Estimating Narrative Durations: Proof of Concept

Mustafa Ocal, Akul Singh and Mark Finlayson

Knight Foundation School of Computing and Information Sciences Florida International University CASE Building, Room 362, 11200 S.W. 8th Street, Miami, FL USA 33199

Abstract

The duration of a narrative—that is, how long the events described in the story world take to unfold—is a feature of general interest to narrative understanding, and has been a topic of interest to theoreticians of narrative for some time. We combine prior work on timeline extraction (the TLEX algorithm) with an event duration estimation method that uses temporal pattern mining from large text corpora, to demonstrate a method for estimating the duration of narratives. We first gathered over 400K event durations mined from nearly 400M words from two corpora (the iWeb corpus and LexusNexis) using 10 hand-crafted temporal patterns. We then apply our approach to 30 selected stories from two corpora that already have annotated gold standard events and times (the ProppLearner corpus and the TimeBank corpus), estimating the duration of each story. We then conducted a preliminary evaluation using human judgements, showing that the durations extracted by the system are judged reasonable by people approximately 70% (for folktales) and 28% (for news stories) of the time. The gap between actual and desired performance reveals several challenges which are of interest, related both to duration estimation and its evaluation.

Keywords

Temporal Reasoning, Temporal Information Retrieval, TimeML, Duration

1. Introduction

Genette et al. [1] defines a *narrative* as "the representation of an event or of a sequence of events". This captures the notion—common to many definitions of the term—that narratives involve events and so can be placed in time. Genette himself was well known for examining the relationship between narrative time and discourse time, i.e., time as it unfolds in the story vs. the time taken to actually read the narrative [2]. More generally, it has been observed that the relationship between narrative and temporality is one of the most popular research areas in narrative theory [3].

Most approaches to narrative time require us to have some sense of the duration of events in the story world. Unfortunately, even the identification of basic event durations has been relatively understudied in NLP, and this is even more true of the duration of a complex sequence of events. Fortunately, several recent results have opened the way for more quantitative studies of narrative time (and, indeed, event sequences generally). In particular, TLEX, or TimeLine EXtraction, is a recently developed algorithm for extracting exact timelines from texts annotated with temporal annotation scheme TimeML [4]. Timelines expose the global order of events in

In: R. Campos, A. Jorge, A. Jatowt, S. Bhatia, M. Litvak (eds.): Proceedings of the Text2Story'24 Workshop, Glasgow (Scotland), 24-March-2024

mocal@fiu.edu (M. Ocal); asing118@fiu.edu (A. Singh); markaf@fiu.edu (M. Finlayson)

a narrative. Beyond this, all that is needed to compute the duration of the narrative are the durations of individual events and time periods.

We demonstrate a proof of concept of an approach to event duration estimation that uses temporal pattern mining from large text corpora, which allows us, in combination with TLEX, to extract durations of specific narratives. The event duration estimation system first extracts the durations of individual events from large text corpora using a set of 10 manually identified temporal patterns related to verbal events. This results in a large dataset of event duration statistics, from which we can compute average event durations for events expressed with verbs. The corpora used in this work included both news and editorial articles drawn from LexisNexus [5], as well as a selection of the intelligent Web (iWeb) corpus [6], resulting in a dataset with over 400,000 specific durations for almost 6,000 types of verbal events. Next, we used the TLEX algorithm [using provided implementations; 7] to extract the timelines of 15 narratives from a small corpus of folktales [the ProppLearner Corpus; 8], and 15 random news stories from the TimeBank corpus [9]. All of these narratives have gold-standard TimeML annotations, which comprise events, times, and temporal relations, but our approach could just as easily be applied on top of automatically computed TimeML annotations using existing automatic temporal annotation systems. Finally, our system combines the event durations with the timeline to estimate the overall duration of the narrative. We then conducted a preliminary evaluation using human judges, the result of which suggests that the technique has promise. It also reveals interesting challenges for evaluation as well as useful next steps.

The paper is organized as follows. First, we review TimeML, prior work on event duration extraction, and approaches for timeline extraction (§2). Next, we present the results of our event duration estimation and explain the overall narrative duration system in detail, as well as how we applied it to the TimeML corpora (§3). We next describe our preliminary evaluation using human raters, which shows reasonable agreement and performance modulo such a small sample (§4). Finally, we discuss the results and propose future work (§5), and provide a summary of the contributions (§6).

2. Related Work

2.1. Event Duration

Estimating event durations is a task that has been examined in prior NLP research. Prior work can be classified into two types. The first is *coarse-grained* classification, which predicts whether an event is more or less than some given amount of time, e.g., more or less than a day. The second type is *fine-grained* classification, which predicts the specific duration of an event, possibly grouping durations into bins, e.g., *seconds*, *minutes*, *hours*, *days*, *weeks*, *months*, or *years*.

Researchers have proposed a number of supervised machine learning-based solutions in both of these tasks. Pan et al. [10] presented a maximum entropy system that uses tokens, lemmas, POS tags, and subject-object relations as features, achieving 73.5% accuracy on coarse-grained classification and 61.9% on fine-grained. Similarly, Gusev et al. [11] also used a maximum entropy classifier but added named entities, verbs, verb types, and verb dependencies as features, achieving 74.8% accuracy on course-grained classification and 66% on fine-grained. Later, Vempala et al. [12] used convolutional neural networks with event class, POS tags, named entities, and dependencies as features, achieving 83.2% accuracy on coarse-grained classification. The most recent prior work presents rule-based systems instead of machine learning-based systems. Zhou et al. [13] defined a list of trigger words such as *for*, *since*, and *on*, which was then used in conjunction with semantic role labeling (SRL) to identify temporal prepositional phrases related to verbal events, and their system achieves 84.1% accuracy on coarse-grained classification. Finally, Yang et al. [14] defined temporal patterns involving keywords such as *for*, *take*, *spend*, *last*, *lasting*, *duration*, and *period*, and performs temporal pattern extraction, achieving 76.9% on coarse-grained classification and 76.2% on fine-grained classification.

These approaches have used two existing duration datasets: the McTaco dataset and the TimeBank Duration dataset. McTaco was built by giving crowdsourcers a sentence and asking them how long each event in the sentence lasted [15], which defined a possible range for each event. However, it should be noted that the dataset contains only a few hundred event durations. The TimeBank Duration dataset contains 58 news articles in which each event is annotated with lower-bound and upper-bound duration [16]. However, the corpus has only a 44% inter-annotator agreement for fine-grained classification, which is quite low; this perhaps is a result of the fact that determining the duration of an event can be highly subjective. The reason why prior approaches have high accuracy when the agreement score is low is that they use relaxed matching to evaluate their accuracy. For example, if the targeted event has [3 hours, 3 months] as lower and upper bound duration, a relaxed matching approach will accept the system's prediction as correct if the system's output is hours, days, weeks, or months.

In addition to low inter-annotator agreement, the TimeBank Duration dataset contains many questionable annotations. First, many of the duration ranges seem to depend strongly on the limited size of the corpus. The following example from the dataset indicates that **think** takes from 3 months to 3 years, **ignorance** and **fear** take from 1 year to 20 years, both oddly specific.

(1) "I **think**$_{[3MO,3Y]}$, a sense of, of **ignorance**$_{[1Y,20Y]}$ about Islam, a **fear**$_{[1Y,20Y]}$ about who Muslims are..."

Other problems include that the corpus assigns durations for words that are not events (e.g., nouns such as *dialect* or *language*). The corpus also contains finite durations for events that should have an unbounded duration, for example, *being* dead. Less problematically, they also assigned durations for the events that never happened, but this no doubt introduces an additional element of subjectivity (e.g., *the minister was not attacked*).

Since the McTaco dataset only provides the duration for a limited number of events, and the TimeBank Duration dataset contains numerous questionable annotations, there is a need for a new dataset that includes reliable and reasonable durations.

2.2. TimeML

We begin with a representation of temporal information in the text. We use TimeML, which is an annotation scheme for temporal information in text [17]. TimeML provides the ability to notate events, temporal expressions (i.e., dates, times, durations), and the relations between them that capture temporal information. TimeML annotations can be used for a variety of NLP tasks, including information extraction, question answering, summarization, and machine translation [18, 19, 20].

There are a number of manually and automatically annotated TimeML corpora. We used two: the ProppLearner corpus and the TimeBank corpus. ProppLearner consists of 15 texts comprising 18,862 words, with 3,438 events, 142 temporal expressions, and 2,778 temporal relationships, all of which were manually annotated [8]. On the other hand, the TimeBank corpus consists of 183 news stories gathered from diverse American news sources [9]. Note that while using texts with gold-standard event and time annotations allows us to evaluate our approach independent of noise or inaccuracies in the TimeML layer, our approach could just as easily be applied on top of automatically computed event and time annotations, using existing automatic temporal annotation systems, e.g., TARSQI [21], CAEVO [22], ClearTK [23], or others [24, 25].

2.3. Timeline Extraction

Timelines are almost never explicit in the texts and rarely can be extracted directly. There have been prior ML-based approaches to extracting timelines from TimeML annotations [26, 27, 28, 29]. However, these approaches have limitations, in particular, they do not handle all possible temporal relations, they ignore subordinated events, and they have less-than-perfect performance even on perfect TimeML annotations. Unlike prior approaches, Finlayson et al. [4] presented a Temporal Constraint Satisfaction Problem (TCSP)-based solution called TLEX (TimeLine EXtraction). TLEX converts TimeML graphs into temporally connected subgraphs, which are then converted to TCSPs, which can be solved using standard CSP techniques. Such solutions, by assigning integers to the start and end time points of temporal intervals in the TimeML graph, reveal a global order of events and times (i.e., timeline). Importantly, TLEX is deterministic, complete, and provably correct, resulting in 100% performance modulo the original TimeML graph. Also, later work presented a reference implementation of TLEX in the form of a Java library jTLEX [7], which we make use of in our implementation.

3. Methodology

Our system implements six steps. First, we preprocess a large dataset of free text by cleaning it and splitting it into sentences. Second, we use an off-the-shelf temporal parser [SynTime; 30] to detect temporal expressions (TIMEXes) in each sentence in the corpus. Third, for every sentence containing at least one TIMEX we identify the verbs and identify their lemmas. Fourth, we apply a set of temporal patterns to extract the duration expressed for the verb. When applied to the whole corpus, this data allows us to extract the mode of the duration of each verb lemma. Fifth, we used jTLEX to extract timelines of the ProppLearner and TimeBank texts and to identify "real-world" timelines and their events. Finally, we assign durations to the verbal events in the ProppLearner and TimeBank texts, which are placed in the timeline (the mode as extracted in step 4), allowing us to estimate the duration of the entire narrative.

3.1. Data Preprocessing

We mined temporal information from a large corpus of web documents, made up of two different collections. The first collection contains 8,059 news and editorial articles sourced from

LexusNexus [5], which we obtained from the authors. This collection ranges widely across topics. The second collection was a selection from the iWeb corpus, which is a web-based corpus and contains nearly 14 billion words from 22 million web pages [6]. Because iWeb is quite large, and we were only trying to develop a proof of concept, we used only the first 14 million sentences of the corpus. We chose these corpora not because they were particularly special or especially suitable for the task but rather because they were already available to us; other large collections of texts would no doubt have served equally well, and larger collections would have also potentially improved the overall accuracy of the duration estimations. We removed irrelevant content such as headlines, usernames, XML tags, phone numbers, and website names. For each text, we split the sentences using Spacy [31] and removed the duplicate sentences. This process resulted in a total of over 13 million unique sentences.

3.2. Temporal Expression Recognition

We next extracted sentences that contain possible event durations. We extracted temporal expressions using SynTime, which is a rule-based temporal expression detector [30]. It identifies the time tokens from raw text, then looks for modifiers and numerals next to time tokens to form time segments, and finally merges the time segments into time expressions. SynTime also identifies specific durations and normalizes them; for example, the temporal expression *six minutes* would result in a normalized duration of `PT6M`, which is a time duration expressed in ISO-8601. After the temporal expression recognition, we removed sentences without any temporal expressions, leaving 510,379 sentences.

3.3. Verb Detection

For each sentence with a temporal duration, we then identified verbs using the POS tagger found in the Stanford CoreNLP library [32]. We also identified lemmas using CoreNLP.

3.4. Temporal Pattern Mining

Although every remaining sentence contains a time expression, not every time expression is relevant to the verb duration in the sentence. For instance, in the sentence *He called me after 5 pm.*, the temporal expression “5 pm” is present but does not reveal anything about how long the call lasted. Indeed, there are many ways in which a TIMEX can be related to a verb. We defined 10 temporal patterns to express what we found were the most common relationships expressed between temporal durations and verbs, shown in Table 1. Note that the first five patterns provide the exact duration of an event, while patterns 6–9 provide an upper bound and pattern 10 a lower bound. Based on the number of durations extracted from the text, we estimate these patterns account for approximately 90–95% of the verb-duration relevant statements. A summary of the temporal patterns found is shown in Table 2.

We estimate the duration of any particular verbal event as follows. First, we extracted the duration of events from the candidates which match patterns 1–5. For 1, 4, and 5 the event duration will be the *normalized TIMEX value*. For 2 and 3, the event duration will be *TIMEX2 value minus TIMEX1 value*. For example, for “He *played* soccer from *1* pm to *2.30* pm.” the duration of *playing soccer* will be *2.30* - *1* = 1.5 hours (PT1.5H). When the boundaries of an event

1. **VERB + for + TIMEX**
He walked for 45 minutes.
2. **VERB + between + TIMEX + and + TIMEX**
The power went out between 4 pm and 6 pm.
3. **VERB + from? + TIMEX1 + “to” + TIMEX2**
He played soccer from 1 pm to 2.30 pm.
4. **VERB + (lasts|lasted) + TIMEX**
The drive lasted 40 minutes.
5. **VERB + PP? + (took|takes) + TIMEX**
Flying to NYC from Miami takes 3 hours.
6. **VERB + (on|in|at) + TIMEX**
I’m flying on Wednesday.
7. **VERB + daily**
He works out daily.
8. **VERB + (each|every) + TIMEX**
He ate donuts every morning in Paris.
9. **VERB + (per|once|twice|NUM times) + TIMEX**
Muslims pray five times a day.
10. **VERB + since + TIMEX**
She’s been doing the puzzle since 2 pm.

Table 1
List of Temporal Patterns

Pattern # →	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	
						on/		**each/**	**per/**		
Corpus	**For**	**Between**	**.. to ..**	**Last**	**Take**	**in/at**	**daily**	**every**	**times**	**since**	**Total**
LexisNexus	3,126	239	825	28	243	18,523	351	745	426	962	**25,468**
iWeb	72,943	4,229	8,228	860	5,439	268,247	10,038	17,739	13,923	18,121	**419,767**
Combined	**76,069**	**4,468**	**9,053**	**888**	**5,682**	**286,770**	**10,389**	**18,484**	**14,349**	**19,083**	**445,235**

Table 2
Results of temporal pattern mining for categories 1–10. The combined row is the statistics for the duration dataset that I created.

have different units (e.g., days versus hours), we first convert both boundaries into seconds, apply subtraction to ascertain the duration, and then convert the result back into the highest relevant time unit for ease of interpretation and consistency.

If the sentence does not match patterns 1–5, we checked for patterns 6–10 to extract the lower and upper-bound durations. For categories 6 and 8, the upper-bound duration is the normalized TIMEX value. For category 7, the upper-bound duration is always *PT24H* because the TIMEX is daily. Category 9 provides the upper-bound duration as the normalized TIMEX value divided by the repeating number. For example, once a day: PT24H / 1 = PT24H, twice a day: PT24H / 2 = PT12H, three times a day: PT24H / 3 = PT8H, and so on. On the other hand, category 10 provides the lower-bound duration, which is obtained by subtracting the normalized TIMEX value from DCT. For instance, in the sentence “The kid has been *watching* the cartoon since *8* AM (DCT = *9* AM)”. Here, the lower-bound duration is 9 AM–8 AM = PT1H.

Once all the sentences have been processed, and each extracted duration is associated with its verb lemma, we extract the most frequent duration (the mode), which is then passed on to the next stage as the duration of the event. To illustrate this, let’s consider the event *walking*. Our duration dataset contains 54 event durations from 37 sentences that fall under categories 1 to 5. Figure 1 shows a bar graph representation of these 54 event durations. Based on our duration reasoning system, we can conclude that *walking* takes between 3 minutes and 3 days, with the most likely duration being 30 minutes, which is the most frequent value. Importantly, we used the mode rather than the average because many verbs had very long upper tails (e.g., the verb *go* had a maximum duration of 40 years), which skewed the judgements.

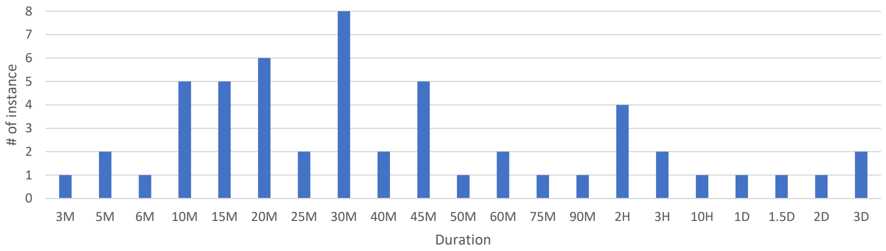

Figure 1: The bar graph of exact duration candidates for the event *walking*. The X-axis is the exact duration (M, minutes; H, hours; D, days), and the Y-axis is the number of instances of the duration.

3.5. Applying TLEX

We used jTLEX to generate timelines for ProppLearner texts[1] and randomly selected 15 TimeBank texts. The TLEX method differentiates events that happen on the "real-world" timeline and events that occur on possible, conditional, or modal world timelines (also called *subordinated* timelines). For instance, in the sentence "I went to the supermarket but forgot to buy milk." the going to the supermarket is represented as occurring in the timeline of the narrative, but the buying of the milk was forgotten, and so appears on a subordinated and hypothetical (negated) timeline. To compute the duration of the narrative, we use only the main, real-world timeline. From the timeline, we can then extract a sequence of events that "covers" the full timeline (i.e., time spans are not duplicated in different events). Combining event durations with the main timeline enables us to estimate the duration of a narrative. For each event (indicated by a verb lemma) in a text, we used the mode of duration as described above.

4. Evaluation

We recruited judges, who are undergraduate and graduate NLP researchers, to rate the reasonableness of the system's output [2]. We assigned nine judges sections of stories paired with the overall duration of the section as predicted by the system and asked to judge whether or not the system's estimation was reasonable. We defined "reasonable" as being within roughly 10% of what the judge would consider the "true" duration. When a judge marked an estimate as unreasonable, we asked them to explain why. We assigned sections of stories rather than whole stories because many of the folktales contained highly ambiguous starting or ending sections, such as "Once upon a time a man and a woman lived together in the woods." or "They lived happily ever after." Judges varied dramatically in their judgements of these types of events, and so we truncated stories to start and end with the first and last "short" event (as judged by us).

For 3 out of 15 sections from the ProppLearner stories, all nine judges marked the system's output reasonable. For 11 texts, the judges partially agreed, and for one text, all nine judges

[1] As described elsewhere [33], we used versions of ProppLearner in which inconsistent annotations were corrected.

[2] While we specifically chose judges with a background in narrative research, it's crucial to acknowledge that there's no method to select judges in a manner that eliminates subjectivity in evaluating event durations, given their inherently subjective nature.

judged the estimation unreasonable. The overall agreement score for our system's narrative duration estimation on ProppLearner stories was 0.696. Notably, the inter-annotator agreement for event duration in the fine-grained task within the TimeBank Duration corpus was 0.44 [10]. While a direct comparison between our narrative-level duration agreement and their event-level duration agreement is not feasible, it is evident that our system's performance was commendable.

On the other hand, for the TimeBank corpus, none of the stories received unanimous agreement from all nine judges in favor of the system's estimation. Surprisingly, seven out of the 15 stories had two or fewer judges marking the estimation as unreasonable. Consequently, the overall agreement score for our system's estimations on TimeBank stories was notably lower at 0.281. The main reason for that is that 7 out of 15 stories had *duration inconsistency*, which is a novel concept that we describe in more detail in Section 5.

5. Discussion & Future Work

We have introduced here a novel, yet simple, methodology for estimating the temporal duration of narratives. Consequently, we opted for an experimental evaluation approach, yielding agreement scores of 0.696 and 0.281. We investigated the lower agreement scores and identified five primary challenges and avenues for future research.

Notably, we initially attempted to use Large Language Models (LLMs), such as ChatGPT [34] or BARD [35], to evaluate our approach. We prompted the LLMs with identical story segments and solicited estimations for the duration of events. We observed that the LLMs struggled to predict temporal durations and often refused to produce an answer, many times requesting additional context. Therefore human evaluation was a more reliable and explainable method.

First, in our error analysis, we uncovered instances of what we term **duration inconsistency** in the temporal annotation. In such cases, the temporal annotation suggests that event *A* includes event *B*, yet the duration of event *A* is shorter than that of event *B*. Notably, this phenomenon was observed in 7 out of the 15 TimeBank stories. Interestingly, certain events lasting mere seconds sometimes included multiple events spanning days, weeks, months, or even years. For instance, in one story, the temporal annotation indicates that the event said, *The Pentagon said today it will re-examine the question of whether the remains inside the Tomb of the Unknown from the Vietnam War,* includes the event re-examine. However, our dataset reflects that the action of *saying* can be accomplished within seconds, while the action of *re-examining* may extend over weeks or months.

One of the most persistent difficulties we encountered is the **highly ambiguous nature of duration judgements**. In prior work by others, even where annotators were asked to judge the specific durations of events in context, the agreement was quite low, with agreement dropped as additional annotators were added. In our own evaluation, we noted similar problems. In an earlier version of our evaluation, we asked our judges to give us the duration of whole stories rather than relatively constrained sections. This resulted in hugely variable duration judgements, from as little as *1 day* to as much as (the oddly specific) *751 years 3 weeks 3 hours and 15 minutes.* The source of these ambiguities is numerous and can include different senses of the same word, differing patients and agents of the events, cultural or other commonsense context, or the level of detail provided.

Another place with room for improvement is how to pick the **specific duration** to use from the duration distribution. Here, we used the mode of the distribution, but as can be seen in Figure 1, the duration distributions are not necessarily normal or described by a common distribution function, and so more sophisticated means of selecting the appropriate duration in a specific context would be useful. In addition, it's important to note that the duration of activities, such as walking, can be highly subjective. For instance, if one were to consider the duration of walking as indicated by the mode of 30 minutes in folktales, it becomes evident that such narratives often depict characters embarking on journeys spanning days or even hours. These divergent values underscore the domain-dependency of duration metrics, indicating that data from one domain may not necessarily generalize well to another. In the case of walking, it might be more prudent to consider multiple values associated with this activity, each representing a different granularity (e.g., minutes, hours, days), and then select the appropriate duration based on the contextual requirements.

Event duration is heavily influenced by the **event arguments** (i.e., subject and object). While *walking* takes minutes, *walking with God* can take years. Another example is that sometimes a narrative can contain information about an action being repeated multiple times, or a single action being applied to multiple items as so implying a longer duration. For instance, in the following example, the seizing of 12 cows potentially takes longer than the seizing of a single cow (perhaps not 12 times as long though, depending on how the dragon does it). This was one of the reasons why two annotators disagreed with the system's duration estimation in this instance.

(2) The dragon flew into a rage and instead of six, seized twelve cows ...

Finally, a clear deficiency of the current work is that we focus only on durations of **events expressed with verbs**. Future work should expand the system to include non-verb events as well, which presumably will involve additional patterns for duration mining.

6. Contributions

Our contributions in this paper are three-fold. **First,** we presented a proof of concept method for narrative duration extraction. **Second,** we presented a large event duration dataset that consists of 445,235 durations for over 8,000 individual verbal event lemmas. **Third,** we note a number of interesting problems, especially with regard to challenges of evaluation, and note a number of directions for future improvements of the approach.

References

[1] G. Genette, A. Sheridan, M.-R. Logan, Figures of literary discourse, Columbia University Press New York, 1982.

[2] G. Genette, Narrative discourse: An essay in method, volume 3, Cornell University Press, 1983.

[3] M. Fludernik, Time in narrative, in: D. Herman, M. Jahn, M.-L. Ryan (Eds.), Routledge Encyclopedia of Narrative Theory, Routledge, New York, NY, 2005, pp. 608–612.

[4] M. A. Finlayson, A. Cremisini, M. Ocal, Extracting and aligning timelines, Computational Analysis of Storylines: Making Sense of Events (2021) 87.

[5] W. V. H. Yarlott, A. Ochoa, A. Acharya, L. Bobrow, D. C. Estrada, D. Gomez, J. Zheng, D. McDonald, C. Miller, M. A. Finlayson, Finding trolls under bridges: Preliminary work on a motif detector, arXiv preprint arXiv:2204.06085 (2022).

[6] M. Davies, J.-B. Kim, The advantages and challenges of" big data": Insights from the 14 billion word iweb corpus, Linguistic Research 36 (2019) 1–34.

[7] M. Ocal, A. Singh, J. Hummer, A. Radas, M. A. Finlayson, jTLEX: A Java Library for TimeLine EXtraction, in: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Dubrovnik, Croatia, 2023, p. 27–34.

[8] M. A. Finlayson, Propplearner: Deeply annotating a corpus of russian folktales to enable the machine learning of a russian formalist theory, Digital Scholarship in the Humanities 32 (2017) 284–300. URL: +http://dx.doi.org/10.1093/llc/fqv067. doi:10.1093/llc/fqv067.

[9] J. Pustejovsky, P. Hanks, R. Saurí, A. See, R. Gaizauskas, A. Setzer, D. Radev, B. Sundheim, D. Day, L. Ferro, M. Lazo, The timebank corpus, Proceedings of Corpus Linguistics (2003).

[10] F. Pan, R. Mulkar-Mehta, J. R. Hobbs, Learning event durations from event descriptions, in: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, 2006, pp. 393–400.

[11] A. Gusev, N. Chambers, D. R. Khilnani, P. Khaitan, S. Bethard, D. Jurafsky, Using query patterns to learn the duration of events, in: Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011), 2011, pp. 145–154.

[12] A. Vempala, E. Blanco, A. Palmer, Determining event durations: Models and error analysis, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), 2018, pp. 164–168.

[13] B. Zhou, Q. Ning, D. Khashabi, D. Roth, Temporal common sense acquisition with minimal supervision, arXiv preprint arXiv:2005.04304 (2020).

[14] Z. Yang, X. Du, A. Rush, C. Cardie, Improving event duration prediction via time-aware pre-training, arXiv preprint arXiv:2011.02610 (2020).

[15] B. Zhou, D. Khashabi, Q. Ning, D. Roth, " going on a vacation" takes longer than" going for a walk": A study of temporal commonsense understanding, arXiv preprint arXiv:1909.03065 (2019).

[16] F. Pan, R. Mulkar-Mehta, J. R. Hobbs, Annotating and learning event durations in text, Computational Linguistics 37 (2011) 727–752.

[17] J. Pustejovsky, J. Castaño, R. Ingria, R. Saurí, R. Gaizauskas, A. Setzer, G. Katz, TimeML: robust specification of event and temporal expressions in text, in: Fifth International Workshop on Computational Semantics (IWCS-5), 2003, pp. 1–11. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.161.8972&rep=rep1&type=pdf.

[18] B. Boguraev, R. K. Ando, Timeml-compliant text analysis for temporal reasoning., in: IJCAI, volume 5, 2005, pp. 997–1003.

[19] E. Saquete, P. Martínez-Barco, R. Muñoz, J. L. Vicedo, Splitting complex temporal questions for question answering systems, in: Proceedings of the 42Nd Annual Meeting

on Association for Computational Linguistics, ACL '04, Association for Computational Linguistics, Stroudsburg, PA, USA, 2004. URL: https://doi.org/10.3115/1218955.1219027. doi:10.3115/1218955.1219027.

[20] S. Liu, M. X. Zhou, S. Pan, Y. Song, W. Qian, W. Cai, X. Lian, Tiara: Interactive, topic-based visual text summarization and analysis, ACM Trans. Intell. Syst. Technol. 3 (2012) 25:1–25:28. URL: http://doi.acm.org/10.1145/2089094.2089101. doi:10.1145/2089094.2089101.

[21] M. Verhagen, I. Mani, R. Saurí, J. Littman, R. Knippen, S. B. Jang, A. Rumshisky, J. Phillips, J. Pustejovsky, Automating Temporal Annotation with TARSQI, in: ACL, 2005, pp. 81–84. URL: http://aclweb.org/anthology/P/P05/P05-3021.pdf.

[22] N. Chambers, T. Cassidy, B. McDowell, S. Bethard, Dense event ordering with a multi-pass architecture, Transactions of the Association for Computational Linguistics 2 (2014) 273–284. URL: https://doi.org/10.1162/tacl_a_00182. doi:10.1162/tacl_a_00182.

[23] S. Bethard, ClearTK-TimeML: A minimalist approach to tempeval 2013, in: Second joint conference on lexical and computational semantics (*SEM), volume 2: proceedings of the seventh international workshop on semantic evaluation (SemEval 2013), 2013, pp. 10–14.

[24] P. Mirza, S. Tonelli, Catena: Causal and temporal relation extraction from natural language texts, in: The 26th international conference on computational linguistics, ACL, 2016, pp. 64–75.

[25] M. Ocal, A. Perez, A. Radas, M. Finlayson, Holistic evaluation of automatic timeml annotators, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022, pp. 1444–1453.

[26] I. Mani, M. Verhagen, B. Wellner, C. M. Lee, J. Pustejovsky, Machine learning of temporal relations, in: Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (ICCL-ACL'06), 2006, pp. 753–760. Sydney, Australia.

[27] Q. X. Do, W. Lu, D. Roth, 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), 2012, pp. 677–687.

[28] O. Kolomiyets, S. Bethard, M.-F. Moens, Extracting narrative timelines as temporal dependency structures, in: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL'12), 2012, pp. 88–97.

[29] A. Leeuwenberg, M. F. Moens, Towards extracting absolute event timelines from english clinical reports, IEEE/ACM Transactions on Audio, Speech, and Language Processing 28 (2020) 2710–2719. doi:10.1109/TASLP.2020.3027201.

[30] X. Zhong, A. Sun, E. Cambria, Time expression analysis and recognition using syntactic token types and general heuristic rules, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017, pp. 420–429.

[31] Y. Vasiliev, Natural Language Processing with Python and SpaCy: A Practical Introduction, No Starch Press, 2020.

[32] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, D. McClosky, The stanford corenlp natural language processing toolkit, in: Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations, 2014, pp. 55–60.

[33] M. Ocal, M. Finlayson, Evaluating information loss in temporal dependency trees, in: Proceedings of The 12th Language Resources and Evaluation Conference, 2020, pp. 2148–2156.

[34] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[35] S. Narang, A. Chowdhery, Pathways language model (palm): Scaling to 540 billion parameters for breakthrough performance, Google AI Blog (2022).