Cross-document coreference: An approach to capturing coreference without context

Kristin Wright-Bettner[1], **Martha Palmer**[1], **Guergana Savova**[2], **Piet de Groen**[3], **Timothy Miller**[2]
[1]Department of Linguistics, University of Colorado, Boulder, CO 80309
[2]Boston Children's Hospital and Harvard Medical School, Boston, MA 02115
[3]Department of Medicine, University of Minnesota, Minneapolis, MN 55455
kristin.wrightbettner@colorado.edu
martha.palmer@colorado.edu
[2]{first.last}@childrens.harvard.edu
degroen@umn.edu

Abstract

In this paper, we discuss a cross-document coreference annotation schema that we developed to further automatic extraction of timelines in the clinical domain. Lexical senses and coreference choices are determined largely by context, but cross-document work requires reasoning across contexts that are not necessarily coherent. We found that an annotation approach that relies less on context-guided annotator intuitions and more on schematic rules was most effective in creating meaningful and consistent cross-document relations.

1 Introduction

The ability to learn cross-document coreference and temporal relationships in clinical text is crucial for the automatic extraction of comprehensive patient timelines of events (Raghavan et al., 2014). To that end, we present a gold corpus of 198 clinical-narrative document sets, where each set consists of three notes for a given patient (594 individual notes total). Each file is annotated with intra-document temporal, coreference, and bridging relations (SET-SUBSET, WHOLE-PART, CONTAINS-SUBEVENT), and each set is annotated with cross-document coreference and bridging relations.

The goal of the current project was to leverage the inherited, intra-document annotations from two prior projects (discussed in Section 2) to capture longer, more developed timelines of patient information. We did this by creating human-annotated cross-document coreference and bridging links and then using inference to combine this information with the knowledge already gained from the intra-document temporal and coreference/bridging links.

In this paper, we discuss the impacts of cross-document-specific phenomena on human annotation and machine learning, most notably the effect of disjunct narratives on cross-document coreference judgments. Cohesive discourse is a crucial linguistic tool for determining coreference, yet the cross-document relations annotation task fundamentally takes place across discontinuous narratives. We found an approach that is governed more by annotation rules than annotator intuition to be most effective, producing an inter-annotator agreement score of 93.77% for identical relations. While an approach that moves away from linguistically-intuitive judgments may seem surprising at first, it is in fact quite fitting for a task that is inherently void of the discourse-level linguistic cues that humans employ to make those intuitive associations.

We also discuss other cross-document phenomena, inter-annotator agreement, and, briefly, areas for future work. Related work is discussed throughout.

2 The THYME colon cancer corpus

This annotation effort merged and expanded on document-level annotations created by two prior projects – a temporal relations project[1] (Styler et al., 2014), and a coreference and bridging relations project.[2] These two projects will be referred to as THYME 1 (Temporal History of Your Medical Events) and Clinical Coreference.

[1] Corpus publicly available from TempEval. Guidelines available at http://clear.colorado.edu/compsem/documents/THYME_guidelines.pdf.
[2] Clinical Coreference Annotation Guidelines available at http://clear.colorado.edu/compsem/documents/coreference_guidelines.pdf.

Proceedings of the 10th International Workshop on Health Text Mining and Information Analysis (LOUHI 2019), pages 1–10
Hong Kong, November 3, 2019.
https://doi.org/10.18653/v1/D19-62

Relation Type	Description	Link
IDENTICAL (IDENT)	M1 refers to the same event/entity as M2	*[M1] IDENT [M2]*
SET-SUBSET (S-SS)	M2 refers to one or more members of a larger group, represented by M1.	$[M1]_{SET}$ - $[M2]_{SUBSET}$
CONTAINS-SUBEVENT (CON-SUB)	M1 temporally contains M2, and M2 is inherently part of the structure of M1.	*[M1] CON-SUB [M2]*
WHOLE-PART (W-P)	M2 is compositionally part of a larger entity, represented by M1.	$[M1]_{WHOLE}$ - $[M2]_{PART}$

Table 1: Gold-annotated cross-document relation types in the THYME colon cancer corpus. M1 refers to Markable 1 and M2 refers to Markable 2. Markables include events, entities, or temporal expressions. W-P was used only for entities; CON-SUB only for events. All four relation types are coreference or bridging links rather than temporal links, except for CON-SUB, which conveys both temporal and structural information and is represented as a temporal link (TLINK) in our annotation tool. This TLINK type is discussed in Section 3.1.

The corpus [3] consists of de-identified physicians' notes on colon cancer patients. The examples used throughout this paper are artificially created; however, we have done our best to replicate the relevant linguistic contexts. Each set of three notes consists of a clinical report, a pathology report, and a second clinical report, in that chronological order and spanning a period of weeks or months. Capturing such temporally-extensive information gives us the ability to track the status of the disease over time and responses (or not) to treatment.

The THYME colon cancer corpus now includes: a) intra-document gold annotations for all markables (events, entities, and temporal expressions) and several types of temporal, coreference, and bridging relations; and b) cross-document gold annotations for four coreference and bridging relation types, which represent a subset of the intra-document types [4] and are described in Table 1.

[3]This corpus has also been annotated according to the Penn Treebank, PropBank, and Unified Medical Language System (UMLS) Semantic Network schemas (Albright et al., 2013), though these data did not influence the current project.

[4] The intra-document relations additionally include the following types: *CONTAINS, BEFORE, OVERLAP, BEGINS-ON, ENDS-ON, NOTED-ON*, and *APPOSITIVE*. These are all temporal relations, except APPOSITIVE, which is a coreference relation. All were used by either Clinical Coreference or THYME 1 (Styler et al., 2014), except for *CONTAINS-SUBEVENT* and *NOTED-ON*, which are new to the current project. All are discussed in detail in our guidelines: `https://www.colorado.edu/lab/clear/projects/computational-semantics/annotation`.

Many prior studies have noted the intractability of creating cross-document gold annotations on large corpora (Day et al., 2000, for example). Each cross-document effort has therefore restricted the scope of their annotations in some way (e.g., Song et al., 2018; Cybulska and Vossen, 2014) and/or developed machine-produced annotations for cross-document relations, rather than human-produced (Raghavan et al., 2014; Dutta and Weikum, 2015; Baron and Freedman, 2008; Gooi and Allen 2004; etc.). We likewise restricted our approach by limiting the cross-document relations to the groups of three files which represent each patient, and by limiting the number of annotated relation types. However, the THYME corpus is the largest dataset of gold-annotated clinical narratives to-date that we are aware of, in terms of types of markables and relations annotated.

We are indebted to the contributions of the projects that preceded us. Much of the technical and conceptual groundwork had already been laid for our task. In particular, the notion of narrative containers (Styler et al., 2014; Pustejovsky and Stubbs, 2011) informed our addition of the CONTAINS-SUBEVENT temporal link and our cross-document annotation process.

However, we found that the segregation of tasks during the creation of the single-file gold annotations caused a variety of technical and conceptual conflicts once their outputs were merged. Furthermore, aspects of the temporal task suffered from focusing only on local-context relations; a global grasp of the text, which coreference annotation facilitates, reveals

temporally significant information that may be otherwise missed or misinterpreted.

Early experiments showed that conflicts in the merged annotations rendered meaningful cross-document annotation untenable. To reconcile these conflicts, we therefore introduced an intra-document corrections-style manual annotation pass prior to cross-document double-annotation and adjudication.

3 Cross-document annotation: Process, assumptions, phenomena

It has been well-attested that determining cross-document relations poses a unique set of challenges for both systems and annotators. Song et al. (2018) discuss the cognitive strain on annotators, and others have observed the decrease in linguistic cues that occurs cross-document (Raghavan et al., 2014; Hong et al., 2016). In this paper, we are most interested in the latter, particularly the impacts of cross-document mode on identical relations.

Many coreference annotation guidelines, including ours, use a straightforward definition of coreference, which may be summarized as two different mentions in a text having the same real- or hypothetical-world referent (e.g., Cybulska and Vossen, 2014; Richer Event Description Annotation Guidelines, 2016 [5]; Cohen et al., 2017). This definition leads to a binary approach to identical judgments – two mentions either refer to the same thing or they do not. Annotators are forced to make a polar choice about representations of meaning, when those representations in fact exist on a spectrum. This is not a new discovery: "There are cases where variant readings of a single lexical form would seem to be more appropriately visualized as points on a continuum – a single fabric of meaning with no clear boundaries" (Cruse, 1986). However, the natural language processing community is still learning how to deal with this.

Others have identified the problems that this oversimplified definition creates for annotation: "Degrees of referentiality as well as relations that do not fall neatly into either coreference or non-coreference—or that accept both interpretations—are a major reason for the lack of inter-coder agreement in coreference annotation" (Recasens, 2010). Hovy et al. (2013) also recognized the need for a more nuanced approach and introduced membership and subevent relations as a result.

Furthermore, both Recasens and Hovy discuss the role that pragmatics plays in determining coreference:

- "Two mentions fully corefer if their activity/event/state DE [discourse element] is identical in all respects, *as far as one can tell from their occurrence in the text*" (Hovy et al., 2013, emphasis added).
- "We redefine coreference as a scalar relation between two (or more) linguistic expressions that refer to discourse entities *considered to be at the same granularity level relevant to the linguistic and pragmatic context*" (Recasens et al., 2011, emphasis added).

Context, therefore, contributes in a crucial way to determining sense for a given lexical unit – and therefore also to determining coreference relations for that unit. We agree with Cruse, Recasens, and Hovy and observe the unique challenge this poses for cross-document annotation, since distinct narratives do not share a coherent discourse context. Recasens et al. (2011) propose that categorization and meaning are constructed in a *temporary*, active process; in cross-document work, we are attempting to create meaningful relations between temporally disconnected discourses. Put differently, the coherence of context is decreased while the number of contexts for lexical senses is increased.

While not surprising, this phenomenon does have interesting consequences for annotation. In fact, by the definitions given above, "doing" coreference between disjunct linguistic and pragmatic contexts could be viewed, on some level, as impossible.

But not all hope is lost. Particularly for our corpus, texts are very closely related and it is possible to create meaningful relations. However, the phenomenon just described requires an approach to cross-document coreference annotation that is unique from within-document. We dealt with this primarily by adding a subevent relation that was governed more by annotation rules and less by annotators' intuitions. We present the reasons for and outcome of this approach in the next section, followed by discussion of other cross-document phenomena and our technical cross-document linking process.

[5] https://github.com/timjogorman/RicherEventDescription/blob/master/guidelines.md

3.1 An approach to coreference across disjunct contexts

Consider the following single-file example:

(1) *October 15th, 2015 – Dr. Wu performed* ***resection*** *of the primary tumor. Ms. Smith's recovery from* ***surgery*** *has been without complication.*

The choice here about whether to link *resection* and *surgery* as coreferential is likely to produce a disagreement. Annotator A may decide they are IDENTICAL (IDENT) since they clearly refer on some level to the same cancer treatment procedure; a significant semantic relationship would be lost if we did not link them. Annotator B, however, may decide *resection* refers only to the literal act of removing the tumor, while *surgery* points to the entire procedure. Essentially, the annotators disagree about whether the two terms are "close enough" on the meaning spectrum to warrant an IDENT link. More precisely, the disagreement stems from different interpretations of semantic granularity – Annotator A's identity "lens" is more coarse-grained, while Annotator B's is more fine-grained.

Consider a second example:

(2) *PLAN:* ***Resection*** *of primary tumor and gallbladder removal. Patient is scheduled for* ***surgery*** *on October 15th, 2015.*

Here, the finer-grained approach to *Resection* and *surgery* is supported – required, in fact – by the context. No coreference relationship is possible since the surgery clearly consists of two subprocedures, the tumor resection and the gallbladder removal.

Now consider the two examples together, where (2) is from the chronologically earlier note and (1) is from the later note in a single set:

(3) **Note A**: *PLAN:* ***Resection*** *of primary tumor and gallbladder removal. Patient is scheduled for* ***surgery*** *on October 15th, 2015.*

- No coreference link
- *surgery* CONTAINS *Resection*
- *surgery* CONTAINS *removal*

Note B: *October 15th, 2015 – Dr. Wu performed* ***resection*** *of the primary tumor. Ms. Smith's recovery from* ***surgery*** *has been without complication.*

- *resection* IDENTICAL *surgery*

The IDENT link shown for Note B represents the original gold annotation in our data, i.e., the more coarse-grained approach to identity described above. This is arguably the better perspective here, based on Recasens' definition of coreference above ("discourse entities considered to be at the same granularity level relevant to the linguistic and pragmatic context"); Note B's narrative is quite broad-brushed and supports what Hovy terms a "wide reading" of *resection* (Hovy et al., 2013). Pragmatically, *resection* and *surgery* are the same in Note B; pragmatically, they are not the same in Note A.

The predicament for cross-document linking is obvious. If we link $resection_A$ to $resection_B$, this entails that $resection_A$ is IDENT to $surgery_B$; if we then link $surgery_A$ to $surgery_B$, this now entails that the procedure temporally contains itself. If we leave $resection_A$ unlinked to $resection_B$ to avoid this conflict, problematically, we miss the relation between identical strings that refer to the same event (*?resection of primary tumor* ~~*IDENT*~~ *resection of primary tumor*), not to mention that leaving these unlinked would be extremely counterintuitive for annotators.

This type of situation is common in cross-document work. Since identity judgments are based on granularity levels that are in turn determined by the pragmatics of the narrative, and since the pragmatic contexts of two or more disjunct narratives are not necessarily coherent, cross-document mode frequently forces annotators to choose between: (a) not linking two mentions that are obviously and significantly semantically related, or (b) linking these mentions and thereby forcing logically-conflicting information as in (3), which in turn renders the existing temporal links much less meaningful.

To account for this variation in context-determined granularity, we introduced the CONTAINS-SUBEVENT (CON-SUB) link, which says that EVENT B is both temporally contained by EVENT A and it composes part of EVENT A's essential structure (modeled after the subevent relation in O'Gorman et al., 2016). We added this new relation intra-document in the corrections pass, as well as in the later cross-document pass. For examples like (3), this meant the Note B IDENT relation was re-interpreted as a subevent relation: *surgery* CON-SUB *resection*. This allowed us to preserve the close semantic connection between the two EVENTs in both narratives, while avoiding the logical conflicts that would have rendered our output much less meaningful and informative. We can also assume the inter-annotator agreement

achieved (discussed in section 4) is much higher than it would have been had we left annotators in the predicament shown in (3).

The consistency noted above was achieved by an approach that relied less on discourse cues and more on general semantic distinctions. Instead of allowing annotators to intuitively judge between wide and narrow readings (borrowing Hovy's terms again) of lexical items based on the context, we required IDENT and CON-SUB relations to be based more on the dictionary definitions of the terms. This is because we could not predict the granularity distinctions that cross-document information would expose, as shown in (3). For example, annotators were required to differentiate between "general" surgery terms (e.g., *surgery*, *procedure*, *operation*, etc.) and "specific" surgery terms (*colectomy*, *resection*, *excision*, etc.), such that the general term nearly always contained the specific term as a subevent. This compensated for the majority of granularity distinctions in the THYME corpus (though not all, since there can always be more fine-grained levels of nuance). This framework therefore facilitated more straightforward cross-document linking, though it did also force annotators to make some counterintuitive within-document choices since senses *are* influenced by the context.

Song et al. (2018) took an opposite approach to cross-document coreference linking through their use of event hoppers, which permit "coreference of two events that are *intuitively* the same although certain features may differ" (emphasis added). We found this approach did not suit our needs since the ultimate goal was to capture a coherent timeline of clinical events, and intuitive coreference linking produced temporal conflicts, as shown above.

While coreference linking is not possible on the cross-document level in the same nuanced and intuitive way that it is within-document, there is still a great deal of important information we can capture. The texts in our corpus are topically very similar and there are typically a lot of corroborating details, such as dates and locations (again, these have been de-identified, but in a consistent fashion). Additionally, the clinically-delineated sections and the note types and structure provide clues about how to interpret the events; for example, due to the date and descriptive details, we can know which procedure in the clinical note the pathology note refers to, even if the overall procedure is not explicitly mentioned in the pathology note.

Time constraints prevented us from adding CON-SUB for all event types. We annotated it for four event categories, chosen based on clinical significance and demonstrated need due to cross-document conflicts like the one in (3): (a) patient treatment events, including surgical procedures and chemotherapy/radiation treatments; (b) cancer events (*cancer*, *adenocarcinoma*, *tumor*, etc.); (c) medications; and (d) chronic disease events.

Due to other conflicts arising from the disconnected contexts that the subevent relation was not able to reconcile, we permitted the cross-document adjudicators (but not annotators) to make within-document annotation changes when absolutely necessary.

In summary, we found that it is possible to capture meaningful cross-document coreference relations, but the approach must differ from intra-document annotation because pragmatically-directed within-document intuitions may conflict in unpredictable ways on the cross-document level.

3.2 Other cross-document phenomena

We have discussed in depth the way identical judgments are affected by disjunct contexts. We discuss two more cross-document phenomena here: (1) the use of inference in linking stative events; and (2) how cross-document work exposes typos and misinformation. Notably, this could be leveraged to identify mistakes in the text, which may contribute to current efforts to reduce medical errors in patient treatment.

Inference and stative events

Cross-document coreference is typically easier for punctual events (such as tests and procedures) and harder for durative events that can change in value over time (for example, a mass that is initially benign but becomes malignant). As with many other cross-document challenges, this issue is also present within-document, but is exacerbated in the cross-document setting because context is reduced. Consider the following example:

(4) Note A (March 24 2012 SECTIONTIME):
Pulse Rate: ***Regular***
Note B (March 26 2012 SECTIONTIME):
Heart: ***Regular*** *rate*

Here we have two clinically-relevant states associated with two different times: the regularity of the patient's heart rate on March 24, 2012, and the regularity of the patient's heart rate on March 26, 2012. The question for a cross-doc annotator is whether these two EVENTs are IDENTICAL.

For the current example, it is likely the regular condition has continued, but the fact is we do not

know, especially since the patient may have a medical condition that causes sporadic irregularity. Furthermore, we might be initially inclined to infer sameness due to the close temporal proximity of the two measurements (two days apart), but that thought trajectory quickly leads to problems: When are two continuous events not temporally near enough to infer sameness? A week? A month? How do we decide?

Song et al. (2018) discuss a similar example across four notes, in which they corefer the first three events because they occur in "about the same time period and same place" (occurring over the timespan of a month), but they do not corefer the fourth event "as it happened at a different time" (about four months after the most recent other mention). However, it is not clear how they determined that a month is a reasonably close enough timespan to infer sameness, while four months is not.

Our approach, therefore, was that when condition or attributive EVENTs – events that vary in value – are measured or identified at two different times, they should not be linked, unless there is explicit linguistic evidence (e.g., use of the present perfect tense) they are the same event. Essentially, we decided that temporal proximity alone was not enough to infer an identical relation for two condition/value events.

Of course, inference is a source of inter-annotator disagreement for other cross-document choices as well. A comprehensive analysis is outside the scope of this paper, but the topic is discussed further in the following point.

How cross-document annotation exposes mistakes in the data

We discuss this in detail not only because it has implications for discovering misinformation in the text, but also because it demonstrates two more significant challenges to cross-document clinical annotation: the heavy cognitive burden on annotators, and the need for clinical knowledge. Consider the following example:

(5) **Note A** (DOCTIME: August 21, 2012):
*We have ordered a **CT** abdomen and pelvis to rule out liver metastases prior to surgery. Mr. Olson will also need an EKG and bloodwork. Testing was negative.*

- *CT* assigned DocTimeRel of *AFTER*, i.e., it occurs after DOCTIME (Aug 21, 2012).

Note B (DOCTIME: September 30, 2012):
***CT** abdomen and pelvis was compared to the prior **study** of August 20, 2012, Mr. Olson had low-anterior resection.*

- *August 20, 2012* CONTAINS *study*

CT_A and $study_B$ are in fact IDENTICAL. Combined with the temporal information noted above, this entails that the same event both occurs after Aug. 21, 2012, and is temporally contained by Aug. 20, 2012 – a logical impossibility.

We know they are the same event based primarily on real-world knowledge of the standard order of medical procedures, as follows: It is clear in Note B that there are two different CT scans. The question facing a cross-document annotator is which one, if either, is IDENT to CT_A? We know explicitly from the text that CT_A occurred prior to the patient's surgery. CT_B occurred after the patient's surgery, since, however cryptically, it references observation of the surgery ("Mr. Olson had low-anterior resection"). Therefore, CT_A and CT_B are not referring to the same scan.

Now the question is whether $study_B$ is IDENT to CT_A. The initial evidence is to the contrary – $study_B$ is explicitly said to occur on Aug. 20, while *CT* is inferably after (or later in the day on) Aug. 21. However: (a) it is unusual to have two CT scans back-to-back, without further discussion; (b) an Aug. 20 CT is not discussed in the Aug. 21 note; (c) in Note A, immediately after noting that several tests have been ordered, the text says, "Testing was negative." Based on the verb tenses in the paragraph, the assumption would likely be that *Testing* here refers to other tests, not the ones just ordered. However, the flow of discourse suggests otherwise, along with the fact that no other prior testing is referred to in the same section. With the additional information we have from Note B, a more reasonable interpretation presents itself: $study_B$ is IDENT to CT_A, and Aug. 20 is the correct date of the scan. The note was likely originally written on Aug. 20, prior to the scan that was done later that day, and was later updated with the test results but without any indication of the update being written at a later time. This analysis was confirmed by review of all notes by our medical expert consultant.

There are several noteworthy observations about this: First, there is quite a bit of oncological knowledge required to notice the conflict above. Furthermore, the non-standard syntax in Note B would make it easy for an annotator to miss the fact that CT_B is after the resection.

Example	Text	Within-doc links	Cross-doc links
File Set 1	Note A: ...*screening tests*… Note B: ...*screening tests*...*MRI*	*tests*$_B$ SET-SUBSET *MRI*$_B$	*tests*$_A$ IDENT *tests*$_B$
File Set 2	Note A: ...*screening tests*… Note B: ...*MRI*...	None	*tests*$_A$ SET-SUBSET *MRI*$_B$
File Set 3	Note A: ...*screening tests*...*MRI* Note B: ...*screening tests*...*MRI*	*tests*$_A$ SET-SUBSET *MRI*$_A$ *tests*$_B$ SET-SUBSET *MRI*$_B$	*tests*$_A$ IDENT *tests*$_B$ *MRI*$_A$ IDENT *MRI*$_B$

Table 2: For File Set 1, there is no cross-doc S-SS link between *tests*$_A$ and *MRI*$_B$ because this can be inferred from the cross-doc IDENT link and the within-doc S-SS link shown. For File Set 3, the fact that *MRI*$_A$ has the same referent as *MRI*$_B$ is not inferable from the intra-document structural links; hence, we create a cross-doc IDENT link. (Crucially, all examples assume that context allows us to know that these mentions do in fact refer to the same testing events.)

Second, even armed with the necessary clinical knowledge, there is still a fair amount of inference involved in making the above choice. However, note that *all* of the annotation options here, including the option to not link at all, require a lot of inference (as is the nature of many cross-document analyses). There are different types of inference based on different kinds of information. While we decided that temporal closeness is not enough by itself to infer a relation for condition/value events, we decided here that medical knowledge of standard processes is enough to infer a relation.

Third, assuming the above observations were made, an impossible annotation choice presents itself: Do we make the coreference link even though it forces a temporal conflict, or do we keep the timeline clean and lose the coreference relation? We decided on the former, and kept track of the noted temporal conflicts in order to inform systems training.

Finally, note the time, attention, and careful thought process required for determining this single cross-document link. While certainly not all decisions are this demanding, the amount of time necessary to produce high-caliber annotations should be apparent. It took highly-experienced annotators about 1.5 hours on average to complete one document set, or an estimated 891 hours total for two annotators and one adjudicator to produce 198 gold sets with a total of 10,560 cross-document links. This does not include time spent on initial annotation experiments, process and guidelines development, annotator training, and post-processing steps.

3.3 Cross-document annotation process

To manage the potentially vast number of cross-document links, we established a set of assumptions about inferable relations that guided the following process and are further discussed in Table 2 (note: "structural links" refers to links that have a hierarchical rather than identical relationship: *CON-SUB, S-SS, W-P*):

(a) Link topmost mention to topmost mention. We assume the other relations can be inferred from within-document chains.

(b) If there is a within-document structural link between two markables, do not create that same link cross-document for the same two events/entities. Put differently, create cross-document structural links only when *both* components of the relation do not have a cross-document IDENT link. Again, we assume that other relations can be inferred.

(c) Always create IDENT links whenever appropriate.

4 Inter-annotator agreement

We scored inter-annotator agreement (IAA) only for annotation categories that were new to the current project, i.e., intra-document CON-SUB links and all cross-document links. Furthermore, we only scored annotator-annotator agreement (not annotator-gold), since adjudicators were permitted to change single-file annotations while annotators were not. The total number of gold markables and relations are shown in Table 3; IAA results are shown in Table 4 and are averaged over all the documents (both tables shown on following page).

The IDENT score is much higher than the structural linking scores because the structural links were only created in cases where neither component of the link had a cross-document IDENT relation (see Section 3.3). These relations were therefore brand-new and had to be identified

Markables (594 documents)	143,147 total	Relations, within-doc and cross-doc (594 documents)	70,572 total	Cross-doc relations (198 documents)	10,762 total
TIMEX3s	7,796	**Temporal links**	35,428 total	**IDENTICAL**	9,102
Entities*	47,355	CONTAINS	14,037	**SET-SUBSET**	405
EVENTs	86,172	CON-SUB**	4,718	**WHOLE-PART**	13
SECTIONTIME	1,230	BEFORE	4,217	**CON-SUB**	1,242
DOCTIME	594	OVERLAP	5,091		
		BEGINS-ON	1,200		
**Entities are referred to as MARKABLEs in our guidelines, due to the naming practice of the prior Clinical Coreference project.*		ENDS-ON	557	***CON-SUB is listed twice under the second column since it's both a temporal link and a bridging link.*	
		NOTED-ON	5,608		
		Aspectual links	873 total		
		INITIATES	259		
		CONTINUES	302		
		TERMINATES	278		
		REINITIATES	34		
		Coreference and bridging links	38,337 total		
		IDENTICAL	23,827		
		SET-SUBSET	5,907		
		WHOLE-PART	3,885		
		CON-SUB**	4,718		

Table 3: Total gold markables and relations for the THYME colon cancer corpus.

without the benefit of a single coherent discourse, as discussed in depth above. On the other hand, annotators were able to draw on the information conveyed in intra-document relations when determining cross-document IDENT relations.

The WHOLE-PART (W-P) IAA score is zero because there were very few cross-document W-P relations in the corpus, under our guidelines. W-P is used only for entities, and we did not do W-P cross-document linking for anatomical entities (due to the massive amount of mentions, the spider-webbed relations, and the number of vague terms – *tissue portions*, etc. – we only created IDENT anatomy relations at the cross-document level). Therefore, the only cross-narrative W-P relations were between organizations/departments and members of those entities, which were only rarely knowable from the text.

The CONTAINS-SUBEVENT (CON-SUB) agreement score is likely higher than the SET-SUBSET (S-SS) score because we applied it to four specific event categories (see Section 3.1) that consist of oft-repeated terms. S-SS, on the other hand, had no such constraints, making this relation much more challenging to identify over the scope of three often-lengthy documents. Furthermore, while some set-member relations are obvious, others are not. For example:

(6) **Note A**: *Pt denies alcohol or tobacco **use**.*
Note B: *He denies **drinking**.*
- use_{NEG} S-SS $drinking_{NEG}$

Intra-document IAA	Cross-document IAA
CON-SUB: 34.14%	IDENTICAL: 93.77% CON-SUB: 36.43% SET-SUBSET: 6.88% WHOLE-PART: 0.00%

Table 4: Intra-document and cross-document inter-annotator agreement scores in terms of percentage agreement.

One of our annotators identified the S-SS link shown, while the other did not. In the future, more examples and/or constraints of fringe S-SS relations in the annotation guidelines could be developed to improve S-SS agreement.

5 Conclusion

As demonstrated, developing an extensive timeline of patient events that occur over multiple weeks and months is an extremely complicated process. Understanding the breadth of complexity and the heavy demands on annotators is necessary for projecting annotation budgets and timelines, and for understanding the nature and quality of the resulting data for predicting machine learning performance. Two of the most pressing areas for future research include: (a) further development and testing of our approach to cross-document linking presented in section 3.1; and (b) development of a comprehensive methodology for incorporating medical expertise, as alluded to in section 3.2 (building on but extending beyond the light-annotation tasks methodology proposed

by Stubbs, 2013). It is critical that wherever possible the annotation process is based on clear rules rather than annotator intuition as the former lends itself to automation whereas the latter at best results in a non-scalable solution with a narrow field of implementation. Developing these rules requires medical domain expertise.

Our results for cross-document coreference annotation leave ample room for improvement. Yet we believe that the approaches discussed here will serve as another significant step in the development of automatic extraction of event timelines in medical data.

Acknowledgments

The work was supported by funding R01LM010090 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 Institute of Health.

We would like to thank: Dana Green especially for annotation and insightful annotation input; Ahmed Elsayed and Dave Harris for medical annotation and advice; James Martin for schema development advice; Wei-Te Chen and Skatje Myers for technical support; Michael Regan, Matthew Oh, Hayley Coniglio, Samuel Beer, and Jameson Ducey for annotation; and Adam Wiemerslage for IAA and post-processing scripts.

References

Daniel Albright, Arrick Lanfranchi, Anwen Fredriksen, William F Styler IV, Colin Warner, Jena D Hwang, Jinho D Choi, Dmitriy Dligach, Rodney D Nielsen, James Martin, Wayne Ward, Martha Palmer, Guergana K Savova. 2013. Towards comprehensive syntactic and semantic annotations of the clinical narrative. *Journal of the American Medical Informatics Association*, *20*(5): 922–930. https://doi.org/10.1136/amiajnl-2012-001317.

Alex Baron and Marjorie Freedman. 2008. Who is Who and What is What: Experiments in Cross-Document Co-Reference. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, pages 274–283. https://www.aclweb.org/anthology/D08-1029.

K. Bretonnel Cohen, Arrick Lanfranchi, Miji Joo-young Choi, Michael Bada, William A. Baumgartner Jr., Natalya Panteleyeva, Karin Verspoor, Martha Palmer, Lawrence E. Hunter. 2017. Coreference annotation and resolution in the Colorado Richly Annotated Full Text (CRAFT) corpus of biomedical journal articles. *BMC Bioinformatics*, 18:372.

D. A. Cruse. 1986. *Lexical Semantics*. Cambridge University Press, Cambridge, UK.

Agata Cybulska and Piek Vossen. 2014. Using a sledgehammer to crack a nut? Lexical diversity and event coreference resolution. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, European Language Resources Association, pages 4545-4552. http://www.lrec-conf.org/proceedings/lrec2014/pdf/840_Paper.pdf.

David Day, Alan Goldschen, John Henderson. 2000. A Framework for Cross-Document Annotation. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, European Language Resources Association. http://www.lrec-conf.org/proceedings/lrec2000/pdf/201.pdf.

Sourav Dutta and Gerhard Weikum. 2015. Cross-Document Co-Reference Resolution using Sample-Based Clustering with Knowledge Enrichment. *Transactions of the Association for Computational Linguistics*, *3*, pages 15–28. https://doi.org/10.1162/tacl_a_00119.

Chung Heong Gooi and James Allan. 2004. Cross-Document Coreference on a Large Scale Corpus. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, Association for Computational Linguistics, pages 9-16. https://www.aclweb.org/anthology/N04-1002.

Yu Hong, Tongtao Zhang, Tim O'Gorman, Sharone Horowit-Hendler, Heng Ji, Martha Palmer. 2016. Building a Cross-document Event-Event Relation Corpus. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, Association for Computational Linguistics, pages 1–6. https://aclweb.org/anthology/W16-1701.

Edward Hovy, Teruko Mitamura, Felisa Verdejo, Jun Araki, Andrew Philpot. 2013. Events are Not Simple: Identity, Non-Identity, and Quasi-Identity. In *Workshop on Events: Definition, Detection, Coreference, and Representation*, Association for Computational Linguistics, pages 21–28. https://www.aclweb.org/anthology/W13-1203.

Tim O'Gorman, Kristin Wright-Bettner, Martha Palmer. 2016. Richer Event Description: Integrating event coreference with temporal, causal and bridging annotation. In *Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016)*, Association for Computational

Linguistics, pages 47–56. https://aclweb.org/anthology/W16-5706.

James Pustejovsky and Amber Stubbs. 2011. Increasing informativeness in temporal annotation. In *Proceedings of the 5th Linguistic Annotation Workshop*, Association for Computational Linguistics, pages 152–160. https://www.aclweb.org/anthology/W11-0419.

Preethi Raghavan, Eric Fosler-Lussier, Noémie Elhadad and Albert M. Lai. 2014. Cross-narrative temporal ordering of medical events. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, pages 998–1008. https://aclweb.org/anthology/P14-1094.

Marta Recasens. 2010. "Coreference: Theory, Annotation, Resolution and Evaluation." PhD Thesis. University of Barcelona. http://stel.ub.edu/cba2010/phd/phd.pdf.

Marta Recasens, Eduard Hovy, M. Antònia Martí. 2011. Identity, Non-identity, and Near-identity: Addressing the complexity of coreference. *Lingua, 121*(6): 1138-1152. https://doi.org/10.1016/j.lingua.2011.02.004.

Marta Recasens, M. Antònia Martí, Constantin Orasan. 2012. Annotating Near-Identity from Coreference Disagreements. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, European Languages Resources Association, pages 165-172. http://www.lrec-conf.org/proceedings/lrec2012/pdf/674_Paper.pdf.

Zhiyi Song, Ann Bies, Justin Mott, Xuansong Li, Stephanie Strassel, Christopher Caruso. 2018. Cross-Document, Cross-Language Event Coreference Annotation Using Event Hoppers. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, European Languages Resources Association, pages 3535-3540. https://www.aclweb.org/anthology/L18-1558.

Amber Stubbs. 2013. "A Methodology for Using Professional Knowledge in Corpus Annotation." PhD Thesis. Brandeis University. http://amberstubbs.net/docs/AmberStubbs_dissertation.pdf.

William F. Styler IV, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C de Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova and James Pustejovsky. 2014. Temporal Annotation in the Clinical Domain. *Transactions of the Association for Computational Linguistics, 2*, pages 143–154. https://doi.org/10.1162/tacl_a_00172.