Temporal Event Knowledge Acquisition via Identifying Narratives

Wenlin Yao and Ruihong Huang
Department of Computer Science and Engineering
Texas A&M University
{wenlinyao, huangrh}@tamu.edu

Abstract

Inspired by the double temporality characteristic of narrative texts, we propose a novel approach for acquiring rich temporal "before/after" event knowledge across sentences in narrative stories. The double temporality states that a narrative story often describes a sequence of events following the chronological order and therefore, the temporal order of events matches with their textual order. We explored narratology principles and built a weakly supervised approach that identifies 287k narrative paragraphs from three large text corpora. We then extracted rich temporal event knowledge from these narrative paragraphs. Such event knowledge is shown useful to improve temporal relation classification and outperform several recent neural network models on the narrative cloze task.

1 Introduction

Occurrences of events, referring to changes and actions, show regularities. Specifically, certain events often co-occur and in a particular temporal order. For example, people often go to *work* after *graduation* with a degree. Such "before/after" temporal event knowledge can be used to recognize temporal relations between events in a document even when their local contexts do not indicate any temporal relations. Temporal event knowledge is also useful to predict an event given several other events in the context. Improving event temporal relation identification and event prediction capabilities can benefit various NLP applications, including event timeline generation, text summarization and question answering.

While being in high demand, temporal event knowledge is lacking and difficult to obtain. Existing knowledge bases, such as Freebase (Bollacker et al., 2008) or Probase (Wu et al., 2012), often contain rich knowledge about entities, e.g., the birthplace of a person, but contain little event knowledge. Several approaches have been proposed to acquire temporal event knowledge from a text corpus, by either utilizing textual patterns (Chklovski and Pantel, 2004) or building a temporal relation identifier (Yao et al., 2017). However, most of these approaches are limited to identifying temporal relations within one sentence.

Michael Kennedy graduated with a bachelor's degree from Harvard University in 1980. He married his wife, Victoria, in 1981 and attended law school at the University of Virginia. After receiving his law degree, he briefly worked for a private law firm before joining Citizens Energy Corp. He took over management of the corporation, a non-profit firm that delivered heating fuel to the poor, from his brother Joseph in 1988. Kennedy expanded the organization goals and increased fund raising.

Beth paid the taxi driver. She jumped out of the taxi and headed towards the door of her small cottage. She reached into her purse for keys. Beth entered her cottage and got undressed. Beth quickly showered deciding a bath would take too long. She changed into a pair of jeans, a tee shirt, and a sweater. Then, she grabbed her bag and left the cottage.

Figure 1: Two narrative examples

Inspired by the double temporality characteristic of narrative texts, we propose a novel approach for acquiring rich temporal "before/after" event knowledge across sentences via identifying narrative stories. The double temporality states that a narrative story often describes a sequence of events following the chronological order and therefore, the temporal order of events matches with their textual order (Walsh, 2001; Riedl and Young, 2010; Grabes, 2013). Therefore, we can easily distill temporal event knowledge if we have identified a large collection of

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers), pages 537–547
Melbourne, Australia, July 15 - 20, 2018.

narrative texts. Consider the two narrative examples in figure 1, where the top one is from a news article of New York Times and the bottom one is from a novel book. From the top one, we can easily extract one chronologically ordered event sequence {graduated, marry, attend, receive, work, take over, expand, increase}, with all events related to the main character Michael Kennedy. While some parts of the event sequence are specific to this story, the event sequence contains regular event temporal relations, e.g., people often {graduate} first and then get {married}, or {take over} a role first and then {expand} a goal. Similarly, from the bottom one, we can easily extract another event sequence {pay, jump out, head, reach into, enter, undress, shower, change, grab, leave} that contains routine actions when people take a shower and change clothes.

There has been recent research on narrative identification from blogs by building a text classifier in a supervised manner (Gordon and Swanson, 2009; Ceran et al., 2012). However, narrative texts are common in other genres as well, including news articles and novel books, where little annotated data is readily available. Therefore, in order to identify narrative texts from rich sources, we develop a weakly supervised method that can quickly adapt and identify narrative texts from different genres, by heavily exploring the principles that are used to characterize narrative structures in narratology studies. It is generally agreed in narratology (Forster, 1962; Mani, 2012; Pentland, 1999; Bal, 2009) that a narrative is a discourse presenting a sequence of events arranged in their time order (the plot) and involving specific characters (the characters). First, we derive specific grammatical and entity co-reference rules to identify narrative paragraphs that each contains a sequence of sentences sharing the same actantial syntax structure (i.e., *NP VP* describing *a character did something*) (Greimas, 1971) and mentioning the same character. Then, we train a classifier using the initially identified seed narrative texts and a collection of grammatical, co-reference and linguistic features that capture the two key principles and other textual devices of narratives. Next, the classifier is applied back to identify new narratives from raw texts. The newly identified narratives will be used to augment seed narratives and the bootstrapping learning process iterates until no enough new narratives can be found.

Then by leveraging the double temporality characteristic of narrative paragraphs, we distill general temporal event knowledge. Specifically, we extract event pairs as well as longer event sequences consisting of strongly associated events that often appear in a particular textual order in narrative paragraphs, by calculating Causal Potential (Beamer and Girju, 2009; Hu et al., 2013) between events.

Specifically, we obtained 19k event pairs and 25k event sequences with three to five events from the 287k narrative paragraphs we identified across three genres, news articles, novel books and blogs. Our evaluation shows that both the automatically identified narrative paragraphs and the extracted event knowledge are of high quality. Furthermore, the learned temporal event knowledge is shown to yield additional performance gains when used for temporal relation identification and the Narrative Cloze task. The acquired event temporal knowledge and the knowledge acquisition system are publicly available[1].

2 Related Work

Several previous works have focused on acquiring temporal event knowledge from texts. VerbOcean (Chklovski and Pantel, 2004) used predefined lexico-syntactic patterns (e.g., "X and then Y") to acquire event pairs with the temporal *happens_before* relation from the Web. Yao et al. (2017) simultaneously trained a temporal "before/after" relation classifier and acquired event pairs that are regularly in a temporal relation by exploring the observation that some event pairs tend to show the same temporal relation regardless of contexts. Note that these prior works are limited to identifying temporal relations within individual sentences. In contrast, our approach is designed to acquire temporal relations across sentences in a narrative paragraph. Interestingly, only 195 (1%) out of 19k event pairs acquired by our approach can be found in VerbOcean or regular event pairs learned by the previous two approaches.

Our design of the overall event knowledge acquisition also benefits from recent progress on narrative identification. Gordon and Swanson (2009) annotated a small set of paragraphs presenting stories in the ICWSM Spinn3r Blog corpus (Burton et al., 2009) and trained a classifier using bag-of-words features to identify more stories. (Ceran

[1] http://nlp.cs.tamu.edu/resources.html

et al., 2012) trained a narrative classifier using semantic triplet features on the CSC Islamic Extremist corpus. Our weakly supervised narrative identification method is closely related to Eisenberg and Finlayson (2017), which also explored the two key elements of narratives, the plot and the characters, in designing features with the goal of obtaining a generalizable story detector. But different from this work, our narrative identification method does not require any human annotations and can quickly adapt to new text sources.

Temporal event knowledge acquisition is related to script learning (Chambers and Jurafsky, 2008), where a script consists of a sequence of events that are often temporally ordered and represent a typical scenario. However, most of the existing approaches on script learning (Chambers and Jurafsky, 2009; Pichotta and Mooney, 2016; Granroth-Wilding and Clark, 2016) were designed to identify clusters of closely related events, not to learn the temporal order between events though. For example, Chambers and Jurafsky (2008, 2009) learned event scripts by first identifying closely related events that share an argument and then recognizing their partial temporal orders by a separate temporal relation classifier trained on the small labeled dataset TimeBank (Pustejovsky et al., 2003). Using the same method to get training data, Jans et al. (2012); Granroth-Wilding and Clark (2016); Pichotta and Mooney (2016); Wang et al. (2017) applied neural networks to learn event embeddings and predict the following event in a context. Distinguished from the previous script learning works, we focus on acquiring event pairs or longer script-like event sequences with events arranged in a complete temporal order. In addition, recent works (Regneri et al., 2010; Modi et al., 2016) collected script knowledge by directly asking Amazon Mechanical Turk (AMT) to write down typical temporally ordered event sequences in a given scenario (e.g., shopping or cooking). Interestingly, our evaluation shows that our approach can yield temporal event knowledge that covers 48% of human-provided script knowledge.

3 Key Elements of Narratives

It is generally agreed in narratology (Forster, 1962; Mani, 2012; Pentland, 1999; Bal, 2009) that a narrative presents a sequence of events arranged in their time order (the plot) and involving specific characters (the characters).

Plot. The plot consists of a sequence of closely related events. According to (Bal, 2009), an event in a narrative often describes a "transition from one state to another state, caused or experienced by actors". Moreover, as Mani (2012) illustrates, a narrative is often "an account of past events in someone's life or in the development of something". These prior studies suggest that sentences containing a plot event are likely to have the actantial syntax "NP VP"[2] (Greimas, 1971) with the main verb in the past tense.

Character. A narrative usually describes events caused or experienced by actors. Therefore, a narrative story often has one or two main characters, called protagonists, who are involved in multiple events and tie events together. The main character can be a person or an organization.

Other Textual Devices. A narrative may contain peripheral contents other than events and characters, including time, place, the emotional and psychological states of characters etc., which do not advance the plot but provide essential information to the interpretation of the events (Pentland, 1999). We use rich Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2015) features to capture a variety of textual devices used to describe such contents.

4 Phase One: Weakly Supervised Narrative Identification

In order to acquire rich temporal event knowledge, we first develop a weakly supervised approach that can quickly adapt to identify narrative paragraphs from various text sources.

4.1 System Overview

The weakly supervised method is designed to capture key elements of narratives in each of two stages. As shown in figure 2, in the first stage, we identify the initial batch of narrative paragraphs that satisfy strict rules and the key principles of narratives. Then in the second stage, we train a statistical classifier using the initially identified seed narrative texts and a collection of soft features for capturing the same key principles and other textual devices of narratives. Next, the classifier is applied to identify new narratives from raw texts again. The newly identified narratives will be used to augment seed narratives and the bootstrapping

[2]NP is Noun Phrase and VP is Verb Phrase.

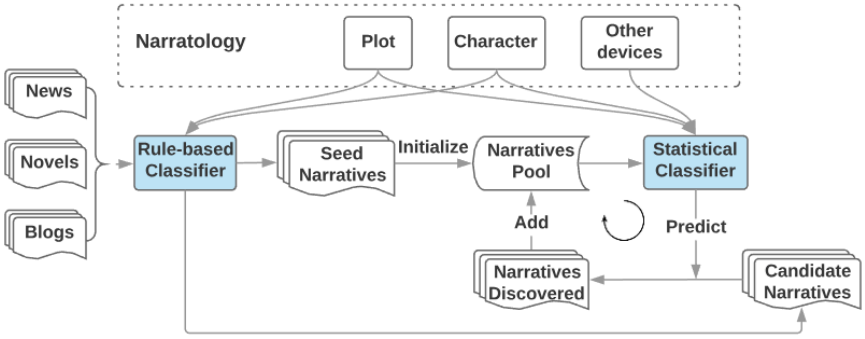

Figure 2: Overview of the Narrative Learning System

learning process iterates until no enough (specifically, less than 2,000) new narratives can be found. Here, in order to specialize the statistical classifier to each genre, we conduct the learning process on news, novels and blogs separately.

4.2 Rules for Identifying Seed Narratives

Grammar Rules for Identifying Plot Events. Guided by the prior narratology studies (Greimas, 1971; Mani, 2012) and our observations, we use context-free grammar production rules to identify sentences that describe an event in an actantial syntax structure. Specifically, we use three sets of grammar rules to specify the overall syntactic structure of a sentence. First, we require a sentence to have the basic active voiced structure "S → NP VP" or one of the more complex sentence structures that are derived from the basic structure considering Coordinating Conjunctions (CC), Adverbial Phrase (ADVP) or Prepositional Phrase (PP) attachments[3]. For example, in the narrative of Figure 1, the sentence *"Michael Kennedy earned a bachelor's degree from Harvard University in 1980."* has the basic sentence structure "S → NP VP", where the "NP" governs the character mention of 'Michael Kennedy' and the "VP" governs the rest of the sentence and describes a plot event.

In addition, considering that a narrative is usually "an account of past events in someone's life or in the development of something" (Mani, 2012; Dictionary, 2007), we require the headword of the VP to be in the past tense. Furthermore, the subject of the sentence is meant to represent a character. Therefore, we specify 12 grammar rules[4] to require the sentence subject noun phrase to have a simple structure and have a proper noun or pronoun as its head word.

For seed narratives, we consider paragraphs containing at least four sentences and we require 60% or more sentences to satisfy the sentence structure specified above. We also require a narrative paragraph to contain no more than 20% of sentences that are interrogative, exclamatory or dialogue, which normally do not contain any plot events. The specific parameter settings are mainly determined based on our observations and analysis of narrative samples. The threshold of 60% for "sentences with actantial structure" was set to reflect the observation that sentences in a narrative paragraph usually (over half) have an actantial structure. A small portion (20%) of interrogative, exclamatory or dialogue sentences is allowed to reflect the observation that many paragraphs are overall narratives even though they may contain 1 or 2 such sentences, so that we achieve a good coverage in narrative identification.

The Character Rule. A narrative usually has a protagonist character that appears in multiple sentences and ties a sequence of events, therefore, we also specify a rule requiring a narrative paragraph to have a protagonist character. Concretely, inspired by Eisenberg and Finlayson (2017), we applied the named entity recognizer (Finkel et al., 2005) and entity coreference resolver (Lee et al., 2013) from the CoreNLP toolkit (Manning et al., 2014) to identify the longest entity chain in a paragraph that has at least one mention recognized as a *Person* or *Organization*, or a gendered pronoun. Then we calculate the normalized length of this entity chain by dividing the number of entity mentions by the number of sentences in the paragraph. We require the normalized length of this longest

[3]We manually identified 14 top-level sentence production rules, for example, "S → NP ADVP VP", "S → PP , NP VP" and "S → S CC S". Appendix shows all the rules.

[4]The example NP rules include "NP → NNP", "NP → NP CC NP" and "NP → DT NNP".

entity chain to be ≥ 0.4, meaning that 40% or more sentences in a narrative mention a character [5].

4.3 The Statistical Classifier for Identifying New Narratives

Using the seed narrative paragraphs identified in the first stage as positive instances, we train a statistical classifier to continue to identify more narrative paragraphs that may not satisfy the specific rules. We also prepare negative instances to compete with positive narrative paragraphs in training. Negative instances are paragraphs that are not likely to be narratives and do not present a plot or protagonist character, but are similar to seed narratives in others aspects. Specifically, similar to seed narratives, we require a non-narrative paragraph to contain at least four sentences with no more than 20% of sentences being interrogative, exclamatory or dialogue; but in contrast to seed narratives, a non-narrative paragraph should contain 30% of or fewer sentences that have the actantial sentence structure, where the longest character entity chain should not span over 20% of sentences. We randomly sample such non-narrative paragraphs that are five times of narrative paragraphs [6].

In addition, since it is infeasible to apply the trained classifier to all the paragraphs in a large text corpus, such as the Gigaword corpus (Graff and Cieri, 2003), we identify candidate narrative paragraphs and only apply the statistical classifier to these candidate paragraphs. Specifically, we require a candidate paragraph to satisfy all the constraints used for identifying seed narrative paragraphs but contain only 30%[7] or more sentences with an actantial structure and have the longest character entity chain spanning over 20%[8] of or more sentences.

We choose Maximum Entropy (Berger et al., 1996) as the classifier. Specifically, we use the MaxEnt model implementation in the LIBLINEAR library[9] (Fan et al., 2008) with default parameter settings. Next, we describe the features used to capture the key elements of narratives.

Features for Identifying Plot Events: Realizing that grammar production rules are effective in identifying sentences that contain a plot event, we encode all the production rules as features in the statistical classifier. Specifically, for each narrative paragraph, we use the frequency of all syntactic production rules as features. Note that the bottom level syntactic production rules have the form of POS tag $\rightarrow$ WORD and contain a lexical word, which made these rules dependent on specific contexts of a paragraph. Therefore, we exclude these bottom level production rules from the feature set in order to model generalizable narrative elements rather than specific contents of a paragraph.

In addition, to capture potential event sequence overlaps between new narratives and the already learned narratives, we build a verb bigram language model using verb sequences extracted from the learned narrative paragraphs and calculate the perplexity score (as a feature) of the verb sequence in a candidate narrative paragraph. Specifically, we calculate the perplexity score of an event sequence that is normalized by the number of events, $PP(e_1, ..., e_N) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(e_i|e_{i-1})}}$, where N is the total number of events in a sequence and e_i is a event word. We approximate $P(e_i|e_{i-1}) = \frac{C(e_{i-1},e_i)}{C(e_{i-1})}$, where $C(e_{i-1})$ is the number of occurrences of e_{i-1} and $C(e_{i-1}, e_i)$ is the number of co-occurrences of e_{i-1} and e_i. $C(e_{i-1}, e_i)$ and $C(e_{i-1})$ are calculated based on all event sequences from known narrative paragraphs.

Features for the Protagonist Characters: We consider the longest three coreferent entity chains in a paragraph that have at least one mention recognized as a *Person* or *Organization*, or a gendered pronoun. Similar to the seed narrative identification stage, we obtain the normalized length of each entity chain by dividing the number of entity mentions with the number of sentences in the paragraph. In addition, we also observe that a protagonist character appears frequently in the surrounding paragraphs as well, therefore, we calculate the normalized length of each entity chain based on its presences in the target paragraph as well as one preceding paragraph and one follow-

[5] 40% was chosen to reflect that a narrative paragraph often contains a main character that is commonly mentioned across sentences (half or a bit less than half of all the sentences).

[6] We used the skewed pos:neg ratio of 1:5 in all bootstrapping iterations to reflect the observation that there are generally many more non-narrative paragraphs than narrative paragraphs in a document.

[7] This value is half of the corresponding thresshold used for identifying seed narrative paragraphs.

[8] This value is half of the corresponding thresshold used for identifying seed narrative paragraphs.

[9] https://www.csie.ntu.edu.tw/~cjlin/liblinear/

	0 (Seeds)	1	2	3	4	Total
News	20k	40k	12k	5k	1k	78k
Novels	75k	82k	24k	6k	2k	189k
Blogs	6k	10k	3k	1k	-	20k
Sum	101k	132k	39k	12k	3k	287k

Table 1: Number of new narratives generated after each bootstrapping iteration

ing paragraph. We use 6 normalized lengths (3 from the target paragraph [10] and 3 from surrounding paragraphs) as features.

Other Writing Style Features: We create a feature for each semantic category in the Linguistic Inquiry and Word Count (LIWC) dictionary (Pennebaker et al., 2015), and the feature value is the total number of occurrences of all words in that category. These LIWC features capture presences of certain types of words, such as words denoting relativity (e.g., motion, time, space) and words referring to psychological processes (e.g., emotion and cognitive). In addition, we encode Parts-of-Speech (POS) tag frequencies as features as well which have been shown effective in identifying text genres and writing styles.

4.4 Identifying Narrative Paragraphs from Three Text Corpora

Our weakly supervised system is based on the principles shared across all narratives, so it can be applied to different text sources for identifying narratives. We considered three types of texts: (1) **News Articles**. News articles contain narrative paragraphs to describe the background of an important figure or to provide details for a significant event. We use English Gigaword 5th edition (Graff and Cieri, 2003; Napoles et al., 2012), which contains 10 million news articles. (2) **Novel Books**. Novels contain rich narratives to describe actions by characters. BookCorpus (Zhu et al., 2015) is a large collection of free novel books written by unpublished authors, which contains 11,038 books of 16 different sub-genres (e.g., Romance, Historical, Adventure, etc.). (3) **Blogs**. Vast publicly accessible blogs also contain narratives because "personal life and experiences" is a primary topic of blog posts (Lenhart, 2006). We use the Blog Authorship Corpus (Schler et al., 2006) collected from the blogger.com website, which consists of 680k posts written by thousands of authors. We applied the Stanford CoreNLP tools (Manning et al., 2014) to the three text corpora to obtain POS tags, parse trees, named entities, coreference chains, etc.

In order to combat semantic drifts (McIntosh and Curran, 2009) in bootstrapping learning, we set the initial selection confidence score produced by the statistical classifier at 0.5 and increase it by 0.05 after each iteration. The bootstrapping system runs for four iterations and learns 287k narrative paragraphs in total. Table 1 shows the number of narratives that were obtained in the seeding stage and in each bootstrapping iteration from each text corpus.

5 Phase Two: Extract Event Temporal Knowledge from Narratives

Narratives we obtained from the first phase may describe specific stories and contain uncommon events or event transitions. Therefore, we apply Pointwise Mutual Information (PMI) based statistical metrics to measure strengths of event temporal relations in order to identify general knowledge that is not specific to any particular story. Our goal is to learn event pairs and longer event chains with events completely ordered in the temporal "before/after" relation.

First, by leveraging the double temporality characteristic of narratives, we only consider event pairs and longer event chains with 3-5 events that have occurred as a segment in at least one event sequence extracted from a narrative paragraph. Specifically, we extract the event sequence (the plot) from a narrative paragraph by finding the main event in each sentence and chaining the main events[11] according to their textual order.

Then we rank candidate event pairs based on two factors, how strongly associated two events are and how common they appear in a particular temporal order. We adopt the existing metric, Causal Potential (CP), which has been applied to acquire causally related events (Beamer and Girju, 2009) and exactly measures the two aspects. Specifically, the CP score of an event pair is calculated using the following equation:

$$cp(e_i, e_j) = pmi(e_i, e_j) + log\frac{P(e_i \rightarrow e_j)}{P(e_j \rightarrow e_i)} \quad (1)$$

where, the first part refers to the Pointwise Mutual Information (PMI) between two events and the

[10] Specifically, the lengths of the longest, second longest and third longest entity chains.

[11] We only consider main events that are in base verb forms or in the past tense, by requiring their POS tags to be VB, VBP, VBZ or VBD.

second part measures the relative ordering or two events. $P(e_i \rightarrow e_j)$ refers to the probability that e_i occurs before e_j in a text, which is proportional to the raw frequency of the pair. PMI measures the association strength of two events, formally, $pmi(e_i, e_j) = log\frac{P(e_i,e_j)}{P(e_i)P(e_j)}$, $P(e_i) = \frac{C(e_i)}{\sum_x C(e_x)}$ and $P(e_i, e_j) = \frac{C(e_i,e_j)}{\sum_x \sum_y C(e_x,e_y)}$, where, x and y refer to all the events in a corpus, $C(e_i)$ is the number of occurrences of e_i, $C(e_i, e_j)$ is the number of co-occurrences of e_i and e_j.

While each candidate pair of events should have appeared consecutively as a segment in at least one narrative paragraph, when calculating the CP score, we consider event co-occurrences even when two events are not consecutive in a narrative paragraph but have one or two other events in between. Specifically, the same as in (Hu and Walker, 2017), we calculate separate CP scores based on event co-occurrences with zero (consecutive), one or two events in between, and use the weighted average CP score for ranking an event pair, formally, $CP(e_i, e_j) = \sum_{d=1}^{3} \frac{cp_d(e_i,e_j)}{d}$.

Then we rank longer event sequences based on CP scores for individual event pairs that are included in an event sequence. However, an event sequence of length n is more than $n-1$ event pairs with any two consecutive events as a pair. We prefer event sequences that are coherent overall, where the events that are one or two events away are highly related as well. Therefore, we define the following metric to measure the quality of an event sequence:

$$CP(e_1, e_2, \cdots, e_n) = \frac{\sum_{d=1}^{3} \sum_{j=1}^{n-d} \frac{CP(e_j,e_{j+d})}{d}}{n-1}. \quad (2)$$

6 Evaluation

6.1 Precision of Narrative Paragraphs

From all the learned narrative paragraphs, we randomly selected 150 texts, with 25 texts selected from narratives learned in each of the two stages (i.e., seed narratives and bootstrapped narratives) using each of the three text corpora (i.e., news, novels, and blogs). Following the same definition "A story is a narrative of events arranged in their time sequence" (Forster, 1962; Gordon and Swanson, 2009), two human adjudicators were asked to judge whether each text is a narrative or a non-narrative. In order to obtain high inter-agreements, before the official annotations, we trained the two annotators for several iterations. Note that the texts we used in training annotators are different from the final texts we used for evaluation purposes. The overall kappa inter-agreement between the two annotators is 0.77.

Narratives	Seed	Bootstrapped
News	0.84	0.72
Novel	0.88	0.92
Blogs	0.92	0.88
AVG	0.88	0.84

Table 2: Precision of narratives based on human annotation

pairs	graduate → teach (5.7), meet → marry (5.3) pick up → carry (6.3), park → get out (7.3) turn around → face (6.5), dial → ring (6.3)
chains	drive → park → get out (7.8) toss → fly → land (5.9) grow up → attend → graduate → marry (6.9) contact → call → invite → accept (4.2) knock → open → reach → pull out → hold (6.0)

Table 3: Examples of event pairs and chains (with CP scores). → represents *before* relation.

Table 2 shows the precision of narratives learned in the two stages using the three corpora. We determined that a text is a correct narrative if both annotators labeled it as a narrative. We can see that on average, the rule-based classifier achieves the precision of 88% on initializing seed narratives and the statistical classifier achieves the precision of 84% on bootstrapping new ones. Using narratology based features enables the statistical classifier to extensively learn new narrative, and meanwhile maintain a high precision.

6.2 Precision of Event Pairs and Chains

To evaluate the quality of the extracted event pairs and chains, we randomly sampled 20 pairs (2%) from every 1,000 event pairs up to the top 18,929 pairs with CP score ≥ 2.0 (380 pairs selected in total), and 10 chains (1%) from every 1,000 up to the top 25,000 event chains[12] (250 chains selected in total). The average CP scores for all event pairs and all event chains we considered are 2.9 and 5.1 respectively. Two human adjudicators were asked to judge whether or not events are likely to occur in the temporal order shown. For event chains, we have one additional criterion requiring that events form a coherent sequence overall. An

[12]It turns out that many event chains have a high CP score close to 5.0, so we decided not to use a cut-off CP score of event chains but simply chose to evaluate the top 25,000 event chains.

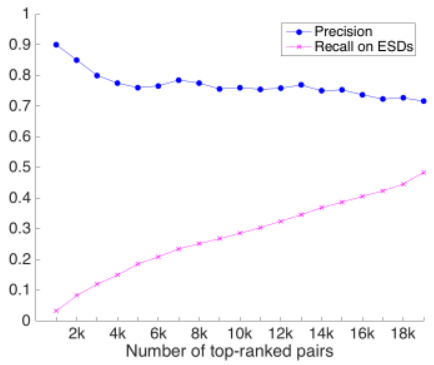

Figure 3: Top-ranked event pairs evaluation

# of top chains	5k	10k	15k	20k	25k
Precision	0.76	0.8	0.75	0.73	0.69

Table 4: Precision of top-ranked event chains

event pair/chain is deemed correct if both annotators labeled it as correct. The two annotators achieved kappa inter-agreement scores of 0.71 and 0.66, on annotating event pairs and event chains respectively.

As we know, coverage on acquired knowledge is often hard to evaluate because we do not have a complete knowledge base to compare to. Thus, we propose a pseudo recall metric to evaluate the coverage of event knowledge we acquired. Regneri et al. (2010) collected Event Sequence Descriptions (ESDs) of several types of human activities (e.g., baking a cake, going to the theater, etc.) using crowdsourcing. Our first pseudo recall score is calculated based on how many consecutive event pairs in human-written scripts can be found in our top-ranked event pairs. Figure 3 illustrates the precision of top-ranked pairs based on human annotation and the pseudo recall score based on ESDs. We can see that about 75% of the top 19k event pairs are correct, which captures 48% of human-written script knowledge in ESDs. In addition, table 4 shows the precision of top-ranked event chains with 3 to 5 events. Among the top 25k event chains, about 70% are correctly ordered with the temporal "after" relation. Table 3 shows several examples of event pairs and chains.

6.3 Improving Temporal Relation Classification by Incorporating Event Knowledge

To find out whether the learned temporal event knowledge can help with improving temporal relation classification performance, we conducted experiments on a benchmark dataset - TimeBank corpus v1.2, which contains 2308 event pairs that are annotated with 14 temporal relations [13].

Models	Acc.(%)
Choubey and Huang (2017)	51.2
+ CP score	**52.3**

Table 5: Results on TimeBank corpus

Method	Acc.(%)
(Chambers and Jurafsky, 2008)	30.92
(Granroth-Wilding and Clark, 2016)	43.28
(Pichotta and Mooney, 2016)	43.17
(Wang et al., 2017)	46.67
Our Results	**48.83**

Table 6: Results on MCNC task

To facilitate direct comparisons, we used the same state-of-the-art temporal relation classification system as described in our previous work Choubey and Huang (2017) and considered all the 14 relations in classification. Choubey and Huang (2017) forms three sequences (i.e., word forms, POS tags, and dependency relations) of context words that align with the dependency path between two event mentions and uses three bi-directional LSTMs to get the embedding of each sequence. The final fully connected layer maps the concatenated embeddings of all sequences to 14 fine-grained temporal relations. We applied the same model here, but if an event pair appears in our learned list of event pairs, we concatenated the CP score of the event pair as additional evidence in the final layer. To be consistent with Choubey and Huang (2017), we used the same train/test splitting, the same parameters for the neural network and only considered intra-sentence event pairs. Table 5 shows that by incorporating our learned event knowledge, the overall prediction accuracy was improved by 1.1%. Not surprisingly, out of the 14 temporal relations, the performance on the relation *before* was improved the most by 4.9%.

6.4 Narrative Cloze

Multiple Choice version of the Narrative Cloze task (MCNC) proposed by Granroth-Wilding and Clark (2016); Wang et al. (2017), aims to eval-

[13] Specifically, the 14 relations are *simultaneous, before, after, ibefore, iafter, begins, begun by, ends, ended by, includes, is included, during, during inv, identity*

uate understanding of a script by predicting the next event given several context events. Presenting a chain of contextual events $e_1, e_2, ..., e_{n-1}$, the task is to select the next event from five event candidates, one of which is correct and the others are randomly sampled elsewhere in the corpus. Following the same settings of Wang et al. (2017) and Granroth-Wilding and Clark (2016), we adapted the dataset (test set) of Chambers and Jurafsky (2008) to the multiple choice setting. The dataset contains 69 documents and 349 multiple choice questions.

We calculated a PMI score between a candidate event and each context event $e_1, e_2, ..., e_{n-1}$ based on event sequences extracted from our learned 287k narratives and we chose the event that have the highest sum score of all individual PMI scores. Since the prediction accuracy on 349 multiple choice questions depends on the random initialization of four negative candidate events, we ran the experiment 10 times and took the average accuracy as the final performance.

Table 6 shows the comparisons of our results with the performance of several previous models, which were all trained with 1,500k event chains extracted from the NYT portion of the Gigaword corpus (Graff and Cieri, 2003). Each event chain consists of a sequence of verbs sharing an actor within a news article. Except Chambers and Jurafsky (2008), other recent models utilized more and more sophisticated neural language models. Granroth-Wilding and Clark (2016) proposed a two layer neural network model that learns embeddings of event predicates and their arguments for predicting the next event. Pichotta and Mooney (2016) introduced a LSTM-based language model for event prediction. Wang et al. (2017) used dynamic memory as attention in LSTM for prediction. It is encouraging that by using event knowledge extracted from automatically identified narratives, we achieved the best event prediction performance, which is 2.2% higher than the best neural network model.

7 Conclusions

This paper presents a novel approach for leveraging the double temporality characteristic of narrative texts and acquiring temporal event knowledge across sentences in narrative paragraphs. We developed a weakly supervised system that explores narratology principles and identifies narrative texts from three text corpora of distinct genres. The temporal event knowledge distilled from narrative texts were shown useful to improve temporal relation classification and outperform several neural language models on the narrative cloze task. For the future work, we plan to expand event temporal knowledge acquisition by dealing with event sense disambiguation and event synonym identification (e.g., drag, pull and haul).

8 Acknowledgments

We thank our anonymous reviewers for providing insightful review comments.

References

Mieke Bal. 2009. *Narratology: Introduction to the theory of narrative*. University of Toronto Press.

Brandon Beamer and Roxana Girju. 2009. Using a bigram event model to predict causal potential. In *CICLing*. Springer, pages 430–441.

Adam L Berger, Vincent J Della Pietra, and Stephen A Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational linguistics* 22(1):39–71.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. AcM, pages 1247–1250.

Kevin Burton, Akshay Java, and Ian Soboroff. 2009. The icwsm 2009 spinn3r dataset. In *Third Annual Conference on Weblogs and Social Media (ICWSM 2009)*. AAAI.

Betul Ceran, Ravi Karad, Steven Corman, and Hasan Davulcu. 2012. A hybrid model and memory based story classifier. In *Proceedings of the 3rd Workshop on Computational Models of Narrative*. pages 58–62.

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*. Association for Computational Linguistics, pages 602–610.

Nathanael Chambers and Daniel Jurafsky. 2008. Unsupervised learning of narrative event chains. In *ACL*. volume 94305, pages 789–797.

Timothy Chklovski and Patrick Pantel. 2004. Verbocean: Mining the web for fine-grained semantic verb

relations. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*.

Prafulla Kumar Choubey and Ruihong Huang. 2017. A sequential model for classifying temporal relations between intra-sentence events. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. pages 1796–1802.

Oxford English Dictionary. 2007. Oxford english dictionary online.

Joshua Eisenberg and Mark Finlayson. 2017. A simpler and more generalizable story detector using verb and character features. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. pages 2698–2705.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *Journal of machine learning research* 9(Aug):1871–1874.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting on association for computational linguistics*. Association for Computational Linguistics, pages 363–370.

Edward Morgan Forster. 1962. Aspects of the novel. 1927. *Ed. Oliver Stallybrass* .

Andrew Gordon and Reid Swanson. 2009. Identifying personal stories in millions of weblog entries. In *Third International Conference on Weblogs and Social Media, Data Challenge Workshop, San Jose, CA*. volume 46.

Hebert Grabes. 2013. Sequentiality. *Handbook of Narratology* 2:765–76.

David Graff and C Cieri. 2003. English gigaword corpus. *Linguistic Data Consortium* .

Mark Granroth-Wilding and Stephen Clark. 2016. What happens next? event prediction using a compositional neural network model. In *AAAI*. pages 2727–2733.

Algirdas Julien Greimas. 1971. Narrative grammar: Units and levels. *MLN* 86(6):793–806.

Zhichao Hu, Elahe Rahimtoroghi, Larissa Munishkina, Reid Swanson, and Marilyn A Walker. 2013. Unsupervised induction of contingent event pairs from film scenes. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. pages 369–379.

Zhichao Hu and Marilyn Walker. 2017. Inferring narrative causality between event pairs in films. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*. pages 342–351.

Bram Jans, Steven Bethard, Ivan Vulić, and Marie Francine Moens. 2012. Skip n-grams and ranking functions for predicting script events. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 336–344.

Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2013. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics* 39(4):885–916.

Amanda Lenhart. 2006. *Bloggers: A portrait of the internet's new storytellers*. Pew Internet & American Life Project.

Inderjeet Mani. 2012. Computational modeling of narrative. *Synthesis Lectures on Human Language Technologies* 5(3):1–142.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *ACL (System Demonstrations)*. pages 55–60.

Tara McIntosh and James R Curran. 2009. Reducing semantic drift with bagging and distributional similarity. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*. Association for Computational Linguistics, pages 396–404.

Ashutosh Modi, Tatjana Anikina, Simon Ostermann, and Manfred Pinkal. 2016. Inscript: Narrative texts annotated with script information. In *LREC*. pages 3485–3493.

Courtney Napoles, Matthew Gormley, and Benjamin Van Durme. 2012. Annotated gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*. Association for Computational Linguistics, pages 95–100.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015. Technical report.

Brian T Pentland. 1999. Building process theory with narrative: From description to explanation. *Academy of management Review* 24(4):711–724.

Karl Pichotta and Raymond J Mooney. 2016. Learning statistical scripts with lstm recurrent neural networks. In *AAAI*. pages 2800–2806.

James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, et al. 2003. The timebank corpus. In *Corpus linguistics*. Lancaster, UK., volume 2003, page 40.

Michaela Regneri, Alexander Koller, and Manfred Pinkal. 2010. Learning script knowledge with web experiments. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 979–988.

Mark O Riedl and Robert Michael Young. 2010. Narrative planning: Balancing plot and character. *Journal of Artificial Intelligence Research* 39:217–268.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. 2006. Effects of age and gender on blogging. In *AAAI spring symposium: Computational approaches to analyzing weblogs*. volume 6, pages 199–205.

Richard Walsh. 2001. Fabula and fictionality in narrative theory. *Style* 35(4):592–606.

Zhongqing Wang, Yue Zhang, and Ching-Yun Chang. 2017. Integrating order information and event relation for script event prediction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. pages 57–67.

Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Q Zhu. 2012. Probase: A probabilistic taxonomy for text understanding. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*. ACM, pages 481–492.

Wenlin Yao, Saipravallika Nettyam, and Ruihong Huang. 2017. A weakly supervised approach to train temporal relation classifiers and acquire regular event pairs simultaneously. In *Proceedings of the 2017 Conference on Recent Advances in Natural Language Processing*. pages 803–812.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*. pages 19–27.

A Appendix

Here is the full list of grammar rules for identifying plot events in the seeding stage (Section 4.2).

Sentence rules (14):

- S → S CC S
- S → S PRN CC S
- S → NP VP
- S → NP ADVP VP
- S → NP VP ADVP
- S → CC NP VP
- S → PP NP VP
- S → NP PP VP
- S → PP NP ADVP VP
- S → ADVP S NP VP
- S → ADVP NP VP
- S → SBAR NP VP
- S → SBAR ADVP NP VP
- S → CC ADVP NP VP

Noun Phrase rules (12):

- NP → PRP
- NP → NNP
- NP → NNS
- NP → NNP NNP
- NP → NNP CC NNP
- NP → NP CC NP
- NP → DT NN
- NP → DT NNS
- NP → DT NNP
- NP → DT NNPS
- NP → NP NNP
- NP → NP NNP NNP