Temporal Anchoring of Events for the TimeBank Corpus

Nils Reimers†‡, Nazanin Dehghani†‡ [*], Iryna Gurevych†‡
†Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science, Technische Universität Darmstadt
‡ Research Training Group AIPHES
Technische Universität Darmstadt
http://www.ukp.tu-darmstadt.de

Abstract

Today's extraction of temporal information for events heavily depends on annotated temporal links. These so called TLINKs capture the relation between pairs of event mentions and time expressions. One problem is that the number of possible TLINKs grows quadratic with the number of event mentions, therefore most annotation studies concentrate on links for mentions in the same or in adjacent sentences. However, as our annotation study shows, this restriction results for 58% of the event mentions in a less precise information when the event took place.

This paper proposes a new annotation scheme to anchor events in time. Not only is the annotation effort much lower as it scales linear with the number of events, it also gives a more precise anchoring when the events have happened as the complete document can be taken into account. Using this scheme, we annotated a subset of the TimeBank Corpus and compare our results to other annotation schemes. Additionally, we present some baseline experiments to automatically anchor events in time. Our annotation scheme, the automated system and the annotated corpus are publicly available.[1]

1 Introduction

In automatic text analysis, it is often important to precisely know when an event occurred. A user might be interested in retrieving news articles that discuss certain events which happened in a given time period, for example articles discussing car bombings in the 1990s. The user might not only be interested in articles from that time period, but also in more recent articles that cover events from that period. Knowing when an event happened is also essential for time aware summarization, automated timeline generation as well as automatic knowledge base creation. In many cases, time plays a crucial role for facts stored in a knowledge base, for example for the facts when a person was born or died. Also, some facts are only true for a certain time period, like being the president of a country. Event extraction can be used to automatically infer many facts for knowledge bases, however, to be useful, it is crucial that the date when the event happened can precisely be extracted.

The TimeBank Corpus (Pustejovsky et al., 2003) is a widely used corpus using the TimeML specifications (Saurí et al., 2004) for the annotations of event mentions and temporal expressions. In order to anchor events in time, the TimeBank Corpus uses the concept of temporal links (TLINKs) that were introduced by Setzer (2001). A TLINK states the temporal relation between two events or an event and a time expression. For example, an event could happen *before*, *simultaneous*, or *after* a certain expression of time. The TimeBank Corpus served as dataset for the shared tasks TempEval-1, 2 and 3 (Verhagen et al., 2007; Verhagen et al., 2010; UzZaman et al., 2013).

In this paper we describe a new approach to anchor every event in time. Instead of using temporal links between events and temporal expressions, we consider the event time as an argument of the event mention. The annotators are asked to write down the date when an event happened in a normalized format for every event mention. The annotation effort is for this reason identical

[*] Guest researcher from the School of Electrical and Computer Engineering, University of Tehran.

[1] https://www.ukp.tu-darmstadt.de/data/timeline-generation/temporal-anchoring-of-events/

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 2195–2204,
Berlin, Germany, August 7-12, 2016. ©2016 Association for Computational Linguistics

to the number of event mentions, i.e. for a document with 200 event mentions, the annotators must perform 200 annotations. When annotating the event mentions, the annotators are asked to take the complete document into account. Section 3 presents our annotation scheme, and section 4 gives details about the conducted annotation study.

The number of possible TLINKs scales quadratic with the number of events and temporal expressions. Some documents of the TimeBank Corpus contain more than 200 events and temporal expressions, resulting in more than 20.000 possible TLINKs. Hand-labeling all links is extremely time-consuming and even when using transitive closures and computational support, it is not feasible to annotate all possible TLINKs for a larger set of documents. Therefore, all annotation studies limited the number of TLINKs to annotate. For example, in the original TimeBank Corpus, only links that are salient were annotated. Which TLINKs are salient is fairly vague and results in a comparably low reported inter-annotator agreement. Furthermore, around 62% of all events do not have any attached TLINK, i.e. for most of the events in the original TimeBank Corpus, no temporal statement can be made.

In contrast to the sparse annotation of TLINKs used in the TimeBank Corpus, the TimeBank-Dense Corpus (Cassidy et al., 2014) used a dense annotation and all temporal links for events and time expressions in the same sentence and in directly succeeding sentences were annotated. For a subset of 36 documents with 1729 events and 289 time expressions, they annotated 12,715 temporal links, which is around 6.3 links per event and time expression. Besides the large effort needed for a dense annotation, a major downside is the limitation that events and time expressions must be in the same or in adjacent sentences. Our annotation study showed that in 58.72% of the cases the most informative temporal expression is more than one sentence apart from the event mention. For around 25% of the events, the most informative temporal expression is even five or more sentences away. Limiting the TLINKs to pairs that are at most one sentence apart poses the risk that important TLINKs are not annotated and consequently cannot be learned by automated systems.

A further drawback of TLINKs is that it can be difficult or even impossible to encode temporal information that originates from different parts in the text. Given the sentence:

> *December 30th, 2015 - During New Year's Eve, it is traditionally very busy in the center of Brussels and people gather for the fireworks display. But the upcoming [display]$_{Event}$ was canceled today due to terror alerts.*

For a human it is simple to infer the date for the event *display*. But it is not possible to encode this knowledge using TLINKs, as the date is not explicitly mentioned in the text.

To make our annotations comparable to the dense TLINK annotation scheme of the TimeBank-Dense Corpus (Cassidy et al., 2014), we annotated the same documents and compare the results in section 5. For 385 out of 872 events (44.14%), our annotation scheme results in a more precise value on which date an event happened.

Section 6 presents a baseline system to extract event times. For a subset of events, it achieves an F_1-score of 49.01% while human agreement for these events is 80.50%.

2 Previous Annotation Work

The majority of corpora on events uses sparse temporal links (TLINKs) to enable anchoring of events in time. The original TimeBank (Pustejovsky et al., 2003) only annotated salient temporal relations. The subsequent TempEval competitions (Verhagen et al., 2007; Verhagen et al., 2010; UzZaman et al., 2013) are based on the original TimeBank annotations, but tried to improve the coverage and added some further temporal links for mentions in the same sentence. The MEANtime corpus (van Erp et al., 2015) applied a sparse annotation and only temporal links between events and temporal expressions in the same and in succeeding sentences were annotated. The MEANtime corpus distinguishes between main event mentions and subordinated event mentions, and the focus for TLINKs was on main events.

More dense annotations were applied by Bramsen et al. (2006), Kolomiyets et al. (2012), Do et al. (2012) and by Cassidy et al. (2014). While Bramsen et al., Kolomiyets et al., and Do et al. only annotated some temporal links, Cassidy et al. annotated all Event-Event, Event-Time, and Time-Time pairs in the same sentence as well as in directly succeeding sentences leading to the densest annotation for the TimeBank Corpus.

A drawback of the previous annotation works is the limitation that only links between expressions in the same or in succeeding sentences are annotated. In case the important temporal expression, that defines when the event occurred, is more than one sentence away, the TLINK will not be annotated. Consequently, retrieving the information when the event occurred is not possible. Increasing this window size would result in a significantly increased annotation effort as the number of links grows quadratic with the number of expressions.

Our annotation is the first for the TimeBank Corpus that does not try to annotate the quadratic growing number of temporal links. Instead, we consider the event time as an argument of the individual event mention and it is annotated directly by the annotators. This reduces the annotation effort by 85% in comparison to the TimeBank-Dense Corpus. This allows an annotator to annotate significant more documents in the same time. Also, all temporal information, independent where it is mentioned in the document, can be taken into account resulting in a much more precise anchoring of events in time, as section 5 shows.

3 Event Time Annotation Scheme

The annotation guidelines for the TimeBank Corpus (Saurí et al., 2004) define an *event* as a cover term for situations that *happen* or *occur*. Events can be *punctual* or last for *a period of time*. They also consider as events those predicates describing *states* or *circumstances* in which something holds true. For the TimeBank Corpus, the smallest extent of text (usually a single word) that expresses the occurrence of an event is annotated.

The aspectual type of the annotated events in the TimeBank Corpus can be distinguished into *achievement events*, *accomplishment events*, and *states* (Pustejovsky, 1991). An achievement is an event that results into an instantaneous change of some sort. Examples of *achievement events* are *to find*, *to be born*, or *to die*. *Accomplishment events* also result into a change of some sort, however, the change spans over a longer time period. Examples are *to build something* or *to walk somewhere*. States on the other hand do not describe a change of some sort, but that something holds true for some time, for example, *being sick* or *to love someone*. The aspectual type of an event does not only depend on the event itself, but also on the context in which the event is expressed.

Our annotation scheme was created with the goal of being able to create a knowledge base from the extracted events in combination with their event times. Punctual events are a single dot on the time axis while events that last for a period of time have a begin and an end point. It can be difficult to distinguish between punctual events and events with a short duration. Furthermore, the documents typically do not report precise starting and ending times for events, hence we decided to distinguish between events that happened at a *Single Day* and *Multi-Day Events* that span over multiple days. We used days as the smallest granularity for the annotation as none of the annotated articles contained any information on the hour, the minute or the second when the event happened. In case a corpus contains this information, the annotation scheme could be extended to include this information as well.

For *Single Day Events*, the event time is written in the format *YYYY-MM-DD*. For *Multi-Day Events*, the annotator annotates the *begin point* and the *end point* of the event. In case no statement can be made on when an event happened, the event will be annotated with the label *not applicable*. This applies only to 0.67% of the annotated events in the TimeBank Corpus which is mainly due to annotation errors in the TimeBank Corpus.

> *He was* ***sent*** *into space on May 26, 1980. He* ***spent*** *six days aboard the Salyut 6 spacecraft.*

The first event in this text, ***sent***, will be annotated with the event time *1980-05-26*. The second event, ***spent***, is a Multi-Day Event and is annotated with the event time *beginPoint=1980-05-26* and *endPoint=1980-06-01*.

In case the exact event time is not stated in the document, the annotators are asked to narrow down the possible event time as precisely as possible. For this purpose, they can annotate the event time with *after YYYY-MM-DD* and *before YYYY-MM-DD*.

> *In 1996 he was* ***appointed*** *military attache at the Hungarian embassy in Washington. [...] McBride was* ***part*** *of a seven-member crew aboard the Orbiter Challenger in October 1984*

The event ***appointed*** is annotated *after 1996-01-01 before 1996-12-31* as the event must have happened sometime in 1996. The Multi-Day Event

part is annotated with *beginPoint=after 1984-10-01 before 1984-10-31* and *endPoint=after 1984-10-01 before 1984-10-31*.

To speed up the annotation process, annotators were allowed to write *YYYY-MM-xx* to express that something happened sometime within the specified month and *YYYY-xx-xx* to express that the event happened sometime during the specified year. Annotators were also allowed to annotate events that happened at the Document Creation Time with the label *DCT*.

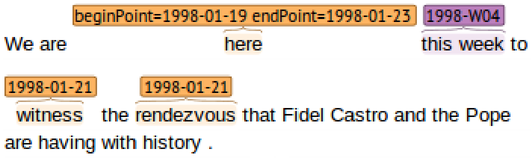
The proposed annotation scheme requires that event mentions are already annotated. For our annotation study we used the event mentions that were already defined in the TimeBank Corpus. In contrast to the annotation of TLINKs, temporal expressions must not be annotated in the corpus.

4 Annotation Study

The annotation study was performed on the same subset of documents as used by the TimeBank-Dense Corpus (Cassidy et al., 2014) with the event mentions that are present in the TempEval-3 dataset (UzZaman et al., 2013). Cassidy et al. selected 36 random documents from the TimeBank Corpus (Pustejovsky et al., 2003). These 36 documents include a total of 1498 annotated events. This allows to compare our annotations to those of the TimeBank-Dense Corpus (see section 5).

Each document has been independently annotated by two annotators according to the annotation scheme introduced above. We used the freely available WebAnno (Yimam et al., 2013). To speed up the annotation process, the existent temporal expressions that are defined in the TimeBank Corpus were highlighted. These temporal expressions are in principle not required to perform our annotations, but the highlighting of them helps to determine the event time. Figure 1 depicts a sample annotation made by WebAnno. The two annotators were trained on 15 documents distinct from the 36 documents annotated for the study. During the training stage, the annotators discussed the decisions they have made with each other.

After both annotators completed the annotation task, the two annotations were curated by one person to derive one final annotation. The curator examined the events where the annotators disagreed and decided on the final annotation. The final annotation might be a merge of the two provided annotations.



beginPoint=1998-01-19 endPoint=1998-01-23 1998-W04
We are here this week to
1998-01-21 1998-01-21
witness the rendezvous that Fidel Castro and the Pope
are having with history .

Figure 1: Sample Annotation made with WebAnno. The violet annotations are existing annotations of temporal expressions from the TimeBank Corpus. The span for the beige annotations, the event mentions, come also from the TimeBank Corpus. Our annotators added the value for the event time for those beige annotations.

4.1 Inter-Annotator-Agreement

We use Krippendorff's α (Krippendorff, 2004) with the nominal metric to compute the Inter-Annotator-Agreement (IAA). The nominal metric considers all distinct labels equally distant from one another, i.e. partial agreement is not measured. The annotators must therefore completely agree.

Using this metric, the Krippendorff's α for the 36 annotated documents is $\alpha = 0.617$. Cassidy et al. (2014) reported a Kappa agreement between $0.56 - 0.64$ for their annotation of TLINKs. Comparing these numbers is difficult, as the annotation tasks were different. According to Landis and Koch (1977), these numbers lie on the border of a moderate and a substantial level of agreement.

4.2 Disagreement Analysis

In 648 out of 1498 annotated events, the annotators disagreed on the event time. In 42.3% of the disagreements, the annotators disagreed on whether the event mention is a Single Day Event or a Multi-Day Event. Such disagreement occurs when it is unclear from the text whether the event lasted for one or for several days. For example, an article reported on a meeting and due to a lack of precise temporal information in the document, one annotator assumed that the meeting lasted for one day, the other that it lasted for several days. A different source for the disagreement has been the annotation of states. They can either be annotated with the date where the text gives evidence that they hold true, or they can be annotated as a Multi-Day Event that begins before that date and ends after that date.

Different annotations for Multi-Day Events account for 231 out of the 648 disagreements (35.6%). In this category, the annotators disagreed on the begin point in 110 cases (47.6%), on the end point in 57 cases (24.7%) and on the begin as well as on the end point in 64 cases (27.7%). The Krippendorff's α for all begin point annotations is 0.629 and for all end point annotations it is 0.737.

A disagreement on Single Day Events was observed for 143 event mentions and accounts for 22.1% of the disagreements. The observed agreement for Single Day Events is 80.5% or $\alpha = 0.799$. Most disagreements for Single Day Events were whether the event occurred on the same date as the document was written or if it occurred before the document was written.

4.3 Measuring Partial Agreement

One issue of the strict nominal metric is that it does not take partial agreement into account. In several cases, the two annotators agreed in principle on the event time, but might have labeled it slightly differently. One annotator might have taken more clues from the text into account to narrow down when an event has happened. One annotator for example, has annotated an event with the label *after 1998-08-01 before 1998-08-31*. The second annotator has taken an additional textual clue into account, which was that the event must have happened in the first half of August 1998 and annotated it as *after 1998-08-01 before 1998-08-15*. Even though both annotators agree in principle, when using the nominal metric it would be considered as a distinct annotation.

To measure this effect, we created a relaxed metric to measure mutual exclusivity:

$$d_{ME}(a,b) = \begin{cases} 1 \text{ if } a \text{ and } b \text{ are mutual exclusive} \\ 0 \text{ else} \end{cases}$$

The metric measures whether two annotations can be satisfied at the same time. Given the event happened on August 5th, 1998, then the two annotations *after 1998-08-01 before 1998-08-31* and *after 1998-08-01 before 1998-08-15* would both be satisfied. In contrast, the two annotations *after 1998-02-01* and *before 1997-12-31* can never be satisfied at the same time and are therefore mutual exclusive.

Out of the 648 disagreements, 71 annotations were mutually exclusive. Computing the Krippendorff's α with the above metric yields a value of $\alpha_{ME} = 0.912$.

4.4 Annotation Statistics

Table 1 gives an overview of the assigned labels. Around 58.21% of the events are either instantaneous events or their duration is at most one day. 41.12% of the events are Multi-Day Events that take place over multiple days. While for Single Day Events there is a precise date for 55.73% of the events, the fraction is much lower for Multi-Day Events. In this category, only in 19.81% of the cases the begin point is precisely mentioned in the article and only in 15.75% of the cases, the end point is precisely mentioned.

The most prominent label for Single Day Events is the Document Creation Time (DCT). 48.28% of Single Day Events happened on the day the article was created, 33.49% of these events happened at least one day before the DCT and 17.43% of the mentions refer to future events. This distribution shows, that the news articles and TV broadcast transcripts from the TimeBank Corpus mainly report on events that happened on the same day.

For Multi-Day Events, the distribution looks different. In 76.46% of the cases, the event started in the past, and in 65.10% of the cases, it is still ongoing.

4.5 Most Informative Temporal Expression

Not all temporal expressions in a text are of the same relevance for an event. In fact, in many cases only a single temporal expression is of importance, which is the expression stating when the event occurred. Our annotations allow us to determine *most informative* temporal expression for an event. We define the most informative temporal expression as the expression that has been used by the annotator to determine the event time. We checked for all annotations whether the event date can be found as a temporal expression in the document and computed the distance to the closest one with a matching value. The distance is measured as the number of sentences. 421 out of 1498 events happened on the Document Creation Time and were excluded from this computation. The Document Creation Time is provided as additional metadata in the TimeBank Corpus, and it is often not explicitly mentioned in the document text.

Figure 2 shows the distance between the most informative temporal expression and the event mention. In 23.68% of the cases, the time expression is in the same sentence, and in 17.59% of the cases, the time expression is either in the

	# Events	%
Single Day Events	**872**	**58.21%**
with precise date	486	55.73%
after + before	145	16.63%
after	124	14.22%
before	117	13.42%
past events	292	33.49%
events at DCT	421	48.28%
future events	152	17.43%
Multi-Day Events	**616**	**41.12%**
precise begin point	122	19.81%
precise end point	97	15.75%
begins in the past	471	76.46%
begins on the DCT	38	6.17%
begins in the future	105	17.05%
ends in the past	179	29.06%
ends on the DCT	26	4.22%
ends in the future	401	65.10%
Not applicable	**10**	**0.67%**

Table 1: Statistic on the annotated event times. Single Day Events happen on a single day, Multi-Day Events take place over multiple days. The event time can either be precise or the annotators used *before* and *after* to narrow down the event time, e.g. the event has happened in a certain month and year. DCT = Document Creation Time.

next or in the previous sentence. It follows that in 58.72%, of the cases the most informative time expression cannot be found in the same or in the preceding or succeeding sentence. This is important to note, as previous shared tasks like TempEval-1,-2, and -3 (Verhagen et al., 2007; Verhagen et al., 2010; UzZaman et al., 2013) and previous annotation studies like the TimeBank-Dense Corpus (Cassidy et al., 2014) only considered the relation between event mentions and temporal expressions in the same and in adjacent sentences. However, for the majority of events, the most informative temporal expression is not in the same or in the preceding / succeeding sentence.

For 7.31% of the annotated events, no matching temporal expression was found in the document. Those were mainly events where the event time was inferred by the annotators from multiple temporal expressions in the document. An example is that the year of the event was mentioned in the beginning of the document and the month of the event was mentioned in a later part of the document.

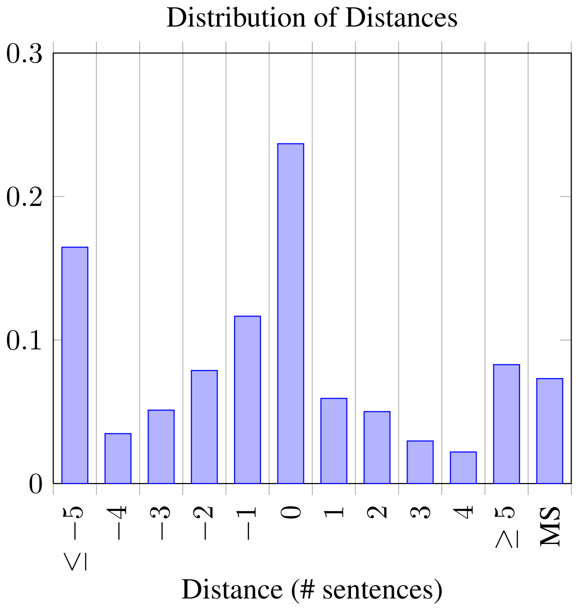


Figure 2: Distribution of distances in sentences between the event mention and the most informative temporal expression. For 58.72% of the event mentions, the most informative time expression is not in the same or in the previous/next sentence. For 7.3% of the mentions, the time expression originates from multiple sources (MS).

5 Comparison of Annotation Schemes

Depending on the application scenario and the text domain, the use of TLINKs or the proposed annotation scheme may be advantageous. TLINKs have the capability to capture the temporal order of events, even when temporal expressions are completely absent in a document, which is often the case for novels. The proposed annotation scheme has the advantage that temporal information, independent where and in which form it is mentioned in the document, can be taken into account. However, the proposed scheme requires that the events can be anchored on a time axis, which is easy for news articles and encyclopedic text but hard for novels and narratives.

In this section, we evaluate the application scenario of temporal knowledge base population and time-aware information retrieval. For temporal knowledge base population, it is important to derive the date for facts and events as precisely as possible (Surdeanu, 2013). Those facts can either be instantaneous, e.g. a person died, or they can last for a longer time like a military conflict.

Similar requirements are given for time-aware information retrieval, where it can be important to know at which point in time something occurred (Kanhabua and Nørvåg, 2012).

We use the TimeBank-Dense Corpus (Cassidy et al., 2014) with its TLINKs annotations and compare those to our event time annotations. The TimeBank-Dense Corpus annotated all TLINKs between Event-Event, Event-Time, and Time-Time pairs in the same sentence and between succeeding sentences as well as all Event-DCT and Time-DCT pairs. Six different link types were defined: `BEFORE`, `AFTER`, `INCLUDES`, `IS_INCLUDED`, `SIMULTANOUS`, and `VAGUE`, where `VAGUE` encodes that the annotators where not able to make a statement on the temporal relation of the pair.

We studied how well the event time is captured by the dense TLINK annotation. We used transitive closure rules as described by Chambers et al. (2014) to deduct also TLINKs for pairs that were not annotated. For example, when $event_1$ happened before $event_2$ and $event_2$ happened before $date_1$, we can infer that $event_1$ happened before $date_1$. Using this transitivity allows to infer relations for pairs that are more than one sentence apart. For all annotated events, we evaluated all TLINKs, including the TLINKs inferred from the transitivity rules, and derived the event time as precisely as possible. We then computed how precise the inferred event times are in comparison to our annotations. Preciseness is measured in the number of days. An event that is annotated with *1998-02-13* has the preciseness of 1 day. If the inferred event time from the TLINKs is *after 1998-02-01 and before 1998-02-15*, then the preciseness is 15 days. A more precise anchoring is preferred.

The TimeBank-Dense Corpus does not have a link type to mark that an event has started or ended at a certain time point. This makes the TLINK annotation impractical for the durative events that span over multiple days. According to our annotation study, 41.12% of the events in the TimeBank Corpus last for longer time periods. For these 41.12%, it cannot be inferred from when to when the events lasted.

In 487 out of the 872 Single Day Events (55.85%), the TLINKs give a result with the same precision as our annotations. For 198 events (22.71%), our annotation is more precise, i.e. the time window where the event might have happened is smaller. For 187 events (21.44%), no event time could be inferred from the TLINKs. This is due to the fact that there was no link to any temporal expression even when transitivity was taken into account.

For the 487 events where the TLINKs resulted in an event time as precise as our annotation, the vast majority of them were events that happened at the Document Creation Time. As depicted in Table 1, 421 events happened at DCT. For those events the precise date can directly be derived from the annotated link between each event mention and the DCT. For all other events that did not happen at the Document Creation Time, the TLINKs result for the most cases in a less precise anchoring in time and for around a fifth of these cases in no temporal anchoring at all while we do anchor them.

We can conclude, that even a dense TLINK annotation gives suboptimal information on when events have happened, and due to the restriction that TLINKs are only annotated in the same and in adjacent sentences, a lot of relevant temporal information gets lost.

6 Automated Event Time Extraction

In this section, we present a baseline system for automatic event time extraction. The system uses temporal relations in which the event is involved and anchors the event to the most precise time. For this purpose, we have defined a two-step process to determine the events' time. Given a set of documents in which the events and time expressions are already annotated, the system first obtains a set of possible times for each of the events. Second, the most precise time is selected or generated for each event.

For the first step, we use the multi-pass architecture introduced by Chambers et al. (2014) that was trained and evaluated on the TimeBank-Dense Corpus (Cassidy et al., 2014). Chambers et al. describe multiple rules and machine learning based classifiers to extract relations between events and temporal expressions. This architecture extracts temporal relations of the type `BEFORE`, `AFTER`, `INCLUDES`, `IS_INCLUDED`, and `SIMULTANOUS`. The classifiers are combined into a precision-ranked cascade of sieves. The architecture presented by Chambers et al. does not produce temporal information that an event has started or ended at a certain time point and can therefore only be used for Single Day Events.

We use these sieves to add the value of the temporal expression and the corresponding relation to a set of possible times for each event. In fact, for each event we generate a set of <`relation, time`> tuples in which the event is involved.

> *Police* ***confirmed Friday*** *that the body found along a highway*

For example, the one sieve adds [`IS_INCLUDED`, $Friday_{1998-02-13}$] and a second sieve adds [`BEFORE`, $DCT_{1998-02-14}$] to the set of possible event times for the *confirmed* event.

Applying the sequence of the sieves will obtain all various temporal links for each event. In the next step, if the event has a relation of type `SIMULTANEOUS`, `IS_INCLUDED` or `INCLUDES`, the system sets the event time to the value of the time expression. If the event has a relation of type `BEFORE` and/or `AFTER`, the system narrows down the event time as precisely as possible. If the sieve determines the relation type as `VAGUE`, the set of possible event times remains unchanged.

Algorithm 1 demonstrates how the event time is selected or generated from a set of possible times.

Algorithm 1 Automatic Event Time Extraction

```
function EventTime(times)
    if times is empty then
        return 'Not Available'                ▷ the event has no non-vague relation
    end if
    min_before_time = DATE.MAX_VALUE
    max_after_time = DATE.MIN_VALUE
    for [relation, time] in times do
        if relation is SIMULTANEOUS or IS_INCLUDED or INCLUDES then
            return time
        else if relation is BEFORE and time < min_before_time then
            min_before_time = time
        else if relation is AFTER and time > max_after_time then
            max_after_time = time
        end if
    end for
    event_time = AFTER + max_after_time + BEFORE + min_before_time
    return event_time
end function
```

Applying the proposed method on the TimeBank-Dense Corpus, we obtained some value for the event time for 593 of 872 (68%) Single Day Events. For 359 events (41%), the system generates the event time with the same precision as our annotations. Table 2 gives statistics of the automatically obtained event times.

Single Day Events	# Events	%
with precise date	260	29.82%
after + before	16	1.84%
after	99	11.35%
before	218	25%
not available	279	31.99%

Table 2: Statistics on the automatically obtained event times for events happened on a single day. The obtained event time can either be precise or the system used *before* and *after* to narrow down the event time. For 279 events, the system cannot infer any event time.

To evaluate the output of the proposed system, we evaluated how precise the automatically obtained event times are in comparison with our annotations. Table 3 shows for 41% of events, the proposed system generates the same event time as our annotations. For 21% events our annotation is more precise, i.e. the time window where the event might have happened is smaller. For 47 events (5.38%), the system infers an event time that is in conflict with the human annotation, for example a disagreement if an event happened before or after DCT. Considering event times that have the same preciseness as our annotations as true positives, the precision of the proposed system is 60.54% and the recall is 41.17% for Single Day Events. As presented in section 4, human annotators agree in 80.50% of the cases on the label for Single Day Events. The less precise and non-inferred event times are mainly due to the fact that temporal expressions, that are more than one sentence apart, are not taken into account by the sieve architecture.

Obtained event time	# Events	%
same as human annotation	359	41.17%
less precise	187	21.44%
conflicting annotations	47	5.38%
cannot infer event time	279	31.99%
Precision		**60.54%**
Recall		**41.17%**
F$_1$-Score		**49.01%**
Human F$_1$-Score		**80.50%**

Table 3: Evaluation results of proposed system in comparison with our annotations.

In this work we focused on the automated anchoring of Single Day Events and presented a baseline system that relies on the work of Chambers et al. (2014). The F_1-score with 49.01% is in comparison to the human score of 80.50% comparatively low. However, only in 5.38% of the cases, the automatically inferred event time is plain wrong. In the most cases, no event time could be inferred (31.99%) or it was less precise than the human annotation (21.44%).

Extending the described approach to Multi-Day-Events is not straight forward. The TimeBank-Dense Corpus and consequently the system by Chambers et al. does not include a TLINK type to note that an event has started or ended at a certain date, hence, extracting the begin point and end point for Multi-Day-Events is not possible. A fundamental adaption of the system by Chambers et al. would be required.

In contrast to Single Day Events, extracting the event time for Multi-Day Events requires more advanced logic. The start date of the event must be before the end date of the event. The relation to events that are included in the Multi-Day Events must be checked to avoid inconsistencies. The development of an automated system for Multi-Day Events is subject of our ongoing work.

7 Conclusion

We presented a new annotation scheme for anchoring events in time and annotated a subset of the TimeBank Corpus (Pustejovsky et al., 2003) using this annotation scheme. The annotation guidelines as well as the annotated corpus are publicly available.[2] In the performed annotation study, the Krippendorff's α inter-annotator agreement was considerably high at $\alpha = 0.617$. The largest disagreement resulted from events in which it was not explicitly mentioned when the event happened. Using a more relaxed measure for Krippendorff's α which only assigns a distance to mutual exclusive annotations, the agreement changed to $\alpha_{ME} = 0.912$. We can conclude that after little training, annotators are able to perform the annotation with a high agreement.

The effort for annotating TLINKs on the other hand scales quadratic with the number of events and temporal expressions. This imposes the often used restriction that only temporal links between events and temporal expressions in the same or in succeeding sentences are annotated. Even with this restriction, the annotation effort is quite significant, as on average 6.3 links per mention must be annotated. As Figure 2 depicts, in more than 58.72% of the cases the most informative temporal expression is more than one sentence apart from the event mention. As a consequence, inferring from TLINKs when an event happened is less precise as temporal information that is more than one sentence away can often not be taken into account.

For the 872 Single Day Events, the correct event time could be inferred from the TLINKs only in 487 cases. For 187 Single Day Events, no event time at all could be inferred, as no temporal expression was within the one sentence window of that event.

A drawback of the proposed scheme is the lack of temporal ordering of events beyond the smallest unit of granularity, which was in our case one day. The scheme is suitable to note that several events occurred at the same date, but their order on that date cannot be encoded. In case the temporal ordering is important for the application scenario, the annotation scheme could be extended and TLINKs could be annotated for events that fall on the same date. Another option is to increase the granularity, but this requires that the information in the documents also allow this more precise anchoring.

Acknowledgement

This work has been supported by the German Research Foundation as part of the Research Training Group Adaptive Preparation of Information from Heterogeneous Sources (AIPHES) under grant No. GRK 1994/1 and by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant No. I/82806. Additional support was provided by the German Federal Ministry of Education and Research (BMBF) as a part of the Software Campus program under the promotional reference 01-S12054.

References

Philip Bramsen, Pawan Deshpande, Yoong Keok Lee, and Regina Barzilay. 2006. Inducing Temporal Graphs. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 189–198, Stroudsburg, PA, USA. Association for Computational Linguistics.

Taylor Cassidy, Bill McDowell, Nathanael Chambers, and Steven Bethard. 2014. An Annotation Framework for Dense Event Ordering. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 501–506, Baltimore, Maryland, USA. Association for Computational Linguistics.

Nathanael Chambers, Taylor Cassidy, Bill McDowell, and Steven Bethard. 2014. Dense Event Ordering

[2] https://www.ukp.tu-darmstadt.de/data/timeline-generation/temporal-anchoring-of-events/

with a Multi-Pass Architecture. *Transactions of the Association for Computational Linguistics*, 2:273–284.

Quang Xuan Do, Wei Lu, and Dan Roth. 2012. Joint Inference for Event Timeline Construction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 677–687, Stroudsburg, PA, USA. Association for Computational Linguistics.

Nattiya Kanhabua and Kjetil Nørvåg. 2012. Learning to Rank Search Results for Time-sensitive Queries. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 2463–2466, New York, NY, USA. ACM.

Oleksandr Kolomiyets, Steven Bethard, and Marie-Francine Moens. 2012. Extracting Narrative Timelines As Temporal Dependency Structures. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 88–97, Stroudsburg, PA, USA. Association for Computational Linguistics.

Klaus Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology (second edition)*. Sage Publications.

J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174.

James Pustejovsky, Patrick Hanks, Roser Sauri, A. See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, D. Day, Lisa Ferro, and Marcia Lazo. 2003. The TIMEBANK Corpus. In *Proceedings of Corpus Linguistics 2003*, pages 647–656, Lancaster, UK.

James Pustejovsky. 1991. The Syntax of Event Structure. *Cognition 41 (1991)*, pages 47–81.

Roser Saurí, Jessica Littman, Robert Gaizauskas, Andrea Setzer, and James Pustejovsky. 2004. TimeML Annotation Guidelines, Version 1.2.1.

Andrea Setzer. 2001. *Temporal Information in Newswire Articles: An Annotation Scheme and Corpus Study*. Ph.D. thesis, University of Sheffield, Sheffield, UK.

Mihai Surdeanu. 2013. Overview of the TAC 2013 Knowledge Base Population Evaluation: English Slot Filling and Temporal Slot Filling. In *Proceedings of the TAC-KBP 2013 Workshop*, Gaithersburg, Maryland, USA.

Naushad UzZaman, Hector Llorens, Leon Derczynski, Marc Verhagen, James F. Allen, and James Pustejovsky. 2013. SemEval-2013 Task 1: TempEval-3: Evaluating Time Expressions, Events, and Temporal Relations. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9, Atlanta, Gerogia, USA.

Marieke van Erp, Piek Vossen, Rodrigo Agerri, Anne-Lyse Minard, Manuela Speranza, Ruben Urizar, Egoitz Laparra, Itziar Aldabe, and German Rigau. 2015. Annotated Data, version 2. Technical report, Amsterdam, Netherlands.

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. 2007. SemEval-2007 Task 15: TempEval Temporal Relation Identification. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 75–80, Stroudsburg, PA, USA. Association for Computational Linguistics.

Marc Verhagen, Roser Saurí, Tommaso Caselli, and James Pustejovsky. 2010. SemEval-2010 Task 13: TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 57–62, Stroudsburg, PA, USA. Association for Computational Linguistics.

Seid Muhie Yimam, Iryna Gurevych, Richard Eckart de Castilho, and Chris Biemann. 2013. WebAnno: A Flexible, Web-based and Visually Supported System for Distributed Annotations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 1–6, Sofia, Bulgaria, August. Association for Computational Linguistics.