Scaling Out and Evaluation of OBSecAn, an Automated Section Annotator for Semi-Structured Clinical Documents, on a Large VA Clinical Corpus

Le-Thuy T. Tran, PhD, Guy Divita, MS, Andrew Redd, PhD, Marjorie E. Carter, MSPH, Matthew Samore, MD, Adi V. Gundlapalli, MD, PhD, MS
IDEAS Center, VA Salt Lake City Health Care System, Salt Lake City, Utah, USA; Departments of Internal Medicine and Biomedical Informatics, University of Utah School of Medicine, Salt Lake City, Utah, USA

Abstract

"Identifying and labeling" (annotating) sections improves the effectiveness of extracting information stored in the free text of clinical documents. OBSecAn, an automated ontology-based section annotator, was developed to identify and label sections of semi-structured clinical documents from the Department of Veterans Affairs (VA). In the first step, the algorithm reads and parses the document to obtain and store information regarding sections into a structure that supports the hierarchy of sections. The second stage detects and makes correction to errors in the parsed structure. The third stage produces the section annotation output using the final parsed tree. In this study, we present the OBSecAn method and its scale to a million document corpus and evaluate its performance in identifying family history sections. We identify high yield sections for this use case from note titles such as primary care and demonstrate a median rate of 99% in correctly identifying a family history section.

Introduction

Information stored in electronic health records (EHRs) such as patient expressed symptoms, physical findings, medications, past medical history, and family history of illnesses has tremendous potential for aiding in the detection of patient care and treatment patterns, risks and outcomes of diseases, or adverse events[1, 2, 3, 4]. Effective clinical text mining methods and techniques are important for the discovery of these valuable health data for clinical, operational and research purposes. Statistical text mining and natural language processing (NLP) are among the popular techniques used in the field [5, 6]. However, the unstructured, heterogeneous, and context-dependent nature of the clinical documents has made these techniques less effective [6, 7]. Recent efforts in the development of a section identifier and labeler for preprocessing clinical documents before applying NLP enhances their effectiveness [6, 8, 9, 10].

The section identifier and labeler OBSecAn, which stands for Object-Based Section Annotator, was developed to preprocess the vast amount of clinical notes stored by the U.S. Department of Veterans Affairs (VA) in the Veterans Health Information Systems and Technology Architecture (VistA) network for various tasks including extracting symptoms and concepts related to specific domains of interest such as homelessness and indwelling urinary catheters. VistA clinical notes are semi-structured documents generated by entering text into pre-defined or user-defined templates[7]. The lack of consistency in the structure of the notes makes it difficult to solely use regular expressions to identify sections since these documents do not adhere to particular grammatical rules. OBSecAn identifies sections and their hierarchical relationships using an ontology that describes the relevant concepts of clinical notes sections and the relationships among these concepts. OBSecAn identifies labeled and unlabeled sections (sections with and without section headers) in semi-structured clinical documents and annotates these sections to be used for further information extraction tasks.

In a separate study, OBSecAn was trained and evaluated on a small corpus of 1000 VistA notes including both inpatient and outpatient encounters from any medical specialty. The results showed reasonable metrics for recovering both labeled and unlabeled sections from semi-structured electronic medical notes. In this study, we use all 1000 notes to further train OBSecAn and evaluate the performance of this method on annotating sections in one million medical documents derived from routine care offered to Veterans at VA medical facilities.

Background

This importance of and motivation for standardization of section headers in clinical notes has been addressed by many authors[6, 8, 9]. This standardization is a key step toward the creation of an ontology for clinical document sections. The existence of such an ontology may be useful for integrating and sharing clinical information among health care professionals and information systems. So far, a terminology of discovered section headers was developed for use with SecTag, a method for automated section header identification[8]. The SecTag method was the

first terminology-based method for identifying and labeling sections. In addition, other methods have identified and/or labeled section headers using regular expressions, rules, and machine learnings[6, 8, 12].

Until now, the SecTag method and its study were the best developed and evaluated of the methods for automated section header identification. However, the development of SecTag was based on notes within the Vanderbilt EMR system and demonstrated some limitations when applied to identify and label sections of VA notes. Seeing the potential of terminology-based approach in identifying and labeling sections in the semi-structured notes within VA, OBSecAn was developed using a similar approach to that of SecTag with modifications to the section header terminology, the annotated structure, and the algorithm to provide an automated section annotator for VA notes. In this paper, we discuss the development of the OBSecAn method and its scale to VA big data.

Methods

The OBSecAn algorithm for section annotation identifies the sections and labels the unlabeled sections within semi-structured clinical documents. The components of the OBSecAn method include:

Clinical Document Sections Ontology (CDSO)

A section header terminology was developed to provide a list of concepts and synonyms for clinical section headings and subsections[10]. The resulting section header terminology contained 1109 concepts with 4332 synonyms and the hierarchical relationships. As mentioned in Denny *et al.*[10], the existence of section header terminology helps to better understand clinical notes. However, we think it is important to not only standardize section headers but also standardize sections. The creation of a clinical document sections ontology (CDSO) is a step toward the standardization of sections for heath data integration. The existence of a CDSO can also aid NLP and data mining tasks. For the OBSecAn method, the CDSO helps to recover unlabeled sections. For example, "regular rate and rhythm" is often used under the "Physical Exam" section of a clinical H&P note to describe a result obtained from examining the heart. In other notes the words "regular rate and rhythm" are placed under the section "heart exam" and many times the section headers for the"heart exam" section are omitted. If the phrase "regular rate and rhythm" is added into the CDSO and related to the "heart exam" section concept, it increases the chance of recognizing the "heart exam" section if it is not labelled. For the OBSecAn method, we created a very preliminary CDSO which was evolved from the section header terminology to aid the OBSecAn algorithm. Currently, we have three different types of elements within the CDSO: Section, Section Header, and Property.

A section as defined in Denny *et al.*[10] is a clinically meaningful grouping of symptoms, history, findings, results, or clinical reasoning that is not itself part of the unique narrative for a patient. In the CDSO, a Section element represents a unit often being used by health care providers to group several elements that have something in common. The way to group the elements in clinical notes varies among health care providers. The CDSO aims to provide an exhaustive list of these groupings. Physicians commonly group into one section several separable sections that can be found in other notes. For example, the section labeled "personal and social history" is a composition of the sections "personal history" and "social history". We excluded from the CDSO those sections which can be decomposed into more than one section. The structure of a Section element in the CDSO consists of a concept identification, a description, and a data type.

A Section Header element in the CDSO is a found header for a Section element. There are variations of headers being used to label the same Section element across clinical notes. There are also cases where the same header string can be used to label different Section elements. For example, the header string "Cardiac" is used to label the section "Cardiovascular Exam" as well as the section "Cardiovascular Review." For the header string "Cardiac", we need to know the context of the document in order to determine which clinical section is being labeled. If this header string is found under the section "Physical Exam" then it is referring to "Cardiovascular Exam"; it refers to "Cardiovascular Review" if placed under the section "Review of Systems." To assist the OBSecAn algorithm in efficiently labeling the sections, we document in the CDSO aheaders with direct reference to the section like "Cardiovascular Exam" as *explicit headers*. The CDSO inherited a subset of header strings found in the section header terminology of Denny *et al*[10]. In order to achieve good coverage of the section headers found in various clinical notes, we extracted and reviewed headers that appeared frequently in more than 50,000 VA boiler-plate templates generated from all stations within the VA for additional document section headers.

A Property element in the CDSO is a term that often appears in the content of a section. One example is the phrase "regular rate and rhythm". As mentioned earlier, Property elements in the CDSO play an important role in recovering the unlabeled sections. To generate the list of possible Property elements for a section, we use the content of the labeled sections to obtain a list of frequently found terms used in that section.

There are different types of relationships among the elements of CDSO. The parent-child relationships are added to the CDSO to represent the hierarchical structure of section concepts. The relationships among the concepts are polyheirarchical parent-child relationships. Children of a document section concept are all the subsections found for that section across the clinical notes. A child section may be a part of several different parent sections. CDSO also contains the is-Header relationships between Section Header elements and Section elements and the is-Property relationships between Section elements and Property elements. The relationships among the elemements of CDSO were obtained from training on a sample of 1000 documents extracted from VistA notes.

Section Annotation Output. With the aim of using OBSecAn to preprocess clinical notes to be used for further focused information extraction, we saved the section annotation output outside of the original document. An example of the format for the section annotation output, viewable with Visual Tagging Tool (VTT)[11] is shown in Figure 1

```
0    |0  |SECTION|1||[SUBJECTIVE]:9965:0-3369
0    |7  |SECTION|2||HISTORY:4041:0-69
72   |11 |SECTION|2||Main issues:1132:72-1810
1812 |3  |SECTION|2||PMH:1260:1812-2237
2240 |12 |SECTION|2||past surgery:8535:2240-2332
2334 |11 |SECTION|2||Medications:5799:2334-2591
2594 |3  |SECTION|2||ALL:358:2594-2602
2604 |3  |SECTION|2||FH :2998:2604-2702
2609 |6  |SECTION|3||MOTHER:6133:2609-2630
2632 |6  |SECTION|3||FATHER:3033:2632-2648
2650 |4  |SECTION|3||AUNT:652:2650-2677
2679 |6  |SECTION|3||sister:10332:2679-2702
2704 |6  |SECTION|2||SOCIAL:10480:2704-3003
3005 |4  |SECTION|2||ROS :9828:3005-3369
3018 |7  |SECTION|3||GENERAL:3358:3018-3175
3183 |9  |SECTION|3||PULMONARY:9049:3183-3232
3241 |7  |SECTION|3||CARDIAC:1011:3241-3271
3279 |5  |SECTION|3||GI/GU:3332,3558:3279-3369
3373 |0  |SECTION|1||[OBJECTIVE]:6768:3373-4737
3373 |13 |SECTION|2||PHYSICAL EXAM:7678:3373-4737
3387 |0  |SECTION|3||[VITAL SIGNS]:12120:3387-3619
3387 |14 |SECTION|4||BLOOD PRESSURE:848:3387-3427
3430 |5  |SECTION|4||PULSE:3735 :3430-3457
3460 |6  |SECTION|4||WEIGHT:12205:3460-3502
3505 |6  |SECTION|4||HEIGHT:3757:3505-3547
3550 |15 |SECTION|4||BODY MASS INDEX:859:3550-3592
3594 |4  |SECTION|4||PAIN:7279:3594-3619
3621 |5  |SECTION|3||HEENT:3741:3621-3808
3635 |4  |SECTION|4||EYES:12450:3635-3685
3695 |4  |SECTION|4||EARS:2441:3695-3753
3763 |4  |SECTION|4||NOSE:6626:3763-3778
3788 |5  |SECTION|4||MOUTH:6161:3788-3800
3810 |4  |SECTION|3||NECK:6380:3810-3865
3867 |5  |SECTION|3||NODES:6600:3867-3967
3970 |5  |SECTION|3||LUNGS:5566:3970-4008
4010 |5  |SECTION|3||HEART:3730:4010-4148
4150 |7  |SECTION|3||ABDOMEN:22:4150-4235
4237 |6  |SECTION|3||RECTAL:9403:4237-4287
4290 |11 |SECTION|3||EXTREMITIES:2885:4290-4506
4508 |5  |SECTION|3||NEURO:6428:4508-4737
4739 |11 |SECTION|1||ASSESSMENT:4:4739-4831
4834 |0  |SECTION|1||[PLAN]:7255:4834-6063
```

Figure 1. An example of section annotation output to be viewed with Visual Tagging Tool (VTT)

For annotating sections in large scale, we simplify the output to contain only the information that is useful for future retrievals of the section content. Once the OBSecAn has annotated sections in a clinical note, the identification of the original document and the annotated section information are saved for future retrieval. The annotated section information includes the section identifications, the section beginning offsets, and the section ending offsets. The section beginning offsets are set to the beginning section *content* offsets, not the beginning section *label* offsets.

OBSecAn Algorithm. The OBSecAn algorithm has three major stages. The first stage reads and parses the document to obtain and store information regarding sections into a structure that supports the hierarchy of sections. The second stage detects and makes correction to errors in the parsed structure. The third stage produces the section annotation output using the final parsed tree. .

The first stage of the algorithm involves processing the documents by sequential chunks of text and storing the processed information into a structure that supports the hierarchy of sections. A clinical note may contain a list of sections and a section may contain a list of nested sections. OBSecAn identifies sections from the note and saves the structure of those identified sections into a supported data structure. This data structure contains a list of nodes representing outermost sections called root nodes. Each node in the list again links to a list of nodes representing the nested sections. OBSecAn sequentially processes a note and creates nodes and inserts them into the data structure. An illustration of data structure used by OBSecAn to store the information of identified section is shown in Figure 2.

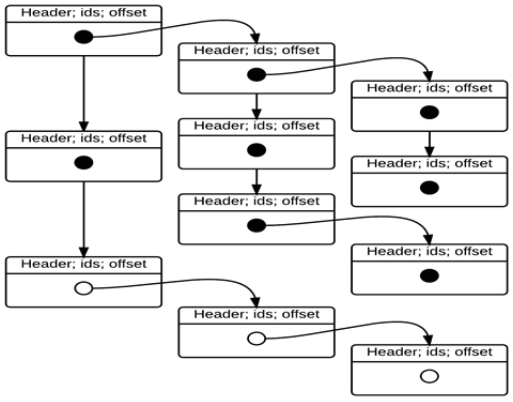

Figure 2. Illustration of data structure used by OBSecAn to store the information of identified sections

A node in Figure 2 represents a section in the document and contains the following section information: section header (Header), section identifications (ids), and offset into the note (offset). The information of the most recently

created root node and its linked nodes provide the context for disambiguating the next created section. Let ids_1 be the concept identification of the section in the most recent created root node. The concept identifications of the most recent created nested sections subsequently are ids_2, ids_3, …, ids_n. The steps for processing a clinical note are as follows:

Select Text for Annotation. The text from a clinical document is annotated by chunks. A chunk of text is formed by reading a line from the document. The next line of the text in the document is read and concatenated to the chunk only if the beginning of the next line is a continuation of the line that has already been read. The resulting chunk of text is then split up into segments using "dot/period (.)" as delimiter. We added an extra step to verify whether the dot/period is not a sentence delimiter but rather a decimal separator or a part of an abbreviation. Each text segment is then selected for annotation. The flow chart for annotating sections in a text segment is shown in Figure 3.

Map Header Candidate String to Section Concepts. A header candidate string of a text segment is either the whole text segment or the string that begins the text segment and ends right before a colon, dash, comma, semicolon, backslash, or dot. There may be several header candidate strings for a text segment. We examine the header candidate strings in sequence until we find maps of the current header candidate string to the section headers in the CDSO. As previously mentioned, a header candidate string may also be mapped to the header string of several Section concepts.

Map the Leading Words to Section Concepts. A section header may appear in a clinical document as leading words of a text segment. For example, the section header "Concentration" of the following text segment "Concentration is very poor", which appears under the section "Mental Status Exam", is not delimited from the rest of the text. In this case, we start from the first word of the text segment and find the maximum-length string which can be mapped to Section Headers in CDSO. This maximum-length string may also be mapped to several Section concepts. We also consider trimming the leading words that are prepositions or articles; for example, trimming the part "On the" in the following text segment "On the examination, temperature of the patient is 97.5."

Choose Best-Mapped Section Concept. For each of the mapped Section concepts, we calculate the distance to the recently created sections ids_1, ids_2, ids_3, …, ids_n. based on the number of unlabeled sections to be inserted into the current section hierarchy. The distance is the minimum number of sections created if the mapped Section concept is to be inserted into the current section hierarchy. The best-mapped Section Concept is the concept among the mapped Section concepts with the minimum distance to the recently created sections. The list of minimal created sections (referred to as unlabeled sections) is also inserted into the current section hierarchy along with the mapped section. In order to reduce the noise in introducing the labels to the unlabeled sections, we set thresholds for the number of created sections. The threshold is set to 3 if the mapped section is derived from mapping to Section Header elements. Otherwise, it is set to 2 if the mapped section is derived from mapping to Property elements. If the number of the unlabeled sections introduced when inserting a mapped section into the current section hierarchy exceeds the threshold for the given case, we omit the insertion of the mapped section and move forward.

Find first occurrence of Header Strings and Property Strings of a Section. Searching for Header Strings and Property Strings embedded in the text segment is used to discover the unlabeled sections. The list of Header Strings and Property String of Section ids_n includes all Header Strings and Property Strings with the relationships is-Header or is-Property to Section ids_n and its child Sections.

Retrieve Mapped Section Concept. The Section concepts associated with the Header String or Property String which first occurs in the text segment that are retrieved. In the case that there are several concepts associated with a Header String or Property String, we consider the best-mapped Section concept.

Insert Mapped Section. After a mapped section has been identified from the text segment, we create a node for the mapped section along with the nodes for unlabeled sections. These nodes are linked to the nodes in the data structure for recently created sections ids_1, ids_2, ids_3, …, ids_n or linked to a newly created root node.

Select remain text for Annotation. The part of the text segment after the text used to identify the mapped Section concept is then being annotated.

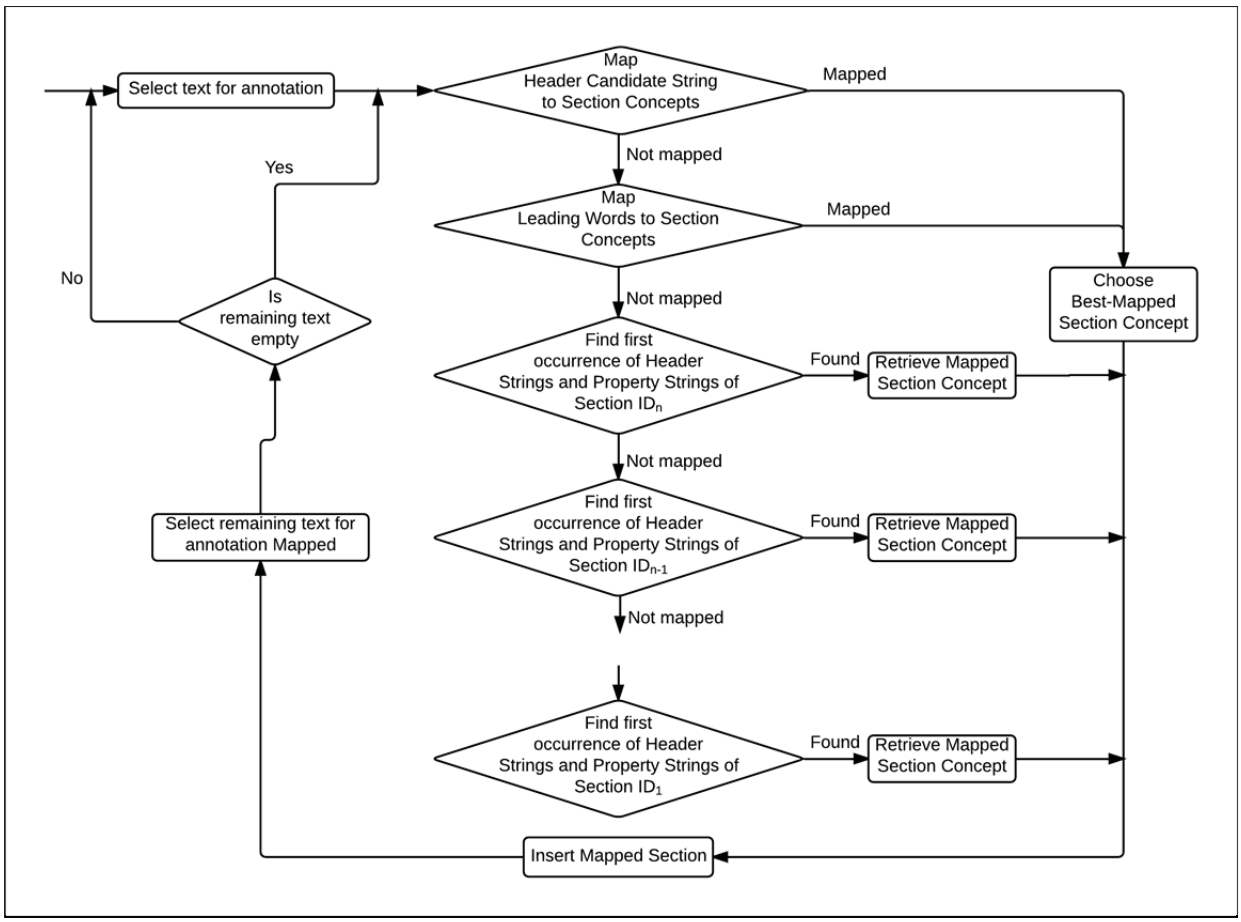

Figure 3. Flow chart for annotating sections in semi-structured electronic medical notes.

Determining beginning and ending offsets of sections. The beginning offset of a section is derived along with the position of the Section Header or Property related to the discovery of the section. For determining the ending offset of a section, we have used two different approaches. The first approach sets the ending offset of a section to the beginning of the next discovered section. The second approach derives the ending offset using machine learning. For the OBSecAn method, we used the first approach in which the ending offset of a section is derived from the beginning offset of the next discovered "same level" section. Though Denny *et al*. have shown that using machine learning could improve the accuracy in setting the section's ending offset, it has not been tested on large corpora.

After the entire document has been processed and the data structure used to store the identified section has been built, the second stage of the algorithm proceeds to detect and make correction to the errors found in the parsed data structure. We recursively traversed the parsed data structure from the top nodes and employed the following validation checks:

- Check if there were cases that sections were discovered by mistake in the middle of other sections. As a result, in the parsed structure, the nodes corresponding to the section discovered by mistake are being deleted.
- Validate the content of the tagged section using its data type.

Once the recognizable errors in the parsed data structure had been corrected, we calculated the ending offset of the sections via the information available in the final parsed data structure and generate the OBSecAn's output.

Evaluation

The evaluation of the performance of the OBSecAn method on a small corpus was done in a separate study. In this paper, the scale out of OBSecAn to a VA corpus of one million record was evaluated in three steps: (1) the estimation of disk space required to save the annotated sections vs. the original note corpus; (2) identification of high yield sections from the various note titles wherein a high yield section was determined as the ratio of the

number of an annotated section to the total number of notes; and (3) the rate of accurately identifying a particular section of notes from OBSecAn output. For this study, we have used the family history section as an illustrative example. We processed the one million records and extracted the content of the annotated family history sections. The content of these annotated sections was then written to a spreadsheet. Each row in the spreadsheet is the content of one annotated family history section. We then sorted the spreadsheet so that the rows with similar contents were clustered. These clustered rows made it easier to review in groups.

Document Corpus

The VA's VistA network stores a large number of clinical notes (over 2 billion). The notes are generated from various types of medical encounters across different facilities and medical specialties; therefore, the list of sections and section structures are also varied among different type of notes. To evaluate the performance of OBSecAn on a representative sample of VA notes, we extracted a random corpus of one million notes that are stored and made available for research in a secure VA database: Veterans Informatics and Computing Infrastructure (VINCI)[14]. These notes organized into approximately 2819 VA Enterprise Note Titles that are further clustered into ~220 higher level Note Titles. The million corpus represents a proportional random sample of documents based on the frequencies of the Enterprise Note titles in VINCI.

Results

About 75% of the notes in the million document corpus are accounted for by the top 57 Note Titles presented in the corpus (**Table 1**).

Table 1. Frequency distribution of top 57 Note Titles presented in the one-million note corpus.

Note Titles	%	Note Titles	%	Note Titles	%
Nursing	23.53	Physical Medicine Rehab	0.65	Anesthesia	0.17
Primary Care	15.36	Nutrition	0.58	Phys	0.17
Mental	5.00	Optometry	0.56	Nephrology	0.15
Pharmac	2.45	Psychology	0.56	Treatment	0.13
Internal	1.88	Urology	0.45	Specialty Care	0.13
Administrative Note	1.84	Orthopedic	0.45	Neurology	0.11
Inpatient	1.75	Recreation Therapy	0.45	Procedure	0.11
Psychiatry	1.37	Dermatology	0.43	Conse	0.11
Social Work	1.31	Occupational	0.41	C&P	0.10
Surgery	1.19	Telephone Encounter	0.37	Team	0.10
Dental	1.09	Ophthalmology	0.36	ENT	0.10
Physical Therapy	1.05	Hematology Oncology	0.34	HBPC	0.09
Podiatry	1.05	Cardiology	0.32	Pain	0.09
Discharge Summary	0.99	Immunization	0.30	Pulmonary	0.09
Emergency	0.92	Respiratory	0.28	Research	0.08
Eye	0.90	Preventive Medicine	0.24	Mental	0.08
Respiratory Therapy	0.87	Critical Care	0.19	Gastroenterology	0.08
Audiology	0.80	Urgent Care	0.18	Individual	0.08
SATP	0.71	Nurse Practitioner	0.17	Primary	0.07

Annotating and storing sections using OBSecAn

The original document's identification, the section annotation output including the identification, and the beginning and ending offsets of the annotated sections were saved for use with further information extraction tasks.

For efficiency in storing and querying data, we divided the one million notes into 20 sets of fifty-thousand notes. All fifty-thousand notes in each set were concatenated and saved into one large file. A note in this large file contains the note identification, the content of the note, and the end of note delimiter. We opened a file stream and annotated sections for all the notes in each of the large file and stored the annotation output into a corresponding file. An output file included the sections annotated for all fifty-thousand notes. The annotated output for each note in the output file includes the note identification, the section annotated information, and the end of note delimiter. The

execution times and disk spaces for storing the original notes and the annotation output for each of the fifty-thousand sets of notes are shown in Figure 4.

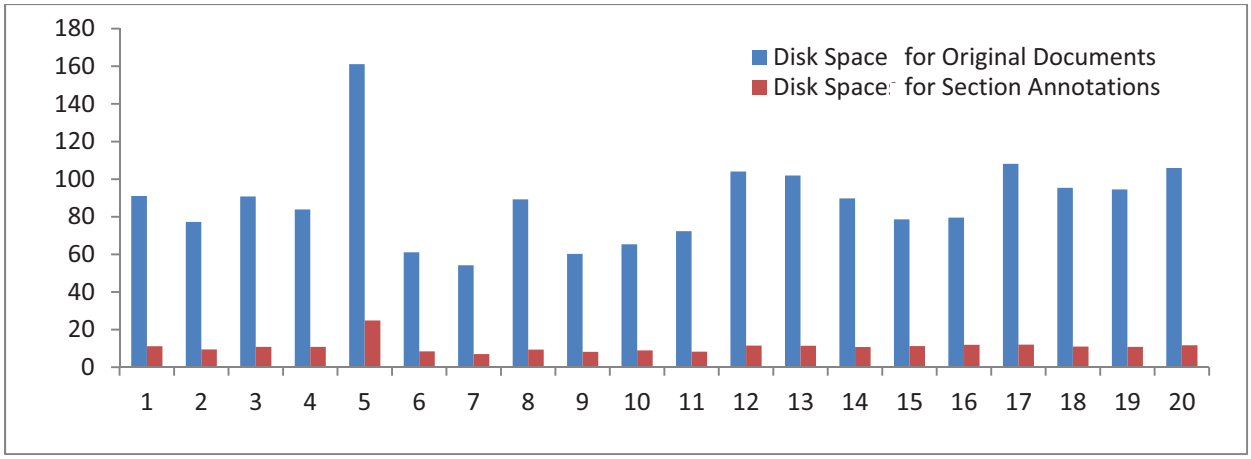

Figure 4. Disk space in Kilobytes (Kb) for storing the original medical notes and the section annotation outputs for 20 sets of fifty-thousand notes each.

High Yield Sections and Corresponding Note Titles

The frequency of a section that appears in notes varies by Note Titles. For each of the Note Titles, the ratio of the number of an annotated section to the total number of notes was used to determine the high yield sections for the corresponding Note Title. There are sections that may appear more frequently in some note title than others. For example, it would be appropriate to expect a completed family history section in a primary care note as opposed to an administrative note. Similarly, sections pertaining to nursing assessments such as the Braden Scale or skin assessment would be expected more frequently in a nursing note. Thus, depending on the information extraction task and domain knowledge, this method would identify a high yield section for natural language processing. For a representative set of 10 of the above Note Titles, we present a list of selected sections and their frequencies (in parentheses):

Nursing: Vital Signs (45007), Preventative Medicine (26874), Hygiene (26261), Mobility (24369), Activity Assessment (19238), Patient Education (14050), Review Of Systems (13510), Chronic Pain History (13242), Active Medications (12862), Physical Exam (12769), Functional Status (12153), Activities Of Daily Living (11514), Skin Reassessment (11484), Mental Status (10537), Pain Assessment (10252), Braden Scale (10218), Nursing Comments (10110), Nutrition (9848), Ambulation Status (9245), Functional Status Prior To Admission (9201), Current Skin Assessment (8556), Dressing (8215), Post Op Care (7960), Lab Data (7615), Chief Complaint (7450), Nursing Activity (7253), Morse Fall Scale (7221), Nursing Assessment (7080)

Primary care: Vital Signs (72565), Physical Exam (64143), Assessment (46917), Active Medications (44390), Pain Exam (43442), Preventative Medicine (39120), Review Of Systems (36975), Lab Data (32048), Hx Present Illness (30968), Chief Complaint (25914), Past Medical History (17748), Family History (7022), Allergies (12388), Alcohol Audit Screening (10330), Screen For Depression (8176), Clinical Impression (6906), Colorectal Cancer Screen (6361), Reason For Visit (6306), Personal Habits (6010),

Mental health: Assessment (13992), Mental Status (7417), Axis I (5524), Axis II (5170), Discharge Diagnosis (4548), Axis III (4367), Axis V (3897), Axis IV (3885), Active Medications (3578), Preventative Medicine (2812), Clinical Impression (2163), Family History (1201), Goal (1040), Gaf (1028), Mental Health/psychiatric (903),Substance Abuse Hx (808), Past Medical History (797), Diagnostic Impression (791), Hx Present Illness (722),

Inpatient: Physical Exam (10620), Vital Signs (10128), Assessment (8869), Active Medications (5274), Active Problems (3323), Allergies (3310), Admitting Diagnosis (3224), Hx Present Illness (3131), Family History (747), Chief Complaint (2913), Vitals Interpretation (2894)

Psychiatry: Assessment (6381), Mental Status (4373), Axis I (3156), Axis III (3021), Axis II (3004), Axis V (2969), Axis IV (2927), Clinical Impression (2738), Discharge Diagnosis (2387), Active Medications (1920), Preventative Medicine (1433), Active Problems (1021), Additional Diagnosis (867), Family History (344), Gaf (591), Past Medical History (587), Hx Present Illness (581), Lab Data (521), Mental Health/psychiatric (333)

Surgery: Preoperative Diagnosis (2812), Procedure (2763), Postoperative Diagnosis (2099), Vital Signs (1692), Review Of Systems (1472), Active Meds (1145), Physical Exam (1095), Reason For Procedure (923), Past Medical History (483), Active Problems (264), Discharge Medications (248), Informed Consent (245), Advance Directive (241)

Dental: Diagnosis (8374), Treatment Status (3358), Problem (3254), Past Medical History (393), Medications (539), Chief Complaint (292), Previous Treatments (263), Active Problems (244)

Physical Therapy: Physical Exam (979), Functional Status (928), Range Of Motion Exam (811), Past Medical History (677), Statement Of Goal (584), Bed Mobility (579), Procedure (543)

Podiatry: Physical Exam (6853), Past Medical History (2345), Musc Exam (2254), Lymp Nodes Exam (1770), Muscle Strength (1669), Range Of Motion (1636), Previous Treatments (1570), Medications (773), Active Problems (567)

Discharge Summary: Assessment (7932), Course In Hospital (7255), Hx Present Illness (6076), Physical Exam (5997), Discharge Medications (5070), Discharge Diagnosis (4772), Past Medical History (3996)

Table 2. Family history annotations by OBSecAn on the million document corpus. Annotated sections refers to those identified by OBSecAn from the respective documents. Labeled refers to those sections that were designated as "Family History" in the note.

Note Titles	Number of Annotated Sections	Labeled Sections (% of total annotations)	Un-labeled Sections	True Family History Annotation (% of total annotations)	Number of Incorrect Annotation	Incorrect due to the Ending Offset
Primary Care	7022	6753 (96)	269	6916 (98)	106	52
Nursing	4384	4315 (98)	69	4306 (98)	78	55
Mental Health	1201	1190 (99)	11	1177 (98)	24	21
Pharmacy	677	664 (98)	13	659 (97)	18	11
Int Medicine	536	530 (99)	6	536 (100)	0	0
Administrative Note	603	597 (99)	6	596 (99)	7	3
Inpatient	747	743 (99)	4	726 (97)	21	15
Psychiatry	344	335 (97)	9	338 (98)	6	0
Social Work	390	384 (98)	6	387 (99)	3	0
Surgery	307	302 (99)	5	279 (91)	28	26
Dental	115	111 (97)	4	112 (97)	3	0
Physical Therapy	236	235 (100)	1	236 (100)	0	0
Podiatry	219	214 (98)	5	211 (96)	8	5
Discharge	660	657 (100)	3	648 (98)	12	7
Emergency	291	289 (99)	2	285 (98)	6	2
Eye	381	380 (100)	1	381 (100)	0	0
Respiratory	129	127 (98)	2	127 (98)	2	0
Audiology	226	222 (98)	4	219 (97)	7	6
SATP	234	231 (99)	3	233 (100)	1	0
Physical Medicine	171	170 (99)	1	166 (97)	5	2

Accuracy of OBSecAn Annotation of Family History Section

As shown in Table 2, the family history section is present in many note titles in varying frequencies. Primary care, nursing, and mental health are the three note titles with the highest yield for family history sections. It was interesting to note that the family history section was present in nearly half the note titles. This section was labeled as such in almost all instances in all note titles. The rate of accurately identifying the family history as such was between 93% and 100% with a median of 99% in the top 20 note titles shown in Table 2. Most of the inaccuracies were due to an incorrect ending offset of the section.

Discussion

We demonstrate the feasibility of accurately identifying a specific section in VA electronic medical notes, using OBSecAn, an automated ontology-based section annotator. OBSecAn was first trained and tested on a small corpus of documents and now has been scaled out to a million documents from the VA.

Several methods have been developed to sectionize a document. Extending current methods to include a robust clinical documents section ontology and a logical sequence of steps to parse, correct, and ultimately annotate a section based on the final parsed tree are contributions of OBSecAn to the field of automated information extraction. The computational intensity of this algorithm, while significant, is feasible and practical. The stored annotations represent only a fraction of the storage needed for the original documents and thus represents economy of scale while dealing with large document corpora.

The identification of 'high yield' sections from specific document note titles represents another conceptual innovation that has the potential to increase the efficiency of information extraction tasks. As noted above, the family history section is present in higher frequency in certain notes, such as primary care. While this is intuitive based on domain knowledge and workflow of clinicians, it offers an objective metric to identify high yield notes based on the specific task and use case. With the example of family history, it was interesting to note that the sections were appropriately labeled as family history in nearly all the top documents identified by their frequency count of the presence of family history sections. The accuracy with which OBSecAn was able to identify a section as family history was nearly perfect in most of the high yield documents reviewed.

We acknowledge several limitations. Even though the million document corpus was representative of the entire VA document corpus, it is possible that there are variations in the documents with respect to their semi-structured format and section beginning and ending offsets. Thus, it is important to further evaluate the sectionizer on a larger scale out. Rather than brute force application of OBSecAn to a larger corpus, we are currently preparing to apply the method to a large corpus and specific use cases such as symptom extraction, homelessness, and detecting the presence of indwelling urinary catheters in hospitalized patients. Accurately identifying a section achieves only one objective. The information extraction task based on specific use cases (such as the family history example discussed above) would need to be evaluated for the relevance of the information extracted. Based on pilot reviews of the textual content of the family history section of several hundred documents, it appears there are several types of templates and free text modalities for recording the family history. Thus, extracting information from this section is a major challenge. We developed this method exclusively on VA documents and, therefore, it may not perform as well in other settings. However, we are interested in extending and applying OBSecAn to electronic notes from other healthcare systems with electronic medical records

Conclusion

There is a growing recognition of the importance of identifying and labeling sections in clinical notes. The main purpose of the task is to provide the document context for efficient and accurate information extraction and retrieval. In this paper, we have improved on the original implementation of the OBSecAn for automated section identification on a large corpus of electronic medical notes. To the best of our knowledge, this is the first ever attempt to apply and evaluate a section annotator on a large scale.

Acknowledgements: This work was supported by U.S. Department of Veterans Affairs, Veterans Health Administration, Office of Research and Development, Health Services Research and Development Projects HIR 10-001 (PI: Samore), HIR-10-002 (PI: Gundlapalli), and U.S. National Library of Medicine, National Institutes of Health training grant T15LM007124 (Tran). We would like to express our gratitude to the administration and staff of the VA Informatics and Computing Infrastructure (VINCI) for their support of our project. We also acknowledge

the staff, resources, and facilities of the VA Salt Lake City IDEAS Center for providing a rich and stimulating environment for NLP research.

Disclaimer: The views expressed in this paper are those of the authors and do not necessarily represent the views of the U.S. Department of Veterans Affairs or the United States Government.

References

1. Pakhomov S, Bjornsen S, Hanson P, Smith S. Quality Performance Measurement using the Text of Electronic Medical Records. Med Decis Making. Jul-Aug 2008; 28(4):462-470.
2. Rao RB, Krishnan S, Niculescu RS. Data Mining for Improved Cardiac Care. SIGKDD Explor. Newsl. 2006;8(1):3-10.
3. Percha B, Garten Y, Altman RB. Discovery and Explanation of Drug-Drug Interactions via Text Ming. Pacific Symposium on Biocomputing 2012; 410-421.
4. Tatonetti NP, Ye PP, Daneshjou R, Altman RB. Data-Driven Prediction of Drug Effects and Interactions. Science Translational Med. 2012 March 14; 4(125): 125ra31. DOI: 10.1126/scitranslmed.3003377
5. Nadkarni PM1, Ohno-Machado L, Chapman WW. Natural Language Processing: An Introduction. J Am Med Inform Assoc. 2011 Sep-Oct; 18(5):544-551.
6. Haug PJ, Wu X, Ferraro JP, Savova GK, Huff SM, and Chute CG. Developing a Section Labeler for Clinical Documents. AMIA Annu Symp Proc 2014; 2014:636-643.
7. Finch DK, McCart JA, and Luther SL. TagLine: Information Extraction for Semi-Structured Text in Medical Progress Notes. AMIA Annu Symp Proc 2014; 2014:534-543.
8. Denny JC, Spickard A, Johnson KB, Peterson NB, Peterson JF, Miller RA. Evaluation of a Method to Identify and Categorize Section Headers in Clinical Documents. J Am Med Info Assoc. 2009;16(6):806-815. DOI 10.1197
9. Hyun S, and Bakken S. Toward the Creation of an Ontology for Nursing Document Sections: Mapping Section Headings to the LOINC Semantic Model. AMIA Annu Symp Proc 2006; 2006:364-368
10. Denny JC, Miller RA, Johnson KB, and Spickard A. Development and Evaluation of a Clinical Note Section Header Terminology. AMIA Annu Symp Proc. 2008;2008:156-160
11. Visual Tagging Tool. (Last updated: 10/17/2013; cited 07/08/2015). http://lsg3.nlm.nih.gov/LexSysGroup/Projects/vtt/current/web/index.html
12. Meystre S, Haug PJ. Automation of a Problem List using Natural Language Processi30ng. BMC Med Inform Decis Mak 2005; 5:30
13. US Department of Veterans Affairs. *VA Informatics and Computing Infrastructure (VINCI)*. 2015 [cited 2015; Available from: http://www.hsrd.research.va.gov/for_researchers/vinci/