Clinical Information Retrieval with Split-layer Language Models

Stephen Wu
Mayo Clinic
200 First St SW
Rochester, MN
wu.stephen@mayo.edu

Dongqing Zhu
University of Delaware
101 Smith Hall
Newark, DE 19716
zhu@cis.udel.edu

William Hersh
Oregon Health & Science University
3181 Sam Jackson Park Road
Portland, OR
hersh@ohsu.edu

Hongfang Liu
Mayo Clinic
200 First St SW
Rochester, MN
liu.hongfang@mayo.edu

ABSTRACT
With the increasing prevalence of electronic medical records (EMRs), search technologies for these systems hold significant promise for improving patient and population care. We present a split-layer language model that embeds linguistic layers from existing NLP systems in retrieving medical documents. On the cohort identification task of the TREC Medical Records Track, our approach shows improvement over baselines, with the best performance achieved by mixing in all tested layers of NLP artifacts.

Categories and Subject Descriptors
H.3.3 [Information Search and Retrieval]: Retrieval models

Keywords
layered language model, medical records, information retrieval, natural language processing

1. INTRODUCTION
With the increasing prevalence of electronic medical records (EMRs), search technologies for these systems hold significant promise for improving patient and population care. The use of EMR text (e.g., progress notes) is particularly important for the clinical domain because some information (e.g., symptoms) is recorded nowhere else. However, raw text alone will not capture some important aspects of meaning. For example, saying that a patient has cancer, history of cancer, may have cancer, or does not have cancer will each have a different effect on relevance to a query. Keyword searches will fail to discover this type of granular information.

Addressing the need for fine-grained analysis, significant work has gone into NLP and information extraction (IE) techniques in the clinical domain, including systems like cTAKES [4], MetaMap [1], and MedTagger [2]. Each of these systems allows medical concepts to be represented. Attributes of those concepts, such as history, uncertainty, and negation, are also typically represented and discovered. These layers of NLP analyses have yet to be fully integrated into clinical retrieval techniques. Thus, the present work is a preliminary step towards this full integration, making NLP-produced concepts and attributes searchable by using language models that account for those artifacts.

In order to evaluate the contribution of this language model, we adopt the retrieval framework of cohort identification from EMR text, as defined in the Text Retrieval Conference (TREC) Medical Records Track [8, 7]. Results show that mixing in different layers of language processing is beneficial in almost every setting, even with naïve weighting schemes.

The rest of this paper proceeds as follows. Section 2 reviews some related work, Section 3 introduces a language model that incorporates layers of NLP output, Section 4 describes the TREC-med evaluation and our system, and Section 5 presents our results.

2. RELATED WORK
Unstructured data in IR has been augmented by structured objects (such as NLP artifacts in this work) in several ways. First, searching structured objects has typically been the realm of relational database searches, and some work has been done in embedding text search in these frameworks. Querying both text and its linked NLP artifacts, however, is a challenging and unsolved task. Second, perhaps the primary NLP artifact to be considered in layered language IR is that of named entities (concepts) and their attributes, and significant recent research has focused on entity search and semantic search. Our proposed work will align with these techniques in many cases, but we will take a language modeling approach that weighs NLP artifacts together with these entities (and other structures) in a statistical framework. While overall there were few early successes with using NLP techniques in IR, some approaches were shown to benefit retrieval [9]. Unlike many of those early systems, the split-layer language model is done in a language mod-
Cohort identification is a well-known problem in medical informatics. However, few approaches have used traditional text-based IR approaches. Notable exceptions are the EMERSE (Electronic Medical Record Search Engine) system and the recent work done on the TREC Medical Records Track. EMERSE is a non-commercial EMR search engine that supports free-text queries and has demonstrated the effectiveness of IR techniques in making chart reviews more efficient [5], but it focused on a sound user interface rather than underlying retrieval techniques. The TREC Medical Records track provided resources for significant innovation in medical IR, with a shareable collection of clinical text, information needs, and judgments [8, 7]. We will use the TREC-med setting as our evaluation framework.

3. MODEL

We introduce the split-layer language model (§3.3), a preliminary instantiation of a general class of layered language models that extend the query likelihood language model (§3.1) with NLP-produced artifacts (§3.2). The intuition is that human language can be described in somewhat overlapping linguistic layers: phonology, morphology, syntax, semantics, and discourse. Search technologies often perform quite well with just surface forms (raw text), but especially in the clinical domain, other layers are important. Access to these other layers is provided through standard NLP techniques such as parsing and named entity recognition (NER).

3.1 Query likelihood language model

In IR language models, it is common to rank according to score($d, q$) = $P(d) \cdot P(q \mid d)$, where the latter conditional probability encapsulates the intuition that an ad hoc user trying to find document $d$ will try to write an effective query $q$. We will represent the document or query as a list of terms, which we write $T_d$ and $T_q$, respectively. The standard query likelihood language model estimates the conditional probability as a Dirichlet-smoothed maximum likelihood estimate.

$$\hat{P}_{QL}(T_q \mid T_d) \overset{\text{def}}{=} (1 - \alpha_D) \prod_{t_q \in T_q} \hat{P}_d(t_q \mid T_d) + \alpha_D \prod_{t_q \in T_q} \hat{P}_D(t_q \mid T_D)$$  

(1)

where the $\alpha_D$ is related to the Dirichlet smoothing parameter $\mu$ according to $\alpha_D = \frac{\mu}{\mu + |T_d|}$. Note that $\hat{P}_d(t_q \mid T_d)$ actually denotes a probability conditioned on a model $\Theta_{T_d}$, but we will drop the model variable $\Theta$ throughout.

3.2 Indexing multiple linguistic layers

In order to incorporate layers of linguistic information, we first assume that these layers can be inferred from text — whether the document or the query. This inference is done by means of standard NLP techniques and stored in index layers. The left side of Figure 1 shows two sentences processed with an NLP pipeline and some output data types; e.g., sentence detection, named entity recognition, and attribute discovery on the text might result in Sentence and Concept types, with Polarity Semantic Group attributes. These output NLP artifacts are explicitly tied to the text with begin and end character offsets. To make the diverse NLP structures retrievable, we build retrieval indexes (Figure 1, right side) using two strategies: annotations and fields.

Fields have separate inverted indexes. Figure 1 (right side) shows an index for the original text, labeled “text field,” alongside an index for an NLP artifact, labeled “artifact field.” NLP artifacts that warrant their own artifact fields are termed content artifacts. In the example, the concept unique identifiers (CUIs, from the Unified Medical Language System (UMLS) Metathesaurus) of medically-relevant concepts may be stored as ‘text’ in a concept index. This is essentially a layer for the shallow semantic representation of NER-produced concepts. The concept-based artifact field at the bottom-right of Figure 1 could, at minimum, be used to support CUI searches.

Annotations provide additional layered information whose meaning is inextricably tied to the items in an index. These annotations retain begin and end offsets with respect to the index they are a part of. In the example, Sentence annotations divide the text into different sentences, and are thus annotations on the text field. However, attribute discovery finds Polarity (negation) values on newly-recognized Concepts (named entities), so Polarity values are first stored in the Concept index as annotations (bottom right, Figure 1). Because the Concepts themselves correspond to spans on the original text, it is possible to populate the text space with the corresponding annotations as well (top right, Figure 1). An alternative strategy for future work would be to consider storing a pointer from items in the artifact field to their corresponding spans in the text field. Annotation values may also be indexed (in a manner similar to fields) in
order to support narrower searches. For example, a search for CUIs could be constrained to only non-negated concepts, or could exclude medications. A search specifying sentence annotations gives the expected behavior of extent retrieval.

3.3 Split-layer language model

We define a query according to its index layers as \( q = (T_q, C_q, A_q) \), where \( T \) is the list of text tokens, \( A \) is the annotations, and \( C \) is the content artifacts associated with the query text. The queries thus deterministically contain all NLP-extracted structures. Similarly, we can write documents as \( d = (T_d, C_d, A_d) \).

For simplicity, our equations below will make the assumption that there is only one type of annotation and one type of content artifact. We will write annotations \( A \) together with their respective fields \( T \) or \( C \). Notationally, we use uppercase variables to represent sets of lowercase variables, so that a corpus \( D \) is a set of documents \( d \). It is not the case that every type of structure discovered by clinical NLP methodologies will be empirically useful in document retrieval; however, the probabilistic framework must account for the multiple layers of artifacts.

Here, we introduce layered language \( T \), \( A \), and \( C \) components. We make independence assumptions between the text field and any content fields, and treat annotations (on either text or content artifacts) jointly.

\[
\hat{P}_{SL}(q|d) \overset{\text{def}}{=} f \left( \hat{P}_{TBR}(T_q|T_d), \hat{P}_{TBRa}(T_A_q|T_A_d), \hat{P}_{CBR}(C_q|C_d), \hat{P}_{CBRa}(C_A_q|C_A_d) \right)
\]

where \( \hat{P}(\cdot) \) represents an estimated distribution that will typically include a Dirichlet term. For example, the first term in Eq. 2 would then be implemented as the query likelihood model \( \hat{P}_{QL}(T_q|T_d) \) of Eqn. 1. The function \( f(\cdot) \) represents the way to combine the probabilities from different layers. In this paper, we define \( f(\cdot) \) as a simple linear combination function whose effect is similar to ranking by combining different document representations [3,10].

We have four basic models of the layered query likelihood, each corresponding to the four terms of Eqn. 2:

- **TBR**: Text-based retrieval, i.e., the query likelihood model.
- **TBRa**: Text-based retrieval with annotations.
- **CBR**: Concept (CUI)-based retrieval.
- **CBRa**: Concept (CUI)-based retrieval with annotations.

Note that each of the layers essentially implement backoff and smoothing. Because there may be insufficient statistics for concepts with annotations \( \hat{P}(C_A_q|C_A_d) \), the estimates for \( \hat{P}(C_q|C_d) \) serve as an backoff model alongside Dirichlet smoothing. The same can be said for the ‘text only’ and ‘text with annotations’ layers.

4. Evaluation

4.1 Cohort identification task

We evaluated the contribution of different split-layer language models on the task of cohort identification, mirroring the Text Retrieval Conference (TREC) Medical Records Track [8,7]. The retrieval collection was the University of Pittsburgh’s BLU repository. Each patient at the University of Pittsburgh would have one or more medical records (documents) associated with one or more of his/her visits to the hospital. The unit of retrieval was defined as a patient visit, since they were broken by the de-identification procedure that made the records shareable. In total, there were 95,702 records that corresponded to 17,198 visits. The largest visit was 418 records, but the mean visit was 5.56 records.

4.2 System and evaluation setup

To isolate the contributions of the split-layer language model, we ignored any metadata (primarily ICD-9 codes, whose availability and reliability are inconsistent) and based retrieval on the text portions only. We processed this text using the MedTagger [2] information extraction system, producing “layers” including: content artifacts (UMLS CUIs), and contextual attributes on those artifacts (semantic group, negation, uncertainty, and experriener).

These artifacts were indexed in Indri [6] using the offset annotation capabilities. Rather than considering all the layers generated by MedTagger, we focused our evaluation on those that were most likely to be semantically relevant, namely, the CUIs and the contextual attributes listed above. Thus, TBR was a basic query likelihood model; TBRa utilized the concept-based annotations (with the offset spans mapped appropriately); CBR was populated with CUIs; CBRa contained both CUIs and associated annotations.

Our tests did not train any parameters, but tested on all 81 queries from the official evaluations of TREC-med 2011–2012. We use mean average precision (MAP) as our main evaluation metric, since the official metrics for TREC-med were different in 2011 and 2012 (bpref and infAP) but MAP scores corresponded to the official metrics in both cases. We tested the contribution of the Dirichlet smoothing parameter \( \mu \) to each of TBR, TBRa, CBR, and CBRa, comparing how each performed. We then examined the possible configurations in which these models could be linearly combined.

5. Results and discussion

5.1 Dirichlet smoothing

Figure 2 shows the effect of the Dirichlet smoothing parameter on the four separate layers. Interestingly, it appears that the performance is inversely proportional to the smoothing parameter. This is surprising given the short length of medical documents, since shorter document lengths typically correspond to more need for collection-level smoothing. These effects may be because medical documents have relevant information concentrated in relatively few notes (such as those within the same specialty or note type); performance is diluted by smoothing with the whole collection. Thus, we use the lowest tested Dirichlet parameters (\( \mu = 2500 \)) in all further analyses.

5.2 Comparison between layers
demonstrate the capabilities of annotations — annotations of TREC-med pools and corresponding metrics do not easily demonstrate the capabilities of annotations — annotations imply weighted relevance for text or content artifacts. Figure 2 also has four lines, corresponding to the TBR, TBRa, CBR, and CBRa models. It is clear that without any mixing of the models, the layer of annotations decreases performance on both the text-based and concept-based retrieval models. When evaluating on the standard metrics such as MAP, this is likely due to the lack of sufficient data to support such specific maximum likelihood models. However, we also hypothesize that the binary relevance judging of TREC-med pools and corresponding metrics do not easily demonstrate the capabilities of annotations — annotations imply weighted relevance for text or content artifacts.

### 5.3 Split-layer combinations

The capabilities of the split-layer language model are evaluated by linearly combining the four basic models in different configurations. In Table 1, we show the effects of a naive linear combination with equal weights (i.e., arithmetic mean of included components). Moving along each row, we see the positive effect of mixing in the different annotations layers (statistical significance as compared to the baselines is notated by the superscripts). While we noted from Figure 2 that annotations-included layers alone (TBRa or CBRa) underperformed their respective content fields (TBR and CBR), it is clear that mixing content layers with annotations tends to improve performance. Furthermore, we can see the benefits of mixing text with a content field by moving down the columns; in every case, TBR+CBR outperforms either TBR or CBR alone. Both of these observations are captured in the bottom-right result, in which all four models (TBR + CBR + TBRa + CBRa) yield the best overall performance.

### 6. CONCLUSIONS

We have introduced the split-layer language model, a model that embeds multiple linguistic layers into IR analysis by incorporating the results of core NLP tasks. We have shown that, on the task of cohort identification from medical records, the split-layer language model improves performance over a query likelihood baseline. A model that mixes all text, content artifacts, and annotations performed the best overall.

The split-layer language model is termed such because of the independence assumptions between the text and content artifacts layers. Other layered language models are possible — multiple layers could be integrated into the component models, such that text and content artifacts are estimated together in probability distributions. This is an active area of future work. Additionally, future work will evaluate the effect of smoothing parameters, and will establish a means by which multiple collocated artifacts can be queried at once.

### 7. REFERENCES


