Statistical Section Segmentation in Free-Text Clinical Records

Michael Tepper[1], Daniel Capurro[2], Fei Xia[1,2], Lucy Vanderwende[3,2], Meliha Yetisgen-Yildiz[2,1]

[1]Department of Linguistics, [2]Biomedical and Health Informatics
University of Washington, Seattle, WA 98195, USA
[3]Microsoft Research, Redmond WA 98052, USA
{mtepper, dcapurro, fxia}@uw.edu, lucy.vanderwende@microsoft.com, melihay@uw.edu

Abstract

Automatically segmenting and classifying clinical free text into sections is an important first step to automatic information retrieval, information extraction and data mining tasks, as it helps to ground the significance of the text within. In this work we describe our approach to automatic section segmentation of clinical records such as hospital discharge summaries and radiology reports, along with section classification into pre-defined section categories. We apply machine learning to the problems of section segmentation and section classification, comparing a joint (one-step) and a pipeline (two-step) approach. We demonstrate that our systems perform well when tested on three data sets, two for hospital discharge summaries and one for radiology reports. We then show the usefulness of section information by incorporating it in the task of extracting comorbidities from discharge summaries.

Keywords: Document segmentation, Clinical NLP, Text classification

1. Introduction

Accessibility to the details of patient data available in clinical records is critical to improve the health care process and to advance clinical research (Friedman and Johnson, 2005; Friedman, 2005). Although clinical records are free text, they are often structured in terms of sections. As health care providers create their reports, they typically use some conceptual or electronic templates to divide their narratives into general sections. There are many report types, which are created for different purposes. For example, admit notes describe the state of a patient at the time of hospital admission, and discharge summaries summarize the overall hospital course and the state of the patient at the time of hospital discharge. These report types have different characteristics in terms of section headings and content (e.g., discharge summaries often include a *medications at discharge* section, which is absent from admit notes). Furthermore, physicians can freely modify the sections defined in a template; as a result, reports of the same report type may have different sections or section headers, and such differences are more prominent for reports coming from different hospital divisions or different hospitals.

Accurate identification of section boundaries and section types in clinical reports can help various automated tasks such as named-entity recognition and sense disambiguation of the identified entities. As an example, the acronym BS has the following three senses (1) *bowel sounds* if found in the abdominal exam section, (2) *breath sounds* if found in the chest exam section, and (3) *blood sugar* if found in the laboratory test section.

In this paper, we describe our work towards building a statistical section segmenter for clinical documents. We report performance results on three datasets and two clinical report types. Finally, as an example application, we demonstrate the usefulness of the sections assigned by our procedure for extracting comorbidity information from discharge summaries.

2. Related Work

The problem of section segmentation for scientific literature is related to our task and has been studied fairly extensively. The basic idea behind this work is to identify and mark the underlying structure of scientific papers and abstracts in order to improve tasks such as information extraction and automatic summarization. In practice this tends to involve training a classifier to determine a section category label for each sentence of a document.

One major research track focuses on recovering section header labels from MedLINE abstracts. The goal of this work is to classify sentences in abstracts into sections such as *Introduction, Method, Result*, and *Conclusion* (McKnight and Srinivasan, 2003; Lin et al., 2006; Hirohata et al., 2008). A sampling of methods that have been tried for this problem include a non-sequential classifier based on SVM (McKnight and Srinivasan, 2003), and sequential classifiers based on HMM (Lin et al., 2006) as well as CRF (Hirohata et al., 2008). Performance on this task has hovered around 90% per-sentence accuracy, with Hirohata et al. (2008) topping performance at 94.3%. Hirohata's approach was notable, demonstrating clear benefits from using sequential classification (CRF) over non-sequential classification (SVM) and from augmenting sequence (section) tags with B- and I- prefixes.

A second major track called *Argumentative Zoning* is focused on uncovering basic discourse structure in scientific documents (Teufel, 1999). Under this framework, each sentence from a scientific paper is classified under seven rhetorical categories such as *background, other (researcher's work)*, and *own (work)*. Sequential tagging approaches have dominated this domain, such as the Naïve Bayes (NB) approach of Teufel and Moens (2002) and the Maximum Entropy (MaxEnt) approach of Merity et al (2009). Merity's MaxEnt approach topped out at 96.88% F-score, which substantially outperformed previous work. This approach showed the benefits of using a discrimina-

Dataset	Report Type	Report Count	Section Count	Avg. Sections / Report	Avg. Words / Report	Annotators
#1	Discharge Summary	191	2527	13.2	958.8	2 w/ med. training
#2	Discharge Summary	183	2067	11.3	775.2	1 w/o med. training
#3	Radiology Report	100	594	5.9	271.5	1 w/o med. training

Table 1: Statistics of the three datasets used in our study.

tive classifier with simple n-gram features over Teufel and Moens's NB approach with a more complex feature set.

2.1. Section Segmentation in the Clinical Domain

Section segmentation (i.e., identification of section boundaries) in the clinical domain is less well studied, and past approaches have focused mainly on section classification (i.e., labeling a section with a pre-defined section category), while relying on hand-coded heuristics for detecting section boundaries. Section classification in clinical records is a difficult problem, as clinicians do not follow strict section naming conventions. While it is common practice to define auto-populated templates for report types in Electronic Medical Record (EMR) systems, clinicians have the flexibility to modify those defined templates or create their own depending on their needs at the time.

Denny et. al (2008, 2009) trained an NB classifier on a set of clinical notes that were annotated based on a manually created hierarchical section terminology. Using 10,767 clinical notes for development and 540 for test, Denny et al. (2009) reported per-section performance at 99.0% recall and 95.6% precision. Li et al. (2010) defined section classification as a sequence-labeling problem and used Hidden Markov Model (HMM) to classify sections in medical records by finding an optimal sequence of section categories. Section headers were mapped to 15 manually selected general section categories (e.g., chief complaint). With a dataset of 9697 clinical notes (78% used for training, 22% used for testing), the classifier achieves a per-section accuracy of 0.93 and a per-note accuracy of 0.70.

While both of the above-mentioned approaches have achieved good results, their main limitation is that they rely on hand-coded heuristics for section segmentation (boundary detection), which may be difficult to replicate or extend. These heuristics are based on conventions like the capitalization of headers and the presence of blank lines between sections (Li et al., 2010). While Denny et al. (2009) use a larger set of targeted heuristics, they mention that formatting and style differences may vary across clinical settings, which is a weakness of their approach. In this study, we propose a heuristics-free machine learning approach for both section segmentation and classification. This approach can be adapted to a new clinical setting simply by annotating new training data, rather than having to commit developer time and resources to extend the hand-coded heuristics.

3. Our Datasets

We used three datasets composed of discharge summaries and radiology reports to develop our statistical section segmenter and test its performance. A detailed summary of the datasets is presented in Table 1. In this section, we will describe each of the three datasets, present the ontology created for annotation, and provide information about the annotation effort.

3.1. Dataset 1 – UW Discharge Summary Corpus

This corpus consists of 430 discharge summaries of 402 patients who had a surgery at UW's medical center in 2010[1]. The retrospective review of those reports was approved by the UW Human Subjects Committee of Institutional Review Board, who waived the need for informed consent. We used 191 randomly selected discharge summaries from this corpus to build the section category ontology and to create the gold standard for the statistical section segmentation task. The whole dataset was then used for an extrinsic evaluation consisting of comorbidity extraction from discharge summaries, to demonstrate the performance of the segmenter in a real-world application.

3.2. Dataset 2 – i2b2 Discharge Summary Corpus

This corpus was created for the 2010 i2b2 natural language processing challenge on medical concept, assertion, and relation extraction (Uzuner et al., 2011). The corpus consists of 835 discharge summaries from three institutions (Partners HealthCare, Beth Israel Deaconess Medical Center, and University of Pittsburgh Medical Center). We used 183 randomly selected discharge summaries from this corpus to test the generalizability of the proposed approach on discharge summaries created by different institutions.

3.3. Dataset 3 – UW Radiology Report Corpus

This corpus consists of 100 radiology reports extracted from the UW Radiology Information System. The reports contain a mixture of imaging modalities including radiographs, CT scans, ultrasounds, and magnetic resonance imaging (MRI). The retrospective review of those reports was approved by the UW Human Subjects Committee of Institutional Review Board, who waived the need for informed consent.

3.4. Section Category Ontology

We constructed an ontology of 33 section categories for discharge summaries (see Table 2) and an ontology of 11 section categories for radiology reports (see Table 3). The ontologies have been designed to cover typical discharge summary sections (across two datasets) and radiology report sections, as advised by a clinical expert.

[1] 24 patients had at least two reports generated for them under the report type discharge summary (one detailed discharge summary and additional notes from other attending clinicians).

Section Categories	Freq (Percentage)	
	Dataset 1	Dataset 2
GENERAL PATIENT INFO		
Admit Date	180 (7.1%)	170 (8.2%)
Discharge Date	181 (7.0%)	182 (8.8%)
Service	10 (0.4%)	33 (1.6%)
PROVIDER INFO		
Attending	66 (2.6%)	103 (5.0%)
Admit Physician	4 (0.2%)	1 (0.1%)
Discharge Physician	2 (0.1%)	0 (0.0%)
CONDITION BEFORE ADMISSION		
Admission Diagnoses	98 (3.9%)	56 (2.7%)
History	81 (3.2%)	81 (3.9%)
Medications	53 (2.1%)	53 (2.1%)
Reason for Admission	9 (0.4%)	29 (1.4%)
CONDITION AT DISCHARGE		
Condition	75 (3.0%)	64 (3.1%)
Disposition	103 (4.1%)	53 (2.5%)
Discharge Diagnoses	156 (6.2%)	148 (7.2%)
Other Diagnoses	5 (0.2%)	62 (3.0%)
Physical Exam on Disch.	39 (1.5%)	0 (0.0%)
MEDICAL HISTORY		
Allergies	56 (2.2%)	64 (3.1%)
Family History	31 (1.2%)	26 (1.3%)
Gynecological History	0 (0.0%)	3 (0.2%)
Past Medical History	73 (2.9%)	66 (3.2%)
Past Surgical History	69 (2.7%)	14 (0.7%)
Social History	44 (1.7%)	49 (2.4%)
HOSPITAL COURSE		
Consultation	148 (5.9%)	17 (0.8%)
Hospital Course	115 (4.6%)	140 (6.8%)
Physical	15 (0.6%)	75 (3.6%)
Procedures	182 (7.2%)	98 (4.7%)
Studies	0 (0.0%)	76 (3.7%)
DISCHARGE INSTRUCTIONS		
Follow up	180 (7.1%)	61 (3.0%)
Diagnostic Studies Rec'd	24 (1.0%)	0 (0.0%)
Discharge Instructions	146 (5.8%)	104 (5.0%)
Discharge Medications	179 (7.1%)	124 (6.0%)
ADDENDA		
Attending Statement	7 (0.3%)	0 (0.0%)
Note	0 (0.0%)	0 (0.0%)
OTHER		
Catchall	42 (1.7%)	19 (0.9%)
Combined	127 (5.0%)	37 (1.8%)
Total	2527 (100%)	2067 (100%)

Table 2: Section category ontology for discharge summaries and category frequencies in Datasets 1 and 2.

To build each ontology, we first created a list of sections by sampling a small subset of twenty discharge summaries from Dataset 1 and ten radiology reports from Dataset 3. Then, with the help of an expert, we (a) grouped similar sections together under general categories, and (b) put rare or atypical sections under a catch-all category. For example, when building the ontology for discharge summaries, we grouped *Procedures*, *Surgical Procedures*, and *Operations* under the section category *Procedures*. When we came across *Impression*, we put it under the catch-all category, as it is not typical in a discharge summary (in fact, it is usually found in a radiology report).

Section Categories	Freq (Percentage)
	Dataset 3
CLINICAL INFORMATION	
Clinical History	99 (16.7%)
EXAM DETAILS	
Exam	8 (1.4%)
Comparison	89 (15.0%)
Contrast	22 (3.7%)
Procedure	70 (11.8%)
FINDINGS	
Findings	100 (16.8%)
IMPRESSION	
Impression	77 (13.0%)
Attending Statement	14 (2.4%)
OTHER	
Document Header	104 (17.5%)
Catchall	6 (1.0%)
Combined	5 (0.8%)
Total	594 (100%)

Table 3: Section category ontology for radiology reports and category frequencies in Dataset 3.

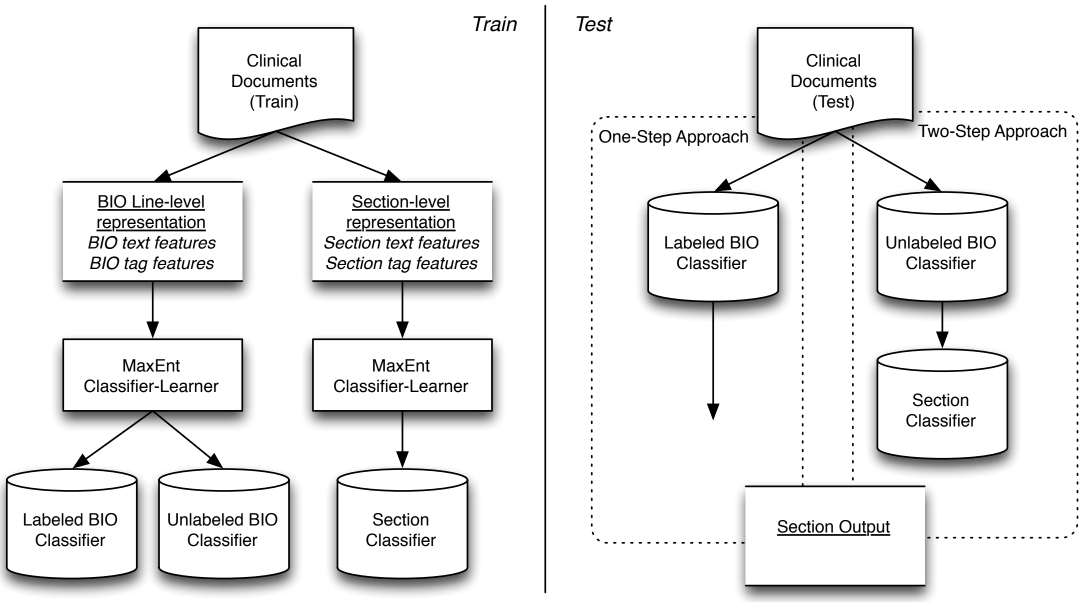
3.5. Annotation Task

There were three annotation tasks conducted as part of this study. Our annotators were instructed to mark the section header and select an appropriate category from the ontology. For combined sections (e.g., History and Physical) annotators were instructed to annotate the header multiple times (e.g., once with History as the section category and again with Physical Exam). Dataset 1 had the most combined sections at 5% (see Table 2), whereas Dataset 3 had the fewest, at 0.8% (see Table 3).

For the first annotation task, two annotators annotated a total of 191 discharge summaries from Dataset 1, of which 24 summaries were double annotated. One of the annotators is a clinical expert who is a medical doctor with 7 years clinical experience. The other annotator is an experienced medical records specialist. Together they annotated 2527 sections in 191 discharge summaries, averaging 13.2 sections per document.

The inter-annotator agreement f-measure was 0.91 when matching section-header boundaries and section category, and 0.95 when matching header boundaries alone. One main source of disagreement was that annotators sometimes overlooked a few headers in a given report. When the annotators agreed on the location of a header, agreement on the section category was high, at 0.97 observed agreement. For the second annotation task, there was a single annotator who annotated 2067 sections in 183 discharge summaries from Dataset 2, averaging 11.3 sections per document. In the third annotation task, we had single annotator who annotated 594 sections in 100 radiology reports from Dataset 3, averaging 5.9 sections per document. For both the second and third annotation tasks, the annotators did not

Figure 1: Training and testing stage in the one-step and two-step approach.



have any clinical background.

4. Methods

Our basic methodology for section segmentation is to classify each line in a document to indicate its membership to a section. Our classifier operates at the line-level rather than the sentence-level, as content of clinical records tends to be fragmentary and list-based. Similar to Hirohata et al. (2008), we relied on BIO tags to differentiate the beginnings of sections (which tend to consist of headers) from the remaining lines.

Under this methodology we tried two approaches: a joint (one-step) approach and a pipeline (two-step) approach. Both approaches are described below.

4.1. One-step approach

This approach uses a section segmentation model that has been enriched with section category labels such that it segments and classifies sections in one step. To be more specific, we have extended the BIO tags with category labels X; that is, B-X and I-X indicate that the current line begins (B) or lies inside (I) a section with category X; O means the current line is not in any section (e.g., a blank line at the beginning of a document).

4.2. Two-step approach

This approach relies on separate models for section segmentation and classification. First, the section boundaries are identified by labeling each line with a B, I, or O tag. Then, the unlabeled sections from the first step are passed to the second step, where a separate classifier is called upon to label each section with the appropriate section category.

4.3. Modeling

For both approaches, we used Maximum Entropy (MaxEnt) models for classification (Berger et al. 1996), and used beam search to find a good tag sequence. We used the MALLET toolkit v2.07 (McCallum, 2002) with L-BFGS parameter estimation and Gaussian prior smoothing. The Gaussian prior variance was left at its default value (=1).

4.4. Features

There were two types of tagger developed for these experiments: one type labels each line in a document; it is used for the one-step approach as well as the 1st step of the two-step approach. The other type labels each section in a document, and it is used for the 2nd step of the two-step approach. The feature sets for each type are described below.

Features for line labeling

Table 4 shows features used in the one-step approach and step 1 of the two-step approach. The text features look at the shape (e.g., capital letters, numbers, blank lines) and content (e.g., first token in a line, any unigram) of current and neighboring lines; the tag features look at the tags of previous lines and how many lines have the same tag (a.k.a. *tagChainLength*).

Type	Features
Text features	*isAllCaps, isTitleCaps, containsNumber, beginsWithNumber, numTokens, numPreBlanklines, numPostBlanklines, firstToken, secondToken, unigram*
Tag features	*prevTag, prevTwoTag, tagChainLength*

Table 4: Features for line labeling.

Features for section labeling

Table 5 shows features used in Step 2 of the two-step approach. Header features are the same as text features in Table 4, and they are extracted only from the first line of a

	Dataset 1		Dataset 2		Dataset 3	
Exp#	Unlabeled Prec/Rec/F1	Labeled Prec/Rec/F1	Unlabeled Prec/Rec/F1	Labeled Prec/Rec/F1	Unlabeled Prec/Rec/F1	Labeled Prec/Rec/F1
1	91.3/90.7/91.0	87.1/86.5/86.8	90.8/82.5/86.4	85.9/78.1/81.8	93.8/88.8/91.2	91.5/86.5/88.9
2	93.4/94.7/94.1	88.2/89.4/88.8	93.2/92.9/93.1	87.6/87.3/87.5	91.8/89.7/90.7	89.3/87.2/88.2
3	93.4/94.7/94.1	88.5/89.7/89.1	93.2/92.9/93.1	82.7/86.9/87.0	91.8/89.7/90.7	88.9/86.8/87.8
4	93.4/94.7/94.1	87.9/89.1/88.5	93.2/92.9/93.1	86.7/86.4/86.5	91.8/89.7/90.7	88.9/86.8/87.8

(a) Section mapping performance.

	Dataset 1		Dataset 2		Dataset 3	
Exp#	Unlabeled Prec/Rec/F1	Labeled Prec/Rec/F1	Unlabeled Prec/Rec/F1	Labeled Prec/Rec/F1	Unlabeled Prec/Rec/F1	Labeled Prec/Rec/F1
1	96.3/95.7/96.0	91.5/90.9/91.2	97.8/88.9/93.1	92.2/83.8/87.8	97.5/92.2/94.7	94.7/89.5/92.0
2	96.6/97.9/97.3	91.2/92.4/91.8	97.0/96.7/96.8	91.1/90.8/91.0	96.8/94.6/95.6	94.1/91.9/92.9
3	96.6/97.9/97.3	91.5/92.7/92.1	97.0/96.7/96.8	90.7/90.4/90.5	96.8/94.6/95.6	93.6/91.5/92.5
4	96.6/97.9/97.3	91.1/92.4/91.8	97.0/96.7/96.8	90.0/89.7/89.8	96.8/94.6/95.6	93.2/91.1/92.1

(b) Header mapping performance.

Table 6: System performance on all three datasets. All the results based on 5-fold cross validation over each of the three datasets. For each dataset, we ran four experiments: Experiment 1 is the one-step approach with features in Table 4. Experiments 2, 3, and 4 are the two-step approach which use the features in Table 4 for the first step, and different features from Table 5 for the second step. Experiment 2 uses header features only; Experiment 3 uses header and tag features; Experiment 4 uses all three features types in Table 5.

Type	Features
Header features	*Same as text features for line labeling, but only the header line is used*
Body features	*avgLineLength, numLines, docPosition, containsList, unigram*
Tag features	*prevTag, tagHistUnigram, tagChainLength*

Table 5: Features for section labeling.

section. Body features look at the shape (e.g., *containsList*, *avgLineLength*) and content (e.g., *unigram*) of the whole section and relative position of the section in the document by quintiles (*docPosition*). Tag features look at tags of previous sections, capturing regularities in section ordering (for example, Discharge Date typically follows Admit Date in the discharge summaries in Dataset 1).

5. Experiments

For evaluation, we ran 5-fold cross evaluations on each dataset and reported the micro-average of the five runs.

5.1. Evaluation measures

For evaluation, we calculated precision, recall, and f-score of header matching and section matching. For header matching, there is a match when a line is marked as the first line of a section by both the gold standard and the system output. For section matching, there is a match when a sequence of lines is marked as a section by both the gold standard and the system output. Section matching is a stricter measure than header matching, as one wrongly identified section header is one error in header matching, but could result in incorrect boundaries for two sections. For both types of matching, an unlabeled match checks only the location of a header or the boundary of a section, whereas a labeled match checks location/boundary as well as the category of the section.

5.2. Results

One-step vs two-step approach

Tables 6(a) and 6(b) show the performance of each system, measured by section and header matching, respectively. For the one-step approach and Step 1 of the two-step approach, we used the features in Table 4 (Experiment 1). For Step 2 of the two-step approach, we ran three experiments (Experiments 2, 3, 4) with different combinations of features in Table 5. For decoding in all experiments, we ran a beam-search with the beam size set to 100.

For Datasets 1 and 2, the performance results show that the two-step approach (Experiments 2, 3, and 4) outperforms the one-step approach (Experiment 1). This is likely because the joint model used by the one-step approach has sparse features, as they must be calculated over 67 categories, while the two-step approach has a segmentation model with just three categories, so it is better trained.

Several recent studies (e.g., Zhang and Clark (2010), which combines word segmentation and POS tagging) have shown that joint models outperform the pipeline approach when the benefits of allowing the two tasks to provide constraints and feedback to each other outweigh potential drawbacks due to multiplication of the two tag sets. This condition does not seem to hold for the current task, and as a result, the two-step, pipeline approach outperforms the one-step, joint approach. For Dataset 3, the two-step approach slightly decreased the performance; however, the differences are too small to draw strong conclusions. One contributing factor may be that the section category ontology for Dataset 3 is comparably much smaller than for the other two Datasets. This means the tagset used in the one-step approach is much smaller, and thus the one-step segmentation model can be better trained on a comparably small dataset.

For Step 2 of the two-step approach, adding tag features to header features helped a little bit for Dataset 1, but not for Datasets 2 or 3. Meanwhile, using all three types of features decreased the performance slightly for all datasets, which is the result of overtraining on somewhat noisy annotations.

Error analysis on Datasets 1 and 2 reveals that broad categories that cover a diverse array of sections perform the worst. The catch-all category had the worst performance. The *Studies* category also performed quite poorly. It was used for any section describing test or lab results (e.g., *Laboratory Data on Admission, Radiology, Radiology / Imaging*). Many sections in Datasets 1 and 2 that should have received these broad categories were not observed in the training data. This is not surprising as the catch-all category was designated for rare and atypical sections, and the *Studies* category encompassed a long tail of rare sections for the myriad of labs and studies that are performed in a clinical setting. In both datasets, there were also many errors with combined categories because frequently the needed combination at test time had not been observed in the training data.[2]

Performance differences across section categories

Table 7 includes f-scores for each section category. The scores are from the best performing systems on Dataset 1 and Dataset 2, respectively.

Although both Dataset 1 and Dataset 2 include discharge summaries, there are major performance differences for some categories. The f-scores for *Admission Diagnoses* are 96.0 for Dataset 1 and 63.0 for Dataset 2. Error analysis revealed that *Admission Diagnoses* was a particularly noisy category for Dataset 2, but not for Dataset 1. Dataset 2 is currently undergoing annotation revisions which may improve performance on this category. Another large performance difference occurs with the aggregate category *Combined*, which includes all cases where a section belongs to multiple categories. Here the f-scores are 84.1 for Dataset 1 and 47.1 for Dataset 2. This is likely the result of sparse training data for Dataset 2, as Dataset 2 has only 37 instances of combined sections, compared to 127 for Dataset 1.

Domain adaptability across datasets

Table 8(a) and 8(b) reveal that the system performance degrades significantly when the training and test data do not come from the same dataset, despite both sets being discharge summaries.

Section Categories	F-score	
	Dataset 1	Dataset 2
General Patient Info		
Admit Date	97.5	98.3
Discharge Date	97.8	98.4
Service	84.2	95.8
Provider Info		
Attending	93.3	93.3
Admit Physician	28.6	0.0
Discharge Physician	0.0	–
Condition Before Admission		
Admission Diagnoses	96.0	63.0
History	95.7	92.4
Medications	95.2	86.5
Reason for Admission	76.9	82.5
Condition at Discharge		
Condition	90.4	96.7
Disposition	96.6	92.5
Discharge Diagnoses	90.9	87.2
Other Diagnoses	76.9	97.6
Physical Exam on Discharge	74.4	–
Medical History		
Allergies	99.1	97.6
Family History	93.5	100
Gynecological History	–	66.7
Past Medical History	93.2	96.3
Past Surgical History	97.8	96.6
Social History	93.2	97.0
Hospital Course		
Consultation	97.1	87.5
Hospital Course	97.4	95.5
Physical	61.5	97.4
Procedures	96.5	95.3
Studies	–	64.3
Discharge Instructions		
Follow-up	92.6	95.1
Diagnostic Studies Rec'd	93.9	–
Discharge Instructions	74.9	85.7
Discharge Medications	96.2	92.4
Addenda		
Attending Statement	–	–
Note	–	–
Other		
Catchall	29.7	27.3
Combined	84.1	47.1
Total	92.1	90.8

Table 7: F1-scores by section category on Datasets 1 and 2. These results are from Experiment 3 for Dataset 1 and Experiment 2 for Dataset 2.

Upon a closer look, it becomes apparent that the two datasets have different characteristics. Clinical language and report style vary across different institutions. Stylistically, Dataset 2 has more header variation. For example, headers can be split between two lines and contain punctuation, while in Dataset 1 this does not occur. Content wise, one difference is that Dataset 2 has been anonymized while Dataset 1 has not. Also, these sets contain different variations for header labels. For example PMH is a common way of writing Past Medical History in Dataset 1, but not in Dataset 2. Finally, the frequency and distribution of combined sections can make a difference as well. Dataset 1 has many more combined sections than Dataset 2, as can be seen on Table 8. Thus, when training on Dataset 1 and

[2] Combined categories exist because some sections fall into multiple categories (e.g., *History*, *Physical*). These categories are concatenated to form single categories during training (e.g., *History & Physical*). This way, the algorithm only needs to assign one category per section. The downside is this greatly increases the label set, and combined categories are frequently undertrained (or unseen).

	Unlabeled Prec/Rec/F1	Labeled Prec/Rec/F1
Train: Dset 1 Test: Dset 2	58.2/73.9/65.1	48.6/61.8/54.4
Train: Dset 2 Test: Dset 1	77.9/52.0/62.4	65.6/43.8/52.5

(a) Section mapping performance.

	Unlabeled Prec/Rec/F1	Labeled Prec/Rec/F1
Train: Dset 1 Test: Dset 2	67.4/85.7/75.5	55.6/70.6/62.2
Train: Dset 2 Test: Dset 1	89.8/60.0/71.9	75.6/50.5/60.5

(b) Header mapping performance.

Table 8: Experiments to measure domain adaptability across datasets. Experiment settings are equivalent to Experiment 2: two-step approach with header features.

testing on 2, many combined headers will not be identified at test time.

6. Application – Comorbidity Extraction from Discharge Summaries

For an extrinsic measure, the section segmenter was applied to the problem of comorbidity extraction from clinical records. In medicine, comorbidity is defined as the presence of one or more disorders or diseases in addition to a primary disease or disorder (Valderas et al., 2009). For instance, a patient accepted for cancer treatment may also have other disorders such as diabetes and hypertension, which are called comorbidities for this parent. As part of a quality improvement (QI) project, annotators affiliated with UW manually abstracted each of the 402 patients in Dataset 1 for a long list of data elements including the presence or absence of the following four types of comorbidities; sleep apnea, diabetes, asthma, and hypertension[3]. The annotators had access to the complete set of clinical notes generated during the patients hospital stays (e.g., admit notes, discharge summaries, operative notes). Because the section segmenter was trained only on discharge summaries, this extraction study was on a subset of Dataset 1 consisting of 435 discharge summaries.

The baseline comorbidity extraction system has two main steps. The first step identifies medical concepts in discharge summaries with MetaMap (Aronson, 2001; Aronson and Lang, 2010) and the second step checks the presence of the comorbidities in the list of identified medical concepts. The results are shown as "baseline" in Table 9. The error analysis indicates that many of the false positives were due to appearance of those comorbidities in sections that are not about the patient (e.g., father with diabetes in the *family history* section).

We extended the baseline approach by introducing section information from the section segmenter. This involved manually identifying 14 section categories that were most likely to contain comorbidity information related to the patient (e.g., admission diagnoses, past medical history) and excluding the rest. Thus comorbidities were only identified in the list of medical concepts found under those 14 section types, and sections such as *Family History* were excluded.

The performance is listed in the "system with section" rows in Table 9. As can be seen from the table, with section information, precision increased for diabetes and hypertension and did not change for sleep apnea and asthma. Recall remained the same except for a slight decrease. Overall, introducing section information increased the micro averaged performance both in terms of precision and f-score.

Comorbidity	System	Prec/Rec/F1
Diabetes	Baseline	87.0/35.1/50.0
	System with section	87.0/35.1/50.0
Asthma	Baseline	82.8/84.1/83.5
	System with section	89.5/81.0/85.0
Hypertension	Baseline	82.1/60.5/69.7
	System with section	82.1/60.5/69.7
Diabetes	Baseline	88.8/75.8/81.8
	System with section	92.2/75.8/83.2
Microaverage	Baseline	85.9/**65.8**/74.5
	System with section	**89.4**/65.0/**75.3**

Table 9: Performance results for comorbidity extraction. The results were collected for 402 patients.

7. Discussion

Our experiments show that the system works well when applied to three different datasets. Compared to previous research, which relies on heuristic based rules, the approach described here requires only a small annotated dataset. The experiments in Section 6 demonstrate that running this system (two-step approach) as a preprocessing step improves the performance of comorbidity extraction.

The experiments also reveal some challenges to the task. First, the ontology depends on report types. Only one of the discharge summary section categories listed in Table 1 applies to radiology reports. Second, even for the same report types, the distribution of the labels and the characteristics of the reports can vary a lot (Table 2). As an example, 5.9% of the sections in Dataset 1 are *Consultation* sections. This value drops to 0.8% in Dataset 2. As a result, system performance degrades significantly when the training and test data come from different datasets.

8. Conclusion

In this paper we proposed a fully statistical system for section segmentation and classification. The system has achieved good performance on three different datasets. We have also shown that automatic section segmentation and classification will lead to improvements in extraction tasks, such as comorbidity identification.

As future work, we plan to expand the current work in several directions. First, we will study whether domain

[3] The annotations created as part of the QI project were only used to automatically evaluate the performance of the proposed comorbidity extraction approach. The authors did not have access to annotations.

adaptation methods might improve the system performance where there is no labeled data for the test domain. Second, we will extend the section category ontology for other report types. Third, in addition to comorbidity extraction, we will use this system for other applications including sense disambiguation of medical concepts and abbreviations, and phenotype extraction from clinical records.

9. Acknowledgement

We would like to thank the reviewers for their valuable feedback. This work was supported by Agency for Healthcare Research and Quality Grant Number 1 R01 HS 20025-01. The Surgical Care and Outcomes Assessment Program (SCOAP) is a Coordinated Quality Improvement Program of the Foundation for Health Care Quality. CERTAIN is a program of the University of Washington, the academic research and development partner of SCOAP. Personnel contributing to this study:

Centers for Comparative and Health Systems Effectiveness (CHASE Alliance), University of Washington, Seattle, WA: David R. Flum, MD, MPH; Rafael Alfonso, MD, PhD (c); Alexander Clowes, MD; E. Beth Devine, PharmD, MBA, PhD; Todd Edwards, PhD, MA; Larry Kessler, ScD; Mark Meissner, MD; Donald Patrick, PhD, MSPH; Sean D. Sullivan, PhD; Peter Tarczy-Hornoch, MD; Erik Van Eaton, MD; N. David Yanez III, PhD; Meliha Yetisgen-Yildiz, PhD, MSc; Allison Devlin Rhodes, MS; Cheryl Armstrong, BSN, MPH; Mitchell Berman; Robin Boland; Daniel Capurro, MD; Augusto Ferreres, MHA; Rosemary Grant, BSN, CCRC, CPHQ; Marisha Hativa, MSHS; Marya Johansen; Sondra Johnson; Wendy Klamp, MPA; Sarah Lawrence, MA; Angela Lloyd, MS; Stephanie Mallahan; Kate Nickel, MPH; Rahma Osman; Catherine Pagoaga; Robert Salazar; Anne Skalicky, MS; Verne Stevens, MBA; Rebecca Gaston Symons, MPH; Michael Tepper, MA; Tomio Tran; Christina Yantsides; Megan Zadworny, MHA.
Providence Everett Regional Medical Center, Everett, WA: Ellen Farrokhi, MD.

10. References

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

Alan R. Aronson. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *Proceedings of the AMIA Annual Symposium*, pages 17–21.

Joshua C. Denny, Randolph A. Miller, Kevin B. Johnson, and Anderson Spickard III. 2008. Development and evaluation of a clinical note section header terminology. In *Proceedings of the AMIA Annual Symposium*, pages 156–160.

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):156–160.

Carol Friedman and Stephen B. Johnson. 2005. Natural language and text processing in biomedicine. In Edward H. Shortliffe and James J. Cimino, editors, *Biomedical Informatics: Computer Applications in Health Care and Medicine*. Springer.

Carol Friedman. 2005. Semantic text parsing for patient records. In Hsinchun Chen, Sherrilynne S. Fuller, Carol Friedman, and William Hersh, editors, *Medical Informatics Knowledge Management and Data Mining in Biomedicine*. Springer.

Kenji Hirohata, Naoaki Okazaki, Sophia Ananiadou, and Mitsuru Ishizuka. 2008. Identifying sections in scientific abstracts using conditional random fields. In *Proceedings of the International Joint Conference on Natural Language Processing*, pages 381–388.

Ying Li, Sharon Lipsky Gorman, and Noemie Elhadad. 2010. Section classification in clinical notes using supervised hidden markov model. In *Proceedings of the 1st ACM International Health Informatics Symposium*, pages 744–750.

Jimmy Lin, Damianos Karakos, Dina Demner-Fushman, and Sanjeev Khudanpur. 2006. Generative content models for structural analysis of medical abstracts. In *Proceedings of the HLT-NAACL BioNLP Workshop*, pages 65–72.

Andrew Kachites McCallum. 2002. MALLET: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Larry McKnight and Padmini Srinivasan. 2003. Categorization of sentence types in medical abstracts. In *Proceedings of the AMIA Annual Symposium*, pages 440–444.

Stephen Merity, Tara Murphy, and James R. Curran. 2009. Accurate argumentative zoning with maximum entropy models. In *Proceedings of the ACL Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, pages 19–26.

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

Simone Teufel. 1999. *Argumentative zoning: Information extraction from scientific text*. Ph.D. thesis, University of Edinburgh, Edinburgh, UK.

Özlem Uzuner, Brett R. South, Shuying Shen, and Scott L. Duvall. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18:552–556.

Jose M. Valderas, Barbara Starfield, Bonnie Sibbald, Chris Salisbury, and Martin Roland. 2009. Defining comorbidity: implications for understanding health and health services. *Annals of Family Medicine*, 7(4):357–363.

Yue Zhang and Stephen Clark. 2010. A fast decoder for joint word segmentation and POS-tagging using a single discriminative model. In *Proceedings of the Conference on Emperical Methods in Natural Language Processing*, pages 843–852.