Automatically Detecting the Position and Type of Psychiatric Evaluation Report Sections

Deya M. Banisakher, Naphtali Rishe, Mark A. Finlayson
School of Computing and Information Sciences
Florida International University
Miami, FL 33199
{dbani001,rishe,markaf}@fiu.edu

Abstract

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. These reports contain free-text structured within sections using a convention of headings. We present a model for automatically detecting the position and type of different psychiatric evaluation report sections. We developed this model using a corpus of 150 sample reports that we gathered from the Web, and used sentences as a processing unit while section headings were used as labels of section type. From these labels we generated a unified hierarchy of labels of section types, and then learned n-gram models of the language found in each section. To model conventions for section order, we integrated these n-gram models with a Hierarchical Hidden Markov Model (HHMM) representing the probabilities of observed section orders found in the corpus, and then used this HHMM n-gram model in a decoding framework to infer the most likely section boundaries and section types for documents with their section labels removed. We evaluated our model over two tasks, namely, identifying section boundaries and identifying section types and orders. Our model significantly outperformed baselines for each task with an F_1 of 0.88 for identifying section types, and a 0.26 WindowDiff (W_d) and 0.20 and (P_k) scores, respectively, for identifying section boundaries.

1 Introduction

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.

We present an approach to automatically model the discourse structure of psychiatric reports as well as segment these reports into various sections. Our model learns the section types, positions, and sequence and can automatically segment unlabeled text in a psychiatric report into the corresponding sections. We 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, we train a Hierarchical Hidden Markov Model (HHMM) that categorizes sections in psychiatric reports into one of 25 pre-defined section labels.

The remainder of this paper is organized as follows: we first introduce psychiatric reports and their various types and conventions (§2). Next, we discuss the task definition in detail (§3). We then describe our approach including the corpus used, and the two main components of our model (§4). Additionally, we present and discuss the baselines and experiments performed as well as the results obtained from those experiments (§5). We follow this with a review of related work on document

Proceedings of the 9th International Workshop on Health Text Mining and Information Analysis (LOUHI 2018), pages 101–110
Brussels, Belgium, October 31, 2018.

section identification and text segmentation (§6). Finally, we conclude and specify our contributions (§7).

2 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 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). Our study focuses on psychiatric evaluation reports. 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 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.

***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 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*).

3 Task Definition

Our 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 we 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. We assume that the reports follow the general guidelines of psychiatric evaluation report writing discussed in (§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*. Figure 1 shows an example. 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, our 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 doctorpatient 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., Treatment Plan), and with various general conventions for what information 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*.

We 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.

4 Approach

Given the sequential nature of the reports' sections, we 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 our problem formulation and approach, we 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 a clinical note documents. We take a supervised learning approach where we learn the HMM parameters using a labeled corpus. Our implementation was generally guided by the work described in Barzilay and Lee (2004) and (Rabiner, 1989).

To overcome the challenges outlined in (§3), we first created a unified hierarchy of standardize section labels types, based on observations in a 150 report corpus that we assembled. Second, while Li et al. (2010) focused on the section level when building their n-gram language models, we focus on the sentence level, similar to Sherman and Liu (2008). Additionally, to model the inclusion of some sections within others as discussed in (§3) we 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) we built a two-level HHMM that models the positions and order of the reports' sections. To tackle the second subtask, we built language models (namely, n-gram models) per section type that describe distinctive lexical information for each of those sections. We 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, we follow a decoding scheme using the Viterbi algorithm (discussed briefly in §4.4).

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

4.1 Corpus

To the best of our knowledge there is no corpus of psychiatric reports annotated with section labels, so we created our own. We collected 150 publicly available psychiatric evaluation report samples by crawling the web through custom search engines (Google Custom Search Engine for Med-

Parent Label	Section Label	# Words	# Sentences	Avg. Sent. Length	% Present	% Implicit
-	*IDENTIFYING DATA*	12	2	6	100	-
	CHIEF COMPLAINT	27	3	9	100	-
MEDICAL HISTORY	*HISTORY OF PRESENT ILLNESS*	232	29	8	95	10
	PSYCHIATRIC HISTORY	85	8	11	82	36
	SUBSTANCE ABUSE HISTORY	98	10	10	88	44
	REVIEW OF SYMPTOMS	150	19	8	96	51
-	*SURGERIES*	28	3	7	33	-
	ALLERGIES	4	2	2	98	-
	CURRENT MEDICATIONS	40	9	4	100	-
FAMILY HISTORY	*BIRTH AND DEVELOPMENTAL HISTORY*	59	5	10	31	51
	ABUSE HISTORY / TRAUMA	110	9	12	79	34
	FAMILY PSYCHIATRIC HISTORY	44	5	9	73	80
	FAMILY MEDICAL HISTORY	48	7	7	92	38
	SOCIAL HISTORY	80	7	11	76	45
	PREGNANCY	29	3	8	47	64
-	*SPIRITUAL BELIEFS*	12	2	5	24	-
	EDUCATION	32	3	8	68	-
	EMPLOYMENT	31	3	9	79	-
	LEGAL	10	1	5	20	-
MENTAL STATUS	*MENTAL STATUS*	155	18	9	95	11
	STRENGTHS AND SUPPORTS	8	1	8	71	43
-	*FORMULATION*	35	4	8	62	-
	DIAGNOSES	63	12	5	100	-
	PROGNOSIS	8	2	3	74	-
	TREATMENT PLAN	121	12	10	100	-

Table 1: List of possible sections in a psychiatric report used in the corpus.

ical Transcriptionists[1] and GoogleMT[2]) and other sources [3]. The reports we selected were complete and adhere to the general guidelines for psychiatric report writing discussed in the previous sections. Some of the reports were anonymized samples of real reports, while others were mock reports written for educational purposes.

We prepared the corpus in two stages. First, we 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 we selected *EDUCATION* as the uniform section label across all reports.

Second, we created a hierarchy for the section names which reflected implicit embedded sections types that we found in the corpus. There were only three section types that included implicit subsections in our data, 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 we 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 1 lists the the parent sections that sometimes included other sections implicitly (first column), the unified list of section types found in the collected reports (second column), word and sentence level statistics (columns 3-5), and percentage of reports containing those sections in the corpus (last two columns). For both of these stages we used all 150 reports.

Following standard procedure for supervised machine learning, we split our 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.

[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/

4.2 Modeling the Section Orders

As discussed before, we built an HHMM where each state corresponds to a distinct section label. We 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, we created a two-level HHMM where *parent states* contained *states* representing the ordered subsections found in the *parent state* section. Thus our model contained 25 *states* and three *parent states* corresponding to information in Table 1. 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 our HHMM, each *parent state* is simply an HMM itself. Thus our discussion of HMM parameter calculation applies to both *states* and *parent states*.

Our model learned transition probabilities from the labeled corpus. The state transition probabilities capture constraints on section orderings. We estimated the probabilities between each state s using Equation 1. Additionally, to account for sparsity (that is, unseen section orders) we smoothed the probabilities by the total number of section labels t_S following Laplace smoothing.

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

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 were 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).

4.3 Modeling Section Language

To tackle the second subtask identified in (§3), we 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. We opted to use bigrams as our training corpus because higher n-gram models were extremely sparse, and had poor performance. This is consistent with significant research showing that in most applications bigrams work well and better than others (Reynar, 1998).

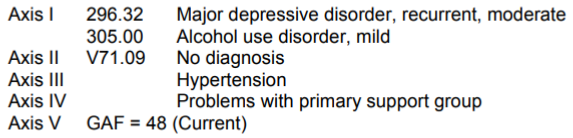
We built independent bigram models for each section type in the reports, using only text from that section type. Additionally, for each of the three section types represented by the *parent states* (discussed above) we built bigram models using text found in all of the contained subsections. A common 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 our relatively small corpus, our models were quite sparse at first, however, we used Laplace Smoothing as a solution.

Similar to transition probabilities, our HHMM learned observation probabilities from the labeled corpus. We trained a bigram model for each state s of the HHMM. Equation 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 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).

$$P(w_0^k|s) = \prod_{0}^{k-1} P_s(w_{i+1}|w_i) \tag{2}$$

$$P_s(w_{i+1}|w_i) = \frac{count_S(w_i^{i+1}) + 1}{count_S(w_i) + |V_S|} \tag{3}$$

We 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 2. We recognize that this standard has been dropped with the introduction of *DSM-5* in 2013, however, our dataset 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 2: Example of *DSM-IV* multiaxal diagnosis assessment.

For the *MEDICATIONS* section we 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 our algorithm would search for the key headers "Axis I, II, III, IV, V".

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

4.4 Decoding

We 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. We 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:

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

5 Results and Discussion

As discussed above, we randomly split the corpus into training and testing sets in a cross-validation setup, using ten folds, resulting in 120 reports for training and 30 for testing in each fold. Our models were trained to learn a total of 25 distinct sections. Here we present our 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.

5.1 Evaluation Methods

There are two problems that our 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. We evaluate our system's performance on these two problems separately.

For the section ordering, we 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) (Pevzner and Hearst, 2002) and P_k (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.

5.2 Baseline Methods

We compared our 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, we followed the primary approach proposed by Li et al. (2010) which is a flat HMM model built similarly to our model as described previously (§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. Our 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.

We compared our 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). LC-

Section	**Independent Bigram**			**Flat HMM**			**HHMM**		
	P	***R***	***F₁***	***P***	***R***	***F₁***	***P***	***R***	***F₁***
IDENTIFYING DATA	0.83	0.81	0.82	0.96	0.94	0.95	0.98	0.95	**0.97**
CHIEF COMPLAINT	0.68	0.65	0.67	0.88	0.74	0.80	0.94	0.89	**0.91**
MEDICAL HISTORY	0.66	0.66	0.65	0.93	0.88	**0.90**	0.93	0.88	**0.90**
HISTORY OF PRESENT ILLNESS	0.69	0.67	0.68	0.91	0.86	0.88	0.94	0.86	**0.90**
PSYCHIATRIC HISTORY	0.65	0.60	0.62	0.74	0.85	0.79	0.93	0.86	**0.89**
SUBSTANCE ABUSE HISTORY	0.69	0.69	0.69	0.88	0.80	0.84	0.95	0.83	**0.89**
REVIEW OF SYMPTOMS	0.8	0.67	0.73	0.79	0.86	0.82	0.94	0.87	**0.90**
SURGERIES	0.4	0.31	0.35	0.79	0.51	0.62	0.85	0.64	**0.73**
ALLERGIES	0.6	0.80	0.69	0.90	0.86	0.88	0.88	0.91	**0.89**
CURRENT MEDICATIONS	0.87	0.74	0.80	0.90	0.84	0.87	0.91	0.93	**0.92**
FAMILY HISTORY	0.60	0.56	0.58	0.92	0.86	**0.89**	0.92	0.86	**0.89**
BIRTH AND DEVELOPMENTAL HISTORY	0.68	0.50	0.57	0.71	0.68	0.69	0.89	0.80	**0.84**
ABUSE HISTORY / TRAUMA	0.42	0.33	0.37	0.87	0.77	0.82	0.96	0.81	**0.88**
FAMILY PSYCHIATRIC HISTORY	0.57	0.59	0.58	0.92	0.87	0.89	0.92	0.90	**0.91**
FAMILY MEDICAL HISTORY	0.65	0.60	0.62	0.92	0.89	0.90	0.94	0.89	**0.91**
SOCIAL HISTORY	0.67	0.69	0.68	0.66	0.89	0.76	0.93	0.81	**0.87**
PREGNANCY	0.6	0.67	0.63	0.89	0.51	0.65	0.92	0.80	**0.86**
SPIRITUAL BELIEFS	0.73	0.46	0.56	0.90	0.9	**0.90**	0.93	0.88	**0.90**
EDUCATION	0.66	0.61	0.63	0.71	0.77	0.74	0.92	0.84	**0.88**
EMPLOYMENT	0.65	0.62	0.63	0.91	0.88	**0.89**	0.92	0.86	**0.89**
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**
MENTAL STATUS EXAM	0.64	0.63	0.64	0.83	0.96	0.89	0.85	0.96	**0.90**
STRENGTHS AND SUPPORTS	0.42	0.82	0.56	0.80	0.92	0.86	0.82	0.92	**0.87**
FORMULATION	0.56	0.71	0.63	0.86	0.78	0.82	0.92	0.82	**0.87**
DIAGNOSES	0.88	0.76	0.81	0.96	0.95	0.96	0.98	0.98	**0.98**
PROGNOSIS	0.66	0.62	0.64	0.84	0.82	0.83	0.90	0.86	**0.88**
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 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.

Seg 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).

5.3 Results

For the section labeling problem, our model equaled or outperformed both baselines in all the sections. Table 2 shows the precision, recall, and F_1 scores for the two baselines and our 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. We 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. Our model performed better for all implicit subsections, and significantly better for two (i.e., *PREGNANCY* and *BIRTH AND DEVELOPMENTAL HISTORY*. Finally, our model performed exactly the same as the Flat HMM baseline for the three parent types, as our 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, we computed both macro- and micro-averaged precision, recall, and F-measure (last two rows in Table 2). Our 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 our HHMM and the Flat-HMM baseline seemed to neither 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), we 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 we allowed the parameter to be chosen by LCSeg and TopicTiling, respectively, while in the second experiment, we provide the algorithms with the correct number of boundaries (i.e., number of sections minus one). Our model however, needs no prior information regarding the number of sections present in a given report. Table 3 shows the W_d and P_k scores for all three approaches. Our 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 our approach by a small margin when provided with the correct parameter value. We note, however, that when running open loop on new text, the number of sections will be unknown, so this result does not reflect how we envision the approach being used.

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

Table 3: Section boundary identification results.

6 Related Work

As discussed above, our 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., 2009), 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 identifying scientific article sections (e.g., abstract, introduction, methodology, etc.) (Teufel, 1999).

Our 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.

7 Contributions

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

Acknowledgments

We thank our colleagues at the Cognition, Narrative, and Culture Laboratory (Cognac Lab), especially Mohammed Aldawsari and Victor Yarlot for their contributions to the overall idea, approach, and evaluation.

References

Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. 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.

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

American Psychiatric Association. 2018. What is psychiatry? Available from: `https://www.psychiatry.org/patients-families/what-is-psychiatry`. (Accessed on Jul 1, 2018).

Regina Barzilay and Lillian Lee. 2004. 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.

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

Hung H. Bui, Dinh Q. Phung, and Svetha Venkatesh. 2004. 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.

Wendy W Chapman, Prakash M Nadkarni, Lynette Hirschman, Leonard W D'Avolio, Guergana K Savova, and Ozlem Uzuner. 2011. 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.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. 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.

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

Munmun De Choudhury, Scott Counts, Eric J. Horvitz, and Aaron Hoff. 2014. 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.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. 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.

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

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

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

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

Michel Galley, Kathleen McKeown, Eric Fosler-Lussier, and Hongyan Jing. 2003. 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.

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

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

Marti A. Hearst. 1994. 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.

Christopher Homan, Ravdeep Johar, Tong Liu, Megan Lytle, Vincent Silenzio, and Cecilia Ovesdotter Alm. 2014. 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.

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

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

Ying Li, Sharon Lipsky Gorman, and Noémie Elhadad. 2010. 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, VA.

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

Paul van Mulbregt, Ira Carp, Lawrence Gillick, Steve Lowe, and Jon Yamron. 1998. Text segmentation and topic tracking on broadcast news via a hidden markov model approach. In *Fifth International Conference on Spoken Language Processing*.

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

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

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

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

R Reeves and R Rosner. 2016. *Forensic Psychiatry and Forensic Psychology: Forensic Psychiatric Assessment*. Elsevier.

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

Martin Riedl and Chris Biemann. 2012. Topictiling: A text segmentation algorithm based on lda. In *Proceedings of ACL 2012 Student Research Workshop*, ACL '12, pages 37–42, Jeju Island, Korea.

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

M. Sherman and Yang Liu. 2008. Using hidden markov models for topic segmentation of meeting transcripts. In *Proceedings of the 2008 IEEE Spoken Language Technology Workshop*, pages 185–188.

Carlo Strapparava and Rada Mihalcea. 2008. Learning to identify emotions in text. In *Proceedings of the ACM Symposium on Applied Computing (SAC)*, SAC '08, pages 1556–1560, Fortaleza, Ceara, Brazil.

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

Simone Teufel, Jean Carletta, and Marc Moens. 1999. 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.

J. P. Yamron, I. Carp, L. Gillick, S. Lowe, and P. van Mulbregt. 1998. A hidden markov model approach to text segmentation and event tracking. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 333–336 vol.1.

Jia Yu, Xiong Xiao, Lei Xie, Chng Eng Siong, and Haizhou Li. 2016. A dnn-hmm approach to story segmentation. In *INTERSPEECH*, San Francisco, USA.