTORY	0.65	0.6	0.62	0.74	0.85	0.79	0.93	0.86	**0.89**
5	SUBST. ABUSE HIST.	0.69	0.69	0.69	0.88	0.80	0.84	0.95	0.83	**0.89**
6	REVIEW OF SYMPTOMS	0.80	0.67	0.73	0.79	0.86	0.82	0.94	0.87	**0.90**
7	SURGERIES	0.40	0.31	0.35	0.79	0.51	0.62	0.85	0.64	**0.73**
8	ALLERGIES	0.60	0.80	0.69	0.90	0.86	0.88	0.88	0.91	**0.89**
9	CURRENT MEDICATIONS	0.87	0.74	0.80	0.90	0.84	0.87	0.91	0.93	**0.92**
	FAMILY HISTORY	0.68	0.56	0.58	0.92	0.86	0.89	0.92	0.86	**0.89**
10	BIRTH AND DVLP. HIST.	0.68	0.50	0.57	0.71	0.68	0.69	0.89	0.80	**0.84**
11	ABUSE HIST. / TRAUMA	0.42	0.33	0.37	0.87	0.77	0.82	0.96	0.81	**0.88**
12	FAMILY PSYCH. HIST.	0.57	0.59	0.58	0.92	0.87	0.89	0.92	0.90	**0.91**
13	FAMILY MED. HISTORY	0.65	0.60	0.62	0.92	0.89	0.90	0.94	0.89	**0.91**
14	SOCIAL HISTORY	0.67	0.69	0.68	0.66	0.89	0.76	0.93	0.81	**0.87**
15	PREGNANCY	0.60	0.67	0.63	0.89	0.51	0.65	0.92	0.80	**0.86**
16	SPIRITUAL BELIEFS	0.73	0.46	0.56	0.90	0.90	0.90	0.93	0.88	**0.90**
17	EDUCATION	0.66	0.61	0.63	0.71	0.77	0.74	0.92	0.84	**0.88**
18	EMPLOYMENT	0.65	0.62	0.63	0.91	0.88	0.89	0.92	0.86	**0.89**
19	LEGAL	0.16	0.62	0.26	0.67	0.61	0.64	0.72	0.68	**0.70**
	MENTAL STATUS	0.56	0.72	0.62	0.85	0.94	0.89	0.85	0.94	**0.89**
20	MENTAL STATUS EXAM	0.64	0.63	0.64	0.83	0.96	0.89	0.85	0.96	**0.90**
21	STRENGTHS AND SUPPORTS	0.42	0.82	0.56	0.80	0.92	0.86	0.82	0.92	**0.87**
	TREATMENT	-	-	-	-	-	-	-	-	
22	FORMULATION	0.56	0.71	0.63	0.86	0.78	0.82	0.92	0.82	**0.87**
23	DIAGNOSES	0.88	0.76	0.81	0.96	0.95	0.96	0.98	0.98	**0.98**
24	PROGNOSIS	0.66	0.62	0.64	0.84	0.82	0.83	0.90	0.86	**0.88**
25	TREATMENT PLAN	0.74	0.83	0.78	0.95	0.93	0.94	0.97	0.93	**0.95**
	Macro-Average	0.62	0.64	0.62	0.85	0.82	0.83	0.91	0.86	**0.88**
	Micro-Average	0.62	0.62	0.62	0.86	0.83	0.84	0.93	0.91	**0.92**

Table 3.2: Section type identification results (precision, recall and F_1 scores) per section as well as micro and macro averages. Parent sections are in bold.

underfit, which is indicated by higher micro-averaged compared to the macro-averaged scores.

As for the boundary detection problem, and similar to the evaluation in Sherman and Liu [2008], I performed two experiments for the baselines since both baselines require a parameter representing the number of boundaries (number of topics minus one). In

the first experiment I allowed the parameter to be chosen by LCSeg and TopicTiling, respectively, while in the second experiment, I provide the algorithms with the correct number of boundaries (i.e., number of sections minus one). The HHMM however, needs no prior information regarding the number of sections present in a given report. Table 3.1.5 shows the W_d and P_k scores for all three approaches. My system again outperformed both baselines indicated by lower W_d and P_k error rates overall. Both baselines performed better when the number of boundaries is known—an expected result. In fact, TopicTiling outperformed my approach by a small margin when provided with the correct parameter value. I note, however, that when running open loop on new text, the number of sections will be unknown, so this result does not reflect how I envision the approach being used.

# Boundaries	Algorithm	P_k	W_d
System Choice	LCSeg	0.29	0.37
	Topic Tiling	0.27	0.33
Provided	LCSeg	0.25	0.33
	Topic Tiling	0.20	0.25
	HHMM	**0.20**	**0.26**

Table 3.3: Section boundary identification results.

3.1.6 Related Work

As discussed above, my work simultaneously solves two problems within a psychiatric evaluation report: identifying section types and identifying section boundaries. The first problem has been referred to as argumentative zoning [Teufel et al., 1999, Li et al., 2010, Denny et al., 2009a], while the second is a type of text segmentation problem [Hearst, 1994, Riedl and Biemann, 2012]. Argumentative zoning refers to classifying text sections into mutually exclusive categories. Work on this task is mostly centered around identi-

fying scientific article sections (e.g., abstract, introduction, methodology, etc.) [Teufel, 1999].

My work is a combination and extension of Li et al. [2010]'s work on identifying section types within clinical notes and Sherman and Liu [2008]'s work on text segmentation of meeting minutes. Both approaches integrated n-gram language models into HMMs. The former modeled HMM emissions at the section level using bigrams, while the later modeled the emissions at the sentence level and used unigrams and trigrams. Other approaches followed similar strategies in segmenting story text and in creating generative models for detecting story boundaries [Mulbregt et al., 1998, Yamron et al., 1998]. More recently, Yu et al. [2016] used a hybrid deep neural network combined with a Hidden Markov Model (DNN-HMM) to segment speech transcripts from broadcast news to a sequence of stories.

More broadly, there has been some work on applying NLP in the mental health domain. However, due to lack of readily available clinical data (e.g. clinical reports), researchers have focused on non-clinical sources (e.g., social media) [Chapman et al., 2011]. Several algorithms were developed to detect specific emotions from suicide notes and online journals [Pestian et al., 2012, Strapparava and Mihalcea, 2008], while twitter data was used to detect distress and suicide ideation [Homan et al., 2014, O'Dea et al., 2015]. Additionally, twitter data was used to measure mood valence and detect depression [Sadilek et al., 2013, De Choudhury et al., 2013, Coppersmith et al., 2015]. Facebook data was used to measure emotion contagion and to predict post-partum depression [Coviello et al., 2014, De Choudhury et al., 2014]. Instead of social media and publicly available, non-clinical data Althoff et al. [2016] used counseling conversations gathered using a messaging service and developed discourse analysis methods to measure the correlation of outcomes with various linguistic aspects.

3.2 Using Conditional Random Fields to Automatically Identify Sections in Clinical Reports

The automatic identification of sections in clinical free text is an important step in automatic understanding of electronic health records, and is useful for information extraction, data mining, and Semantic analysis. In this section I describe an improved supervised approach to identifying sections within semi-structured clinical reports that uses conditional random fields (CRFs), thus extending the HHMM approach described in the previous section. I developed and tested the CRF approach on three different clinical report types: psychiatric evaluations, hospital discharge summaries, and radiology reports.

3.2.1 Background

With the exponential growth of electronic health records (EHRs), it is ever more important to develop natural language processing (NLP) systems that can automatically understand and parse the free text contained within those reports. When combined with other systems, these NLP models can aid in a variety of useful medical tasks, such as clinical decision support [Hripcsak et al., 2003], trend analysis, and precision medicine [Demner-Fushman et al., 2009].

The majority of clinical NLP work has focused on semantic parsing and information retrieval of clinical notes found in EHRs [Reyes-Ortiz et al., 2015, Névéol et al., 2018]. In contrast, the work I present here concerns learning and using a model of *section structure*, that is, the types of sections used in for a given document type, their common ordering, and the language typically found in that section. Automatically learning and applying such a model to detect sections can be beneficial to overcome a number of challenges in automatically understanding unstructured clinical text, encompassing many levels of

language processing: identifying document layouts, determining their discourse organization, mapping lexical information to semantic concepts found in biomedical ontologies, resolving co-references, and extracting temporal relationships [Wang et al., 2018, Roberts et al., 2016, Filannino and Uzuner, 2018, Li et al., 2010]. For example, understanding the section structure can greatly aid semantic concept extraction when certain concepts can be mapped to specific sections. This, in turn, can aid further tasks in information extraction and semantic search.

I demonstrate my improved approach on three different types of clinical reports: psychiatric evaluations, radiology reports, and discharge summaries (Corpora 2.1-2.3). The models and results presented in this section are extensions of my work for psychiatric evaluations presented in a prior workshop paper [Banisakher et al., 2018a]. I show that the extended approach provides significantly better performance for the task across all three report types and can further label implicit sections (that is, sections included within others with no explicit section headings indicating so). Figure 2.1 shows a snippet example of a typical psychiatric evaluation report, while other reports also follow a similar structure. I describe the corpus in detail in the following section.

I define the section identification task as follows, which is identical to the definition found in the previous section (§3.1.3). First, I assume I am given a corpus of reports, all of the same type, as well as an ontology of section types found in that report type. The corpus is assumed to be labeled with the ontology, in that each sentence in every document is labeled with its section type. The corpus is then split into training and testing portions. The training portion of the corpus is then use to train a model. The test portion of the corpus is stripped of all section headings. The model is then expected to be able to simultaneously identify the (unlabeled) section boundaries and label their types. Thus, the task comprises two subtasks: (1) determining the section ordering for each specific report, including identifying when sections are missing, and (2) locating section boundaries.

As discussed in the previous section (§3.1), this task has four main challenges. First, there is great ambiguity and variety in the section headings present in the data [Li et al., 2010]. Using psychiatric reports as an example, a section labeled *IDENTIFICATION OF PATIENT* by one psychiatrist might be labeled *REFERRAL DATA* or *IDENTIFYING INFORMATION* by another. Second, some sections are included inside others; for example, the section *MEDICAL HISTORY* might include *REVIEW OF SYMPTOMS* and *PSYCHIATRIC HISTORY* subsections, while the section *FAMILY HISTORY* might include a subsection addressing *PREGNANCY*. Like sections, these subsections can either be explicitly labeled (heading present) or just implicit (heading omitted). Figures 2.1 and 3.1 show snippets of psychiatric reports that demonstrate implicit sections. Third, the section ordering can differ between reports, again, depending on the psychiatrist. And fourth, sections may be omitted, especially when that information in not relevant to the patient in question. For example a report regarding a male patient would likely not contain a *PREGNANCY* section. These challenges apply equally to many other types of clinical reports, including the discharge summaries and radiology reports used here.

In my prior work I developed a model to solve this task which combined Hierarchical Hidden Markov Models (HHMMs) and n-grams. The states in the HHMM represented sections types listed in the provided ontology, with the transition probabilities learned from the labeled corpus. Each state was associated with an n-gram model which was trained on the language found in corresponding sections of that type. I then applied this combined model to a report using Viterbi decoding to simultaneously locate section boundaries and label the section types.

In this work, I developed a more robust model based on conditional random fields (CRFs) which takes a discriminative rather than generative approach. While generative approaches are general and more interpretable, discriminative approaches generally have better performance. I also demonstrate that this technique generalizes beyond psychiatric

evaluations to at least two other clinical report types, namely, discharge summaries and radiology reports.

The section proceeds as follows. First I describe the datasets that I use for training and testing (§3.2.2). Then I detail the methods, including the CRF model and how it captures both section ordering and section content, how the model is trained, and how it is used to locate section boundaries and determine section labels (§3.2.3). I next compare the performance of the CRF model with various baselines, demonstrating that it performs better than prior models (§3.2.4). I then discuss limitations and future directions (§3.2.5). Finally, I conclude this section with a discussion of related work (§3.2.6).

Corpus	Report Type	Report Count	Section Count	Avg. Sections per Report	Avg. Words per Report
1	Psychiatric Evaluation	150	2824	18.8	1521
2	Radiology Report	150	900	6	463
3	Discharge Summary	150	2977	19.8	1829

Table 3.4: Summary of corpora statistics.

3.2.2 Data

I used three corpora of clinical reports to test the improved section identification model. Each corpus was paired with an ontology of section types specific to that report type. These ontologies were drawn from prior work and are described in more detail in chapter 2. As described previously, the data was doubly annotated. For this, various section names were tagged with an unified heading from the respective ontology first. Second, each sentence was tagged with a section heading. I calculated the inter-rater reliability using Cohen's κ statistic, achieving 0.90, 0.88, and 0.84 for each corpus, respectively. These agreement values are considered "perfect" agreement [Artstein and Poesio, 2008]. Table 3.2.1 shows a summary of the three corpora statistics I used in this study. I list the

corpora statistics and ontologies from chapter 2 again in this chapter for ease of reference (Tables 3.1.2, 3.2.2, 3.2.2).

Section	# Words	# Sent.	Sent. Length	% Present	% Implicit
CLINICAL INFORMATION					
CLINICAL HISTORY	80	8	10	100	0
EXAM DETAILS					
EXAM	16	2	8	100	0
COMPARISON	16	2	8	86	10
CONTRAST	14	2	7	14	53
PROCEDURE	12	2	6	100	60
FINDINGS					
FINDINGS	192	24	8	100	0
IMPRESSION					
IMPRESSION	133	19	7	100	0
ATTENDING STATEMENT	-	-	-	0	-

Table 3.5: Section ontology for radiology reports and corpus statistics.

3.2.3 Methods

I treated section identification as a sequence modeling task. Formally, the task is as follows: given a clinical report with n sections and m sentences, where the sections are unlabeled and n is not known, identify the optimal sequence (order) of section labels $H^* = (L_1^*, \ldots, L_n^*)$ from among all possible section sequences, and assign every sentence a section label $H^* = (H_1, \ldots, H_m)$ consistent with L^*. Sequence labeling problems in NLP and medical informatics have been solved by both generative and discriminative approaches, including Hidden Markov Models (HMMs; generative) and Conditional Random Fields (CRFs; discriminative). Li *et al.* [Li et al., 2010] used HMM and n-gram models to detect the orders or labels of sections within clinical reports, while modeling the observation probabilities at the section level. Sherman and Liu [Sherman and Liu,

Section	# Words	# Sent.	Sent. Length	% Present	% Implicit
GENERAL PATIENT INFO					
ADMIT DATE	3	1	3	100	0
DISCHARGE DATE	3	1	3	100	0
SERVICE	4	2	2	100	0
PROVIDER INFO					
ATTENDING	2	1	2	82	0
ADMIT PHYSICIAN	2	1	2	100	0
DISCHARGE PHYSICIAN	2	1	2	100	0
CONDITION BEFORE ADMISSION					
ADMISSION DIAGNOSES	96	12	8	100	0
HISTORY	135	15	9	76	58
MEDICATIONS	55	11	5	100	0
REASON FOR ADMISSION	162	18	9	100	0
CONDITION AT DISCHARGE					
CONDITION	4	2	2	100	0
DISPOSITION	2	1	2	34	10
DISCHARGE DIAGNOSES	144	18	8	89	37
OTHER DIAGNOSES	-	-	-	0	-
PHYSICAL EXAM ON DISCH.	45	9	5	40	38
MEDICAL HISTORY					
ALLERGIES	12	3	4	100	0
FAMILY HISTORY	81	9	9	43	20
GYNECOLOGICAL HISTORY	-	-	-	0	-
PAST MEDICAL HISTORY	144	16	9	100	41
PAST SURGICAL HISTORY	32	4	8	100	58
SOCIAL HISTORY	84	7	12	37	66
HOSPITAL COURSE					
CONSULTATION	88	11	8	6	0
HOSPITAL COURSE	168	14	12	85	0
PHYSICAL	66	11	6	28	13
PROCEDURES	15	5	3	65	10
STUDIES	-	-	-	0	0
DISCHARGE INSTRUCTIONS					
FOLLOW UP	-	-	-	0	-
DIAGNOSTIC STUDIES REC'D	-	-	-	0	-
DISCHARGE INSTRUCTIONS	408	34	12	100	0
DISCHARGE MEDICATIONS	72	12	6	100	0

Table 3.6: Section ontology for discharge summary reports and corpus statistics.

2008] used HMMs as well as n-gram models to detect topic shifts in meeting minutes, and, in contrast to Li *et al.*, modeled the observation probabilities on the sentence level.

In my prior work [Banisakher et al., 2018a] (described in the §3.1) I combined the approaches of Li *et al.* [Li et al., 2010] and Sherman and Liu [Sherman and Liu, 2008] to learn the section structure and content and then used that model to both determine the most likely section sequence and locate section boundaries, that is, segment the sections. My approach combined a *Hierarhical* Hidden Markov Model (HHMM)—which used section statistics as the model's transition probabilities—with n-grams for the observation probabilities of words. In this work I substituted CRFs for the HHMM. As noted previously, generative models such as HMMs have more explanatory power when compared with their discriminative counterparts such as CRFs. However, HMMs, rely on the assumption that observations are statistically independent from one another. For this example, this means that the HMM assumes that the presence of certain sentences within section A is independent from other sentences within another section B. In practice, however, this is not the case: for example, the fact that the *CHIEF COMPLAINT* section in a psychiatric evaluation mentions depression means that the *DIAGNOSIS* and *TREATMENT* sections will also likely mention depression. Following this intuition, I hypothesize that section structure and language can be better modeled if the independence assumption is relaxed, which motivates the move to CRFs.

My new approach differs from my prior work in five ways: first, I used a discriminative CRF (namely, linear chain CRF) instead of a generative HMM to capture more features encoded within the sections' content and to model the dependencies between sections. Second, I trained the model to learn not only n-gram features, as it is the case with my previous approach, but also other lexical and positional features. This is possible in CRFs (as opposed to HMMs) because they do not have restrictions on variable dependence. Third, my CRF model is flat, as opposed to the former hierarchical HMM. This does not limit the CRF approach from detecting implicit sections. Fourth, our prior system used a rule-based approach instead of n-grams for three section types (*MEDICATIONS*, *ALLERGIES*, and *DIAGNOSIS*). In the this approach I eliminated the dependence

Approach	Features	Hierarchical?	#Model Layers	Rule-Based Features?	Corpus specific?
HHMM	n-grams only	Yes	2	Yes (for some sections)	Yes
CRF	n-grams, lexical, and positional	No	1	Completely Automatic	No

Table 3.7: Summary of differences between the previous (HHMM) and current (CRF) approaches.

on these hand-crafted rules, making the model fully automatically learned from the data. Fifth, because of the elimination of the rule-based approach to detecting certain sections, this new model is generalizable and is not corpus specific, and thus can be applied to other document types. Table 3.2.3 summarizes these differences.

Linear Chain Conditional Random Fields

Conditional Random Fields (CRFs) are undirected graphical models [Lafferty et al., 2001, Konkol and Konopík, 2013] that can be used for discriminative sequence labeling. CRFs have proved useful for many sequence labeling problems in NLP and computer vision [Lin and Wu, 2009], including Named Entity Recognition (NER) and image classification. There are several CRF variations such as the tree CRF and the hierarchical CRF which are mostly used for computer vision related tasks. Linear chain CRFs are the most popular among CRF approaches for sequence labeling tasks largely due to its relative simplicity and low computational cost when compared with other CRF models.

I built and trained a linear chain CRF analogous to the prior HHMM approach. In contrast to an HHMM, the CRF encodes sections as nodes in the CRF graphical representation (instead of HMM states), and uses weighted feature functions for transitions between nodes (instead of the HMM transition and emission probabilities). Additionally, the CRF model captures the "true" desired probability distribution, that is the *conditional distribution* of labels given the observations $P(Y|X)$, instead of modeling the joint distribution of observations and labels $P(X, Y)$. This a known advantage of CRFs in general over HMMs and is mainly due to, again, removing the independence assumption. Thus,

CRFs can have an arbitrary number of dependencies as opposed to the limited dependency structure of HMMs. The CRF model benefits from this as it does not only record the dependence of a section only on its predecessor and observations, but on additional dependencies given the entire sequence of labels (i.e., section types) and observations (i.e., sentences).

The CRF probability distribution is defined by Equation 3.7. Let $\overline{l}$ be the sequence of section labels, $\overline{s}$ be the sequence of sentences (i.e., the observations) in a given report, and L be the possible label sequences. The model follows a typical linear chain CRF where the conditional distribution is:

$$P(\overline{l}|\overline{s},\lambda) = \frac{\exp(\sum_i \sum_j \lambda_j F_j(l_{i-1}, l_i, \overline{s}, i))}{\sum_{l' \in L} \exp(\sum_i \sum_j \lambda_j F_j(l'_{i-1}, l'_i, \overline{s}, i))} \tag{3.7}$$

Where λ is a set of model parameters, and each λ_j is a parameter (or weight) associated with each feature function F_j. Each feature function represents a dependency within the model. I used the L-BFGS method to estimate each λ_j [Nocedal, 1980]. The model's probability distribution is thus generated by summing over the entire observation sequence, where each observation is indexed by the variable i and the entire feature function space index by the variable j. The denominator sums over all possible label sequences L.

The most critical component in the design of CRF models is the feature function space. In this model, each feature function is:

$$F_j(l_{i-1}, l_i, \overline{s}, i) = H_j(l_{i-1}, l_i, \overline{s}, i) \cdot SF_j(l_{i-1}, l_i, \overline{s}, i) \tag{3.8}$$

Where H_j models the section order, and SF_j models the section content. These are similar to an HMM's transition and emission probability distributions, respectively. In contrast to HMMs, however, the feature functions are evaluated over the entire observation sequence $\overline{s}$ taking into account the neighboring labels (or sections) l_i and l_{i-1}. This thus conditions the probability of a given section type on the content and order of the entire sequence. I

outline the intuition behind and implementation of the feature functions in the following sections.

Section Order Modeling

The feature function F_j incorporates section ordering through the section ordering function $H(l_{i-1}, l_i, \overline{s}, i)$. As discussed above, there is a feature function for each of the dependencies defined in the model. Analogous to the state transition probabilities in the prior HHMM approach [Banisakher et al., 2018a], I encode the interdependent order of sections (i.e., which sections depend upon each other) using a binary matrix. To achieve this, I first used the distinct section labels from the ontologies shown in Tables 3.1.2, 3.2.2, 3.2.2, and discussed in the Data section and chapter 2. Then I created a binary matrix V_{l_{i-1},l_i} whose entries represent whether a section follows another or not. For example if section *SOCIAL HISTORY* (indexed as section 4) was observed in the data directly before *PREGNANCY* (indexed as section 5), then the entry $V_{4,5}$ would contain a value of 1. The matrix contained N^2 entries, where N is the total number of sections for each report type as shown in the section label ontologies presented in Tables 3.1.2, 3.2.2, 3.2.2. This was performed for each corpus separately, and I only modeled sections that were present in the data. Thus the CRF models contained 25 nodes each for psychiatric evaluations and discharge summaries, and 7 for radiology reports. The section order feature function was formulated as follows:

$$H_j(l_{i-1}, l_i, \overline{s}, i) = V_{l_{i-1},l_i} \tag{3.9}$$

Note that for each section (or label) s_i, the model sums the total entries for the entire sequence of labels and observations as shown in Equation 3.7, thus conditioning each section on the entire sequence.

Section Content Modeling

Similarly, the feature function F_j incorporates section content via the section feature function $SF(l_{i-1}, l_i, \overline{s}, i)$. This function is analogous to the emission probabilities in my prior HHMM approach [Banisakher et al., 2018a]. These functions model the dependency between a section and its content. Importantly, the feature function should not be confused with the linguistic features extracted and input into the section feature function. To capture section content (i.e., to model section-specific language) I extracted three sets of lexical and positional features: (1) lexical features comprising n-grams (specifically, unigrams and bigrams), (2) sentence position and length features comprising local (relative to section) and global (relative to an entire report) sentence positions as well as the sentence length, and (3) the top three key terms per section type extracted using the TF-IDF method [Church and Gale, 1999]. I combined these features into a feature vector X which was then normalized and summed. I then used a threshold function T_k calculated for each section type k to assign a binary value for the section feature function $SF(l_{i-1}, l_i, \overline{s}, i)$ as follows.

$$SF_j(l_{i-1}, l_i, \overline{s}, i) = \begin{cases} 1 & T_k \leq \sum X_i \\ 0 & \text{otherwise} \end{cases} \tag{3.10}$$

The threshold function T_k is defined by Equation 3.11 below. The first part in the equation is the average sum of feature vectors for section type k, given n_k total sections of type k. The parameter α restricts or relaxes the threshold, and σ is the standard deviation of all the sums of feature vectors for each section type k. This function was used as a constraint in a grouping genetic algorithm for a clustering task [Falkenauer, 1992, Agustı et al., 2012]. Finally, I estimated α to be 0.60 using a similar technique to the Wu-Palmer score following [Warin and Volk, 2004].

$$T_k = \frac{\sum X_k}{n_k} + \alpha * \sigma(\sum X_k) \tag{3.11}$$

Inference

To apply the model, I applied the usual inference process for linear chain CRFs. For the CRF model this is equivalent to simultaneously locating section boundaries and labeling each section with a section type. Inference in linear chain CRFs follows a similar algorithm to Viterbi [Forney, 1973], which is used in decoding HMM models. While not stated explicitly in the Equation 3.7 above, the normalization factor $Z(S)$ is calculated as is often done using the Gaussian prior as it was introduced in [Chen and Rosenfeld, 1999].

3.2.4 Results and Discussion

In order to test the CRF models, I randomly split each corpus into training and testing sets in a cross-validation setup, using five folds, resulting in 120 reports for training and 30 for testing in each fold for each document type. The models were trained to learn a total of 25 distinct sections for psychiatric reports and discharge summaries, and 7 for radiology reports. In this section I describe the evaluation metrics, baseline comparisons, overall experiments and results.

Evaluation Metrics

As discussed above, the CRF system simultaneously solves two tasks: (1) determining the section order—applying the correct section label to each section—and (2) locating the section boundaries. I evaluated the system's performance on these two tasks separately.

For section ordering, I evaluated performance using the F_1 measure averaged across all folds. In this evaluation, the section boundaries were known to the system in order to independently evaluate the performance of the system for this subtask. The output of the models were compared to the annotated labels (i.e., ground truth) in the test sets. A

section was considered to be identified correctly if it was outputted at the same position as that in the ground truth.

Similar to the previous HHMM study, I evaluated performance for boundary detection using the WindowDiff (W_d) (Equation 3.5) [Pevzner and Hearst, 2002] and P_k (Equation 3.6) [Beeferman et al., 1999] metrics. A detailed description of these metrics including their respective calculations is listed in §3.1.5.

Baseline Methods

I compared the CRF model's performance in determining section ordering to two baselines: (1) my prior HHMM model, and (2) an n-gram-only model (specifically, bigrams) to independently classify each section, which disregards any section order information.

I compared the system's performance in locating section boundaries against four baselines. The first, was a HMM-LSA model implemented as descried by Ginter *et al.* [Ginter et al., 2009]. This model treated sections in ICU notes as topics and performs segmentation and labeling of those sections as topics. The method is unsupervised and is based on a combination of Hidden Markov Models and latent semantic indexing which allows the topics of interest to be defined freely, without the need for data annotation, and can identify short segments. The second baseline was LCSeg, a popular text segmentation baseline [Galley et al., 2003]. LCSeg assumes that a topic (section) change in written text occurs when chains of frequent repetitions of words begin and end. It rewards shorter chains over longer ones and further rewards chains with more repeated terms. The lexical cohesion between two chains is evaluated using a cosine similarity. The third baseline was TopicTiling, an augmentation of the well-known TextTiling algorithm [Hearst, 1994]. TopicTiling [Riedl and Biemann, 2012] is LDA-based and represents segments as dense vectors of terms contained in dominant topics (as opposed to sparse term vectors). Finally, I compared the CRF model to my previous HHMM approach as a fourth baseline.

Experiments and Results

For the section ordering task, my model equaled or outperformed both baselines in all sections across all three report types. Tables 3.2.4, 3.2.4, 3.2.4 show the precision, recall, and F_1 scores for the two baselines as well as the CRF model. I omitted sections that did not exist in the corpus from the results tables. Similar to my prior work [Banisakher et al., 2018a], the CRF model performed better for sections with higher prevalence in the corpus (e.g. *IDENTIFYING DATA* in psychiatric reports, *DISCHARGE INSTRUCTIONS* in discharge summaries, and *CLINICAL HISTORY* in radiology reports). Additionally, the CRF models showed significantly better performance for sections with highly distinctive content such as the *DIAGNOSIS* in psychiatric reports, *ADMIT DATE* in discharge summaries, and *MEDICATIONS* in radiology reports.

Both the HHMM and the CRF models performed better in leading and ending sections across all three report types. This is because those sections typically display minimal variability in position (e.g., *TREATMENT PLAN* in psychiatric reports, and *DISCHARGE MEDICATIONS* in discharge summaries). This, in turn, increased performance on the surrounding sections, as fewer errors were propagated through the models. The CRF models equaled or outperformed the HHMM models in identifying sections for which my prior approach used hand-crafted rules specifically *ALLERGIES*, *CURRENT MEDICAITONS*, and *DIAGNOSIS* sections in psychiatric reports.

Additionally, the HHMM and CRF models identified implicit sections successfully. The CRF model performed better for all implicit sections and significantly better for sections *BIRTH AND DEVELOPMENTAL HISTORY* in psychiatric reports, *PAST MEDICAL HISTORY* and *PAST SURGICAL HISTORY* in discharge summary reports, and *PROCEDURE* in radiology reports.

My models performed worst on the sections *SURGERIES* and *LEGAL* in psychiatric reports, *CONSULTATION* in discharge summaries, and *CONTRAST* in radiology reports.

Section	Independent Bigram			HHMM			CRF		
	P	R	F_1	P	R	F_1	P	R	F_1
IDENTIFYING DATA	0.83	0.81	0.82	0.98	0.95	0.97	0.99	0.99	**0.99**
CHIEF COMPLAINT	0.68	0.65	0.67	0.94	0.89	0.91	0.99	0.97	**0.98**
HIST. OF PRSNT. ILLNSS.	0.69	0.67	0.68	0.94	0.86	0.90	0.96	0.90	**0.93**
PSYCHIATRIC HISTORY	0.65	0.6	0.62	0.93	0.86	0.89	0.95	0.93	**0.94**
SUBST. ABUSE HIST.	0.69	0.69	0.69	0.95	0.83	0.89	0.96	0.92	**0.94**
REVIEW OF SYMPTOMS	0.80	0.67	0.73	0.94	0.87	0.90	0.94	0.94	**0.94**
SURGERIES	0.40	0.31	0.35	0.85	0.64	0.73	0.86	0.84	**0.85**
ALLERGIES	0.60	0.80	0.69	0.88	0.91	0.89	0.90	0.88	**0.89**
CURRENT MEDICATIONS	0.87	0.74	0.80	0.91	0.93	0.92	0.96	0.96	**0.96**
BIRTH AND DVLP. HIST.	0.68	0.50	0.57	0.89	0.80	0.84	0.96	0.94	**0.95**
ABUSE HIST. / TRAUMA	0.42	0.33	0.37	0.96	0.81	0.88	0.95	0.91	**0.93**
FAMILY PSYCH. HIST.	0.57	0.59	0.58	0.92	0.90	0.91	0.97	0.95	**0.96**
FAMILY MED. HISTORY	0.65	0.60	0.62	0.94	0.89	0.91	0.96	0.94	**0.95**
SOCIAL HISTORY	0.67	0.69	0.68	0.93	0.81	0.87	0.95	0.91	**0.93**
PREGNANCY	0.60	0.67	0.63	0.92	0.8	0.86	0.94	0.88	**0.91**
SPIRITUAL BELIEFS	0.73	0.46	0.56	0.93	0.88	0.90	0.94	0.92	**0.93**
EDUCATION	0.66	0.61	0.63	0.92	0.84	0.88	0.95	0.91	**0.93**
EMPLOYMENT	0.65	0.62	0.63	0.92	0.86	0.89	0.97	0.93	**0.95**
LEGAL	0.16	0.62	0.26	0.72	0.68	0.70	0.86	0.84	**0.85**
MENTAL STATUS	0.64	0.63	0.64	0.85	0.96	0.90	0.91	0.93	**0.92**
STRENGTHS AND SUPPORTS	0.42	0.82	0.56	0.82	0.92	0.87	0.93	0.91	**0.92**
FORMULATION	0.56	0.71	0.63	0.92	0.82	0.87	0.93	0.91	**0.92**
DIAGNOSES	0.88	0.76	0.81	0.98	0.98	0.98	0.99	0.99	**0.99**
PROGNOSIS	0.66	0.62	0.64	0.90	0.86	0.88	0.97	0.93	**0.95**
TREATMENT PLAN	0.74	0.83	0.78	0.97	0.93	0.95	0.98	0.96	**0.97**
Macro-Average	0.62	0.64	0.62	0.91	0.86	0.88	0.95	0.92	**0.94**
Micro-Average	0.62	0.62	0.62	0.93	0.91	0.92	0.98	0.96	**0.97**

Table 3.8: Section identification results for psychiatric evaluation reports.

I suspect that this is due to low prevalence of these sections and their content in the corpora. However, while the HHMM approach struggled in identifying sections with low prevalence, the CRF model was able to model those sections well with higher than 0.83 F_1 score (with the exception of the *CONTRAST* which only existed in 14$ of the radiology corpus). I hypothesize that this is due to the independence assumption in the HHMM model. Each state in the HHMM is dependent only on the previous state as well as its observations (i.e., the section content), while the CRF models the entire observation sequence at every stage. Thus, the CRF is less sensitive to section prevalence. Moreover,

Section	Independent Bigram			HHMM			CRF		
	P	R	F_1	P	R	F_1	P	R	F_1
ADMIT DATE	0.88	0.88	0.88	0.97	0.98	0.99	0.99	0.99	**0.99**
DISCHARGE DATE	0.88	0.86	0.87	0.97	0.98	0.99	0.99	0.99	**0.99**
SERVICE	0.68	0.84	0.75	0.95	0.95	0.95	0.99	0.97	**0.98**
ATTENDING	0.39	0.30	0.34	0.85	0.88	0.91	0.94	0.94	**0.94**
ADMIT PHYSICIAN	0.39	0.35	0.37	0.79	0.86	0.94	0.98	0.92	**0.95**
DISCHARGE PHYSICIAN	0.37	0.35	0.36	0.80	0.86	0.93	0.97	0.93	**0.95**
ADMISSION DIAGNOSES	0.72	0.70	0.71	0.90	0.92	0.94	0.99	0.93	**0.96**
HISTORY	0.71	0.73	0.72	0.89	0.89	0.89	0.94	0.90	**0.92**
MEDICATIONS	0.59	0.65	0.62	0.85	0.87	0.86	0.97	0.91	**0.94**
REASON FOR ADMISSION	0.67	0.61	0.64	0.82	0.88	0.95	0.97	0.95	**0.96**
CONDITION	0.58	0.52	0.55	0.80	0.86	0.93	0.98	0.87	**0.92**
DISPOSITION	0.34	0.34	0.34	0.71	0.77	0.84	0.82	0.88	**0.85**
DISCHARGE DIAGNOSES	0.65	0.69	0.67	0.88	0.91	0.94	0.98	0.94	**0.96**
PHYSICAL EXAM ON DISCH.	0.71	0.69	0.70	0.88	0.86	0.84	0.91	0.89	**0.90**
ALLERGIES	0.63	0.59	0.61	0.85	0.87	0.89	0.94	0.88	**0.91**
FAMILY HISTORY	0.60	0.60	0.60	0.84	0.87	0.90	0.94	0.90	**0.92**
PAST MEDICAL HISTORY	0.68	0.70	0.69	0.85	0.89	0.87	0.97	0.95	**0.96**
PAST SURGICAL HISTORY	0.69	0.65	0.67	0.90	0.89	0.88	0.93	0.93	**0.93**
SOCIAL HISTORY	0.55	0.64	0.59	0.79	0.81	0.83	0.89	0.85	**0.87**
CONSULTATION	0.25	0.35	0.29	0.67	0.68	0.69	0.84	0.82	**0.83**
HOSPITAL COURSE	0.78	0.72	0.75	0.91	0.93	0.95	0.96	0.96	**0.96**
PHYSICAL	0.65	0.71	0.68	0.81	0.82	0.83	0.95	0.84	**0.89**
PROCEDURES	0.73	0.71	0.72	0.86	0.88	0.90	0.98	0.87	**0.92**
DISCHARGE INSTRUCTIONS	0.88	0.80	0.84	0.96	0.98	0.98	0.99	0.99	**0.99**
DISCHARGE MEDICATIONS	0.79	0.79	0.79	0.95	0.97	0.99	0.99	0.97	**0.98**
Macro-Average	0.63	0.63	0.63	0.86	0.88	0.90	0.95	0.92	**0.93**
Micro-Average	0.64	0.63	0.64	0.90	0.94	0.92	0.95	0.94	**0.95**

Table 3.9: Section identification results for hospital discharge summaries.

of the three corpora, the radiology report corpus saw the worst performance by all models. This again is probably due to the relatively small amount of data since radiology reports tend to be shorter and thus contain less content.

Since the report sections vary in size, I computed both macro- and micro-averaged precision, recall, and F_1-measure (last two rows in Tables 3.2.4, 3.2.4, 3.2.4). The CRF model's micro-averaged F_1-measure was 93%, 95%, and 97% for radiology reports, hospital discharge summaries, and psychiatric reports respectively. Similar to my prior work [Banisakher et al., 2018a], both the CRF model and the HHMM baseline seemed to nei-

Section	Independent Bigram			HHMM			CRF		
	P	R	F_1	P	R	F_1	P	R	F_1
CLINICAL HISTORY	0.72	0.70	0.71	0.96	0.89	0.92	0.98	0.94	**0.96**
EXAM	0.50	0.48	0.49	0.83	0.83	0.83	0.91	0.95	**0.93**
COMPARISON	0.38	0.38	0.38	0.81	0.70	0.75	0.89	0.81	**0.85**
CONTRAST	0.20	0.22	0.21	0.68	0.68	0.68	0.75	0.78	**0.76**
PROCEDURE	0.50	0.41	0.45	0.87	0.76	0.81	0.96	0.87	**0.91**
FINDINGS	0.77	0.73	0.75	0.93	0.89	0.91	0.95	0.91	**0.93**
IMPRESSION	0.68	0.59	0.63	0.84	0.80	0.82	0.95	0.92	**0.94**
Macro-Average	0.54	0.50	0.52	0.85	0.79	0.82	0.91	0.88	**0.90**
Micro-Average	0.52	0.50	0.51	0.87	0.83	0.85	0.95	0.90	**0.93**

Table 3.10: Section identification results for radiology reports.

ther overfit nor underfit, which is indicated by higher micro-averaged compared to the macro-averaged scores.

As for the boundary detection problem, and similar to the evaluation in [Sherman and Liu, 2008, Banisakher et al., 2018a], I performed two experiments for the first two baselines since both baselines require a parameter representing the number of boundaries (number of topics minus one). In the first experiment I allowed the parameter to be chosen by LCSeg and TopicTiling, respectively, while in the second experiment, I provided the algorithms with the correct number of boundaries (i.e., number of sections minus one). The CRF, the HMM-LSA as well as the HHMM models, however, need no prior information regarding the number of sections present in a given report. Table 3.2.4 shows the W_d and P_k scores for all five approaches. My system again outperformed all the baselines indicated by lower W_d and P_k error rates overall. Both the text segmentation baselines performed better when the number of boundaries is known—an expected result.

Finally, I conducted three feature combination experiments for both subtasks. In the first experiment, I used n-gram-only features, while in the second experiment I added sentence position and length features, and finally in the third, I used all the feature sets. Table 3.2.4 shows the section ordering (identification) results for each of those experi-

# of Boundaries	Algorithm	Psychiatric		Discharge		Radiology	
		P_k	W_d	P_k	W_d	P_k	W_d
System Choice	LCSeg	0.29	0.37	0.31	0.40	0.24	0.34
	TopicTiling	0.27	0.33	0.28	0.35	0.22	0.30
Provided	LCSeg	0.25	0.33	0.27	0.34	0.22	0.31
	TopicTiling	0.20	0.25	0.21	0.26	0.20	0.23
	HMM-LSA	0.25	0.32	0.26	0.35	0.25	0.30
	HHMM	0.20	0.26	0.19	0.26	0.21	0.28
	CRF_Bigram_Only	0.20	0.26	0.19	0.26	0.21	0.28
	CRF_Bigram_Position	0.20	0.26	0.19	0.26	0.21	0.28
	CRF_All	**0.17**	**0.20**	**0.16**	**0.18**	**0.18**	**0.20**

Table 3.11: Section boundary identification results.

	Psychiatric			Discharge			Radiology		
Algorithm	P	R	F_1	P	R	F_1	P	R	F_1
CRF_Bigram_Only	0.91	0.87	0.89	0.93	0.89	0.91	0.84	0.80	0.82
CRF_Bigram_Position	0.94	0.90	0.92	0.93	0.91	0.92	0.91	0.85	0.88
CRF_All	**0.95**	**0.92**	**0.94**	**0.95**	**0.92**	**0.93**	**0.91**	**0.88**	**0.90**

Table 3.12: Feature combination experiments for section identification.

ments. I report macro-F_1 scores for this subtask. Table 3.2.4 also shows the results of the feature experiments for the boundary location subtask. The CRF models outperformed the HHMM in all experiments including the *CRF_Bigram_Only* with the *CRF_ALL* model , which included all features, achieving the best performance. Adding the sentence position and length features to the *CRF_Bigram_Position* model significantly improved the results for both subtasks. This further confirms that section content and sentences within a clinical report are dependent upon the entire report in which they are found.

3.2.5 Future Directions

My work demonstrates the feasibility of learning a section structure for documents containing section-ordered free text. Thus, several next steps can be taken to both improve and build upon the problem and models I demonstrated. First, the data I used in my study can be expanded in two dimensions: quantity and type. Additional medical documents would help in training and honing the supervised models I built and would allow for further evaluation and analysis. This however is difficult due to the limitations inherent to medical data access as well as due to the time-intensive annotation process needed to produce the necessary data. Additionally, my section learning models are not limited to medical document applications, as they can be further applied to other documents types such as patent documents and scientific articles or any other structurally-similar document types.

Second, although statistical models carry several advantages as they are highly interpretable and are simple to develop and replicate, they may not be the best solution in a practical setting when compared to deep learning models as their computation time can be longer and their fine tuning limited. Thus further development of models such as Gated Recurrent Units (GRU) and Long-Short Term Memory (LSTM) for this task may lead to better results and would be a question worthy of evaluation. This however, is again limited by the size of data available for training. Opting for a statistical based model was thus justified in this study given the small amount of data available.

Third, and most importantly, this study opens the door for the introduction of a new problem: section type discovery. That is, identifying and clustering sections for a given document type (e.g. psychiatric evaluations) in an unsupervised manner. Automatically labeling sections with a pre-defined ontology of section types is useful for document understanding, and has been shown to improve tasks as varied as information extraction [Hofmann et al., 2009, Tepper et al., 2012], data mining [Dou et al., 2015, Repta et al.,

2018], and document search [Wu et al., 2015a, Xu and Croft, 2017, Doucet, 2018]. But where does the ontology come from? Automatically labeling sections with their types requires not just a list of possible sections, but also what different headers are used for each, their usual order (with possible exceptions), and the type of language normally found within. As discussed above, manually creating this knowledge is laborious and error prone, and so a solution to automatically discovering it from examples would be preferred. Therefore, I propose and are currently working towards an approach to discovering section types for a given document type in a data-driven manner using a combination of a modified Bayesian model merging algorithm [Stolcke and Omohundro, 1994], and the Analogical Story Merging (ASM) algorithm presented by Finlayson *et al.*

3.2.6 Related Work

There are several tasks related to the subtasks I outlined that have been investigated by other researchers. The section ordering subtask has been referred to as argumentative zoning [Teufel et al., 1999, Li et al., 2010, Denny et al., 2009a]. Argumentative zoning refers to classifying text sections into mutually exclusive categories. Work on this task is mostly centered around identifying scientific article sections (e.g., abstract, introduction, methodology, etc.) [Teufel, 1999]. The boundary location subtask can be considered as a type of text segmentation problem [Hearst, 1994, Riedl and Biemann, 2012]. There has been an extensive amount of research in general text segmentation tasks [Simmons et al., 2016, Eskenazi et al., 2017]. However, work on text segmentation of clinical notes has been limited [Ganesan and Subotin, 2014]. Most prior approaches have either treated this task as a section classification task (as opposed to a sequence labeling task) thus discarding contextual information [Pomares-Quimbaya et al., 2019]. Additionally, several approaches have employed heuristics and regular expressions specific to a document type

or a source in detecting section headers. Thus failing when faced with unseen data from different sources [Ganesan and Subotin, 2014].

In this work, I extend an earlier study on section identification of psychiatric evaluation reports that combined the work of Li *et al.* [Li et al., 2010] on identifying section types within clinical reports and that of Sherman and Liu [Sherman and Liu, 2008] on text segmentation of meeting minutes. Li *et al.* modeled HMM emissions at the section level using bigrams, while Sherman and Liu modeled the emissions at the sentence level and used unigrams and trigrams. Other approaches followed similar strategies in segmenting story text and in creating generative models for detecting story boundaries [Mulbregt et al., 1998, Yamron et al., 1998]. More recently, Yu *et al.* [Yu et al., 2016] used a hybrid deep neural network combined with a Hidden Markov Model (DNN-HMM) to segment speech transcripts from broadcast news to a sequence of stories.

Further, there are several studies that have demonstrated approaches for the identification, classification, and segmentation of clinical notes. Most of which however, focused on the identification of section headers rather than content and used source- and note-specific heuristics. Denny *et al.* [Denny et al., 2009b] developed the SecTag algorithm which uses terminology-based rules, and naive Bayesian scoring methods to identify note section headers in "history and physical examination documents" (H&P notes). Their approach relied on data from a single source and a specific clinical note type, limiting its generalizability. My work however addresses multiple clinical note types and while the discharge summaries and radiology corpora were pulled from the same source, the psychiatric evaluation reports corpus was collected from various sources that followed different formats.

Using machine learning methods, Apostolova *et al.* [Apostolova et al., 2009] and Tepper *et al.* [Tepper et al., 2012] demonstrated automatic supervised approaches for detecting section headers and boundaries but showed low adaptability when faced with

various clinical note types [Ganesan and Subotin, 2014]. Haug *et al.* [Haug et al., 2014] and Ganesan and Subotin [Ganesan and Subotin, 2014] also developed machine learning methods to secgment clinical notes sections. While these studies were extensive in applying their methods to both a multitude of clinical note types and sizes, both neglected implicit (unlabeled sections) sections or relaxed the problem by collapsing these subsections into their parent sections. Most closely, Dai *et al.* [Dai et al., 2015] used CRFs to segment clinical notes at the token-level. They report a minor improvement over the sentence-level approach which objectively does not improve section identification especially when considering the computational overhead for processing notes at the token-level. Additionally, although they mention implicit sections in their study, these sections are not truly implicit as they consider sections to be implicit only when their headers are surrounded by other text rather than appearing stand-alone on an isolated line in text. In this study, I consider sections to be implicit if and only if the header is completely missing and related information is included within other sections.

3.3 Improving the Identification of the Discourse Function of News Article Paragraphs

Identifying the discourse structure of documents is an important task in understanding written text. Building on prior work, I demonstrate an improved approach (using CRFs extended from the previous section) to automatically identifying the discourse function of paragraphs in news articles. I start with the hierarchical theory of news discourse developed by van Dijk [1988] which proposes how paragraphs function within news articles. This discourse information is a level intermediate between phrase- or sentence-sized discourse segments and document genre, characterizing how individual paragraphs convey information about the events in the storyline of the article. Specifically, the theory cat-

egorizes the relationships between narrated events and (1) the overall storyline (such as MAIN EVENTS, BACKGROUND, or CONSEQUENCES) as well as (2) commentary (such as VERBAL REACTIONS and EVALUATIONS).

3.3.1 Introduction

News articles usually follow strong principles of journalistic structure. By design, they often begin with a introductory summary of main events, followed by detailed exposition of the main events and consequences, interspersed in a stereotyped fashion with relevant background information, current and past evidence, and reported speech. Yarlott et al. [2018] demonstrated the feasibility of detecting this type of discourse structure for news articles using an established hierarchical theory of news discourse [van Dijk, 1988]. In their study, they showed that it was feasible to identify the discourse function of news paragraphs using a support vector machine (SVM) model and a small set of simple linguistic features, with a performance of 0.54 F_1.

Similar to Yarlott et al.'s [2018] approach, I demonstrate an improved approach to automatically labeling news article paragraphs with the van Dijk discourse functions Yarlott et al. [2018] applied in their study. My work uses a conditional random field (CRF) model, along with new features, to obtain an improved performance of 0.71 F_1. Most importantly, my model is able to precisely capture the interdependencies between the various discourse label types, which flows from my hypothesis that each paragraph in an article is dependent not only on the previous one but rather on a longer sequence of previous paragraphs.

The remainder of this section is structured as follows. I first provide a definition of van Dijk's theory as was presented in [Yarlott et al., 2018] (§3.3.2). Second, I describe the dataset I used in training and testing my CRF model (§3.3.3). I then detail the dis-

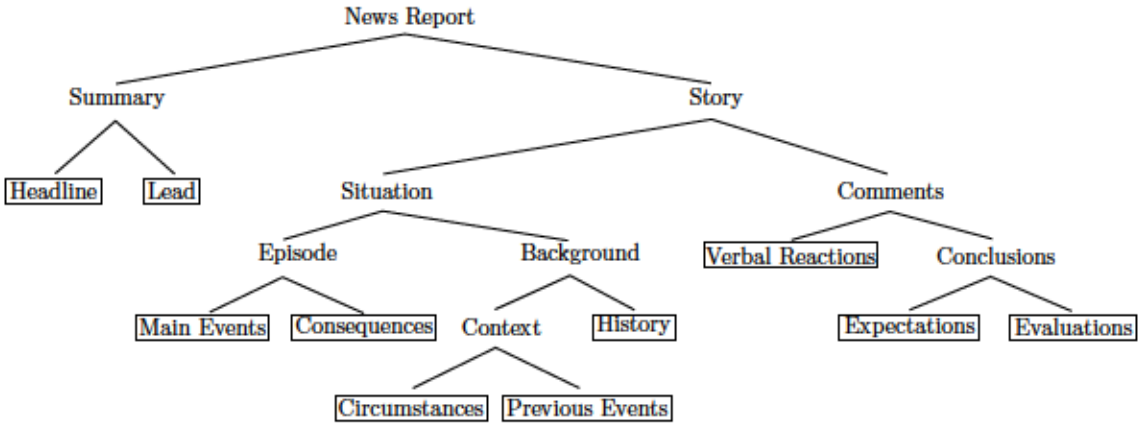


Figure 3.3: The hierarchical discourse structure of news proposed by van Dijk [van Dijk, 1988]. Boxes indicate labels that were directly annotated on the documents; other labels can be inferred. From Yarlott et al. [2018], Figure 1.

course label identification methods, including the CRF model and how it captures both section ordering and section content, how the model is trained, and the features it leverages (§3.3.4). I next compare the performance of the CRF model with various baselines, demonstrating that it performs better than prior models (§3.3.5). I finally conclude this section with a discussion of the related work (§5.2.2).

3.3.2 Van Dijk's Theory of News Discourse

Van Dijk [1988] described a hierarchical theory of news discourse, the categories of which are shown in Figure 3.3, which I apply to a subset of the news articles of the ACE Phase 2 corpus. In this section, I repeat the descriptions of the leaf categories from Yarlott et al.'s prior paper, as well as their parent categories when appropriate, for ease of reference.

SUMMARY elements express the major subject of the article, with the HEADLINE being the actual headline of the article, and the LEAD being the first sentence, which is often a summary of the main events of the article.

SITUATION elements are the actual events that comprise the major subject of the article. EPISODES concern MAIN EVENTS, which are those events that directly relate to the

major subject of the article, and the CONSEQUENCES of those events. The BACKGROUND provides important information about the relation of each paragraph with respect to the central events of a news story. Background includes the CONTEXT, of which CIRCUMSTANCES are temporally or spatially non-specific states that contribute to understanding the subject, while PREVIOUS EVENTS are specific recent events that enhance understanding of the main events. HISTORY paragraphs are another type of Background describing events that have not occurred recently, typically referenced in terms of years prior, rather than months, weeks, or days.

COMMENTS provide further supporting context for the central events of an article. Comments may include VERBAL REACTIONS solicited from an external source, such as a person involved in the events, or an expert. CONCLUSIONS, by contrast, are comments made by a journalistic entity (the newspaper, reporter, etc.) regarding the subject. Conclusions can be separated into EXPECTATIONS about the resolution or consequences of an event, or EVALUATIONS of the current situation.

3.3.3 Dataset

Following the discussion in chapter 2, I used a gold-standard corpus previously developed by Yarlott et al. [2018] of van Dijk's labels applied to a subset of the Automated Content Extraction (ACE) Phase 2 corpus [NIST, 2002] (Corpus 2.6). The ACE Phase 2 corpus is a major standard corpora of news articles that boasts three advantages: it is widely-used, has relevance to other tasks, and was readily available to researchers. This dataset comprises 50 documents containing 28,236 words divided in 644 paragraphs. I include Table 3.13 below (repeated from chapter 2 for ease of reference), which shows the corpus-wide statistics for the number of words and paragraphs, where each paragraph is given a single type in accordance to van Dijk's theory.

	Words	Paragraphs
Total	28,236	644
Average	564.7	12.9
Std. Dev.	322.1	4.9

Table 3.13: Corpus-wide statistics for Corpus 5: News Articles. Adapted from Yarlott et al. [2018], Table 1.

Yarlott et al. [2018] doubly annotated 50 randomly selected news articles, divided into ten sets of five documents each. Within these sets, documents were swapped or replaced in order to obtain uniform sets in terms of total document lengths. The majority of texts were already divided into paragraphs in an obvious manner, either with empty lines or with indentation. The remaining texts were divided by the adjudicator based on either contextual or structural clues, such as abrupt change in topic or unnatural line breaks. The authors report an all-around high agreement with the gold standard ($F_1 = 0.85$, $\kappa = 0.75$) which demonstrates that the gold-standard was not dominated by a single annotator.

Although the dataset discussed was annotated for all labels discussed here, the HEADLINE label could be computed automatically from the structure of ACE Phase 2 corpus, as the files has the headline separate as part of its markup scheme.

Table 3.14 provides the resulting distribution of van Dijk's labels. Verbal reactions and circumstances dominate the labels. Although the distribution of labels is highly skewed, I find that this is roughly in-line with the style of reporting featured in the ACE Phase 2 corpus, which seeks comments and analysis from experts within the field as well as explaining the immediate context that has an effect on the main event.

3.3.4 Identifying Discourse Labels

In contrast to the approach reported by [Yarlott et al., 2018], I treated the identification of paragraph functions as a sequence modeling task. Formally, the task is as follows:

Label	Count	Label	Count
HEADLINE	50	LEAD	42
MAIN EVENTS	60	CONSEQUENCES	19
CIRCUMSTANCES	103	PREVIOUS EVENTS	64
HISTORY	27	VERBAL REACTIONS	252
EXPECTATIONS	21	EVALUATIONS	56

Table 3.14: Distribution of the labels within the annotated corpus, with 644 labels total. The majority of paragraphs fall under the categories of verbal reactions or circumstances. From [Yarlott et al., 2018]

given a news report with n discourse labels and m paragraphs, where the paragraphs are unlabeled, identify the optimal sequence (order) of discourse labels $H^* = (L_1^*, \ldots, L_n^*)$ from among all possible label sequences, and assign every paragraph a discourse label $H^* = (H_1, \ldots, H_m)$ consistent with L^*. Sequence labeling problems in NLP, medical informatics, and discourse parsing have been studied by both generative and discriminative approaches, including Hidden Markov Models (HMMs; generative) and Conditional Random Fields (CRFs; discriminative). Li et al. [2010] used HMM and n-gram models to detect the orders or labels of sections within clinical reports, while modeling the observation probabilities at the section level. Sherman and Liu [2008] used HMMs as well as n-gram models to detect topic shifts in meeting minutes, and, in contrast to Li et al., modeled the observation probabilities on the sentence level.

My approach was inspired by the method described in Banisakher et al. [2018a] (discussed previously in §3.1) which identifies section labels in clinical psychiatric reports. As discussed previously, my previous approach combined a *Hierarchical* Hidden Markov Model (HHMM)—which used section statistics as the model's transition probabilities—with n-grams for the observation probabilities of words. In this study, I substitute a CRF for the HHMM (similar to the approach I outlined in §3.2). Generative models such as HMMs have more explanatory power when compared with their discriminative counterparts such as CRFs. However, HMMs, rely on the assumption that observations are

statistically independent from one another. For this problem, this means that an HMM assumes that the presence of certain paragraphs corresponding to a certain discourse label or function A is independent from other paragraphs within another section B. In practice, however, this is not the case: for example a paragraph following the MAIN EVENTS are often either CONSEQUENCES or CIRCUMSTANCES.

Linear Chain Conditional Random Fields

I built and trained a linear chain CRF modeled on Banisakher et al.'s HHMM approach. In contrast to an HHMM, the CRF encodes labels as nodes in the CRF graphical representation (instead of HMM states), and uses weighted feature functions for transitions between nodes (instead of the HMM transition and emission probabilities). Additionally, the CRF model captures the "true" desired probability distribution, that is the *conditional distribution* of labels given the observations $P(Y|X)$, instead of modeling the joint distribution of observations and labels $P(X, Y)$. This a known advantage of CRFs in general over HMMs and is mainly due to, again, removing the independence assumption. Thus, CRFs can have an arbitrary number of dependencies as opposed to the limited dependency structure of HMMs. My model benefits from this as it does not only record the dependence of a discourse label only on its predecessor and observations, but on additional dependencies given the entire sequence of labels (i.e., paragraph discourse functions) and observations (i.e., paragraphs).

I built and trained a linear chain CRF analogous to the prior HHMM approach. From the earlier discussion in §3.2, in contrast to an HHMM, the CRF encodes sections as nodes in the CRF graphical representation (instead of HMM states), and uses weighted feature functions for transitions between nodes (instead of the HMM transition and emission probabilities). Additionally, the CRF model captures the "true" desired probability distribution, that is the *conditional distribution* of labels given the observations $P(Y|X)$,

instead of modeling the joint distribution of observations and labels $P(X, Y)$. This a known advantage of CRFs in general over HMMs and is mainly due to, again, removing the independence assumption. Thus, CRFs can have an arbitrary number of dependencies as opposed to the limited dependency structure of HMMs. The CRF model benefits from this as it does not only record the dependence of a section only on its predecessor and observations, but on additional dependencies given the entire sequence of labels (i.e., section types) and observations (i.e., sentences).

The rest of the CRF model details follow exactly as discussed in §3.2.3.

Modeling the Discourse Labels' Order

The feature function F_j (from Equation 3.7) incorporates section ordering through the section ordering function $H(l_{i-1}, l_i, \overline{p}, i)$ (Equation 3.9). As discussed in detail in §3.2.3, there is a feature function for each of the dependencies defined in the model. I encode the interdependent order of labels (i.e., which labels depend upon each other) using a binary matrix following the same setup.

Modeling the Discourse Labels' Content

Similarly, the feature function F_j (from Equation 3.7) incorporates the discourse label type content via the feature function $SF(l_{i-1}, l_i, \overline{p}, i)$. These functions model the dependency between a label type and its content. Importantly, the feature function should not be confused with the linguistic features that are extracted from the text and input into the section feature function. To capture label content (i.e., to model discourse label type-specific language) I extracted the following set of features:

Features from Yarlott et al. [2018]: *Unigrams* (i.e., bag of words), the *tf-idf* count vector of the top 3 words (across the corpus) per label type, bag-of-words, and *paragraph vectors* using the Doc2Vec approach [Le and Mikolov, 2014]. As pointed out by Yarlott et al., the

tf-idf and *paragraph vectors* approximate topics within a given paragraph. Yarlott et al. also used the previous paragraph's label as an explicit feature; this is included by default in the CRF model.

Lexical: *Bigrams* to capture the type of language per discourse label type.

Positional: *Size of paragraphs* represented by number sentences present. As well as the *paragraph position* relative to the document head.

Syntactic: A *POS count vector* which encodes the number of times each part of speech (POS) (specifically, nouns, verbs, adjectives, and adverbs) appears in the paragraph.

Semantic: Here I incorporated four additional features: a *reported speech* feature, a *majority event tense* feature, a *subevent relation* count vector, and *NER vectors* representing a select set of named entities. For the *reported speech* feature, I extracted quotations and sentences with tagged as reported speech by the `textacy` library [DeWilde, 2020] and labeled the containing paragraph as VERBAL REACTIONS. For the *majority event tense* feature, I extracted the events in each paragraph using the CAEVO event extraction system [Chambers et al., 2014], noted their POS tags using a dependency tree, and recorded the majority verb tense in that paragraph. For the *subevent relation* feature, I used Aldawsari and Finlayson's subevent extraction system (2019) to capture relationships between paragraphs. For this, I used a vector for each paragraph corresponding to the number of paragraphs of the article with the maximum number of paragraphs in the corpus. Aldawsari and Finlayson [2019] presented a supervised model for automatically identifying when one event is a subevent of another using narrative and discourse features. For each event relation found by this system between two distinct paragraphs, I recorded a +1 in that corresponding vector cell, while I discarded relationships found within a single paragraph. For the *NER vectors*, I applied Named Entity Recognition (NER) and extracted the first 13 named entity types found by the Spacy library [AI, 2020] including PERSON, LOCATION, DATE, and TIME. These 13 types were represented in

a numerical vector for each discourse label type such that, for each type, I recorded the number of entity occurrences.

Inference

I applied the usual inference process for linear chain CRFs operating at the paragraph level [Forney, 1973]. Inference in linear chain CRFs follows a similar algorithm to Viterbi, which is used in decoding HMM models. While not stated explicitly in the Equation 3.7 above, the normalization factor $Z(S)$ is calculated as is often done using the Gaussian prior as it was introduced in [Chen and Rosenfeld, 1999].

3.3.5 Results and Discussion

In order to test the CRF model, I randomly split each corpus into training and testing sets in a cross-validation setup, using five folds, resulting in 40 news reports for training and 10 for testing in each fold. The CRF model was trained to learn a total of 9 distinct discourse label types as represented in 3.14 (all leaf labels minus HEADLINE). In this section I describe the baseline comparisons and overall experiments and results.

Baseline Methods

I followed and extened Yarlott et al. [2018] in their baseline comparisons. I compared my model's performance against six other methods: two baselines including the most frequent class (MFC) and a support vector machine using bag-of-words (SVM+BoW); third, a decision tree classifier; fourth, a random forest classifier; and fifth, Yarlott et al. [2018]'s best performing model, a support vector machine. As described above, the latter three models incorporate a the following set of four features: bag-of-words, *tf-idf*, paragraph vectors, and previous paragraph labels. I used the same experimental setup for all of

these models. Yarlott et al. [2018] obtained the best experimental results using grid search to maximize the micro-averaged performance of each classifier, as measured across five folds. Following Yarlott et al. [2018], the SVM classifier uses a linear kernel with $C = 10$ and the class weights balanced based on the training data; the decision tree classifier uses the default parameters with the class weights balanced; the random forest uses 50 estimators with balanced class weights. Additionally, I included a sixth baseline (an HHMM) following from my earlier work in Banisakher et al. [2018a] and as discussed in §3.1.4.

Results

The CRF model outperformed all other classifiers and baselines achieving a 0.71 F_1 score. Table 3.15 shows the micro-averaged precision (P), recall (R), and F_1 scores for the five models from [Yarlott et al., 2018] as well as our current CRF approach. The experimental results show that our CRF approach is a substantial improvement over the previously best performing model.

For the CRF model, I performed 8 feature combination experiments (shown in Table 3.15) to evaluate the effect of feature classes as well as the individual semantic features. As discussed before, the SVM as well as the decision tree and random forest classifiers only leveraged Yarlott et al.'s original four features: bag-of-words, *tf-idf*, paragraph vectors, and previous paragraph labels. While our CRF approach uses a more sophisticated set of features leveraging additional syntactic and semantic features as outlined in 5.2.4. Most importantly, my model treats the problem as a sequence labeling task and therefore captures the sequential dependencies between the paragraphs as well as the labels within each report. This is evidenced by the CRF model that uses only Yarlott et al.'s features, which achieves a higher performance than the original SVM classifier.

The CRF model achieved the largest increase in performance after adding the semantic features. This was expected: I anticipated a boost in performance on the VERBAL RE-

Model	Features	P	R	F_1
MFC	-	0.39	0.39	0.39
HHMM	Bigrams	0.42	0.45	0.43
SVM	BoW	0.46	0.46	0.46
DT	Yarlott et al.	0.41	0.41	0.41
RDF	Yarlott et al.	0.43	0.43	0.43
SVM	Yarlott et al.	0.54	0.54	0.54
CRF	Yarlott et al.	0.58	0.60	0.59
CRF	+Lexical	0.61	0.63	0.62
CRF	+Positional	0.62	0.66	0.64
CRF	+Syntactic	0.65	0.69	0.67
CRF	*+subevent relation*	0.65	0.0.70	0.67
CRF	*+majority event tense*	0.67	0.71	0.68
CRF	*+reported speech*	0.68	0.72	0.70
CRF	All (+Remaining Sem.)	0.69	0.73	**0.71**

Table 3.15: Experimental results for discourse label identification. All results are micro-averaged across instances, including precision (P), recall (R), and balanced F-measure (F_1). The HHMM used Bigram features as discussed in [Banisakher et al., 2018a] and §3.1. The Decision Tree, Random Forest, and SVM classifiers used the features outlined in [Yarlott et al., 2018]. For the middle three lines of the CRF section, these indicate features groups added to the previous line's model. I present the results for the semantic features individually (the last four lines). The CRF model in the last line (CRF with ALL features) includes all the features from the previous lines as well as all remaining semantic features.

ACTIONS class given detection of reported speech, and a similar increase in performance on the MAIN EVENTS and PREVIOUS EVENTS classes given the addition of event and subevent features. Of the semantic features, the *reported speech* feature had the biggest impact on the model's performance as the verbal reactions section was predominant in the dataset. Here `textacy` performed quite well in automatically identifying reported speech as the model achieved a 0.91 F_1 score for the VERBAL REACTIONS class. The *subevent relation* and *majority event tense* features improved the performance by about one point F_1 each, with the second contributing slightly more to the overall performance. The *majority event tense* feature contributed heavily to the PREVIOUS EVENTS and HISTORY, I suspect due to the relatively more frequent use of past tense verbs in paragraphs

belonging to those classes. As discussed before, I used automated systems to detect events and subevent relations. Naturally, these systems do not boast a perfect performance and therefore error propagation is expected. Thus, I expect that my model can further achieve higher performance using more refined event detection solutions, as well as a larger corpus.

Table 3.16 presents the per-label results from our experiments. The relatively strong performance on CIRCUMSTANCES and VERBAL REACTIONS is not surprising, given their relative prevalence in the news articles corpus. Similarly it is not surprising to see the low performance on labels that occur, on average, about once (or less) a document (HISTORY, EXPECTATIONS). However, these label types saw a significant performance boost in my model compared to the previous approaches as our features have captured more of their distinct language. For CONSEQUENCES HISTORY, EXPECTATIONS, and EVALUATIONS, the syntactic and positional features were most helpful. Similar to [Yarlott et al., 2018], I observe an unexpected–but not surprising–level of performance on LEAD paragraphs, given their relative scarcity in the dataset: I find that leads, with a single exception, occur once at the start of the document.

Again, similar to [Yarlott et al., 2018], I expected the tree-oriented methods—decision trees and random forests—to at least outperform the SVM classifier. However, this was not the case in practice and they were outperformed by one of the baselines. I believe that this partially attributed to the fact that these models did not leverage the full set of hierarchical labels in van Dijk's discourse theory: they were only presented with the leaf labels.

Label Type	F_1	Label Type	F_1
HEADLINE	-	LEAD	0.95
MAIN EVENTS	0.69	CONSEQUENCES	0.29
CIRCUMSTANCES	0.72	PREVIOUS EVENTS	0.51
HISTORY	0.24	VERBAL REACTIONS	0.91
EXPECTATIONS	0.26	EVALUATIONS	0.51
		Macro Average	0.56

Table 3.16: Per-label F_1 results. The last row shows the macro average over all label types. Best performance occurs for the LEAD, MAIN EVENTS, CIRCUMSTANCES, and VERBAL REACTIONS.

3.3.6 Related Work

There has been substantial work describing how the structure of news operates with regards to the chronology of real-world events. Much news follows an inverted chronology—called the inverted pyramid [Bell, 1998, Delin, 2000] or relevance ordering [Van Dijk, 1986]—where the most important and typically the most recent events come first. Bell claims that *"news stories... are seldom if ever told in chronological order"* [Bell, 1994, p. 105], which is demonstrated by Rafiee et al. for both Western (Dutch) and non-Western (Iranian) news (2018). Rafiee et al. also show that many stories follow a hybrid structure, which combines characteristics from both inverted and chronological structures.

Our approach was inspired by Banisakher et al. [2018a]'s HHMM approach to section identification in clinical notes. In turn, their work extend an earlier study on section identification of psychiatric evaluation reports that combined the work of Li et al. [2010] on identifying section types within clinical reports and that of Sherman and Liu [2008] on text segmentation of meeting minutes. Li et al. modeled HMM emissions at the section level using bigrams, while Sherman and Liu modeled the emissions at the sentence level and used unigrams and trigrams. Other approaches followed similar strategies in segmenting story text and in creating generative models for detecting story boundaries [Mulbregt et al., 1998, Yamron et al., 1998]. More recently, Yu et al. [2016] used a

hybrid deep neural network combined with a Hidden Markov Model (DNN-HMM) to segment speech transcripts from broadcast news to a sequence of stories. Similar to my approach, [Sprugnoli et al., 2017] used CRFs and SVMs for the classification of automatic classification of Content Types, a novel task that was introduced to provide cues to access the structure of a document's types of functional content.

Discussing van Dijk's theory of news discourse, Bekalu stated that analysis of *"the processes involved in the production of news discourses and their structures will ultimately derive their relevance from our insights into the consequences, effects, or functions for readers in different social contexts, which obviously leads us to a consideration of news comprehension"* [2006, p. 150]. The theory proposed by van Dijk has also been proposed for use in annotating the global structure of elementary discourse units in Dutch news articles [van der Vliet et al., 2011].

Pan and Kosicki [1993], in a similar analysis, presented a framing-based approach that provides four structural dimensions for the analysis of news discourse: syntactic structure, script structure, thematic structure, and rhetorical structure. Of these, the syntactic structure is most closely aligned with van Dijk's theory. In this study, I chose to focus on van Dijk's theory as Pan and Kosicki do not provide a list or description of the structure that could be readily translated into an annotation scheme.

While White [1998] treats the structure of news as being centered around the headline and lead. White suggests that the headline and lead, which act as a combination of both synopsis and abstract for the news story, serve as the nucleus for the rest of the text: *"the body which follows the headline/lead nucleus—acts to specify the meanings presented in the opening headline/lead nucleus through elaboration, contextualisation, explanation, and appraisal"* [1998, p. 275]. I focus on van Dijk's theory for this study as I find it to provide a higher degree of specificity: White's specification modes serve roughly the same purpose as higher-level groupings in van Dijk's theory.

CHAPTER 4

AUTOMATIC SECTION STRUCTURE CLUSTERING

Labeling document sections (e.g., *Introduction*, *Methods*, *Conclusion*, etc.) is an important step in automatic document understanding and is useful for information extraction, data mining, and document search. In the absence of explicit headings however, labeling requires knowledge of the section types: what sections should be present, in what order, their various possible headings, and containing what kind of language. In this chapter, I describe an approach to automatically discovering section type knowledge for a document class in a data-driven fashion using a modified Bayesian model merging algorithm. I tested my approach on five different document classes from three domains: psychiatric evaluations, radiology reports, and discharge summaries (Corpora 2.1-2.3) in the clinical domain; patent documents (Corpus 2.4) in the intellectual property (IP) domain, and environmental scientific articles (Corpus 2.5) from the scientific domain.

4.1 Introduction

Many types of documents have explicit section structure, that is, headers which delimit blocks of text and set expectations about the content and purpose of that block. In the absence of explicit headers, or in the face of non-standard headers, automatically labeling sections with a pre-defined ontology of section types is useful for document understanding, and has been shown to improve tasks as varied as information extraction [Hofmann et al., 2009, Tepper et al., 2012], data mining [Dou et al., 2015, Repta et al., 2018], and document search [Wu et al., 2015a, Xu and Croft, 2017, Doucet, 2018]. Additionally, automatic section labeling has been shown to be tractable and has been demonstrated on psychiatric document understanding [Banisakher et al., 2018a]. But where does the section type ontology come from? Automatically labeling sections with their types requires not just a list of possible sections, but also their usual order (with possible exceptions),

what different headers might be used for each type, and the kind of language normally found within. Manually creating this knowledge is laborious and error prone, and so a solution to automatically discovering it from examples would be preferred.

Interestingly, automatically discovering the types is challenging: for a document class (e.g., a *psychiatric evaluation* or a *U.S. patent*), the presence of a particular section type is often ambiguous. First, there is great variety and ambiguity in the section headers; second, sections are sometimes included within other sections; third, the section order might not be strict; and finally, sections may be omitted for a variety of reasons.

Here I describe an approach to discovering section types for a given document class in a data-driven manner. My approach uses a modified Bayesian model merging algorithm [Stolcke and Omohundro, 1994], as inspired by the Analogical Story Merging (ASM) algorithm presented by Finlayson [2016]. I demonstrate this approach on five different corpora from two domains: psychiatric evaluations, radiology reports, and discharge summaries (Corpora 2.1-2.3) in the clinical domain; patent documents (Corpus 2.4) in the intellectual property (IP) domain, and environmental scientific articles (Corpus 2.5) from the scientific domain. I show that it is feasible to learn the section structure of documents without a pre-existing ontology of sections.

The chapter is organized as follows. I first describe the datasets I used and the challenges of discerning section types for each corresponding document class (§4.4). I then define the task (§4.3) and describe my approach and its steps (§4.4). Next I compare the performance of my approach with various baselines (§4.5), demonstrating that it performs better than existing document clustering approaches for my task (§3.3.5). Finally, I discuss related work (§4.7) as well as the limitations and future directions of my approach (§4.8).

Corpus	Document Class	# of Docs.	# of Secs.	Secs./Doc.	Words/Doc.
1	Psychiatric Evaluations	150	2,824	18.8	1,521
2	Radiology Reports	423	2,538	6.0	463
3	Discharge Summaries	150	2,977	19.8	1,829
4	U.S. Patents	464	3,249	7.0	18,351
5	Scientific Articles	19	111	7.4	4,741

Table 4.1: Summary of corpora statistics.

4.2 Data and Challenges

I tested my approach on five different document classes from three domains: psychiatric evaluations, radiology reports, and discharge summaries (Corpora 2.1-2.3) in the clinical domain; patent documents (Corpus 2.4) in the intellectual property (IP) domain, and environmental scientific articles (Corpus 2.5) from the scientific domain. For each corpus I manually created or found an ontology of distinct section types. As discussed in chapter 2, I conducted five annotation studies (one for each corpus) in a double-blinded manner, and calculated inter-annotator agreements resulting in a Cohen's κ of 0.90, 0.98, 0.94, 0.92, and 0.90 for each corpus, respectively. These agreement values are considered "perfect" agreement[Artstein and Poesio, 2008]. The ground truth data was only used for evaluating the approach. In the following sections I describe each corpus briefly (repeated from chapter 2 for ease of reference) followed by the challenges in section type discovery. Table 4.1 shows a summary of these corpora and their corresponding section and word statistics.

4.2.1 Corpus 1: Psychiatric Evaluations

Psychiatric evaluations consist of long-form unstructured text. They are the end product of an assessment in which a psychiatrist summarizes the information they have gathered, integrating the patient history, evaluation, diagnosis, and suggested treatments or future

steps [Groth-Marnat, 2009, Goldfinger and Pomerantz, 2013]. Although there is no strict format, there are general guidelines for writing these reports, typically structured as an ordered list of headed sections [Association, 2006].

As discussed in detail in chapter 2, I used a corpus of psychiatric evaluations and a corresponding ontology of section types previously collected and developed by Banisakher et al. [2018a]. The corpus contains 150 publicly available psychiatric evaluations collected by crawling the web and querying custom search engines. The reports in the corpus were anonymized samples of either real or synthetic psychiatric evaluations written and published for educational purposes. Each document is complete, and adheres to the general writing guidelines for psychiatric evaluations discussed in prior work [Banisakher et al., 2018a]. Table 4.2 (adapted from Table 2.1 for ease of reference) lists the main section types in their usual order of appearance as well as how often they appear in my corpus and their relevant statistics.

4.2.2 Corpus 2: Radiology Reports

A radiology report is a summary of a radiology scan such as an X-Ray or an MRI, where a radiologist communicates findings and an analysis of the results [of Radiology, 2019, Pool and Siemienowicz, 2019]. Similar to the previous two clinical document classes, radiologists are typically trained to follow a general guideline. This is not a strict format, as reports vary in their section structure and content based on the procedure performed, the patient's specific case, and the radiologist's and medical institution's writing styles.

As discussed in detail in chapter 2, I randomly extracted 423 radiology reports from MIMIC-III that were complete and adhered to the general radiology writing guidelines [of Radiology, 2019, Pool and Siemienowicz, 2019]. These reports covered a variety of procedures and scan types, including X-Rays, MRIs, and ultrasound. I used the ontology

#	Section	# Words	# Sents.	Sent. Length	% Present
	GENERAL PATIENT INFO				
1	IDENTIFYING DATA	12	2	6	100
2	CHIEF COMPLAINT	27	3	9	100
	MEDICAL HISTORY				
3	HIST. OF PRSNT. ILLNSS.	232	29	8	95
4	PSYCHIATRIC HISTORY	85	8	11	82
5	SUBST. ABUSE HIST.	98	10	10	88
6	REVIEW OF SYMPTOMS	150	19	8	96
7	SURGERIES	28	3	7	33
8	ALLERGIES	4	2	2	98
9	CURRENT MEDICATIONS	40	9	4	100
	FAMILY HISTORY				
10	BIRTH AND DVLP. HIST.	59	5	10	31
11	ABUSE HIST. / TRAUMA	110	9	12	79
12	FAMILY PSYCH. HIST.	44	5	9	73
13	FAMILY MED. HISTORY	48	7	7	92
14	SOCIAL HISTORY	80	7	11	76
15	PREGNANCY	29	3	8	47
16	SPIRITUAL BELIEFS	12	2	5	24
17	EDUCATION	32	3	8	68
18	EMPLOYMENT	31	3	9	79
19	LEGAL	10	1	5	20
	MENTAL STATUS				
20	MENTAL STATUS	155	18	9	95
21	STRENGTHS AND SUPPORTS	8	1	8	71
	TREATMENT				
22	FORMULATION	35	4	8	62
23	DIAGNOSES	63	12	5	100
24	PROGNOSIS	8	2	3	74
25	TREATMENT PLAN	121	12	10	100

Table 4.2: Section ontology and relevant statistics for Corpus 1: Psychiatric Evaluation Reports.

of section types presented in [Tepper et al., 2012]. Table 4.6 lists the main section types in their usual order of appearance as well as how often they occur in my corpus and their relevant statistics.

#	Section	# Words	# Sents.	Sent. Length	% Present
	CLINICAL INFORMATION				
1	CLINICAL HISTORY	80	8	10	100
	EXAM DETAILS				
2	EXAM	16	2	8	100
3	COMPARISON	16	2	8	86
4	CONTRAST	14	2	7	14
5	PROCEDURE	12	2	6	100
	FINDINGS				
6	FINDINGS	192	24	8	100
	IMPRESSION				
7	IMPRESSION	133	19	7	100
8	ATTENDING STATEMENT	-	-	-	0

Table 4.3: Section ontology and relevant corpus statistics for Corpus 2: Radiology Reports.

4.2.3 Corpus 3: Discharge Summaries

A discharge summary is the final documentation of a hospital stay. These reports summarize the course of hospital treatment by listing the various events during hospitalization [Horwitz et al., 2013]. Similar to psychiatric evaluations, discharge summaries are governed by general writing guidelines that suggest the information that should be included. In practice, different hospital networks and even different medical professionals within the same hospital often write these reports differently, tailoring them to specific patient cases.

As discussed in detail in chapter 2, and similar to radiology reports, I randomly extracted 150 discharge summaries from the MIMIC-III database [Johnson et al., 2016]. I selected summaries that were complete and that adhere to the general clinical note writing guidelines. As with all MIMIC-III data, the summaries are anonymized. I used the ontology of section types presented in [Tepper et al., 2012]. Table 4.4 lists the main section

#	Section	# Words	# Sents.	Sent. Length	% Present
	GENERAL PATIENT INFO				
1	ADMIT DATE	3	1	3	100
2	DISCHARGE DATE	3	1	3	100
3	SERVICE	4	2	2	100
	PROVIDER INFO				
4	ATTENDING	2	1	2	82
5	ADMIT PHYSICIAN	2	1	2	100
6	DISCHARGE PHYSICIAN	2	1	2	100
	COND. BEFORE ADMISSION				
7	ADMISSION DIAGNOSES	96	12	8	100
8	HISTORY	135	15	9	76
9	MEDICATIONS	55	11	5	100
10	REASON FOR ADMISSION	162	18	9	100
	COND. AT DISCHARGE				
11	CONDITION	4	2	2	100
12	DISPOSITION	2	1	2	34
13	DISCHARGE DIAGNOSES	144	18	8	89
14	OTHER DIAGNOSES	-	-	-	0
15	PHYSICAL EXAM ON DISCH.	45	9	5	40
	MEDICAL HISTORY				
16	ALLERGIES	12	3	4	100
17	FAMILY HISTORY	81	9	9	43
18	GYNECOLOGICAL HISTORY	-	-	-	0
19	PAST MEDICAL HISTORY	144	16	9	100
20	PAST SURGICAL HISTORY	32	4	8	100
21	SOCIAL HISTORY	84	7	12	37
	HOSPITAL COURSE				
22	CONSULTATION	88	11	8	6
23	HOSPITAL COURSE	168	14	12	85
24	PHYSICAL	66	11	6	28
25	PROCEDURES	15	5	3	65
26	STUDIES	-	-	-	0
	DISCHARGE INSTRUCTIONS				
27	FOLLOW UP	-	-	-	0
28	DIAGNOSTIC STUDIES REC'D	-	-	-	0
29	DISCHARGE INSTRUCTIONS	408	34	12	100
30	DISCHARGE MEDICATIONS	72	12	6	100

Table 4.4: Section ontology and relevant statistics for Corpus 3: Discharge Summaries.

types in their usual order of appearance as well as how often they occur in my corpus and their relevant statistics.

4.2.4 Corpus 4: Patent Documents

Patent documents are the result of a successful patent application. Many of a patent's sections are mandatory, e.g., the claims section [WIPO, 2007]. Similarly, the description section in these documents is further composed of subsections, some of which are mandatory, while others are optional and depend on the authors' preferences as well as the patent's technical topics. In their work on patent section segmentation, Brügmann et al. [2015] outlined the structure of the description section in a patent document into five mandatory and two optional segments.

#	Section	# Words	# Sents.	Sent. Length	% Present
1	TECHNICAL FIELD	85	3	8	100
2	BACKGROUND ART	267	57	11	100
3	SUMMARY OF THE INVENTION	1,286	89	10	100
4	DESCRIPTION OF DRAWINGS	975	19	8	100
5	PREFERRED EMBODIMENTS	4,106	208	7	100
6	INDUSTRIAL APPLICABILITY	2,731	96	2	31
7	EXAMPLES	1,258	82	4	14

Table 4.5: Section ontology and relevant statics for Corpus 4: U.S. Patent Documents.

For this work (as discussed in detail in chapter 2) I focus on the description section of patent documents and refer to those as patent documents in my discussion throughout this paper. I randomly collected 464 U.S. patent documents using the PATENTSCOPE database [WIPO, 2019] provided by the World Intellectual Property Organization (WIPO). The documents spanned the period between 1954 and 2010. I then extracted the description sections from the original patent documents to construct my corpus. Finally, I used the ontology of section types presented in [Brügmann et al., 2015]. Table 4.5 lists the main section types in their usual order of appearance as well as how often they occur in my corpus and their relevant statistics.

#	Section	# Words	# Sent.	Sent. Length	% Present
	INTRODUCTION				
1	BACKGROUND	800	35	23	100
2	PROBLEM	400	19	21	100
3	METHOD	1,413	53	27	100
4	RESULT	1,925	84	23	100
	RELATED WORK				
5	CONNECTION	356	21	17	100
6	DIFFERENCE	281	14	20	100
7	FUTURE WORK	350	20	18	40
8	CONCLUSION	205	10	21	100

Table 4.6: Section ontology and relevant statics for Corpus 6: Environmental Scientific Articles. All columns represent averages. The last three rows are the max, average, and min of averages.

4.2.5 Corpus 5: Environmental Scientific Articles

As discussed in chapter 2, this corpus was the result of an interdisciplinary collaborative project between computer scientists (including myself and other colleagues at the School of Computing and Information Sciences) and environmental scientists at Florida International University's Earth and Environment department. To the best of our knowledge there was no corpus of scientific articles annotated with ENVO concepts, so we created our own. We collected a total of 19 articles (90,074 total words) using four search queries that were created by three domain experts (two PhD students and a professor of Hydrology). Our domain experts ran the queries through Google Scholar and examined from the several hundred results returned, identifying the top four or five most relevant articles for each query. Importantly, several of the articles were not ranked near the top of Google's results, and were rather found many pages deep. Table 4.6 lists the main section types in their usual order of appearance as well as how often they occur in my corpus and their relevant statistics.

4.2.6 Challenges in Section Type Discovery

There are several challenges in discovering section types within a given document class. First, section headings are varied and ambiguous [Li et al., 2010, Banisakher et al., 2018a]. Using psychiatric reports as an example, a section labeled *IDENTIFICATION OF PATIENT* by one psychiatrist might be labeled *REFERRAL DATA* or *IDENTIFYING INFORMATION* by another. Second, some sections are included inside others; for example, the section *MEDICAL HISTORY* might include *REVIEW OF SYMPTOMS* and *PSYCHIATRIC HISTORY* subsections, while the section *FAMILY HISTORY* might include a subsection addressing *PREGNANCY*. Like top-level sections, subsections can either be explicitly labeled (heading present) or just implicit (heading omitted). Third, the section ordering can differ between reports, again, depending on the psychiatrist. And fourth, sections may be omitted, especially when that information in not relevant to the patient in question. For example a report regarding a male patient would likely not contain a *PREGNANCY* section [Banisakher et al., 2018a]. These challenges apply equally to many other types of clinical reports, including the discharge summaries and radiology reports used in my study.

Some document classes have stricter expectations about section structure than others. For example, while patent documents are more uniformly structured than clinical documents, they still suffer from inconsistencies between different authors, and especially between different countries [Diallo and Lupu, 2017]. In an effort to minimize these inconsistencies and to increase interoperability of patent analysis and discovery systems, WIPO outlined writing guidelines for patent documents in its patent drafting manual [WIPO, 2007]. Even so, the manual itself discusses and accepts the possibility of different formatting and structuring of the sections of patent documents. Thus, the challenges outlined above apply, but to a lesser degree, to U.S. patent documents. Scientific articles

are similarly ambiguous in their section structure, as many journals and conferences have their own required structure that differ from each other.

4.3 Task Definition

Given a corpus of documents from a single document class (e.g., psychiatric reports), I aim to identify a section structure that reflects the underlying statistics of the corpus. That is, a solution must identify a distinct list and general order of section types regardless of the section labels found within the documents. For example, a section originally labeled as *IDENTIFYING DATA* by one psychiatrist and *IDENTIFICATION OF PATIENT* by another, in two different reports, must be identified as a single distinct section type.

4.4 Approach

I treat this task as an unsupervised clustering problem that can be appropriately tackled by Bayesian model merging [Stolcke and Omohundro, 1994]. My approach is inspired by the *Analogical Story Merging* (ASM) approach which applied model merging to clustering events from natural language text, as introduced by Finlayson [2016]. In that prior work, events correspond to model states, and deriving the clustering involves four steps: (1) creating an initial model incorporating the sequence of events in each document in the corpus, (2) defining a merge operation over the events, (3) defining a prior over the models created, and (4) searching the merge space. The event clustering task defined there is analogous to my section discovery problem, where events are replaced by sections.

In this approach to section discovery, I adapted and extended the ASM steps. Given each corpus, I built an initial HMM-like model M_0, where each document is represented as a linear chain of states, with each state corresponding to a section of unknown type in the same order as found in the document. For example, a document containing ten

sections will be represented with a chain of ten states. Thus for the 150 reports in the psychiatric reports corpus, for example, I started with 150 linear branches composed of states that represented the sections. The model also incorporates single start and end states that link to all the first and last states of each of the linear branches, respectively. The goal in my approach is to iteratively merge similar section states to generate a succession of models M_i, seeking to maximizing the posterior probability $P(M_i|D)$, the probability of a model given the data. Figure 4.1 shows a toy example of this approach. I discuss the details of that example later in this section. The next three steps are non-trivial, and thus I discuss them separately as follows.

4.4.1 The Merge Operation

The merge operation merges two states in one model M_i to generate a new model M_{i+1}. The states' content are represented as bigram models of their corresponding section(s)' free text. A merged state's emission and transition probabilities are obtained from the weighted sum of their parent states, thus modeling the order of section types. I added two restrictions on candidate merged models. First, no cycles are allowed in a merged model, which maintains a directed order of sections and disallows repeated section types in a single linear chain. Second, I merge only sections with section-to-document-size ratios with one standard deviation of each other. This ratio is the number of tokens in a section to the number of tokens in the document. The intuition of this restriction is that, given a document class, I have a general expectation on the size of a specific section type relative to its document size. In scientific articles, for example, an introduction section in an 8-page long scientific article is typically about a page long while it would be two to three pages long in a 30- or 40-page article. Placing this as a restriction on the model

rather than a weight can be seen as favoring precision over recall as I favor models with states containing more similar sections.

After the search converges to a model with maximum probability, I obtain the most likely label (header) for each state by computing a majority vote over the headers of the sections merged into that state. Thus my approach can be further used to identify actual section headers for a document class in a given corpus.

4.4.2 Defining the Prior Over Linear Models

The posterior probability guides the search in model merging, but a prior probability is needed to compute it. A prior probability distribution represents my initial belief over the size and structure of the models. I first assume a normal distribution over the number of sections present in the model. For instance, in clinical notes this follows intuition in that: (1) patients share similar characteristics overall, (2) most patients treated fall under an umbrella of a small subset of medical issues (e.g., depression, anxiety, and ADHD in mental health), and (3) most medical professionals share a similar report writing and structuring style given that they follow the general medical writing guidelines. A similar intuition follows for patent documents as well. I further verified my intuition through examination of the corpora.

Additionally, I disallow models that merge sections with dissimilar content. This is achieved by setting a similarity threshold and setting the prior probability to zero if a state merges two sections with content less than a threshold T. The similarity function is defined in the next section. The resulting prior $P(M)$ is thus formulated as follows:

$$P(M) = N(\mu, \sigma^2) \prod_i G(S_i) \tag{4.1}$$

$$G(S_i) = \begin{cases} 1 & \forall s_j, s_k \in S_i,\ Sim(s_j, s_k) > T \\ 0 & \text{otherwise} \end{cases} \tag{4.2}$$

In Equation 4.1, the Normal distribution $N(\mu, \sigma^2)$ is multiplied by the product of a threshold function for each state in the model M. S_i is the i^{th} state in M. In Equation 4.2, s_j and s_k are section contents (i.e., text blocks) that have been merged into state S_i, Sim is the similarity function, and T is a similarity threshold. Following a parameter sweep using grid search, and for a strict similarity threshold T is set to be 1.5 standard deviation from the the mean similarity of all candidate sections to be merged, and therefore is tuned to the data.

4.4.3 The Similarity Function

The similarity function Sim takes the content of two candidate sections s_j and s_k (or collection of sections in the case of merged states), and computes the cosine similarity of their vector representations. These vector representations are computed from a set of extracted features that are used to model a section's free text content. I extracted the following sets of lexical, positional, and semantic features: (1) unigrams and bigrams; (2) the top three key terms per section as indicated by *tf-idf* [Church and Gale, 1999]; (3) the section position relative to its document; (4) the length of the section in tokens; (5) extracted named entities, their types, and counts; and (6) the Wu-Palmer similiarity score [Wu and Palmer, 1994]. Additionally, although not shown in Equation 4.2, if the headers of all sections in the merged states are exactly the same, $G(S_i)$ is set to 1.

4.4.4 Searching the Merge Space

As discussed earlier, the posterior probability $P(M_i|D)$ drives the search, as maximizing it will result in a generalizable model that fits the given data. Prior work with model merging used greedy, best-first search [Stolcke and Omohundro, 1994, Finlayson, 2016] because of the size of the merge space, and I follow the same approach. As in prior approaches, I do not compute $P(M_i|D)$ directly, but rather seek to compute $P(M_i)P(D|M_i)$, which is proportional to it. Further, because computing $P(D|M_i)$ is costly, I estimate it following approximations described by Stolcke and Omohundro [1993] that compute heuristics for finding a maximum a posteriori probability (MAP).

Figure 4.1 shows a toy example of the section merging approach over two small psychiatric reports. Each composed of four sections: *IDENTIFYING DATA*, *REVIEW OF SYMPTOMS*, *PREGNANCY*, and *TREATMENT* in the first report, and *PATIENT*, *EDUCATION*, *MEDICAL HISTORY*, and *PLAN FORMULATION* in the second. In the first model M_0, the model is initialized such that each report is an HMM-like linear chain of states which in turn correspond to sections in their original order of appearance. The figure shows a series of merges leading to the model that maximizes the posterior under the described parameters. In M_1 *IDENTIFYING DATA* and *PATIENT* are merged into a single state, the transitions are inherited as well as the section headers and content. Similarly this is shown for *REVIEW OF SYMPTOMS* and *MEDICAL HISTORY* in M_2, and for the last two sections in each report in M_3. The final model M_3 can generate not only the two input reports (i.e., two distinct section sequences), but an additional two section sequences that alternatively include or exclude both states 3 and 6. Thus the model has generalized beyond the two input examples. Most importantly, I can obtain a distinct list of section types and ordering for the input data from the generalizing final model.

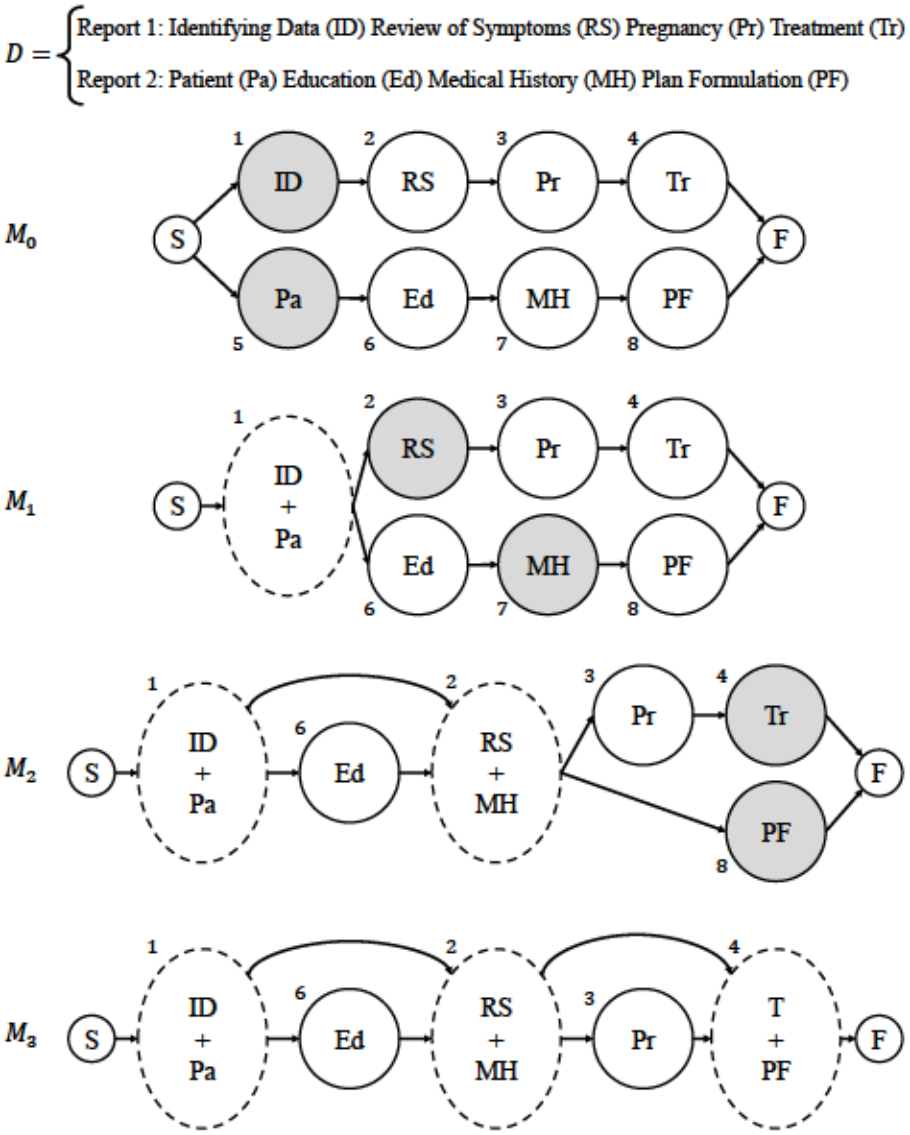


Figure 4.1: Example of the approach on two toy psychiatric reports D. Each report comprises four sections. States are represented by circles and transitions by arrows. Abbreviated section headings inside states indicate that the state can emit that section content. Shaded states are merged into the dashed state in the next step.

4.5 Evaluation Methods and Metrics

My approach aims to identify a section structure that reflects the underlying statistics of the corpus. Thus the output model results in a (1) a set of proposed section types and (2) a finite state machine the structure of which models the order for those sections. I evaluate

these two results separately. As discussed in §4.2 I annotated each corpus with ontologies of section types, and this ground truth (which was not provided to my models) was used for the evaluation. Additionally, as discussed in §4.4.3, section types are given a preferred label (header) following a majority vote of merged sections in each state.

To evaluate section type discovery (i.e., identifying the set of possible section types) I treated it as a document clustering task, with each section a separate document. I compared the models against four document clustering baselines: Non-negative Matrix Factorization (NMF) using *tf-idf* vectors; K-means over *tf-idf* vectors; K-means over latent Dirichlet allocation (LDA) topic vectors; and k-means over word2vec embeddings. Similar to the experimental setup in [Hosseini-Asl and Zurada, 2014] for NMF and setup in [Xie and Xing, 2013] for K-means, I provided these algorithms with the correct number of clusters k for each corpus. This is not possible in the general case and therefore I would expect worse results than shown here. Additionally, to maximize the performance of the baselines, I disallowed clustering of sections within the same document, as sections from the same document will often be grouped because they share similar topic and term distributions. I evaluated the clustering using two metrics: the chance-adjusted Rand index (Rand) to evaluate the overall clustering quality, and the F_1 measure to each section type independently.

To evaluate the section ordering, I computed an F_1 measure for each section type that compared the proportions of succeeded sections in the model to that in the ground truth annotations.

4.6 Results and Discussion

I evaluated my models and baselines over the five corpora presented in §4.2. My approach significantly outperformed all four baselines when discovering section types, with

#	Section	% Present	P	R	F_1
	GENERAL PATIENT INFO				
1	IDENTIFYING DATA	100	0.96	0.96	**0.96**
2	CHIEF COMPLAINT	100	0.94	0.92	**0.93**
	MDEICAL HISTORY				
3	HIST. OF PRSNT. ILLNSS.	95	0.96	0.94	**0.95**
4	PSYCHIATRIC HISTORY	82	0.89	0.89	**0.89**
5	SUBST. ABUSE HIST.	88	0.90	0.88	**0.89**
6	REVIEW OF SYMPTOMS	96	0.95	0.95	**0.95**
7	SURGERIES	33	0.80	0.71	**0.75**
8	ALLERGIES	98	0.92	0.94	**0.93**
9	CURRENT MEDICATIONS	100	0.94	0.90	**0.92**
	FAMILY HISTORY				
10	BIRTH AND DVLP. HIST.	31	0.73	0.67	**0.70**
11	ABUSE HIST. / TRAUMA	79	0.90	0.86	**0.88**
12	FAMILY PSYCH. HIST.	73	0.92	0.90	**0.91**
13	FAMILY MED. HISTORY	92	0.95	0.93	**0.94**
14	SOCIAL HISTORY	76	0.90	0.88	**0.89**
15	PREGNANCY	47	0.75	0.69	**0.72**
16	SPIRITUAL BELIEFS	24	0.72	0.68	**0.70**
17	EDUCATION	68	0.83	0.77	**0.80**
18	EMPLOYMENT	79	0.88	0.86	**0.87**
19	LEGAL	20	0.74	0.65	**0.69**
	MENTAL STATUS				
20	MENTAL STATUS	95	0.95	0.91	**0.93**
21	STRENGTHS AND SUPPORTS	71	0.85	0.83	**0.84**
	TREATMENT				
22	FORMULATION	62	0.83	0.79	**0.81**
23	DIAGNOSES	100	0.97	0.95	**0.96**
24	PROGNOSIS	74	0.85	0.83	**0.84**
25	TREATMENT PLAN	100	0.95	0.89	**0.92**
		Average	**0.88**	**0.85**	**0.86**

Table 4.7: Section ontology and merging results for Corpus 1: Psychiatric Evaluations. Column 3 shows the percentage of documents that contain that section type. Columns 4-6 show the precision, recall, and F_1 scores for section merging.

improvements of 68%, 77%, 67%, 58%, and 90% respectively for each corpus, over the best performing baseline (Word2Vec+K-means). Table 4.11 shows these results. The chance-adjusted Rand index is analogous to accuracy, which suggests that most states

in my model had a relatively small number of dissimilar sections. I also performed five feature combination experiments. Adding section positional and length features had a significant positive impact on the models' performance, achieving above 80% on the Rand index, while semantic features helped, but by a lower factor. Further, I tested the impact of using exact header matching (§4.4.1) by relaxing that rule. Under that condition, my models only lost 3% performance on average between all the corpora which shows that my model can be effective even when a corpus contains no section header information at all.

My approach also significantly outperformed all baselines even when only using lexical features. Careful inspection of the baseline results revealed that sections were grouped based on topics—an expected result. For example, LDA-K-means created a cluster for ADHD in the psychiatric evaluations corpus and thus grouped sections regardless of type into that cluster. This confirms that topical models and classical document clustering techniques are inefficient in discriminating "types" of text rather than "topics".

#	Section	% Present	P	R	F_1
	CLINICAL INFORMATION				
1	CLINICAL HISTORY	100	0.96	0.92	**0.94**
	EXAM DETAILS				
2	EXAM	100	0.92	0.92	**0.92**
3	COMPARISON	86	0.84	0.76	**0.80**
4	CONTRAST	14	0.71	0.65	**0.68**
5	PROCEDURE	100	0.91	0.89	**0.90**
	FINDINGS				
6	FINDINGS	100	0.89	0.83	**0.86**
	IMPRESSION				
7	IMPRESSION	100	0.90	0.84	**0.87**
8	ATTENDING STATEMENT	0	-	-	-
		Average	**0.88**	**0.83**	**0.85**

Table 4.8: Section ontology and merging results for Corpus 2: Radiology Reports. Columns are organized as in Table 4.5.

#	Section	% Present	P	R	F_1
	GENERAL PATIENT INFO				
1	ADMIT	100	0.95	0.91	**0.93**
2	DISCHARGE	100	0.95	0.91	**0.93**
3	SERVICE	100	0.91	0.91	**0.91**
	PROVIDER INFO				
4	ATTENDING	82	0.82	0.82	**0.82**
5	ADMITTING	100	0.75	0.77	**0.76**
6	DISCHARGING	100	0.80	0.74	**0.77**
	COND. BEFORE ADMISSION				
7	ADMISSION DIAGNOSES	100	0.90	0.82	**0.86**
8	HISTORY	76	0.90	0.81	**0.85**
9	MEDICATIONS	100	0.83	0.81	**0.82**
10	REASON FOR ADMISSION	100	0.80	0.78	**0.79**
	COND. AT DISCHARGE				
11	CONDITION	100	0.78	0.76	**0.77**
12	DISPOSITION	34	0.65	0.71	**0.68**
13	DISCHARGE DIAGNOSES	89	0.85	0.83	**0.84**
14	OTHER DIAGNOSES	0	-	-	-
15	PHYSICAL EXAM ON DISCH.	40	0.84	0.84	**0.84**
	MEDICAL HISTORY				
16	ALLERGIES	100	0.85	0.79	**0.82**
17	FAMILY HISTORY	43	0.82	0.80	**0.81**
18	GYNECOLOGICAL HISTORY	0	-	-	-
19	PAST MEDICAL HISTORY	100	0.81	0.83	**0.82**
20	PAST SURGICAL HISTORY	100	0.88	0.84	**0.86**
21	SOCIAL HISTORY	37	0.77	0.75	**0.76**
	HOSPITAL COURSE				
22	CONSULTATION	6	0.64	0.64	**0.64**
23	HOSPITAL COURSE	85	0.89	0.85	**0.87**
24	PHYSICAL	28	0.79	0.77	**0.78**
25	PROCEDURES	65	0.84	0.82	**0.83**
26	STUDIES	0	-	-	-
	DISCHARGE INSTRUCTIONS				
27	FOLLOW UP	0	-	-	-
28	DIAGNOSTIC STUDIES REC'D	0	-	-	-
29	DISCHARGE INSTRUCTIONS	100	0.92	0.92	**0.92**
30	DISCHARGE MEDICATIONS	100	0.91	0.93	**0.91**
		Average	**0.83**	**0.81**	**0.82**

Table 4.9: Section ontology for the discharge summary corpus and merging results for Corpus 3: Discharge Summaries. Columns are organized as in Table 4.2.

I also evaluated the models' performance on each section type individually using precision, recall, and F_1 (Tables 4.5-4.5). Compared against ground truth, my models performed significantly better for sections with highly distinctive content (compared with

#	Section	% Present	P	R	F_1
1	TECHNICAL FIELD	100	0.87	0.83	**0.85**
2	BACKGROUND ART	100	0.93	0.89	**0.91**
3	SUMM. OF THE INVENTION	100	0.94	0.92	**0.93**
4	DESC. OF DRAWINGS	100	0.96	0.94	**0.95**
5	PREF. EMBODIMENTS	100	0.96	0.96	**0.96**
6	INDUST. APPLICABILITY	41	0.85	0.72	**0.78**
7	EXAMPLES	16	0.80	0.71	**0.75**
		Average	**0.90**	**0.85**	**0.88**

Table 4.10: Section ontology and merging results for Corpus 4: U.S. Patent Documents. Columns are organized as in Table 4.2.

other sections): e.g., *DIAGNOSIS* in psychiatric evaluations, *DISCHARGE INSTRUCTIONS* in discharge summaries, *EXAM* in radiology reports, and *DESCRIPTION OF DRAWINGS* in patent documents. Similarly, my models performed better in beginning and ending sections in general (e.g., *TREATMENT PLAN* in psychiatric reports, and *DISCHARGE MEDICATIONS* in discharge summaries). I suspect that this is because those sections typically display minimal variability in position. On average the precision was higher than recall reflecting my explicit choice to bias toward precision (c.f., §4.4.2).

I computed the the confusion matrix counting the correct (TP), incorrect (FP), and missing (FN) forward transitions for each section type in comparison with the ground truth annotation, and then used these to compute the precision (P), recall (R), and F_1 scores. Average P, R and F_1 were then obtained by weighing the scores by the number of sections for each section type (Table 4.12). The model achieved high performance for all four corpora, while again performing best on the patent corpus and achieving a 0.95 weighted F_1 score. This can be partially attributed to the fact that patent documents have a more uniform section structure compared to the other document classes (c.f. §4.2.6).

Overall, my approach performed best on patents, followed by the psychiatric and radiology corpora, and worst on discharge summaries. A further analysis of the results and data led us to characterize my approach in four ways as the resulting models favor

Algorithm	Psychiatric Reports.	Dischcharge Summaries	Radiology Reports.	Patent Documents	Scientific Articles.
NMF	0.40	0.39	0.43	0.42	0.35
tf-idf+K-means	0.47	0.40	0.52	0.49	0.39
LDA+K-means	0.50	0.41	0.54	0.51	0.40
Word2Vec+K-means	0.53	0.48	0.52	0.59	0.44
My Approach					
Lexical Only	0.74	0.72	0.77	0.81	0.75
Semantic+Lexical	0.80	0.75	0.78	0.83	0.80
Positional+Lexical	0.84	0.80	0.81	0.85	0.81
All_Features_No Header	0.87	0.83	0.85	0.87	0.82
Section Merging_ALL	**0.89**	**0.85**	**0.87**	**0.93**	**0.84**

Table 4.11: Rand results for section type discovery of the baseline algorithms and my approach. My approach's results are also shown for different combinations of features

document classes with (1) higher variance in section content distinctiveness, (2) lower average section-to-document ratio, (3) higher average word-to-section ratio, and (4) more uniform ordering

4.7 Related Work

To the best of my knowledge there have not been any prior attempts to automatically discover section types. Bayesian model merging [Stolcke and Omohundro, 1993] has been adapted for various tasks including induction of probabilistic programming languages [Hwang et al., 2011], induction of stochastic grammars for page classification [Frasconi et al., 2001], lexical categorization and word grouping [Brent and Cartwright, 1996], and inference of story grammars [Finlayson, 2016]. Bayesian model merging has also been extended for clustering sequence data [Li and Biswas, 1999]. There, model merging is used to search for the HMM topology that best represents sequence data; clustering is done for an entire sequence, rather than parts of a sequence as in my approach.

Several studies have demonstrated approaches for the identification and segmentation of documents such as clinical notes and scientific articles. Most approaches have focused

Corpus	P	R	F_1
Psychiatric Evaluations	0.87	0.92	**0.89**
Radiology Reports	0.87	0.91	**0.89**
Discharge Summaries	0.83	0.88	**0.85**
US Patents	0.94	0.96	**0.95**
Scientific Articles	0.82	0.86	**0.84**

Table 4.12: Section Ordering results for section type discovery.

on the identification of section headers rather than content. Denny et al. [2009b] developed the SecTag algorithm which uses terminology-based rules, and naive Bayes scoring methods to identify clinical note section headers. Using machine learning methods, Apostolova et al. [2009] and Tepper et al. [2012] demonstrated supervised approaches for detecting section headers and boundaries but showed low adaptability when faced with various clinical note documents [Ganesan and Subotin, 2014]. More recently, Li et al. [2010] and Banisakher et al. [2018a] demonstrated HMM-based approaches to learning the section structure and ordering in clinical note documents.

Finally, the output model in my approach contains a distinct list of section types (i.e., an ontology). In ontology learning and extraction however, there has been no efforts to learn documents' section structure. Rather most approaches focused on learning semantic concepts and relations [Dou et al., 2015], some of which used the document structure as input instead [Rimale et al., 2016, Diallo and Lupu, 2017].

4.8 Limitations and Future Work

My approach finds a general model of section types and their orders, given corpus of documents from the same class. The evaluation suggest four limitations of my current approach which point toward next steps. First, although I tested on four corpora, an evaluation over a more diverse set of larger corpora would help better characterize the approach.

Second, although my approach requires no training, the prior still requires knowledge that the document class is governed by some at least rough writing guidelines. I hypothesize, however, that a normal distribution with default parameters can be assumed, regardless of the document class, without a significant performance loss. Third, my approach does not use discourse features which likely would be beneficial in certain cases. Discourse features would be a natural extension of the positional and length features that I use already.

Finally, my approach assumed knowledge of the section boundaries, which is not guaranteed in practice. A fully general approach would require a companion system able to segment the sections. I did not investigate this as numerous approaches have successfully tackled segmentation for both general and specific classes of documents. Modifying my approach to operate on the sentence level would conceivably be possible, which potentially could transform the approach into a complete section structure extraction system.

CHAPTER 5

ONTOLOGICAL SEMANTIC SEARCH

5.1 Survey of Academic Search Approaches

Academic search is the process of using specialized search engines or bibliographic databases to find academic articles, often involving highly specific academic concepts. It is more specialized than general web or database search, and is a critical first step in any research project. Academic search has become increasing challenging in the past few decades as the academic literature has grown exponentially, with a proliferation of new venues and subfields which may contain relevant material and yet are unknown to even well-read researchers or scholars. This review presents state-of-the-art approaches to academic search, specifically focusing on *semantic* academic search. Semantic search contrasts with traditional keyword-based search by attempting to analyze conceptual meaning behind user queries and match them to concepts in target documents. Evidence suggests that this highly informed search has significantly better performance for academic search, and represents one of the most promising future directions.

5.1.1 Introduction

Academic search is the process in which an individual uses specialized search engines or bibliographic databases to query and search through a database of academic articles for relevant scientific literature. Academic search is the first step in any research project. Be it a novice or an experienced researcher, the major concerns accompanying academic search are: first, the efficiency of research methodologies; second, the comprehensiveness of research materials available; third, the volume of research undertaken; and fourth, the time needed to perform the desired research using digital media [Redfern, 2011]. Academic

search engines aim to address these four major concerns through various methods involving content inclusion and exclusion or content selection as well as search and retrieval approaches and user interface design.

Numerous studies have surveyed and evaluated general search engines with studies concerning keyword-based and semantic search approaches [Mangold, 2007, Yu et al., 2010, Wilson et al., 2010, Dou et al., 2015, Laddha et al., 2015, Klusch et al., 2016, Zhou et al., 2016]. However, relatively much less research has been conducted on academic search engines and the impact of different approaches on systems' retrieval output [Amolochitis, 2014, Rodrigues and Prates, 2016, Khabsa et al., 2016]. In this section, I discuss state-of-the-art in the field of academic search. I present the approaches, technologies, and tools developed over the last decade for searching the academic literature for both open and closed domains. Throughout my discussion, I group these technologies into two main approaches, namely, keyword-based search and semantic search while also discussing other novel approaches.

5.1.2 History

Search is a core problem in the theory of computation and computability. Numerous algorithms were developed over the past half century that paved the path to current sophisticated search systems. With the growth of digitized data, search problems became even more important and pressing. Similarly, with the growth of knowledge and published science in the form of scientific articles, academic search is evermore important [Amolochitis, 2014].

In 1945, the closest idea of an Internet was introduced by Bush [1945] where he proposed a future where individuals could store their books, articles, and communications as well as link them and allow them to be searched and retrieved. Garfield introduced

the idea of citation indices for scientific articles in 1955 [Garfield, 1955] and later followed that with the introduction of the Science Citation Index (SCI) [Garfield, 1964]. Decades later, the Institute for Science Information introduced the Web of Knowledge, a subscription-based scientific citation indexing service which is currently maintained by Clarivate Analytics (previously known as Thomson Reuters) [Analytics, 2016].

Web of Science (WoS) (then Web of Knowledge) was launched in 2002 as an academic search engine backed by a database that included high quality journals, patents, and proceedings. Its content included published work in the sciences, social sciences, arts, and humanities dating back to 1945. This was followed by the introduction of several relevant tools for cross search, cited reference search, text linking, and automatic citation engines [cla, 2018]. Google Scholar followed in 2004 as another academic search engine allowing users to search through a specialized database of scientific journals. After GS, numerous academic search engines were developed and made available such as Windows Live Academic Search (currently known as Microsoft Academic) [Harzing, 2016], Elsevier's Scopus [sco], and CiteSeer^X [Li et al., 2006]. Among many others, these academic search engines make up the standard method for academic search today utilizing different search, indexing, and ranking approaches and algorithms most of which I will discuss in this section.

Historically, two approaches have been proposed to conduct document and text search given a query: keyword-based search, and semantic search. Fundamentally, given a search query (that is, a text span an individual uses to find relevant or desired documents), keyword-based search seeks to retrieve what the individual "said" (i.e. word spans), while semantic search seeks to retrieve what the individual actually "wants" (i.e. concepts behind word spans). Following, I discuss these two approaches and their respective uses in academic search.

5.1.3 Keyword-based Search

Keyword-based search is the de facto approach for most text-based search engines [Laddha et al., 2015]. It relies on lexical forms (that is, words and collections of words) to determine whether a document is relevant to the search query or not. A core–classical– method for implementing this approach is the Term Frequency-Inverse Document Frequency (TF-IDF) Sparck Jones [1972], Jones [1973]. TF-IDF aims to find how important a word is in a collection of documents (i.e. a corpus) using numerical statistics. Keyword-based search engines typically use TF-IDF to index documents in their databases linking each document to its respective representative words. These search engines then use the terms in a user's query to search through the index and retrieve the documents with the highest TF-IDF weights corresponding to those terms [Tümer et al., 2009, Laddha et al., 2015].

Most search engines nowadays however, use more sophisticated keyword-based approaches for document indexing, retrieval, and ranking. Google Scholar,PubMed, Science Direct, and earlier versions of Microsoft Academic are prime examples of keyword-based academic search engines and searchable bibliographic databases. Although popular for scientific research and literature search, these examples have been shown to be insufficient for retrieving what users actually look for [Falagas et al., 2008, Boeker et al., 2013, Giustini and Boulos, 2013, Mangold, 2007]. This is relatable, as many of us have to learn the "art of search" in order to achieve efficient and accurate search. Researchers have discussed and consistently complain about these search engines in their writings, meetings, and over online discussion forms among others [Amolochitis, 2014, Laddha et al., 2015, Schoormann et al., 2018].

Keyword-based search systems have been improved recently by including several Natural Language Processing (NLP) tasks in preprocessing and search steps [Laddha et al., 2015]. Such tasks include word repetition, multiword, and named entity recognition. In

a sense, these search engines are no longer pure keyword-based systems. Although early approaches were successful and sufficient at the time, this approach diversion is due to the recognition that keyword-based search is inefficient and ineffective in today's vast knowledge base [Tümer et al., 2009]. Keyword-based search performs well for small databases and niche research topics, however with the growth of published research these databases are becoming larger than ever with Google Scholar estimated to contain over 160 million documents as an example. Additionally, keyword-based search can be considered advantageous due to its simple and computationally inexpensive nature.

Here, I highlight some of the well-known deficiencies typically inherited with keyword-based search [Zou et al., 2008, De Virgilio et al., 2012]:

- retrieval of an overwhelming number of results,
- rankings that do not precisely reflect true relevance,
- the omission of relevant results because they do not contain the idiosyncratic keywords of the query

5.1.4 Semantic Search

Semantics represent the underlying meaning behind words—that is, the true conceptual meaning [Cruse, 2004]. In linguistics, semantics consists mainly of two areas: logical semantics and lexical semantics. The former is concerned with the words presupposition, causality, sense, and reference, while the latter is concerned with the words' senses and their interrelations [Cruse and Cruse, 1986, Geeraerts, 2002, Mangold, 2007]. In academic search, semantic search uses concepts in the query and documents' text to drive document retrieval, indexing, and ranking. Semantic search often leverages domain-specific knowledge, typically encoded in ontologies, to help rank the relevance of documents relative to a search query [Mangold, 2007].

Semantic search is difficult, however, because the required process entails significant knowledge engineering as well as sophisticated natural language processing (NLP). Despite these problems, search engines today do boast high performance compared to prior decades precisely due to the inclusion of minor semantic knowledge in their search algorithms; Latent Semantic Indexing (LSI) [Deerwester et al., 1989], for example, has been used in various academic search engines. LSI uses synonyms and relationships between page headers, document titles, and content to assist ranking [Deerwester et al., 1990]. Nevertheless, we are still far from the full realization of true semantic search which uses deep semantic techniques fully integrated into back-end algorithms. For these reasons semantic search research is experiencing a rise in interest among various groups [Wu et al., 2015b].

Semantic search was proposed as a possible search solution over four decades ago, while academic semantic search was sought as a solution in the early 2000s after the rise of keyword academic search engines. Additionally, the introduction of the ideas of the Semantic Web which was proposed by Berners-Lee et al. [2001] and the Web Ontology Language (OWL) [Antoniou and Van Harmelen, 2004, McGuinness et al., 2004] pushed semantic academic search further by allowing ontologies (which represent the knowledge base for academic search systems) to be created and leveraged in a highly organized structure [blo, 2007]. Academic search systems have come a long way from the initial ideas; Currently, systems like Semantic Scholar, Microsoft Academic, Textpresso, and even Google Scholar integrate semantic search techniques in their backend algorithms in various degrees [Kearl et al., 2017].

Although advantageous to keyword-based search, semantic search has some disadvantages and limitations that must be taken into account when considering building systems that rely on such algorithms [De Virgilio et al., 2012]. As discussed earlier, the most significant limitation to semantic search is its theoretical complexity and difficulty. Another

limitation, although can be overcome with the advancement of other sectors in storage and communication technologies, is the time and space complexity semantic search entails especially when dealing with a large knowledge base. Finally, true semantic search may sometimes lead to results that are too specific (achieving its goal of favoring precision over recall)—users could be starved from being exposed to results that they did not intend to look for but that may be of aid to their overall search [Latard et al., 2017].

5.1.5 Academic Search Approaches

Academic search approaches, as discussed earlier, are mainly split into two domains, (1) keyword-based search and (2) semantic search. Due to its advantages centered around simplicity and effectivity when dealing with small datasets, keyword-based search was (and still is) the first choice for most search engines. Specifically, in the past decade, academic search was heavily dominated by this approach. Following, I first briefly discuss Google Scholar and its backbone algorithms as they represent the general approach to keyword-based search. I follow by discussing semantic search, document ranking approaches for document retrieval, as well as other approaches.

In 2004, Google Scholar launched, a web-based search engine focusing on the retrieval of academic articles. This search engine was well known for eliminating reliance on sorting by date or citation and instead retrieved articles based on relevance. For Google Scholar and many other search engines, relevance was–and, in large, still is–being determined by exact keyword matching, while heavily relying on citation counts as a ranking method [Tümer et al., 2009, Malve and Chawan, 2015]. This ranking approach was chosen to make for a quicker search for the most significant papers as opposed to the traditional bibliographic databases at the time (i.e. PubMed, Web of Science, and Scopus) [Martín-Martín et al., 2016]. Two years later, Microsoft released its own academic search

engine as a direct competitor to Google Scholar. Windows Live Academic Search was launched in 2006 and was later renamed twice in 2008 and late 2009 to Live Search Academic [Jacsó, 2008] and Microsoft Academic Search, respectively Martín-Martín et al. [2016]. Microsoft Academic Search enjoyed a steady growth since its launch, introducing helpful visualization tools [Orduña-Malea et al., 2014]. Nonetheless, Google Scholar remained to be the most favored tool due to its wide literature coverage, its focus on the academic citations and metrics, and its clean bibliographic knowledge base which linked and removed duplicated records.

Although Google Scholar has become an essential tool for researchers in academia and industry, it has its limitations. It makes use of the PageRank algorithm [Page et al., 1999a], adapted from Google's web search engine, to measure the relevance of a particular paper using exact-keyword matching and heavily relying on the documents' citation counts. When it comes to web search, the PageRank algorithm works on a graph created by treating all sites in the World Wide Web as nodes and their hyperlinks as edges. It indicates the importance of a web page by favoring older pages since new pages typically do not contain as many links [Al-Hattab, 2016]. PageRank works best when looking for standard papers in a certain field. However, it is not the best choice when trying to explore novel ideas and what some researchers might call "hidden gems" [Langville and Meyer, 2011].

Other than citation counts, Google Scholar also relies heavily on terms found in the document's title. Its methods for retrieval and ranking of relevant papers has also been shown to suffer from cases of academic search engine spam. By learning the approach for Google Scholar to rank the papers, users are able to exploit the algorithm to obtain a higher rankings for specific papers in relation to certain specified search queries [Beel and Gipp, 2010].

Semantic Search Approaches

Semantic search systems were introduced following the standards established by the World Wide Web Consortium (W3C) for the Semantic Web. In the Semantic Web, resources that have some type of relationships between each other need to have an explicit set of connections for effective navigation and discovery of new information. The goal for the creation of semantic web was to: (1) provide information that is well-defined and structured, (2) provide data that is readily interpretable by machines, and (3) facilitate the exchange and reuse of data. With the ever-increasing number of papers published each day, there is a need for a more effective, efficient, and specific search that focuses on the retrieval of the most relevant articles (that is, a focus on precision over recall). Following I highlight the various components of an academic search engine and their integration in semantic search approaches.

Document and Query Pre-processing

Pre-processing is a general and essential step in any data mining task which serves to organize and structure raw data or free-text. Document and query pre-processing involves common NLP tasks such as tokenization, sentence segmentation, and stop-word removal (i.e. removing language-specific common words such as “the”, ”that”, “a”). This step aims at preparing the free-text for higher syntactic and semantic NLP tasks:

Word normalization involves transforming words into their most basic form (i.e. roots or lemmas). In natural text, words are found in their inflectional forms (e.g. studying, systems, subjects). This presents a problem to search algorithms as each word can be inflected in several ways while its core meaning would be unchanged. Thus searching for inflected words would be redundant and ineffective. Stemming and lemmatization are two NLP techniques used for word normalization. Both achieve similar results; in that,

they allow systems to identify all words that share the same base or root. For example, when searching "study", all other forms must be matched including "studies" and "studying". The main difference between both techniques is that stemming is a crude heuristic provcess that reduces inflection by chopping off the ends of words, while lemmatiztion is an accurate inflection reduction process that makes use of vocabularies and morphological analysis of words. This step allows search engines to properly search documents for words found in a search queries.

Part-of-speech (POS) tagging is the process of assigning words to defined sets of grammatical categories or tags (e.g. nouns, pronouns, vervs, adjectives, adverbs). This process involves identifying word definitions as well as word context (i.e. surrounding words, sentences, or paragraphs). Several automatic taggers were proposed and are used by search engines. For example,Texpresso makes use of the brill tagger to attach syntactic categories to each tokenized word [Müller et al., 2004]. Brill tagger is a well-known and publicly available simple rule-based POS tagger which automatically acquires its rules and tags [Brill, 1992]. Used transformation-based learning, with rule-templates referring to neighboring words, POS tags, and chunk tags (up to a distance of 3 for words or POS tags, and 2 for chunk tags). Additionally, POS tagging is also considered a preparatory step for several semantic NLP tasks such as Named Entity Recogition and Word Sense Disabiguation. GeneView has also demonstrated the use of POS tagger as influential in determining the relationship between key concepts [Thomas et al., 2013].

Another syntactic pre-processing step is *Parsing*, which entails analyzing strings of words (i.e. sentences) for their constituents following a formal grammar and resulting in a parse tree. This step is important for analyzing referential phrases, sentences, and text chuncks. Several search engines including Textpresso and SemanticScholar make use of this technique in order to analyze both, search queries and documents.

Finally, recognizing the *document structure* for scientific articles is an important pre-processing task. A growing focus on identifying the physical and logical document structure of scientific articles is present among academic search engines. For example, the detection of section headings and abbreviation to long-form mappings allows academic search engines to better identify and search specific parts of the documents such as the titles, headings, or abstracts in isolation of the rest of the test. This step can also aid in the visualization of documents (e.g. retrieving abstracts or summaries) [Thomas et al., 2013].

Entity Recognition

Named Entity Recognition (NER) is a semantic task that concerns the identification and mapping of words to named entity categories that correspond to real-world objects such as persons, organizations, locations, and products. Semantic search engines such as GoPubMed [Doms and Schroeder, 2005] and GoWeb [Dietze and Schroeder, 2009], for example, utilize entity recognition techniques that can identify protein and gene names. These systems use a simple approach which implements the named entity recognition process by matching the term against a pre-defined synonym list [Hakenberg et al., 2007]. Some systems demonstrate the fusion of a large variety of named-entity recognition tools which can automatically annotate different entity classes in academic literature [Thomas et al., 2012]. In the biomedical domain, various tools are used to annotate genes, species, chemicals, histone modifications, protein-protein interactions (PPIs) and other entities. These tools use a range of methods spanning from sophisticated machine learning approaches such as conditional random fields (CRFs) and other deep learning methods to to simple sting and regular expression matching.

Word Sense Disambiguation (WSD) [Agirre, 2006] is a challenging task often used for NER as well as other higher semantic tasks. WSD involves mapping words to their true senses (i.e. meaning). Methods implementing WSD rely on extracting the con-

textual meaning of words using surrounding word groups, spans of sentences, and paragraphs[Navigli, 2009]. BabelNet is a multilingual semantic network [Navigli and Ponzetto, 2012], developed to perform both monolingual and cross-lingual WSD and automatically map encyclopedic entries to a computational lexicon. Because of its extensive functionality, it has been proposed as a good knowledge database for semantic search engines for scientific literature as well as general search engines [Latard et al., 2017].

Additionally, it is important to recognize domain-specific entities especially when developing domain-specific systems. This task is typically referred to as *terminology extraction*[Pazienza et al., 2005]. The process is a subtask of information extraction and involves extracting terms specific to a given corpus often by leveraging POS and WSD tasks as a precursory step [Alrehamy and Walker, 2017]. The first step in terminology extraction is collecting a vocabulary of domain-relevant terms that is representative of the domain's concepts (e.g. a list of drug names). Because of their low ambiguity and high specificity, these terms are especially useful in conceptualizing a knowledge domain or a terminology base (e.g. an ontology). Several approaches have been proposed [Pazienza et al., 2005], many of which relying on pre-built ontologies [Park et al., 2002, Lossio-Ventura et al., 2016a, Spasić et al., 2015], while others aim at building the ontologies themselves [Navigli and Velardi, 2004, Wong et al., 2007, Lossio-Ventura et al., 2016b]. Semantic academic search engines, especially domain-specific engines, such as GeneView, Semantic Scholar, and SEMEDICO have used terminology extraction to tailor their algorithms to be more specific in document retrieval [Mangold, 2007, Thomas et al., 2012].

Multiword expressions are prevalent in text (e.g. "search engine", "academic search", "Word Sense Disambiguation"). *Multiword recognition* is essential for terminology extraction [Alrehamy and Walker, 2017] and WSD tasks [Finlayson and Kulkarni, 2011]. Additionally, recognizing such expressions is paramount to search engines due to the

prevalence of multiwords in both, search queries and scientific articles [Karttunen et al., 1996, Jacquemin et al., 1997]. Thus, almost all search engines (Google Scholar, Microsoft Academic, Semantic Scholar, Textpresso, among otters) make use of some approach to detect connected word spans and multiword expressions. Historically, linguistic approaches had dominated the field of multi-word recognition (as is the case with many other NLP tasks), however, more recent approaches (i.e. early 2000s to present day) have employed statistical as well as linguistic knowledge [Heid, 1998], building models using supervised and semi-supervised techniques [Frantzi et al., 1998, 2000, Kulkarni and Finlayson, 2011, Oliver and Vàzquez, 2015].

Relation Extraction

Once the key entities are identified from the text, finding the relationship between each entity is essential when searching for underlying meanings—that is, semantics. Typically, this step involves identifying keywords such as "located_in", "part of", "is a", which are recognized phrases in the English language that serve to unite two or more entities. Other abstract relationships are also possible such as "ecologically_related_to" and "produced_by" which are commonly found in subfields of life sciences [Faessler and Hahn, 2017]. The use of ontologies in identifying the relationships between entities has shown promising results in many applications of relation extraction. Ontologies are essential in the portrayal of a hierarchical relationships between different entities. Especially prominent in the biomedical field, ontologies have proven to be essential when identifying key relationships such as protein-protein interactions and linking protein symptoms and genomic entities to diseases [Müller et al., 2004, Delfs et al., 2004, Dietze and Schroeder, 2009, Faessler and Hahn, 2017, Hu et al., 2017]. Some methodologies have made use of machine learning techniques to identify the relations and properties [Thomas et al.,

2013]. Machine learning methods such as Support Vector Machines (SVMs) can been used to classify pairs of entities found in sentences based on large feature vectors that include Bag of Words (BoWs) of the text surrounding the entities.

Topic Discovery

In NLP and machine learning, topic modeling concerns the development of statistical models for the discovery of abstract topics present in a collection of documents or a corpus. This process is frequently used in text-mining for the discovery of hidden semantic structures in unstructured text [Steyvers and Griffiths, 2007]. This is particularly useful for multidisciplinary search engines as documents are classified into buckets of topics which aids in the retrieval and indexing of documents [Brophy and Bawden, 2005]. For example given a query (e.g. "Natural Language Processing"), a user does not typically mean to find papers containing those three words, but rather is looking to find papers in the field of NLP—that is, the topic [Tang et al., 2008a]. Latent Semantic Indexing or Analysis (LSI) and Latent Dirichlet Allocation (LDA) are two prominent algorithms that are used for the discovery and clustering of documents based on topics. While both LSI and LDA are wildly used by academic search engines, they are typically augmented or extended to fit the academic search domain. For example, Tang et al. [2008a] proposed a unified probalistic topic model, namely, Author-Conference-Topic (ACT) model, specific for academic search engines and frameworks. The ACT model simultaneously models papers, authors, and publication venues. Developing topic modeling methods for the academic search domain typically concerns the retrieval accuracy, specificity, and precision of documents.

Indexing

While some search engines such as PubMed and GeneView index their own database, most search engines with a multi-disciplinary focus, including Google Scholar, Microsoft Academic, and Semantic Scholar, provide the capability of indexing multiple databases on the web. Having a central site to search multiple sources from is useful as it reduces the time spent searching for relevant documents.However, it can be a limitation in the creation of semantic search systems as they typically rely on source and domain-specific semantics and thus need a uniform representation of data. Some systems proposed the usage of a exhaustive mappings of keywords or concepts with semantic categories to index an entire corpus [M¨uller et al., 2004]. While others follow a more popular approach by building on top of Apache Lucene and taking advantage of its inverted indexing. Lucene's indexing method retrieves sections within a document related to a keyword by first searching an article's index, as opposed to searching the words within each document. Lucene provides an efficient searching methodology and is used by GeneView, for example, in its text storage, query processor, and ranking engine modules. A more recent search tool based on Lucene, known as Elasticsearch, has seen more popularity among some of the latest proposed semantic search systems [Liu et al., 2015, Faessler and Hahn, 2017]. Additionally, Elasticsearch's TF-IDF scoring, is often treated as a basic document scoring algorithm.

Query Expansion

Query expansion is a technique that reformulates the user's original query, to improve the retrieval performance. After analyzing the input query, the system generates alternative queries on a lexical and semantic level. Query expansion provides the advantage of no vocabulary mismatch problem, or expensive human-constructed knowledge bases. Two ways, either updating the query keywords or their weights to enhance efficiency of in-

formation retrieval. It is an indispensable for solving ambiguous queries [Vechtomova and Wang, 2006]. A recently proposed Proximity Relevance Model (PRM) adapted the relevance model by incorporating contextual proximity information. The closer a term is to a query term, the better the QE term candidate; moreover, proximity is directly captured in terms of sentences rather than tokens [Ermakova et al., 2016]. An example of Query Expansion is demonstrated by the search engine BioSearch, which leverages the Semanticscience Integrated Ontology (SIO) for the mediating ontology and use hierarchical relations between ontology classes to conduct query expansion [Hu et al., 2017].

Visualization and Organization of Results

In search engines following a keyword-search approach, the visualizations of results are often pretty simple. An interface for academic search must make sure to display the critical metadata of each document (e.g. title, authors, abstract) and provide the user with the enough information for he/she to be able to identify if the paper is relevant to what they are searching. After searching for a keyword in GS, the interface displays the results by giving focus to the title, author names, year of publication, citation counts, publisher and a snippet of the paper's abstract. Google Scholar will also highlight the keyword the user searched within the paper's title and the abstract snippet. The Google Scholar interface is easy to navigate, providing a user-friendly functionality. However, because of it's simplicity, users often struggle to make more advanced searches. In graphical user interfaces for semantic search systems, the aim is often to show more advanced visuals but in a simple approach. Systems such as Textpresso, the focus is a side menu with links to informative pages about the ontology, a document type definition, a user guide, and example searches, as well as the two retrieval and customization interfaces. The result list retrieved by a query can be customized in such a way that the user can choose how to

display the information [Müller et al., 2004]. GoPubMed follows a similar approach by showing the Gene Ontology in the interface on the left-side of the page. The part of the Gene Ontology (GO) relevant to the query is highlighted and on the right there are listed the abstracts for a selected GO term. GoPubMed also uses color as visual cues where the search terms are highlighted in orange and the GO terms in green [Thomas et al., 2013].

Document Ranking Approaches

Ranking algorithms for academic search engines usually fall into two categories (1) sorting based on the paper's prestige (i.e. citation count, publishing venue, impact factor) and (2) sorting based on the paper's level of relevance to the user's query. Most search engines, nowadays, use a combination of both approaches.

Ranking based on importance or prestige

There are many factors that can be used to measure the importance of a paper in a specific research field. The most popular approach is a sorting that depends on the number of times the paper has been cited by other academic works. Google Scholar is well-known for measuring a paper's importance by heavily relying on citation counts [Beel and Gipp, 2009]. The method is beneficial when looking for standard papers that have had a high influence on a certain research field throughout the years. However, it makes it difficult to find recently proposed works that advance a certain research field. Other factors that could be considered as prestigious (e.g. the journal's impact index) seems to not have any influence on the paper rank in Google Scholar.

The *PageRank* algorithm, adopted by Google's search engine, is a well-known method to compute the ranking of every page in the web [Page et al., 1999b]. It is an algorithm that works on a graph created by all world wide web pages as nodes and hyperlinks as edges. It indicates the importance of a web page by favoring the older pages since a new page

will not contain many links [Al-Hattab, 2016]. The following is the simplified formula of how the algorithm works. Assuming we have a series of web pages $\{w_1, w_2, w_3, w_4, ...\}$, the PageRank of a page is recursively defined, and depend on the number of incoming links. PageRank algorithm depends on the relation between pages.

$$PR(w_1) = \frac{1-d}{N} + d(\frac{PR(w_2)}{L(w_2)} + \frac{PR(w_3)}{L(w_3)} + \frac{PR(w_4)}{L(w_4)} + ...),$$

where L is number of outbound links, d the damping factor, N is the total number of pages on the web. It is strongly believed that Google Scholar adopted the PageRank algorithm along with other undefined methodologies to aid on its article retrieval [Al-Hattab, 2016].

One major concern for PageRank algorithm is new papers will never be cited and remain buried under the pile of older and highly-cited papers. Hasson *et al.* [Hasson et al., 2014] addresses these issues by proposing *Paper Time Ranking Algorithm (PTRA)*, an approach that depends on the paper's age, citation index, and publication venue. The algorithm gives a different level of priority for each of these parameters, giving more weight to the publication date such that new papers come out at the highest rank. The weight of each paper is calculated using the following formula.

$$Paperweight = A + T + C$$

where A is either the age of the conference or publishing journal's impact factors, T is the age of the paper and C is the citation index. Depending on whether the paper is a conference of journal paper, the algorithm may calculate A as either $A = M * d_1$ where M is the journal impact factor, or $A = Q * d_1$, where Q is the year of the conference and d_1 is the coefficient. The value for date of publication T, is regarded as the most important metric in the equation. It is calculated by $T = (CurrentYear - PublicationYear) * d_2$, where d_2 is the coefficient. And lastly, the citation index value C is calculated by $C = T * d_3$, where T is the value of citation index and d_3 is the coefficient. Although the approach showed a significant improvement compared to other ranking algorithms, it's

performance was not as impressive compared to other similar methodologies [Brisebois et al., 2017].

Reputation-Based Ranking [Ribas et al., 2015] The authors propose a random walk model to identify the most reputable entities of a domain based on a conceptual framework of reputation flows The focus on the paper is not ranking articles, but instead identifying and ranking venues and research groups. Model research groups as reputation sources and publication venues as reputation targets, with edges running from a source to a target and back again to indicate the transference of reputation through one or more publications. The paper interprets the relative reputation from entities such as author, research group and publication venue. Each entity reflects a steady state probability, which can be further propagated to other comparable entities depending on the matrix $P^{\langle TC \rangle}$ of size $|T| \times |C|$ representing the transitions from reputation targets to collateral entities. The P-score of of an entity e is defined as

$$P - score(e) = \begin{cases} \sum_{t \in T} P_{te}^{\langle TC \rangle} \pi_t & if e \in C \\ \pi_e & otherwise \end{cases}$$

$P^{\langle TC \rangle}$ The calculated P-score is then used to produce an overall reputation-oriented ranking of the entities.

Ranking based on content relevance

These ranking methodologies are query-dependent, where the top-ranked papers must include concepts that are closely related to the user's request. The most standard ranking algorithm is based on a simple keyword-count. After keywords have been identified in the user's query, the system will look through the . For a long time, literature has acknowledged the limitations and criticized the method of keyword-based indexing for academic articles [Lee et al., 1997]. Nonetheless, it remains the preferred approach by most search engines.

Ranking based on concept *TF-IDF* (term frequency-inverse document frequency) is a frequency-based approach and one of the most popular schemes for document term-weighting [Breitinger et al., 2015].

$$tfidf(t, d, F) = tf(t, d) \times idf(t, D)$$

Term frequency is defined by

$$tf(t, d) = 0.5 + 0.5 \times \frac{f_{t,d}}{\max\{f_{t',d} : t' \in d\}}$$

Meanwhile, inverse document frequency is defined by

$$idf(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

where N is the number of documents in the corpus. After determining the term weights, a ranking function serves to measure the similarity between the query and document vectors. As shown by Lee *et al.*[Lee et al., 1997] cosine measure is a commonly known similarity measure which calculates the angle between the document vectors and the query vector that are represented in a V-dimensional Euclidean space, where V is the vocabulary size. The similarity between a document D_i and a query Q is defined as

$$Sim(Q, D_i) = \frac{\sum_{j=1}^{V} w_{Q,j} \times w_{i,j}}{\sqrt{\sum_{j=1}^{V} w_{Q,j}^2 \times \sum_{j=1}^{V} w_{i,j}^2}}$$

where $w_{Q,j}$ is the weight of term j in the query, and is defined similar to $w_{i,j}$. The denominator in this equation, known as the normalization factor, discards the effect of document lengths on document scores.

Explicit Semantic Ranking (ESR) [Xiong et al., 2017] is a proposed ranking technique that leverages knowledge graph embedding to analyze query logs from Semantic Scholar (S2). It is an approach proposed to improve S2's current ranking system, at the time built on top of ElasticSearch's vector space model. It works by creating an estimation

of the ranking score for the query terms and bi-grams on the papers' title, abstract, body text, and citation context. Entity embeddings are used to obtain "soft-match" feature of each query, document pair. First, the paper builds knowledge graphs, that stores semantic information. Given a query q, and a set of candidate documents $D = \{d_1, ..., d_n\}$, ESR finds a ranking function $f(q, d|G)$, that better ranks D using the explicit semantics stored in the knowledge graph G. The explicit semantics include entities ($E = e_1, ..., e_{|E|}$) and edges (represented by predicates P and tails T). ESR creates a representation for query and documents through their bag-of-entities. ESR matches query and documents' entity representations using the knowledge graph embedding. Each element in the matrix is the connection strength between a query entity e_i and a document entity e_j, calculated by their embeddings' cosine similarity. The ranking with semantic evidence is defined as

$$f(q, d|G) = w_0 f_{s2}(q, d) + W^T B(q, d)$$

where $f_{s2}(q, d)$ is the score produced by S2, w_0 and W as the parameters to learn, and $f(q, d|G)$ is the final ranking score. B is the bin score produced by the bin-pooling (histogram), count the matches at different strengths. One of the limitations of ESR is its requirement for training data to combine word-based and entity-based relevance scores and to select parameter settings, addressed by *Dual Embedding Space Model (DESM)* [Mitra et al., 2016]. The paper investigates neural word embeddings as a source of evidence in document ranking. DESM trains a word2vec embedding model on an unlabeled query corpus, retaining both the input and the output projections and identifying whether a document is about a query term in addition to what is modeled by traditional term-frequency based approaches. Part of the ranking process, the system maps the query words into the input space and the document words into the output space, then estimates query-document relevance score by aggregating the cosine similarities across all the query-document word pairs.

Other Approaches

Compared to the methodologies we discussed thus far, these approaches are more innovative in nature. Arnetminer (AMiner) is a novel online academic search and mining system, with the goal to provide a systematic modeling approach and extract researchers' profiles automatically from the Web and integrates them with published papers, first performing name disambiguation. Generative probabilistic model is devised to simultaneously model the different entities while providing a topic-level expertise search [Tang et al., 2008b]. AMiner is devised as a unified topic modeling approach to modeling the different entities (authors, papers, venues) simultaneously and providing a topic-level expertise search. It's focus fall into social influence analysis, influence visualization, collaboration recommendation, relationship mining, similarity analysis, and community evolution [Tang, 2016]. Another approach is random walk model to infer the reputation of a target set of entities with respect to suitable sources of reputation. Utilized for ranking a target set of entities with respect to suitable sources of reputation [Ribas et al., 2015].

5.1.6 Conclusion

In this section, I introduced academic search, its history, segments, and approaches. The advancement in technology and its various sectors brought about an exponential growth in digitized data of all forms. Human knowledge is no exception. In fact, the "knowledge doubling curve" Fuller and Kuromiya [1981] once created by Buckminster Fuller as an indicative to the average growth of human knowledge is no longer valid Schilling [2013]. By the end of World War II, human knowledge was doubling every 25 years, while more recently it has been doubling every 13 months on average. According to IBM, the rise in adoption of technologies within the Internet of Things (IOT) is predicted to double human knowledge every 12 hours [Coles et al., 2006] indicating a shift from linear to

exponential growth of knowledge. The academic search space and market have seen an increase in demand followed–naturally–by an increase in offer due to the explosion of published human knowledge and digitized data. This was especially the case over the past decade with the initial spark in the early 2000s.

Since the launch of early solutions such as Web of Science (WoS) and Google Scholar (GS), the area has been dominated with a simple yet insufficient approach—that is, keyword-based search. This is not to state that this approach has been static. In fact, it has been significantly improved and hybridized such that it is becoming inaccurate to compare current practices to their predecessors. Semantic search has always been a desired approach for most academics and researchers working in the field. However on one hand, its inherited difficulties and complexities, and on the other hand, keyword-based search's simplicity and effectiveness (when dealing with a small knowledgebase) has pushed it further away from widespread adaptation and realization.

5.2 Ontology-Based Supervised Concept Learning for the Biogeochemical Literature

Academic literature search is a vital step of every research project, especially in the face of the increasingly rapid growth of scientific knowledge. *Semantic* academic literature search is an approach to scientific article retrieval and ranking using concepts in an attempt to address well-known deficiencies of keyword-based search. The difficulty of semantic search, however, is that it requires significant knowledge engineering, often in the form of conceptual ontologies tailored to a particular scientific domain. It also requires non-trivial tuning, in the form of domain-specific term and concepts weights. In this section I present an ontology-based supervised concept learning approach for the biogeo-

chemical scientific literature. This study was part of an ongoing project (ENVOSCHOLAR) seeking to build a domain-specific semantic search system.

5.2.1 Introduction

The first step of most scientific research projects is a review of the existing literature. *Academic literature search* allows a researcher to understand what hypotheses have been proposed, what methods or procedures have been tried or tested, and what results have been achieved. In most cases, indexing and retrieval of relevant articles is done using keywords [Lewandowski, 2015]. Although simple and computationally inexpensive, keyword-based search has serious limitations considering the complexity of human language [Lewandowski, 2015, Martínez-Sanahuja and Sánchez, 2016]. Furthermore, as scientific knowledge grows exponentially larger, these limitations become more serious and serve to inhibit the ability of researchers to use existing tools to find relevant scientific literature [Brophy and Bawden, 2005].

A solution to this problem that has often been proposed is *semantic search*, that is, systems that can infer the meaning of a user's query and therefore retrieve articles of greater relevance [Leyba, 2016]. Ontologies are a key component of this approach, as they provide a specific lists of terms and concepts as well as relationships between those items [Huang et al., 2016]. The challenge, however, lies in mapping articles and their constituent parts to the relevant parts of the ontologies [Dang et al., 2017].

Early work on ontology-based concept extraction used regular expressions or exact keywords matching [Müller et al., 2004, Allahyari et al., 2014]. However, this requires encoding knowledge of all possible tokens that can map to specific ontology entities [Sriharee, 2015], a problematic task because of the ambiguity of language. Because of this, keyword approaches often miss essential concepts during the recognition and extraction

steps. More recent work tackles the problem using matches driven by supervised machine learning (ML), which can automatically learn and judge which ontology concept is indicated by observed text.

The work presented here demonstrates the latter approach specifically for the biogeochemical domain. This was part of a larger domain-specific semantic search engine for the biogeochemical academic literature. In a prior report, my co-authors and I demonstrated the efficacy and feasibility of using ontological concepts to rank articles based on a search query [Eisenberg et al., 2017]. In this section, I demonstrate the development of a supervised machine learning (ML) approach that automatically learns ontological concepts, and labels sentences from biogeochemical articles with those concepts using features extracted from the unstructured text. I discuss the features necessary to build such systems and the process by which those features are extracted.

The remainder of this section is organized as follows: I first review related work on ontology-based concept extraction (§5.2.2). Next, I describe my approach including the task definition, the ontology used, as well as the dataset created (§5.2.4). I finally present and discuss the experiments performed as well as the results obtained from those experiments (§5.2.5).

5.2.2 Related Work

An ontology provides formal and explicit specifications of conceptualizations, usually with a focus on a particular domain. Ontologies are one of the most recognized methodology of knowledge representation, providing definitions for a particular entities, relationships between entities, and classification of an entity on a class hierarchy. Ontology-based information extraction (OBIE) has been recently coined as a subfield of information extraction. In OBIE, ontologies play a crucial role in providing knowledge representation.

The process is a core building block for the implementation of semantic search for large document repositories as well as the development of the Semantic Web [Dou et al., 2015, Wimalasuriya and Dou, 2010].

Ontologies have been useful for semantic data mining and search tasks. Ontology-based semantic data mining and search approaches and task include: association rule mining, classification, clustering, information extraction, recommendation systems, and link prediction for social networks [Dou et al., 2015]. Classification is a common task in data mining as well as other fields which aims at finding a model (or function) to describe and distinguish data classes or concepts [Jaiwei and Kamber, 2006]. Typical use of classification in ontology-based semantic search is the annotation of classification labels using entities and relations defined within the ontology. Setchi *et al.* [Setchi and Tang, 2007] proposed a concept indexing algorithm that makes use of general-purpose ontologies. Although their work uses a supervised approach, the ontology tagging process was done automatically instead of manually. Therefore, the accuracy of the tagged terms is only an approximate.

Some approaches to ontology-based classification of documents or topic modeling use the similarity of semantic graphs. The HITS algorithm [Kleinberg et al., 1999] works over semantic graphs to identify core entities. Using DBpedia-based ontologies, Allahyari *et al.* [Allahyari et al., 2014] identified entities and their relations from test documents. By contrast, for this work, I focus on indexing ontology concepts at the sentence level, other approaches have indexed concepts at the word or the document level [Wimalasuriya and Dou, 2010].

Most related to this work is Textpresso [Müller et al., 2004], a search engine which promises to enhance the retrieval of biological literature (as opposed to the biogeochemical here) by using an ontology-based approach. In Textpresso, multiple ontologies play essential roles in the retrieval of pertinent information from documents, resulting in sig-

nificant acceleration of extraction of biological facts. The user can retrieve a set of documents by searching one or a combination of keywords. Ontologies make it possible to create semantic queries, facilitating the search the corpus of text by meaning instead of keyword-match. Textpresso achieves this by first identifying and matching the terms against pre-defined regular expressions.

Additionally, the creation and use of ontologies have been especially relevant in the biomedical domain where they were used for the identification of biological terms within raw text—such as scholarly publications and medical records [Žitnik et al., 2015, Gurulingappa et al., 2012, Moens, 2006]. The first step in the extraction of such terms is named entity recognition (NER), where the system can recognize and extract names of genes, drugs, chemical compounds, diseases, and so on. After these terms have been listed and formally defined via ontologies, the next step is defining the relationships between different entities (i.e., identify gene-gene or protein-protein interaction) [Moens, 2006].

5.2.3 Dataset

As discussed in chapter 2, In a prior study with my colleagues [Eisenberg et al., 2017] we determined that the most useful ontology for our academic search engine project's purposes was the Environment Ontology (ENVO), a community-led, open ontology for various life science disciplines [Buttigieg et al., 2013]. According to its creators, ENVO is an attempt at establishing a standard annotation scheme for several co-dependent or related disciplines, including, but not limited to, ecology, hydrology, environmental biology, and the geospatial sciences. ENVO contains concepts corresponding to a wide range of natural environments and environmental conditions. It is encoded in the Open Biomedical Ontologies (OBO) syntax, which is a subset of the Web Ontology Language

(OWL). ENVO can be populated, managed, and maintained using the OBO-Edit ontology development tool.

ENVO, like many ontologies, is hierarchical in design. Three of its top-level, most developed branches are *environmental system*, *environmental feature*, and *environmental material*. It's hierarchical structure allows for it to include not only entities, but also higher-level relationships between various concepts, including many standard ontological relationships such as `is-a`, `part-of`, `contained-in`, `connects`, and `has-condition`. ENVO also contains scientific and domain-specific relationships such as `derives-from`, `input-of`, `output-of`, `has-habitat`, and `biomechanically-rel` Furthermore, the ontology boasts a well-connected graph of synonymy relationships, encoded using different granularities including `broad`, `exact`, and `narrow`.

ENVO has seen quite a bit of success in adoption and use. It has served as the foundation for the creation and expansion of a number of other ontologies, as well as applied in several annotation projects such as the International Census of Marine Microbes (ICOMM) and the International Nucleotide Sequence Database Collaboration (INSDC) [Field et al., 2011]. Additionally, ENVO has been used in data retrieval and query-based systems such as the Genomic Metadata for Infectious Agents Database (GEMINA) [Schriml et al., 2010], while the National Institute for Allergy and Infectious Diseases Bioinformatics Resource Centers (NIAID BRCs) employ ENVO in metadata formulation and manipulation [NIH NAIDS].

To the best of our knowledge there is no corpus of scientific articles annotated with ENVO concepts, so we created our own. For this study we collected a total of 14 articles (62,015 total words) using three search queries that were created by two domain experts. Our domain experts ran the queries through Google Scholar and examind from the several hundred results returned, identifying the top four or five most relevant articles for each query. Importantly, several of the articles were not ranked near the top of Google's results,

Query	Title	Citation	Tokens	Sentences	Unique Concepts	κ
Methyl-Mercury concentrations in Everglades water and sediment	Mercury in the Aquatic Environment ...	[Ullrich et al., 2001]	5,081	162	26	n/a
	Sulfide Controls on Mercury Speciation ...	[Benoit et al., 1999]	4,133	168	13	n/a
	Sulfate Stimulation of Mercury Methylation ...	[Gilmour et al., 1992]	3,642	160	18	n/a
	Effect of Salinity on Mercury Activity ...	[Compeau and Bartha, 1987]	3,421	150	22	n/a
Sulfate reduction occurring in Everglades pore waters and sediments	Anaerobic Microflora of Everglades Sediments ...	[Drake et al., 1996]	4,651	179	35	0.64
	Constants for mercury binding ...	[Benoit et al., 2001]	4,629	173	17	0.62
	Mercury methylation in periphyton ...	[Cleckner et al., 1999]	3,839	159	18	0.75
	Methylmercury Concentrations ...	[Gilmour et al., 1998]	4,295	183	26	0.30
	Bacterial Methylmercury Degradation ...	[Marvin-DiPasquale and Oremland, 1998]	3,696	199	27	0.44
Sulfur reduction affecting South Florida Everglades soils	Groundwater's significance to changing ...	[Harvey and McCormick, 2009]	9,650	300	73	0.63
	Variation in Soil Phosphorus ...	[Chambers and Pederson, 2006]	3,032	103	39	0.71
	Sulfur in the South Florida ecosystem ...	[Orem et al., 2011]	3,485	149	37	0.69
	Sulfur in peat-forming systems ...	[Casagrande et al., 1977]	3,998	165	35	0.71
	Effects of sulfate amendments ...	[Dierberg et al., 2011]	4,463	160	42	0.62
		Max	9,650	300	73	0.75
		Average	4,430	172	31	0.61
		Min	3,032	103	13	0.30
		Standard Deviation	1,604	43	15	0.14

Table 5.1: Articles in the test set. Listed are the number of tokens in each article, the number of sentences overall, the number of unique concepts, and the annotator agreement expressed as Cohen's κ.

and were rather found many pages deep. We then manually annotated articles at the sentence level using concepts from ENVO (chapter 2 and §5.2.3 discusses the annotation study in detail). Table 5.1 lists the queries, the corresponding articles returned from the search results, as well as article-specific statistics. The articles have an average of 4,430 tokens, 172 sentences, 192 unique ENVO concepts. Table 5.1 presents detailed statistics on the test set.

Annotation Study

As discussed in chapter 2, the purpose of manually annotating concepts from the ontology was twofold: first, to show that the ontological concepts appear in the target texts and, second, to show that it is possible to automatically learn domain-specific concepts from a relevant ontology. Because developing concept detectors is a non-trivial task, in prior work we tested the utility of the ontology, as well as verified that it is feasible to automatically rank articles using detected ontological concepts [Eisenberg et al., 2017]. The current work expands that effort by creating a larger gold-standard corpus and demonstrating that we can identify the concepts in the articles automatically.

As discussed above, we collected a corpus of 14 articles from the biogeochemical domain, aligned with three search queries. Our team of domain trained annotators then annotated the queries and the articles for concepts from ENVO. For each article, annotations were collected at the sentence level. The resulting micro-averaged inter-annotator measure agreement over all annotator groups using Cohen's κ is 0.61 which is "substantial" agreement [Artstein and Poesio, 2008].

5.2.4 Approach

The goal of the work presented here was to label the sentences of scientific articles—drawn from the biogeochemical academic literature—with concepts derived from a domain-specific ontology (specifically the *Environment Ontology*, or ENVO). I treated this as a supervised classification problem where I train a classifier using sentences that have been manually labeled (annotated) for their concepts; then, this classifier takes individual sentences found in a new article as input, outputting ontology concepts.

In this section I first describe the task in detail, next I discuss the classification training process, starting with data preprocessing, followed by feature extraction, and ending with classifier construction.

Task Definition

As noted above, the task was to index academic articles in the biogeochemical domain with concepts derived from ENVO. That is, given a set of academic articles and a domain-specific ontology, the solution is a supervised classification model that can assign ontology concepts to the sentences found in the articles. As discussed above, we created a dataset of articles which was manually labeled and indexed with concepts from ENVO. Each sentence may have any number of concepts and therefore the labels are not mutually

exclusive and my solution must admit a multi-label classification, including possibly no label. identified a set of distinctive features to support this classification, and designed feature extractors to compute these features over article text.

Data Preprocessing

In addition to annotating the data with ENVO concepts as described in the previous section, I performed standard NLP preprocessing tasks to prepare the data for feature extraction and supervised learning. First, I encoded document structure and formatting information such as section and paragraph headers, as well as sentence counts and relative positions of sentences within sections. Next, I cleaned the text by removing in-text citations and stand-alone mathematical, chemical, and biological formulas. I then tagged each token with its part-of-speech [Bird and Loper, 2004], lemmatized tokens using WordNet [Fellbaum, 1998], filtered known stop words using PubMed's list [PubMed Help, 2005], and used the pywsd module to perform word-sense disambiguation [Tan, 2014] to tag words with WordNet senses.

Data Balancing

The articles included 192 unique concepts across 3,434 occurrences. More than half of these occurrences (2,049) represented only 10 concepts, while the most frequent 50 concepts (26% of the total) occurred 3,091 times in total. Additionally, 61 concepts (32%) appeared only once. When supervised ML is performed over such distributions, they tend to overfit the classes with higher number of examples. Several solutions have been proposed and used for the problem of imbalanced data such as sampling (undersampling and oversampling) and weight assignment. These techniques are used to help supervised ML classifiers learn more about a class that has a significantly smaller number of examples relative to others. In this case I opted to use the Synthetic Minority Over-sampling

Technique (SMOTE) [Chawla et al., 2002]. SMOTE is a hybrid sampling technique that oversamples the minority classes while undersampling the majority classes. I applied resampling to the training set only, leaving the testing set with the original distribution.

Feature Extraction

Identifying a useful set of features is integral for an accurate machine learning model. For this task I extracted lexical, syntactic, and semantic features from the articles and their sentences. For lexical features, I used the most frequent distinctive terms for each article using *term frequency-inverse document frequency* (tf-idf) [Church and Gale, 1999]. I used the top 10% of the resulting lists. Additionally, I used global and local sentence positions as features—i.e., the relative position of a sentence in both its section and article, expressed as a real number between 0 and 1, inclusive. Further, I extracted named entities from each sentence by examining parts-of-speech (looking for runs of tokens tagged NNP or NNPS), and used these entities as features. As discussed earlier, recognizing named entities is useful for many IR and NLP tasks. An example of this from my study is the term *Everglades* which is found encoded in ENVO as a synonym and part definition for *peat swamp*.

Finally, for semantic features, I mapped the words in each sentence to a semantic embedding space. As an example of an embedding approach, word2vec [Mikolov et al., 2013] is a popular and powerful method to represent high-dimensional word embeddings which reduce the complexity and size of the feature set as opposed to a bag of words (BoW) approach. However, word2vec does not consider words that have multiple senses, mapping them to the same position in the vector space. To address this limitation, I used sense2vec [Trask et al., 2015], where different senses of the same word are placed differently in the embedding space. I used Sense2vec as implemented in the SpaCy python module [AI, 2015], and followed the algorithm described in [Trask et al., 2015] by us-

ing the part-of-speech tags and named entity labels assigned to the tokens. Additionally, I merged named entities into single tokens (using hash symbols), so that they were assigned a single vector.

In addition features extracted directly from the raw text, I also used other concepts as features. First, I used concepts identified in the abstract of each article as features for the body of the article. Second, I used the concepts present in a the immediately preceding sentence as features for determining the next sentence's concepts. This feature engineering led to several interesting observations; first that concepts found in the abstract of an article can improve concept labeling performance for the article body; and further, that knowing which concepts came before a sentence (i.e., in sentences preceding the sentence in question) also improves concept labeling performance.

Concept Learning

The first stage of classification is model training, followed by a stage of testing on separate (unseen) data. The original data was randomly split into two portions ten different times (ten folds), 80% in the training set and 20% in testing set (11 and 3 articles, respectively). I built and trained the concept learning models using random decision forest models (RDFs). RDFs are ensemble learning methods and are employed in regression and classification applications [Ho, 1995]. They operate through the construction of numerous decision trees during the training stage. The technique outputs the class that contains the mode of the classes of the collection of collection of tress. This technique is influential, especially in data mining applications [Franklin, 2005]. A major advantage of RDF over regular decision trees is that the RDF avoids overfitting the training set [Criminisi et al., 2012].

I built and trained two separate models using the features discussed in the previous section—a *body-only* model, which used all features, and an *abstract-only* model, which

omitted the abstract concept features as well as the sentence counts and position features. This two-model approach attempts to mimic how human read scientific articles, namely, using the concepts found in the abstract to better guide the understanding concepts found in the rest of the text.

With regard to the parameters of the RDF classifiers, *max_features* was set to the square root of the total number of features in an individual run, *number_of_trees* was arbitrarily set to 50, where this is referring to the number of trees built before taking the average tree votes for predictions. Finally, *min_sample_leaf* was set to 50. To implement these models I used the scikit-learn python ensemble module [Pedregosa et al., 2011].

5.2.5 Experiments and Results

As discussed above, I randomly split the dataset into training and testing sets across ten folds, resulting in 11 articles for training and 3 for testing in each fold. The models learned a total of 192 unique concepts. For all experiments, I evaluated the performance of the models on each concept using the F_1 measure averaged across all folds. Here I present the evaluation methods and results, describing the baseline approaches, as well as the performance of both the baselines and my method average, averaged across the test sets.

Baseline Methods

I compared my approach to two baseline methods. The first baseline was a keyword-based approach, where I matched sentence words directly to the names of ontology concepts. All previously mentioned preprocessing steps were performed on both the text and the ontology, such as lemmatization of both concepts and words in the sentences. This model needed no training. The second baseline was a Bag of Words (BoW) supervised classifier.

For this classifier, I trained and tested a support vector machine (SVM) [Cortes and Vapnik, 1995] following the same cross-validation splits and multi-label fashion as used for my proposed approach. The SVM classifier was trained using the RBF kernel function and a soft margin C of 10,000—a common setup.

Experiments

As noted above I built two models: (1) an *abstract-only* model, and (2) a *body-only* model. Both the models learn concepts using all sentences in the text (including the abstract), but as the names suggest, they only used to label the abstract sentences and the body sentences respectively. Additionally, the the *body-only* model uses the labels produced by the *abstract-only* model as features for labeling the body of an article. In order to compare the efficacy of using the sense2vec approach as a feature, I built trained and tested the same models using a word2vec approach instead.

Table 5.2 shows three average F_1 scores over different sets of concepts for all discussed approaches. The first column shows the average F_1 score for the concepts with single occurrence in the original data (61 concepts), while the second column shows the average scores for the top 50 concepts in terms of total occurrences over all the articles. The last column shows the results over all concepts. The proposed approach (RDF_sense2vec) outperforms both baselines as well as the RDF_word2vec models across all concepts. Additionally, Figure 5.1 shows the frequency of the top 50 ENVO concepts as well as the average F_1 score of each for each of the concepts. As shown, the score drops with the frequency of the concept in the dataset, although not dramatically. This is expected as it is a result of the original class imbalance. Finally, the *abstract-only* model performed similarly well with a 0.69 F_1 over all concepts present in the abstract sections, which were relatively small in number.

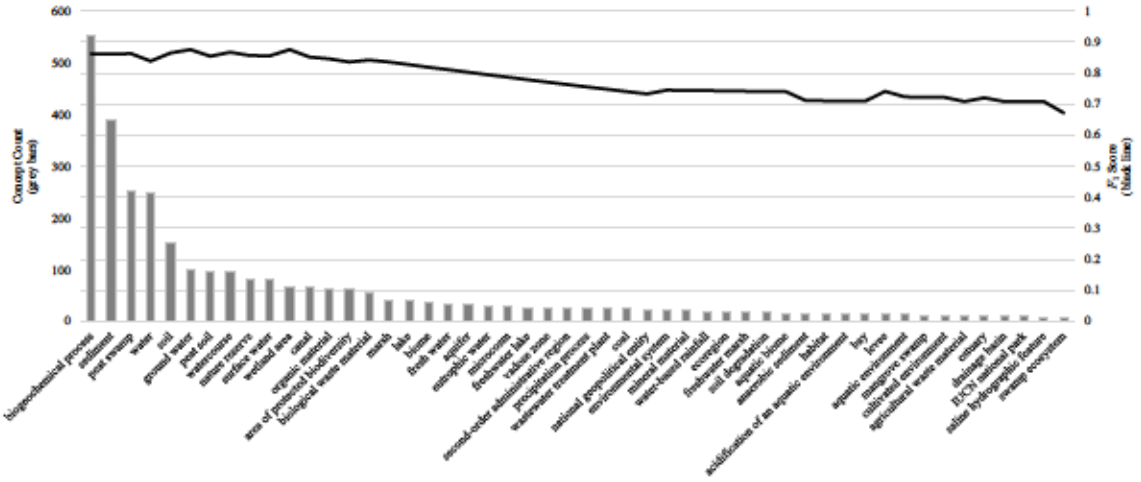


Figure 5.1: Frequency of the top 50 ENVO concepts (grey bars) and the average 10-fold testing results (F_1 scores) for each of the concepts (black line).

	Unique Concepts		
Approach	Single	Top 50	All
Keyword Search	0.39	0.35	0.38
SVM_BoW	0.45	0.56	0.50
RDF_word2Vec	0.54	0.69	0.61
RDF_sense2Vec	0.67	0.78	**0.76**

Table 5.2: Average F_1 scores per approach over all concepts, the 50 most frequent concepts, and the 61 least frequent concepts with single occurrences.

As discussed in §5.2.4, I proposed that the model's performance would improve when (1) abstract concepts are used as features for the body concept extraction, and (2) sentence positions are also included as features (i.e., sentence positions relative to the article as a whole and individual sections). To evaluate this, I performed four experiments, testing the inclusion of abstract concepts and sentence position features. Table 5.3 shows three average F_1 scores over different sets of concepts per experiment.

In the last row in Table 5.3, both abstract concepts and sentence positions were included as features in the models. The results confirm my hypothesis, in that the inclusion of both of those features yields better labeling results across all concepts. The second-to-last row shows the results for including the abstract concepts as features, but omitting the

Features		Unique Concepts		
Abstract	Position	Single	Top 50	All
Omitted	Omitted	0.52	0.69	0.65
Omitted	Included	0.63	0.71	0.68
Included	Omitted	0.63	0.76	0.70
Included	Included	0.67	0.78	**0.76**

Table 5.3: Average F_1 scores for the feature combination experiments. The first two columns indicate whether the abstract concepts and sentence positions were included or omitted as features in the models.

position features. This resulted in lower results overall, but significantly impacted the average score for the single occurrence concepts. Interestingly, I investigated this and found that deeper concepts (i.e., in terms of the ontology hierarchy) are found at higher densities close to the middle of the articles as well as the centers of article sections. In retrospect this makes sense, as the methodology section of a scientific paper (located around the middle) would normally contain detailed concepts rather than abstract ones. To put this together, most of the single occurrence concepts are deeper, low-level concepts, hence the low occurrence frequency in the original data. The second row shows the results for only omitting the abstract concepts as features when labeling the rest of the text in the articles. Again, the models' performance dropped overall, but less so than for single occurrence concepts. This can be attributed again to including the sentence position features, which aid the labeling for less frequent concepts. Finally, the first row shows the results for omitting both features with the models performing the worst across all concepts.

5.3 Using Document Structure for Ontology-Based Concept Learning

Ad discussed before, academic literature search is a vital step of every research project, especially in the face of the increasingly rapid growth of scientific knowledge. Ontology-

based *semantic* search is a an approach to scientific article retrieval using concepts in an attempt to address well-known deficiencies of keyword-based search. Leveraging domain-specific ontologies is useful for an accurate retrieval of scientific articles. However, many solutions often miss leveraging key information that is typically encoded in the document structure of scientific articles. In this section, I present an extended approach to ontology-based supervised concept learning for the biogeochemical scientific literature that uses the section structure of scientific articles.

5.3.1 Introduction

We all have had the experience of searching the scientific literature using a keyword-based search engine. You probably started with a general query, which returned thousands of articles that only tangentially related to your interests. Because no researcher would have time to even skim all the results, you returned to the original query, rewording it multiple times in different ways until highly relevant articles were ranked at the top of the search. These are long-known deficiencies of keyword-based search, namely: (1) retrieval of an overwhelming number of results, (2) rankings that do not precisely reflect true relevance, and (3) the omission of relevant results because they do not contain the idiosyncratic keywords of the query [Tümer et al., 2009].

A long-proposed solution to this problem is *semantic search*, which uses concepts in the query rather than just keywords to drive document retrieval and ranking. Semantic search often leverages domain-specific knowledge, usually encoded in ontologies, to help rank the relevance of documents relative to a search query. Semantic search is difficult, however, because the required knowledge entails significant knowledge engineering or sophisticated natural language processing (NLP). Despite these problems, search engines today do boast high performance compared to prior decades precisely because they in-

clude minor semantic knowledge in their search algorithms; one approach, for example, is Latent Semantic Indexing (LSI), which uses synonyms and relationships between page headers, document titles, and content to assist ranking [Li et al., 2014]. Nevertheless, we are still far from the full realization of true semantic search that uses deep semantic techniques fully integrated into back-end algorithms. For these reasons semantic search research is experiencing a rise in interest among various groups [Bast et al., 2016, Jindal et al., 2014, Li et al., 2014].

Scientific articles usually follow strong principles of scientific writing structure. By design, they often contain a certain number of standard sections that convey different sets of ideas and solutions that support a central hypothesis or approach. These documents typically start with an introduction or motivation which discusses the research problem, followed by the research methodology, results and experiments, related work, and finally a conclusion. Semantic search approaches often miss utilizing this structure in identifying the key information to be retrieved. In the previous section (§5.2), I showed the feasibility of learning domain-specific concepts for the biogeochemical scientific literature. In this section, I present an improved ontology-based approach that uses a simple linear chain conditional random fields model to learn domain-specific semantic concepts. In my approach I use the automatic section discovery solution discussed in chapter 4 to embed the section structure of scientific articles as a feature in automatic concept learning. **XXX** mention

The remainder of this section is organized as follows: I discuss the dataset I ussed in results this study (§5.3.2), then, I describe my approach including the task definition, the CRF apporach, the features extracted, as well as how I used the section structure for learning semantic concepts (§5.3.3).

Query	Title	Citation	Tokens	Sentences	Unique Concepts	κ
Methyl-Mercury concentrations in Everglades water and sediment	Mercury in the Aquatic Environment ...	[Ullrich et al., 2001]	5,081	162	26	n/a
	Sulfide Controls on Mercury Speciation ...	[Benoit et al., 1999]	4,133	168	13	n/a
	Sulfate Stimulation of Mercury Methylation ...	[Gilmour et al., 1992]	3,642	160	18	n/a
	Effect of Salinity on Mercury Activity ...	[Compeau and Bartha, 1987]	3,421	150	22	n/a
Sulfate reduction occurring in Everglades pore waters and sediments	Anaerobic Microflora of Everglades Sediments ...	[Drake et al., 1996]	4,651	179	35	0.64
	Constants for mercury binding ...	[Benoit et al., 2001]	4,629	173	17	0.62
	Mercury methylation in periphyton ...	[Cleckner et al., 1999]	3,839	159	18	0.75
	Methylmercury Concentrations ...	[Gilmour et al., 1998]	4,295	183	26	0.30
	Bacterial Methylmercury Degradation ...	[Marvin-DiPasquale and Oremland, 1998]	3,696	199	27	0.44
Sulfur reduction affecting South Florida Everglades soils	Groundwater's significance to changing ...	[Harvey and McCormick, 2009]	9,650	300	73	0.63
	Variation in Soil Phosphorus ...	[Chambers and Pederson, 2006]	3,032	103	39	0.71
	Sulfur in the South Florida ecosystem ...	[Orem et al., 2011]	3,485	149	37	0.69
	Sulfur in peat-forming systems ...	[Casagrande et al., 1977]	3,998	165	35	0.71
	Effects of sulfate amendments ...	[Dierberg et al., 2011]	4,463	160	42	0.62
Everglades groundwater surface water interaction	Coastal groundwater discharge – an additional ...	[Price et al., 2006]	4,445	198	32	0.85
	The Influence of Hydrologic Restoration ...	[Sullivan et al., 2014]	5,860	220	28	0.81
	Ground Water Recharge and Discharge ...	[Harvey et al., 2004]	6,257	223	36	0.88
	Estimates of groundwater discharge ...	[Zapata-Rios and Price, 2012]	6,480	307	48	0.83
	Quantifying time-varying ground-water ...	[Choi and Harvey, 2000]	4,747	186	46	0.81
		Max	9,650	307	73	0.88
		Average	4,741	186	33	0.69
		Min	3,032	103	13	0.30
		Standard Deviation	1,604	43	15	0.15

Table 5.4: Articles used in the corpus. Listed are the number of tokens in each article, the number of sentences overall, the number of unique concepts, and the annotator agreement expressed as Cohen's κ.

5.3.2 Dataset

As discussed in chapter 2 and in the previuos section (§5.2), a team of reserachers and I collected and annotated a corpus of environmental scientific articles with the help of domain experts. While in the previous study (§5.2) I only utilized 14 articles for the development of the concept learning model, I used the full expanded corpus discussed in chapter 2. Again, similar to the previous study, I used the ENVO ontology and the annotation results from our annotation study (discussed in detail in chapter 2. The dataset consisted of 19 articles (90,074 total words) collected using four search queries that were created by three domain experts (two PhD students and a professor of Hydrology). Our domain experts ran the queries through Google Scholar and examined from the several hundred results returned, identifying the top four or five most relevant articles for each query. Importantly, several of the articles were not ranked near the top of Google's results, and were rather found many pages deep. I list detailed statistics of the dataset I used as presented in chapter 2 for ease of reference in Table 5.4.

5.3.3 Approach

Similar to my previous study (presented in section 5.2), the goal of the work presented here was to label the sentences of scientific articles—drawn from the biogeochemical academic literature—with concepts derived from a domain-specific ontology (specifically the *Environment Ontology*, or ENVO). I treated this as a supervised classification problem where I train a classifier using sentences that have been manually labeled (annotated) for their concepts; then, this classifier takes individual sentences found in a new article as input, outputting ontology concepts.

In this section I first describe the task in detail, next I discuss the classification training process, starting with data preprocessing, followed by feature extraction, document structure encoding, and ending with classifier construction.

Task Definition

As noted above, the task was to index academic articles in the biogeochemical domain with concepts derived from ENVO. That is, given a set of academic articles and a domain-specific ontology, the solution is a supervised classification model that can assign ontology concepts to the sentences found in the articles. Additionally, the task included testing the efficacy and efficiency of using the section strucutre of scientific articles for automatic concept learning. As discussed before, we created a dataset of articles which was manually labeled and indexed with concepts from ENVO. Each sentence may have any number of concepts and therefore the labels are not mutually exclusive and my solution must admit a multi-label classification, including possibly no label. I identified a set of distinctive features to support this classification, and designed feature extractors to compute these features over article text and included the document section structure as a feature.

Data Preprocessing

In addition to annotating the data with ENVO concepts as described in the previous sections and in chapter 2, I performed standard NLP preprocessing tasks to prepare the data for feature extraction and supervised learning. Similar to the previous study [Banisakher et al., 2018b] (discussed in §5.2), I cleaned the text by removing in-text citations and stand-alone mathematical, chemical, and biological formulas. I then tagged each token with its part-of-speech [Bird and Loper, 2004], lemmatized tokens using WordNet [Fellbaum, 1998], filtered known stop words using PubMed's list [PubMed Help, 2005], and used the pywsd module to perform word-sense disambiguation [Tan, 2014] to tag words with WordNet senses.

Data Balancing

The articles included 261 unique concepts across 5,562 occurrences. More than half of these occurrences (3,318) represented only 10 concepts, while the most frequent 50 concepts (26% of the total) occurred 4,728 times in total. Additionally, 98 concepts (37%) appeared only once. When supervised ML is performed over such distributions, they tend to overfit the classes with higher number of examples. Several solutions have been proposed and used for the problem of imbalanced data such as sampling (undersampling and oversampling) and weight assignment. These techniques are used to help supervised ML classifiers learn more about a class that has a significantly smaller number of examples relative to others. Similar to the previous study [Banisakher et al., 2018b], I opted to use the Synthetic Minority Over-sampling Technique (SMOTE) [Chawla et al., 2002]. SMOTE is a hybrid sampling technique that oversamples the minority classes while undersampling the majority classes. I applied resampling to the training set only, leaving the testing set with the original distribution.

Feature Extraction

Identifying a useful set of features is integral for an accurate machine learning model. In my previous study, I extracted lexical, syntactic, and semantic features (§5.2.4). In this extended study, I used the same features as well as additional features in each feature class as follows:

Features from Banisakher et al. [2018b] (§5.2.4)**:** *tf-idf* (term frequency-inverse document frequency) [Church and Gale, 1999] I used the top 10% of the most frequent distinctive terms for each article using; *global-position* and *local-position* corresponding to the relative position of a sentence in both its section and article; *named-entities* for which I extracted named entities from each sentence by examining parts-of-speech examining parts-of-speech (looking for runs of tokens tagged NNP or NNPS); *sentence-embeddings* for which I used sense2vec [Trask et al., 2015], where different senses of the same word are placed differently in the embedding space. I used Sense2vec as implemented in the SpaCy python module [AI, 2015], and followed the algorithm described in [Trask et al., 2015] by using the part-of-speech tags and named entity labels assigned to the tokens. Additionally, I merged named entities into single tokens (using hash symbols), so that they were assigned a single vector; *abstract-concepts* where I used concepts identified in the abstract of each article as features for the body of the article. Banisakher et al. (§5.2.4) also used the previous sentence's concept as an explicit feature; this is included by default in the CRF model.

Lexical: I added the *bigrams* feature to capture the type of language per section type; and *reference* where the number of citations or referred tables and figures were counted in a given sentence.

Syntactic: I added *gramatical-relation* where I captured the following: Subject (ncsubj), direct object (dobj), indirect object (iobj) and second object (obj2) relations involving verbs, e.g. (ncsubj observed difference obj).

Structural: To capture and encode the section structure of the scientific articles, I used my earlier work for section type discovery demonstrated in chapter 4 which tags runs of sentences with a section heading. After running the model merging approach and obtaining a section heading for each sentence, I tagged each concept to be learned with a section heading number.

Concept Learning

The first stage of classification is model training, followed by a stage of testing on separate (unseen) data. As I did in the previous study, I randomly split the dataset into into five folds, 80% in the training set and 20% in testing set (15 and 4 articles, respectively). I built and trained the concept learning models using a simple linear chain conditional random field (CRF) model.

Conditional Random Fields (CRFs) are undirected graphical models [Lafferty et al., 2001, Konkol and Konopík, 2013] that can be used for discriminative sequence labeling. CRFs have proved useful for many sequence labeling problems in NLP and computer vision [Lin and Wu, 2009], including Named Entity Recognition (NER) and image classification. There are several CRF variations such as the tree CRF and the hierarchical CRF which are mostly used for computer vision related tasks. Linear chain CRFs are the most popular among CRF approaches for sequence labeling tasks largely due to its relative simplicity and low computational cost when compared with other CRF models.

I built and trained two separate models using the features discussed in the previous section—a *body-only* model, which used all features, and an *abstract-only* model, which omitted the abstract concept features as well as the sentence counts and position features. This two-model approach attempts to mimic how human read scientific articles, namely, using the concepts found in the abstract to better guide the understanding concepts found in the rest of the text.

With regard to the CRF implimentation, I used the python CRFsuite [Okazaki, 2007] utlizing the 1st-order Markov CRF. As for model optimization and parameter estimiation, I used the L-BFGS [Nocedal, 1980] which is a quazi-Newton method that computes an approximation to the Hessian from only the first derivative of the objective function, and has been shown to be successful in parameter estimation and optimization.

5.3.4 Results and Discussion

As discussed above, I randomly split the dataset into training and testing sets across five folds, resulting in 15 articles for training and 4 for testing in each fold. The models learned a total of 261 unique concepts. For all experiments, I evaluated the performance of the models on each concept using the F_1 measure averaged across all folds. Here I present the evaluation methods and results, describing the baseline approaches, as well as the performance of both the baselines and my method average, averaged across the test sets.

Baseline Methods

I compared my CRF approach against six other models: three baselines including two from my previous study [Banisakher et al., 2018b] (§5.2.5), namely, a keyword search approach, where I matched sentence words directly to the names of ontology concepts, and a support vector machine (SVM) [Cortes and Vapnik, 1995] using Bag of Words (BoW) as the sole feature. I added a third baseline using another SVM model that incorported sense2vec as it's base feature (see discussion on sense2vec in the previous section and in section 5.2.4). I trained and tested the SVM models following the same cross-validation splits and multi-label fashion as used for my proposed approach. The SVM classifier was trained using the RBF kernel function and a soft margin C of 10,000—a common setup.

Additionally, I compared my approach against two random decision tree models: for the first, I used the same set of features outlined from my previous RDF study [Banisakher et al., 2018b], while for the second, I used the same features as the CRF model. This is to test wether the CRF performs better at capturing the inter-dependencies between sequences of sentences and concepts. Finally, for the the sixth comparison, I trained and tested a Long Short-Term Memory (LSTM) recurrent nueral network (RNN) with the same features as the CRF model as discussed previously in feature extraction.

Results

As noted above I built two models: (1) an *abstract-only* model, and (2) a *body-only* model. Both the models learn concepts using all sentences in the text (including the abstract), but as the names suggest, they were used to label the abstract sentences only and the body sentences only, respectively. Additionally, the the *body-only* model uses the labels produced by the *abstract-only* model as features for labeling the body of an article.

Table 5.5 shows three average F_1 scores over different sets of concepts for all discussed approaches. The first column shows the average F_1 score for the concepts with single occurrence in the original data (98 concepts), while the second column shows the average scores for the top 50 concepts in terms of total occurrences over all the articles. The last column shows the results over all concepts.

The CRF approach with the additional features outperforms the three baselines as well as the previous RDF model with the previous set of features across all concepts. Notably, the CRF model outperforms the RNN-LSTM model—an expected result as the size of data is not enough for a deep learning model. Additionally, the results show that the new features in this study perform much better that the previously discussed ones in [Banisakher et al., 2018b]. This is evident in the RDF model that uses the current features which achieved an 11% increase in performance over the previous RDF model.

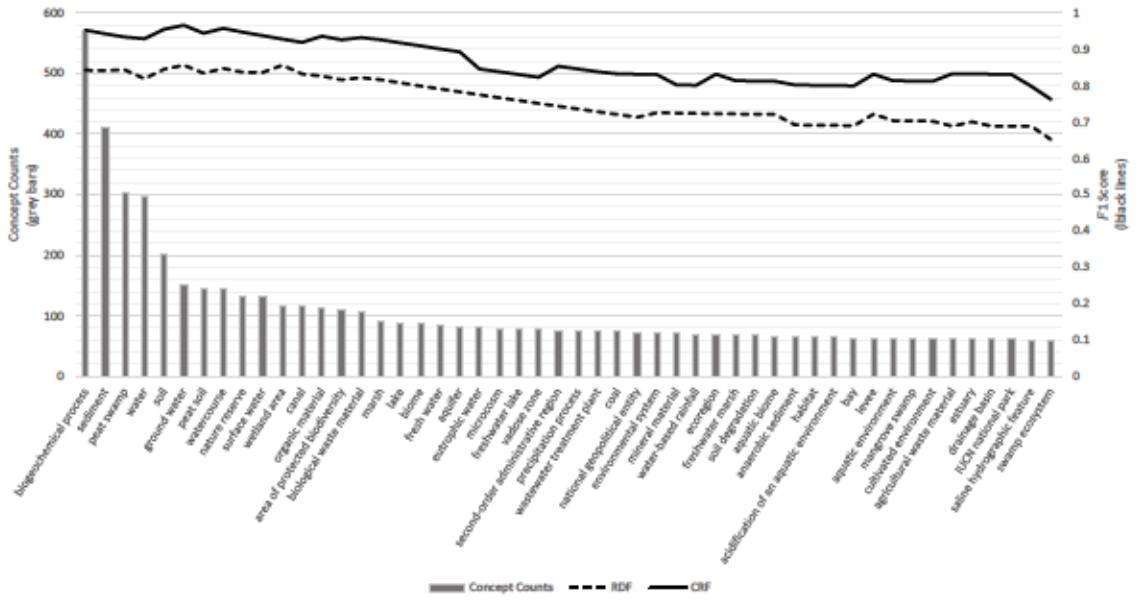


Figure 5.2: Frequency of the top 50 ENVO concepts (grey bars) and the average 5-fold testing results (F_1 scores) for each of the concepts (black lines). The solid line represents the CRF model, while the dotted line represents the RDF model with features from the previous study [Banisakher et al., 2018b]

Additionally, Figure 5.2 shows the frequency of the top 50 ENVO concepts as well as the average F_1 score for both the CRF model and the previous RDF model with features from [Banisakher et al., 2018b] for each of the concepts. As shown, the scores drops with the frequency of the concept in the dataset, although not dramatically. This is expected as it is a result of the original class imbalance. Finally, the *abstract-only* model performed similarly well with a 0.79 F_1 over all concepts present in the abstract sections, which were relatively small in number.

As discussed above, I expected that the model's performance would improve when (1) abstract concepts are used as features for the body concept extraction, and (2) structural features (i.e the section structure encoding and sentence positional features) are encoded as features in the model. To evaluate this, I performed four experiments, testing the inclusion of abstract concepts and structural features. Table 5.6 shows three average F_1 scores over different sets of concepts per experiment.

Approach	Features	Unique Concepts		
		Single	Top 50	All
Keyword Search	-	0.39	0.36	0.38
SVM	BoW	0.43	0.55	0.49
SVM	[Banisakher et al., 2018b]	0.69	0.75	0.73
RDF	[Banisakher et al., 2018b]	0.70	0.76	0.74
RNN-LSTM	Current features (§5.3.3)	0.72	0.78	0.76
RDF	Current features (§5.3.3)	0.78	0.84	0.82
CRF	Current features (§5.3.3)	0.81	0.87	**0.85**

Table 5.5: Average F_1 scores per approach over all concepts, the 50 most frequent concepts, and the 61 least frequent concepts with single occurrences.

Features		Unique Concepts		
Abstract	Structural	Single	Top 50	All
Omitted	Omitted	0.56	0.69	0.64
Omitted	Included	0.70	0.78	0.75
Included	Omitted	0.70	0.82	0.77
Included	Included	0.81	0.87	**0.85**

Table 5.6: Average F_1 scores for the feature combination experiments over the CRF model. The first two columns indicate whether the abstract concepts and structural features were included or omitted as features in the models.

Similar to my results discussion in §5.2.5, in the last row in Table 5.6, both abstract concepts and sentence positions were included as features in the models. The results confirm my hypothesis, in that the inclusion of both of those features yields better labeling results across all concepts.

The second-to-last row shows the results for including the abstract concepts as features, but omitting the structural features. This resulted in lower results overall, but significantly impacted the average score for the single occurrence concepts. From further investigation, I found that deeper concepts (i.e., in terms of the ontology hierarchy) are found at higher densities close to the middle of the articles as well as the centers of article sections. In retrospect this makes sense, as the methodology section of a scientific paper (located around the middle) would normally contain detailed concepts rather than abstract

ones. To put this together, most of the single occurrence concepts are deeper, low-level concepts, hence the low occurrence frequency in the original data.

The second row shows the results for only omitting the abstract concepts as features when labeling the rest of the text in the articles. Again, the models' performance dropped overall, but less so than for single occurrence concepts. This can be attributed again to including the structural features, which further shows the efficacy for the labeling of less frequent concepts.

Finally, the first row shows the results for omitting both features with the models performing the worst across all concepts.

CHAPTER 6

CONCLUSION

As discussed in chapter 1 My research problem, in large, concerns the analysis and development of a framework for using logical document structure knowledge (that is, *section structure*) in detecting semantic concepts within documents in various domains. This entailed four abstract conceptual components which have driven my research, namely, corpora collection and annotation, modeling the logical document structure, detecting semantic concepts through the use of domain-specific ontologies, and incorporating section structure in the detection of semantic concepts.

In chapter 2, I presented the corpora I used for the development and evaluation of the models I discussed in the dissertation. These corpora consisted of documents from six different datasets spanning four domains: medical, legal,scientific, and news reporting.The document classes are as follows: psychiatric report evaluations, hospital discharge summaries, and radiology reports in the medical domain; Patent documents in the legal domain; environmental journal articles in the scientific domain; Finally, business and politics news articles. There, I discussed each of these document classes, and the specific corpora I used or collected. I also discussed the corpora ontologies and report detailed statistics as well as the annotation process, agreement metrics, and annotation results for each corpus.

In chapter 3, I presented three studies for section structure identification. The first (§3.1), uses an Hierarchical Hidden Markov Model (HHMM) that was developed using the psychiatric evaluation reports (Corpus 2.1). The second (§3.2), extends the HHMM approach by using Conditional Random Fields (CRFs) which I developed using three corpora: psychiatric evaluation reports, radiology reports, and discharge summaries (Corpora 2.1-2.3). Finally, in the third (§3.3), I present an extended application of the CRF

approach to improving the detection of paragraph functions in news article paragraphs (Corpus 2.6).

To the best of my knowledge, my work presented in section 3.1 represents the only attempt at detecting the position and type of psychiatric report sections. In that section I presented an approach that applies and extends earlier work on document section discovery and segmentation. I collected a corpus of psychiatric documents and created a unified hierarchy of section labels. I built an n-gram-based HHMM model that successful detects the order of sections as well as their boundaries within a given report. I evaluated the model's performance over two separate tasks, namely the section ordering task and the section boundary identification. My model outperformed baselines for both of those tasks. Finally, my approach further confirms that learning the section ordering of a psychiatric report yields better performance for boundary identification and text segmentation.

I also presented an approach that extends the HHMM work on section identification of clinical reports in section 3.2. My CRF model of section structure can be used to identify sections in a variety clinical reports types regardless of whether headings are present: that is, whether the section headings are explicit or implicit. We built a linear chain CRF model incorporating n-gram features that successfully detects the order of sections as well as their boundaries within a given report. I evaluated my model's performance with regard to to two subtasks, namely determining the section ordering and locating the section boundaries, using different combinations of features. Additionally, my approach further confirms that learning a combined model of section ordering and section content yields better performance on the overall task. Finally, I demonstrated that modeling dependencies between sections' presence, order, and content across the entire report yields significantly better performance.

Finally, I presented a study that extended earlier work on news paragraph discourse function labeling. I built a linear chain CRF model incorporating various lexical, posi-

tional, syntactic, and semantic features that improves detection of the order of discourse labels in a news article at the paragraph level as well as models the paragraph content of each label type. I evaluated my model's performance against two baselines and three existing models with various subsets of features. I showed that the CRF model represents a significant improvement in this task. Most importantly, my work demonstrated the importance of modeling paragraph and discourse label type inter-dependencies.

In chapter 4 I demonstrated the first approach to discovering, in a data-driven manner, the section structure for a document class. My approach uses a modified Bayesian model merging algorithm [Stolcke and Omohundro, 1994, Finlayson, 2016], which outperforms three baselines across five different document classes by significant margins.

On another tangent, in chapter 5, I first presented a literature review of state-of-the-art approaches to academic search, specifically focusing on *semantic* academic search and contrasting that with keyword search approaches. I then discussed two studies on ontology-based semantic concept detection of scientific articles. In section 5.2, I presented a system for learning to identify domain-specific ontology concepts in the academic literature, specifically for the biogeochemical domain. I created a dataset of academic articles that a team of researchers and I manually annotated. I then used the annotated dataset to build a supervised machine learning model—a random decision forest classifier—which was trained and tested using cross-validation. Further, I identified a set of useful features and evaluated their efficacy in training and testing the models. The RDF model significantly outperformed the the baseline methods discussed. Finally, in section 5.3, I demonstrated an extended approach for semantic concept learning for the scientific literature. I used a conditional random fields model with an extended set of features that demonstrated the efficacy and efficiency of structural features for concept learning. Importantly, I incorporated my earlier approach on section type discovery of scientific articles and demonstrated that it significantly improves the classification results.

This highlights the importance of using information embedded in document structure in semantic search tasks.

Finally, my research advances NLU and automatic semantic extraction research through the creation of novel domain-specific document structure understanding and concept detection models. As discussed earlier digitized data is ever-growing, and that includes medical, legal, journalistic and scientific documents [Feldman et al., 2012]. Published scientific articles and human knowledge is no different. In fact, according to IBM the human knowledge curve has moved from being linear to exponential and it is expected that soon (a few years) human knowledge will be doubling every 12 hours [Coles et al., 2006].

More specifically, this research has an expected impact in several NLP subdomains and applications. To name a few: semantic search is a direct example of the applications that benefit from such research as keyword-based search is becoming obsolete given the amount of available information and its well-known deficiencies in identifying and retrieving relevant results. While automatic summarization of documents is another example. Applications in this subdomain and relevant to this work would include automatic summarization of psychiatric reports, for example, to allow medical practitioners to have more time listening to patients rather than reading documents describing them. While scientific article summarization models can automatically generate abstracts as well as be used to semantic plagiarism analyzers.

BIBLIOGRAPHY

Ronald T Kellogg. *The psychology of writing*. Oxford University Press, 1999.

Ruogu Fang, Samira Pouyanfar, Yimin Yang, Shu-Ching Chen, and SS Iyengar. Computational health informatics in the big data age: a survey. *ACM Computing Surveys (CSUR)*, 49(1):12, 2016.

Richard Power, Donia Scott, and Nadjet Bouayad-Agha. Document structure. *Computational Linguistics*, 29(2):211–260, 2003.

Scott G Paris, Barbara A Wasik, and Julianne C Turner. The development of strategic readers. 2016.

Rob Waller. What makes a good document? *University of Reading, United Kingdom, Tech. Rep*, 2011.

Walter Kintsch. The representation of meaning in memory. 1974.

Timothy McNamara, Diana L Miller, and John D Bransford. Mental models and reading comprehension. 1991.

Mary B McVee, Kailonnie Dunsmore, and James R Gavelek. Schema theory revisited. *Review of educational research*, 75(4):531–566, 2005.

Bengt Nordström. Towards a theory of document structure. 2008.

Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D Trippe, Juan B Gutierrez, and Krys Kochut. Text summarization techniques: a brief survey. *arXiv preprint arXiv:1707.02268*, 2017.

Lucas Drumond and Rosario Girardi. A survey of ontology learning procedures. *WONTO*, 427:1–13, 2008.

Dejing Dou, Hao Wang, and Haishan Liu. Semantic data mining: A survey of ontology-based approaches. In *Proceedings of the 2015 IEEE International Conference on Semantic Computing (ICSC)*, pages 244–251, Anaheim, CA, 2015.

Roberto Navigli and Paola Velardi. Learning domain ontologies from document warehouses and dedicated web sites. *Computational Linguistics*, 30(2):151–179, 2004.

Deya Banisakher, Naphtali Rishe, and Mark A. Finlayson. Automatically Detecting the Position and Type of Psychiatric Evaluation Report Sections. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 101–110, Brussels, Belgium, October 2018a. Association for Computational Linguistics. doi: 10.18653/v1/W18-5612. URL `https://www.aclweb.org/anthology/W18-5612`.

R Reeves and R Rosner. Forensic psychiatry and forensic psychology: Forensic psychiatric assessment. 2016.

American Psychiatric Association. What is psychiatry?, 2018. URL `https://www.psychiatry.org/patients-families/what-is-psychiatry`. (Accessed on Feb 10, 2019).

Gary Groth-Marnat. *Handbook of Psychological Assessment*. John Wiley & Sons, Hoboken, NJ, 2009.

Karen Goldfinger and Andrew M Pomerantz. *Psychological Assessment and Report Writing*. Sage, Thousand Oaks, CA, 2013.

American Psychiatric Association. *American Psychiatric Association Practice Guidelines for the Treatment of Psychiatric Disorders: Compendium 2006*. American Psychiatric Association Publishing, Washington, DC, 2006.

American Board of Radiology. Diagnostic Radiology, 2019. URL `https://www.theabr.org/diagnostic-radiology`. (Accessed: 20 Feb. 2020).

Michael Tepper, Daniel Capurro, Fei Xia, Lucy Vanderwende, and Meliha Yetisgen-Yildiz. Statistical section segmentation in free-text clinical records. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 2001–2008, 2012.

Leora I Horwitz, Grace Y Jenq, Ursula C Brewster, Christine Chen, Sandhya Kanade, Peter H Van Ness, Katy LB Araujo, Boback Ziaeian, John P Moriarty, Robert L Fogerty, et al. Comprehensive quality of discharge summaries at an academic medical center. *Journal of hospital medicine*, 8(8):436–443, 2013.

Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035–160035, 2016.

WIPO. *WIPO Patent Drafting Manual*. World Intellectual Property Organization, Geneva, Switzerland, 2007.

Sören Brügmann, Nadjet Bouayad-Agha, Alicia Burga, Serguei Carrascosa, Alberto Ciaramella, Marco Ciaramella, Joan Codina-Filba, Enric Escorsa, Alex Judea, Simon Mille, et al. Towards content-oriented patent document processing: intelligent patent analysis and summarization. *World Patent Information*, 40:30–42, 2015.

WIPO. PATENTSCOPE - World Intellectual Property Organization. `https://www.wipo.int/patentscope/en/`, 2019. Accessed: 20 May 2019.

Susanne M Ullrich, Trevor W Tanton, and Svetlana A Abdrashitova. Mercury in the Aquatic Environment: A Review of Factors Affecting Methylation. *Critical Reviews in Environmental Science and Technology*, 31(3):241–293, 2001.

Janina M Benoit, Cynthia C Gilmour, Robert P Mason, and Andrew Heyes. Sulfide controls on mercury speciation and bioavailability to methylating bacteria in sediment pore waters. *Environmental Science & Technology*, 33(6):951–957, 1999.

Cynthia C Gilmour, Elizabeth A Henry, and Ralph Mitchell. Sulfate stimulation of mercury methylation in freshwater sediments. *Environmental Science & Technology*, 26 (11):2281–2287, 1992.

Geoffrey C Compeau and Richard Bartha. Effect of salinity on mercury-methylating activity of sulfate-reducing bacteria in estuarine sediments. *Applied and Environmental Microbiology*, 53(2):261–265, 1987.

Harold L Drake, Nicholas G Aumen, Carla Kuhner, Christine Wagner, Anja Griesshammer, and Martina Schmittroth. Anaerobic microflora of Everglades sediments: Effects of nutrients on population profiles and activities. *Applied and Environmental Microbiology*, 62(2):486–493, 1996.

JM Benoit, Robert P Mason, Cynthia C Gilmour, and George R Aiken. Constants for mercury binding by dissolved organic matter isolates from the Florida Everglades. *Geochimica et cosmochimica acta*, 65(24):4445–4451, 2001.

Lisa B Cleckner, Cynthia C Gilmour, James P Hurley, and David P Krabbenhoft. Mercury methylation in periphyton of the Florida Everglades. *Limnology and Oceanography*, 44(7):1815–1825, 1999.

Cynthia C Gilmour, GS Riedel, MC Ederington, JT Bell, GA Gill, and MC Stordal. Methylmercury concentrations and production rates across a trophic gradient in the northern Everglades. *Biogeochemistry*, 40(2-3):327–345, 1998.

Mark C Marvin-DiPasquale and Ronald S Oremland. Bacterial methylmercury degradation in Florida Everglades peat sediment. *Environmental Science & Technology*, 32 (17):2556–2563, 1998.

Judson W Harvey and Paul V McCormick. Groundwater's significance to changing hydrology, water chemistry, and biological communities of a floodplain ecosystem, Everglades, South Florida, USA. *Hydrogeology Journal*, 17(1):185–201, 2009.

Randolph M Chambers and Kristin A Pederson. Variation in soil phosphorus, sulfur, and iron pools among South Florida wetlands. *Hydrobiologia*, 569(1):63–70, 2006.

William Orem, Cynthia Gilmour, Donald Axelrad, David Krabbenhoft, Daniel Scheidt, Peter Kalla, Paul McCormick, Mark Gabriel, and George Aiken. Sulfur in the South Florida ecosystem: Distribution, sources, biogeochemistry, impacts, and management for restoration. *Critical Reviews in Environmental Science and Technology*, 41 (S1):249–288, 2011.

Daniel J Casagrande, Kristine Siefert, Charles Berschinski, and Nell Sutton. Sulfur in peat-forming systems of the Okefenokee swamp and Florida Everglades: Origins of sulfur in coal. *Geochimica et Cosmochimica Acta*, 41(1):161–167, 1977.

Forrest E Dierberg, Thomas A DeBusk, Nichole R Larson, Michelle D Kharbanda, Nancy Chan, and Mark C Gabriel. Effects of sulfate amendments on mineralization and phosphorus release from South Florida (USA) wetland soils under anaerobic conditions. *Soil Biology and Biochemistry*, 43(1):31–45, 2011.

Rene M Price, Peter K Swart, and James W Fourqurean. Coastal groundwater discharge–an additional source of phosphorus for the oligotrophic wetlands of the everglades. *Hydrobiologia*, 569(1):23–36, 2006.

Pamela L Sullivan, René M Price, Jessica L Schedlbauer, Amartya Saha, and Evelyn E Gaiser. The influence of hydrologic restoration on groundwater-surface water interactions in a karst wetland, the everglades (fl, usa). *Wetlands*, 34(1):23–35, 2014.

Judson W Harvey, Steven L Krupa, and James M Krest. Ground water recharge and discharge in the central everglades. *Groundwater*, 42(7):1090–1102, 2004.

Xavier Zapata-Rios and René M Price. Estimates of groundwater discharge to a coastal wetland using multiple techniques: Taylor slough, everglades national park, usa. *Hydrogeology Journal*, 20(8):1651–1668, 2012.

Jungyill Choi and Judson W Harvey. Quantifying time-varying ground-water discharge and recharge in wetlands of the northern florida everglades. *Wetlands*, 20(3):500–511, 2000.

Joshua D Eisenberg, Deya Banisakher, Maria Presa, Kalli Unthank, Mark A Finlayson, Rene Price, and Shu-Ching Chen. Toward semantic search for the biogeochemical literature. In *Proceedings of the 2017 IEEE International Conference on Information Reuse and Integration (IRI)*, pages 517–525, San Diego, CA, 2017.

Pier Luigi Buttigieg, Norman Morrison, Barry Smith, Christopher J Mungall, and Suzanna E Lewis. The environment ontology: Contextualizing biological and biomedical entities. *Journal of Biomedical Semantics*, 4(1):43, 2013.

Teun A van Dijk. *News as Discourse*, chapter Structure of News, pages 52–57. Lawrence Erlbaum Associates, Inc., Hillsdale, New Jersey, USA, 1988.

W. Victor Yarlott, Cristina Cornelio, Tian Gao, and Mark Finlayson. Identifying the discourse function of news article paragraphs. In *Proceedings of the Workshop Events and Stories in the News 2018*, pages 25–33, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/W18-4304`.

NIST. Ace phase 2, 2002. URL `https://www.ldc.upenn.edu/collaborations/past-projects/ace/annotation-tasks-and-specifications`.

American College of Radiology et al. Practice parameters and technical standards, 2018.

Mark A. Musen. The Protégé project: A look back and a look forward. *AI Matters*, 1(4): 4–12, 2015.

Simone Teufel and Marc Moens. Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational linguistics*, 28(4):409–445, 2002.

Yoko Mizuta, Anna Korhonen, Tony Mullen, and Nigel Collier. Zone analysis in biology articles as a basis for information extraction. *International journal of medical informatics*, 75(6):468–487, 2006.

Yufan Guo, Ilona Silins, Ulla Stenius, and Anna Korhonen. Active learning-based information structure analysis of full scientific articles and two applications for biomedical literature review. *Bioinformatics*, 29(11):1440–1447, 04 2013. ISSN 1367-4803. doi: 10.1093/bioinformatics/btt163. URL `https://doi.org/10.1093/bioinformatics/btt163`.

J Richard Landis and Gary G Koch. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, pages 363–374, 1977.

Ron Artstein and Massimo Poesio. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596, 2008. doi: 10.1162/coli.07-034-R2.

Cornelis J. van Rijsbergen. *Information retrieval*. Butterworths, London Boston, 1979. ISBN 0-408-70929-4.

Jacob Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213, 1968.

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

George Hripcsak, Suzanne Bakken, Peter D Stetson, and Vimla L Patel. Mining complex clinical data for patient safety research: A framework for event discovery. *Journal of Biomedical Informatics*, 36(1-2):120–130, 2003.

Ying Li, Sharon Lipsky Gorman, and Noémie Elhadad. Section Classification in Clinical Notes Using Supervised Hidden Markov Model. In *Proceedings of the 1st ACM International Health Informatics Symposium IHI*, pages 744–750, Arlington, Virgina, USA, 2010.

M. Sherman and Yang Liu. Using Hidden Markov Models for Topic Segmentation of Meeting Transcripts. In *Proceedings of the 2008 IEEE Spoken Language Technology Workshop*, pages 185–188, Goa, India, 2008.

Regina Barzilay and Lillian Lee. Catching the drift: Probabilistic content models, with applications to generation and summarization. In *Proceedings of the 2004 North American Chapter of the Association for Computational Linguistics: Human Language Technologies Conference (HLT-NAACL)*, pages 113–120, 2004.

L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989. doi: 10.1109/5.18626.

Hung H. Bui, Dinh Q. Phung, and Svetha Venkatesh. Hierarchical hidden markov models with general state hierarchy. In *Proceedings of the 19th National Conference on Artifical Intelligence*, AAAI'04, pages 324–329, San Jose, California, 2004. ISBN 0-262-51183-5. URL `http://dl.acm.org/citation.cfm?id=1597148.1597202`.

Shai Fine, Yoram Singer, and Naftali Tishby. The hierarchical hidden markov model: Analysis and applications. *Machine Learning*, 32(1):41–62, 1998. doi: 10.1023/A:1007469218079.

Kush Jain, Priya Khatri, and Garima Indolia. Chunked n-grams for sentence validation. *Procedia Computer Science*, 57:209–213, 2015.

Jeffrey C Reynar. *Topic Segmentation: Algorithms and Applications.* PhD thesis, University of Pennsylvania, Philadelphia, PA, 1998.

eMedicineHealth. Medications and drugs listing. `https://www.emedicinehealth.com/medications-drugs/article_em.htm`, 2018. (Accessed on Feb 18, 2018).

S. Liu, Wei Ma, R. Moore, V. Ganesan, and S. Nelson. Rxnorm: Prescription for electronic drug information exchange. *IT Professional*, 7(5):17–23, Sept 2005. doi: 10.1109/MITP.2005.122.

Lev Pevzner and Marti A Hearst. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36, 2002.

Doug Beeferman, Adam Berger, and John Lafferty. Statistical models for text segmentation. *Machine Learning*, 34(1):177–210, 1999.

Michel Galley, Kathleen McKeown, Eric Fosler-Lussier, and Hongyan Jing. Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics (ACL)*, volume 1, pages 562–569, Sapporo, Japan, 2003.

Marti A. Hearst. Multi-paragraph segmentation of expository text. In *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics*, ACL '94, pages 9–16, Las Cruces, NM, 1994.

Martin Riedl and Chris Biemann. Topictiling: A text segmentation algorithm based on lda. In *Proceedings of ACL 2012 Student Research Workshop*, ACL '12, pages 37–42, Jeju Island, Korea, 2012. URL `http://dl.acm.org/citation.cfm?id=2390331.2390338`.

Simone Teufel, Jean Carletta, and Marc Moens. An annotation scheme for discourse-level argumentation in research articles. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics (EACL)*, pages 110–117, Bergen, Norway, 1999.

Joshua C. Denny, Anderson Spickard, III, Kevin B. Johnson, Neeraja B. Peterson, Josh F. Peterson, and Randolph A. Miller. Evaluation of a method to identify and categorize section headers in clinical documents. *Journal of the American Medical Informatics Association*, 16(6):806–815, 2009a.

Simone Teufel. *Argumentative zoning: Information Extraction from Scientific Text.* PhD thesis, University of Edinburgh, Edinburgh, Scotland, UK, 1999.

Paul van Mulbregt, Ira Carp, Lawrence Gillick, Steve Lowe, and Jon Yamron. Text Segmentation and Topic Tracking on Broadcast News Via a Hidden Markov Model Approach. In *Fifth International Conference on Spoken Language Processing*, ICSLP '98, Sydney, Australia, 1998.

J. P. Yamron, I. Carp, L. Gillick, S. Lowe, and P. van Mulbregt. A Hidden Mrkov Model Approach to Text Segmentation and Event Tracking. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP '98, pages 333–336 vol.1, Seattle, Washington, USA, May 1998.

Jia Yu, Xiong Xiao, Lei Xie, Chng Eng Siong, and Haizhou Li. A DNN-HMM Approach to Story Segmentation. In *INTERSPEECH 2016*, San Francisco, California, USA, 2016.

Wendy W Chapman, Prakash M Nadkarni, Lynette Hirschman, Leonard W D'Avolio, Guergana K Savova, and Ozlem Uzuner. Overcoming barriers to NLP for clinical text: The role of shared tasks and the need for additional creative solutions. *Journal*

of the American Medical Informatics Association, 18(5):540–543, 2011. doi: 10.1136/amiajnl-2011-000465.

John P. Pestian, Pawel Matykiewicz, Michelle Linn-Gust, Brett South, Ozlem Uzuner, Jan Wiebe, K. Bretonnel Cohen, John Hurdle, and Christopher Brew. Sentiment analysis of suicide notes: A shared task. *Biomedical Informatics Insights*, 5s1:BII.S9042, 2012.

Carlo Strapparava and Rada Mihalcea. Learning to identify emotions in text. In *Proceedings of the ACM Symposium on Applied Computing (SAC)*, SAC '08, pages 1556–1560, Fortaleza, Ceara, Brazil, 2008. URL `http://doi.acm.org/10.1145/1363686.1364052`.

Christopher Homan, Ravdeep Johar, Tong Liu, Megan Lytle, Vincent Silenzio, and Cecilia Ovesdotter Alm. Toward macro-insights for suicide prevention: Analyzing fine-grained distress at scale. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 107–117, 2014.

Bridianne O'Dea, Stephen Wan, Philip J. Batterham, Alison L. Calear, Cecile Paris, and Helen Christensen. Detecting suicidality on twitter. *Internet Interventions*, 2:183–188, 4 2015. doi: 10.1016/j.invent.2015.03.005.

Adam Sadilek, Christopher Homan, Walter S Lasecki, Vincent Silenzio, and Henry Kautz. Modeling fine-grained dynamics of mood at scale. *WSDM, Rome, Italy*, pages 3–6, 2013.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting depression via social media. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM)*, volume 13, pages 1–10, Boston, MA, 2013.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. CLPsych 2015 shared task: Depression and PTSD on twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality (CLPsych)*, pages 31–39, 2015.

Lorenzo Coviello, Yunkyu Sohn, Adam D. I. Kramer, Cameron Marlow, Massimo Franceschetti, Nicholas A. Christakis, and James H. Fowler. Detecting emotional contagion in massive social networks. *PLOS ONE*, 9(3):1–6, 2014. doi: 10.1371/journal.pone.0090315.

Munmun De Choudhury, Scott Counts, Eric J. Horvitz, and Aaron Hoff. Characterizing and predicting postpartum depression from shared facebook data. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '14, pages 626–638, Baltimore, MD, 2014.

Tim Althoff, Kevin Clark, and Jure Leskovec. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics*, 4:463–476, 2016.

José A Reyes-Ortiz, Beatriz A González-Beltrán, and Lizbeth Gallardo-López. Clinical decision support systems: a survey of nlp-based approaches from unstructured data. In *2015 26th International Workshop on Database and Expert Systems Applications (DEXA)*, pages 163–167. IEEE, 2015.

Aurélie Névéol, Hercules Dalianis, Sumithra Velupillai, Guergana Savova, and Pierre Zweigenbaum. Clinical natural language processing in languages other than english: opportunities and challenges. *Journal of biomedical semantics*, 9(1):12, 2018.

Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, et al. Clin-

ical information extraction applications: a literature review. *Journal of biomedical informatics*, 77:34–49, 2018.

Kirk Roberts, Matthew Simpson, Dina Demner-Fushman, Ellen Voorhees, and William Hersh. State-of-the-art in biomedical literature retrieval for clinical cases: a survey of the trec 2014 cds track. *Information Retrieval Journal*, 19(1-2):113–148, 2016.

Michele Filannino and Özlem Uzuner. Advancing the state of the art in clinical natural language processing through shared tasks. *Yearbook of medical informatics*, 27(01): 184–192, 2018.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1558607781.

Michal Konkol and Miloslav Konopík. CRF-Based Czech named Entity Recognizer and Consolidation of Czech NER Research. In Ivan Habernal and Václav Matoušek, editors, *Text, Speech, and Dialogue*, pages 153–160, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

Dekang Lin and Xiaoyun Wu. Phrase Clustering for Discriminative Learning. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1030–1038, Suntec, Singapore, August 2009. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/P09-1116`.

Jorge Nocedal. Updating Quasi-Newton Matrices with Limited Storage. *Mathematics of computation*, 35(151):773–782, 1980.

Kenneth Church and William Gale. Inverse document frequency (idf): A measure of deviations from Poisson. In *Natural Language Processing Using Very Large Corpora*, pages 283–295. Springer, New York, 1999.

Emanuel Falkenauer. The grouping genetic algorithms-widening the scope of the gas. *Belgian Journal of Operations Research, Statistics and Computer Science*, 33(1):2, 1992.

LE Agustı, Sancho Salcedo-Sanz, Silvia Jiménez-Fernández, Leopoldo Carro-Calvo, Javier Del Ser, José Antonio Portilla-Figueras, et al. A new grouping genetic algorithm for clustering problems. *Expert Systems with Applications*, 39(10):9695–9703, 2012.

Martin Warin and HM Volk. Using wordnet and semantic similarity to disambiguate an ontology. *Retrieved January*, 25:2008, 2004.

G.D. Forney. The Viterbi Algorithm. *Proceedings of the IEEE*, 61(3):268–278, 1973.

Stanley F Chen and Ronald Rosenfeld. A Gaussian Prior for Smoothing Maximum Entropy Models. Technical report, Carnegie-Mellon University, School of Computer Science, Pittsburgh, Pennsylvania, USA, 1999.

Filip Ginter, Hanna Suominen, Sampo Pyysalo, and Tapio Salakoski. Combining hidden markov models and latent semantic analysis for topic segmentation and labeling: Method and clinical application. *International journal of medical informatics*, 78 (12):e1–e6, 2009.

Katja Hofmann, Manos Tsagkias, Edgar Meij, and Maarten De Rijke. The impact of document structure on keyphrase extraction. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 1725–1728. ACM, 2009.

Dragos Repta, Ioan Stefan Sacala, Mihnea Alexandru Moisescu, and Ioan Dumitrache. Towards document flow discovery in e-health systems. In *2018 International Conference on Intelligent Systems (IS)*, pages 267–271. IEEE, 2018.

Jian Wu, Kyle Mark Williams, Hung-Hsuan Chen, Madian Khabsa, Cornelia Caragea, Suppawong Tuarob, Alexander G Ororbia, Douglas Jordan, Prasenjit Mitra, and C Lee Giles. CiteSeerX: AI in a digital library search engine. *AI Magazine*, 36 (3):35–48, 2015a.

Jinxi Xu and W Bruce Croft. Quary expansion using local and global document analysis. In *Acm sigir forum*, number 2, pages 168–175. ACM, 2017.

Antoine Doucet. Logical structure extraction from digitized books. *Document Analysis And Text Recognition: Benchmarking State-of-the-art Systems*, 82:1, 2018.

Andreas Stolcke and Stephen Omohundro. Inducing probabilistic grammars by bayesian model merging. In *International Colloquium on Grammatical Inference*, pages 106–118. Springer, 1994.

Michael Simmons, Ayush Singhal, and Zhiyong Lu. Text mining for precision medicine: bringing structure to ehrs and biomedical literature to understand genes and health. In *Translational Biomedical Informatics*, pages 139–166. Springer, 2016.

Sébastien Eskenazi, Petra Gomez-Krämer, and Jean-Marc Ogier. A comprehensive survey of mostly textual document segmentation algorithms since 2008. *Pattern Recognition*, 64:1–14, 2017.

Kavita Ganesan and Michael Subotin. A general supervised approach to segmentation of clinical texts. In *2014 IEEE International Conference on Big Data (Big Data)*, pages 33–40, 2014.

Alexandra Pomares-Quimbaya, Markus Kreuzthaler, and Stefan Schulz. Current approaches to identify sections within clinical narratives from electronic health records: a systematic review. *BMC medical research methodology*, 19(1):155, 2019.

Joshua C. Denny, III Spickard, Anderson, Kevin B. Johnson, Neeraja B. Peterson, Josh F. Peterson, and Randolph A. Miller. Evaluation of a Method to Identify and Categorize Section Headers in Clinical Documents. *Journal of the American Medical Informatics Association*, 16(6):806–815, 11 2009b. doi: 10.1197/jamia.M3037.

Emilia Apostolova, David S. Channin, Dina Demner-Fushman, Jacob Furst, Steven Lytinen, and Daniela Raicu. Automatic segmentation of clinical texts. In *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, volume 2009, pages 5905–5908, 2009.

Peter J. Haug, Xinzi Wu, Jeffery P. Ferraro, Guergana K. Savova, Stanley M. Huff, and Christopher G Chute. Developing a section labeler for clinical documents. In *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, volume 2014, pages 636–644, 2014.

Hong Jie Dai, Shabbir Syed-Abdul, Chih Wei Chen, and Chieh Chen Wu. Recognition and evaluation of clinical section headings in clinical documents using token-based formulation with conditional random fields. *BioMed Research International*, 2015: 873012–873012, 2015.

Quoc Le and Tomas Mikolov. Distributed Representations of Sentences and Documents. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1188–1196, Beijing, China, 2014.

Burton DeWilde. textacy, Mar 2020. URL `https://pypi.org/project/textacy/`.

Nathanael Chambers, Taylor Cassidy, Bill McDowell, and Steven Bethard. Dense Event Ordering with a Multi-Pass Architecture. *Transactions of the Association for Computational Linguistics*, 2:273–284, 2014. doi: 10.1162/tacl_a_00182. URL https://www.aclweb.org/anthology/Q14-1022.

Mohammed Aldawsari and Mark Finlayson. Detecting Subevents using Discourse and Narrative Features. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4780–4790, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1471. URL https://www.aclweb.org/anthology/P19-1471.

Explosion AI. Annotation Specifications-SpaCy API Documentation, 2020. URL https://spacy.io/api/annotation#named-entities.

Allan Bell. The Discourse Structure of News Stories. In *Approaches to Media Discourse*, pages 64–104. Blackwell Oxford, 1998.

Judy Delin. *The Language of Everyday Life: An Introduction*. Sage, London, UK, 2000.

Teun A Van Dijk. *Studying Writing: Linguistic Approaches. Written Communication Annual: An International Survey of Research and Theory Series, Volume 1.*, chapter News Schemata, pages 155–185. Sage, Beverly Hills, California, USA, 1986.

Allan Bell. Telling stories. In David Graddol and Oliver Boyd-Barrett, editors, *Media texts: Authors and readers*, pages 100–118. Multilingual Matters, Clevedon, U.K., 1994.

Afrooz Rafiee, Wilbert Spooren, and José Sanders. Culture and Discourse Structure: A Comparative Study of Dutch and Iranian News Texts. *Discourse & Communication*, 12(1):58–79, 2018. doi: 10.1177/1750481317735626. URL https://doi.org/10.1177/1750481317735626.

Rachele Sprugnoli, Tommaso Caselli, Sara Tonelli, and Giovanni Moretti. The content types dataset: a new resource to explore semantic and functional characteristics of texts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 260–266, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/E17-2042.

Mesfin Awoke Bekalu. Presupposition In News Discourse. *Discourse & Society*, 17(2): 147–172, 2006. doi: 10.1177/0957926506060248. URL https://doi.org/10.1177/0957926506060248.

Nynke van der Vliet, Ildikó Berzlánovich, Gosse Bouma, Markus Egg, and Gisela Redeker. Building a Discourse-Annotated Dutch Text Corpus. *Bochumer Linguistische Arbeitsberichte*, 3:157–171, 2011.

Zhongdang Pan and Gerald M Kosicki. Framing Analysis: An Approach to News Discourse. *Political Communication*, 10(1):55–75, 1993.

Peter R White. *Telling Media Tales: The News Story as Rhetoric*. Department of Linguistics, Faculty of Arts, University of Sydney, Sydney, Australia, 1998.

Mark Alan Finlayson. Inferring propp's functions from semantically annotated text. *The Journal of American Folklore*, 129(511):55–77, 2016. doi: 10.5406/jamerfolk.129.511.0055.

Felicity Jane Pool and Miranda Lynette Siemienowicz. New RANZCR clinical radiology written report guidelines. *Journal of Medical Imaging and Radiation Oncology*, 63 (1):7–14, 2019.

Barrou Diallo and Mihai Lupu. Future patent search. In *Current Challenges in Patent Information Retrieval*, volume 37, pages 433–455. Springer, 2017. doi: 10.1007/978-3-662-53817-3_17.

Zhibiao Wu and Martha Palmer. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138. Association for Computational Linguistics, 1994.

Andreas Stolcke and Stephen Omohundro. Hidden markov model induction by bayesian model merging. In *Advances in neural information processing systems*, pages 11–18, 1993.

Ehsan Hosseini-Asl and Jacek M Zurada. Nonnegative matrix factorization for document clustering: A survey. In *International Conference on Artificial Intelligence and Soft Computing*, pages 726–737. Springer, 2014.

Pengtao Xie and Eric P Xing. Integrating document clustering and topic modeling. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, pages 694–703. AUAI Press, 2013.

Irvin Hwang, Andreas Stuhlmüller, and Noah D Goodman. Inducing probabilistic programs by bayesian program merging. *arXiv preprint arXiv:1110.5667*, 2011.

Paolo Frasconi, Giovanni Soda, and Alessandro Vullo. Text categorization for multi-page documents: A hybrid naive bayes hmm approach. In *Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries*, pages 11–20. ACM, 2001.

Michael R Brent and Timothy A Cartwright. Lexical categorization: Fitting template grammars by incremental mdl optimization. In *International Colloquium on Grammatical Inference*, pages 84–94. Springer, 1996.

Cen Li and Gautam Biswas. Clustering sequence data using hidden markov model representation. In *Data Mining and Knowledge Discovery: Theory, Tools, and Technology*, volume 3695, pages 14–22. International Society for Optics and Photonics, 1999.

Zouhair Rimale, Abderrahim Tragha, et al. An approach for the automatic generation af a content type of a semantic learning object from ontology. In *2016 11th International Conference on Intelligent Systems: Theories and Applications (SITA)*, pages 1–6. IEEE, 2016.

Victoria Lesley Redfern. Enhanced searching using a thesaurus, September 27 2011. US Patent 8,027,991.

Christoph Mangold. A survey and classification of semantic search approaches. *International Journal of Metadata, Semantics and Ontologies*, 2(1):23–34, 2007.

Jeffrey Xu Yu, Lu Qin, and Lijun Chang. Keyword search in relational databases: A survey. *IEEE Data Eng. Bull.*, 33(1):67–78, 2010.

Max L Wilson, Bill Kules, Ben Shneiderman, et al. From keyword search to exploration: Designing future search interfaces for the web. *Foundations and Trends® in Web Science*, 2(1):1–97, 2010.

Shilpa S Laddha, Anurag R Laddha, and Pradip M Jawandhiya. New paradigm to keyword search: A survey. In *Green Computing and Internet of Things (ICGCIoT), 2015 International Conference on*, pages 920–923. IEEE, 2015.

Matthias Klusch, Patrick Kapahnke, Stefan Schulte, Freddy Lecue, and Abraham Bernstein. Semantic web service search: a brief survey. *KI-Künstliche Intelligenz*, 30(2): 139–147, 2016.

Yuchao Zhou, Suparna De, Wei Wang, and Klaus Moessner. Search techniques for the web of things: A taxonomy and survey. *Sensors*, 16(5):600, 2016.

Emmanouil Amolochitis. *Algorithms for Academic Search and Recommendation Systems*. PhD thesis, Aalborg Universitet, 2014.

Pricila R Rodrigues and Raquel O Prates. A semiotic study on academic search interfaces. In *Proceedings of the 15th Brazilian Symposium on Human Factors in Computing Systems*, page 53. ACM, 2016.

Madian Khabsa, Zhaohui Wu, and C Lee Giles. Towards better understanding of academic search. In *Digital Libraries (JCDL), 2016 IEEE/ACM Joint Conference on*, pages 111–114. IEEE, 2016.

Vannevar Bush. As we may think, July 1945. URL `https://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/`.

Eugene Garfield. Citation indexes for science; a new dimension in documentation through association of ideas. *Science (New York, NY)*, 122(3159):108, 1955.

Eugene Garfield. Science citation index"-a new dimension in indexing. *Science*, 144 (3619):649–654, 1964.

Clarivate Analytics. Acquisition of the thomson reuters intellectual property and science business by onex and baring asia completed, Oct 2016. URL `https://www.prnewswire.com/news-releases/acquisition-of-the-thomson-reuters-intellectual-property-and-scien` `html`.

Libguides: Librarian toolkit: Our history, October 2018. URL `https://clarivate.libguides.com/newlibrarian/history`.

Anne-Wil Harzing. Microsoft academic (search): a phoenix arisen from the ashes? *Scientometrics*, 108(3):1637–1647, 2016.

Scopus preview. URL `http://www.scopus.com/`.

Huajing Li, Isaac Councill, Wang-Chien Lee, and C Lee Giles. Citeseerx: an architecture and web service design for an academic document search engine. In *Proceedings of the 15th international conference on World Wide Web*, pages 883–884. ACM, 2006.

Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21, 1972.

Karen Sparck Jones. Index term weighting. *Information storage and retrieval*, 9(11): 619–633, 1973.

Duygu Tümer, Mohammad Ahmed Shah, and Yiltan Bitirim. An empirical evaluation on semantic search performance of keyword-based and semantic search engines: Google, yahoo, msn and hakia. In *2009 Fourth International Conference on Internet Monitoring and Protection*, pages 51–55. IEEE, 2009.

Matthew E Falagas, Eleni I Pitsouni, George A Malietzis, and Georgios Pappas. Comparison of pubmed, scopus, web of science, and google scholar: strengths and weaknesses. *The FASEB journal*, 22(2):338–342, 2008.

Martin Boeker, Werner Vach, and Edith Motschall. Google scholar as replacement for systematic literature searches: good relative recall and precision are not enough. In *BMC medical research methodology*, 2013.

Dean Giustini and Maged N Kamel Boulos. Google scholar is not enough to be used alone for systematic reviews. *Online journal of public health informatics*, 5(2):214, 2013.

Thorsten Schoormann, Dennis Behrens, Michael Fellmann, and Ralf Knackstedt. ¡¡ sorry, too much information¿¿ design principles for supporting rigorous search strategies in literature reviews. In *2018 IEEE 20th Conference on Business Informatics (CBI)*, pages 99–108. IEEE, 2018.

Guobing Zou, Bofeng Zhang, Yanglan Gan, and Jianwen Zhang. An ontology-based methodology for semantic expansion search. In *2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery*, volume 5, pages 453–457. IEEE, 2008.

Roberto De Virgilio, Francesco Guerra, and Yannis Velegrakis. *Semantic search over the web*. Springer Science & Business Media, 2012.

Alan Cruse. *Meaning in Language: An introduction to Semantics and Pragmatics*. Oxford: Oxford University Press, 2004.

D Alan Cruse and David Alan Cruse. *Lexical semantics*. Cambridge university press, 1986.

Dirk Geeraerts. The theoretical and descriptive development of lexical semantics. *The lexicon in focus. Competition and convergence in current lexicology*, pages 23–42, 2002.

Scott C Deerwester, Susan T Dumais, George W Furnas, Richard A Harshman, Thomas K Landauer, Karen E Lochbaum, and Lynn A Streeter. Computer information retrieval using latent semantic structure, June 13 1989. US Patent 4,839,853.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

Jian Wu, Jason Killian, Huaiyu Yang, Kyle Williams, Sagnik Ray Choudhury, Suppawong Tuarob, Cornelia Caragea, and C Lee Giles. Pdfmef: A multi-entity knowledge extraction framework for scholarly documents and semantic search. In *Proceedings of the 8th International Conference on Knowledge Capture*, page 13. ACM, 2015b.

Tim Berners-Lee, James Hendler, Ora Lassila, et al. The semantic web. *Scientific american*, 284(5):28–37, 2001.

Grigoris Antoniou and Frank Van Harmelen. Web ontology language: Owl. In *Handbook on ontologies*, pages 67–92. Springer, 2004.

Deborah L McGuinness, Frank Van Harmelen, et al. Owl web ontology language overview. *W3C recommendation*, 10(10):2004, 2004.

Q&a with tim berners-lee: The inventor of the web explains how the new semantic web could have profound effects on the growth of knowledge and innovation, Apr 2007. URL https://www.bloomberg.com/news/articles/2019-03-08/tesla-bull-left-shares-photo-of-what-he-says-is-the-new-model-s.

Matthew Russell Kearl, Cherie Bakker Noteboom, and Deb Tech. A novel improvement to google scholar ranking algorithms through broad topic search. 2017.

Bastien Latard, Jonathan Weber, Germain Forestier, and Michel Hassenforder. Towards a semantic search engine for scientific articles. In *International Conference on Theory and Practice of Digital Libraries*, pages 608–611. Springer, 2017.

Ankita Malve and PM Chawan. A comparative study of keyword and semantic based search engine. *International Journal of Innovative Research in Science, Engineering and Technology*, 4(11):11156–11161, 2015.

Alberto Martín-Martín, Enrique Orduna-Malea, Juan M Ayllón, and Emilio Delgado López-Cózar. Back to the past: on the shoulders of an academic search engine giant. *Scientometrics*, 107(3):1477–1487, 2016.

Péter Jacsó. Google scholar revisited. *Online information review*, 32(1):102–114, 2008.

Enrique Orduña-Malea, Alberto Martín-Martín, Juan M. Ayllon, and Emilio Delgado Lopez-Cozar. The silent fading of an academic search engine: the case of microsoft academic search. *Online Information Review*, 38(7):936–953, 2014.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999a.

Fathi Mahmoud Fathi Al-Hattab. An efficient ranking algorithm for scientific research papers. *Zarqa University-Jordan*, 2016.

Amy N Langville and Carl D Meyer. *Google's PageRank and beyond: The science of search engine rankings*. Princeton University Press, 2011.

Joeran Beel and Bela Gipp. Academic search engine spam and google scholar's resilience against it. *Journal of electronic publishing*, 13(3), 2010.

Hans-Michael Müller, Eimear E Kenny, and Paul W Sternberg. Textpresso: An ontology-based information retrieval and extraction system for biological literature. *PLoS Biology*, 2(11):e309, 2004.

Eric Brill. A simple rule-based part of speech tagger. In *Proceedings of the third conference on Applied natural language processing*, pages 152–155. Association for Computational Linguistics, 1992.

Philippe E Thomas, Johannes Starlinger, and Ulf Leser. Experiences from developing the domain-specific entity search engine geneview. In *BTW*, pages 225–239, 2013.

Andreas Doms and Michael Schroeder. Gopubmed: exploring pubmed with the gene ontology. *Nucleic acids research*, 33(suppl_2):W783–W786, 2005.

Heiko Dietze and Michael Schroeder. Goweb: a semantic search engine for the life science web. *BMC bioinformatics*, 10(10):S7, 2009.

Jörg Hakenberg, Loic Royer, Conrad Plake, Hendrik Strobelt, and Michael Schroeder. Me and my friends: gene mention normalization with background knowledge. In *Proc 2nd BioCreative Challenge Evaluation Workshop*, pages 1–4, 2007.

Philippe Thomas, Johannes Starlinger, Alexander Vowinkel, Sebastian Arzt, and Ulf Leser. Geneview: a comprehensive semantic search engine for pubmed. *Nucleic acids research*, 40(W1):W585–W591, 2012.

Eneko Agirre. Word sense disambiguation. *Text, Speech and Language Technology*, 2006.

Roberto Navigli. Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*, 41(2):10, 2009.

Roberto Navigli and Simone Paolo Ponzetto. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250, 2012.

Maria Teresa Pazienza, Marco Pennacchiotti, and Fabio Massimo Zanzotto. Terminology extraction: an analysis of linguistic and statistical approaches. In *Knowledge mining*, pages 255–279. Springer, 2005.

Hassan H Alrehamy and Coral Walker. Semcluster: unsupervised automatic keyphrase extraction using affinity propagation. In *UK Workshop on Computational Intelligence*, pages 222–235. Springer, 2017.

Youngja Park, Roy J Byrd, and Branimir K Boguraev. Automatic glossary extraction: beyond terminology identification. In *COLING 2002: The 19th International Conference on Computational Linguistics*, 2002.

Juan Antonio Lossio-Ventura, Clement Jonquet, Mathieu Roche, and Maguelonne Teisseire. Biomedical term extraction: overview and a new methodology. *Information Retrieval Journal*, 19(1-2):59–99, 2016a.

Irena Spasić, Bo Zhao, Christopher B Jones, and Kate Button. Kneetex: an ontology–driven system for information extraction from mri reports. *Journal of biomedical semantics*, 6(1):34, 2015.

Wilson Wong, Wei Liu, and Mohammed Bennamoun. Determining termhood for learning domain ontologies using domain prevalence and tendency. In *Proceedings of the sixth Australasian conference on Data mining and analytics-Volume 70*, pages 47–54. Australian Computer Society, Inc., 2007.

Juan Antonio Lossio-Ventura, Clement Jonquet, Mathieu Roche, and Maguelonne Teisseire. A way to automatically enrich biomedical ontologies. In *EDBT: Extending Database Technology*, volume 1. ACM, 2016b.

Mark Alan Finlayson and Nidhi Kulkarni. Detecting multi-word expressions improves word sense disambiguation. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 20–24. Association for Computational Linguistics, 2011.

Lauri Karttunen, Jean-Pierre Chanod, Gregory Grefenstette, and Anne Schille. Regular expressions for language engineering. *Natural Language Engineering*, 2(4):305–328, 1996.

Christian Jacquemin, Judith L Klavans, and Evelyne Tzoukermann. Expansion of multi-word terms for indexing and retrieval using morphology and syntax. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 24–31. Association for Computational Linguistics, 1997.

Ulrich Heid. A linguistic bootstrapping approach to the extraction of term candidates from german text. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 5(2):161–181, 1998.

Katerina T Frantzi, Sophia Ananiadou, and Junichi Tsujii. The c-value/nc-value method of automatic recognition for multi-word terms. In *International Conference on Theory and Practice of Digital Libraries*, pages 585–604. Springer, 1998.

Katerina Frantzi, Sophia Ananiadou, and Hideki Mima. Automatic recognition of multi-word terms:. the c-value/nc-value method. *International journal on digital libraries*, 3(2):115–130, 2000.

Nidhi Kulkarni and Mark Finlayson. jmwe: A java toolkit for detecting multi-word expressions. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 122–124, 2011.

Antoni Oliver and Mercè Vàzquez. Tbxtools: a free, fast and flexible tool for automatic terminology extraction. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 473–479, 2015.

Erik Faessler and Udo Hahn. Semedico: a comprehensive semantic search engine for the life sciences. *Proceedings of ACL 2017, System Demonstrations*, pages 91–96, 2017.

Ralph Delfs, Andreas Doms, Alexander Kozlenkov, and Michael Schroeder. Gopubmed: ontology-based literature search applied to gene ontology and pubmed. In *German Conference on Bioinformatics*, volume 169, page 178, 2004.

Wei Hu, Honglei Qiu, Jiacheng Huang, and Michel Dumontier. Biosearch: a semantic search engine for bio2rdf. *Database*, 2017, 2017.

Mark Steyvers and Tom Griffiths. Probabilistic topic models. *Handbook of latent semantic analysis*, 427(7):424–440, 2007.

Jan Brophy and David Bawden. Is google enough? Comparison of an internet search engine with academic library resources. In *Aslib Proceedings*, volume 57, pages 498–512. Emerald Group Publishing Limited, 2005.

Jie Tang, Ruoming Jin, and Jing Zhang. A topic modeling approach and its integration into the random walk framework for academic search. In *2008 Eighth IEEE International Conference on Data Mining*, pages 1055–1060. IEEE, 2008a.

Yifeng Liu, Yongjie Liang, and David Wishart. Polysearch2: a significantly improved text-mining system for discovering associations between human diseases, genes, drugs, metabolites, toxins and more. *Nucleic acids research*, 43(W1):W535–W542, 2015.

Olga Vechtomova and Ying Wang. A study of the effect of term proximity on query expansion. *Journal of Information Science*, 32(4):324–333, 2006.

Liana Ermakova, Josiane Mothe, and Elena Nikitina. Proximity relevance model for query expansion. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing*, pages 1054–1059. ACM, 2016.

Jöran Beel and Bela Gipp. Google scholar's ranking algorithm: an introductory overview. In *Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI'09)*, volume 1, pages 230–241. Rio de Janeiro (Brazil), 2009.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999b. Previous number = SIDL-WP-1999-0120.

Mushtaq A Hasson, Song Feng Lu, and Basheer A Hassoon. Scientific research paper ranking algorithm ptra: A tradeoff between time and citation network. In *Applied Mechanics and Materials*, volume 551, pages 603–611. Trans Tech Publ, 2014.

Ronald Brisebois, Alain Abran, Apollinaire Nadembega, and Philippe N'techobo. An assisted literature review using machine learning models to identify and build a literature corpus. *International Journal of Engineering Science Invention*, 6(7):72–84, 2017.

Sabir Ribas, Berthier Ribeiro-Neto, Rodrygo LT Santos, Edmundo de Souza e Silva, Alberto Ueda, and Nivio Ziviani. Random walks on the reputation graph. In *Proceedings of the 2015 international conference on the theory of information retrieval*, pages 181–190. ACM, 2015.

Dik L Lee, Huei Chuang, and Kent Seamons. Document ranking and the vector-space model. *IEEE software*, 14(2):67–75, 1997.

Corinna Breitinger, Bela Gipp, and Stefan Langer. Research-paper recommender systems: a literature survey. *International Journal on Digital Libraries*, 17(4):305–338, 2015.

Chenyan Xiong, Russell Power, and Jamie Callan. Explicit semantic ranking for academic search via knowledge graph embedding. In *Proceedings of the 26th international conference on world wide web*, pages 1271–1279. International World Wide Web Conferences Steering Committee, 2017.

Bhaskar Mitra, Eric Nalisnick, Nick Craswell, and Rich Caruana. A dual embedding space model for document ranking. *arXiv preprint arXiv:1602.01137*, 2016.

Jie Tang, Jing Zhang, Limin Yao, and Juanzi Li. Extraction and mining of an academic social network. In *Proceedings of the 17th international conference on World Wide Web*, pages 1193–1194. ACM, 2008b.

Jie Tang. Aminer: Toward understanding big scholar data. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pages 467–467. ACM, 2016.

R Buckminster Fuller and Kiyoshi Kuromiya. *Critical path*. Macmillan, 1981.

David Russell Schilling. Knowledge doubling every 12 months, soon to be every 12 hours, Apr 2013. URL `http://www.industrytap.com/knowledge-doubling-every-12-months-soon-to-be-every-12-hours/3950`.

P Coles, T Cox, C Mackey, and S Richardson. The toxic terabyte-how datadumping threatens business efficiency. *IBM Global Technical Services*, 2006.

Dirk Lewandowski. Evaluating the retrieval effectiveness of web search engines using a representative query sample. *Journal of the Association for Information Science and Technology*, 66(9):1763–1775, 2015.

Laura Martínez-Sanahuja and David Sánchez. Evaluating the suitability of web search engines as proxies for knowledge discovery from the web. *Procedia Computer Science*, 96:169–178, 2016.

Todd Leyba. Semantic search by means of word sense disambiguation using a lexicon, Apr 19, 2016. US Patent 9,317,589.

Jingshan Huang, Fernando Gutierrez, Harrison J. Strachan, Dejing Dou, Weili Huang, Barry Smith, Judith A. Blake, Karen Eilbeck, Darren A. Natale, Yu Lin, Bin Wu, Nisansa de Silva, Xiaowei Wang, Zixing Liu, Glen M. Borchert, Ming Tan, and Alan Ruttenberg. Omnisearch: A semantic search system based on the ontology for microRNA target (OMIT) for microRNA-target gene interaction data. *Journal of Biomedical Semantics*, 7(1):25, 2016.

Jlangbo Dang, Murat Kalender, Candemir Toklu, and Kenneth Hampel. Semantic search tool for document tagging, indexing and search, Jun 20, 2017. US Patent 9,684,683.

Mehdi Allahyari, Krys J Kochut, and Maciej Janik. Ontology-based text classification into dynamically defined topics. In *Proceedings of the 2014 IEEE International Conference on Semantic Computing (ICSC)*, pages 273–278, Newport Beach, CA, 2014.

Gridaphat Sriharee. An ontology-based approach to auto-tagging articles. *Vietnam Journal of Computer Science*, 2(2):85–94, 2015.

Daya C. Wimalasuriya and Dejing Dou. Ontology-based information extraction: An introduction and a survey of current approaches. *Journal of Information Science*, 36 (3):306–323, 2010. doi: 10.1177/0165551509360123.

Han Jaiwei and Micheline Kamber. *Data mining: concepts and techniques*. Morgan Kaufmann, San Francisco, 2006.

Rossitza Setchi and Qiao Tang. Concept indexing using ontology and supervised machine learning. *Transactions on Engineering, Computing and Technology*, 19:221–226, 2007.

Jon M Kleinberg, Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew S Tomkins. The web as a graph: Measurements, models, and methods. In *Proceedings of the International Computing and Combinatorics Conference*, pages 1–17, Tokyo, Japan, 1999.

Slavko Žitnik, Marinka Žitnik, Blaž Zupan, and Marko Bajec. Sieve-based relation extraction of gene regulatory networks from biological literature. *BMC Bioinformatics*, 16(16):S1, 2015.

Harsha Gurulingappa, Abdul Mateen-Rajpu, and Luca Toldo. Extraction of potential adverse drug events from medical case reports. *Journal of Biomedical Semantics*, 3 (1):15, 2012.

Marie-Francine Moens. *Information extraction: Algorithms and prospects in a retrieval context*. Springer Netherlands, Dordrecht, The Netherlands, 2006.

Dawn Field, Linda Amaral-Zettler, Guy Cochrane, James R Cole, Peter Dawyndt, George M Garrity, Jack Gilbert, Frank Oliver Glöckner, Lynette Hirschman, and Ilene Karsch-Mizrachi. The genomic standards consortium. *PLoS Biology*, 9(6): e1001088, 2011.

Lynn M Schriml, Cesar Arze, Suvarna Nadendla, Anu Ganapathy, Victor Felix, Anup Mahurkar, Katherine Phillippy, Aaron Gussman, Sam Angiuoli, Elodie Ghedin, Owen White, and Neil Hall. GeMInA, genomic metadata for infectious agents, a geospatial surveillance pathogen database. *Nucleic Acids Research*, 38, Suppl. 1: D754–D764, 2010.

NIH NAIDS. The national institute for allergy and infectious diseases (NIAID), microbiology and infectious diseases resources, DMID metadata standards core sample. `https://www.niaid.nih.gov/research/`

dmid-metadata-standards-core-sample, 2017. Retrieved on Jun 19, 2018.

Steven Bird and Edward Loper. NLTK: the natural language toolkit. In *Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions*, page 31, Barcelona, Spain, 2004.

Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, 1998.

PubMed Help. Stopwords table. https://www.ncbi.nlm.nih.gov/books/NBK3827/table/pubmedhelp.T.stopwords/, National Center for Biotechnology Information, 2005. accessed on Jun 19, 2018.

Liling Tan. Pywsd: Python implementations of word sense disambiguation (WSD) technologies [software]. https://github.com/alvations/pywsd, 2014. accessed on Jun 19, 2018.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

Andrew Trask, Phil Michalak, and John Liu. sense2vec – A fast and accurate method for word sense disambiguation in neural word embeddings. *arXiv Computing Research Repository (CoRR)*, 2015. abs/1511.06388.

Explosion AI. SpaCy: A library for advanced natural language processing in python and cython [software]. https://github.com/explosion/spaCy, 2015. accessed on Jun 19, 2018.

Tin Kam Ho. Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 1, pages 278–282, Montreal, Canada, 1995. doi: 10.1109/ICDAR.1995.598994.

James Franklin. The elements of statistical learning: Data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85, 2005.

Antonio Criminisi, Jamie Shotton, Ender Konukoglu, et al. Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and Trends in Computer Graphics and Vision*, 7(2–3):81–227, 2012.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20 (3):273–297, 1995.

Wenwen Li, Michael F Goodchild, and Robert Raskin. Towards geospatial semantic search: exploiting latent semantic relations in geospatial data. *International Journal of Digital Earth*, 7(1):17–37, 2014.

Hannah Bast, Björn Buchhold, and Elmar Haussmann. Semantic Search on Text and Knowledge Bases. *Foundations and Trends in Information Retrieval*, 10(2-3):119–271, 2016.

Vikas Jindal, Seema Bawa, and Shalini Batra. A review of ranking approaches for semantic search on web. *Information Processing & Management*, 50(2):416–425, 2014.

Deya Banisakher, Maria E Presa Reyes, Joshua D Eisengberg, Joshua Allen, Mark A Finlayson, Rene Price, and Shu-Ching Chen. Ontology-based supervised concept

learning for the biogeochemical literature. In *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, pages 402–410. IEEE, 2018b.

Naoaki Okazaki. Crfsuite: a fast implementation of conditional random fields (crfs). 2007. URL `https://github.com/chokkan/crfsuite`.

Bonnie Feldman, Ellen M Martin, and Tobi Skotnes. Big data in healthcare hype and hope. *Dr. Bonnie*, 360:122–125, 2012.

VITA

DEYA BANISAKHER

April 23, 1992	Born, Amman, Jordan
2014	B.S., Computer Engineering Bethune-Cookman University Daytona Beach, Florida
2014	B.S., Computer Science Bethune-Cookman University Daytona Beach, Florida
2019	M.S., Computer Science Florida International University Miami, Florida

PUBLICATIONS

Banisakher, D., Yarlott, W. V., Aldawsari, M., Rishe, N., Finlayson, M. A. (2020). *Improving the Identification of the Discourse Function of News Article Paragraphs.* In Proceedings of the First Joint Workshop on Narrative Understanding, Storylines, and Events (NUSE 2020).

Banisakher, D., Rishe, N., Finlayson, M. A. (2018). *Automatically Detecting the Position and Type of Psychiatric Evaluation Report Sections.* In Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis (LOUHI 2018), pp. 101–110. Brussels, Belgium.

Banisakher, D., Reyes, M. E. P., Eisenberg, J. D., Allen, J., Finlayson, M. A., Price, R., Chen, S. C. (2018). *Ontology-Based Supervised Concept Learning for the Biogeochemical Literature.* In Proceedings of the 2018 IEEE International Conference on Information Reuse and Integration (IRI) pp. 402–410. Salt Lake City, UT.

Banisakher, D., Reyes, M. E. P., Eisenberg, J. D., Allen, J., Finlayson, M. A., Price, R., Chen, S. C. (2018). *Ontology-Based Supervised Concept Learning for the Biogeochemical Literature.* In Proceedings of the 2018 IEEE International Conference on Information Reuse and Integration (IRI) pp. 402–410. Salt Lake City, UT.

Eisenberg, J.D., Banisakher, D. M., Presa, M., Unthank, K., Finlayson, M.A., Price, R., Chen, S. (2017) *Toward Semantic Search for the Biogeochemical Literature.* In Proceedings of the 18th IEEE International Conference on Information Reuse and Integration (IRI), pp. 517-525. San Diego, CA.

Alexenko, T., Biondo, M., Banisakher, D. M., Skubic, M. (2013). *Android-based Speech Processing for Eldercare Robotics.* In Proceedings of the Companion Publication of the 2013 International Conference on Intelligent User Interfaces (IUI), pp. 87–88. Los Angeles, CA.

Cho, H. J., Ogashawara, I., Mishra, D., White, J., Kamerosky, A., Morris, L., ... Banisakher, D. M. (2014) *Evaluating Hyperspectral Imager for the Coastal Ocean (HICO) data for seagrass mapping in Indian River Lagoon, FL.* GIScience Remote Sensing, 51(2), 120-138.

Aziz, H., Banisakher, D. M., Lee, S., Chen, L., Chinthavali, S., Duan, S., (2018) *Stochastic Graph Modeling for Large Heterogeneous Graphs – An unsupervised approach.* Manuscript submitted for publication to Journal of Transportation Health.

Banisakher, D., Rishe, N., Finlayson, M. A. (2019) *Using Conditional Random Fields to Automatically Identify Sections in Clinical Reports*. Manuscript submitted for publication to Journal of Biomedical Semantics.