Section Classification in Clinical Notes with Multi-task Transformers

Fan Zhang and **Itay Laish** and **Ayelet Benjamini** and **Amir Feder**
{zhanfan,itaylaish,ayelet,afeder}@google.com
Google Research

Abstract

Clinical notes are the backbone of electronic health records, often containing vital information not observed in other structured data. Unfortunately, the unstructured nature of clinical notes can lead to critical patient-related information being lost. Algorithms that organize clinical notes into distinct sections are often proposed in order to allow medical professionals to better access information in a given note. These algorithms, however, often assume a given partition over the note, and classify section types given this information. In this paper, we propose a multi-task solution for note sectioning, where a single model identifies context changes and labels each section with its medically-relevant title. Results on in-distribution (MIMIC-III) and out-of-distribution (private held-out) datasets reveal that our approach successfully identifies note sections across different hospital systems.

1 Introduction

The increasing role of free-text narrative in Electronic Health Records (EHR) is both a blessing and a curse. It allows much more nuanced information about patients' conditions being saved and documented (Uzuner et al., 2010; Jensen et al., 2012; Wang et al., 2018; Feder et al., 2020). However, the unstructured nature of this data can also make it unavailable to medical care givers interested in searching for specific patient-related information (Walsh, 2004; Ford et al., 2016).

To better organize free-form clinical notes and allow researchers and practitioners to quickly search over them, many solutions were proposed, mainly focusing on sectioning notes to correspond to headers described within the note (Pomares-Quimbaya et al., 2019). These solutions were often rule-based (Savova et al., 2010), identifying common section headers in the data. Unfortunately, this approach often failed to correctly classify sections across different hospital departments, care providers and EHR systems. For brevity throughout this paper, we refer these as different *data sources* or *distributions* interleaving. Alternatively, machine learning methods were proposed to classify individual text-spans and map them into a pre-existing list of possible sections. This approach successfully outperformed rule-based approaches, but was often not deployed because of its inability to identify section-boundaries.

With the recent success of transformer-based models in natural language understanding, we identify an opportunity to tackle the section boundary detection problem alongside section classification, and propose a unified solution. Our approach is based on pre-trained encoder-only transformer models, which were shown to produce superior results on natural language understanding (NLU) tasks broadly (Vaswani et al., 2017; Devlin et al., 2018), and specifically on clinically-relevant data (Alsentzer et al., 2019; Lee et al., 2020).

We start by exploring current section classification methods (§2). Then, we introduce our baseline, a marker-based section header extraction system, and describe how to use it to generate training labels for ML-based methods (§3). We then pose hypotheses for when should ML systems outperform rule-based approaches, and propose solutions based on the hypotheses (§4). We continue by proposing a dataset for training multi-task transformers from rule-based labels (§5) and demonstrate how such models can outperform rule-based approach on in-distribution and out-of-distribution data (§6). Finally, we conclude our work in light of our posed hypotheses (§7).

2 Related Work

Identifying section headers in free-form clinical notes is long identified as a crucial task for organizing patient-level data in biomedical informatics (Li et al., 2010). Both ML-based and rule-based solutions were proposed in the last decade to solve

Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI), pages 54 - 59

the problem (Pomares-Quimbaya et al., 2019). Unfortunately, existing solutions focus on solving the relatively narrowly-defined task of classifying predefined sections into section types, assuming that section borders are already given (Li et al., 2010; Tepper et al., 2012; Dai et al., 2015; Pomares-Quimbaya et al., 2019). In practice, however, we often observe complete notes, and are tasked with identifying distinct paragraphs and only then classifying them into individual sections.

Recently, there has been an influx of research demonstrating the power of pre-trained language models in solving multi-task problems (Peng et al., 2020; Radford et al., 2019; Wolf et al., 2020), including on long texts (Beltagy et al., 2020). Following this newly-formed conventional wisdom, we embrace this approach here, and propose an ML architecture that attempts to jointly detect section boundaries and classify individual sections.

3 Marker-based Section Header Extraction

We start by developing a *marker*-based section header extractor. This extractor will then be used for labeling our training data in §5 and as a baseline in §6. In this approach, a *marker* corresponds to a word that is usually used as the header of section. E.g. **PMH** is a typical marker word that represents the section Past Medical History. After examining patterns in the data, we discover hundreds of such markers in the MIMIC-III dataset (Johnson et al., 2016). Lines that start with these markers are extracted and are labeled as section headers. These headers mark the boundary between two sections and the text between two headers is then treated as one single section.

During our exploration, we recognized that there exists correlations between the type of the notes and the structure of the sections in the note. With that in regard, we customized our markers to the type of notes and certain markers will only be applied when the type of the note matches our definition. We identified 5 core note types that are most important for our usage: *History and Physical*, *Progress*, *Discharge summary*, *Consult* and *Operative*.

Building on the MIMIC-III dataset, we use an iterative approach to collect markers. A bootstrapping marker set is first developed on a sampled set of notes from the MIMIC-III dataset. The marker set is then used to extract sections on the sampled set and the extracted sections are then sent to experienced clinicians for rating. New markers are then added according to the errors collected from the raters and then used on a new set of randomly collected notes. This process is repeated until no more errors are reported from the raters. In practice, we found that this method shows both high precision and high coverage in recognizing the sections. However, this approach does not work well when we try to transfer it to a new dataset where the medical notes come from a different healthcare provider, where we see the recall numbers dropping significantly (see §6 for complete results).

By analyzing the errors, we are seeing the following patterns:

- Plurals. E.g. “complaint” and “complaints”
- Abbreviations. E.g. “ALL” for “allergy”, “Hx” for “history”.
- Mutation of marker orders. E.g. “PMH/PSH” and “PSH/PMH”.
- Additional punctuations. E.g “** Marker **”
- Character splits, e.g. “P H Y S I C A L E X A M I N A T I O N”.

By comparing with MIMIC-III, we observe that while the headers are semantically similar across different healthcare providers, many cases are actually non-identical and can therefore cause recall losses. Additionally, this approach does not take the context information into consideration, and is not able to recognize many cases above even if the section contents look similar to each other.

4 Section Classification Methods

To build solutions that are robust across different distributions or require minimum learning efforts to adapt, we need to understand what is the transferable knowledge that applies. Based on our experiences in building the marker-based approach, we have the following hypotheses:

- **Section titles are shared across different sources.** This means that we expect the same terminology is shared across different sources. For example, we would expect “assessment and plan” is a common terminology shared across different sources. There might be some slight variations, for example, “chief complaint” vs. “chief complaints”.

- **Section contents are similar across different sources.** We are expecting that the same

section type would have similar content even if they are from different healthcare systems.

- **Structure of the sections is different for different types of notes.** For example, we would expect the discharge summary notes to have a different set of sections in comparison to operative notes.

For the first hypothesis, we want to understand if we can build source-agnostic solutions by just expanding the markers used in the baseline. For the second hypothesis, we want to check if we can improve the accuracy of section type identification with additional information from the surrounding text of the section titles. For the last hypothesis, we propose to take advantage of the note type information within a multi-task framework.

4.1 Expanding section titles

We first explore the approach using the same mechanism as the baseline approach, where we identify section titles as section boundaries and categorize sections according to the marker types. Instead of the exact match used in the baseline approach, we modify the method to fuzzy-match with embedding-based similarity calculation. Here, we use embeddings from the Universal Sentence Encoder (Cer et al., 2018) to generate a sentence embedding for each section marker. Using the sentence embeddings, we calculate the cosine similarity and use it to filter out section markers. Using the dev set to select the best threshold in terms of both precision and recall, we find that 0.98 cosine similarity is the best for filtering potential markers.

4.2 Using context information

We conducted three types of experiments regarding the use of context information: (1) Section title only. For this, we only use the text of the target sentence itself as the input feature for our model and generate the input feature as <CLS><Target>. (2) Context information only. We exclude the section titles from the input feature of our model to see if we can achieve good enough performance with only context information. We generate the feature as <CLS><Text before><SEP><Text after>. (3) Title + Context. For this we use the entire segment of text including title + context for prediction and generate the feature as <CLS><Text before><SEP><Target><Text after>.

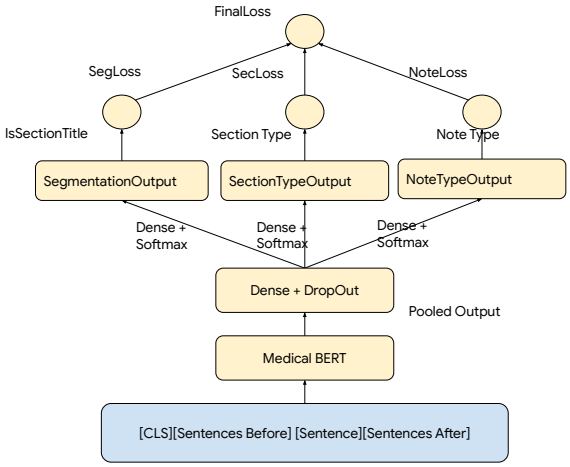

Figure 1: Structure of our multi-task BERT-based transformer model.

4.3 Multi-task BERT model

We propose the multi-task bert model[1] as shown in Fig 1[2]: We split the text into text spans according to line breaks and treat each text span as a training example. For each example, we create three separate losses for different tasks and use a combined loss as the final loss function.

- Segmentation Loss: This task does a binary classification regarding whether the target sentence is a section title or not.
- Section Type Loss: This task does a multi-class classification regarding the section type of the target sentence. We end up with a 19-way softmax by identifying 18 most important section type sand treat the rest as others. The details of thse 18 section types can be seen in Appendix A.
- Note Type Loss: This task predicts which type of the note the target sentence comes from. We end up with a 7-way softmax, including 5 core types as mentioned in Section 3 + 1 unspecified type for notes with no obvious structures + 1 others.

The combined loss is calculated as a weighted sum of all losses. We tested on our dev set and set an equal weight for each loss in our experiment. To verify whether the use of note type information is actually helpful, we added the experiment where we set the weight for note type loss to 0.

[1]For BERT, we are using medical-bert fine-tuned on pubmed data.

[2]Dense layers set as (128 - 32 - Final prediction towers) with 0.1 dropout

Method	Description	P	R
Embedding-based	Title only	0.82	1
BERT (target only)	Title only	**0.94**	**0.99**
BERT (context only)	Context Only	0.88	0.94
BERT (target + context)	Title + Context	**0.94**	**0.99**
BERT (no note loss)	Title + Context	**0.92**	**0.99**

Table 1: MIMIC-III (in-distribution) segmentation results. We only report segmentation results here as we found that the section type accuracy is usually high when we can recognize the correct section title.

Method	Description	P	R
MIMIC3 Markers	Baseline	0.98	0.65
Embedding-based	Title Only	0.66	0.84
BERT (target only)	Title Only	**0.72**	0.88
BERT (context only)	Context Only	0.66	0.80
BERT (target + context)	Title + Context	0.70	**0.95**

Table 2: out-of-distribution validation results

5 Data

To have enough data for training/evaluation, the output of the baseline system (Section 3) is used as the golden data. Due to the nature of the baseline algorithm, we can expect the generated data to have high precision/recall for training models on MIMIC-III and also high precision but low recall for validation on the held-out private dataset.

Test data For MIMIC-III data, we use the data described above. For the held-out private data, we use the same approach as described above and use all the extracted data as the test data. We randomly selected 500 notes for validation.

Data pre-processing The baseline approach uses the following rules for identifying the potential section titles which satisfy the following two constraints: (1) Sentence at the start of the text. (2) Sentence that ends with title endings (":", "-", "("). We followed similar ideas and relaxed this constraint in our data processing, where we split the text into spans of text when there is line break or a title ending. The section type information are then assigned the text spans according to the output of the baseline approach.

Training data We use MIMIC-III dataset as the training data. We randomly selected 4,000 notes for each of the five core note types, 4,000 notes where the note type is not specified and 8,000 notes randomly sampled from the entire dataset. As we don't have enough data for some categories, we end up having 20,000 notes with 3M text spans (among which there are 200k section titles). We split these examples to training/validation/testing in the ratio of 8:1:1.

6 Experiments and Results

We first conducted experiments on MIMIC-III dataset and Table 1 demonstrates the results. As our approaches are based on MIMIC-III markers and we are evaluating on the results extracted from the markers, we expected to see good recall performance for all our approaches. We are seeing that the embedding-based approach and BERT models that use title information were able to get a recall of more than 0.99. To our surprise, we also see that we are able to get a recall of 0.94 with just context information, proving that context information is useful even if used alone. However, we did not see better results with both title and context information, probably because that there exists limited headroom for improvement. In the meanwhile, we do see a small boost in precision with the inclusion of note type classification loss.

We applied the models trained on MIMIC-III and then to a new held-out dataset and results are shown in Table 2. The MIMIC3 markers-based approach was used as a baseline for comparison. We can see that while the markers-based approach still has a high precision due to its exact-match nature while its recall dropped to 0.65. With fuzzy title matching, the embedding-based approach improved the recall to 0.84 at the cost of dropping the precision to 0.66. Again, we see a reasonable performance with BERT + only context information. The BERT model with only title information reached a precision of 0.72 and recall of 0.88. With the addition of context information, the model's recall improves to 0.95 without much loss in precision.

7 Conclusion

In this work, we explored approaches for recognizing sections in free-form clinical notes. Our approach is based on the hypothesis that section content is similar across distributions and can be used to generate a robust section classifier. Our results demonstrate that our BERT-based model trained on MIMIC-III has very good performance on MIMIC-III and on our held-out private data, outperforming strong baselines.

References

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder for english. In *Proceedings of the 2018 conference on empirical methods in natural language processing: system demonstrations*, pages 169–174.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Amir Feder, Danny Vainstein, Roni Rosenfeld, Tzvika Hartman, Avinatan Hassidim, and Yossi Matias. 2020. Active deep learning to detect demographic traits in free-form clinical notes. *Journal of Biomedical Informatics*, 107:103436.

Elizabeth Ford, John A Carroll, Helen E Smith, Donia Scott, and Jackie A Cassell. 2016. Extracting information from the text of electronic medical records to improve case detection: a systematic review. *Journal of the American Medical Informatics Association*, 23(5):1007–1015.

Peter B Jensen, Lars J Jensen, and Søren Brunak. 2012. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6):395–405.

Alistair EW 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.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

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

Yifan Peng, Qingyu Chen, and Zhiyong Lu. 2020. An empirical study of multi-task learning on bert for biomedical text mining. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 205–214.

Alexandra Pomares-Quimbaya, Markus Kreuzthaler, and Stefan Schulz. 2019. Current approaches to identify sections within clinical narratives from electronic health records: a systematic review. *BMC medical research methodology*, 19(1):1–20.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

Michael Tepper, Daniel Capurro, Fei Xia, Lucy Vanderwende, and Meliha Yetisgen-Yildiz. 2012. Statistical section segmentation in free-text clinical records. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2001–2008.

Özlem Uzuner, Imre Solti, and Eithon Cadag. 2010. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, 17(5):514–518.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Stephen H Walsh. 2004. The clinician's perspective on electronic health records and how they can affect patient care. *Bmj*, 328(7449):1184–1187.

Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, et al. 2018. Clinical information extraction applications: a literature review. *Journal of biomedical informatics*, 77:34–49.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

A Appendix: Section Types

Table 3 shows a list of section types covered in this paper.

Section Type	Example Markers
CHIEF COMPLAINT	Chief Complaint, CC, Presenting Problem
PAST MEDICAL HISTORY	Pmh, Past Medical Problem
REVIEW OF SYSTEMS	ROS, Review of Systems
SOCIAL HISTORY	Family/Social History, Social Hx, SH
OTHER SUBJECTIVE	Subjective, health maintenance, Influenza vaccine screening
IMAGING	Image Result, IMAGING STUDIES
MEDICATION	Allergies/Medication List, med list, Infusions
PHYSICAL EXAMINATION	Physical Exam, Phys exam, PEx, Height And Weight
LAB RESULTS	Review of Laboratory Data, Labs and Reports, Blood Chemistry Studies
OTHER OBJECTIVE	Stress test, pathology
ASSESSMENT AND PLAN	A&P, Impression and Plan, Plan
PROBLEM LIST	Problem list, Problems (Active), Diagnoses
HOSPITAL COURSE	Brief history of hospital course, Hospital Summary
DISCHARGE TRANSFER DIAGNOSIS	Discharge/Transfer Diagnoses, Primary Diagnosis
DISCHARGE TRANSFER MEDICATION	Medications on discharge, Transfer Meds
FOLLOW UP	Discharge instructions and followup, Follow-up Plan, Followup Instructions
OTHER DISCHARGE INFORMATION	Discharge activity, Discharge Diet
INTERVAL EVENTS	Interval events, 24 hour events, o/n

Table 3: 18 core section types used in the study.