Cross-narrative temporal ordering of medical events

Preethi Raghavan*, **Eric Fosler-Lussier***, **Noémie Elhadad**† **and Albert M. Lai***
*The Ohio State University, Columbus, Ohio
†Columbia University, New York, NY
{raghavap, fosler}@cse.ohio-state.edu
noemie.elhadad@columbia.edu, albert.lai@osumc.edu

Abstract

Cross-narrative temporal ordering of medical events is essential to the task of generating a comprehensive timeline over a patient's history. We address the problem of aligning multiple medical event sequences, corresponding to different clinical narratives, comparing the following approaches: (1) A novel weighted finite state transducer representation of medical event sequences that enables composition and search for decoding, and (2) Dynamic programming with iterative pairwise alignment of multiple sequences using global and local alignment algorithms. The cross-narrative coreference and temporal relation weights used in both these approaches are learned from a corpus of clinical narratives. We present results using both approaches and observe that the finite state transducer approach performs performs significantly better than the dynamic programming one by 6.8% for the problem of multiple-sequence alignment.

1 Introduction

Discourse structure, logical flow of sentences, and context play a large part in ordering medical events based on temporal relations within a clinical narrative. However, cross-narrative temporal relation ordering is a challenging task as it is difficult to learn temporal relations among medical events which are not part of the logically coherent discourse of a single narrative. Resolving cross-narrative temporal relationships between medical events is essential to the task of generating an event timeline from across unstructured clinical narratives such as admission notes, radiology reports, history and physical reports and discharge summaries. Such a timeline has multiple applications in clinical trial recruitment (Luo et al., 2011), medical document summarization (Bramsen et al., 2006, Reichert et al., 2010) and clinical decision making (Demner-Fushman et al., 2009).

Given multiple temporally ordered medical event sequences generated from each clinical narrative in a patient record, how can we combine the events to create a timeline across all the narratives? The tendency to copy-paste text and summarize past information in newly generated clinical narratives leads to multiple mentions of the same medical event across narratives (Cohen et al., 2013). These cross-narrative coreferences act as important anchors for reasoning with information across narratives. We leverage cross-narrative coreference information along with confident cross-narrative temporal relation predictions and learn to align and temporally order medical event sequences across longitudinal clinical narratives. We model the problem as a sequence alignment task and propose solving this using two approaches. First, we use weighted finite state machines to represent medical events sequences, thus enabling composition and search to obtain the most probable combined sequence of medical events. As a contrast, we adapt dynamic programming algorithms (Needleman et al., 1970, Smith and Waterman, 1981) used to produce global and local alignments for aligning sequences of medical events across narratives. We also compare the proposed methods with an Integer Linear Programming (ILP) based method for timeline construction (Do et al., 2012). The cross-narrative coreference and temporal relation scores used in both these approaches are learned from a corpus of patient narratives from The Ohio State University Wexner Medical Center.

The main contribution of this paper is a general framework that allows aligning multiple event sequences using cascaded weighted finite state transducers (WFSTs) with the help of efficient composition and decoding. Moreover, we demonstrate that this method can be used for more accurate multiple sequence alignment when compared to

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pages 998–1008,
Baltimore, Maryland, USA, June 23-25 2014.

dynamic programming or other ILP-based methods proposed in literature.

2 Related Work

In the areas of summarization and text-to-text generation, there has been prior work on several ordering strategies to order pieces of information extracted from different input documents (Barzilay et al., 2002, Lapata, 2003, Bollegala et al., 2010). In this paper, we focus on temporal ordering of information, as discussed next.

Recent state-of-the art research has focused on the problem of temporal relation learning within the same document, and in many cases within the same sentence (Mani et al., 2006, Verhagen et al., 2009, Lapata and Lascarides, 2011). Chambers and Jurafsky (2009) describe a process to induce a partially ordered set of events related by a common protagonist by using an unsupervised distributional method to learn relations between events sharing coreferring arguments, followed by temporal classification to induce partial order. The task was carried out on the Timebank newswire corpus, but was limited to an intra-document setting. More recently, (Do et al., 2012) proposed an ILP-based method to combine the outputs of an event-interval and an event-event classifier for timeline construction on the ACE 2005 corpus. However, this approach is also restricted to events within documents and requires annotations for event intervals. We empirically compare our methods for timeline creation from longitudinal clinical narratives to such an ILP-based approach in Section 7. While a lot of this work has been done in the news domain, there is also some recent work in rule-based algorithms (Zhou et al., 2006) and machine learning (Roberts et al., 2008) applied to temporal relations between medical events in clinical text. Clinical narratives are written in a distinct sub-language with domain specific terminology and temporal characteristics, making them markedly different from newswire text.

There is limited prior work in learning relations across documents. Ji and Grishman (2008) extended the one sense per discourse idea (Yarowsky, 1995) to multiple topically related documents and propagate consistent event arguments across sentences and documents. Barzilay and McKeown (2005) propose a text-to-text generation technique for synthesizing common information across documents using sentence fusion. This involves multisequence dependency tree alignment to identify phrases conveying similar information and statistical generation to combine common phrases into a sentence. Along with syntactic features, they combine knowledge from resources like WordNet to find similar sentences. In case of clinical narratives and medical event alignment, the objective is to identify a unique sequence of temporally ordered medical events from across longitudinal clinical data.

To the best of our knowledge, there is no prior work on cross-document alignment of event sequences. Multiple sequence alignment is a problem that arises in a variety of domains including gene/protein alignments in bioinformatics (Notredame, 2002), word alignments in machine translation (Kumar and Byrne, 2003), and sentence alignments for summarization (Lacatusu et al., 2004). Dynamic programming algorithms have been popularly leveraged to produce pairwise and global genetic alignments, where edit distance based metrics are used to compute the cost of insertions, deletions and substitutions. We use dynamic programming to compute the best alignment, given the temporal and coreference information between medical events across these sequences. More importantly, we propose a cascaded WFST-based framework for cross-document temporal ordering of medical event sequences. Composition and search operations can be used to build a single transducer that integrates these components, directly mapping from input states to desired outputs, and obtain the best alignment (Mohri et al., 2000). In natural language processing, WFSTs have seen varied applications in machine translation (Kumar and Byrne, 2003), morphology (Sproat, 2006), named entity recognition (Krstev et al., 2011) and biological sequence alignment / generation (Whelan et al., 2010) among others. We demonstrate that the WFST-based approach outperforms popularly used dynamic programming algorithms for multiple sequence alignment.

3 Problem Description

Medical events are temporally-associated concepts in clinical text that describe a medical condition affecting the patient's health, or procedures performed on a patient. We represent medical events by splitting each event into a start and a stop. When there is insufficient information to discern the start or stop of an event, it is represented as a single concept. If only the start is known then the stop is set to $+\infty$, whereas when only the stop is known , the start is set to the date of birth of the

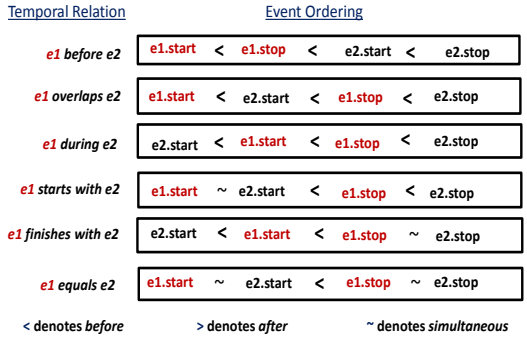

Figure 1: Medical event start / stop representation mapped to Allen's temporal relations (Allen, 1981). Temporal ordering of event starts and stops using $\{before, after, simultaenous\}$ (shown on the right) allows us learn temporal relations between the medical events (shown on the left). $e1_{start} = e2_{start}$ and $e1_{stop} = e2_{stop}$, when $e1$ and $e2$ corefer.

patient.[1] Often, for chronic ailments like *hypertension*, we would only associate a start with the medical event and set the stop to $+\infty$. The start of *hypertension* may be associated with the temporal expression *history of* in the narrative. This, when considered along with the admission date, allows us to relatively order *hypertension* with respect to other medical events. A medical event occurrence like *chest pain* may be associated with a start and a stop, where the start may be determined by the mention of "patient was complaining of chest pain *yesterday*" in the narrative text. Further, the narrative may state that "he *continued* to have chest pain *on admission*, but *currently* he is chest pain free"; this may be used to infer the relative stop of *chest pain*. Medical events may also be instantaneous, for e.g., *injected with antibiotic*. Such events are represented with the start and stop as being the same. Temporal relations exist between the start and stop of events as shown in Figure 1. Learning temporal relations *before, after* and *simultaneous* between the medical event starts and stops corresponds to learning all of Allen's temporal relations (Allen, 1981) between the medical events. Following our previous work (Raghavan et al., 2012c), such a representation allows us to temporally order the event starts and stops within each clinical narrative by learning to rank them in relative order of time. The problem definition is as follows:

[1]Patient date of birth, admission/ discharge date are usually available in the metadata associated with a clinical narrative.

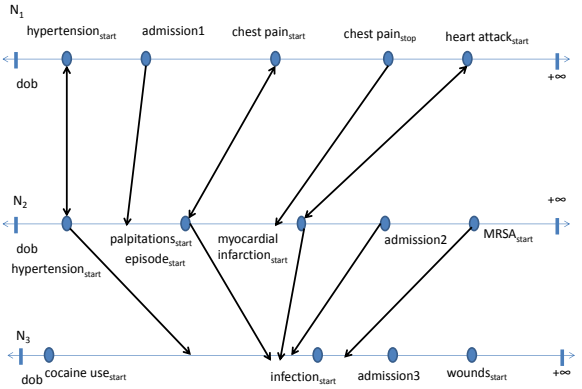

Figure 2: Given temporally ordered medical event sequences, N_1, N_2, N_3, we address the task of combining events across these sequences by merging or ordering them to create a single comprehensive timeline.

Input: Sequences of temporally ordered medical event starts and stops. This corresponds to N_1, N_2, and N_3 in Figure 2. Each sequence corresponds to a clinical narrative. The total number of sequences correspond to the number of clinical narratives for a patient.

Problem: Combine medical events across these sequences to generate a timeline i.e., a single comprehensive sequence of medical events over all clinical narratives of the patient.

Expected Output: In the example shown in Figure 2, the output would be as follows: Timeline (N_1, N_2, N_3)= $\{$cocaine use$_{start}$ $<$ hypertension$_{start}$ $=$ hypertension$_{start}$ $<$ admission1 $<$ chest pain$_{start}$ $\sim$ palpitations$_{start}$ $<$ chest pain$_{stop}$ $<$ heart attack$_{start}$ $=$ myocardial infarction$_{start}$ $<$ admission2 $<$ infection$_{start}$ $<$ MRSA$_{start}$ $<$ admission3 $<$ wounds$_{start}\}$.

The goal of multiple sequence alignment is to find an alignment that maximizes some overall alignment score. Thus, in order to align event sequences, we need to compute scores corresponding to cross-narrative medical event coreference resolution and cross-narrative temporal relations.

4 Cross-Narrative Coreference Resolution and Temporal Relation Learning

The first approach to learning a temporal ordering of medical events across all clinical narratives is to consider all pairs of events across all narratives and learn to classify them as sharing one of Allen's temporal relations (Allen, 1981) using a single learning model. Alternatively, a ranking ap-

proach, similar to the one used to generate intra-narrative temporal ordering, can also be extended to the cross-narrative case. However, the features related to narrative structure and relative and implicit temporal expressions used for temporal ordering within a clinical narrative may not be applicable across narratives. For instance, a history and physical report may have sections like "past medical history", "history of present illness", "assessment and plan", and a certain logical pattern to the flow of text within and across these sections. Further, temporal cues like "thereafter", "subsequently", follow from the context around an event mention. The absence of such features in the cross-narrative case does not allow such a model to generate accurate temporal relation predictions.

Thus, for use in our sequence alignment models, we learn two independent classifiers for medical event coreference and temporal relation learning across narratives. We train a classifier to resolve cross-narrative coreferences by extracting semantic and temporal relatedness feature sets for each pair of medical concepts. Extracting these feature sets helps us train a classifier to predict medical event coreferences (Raghavan et al., 2012a). Another classifier is then trained to classify pairs of medical event starts and stops across narratives as sharing temporal relations {before, after, overlaps}. The learned cross-narrative coreference predictions can then be used along with confident temporal relation predictions to derive a joint probability to enable cross-narrative temporal ordering.

5 Narrative Sequence Alignment for Cross-narrative Temporal Ordering

Sequence alignment algorithms have been developed and popularly used in bioinformatics. However, multiple sequence alignment (MSA) has been shown to be NP complete (Wang and Jiang, 1994) and various heuristic algorithms have been proposed to solve this problem (Notredame, 2002). We propose a novel WFST-based representation that enables accurate decoding for MSA when compared to popularly used dynamic programming algorithms (Needleman et al., 1970, Smith and Waterman, 1981) or other state of the art methods (Do et al., 2012).

In the problem of aligning events across multiple narrative sequences, we want to align temporally ordered medical events corresponding to clinical narratives of a patient. Unlike problems in biological sequence alignment where the symbols to be aligned across sequences are restricted to a fixed set, our symbol set is not fixed or certain because the symbols correspond to medical events in clinical narratives. Moreover, we cannot have fixed scores for symbol transformations since our transformations correspond to coreference and temporal relations between the medical events across sequences. The computation of these scores is described next.

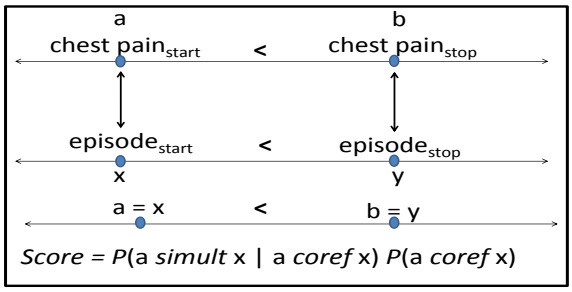

Figure 3: Score computation for aligning events across temporally ordered event sequences chest pain$_{start}$ = episode$_{start}$ < chest pain$_{stop}$ = episode$_{stop}$, where events across the sequences occur simultaneously and corefer.

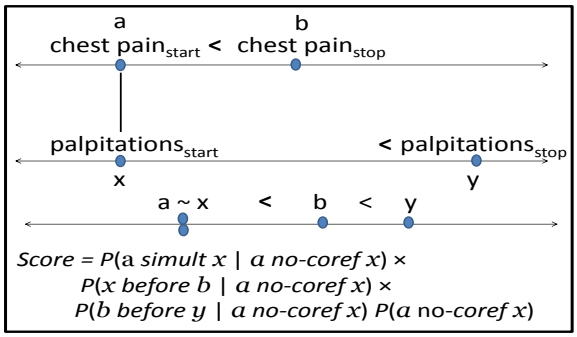

Figure 4: Score computation for aligning events across temporally ordered event sequences chest pain$_{start}$ ∼ palpitations$_{stop}$ < chest pain$_{stop}$ < palpitations$_{stop}$, where some events across the sequences occur simultaneously but do not corefer.

5.1 Scoring Scheme

Let us assume a, b are medical events in the first clinical narrative and have been temporally ordered so $a < b$. Similarly, x, y are medical events in the second clinical narrative such that $x < y$. There exists a match or an alignment between a pair of medical events, across the sequences, in the following cases:

1. If the medical events are simultaneous and coreferring, denoted as $a = x$.
2. If the medical events are simultaneous and non-coreferring, denoted as $a \sim x$.

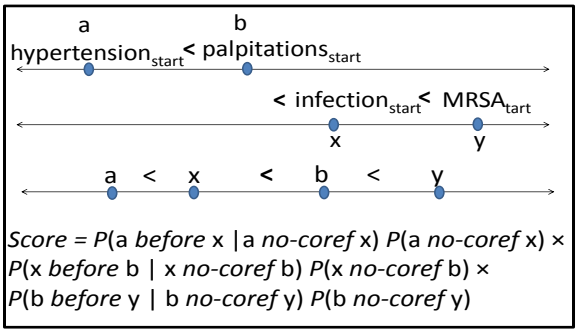

Figure 5: Score computation for aligning events across temporally ordered event sequences hypertension$_{start}$ $<$ palpitations$_{start}$ $<$ infection$_{start}$ $<$ MRSA$_{start}$, where events across the sequences do not occur simultaneously and do not corefer.

3. If the a medical event from one sequence is before a medical event from another sequence, denoted as $a < x$.
4. If the a medical event from one sequence is after a medical event from another sequence, denoted as $a > x$.

We now illustrate how the scores for candidate aligned sequences are computed using the learned cross-narrative coreference and temporal probabilities for the following three scenarios:

- The medical events across sequences are simultaneous and corefer as illustrated in Figure 3. The joint score considers the probability of event temporal relations *simultaneous* conditioned on *coreference*.
- Some medical events across sequences are simultaneous but do not corefer as illustrated in Figure 4. Here, the joint score considers the joint probability of temporal relations *simultaneous* or *before* and *no-coreference*.
- The medical events across sequences are not simultaneous and do not corefer as illustrated in Figure 5. In this case, the joint score considers the probability of the temporal relation *before* and *no coreference*.

Thus, the coreference and temporal relation scores can be leveraged for aligning sequences of medical events. These scores are used in both the WFST-based representation and decoding, as well as for dynamic programming.

5.2 Alignment using a Weighted Finite State Representation

A weighted finite-state transducer (WFST) is an automaton in which each transition between states is associated with an input symbol, an output symbol, and a weight (Mohri et al., 2005). WFSTs can be used to efficiently represent and combine sequences of medical events based coreference and temporal relation information. The WFST representation gives us the ability to talk about the global joint probability derived from coreference and temporal relation scores described in Section 5.1. It allows us to build a weighted lattice of sequences that can be searched for the most probable sequence of medical events from across all clinical narratives of a patient. We use unweighted FSAs to represent the input described in Section 3, i.e. temporally ordered sequences of medical events corresponding to clinical narratives. This corresponds to N_1 and N_2 in Figure 6.

Based on whether we want to align the sequences purely based on coreference scores or both coreference and temporal relation scores, the arc weights for the WFST can be determined. M_{12}^{c} is a WFST that maps input symbols from N_1 to output symbols in N_2 and is weighted by the probability of coreference or no-coreference between medical events across N_1 and N_2. The representation in WFST M_{12}^{c+t} shown in Figure 7 allows us to align N_1 and N_2 based on both coreference as well as temporal relation probabilities. The WFST has ϵ transitions to accommodate insertion and deletion of medical events when combining the sequences. Deletions correspond to the case when an event in the first sequence does not map to any event in the second sequence; similarly insertions correspond to the case where an event in the second sequence does not map to any event in the first sequence. The WFST composition operation allows the outputs of one WFST to be fed to the inputs of a second WFST or FSA. Thus, we build our final machine by composing the three sub-machines as,

$$D = N_1 \circ M_{12}^{i} \circ N_2. \quad (1)$$

where $i = c$ or $i = c + t$. This gives us a combined weighted graph by mapping the output symbols of the first medical event sequence to the input symbols of the second medical event sequence. The scores on the decoding graph are derived from only the coreference probabilities if $i = c$ and both coreference and temporal relation probabilities if $i = c + t$.

In the medical event sequence alignment problem, we want to align multiple sequences of medical events that correspond to multiple clinical narratives of a patient. Since we want to now combine

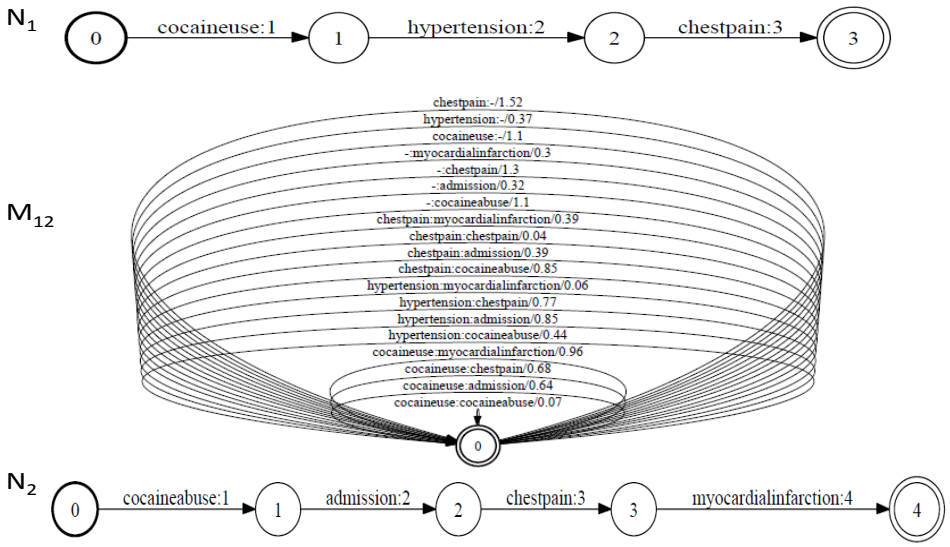

Figure 6: N_1 and N_2 are medical event sequences represented using FSAs. M^c_{12} maps medical events across N_1 and N_2 and is weighted only by the probability of coreference between events across N_1 and N_2.

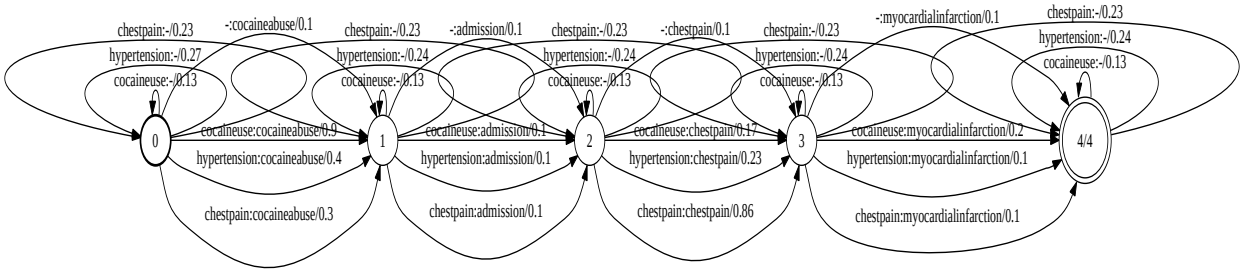

Figure 7: M^{c+t}_{12} is a WFST representation used for mapping medical events between N_1 and N_2 (from Figure 2) and is weighted by both the coreference and temporal relation probabilities

all narrative chains belonging to the same patient, the composition cascade to build the final combined sequence will be as,

$$D_f = N_1 \circ M^i_{12} \circ N_2 \circ M^i_{23} \circ N_3 \circ M^i_{34} ... \circ N_n \quad (2)$$

where $i = c$ or $i = c + t$ and n is the number of medical event sequences corresponding to clinical narratives for a patient. During composition we retain intermediate paths like M^i_{23} utilizing the ability to do lazy composition (Mohri and Pereira, 1998) in order to facilitate beam search through the multi-alignment. The best hypothesis corresponds to the highest scoring path which can be obtained using shortest path algorithms like Djikstra's algorithm. The best path corresponds to the best alignment across all medical event sequences based on the joint probability of cross-narrative medical event coreferences and temporal relations across the narrative sequences.

The complexity of decoding increases exponentially with the number of narrative sequences in the composition, and exact decoding becomes infeasible. One solution to this problem is to do the alignment greedily pairwise, starting from the most recent medical event sequences, finding the best path, and iteratively moving on to the next sequence, and proceeding until the oldest medical event sequence. The disadvantage of such a method is that it does not take into account constraints between medical events across multiple event sequences and may lead to a less accurate solution.

An alternative method is to use lazy composition to perform more efficient composition as it allows practical memory usage. We also use beam search to make for an efficient approximation to the best-path computation (Mohri et al., 2005). This allows accommodating constraints from across multiple sequences and generates a more accurate best path. Thus, this method generates more accurate alignments when we have more than two sequences to be aligned.

For instance, instance say $a, b \in N_1, x, y \in N_2$, and $m, n \in N_3$ are temporally medical event sequences corresponding to narratives N_1, N_2 and N_3. Based on the learned pairwise temporal relations, if we have the following constraints $a < x$, $m > x$, $m < a$. Aligning N_1 and N_2 greedily pairwise may give us the best combined sequence as $a, x, b, y \in N_{12}$. Now in aligning N_{12} with N_3, we won't be able to accommodate $m > x$ and $m < a$. However, performing a beam search over the composed WFST in equation 2 allows us to accommodate such constraints across multiple sequences. The complexity of composing two transducers is $O(V_1 V_2 D_1 (logD_2 + M_2))$ where each edge from the first sequence matches every edge in the second sequence and V_i is the number of states, D_i is the maximum out-degree and M_i maximum multiplicity for the i^{th} FST (Mohri et al., 2005).

We also use popular dynamic programming algorithms (Needleman et al., 1970, Smith and Waterman, 1981) for sequence alignment of medical events across narratives and compare it to the WFST-based representation and decoding.

5.3 Pairwise Alignment using Dynamic Programming

As a contrast, we adapt two dynamic programming algorithms for sequence alignment: global alignment using the Needleman Wunsch algorithm (NW) (Needleman et al., 1970) and local alignment using the Smith-Waterman algorithm (SW) (Smith and Waterman, 1981). NW allows us to align all events in one sequence with all events in another sequence. A drawback of NW is that short and highly similar sequences maybe missed because they get overweighted by the rest of the sequence. NW is suitable when the two sequences are of similar length with significant degree of similarity throughout. On the other hand, SW gives the longest sub-sequence pair that yields maximum degree of similarity between the two original sequences. It does not force all events in a sequence to align with another sequence. SW is useful in aligning sequences that differ in length and have short patches of similarity. The time complexity of these methods for sequences of length m and n are $O(mn)$.

The scoring scheme described earlier is used to update the scoring matrix for dynamic programming. In order to accommodate the temporal relations before and after, we insert a null symbol after every medical event in each sequence in the scoring matrix. A vertical or horizontal gap arises when cases 1, 2, 3 and 4 in Section 5.1 mentioned above are not true. If the medical events are not simultaneous, not before or not after, the medical events will not align. Thus, the value of each cell in the scoring matrix is determined by computing the maximum score at each position $C(i, j)$ as,

$$\max\{(C(i-1, j-1) + S_{ij}), (C(i, j-1) + w), (C(i-1, j) + w)\} \quad (3)$$

where, $S_{ij} = \max\{P(i = j), P(i < j), P(i > j)\}$, and $w = \max\{(1 - P(i = j)), (1 - P(i < j)), (1 - P(i > j))\}$. Here, $C(i-1, j-1)$ corresponds to a match, whereas $C(i, j-1)$ and $C(i-1, j)$ correspond to a gaps in sequence one and two.

In case of the SW algorithm, the negative scoring matrix cells are set to zero, thus making the positively scoring local alignments visible. Backtracking starts at the highest scoring matrix cell and proceeds until a cell with score zero is encountered, yielding the highest scoring local alignment.

The time and space complexity grows exponentially with the number of sequences to be aligned and finding the global optimum has been shown to be a NP-complete problem. The time complexity of aligning N sequences of length L is $O(2^N L^N)$ (Wang and Jiang, 1994). Thus, for MSA using dynamic programming, we use a heuristic method where we combine pairwise alignments iteratively starting with the latest narrative and progressing towards the oldest narrative.

6 Experiments and Evaluation

Corpus Description. The corpus consists of a dataset of clinical narratives obtained from the [redacted] medical center. The corpus has a total of 2060 patients, and 100704 clinical narratives. We gathered a gold standard set of seven patients (80 clinical narratives overall) with manual annotation of all medical events mentioned in the narratives, coreferences, and medical event sequence information. The annotation agreement across annotators is high, with 89.5% agreement corresponding to inter-annotator Cohen's kappa statistic of 0.86 (Raghavan et al., 2012b). The types of clinical narratives included 27 discharge summaries, 30 history and physical reports, 15 radiology reports and 8 pathology reports. The distribution of the number of medical event sequences and unique medical events across patients is shown in Table 1. The annotated dataset is used to cross-validate and train our coreference and temporal relation learning models and to evaluate our cross-narrative medical event timeline.

	p1	p2	p3	p4	p5	p6	p7	
No. of Narrative Sequences	5	9	20	13	8	10	15	
No. of Medical events	68	90	119	82	79	72	95	
	% Accuracy							% Avg.
WFST-framework (lazy composition and beam search)[c+t]	**76.1**	**73.2**	**81.2**	**83.5**	**76.4**	**82.5**	**79.7**	78.9
WFST-framework (Iterative pairwise)[c+t]	70.4	67.1	73.5	74.1	61.8	75.5	62.9	69.3
Smith Waterman (Iterative pairwise)[c+t]	71.2	69.7	75.5	75.6	66.3	77.4	68.3	72.1
Needleman-Wunsch (Iterative pairwise)[c+t]	68.1	66.3	72.1	74.4	61.1	75.5	63.6	68.7
WFST-framework (lazy composition and beam search)[c]	**68.5**	**65.3**	**72.3**	**74.4**	**67.2**	**71.3**	**69.1**	69.7
WFST-framework (Iterative pairwise)[c]	61.2	63.3	61.9	60.4	59.8	64.8	60.5	61.7
Smith Waterman (Iterative pairwise)[c]	60.3	63.7	68.2	62.3	58.6	66.7	60.2	62.8
Needleman-Wunsch (Iterative pairwise)[c]	56.6	60.1	59.3	65.6	54.7	63.1	58.2	59.6

Table 1: The distribution of medical events across narrative sequences and sequences across patients and multiple sequence alignment results for the WFST-based framework, and dynamic programming using just coreference scores [c] and using coreference as well as temporal relation scores [c+t].

Evaluation Metric. For each patient and each method (WFST or dynamic programming), the output timeline to evaluate is the highest scoring candidate hypothesis derived as described above. Accuracy of the timeline is calculated as the number of transformations required to obtain the reference sequence in the annotated gold-standard from the one generated by our system. Transformations are measured in terms of the minimum edit distance, insertions, deletions, and substitutions of medical events.

Experiments and Results. We first temporally order medical events within each clinical narrative by learning to rank them in relative order of occurence as described in our previous work (Raghavan et al., 2012c). The overall accuracy of ranking medical events using leave-one-out cross validation is 82.1%. The resulting medical event sequences serve as the input to the problem of cross-narrative sequence alignment.

The cross-narrative coreference and temporal relation pairwise classification models described in Section 4 are trained using a Maximum entropy classifier. The coreference resolution performs with 71.5% precision and 82.3% recall. The temporal relation classifier performs with 60.2% precision and 76.3% recall. The learned pairwise coreference and temporal relation probabilities are now used to derive the score for the WFST and dynamic programming approaches.

WFST representation and decoding. We build finite-state machines using the open source OpenFST library.[2] We use a tropical semi-ring weighted using the negative log-likelihood of the computed scores. OpenFST provides tools that can search for the highest scoring sequences accepted by the machine, and can sample from high-scoring sequences probabilistically, by treating the scores of each transition within the machine as a negative log probability. The decoding process to compute the most likely combined medical event sequence can be defined as searching for the best path in the combined graph representation (Equation 2). The best path is the one that minimizes the total weight on a path (since the arcs are negative log probabilities). In searching for the best path, the beam size is set to 5. The accuracy of the WFST-based representation and beam search across all sequences using the coreference and temporal relation scores to obtain the combined aligned sequence is 78.9%.

Dynamic Programming. We use the NW and SW algorithms described in Section 5.3 to produce local and global alignments respectively. We use the scoring scheme described in Section 5.1 to update the cost matrix for dynamic programming and implement the algorithms as described in Section 5.3. The overall accuracy of sequence alignment with both coreference and temporal relation scores using NW is 68.7% whereas SW gives an accuracy of 72.1%. In case of aligning just two sequences, both methods yield the same results. The accuracy of cross-narrative MSA for each patient, for each method, using cross validation, is shown in Table 1. Results indicate that the WFST-based method outperforms the dynamic programming approach for multi-sequence alignment (statistical significance $p<0.05$). Morever, the results using both coreference and temporal realtion scores for alignment outperform using only coreference scores for alignment using all approaches. This indicates that cross-narrative temporal relations are important for accurately aligning medical event sequences across narratives.

7 Discussion

We propose and evaluate different approaches to multiple sequence alignment of medical events.

[2] www.openfst.org

Approaches to multi-alignment. We address the problem of aligning medical event sequences using a novel WFST-based framework and empirically demonstrate that it outperforms pairwise progressive alignment using dynamic programming. This is mainly because the WFST-based allows us to consider temporal constraints from across multiple sequences when performing the alignment.

Moreover, it also outperforms the integer linear programming (ILP) method for timeline construction proposed in (Do et al., 2012). We implemented the proposed method that also allows combining the output of classifiers subject to some constraints. We derive intervals from event starts and stops and learn two perceptron classifiers for classifying the temporal relations between events and assigning events to intervals. The classifier probabilities are then used to solve the optimization problem using the lpsolve solver.[3] We also use intra-document coreference information to resolve coreference before performing the global optimization. We observe that in case of MSA, the optimal solution using ILP is still intractable as the number of constraints increases exponentially with the number of sequences. Aligning pairwise iteratively gives us an overall average accuracy of 68.2% similar to dynamic programming. While this is comparable to the dynamic programming performance, the WFST-based method significantly outperforms this in case of multi-alignments for cross-narrative temporal ordering.

Performance and error analysis. We perform multi-alignments over medical event sequences for a patient, where each sequence corresponds to temporally ordered medical events in a clinical narrative generated using the ranking model described in (Raghavan et al., 2012c). The accuracy of intra-narrative temporal ordering is 82.1%. The errors in performing this intra-narrative ordering may propagate to the cross-narrative model resulting in reduced accuracy. This may be addressed by considering n-best temporally ordered medical event sequences, generated by the ranking process, and aligning the n-best sequences using the WFST-based framework. This could be feasible as, practically, the WFST-based method for multi-alignment takes only a few secs to align a pair of medical event sequences with average length 40.

The accuracy of alignments across multiple medical event sequences is also affected by the error induced by the coreference and temporal relation scores. Often, insufficient temporal cues leads to misclassification of events incorrectly as sharing the "simultaneous" temporal relation and often as coreferring. This induces errors in the score calculation and hence the alignments. Better methods to address the challenging problem of cross-document temporal relation learning, perhaps with the help of structured data from the patient record, could improve the accuracy of alignments.

There is no clear trend with respect to the number of medical events and narratives for a patient (Table 1.), and the alignment accuracy. In future work, it would be interesting to examine any such correlation and also study the scalability of the WFST-based method for sequence alignment on longer medical event sequences and a larger dataset of patients. Further, the WFST-based method may be used to model multi-alignment tasks in other speech and language problems as well.

8 Conclusion

We propose a novel framework for aligning medical event sequences across clinical narratives based on coreference and temporal relation information using cascaded WFSTs. FSTs provide a convenient and flexible framework to model sequences of temporally ordered medical events and compose them into a combined graph representation. Decoding this graph allows us to jointly maximize coreference as well as temporal relation probabilities to derive a timeline of the most likely temporal ordering of medical events. This approach to aligning multiple sequences of medical events significantly outperforms other approaches such as dynamic programming. Moreover, we demonstrate the importance of learning temporal relations for the task timeline generation from across multiple clinical narratives by empirically proving that decoding using both coreference and temporal relation scores is far more accurate than decoding with only coreference scores.

Acknowledgments

The project was supported by Award Number Grant R01LM011116 from the National Library of Medicine. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Library of Medicine or the National Institutes of Health. The authors would like to thank Yanzhang He for his input on the WFST-based model.

[3] http://lpsolve.sourceforge.net/5.5/

References

James F. Allen. 1981. An interval-based representation of temporal knowledge. In *IJCAI*, pages 221–226.

Regina Barzilay and Kathleen R. McKeown. 2005. Sentence fusion for multidocument news summarization. *Comput. Linguist.*, 31(3):297–328, September.

Regina Barzilay, Noemie Elhadad, and Kathleen McKeown. 2002. Inferring strategies for sentence ordering in multidocument summarization. *Journal of Artificial Intelligence Research (JAIR)*, 17:35–55.

Danushka Bollegala, Naoaki Okazaki, and Mitsuru Ishizuka. 2010. A bottom-up approach to sentence ordering for multi-document summarization. *Information processing & management*, 46(1):89–109.

Philip Bramsen, Pawan Deshpande, Yoong Keok Lee, and Regina Barzilay. 2006. Inducing temporal graphs. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 189–198.

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *ACL/AFNLP*, pages 602–610.

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(1):10.

Dina Demner-Fushman, Wendy Webber Chapman, and Clement J. McDonald. 2009. What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics*, 42(5):760–772.

Quang Xuan Do, Wei Lu, and Dan Roth. 2012. Joint inference for event timeline construction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 677–687. Association for Computational Linguistics.

Heng Ji and Ralph Grishman. 2008. Refining event extraction through cross-document inference. In *Association for Computational Linguistics*.

Cvetana Krstev, Duško Vitas, Ivan Obradović, and Miloš Utvić. 2011. E-dictionaries and Finite-state automata for the recognition of named entities. In *Proceedings of the 9th International Workshop on Finite State Methods and Natural Language Processing*, pages 48–56.

Shankar Kumar and William Byrne. 2003. A weighted finite state transducer implementation of the alignment template model for statistical machine translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, pages 63–70.

V Finley Lacatusu, Steven J Maiorano, and Sanda M Harabagiu. 2004. Multi-document summarization using multiple-sequence alignment. In *LREC*.

Mirella Lapata and Alex Lascarides. 2011. Learning sentence-internal temporal relations. *CoRR*, abs/1110.1394.

Mirella Lapata. 2003. Probabilistic text structuring: Experiments with sentence ordering. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 545–552. Association for Computational Linguistics.

Zhihui Luo, Stephen B. Johnson, Albert M. Lai, and Chunhua Weng. 2011. Extracting temporal constraints from clinical research eligibility criteria using conditional random fields. In *Proc of AMIA Symposium*.

Inderjeet Mani, Marc Verhagen, Ben Wellner, Chong Min Lee, and James Pustejovsky. 2006. Machine learning of temporal relations. In *ACL*.

Mehryar Mohri and Fernando CN Pereira. 1998. Dynamic compilation of weighted context-free grammars. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 2*, pages 891–897. Association for Computational Linguistics.

Mehryar Mohri, Fernando C. N. Pereira, and Michael Riley. 2000. The design principles of a weighted finite-state transducer library. *Theoretical Computer Science*, 231(1):17–32.

Mehryar Mohri, Fernando Pereira, and Michael Riley. 2005. Weighted automata in text and speech processing. *CoRR*, abs/cs/0503077.

S.B. Needleman, C.D. Wunsch, et al. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453.

Cédric Notredame. 2002. Recent progress in multiple sequence alignment: a survey. *Pharmacogenomics*, 3(1):131–144.

Preethi Raghavan, Eric Fosler-Lussier, and Albert M. Lai. 2012a. Exploring semi-supervised coreference resolution of medical concepts using semantic and temporal features. In *North American Association for Computational Linguistics Annual Meeting - Human Language Technologies Conference*. Association for Computational Linguistics.

Preethi Raghavan, Eric Fosler-Lussier, and Albert M. Lai. 2012b. Inter-annotator reliability of medical events, coreferences and temporal relations in clinical narratives by annotators with varying levels of clinical expertise. In *To appear in Proceedings*

of the American Medical Informatics Association. American Medical Informatics Association.

Preethi Raghavan, Eric Fosler-Lussier, and Albert M. Lai. 2012c. Learning to temporally order medical events in clinical text. In *ACL short paper*. Association for Computational Linguistics.

Daniel Reichert, David Kaufman, Benjamin Bloxham, Herbert Chase, and Noémie Elhadad. 2010. Cognitive analysis of the summarization of longitudinal patient records. In *AMIA Annual Symposium Proceedings*, volume 2010, page 667. American Medical Informatics Association.

A. Roberts, R. Gaizauskas, M. Hepple, G. Demetriou, Y. Guo, and A. Setzer. 2008. Semantic Annotation of Clinical Text: The CLEF Corpus. In *Proceedings of the LREC 2008 Workshop on Building and Evaluating Resources for Biomedical Text Mining*, pages 19–26.

T.F. Smith and M.S. Waterman. 1981. Identification of common molecular subsequences. *Journal of molecular biology*, 147(1).

Richard Sproat. 2006. *A Computational Theory of Writing Systems (Studies in Natural Language Processing)*. Cambridge University Press.

Marc Verhagen, Robert J. Gaizauskas, Frank Schilder, Mark Hepple, Jessica Moszkowicz, and James Pustejovsky. 2009. The tempeval challenge: identifying temporal relations in text. *Language Resources and Evaluation*, 43(2):161–179.

Lusheng Wang and Tao Jiang. 1994. On the complexity of multiple sequence alignment. *Journal of computational biology*, 1(4):337–348.

Christopher Whelan, Brian Roark, and Kemal Sonmez. 2010. Designing antimicrobial peptides with weighted finite-state transducers. In *Proceedings of IEEE Engineering in Medical Biology Society*, page 764.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Association for Computational Linguistics*, pages 189–196.

Li Zhou, Genevieve B. Melton, Simon Parsons, and George Hripcsak. 2006. A temporal constraint structure for extracting temporal information from clinical narrative. *Journal of Biomedical Informatics*, pages 424–439.