Annotating Inter-Sentence Temporal Relations in Clinical Notes

Jennifer D'Souza, Vincent Ng

Human Language Technology Research Institute
University of Texas at Dallas
Richardson, TX 75083-0688
{jld082000,vince}@hlt.utdallas.edu

Abstract

Owing in part to the surge of interest in temporal relation extraction, a number of datasets manually annotated with temporal relations between event-event pairs and event-time pairs have been produced recently. However, it is not uncommon to find missing annotations in these manually annotated datasets. Many researchers attributed this problem to ”annotator fatigue”. While some of these missing relations can be recovered automatically, many of them cannot. Our goals in this paper are to (1) manually annotate certain types of missing links that cannot be automatically recovered in the i2b2 Clinical Temporal Relations Challenge Corpus, one of the recently released evaluation corpora for temporal relation extraction; and (2) empirically determine the usefulness of these additional annotations. We will make our annotations publicly available, in hopes of enabling a more accurate evaluation of temporal relation extraction systems.

Keywords: temporal relations, predicate-argument relations, discourse relations

1. Introduction

Recent years have seen a surge of interest in temporal information extraction (IE) in the natural language processing community. In order to facilitate corpus-based approaches to temporal relation extraction, several datasets manually annotated with temporal relations were produced in the past decade, including the TimeBank corpus (Pustejovsky et al., 2003), as well as those produced as part of TempEval-1 (Verhagen et al., 2007), TempEval-2 (Verhagen et al., 2010), and the 2012 i2b2 shared task on temporal relation extraction (Sun et al., 2013).

Manually annotating documents with temporal relations is a fairly sophisticated task, however. Many researchers, including the organizers of the aforementioned shared tasks, have noticed that it is not uncommon to find missing temporal relations in these manually annotated documents. In other words, many temporal relations that must exist between events are in fact not annotated in these documents. Verhagen (2005), Verhagen et al. (2006), and UzZaman et al. (2012), among others, attribute the reason behind this ”missing annotation” problem to annotator fatigue: since the number of temporal relations present in a document, especially one that is long, can potentially be large, it is possible for an annotator to miss some of them during the annotation process. The presence of missing links presents problems to both the training and evaluation of temporal relation extraction systems: not only do they cause a temporal relation extraction system to be unfairly evaluated, but they also cause systems to be trained on instances with incorrect temporal relation labels.

Fortunately, some of these missing temporal links can be recovered automatically. For example, if a human annotator determines that A occurs before B and B occurs before C, then even if she forgets to annotate A as occurring before C, we can automatically recover such missing link by transitivity. On the other hand, there are many cases where missing links cannot be automatically recovered (see Section 3 for details).

Our goal in this paper is to manually annotate certain types of missing links that cannot be automatically recovered in the i2b2 Clinical Temporal Relations Challenge Corpus (henceforth i2b2 corpus), one of the recently released evaluation corpora for temporal relation extraction. To determine the usefulness of our annotations, we train and evaluate two temporal relation extraction systems, one on only the existing annotated temporal relations, and the other on those augmented with our new annotations. In addition, to enable accurate evaluation of temporal relation extraction systems on this corpus, we will make our annotations publicly available. [1]

2. Corpus

For annotation, we use the i2b2 corpus, which consists of 310 de-identified discharge summaries pre-partitioned into a training set (190 summaries) and a test set (120 summaries). The original corpus is marked up with annotations for event expressions, time expressions, and temporal relations between some of the event pairs and event-time pairs. Being outside the scope of this work, details on time expressions and temporal relations between event-time pairs is excluded from further discussion. In the i2b2 corpus, an event refers to clinically relevant patient-related actions, and contains various attributes, including the type of event [2] , polarity, and modality. For the event pairs that are annotated with temporal relations, the temporal order of the events in the pair is reflected by the *type* of the temporal relation they are assigned. The i2b2 corpus has 12 relation types defined in all. They are **Simultaneous**, **Overlap**,

[1] Downloadable from `http://www.hlt.utdallas.edu/~jld082000/temporal-relations/`

[2] Six types of events are defined, including TEST (e.g., *CT scan*), PROBLEM (e.g., *the tumor*), TREATMENT (e.g., *operation*), CLINICAL DEPARTMENTS (e.g., *ICU*), EVIDENTIAL information (e.g., *complained*), and clinically relevant OCCURRENCE (e.g., *discharge*).

Before, **After**, **Before_Overlap**, **Overlap_After**, **During**, **During_Inv**, **Begins**, **Begun_By**, **Ends**, and **Ended_By**.

	% Temporal Relation Annotations	
Corpus	Intrasentence	Intersentence
Training data	64.4%	35.6%
Test data	61.2%	38.8%

Table 1: i2b2 corpus annotation percentages of intrasentence and intersentence temporal relations

These event pairs chosen for annotating with temporal relations can either be intrasentential or intersentential. Table 1 provides a breakdown of the percentage of annotated temporal relations in the i2b2 corpus for intrasentential event pairs versus for intersentential event pairs. From these percentages, we see that the number of annotations for intersentence event pairs is much lower than the number of annotations made for intrasentence event pairs. This sparsity in intersentence links is a problem because it creates gaps in a patient's clinical timeline. As an example, consider:

> [sample1.txt]
> [$_{\text{event}}$*Cardiac catheterization*] done on 01-07 showed [$_{\text{event}}$*an aortic valve area*] of .38 cm.sq. , with [$_{\text{event}}$*a mean gradient*] of 62 . [$_{\text{event}}$*PA pressure*] 48/24 , [$_{\text{event}}$*1+ mitral regurgitation*] , and [$_{\text{event}}$*an ejection fraction*] of 43% .

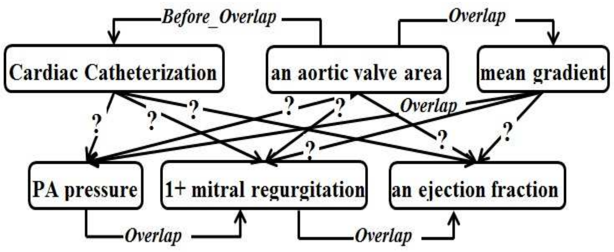

Figure 1: Timeline of events constructed from sample.txt

From Figure 1 we observe that the timeline created from existing temporal relation annotations has many missing links, particularly intersentential ones. Since the events of type TEST discussed in the sample1.txt are all conducted together, roughly overlapping in time, it is important for a medical timeline constructed from these events to not leave out this vital piece of information. Such instances of missing temporal relations can cause gaps in constructing a timeline of clinical events. Therefore, the main aim of this work is to fill in the gaps that arise due to missing intersentence temporal relation annotations by providing human annotations for missing unrecoverable links between event pairs from adjacent sentences.

3. Transitive Closure

A discharge summary has temporal relations expressed in the following two ways: 1) explicit relations that are annotated in the document; and 2) implicit relations that are not annotated in the document, but can be generated by computing a transitive closure over the explicit temporal relations. Computing transitive closure over a collection of related event pairs entails applying a set of transitivity rules that have originated from Allen's (1983) interval algebra. In the closure process, the satisfaction of any rule results in a new relation implied by the rule. As an illustrative example, consider the events from sample2.txt represented in Figure 2. In this example, we are given that event e_1 *his overall status* is **Before** event e_2 *transferred* and that e_2 *transferred* **Begins** event e_3 *the floor*. So even though we are not given any explicit relation annotation between e_1 *his overall status* and e_3 *the floor*, their temporal relation can be inferred as **Before** by applying the following transitivity rule: $((e_1 \mathbf{Before} e_2) \wedge (e_2 \mathbf{Begins} e_3)) \implies (e_1 \mathbf{Before} e_3)$.

> [sample2.txt]
> His oxygen requirement decreased, and [$_{\text{event1}}$*his overall status*] improved quickly . He is being [$_{\text{event2}}$*transferred*] to [$_{\text{event3}}$*the floor*] for ongoing care .

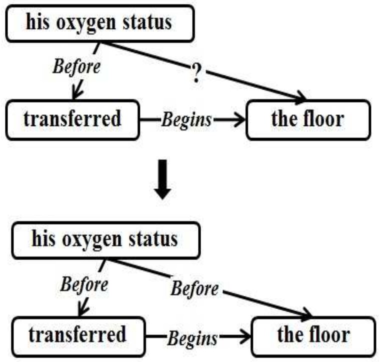

Figure 2: Example showing automatic temporal relation inference.

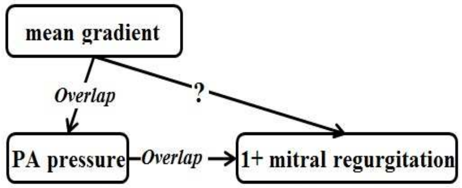

Figure 3: Example where automatic temporal relation inference is not possible.

However, automatic inference of temporal relations that must exist between events but which are missing from annotations is not always possible. Let us reconsider some of the event pairs from Figure 1, as re-depicted in Figure 3. In the figure, we see that event *mean gradient* of type TEST **Overlaps** with another TEST event *PA pressure*, event *PA pressure* in turn **Overlaps** with event *1+ mitral regurgitation* which is also of type TEST, and that there is a missing link between events *mean gradient* and *1+ mitral regurgitation*. Unlike in the previous example, however, in this case there are no applicable rules of transitivity to enable inferring the relation between event pair (e_1,e_3). Such situations call for human intervention in making the correct choice of applicable relation from those present in the relations' set. Therefore, in this work we rely on human judgement to annotate the dataset with inter-sentence missing links that must be present but aren't, and which are also not automatically recoverable from existing explicit annotations.

4. Annotation of Temporal Relations

4.1. Annotation Guidelines

To annotate these missing links, we must provide guidelines for identifying *eligible* links. The original i2b2 corpus annotation guidelines describes *eligibility* as when the paired events have an explicit temporal influence on each other in a patient's medical timeline. In this setting, annotators are called to exercise two levels of judgement: 1) decide if the paired events influence each other medically; and if they do, then 2) choose from the given set of temporal relations the one applicable to the pair. We recognize that this annotation setting is susceptible to annotation inconsistencies, especially in a multiple annotator setting, on both levels of judgement: 1) in deciding *eligibility*; and 2) in selecting temporal relation type. Our guidelines differ from the original i2b2 corpus guidelines in that instead of using the more generic description of *eligibility*, we provide annotators with 10 cases within which the original eligibility statement is true qualifying an event pair that is classified into 1 of the 10 cases as *eligible* to have a relation. This aids in alleviating annotation inconsistencies arising from the first level of judgements required to be made when using the generic description of *eligibility*. Once identified as *eligible,* temporal relation type choice is based on annotator judgement.

Next, we describe the 10 criteria for *eligibility*. For each event pair (e_1,e_2) where e_1 and e_2 are from adjacent sentences in the text, the event pair has an explicit temporal relation if it satisfies any one of the following cases:

Case 1

Prior to and surrounding admission events:

is true when,

- one of the events is an admission OCCURRENCE, and the other event is a TREATMENT/TEST event that happened before the patient's admission; or
- one of the events is an admission OCCURRENCE for the purpose of the other event which is a TREATMENT; or
- one of the events is an admission OCCURRENCE because of a PROBLEM which is the other event; or
- one of the events is a PROBLEM from the patient's past and the other event is a PROBLEM because of which the patient is admitted.

consider passages,

...with [$_{\text{event}}$*metastatic cervical cancer*] [$_{\text{event}}$*admitted*] with a question of [$_{\text{event}}$*malignant pericardial effusion*] .
Patient underwent [$_{\text{event}}$*a total abdominal hysterectomy*] in the past for a 4x3.6x2 cm [$_{\text{event}}$*cervical mass*] ...

eligible pair: (*admitted*, *a total abdominal hysterectomy*)

The patient was [$_{\text{event}}$*admitted*] to [$_{\text{event}}$*the Tau Memorial Hospital*] .
She was seen by Dr. Freiermthalskush of [$_{\text{event}}$*the renal service*] for [$_{\text{event}}$*management*] of her [$_{\text{event}}$*chronic renal insufficiency*] .

eligible pairs: (*admitted*, *management*), (*admitted*, *chronic renal insufficiency*)

...woman who is [$_{\text{event}}$*HIV positive*] for two years .
She [$_{\text{event}}$*presented*] with [$_{\text{event}}$*left upper quadrant pain*] as well as [$_{\text{event}}$*nausea*] and [$_{\text{event}}$*vomiting*] ...

eligible pairs: (*HIV positive*, *left upper quadrant pain*), (*HIV positive*, *nausea*), (*HIV positive*, *vomiting*)

Case 2

Coreferent events:

is true when,

- both events are actually the same event just referred to by distinct phrases, or by the same phrase in two different places.

consider passage,

...his usual monthly change of [$_{\text{event}}$*his suprapubic catheter*] and felt [$_{\text{event}}$*discomfort*] ...
...he was [$_{\text{event}}$*evaluated*] by Urology and had [$_{\text{event}}$*his catheter*] changed .

eligible pair: (*his suprapubic catheter*, *his catheter*)

Case 3

Recurrent events:

is true when,

- one of the events is a recurrence of the other event.

consider passage,

She has complaints of ...[$_{\text{event}}$*left upper quadrant pain*] on and off getting progressively worse ...
She has had [$_{\text{event}}$*similar pain*] intermittently for last year .

eligible pair: (*left upper quadrant pain*, *similar pain*)

	Relation	Number of Intra-Sentence Instances							
		Training Data				Test Data			
		Original	After Annotation	Additional Links	% Additional links	Original	After Annotation	Additional Links	% Additional links
1	**Simultaneous**	435	634	199	45.7%	353	560	207	58.6%
2	**Overlap**	335	2715	2380	710.4%	188	1573	1385	736.7%
3	**Before**	65	2189	2124	3267.7%	44	1548	1504	3418.2%
4	**After**	10	89	79	790%	6	52	46	766.7%
5	**Before_Overlap**	74	1727	1653	2233.8%	70	1385	1315	1878.6%
6	**Overlap_After**	0	83	83	-	0	90	90	-
7	**During**	15	166	151	1006.7%	11	150	139	1263.6%
8	**During_Inv**	0	96	96	-	0	65	65	-
9	**Begins**	0	31	31	-	0	19	19	-
10	**Begun_By**	9	12	3	33.3%	7	13	6	85.7%
11	**Ends**	9	12	3	33.3%	7	13	6	85.7%
12	**Ended_By**	49	97	48	98.0%	48	97	49	102.1%

Table 2: Number and Percentages of additional links added to the training and test set of the i2b2 corpus.

Case 4

Sequence of events:

is true when,

- both events are either TREATMENT or TEST which are administered in an asynchronous sequence.
- both events are CLINICAL DEPARTMENTs which the patient visited at different times.
- both events refer to specific admission, discharge, or transference OCCURRENCE with at least one of them being a verb.

consider passages,

She underwent [$_{\text{event}}$*pre-operative testing*] along with [$_{\text{event}}$*carotid ultrasound*] .
CNIS revealed carotid stenosis and she ultimately underwent [$_{\text{event}}$*left carotid stenting*] by [$_{\text{event}}$*vascular surgery*] on 10-30 .

eligible pairs: (*carotid ultrasound*, *left carotid stenting*), (*carotid ultrasound,vascular surgery*)

While in [$_{\text{event}}$*Oaksgekesser/ Memorial Hospital*] the patient was never able to get out of bed . . .
. . . when he is [$_{\text{event}}$*transferred*] to [$_{\text{event}}$*a rehab facility*] that he will continue with [$_{\text{event}}$*physical therapy*] . . .

eligible pair: (*Oaksgekesser/ Memorial Hospital*, *a rehab facility*)

He was most recently [$_{\text{event}}$*discharged*] from [$_{\text{event}}$*Oaksgekesser/ Memorial Hospital*] on 03/06/99 and was then [$_{\text{event}}$*transferred*] to [$_{\text{event}}$*Linghs County Medical Center*] . . .
. . . doing relatively [$_{\text{event}}$*well*] at home until one day prior to [$_{\text{event}}$*admission*] he complained [$_{\text{event}}$*discomfort*] . . .

eligible pairs: (*discharged*, *admission*), (*transferred*, *admission*)

Case 5

PROBLEM causing events:

is true when,

- one of the events is an OCCURRENCE because of the other event of type PROBLEM.
- one of the events is a TEST conducted because of the other event of type PROBLEM.
- one of the events is a TREATMENT administered because of the other event of type PROBLEM.

consider passages,

She was [$_{\text{event}}$*noted*] , on 06/16 , to have [$_{\text{event}}$*numerous erythematous maculopapules*] . . . [$_{\text{event}}$*Dermatology*] was [$_{\text{event}}$*consulted*] and they felt that this was most likely [$_{\text{event}}$*steroid acne*] .

eligible pair: (*numerous erythematous maculopapules*, *consulted*)

[$_{\text{event}}$*The patient's shortness of breath*] and [$_{\text{event}}$*wheezing*] continued but without change . [$_{\text{event}}$*Her cardiac examination*] remained the same and there continued to be no evidence of [$_{\text{event}}$*tamponade*] .

eligible pairs: (*The patient's shortness of breath*, *Her cardiac examination*), (*wheezing*, *Her cardiac examination*)

She described [$_{\text{event}}$*the pain*] as a burning pain which is positional . . .
She has no relief from *antacids* or *H2 blockers* .

eligible pairs: (*the pain*, *antacids*), (*the pain*, *H2 blockers*)

Case 6

TEST reveals PROBLEM:

is true when,

- one of the events is a TEST which reveals the other event a PROBLEM that the patient faces.

consider passage,

[$_{\text{event}}$*A follow-up CT scan*] was done which did not [$_{\text{event}}$*show*] any evidence for [$_{\text{event}}$*splenomegaly*] . . . [$_{\text{event}}$*The 1 cm cyst*] which was seen in 10/92 was still present .

eligible pair: (*A follow-up CT scan*, *The 1 cm cyst*)

Case 7

IS-A relation between events:

is true when,

- both events are PROBLEMs or TREATMENTs or TESTs in a hierarchical IS-A relation.

consider passages,

Pathology revealed [$_{\text{event}}$*poorly differentiated squamous cell carcinoma of the cervix*] and [$_{\text{event}}$*metastatic squamous cell carcinoma in the cardinal ligaments*] with [$_{\text{event}}$*extensive lymphatic invasion*] .
Patient was felt to have [$_{\text{event}}$*stage 2B disease*] and post-operatively , she was treated . . .

eligible pairs: (*poorly differentiated squamous cell carcinoma of the cervix*, *stage 2B disease*), (*metastatic squamous cell carcinoma in the cardinal ligaments*, *stage 2B disease*), (*extensive lymphatic invasion*, *stage 2B disease*)

. . . where he continued the rest of [$_{\text{event}}$*his treatment*] .
This consisted of continued [$_{\text{event}}$*nebulizer treatments*] as well as [$_{\text{event}}$*intravenous antibiotics*] and [$_{\text{event}}$*intravenous Solu-Medrol*] .

eligible pairs: (*his treatment*, *nebulizer treatments*), (*his treatment*, *intravenous antibiotics*), (*his treatment*, *intravenous Solu-Medrol*)

Case 8

TREATMENT/TEST/OCCURRENCE events in a CLINICAL_DEPARTMENT:

is true when,

- one of the events is a TREATMENT/TEST/OCCURRENCE that happens in a CLINICAL_DEPARTMENT which is the other event.

consider passage,

He stayed in [$_{\text{event}}$*the unit*] for about one day .
During that period of time he received [$_{\text{event}}$*nebulizer treatments*] with [$_{\text{event}}$*Albuterol*] .

eligible pairs: (*the unit*, *nebulizer treatments*), (*the unit*, *Albuterol*)

Case 9

Event persists during another event:

is true when,

- one of the events persists for the duration of the other event.

consider passage,

. . . towards the last 2-3 days of [$_{\text{event}}$*his hospitalization*] the amount of improvement was minimal .
It was felt that [$_{\text{event}}$*his respiratory status*] has been [$_{\text{event}}$*maximally treated*] by then .

eligible pair: (*his hospitalization*, *maximally treated*)

Case 10

EVIDENTIAL nature of paired events:

is true when,

- one of the events is a TEST whose EVIDENTIAL function is the other event; or
- both are EVIDENTIAL events from separate sentences where the events in one sentence causally influence the other sentence events.

consider passage,

She underwent usual [$_{\text{event}}$*pre-operative testing*] along with [$_{\text{event}}$*carotid ultrasound*] . [$_{\text{event}}$*CNIS*] [$_{\text{event}}$*revealed*] [$_{\text{event}}$*carotid stenosis*] . . .

eligible pair: (*carotid ultrasound*, *revealed*)

4.2. Annotation Tool

In order to annotate the dataset, annotators were provided with a Java-based annotation tool. [3] The interface of the tool presented them with the text of the discharge report, marked up event expressions that were highlighted in a different color for ease of identification, and existing annotations for temporal relations in the report. From the tool the annotators had two views of the data which were split horizontally, with the discharge report text along with marked up event expressions on the top half of the screen, and TLINK information between events on the bottom half

[3] This tool was made available as part of the resources for the 2012 i2b2 Temporal Relations Challenge by the challenge organizers.

of the screen. In order to annotate a temporal relation, two events needed to be clicked on, after which a dialog prompt appears to select a relation to assign. This newly created relation then gets added to the set of existing TLINKs on the bottom half of the screen. Thus annotators could annotate the temporal relation between two events with just 3 mouse clicks.

4.3. Agreement on Annotation Scheme

Table 2 shows the number of additional links that were added to the training and the test dataset in the i2b2 corpus. While the entire dataset is annotated by one annotator, to evaluate the annotation scheme, 50 randomly chosen training set files were annotated by a second annotator. We then compute two types of agreement based on Cohen's Kappa (Carletta, 1996): agreement on which events to pair across adjacent sentences, and agreement on both pairing events and assigning a temporal relation label to the event pair. We obtain a fairly high inter-annotator agreement on linking events of 0.82, and a slightly lower agreement of 0.65 on linking and assigning relation labels. However, considering that ours is a 12-class relation annotation scheme, a score of 0.65 suggests that the annotators judgement is fairly agreeable despite the fine-grained nature of this annotation task. Our hypothesis is that providing annotators with specific guidelines (elicited in Section 4) for linking events has a significant contribution in the next stage of annotation which involves choosing a relation. In order to gain insights into the difficult annotation classes as faced by the annotators, we obtain class-based inter-annotator agreement scores (shown in Table 3). From the numbers in Table 3, we see that the classes where annotators differed most in agreement were classes which were temporally overlapping in nature such as **Overlap**, **Before_Overlap**, and **Overlap_After**. This can be in part attributable to differences in perception of overlapping events between annotators. For example, if a treatment administered to a patient is started right before the patient is admitted to a clinical department, then discrepancies in judgement can occur in deciding whether to use the **Overlap** or **Before_Overlap** relations. However, given the overall annotation agreement score, these discrepancies in annotator judgement seem to have a very minor impact on the overall agreement quality of the expanded corpus.

5. Evaluation

5.1. Experimental Setup

In this section, we will conduct experiments using an existing temporal relations identification and classification system (D'Souza and Ng, 2013). The goal of these experiments is to practically examine changes that are caused in classifier performance after augmenting the i2b2 corpus with the new temporal relation annotations discussed in the preceding sections.

Dataset. We use the 190 training documents from the i2b2 corpus for classifier training and and reserve the 120 test documents for evaluating system performance. The additional inter-sentence temporal relation annotations cause us to have two versions of the i2b2 corpus: 1) original corpus (Original$_{corpus}$); and 2) original corpus augmented with additional inter-sentence temporal relation annotations (Expanded$_{corpus}$). Therefore, since we will run our classification system on both versions of the corpus, for each experiment there will be two sets of results shown.

	Relation	Agreement
1	**Overlap**	0.6
2	**Simultaneous**	0.66
3	**Before**	0.76
4	**After**	0.69
5	**Before_Overlap**	0.6
6	**Overlap_After**	0.55
7	**During**	0.71
8	**During_Inv**	0.7
9	**Begun_By**	0.72
10	**Begins**	0.67
11	**Ended_By**	0.68
12	**Ends**	0.71

Table 3: Per-class annotator agreement (Cohen's Kappa) for temporal relation annotation based on which temporal relation from the 12 relations defined in the i2b2 corpus to assign to each pair

Evaluation metrics. We employ the standard metrics viz. *Recall* (R), *Precision* (P), and *micro F-score* (F^{mi}) to evaluate our 12-class temporal relations classifier.

5.2. Results and Discussion

Tables 4 and 5 shows the results for our 12-class temporal relation identification and classification task over all instances from that dataset and over only the inter-sentence annotated instances, respectively. Since this work deals with augmenting the corpus with additional links, the overall classifier performance (in Table 4) is affected by the inter-sentence classification performance (in Table 5). As mentioned earlier, we show experiment results on both versions of the i2b2 corpus i.e. before and after augmenting the corpus with additional inter-sentence temporal relation annotations. The results in row (1) show classifier performance on the original i2b2 corpus. And the results in row (2) are from applying the classifier on the expanded version of the i2b2 corpus. In order to explain the results that rows (i) and (ii) represent, we first need to clarify the concept of data skew which we do next.

Data instances extracted from a corpus with a large number of gaps in temporal relation annotations tends to be skewed towards the class meant to represent a valid "no temporal relation" between an event pair. Skew in data is a problem because in such situations the classifier is unduly influenced by a single class thus making it incapable of learning representative features of the other classes. We deal with data skew in the original i2b2 corpus by pruning a majority of the instances belonging to the "no temporal relation" class based on certain conditions. While using a pruning component does tend to reduce data skew, the heuristics therein also tend to prune some of the positive data instances as a result which is an unwanted side-effect. One of the main motivating factors of this annotation work was to alleviate this problem of data skew. With the additional annotations in place, it is now possible to examine whether this problem has indeed been dealt with. We do this by training and testing the classifier with and without pruning heuristics on both versions of the corpus. If the classifier performance reduces significantly without pruning heuristics, then this is an indication of skew in data. But if classifier performs remains the same or shows only a slight variation, then this is an indication of no skew in the corpus being used. Rows (i) and (ii) represents classifier performance with and without pruning heuristics, respectively, when trained and tested on the two different corpus versions.

	Corpus		R	P	F^{mi}
1	Original$_{corpus}$	(i)	52.9	21.0	30.0
		(ii)	54.1	12.5	20.3
2	Expanded$_{corpus}$	(i)	36.0	32.5	34.1
		(ii)	37.6	30.9	33.9

Table 4: Classification of automatically identified temporal relations on the original and expanded versions of the i2b2 corpus

	Corpus		R	P	F^{mi}
1	Original$_{corpus}$	(i)	52.7	6.7	11.9
		(ii)	60.9	3.0	5.6
2	Expanded$_{corpus}$	(i)	13.0	35.9	19.0
		(ii)	16.7	27.0	20.6

Table 5: Classification of automatically identified temporal relations inter-sentence temporal relations on the original and expanded versions of the i2b2 corpus

In examining the results in rows (1) and (2) of Tables 4 and 5, we see that classifier performance has a higher *reliability* factor on the *Expanded* version of the corpus than on the *Original* i2b2 corpus. This is reflected by the improvements in precision in the identification and classification system's output in row (2)'s results when compared with row 1's results. Next we also see that the data skew problem which existed in the *Original* corpus as evidenced by the drop in Original$_{corpus}$ f-score in row (ii) when compared with row (i)'s f-score, no longer exists in the *Expanded* version of the corpus, as the f-scores in rows (i) and (ii) remain relatively unchanged. This shows that the additional annotations alleviate skew in original data.

		Original$_{corpus}$			Expanded$_{corpus}$		
	Relation	R	P	F^{mi}	R	P	F^{mi}
1	**Overlap**	69.7	2.4	4.8	19.1	25.8	22.0
2	**Simultaneous**	69.2	9.9	17.3	49.2	90.4	63.7
3	**Before**	56.1	1.8	3.4	17.7	18.9	18.3
4	**After**	32.5	3.2	5.9	4.3	18.8	70.2
5	**Before_Overlap**	38.6	1.6	3.1	13.2	18.1	15.3
6	**Overlap_After**	41.2	1.7	3.3	7.3	13.7	9.4
7	**During**	0.0	0.0	0.0	18.1	18.2	18.2
8	**During_Inv**	58.1	3.0	5.6	24.1	26.6	25.3
9	**Begun_By**	75.0	18.8	30.0	25.0	75.0	37.5
10	**Begins**	4.3	0.6	2.7	7.4	19.0	5.8
11	**Ended_By**	40.0	7.6	12.7	74.8	53.3	13.1
12	**Ends**	30.8	25.0	27.6	7.1	75.0	13.0

Table 6: Per-class inter-sentence temporal relation classification without pruning heuristics on both versions of the i2b2 corpus

Finally, for more detailed insights into the task of automatically linking and classifying temporal relations, we present in Table 6 per-class classification results of the 12-class system without pruning heuristics on both version of the corpus. From these results we see that the difficult classes for the classifier to classify are **Overlap**, **Before_Overlap**, and **Overlap_After** which happens to be exactly the same set of classes which the human annotators had least agreement scores on. We can now safely conclude that the automatic identification and classification of temporal relations closely models after human annotation of temporal relations.

6. Conclusion

In conclusion, we have augmented the i2b2 corpus with a significant proportion of additional temporal relation links that are not automatically recoverable. We have shown that the data produced as result of this work is consistent in choice of temporal relations to add to the corpus owing to fair inter-annotator agreement scores. In addition, for any annotation work dealing with temporal relations in medical data we provide an extended version of the original i2b2 guidelines by eliciting specific criteria for eligibility of temporal relations between events. Experiments on the expanded version of the corpus shows that the additional annotations added to the i2b2 corpus has a positive effect on classifier performance and also eliminates data skew thus facilitating a more natural experimental setting.

Acknowledgements

We would like to thank the anonymous reviewers for their constructive comments. We would also like to thank Isaac Persing whose invaluable annotation work made it possible for us to obtain the inter-annotator agreement scores.

7. References

Allen, James F. (1983). Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843.

Carletta, Jean. (1996). Assessing agreement on classification tasks: The Kappa statistic. *Computational Linguistics*, 22(2):249–254.

D'Souza, Jennifer and Ng, Vincent. (2013). Temporal relation identification and classification in clinical notes. In *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*, page 392. ACM.

Pustejovsky, James, Hanks, Patrick, Sauri, Roser, See, Andrew, Day, David, Ferro, Lisa, Gaizauskas, Robert, Lazo, Marcia, Setzer, Andrea, and Sundheim, Beth. (2003). The TimeBank corpus. In *Corpus Linguistics*, pages 647–656.

Sun, Weiyi, Rumshisky, Anna, and Uzuner, Ozlem. (2013). Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association*, 20(5):806–813.

UzZaman, Naushad, Llorens, Hector, Allen, James, Derczynski, Leon, Verhagen, Marc, and Pustejovsky, James. (2012). Tempeval-3: Evaluating events, time expressions, and temporal relations. *arXiv preprint arXiv:1206.5333*.

Verhagen, Marc, Knippen, Robert, Mani, Inderjeet, and Pustejovsky, James. (2006). Annotation of temporal relations with Tango. In *Proceedings of LREC*.

Verhagen, Marc, Gaizauskas, Robert, Schilder, Frank, Hepple, Mark, Katz, Graham, and Pustejovsky, James. (2007). SemEval-2007 Task 15: TempEval temporal relation identification. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 75–80.

Verhagen, Marc, Sauri, Roser, Caselli, Tommaso, and Pustejovsky, James. (2010). SemEval-2010 Task 13: TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 57–62.

Verhagen, Marc. (2005). Temporal closure in an annotation environment. *Language Resources and Evaluation*, 39(2):211–241.