Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2223–2230
Marseille, 11–16 May 2020

Temporal Histories of Epidemic Events (THEE): A Case Study in Temporal Annotation for Public Health

Jingcheng Niu[1], **Victoria Ng**[2], **Gerald Penn**[1], **Erin E. Rees**[2]
[1]Department of Computer Science, University of Toronto
[2]Public Health Risk Sciences Division, Public Health Agency of Canada
{niu, gpenn}@cs.toronto.edu, {victoria.ng, erin.rees}@canada.ca

Abstract

We present a new temporal annotation standard, THEE-TimeML, and a corpus TheeBank enabling precise temporal information extraction (TIE) for event-based surveillance (EBS) systems in the public health domain. Current EBS must estimate the occurrence time of each event based on coarse document metadata such as document publication time. Because of the complicated language and narration style of news articles, estimated case outbreak times are often inaccurate or even erroneous. Thus, it is necessary to create annotation standards and corpora to facilitate the development of TIE systems in the public health domain to address this problem. We will discuss the adaptations that have proved necessary for this domain as we present THEE-TimeML and TheeBank. Finally, we document the corpus annotation process, and demonstrate the immediate benefit to public health applications brought by the annotations.

Keywords: Temporal Information Extraction, Event-based Surveillance System, TimeML, Public Heath

1. Introduction

Event-based surveillance systems gather publicly available news articles, blogs and social media posts, and then extract epidemic-related information to detect evidence of emerging health threats due to infectious diseases. Most EBS systems (Brownstein et al., 2008; Mawudeku et al., 2013; Odlum and Yoon, 2018; Jordan et al., 2019) must rely on coarse temporal information, such as the document fetch time or import time, to estimate the occurrence time of the events extracted. This compromise often leads to inaccuracies or even errors when determining the temporal information of individual events, because of the complicated layering of temporal references used in reportage about epidemics.

Meanwhile, fine-grained information extraction is a well-researched area in the information extraction community, especially for temporal information extraction (TIE), which is the primary focus of this work. Several data annotation standards have been developed, including TimeML[1] (Pustejovsky et al., 2005), ISO-TimeML (Pustejovsky et al., 2010), which is based upon it, and the clinical, domain-specific THYME standard (Styler et al., 2014b). Based on those annotation standards, SemEval has held multiple shared tasks (Verhagen et al., 2007; Verhagen et al., 2010; UzZaman et al., 2013; Bethard et al., 2015; Bethard et al., 2016). An abundance of TIE systems have been developed for the shared tasks, reaching near-human performance (Chang and Manning, 2012; Chambers, 2013; Lee et al., 2016; Strötgen et al., 2013; Lin et al., 2019).

We created an annotation standard (THEE-TimeML) and a corpus (TheeBank) based upon it to facilitate the development of TIE systems for the public health domain. The standard and the corpus are aimed at three primary use cases but should be comprehensive enough for future applications. The first immediate use case would be to extract and infer precise temporal information about public health events, such as infectious disease outbreaks, for EBS. Second, THEE and TheeBank could be used to extract previously unavailable epidemiological data for epidemic risk modelling. Finally, interactive timelines of disease outbreaks could be created to assist epidemiological researchers.

In this paper, we present: (1) a new temporal information annotation schema created for the public health domain based on well-established standards such as TimeML, ISO-TimeML, and THYME; (2) a detailed description of the adjustments made for adapting these standards to the public health domain; (3) a corpus of 394 temporally-annotated Zika virus (ZIKV)-related articles; (4) a detailed description of the entire annotation process; and (5) an exemplification (in Section 6.) of one potential use case of both the corpus and the standard.

2. News Articles Source

2.1. EIOS

Information retrieval plays an important role in creating better EBS systems. The Epidemic Intelligence from Open Sources (EIOS)[2] (Abdelmalik et al., 2018) initiative is a collaboration between the World Health Organization (WHO) and other health organizations from member states to consolidate different initiatives, networks, and systems aiming for early detection of threats to public health. EIOS inherits a massive database of open-sourced online articles from five EBS systems, namely the Global Public Health Intelligence Network (GPHIN) (Mawudeku et al., 2013), HealthMap (Brownstein et al., 2008), ProMED (Morse et al., 1996), the Europe Media Monitor (EMM) (Steinberger et al., 2013), and MediSys (Rortais et al., 2010), that include comprehensive metadata such as article import or fetch time, original URL and language. New metadata such

[1] In the rest of the paper, we are going to use TimeML to refer to the collection of TimeML-based general domain annotation standards.

[2] https://www.who.int/eios

as a list of mentioned countries in the article have been extracted as part of EIOS itself. Apart from compiling historical data, EIOS monitors the internet 24/7 and keeps populating the sizeable collection of epidemic-related articles. EIOS categorizes those articles using keyword patterns based on a taxonomy and ontology of the infectious disease.

Overall, the language of EIOS articles is similar to that of the news articles in TimeBank (Pustejovsky et al., 2003) since the majority of the sources are mainstream-media news articles. Those news articles usually use complex sentence structures for temporal information referencing. For example, in sentence (1), there are two time expressions (TIMEX3s) in a single sentence for two different events. *Friday* is the time when the health minister is reporting the event to the media, but *last week* is the actual time of the death of the patients. In sentence (2), a temporal relation algebra such as the Allen algebra (Allen, 1983) is needed to infer the actual appearance time of the symptoms.

(1) Colombia's health minister Alejandro Gaviria told reporters **Friday** that three patients with the syndrome died **last week** at a clinic in Medellin, the country's second-largest city.

(2) Health Minister S Subramaniam said in a statement Wednesday that the 27-year-old female was confirmed infected with Zika on **Sept. 6**, after developing a rash, fever and body aches **four days before**.

2.2. Article Filtering

Our project has used articles related to the Zika virus (ZIKV) outbreaks during their peak from February 2015 to October 2016. By applying the EIOS keyword pattern filter to the EIOS article feed, we obtained 155,583 ZIKV-related news articles within that period in which the language of reportage is labelled by metadata as English. Keyword-based filtering algorithms are insufficient, however, for removing irrelevant or low-quality articles. When we randomly sampled ten articles, five of them were irrelevant or poor translations of non-English-language originals. Therefore, we used an article clustering algorithm (Trstenjak et al., 2014) based upon Term Frequency-Inverse Document Frequency (`tf.idf`) scores to filter out low-quality or irrelevant articles.

The algorithm works as follows. First, we kept the initial ten randomly sampled articles, including the five problematic ones. This sample was then supplemented by 20 groups of 1000 EIOS articles, randomly sampled from those that passed a keyword-based filter, and all of these were clustered into seven classes. The class of articles with the largest number of usable probing samples was chosen. With this process, we collected 2778 articles in total. As the immediate purpose of this corpus is the advancement of data analysis tools for EBS systems, we applied another filter from Nasheri et al. (2019) that detects articles containing infection case-counting information. This narrowed the number of articles down to 483. During the annotation process, we further narrowed the number of articles down to 394 by requiring them to have at least one high-priority event according to the taxonomy described in Section 3.1.. There are a total of 8932 sentences in the resulting subsample.

3. Annotation Standard

As proven by the success of THYME, standards adaptations that utilize domain-specific knowledge could be crucial to improving the ease and quality of the annotation process, as well as the utility of the corpus in building applications. THYME adapted ISO-TimeML towards medical records, which are considerably different from the news genres that were the focus of AQUAINT, the programme that produced TimeML and TimeBank. Apart from the clinical knowledge added, a few annotation standard changes were made to reflect the medical-records domain, such as the removal of subordination links (SLINKs). As described in section 2.1., the language style of EIOS articles is in fact similar to that of the news articles in TimeBank. Therefore, we only implemented changes for the public health domain, which will be laid out in the following section.[3]

3.1. Determining Relevancy

Domain-specific annotation standards face the problem of handling out-of-domain content. TimeML aims for a general temporal annotation standard, and therefore any corpus based on TimeML, such as TimeBank, will consider every event in every news article to be relevant. THYME ignores entities "that persist throughout the relevant temporal period of the clinical timeline (*endurants* in ontological circles)" (Styler et al., 2014b). THEE must face an even greater proportion of data in news articles that serve no constructive purpose from an EBS perspective, such as the financial or political aspects of an outbreak, because the articles target more general audiences than medical records. It is common to observe both out-of-domain or background information in relevant articles as well as relevant information in out-of-domain articles. For instance the article [4] entitled *Zika Virus Symptoms and the Link of the Aedes Mosquitoes*, uses the first paragraph to introduce the history of the discovery of ZIKV in Uganda, and the last paragraph to address an advisory on ZIKV prevention. Those are ZIKV-related but are irrelevant to early warning and monitoring of the disease. Another article[5] entitled *Time to talk about abortion*, focuses on a debate on abortion but mentions a case of ZIKV infection in Trinidad. We therefore developed a ZIKV-specific taxonomy to quantify the relevance of EIOS articles. Events are categorized into the following four types:

- **High Priority:** Case counts of ZIKV, occurrences of symptoms of ZIKV (birth defects, fever), and vectors of transmission or other details concerning the spread of ZIKV;

[3] The THEE-TimeML annotation standard and the TheeBank Corpus will be publicly available at https://doi.org/10.5683/SP2/AXIMY1.

[4] http://www.ecanadanow.com/health/2016/01/24/zika-virus-symptoms-and-the-link-of-the-aedes-mosquitoes/

[5] https://trinidadexpress.com/news/local/time-to-talk-about-abortion/article_d0215c98-9d22-55b2-a9ce-21e09f344266.html

- **Low Priority:** Organizational reactions, research progress, long-term complications from the disease (Guillain-Barré Syndrome), treatment of ZIKV, prevention of ZIKV, and testing for ZIKV;
- **Irrelevant:** Politics, finance, sport, funding, and fundraising;
- **Relevant but Too Generic:** Advisory content, research details about generic individual infection processes, general background information of the diseases.

Articles with mostly irrelevant or generic content and no high priority events were discarded. Sections of articles that are irrelevant or generic and have no high priority events are ignored. There has been some work on automatically identifying generic content in the public health literature, e.g., Nasheri et al. (2019). Only events with high and low priority are labelled, but all time expressions are labelled regardless of context.

3.2. Adapting TimeML

Although THYME-TimeML (Styler et al., 2014b) says that its definition of events is identical to that of ISO-TimeML, the THYME corpus does drop events from its annotation both for the sake of "expediency and ease of annotation," and because of "the needs of the clinical domain." By contrast, we drop a greater percentage of events from our own corpus relative to what ISO-TimeML would have annotated, but are guided by the domain-specific taxonomy in Section 3.1. in determining when we do so.

There are other important deviations from ISO-TimeML than the definitions of events, too. As advised by Pustejovsky and Stubbs (2011), both THYME and THEE use a system of narrative containers that had not been introduced into the original ISO-TimeML standard. It greatly expands upon the more widespread practice of temporally relating events as occurring before, overlapping with or after the document's creation time, and includes the possibility of nesting containers, and anchoring containers to events instead of the DocTimeRel attribute, even in the absence of TIMEX3s. See Section 3.2.4. and also Styler et al. (2014a), Section 5. Both THYME and THEE also retain ALINKs (aspectual links, e.g., *X initiates Y*. On the other hand, THYME eschews SLINKs (subordination links), although many subordinated clauses are connected with temporal TLINKs when they do not involve *irrealis* or possible-world reasoning. THEE uses SLINKs.

The balance of our domain adaptations have been motivated by characteristics of the public health domain and the language of the EIOS articles that we sampled from.

3.2.1. Syntactic Heads and Event Span

To avoid tagging discontinuous sequences, TimeML applies a series of strategies to select words representative of a whole event. THYME summarizes the procedure and further restricts labelling to the best single-word, SYNTACTIC HEAD of an event. We initially followed this strategy but found that we needed to accommodate several of what the TempEval-2 version of TimeML (Verhagen et al., 2010) calls "complex event constructions." Therefore, we extended the taxonomic approach documented in TE2-TimeML. This standard first specifies the handling of several event-denoting expressions, sorted by the part of speech of their heads, and then the handling of several complex event constructions. We carried over the handling of the former without change, but made modifications to the handling of complex event constructions.

TE2-TimeML requires the annotation of six complex constructions: copulative, aspectual, inchoative, light verb, causal, or functional nouns. Although the THYME annotation guidelines do not mention functional noun or causal constructions,[6] they instruct annotators to ignore copulative and "semantically light" predications, which seem to include both light and inchoative verbs. During a pilot study, in which our annotators read these guidelines from THYME, we found that the instructions to ignore semantically light predications actually created ambiguities, because of the uncertainty surrounding what qualifies as "light." Sentence (3), for example, was not considered to be light and therefore labelled as the head of an event, but sentence (4) was regarded as light by some annotators, even though we (the authors) consider both sentences to be semantically equivalent. THYME also explicitly mentions that verbs of experiencing, as in Sentence (5), are included in the class of semantically light predications. This example, however, comes from an EIOS article. Note the attribution to an email message, which means that the source being paraphrased is itself a clinical record resembling the THYME domain. Tokens of *experience* in EIOS articles are nevertheless not always so easily ignored. In Sentence (6), for example, *continues to experience* bears important aspectual information in relation to the rate of new travel-related Zika cases that would preclude simply connecting *continues* and *number* (the meaning of which would be consistent with Florida having the same high number of cases this week as it had in the previous week).

(3) Two cases were confirmed on Tuesday.

(4) Two cases were found on Tuesday.

(5) Gambineri said in an email the cases of individuals in Wynwood experienced Zika symptoms in mid-July, prior to the start of an aerial spraying campaign.

(6) However, while Florida continues to experience a high number of travel-related Zika cases, ...

Our subsequent instructions, to read the enumeration of complex constructions in the TE2 guidelines and to ignore only copulative and inchoative predications, eliminated these disagreements. We nevertheless added two new types of constructions to the list that are common in EIOS articles:

- **Light nominalization:** This newly introduced construction refers to a situation (e.g., Sentence 7) in

[6] THYME uses ALINKs, and its guidelines do instruct annotators to label aspectual events.

which the head of an event phrase (*practice*) is semantically light but nominal, and a gerund *isolating* conveys the actual meaning of a policy or process (as opposed to the deverbalized noun *isolation*, which, without *practice*, might refer to specific cases in which patients had been isolated). Therefore, it should be handled in a way similar to the handling of light verb constructions that wrap meaning-bearing nouns (e.g., Sentence 8).

(7) We are reviewing the [**practice**] of [*isolating*] Zika-positive patients who are actually clinically well.

(8) ...[**taking**] into [*consideration*] the potential serious health consequences of Zika virus infection.

- **Comparative:** When it comes to reporting case counts, comparisons are often used to characterize the trend of infections. There are two primary situations in which comparatives are used: (1) when comparing two events (EVENT compare EVENT), as seen in example (9); (2) when a functional noun is followed by a change-of-state, an optional quantity, the word *compared*, and a TIMEX3 (N_{funct} change-of-state X *compared* to TIMEX3), as seen in example (10).

  The first situation can be annotated using the standard procedure of annotating and linking both events with a TLINK. The problem with the second situation is that there exists an event that is anchored on TIMEX3. For instance, in example (10), the dengue cases in the week before, (*March 20 to 26*), constitute an actual event. The TE2-TimeML annotation format implicitly does not allow event and TIMEX3 tags to overlap. In its handling of causal constructions, TE2 uses the word *cause* to anchor the logical subject when it is an entity without eventuality, as in Sentence (11). Inspired by that practice, we used the word *"compared"* to anchor the implied event that occurs during the TIMEX3.

(9) This year, 87 [*deaths*] were reported compared to [*108*] during the same period in 2015.

(10) The [*number*] of dengue cases in the country from March 27 to April 2 [*dropped*] by five cases as [*compared*] to [**the week before**].

(11) Twelve babies have been born in the U.S. with birth defects caused by Zika virus, ...

We likewise found that instructions to annotate causal constructions also avoided ambiguity, although in practice they have not been very common. Documenting functional noun constructions turns out to be very important in the public health domain, because they are often used in descriptions of case counts.

These changes to the specification were sufficient to handle the majority of our sampled articles without serious variance between annotators.

3.2.2. Event Attributes

TimeML specifies that each event has nine attributes: class, part-of-speech (POS), tense, aspect, polarity, modality, type, genericity, and cardinality. THYME added some attributes for the clinical domain in addition to combining or discarding some of these. Each event in THYME has seven attributes: DocTimeRel, Type, Polarity, Degree, ContextualModality, ContextualAspect, and Permanence. The attributes are used to capture different distinctions between events. Those distinctions could be syntactic, semantic, contextual, domain-specific, or used to anchor narrative containers. We discarded most of the syntactic and clinical attributes, and kept some of the semantic and contextual attributes: Polarity, Novelty, Genericity, and DocTimeRel. We also added the TE2 attribute, Cardinality, in view of its importance to case counting.

3.2.3. Event Classes

TimeML-compliant annotation maintains a set of seven classes: Reporting, Perception, Aspectual, I_Action, I_State, State, and Occurrence. THYME discarded the class hierarchy because of characteristics of clinical notes. But the news articles used by THEE contain complicated subordinated structures that imply temporal information. Therefore we decide to keep the class hierarchy but made two significant changes:

Occurrence vs. State TimeML drew a distinction between an Occurrence and a State. This distinction is often hard to make in epidemic-related news articles. For example, in sentence (12),

(12) An adolescent girl was *infected* after traveling to El Salvador in 2015.

the event *'infected'* could be regarded as a state of infection by the girl or could be considered as an occurrence of infection. We have found that this distinction was not significant for further algorithm development. For instance, in example (13), the event with *caught* is an occurrence, whereas in example (14), the event with *has* is a state. The two sentences should be labelled differently, but an epidemic monitoring EBS will scrutinize them in the same way. Therefore, we merged the two classes into one and used Occurrence as the name of the resulting class.

(13) The researchers confirmed that the girl *caught* Zika.

(14) The researchers confirmed that the girl *has* Zika.

Perception TimeML defines PERCEPTION as "...events involving the physical perception of another event." All of the perception words in the corpus, however, are metaphorical. For example, in the sentence (15),

(15) The Morris De Castro Clinic in Cruz Bay had seen services suspended beginning in mid-2012 because of budget shortfalls.

the word *seen* does not entail a physical perception, but rather something similar to *experienced* or *undergone*, and therefore should be classified as an I_ACTION. After a comprehensive search of words that might express a physical perception, we did not find any cases of true perception events. Therefore, the class was discarded.

3.2.4. TLINKs and Narrative Containers

Narrative containers (Pustejovsky and Stubbs, 2011) were introduced to increase informativeness and improve annotation accuracy. THYME adopted narrative containers for three primary benefits. The first is that using narrative containers can bind events in the same narrative together and can make temporal relations between events in different containers inferable without being explicitly annotated. The second is that narrative containers nicely fit the structure of story telling. Both the general domain (TimeML) and the clinical domain (THYME) cluster discussions about the same topic or in the same time period together. This pattern also holds in the public health domain. The third benefit is that narrative containers provide a framework to handle sub-events.

We adopted the use of narrative containers in THEE-TimeML. First, the annotators need to classify events into the three broad narrative containers relative to the document creation time: BEFORE, AFTER, and OVERLAP. In the clinical domain, however, there is much more concern over how long durative events such as symptoms have lasted, which necessitates a distinction between OVERLAP and BEFORE-OVERLAP. In the epidemiological domain, what matters more is the punctual emergence of symptoms. Therefore, we merged the two containers into a single OVERLAP. The remaining narrative container relations are specified by temporal links (TLINK). TimeML has 13 TLINKs based upon the interval calculus (Allen, 1983). THYME is able to reduce the number of TLINKs to five because of the simplicity brought by applying the narrative container.

Our own attempt so simplify the types of TLINKs yielded eight types: BEFORE, AFTER, OVERLAP_BEFORE, OVERLAP_AFTER, IS_INCLUDED, DURING, SIMULTANEOUS and IDENTITY, partly because of our decision to use the BRAT annotation tool (Stenetorp et al., 2012). ISINCLUDED is the symmetric dual of the THYME CONTAINS link, which we use because BRAT forces link points to originate from events only. The THYME links OVERLAP, BEGINS-ON and ENDS-ON are all covered by our OVERLAP_BEFORE and OVERLAP_AFTER. THYME has no explicit link for AFTER because it can directionally reverse its BEFORE links, but BRAT does not allow a link from a TIMEX3 to an event, and so we needed to add AFTER to describe occurrences of events after a TIMEX3. DURING was added back because EIOS articles have many durative TIMEX3s. IDENTITY and DURING are both very important for counting cases. For the same reason, SIMULTANEOUS must be distinguished from IDENTITY, and needs to appear in addition to BEFORE and AFTER in order to accurately reconstruct timelines.

Narrative containers have not been easy for our annotators to accommodate. In part, this is due to a general lack of exemplary annotated texts with NCs for us to use as a guide. The medical records in THYME contain personally identifiable information, and so are not publicly available, for example. Relations in medical records are also relatively less heterogeneous than in EIOS articles. Another potential reason is that narrative containers in TimeBank may generally be range-bound to relatively shorter time spans. In EIOS articles, there can be as many as 300 cases mentioned over a one-month period.

	Training Set		Test Set		Total
	Count	%	Count	%	Count
Articles	347	88.1	47	11.9	394
Tokens	195125	88.2	26079	11.8	221204

Table 1: Corpus overview

3.2.5. Event Collections

One particular dilemma that has not been easy to resolve concerns collections of events:

(16) Three of the country's five Zika patients caught the virus while on vacation.

We believe that the intention behind narrative containers has been to place all of the vacations in one container V and all of the catching events in a different container C. We, too, have followed this practice. Yet under the reading in which each of the three patients mentioned caught Zika on a different vacation, neither V nor C may refer to a continuous interval of time.

4. Annotation Process

The 394 articles were partitioned into 347 training articles and 47 test articles that were annotated by four annotators, including three linguistics graduate students and one computational linguistics undergraduate student. Each training set article was annotated by a single annotator, while two annotators annotated each test article in parallel. The annotators were given instructions and strategies about how to apply the annotation standard and used the BRAT annotation tool. The annotators were instructed to finish the training set annotation before moving on to the test set. They were allowed to discuss, ask questions, and provide feedback during the training set annotation process to ensure a common understanding of the standard. Periodically, the annotators were instructed to annotate articles collectively, and their inter-annotator agreement (IAA) was calculated to ensure the quality of the annotation.

No discussion was allowed during the annotation of the test set, in order to provide an accurate IAA. When there were any conflicts in annotations between two annotators annotating the same test set article, a third adjudicator was forced to reconcile the disagreement.

5. TheeBank Results

5.1. Corpus Statistics

We present here the statistics of the annotated corpus. Table 1 shows an overview of the corpus; table 2 shows the distribution of different types of TIMEX3s, events, and event attributes; and table 3 shows the distribution of different types of links.

	Training Set		Test Set		Total
	Count	%	Count	%	Count
TIMEX	2266	86.7	347	13.2	2005
Date	1740	86.8	265	30.0	60
Time	42	70.0	18	9.8	488
Duration	440	90.2	48	26.7	60
Set	44	73.3	16	13.3	2613
EVENTS	15504	92.2	1307	7.8	16811
Occurrence	10622	92.2	902	7.8	11524
IAction	2138	91.9	188	8.1	2326
Reporting	1617	91.8	145	8.2	1762
Aspectual	349	91.4	33	8.6	382
IState	778	95.2	39	4.8	817
Event Attributes	15553	91.9	1373	8.1	16926
After	895	92.3	75	7.7	970
Overlap	4419	83.9	845	16.1	5264
Before	7798	99.2	63	0.8	7861
Generic	1633	97.0	50	3.0	1683
Negative	553	96.0	23	4.0	576
Novel	255	44.6	317	55.4	572

Table 2: Distribution of TIMEX3s and Events

	Training Set		Test Set		Total
	Count	%	Count	%	Count
TLINK	11523	91.8	1033	8.2	12556
Before	4201	90.7	433	9.3	4634
After	149	91.4	14	8.6	163
OBefore	1200	95.5	57	4.5	1257
OAfter	117	90.0	13	10.0	130
Simultaneous	1127	90.9	113	9.1	1240
During	190	90.9	19	9.1	209
Identity	1973	93.8	130	6.2	2103
IsIncluded	2566	91.0	254	9.0	2820
SLINK	4964	91.7	449	8.3	5413
Modal	1619	95.0	85	5.0	1704
Evidential	2089	92.1	180	7.9	2269
Negative Evidential	44	95.7	2	4.3	46
Factive	855	83.7	167	16.3	1022
Counter Factive	268	97.5	7	2.5	275
Conditional	89	91.8	8	8.2	97
ALINK	351	91.4	33	8.6	384
Continues	123	93.9	8	6.1	131
Initiates	174	90.6	18	9.4	192
Reinitiates	5	71.4	2	28.6	7
Terminates	33	91.7	3	8.3	36
Culminates	16	88.9	2	11.1	18

Table 3: Distribution of Different Link Types.

5.2. Inter-Annotator Agreement

We follow the IAA evaluation protocol proposed by Styler et al. (2014b), namely an F1-score[7] for entity identification agreement and Krippendorff's Alpha (Krippendorff, 2018) for classification agreement evaluation. Table 4 shows the IAA of entity annotation, and table 5 shows the IAA of even-attribute annotation.

For link annotation IAA, we calculate the F1 score and Krippendorff's Alpha on links connecting entities that are recognized by both annotators. That is, if an entity is only recognized by one annotator, all the links associated with that entity will be ignored during link IAA calculation, because that part of the IAA is already accounted for during entity IAA evaluation. Table 6 shows the evaluation result of link annotation.

Annotation Type	F1	Alpha
EVENT	0.719	0.728
TIMEX3	0.780	0.809

Table 4: Entity IAA

Attribute	F1	Alpha
DocTimeRel	0.738	0.511
Generic	0.279	0.230
Polarity	0.914	0.911
Novel	0.450	0.442

Table 5: Entity attribute IAA

Link type	partic. only F1	partic.+type F1	Alpha
TLINK	0.716	0.648	0.852
SLINK	0.964	0.928	0.629
ALINK	0.968	0.968	1.0

Table 6: Link IAA by participant only and by participant and types.

5.3. Baseline Algorithm

ClearTK-TimeML (Bethard, 2013), a support vector machine (SVM) based algorithm from TempEval-3 (UzZaman et al., 2013), is used as a baseline system because it was evaluated on both TempEval-3 and THYME (table 7). The results cover the same tasks as defined for the IAA evaluation in section 5.2..

6. Case Counting Example

Infection case counting is one of the quintessential tasks of epidemic monitoring EBS systems, and a likely beneficiary of having a good TIE system. Without any available TIE system for the public health domain, research conducted using case counting data must assume that all cases occurred at the publication time of the article.

We labelled all of the case events in the test set articles. We found that the article publication time falls inside the events' inferred occurrence time interval in only 93 out of 152 (61.2%) cases, based on our annotation.

7. Future Work

This corpus solely focuses on ZIKV-related news articles. Therefore, the next steps of this work will be creating a new corpus based on THEE-TimeML that will contain articles covering multiple additional diseases to continue advancing the abilities of EBS system data extraction. Such a corpus can support the development and evaluation of transfer-learning-based (Pan and Yang, 2010) TIE systems, allowing EBS to monitor newly emerging diseases.

8. Acknowledgements

This study is funded by the Canadian Safety and Security Program (CSSP) from Defence Research and Development Canada (DRDC) and the Ontario Big data Research, Analytics and Information Network (BRAIN). We thank Emily Clare, Andrei Munteanu, Owen Sawyer, and Luke Zhou for

[7] This should properly be viewed as the harmonic mean of two opposite-facing recall scores, but the computation is the same as an asymmetric F1-score, in which one annotator's work is viewed as the true annotation. The practice of using F1 began with UzZaman and Allen (2011), who were justified in calling it this because they were evaluating system outputs against human-annotated gold standards.

	TempEval-3			THYME Corpus			TheeBank		
	P	R	F1	P	R	F1	P	R	F1
TIMEX3	83.2	71.7	77.0	59.3	42.8	49.7	79.7	74.6	77.1
Event	81.4	76.4	78.8	78.9	23.9	36.6	66.9	75.5	70.9
DocTimeRel	-	-	-	47.4	47.4	47.4	45.0	50.8	47.7
Link	28.6	30.9	26.6	22.7	18.6	20.4	18.1	25.1	21.0
Event-TIMEX3	-	-	-	32.3	60.7	42.1	40.6	19.7	26.5
Event-Event	-	-	-	7.0	3.0	4.2	11.6	30.2	16.7

Table 7: ClearTK-TimeML performance. For TheeBank, we used the TempEval-3 evaluation script. In the case of DocTimeRel, we accomplished this by renaming it as "part-of-speech," which is an event attribute that the script evaluates. The TIMEX3 and Event rows use strict matching, and the link rows are evaluated for temporal awareness (UzZaman and Allen, 2011). The THYME corpus numbers are quoted from Styler et al. (2014b), which we assume made these same decisions, except for DocTimeRel, for which a more complicated heuristic involving TLINKs was used.

their work on the annotation. We would also like to thank the Factiva, GPHIN and EIOS teams for providing us the source data and consultation.

9. Bibliographical References

Abdelmalik, P., Peron, E., Schnitzler, J., Fontaine, J., Elfenkampera, E., and Barbozaa, P. (2018). The Epidemic Intelligence from Open Sources initiative: A collaboration to harmonize and standardize early detection and epidemic intelligence among public health organizations. *Weekly Epidemiological Record*, 93(20):267–270, May.

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

Bethard, S., Derczynski, L., Savova, G., Pustejovsky, J., and Verhagen, M. (2015). SemEval-2015 Task 6: Clinical TempEval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 806–814, Denver, Colorado, June. Association for Computational Linguistics.

Bethard, S., Savova, G., Chen, W.-T., Derczynski, L., Pustejovsky, J., and Verhagen, M. (2016). SemEval-2016 Task 12: Clinical TempEval. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1052–1062, San Diego, California, June. Association for Computational Linguistics.

Bethard, S. (2013). ClearTK-TimeML: A minimalist approach to TempEval 2013. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 10–14, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Brownstein, J. S., Freifeld, C. C., Reis, B. Y., and Mandl, K. D. (2008). Surveillance Sans Frontières: Internet-Based Emerging Infectious Disease Intelligence and the HealthMap Project. *PLoS Medicine*, 5(7), July.

Chambers, N. (2013). NavyTime: Event and Time Ordering from Raw Text. Technical report, NAVAL ACADEMY ANNAPOLIS MD, June.

Chang, A. X. and Manning, C. (2012). SUTime: A library for recognizing and normalizing time expressions. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3735–3740, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Jordan, S. E., Hovet, S. E., Fung, I. C.-H., Liang, H., Fu, K.-W., and Tse, Z. T. H. (2019). Using Twitter for Public Health Surveillance from Monitoring and Prediction to Public Response. *Data*, 4(1):6, March.

Krippendorff, K. (2018). *Content Analysis: An Introduction to Its Methodology*. SAGE Publications, May.

Lee, H.-J., Xu, H., Wang, J., Zhang, Y., Moon, S., Xu, J., and Wu, Y. (2016). UTHealth at SemEval-2016 Task 12: An End-to-End System for Temporal Information Extraction from Clinical Notes. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1292–1297, San Diego, California, June. Association for Computational Linguistics.

Lin, C., Miller, T., Dligach, D., Bethard, S., and Savova, G. (2019). A BERT-based Universal Model for Both Within- and Cross-sentence Clinical Temporal Relation Extraction. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 65–71, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Mawudeku, A., Blench, M., Boily, L., John, R. S., Andraghetti, R., and Ruben, M. (2013). The Global Public Health Intelligence Network. In *Infectious Disease Surveillance*, pages 457–469. John Wiley & Sons, Ltd.

Morse, S. S., Rosenberg, B. H., and Woodall, J. (1996). ProMED global monitoring of emerging diseases: Design for a demonstration program. *Health Policy (Amsterdam, Netherlands)*, 38(3):135–153, December.

Nasheri, N., Vester, A., and Petronella, N. (2019). Foodborne viral outbreaks associated with frozen produce. *Epidemiology and Infection*, 147, October.

Odlum, M. and Yoon, S. (2018). Health Information Needs and Health Seeking Behavior During the 2014-2016 Ebola Outbreak: A Twitter Content Analysis. *PLoS currents*, 10, March.

Pan, S. J. and Yang, Q. (2010). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, October.

Pustejovsky, J. and Stubbs, A. (2011). Increasing Informativeness in Temporal Annotation. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 152–160, Portland, Oregon, USA, June. Association for Computational Linguistics.

Pustejovsky, J., Hanks, P., Saurí, R., See, A., Gaizauskas, R., Setzer, A., Radev, D., Sundheim, B., Day, D., Ferro, L., and Lazo, M. (2003). The TimeBank corpus. *Proceedings of Corpus Linguistics*, January.

Pustejovsky, J., Ingria, B., Sauri, R., Castano, J., Littman, J., Gaizauskas, R., Setzer, A., Katz, G., and Mani, I. (2005). The specification language TimeML. *The language of time: A reader*, pages 545–557.

Pustejovsky, J., Lee, K., Bunt, H., and Romary, L. (2010). ISO-TimeML: An International Standard for Semantic Annotation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

Rortais, A., Belyaeva, J., Gemo, M., van der Goot, E., and Linge, J. P. (2010). MedISys: An early-warning system for the detection of (re-)emerging food- and feed-borne hazards. *Food Research International*, 43(5):1553–1556, June.

Steinberger, R., Pouliquen, B., and van der Goot, E. (2013). An introduction to the Europe Media Monitor family of applications. *arXiv:1309.5290 [cs]*, September.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). Brat: A Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France, April. Association for Computational Linguistics.

Strötgen, J., Zell, J., and Gertz, M. (2013). HeidelTime: Tuning English and Developing Spanish Resources for TempEval-3. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 15–19, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Styler, W., Savova, G., Palmer, M., Pustejovsky, J., O'Gorman, T., and de Groen, P. C., (2014a). *THYME Annotation Guidelines*. The Temporal Histories of Your Medical Event (THYME) project, February.

Styler, W. F., Bethard, S., Finan, S., Palmer, M., Pradhan, S., de Groen, P. C., Erickson, B., Miller, T., Lin, C., Savova, G., and Pustejovsky, J. (2014b). Temporal Annotation in the Clinical Domain. *Transactions of the Association for Computational Linguistics*, 2:143–154.

Trstenjak, B., Mikac, S., and Donko, D. (2014). KNN with TF-IDF based Framework for Text Categorization. *Procedia Engineering*, 69:1356–1364, January.

UzZaman, N. and Allen, J. F. (2011). Temporal evaluation. In *Proceedings of the 49th Annual Meeting of the ACL*, pages 351–356.

UzZaman, N., Llorens, H., Derczynski, L., Allen, J., Verhagen, M., and Pustejovsky, J. (2013). SemEval-2013 Task 1: TempEval-3: Evaluating Time Expressions, Events, and Temporal Relations. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Verhagen, M., Gaizauskas, R., Schilder, F., Hepple, M., Katz, G., and Pustejovsky, J. (2007). SemEval-2007 Task 15: TempEval Temporal Relation Identification. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 75–80, Prague, Czech Republic, June. Association for Computational Linguistics.

Verhagen, M., Saurí, R., Caselli, T., and Pustejovsky, J. (2010). SemEval-2010 Task 13: TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 57–62, Uppsala, Sweden, July. Association for Computational Linguistics.