Distribution Agreement

In presenting this thesis as a partial fulfillment of the requirements for a degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis in whole or in part in all forms of media, now or hereafter now, including display on the World Wide Web. I understand that I may select some access restrictions as part of the online submission of this thesis. I retain all ownership rights to the copyright of the thesis. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

Yingying Chen March 28, 2022

Analysis of Temporal Relations in Various Types of Text Data

by

Yingying Chen

Jinho D. Choi
Adviser

Quantitative Theory and Methods

Jinho D. Choi

Adviser

Jason McLarty

Committee Member

Marjorie Pak

Committee Member

2022

Analysis of Temporal Relations in Various Types of Text Data

By

Yingying Chen

Jinho D. Choi

Adviser

An abstract of
a thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Quantitative Theory and Methods

2022

Abstract

Analysis of Temporal Relations in Various Types of Text Data
By Yingying Chen

Detecting the temporal relations of events in a text is a complicated natural language understanding task. However, figuring out the timeline of events is key to improving machine comprehension. Previous work specified approaches to identifying events in texts, proposing appropriate temporal relations and ways to order events with respect to one another. However, the vast majority of existing temporal dependency annotation has been carried out on simple narrative text or news sources. The annotation schemes are not always applicable to noisy, highly variable, social media texts such as Reddit posts. We devise a more generalized and robust scheme to support a broader range of text annotation. In this research, we aim to 1) improve existing annotation guidelines for more complex sentence structures, 2) evaluate the annotation performance among student annotators to achieve competitive inter-annotator agreement scores, 3) quantify the characteristics unique to Reddit text and provide a statistical analysis of the difficulties encountered when annotating Reddit data, and 4) compare and contrast the effectiveness of our temporal annotation scheme across three diverse sources: children's stories, social media texts, and news articles. The results show that our annotation scheme is effective in identifying events with high-level inter-annotator agreement scores, but there is still space to improve for identifying timelines of events. Besides, our results show the challenges of generating a unifying temporal relations scheme for different types of text. These challenges lead to the discussion of how to evaluate the effectiveness of temporal relation schemes.

Analysis of Temporal Relations in Various Types of Text Data

By

Yingying Chen

Jinho D. Choi

Adviser

A thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Quantitative Theory and Methods

2022

Acknowledgements

As a student who transferred to Emory in my third year, I really appreciate the people I met and the opportunities I got that enriched the second half of my undergraduate experience. I want to thank my advisor Dr. Choi. His Computational Linguistics class inspired me to explore the intersection of Linguistics and Computer Science. I was grateful to join the Emory NLP Lab under his guidance. I want to thank our Lab's postdoctoral researcher Gregor Williamson for his generous support of linguistic and analytics mentoring for conducting this research project from the ground up. I want to thank the other two seniors Angela Cao and Jessica Ji who are also on the Linguist Team in our lab. They conduct different research directions from me, but I can always learn their insights and resilience in a one-year companionship in our senior research journey. I want to thank my two other committee members Dr. Pak and Dr. McLarty for leading me to the amazing world of Linguistics, which has already become my interest rather than merely a subject. Their courses are the unforgettable parts of my Emory experience. Last but not the least, I want to thank my parents for supporting my decision to go to Emory University and my decision to major in Quantitative Sciences with a concentration in Linguistics. They always encourage me to "trust the process", which makes me more courageous in facing difficulties in life.

Contents

4 Discussion & Future Work

4.1 Annotation Round 1 (Social Media Texts)

4.2 Annotation Round 2 (Children's Stories)

4.3 Annotation Round 3 (Small-Scale Children's Stories)

4.4 Annotation Round 4 (Large-Scale in Social Media Texts, Children's Stories, News Report)

5 Discussion & Future Work

6 Conclusion

Bibliography

Chapter 1

Introduction

Many languages have tenses to indicate time reference that makes both the speakers and listeners understand the context. In English, tenses are manifested by the conjugation of verbs. “I did my homework”, “I am doing my homework”, and “I will do my homework” represent the past, present, and future for one event “doing homework” respectively. While it is straightforward to identify when one specific event occurs at the sentence level, it requires more reasoning and comprehension to identify the time reference of a few events mentioned in a sentence or a paragraph. For example, “I did my homework after my teacher assigned it last Friday. I received my grade yesterday.” includes a few events and constructs a timeline, in which “I received my grade” comes after “I did my homework” that follows “my teacher assigned it”.

Semantic understanding regarding temporal reasoning is a growing field of computational linguistics. Figuring out the number of events, the order, and the relations in which these events happened is the key to machine reading comprehension. Thus, temporal dependency rules are needed to construct accurate timelines of events in the text.

TimeML annotation scheme (Pustejovsky, et al., 2003) is the first work in the field and is considered the golden standard to annotate timelines in text. Later studies suggest schemes to further specify the subtasks of temporal relation annotations (Verhagen et al., 2007; Verhagen et al., 2010; UzZaman et al., 2012) and advance annotation approaches for different kinds of text with a high inter-annotator agreement in event and relation identification (Bethard et al., 2012; Mostafazadeh et al., 2016; Araki et al., 2018). In this research, we aim to further improve the existing schemes by conducting large-scale annotation tasks in new types of data and providing an analysis of challenges annotating different types of text.

1.1 Motivation

There are three reasons to work on the analysis of temporal relations in different types of text. First, while previous work has provided annotation schemes such as the golden standard TimeML (Pustejovsky et al., 2003), there are no specific guidelines for human annotators to learn. The performance of human annotation is crucial and necessary in examining the validity of schemes before designing and conducting automatic annotation by machines. Given that the instruction materials are scarce, we decide to develop an informative, learnable guideline for temporal relation annotation.

Second, this research explores a new type of text social media post that is new to the domain. Previous temporal dependency work has researched types of text with professional expressions, simple temporal structures, and limited length, which can ease the level of difficulty in annotation tasks. This research aims to not only examine the effectiveness of existing schemes applying to social media posts but also analyze and quantify the challenges of annotating noisy textual data.

Third, this work conducts a comprehensive analysis of three different types of text, including children's stories, social media posts, and news stories, regarding their style, narrative, tenses, and testament that can affect the performance of annotation. Comparing and contrasting the characteristics of texts can provide insight that helps evaluate the challenges of annotation and offer suggestions for future improvement.

1.2 Thesis Statement

Given that Bethard et al.'s paraphrasing rule with a relation set of Before/After, Contains/Contained, and Overlap (2012) reach high inter-annotator agreement scores by limiting the scope to identify events and relations, we aim to replicate and develop his scheme for new types of text. Besides children's stories, we offer the hypothesis that while their rules are effective, they are not replicable in identifying the timeline in social media posts and news reports. In temporal annotation, each type of text with its unique characteristics can pose difficulties for temporal relation annotation, especially social media posts that can be highly

random and noisy. In this research, we will develop guidelines based on their scheme, design training materials, train student annotators in the research group to meet the standard, conduct several rounds of annotations in different types of text, and calculate the inter-annotator agreement score to examine the effectiveness of our scheme. In the analysis of different types of text, we will categorize the difficulties and provide possible solutions to improve the annotation performance for each round of annotation in both linguistic and quantitative approaches. These challenges can lead to further research on how to evaluate the effectiveness of temporal relation schemes.

1.3 Objectives

There are four objectives for this research.

- 1) Improve existing annotation guidelines for more complex sentence structures.
- 2) Evaluate the annotation performance among student annotators to achieve competitive inter-annotator agreement scores.
- 3) Quantify the characteristics unique to Reddit text and provide a statistical analysis of the difficulties encountered when annotating Reddit data.
- 4) Compare the effectiveness of our temporal annotation scheme across three diverse sources: children's stories, social media texts, and news articles.

It is the first time that researchers introduce a concrete temporal relation annotation guideline, and the first time that our work examines the effectiveness in three different types of texts, especially pioneering in the study of Reddit posts.

Chapter 2

Background

2.1 Related Work

In the 2002 Time and Event Recognition for Question Answering System Workshop (Pustejovsky et al., 2003), the ability of the machine to answer temporally based questions brought attention to computational linguistics. For example, in answering questions with rich temporal context like “Is Bill Gates currently the CEO of Microsoft?”, only knowing dates and times is not enough. The ability to understand how many events are in the text and what are the temporal relations between them is more important to make sense of the context. After that, a variety of annotation schemes and corpus came out to advance the information extraction and temporal reasoning tasks. The ultimate goal of temporal relation reasoning is the automatic identification of events and relations among them in the text, and having a linguistic annotation scheme for human annotators is the first step to getting closer to a reliable method.

TimeML annotation scheme (Pustejovsky, et al., 2003) lays the first theoretical foundation for temporal relation annotation by anchoring event predicates. This work defines an event as a cover term for situations that happen or occur. An event can be punctual or last for some time, and it can be a verb, adjective, predicative clause, or prepositional phrase. The work also provides approaches to identifying what constitutes relevant events, different kinds of temporal relations, and ways to order events to one another. Annotating based on the TimeML scheme, TimeBank (Pustejovsky et al., 2003) as a gold-strand corpus was created.

TempEval-1 (Verhagen et al., 2007) is the initial attempt to evaluate the effectiveness of the TimeML scheme through conducting three subtasks that identify relations between events and time reference in the same sentence, events, and document-creation-time, and main events among sentences. TempEval-2 (Verhagen et al., 2010) extends the former work by conducting

three more subtasks that identify events and time reference and allows examinations in data from six languages. Following TempEva-1 and 2, TempEval-3 (UzZaman et al., 2012) uses the largest dataset adopting the TimeML scheme and supports a new measure to rank systems in each subtask.

While previous works based on TimeML define events and identify a variety of relations, the applicability for texts and learnability for human annotators of previous schemes are relatively low because events are nonspecific and exhaustive to capture and relations are too fine-grained. In recent works researchers start to extract the non-overlapping parts from the previous temporal annotation rules and introduce models to help visualize the timeline to make the annotation schemes more effective. Ning et al. (2018) introduce a multi-axis annotation scheme to focus on annotating event pairs that are considered relevant. Zhang and Xue (2018) propose a dependency tree structure that allows every event to have a time reference and have more than one following events. Yao et al. (2020) propose a temporal dependency graph to better capture the completeness of temporal orderings than hierarchical dependency tree structures by allowing multiple time references for a single event. Van Gysel et al. (2021) incorporate temporal relations as parts of the Uniform Meaning Representation, in which Before, After, Contained, and Overlap relations are used.

In this field, previous work has introduced modified guidelines and done annotations in children's stories, everyday life stories, news reports, and Wikipedia articles. Bethard et al. (2012) suggest a new annotation scheme for representing timelines in Aesop's fables. This work provides more examples of annotation cases and creates an annotation guideline that human annotators can learn from. It highlights the rules of no speech, no modal, and paraphrasing to limit the range of event identification. Both event and temporal relation identification are evaluated and reach high Krippendorff's Alpha scores. For news data, O'Gorman et al. (2016) generate temporal relation rules regarding the coreference reasoning logic by specifying two subtypes of the Contains relations. They also consider that subevents do not need to be annotated if temporal reasoning can tell it is inferable to the main event. Mostafazadeh et al. (2016) propose a temporal annotation scheme concerning a casual framework to annotate temporal

events and relations in 5-sentence everyday stories from the ROCStories corpus. Araki et al. (2018) provide a temporal relation annotation scheme for Simple Wikipedia articles in ten domains ranging from architecture to politics and achieve high inter-annotator agreement on event annotations. Despite that the agreement score in the relation identification is low, they point out that the event granularity and ambiguity in simultaneity and event sequence could lead to a different perception of human annotators.

2.2 The Use of Social Media Texts (Reddit Corpus)

Computational linguists have studied semantic understanding of social media posts, especially on Reddit, for decades. As a network of communities for users to connect with people with similar interests, Reddit is one of the top ten social media platforms in the U.S. (Audier & Anderson, 2021) with 53 million daily active users worldwide in 2020 (Patel, 2020). Previous work has researched semantic analysis of anxiety (Shen et al., 2017), persuasiveness (Hidey et al., 2017), offensiveness (Hata et al., 2021), and narrative timelines in Reddit posts have not been explored. Investing the natural language understanding in social media text is also one of the goals for Emory NLP Lab in recent years because the linguistic style in social media is closer to people's chat behavior that helps build a chatbot for the college. Therefore, in addition to children's stories (Bethard et al., 2012), everyday life stories (Mostafazadeh et al., 2016), news reports (O'Gorman et al., 2016), and Wikipedia articles (Araki et al., 2018), investigating the effectiveness of temporal relation schemes is an important objective in this research.

2.3 Previous Attempts of This Research

In the first half-year of this research, our direction focused on improving the temporal relation annotation in Abstract Meaning Representation (AMR) (Banarescu et al., 2013), a semantic representation language that is expressed in a rooted, labeled graph. AMR only captures the concepts based on PropBank (Paul et al., 2005) corpus. As a minimal representation

structure, AMR does not incorporate a systematic approach to annotating tense. To improve the tense and aspect marking in AMR, Donatelli et al. (2018, 2019) propose an augmented framework with additional temporal distinction, including event time Before, Up To, At, and After the speech time. Van Gysel et al. (2021) design a Uniform Meaning Representation (UMR) to extend tense reasoning in AMR from the sentence level to a document level.

We aimed to build our work upon the current UMR structure and specify more temporal relations that it does not mention. However, two main reasons caused our changes in direction. First, AMR heavily relies on the availability of lexicons in the PropBank (Paul et al., 2005), which captures the majority of English lexicons but not all of them. Such deficiency increases the difficulty for annotators to select the most appropriate semantic representation for the arguments in the existing lexicons. Second, although UMR allows annotators to identify the general temporal relations between sentences, it fails to capture the temporal relations between or among events mentioned in every single sentence given that a sentence can be complex. UMR is not effective in representing temporal relations among events and time references as it claims.

Aiming to improve the identification of events and temporal relations at the document level, we later found that the annotation scheme in Bethard et al. (2012) is more applicable to natural language understanding. For one thing, it refrains from the limit of PropBank lexicons so that every possible event can be annotated. For another, combining the main idea of TimeML (Pustejovsky et al., 2003) and some practical cases, its scheme is instructive and user-friendly for human annotators to study. Thus, inspired by Bethard's work, we redirect our research direction and further develop our work based on it.

Chapter 3

Methodology

3.1 Data

We evaluate the effectiveness of our temporal relation guideline in three types of text: children's stories (Aesop's Fables), news reports (CNN articles), and social media texts (Reddit posts related to college lives). Children's stories come from Project Gutenberg. News reports come from the cnn_dailymail corpus. Social media texts come from a popular college subreddit on Reddit accessed on 14th Feb 2022.

Category	Text	Length
Children's Stories	Aesop's Fables	148.1 tokens
News Report	CNN articles	160.5 tokens
Social Media Text	Reddit posts	145.3 tokens

Table 1: Statistics of different types of text

The examples of each type of text are as follows.

Aesop's Fables: When people go on a voyage they often take with them lap-dogs or monkeys as pets to wile away the time. Thus it fell out that a man returning to Athens from the East had a pet Monkey on board with him. As they neared the coast of Attica a great storm burst upon them, and the ship capsized. All on board were thrown into the water, and tried to save themselves by swimming, the Monkey among the rest. A Dolphin saw him, and, supposing him to be a man, took him on his back and began swimming towards the shore. When they got near the Piraeus, which is the port of Athens, the Dolphin asked the Monkey if he was an Athenian. The Monkey replied that he was, and added that he came from a very distinguished family. "Then, of course, you know the Piraeus," continued the Dolphin. The Monkey thought he was

referring to some high official or other, and replied, "Oh, yes, he's a very old friend of mine." At that, detecting his hypocrisy, the Dolphin was so disgusted that he dived below the surface, and the unfortunate Monkey was quickly drowned.

CNN articles: Armed militants in southwest Pakistan torched two oil tankers carrying fuel for U.S. and NATO forces in Afghanistan, a government official told CNN. Azam Shahwani, a senior government official in Balochistan province, said four gunmen riding on motorcycles opened fire on a convoy of five oil tankers in the area of Mithril, a village in the district of Bolan. Shahwani said the oil tankers were heading toward Afghanistan. No one was injured, but two of the oil tankers were destroyed, Shahwani said. "I could see roaring flames of fire even three hours after the attack," Shahwani said.

Reddit posts: I'm so excited! I've been doing a PhD program in physical chemistry for about 2 years, and today I had my first "holy shit, these results are so exciting" moment! To make a long story short, I'm working on catalysts, and up until now, for like a year, we thought that one of the elements was not involved in catalysis. We used it as part of the support material, to make the Pt more effective. Yesterday, I read a new paper about that element in a similar compound as an active catalyst for a similar reaction that we use. So I presented this paper to my advisor, and suggested we look at the data (I had just run catalysis for the support sample by itself without Pt, just as a blank and I hadn't looked at the products for it because I assumed there just wouldn't be anything). I just looked at it, and holy shit I was right!!! It's forming products without platinum! This is huge for my research!!

3.2 Annotation Procedure

Annotators: There are four annotators in this project. Three of them are undergraduate students and one of them is a postdoctoral researcher.

Training Process: To ensure the quality of annotation, each of the annotators is required to do extensive annotation training for 3 weeks, including 1) studying the guideline, 2) watching

1-hour instruction video, 3) taking online tests consisting of 100 questions regarding event identification, pair identification, and relation identifications, and 4) annotate 10 test documents before the official annotation.

Annotation Platform: We use INCEpTION (Klie et al., 2018), a semantic annotation platform established by UKP Lab at TU Darmstadt, to conduct our annotation experiment.

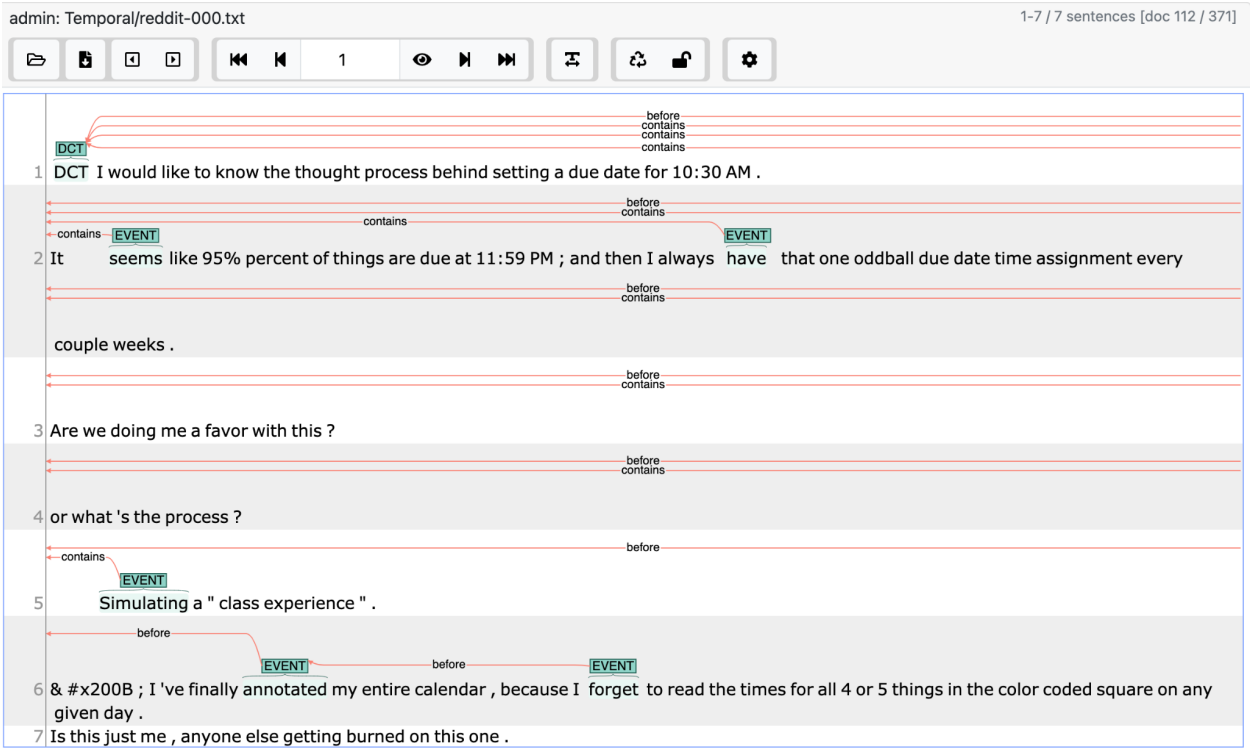

Figure 1: An annotation example from INCEpTION

Rounds of Annotation: We conducted 4 rounds of annotation in examining three types of data. The first three rounds of annotation are small-scale annotations. We call them rules-generation stages, in which we quantify the analysis and come up with possible solutions to improve the annotation performance. The fourth round of annotation is a large-scale annotation examining the effectiveness of the finalized guideline applying to social media text, children's stories, and news reports.

3.3 Evaluation Metrics

We use the F1 score (UzZaman et al., 2012; O'Gorman et al., 2016) to evaluate the inter-annotator agreement for event identification and relation identifications. F1 measurement includes precision and recall metrics. Besides, we also measure the number of events in the text, the number of tokens per event, the number of the five temporal relations, and the number of each relation linked to the Document Creation Time.

3.4 Three-Stage Annotation Method

1) Event Identification: This is the step to highlight the text spans of events in the text. The punctuation is not included in the event.

2) Pair Identification: Before linking the events or time reference with temporal relations, we need to make sure of the order of the events. In most cases, the annotation order follows the reading order by considering the temporal relation between a word and its following word. When it comes to a complex sentence that contains a main clause and a dependent clause, the event in the main clause should be considered first and linked to the main event in the preceding sentence.

3) Relation Identification: After identifying the events and their linking order of them, we can choose the appropriate relations to link between events or between a time reference such as Document Creation Time and an event.

3.5 Annotation Core Rules

Bittar et al. (2012) stress the importance of taking into account the full context of expression to do temporal annotation. The dominant scheme TimeML only supports surface-based annotation that would bring too many words into the consideration of events. We also notice that in social media texts there can be unforeseen cases of non-standard,

ungrammatical expressions that the guideline may not cover. In our temporal annotation guideline, we consider following the context as our golden standard if we encounter sentences with no available rules to apply.

3.5.1 The Definition of Event

TimeML (Pustejovsky et al., 2003) defines the event as “a cover term for situations that happen or occur, can be punctual or last for a period of time, can be those predicates describing states or circumstances in which something obtains or holds are true.” In our research, we follow the definition of events from Pustejovsky et al.

In most cases, the event is the main predicate and verb in a sentence. An event can happen in the past, in the present, or in the future. We value the certainty of events, and not every predicate and verb should be annotated. Main predicate or verbs with a high likelihood of happening based on the context are considered events.

3.5.2 The Definition of Time Reference (Document Creation Time)

TimeML (Pustejovsky et al., 2003) uses Document Creation Time (DCT) to represent the time that the text was written. DCT acts as a temporal anchor to measure the chronological relations among events. To complement the lack of use in time reference from Bethard et al.’s work (2012), we make a rule that every text uses Document Creation Time as a reference in the timeline so that events in the narrative can link relations based on this. For the practicality of annotation, annotators should mark the Document Creation time at the beginning of each text.

Only in the Reddit data, do we find that some texts can have more than one Document Creation Time. Reddit post authors sometimes edit the original posts as a follow-up to update some new information or to reply to some common questions from the comments. Some posts even have three edits. In these cases, we consider each edit as a new Document Creation Time,

and the content that follows could only link to the Document Creation Time ahead. Each Document Creation Time has a Before/After relation to the previous one.

3.5.3 The Annotation of Event (Basic Type)

The square brackets of texts represent the selected text span for events. In the following examples, predicates are highlighted to represent the events that happen with great certainty.

- I [do] my homework.
- I [did] my homework.
- I will [do] my homework.

To make the event annotation more structural, we propose a basic formula for annotating the event when it follows a copula: Copula + noun/verb/adjective/preposition. Copula includes tenses that indicate temporal orders, unlike previous temporal schemes, we incorporate copula as a part of the event annotation. For example, the annotation of events in the form of expressions is provided as follows.

- **Copula + Noun**
    - It has [been a month].
    - It [is a piece of apple].
- **Copula + Verb**
    - I [am doing] my homework.
- **Copula + Adjective**
    - He [was sad].
    - She [is always good].
- **Copula + Preposition**
    - Tony [is in front of] you.
    - Tomy [is always in front of] you.
    - Tony [is always standing] in front of you.

- **Existential Construction**
    - There [are children swimming] in the lake.
    - There [are birds in] the garden.

Prepositions are not always included. We do not annotate infinitives such as “to” and locative prepositions.

- We [went] to the supermarket yesterday.
- He [is standing] on the table.

But in other cases, we include prepositions in phrasal expressions.

- He [walked away] in an air of dignity.
- I [care about] you.
- She [was thinking of] becoming a zoologist.

In the Reddit posts, we find out a number of colloquial expressions such as the use of “be like” and “feel like”. To make the annotation consistent, we generate rules specifically regarding these situations.

- If “be like” expresses the equivalent meaning as the verb “say”, we annotate “be like” as an event.
    - She['s like] "now we need to go to Boston." = She [says] "now we need to go to Boston."
- For the expression "feel like" with nominal complements, we annotate "feel like" because it means "want" as in the dictionary.
    - I [feel like] a burger. = I [want] a burger.
- For the expression "feel like" with clausal complements, we only annotate "feel".
    - I [feel] like everything I did was totally wrong.
- For the emphasized word "do", we annotate the predicate that comes after it.
    - I did [do] the review for the exam!
- For the expression "wanna", we treat it as the verb "want".
    - I [wanna] eat. = I [want] to eat.

- For the expression "gonna", we treat it as "going to" and annotate the verb that comes after it.
    - I am gonna [buy] that ticket. = I am going to [buy] that ticket.
- For the expression "gotta", we treat it as a modal verb such as "have to" and "must". So we do not annotate it due to the no modal rule, which is explained below.
    - I gotta go. = I must leave.

3.5.4 The Annotation of Event (Considering Context in Discourse)

No Speech & No Content after Speech Verbs: Bethard et al. (2012) introduce the no speech rule because direct speech adds difficulties for readers to comprehend and the content is less essential than that in the narrative. It is often unclear when the event described by a direct quote is meant to take place.

- He [says]: “I hope you can bring some food to me.”

Based on the no speech rule for direct speech, we extend such a rule to indirect speech. We do not annotate the content after the speech verbs because it poses confusion in readers’ understanding and the events can be subjective with less certainty. Speech verbs include but are not limited to “say”, “claim”, and “argue”.

- I [am saying] that the school was a sham taking advantage of minority children.

No Hypothetical: Bethard et al. (2012) introduce a hypothetical rule in their guideline, but there are limited examples for instruction. Based on the idea of event certainty, we complement more annotation cases for human annotators to learn.

- Conditionals are a clear case in which events are hypothetical.
    - If it rains, the grass will be wet.

- We do not annotate content after verbs such as "hope", "want", "wish", and "feel" because the content is expressed in hypothetical situations that may not definitely occur.
    - I [hope] to make some great progress.
    - He [wants] to achieve a higher score in the annotation.
    - She [wished] her mother would call her.
    - It [seems] that this is the evidence.
    - I [felt] that the midterm was so hard.
- When it comes to some adverbs that contain hypothetical meanings, we should be careful about the context that expresses uncertainty. In the following case, there is no event.
    - Apparently, from what I have heard bad things were taking place.

No Modal: Modal auxiliaries should not be considered (Bethard et al., 2012) because they indicate a possibility rather than certainty. Keywords include but are not limited to “might”, “can”, “must”, “have to”, and “need”. We do not annotate modal words in the sentences.

- You should go to school right now.

No Negated: Bethard et al. (2012) mention that negated events are hard to place along the timeline. We do not annotate negated expressions since it means the denial of the events. Previous work does not offer enough examples for instruction, we expand the no negated rule considering situations such as negated keywords, adjectives with a negated prefix, and the distinction between negated and factual meanings.

- We do not annotate keywords including "no", "not", "nothing", "none", and "never".
    - His plans for the interior of the building were not completed.
- Some words such as "barely", "hardly", and "little" seem negative, but they do not always negate the event. We should make judgments based on the context.
    - I barely [passed] the exam.

- Some adjectives involve a negated prefix. For consistency of annotation, they should not be annotated either. But adjectives with negative meanings are excluded.
    - I am unhappy.
    - I [am sad].
- Double negation should still refrain from annotating the event. This is to maintain consistency because often it is still unclear the temporal ordering of an event is in a doubly negated sentence.
    - It's not that I am not happy, but I [am anxious] about work.
- When it comes to some combined cases of negation sentences, we only annotate the part with the event(s). Having a negation word does not mean we have to give up annotating the whole sentence.
    - I don’t like this because it [looks] bad.
    - I [like] this because it does not look bad.

No Question: Unlike news reports and children’s stories, questions or rhetorical questions appear frequently in social media texts to express the feelings of the authors. To reduce confusion, we make the following rules for event identification.

- We do not annotate events in "normal questions" since they can be hypothetical.
    - Did he go to bed on time?
- For sentences with question marks, we should capture the events in the sentence if there are any based on the context. The following sentence is written in the context that the author is scolding another person for the things that he did.
    - You [are the one] who [waited] until the last second to [let] anyone know about the textbook and you [expect] us to be ready to use it the next day?
- Sometimes the authors make mistakes in punctuation. So we should annotate them case by case based on context.

No Imperative: Imperative is used to make suggestions and requests. But imperative expressions do not appear in children’s stories such as Aesop’s Fables (Bethard et al., 2012).

Considering the high frequency of imperative expressions in Reddit data, we make some rules not to annotate imperative because it is uncertain if the event happens or not.

- Play this game with me!

No exclamatives, Swear words, Exaggeration: In Reddit texts, it is common to see such unofficial expressions that are unique to social media posts. For clarity, we make some rules to not annotate the following types of expression.

- Exclamatives
    - Thank you!
- Swear words
    - Fuck my life.
- Exaggeration
    - God fucking kills me.

Paraphrasing Rule: Paraphrasing rule means to select words that best capture the meaning in phrasal events. Bethard et al. (2012) suggest the first two sub-rules in their work. We develop the third sub-rule to adjust some more complex situations that require contextual understanding.

- Aspectual verbs include but are not limited to "begin", "start", "stop", "remain", "end up", "proceed", "continue", "finish", and "keep". We only annotate the verbs or adjectives that come after the aspectual verbs.
    - start [doing] my homework
    - stop to [rest]
    - the reason remains [mysterious]
    - end up [marrying] his childhood friend
    - proceed to [read]
    - continue [working]
    - finish [planting] the seed
    - keep [raising] his questions

- Paraphrasing rule applying in expressions or sentences.
    - a dog who used to [snap at] other people
    - He managed to [scramble on] to dry ground from a backwater.
    - He got [called] yesterday.
    - She gets [better] now.
    - He did his best to reach them by [jumping].
    - This experience makes me [realize]...
- Paraphrasing rules should focus on the context rather than lexical rules. There are two examples related to the verb “let”. In the first sentence, we do not annotate "let" because "hang down" already represents the event that happened. In the second sentence, we annotate “let” rather than “play” because “let” means permission and “are done” might not happen.
    - She let herself [hang down] by her hind legs from a peg.
    - The mother [let] her children play Nintendo Switch after their assignments are done.

3.5.5 The Definition of Temporal Relation

Considering the flexibility in temporal relation annotations, we make Before/After and Contains/Contained two reversed pairs that annotators can use either to express the same meaning. In the following table, A and B can both represent an event or a time reference such as Document Creation Time.

Relation	**Annotation**	**How to Read**	**Definition**
Before	A ← before B	B is before A	B is totally before A (B finished before A started)
	= B ← after A	= A is after B	A is totally after B (A started after B finished)

After	A ← after B	B is after A	B is totally after A (B started after A finished)
	= B ← before A	= A is before B	A is totally before B (A finished before B started)
Contains	A ← contains B	B contains A	The time of B properly contains the time of A (B started before A started and finished after A finished)
	= B ← contained A	= A is contained in B	The time of A is properly contained in the time of B (A started after B started and finished before B finished)
Contained	A ← contained B	B is contained in A	The time of B is properly contained in the time of A (B started after A started and finished before A finished)
	= B ← contains A	= A contains B	The time of A properly contains the time of B (A started before B started and finished after B finished)
Overlap	A ← overlap B	B overlaps A.	1) A and B share a time span (may start and end differently) 2) A and B are identical (start and end at the same time)
	B ← overlap A	= A overlaps B	

Table 2: Temporal Relations Table

3.5.6 The Annotation of Relation

Before/After: Before and After relations are used to annotate simple temporal precedence and subsequence relations between two events or between a time reference such as Document Creation Time and an event. The following example sentence comes from Yao et al. (2020). “Went” and “had” are in chronological order.

- [DCT] Yesterday, I [went] to the museum, then [had] dinner with my friends.
    - Relation: DCT ← before [went]

      [went] ← after [had]

Contains/Contained: Contains and Contained are used to annotate events within a containment relationship between two events or between a time reference such as Document Creation Time and an event. The following example sentence comes from Bethard et al. (2012). The “camp up” event of the bear occurred when the protagonist “pretended” himself to be dead.

- He [threw] himself on the ground and [pretended] to be dead. The bear [came up] and [sniffed all around] him.
    - Relation: DCT ← before [threw]

      [threw] ← after [pretended]

      [pretended] ← contained [came up]

      [came up] ← after [sniff all around]

Another example sentence comes from Kolomiyets et al. (2012). The event of the boy being “strung” by a nettle occurred when he was gathering berries.

- A boy [was gathering] berries from a hedge when his hand [was stung] by a nettle. [Smarting] with the pain, he [ran] to tell his mother.
    - Relation: DCT ← before [was gathering]

      [was gathering] ← contained [was stung]

      [was stung] ← after [ran]

[smarting] ← contained [ran]

Overlap: Overlap means two events may share an interval of time. It is also compatible with the events sharing all their times such as happening and ending at identical times. The following example sentence comes from Pimentel et al. (2020). The loss in value and the falling happen at the same time, so the overlap relation can express such meaning.

- The insurer's earnings from commercial property/casualty lines [fell] 59% in the latest quarter, while it [lost] $7.2 million in its personal property/casualty business.
    - Relation: DCT ← before [fell]

[fell] ← overlap [lost]

This example sentence comes from Ross et al. (2020). "Call for" event and "saying" event are almost identical, so we can use Overlap to capture the relationship.

- Yeltsin and Kuchma [called for] the ratification of the treaty, [saying] it would create a "strong legal foundation".
    - Relation: DCT ← before [called for]

[called] ← overlap [saying]

The Distinction between Overlap and Contains in Parallel Structure: The following example uses a parallel structure. It seems that three events happen at the same time and overlap each other. But we prefer to use Contains/Contained relation in this case because containment relation is more informative. If each event contains the Document Creation Time, they must overlap each other. The logic is already entailed in Contains/Contained relation.

- People [are anxious], the world [is heating up], and the ice caps [are melting].
    - Relation: DCT ← contains [are anxious]

DCT ← contains [is heating up]

DCT ← contains [are melting]

3.5.7 Pair Identification

The order of annotation is normally linear and follows the reading order. But in a complex sentence that contains a main clause and a dependent clause, we should consider the event in the main clause first and then link to the main event in the preceding sentence. The following example sentences come from our Reddit data.

- [DCT] Before I [realized] this scam, I firmly [believed] it without any doubt. I [am confused] now.

Before we introduce the pair identification rule, we may link the events in the following way:

- (Inaccurate) Relation: [DCT] ← before [realized]

  [realized] ← before [believed]

  [believed] ← after [am confused]

After we have the pair identification rule, we can better make sense of the logic in the timeline because the "believed" event is more emphasized in the main clause. The difference between the two ways of annotation is that now there is no relation between "realized" and "believed" events.

- Relation: [DCT] ← before [believed]

  [realized] ← before [believed]

  [believed] ← after [am confused]

Chapter 4

Results & Analysis

4.1 Annotation Round 1 (Social Media Texts)

Round 1 represents our first attempt to apply Bethard et al.'s temporal guideline (2012) with some of our improvements to the Reddit data. There are a total of 40 documents. Each of the 4 annotators is assigned 10 documents to annotate. The research has 6 groups of double annotation pairs to compare and calculate the result.

Annotators	3 undergraduates, 1 postdoctoral researcher
Type of the text	Social media text (Reddit posts)
The number of texts	40 in total, 10 for each annotator
The average length of the posts	225 tokens
Steps to annotate	1) Event identification 2) Relation identification
Evaluation Metrics	F1 scores for 1) event identification and 2) relation identification

Table 3: Annotation Round 1 Background

The results show that about half of the relations are linked to Document Creation Time, implying that there are few relations linked between events in the narrative. This is because unlike the storytelling in children's stories where the plot is well-knit, Reddit posts are more like a stream of consciousness.

Numbers of Events per text	19.4
Before/After per text	8.7
Contains/ Contained per text	7.7

Overlap per text	2.0
Relations linked to DCT	11.0

Table 4: Annotation Round 1 Statistics

While the score of event identification of 0.719 is fair compared to the result from Bethard et al. (2012), the score of relation identification of 0.351 is low. The low consistency in relation identification indicates that Bethard et al.'s temporal guideline written for children's stories is not sufficient to represent the timeline for the social media texts, and we need to make improvements on the existing guidelines. In the next round of annotation, we expect to replicate Bethard et al. 's research by annotating Aesop's Fables and find out the difficulties in annotating two types of texts.

	Event Identification F1	**Relation Identification F1**
Mean	0.719	0.351
Highest	0.968	0.634
Lowest	0.471	0

Table 5: Annotation Round 1 Event & Relation Identification F1 Results

First time annotating Reddit posts, we find out that some authors would edit the original post as follow-ups to the story or comments replying to everyone. A text with many times of edits implies that there can be multiple Document Creation Times. Based on the existing rules of time reference, we allow more than one Document Creation Time to exist in one document. Each Document Creation Time has a Before/After relation to the previous one.

Another challenge is that the style of Reddit posts is colloquial. Since these texts are crawled from college subreddits, the authors are college students who tend to use informal, everyday language to write their posts for an audience who is also of the same age and same identity. Colloquial expressions such as "be like" and "feel like" appear frequently. To make the annotation consistent, we make new rules regarding some colloquial expressions. For example, If "be like" expresses the equivalent meaning as the verb "say", we annotate "be like" as an event.

For the expression "feel like" with nominal complements, we annotate "feel like" because it means "want" as in the dictionary. For the expression "feel like" with clausal complements, we only annotate "feel".

4.2 Annotation Round 2 (Children's Stories)

Round 2 represents our second attempt to replicate Bethard et al.'s annotation of children's stories (2012). There are a total of 10 documents. Each of the 4 annotators is assigned 5 documents to annotate. The research has 6 groups of double annotation pairs to compare and calculate the result.

Annotators	3 undergraduates, 1 postdoctoral researcher
Type of the text	Children's stories (Aesop's Fables)
The number of texts	10 in total, 5 for each annotator
The average length of the posts	189 tokens
Steps to annotate	1) Event identification 2) Relation identification
Evaluation Metrics	F1 scores for 1) event identification and 2) relation identification

Table 6: Annotation Round 2 Background

While the numbers of Before/After and Contains/Contained relations in the Reddit posts are similar, the numbers of Before/After and Contains/Contained relations in Aesop's Fables are contrasting. The number of Before/After relations per text is around 12.55, but the number of Contains/Contained relations per text is less than 1. Besides, while the number of relations linked to DCT in the Reddit posts is 11, the number of relations linked to DCT in Aesop's Fables is only 1.3. These sharp contrasts imply that the majority of the relations are linked between events in the narrative, which shows a strong chronological order. This makes sense in the children's stories because the texts tell stories from the beginning to the end.

Numbers of Events per text	16.70
Before/After per text	12.55
Contains/ Contained per text	0.85
Overlap per text	2.20
Relations linked to DCT	1.30

Table 7: Annotation Round 2 Statistics

The score of event identification in Aesop's Fables is about 0.11 higher than that in Reddit posts, and the score of relation identification in Aesop's Fables is 0.16 higher than that in Reddit posts. While the texts are of the same length, the level of annotation difficulty for Aesop's Fables is lower than that for the Reddit posts.

The event identification score is as high as Bethard et al.'s work (2012), but there is still space for relation identification scores to improve. Both types of data contain complex sentences with main clauses and dependent clauses. There is a comma in a sentence to break it into two parts with their own predicates. Previously, we annotated in normal reading order from left to right as one event linked to its previous one. But when it comes to such complex sentences, the events in the main clauses would be more emphasized than the events in the dependent clauses. Take a sentence from the Reddit post as an example. "Although it has been three years, I still get ashamed of seeing that grade on my transcript". The event that the author still gets ashamed is more emphasized than the time that has passed. The order of annotation could change to annotating "ashamed" first and then annotating "been three years" because it shows the logic more explicitly in the timeline. Take another sentence from Aesop's Fables as an example. "But as soon as the man had fitted the handle to his Ax, he went to work to chop down all the best Trees in the forest." The event of going to work is more emphasized than the event of fitting the handle of the Ax, which cannot stand by itself as a sentence. Thus, the event annotation order should not always follow a linear order, which causes the loss of information. So for the next round of annotation, we develop pair identification rules for complex sentence cases to help identify the main clauses and dependent clauses.

	Event Identification F1	Relation Identification F1
Mean	0.83	0.51
Highest	1	0.69
Lowest	0.68	0.27

Table 8: Annotation Round 2 Event & Relation Identification F1 Results

The low disagreement score in the relation identification implies that the guideline may still be far from ideal, but it could also be the intrinsic differences between the two types of texts. Having annotated both Reddit posts and Aesop's Fables, we find the following challenges that pose the difficulty for annotation.

First Person vs. Third Person Point of View: Reddit posts are written by college students to talk about their diverse experiences in college lives from application, academics, internship, to graduation. These social media posts typically adopt the first-person point of view to tell the stories mixed with personal feelings. Few of them use the second-person point of view to write some college tips and reminders for readers. However, Aesop's Fables are based on the third-person point of view, as characters in each of the stories are animals such as ass, lion, and fox and human beings such as cobbler, farmer, and doctor. The difference in point of view can affect the following aspects to annotate temporal relations.

The Mix of Narrative and Personal Feelings: A typical fable includes a narration that has a start, climax, and a resolution of life lessons, whereas the theme and content of the Reddit post are not consistent. Most Reddit posts mix both storytellings and authors' subjective feelings. But it is not uncommon to see a post only with personal feelings or with unstructured, unorganized information such as links. Annotators can get more familiar as they annotate more fables, but they may not be more familiar with the Reddit posts since each of them are different.

Past Tense vs. Mix of Tenses and Frequency Shift of Temporal Focus: Fables describe stories that happened in the past and only use past tense. The fables are in chronological order and events have an obvious timeline. They mostly happened in the past, as the plots go smoothly and linearly. However, Reddit posts can include different tenses such as past, present, perfect,

and future in one text. This is because the author may want to describe his or her experience, make comments on the present times, and may express comments or determination about the future. So the mixed-use of tense increases inconsistency in annotating temporal relations because events in the narrative are not coherent and connected.

Storytelling Style: Fables are grammatically standard and educationally meaningful for English teaching. It aims to enlighten the readers by giving a story of the past. But Reddit is a social media platform to express feelings and share information. Authors are not restricted to writing their content so most posts are causal, emotional, contain typos and ideas of stream of consciousness. Sentences are likely to include hypothetical, modal, and negated expressions, which should be excluded as events. In a post full of personal feelings but no informative narrative, annotators can hardly capture events that happen with great certainty.

Cases of Multiple Document Creation Times: Several Reddit posts include authors' edits as both updates to the original story and interactions with the readers. The follow-ups somehow create new document creation times in addition to the one that already exists for setting the background of the post. To solve this issue, we allow multiple Document Creation Times annotated when it comes to the sign of "edit/update" in the post. The annotators should attach relations between the newly added Document Creation Times and the predicates that come after it. Among 40 Reddit posts, there are 3 posts that have such a situation. This is a special situation that only happens in social media texts, making the annotation guideline unlikely to be intuitive.

Restatements: Restatement almost does not exist in fables but it can be normal in Reddit posts. An author may mention one event several times for emphasis. Some annotators use Overlap to link the repeated events together. Some annotators link them to their previous events or time reference, as there are two separate events to annotate. Repeated events make the annotation results hard to read.

4.3 Annotation Round 3 (Small-Scale Children's Stories)

Round 3 represents our third attempt to examine the new rules still applying to Aesop's Fables, including pair identification, multiple Document Creation Times, and colloquial cases, that we introduced to the guideline. It is a small-scale annotation using 6 documents. Two annotators are required to annotate all of them.

Annotators	1 undergraduates, 1 postdoctoral researcher
Type of the text	Children's stories (Aesop's Fables)
The number of texts	6 in total, 6 for each annotator
The average length of the posts	194 tokens
Steps to annotate	1) Event identification 2) Pair identification 3) Relation identification
Evaluation Metrics	F1 scores for 1) event identification and 2) relation identification

Table 9: Annotation Round 3 Background

Numbers of Events per text	19.01
Before/After per text	14.50
Contains/ Contained per text	0.50
Overlap per text	2.58
Relations linked to DCT	1.08

Table 10: Annotation Round 3 Statistics

While the performance in event identification remains stable at over 0.8, the score of relation identification improves 12% compared to annotation round 3 by following the newly added rules. We still believe that there is space for improvement.

During the annotation, we found out the basic rules to annotate events still remain unclear because the annotators often do not know whether they should include prepositions following a

verb. For example, in the sentence “He walked away in an air of dignity” the annotators do not know if they should annotate “walked” or “walked away”. There are inconsistencies among them even after a few rounds of training. We expect that the score of event identification could still go higher if there is a standard. Based on this situation, we make a clarification about how to annotate the prepositions that come after the predicate verbs: Except for infinitives such as “to” and locative prepositions, we include prepositions in phrasal expressions as these prepositions can specify the meaning of the verbs and make the annotation more intuitive.

	Event Identification F1	**Relation Identification F1**
Mean	0.81	0.58 (+12.1%)
Highest	0.92	0.74
Lowest	0.65	0.34

Table 11: Annotation Round 3 Event & Relation Identification F1 Results

There are 2 among the 6 documents that the two annotators treat differently regarding the choice of prepositions. Once the rule came out, the two annotators re-annotated those 2 documents by including the prepositions in phrasal expression and re-calculated the inter-annotator agreement scores. The results show that both event identification and relation identification scores improved effectively through the adjustment of the existing guidelines. We expect to apply the guideline to larger-scale texts and include other types of text news reports to test the effectiveness of the temporal scheme.

	Event Identification F1	**Relation Identification F1**
Mean	0.85 (+4.9%)	0.67 (+15.5%)
Highest	0.98	0.68
Lowest	0.69	0.6

Table 12: Annotation Round 3 Event & Relation Identification F1 Results (after applying the preposition containment rule)

4.4 Annotation Round 4 (Large-Scale in Social Media Texts, Children's Stories, News Report)

Round 4 represents our fourth attempt to apply Bethard et al.'s temporal guideline (2012) with a number of our improvements to a large scale of three types of data. There are a total of 150 documents, and each type of text includes 50 documents. Each of the 3 annotators is assigned 150, 100, and 50 documents to annotate respectively. The research has 2 groups of double annotation pairs to compare and calculate the result.

Annotators	2 undergraduates, 1 postdoctoral researcher
Type of the text	Social media texts (Reddit posts) Children's stories (Aesop's Fables) News Report (CNN articles)
The number of texts	150 in total, 50 each type of text, 150, 100, 50 for each annotator respectively
Steps to annotate	1) Event identification 2) Pair identification 3) Relation identification
Evaluation Metrics	F1 scores for 1) event identification and 2) relation identification

Table 13: Annotation Round 4 Background

We notice that news reports include both direct speech and indirect speech to show evidence and trustworthiness. However, our guidelines created for Reddit posts and children's stories follow no speech and no content after speech verb rules because it is often unclear when the event described by a quote is meant to occur. We exclude event annotations when it comes to quotations, but it does not apply to news reports, which should be written based on facts. If we blindly apply the existing guideline to news reports which are full of quotations, much information would be lost. To compare the results of both annotating and not annotating the events in quotations in news reports, we want to annotate the same news data twice in different

ways. In the first version of annotation, we follow the existing temporal guideline not to annotate any events indirect speech and indirect speech. In the second version of annotation, we will annotate both direct and indirect quotes because they are from reliable sources. But we do not annotate quotation verbs such as “a spokesperson said”. We want to compare which version of the guideline is more effective in temporal relation annotation for news reports, and more importantly, what could be the challenges for annotating news reports.

The results show interesting results that both confirm the previous analysis and bring new thoughts. CNN version 1 has both the highest event identification and relation identification score, but this is not because of the high consistency among annotators. Since in version one we exclude events in direct and indirect speech, there are few events left for annotation. This can be explained by the contrast that CNN version 1 only has 9.846 events per document, while the other three types of text have more than 11.2 events per document. Also, the tokens per event for CNN version 1 is 17.230, which is the highest among the other types of text and indicates the sparsity of events led by the application of the original guideline.

The performance of relation identification in CNN version 2 is only 0.387, which is the lowest among each type of text and makes a sharp contrast to that of CNN version 1 at 0.564. The adjustment of rules that include events in direct and indirect quotes, which we consider more informative, makes the annotation even harder. The difficulty of annotating news reports does not come from the expression level as Reddit posts, which use colloquial, everyday language. It comes from the narrative structure of news reports. In journalism, news reports use an inverted pyramid structure to deliver information, in which the most important summary or conclusion including who, what, when, where, and why would be presented at the beginning of the text (Schade, 2018). After identifying the key information at the beginning, the rest of the passages expand the stories. But the body parts of the news reports do not follow a linear order as the stories progress. Unlike Aesop’s stories that tell coherent stories from the start, climax, to a resolution, news reports consist of a large number of direct and indirect speech to construct the stories, which are less coherent due to the changes of perspectives of different sources. For example, a news report on a crime may adopt quotations from the police department, the suspect,

the eyewitness, and the lawyer. Although presenting comprehensive sources to show the unbiasedness of the report, too many perspectives may interrupt the flow of a coherent narrative. Following the inverted pyramid structure, news reports would inevitably contain repetition of events. In these cases, some annotators use Overlap to link the repeated events together, but some link them to their previous events or time reference, as there are two separate events to annotate.

The temporal relation structure for Aesop's Fables is the most linear among each type of text because they have the highest number of Before/After relations at 7.296 and the lowest number of Before/After relations linked to Document Creation Time at 1.104. The number of Contains/Contained relations in Aesop's Fables is even 0.563, which is lower than 1. Reddit posts have the lowest number of Before/After relations at 4.458, but they have the highest number of Contains/Contained relations at 4.654 among other types of text, especially more than 90% (4.238/4.654) of the Contains/Contained relations are linked to Document Creation Times. Unlike Aesop's Fables which tell stories with an organizational framework, Social media posts are like streams of consciousness in which each sentence could be relatively independent. Thus there are few relations linked between events, and most events in the Reddit posts are linked to Document Creation Time by Contains relations. Aesop's Fables do the opposite. The high number of Contains/Contained relations in the Reddit posts can be explained by the fact that most Reddit posts are written in the present tense to describe authors' daily routines in academic lives, or they tend to use "I think/ I want/ I feel like" to express their current feelings towards things of the past.

	Reddit	**Aesop's Fables**	**CNN-Ver 1**	**CNN-Ver 2**
Event Identification F1	0.759	0.847	**0.864**	0.753
Relation Identification F1	0.473	0.477	**0.564**	**0.387**
Tokens	145.308	148.083	160.483	

Events	11.617	11.225	**9.846**	**11.413**
Tokens per Event	13.756	14.526	**17.230**	14.819
Before/After	4.458	**7.296**	6.346	6.525
Contains/Contained	**4.654**	0.563	1.367	1.971
Overlap	1.313	2.079	1.054	1.750
All Relations to DCT	6.371	1.200	5.575	4.954
Before/After to DCT	2.133	**1.104**	4.404	3.404
Contains/Contained to DCT	**4.238**	0.096	1.129	1.517

Table 14: Annotation Round 4 Statics and Event & Relation Identification F1 Results

Chapter 5

Discussion & Future Work

The results of our research show that our temporal guidelines reach high inter-annotator agreement scores in event identifications, but there is still space for relation identifications to improve. The low relation identification scores make us reconsider the validity of temporal relations and the effectiveness of temporal guidelines.

One difficulty to annotate temporal relations is that there can be multiple interpretations of one document. There is no standard answer. For example, in the sentence “Terrified by the fear of being poisoned, the Cobbler confessed that he knew nothing about the medicine.” Some annotators think there is a Contains/Contained relation between the “terrified” event and “confessed” event because the terrifying feeling is a longer duration that contains the confession event. But some annotators think there is an Overlap relation between the “terrified” event and “confessed” event because these two events happen at the same time and share a time span. Even given the context, it is hard for the annotators to know when is the exact starting time and ending time for a specific event. There is a lack of information that even the text could not provide. The nuances between these two relations increase the level of difficulty to annotate temporal relations.

Another aspect is the learnability of the guideline. To make Bethard et al.’s guidelines (2012) more applicable to Reddit posts, we create more rules to identify events and temporal relations. The language for social media text is unprofessional and the posts are written by different authors. The performance of annotations can hardly improve even after rounds of guideline training as they always encounter new cases that the existing guidelines cannot cover. Guidelines should be user-friendly and concise for annotators to learn. Otherwise, it will take too much time training human annotators to reach a satisfying standard. The process of improving the guideline would be endless if we add more rules once we capture an irregular case. Previous work has done little in providing annotation solutions to unprofessional expressions. Our

research is a start in dealing with text with high randomness. Still, it is hard to generalize a rule for every unusual expression in the text. The result also leads to the question of how to enrich the content of the guideline without losing its learnability for the annotators.

Our current study also shows the challenges of generating a basic temporal relation scheme applying to different types of text. The annotation result shows that social media posts, children's stories, and news articles are three distinct types of texts. Our selected social media posts are written by college students using unprofessional expressions to talk about academic lives. The themes and the styles vary. Fables are edited by professionals and used for instructional purposes for children to read and learn. The plot is well-knit and the structure is organized with the main plot and a life lesson at the end. New reports are written on facts and strictly follow the inverted pyramid structure that emphasizes key information at the beginning and expands the entire story later. The events mentioned in different sources may lead to repetition of events that makes the temporal relations hard to identify. If we only focus on improving the annotation scores in each type of text, more rules can be created to fit into every single situation to improve the preciseness of annotation. But a guideline with fine-grained improvements specific to one type of text may not apply to another type of text. So in the research, we made some adjustments to the original rules to apply to the CNN news report, which we call Version 2. It is more suitable to develop a guideline with a specific focus on a type of text.

In future research, we can explore more types of text that have not been studied such as screenplay and speech. We can build a classification system to evaluate a variety of texts through style, point of view, fiction/non-fiction, structure, and tenses, which could be useful for later establishing temporal annotation rules for texts with similar categorizations. Also, we can explore more possibilities of combining different annotation schemes such as combining causal, aspectual, and temporal relations together to generate guidelines to make the annotation more precise and informative.

Chapter 6

Conclusion

In this study, we have shown that our temporal relation annotation guideline is possible with a high inter-annotator agreement in event identification applying to social media posts, children's stories, and news reports. Along with Bethard et al. (2012)'s no speech, no modal, no hypothetical, no negated, and paraphrasing rules, we introduce news rules of annotating events with copulas and prepositions, no imperative, no question, no exclamatives/swear words/exaggeration, and pair identification rules with a relation set of Before/After, Contains/Contained, and Overlap that efficiently improves both event and relation annotation performance in 4 rounds of annotation. While the score of temporal relation identification is below our expectations, we note that the difficulties come from the nuances of temporal relations such as the similarity in Overlap and Contains/Contained and the characteristics of different types of text including the mix of narrative, tenses, style, restatement, and points of views. We also quantify annotators' disagreement in event and relation identifications that can give some insights for future study in annotating different types of text. To our knowledge, it is the first work that researches temporal relations in a highly noisy type of text: social media posts. Also, it is an initial attempt to examine the effectiveness of a guideline applying to three different types of text in one work.

Bibliography

Alejandro Pimentel, Gemma Bel Enguix, Gerardo Sierra Martínez, and Azucena Montes. 2020. Temporal Relations Annotation and Extrapolation Based on Semi-intervals and Boundig Relations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3313–3323, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Amy Schade. Inverted Pyramid: Writing for Comprehension. Nielsen Norman Group. Accessed March 28, 2022. https://www.nngroup.com/articles/inverted-pyramid/#:~:text=In%20journalism%2C%20the%20inverted%20pyramid,supporting%20details%20and%20background%20information.

André Bittar, Caroline Hagège, Véronique Moriceau, Xavier Tannier, and Charles Teissèdre. 2012. Temporal Annotation: A Proposal for Guidelines and an Experiment with Inter-annotator Agreement. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3741–3745, Istanbul, Turkey. European Language Resources Association (ELRA).

Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathy McKeown. 2017. Analyzing the Semantic Types of Claims and Premises in an Online Persuasive Forum. In *Proceedings of the 4th Workshop on Argument Mining*, pages 11–21, Copenhagen, Denmark. Association for Computational Linguistics.

Hayley Ross, Jonathon Cai, and Bonan Min. 2020. Exploring Contextualized Neural Language Models for Temporal Dependency Parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8548–8553, Online. Association for Computational Linguistics.

James Pustejovsky, Jose Castano, Robert Ingria, Roser Sauri, Robert Gaizauskas, Andrea Setzer, and Graham Katz. 2003a. TimeML: Robust specification of event and temporal expressions in text. In IWCS-5 Fifth International Workshop on Computational Semantics.

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

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented

Interactive Annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico. Association for Computational Linguistics.

Jiarui Yao, Haoling Qiu, Bonan Min, and Nianwen Xue. 2020. Annotating Temporal Dependency Graphs via Crowdsourcing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5368–5380, Online. Association for Computational Linguistics.

Judy Hanwen Shen and Frank Rudzicz. 2017. Detecting Anxiety through Reddit. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*, pages 58–65, Vancouver, BC. Association for Computational Linguistics.

Jun Araki, Lamana Mulaffer, Arun Pandian, Yukari Yamakawa, Kemal Oflazer, and Teruko Mitamura. 2018. Interoperable Annotation of Events and Event Relations across Domains. In *Proceedings 14th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*, pages 10–20, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for Sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Lucia Donatelli, Michael Regan, William Croft, and Nathan Schneider. 2018. Annotation of Tense and Aspect Semantics for Sentential AMR. In Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018), pages 96–108, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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 Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 75–80, Prague, Czech Republic. 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*, pages 57–62, Uppsala, Sweden. Association for Computational Linguistics.

Nasrin Mostafazadeh, Alyson Grealish, Nathanael Chambers, James Allen, and Lucy Vanderwende. 2016.

CaTeRS: Causal and Temporal Relation Scheme for Semantic Annotation of Event Structures. In *Proceedings of the Fourth Workshop on Events*, pages 51–61, San Diego, California. Association for Computational Linguistics.

Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 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. Association for Computational Linguistics.

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 (Volume 1: Long Papers)*, pages 88–97, Jeju Island, Korea. Association for Computational Linguistics.

Paul Kingsbury and Martha Palmer. 2002. From TreeBank to PropBank. *In Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

Qiang Ning, Hao Wu, and Dan Roth. 2018. A Multi-Axis Annotation Scheme for Event Temporal Relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1318–1328, Melbourne, Australia. Association for Computational Linguistics.

Sahil Patel. Reddit Claims 52 Million Daily Users, Revealing a Key Figure for Social-Media Platforms. *The Wall Street Journal*. Dow Jones & Company, December 1, 2020. https://www.wsj.com/articles/reddit-claims-52-million-daily-users-revealing-a-key-figure-for-social-media-platforms-11606822200.

Steven Bethard, Oleksandr Kolomiyets, and Marie-Francine Moens. 2012. Annotating Story Timelines as Temporal Dependency Structures. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2721–2726, Istanbul, Turkey. European Language Resources Association (ELRA).

Tim O'Gorman, Kristin Wright-Bettner, and 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), pages 47–56, Austin, Texas. Association for Computational Linguistics.

Uniform Meaning Representation (UMR) 0.8 Specification. GitHub. Accessed March 28, 2022. https://github.com/umr4nlp/umr-guidelines/blob/master/guidelines.md#part-3-2-2-Non-participant-role-UMR-relations.

Van Gysel, Jens E. L., Meagan Vigus, Jayeol Chun, Kenneth Lai, Sarah Moeller, Jiarui Yao, Tim O'Gorman, et al. 2021. Designing a Uniform Meaning Representation for Natural Language Processing - Ki - Künstliche Intelligenz. *SpringerLink. Springer Berlin Heidelberg,* April 30, 2021. https://link.springer.com/article/10.1007/s13218-021-00722-w.

Yuchen Zhang and Nianwen Xue. 2018. Neural Ranking Models for Temporal Dependency Structure Parsing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3339–3349, Brussels, Belgium. Association for Computational Linguistics.