EXPLOITING DOCUMENT-LEVEL TEMPORAL RHYTHMS FOR EVENT TEMPORAL STATUS IDENTIFICATION

A Thesis

by

JUSTIN WILLIAM HILL

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	Ruihong Huang
Committee Members,	James Caverlee
	Laura Mandell
Head of Department,	Dilma Da Silva

August 2019

Major Subject: Computer Science

ABSTRACT

Previous research shows that it is a challenging task to determine the temporal statuses of event mentions relative to the document creation time because explicit temporal status cues, such as tense and aspect, are often lacking and an event mention's local context may be ambiguous. To further improve temporal status identification, we exploit the observation that document-level temporal rhythms reflective of story narrative structures exist as sequential patterns among the statuses of event mentions in a document. For example, a news article often starts by introducing the newsworthy event that may overlap with the document creation time, then describes precursory events, and closes by describing future implications. Experiments on the Richer Event Description and TimeBank corpora show that a simple neural network model aware of an event mention's position in a document significantly improves the performance of event temporal status identification. We also demonstrate that exploiting temporal rhythms enables data efficient transfer learning across document domains.

ACKNOWLEDGEMENTS

First, I must thank my committee chair, Dr. Huang, for her advisement and guidance in my research endeavors. I would also like to thank Dr. Mandell and Dr. Caverlee for sitting on my committee as well as Dr. Shipman for participating in my thesis defense. I would especially like to thank Kalain for his encouragement to always pursue the most interesting opportunities. Finally, I would like to thank my parents for making this possible.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a thesis committee consisting of Professor Ruihong Huang and Professor James Caverlee of the Department of Computer Science and Professor Laura Mandell of the Department of English.

All work conducted for the thesis was completed by the student independently, under the advisement of Professor Ruihong Huang of the Department of Computer Science.

Funding Sources

There are no outside funding contributions to acknowledge related to the research and compilation of this document.

NOMENCLATURE

CNN	Convolutional Neural Network
DCT	Document Creation Time
GloVe	Global Vectors for Word Representations
LSTM	Long Short-Term Memory
RED	Richer Event Description

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

1. INTRODUCTION

1.1 Overview

Our ability to discern whether an event mentioned in a document has already happened, is currently ongoing, or may happen sometime in the future allows us to reason about the chronology of events and gives way to our understanding of events' causal and coreference relationships. This skill can also benefit tasks such as news summarization (See et al., 2017) and timeline generation (Li and Cardie, 2014; Yan et al., 2011). However, discerning an event's temporal status requires complex semantic understanding of language and verse and remains a challenging task for state-of-the-art natural language processing methods.

Formally, we define the temporal status of an event as the status of that event at the time of its encompassing document's writing. An event's status can take on three possible values. The BEFORE status indicates the event has already occurred. The OVERLAP status indicates the event is occurring at the time of the document's writing. The AFTER status indicates the event has yet to occur. Consider the example in Figure 1 taken from a news article in the Richer Event Description corpus (O'Gorman et al., 2016). Events with the OVERLAP status are italicized and in orange.

The law *governs* the *privacy* of practically everything *entrusted* to the Internet—family photos *stored* with a Web service, journal entries *kept* online, company documents *uploaded* to the cloud, and the flurry of *emails exchanged* every day.

Figure 1: Excerpt from a News Article in the Richer Event Description Corpus Showing Events with the OVERLAP Status. Events with the OVERLAP Status are Italicized and in Orange.

A simple systematic method relying on word tense would incorrectly identify five of these events as having a BEFORE status. In fact, the only lexical clue in the local context that could indicate these events are ongoing is the “governs" event at the beginning of the sentence. In this case, local context is insufficient for determining these events’ statuses without advanced natural language understanding.

Alternatively, consider a model with knowledge that the excerpt is positioned in the middle of a larger section of text that almost exclusively discusses ongoing events. By identifying the location of the events in the document’s narrative, we can use this broader context to infer their status. In this new approach, we are taking advantage of an event’s status inherently being a relationship between the event and the document-at-large. Specifically, we use an event mention's position in a document's text to link it to the global document context and to ultimately make more informed event status identification decisions.

To better understand the relationship between an event’s position in a document and its status, we introduce the concept of a document’s temporal rhythm. We define a temporal rhythm to be a discernible sequential pattern among the statuses of event mentions in a document. For example, a news article about an election on election night may describe the election results coming in, summarize the precursory campaigns, and close by discussing the election’s future implications. While this is an oversimplification of the structure of most documents, it provides an illustration of what would be an OVERLAP-BEFORE-AFTER temporal rhythm.

In this thesis, we develop a neural network model that uses events’ positional information to learn aggregate temporal rhythms across documents. The model exploits

these rhythms to make more informed status identifications. The model also uses a self-attentive network to capture the local semantic context of an event mention. We perform experiments on the Richer Event Description (O'Gorman et al., 2016) and TimeBank (Pustejovsky et al., 2003) corpora to identify relative performance improvements resulting from the exploitation of temporal rhythms. We also perform experiments on the Richer Event Description corpus (O'Gorman et al., 2016) to evaluate the impact the presence of temporal rhythms has on transfer learning across document domains.

1.2 Related Work

Much of the prior research into the temporal properties of events focuses on classifying multiple temporal relationships such as the temporal order of event mentions, time expressions, and document creation times in tasks such as TempEval (Pustejovsky and Verhagen, 2009). Because these various temporal relationships have historically been considered together, systems capable of status identification have been designed using somewhat complex parsing schemes (UzZaman and Allen, 2010; Llorens et al., 2010). For example, the TIPSem system (Llorens et al., 2010) uses lexical, syntactic, and semantic features obtained via syntactic parsers (Charniak and Johnson, 2005), semantic role labelers (Punyakanok et al., 2004), and other handcrafted rules. Our methods differ with these prior approaches by focusing specifically on status identification and by using almost entirely deterministic inputs that do not suffer from upstream error propagation.

Event temporal status is a component of natural language that requires semantic knowledge of text to identify. Recent work using the EventStatus Corpus has shown the effectiveness of convolutional neural networks for capturing temporal compositionality of local context (Huang et al., 2016). More recent work improved upon these findings by

filtering the inputs of the neural network to be the words appearing between the event mention and the root of a dependency parse tree as well as the words that are governed by the mention (Dai et al., 2017). The modified input was intended to capture longer-distance information to improve status identification. Both works acknowledged the insufficiency of local context for status identification especially with respect to the OVERLAP and AFTER statuses, which often require the wider discourse context to resolve (Huang et al., 2016; Dai et al., 2017). In this thesis, we address these prior methods' contextual deficiencies by emphasizing the global relationship between an event mention and its encompassing document.

1.4 Temporal Rhythms

Temporal rhythms are emergent phenomena in the narrative structure of documents that increase the likelihood a given event has a specific status simply due to where it is located in a document. In Figure 2, we include the first quintile of a New York Times article from the validation set of the Richer Event Description corpus (O'Gorman et al., 2016) to illustrate a temporal rhythm. Event mentions are highlighted in blue, orange, and green for BEFORE, OVERLAP, and AFTER statuses, respectively.

UPDATING AN EMAIL LAW FROM THE PAST CENTURY

Steven Warshak, a Cincinnati businessman who built an empire selling male sexual enhancement drugs, was convicted of wire fraud several years ago, based in large part on his email correspondence, which authorities had extracted via a subpoena under a 1986 law governing electronic privacy.

A federal appeals court in Ohio later found, however, that the government had violated Warshak's constitutional right to privacy. The court said investigators should have convinced a judge that there was probable cause and obtained a search warrant, as though his messages had been stashed in a desk drawer. Although the court let the conviction stand, the case highlighted the conflicting legal rules that govern electronic privacy.

Congress is now set to clarify those rules, bringing that quarter-century-old law, the Electronic Communications Privacy Act, or ECPA, in line with the Internet age.

On Thursday, the Senate Judiciary Committee will start deliberating a measure that would require the government to get a search warrant, issued by a judge, to gain access to personal emails and all other electronic content held by a third-party service provider.

The current statute requires a warrant for emails that are less than 6 months old, but it lets the authorities gain access to older communications – or bizarrely, emails that have already been opened – with just a subpoena and no judicial review.

The law governs the privacy of practically everything entrusted to the Internet – family photos stored with a Web service, journal entries kept online, company documents uploaded to the cloud, and the flurry of emails exchanged every day. The problem is that it was written when the cloud was just a cumulus in the sky.

Figure 2: The First Quintile of a New York Times Article from the Richer Event Description Corpus Showing Events and Their Statuses. Events are Highlighted in Blue, Orange, and Green for BEFORE, OVERLAP, and AFTER Statuses, Respectively.

The document begins with the conviction of a businessman for wire fraud. The author then discusses deliberations of privacy laws. This is followed by details of the application of privacy laws to information on the Internet. Without understanding the chronology of events, it would be difficult to devise the overarching narrative of the article using such a general description. Knowing the first section of the excerpt is primarily composed of BEFORE events, the next section is primarily AFTER events, and the final section is primarily OVERLAP events clearly depicts a narrative centered around conflicting privacy laws, an upcoming discussion about potentially changing them, and the laws' issues that are ongoing at the time of the document's writing. The narrative structure in the excerpt could be described as a BEFORE-AFTER-OVERLAP temporal rhythm.

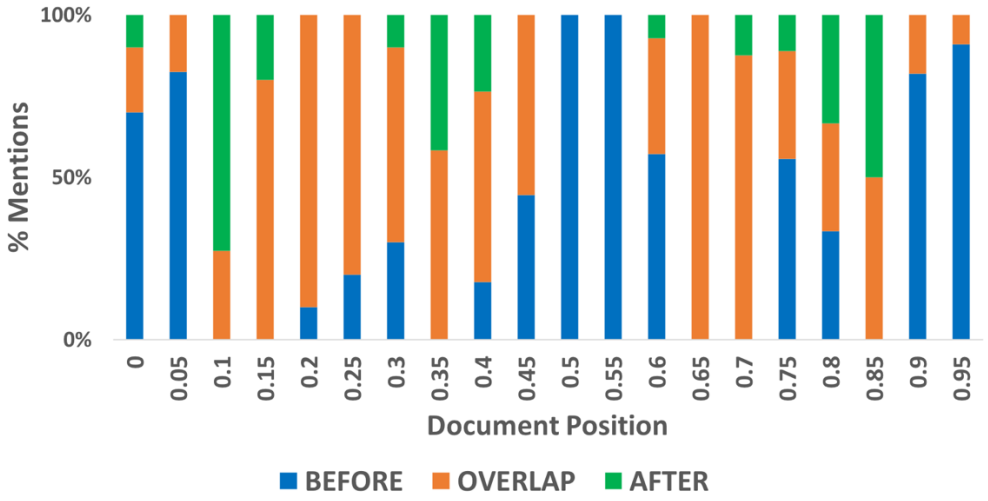

Figure 3: Temporal Rhythms for a New York Times Article from the Validation Set of the Richer Event Description Corpus.

For a temporal rhythm to be useful, contiguous blocks of only one status are not necessarily required. For example, there are some OVERLAP events dispersed amongst sections of primarily BEFORE and AFTER events. Rather, we are concerned with the presence of each status in each section of the narrative without any single status having to be dominant in a section. The normalized temporal rhythm for the entire document, of which the above excerpt is a part, is shown in Figure 3. The general BEFORE-AFTER-OVERLAP rhythm identified in the excerpt is visible in the 0 to 0.2 range of the figure.

The analysis of a single document suggests the position of an event mention in the text is a meaningful indicator of the event's likely status. While this provides valuable insight into the manifestation of a temporal rhythm in a document, we also want to analyze the aggregate temporal rhythms across multiple documents as they are more directly representative of the information being exploited by our proposed model for improving status identification.

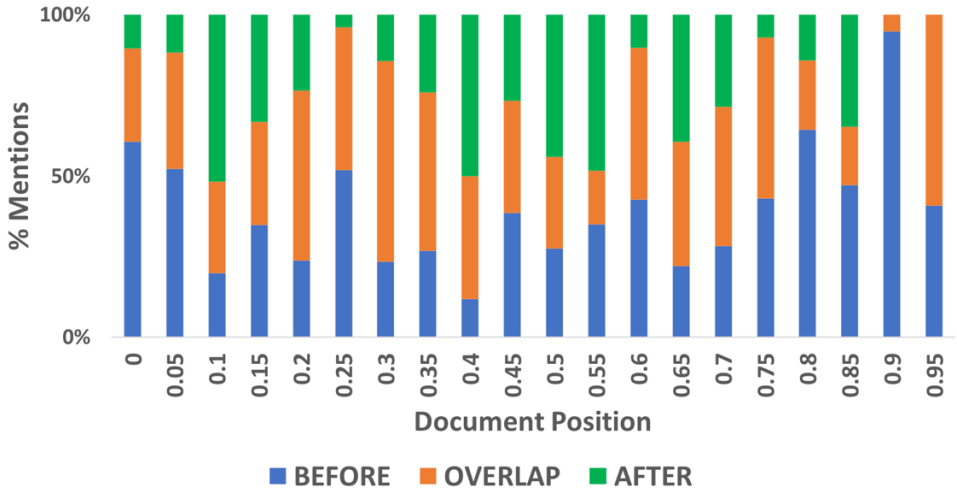

Figure 4: Aggregate Temporal Rhythms Present in Validation News Documents in the Richer Event Description Corpus.

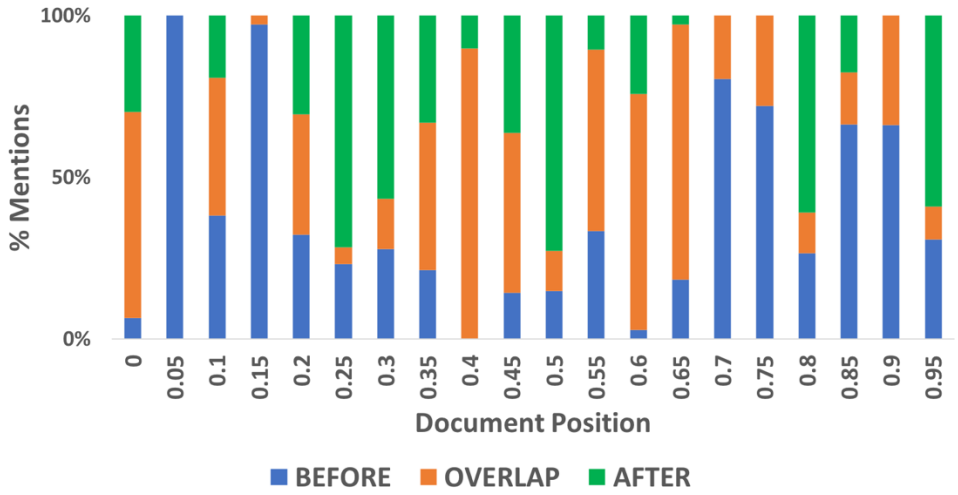

Figure 5: Aggregate Temporal Rhythms Present in Validation Forum Documents in the Richer Event Description Corpus.

Figure 4 shows the aggregate temporal rhythms present in news documents in the validation set of the Richer Event Description corpus (O'Gorman et al., 2016). The data is normalized for each position partition because we are more interested in the likelihood of an event's status in a given position than we are in the absolute number of events that are mentioned in that range. Figure 5 shows the same information as Figure 4 but for forum documents in the validation set.

1.3 Contributions and Outline

In Chapter 2, we develop a model capable of exploiting documents' temporal rhythms for improved event temporal status identification. We first describe how to represent language and position as input to the model, then we describe the model's architecture and how it differs from a model that does not use positional information.

In Chapter 3, we perform a series of quantitative experiments on the Richer Event Description (O'Gorman et al., 2016) and TimeBank (Pustejovsky et al., 2003) corpora.

We identify relative differences in performance between a model that exploits temporal rhythms and a model that does not. We also perform cross-validation experiments to evaluate the robustness of our proposed methods. Finally, we assess any benefits the presence of temporal rhythms may have on transfer learning across document domains.

In Chapter 4, we summarize our findings and discuss possible future directions for this area of research.

2. MODEL

Our goal is to develop a model that can exploit temporal rhythms for improved event status identification. This thesis consists of a series of quantitative experiments based on two status identification models. The first is a "Normal" model that does not use any positional information. The second is an "Enriched" model that includes positional information as input to exploit temporal rhythms. In this chapter, we describe how we create language and position representations for our models, and we describe the architecture of the Normal and Enriched models. We then provide details for training and optimization.

2.1 Representing Language

In order to capture semantic knowledge present in an event mention's local context, we include as input to our models a sequence of tokens centered around the event mention whose status we want to identify. Each token consists of the word itself and a corresponding part-of-speech tag, both of which are passed through separate embedding layers to create dense vector representations. The concatenation of the word embedding and the part-of-speech embedding produce a token embedding, e_t, for each time step t. The entire sequence of token embeddings in an event mention's local context window is denoted by E.

2.2 Representing Position

We define a mention's position to be the location of the mention among all tokens in a document. A mention's position vector, p, is a 1-hot vector where each dimension corresponds to a contiguous range of the document. Position is normalized by dividing the

index of the token by the number of tokens in the document and subsequently multiplying by the number of dimensions in p. The number of dimensions in p varies based on document domain and is determined experimentally.

2.3 Model Architecture

The architecture of both the Normal and Enriched models consists of two main components (see Figure 6). The context network produces a distributed representation, u, of an event mention's local context. The classification network takes that representation, and the position vector in the case of the Enriched model and produces a probability distribution across the three event status classes.

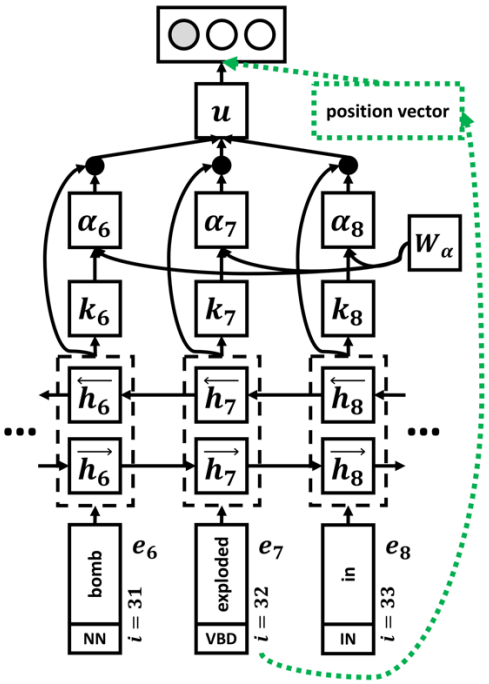

Figure 6: Model Architecture. Green Dotted Elements are Only Present in the Enriched Model.

We use a self-attentive network (Lin et al., 2017) for the context network due to its ability to capture multiple semantic views of a sequence. It uses an attention mechanism across the hidden representations, H, produced by a recurrent neural network. Specifically, we have chosen a bidirectional recurrent neural network that uses long short-term memory (LSTM) units (Hochreiter and Schmidhuber, 1997). For a sequence of length l, the context network is defined as

$$H = (h_1, h_2, \dots, h_l) = [\overrightarrow{\text{LSTM}}(E), \overleftarrow{\text{LSTM}}(E)]_t$$

$$K = (k_1, k_2, \dots, k_l) = \tanh\,(W_k H)$$

$$A = (\alpha_1, \alpha_2, \dots, \alpha_l) = \text{softmax}(W_\alpha K)$$

$$u = \text{flatten}(AH^T)$$

where K is a set of keys produced by an affine transformation, W_k, and a tanh nonlinearity and where A is a set of similarity scores between the keys in $K$and the r learned context vectors in W_α. The matrix resulting from $(h_1, h_2, \dots, h_l)$ being linearly combined using weights $(\alpha_1, \alpha_2, \dots, \alpha_l)$ is flattened into a vector, u.

The classification network is a simple 1-layer neural network with a SoftMax nonlinearity. The network's input in the Normal model is the distributed representation, u, produced by the context network. The network's input in the Enriched model is $[u, p]$, the concatenation of the context vector and the mention's position vector. In the case of an Enriched model being trained on multiple domains (e.g. the Richer Event Description corpus has two domains), p may be the concatenation of multiple position vectors where

each vector corresponds to one domain. Given a context vector, u, and a position vector, p, the classification network is defined as

$$s = \text{softmax}(W_s[u, p])$$

where s is a 3-dimensional vector representing a probability distribution for the given event across the three event status classes.

2.4 Training Details

The input sequence to the context network is composed of $l = 15$ lowercase tokens centered around the event mention under consideration. The words, part-of-speech tags, and indices for the tokens are obtained using Stanford CoreNLP (Manning et al., 2014). Zero-padding is used for those token sequences where the window size is smaller than l. We use 300-dimensional GloVe vectors pretrained on 42 billion Common Crawl tokens for word embeddings (Pennington et al., 2014). Part-of-speech embeddings are 20 dimensions and learned from scratch during training.

All models are trained using cross-entropy loss and the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 0.005. Each time the performance on the validation set drops below the best recorded performance, the batch size is doubled (Smith et al., 2018). The initial batch size is 32. After performance has dropped below the best recorded performance four times, training stops early. The context and classification networks have dropout (Srivastava et al., 2014) of 0.3 and 0.5, respectively.

Hidden states for the recurrent neural network are 50 dimensions in each direction so that each h_t has 100 dimensions. Because W_k is an affine transformation, each k_t also

has 100 dimensions. There are $r = 4$ hops of attention in W_α leading to u being a 400-dimensional vector. Position vectors have 15 and 30 dimensions for news and forum documents, respectively. All hyperparameters are tuned using Richer Event Description (O'Gorman et al., 2016) validation documents.

3. EXPERIMENTS

We have designed a series of experiments to evaluate the relative performances of Normal and Enriched models for event temporal status identification. In Section 3.1 and Section 3.2, we describe the corpora and the metrics we use to train and evaluate the models. In Section 3.3 and Section 3.4, we describe experiments for discovering the presence and robustness of any performance improvements that result from exploiting temporal rhythms. Finally in Section 3.5, we detail an experiment to uncover any advantages in data efficiency for transfer learning made possible by temporal rhythms.

3.1 Data

We train, tune, and evaluate the Normal and Enriched models on 8,568 event mentions from the Richer Event Description corpus (O'Gorman et al., 2016) and 1,105 mentions from the TimeBank corpus (Pustejovsky et al., 2003). The breakdown of both corpora by status is shown in Table 1. The Richer Event Description corpus consists of 45 forum and 50 news documents. We use the train, validation, and test splits suggested in the corpus. The TimeBank corpus consists of 183 news documents.

Status	# Mentions	% Mentions
Richer Event Description		
BEFORE	4,375	51.0
OVERLAP	2,977	34.7
AFTER	1,234	14.4
TimeBank		
BEFORE	582	52.7
OVERLAP	421	38.1
AFTER	102	9.2

Table 1: Breakdown of the Richer Event Description and TimeBank Corpora by Status.

Some preprocessing of both corpora is required to make their annotations suitable for the event status task. Each event in the Richer Event Description corpus (O'Gorman et al., 2016) has been annotated with a "DocTimeRel" that links the event to the document creation time. Notably, the dataset distinguishes between an OVERLAP DocTimeRel, which refers to events happening at the time of the document's writing, and a BEFORE/OVERLAP DocTimeRel, which refers to events that have some clearly indicated start time before the document's time of writing and continue through the document's time of writing. We combine the OVERLAP and BEFORE/OVERLAP DocTimeRel categories into a single OVERLAP category. The BEFORE and AFTER DocTimeRel annotations are equivalent to the BEFORE and AFTER statuses used in our experiments.

The TimeBank corpus (Pustejovsky et al., 2003) similarly annotates temporal relationships between events and their encompassing document's DCT, or document creation time. However, the DCT annotation is restricted to only those events whose stems occur twenty or more times in the corpus. The mapping between TimeBank DCT relations and our status categories is shown in Table 2.

Status	Event to DCT Temporal Relation Labels
BEFORE	BEFORE / IBEFORE / ENDED_BY
OVERLAP	DURING / INCLUDES / IS_INCLUDED / SIMULTANEOUS
AFTER	AFTER / IAFTER

Table 2: TimeBank Annotation Mappings Between Events and the Document Creation Time.

3.2 Evaluation Metrics

The performance of each model for each event status category is evaluated using recall, precision, and F1 scores defined as

$$\text{recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

$$\text{precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

$$\text{F1 score} = 2 \times \frac{\text{recall} \times \text{precision}}{\text{recall} + \text{precision}}$$

for an individual status. To compute a model's performance across the three classes, both micro and macro averages for recall, precision, and F1 scores are computed. Because micro-average metrics are affected by class imbalances, we use the macro-average F1 scores, the average of the computed F1 scores across the three status classes, as the evaluation metric for tuning hyperparameters and determining whether to double the batch size during training as detailed in Section 2.4.

3.3 The Impact of Positional Information

The primary objective of this experiment is to identify any relative performance improvements that result from exploiting documents' temporal rhythms. The Normal and Enriched model are both trained and evaluated on the Richer Event Description corpus (O'Gorman et al., 2016). In the first part of the experiment, the models are trained on both the news and forum domains present in the corpus. The Enriched model uses a concatenated position vector to learn separate aggregate rhythms for each domain. In the

second part of the experiment, we evaluate four distinct model instances by training the Normal and Enriched models independently on each domain.

3.3.1 Results

As shown in Table 3, our Normal model achieves a comparable and slightly better macro F1-score than the convolutional neural network (CNN) baseline used in previous work (Huang et al., 2016). We also observe the CNN overfits the BEFORE class resulting in higher micro performance than the Normal model due to the class imbalance favoring the BEFORE status. The Enriched event status model that exploits temporal rhythms significantly outperforms the Normal model, especially on identifying OVERLAP and AFTER statuses. By breaking out the results by domain, we observe the Enriched model is able to effectively exploit the rhythms in both the news and forum domains. On both domains, the Enriched model outperforms the Normal model across all metrics and status classes, save a single metric (news BEFORE precision with a 1.2% drop).

	BEFORE	**OVERLAP**	**AFTER**	**MACRO**	**MICRO**
CNN	**87.2**/76.0/81.2	46.6/63.0/53.6	38.4/**40.9**/39.6	57.4/60.0/58.1	69.2/69.2/69.2
Normal	73.1/**85.9**/79.0	54.4/55.1/54.8	57.1/32.9/41.8	61.6/58.0/58.5	65.6/65.6/65.6
Enriched	79.3/85.7/**82.4**	**55.4**/**63.3**/**59.1**	**61.6**/36.7/**46.0**	**65.4**/**61.9**/**62.5**	**70.0**/**70.0**/**70.0**
News Domain Only					
Normal	75.2/**88.3**/81.2	39.6/37.7/38.6	55.0/24.4/33.8	56.5/50.2/51.2	66.6/66.6/66.6
Enriched	**80.8**/87.1/**83.8**	**46.4**/**55.3**/**50.5**	**58.5**/**27.6**/**37.5**	**61.9**/**56.7**/**57.3**	**71.9**/**71.9**/**71.9**
Forum Domain Only					
Normal	65.3/73.6/69.2	60.9/68.1/64.3	59.6/41.5/48.9	62.0/61.0/60.8	62.2/62.2/62.2
Enriched	**73.5**/**78.3**/**75.8**	**63.5**/**69.6**/**66.4**	**60.3**/**46.1**/**52.2**	**65.8**/**64.6**/**64.8**	**66.2**/**66.2**/**66.2**

Table 3: Experimental Results on Richer Event Description Test Data for Models Trained Jointly on Both Document Domains. Each Cell Shows Recall/Precision/F1-Score.

The results of the second part of the experiment, shown in Table 4, indicate the Enriched model outperforms the Normal model when trained independently on each domain as well. We first note that performance across all statuses is lower in this individual scenario than in the previous joint scenario. This suggests the context network benefits from the larger amount of data available to the models when they are trained on both domains.

	BEFORE	OVERLAP	AFTER	MACRO	MICRO
News Domain Only					
Normal	74.8/84.5/79.3	35.7/36.4/36.0	56.1/26.4/35.9	55.5/49.1/50.4	65.2/65.2/65.2
Enriched	**77.1/86.8/81.7**	**39.3/40.0/39.6**	**58.5/27.9/37.8**	**58.3/51.6/53.0**	**67.8/67.8/67.8**
Forum Domain Only					
Normal	75.5/65.5/70.1	51.8/65.1/57.7	**50.0**/40.8/45.0	59.1/57.2/57.6	59.4/59.4/59.4
Enriched	**76.5/70.1/73.2**	**66.4/66.4/66.4**	46.6/**55.1/50.5**	**63.2/63.9/63.4**	**65.9/65.9/65.9**

Table 4: Experimental Results on Richer Event Description Test Data for Models Trained Independently Across Both Document Domains. Each Cell Shows Recall/Precision/F1-Score.

Comparing Table 3 to Table 4 also gives us a more granular understanding of the interplay in the Enriched model between a mention's local context and its position. Notably, there is a 3.5% relative underperformance in forum AFTER recall in the independent scenario (see Table 4) that is not present in the joint scenario (see Table 3). Conversely, the relative overperformance in news BEFORE precision present in the independent scenario (see Table 4) does not occur in the joint scenario (see Table 3). These observations suggest training in a single domain may result in the Enriched model being especially precise compared to the Normal model. This is to say, the Enriched model trained on both domains favors increased recall more so than the Normal model.

3.4 The Robustness of Temporal Rhythms

Because the test set of the Richer Event Description corpus (O'Gorman et al., 2016) contains only ten of the ninety-five documents in the corpus, we conduct additional experiments to confirm the broad applicability of exploiting temporal rhythms. We perform a 5-fold cross-validation on the training documents from the corpus. The folds are stratified so that they contain an approximately equal number of documents from both the news and forum domains. We also perform an additional 5-fold cross-validation using the TimeBank corpus (Pustejovsky et al., 2003).

3.4.1 Results

The results of the 5-fold cross-validation experiments shown in Table 5 verify the exploitability of temporal rhythms is robust and not specific only to the Richer Event Description (O'Gorman et al., 2016) test set. On both corpora, the Enriched model outperforms the Normal model across all metrics and status classes. In fact, the relative performance improvements for BEFORE and OVERLAP are nearly identical between the two corpora while the relative improvements in AFTER metrics for TimeBank (Pustejovsky et al., 2003) are even greater in magnitude.

RED	**BEFORE**	**OVERLAP**	**AFTER**	**MACRO**	**MICRO**
CNN	**74.5**/75.4/75.0	51.1/61.4/55.8	48.9/35.2/41.0	58.2/57.3/57.3	62.3/62.3/62.3
Normal	72.9/77.1/74.9	54.2/60.9/57.4	50.2/34.9/41.2	59.1/57.6/57.8	62.7/62.7/62.7
Enriched	73.9/**78.9**/**76.3**	**59.9**/**62.2**/**61.0**	**50.9**/**40.5**/**45.1**	**61.6**/**60.5**/**60.3**	**65.8**/**65.8**/**65.8**
TimeBank	**BEFORE**	**OVERLAP**	**AFTER**	**MACRO**	**MICRO**
Normal	74.2/77.9/76.0	67.1/61.8/64.3	41.8/47.0/44.2	61.1/62.2/61.5	68.7/68.7/68.7
Enriched	**75.5**/**81.2**/**78.2**	**73.5**/**63.7**/**68.3**	**47.8**/**64.0**/**54.7**	**65.6**/**69.6**/**67.1**	**72.2**/**72.2**/**72.2**

Table 5: Experimental Results of 5-Fold Cross-Validation on the Richer Event Description and TimeBank Corpora. Each Cell Shows Recall/Precision/F1-Score.

3.5 Transfer Learning Across Domains

Transfer learning is a valuable technique in machine learning that can reduce the amount of data required to train a model by initializing its weights with those from another model that has already been trained in a similar task or domain. Recent work in language modeling has shown the effectiveness of using pre-trained models for improved performance on a variety of natural language processing tasks (Devlin et al., 2018; Radford et al., 2018). With respect to status identification, we seek to uncover if the exploitation of temporal rhythms can enable data efficient transfer learning across document domains.

By separately encoding domain-specific temporal rhythms, our proposed Enriched model has the potential to enable data efficient transfer learning. We first train an Enriched model using all data from a source Domain X. We then clear the model's position weights and continue training using data from a target Domain Y. We determine the amount of data from Domain Y necessary for the adapted Enriched model to achieve performance parity with another Enriched model trained from scratch on all data in Domain Y. We perform this experiment on the Richer Event Description corpus (O'Gorman et al., 2016) in both directions where (X = news, Y = forum) and where (X = forum, Y = news). The experiment is performed twice in each direction so Domain Y documents are chosen by document length in either ascending or descending order.

The weights used to initialize the models in Domain X before continuing training in Domain Y are those obtained from training in the individual scenario in Section 3.3.

3.5.1 Results

Table 6 shows the Enriched model can quickly adapt across document domains by requiring fewer in-domain documents. It is notable that the Enriched model can adapt to the news domain much more efficiently than it can adapt to the forum domain. This is unsurprising as news documents are often much more structured and contain fewer narratives compared to forum documents which can contain disjoint statements from multiple authors.

Forum → News	**# Docs**	**% Mentions**
Longest Docs 1st	5	20.9
Shortest Docs 1st	25	47.8
News → Forum	**# Docs**	**% Mentions**
Longest Docs 1st	11	61.8
Shortest Docs 1st	34	80.1

Table 6: The Number of Documents and Percentage of Mentions Required to Adapt from One Domain to Another in the Richer Event Description Corpus.

The results also show using longer documents is an important factor for the efficiency of transfer learning. For example, using the shortest documents in the news domain requires 5x as many documents, or 2.3x as much data, as using the longest documents to adapt the model. The advantage of using longer documents lies in the increased positional granularity that corresponds to more detailed temporal rhythms than what are available in shorter documents.

4. CONCLUSIONS

We introduced the concept of document-level temporal rhythms and the use of an event mention's position as a way of connecting the event to the global document context. Our experiments showed using positional information allows a neural network model to exploit temporal rhythms for improved event temporal status identification. These improvements were found to be robust across document domains and across multiple corpora as well as when the models were trained both jointly and independently across document domains.

We also showed temporal rhythms allow for efficient transfer learning across document domains. As a result of our transfer learning experiments, it was revealed using longer documents is more effective for learning temporal rhythms due to their increased positional granularity. Finally, models more easily adapted to news documents than forum documents which suggests increased narrative structure corresponds to more easily learned temporal rhythms.

4.1 Future Work

In this thesis, we connected events to the global document context by using their position and our models used this positional information to learn aggregate temporal rhythms. The next step in applying these methods might be to develop a model that can learn multiple unique, rather than aggregate, temporal rhythms. However, this and additional methods that use discourse-level properties to identify events' statuses would require the more challenging joint consideration of events across a document's narrative.

REFERENCES

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine *n*-best Parsing and MaxEnt Discriminative Reranking. In *Proceedings of the 43rd Annual Meeting of the ACL*, pages 173-180, Ann Arbor, Michigan. Association for Computational Linguistics.

Zeyu Dai, Wenlin Yao, and Ruihong Huang. 2017. Using Context Events in Neural Network Models for Event Temporal Status Identification. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 234-239, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR*, abs/1810.04805.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-term Memory. *Neural Computation*, 9(8):1735-1780.

Ruihong Huang, Ignacio Cases, Dan Jurafsky, Cleo Condoravdi, and Ellen Riloff. 2016. Distinguishing Past, On-going, and Future Events: The EventStatus Corpus. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 44-54, Austin, Texas. Association for Computational Linguistics.

Diederik Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations*.

Jiwei Li and Claire Cardie. 2014. Timeline Generation: Tracking Individuals on Twitter. In *Proceedings of the 23rd International Conference on World Wide Web*. Seoul, Korea.

Zhouhan Lin, Minwei Fang, Cícero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A Structured Self-attentive Sentence Embedding. *CoRR*, abs/1703.03130.

Hector Llorens, Estela Saquete, and Borja Navarro. 2010. TIPSem (English and Spanish): Evaluating CRFs and Semantic Roles in TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 284-291, Stroudsburg, Pennsylvania. Association for Computational Linguistics.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55-60, Baltimore, Maryland. Association for Computational Linguistics.

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.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on*

Empirical Methods of Natural Language Processing, pages 1532-1543, Doha, Qatar. Association for Computational Linguistics.

Vasin Punyakanok, Dan Roth, Wen-tau Yih, Dav Zimak and Yuancheng Tu. 2004. Semantic Role Labeling via Generalized Inference Over Classifiers. In *Proceedings of the Eight Conference on Computational Natural Language Learning (CoNLL)*, pages 130-133, Boston, Massachusetts. Association for Computational Linguistics.

James Pustejovsky, Patrick Hanks, Roser Saur, Andrew See, Rob Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, and Marcia Lazo. 2003. The TimeBank Corpus. *Proceedings of Corpus Linguistics*.

James Pustejovsky and Marc Verhagen. 2009. Evaluating Events, Time Expressions, and Temporal Relations (TempEval-2).

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-training.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the Point: Summarization with Pointer-Generator Networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1073-1083, Vancouver, Canada. Association for Computational Linguistics.

Samuel L. Smith, Pieter-Jan Kindermans, and Quoc V. Le. 2018. Don't Decay the Learning Rate, Increase the Batch Size. In *International Conference on Learning Representations*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15:1929-1958.

Naushad UzZaman and James F. Allen. 2010. TRIPS and TRIOS System for TempEval-2: Extracting Temporal Information from Text. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 276-283, Stroudsburg, Pennsylvania. Association for computational Linguistics.

Rui Yan, Liang Kong, Congrui Huang, Xiaojun Wan, Xiaoming Li, and Yan Zhang. 2011. Timeline Generation through Evolutionary Trans-temporal Summarization. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 433-443, Edinburgh, Scotland. Association for Computational Linguistics.