Finding Temporal Order in Discharge Summaries

Philip Bramsen†, Pawan Deshpande†, Yoong Keok Lee‡, MS and Regina Barzilay†, PhD
Massachusetts Institute of Technology (MIT), Cambridge, MA†
DSO National Laboratories, Singapore‡

Abstract

A method for automatic analysis of time-oriented clinical narratives would be of significant practical import for medical decision making, data modeling and biomedical research. This paper proposes a robust corpus-based approach for temporal analysis of medical discharge summaries. We characterize temporal organization of clinical narratives in terms of temporal segments and their ordering. We consider a temporal segment to be a fragment of text that does not exhibit abrupt changes in temporal focus. Our method derives temporal order based on a range of linguistic and contextual features that are integrated in a supervised machine-learning framework. Our learning method achieves 83% F-measure in temporal segmentation, and 78.3% accuracy in inferring pairwise temporal relations.

I. Introduction

Temporal analysis plays a key role in automatic processing of clinical data. A large body of research is concerned with reasoning about time-oriented clinical data in the context of medical decision making, data modeling and biomedical research.[1]

In some cases, relevant temporal information is directly available in a medical record and can be readily used for subsequent analysis. Consider, for instance, a database record of a laboratory test; time information is likely to be stored in one of the fields of this record, and can be fetched on demand. This access strategy, however, is clearly not suitable for analyzing temporal flow in discharge summaries and other clinical narrative records. In a typical narrative, absolute temporal markers are sparse, and understanding of a temporal flow requires inference over subtle contextual cues.

One possible strategy for interpreting temporal information in a clinical narrative is to employ existing tools for temporal analysis developed in the natural language community. The functionality of these tools include extraction of temporal expressions,[2] time stamping of event clauses,[3] and temporal ordering of events.[4,5] In practice, however, the accuracy of these systems is not sufficient for robust analysis of clinical data. Moreover, most of these methods have been developed for newspaper documents. A marked difference in text organization between clinical narratives and standard newspaper collections calls for new methods of temporal analysis. First steps in this direction have been taken by Zhou et al. who develop a rule-based system for temporal information extraction and reasoning.[6]

This paper proposes a robust machine-learning approach for temporal analysis of medical discharge summaries. We characterize temporal organization of clinical narratives in terms of *temporal segments* and their ordering. We consider a temporal segment to be a fragment of text that does not exhibit abrupt changes in temporal focus. For instance, a medical discharge summary may contain segments describing a patient's admission, his previous hospital visit, and original symptoms' onset. Each of these segments corresponds to a different time frame, and is clearly delineated as such in a text.

Our ultimate goal is to automatically identify temporal segmentation in clinical narratives and induce their ordering. Our key assumption is that temporal progression is reflected in a wide range of linguistic features and contextual dependencies. For instance, given a pair of adjacent segments, the temporal adverb *next hospital visit* in the second segment is a strong predictor of precedence relation. We hypothesize that temporal segmentation and ordering can be learned from a corpus of annotated summaries, based on a set of automatically extracted features.

In the following section, we introduce our temporal annotation scheme and motivate its benefits. Next we describe a corpus of discharge summaries annotated with temporal information, and present our methods for temporal segmentation and ordering. We conclude the paper by presenting and discussing our results.

II. Temporal Annotation Scheme

We view a clinical narrative as a linear sequence of temporal segments. Temporal focus is retained within a segment, but radically changes in between segments.[7] The length of a segment can range from a single clause to a sequence of adjacent sentences. Table 1 shows a discharge summary marked with temporal segment boundaries. Consider as an example the last segment in this text. This segment describes an examination of a patient, and includes several events and states (i.e., an abdominal and neurologic examination) which belong to the same time

frame. Temporal progression of these events is not outlined in the text, and therefore we treat them as one segment.

Given a pair of segments, we categorize it to one of the three ordering relations: BEFORE, AFTER or INCOMPARABLE. The first two relations capture temporal precedence. For instance, the pair of segments (S1, S5) from Table 1 belongs to the AFTER category because the event of admission (S1) happened after the tests described in S5. The pair (S1, S13) stands in the BEFORE relation, since the patient was first admitted to the hospital (S1) and then examined (S13). We cannot always induce the precedence relation between two segments. For instance, consider the segments S5 and S7, which describe patient's previous tests and the onset of eczema. Any order between the two events is consistent with our interpretation of the text; therefore we cannot determine the precedence relation. In such cases, a pair stands in the INCOMPARABLE relation.

In contrast to many existing temporal representations,[8,9] our annotation scheme is coarse: it does not capture event overlap, and distinguishes only a subset of commonly used ordering relations. Our choice of this representation, however, is not arbitrary. These relations are shown to be useful in processing medical discourse and can be reliably recognized by humans. Moreover, the distribution of event ordering links under a more refined annotation scheme, such as TimeML, shows that our subset of relations covers a vast majority of annotated links. [9]

Table 1: Fragment from a discharge summary with four segments

S1: A 32-year-old woman was admitted to the hospital because of left subcostal pain, bouts of fever, and a mass in the left hepatic lobe.
S5: Three months before admission an evaluation elsewhere included an ultrasonographic examination, a computed tomographic (CT) scan of the abdomen, and a magnetic resonance imaging (MRI) scan.
S7: She had a history of eczema and of asthma.
S13: On examination the patient was slim and appeared well. Scaling eczematous plaques were present on the hands and elbows. An abdominal examination revealed a soft systolic bruit in the midepigastrium that overlay a firm, nontender mass. A rectal examination was normal; a stool specimen was negative for occult blood. The arms and legs were normal, and a neurologic examination was normal.

III. Methods

A. *Data*

We compiled a corpus of medical discharge summaries from the on-line edition of The New England Journal of Medicine (NEJM).[10] The summaries are written by physicians of Massachusetts General Hospital. The summaries are edited to follow a particular grammatical and narrational style, which distinguish them from a typical discharge summary written by a physician. While processing an unedited discharge summary is a more challenging task, the ability to temporally order an NEJM summary is the first step towards this goal.

A typical summary describes an admission status, previous diseases related to the current conditions, treatments, family history, and the current course of treatment. For privacy protection, names and dates are removed from the summaries before publication.

The average length of a summary is 47 sentences. The summaries are written in the past tense, and a typical summary does not include instances of the past perfect. The summaries do not follow a chronological order. The ordering of information in this domain is guided by stylistic conventions (i.e., symptoms are presented before treatment) and the relevance of information to the current conditions (i.e., previous onset of the same disease is summarized before the description of other diseases).

i. *Annotating Temporal Segmentation*

A machine-learning approach for temporal segmentation requires annotated data for supervised training. We first conducted a pilot study to assess the human agreement on the task. We hired two annotators to manually segment a portion of our corpus. The annotators were provided with two-page instructions that defined the notion of a temporal segment and included examples of segmented texts. Each annotator segmented eight summaries which on average contained 49 sentences. The first annotator created 168 boundaries, while the second created 224. We computed agreement using the Kappa coefficient. The observed value of Kappa 0.71 indicates a reasonable inter-annotator agreement, and confirms our hypothesis about reliability of temporal segmentation.

Once we established high inter-annotator agreement on the pilot study, one annotator segmented the remaining 52 documents in the corpus. The annotator on average marked 20 boundaries per document. The average length of a segment is three sentences. While 80% of the boundaries occurred at the end of sentences, the rest of the boundaries were placed within a sentence, on the boundary of syntactic clauses. Among 3,297 potential boundaries, 1,178 (36%) were identified as segment boundaries by the annotator.

ii. *Annotating Temporal Ordering*

First we investigated the inter-annotator agreement on the ordering task. Two human annotators ordered segments from five manually segmented summaries,

with an average length of 20 segments. We computed the agreement between human judges by comparing the transitive closure of the derived orderings which consists of 1,331 ordered pairs. The annotators achieved a surprisingly high agreement with a Kappa value of 0.98.

After verifying a human agreement on this task, one of the annotators ordered segments in another 25 summaries. Our corpus consists of 6,544 ordered segment pairs, including pairs derived through transitive closure. The BEFORE relation is prevalent – it accounts for 72% of the pairs. The AFTER relation covers 12% of the pairs, and 16% of the pairs are not comparable.

B. *Method for Temporal Segmentation*

Our first goal is to automatically predict shifts in temporal focus indicative of segment boundaries. Linguistic studies show that speakers and writers employ a wide range of language devices to signal change in temporal discourse.[11] For instance, the presence of the temporal anchor *last year* indicates the lack of temporal continuity between the current and the previous sentence. However, many of these predictors are heavily context-dependent, and thus cannot be considered independently. Instead of manually crafting complex rules controlling feature interaction, we opt to learn them from the data.

We model temporal segmentation as a binary classification task. Given a set of candidate boundaries (i.e., sentence boundaries), our task is to select a subset of the boundaries that delineate temporal segment transitions. We assume that every boundary is represented by a vector of features that are relevant to the segmentation decision. We learn the weight of each feature and their optimal combination in a supervised discriminative framework. In our experiments we used the publicly available BoosTexter classifier.[12]

To implement this approach, we first identify a set of potential boundaries. Our analysis of the manually-annotated corpus reveals that boundaries can occur not only between sentences, but also within a sentence, at the boundary of syntactic clauses. We automatically segment sentences into clauses using a robust statistical parser.[13] Next, we encode each boundary as a vector of features (see descriptions below). Given a set of annotated examples, we train a classifier to learn the feature weights. Once the classifier is trained, we can use it to predict boundaries in new, unseen narratives. In the rest of the section, we describe features used for boundary representation.

Lexical Features Temporal expressions, such as *tomorrow* and *earlier*, are among the strongest markers of temporal discontinuity.[11] In addition to a well-studied set of domain-independent temporal markers, there are a variety of temporal markers specific to the medical discourse. For instance, the phrase *initial hospital visit* functions as a time anchor in a discharge summary. See Table 2 for an example of highly-ranked lexical features learned by our algorithm.

Table 2: Highly ranked features selected automatically by the classifier for temporal segmentation. For presentation purposes, the features are manually grouped based on their semantic role.

Type	Preceding	Following
Relative temporal anchors	*until, when*	*later, next, subsequently, after, followed by*
Absolute temporal anchors	*pm, day, during*	*days, years, one month, hospital day*
Other terms	*showed, recurred, was transferred*	*admission, physical examination*

To automatically extract these expressions, we provide a classifier with unigrams, bigrams and trigrams from each of the candidate sentences preceding and anteceding the candidate segment boundary.

Topical Continuity Temporal segmentation is closely related to topical segmentation.[14] A typical discharge summary covers several topics that span over different time periods, ranging from family history to the current treatment. Thus, predicting these topical transitions can help us to locate boundaries of temporal segments.

We quantify the strength of a topic change by computing a cosine similarity between sentences bordering the proposed segmentation. This measure is commonly used in topic segmentation under the assumption that change in lexical distribution corresponds to topical change.

Positional Features We observe a correlation between the position of the sentence in a discharge summary and the likelihood that it constitutes a boundary. This property is related to patterns in discourse organization of a document as a whole. Some parts of the document are more likely to exhibit temporal change than others. For instance, a medical discharge summary first discusses various developments in the patient's clinical course, and then focuses on his current conditions. Thus, the first part of the summary contains many short temporal segments.

We encode positional features by recording the relative position of a sentence in a discharge summary. In addition, we include the half and the quarter to which the sentence belongs.

Syntactic Features Because our segment boundaries are considered at the clausal, rather than at the sentence level, the syntax surrounding a hypothesized boundary may be indicative of temporal shifts. We represent this feature with a sequence of part-of-speech tags surrounding the potential boundary.

All the feature vectors are available on the website.[15]

C. *Learning to Order Segments*

Our next goal is to find an acceptable order among temporal segments. One possible approach is to cast this task as a standard classification task: predict an ordering for each segment pair based on their attributes alone. If a pair contains a temporal marker, like *two days ago*, then accurate prediction is feasible. In fact, this method is commonly used in event ordering.[4,5] However, many segment pairs lack temporal markers and other explicit cues for ordering. Determining their relation out of context can be difficult, even for humans. Moreover, by treating each segment pair in isolation, we cannot guarantee that all the pairwise assignments are consistent with each other, and yield an *acyclic order*.

Therefore, our ordering algorithm has two steps. First, we employ a classifier to predict an order between a pair of temporal segments. Next, we look for a consistent ordering that combines pairwise predictions, and disallows cycles.

i. *Learning Pairwise Ordering*

Given a pair of segments *(i,j)*, our goal is to assign it to one of three classes: BEFORE, AFTER or INCOMPARABLE. We generate the training data by using all pairs of segments *(i,j)* that belong to the same document, such that *i* appears before *j* in the text. We represent each pair of segments by a vector of automatically extracted features. Similarly to the temporal segmentation task, we apply a discriminative classifier to train the model. Features used in our model are summarized below:

Lexical Features This class of features captures temporal markers and other phrases indicative of order between two segments. Representative examples in this category include domain-independent cues like *years earlier* and domain-specific markers like *during next visit*.

To automatically identify these phrases, we provide a classifier with two sets of *n*-grams extracted from the first and the second segment. The classifier automatically learns phrases with high predictive power.

Temporal Anchor Comparison Temporal anchors are one of the strongest predictors of the event ordering in a summary. For instance, medical discharge summaries use phrases like *two days before admission* to express temporal progression. If the two segments contain temporal anchors, we can determine their ordering by comparing these anchors. We identified a set of temporal anchors commonly used in the discharge summaries, and devised rules for their comparison.

Segment Adjacency Feature By analyzing a corpus of discharge summaries, we found that pairs of adjacent segments are likely to follow a chronological progression.[11] To encode this information, we include a binary feature that captures the adjacency relation between two segments.

ii. *Finding Consistent Ordering*

Given the scores produced by a pairwise classifier, our task is to construct a consistent ordering. The need to perform this additional step arises because pairwise decisions may be inconsistent with each other. For instance, the classifier may predict the BEFORE relation for the pairs (S1, S2) and (S2, S3), but assign the AFTER relation for (S1, S3), thus yielding an ordering cycle. When resolving such conflicts, our priority is to retain relations that the classifier predicted with high confidence. The confidence of the classifier is reflected in its score. Therefore, our goal is to find a consistent assignment with the highest score.

The consistent ordering assignment can be encoded as a directed acyclic graph wherein nodes correspond to temporal segments, and the direction of the edge captures the ordering relation. We search for an optimal ordering graph using a greedy strategy. The algorithm begins by sorting pairwise relations (edges) based on their score. Starting with an empty graph, we add one edge at a time, without violating the consistency constraints. At each step we expand the graph with its transitive closure. We continue this process until all the edges have been considered. While this greedy strategy is not guaranteed to find the optimal solution, it finds a close approximation to the optimal graph.[16]

IV. Evaluation Set-Up

Given the limited size of the available annotated data, we evaluated our system in the leave-one-out cross-

validation scenario. In this framework, a model is tested on one narrative, while trained on the rest of the documents. This process is repeated for each document in a corpus, and the accuracy is averaged over all the splits.

Using cross-validation, we evaluated the temporal segmentation algorithm on the corpus of 60 manually-segmented summaries. The segmentation algorithm achieves the performance of 83% F-measure (recall 78% and precision 89%).

Using the same cross-validation framework, we assessed the accuracy of the ordering component. The algorithm achieves 78.3% accuracy on the ternary relation classification. As a point of comparison, we consider a baseline which assigns the BEFORE relation to all the pairs. In other words, the baseline assumes that a discharge summary follows chronological order. The accuracy of this baseline is 72.2%. Our automatic method outperforms the majority baseline by 6.1%

V. Discussion and Conclusions

This paper introduced a new method for temporal ordering of discharge summaries. Temporal analysis in this domain is challenging in several respects: a typical summary exhibits no significant tense and aspect variations, and contains few absolute time markers. We demonstrate that humans can reliably mark temporal segments, and determine segment ordering in this domain. Our learning method achieves 83% F-measure in temporal segmentation, and 78.3% accuracy in inferring temporal relations between two segments.

Most work on temporal analysis is performed on a finer granularity than proposed here. [6] However, the granularity of our representation facilitates temporal analysis, and is especially suitable for domains with sparse temporal anchors. The output of our algorithm provides useful information for processing clinical narratives. It can be also used as a preprocessing step for more refined methods to reduce the complexity of their analysis.

The strength of our approach lies in its ability to simultaneously optimize pairwise ordering preferences and global constraints on the consistency of the ordering. While the importance of global constraints has been previously validated in symbolic systems for temporal analysis,[6] existing corpus-based approaches operate at the local level.[4,5] The improvements achieved by a global model motivate its use as an alternative to existing pairwise methods. For a more detailed investigation of global inference strategies for segment ordering, see the following paper.[17]

Finally, we collected and annotated corpus of medical discharge summaries. This is the first publicly available corpus of clinical narratives annotated with temporal information.[15] We believe that this corpus can be used as a benchmark for evaluating ordering algorithms in the medical domain, thereby facilitating the development of corpus-based methods for temporal analysis of clinical narratives.

Acknowledgments

This work was supported in part by the National Institutes of Health through the NIH Roadmap for Medical Research, Grant U54LM008748. Information on the National Centers for Biomedical Computing can be obtained from: http://nihroadmap.nih.gov/bioinformatics

References

1. Carlo Combi and Yuval Shahar. Temporal reasoning and temporal data maintenance in medicine: Issues and challenges. Computers in Biology and Medicine. 1997; 27(5):353--368.
2. George Wilson, Inderjeet Mani, Beth Sundheim, and Lisa Ferro. A multilingual approach to annotating and extracting temporal information. Proc. ACL 2001 Workshop on Temporal and Spatial Information Processing. 2001; 81-7.
3. Elena Filatova and Eduard Hovy. Assigning time-stamps to event-clauses. Proc. ACL 2001 Workshop on Temporal and Spatial Information Processing. 2001; 104--111.
4. Inderjeet Mani, Barry Schiffman, and Jianping Zhang. Inferring temporal ordering of events in news. Proc. HLT-NAACL 2003.
5. Branimir Boguraev and Rie Kubota Ando. TimeML-compliant text analysis for temporal reasoning. Proc. IJCAI. 2005; 997-1003.
6. Li Zhou, Carol Friedman, Simon Parsons, and George Hripcsak. System architecture for temporal information extraction, representation and reasoning in clinical narrative reports. Proc. AMIA. 2005; 869-873.
7. Bonnie L. Webber. Tense as discourse anaphor. Computational Linguistics. 1988; 14(2):61-73.
8. James F. Allen. Towards a general theory of action and time. Artificial Intelligence. 1984; 23(2):123-154.
9. James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, David Day, Lissa Ferro, Robert Gaizauskas, Marcia Lazo, Andrea Setzer, and Beth Sundheim. The timebank corpus. Corpus Linguistics. 2003; 647--656.
10. http://content.nejml.org
11. Yves Bestgen and Wietske Vonk. The role of temporal segmentation markers in discourse processing. Discourse Processes. 1995; 19:385-406.
12. R. Schapire and Y. Singer. Boostexter: A boosting-based system for text categorization. Machine Learning. 2000; 39(2/3):135-168.
13. Michael Collins. Head-Driven Models for Natural Language Parsing. Ph.D. thesis, University of Pennsylvania. 1999.
14. Wallace Chafe. The flow of thought and the flow of language. In Talmy Givon, editor, Syntax and Semantics: Discourse and Syntax. 1979; 12: 159-182. Academic Press.
15. http://people.csail.mit.edu/regina/temp
16. William Cohen, Robert Schapire, and Yoram Singer. Learning to order things. Journal of Artificial Intelligence, 1999; 10:243--270.
17. Philip Bramsen, Pawan Deshpande, Yoong Keok Lee, and Regina Barzilay. Inducing Temporal Graphs. Proc. EMNLP. 2006.