Hospital Discharge Summarization Data Provenance

Paul Landes[†], Aaron J. Chaise[♢], Kunal P. Patel[♠],
Sean S. Huang[♣], and Barbara Di Eugenio[†]

[†]Department of Computer Science, University of Illinois Chicago
[♠]Department of Emergency Medicine, University of Illinois Chicago
[♢]Department of Family and Community Medicine, University of Illinois Chicago
[♣]Department of Medicine, Vanderbilt University Medical Center, Nashville, TN
{plande2,achais2,kpate318}@uic.edu, sean.huang@vumc.org, bdieugen@uic.edu

Abstract

Summarization of medical notes has been studied for decades with hospital discharge summaries garnering recent interest in the research community. While methods for summarizing these notes have been the focus, there has been little work in understanding the feasibility of this task. We believe this effort is warranted given the notes' length and complexity, and that they are often riddled with poorly formatted structured data and redundancy in copy and pasted text. In this work, we investigate the feasibility of the summarization task by finding the origin, or data provenance, of the discharge summary's source text. As a motivation to understanding the data challenges of the summarization task, we present DSProv, a new dataset of 51 hospital admissions annotated by clinical informatics physicians. The dataset is analyzed for semantics and the extent of copied text from human authored electronic health record (EHR) notes. We also present a novel unsupervised method of matching notes used in discharge summaries, and release our annotation dataset[1] and source code to the community.

1 Introduction

A discharge summary is written by physicians when a patient is discharged from the hospital and provides clinicians with an overview of the patient's hospital stay. Physicians often reference or copy EHR notes to the discharge summary. These copied notes include progress notes, consult notes, and test results. We call these previously written medical documents *note antecedents* since they are written prior to the discharge summary and can be used as source text as input to an automatic summarization (AS) system. The flow of information from the note antecedent to the discharge summary is traced by annotating semantically similar content and copied text across notes and sections (such as *Brief Hospital Course*), which can be copied or paraphrased text. These annotations, called *note matches*, "tie" the notes' discrete lexical spans of text. Each annotation is a text span in the discharge summary that carries a similar or identical meaning to its matching note antecedent as a span, which is a clinical medical note containing the original information that contributed to the discharge summary. In many cases, the text annotated by the note antecedents is later paraphrased or copied into the discharge summary.

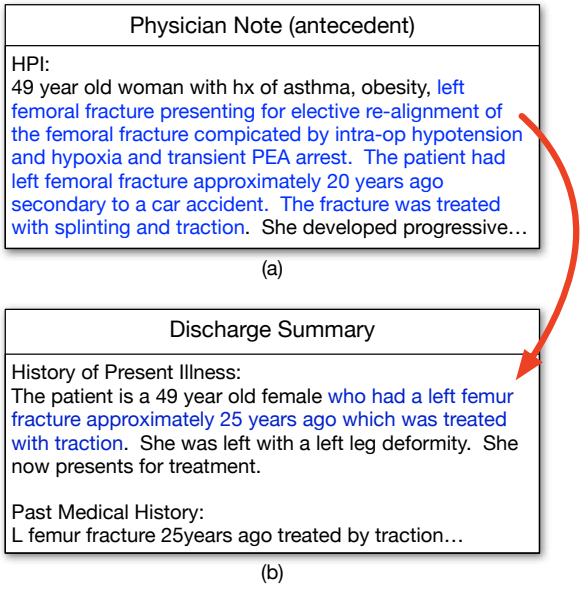

Figure 1: **A note match annotation.** The annotation ties a physician note antecedent to a discharge summary with the matching text spans in blue and their coupling represented with the red arrow.

Figure 1a shows a note match annotation of a physician note and the linked text in the discharge summary given in Figure 1b. In this example, the patient's encounter is documented in the *History of Present Illness* section of the note antecedent and later paraphrased in the summary at the time of the patient's discharge. The red arrow represents the connection between these two lexical spans within an admission. This connection always starts from

[1] https://github.com/uic-nlp-labs/dsprov

The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, pages 439–448

the original text in any clinical note and terminates in the discharge summary for the respective admission. For this work, both ends of the link are text spans written by physicians (see Section 3 for more detail on the annotation process).

Understanding the extent of meaningful data as opposed to low quality copied structured data is key in determining the feasibility and choice of methods needed to generate discharge summaries. In many instances, the copied text comes from high quality sources such as the *History of Present Illness* section of the physician note. However, redundancy, errors and data incoherence is pervasive in the vast amounts of medical data taken from patients during a hospital stay (Cohen et al., 2013). As much as 46% of the discharge summary is copied and pasted, 36% imported from structured data sources, leaving only 18% manually entered (Adams et al., 2021).

The extent of copied text in notes is well known, but the origins of discharge summary text is not. Since it is unclear how concepts arrive into the summary, extrapolating anything more than the measure of copied text and their semantics would necessitate understanding the decisions made by the physician while writing the summary. However, we can infer what is missing from a summarization by subtracting the portion represented by the note antecedents. Knowing what is missing from the summary and unique to the physician's direct individual experience is fundamental for gauging the summarization performance upper bound.

For this reason, our primary goal for this work is to find and analyze the flow of data from notes written prior to the discharge summary, as shown in Figure 1. It is our position that a better understanding of the provenance of data, by note and section, is a crucial first step before AS can be successfully applied to discharge summarization future work. A secondary goal is to produce an unsupervised baseline and model for the research community.

To accomplish these goals three clinical physicians annotated admission records from the freely available Medical Information Mart for Intensive Care III (MIMIC-III) version 1.4 (Johnson et al., 2016) corpus[2]. These annotations uncovered where notes overlap and offer high quality human examples for supervised learning methods. Our contributions are the annotated dataset (see Section 3), its analysis (see Section 3.2), and a novel unsupervised method using the word mover algorithm (Kusner et al., 2015) with clustering to assist in the annotation process (see Section 4).

[2]Access to the MIMIC-III corpus requires creating a PhysioNet account and finishing a training course.

2 Related Work

Most clinical text data analysis has been in the medical field (rather than NLP) with early work in the discovery of data flow using visual summarization methods for clinical psychiatric data (Powsner and Tufte, 1997). Redundant text detection in same-category EHR notes using latent Dirichlet allocation (Blei et al., 2003) has been explored (Cohen et al., 2013). Duplication in medical EHR notes were studied in recent work to find root causes of copied text using 10-gram token spans, which found that 58% of a physician note is copied from previous notes (Steinkamp et al., 2022).

Recent examples of cross discipline collaboration include exploring language and terminology differences between nurses and physicians (Boyd et al., 2018) and patient centric summarization (Acharya et al., 2018). SOAP (Subjective, Objective, Assessment and Plan) note generation (a type of physician note) is another example of an area of mutual interest between the NLP and medical domains (Krishna et al., 2021).

Provenance of data as it relates to the task of summarization has included multi-level recurrent neural networks (RNNs) used for extractive summarization (Zhou et al., 2018). Soon after, BART (Lewis et al., 2020) was used to abstractively condense and smooth the summaries. A large corpus was analyzed and a hybrid extractive/abstractive neural method was used to summarize the *Brief Hospital Course* section (Adams et al., 2021).

The summarization of MIMIC-III physician notes is perhaps the most interesting comparison and potentially most impactful (Gao et al., 2022). Clinical notes were summarized by fine-tuning the T5 (Raffel et al., 2020) and BART (Lewis et al., 2020) state of the art seq2seq models and evaluated using the BERTSCORE (Zhang et al., 2020) and ROUGE (Lin, 2004) scoring methods. In addition, co-occurring extracted Unified Medical Language System (UMLS) medical concepts (Bodenreider, 2004) were also reported. Bodenreider concludes with their intention of providing a foundation for future summarization work and acknowledgment of the challenges facing the clinical note summarization, which echoes the motivation of this work.

3 Dataset

Three attending physicians annotated 51 admissions from the MIMIC-III version 1.4 (Johnson et al., 2016) corpus for overlap with the EHR note antecedents. The overlap was annotated by selecting the semantically similar portions of text from the discharge summary as shown in Figure 1. This provides statistics of overlap both at a note and a section level. To find the section overlap, we use the annotated MedSecId (Landes et al., 2022) corpus for most notes. In the few cases where note section annotations were not available, we used the pretrained MedSecId model to automatically section notes. Such sections are helpful to medical clinicians when finding topical information and useful for billing insurance companies. We direct the reader to Landes et al. for further reading on the function and utility of sections in medical notes.

The annotation selection criteria by admission was the note category and the highest note count from the MedSecId corpus. More specifically, selection $\mathcal{S}_{\mathcal{C}}$ for admission A was chosen as $\mathcal{S}_{\mathcal{C}} = \arg\max_A \sum_{A \in \mathcal{C}} |A|$, where $\mathcal{C}$ is the MIMIC-III corpus and $|A|$ are the number of notes in the admission. Our early findings showed very little overlap of data with the exception of Consult (documentation by consulting physicians across departments) and Physician (typically written daily by the physician) notes. For this reason, admissions with these notes took priority in our annotation process to maximize note match overlap. While it could be argued that these notes' statistics may be combined given their likeness in purpose, it was decided to keep them separate for summarization future work.

Admissions were annotated with note matches using the following process:

1. Extract $\mathcal{S}_{\mathcal{C}}$ admissions from MIMIC-III.
2. For each admission $A \in \mathcal{S}_{\mathcal{C}}$:
   (a) Read A's discharge summary.
   (b) For each note antecedent of the admission A, semantically similar or verbatim copied text was identified and each annotated note match annotated as:
      i. A single span as character offsets tuples in the note antecedent (call it $\boldsymbol{n}$).
      ii. A single span as character offsets tuples in the discharge summary (call it $\boldsymbol{d}$).
      iii. A link between $\boldsymbol{n}$ and $\boldsymbol{d}$.

Each note antecedent may have several note matches (as can a discharge summary), but each has a single link across both note documents.

An annotation guide was created by the lead physician annotator, who then trained the other two physician annotators. Agreement on the criteria for note matches was decided on after each annotator completed one admission. The first admission annotations were then updated per the consensus agreed upon by all annotators. The same process of annotation, discovery, and agreement repeated for three additional admissions; one for each annotator. The remaining 39 were then split among the three annotators.

The 51 admission count might give the mistaken impression of a small dataset. However, as shown in Table 1, the extent of the annotation set is comprehensive with a total of 569 note matches from 291 notes that encompassed a little over 3 million tokens and 11.7 million characters (11.65MB). The number of admissions annotated was an aspect of the time consuming nature of the task. Each admission took an average 20 minutes for the physician to review and annotate as some admissions contained up to 494 notes. However, admissions had an average of 11.16 ($\sigma = 7.3$) notes and one admission had a single note antecedent. Statistics across discharge summaries and note antecedents are separate and independent; for example the 240 note antecedent count in Table 1 does not include counts or other information reflected from discharge summaries.

Description	Count
Admissions	51
Match pairs	569
Discharge summary notes	51
Antecedent notes	240
Total Notes	291
Discharge summary tokens	1,695,466
Antecedent note tokens	1,323,422
Total tokens	3,018,888
Discharge summary characters	7,872,052
Antecedent characters	3,780,025
Total characters	11,652,077

Table 1: **DSProv Corpus Statistics.** The annotation statistics include discharge summary and note antecedent span, token and character counts.

3.1 Limitations

The MIMIC-III corpus includes a discharge summary for each admission. However, it is limited

Note Category	DS Portion	Ant Portion	LS	BERTScore	ROUGEL	BLEU	Count
Physician	**28%**	4%	49.50	68.14	44.10	19.34	310
Radiology	8%	14%	70.10	80.75	64.63	41.03	148
Echo	6%	29%	65.72	**85.11**	66.30	43.65	48
General	5%	4%	50.14	67.71	45.12	19.81	27
Consult	5%	4%	60.24	70.44	54.33	22.36	17
Nursing	2%	7%	23.43	44.56	15.25	0.91	8
ECG	1%	**93%**	**80.94**	79.07	**74.08**	**52.33**	5
Nursing/other	19%	5%	11.38	54.62	7.02	0.00	3
Rehab Services	2%	2%	26.31	60.04	21.26	0.00	2
Case Management	0%	4%	14.29	53.22	18.18	0.00	1

Table 2: **Statistics by MIMIC note category.** "DS Portion" is the ratio of discharge summary tokens to total discharge summary tokens and "Ant Portion" is the ratio of note antecedent tokens to total note antecedent tokens. The "Count" column gives the number of notes annotated.

to patient's time in the intensive care unit (ICU), meaning that the patient's history for any time after transfer from the department is lost. Given most patients progress to lower severity departments as they recover from intensive care, a large cross section of the patient's notes are missing from our analysis. In Section 3.2 we discuss the statistics that justify this conclusion.

3.2 Data Analysis

The DSProv annotation dataset was created with the task of summarization in mind for future work. However, the dataset also provides insight into how EHR note antecedents are used by physicians to write discharge summaries and for the practicality of automatically summarizing them. This analysis provides a quantitative justification for qualitative hypotheses based on clinician's experience writing notes with data observed during annotation.

The human annotated note match text spans were tokenized to compute overlap, ROUGEL (Lin and Hovy, 2003; Lin, 2004), BLEU (Papineni et al., 2002), and BERTSCORE (Zhang et al., 2020). Each note match annotation includes the unique MIMIC-III note antecedent and discharge summary note identifier, absolute character offset in both notes, and the section they span. An additional set of annotations were automatically generated that break spans that overlap sections.

A normalized Levenshtein edit distance (Levenshtein, 1966) was used to measure the extent of copied and pasted text. Since the distance counts the minimum number of edits, rather than a relative measure to the note match span character length, it was normalized with:

$$\text{levsim}(\boldsymbol{w_1}, \boldsymbol{w_2}) = 1 - \frac{\text{lev}(\boldsymbol{w_1}, \boldsymbol{w_2})}{\max |\boldsymbol{w_1}|, |\boldsymbol{w_2}|} \quad (1)$$

where lev is the Levenshtein edit distance, and $|w_1|$ and $|w_2|$ are lengths of the words in characters.

3.2.1 Note Category

Table 2 provides statistics and similarity measures on the DSProv annotation set. The portion columns show the token overlap between each category of note antecedent ("Ant Portion") with the discharge summary ("DS Portion"). The "LS" is the Levenshtein edit similarity as computed with Equation 1. All similarity scores ("LS", "BERTSCORE", "ROUGE" and "BLEU") are computed between the note antecedent and discharge summary span for each match.

The highest discharge summary token overlap is with physician notes (28%), which we consider surprisingly high considering the MIMIC-III corpus only includes ICU notes as mentioned in Section 3.1. We expect the other discharge summary overlap statistics to be underrepresented for the same reason. This high token overlap supports the conjecture that daily progress notes are highly summarized with little copied text. The relatively low edit distance with a high BERTSCORE further supports this conclusion as surface similarity of copied text is low but the semantic similarity is high.

ECG (electrocardiogram notes), Radiology and Echo (echocardiogram data and analysis) notes show a higher similarity (80.94, 70.1, and 65.72 respectively) with the discharge summary since they are frequently copied and pasted. However, these statistics should be higher than presented since some spans include page breaks that result in counting superfluous header and footer tokens. Also note that ECG has the highest note antecedent portion, implying that most of these short reports make it into the discharge summary.

Section Type	DS Portion	Ant Portion	LS	BERTScore	ROUGEL	BLEU	Count
Hospital course	**14%**	4%	41.13	63.39	33.35	11.54	125
Labs imaging	14%	14%	71.55	82.59	67.60	44.08	194
History of present illness	10%	7%	58.51	70.94	51.97	30.22	57
Physical examination	5%	3%	66.23	67.50	56.27	36.50	16
Addendum	5%	10%	54.89	74.53	43.05	19.00	2
Past medical history	4%	4%	44.60	68.54	40.22	20.32	64
Medication history	4%	4%	90.64	84.25	80.47	**55.35**	11
Imaging	2%	10%	**97.92**	**86.99**	**88.06**	54.16	1
Review of systems	2%	14%	24.84	58.46	21.08	2.41	1
Social history	1%	1%	59.55	71.79	49.95	22.66	15
Major surgical or invasive proc.	1%	0%	23.95	55.65	24.13	0.00	7
Family history	1%	1%	56.25	72.34	61.59	35.45	9
Default	0%	**24%**	25.00	28.58	14.29	0.00	1
Chief complaint	0%	0%	55.65	74.68	51.56	9.25	26
Discharge diagnosis	0%	1%	31.87	57.74	10.26	0.00	1

Table 3: **Statistics grouped by discharge summary section type.** "DS Portion" is the ratio of discharge summary tokens to total discharge summary tokens and "Ant Portion" is the ratio of note antecedent tokens to total note antecedent tokens. The "Count" column gives the number of notes annotated. Only the top 15 sections with the highest discharge summary overlap are reported.

3.2.2 Section by Discharge Summary

Statistics by discharge summary section are given in Table 3. Labs and Imaging (71.55%) is a highly copied and pasted section from Radiology and Echo notes. This section takes into account cultures, blood results, and lab tests, and is copied and pasted as structured data.

Hospital Course has the highest discharge summary (14%) section representation. While more analysis is needed, we believe this section has a high overlap due to transfer notes that describe patients moving between departments. These notes have an impact on summarization as they describe what happened to the patient during the hospital visit, including time of death. Generally speaking, good sections for summarization are those that have a low edit score (minimal copying and pasting) but a high similarity. In this case the comparatively low Levenshtein edit similarity (41.13) and somewhat high BERTSCORE (63.39) implies this section is a good target for AS as previously investigated in prior work (Adams et al., 2021).

3.2.3 Section by Note Antecedent

Table 4 shows the overlap from the perspective of the note antecedent. The *Assessment and Plan* section (the overall impression of the patient and how to treat them) has a high (15%) antecedent overlap. We found that the majority of the discharge summary content comes from the *Brief Hospital Course* section. The Echo note's *Conclusions* section has a high portion of text that is summarized from note antecedents. This section's echocardiogram information only consists of 8% on average of the notes annotated.

4 Methods

Automatic methods to assist in bootstrapping the corpus were considered given the large amount of data the physician needs to sift through to find note matches[3]. However, the task is challenging given the pages long document lengths, and precludes transformer pre-training methods. While the task has much in common with paraphrase matching and information retrieval tasks, it is fundamentally different in the way sections of text are matched. For example in question/answer systems, a query matches to an answer in the source text. However, our task requires the correct text span in both note documents. Both the discharge summary and the note antecedent predicted text spans for the respective linked note matches are used when evaluating.

4.1 Evaluation

The human annotations were used for comparison since there is no previous baseline for this task. The unsupervised methods were then used to estimate the overlap portion to trace the origins of the discharge summary to the note antecedent, and then compared against the human annotations. The task is framed as a token classification task (without a label) since spans are token boundary demarcated.

[3]The highest note count for an admission in MIMIC-III corpus is 1,233 notes.

Section Type	DS Portion	Ant Portion	LS	BERTScore	ROUGEL	BLEU	Count
Indication	3%	**28%**	84.88	83.83	**81.00**	53.58	1
Conclusions	8%	26%	79.49	87.21	78.73	**57.31**	33
Findings	9%	15%	81.53	83.11	74.51	48.53	78
Technique	10%	9%	18.38	84.84	29.91	1.15	1
Impression	5%	8%	65.47	81.30	61.70	36.75	54
Review of systems	5%	8%	20.74	57.12	16.31	1.21	2
Wet read	4%	6%	**98.28**	**91.18**	76.19	52.69	2
Addendum	3%	6%	25.77	56.82	14.41	0.00	3
History	7%	5%	12.63	82.96	20.55	0.09	2
Hospital course	6%	5%	40.37	61.23	31.60	16.25	12
History of present illness	11%	5%	59.68	71.75	51.03	29.06	58
Comparison	5%	5%	13.79	80.72	21.58	0.55	8
Labs imaging	10%	4%	85.37	79.32	73.21	44.22	10
Assessment and plan	**15%**	4%	44.90	65.01	38.94	14.06	86
Discharge instructions	0%	4%	14.29	53.22	18.18	0.00	1

Table 4: **Statistics grouped by note antecedent section type.** "DS Portion" is the ratio of discharge summary tokens to total discharge summary tokens and "Ant Portion" is the ratio of note antecedent tokens to total note antecedent tokens. The "Count" column gives the number of notes annotated. Only the top 15 sections with the highest discharge summary overlap are reported.

Our methods were evaluated using the SemEvel 2013 Task 9.1 (Segura-Bedmar et al., 2013) entity extraction scoring method since it is flexible in its strictness as a score. Given the novelty and difficulty of the task, we used the *partial boundary matching* metric with ROUGE as an additional reference point. However, even though the SemEval measure is flexible for partial token matching, it is not ideal since this task aims to classify single token spans on a match-by-match basis.

4.2 Word Mover

The word mover algorithm (Kusner et al., 2015) was used in the first step of our method. Our method frames the task as a transportation problem by using the earth mover algorithm (Pele and Werman, 2009). The intuition follows from modeling probability distributions as piles of dirt that are moved from one location to another. The algorithm treats high dimensional word embeddings of a source document as the probability distribution that "moves" words to the target document embedded distribution. The optimization algorithm minimizes the objective on words $\boldsymbol{w_t}, \ldots, \boldsymbol{w_T}$:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{j\in\mathcal{N}(t)}\log p(\boldsymbol{w_j}|\boldsymbol{w_t}) \qquad (2)$$

where $\mathcal{N}(\boldsymbol{n})$ are the neighboring words, and $p(w_j|w_t)$ is the word vector's hierarchical softmax values. Word embeddings are made unit vectors and distance measures are euclidean. Our experiments include non-contextual word embeddings (Mikolov et al., 2013a,b) and static contextual embeddings from transformer architectures (Devlin et al., 2019). Because document word frequencies are used as the histogram weight to the earth mover algorithm, surface word forms were associated with wordpiece groups (Wu et al., 2016). The centroid of each wordpiece group was used for the embeddings.

4.3 Hybrid Semantic Positional Token Clustering

The word mover algorithm maps words from the discharge summary onto the note antecedent efficiently, but it does not help us link the note matches. A naïve approach would be to chunk tokens based on a metric such as cosine similarity. However, words with little similarity would frequently result in too many span breaks. We still need to cluster the word embeddings computed in Section 4.3 to group concepts in each document separately. However, this still does not address the "Swiss cheese" problem of note matches with too many "holes" (span breaks). We propose to simply add a component with a normalized scaled value to the word embedding. More specifically, we defined the concatenated position vector with:

$$\text{posemb}(w) \triangleq \left[\text{emb}(w); \left(\frac{\alpha_{\mathbf{t}}\cdot i}{|T|}\right)\right] \qquad (3)$$

where $\text{emb}_i(w)$s the word embedding as an application of the language model; $\alpha_{\mathbf{t}}$ is the token

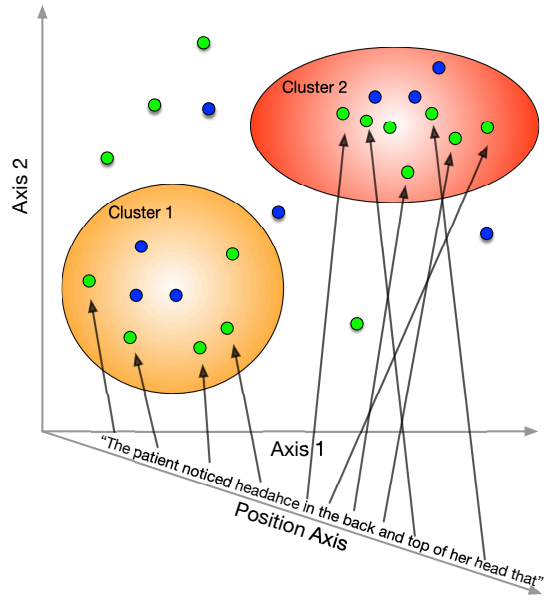

Figure 2: **Hybrid Semantic Positional Token Clustering .** Position embeddings on a third axis shows data blue word embeddings moving from cluster 1 to cluster 2. Cluster spans the discharge summaries (orange), the note antecedent (green) and arrows connecting the tokens to word points.

position component scaler; i is the index the i^{th} word w_i, and $|T|$ is the document token length.

The higher the token position component scaler ($\alpha_{\mathbf{t}}$) value is set, the more the word position is prioritized. This means that high values of the hyperparameter will create longer contiguous token spans, but at the cost of semantic similarity. This effect can be visually explained as a simple 2D word embedding with an additional token position axis. Figure 2 shows such a coordinate system with an example of an embedded span. On the positional axis, each token is spaced at even intervals. Because the positional components are proportionally scaled up for higher values of $\alpha_{\mathbf{t}}$, their relative distances shrink. On the other hand, if this value is lowered, their positional components diminish, effectively reverting the word points (word location in the embedded space) to their pre-trained vectors.

Once clustering of each document's word points is complete, each cluster's points are assigned to note matches. For each iteration over the Cartesian product, each document's points are added to matches by associated cluster (see Algorithm 1). The source and target documents are swapped and Algorithm 1 is run again to create flows from the target to the source. Lexical overlapping matches are combined and their flows added together and sorted to create a ranking of matches.

Since each note match is assigned a flow (as a function of work to transport the word embedding) in both directions, they are combined as a single flow, which represents the highest similarities having the most information between the notes. Finally, matches are sorted in descending order by their flow values, so those with the maximum amount of information gain are ranked first.

Algorithm 1: Matching Algorithm

Input: Documents source A and target B
Output: N note match spans

Function `MatchNoteSpans`(A, B)
  `// assign word vectors and normalize to unit`
  $\mathcal{V}_a \leftarrow \frac{\text{emb}(w)}{||\text{emb}(w)||_2}; \forall w \in A$;
  $\mathcal{V}_b \leftarrow \frac{\text{emb}(w)}{||\text{emb}(w)||_2}; \forall w \in B$;
  `// assign position embeddings`
  $\mathcal{P}_a \leftarrow \text{posemb}(\mathcal{V}_a)$;
  $\mathcal{P}_b \leftarrow \text{posemb}(\mathcal{V}_b)$;
  `// assign word flows`
  $(\mathcal{F}_a, \mathcal{F}_b) \leftarrow$ `WordMover`$(\mathcal{V}_a, \mathcal{V}_b)$;
  `// cluster word points`
  $\mathcal{C}_a \leftarrow$ `Cluster`$(\mathcal{P}_a)$;
  $\mathcal{C}_b \leftarrow$ `Cluster`$(\mathcal{P}_b)$;
  `// add matches`
  $\mathcal{M} \leftarrow \{\emptyset\}$;
  for $f_a \in \mathcal{F}_a$ **do**
    for $f_b \in \mathcal{F}_b$ **do**
      `// get cluster from flow`
      $c_a \leftarrow \mathcal{C}_a[f_a]$;
      $c_b \leftarrow \mathcal{C}_b[f_b]$;
      `// add the match and`
      `// token points`
      if $\{(c_a, c_b)\} not \in \mathcal{M}$ **then**
        $\mathcal{M} \leftarrow \mathcal{M} \cup \{(c_a, c_b)\}$;
      end
    end
  end
  return $\mathcal{M}$;
end
Function `BiMatchNoteSpans`(A, B)
  $\mathcal{M}_{a \to b} \leftarrow$ `MatchNoteSpans`(A, B) ;
  $\mathcal{M}_{b \to a} \leftarrow$ `MatchNoteSpans`(B, A) ;
  $\mathcal{M}_{bi} \leftarrow$ `Sort`$(\mathcal{M}_{a \to b}, \mathcal{M}_{b \to a})$;
end

5 Results and Discussion

The DSProv dataset was provided to the unsupervised algorithm described in Section 4 and evaluated against the human annotated note matches. The match spans with the highest flow (see Section 4.3) were compared and scored using the measures listed in Section 4.1. Before the evaluation, Bayesian hyperparameter optimization (Bergstra et al., 2013) was used on the human annotated dataset on a subset of the data. The model's hyperparameters were set to the Bayesian optimized values and evaluated. Of the 569 note matches annotated, an additional 359 were optimized on 500 iterations. This process was repeated on each word embedding for each note match.

While the hybrid method explained in Section 4.3 had a relatively high SemEval partial recall of 69.06 for matching discharge summaries spans, it suffered a low precision score. This implies finding spans is not an issue, but finding correct span boundaries as more difficult. We report both the good recall but poor precision to help explain the kinds of challenges in matching spans between discharge summaries with note antecedents.

The performance for the note antecedents matching as shown Table 5 tells a better story. We see a similar pattern with a low precision, but a high recall with both netting a higher SemEval partial match harmonic mean of 15.85. Surprisingly the SapBERT (Liu et al., 2021), which models semantic relationships of biomedical domain entities, performed worse than Sentence-BERT (Reimers and Gurevych, 2019). This suggests models trained for embedding and clustering provide better embeddings for our task. The non-contextual word embeddings do not perform as well with the exception of GloVE (Pennington et al., 2014) having the best ROUGE1 for discharge summaries.

As mentioned in Section 3.1, only ICU notes and discharge summaries are provided in MIMIC-III. This has the effect of decreasing available information to potentially summarize and also has a negative impact on results as there are fewer examples to match between note documents. However, it has an even greater impact on summarization since the machine cannot generate what is not available in the EHR. While much of this data is available as structured data that can be textually formatted, additional data provenance based methods are needed to use these information sources. For example, it is feasible to summarize sections from structured data such as tabular prescribed medicine to generate the *Medication History* section, which we leave as future work.

6 Conclusions and Future Work

We have presented DSProv, a new freely available dataset of 569 textual span matches between discharge summaries and note antecedents annotated by clinical informatics physicians. The analysis of our dataset presents new qualitative and quantitative findings of EHR notes. We have also presented a novel unsupervised method for annotating note matches using models tuned on human examples from our dataset. We believe our baseline results and our dataset analysis provide insights necessary for assessing the feasibility of traceable and faithful automatic discharge summarization in future work.

While the purpose of this work was to investigate the provenance of discharge summaries from EHR note antecedents, our secondary goal was to augment our dataset with semi-supervised annotations. Given the unsupervised baseline models have room for improvement, we leave supervised methods (such as pretrained biomedical language models fine-tuned span tagging) as future work.

7 Acknowledgments

This work was supported by award R01 CA225446 from the National Institutes of Health.

Model	SE P	SE R	SE F1	SE CO	ROUGE1	ROUGE2	ROUGEL	EM
BioBERT	7.82	2.05	2.62	168.75	8.43	3.38	7.60	0.000
ClinicalBERT	8.61	2.89	3.91	330.60	10.76	5.07	9.39	0.216
Glove 300D	7.80	5.16	4.88	460.78	12.59	5.99	10.62	0.000
word2vec	9.76	34.15	12.30	3277.37	17.04	13.14	15.43	0.216
SapBERT	10.74	35.90	13.44	3316.16	19.36	14.69	16.95	0.216
SBERT	**11.79**	**50.04**	**15.85**	4635.78	**19.62**	**16.48**	**18.03**	0.216

Table 5: **Note antecedent scores** Performance of the unsupervised method for each word embedding model. Score methods include (SE)MEVAL-2013 (P)recision, (R)ecall, F1 and (Co)rrect mean. ROUGE1, ROUGE2, and ROUGEL F1 scores also provided with an (E)xact (M)atch score.

References

Sabita Acharya, Barbara Di Eugenio, Andrew Boyd, Richard Cameron, Karen Dunn Lopez, Pamela Martyn-Nemeth, Carolyn Dickens, and Amer Ardati. 2018. Towards Generating Personalized Hospitalization Summaries. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 74–82. Association for Computational Linguistics.

Griffin Adams, Emily Alsentzer, Mert Ketenci, Jason Zucker, and Noémie Elhadad. 2021. What's in a Summary? Laying the Groundwork for Advances in Hospital-Course Summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4794–4811.

James Bergstra, Daniel Yamins, and David Cox. 2013. Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. In *Proceedings of the 30th International Conference on Machine Learning*, pages 115–123. PMLR.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*, 32:D267–D270.

Andrew D. Boyd, Karen Dunn Lopez, Camillo Lugaresi, Tamara Macieira, Vanessa Sousa, Sabita Acharya, Abhinaya Balasubramanian, Khawllah Roussi, Gail M. Keenan, Yves A. Lussier, Jianrong 'John' Li, Michel Burton, and Barbara Di Eugenio. 2018. Physician nurse care: A new use of UMLS to measure professional contribution: Are we talking about the same patient a new graph matching algorithm? *International Journal of Medical Informatics*, 113:63–71.

Raphael Cohen, Michael Elhadad, and Noémie Elhadad. 2013. Redundancy in electronic health record corpora: Analysis, impact on text mining performance and mitigation strategies. *BMC bioinformatics*, 14:10.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Yanjun Gao, Dmitriy Dligach, Timothy Miller, Dongfang Xu, Matthew M. M. Churpek, and Majid Afshar. 2022. Summarizing Patients' Problems from Hospital Progress Notes Using Pre-trained Sequence-to-Sequence Models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2979–2991. International Committee on Computational Linguistics.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):1–9.

Kundan Krishna, Sopan Khosla, Jeffrey Bigham, and Zachary C. Lipton. 2021. Generating SOAP Notes from Doctor-Patient Conversations Using Modular Summarization Techniques. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4958–4972. Association for Computational Linguistics.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From Word Embeddings to Document Distances. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 957–966. JMLR.org.

Paul Landes, Kunal Patel, Sean S. Huang, Adam Webb, Barbara Di Eugenio, and Cornelia Caragea. 2022. A New Public Corpus for Clinical Section Identification: MedSecId. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3709–3721. International Committee on Computational Linguistics.

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet Physics Doklady*, volume 10, pages 707–710.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81. Association for Computational Linguistics.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using N-gram co-occurrence statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 71–78. Association for Computational Linguistics.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. Self-Alignment

Pretraining for Biomedical Entity Representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. arXiv: 1301.3781 (Only available as arXiv preprint).

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

O. Pele and M. Werman. 2009. Fast and robust Earth Mover's Distances. In *2009 IEEE 12th International Conference on Computer Vision*, pages 460–467.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Association for Computational Linguistics.

S. M. Powsner and E. R. Tufte. 1997. Summarizing clinical psychiatric data. *Psychiatric Services (Washington, D.C.)*, 48(11):1458–1461.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):140:5485–140:5551.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992. Association for Computational Linguistics.

Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo. 2013. SemEval-2013 Task 9 : Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 341–350. Association for Computational Linguistics.

Jackson Steinkamp, Jacob J. Kantrowitz, and Subha Airan-Javia. 2022. Prevalence and Sources of Duplicate Information in the Electronic Medical Record. *JAMA Network Open*, 5(9):e2233348.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. arXiv: 1609.08144 (Only available as arXiv preprint).

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *Proceedings of the 8th International Conference on Learning Representations*.

Qingyu Zhou, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao. 2018. Neural Document Summarization by Jointly Learning to Score and Select Sentences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–663. Association for Computational Linguistics.