NEWS DISCOURSE STRUCTURE-GUIDED APPROACHES FOR EVENT COREFERENCE RESOLUTION

A Dissertation

by

PRAFULLA KUMAR CHOUBEY

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

DOCTOR OF PHILOSOPHY

Chair of Committee,	Ruihong Huang
Committee Members,	James Caverlee
	Yoonsuck Choe
	Laura Mandell
Head of Department,	Scott Schaefer

May 2021

Major Subject: Computer Science

ABSTRACT

Event coreference resolution aims to determine and cluster event mentions that refer to the same real-world event. It is a relatively less studied natural language processing (NLP) task despite being crucial for various NLP applications such as topic detection and tracking, question answering, and summarization.

A typical event coreference resolution system relies on scoring similarity between two event mentions in a document followed by clustering. However, event coreference chains are sparsely distributed and only certain key events that connect other peripheral events in a document are repeated to organize content and produce a coherent story. This makes manually labeling many event coreference relations very time-consuming. Furthermore, event mentions tend to appear in diverse contexts and few are accompanied by a full set of their arguments. The three challenges, the distributional sparsity of coreferential event mentions, the absence of abundant human-annotated event coreference data, and the high diversity of contexts containing coreferential event mentions, make it hard to build effective event coreference resolution systems.

The primary goal of this dissertation is to develop a holistic approach that can successfully model document-level content structures to overcome the problems arising due to the sparse distribution of event coreference chains. To that end, we first study the discourse-level significance of an event that has many coreferential mentions in a document and devise a heuristics-based approach that captures several specific distributional patterns of coreferential event mentions. Inspired by the empirical improvement of the heuristics-based approach, we propose a new task of news discourse profiling, grounded in the news discourse theories, to identify document-level content structures and present a systematic method to incorporate them into an event coreference resolution system. Besides outperforming the heuristics-based model, the news discourse profiling-based system is capable of explaining the nature of correlations between coreferential event mentions and content structures. Consequently, we leverage the correlations between news discourse profiling and event coreference relations and define several rules to automatically collect event pairs from unlabeled

news documents. Through both manual validation and empirical evaluations, we show that news discourse profiling additionally enables us to overcome the annotational sparsity.

Overall, this dissertation contributes to the current literature on event coreference resolution by adopting news discourse structure-centric approaches that are orthogonal to supervised feature-based pairwise classifiers. News discourse structure, when incorporated through explicit constraints or used to automatically acquire data from unlabeled news documents, adds to the performance of pairwise event coreference classifiers. I hope that the work done in this dissertation potentially inspires new work on analyzing and modeling discourse structure theories to improve event coreference resolution across text genres and languages.

To my family.

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor, Dr. Ruihong Huang, for taking me as her student. I am grateful for her support and guidance during the last five years. Her deep insight into language has helped immensely in my research. Thanks for always encouraging me to focus on the linguistics component of NLP research rather than black-box modeling.

I am extremely grateful to my advisory committee, Dr. James Caverlee, Dr. Yoonsuck Choe, and Dr. Laura Mandell for their mentorship and valuable input on my research. Thanks to our graduate advisor, Dr. Duncan M. Walker, and Karrie Bourquin for being so approachable and providing all the administrative support.

I would also like to acknowledge the support of our NLP group members. Thanks to Wenlin Yao, Ajit Jain, Zeyu Dai, Pei Chen, Yuanyuan Lei, Ayesha Qamar, Zhuoer Wang, Lei Gao, Kaushik Raju, Girish Kasiviswanathan, Sanuj Sharma, Aaron Lee, James Motes, Amulya Agarwal, Bohan Zhang, Sanjeev Kumar Singh, Rizu Jain, Justin Hill, Jack Brady and Abhinav Chandar for their feedback on my projects.

A special thanks to our department staff for their assistance over the years.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a dissertation committee consisting of Dr. Ruihong Huang (advisor), Dr. James Caverlee, and Dr. Yoonsuck Choe from the Department of Computer Science and Engineering, and Dr. Laura Mandell from the Department of English.

The data for chapter 3 was annotated in part by Kaushik Raju, and the data for chapter 5 was annotated in part by Aaron Lee. Further, the baseline random walk-based ranking system for main event identification (chapter 3) was implemented by Kaushik Raju. Both Kaushik Raju and Aaron Lee were from the Department of Computer Science and Engineering. All remaining work for the dissertation was completed by the student, in collaboration with Dr.Ruihong Huang.

Funding Sources

This work was partially supported by the National Science Foundation via NSF Awards IIS-1755943 and IIS-1942918. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of NSF or the U.S. Government.

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

A story is formed by a series of related events and their participants. For deep story understanding, we require collective interpretation of all information related to events and their participants in a document. However, the sequential nature of language, which results in a scattered distribution of information throughout the document, makes it extremely difficult to automatically aggregate and process every piece of information.

Consider the example news article in Figure 1.1. Here, a system modeling local semantics may be able to extract information such as *"Lovers' Lane killer was executed, he tried robbing a man in 2003, and he was convicted for killing a man"*. However, questions such as *"who was Lovers' Lane killer? whom did he kill? how did he kill that man? how was he executed?"* are not directly identifiable and require an additional mechanism to stitch together complementary information from multiple sentences. Event coreference resolution provides one such mechanism to combine information from multiple sentences. For instance, we can infer that Juan Castro was called Lovers' Lane killer and he was executed by lethal injection by identifying that **executes** in "H" and **put to death** in "S2" are referring to the same real-world event. Similarly, if we know that **shot** in "S1, S5" and **killing** in "S4" are referring to the same real event, we can infer that Juan Castro shot Tommy Garcia seven times. Here, event coreference resolution can be utilized to aggregate information and accordingly can help downstream applications such as question answering [Narayanan and Harabagiu, 2004, Bikel and Castelli, 2008] and information extraction [Humphreys et al., 1997].

Secondly, in Figure 1.1, we observe that the sentences describing coreferential events *"execution of Juan Castillo"* and *"killing of Tommy Garcia"* describe information that is more relevant to the story than other sentences. For NLP tasks that require relevance information, such as text summarization [Li et al., 2006] and event saliency identification [Choubey et al., 2018], we can leverage cues from event coreference chains to improve systems' performance. For instance, we can directly extract S1, S2, S8, and S11 from the document in Figure 1.1 and generate either of

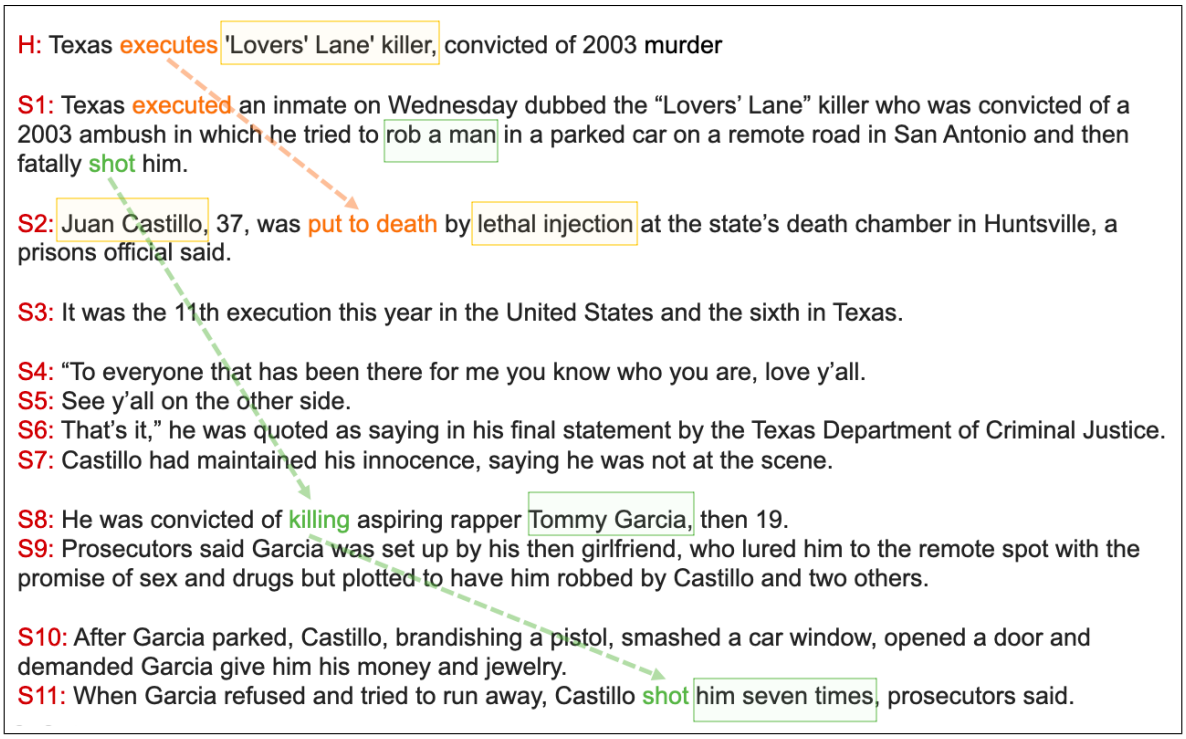

Figure 1.1: An example document to illustrate the role of event coreference resolution in downstream applications.

the extractive or abstractive summary that succinctly describes the news article. Besides, event coreference chains can also help with building document-level *events and time expressions relation graph*, incorporating all relations such as temporal, causal [Gao et al., 2019] and subevents [Araki et al., 2014]. Other notable applications [Lu and Ng, 2018], where event coreference has been shown useful, include knowledge base population [Ji and Grishman, 2011], topic detection and tracking [Allan et al., 2018] and contradiction detection [de Marneffe et al., 2008].

Despite many high-level NLP applications, event coreference is a relatively less studied problem compared to the counterpart entity coreference resolution task. The collective effort has largely been focused on improving event representations by learning better lexical, argument, or semantic features in a pairwise classifier. While the earliest work relied on feature-based classifiers, more recent works take advantage of large neural network-based pre-trained language models [Devlin et al., 2019] that have shown to improve the performance of previous classifiers. But, despite being

1. Himalayan glacial flood killed at least 18 people in India's northern Uttarakhand state.
2. It is too early to conclusively determine how the disaster began, experts said.
3. Past deadly or highly destructive glacial floods have occurred in Peru and Nepal.

Figure 1.2: Example sentences showing challenges of resolving event coreference links due to local lexical and contextual variations.

powerful, even the recent neural classifiers suffer from poor generalization and fail to perform well in practice. In particular, they are affected by at least three major challenges:

- Coreferential event mentions exist in a variety of surface forms such as verb, nominalized, or even pronominalized forms and often use different lemmas. Simultaneously, many non-coreferential event mentions share the synonymous lemma. For instance, in Figure 1.2, **glacial flood** in first sentence is coreferential with **disaster** in second sentence but not with **glacial floods** in third sentence. Thus, lexical semantics, while being important, often are insufficient to determine coreference relations.

- Event argument features are often omitted from the local context to inhibit unnecessary repetitions of information. Again, the second sentence in Figure 1.2 describes the information that compliments the first sentence and does not include the same set of arguments. Further, upstream tasks, such as argument extraction, are error-prone which inhibits the event coreference model from appropriately using argument-related features.

- Most importantly, coreferential event mentions are fewer in a document as most of the events are mentioned just once. Only certain events that have many related events and participants are referred back, when describing further information about those events and participants, to build a coherent story. The scarce distribution of coreferential event mentions also makes manually labeling and creating a large human-annotated corpus laborious and impractical.

1.1 Contributions of this Dissertation

The scarce distribution of event coreference links along with the absence of abundant annotated corpora have severely slowed the progress on event coreference resolution. In this dissertation, we primarily focus on overcoming both distributional and annotational sparsity by following discourse structure-guided approaches. Considering that human-annotated corpora for event coreference mainly include news and discussion forum articles, and a discourse forum article lacks coherent discourse structure as a regular document, we use news articles in our study and explore both heuristics and systematic approaches for modeling news discourse structure for event coreference resolution.

First, we study the association between main news event and event coreference chains [Choubey et al., 2018]. We found that main events tend to have many coreferential mentions that are often mentioned in the headline or lede paragraph and are then spread throughout the document. This is intuitive, given that the main event participates in many relations and has discourse-level relevance. Through repetitions, it helps to connect all peripheral events and produce a coherent news story. Next, to directly model frequent and extended repetitions of the main event in a document, we model the correlation between event coreference chains and document topical structures. Through heuristics, we identify document transition sentences that define topical boundaries and are expected to mention the main event. Then, we encourage coreferential event mentions in transition sentences through integer linear programming (ILP)-based constraints over pairwise coreference scores obtained from a classifier [Choubey and Huang, 2018]. In addition to the main event, our holistic approach incorporates additional correlations between semantically related events, news genre-specific distributional patterns, and subevents structures.

Our heuristics-based approach works by encouraging or discouraging coreference links between event pairs depending on the assumed topical roles or the position of their sentences. However, heuristics fail to explain how the extracted topical structures ground to any known discourse theories. Therefore, in a follow-up work, we use the Van Dijk's theory of News Discourse [Teun A, 1986, Van Dijk, 1988a,b] and created a new human-annotated dataset for news discourse profiling

[Choubey et al., 2020]. The discourse profiling task assigns pre-defined discourse roles to sentences in news articles based on their functions in describing the main news story. The choice of using the main event as a reference is natural given the empirical gain we observe from the event coreference-based main event identification system as well as heuristics over the main event-based event coreference system. We evaluate the computational feasibility of automatically profiling news discourse structure by designing a hierarchical neural network model on our annotated data. We also analyze correlations between event coreference links and human annotations for news discourse profiling and use the hierarchical model to identify discourse structure and incorporate that in an event coreference resolution system [Choubey et al., 2020]. Overall, we observe event coreference resolution system that systematically incorporates news discourse structure outperforms the heuristics-based system besides having well-grounded theoretical interpretation.

The heuristics or news discourse profiling-based models overcome distributional sparsity by using discourse structure-guided cues to make up for non-local and scattered event information. Finally, to overcome annotational sparsity, we use several rules over news discourse profiling to automatically acquire both coreferential and non-coreferential event pairs [Choubey and Huang, 2021] from unannotated news documents that are found to be empirically useful for evaluation datasets of news articles, either within or outside the training domain. However, we observe that acquired datasets are not consistently effective on discussion forum articles. This is predictable given discussion forum documents exhibit discourse structure different from news documents and require discussion forum genre-specific modeling to induce relevant structure for event coreference resolution.

In addition to the empirical analyses and evaluations, this dissertation contributes two human annotated-datasets, one for main news event identification and the other for news discourse profiling, and an automatically acquired dataset for event coreference resolution.

1.2 Dissertation Outline

The outline of this dissertation is summarized as follows.

- In chapter 2, we discuss prior methods for event coreference resolution, existing human-

annotated corpora and their corresponding definitions for event and event coreference resolution, and evaluation metrics.

- In chapter 3, we propose a new task of identifying the most dominant event of a news document, which governs and connects other foreground and background events in the document. We observe that the main event of a document usually has many coreferential event mentions that are scattered throughout the document for enabling a smooth transition of subtopics. Our empirical experiments, using gold event coreference relations, have shown that the main event of a document can be well identified by mining properties of event coreference chains. In addition, we found that the main event can be more accurately identified by further considering the number of sub-events as well as the realis status of an event.

- In chapter 4, we propose a novel approach for event coreference resolution that models correlations between event coreference chains and document topical structures through an ILP formulation. We explicitly model correlations between the main event chains of a document with topic transition sentences, inter-coreference chain correlations, event mention distributional characteristics and sub-event structure, and use them with scores obtained from a local coreference relation classifier for jointly resolving multiple event chains in a document. Our experiments on the benchmark evaluation datasets suggest that each of the structures contributes to improving event coreference resolution performance.

- In chapter 5, we propose a new task of news discourse profiling that helps to understand discourse structures of news articles and effectively contextualize the occurrence of news events. To enable computational modeling of news structures, we apply an existing theory of functional discourse structure for news articles that revolve around the main event and create a human-annotated corpus of 802 documents spanning over four domains and three media sources. Finally, we propose several document-level neural-network models to automatically construct news content structures.

- In chapter 6, we propose to systematically identify and incorporate content structures based

on the news discourse profiling model proposed in chapter 5, into an event coreference resolution system. We use content structures to dissociate sentences that favor coreferential mentions from the ones that favor singletons, and further build a specialized classifier that identifies singletons mentions in a document. We then apply constraints in inferences through an ILP formulation that empirically outperforms the typical pairwise classifier-based system as well as the heuristics-based system from chapter 4 on the benchmark evaluation datasets.

- In chapter 7, we propose to leverage lexical paraphrases and high precision rules informed by news discourse profiling structure to automatically collect coreferential and non-coreferential event pairs from unlabeled English news articles. We perform both manual validation and empirical evaluation on multiple evaluation datasets with different event domains and text genres to assess the quality of our acquired event pairs. We found that a model trained on our acquired event pairs performs comparably as the supervised model when applied to new data out of the training data domains. Further, augmenting human-annotated data with the acquired event pairs provides empirical performance gains on both in-domain and out-of-domain evaluation datasets.

- In chapter 8, we summarize our conclusion from this dissertation. Further, I propose some extensions to this dissertation that can benefit event coreference resolution as well as other downstream NLP tasks.

2. BACKGROUND

2.1 Literature Review

This section gives an overview of the widely used models and linguistics features for event coreference resolution, and common discourse analysis frameworks studied in NLP.

2.1.1 Supervised Models for Event Coreference Resolution

Given a document, an event coreference resolution model detects all event mentions and then links them into event clusters [Jurafsky and Martin, 2008, Lu and Ng, 2018]. While the exact definitions for event mentions and coreference links vary across datasets (described in details in section 2.2), models on all datasets follow one of the following approaches:

- **Mention-Pair Architecture** is the most commonly used model for both event and entity coreference resolution. Typically, a pairwise classifier is trained over features (discussed in section 2.1.2) for a mention pair to predict binary label indicating whether the given mention pair is coreferential or not. The resulting pairwise classifier is used to cluster event mentions. The commonly used strategies include agglomerative clustering that selects the antecedent closest in mention distance that is classified as coreferent or the antecedent with highest coreference likelihood [Chen et al., 2009, Chen and Ng, 2014, Peng et al., 2016]. Alternatively, given pairwise scores, graph-based approaches, such as spectral clustering algorithm, have also been used [Chen and Ji, 2009, Chen et al., 2009, Sangeetha and Arock, 2012].

- **Mention-Ranking Architecture** learns to collectively rank all the antecedents [Lu and Ng, 2017]. It first appends a dummy mention node to the candidate antecedents which act as the parent for singletons or first mention of an event cluster. Then, for generating coreference clusters, it links every event mention to the most likely candidate antecedent. During training, however, multiple antecedents can represent correct event coreference links, and to choose one of the several correct links, the most general approach is to pick the nearest

antecedent as the parent.

- **Event-Based Systems** learns to link an event mention to a previous event cluster representation instead of an event mention representation. The event representation is built incrementally by aggregating information from all the event mentions that form the given event coreference cluster. A common event-based system employs an easy first approach, where it first decides easy coreference links and uses that to accumulate information and obtain event representations, and resolve more difficult coreference links [Liu et al., 2014].

The above three approaches generally use pipeline architecture where event mentions and other features, such as event arguments, are first extracted. Then, either of the mention-pair, mention-ranking, or event-based classifier is trained on the extracted features and later used to determine whether the given event mention pair is coreferential. The pipeline approach accumulates errors from the upstream event and feature extraction systems and consequently joint learning over component tasks, such as event extraction, argument extraction, entity coreference, and event coreference, have also been explored [Chen and Ng, 2016, Lu et al., 2016, Lu and Ng, 2017].

In this dissertation, we follow the pipeline architecture where the event extraction step is kept fixed across all experiments. This allows us to directly evaluate the improvement obtained from incorporating news discourse structure. Further, we use either the mention-pair or joint inference over multiple event coreference clusters using ILP, depending on the modeling objectives.

2.1.2 Linguistic Features for Event Coreference Resolution

The existing literature on event coreference resolution primarily focuses on surface linguistic features [Chen et al., 2009, Bejan and Harabagiu, 2010, Lee et al., 2012, McConky et al., 2012, Sangeetha and Arock, 2012, Adrian Bejan and Harabagiu, 2014, Liu et al., 2014, Yang et al., 2015, Yu et al., 2016, Lu et al., 2016, Lu and Ng, 2017, Choubey and Huang, 2017a], such as:

- **Lexical features** include different string similarity measures (e.g. minimum edit distance) between event mention lemma or simply a binary feature indicating whether two event mention lemma are same or different.

- **Syntactic features** include parts of speech tags and dependency parse tree-based features consisting of modifiers or governor lemmas and their relation type with event mention lemma.

- **Semantic features** include WordNet or distributional embeddings based similarity scores, semantic frames, event mention types, etc.

- **Argument-based features** include binary features indicating the existence of arguments, similarity scores between surface lexical forms of arguments, or features defined following entity coreference resolution of arguments. Separately, features defined over the spatial or temporal location of the event mention are also used.

- **Discourse-based features** are mainly based on the position of event mentions. They include sentence distance (number of sentences between two event mentions), event distance (number of event mentions between two event mentions), and binary features indicating whether an event was mentioned in the headline or the first sentence.

Besides, most recent works [Kenyon-Dean et al., 2018, Barhom et al., 2019, Zuo et al., 2019, Pandian et al., 2020, Sahlani et al., 2020, Lee et al., 2017] simply use contextualized word embeddings to represent an event mention and use them in neural models to design a pairwise classifier.

2.1.3 Discourse Analysis Studies in NLP

Discourse structures have been analyzed from different perspectives and following different objectives. The Rhetorical Structure Theory (RST) [Mann and Thompson, 1988] and Penn Discourse Treebank Project (PDTB) style [Prasad et al., 2008] discourse parsing tasks identify discourse units that are logically connected with a predefined set of rhetorical relations. Text segmentation [Hearst, 1994] is another well studied discourse analysis task that aims to divide a text into a sequence of topically coherent segments.

Different from the above three general discourse theories, function of different discourse units, specific to a text-genre, have also been studied based on different genre-attributes. The majority

of those studies though relate to genres other than news articles. Liddy [1991], Kircz [1991] and Teufel et al. [1999] used rhetorical status and argumentation type to both define functional theories and create corpora for scientific articles. Mizuta et al. [2006], Wilbur et al. [2006], Waard et al. [2009] and Liakata et al. [2012] extensively studied functional structures in biological domain with multiple new annotation schemata.

Studies on functional structures of news articles have been mainly theoretical. Apart from Van Dijk's theory of news discourse [Teun A, 1986, Van Dijk, 1988b], Pan and Kosicki [1993] proposed framing-based approach along four structural dimensions: syntactic, script, thematic and rhetorical, of which syntactic structure is similar to the Dijk's theory. The computational studies on functional structure include Baiamonte et al. [2016] that coarsely separates narration from descriptive contents and Friedrich and Palmer [2014] that classify clauses based on their aspectual property.

2.2 Datasets and Definitions

We use the Knowledge Base Population (KBP) [Mitamura et al., 2015, 2017] and the Richer Event Description (RED) [O'Gorman et al., 2016] corpora for building and evaluating our event coreference resolution systems. The ECB+ corpus Cybulska and Vossen [2014] is another commonly used dataset for evaluating cross-document event coreference resolution performance. However, we determined that the ECB+ corpus is not appropriate for evaluating the models proposed in this dissertation, that explicitly focuses on using discourse-level topic structures for within document event coreference resolution. Particularly, the ECB+ corpus was created to facilitate both cross-document and within-document event coreference resolution research. Thus, the documents in the corpus were grouped based on several common topics and in each document, event mentions and coreference relations were annotated selectively in only sentences that are on a common topic. When the annotated sentences in each document are stitched together, they do not well reveal the original document structure, which makes the ECB+ corpus a bad choice for evaluating our work.

We briefly describe the definitions for both events and event coreference relations that were used to annotate KBP and RED corpora below.

2.2.1 KBP (2015-2017)

The KBP corpora follow richERE [Song et al., 2015] guidelines for event and event coreference annotations. It defines an event as an explicit occurrence of something that happens, and may or may not involve participants. It selectively annotates events of eight types: life, movement, business, conflict, contact, manufacture, personnel, transaction, and justice. For annotating event coreference chains, it uses a relaxed notion of event hopper which includes events that feel "coreferential" but do not necessarily meet the strict criteria of the same arguments. Specifically, two event mentions that are intuitively the same event, have the same attested scope, types, and subtypes but not necessarily have the same arguments, trigger, and realis belong to the same hopper.

2.2.2 RED

The RED guideline [O'Gorman et al., 2016] defines an event as any occurrence, action, process, or event state which deserves a place upon timeline. It defines event coreference as two or more event mentions that refer to the same event in space and time. RED documents are comprehensively annotated with event coreference relations with no restriction on event types or subtypes, thus, allowing us to evaluate coreference resolution performance on a broad range of events.

2.3 Coreference Evaluation Metrics

Evaluating event coreference resolution is non-trivial where we need to compare gold human-annotated event clusters (key K_i) with the predicted events (response R_i). Several evaluation metrics have been proposed to evaluate coreference resolution, but all of them exhibit some known issues. Secondly, neither of them are directly interpretable nor do they empirically agree with each other. Therefore, we follow the past research and use four coreference scoring measures, namely MUC [Vilain et al., 1995], B^3 [Bagga and Baldwin, 1998], $CEAF_e$ [Luo, 2005] and BLANC [Recasens and Hovy, 2011] and the unweighted average of their F1 scores (AVG_{F1}) for evaluating our systems. We report results based on version 1.8 of the official KBP 2017 scorer[1]. Note that

[1]The official KBP 2017 scorer is based on the standard reference scorer [Pradhan et al., 2014]. We obtain the evaluation script from `https://github.com/hunterhector/EvmEval`.

the official KBP 2017 event coreference resolution scorer considers a mention pair coreferent if they strictly match on the event type and subtype, which has been discussed recently to be too conservative [Mitamura et al., 2017]. Since improving event mention type detection is not our main goal, we therefore relax the constraints and do not consider event mention type match while evaluating event coreference resolution systems. This allows us to directly interpret the influences of different document structures in the event coreference resolution task by minimizing any bias from upstream tasks. The four evaluation metrics as well as their drawbacks are described below.

2.3.1 MUC

MUC is a link-based metric that directly compares the coreference links in responses to the links in keys. To calculate recall, it first creates partitions ($p(K_i)$) for each key (K_i) relative to the responses that overlap with the given key. Then, recall is calculated following:

$$Recall = \frac{\sum_{K_i \in K}(|K_i| - |p(K_i)|)}{\sum_{K_i \in K}(|K_i| - 1)}$$

To calculate precision, it simply switches the roles of key and response events. As noted in previous works [Bagga and Baldwin, 1998, Moosavi and Strube, 2016], MUC is the least discriminative of all coreference evaluation metrics and does not distinguish an incorrect link that merges two single mentions from another incorrect link that merges two larger clusters. Secondly, it favors a system that aggressively merges many unrelated events [Luo, 2005], and a trivial system that merges all mentions into one cluster achieves 100% recall with reasonably high precision.

2.3.2 B^3

B^3 is a mention-based metric. Between each pair of key (K_i) and response (R_j), it first calculates scores that are proportional to the square of the number of key mentions included in the response. Then, it calculates the recall by summing over scores corresponding to all keys normal-

ized by the number of total key mentions. The recall is calculated following:

$$Recall = \frac{\sum_{K_i \in K} \sum_{R_j \in R} \frac{|K_i \cap R_j|^2}{|K_i|}}{\sum_{K_i \in K} |K_i|}$$

Similar to MUC, precision is calculated by switching the role of keys and response events. Contrary to MUC, B^3 favors a system that does not link any mention [Luo, 2005], scoring it with 100% precision. Secondly, B^3 considers a resolved mention correct even when it is linked to a wrong event [Luo, 2005, Moosavi and Strube, 2016].

2.3.3 CEAF$_e$

CEAF$_e$ maps every key event cluster (K_i) with exactly one response event cluster ($g^*(K_i)$) using a similarity measure ϕ, and then calculate recall following:

$$Recall = \frac{\sum_{K_i \in K^*} \phi(K_i, g^*(K_i))}{\sum_{K_i \in K} \phi(K_i, K_i)}, \quad where\ \phi(K_i, R_j) = \frac{2 \times |K_i \cap R_j|}{|K_i| + |R_j|}$$

When calculating precision, it changes denominator in above equation to $\sum_{R_j} \phi(R_i, R_i)$. Since CEAF$_e$ first performs best one-to-one mapping and only considers truly aligned response clusters when calculating recall and precision, it ignores all correct responses that are not mapped to any key [Denis and Baldridge, 2009, Moosavi and Strube, 2016].

2.3.4 BLANC

Let C_K, C_R, N_K and N_R be the sets of coreference links in key, coreference links in response, non-coreference links in key and non-coreference links in response, respectively. Then, BLANC precision and recall are calculated as average of precision and recall of coreference (P_C, R_C) and non-coreference (P_N, R_N) links, where,

$$R_C = \frac{|C_K \cap C_R|}{|C_K|}, P_C = \frac{|C_K \cap C_R|}{|C_R|}, R_N = \frac{|N_K \cap N_R|}{|N_K|}, P_N = \frac{|N_K \cap N_R|}{|N_R|}$$

As noted by Moosavi and Strube [2016], BLANC suffers from the mention identification effect and may achieve high precision and recall even when many gold event mentions are unresolved.

3. IDENTIFYING THE MOST DOMINANT EVENT IN A NEWS ARTICLE BY MINING EVENT COREFERENCE RELATIONS[1]

According to the grounding principles [Grimes, 1972], a document consists of foreground events that form the skeleton of the story and move the story forward, and background events that add supportive information. Studies have shown that a foreground event tends to be the most important event in a sentence, which is usually the event that appears in the main clause, is active voiced, and has a high transitivity[2] [Decker, 1985]. But among multiple foreground events, which one is most central to the overall story? In this chapter, we propose a new task of detecting the most dominant event (main event) in a news article, which is an event assumed to govern and connect other foreground events and background events. In other words, removal of the main event can break the entirety of a document and decompose the document into sections describing disjoint sets of situations.

Our intuitive observation is that the main event of a document usually has a large number of coreferential event mentions and those coreferential mentions are spread throughout the document. In Figure 4.1, paragraphs 1-4 each describe a relatively independent subtopic, and the repeated mentions of the main event "demonstration" throughout the document enable a smooth flow of information. For the same reason, identifying the main event facilitates partitioning text into coherent segments. But note that, the main event may not be the most newsworthy event that serves as the trigger for writing an article, and thus may not appear in the title or in the first sentence of a news article. As illustrated in this example, the trigger event is "protesters leave capitol", while the main event is "demonstration", the event that effectively connects other foreground events and background events and makes the story an entirety.

To systematically verify these observations, we annotated main events in news articles taken

[1]Reprinted with permission from "Identifying the Most Dominant Event in a News Article by Mining Event Coreference Relations" by Prafulla Kumar Choubey, Kaushik Raju, and Ruihong Huang. Proceedings of NAACL-HLT 2018, pages 340–345, New Orleans, Louisiana, June 1 - 6, 2018. Copyright 2018 by Association for Computational Linguistics.

[2]High transitivity events have certain properties, are volitional, affirmative, realis, etc.

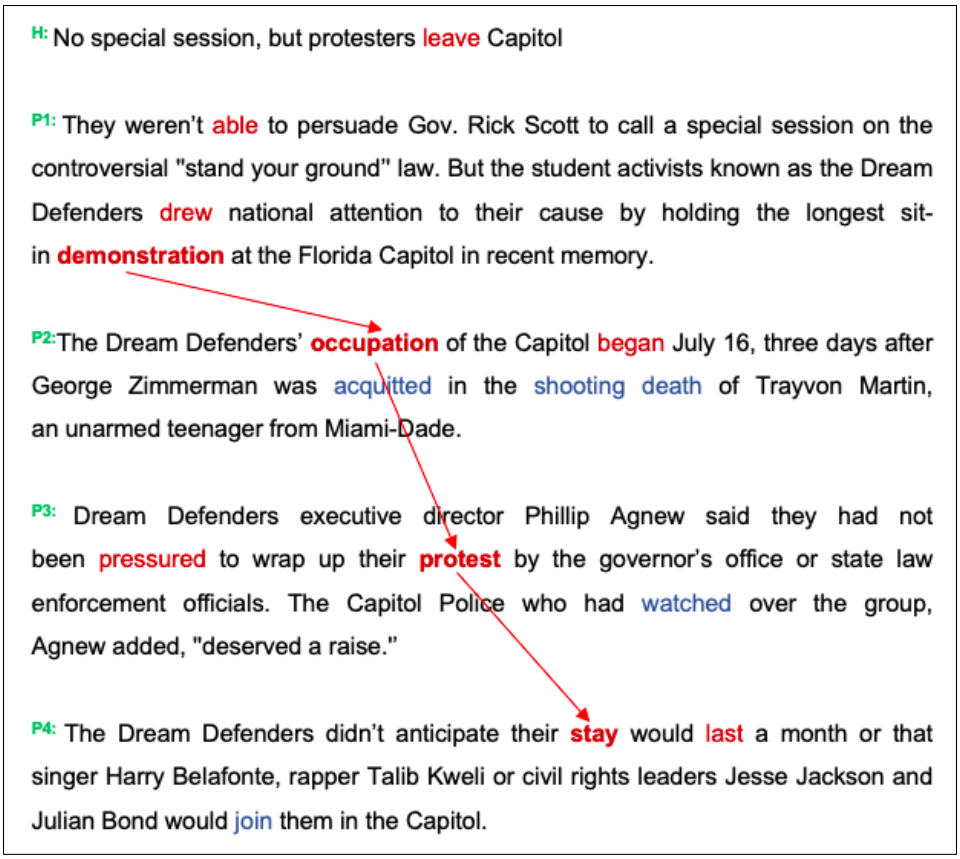

Figure 3.1: An example document to illustrate the main event of a document. Red colored words are foreground events, blue colored words are background events and mentions of the main event are in bold. Reprinted with permission from Choubey et al. [2018].

.

from two publicly available datasets, the richer event description (RED) [O'Gorman et al., 2016] and KBP 2015 [Mitamura et al., 2015] corpora. While whether each news article has only one main event is arguable, our two annotators agreed on the same main event in 97 out of 104 (93%) documents that we annotated. We then designed several rule-based methods to identify the main event by exploiting human-annotated event coreference relations. Experimental results showed that indeed in around 75% of the documents in both corpora, the main event either has the largest number of coreferential event mentions or has the largest stretch size (i.e., the number of sentences between the first mention and the last mention of the main event) in the discourse. In addition, we found that the main event can be more accurately identified by further considering the number of

sub-events as well as the realis status of an event, which indicates if an event is an actual specific event or a generic event, etc. The evaluation shows that the insightful rules outperform several strong baseline approaches, including several heuristic-based methods and random walk-based event ranking methods, as well as two regression classifiers that integrate these rules as features.

3.1 Main Event Annotations

We annotated main events for 30 news articles from the RED corpus and 74 news articles from the KBP 2015 corpus. The RED corpus contains 95 documents in total. However, 65 of those documents are news summaries, discussion forum posts, or web posts. The main event as defined should only be considered for natural coherent texts, therefore, we skipped 65 non-news documents and annotated only 30 news articles in the corpus. Similarly, the KBP 2015 corpus contains 158 documents, where 81 are news articles and the remaining are discussion forum posts. In 7 out of the 81 news articles, annotators unanimously found that the main event was not of one of the interested event types in KBP and was not tagged in the KBP annotations. Therefore, we exclude those 7 documents and annotate 74 news articles.

We asked two annotators to identify the most dominant event that connects other foreground and background events. Both the documents and the gold event mentions for each document inherited from the previous RED and KBP annotations were provided to annotators. The annotators were instructed to select only one event as the main event. For 26 documents from the RED corpus and 71 documents from the KBP corpus, both annotators identified the same main event. For the other 7 documents, where the two annotators disagreed on the main event, we kept the annotations from the first annotator.

3.2 Characteristics of Main Events

We analyzed the distributional properties of main events in the first 10 documents from the RED corpus. The findings are summarized below.

Frequent and Extended Repetitions: Similar to the example document in Figure 4.1, the main event is usually repeated throughout the document. This observation can also be accounted to the

way humans produce and comprehend language. Language is inherently sequential and a writer repeats the same event to remind the readers about the main event. Therefore, the frequent and extended repetitions of the main event facilitate to minimize the cognitive effort needed by the reader for understanding a text.

Early Presences: News articles mostly begin with a summary of important events and continue to elaborate them in subsequent paragraphs. To some extent, the objective of initial paragraphs is to direct readers' attention toward the main subject. Therefore, while the main event may not always appear in the title or in the first sentence of a news article, the main event often appears early in the beginning paragraphs.

Sub-events: Being the most dominant event in a document, the main event often has many sub-events that are present to elaborate and support the main event.

Event Realis Status: Main events are usually specific and have actually occurred. This event attribute has been defined as the contextual modality in RED corpus[3] and realis status in KBP corpus[4] and we observed that this attribute is "Actual" for the majority of main events.

3.3 Main Event Identification

Inspired by the identified characteristics of main events, we designed rule-based classifiers that rely on the following four ranking criteria.

Size Rank: calculated using the number of coreferential event mentions in a event coreference chain. The event having the largest number of coreferential mentions is ranked the highest.

Stretch Rank: based on the number of sentences between the first and the last mention of an event. The event with the largest stretch size is ranked the highest.

Position Rank: based on the sentence number in which an event was first mentioned. This measure is to capture the characteristic that main events tend to appear early in a document.

Enriched Size Rank: based on the sum of the number of coreferential mentions for an event and the number of its sub-events.

[3]defines 4 types of contextual modality, namely, actual, hypothetical, uncertain/ hedged and generic

[4]defines 3 realis status types, namely, actual, generic and other

Algorithm 1 Rule-based Systems

Input: Main Event Candidates, E_Z, E_T, E_P, E_E, E_R

Output: E_{center}

Coreference

$E_{center} := E_Z \cap E_T \cap E_P$

if $E_{center} == \phi$: $E_{center} := (E_Z \cup E_T) \cap E_P$

if $E_{center} == \phi$: $E_{center} := E_P$

Coreference + Subevent

$E_{center} := E_Z \cap E_T \cap E_P \cap E_E$

if $E_{center} == \phi$: $E_{center} := (E_Z \cup E_T) \cap E_P \cap E_E$

if $E_{center} == \phi$: $E_{center} := E_P \cap E_E$

if $E_{center} == \phi$: $E_{center} := E_P$

Coreference + Subevent + Realis

$E_{center} := E_Z \cap E_T \cap E_P \cap E_E \cap E_R$

if $E_{center} == \phi$: $E_{center} := (E_Z \cup E_T) \cap E_P \cap E_E \cap E_R$

if $E_{center} == \phi$: $E_{center} := E_P \cap E_E \cap E_R$

if $E_{center} == \phi$: $E_{center} := E_P \cap E_E$

if $E_{center} == \phi$: $E_{center} := E_P$

3.3.1 Rule Based Classifiers

First, we identify main event candidates by requiring their size rank in the top three positions. Note that more than three events may be selected if there are ties in any of the top three positions. Then, we identify the main event in the candidate set by requiring different combinations of the highest ranks, including the highest size rank E_Z, highest stretch rank E_T, highest position rank E_P and highest enriched size rank E_E. In addition, we identify an event set E_R which includes events whose contextual modality or realis status is "Actual", and use the set for constraining main event

identification. Specifically, we define three rule-based classifiers, described in Algorithm 1, which begin with strict rules followed by relaxed rules in subsequent passes. The systems **Coreference** uses size, stretch and position ranks, **Coreference + Subevent** considers enriched size rank as well, and **Coreference + Subevent + Realis** further combines realis status with each rank in favor of specific events.

3.3.2 Statistical Regression Classifiers

We trained a linear as well as a nonlinear regression classifier to integrate the same set of ranking rules as features for identifying main events, by using the standard ordinary least squares **linear regression** [Galton, 1886] model and the epsilon-support vector regression (**SVR**) [Vapnik, 1995] model with radial basis function kernel respectively. Input to both the linear and nonlinear regression classifiers consists of a 20 (19) dimensional vector, 4-dimensional categorical vector for each of the size, stretch, position, and enriched size ranks, and 4 (3) dimensional categorical vector for realis attribute for RED (KBP) corpus. The models were implemented using scikit-learn module [Pedregosa et al., 2011]. The SVR classifier uses **rbf** kernel with γ coefficient of 0.05 and all other parameters are left to be the default values.

3.4 Evaluation

3.4.1 Baseline Systems

Three Heuristics Based Classifiers: The three systems *Main event: Headline*, *First event: First sentence* and *Main event: First sentence* select the main event (syntactic root) in the headline, the first event in the first sentence and the main event (syntactic root) in the first sentence as the center event respectively.

Random Walk Based Ranking Systems: implemented the random walk-based vertex ranking algorithm [Mihalcea and Tarau, 2004] on graphs generated using human-annotated event relations. The motivation is to decide the importance of an event mention within an event graph of a document based on the importance of its related event mentions. We first build an event graph for a document by using undirected edges for coreference relations and directed edges for other relations

including set/ member, sub-event, temporal and causal relations. This is mainly meant to retain the symmetrical property of coreference relations. Moreover, since coreference links can easily create cycles in the graph, we utilize its transitivity property and link all the coreferent event mentions to its first instance in the document only. Then, we rank event mentions by using the vertex scoring algorithm proposed in Brin and Page [1998].

$$S(V_i) = (1-d) + d \sum_{j=IN(V_i)} \frac{1}{|OUT(V_j)|} S(V_j) \tag{3.1}$$

where $IN(V_i)$ and $OUT(V_j)$ represent the set of event mentions that are predecessors and successors to V_i respectively. Also, d is a damping vector that is kept 0.85 in our experiments. We initially assign random values to all the event mentions in an event graph and then update scores for all event nodes using equation 3.1 after each iteration. Computation stops when the sum of differences between the scores computed for all event mentions at two successive iterations reduces below 0.01.

The system *Random walk: All Relations* uses coreference, sub-event, set/ member, temporal and causal relations to build the graph while the system *Random walk: Coref+SE* only considers event coreference and sub-event relations. We evaluate both systems on documents from the RED corpus only as it extensively annotates event relations which yields a connected graph for each document. However, the graphs generated for documents in the KBP corpus often contain many disconnected components and thus are not suitable for these systems.

3.4.2 Results

We evaluated all the systems using the rest 20 documents from the RED corpus and all the 74 documents from the KBP 2015 corpus. The two regression classifiers were evaluated using 5-fold cross-validation on each corpus. We expect a system to identify only one main event for each document. If a system predicts more than one main event, we will penalize the system on precision strictly and treat each wrongly predicted event as a false hit. Table 7.3 shows the comparison results.

Model	Recall	Precision	F1
Richer Event Description (RED)			
Main event: Headline	45.00	45.00	45.00
First event: First sentence	10.00	10.00	10.00
Main event: First sentence	40.00	40.00	40.00
Random walk: All Relations	40.00	40.00	40.00
Random walk: Coref+SM	45.00	45.00	45.00
Coreference	75.00	55.55	63.82
Coreference + Subevent	75.00	62.50	68.18
Coreference + Subevent + Realis	80.00	66.67	72.73
Linear Regression	63.33	63.33	63.33
SVR	66.67	66.67	66.67
Coreference: Predicted	45.00	45.00	45.00
KBP 2015			
Main event: Headline	45.94	45.94	45.94
First event: First sentence	39.19	39.19	39.19
Main event: First sentence	39.19	39.19	39.19
Coreference	77.03	54.81	64.04
Coreference + Subevent	77.03	60.00	67.46
Coreference + Subevent + Realis	78.37	66.67	72.05
Linear Regression	66.21	61.25	63.63
SVR	67.56	62.50	64.93
Coreference: Predicted	48.65	45.56	47.05

Table 3.1: Performance comparison of different main event identification systems on the RED and KBP 2015 datasets. Reprinted with permission from Choubey et al. [2018].

The heuristic-based systems obtained a low recall on both corpora, which indicates that simple heuristics miss a large proportion of cases. Both random walk-based systems suffered from a low recall of 40-45% as well when applied to the RED corpus, due to the fact that graph-based ranking models do not effectively capture discourse layout features of co-referential event mentions.

In contrast, the rule-based system **Coreference** achieved the recall above 75% on both corpora when using annotated event coreference relations. The system **Coreference + Subevent + Realis** further improves the precision of main event identification by over 11% on both corpora after considering subevents and the realis status in the rules, which facilitate accurate identification of the main event among multiple foreground events. The high recall and precision indicate that the

insightful rules exploiting properties of event chains can capture the overall texture in the discourse. Then compared with rule-based systems, the two statistical classifiers that integrate the same set of rules as features do not further improve the main event identification performance.

3.5 Conclusions

We have presented a new task of identifying the main event for a document. Based on our annotations, we discussed the role of main events in enabling a coherent discourse and the distributional characteristics of main events. We especially emphasized on the importance of event coreference in identifying main events. Inspired by these observations, we designed a rule-based classifier that achieved high recall and precision in identifying main events.

4. IMPROVING EVENT COREFERENCE RESOLUTION BY MODELING CORRELATIONS BETWEEN EVENT COREFERENCE CHAINS AND DOCUMENT TOPIC STRUCTURES[1]

Event coreference resolution presents unique challenges. Compared to entities, coreferential event mentions are fewer in a document and much more sparsely scattered across sentences. Figure 4.1 shows a typical news article. Here, the main entity, "President Chen", appears frequently in every sentence, while the main event "hearing" and its accompanying event "detention" are mentioned much less frequently. If we look more closely, referring back to the same entity serves a different purpose than referring to the same event. The protagonist entity of a story is involved in many events and relations; thus, the entity is referred back each time such an event or relation is described. In this example, the entity was mentioned when describing various events he participated or was involved in, including "detention", "said", "pointed out", "remitted", "have a chance", "release", "cheating", "asked" and "returned", as well as when describing several relations involving him, including "former president", "his family" and "his wife". In contrast, most events only appear once in a text, and there is less motivation to repeat them: a story is mainly formed by a series of related but different events. Essentially, (1) the same event is referred back only when a new aspect or further information of the event has to be described, and (2) repetitions of the same events are mainly used for content organization purposes and, consequently, correlate well with topic structures.

Table 4.1 further shows the comparisons of positional patterns between event coreference and entity coreference chains, based on two benchmark datasets, ERE [Song et al., 2015] and ACE05 [Walker et al., 2006], where we paired each event (entity) mention with its nearest antecedent event (entity) mention and calculated the percentage of (event vs. entity) coreferent mention pairs

[1]Reprinted with permission from "Improving Event Coreference Resolution by Modeling Correlations between Event Coreference Chains and Document Topic Structures" by Prafulla Kumar Choubey, and Ruihong Huang. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers), pages 485–495, Melbourne, Australia, July 15 - 20, 2018. Copyright 2018 by Association for Computational Linguistics.

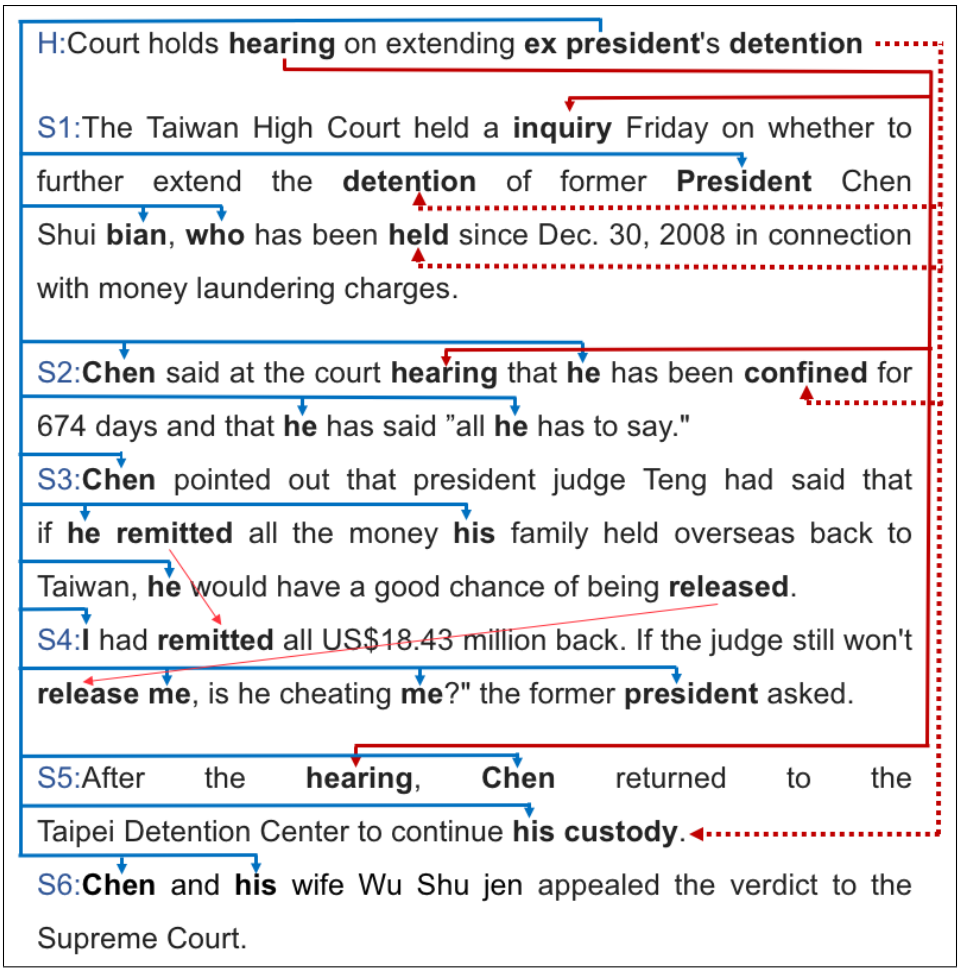

Figure 4.1: An example document to illustrate the characteristics of event (red) and entity (blue) coreference chains. Reprinted with perimission from Choubey and Huang [2018].

based on the number of sentences between two mentions. Indeed, for entity coreference resolution, centering and nearness are striking properties [Grosz et al., 1995], and the nearest antecedent of an entity mention is mostly in the same sentence or the immediately preceding sentence (70%). This is especially true for nominals and pronouns, two common types of entity mentions, where the nearest preceding mention that is also compatible in basic properties (e.g., gender, person, and number) is likely to co-refer with the current mention. In contrast, coreferential event mentions are rarely from the same sentence (10%) and are often sentences apart. The sparse distribution of coreferent event mentions also applies to the three KBP corpora used in this work.

To address severe sparsity of event coreference relations in a document, in this chapter, we

Dataset	Type	0	1	2	3	4	> 4
richERE	event	11	**34**	**20**	**9**	**7**	**19**
	entity	**34**	33	14	6	3	10
ACE-05	event	5	**33**	**19**	**10**	**9**	**24**
	entity	**37**	28	12	7	4	13
KBP 2015	event	15	**34**	**12**	**9**	**6**	**24**
KBP 2016	event	8	**43**	**15**	**7**	**6**	**21**
KBP 2017	event	12	**49**	**13**	**7**	**4**	**15**

Table 4.1: Percentages of adjacent (event vs. entity) mention pairs based on the number of sentences between two mentions in richERE, ACE-05 and KBP corpora. Reprinted with perimission from Choubey and Huang [2018].

propose a holistic approach to identify coreference relations between event mentions by considering their correlations with document topic structures. Our key observation is that event mentions make the backbone of a document and coreferent mentions of the same event play a key role in achieving a coherent content structure. For example, in figure 4.1, the events "hearing" and "detention" were mentioned in the headline (H), in the first sentence (S1) as a story overview, in the second sentence (S2) for transitioning to the body section of the story describing what happened during the "hearing", and then in the fifth sentence (S5) for transitioning to the ending section of the story describing what happened after the "hearing". By attaching individual event mentions to a coherent story and its topic structures, our approach recognizes event coreference relations that are otherwise not easily seen due to a mismatch of two event mentions' local contexts or long distances between event mentions.

We model several aspects of correlations between event coreference chains and document level topic structures, in an ILP joint inference framework. Experimental results on the benchmark event coreference resolution dataset KBP-2016 [Ellis et al., 2016] and KBP 2017 [Getman et al., 2017] show that the ILP system greatly improves event coreference resolution performance by modeling different aspects of correlations between event coreferences and document topic structures, which outperforms the previous best system on the same dataset consistently across several event coreference evaluation metrics.

4.1 Correlations between Event Coreference Chains and Document Topic Structures

We model four aspects of correlations.

4.1.1 Correlations between Main Event Chains and Topic Transition Sentences

The main events of a document, e.g., "hearing" and "detention" in this example 4.1, usually have multiple coreferent event mentions that span over a large portion of the document and align well with the document topic layout structure [Choubey et al., 2018]. While fine-grained topic segmentation is a difficult task in its own right, we find that topic transition sentences often overlap in content (for reminding purposes) and can be identified by calculating sentence similarities. For example, sentences S1, S2, and S5 in Figure 4.1 all mentioned the two main events and the main entity "President Chen". We, therefore, encourage coreference links between event mentions that appear in topic transition sentences by designing constraints in ILP and modifying the objective function. In addition, to avoid fragmented partial event chains and recover complete chains for the main events, we also encourage associating more coreferent event mentions to a chain that has a large stretch (the number of sentences between the first and the last event mention based on their textual positions).

4.1.2 Correlations across Semantically Associated Event Chains

Semantically associated events often co-occur in the same sentence. For example, mentions of the two main events "hearing" and "detention" co-occur across the document in sentences H, S1, S2, and S5. The correlation across event chains is not specific to global main events, for example, the local events "remitted" and "release" have their mentions co-occur in sentences S3 and S4 as well. In ILP, we leverage this observation and encourage creating coreference links between event mentions in sentences that contain other already known coreferent event mentions.

4.1.3 Genre-specific Distributional Patterns

We model document level distributional patterns of coreferent event mentions that may be specific to news genre in ILP. Specifically, news article often begins with a summary of the overall

story and then introduces the main events and their closely associated events. In subsequent paragraphs, detailed information of events may be introduced to provide supportive evidence to the main story. Thereby, a majority of event coreference chains tend to be initiated in the early sections of the document. Event mentions in the later paragraphs may exist as coreferent mentions of an established coreference chain or as singleton event mentions which, however, are less likely to initiate a new coreference chain. Inspired by this observation, we simply modify the objective function of ILP to encourage more event coreference links in early sections of a document.

4.1.4 Subevents

Subevents exist mainly to provide details and evidence for the parent event, therefore, the relation between subevents and their parent event presents another aspect of correlations between event relations and hierarchical document topic structures. Subevents may share the same lexical form as the parent event and cause spurious event coreference links [Araki et al., 2014]. We observe that subevents referring to specific actions were seldomly referred back in a document and are often singleton events. Following the approach proposed by Badgett and Huang [2016], we identify such specific action events and improve event coreference resolution by specifying constraints in ILP to discourage coreference links between a specific action event and other event mentions.

4.2 Modeling Event Coreference Chain - Topic Structure Correlations Using Integer Linear Programming

We model discourse-level event-topic correlation structures by formulating the event coreference resolution task as an ILP problem. Our baseline ILP system is defined over pairwise scores between event mentions obtained from a pairwise neural network-based coreference resolution classifier.

4.2.1 The Local Pairwise Coreference Resolution Classifier

Our local pairwise coreference classifier uses a neural network model based on features defined for an event mention pair. It includes a common layer with 347 neurons shared between two event mentions to generate embeddings corresponding to word lemmas (300) and parts-of-speech (POS)

tags (47). The common layer aims to enrich event word embeddings with the POS tags using the shared weight parameters. It also includes a second layer with 380 neurons to embed suffix[2] and prefix[3] of event words, distances (euclidean, absolute, and cosine) between word embeddings of two event lemmas and common arguments between two event mentions. The output from the second layer is concatenated and fed into the third neural layer with 10 neurons. The output embedding from the third layer is finally fed into an output layer with 1 neuron that generates a score indicating the confidence of assigning the given event pair to the same coreference cluster. All three layers and the output layer use the sigmoid activation function.

4.2.2 The Basic ILP for Event Coreference Resolution

Let λ represents the set of all event mentions in a document, Λ denotes the set of all event mention pairs i.e. $\Lambda = \{< i, j > \mid < i, j > \in \lambda \times \lambda \; and \; i < j\}$ and $p_{ij} = p_{cls}(coref|i, j)$ represents the cost of assigning event mentions i and j to the same coreferent cluster, we can formulate the baseline objective function that minimizes equation 6.3. Further we add constraints (equation 4.2) over each triplets of mentions to enforce transitivity [Denis and Baldridge, 2007, Finkel and Manning, 2008]. This guarantees legal clustering by ensuring that $x_{ij} = x_{jk} = 1$ implies x_{ik} = 1.

$$\Theta_B = \sum_{i,j \in \Lambda} -log(p_{ij})x_{ij} - log(1 - p_{ij})(\neg x_{ij}) \qquad s.t.\; x_{ij} \in \{0, 1\} \tag{4.1}$$

$$\neg x_{ij} + \neg x_{jk} \geq \neg x_{ik} \tag{4.2}$$

We then add constituent objective functions and constraints to the baseline ILP formulation to induce correlations between coreference chains and topical structures (Θ_T), discourage fragmented chains (Θ_G), encourage semantic associations among chains (Θ_C), model genre-specific distributional patterns (Θ_D) and discourage subevents from having coreferent mentions (Θ_S).

[2] te, tor, or, ing, cy, id, ed, en, er, ee, pt, de, on, ion, tion, ation, ction, de, ve, ive, ce, se, ty, al, ar, ge, nd, ize, ze, it, lt
[3] re, in, at, tr, op

4.2.3 Modeling the Correlation between Main Event Chains and Topic Transition Sentences

As shown in the example Figure 4.1, main events are likely to have mentions appear in topic transition sentences. Therefore, We add the following objective function (equation 4.3) to the basic objective function and add the new constraint 4.4 in order to encourage coreferent event mentions to occur in topic transition sentences.

$$\begin{aligned} \Theta_T = \sum_{m,n \in \Omega} & -log(s_{mn}) w_{mn} - log(1 - s_{mn})(\neg w_{mn}) \\ & s.t.\ w_{mn} \in \{0, 1\} \\ & (n - m) \geq |S|/\theta_s \end{aligned} \tag{4.3}$$

$$\sum_{i' \in \xi_m, j' \in \xi_n} x_{i'j'} \geq w_{mn} \tag{4.4}$$

Specifically, let ω represents the set of sentences in a document and Ω denotes the set of sentence pairs i.e. $\Omega = \{< m, n > | < m, n > \in \omega \times \omega \text{ and } m < n\}$. Then, let $s_{ij} = p_{sim}(simscore|m, n)$, which represents the similarity score between sentences m and n and |S| equals to the number of sentences in a given document. Here, the indicator variable w_{mn} indicates if the two sentences m and n are topic transition sentences. Essentially, when two sentences have a high similarity score (> 0.5) and are not near (with $|S|/\theta_s$or more sentences apart, in our experiments we set θ_s to 5), this objective function Θ_T tries to set the corresponding indicator variable w_{mn} to 1. Then, we add constraint 4.4 to encourage coreferent event mentions to occur in topic transition sentences. Note that ξ_m refers to all the event mentions in sentence m, and x_{ij} is the indicator variable which is set to 1 if event mentions defined by index i and j are coreferent. Thus, the above constraint ensures that two topic transition sentences contain at least one coreferent event pair.

4.2.3.1 Identifying Topic Transition Sentences Using Sentence Similarities:

First, we use the unsupervised method based on the weighted word embedding average proposed by Arora et al. [2016] to obtain sentence embeddings. We first compute the weighted average of words' embeddings in a sentence, where the weight of a word w is given by $a/(a+p(w))$. Here,

p(w) represents the estimated word frequency obtained from English Wikipedia and a is a small constant (1e-5). We then compute the first principal component of averaged word embeddings corresponding to sentences in a document and remove the projection on the first principal component from each averaged word embeddings for each sentence.

Then using the resulted averaged word embedding as the sentence embedding, we compute the similarity between two sentences as cosine similarity between their embeddings. We particularly choose this simple unsupervised model to reduce the reliance on any additional corpus for training a new model for calculating sentence similarities. This model was found to perform comparably to supervised RNN-LSTM based models for the semantic textual similarity task.

4.2.3.2 Constraints for Avoiding Fragmented Partial Event Chains:

The above equations (4.3-4.4) consider a pair of sentences and encourage two coreferent event mentions to appear in a pair of topic transition sentences. But the local nature of these constraints can lead to fragmented main event chains. Therefore, we further model the distributional characteristics of global event chains and encourage the main event chains to have a large number of coreferential mentions and a long stretch (the number of sentences that are present in between the first and last event mention of a chain), to avoid creating partial chains. Specifically, we add the following objective function (equation 4.5) and the new constraints (equation 4.6 and 4.7):

$$\Theta_G = -\sum_{i,j \in \mu} \gamma_{ij} \tag{4.5}$$

$$\begin{aligned} \sigma_{ij} = \sum_{k<i} \neg x_{ki} \wedge \sum_{j<l} \neg x_{jl} \wedge x_{ij} \\ \sigma_{ij} \in \{0,1\} \end{aligned} \tag{4.6}$$

$$\begin{aligned} \Gamma_i = \sum_{k,i \in \Lambda} x_{ki} + \sum_{i,j \in \Lambda} x_{ij} \\ M(1-y_{ij}) \geq (\varphi[j] - \varphi[i]).\sigma_{ij} - \lceil 0.75\ (|S|) \rceil \\ \gamma_{ij} - \Gamma_i - \Gamma_j \geq M.y_{ij} \\ \Gamma_i, \Gamma_j, \gamma_{ij} \in Z;\ \Gamma_i, \Gamma_j, \gamma_{ij} \geq 0;\ y_{ij} \in \{0,1\} \end{aligned} \tag{4.7}$$

First, we define an indicator variable σ_{ij} by equation 4.6 [4], corresponding to each event mention pair, that takes value 1 if (1) the event mentions at index i and j are coreferent; (2) the event mention at index i doesn't corefer to any of the mentions preceding it; and (3) mention at index j doesn't corefer to any event mention following it. Essentially, setting σ_{ij} to 1 defines an event chain that starts from the event mention i and ends at the event mention j.

Then with equation 4.7, variable σ_{ij} is used to identify main event chains as those chains which are extended to at least 75% of the document. When a chain is identified as a global chain, we encourage it to have more coreferential mentions. Here, Γ_i (Γ_j) equals the sum of indicator variables x corresponding to event pairs that include the event mention at index i (j) i.e. the number of mentions that are coreferent to i (j), $\varphi[i]$ ($\varphi[j]$) represents the sentence number of event mention i (j), M is a large positive number and y_{ij} represents a slack variable that takes the value 0 if the event chain represented by σ_{ij} is a global chain. Given $\sigma_{i,j}$ is identified as a global chain, variable γ_{ij} equals the sum of variables Γ_i and Γ_j and is used in the objective function Θ_G (equation 4.5) to encourage more coreferential mentions.

4.2.4 Cross-chain Inferences

As illustrated through Figure 4.1, semantically related events tend to have their mentions co-occur within the same sentence. So, we define the objective function (equation 4.8) and constraints (4.9) to favor a sentence with a mention from one event chain to also contain a mention from another event chain, if the two event chains are known to have event mentions co-occur in several other sentences.

$$\Theta_C = -\sum_{m,n \in \Omega} \Phi_{mn} \tag{4.8}$$

[4] Equation 4.6 can be implemented as

$$n_p + n_s \geq \sum_{k<i} x_{ki} + \sum_{j<l} x_{jl} - x_{ij} + (n_p + n_s + 1).\sigma_{ij}$$
$$\sum_{k<i} x_{ki} + \sum_{j<l} x_{jl} - x_{ij} + (n_p + n_s + 1).\sigma_{ij} \geq 0$$

where n_p, n_s represent the number of event mentions preceding event mention i and the number of event mentions following event mention j respectively.

$$\Phi_{mn} = \sum_{i \in \xi_m, j \in \xi_n} x_{ij}$$
$$|\xi_m| > 1;\ |\xi_n| > 1;\ \Phi_{mn} \in Z;\ \Phi_{mn} \geq 0 \tag{4.9}$$

To do so, we first define a variable ϕ_{mn} that equals the number of coreferent event pairs in a sentence pair, with each sentence having more than one event mention. We then define Θ_C to minimize the negative sum of ϕ_{mn}. Following the previous notations, ξ_m in the above equation represents the event mentions in sentence m.

4.2.5 Modeling Segment-wise Distributional Patterns

The position of an event mention in a document has a direct influence on event coreference chains. Event mentions that occur in the first few paragraphs are more likely to initiate an event chain. On the other hand, event mentions in later parts of a document may be coreferential with a previously seen event mention but are extremely unlikely to begin a new coreference chain. This distributional association is even stronger in the journalistic style of writing. We model this through a simple objective function and constraints (equation 7.1).

$$\Theta_D = -\sum_{i \in \xi_m, j \in \xi_n} x_{ij} + \sum_{k \in \xi_p, l \in \xi_q} x_{kl}$$
$$s.t.\ m, n < \lfloor \alpha |S| \rfloor;\ p, q > \lceil \beta |S| \rceil \tag{4.10}$$
$$\alpha \in [0, 1];\ \beta \in [0, 1]$$

Specifically, for the event pairs that belong to the first α (or the last β) sentences in a document, we add the negative (positive) sum of their indicator variables (x) in objective function Θ_D.

The equation 7.1 is meant to inhibit coreference links between event mentions that exist within the latter half of document. They do not influence the links within event chains that start early and extend till the later segments of the document. It is also important to understand that position-based features used in entity coreference resolution [Haghighi and Klein, 2007] are usually defined for an entity pair. However, we model the distributional patterns of an event chain in a document.

4.2.6 Restraining Subevents from Being Included in Coreference Chains

Subevents are known to be a major source of false coreference links due to their high surface similarity with their parent events. Therefore, we discourage subevents from being included in coreference chains in our model and modify the global optimization goal by adding a new objective function (equation 4.11).

$$\Theta_S = \sum_{s \in S} \Gamma_s \tag{4.11}$$

where S represents the set of subevents in a document. We define the objective function Θ_S as the sum of Γ_s, where Γ_s equals the number of mentions that are coreferent to s. Then our goal is to minimize Θ_S and restrict the subevents from being included in coreference chains.

We identify probable subevents by using surface syntactic cues corresponding to identifying a sequence of events in a sentence [Badgett and Huang, 2016]. In particular, a sequence of two or more verb event mentions in a conjunction structure are extracted as subevents.

4.2.7 The full ILP Model and the Parameters

The equations 4.3-4.11 model correlations between non-local structures within or across event chains and document topical structures. We perform ILP inference for coreference resolution by optimizing a global objective function(Θ), defined in equation 4.12, that incorporates prior knowledge by means of hard or soft constraints.

$$\Theta = \kappa_B \Theta_B + \kappa_T \Theta_T + \kappa_G \Theta_G + \kappa_C \Theta_C + \kappa_D \Theta_D + \kappa_S \Theta_S \tag{4.12}$$

Here, all the κ parameters are floating point constants. For the sake of simplicity, we set κ_B and κ_T to 1.0 and $\kappa_G = \kappa_C$. Then we estimate the parameters $\kappa_G(\kappa_C)$ and κ_D through 2-d grid search in range [0, 5.0] at the interval of 0.5 on a held out training data. We found that the best performance was obtained for $\kappa_C = \kappa_G = 0.5$ and $\kappa_D = 2.5$. Since, Θ_S aims to inhibit subevents from being included in coreference chains, we set a high value for κ_S and found that, indeed, the performance remained same for all the values of κ_S in range [5.0,15.0]. In our final model, we keep κ_S = 10.0. Also, we found that the performance is roughly invariant to the parameters κ_G and κ_C if they are set to values between 0.5 and 2.5.

In our experiments, we process each document to define a distinct ILP problem which is solved using the PuLP library [Mitchell et al., 2011].

4.3 Evaluation

4.3.1 Experimental Setup

We trained our ILP system on the KBP 2015 English dataset and evaluated the system on KBP 2016 and KBP 2017 English datasets. All the KBP corpora include documents from both discussion forum and news articles. Each discussion forum document consists of a series of posts in an online discussion thread, which lacks coherent discourse structures as a regular document. Since the goal of this study is to leverage discourse-level topic structure in a document for improving event coreference resolution performance, we only evaluate the ILP system using regular documents (news articles) in the KBP corpora. Specifically, we train our event extraction system and local coreference resolution classifier on 310 documents from the KBP 2015 corpus that consists of both discussion forum documents and news articles, tune the hyper-parameters corresponding to ILP using 50 news articles[5] from the KBP 2015 corpus and evaluate our system on news articles from the official KBP 2016 and 2017 evaluation corpora[6] respectively. For direct comparisons, the results reported for the baselines, including the previous state-of-the-art model, were based on news articles in the test datasets as well.

4.3.2 Event Mention Identification

We use an ensemble of multi-layer feed forward neural network classifiers to identify event mentions [Choubey and Huang, 2017b]. All basic classifiers are trained on features derived from the local context of words. The features include the embedding of word lemma, absolute difference between embeddings of word and its lemma, prefix and suffix of word and pos-tag and dependency relation of its context words, modifiers and governor. We trained 10 classifiers on same feature sets with slightly different neural network architectures and different training parameters including

[5]KBP 2015 dataset consists of 181 and 179 documents from discussion forum and news articles respectively. We randomly picked 50 documents from news articles for tuning ILP hyper-parameters and the remaining 310 documents for training classifiers.

[6]There are 85 and 83 news articles in KBP 2016 and 2017 corpora respectively.

	Lu and Ng [2017]		Ours	
Corpus	Untyped	Typed	Untyped	Typed
KBP 2016	60.13	49.00	60.03	45.45
KBP 2017	-	-	62.89	49.34

Table 4.2: F1 scores for event mention extraction system [Choubey and Huang, 2017b] on the KBP 2016 and 2017 corpus. Reprinted with perimission from Choubey and Huang [2018].

dropout rate, optimizer, learning rate, epochs and network initialization. All the classifiers use relu, tanh and softmax activations in the input, hidden and output layers respectively. We use GloVe vectors [Pennington et al., 2014] for word embeddings and one-hot vectors for pos-tag and dependency relations in each individual model. Pos-tagging, dependency parsing, named entity recognition and entity coreference resolution are performed using Stanford CoreNLP [Manning et al., 2014]

Table 4.2 shows the event mention identification results. We report the F1 score for event mention identification based on the KBP scorer, which considers a mention correct if its span, type, and subtype are the same as the gold mention and assigns a partial score if span partially overlaps with the gold mention. We also report the event mention identification F1 score that only considers mention spans and ignores mention types. We can see that compared to the recent system by [Lu and Ng, 2017] which conducts joint inferences of both event mention detection and event coreference resolution, detecting types for event mentions is a major bottleneck to our event extraction system.

4.3.3 Baseline Systems

We compare our document-structure guided event coreference resolution model with three baselines.

Local classifier performs greedy merging of event mentions using scores predicted by the local pairwise coreference resolution classifier. An event mention is merged to its best matching antecedent event mention if the predicted score between the two event mentions is highest and

greater than 0.5.

Clustering performs spectral graph clustering [Pedregosa et al., 2011], which represents commonly used clustering algorithms for event coreference resolution. We used the relation between the size of event mentions and the number of coreference clusters in training data for pre-specifying the number of clusters. Its low performance is partially accounted to the difficulty of determining the number of coreference clusters.

Joint learning uses a structured conditional random field model that operates at the document level to jointly model event mention extraction, event coreference resolution and an auxiliary task of event anaphoricity determination [Lu and Ng, 2017].

4.3.4 Our Systems

We gradually augment the ILP baseline with additional objective functions and constraints described in sub-sections 4.2.3, 4.2.4, 4.2.5 and 4.2.6. In all the systems below, we combine objective functions with their corresponding coefficients (as described in sub-section 4.2.7).

The Basic ILP System formulates event coreference resolution as an ILP optimization task. It uses scores produced by the local pairwise classifier as weights on variables that represent ILP assignments for event coreference relations. (Equations 6.3, 4.2).

+Topic structure incorporates the topical structure and the characteristics of main event chains in baseline ILP system (Equations 6.3-4.5).

+Cross-chain adds constraints and objective function defined for cross-chain inference to the Topical structure system (Equations 6.3-4.8).

+Distribution further adds distributional patterns to the Cross-chain system (Equations 6.3-7.1).

+Subevent (Full) optimizes the objective function defined in equation 4.12 by considering all the constraints defined in 6.3-4.11, including constraints for modeling subevent structures.

4.3.5 Results and Analysis

Table 4.3 shows performance comparisons of our ILP systems with other event coreference resolution approaches including the joint learning approach [Lu and Ng, 2017] which is one of the

Model	B^3	$CEAF_e$	MUC	BLANC	AVG
	KBP 2016				
Local classifier	51.47	47.96	26.29	30.82	39.13
Clustering	46.97	41.95	18.79	26.88	33.65
Basic ILP	51.44	47.77	26.65	30.95	39.19
+Topic structure	51.44	47.94	28.86	31.87	40.03
+Cross-chain	51.09	47.53	31.27	33.07	40.74
+Distribution	51.06	48.28	33.53	33.63	41.62
+Subevent	51.67	49.1	34.08	34.08	42.23
Joint learning	50.16	48.59	32.41	32.72	40.97
	KBP 2017				
Local classifier	50.24	48.47	30.81	29.94	39.87
Clustering	46.51	40.21	23.10	25.08	33.72
Basic ILP	50.4	48.49	31.33	30.58	40.2
+Topic structure	50.39	48.23	33.08	31.26	40.74
+Cross-chain	50.39	47.67	35.15	31.88	41.27
+Distribution	50.42	48.67	37.52	32.08	42.17
+Subevent	50.35	48.61	37.24	31.94	42.04

Table 4.3: Results for our heuristics-based event coreference resolution systems on the KBP 2016 and 2017 corpus. Joint learning results correspond to the actual result files evaluated in [Lu and Ng, 2017]. Reprinted with perimission from Choubey and Huang [2018].

best performing models on the KBP 2016 corpus. For both datasets, the full discourse structure augmented model achieved superior performance compared to the local classifier-based system. The improvement is observed across all metrics with an average F1 gain of 3.1 for KBP 2016 and 2.17 for KBP 2017. Most interestingly, we see over 28% improvement in the MUC F1 score which directly evaluates the pairwise coreference link predictions. This implies that the document level structures, indeed, help in linking more coreferent event mentions, which otherwise are difficult with the local classifier trained on lexical and surface features. Our ILP based system also outperforms the previous best model Lu and Ng [2017] on the KBP 2016 corpus consistently using all the evaluation metrics, with an overall improvement of 1.21 based on the average F1 scores.

In Table 4.3, we also report the F1 scores when we increasingly add each type of structure in the ILP baseline. Among different scoring metrics, all structures positively contributed to the MUC and BLANC scores for KBP 2016 corpus. However, subevent-based constraints slightly reduced

the F1 scores on KBP 2017 corpus. Based on our preliminary analysis, this can be accounted to the simple method applied for subevent extraction. We only extracted 31 subevents in KBP 2017 corpus compared to 211 in KBP 2016 corpus.

4.4 Conclusions

We have presented an ILP based joint inference system for event coreference resolution that utilizes scores predicted by a pairwise event coreference resolution classifier, and models several aspects of correlations between event coreference chains and document level topic structures, including the correlation between the main event chains and topic transition sentences, interdependencies among event coreference chains, genre-specific coreferent mention distributions and subevents. We have shown that these structures are generalizable on news documents by conducting experiments on both the KBP 2016 and KBP 2017 datasets. Our model outperformed the previous state-of-the-art model across all coreference scoring metrics.

5. DISCOURSE AS A FUNCTION OF EVENT: PROFILING DISCOURSE STRUCTURE IN NEWS ARTICLES AROUND THE MAIN EVENT[1]

Detecting and incorporating discourse structures is important for achieving text-level language understanding. Several well-studied discourse analysis tasks, such as RST [Mann and Thompson, 1988] and PDTB style [Prasad et al., 2008] discourse parsing and text segmentation [Hearst, 1994], generate rhetorical and content structures that have been shown useful for many NLP applications. But these widely applicable discourse structures overlook genre specialties. In this chapter, we focus on studying content structures specific to *news articles*. We believe that genre-specific discourse structures can effectively complement genre independent discourse structures and are essential for achieving deep story-level text understanding.

What is in a news article? Normally, we expect a news article to describe well-verified facts of newly happened events, aka the main events. However, almost no news article limits itself to reporting only the main events. Most news articles also report context-informing contents, including recent precursor events and current general circumstances, that are meant to directly explain the cause or the context of main events. In addition, they often contain sentences providing further supportive information that is arguably less relevant to main events, comprising of unverifiable or hypothetical anecdotal facts, opinionated statements, future projections, and historical backgrounds. Apparently, the *relevance* order of sentences is not always aligned with their *textual* order, considering that sentences in a news article are ordered based on their vague importance that is generally determined by multiple factors, including content relevance as well as other factors such as the focus of an article, the author's preferences, and writing strategies.

While a number of theoretical studies for news discourse exist, little prior effort has been put into computational modeling and automatic construction of news content structures. We introduce

[1]Reprinted with permission from "Discourse as a Function of Event: Profiling Discourse Structure in News Articles around the Main Event" by Prafulla Kumar Choubey, Aaron Lee, Ruihong Huang, and Lu Wang. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5374–5386, July 5 - 10, 2020. Copyright 2020 by Association for Computational Linguistics.

a new task and a new annotated text corpus for profiling news discourse structure that categorizes contents of news articles around the main event. The NewsDiscourse corpus consists of 802 news articles (containing 18,155 sentences), sampled from three news sources (*NYT*, *Xinhua* and *Reuters*), and covering four domains (*business*, *crime*, *disaster* and *politics*). In this corpus, we label each sentence with one of eight content types reflecting common discourse roles of a sentence in telling a news story, following the news content schemata proposed by Van Dijk [Teun A, 1986, Van Dijk, 1988a,b] with several minor modifications.

Next, we present several baselines for automatically identifying the content type of sentences. The experimental results show that a decent performance can be obtained using a basic neural network-based multi-way classification approach. The sentence classification performance can be further improved by modeling interactions between sentences in a document and identifying sentence types in reference to the main event of a document.

5.1 Elements of Discourse Profiling

We consider sentences to be units of discourse and define eight schematic categories to study their roles within the context of the underlying main event. The original Van Dijk's theory was designed for analyzing discourse functions of individual paragraphs with respect to the main event and the pilot study done by Yarlott et al. [2018] also considered paragraphs as units of annotations. Observing that some paragraphs contain more than one type of content, we decided to conduct sentence-level annotations instead to minimize disagreements between annotators and allow consistent annotations[2].

Table 5.1 contains an example for each content type. Consistent with the theory presented by Van Dijk [1988a], the categories are theoretical and some of them may not occur in every news article.

[2]Our two annotators agreed that the majority of sentences describe one type of content. For a small number of sentences that contain a mixture of contents, we ask our annotators to assign the label that reflects the main discourse role of a sentence in the bigger context.

Main Content	**Fine-grained type**
(1) U.S. President Donald Trump tried on Tuesday to calm a storm over his failure to hold Russian President Vladimir Putin accountable for meddling in the 2016 U.S. election, saying he misspoke in a joint news conference in Helsinki.	Main Event
(2) The rouble fell 1.2 percent on Tuesday following Trump's statement.	Consequence
Context-informing Content	**Fine-grained type**
(3) Trump praised the Russian leader for his "strong and powerful" denial of the conclusions of U.S. intelligence agencies that the Russian state meddled in the election.	Previous Event
(4) Special Counsel Robert Mueller is investigating that allegation and any possible collusion by Trump's campaign.	Current Context
Additional Supportive Content	**Fine-grained type**
(5) Congress passed a sanctions law last year targeting Moscow for election meddling.	Historical Event
(6) "The threat of wider sanctions has grown," a businessman told Reuters, declining to be named because of the subject's sensitivity.	Anecdotal Event
(7) Republicans and Democrats accused him of siding with an adversary rather than his own country.	Evaluation
(8) McConnell and House Speaker Paul Ryan, who called Russia's government "menacing," said their chambers could consider additional sanctions on Russia.	Expectation

Table 5.1: Examples for eight fine-grained content types used for news discourse profiling. Reprinted with permission from Choubey et al. [2020].

5.1.1 Main Contents

Main content describes what the text is about, the most relevant information of the news article. It describes the most prominent event and its consequences that render the highest level topic of the news report. **Main Event** (M1) introduces the most important event and relates to the major subjects in a news report. It follows strict constraints of being the most recent and relevant event and directly monitors the processing of remaining document. Categories of all other sentences in the document are interpreted with respect to the main event. **Consequence** (M2) informs about the events that are triggered by the main news event. They are either temporally overlapped with the main event or happens immediately after the main event.

5.1.2 Context-informing Contents

Context-informing sentences provide information related to the actual situation in which the main event occurred. It includes the previous events and other contextual facts that directly explain the circumstances that led to the main event. **Previous Event** (C1) describes the real events that preceded the main event and now act as possible causes or pre-conditions for the main event. They are restricted to events that have occurred very recently, within the last few weeks. **Current Context** (C2) covers all the information that provides context for the main event. They are mainly used to activate the situation model of current events and states that help to understand the main event in the current social or political construct. They have temporal co-occurrence with the main event or describe the ongoing situation.

5.1.3 Additional Supportive Contents

Finally, sentences containing the least relevant information, comprising of unverifiable or hypothetical facts, opinionated statements, future projections, and historical backgrounds, are classified as distantly-related content. **Historical Event** (D1) temporally precedes the main event in months or years. It constitutes the past events that may have led to the current situation or indirectly relates to the main event or subjects of the news article. **Anecdotal Event** (D2) includes events with specific participants that are difficult to verify. It may include fictional situations or personal accounts of incidents of an unknown person especially aimed to exaggerate the situation. **Evaluation** (D3) introduces reactions from immediate participants, experts, or known personalities that are opinionated and may also include explicit opinions of the author or those of the news source. They are often meant to describe the social or political implications of the main event or evaluation of the current situation. Typically, it uses statements from influential people to selectively emphasize their viewpoints. **Expectation** (D4) speculates on the possible consequences of the main or contextual events. They are essentially opinions, but with far stronger implications where the author tries to evaluate the current situation by projecting possible future events.

	M1	M2	C1	C2
Business	336(8.5)	40(1.0)	225(5.8)	1,041(26.6)
Crime	374(10.4)	78(2.2)	271(7.5)	941(26.1)
Disaster	407(10.6)	206(5.3)	223(5.8)	1,032(26.8)
Politics	475(10.4)	21(0.4)	218(4.8)	954(20.9)

	D1	D2	D3	D4	N/A
Business	238(6.1)	70(1.8)	1,049(26.8)	545(13.9)	368(9.4)
Crime	510(14.2)	77(2.1)	816(22.7)	204(5.7)	328(9.1)
Disaster	139(3.6)	330(8.6)	741(19.2)	405(10.5)	368(9.5)
Politics	228(5.0)	85(1.9)	1,492(32.7)	679(14.9)	414(9.1)

Table 5.2: Distribution of content type labels (discourse profiling) across domains, with percentages shown within parentheses. Reprinted with permission from Choubey et al. [2020].

	M1	M2	C1	C2
NYT	492(8.4)	97(1.7)	342(5.8)	1401(24.0)
Xinhua	667(13.6)	95(1.9)	361(7.4)	1249(25.5)
Reuters	624(8.4)	195(2.6)	391(5.1)	1837(24.8)
NYT_KBP	191(8.6)	42(1.9)	157(7.0)	519(23.3)

	D1	D2	D3	D4	N/A
NYT	714(12.2)	197(3.4)	1876(32.1)	532(9.1)	197(3.3)
Xinhua	214(4.4)	96(2.0)	953(19.5)	525(10.7)	736(15.0)
Reuters	571(7.7)	316(4.3)	1867(25.2)	924(12.5)	686(9.3)
NYT_KBP	384(17.3)	47(2.1)	598(26.9)	148(6.7)	141(6.3)

Table 5.3: Distribution of content type labels (discourse profiling) across media sources, with percentages shown within parentheses. Reprinted with permission from Choubey et al. [2020].

5.1.4 Speech vs. Not Speech

In parallel with discourse profiling annotations, we also identify sentences that contain direct quotes or paraphrased comments stated directly by a human and label them as Speech. We assign a binary label, Speech vs. Not Speech, to each sentence independently from the annotations of the above eight schematic discourse roles. Note that Speech sentences may perfectly be annotated with any of the eight news discourse roles based on their contents, although we expect Speech sentences to serve certain discourse roles more often, such as evaluation and expectation.

5.1.5 Modifications to the Van Dijk Theory

The Van Dijk's theory was originally based on case studies of specific news reports. To accommodate wider settings covering different news domains and sources, we made several minor modifications to the original theory. First, we label both comments made by external sources (labeled as "verbal reactions" in the original theory) and comments made by journalistic entities as speech and label speech with content types as well. Second, we added a new category, *anecdotal event* (D2), to distinguish unverifiable anecdotal facts from other contents. Anecdotal facts are quite prevalent in the print media. Third, we do not distinguish news *lead* sentences that summarize the main story from other Main Event (M1) sentences, considering that lead sentences pertain to the main event and major subjects of news.

5.2 Dataset Creation and Statistics

The NewsDiscourse corpus consists of 802 openly accessible news articles containing 18,155 sentences[3] annotated with one of the eight content types or *N/A* (sentences that do not contribute to the discourse structure such as photo captions, text links for images, etc.) as well as Speech labels. The documents span across the domains of business, crime, disaster, and politics from three major news sources that report global news and are widely used: NYT (USA), Reuters (Europe), and Xinhua (China). We include 300 articles each (75 per domain) from Reuters and Xinhua that are collected by crawling the web and cover news events between 2018-'19. NYT documents are taken from existing corpora, including 102 documents from KBP 2015[4] [Ellis et al., 2015] and 100 documents (25 per domain) from the annotated NYT corpus [Evan, 2008].

We trained two annotators for multiple iterations before we started the official annotations. In the beginning, each annotator completed 100 common documents (Eight from each of the domains and sources and four from the KBP) within the corpus to measure the annotator's agreement.

[3]Note that only sentences within the body of the news article are considered for annotation and headlines are considered as independent content. We used NLTK [Bird et al., 2009] to identify sentence boundaries in the body text. Occasionally, one sentence is wrongly split into multiple sentences, the annotators were instructed to assign them with the same label.

[4]KBP documents are not filtered for different domains due to the small size of the corpus.

The two annotators achieved Cohen's κ score [Cohen, 1968] of 0.69144, 0.72389, and 0.87525 for the eight fine-grained, three coarse-grained, and Speech label annotations respectively. Then, the remaining documents from each domain and news source were split evenly between the two annotators.

Detailed distributions of the created corpus, including distributions of different content types across domains and media sources, are reported in Tables 5.2 and 5.3 respectively. We find that distributions of content types vary depending on either domains or media sources. For instance, *disaster* documents report more consequences (M2) and anecdotal events (D2), *crime* documents contain more previous events (C1) and historical events (D1), while *politics* documents have the most opinionated contents (sentences in categories D3 and D4) immediately followed by *business* documents. Furthermore, among different sources, NYT articles are the most opinionated and describe historical events most often, followed by Reuters. In contrast, Xinhua articles have relatively more sentences describing the main event.

Speech labels and content type labels are separately annotated and each sentence has both a content-type label and a speech label (binary, speech vs. not speech). In the created corpus, 5535 out of 18,155 sentences are labeled as speech.

5.3 Document-level Neural Network Model for Discourse Profiling

A wide range of computational models has been applied for extracting different forms of discourse structures. However, across several tasks, neural network methods [Ji and Eisenstein, 2015, Becker et al., 2017] are found the most effective, with relatively superior performance obtained by modeling discourse-level context [Dai and Huang, 2018a,b].

As an initial attempt, we use a hierarchical neural network to derive sentence representations and document encoding, and model associations between each sentence and the main topic of the document when determining content types for sentences. Shown in Figure 5.1, it first uses a word-level bi-LSTM layer [Hochreiter and Schmidhuber, 1997] with soft-attention over word representations to generate intermediate sentence representations which are further enriched with the context information using another sentence-level bi-LSTM. Enriched sentence representations

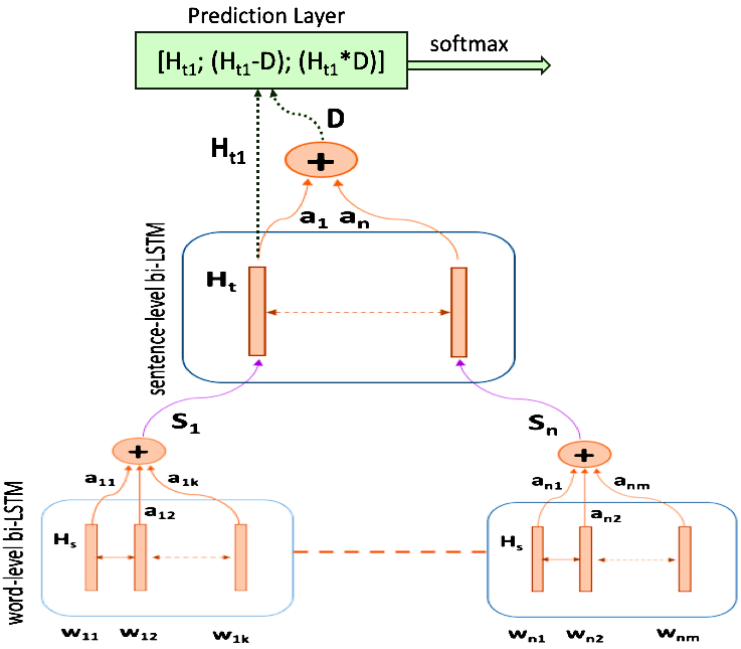

Figure 5.1: Neural-network architecture incorporating document encoding for news discourse profiling. Reprinted with permission from Choubey et al. [2020].

are then averaged with their soft-attention weights to generate document encoding. The final prediction layers model associations between the document encoding and each sentence encoding to predict sentence types.

Context-aware sentence encoding: Let a document be a sequence of sentences $\{s_1, s_2..s_n\}$, which in turn are sequences of words $\{(w_{11}, w_{12}..)\ ..\ (w_{n1}, w_{n2}, ..)\}$. We first transform a sequence of words in each sentence to contextualized word representations using ELMo [Peters et al., 2018] followed by a word-level biLSTM layer to obtain their hidden state representations H_s. Then, we take weighted sums of hidden representations using soft-attention scores to obtain intermediate sentence encodings (S_i) that are uninformed of the contextual information. Therefore, we apply another sentence-level biLSTM over the sequence of sentence encodings to model interactions among sentences and smoothen context flow from the headline until the last sentence in a document. The hidden states (H_t) of the sentence-level bi-LSTM are used as sentence encodings.

$$
\begin{aligned}
H_{s_i} &= biLSTM_{word}(s_i) \in R^{n_i \times 2d_{rnn}} \\
\alpha_{s_i}[k] &= W_{s1}(tanh(W_{s2}h_{s_i}[k] + b_{s2})) + b_{s1} \in R \\
A_{s_i} &= softmax(\alpha_{s_i}) \in R_i^n \\
S_i &= \sum_k A_{s_i}[k].H_{s_i}[k] \in R^{2d_{rnn}} \\
H_t &= biLSTM_{sentence}([S_1, S_2..S_n]) \in R^{n \times 2d_{rnn}}
\end{aligned}
\tag{5.1}
$$

Document Encoding: We generate a reference document encoding, as a weighted sum over sentence encodings using their soft-attention weights.

$$
\begin{aligned}
\alpha_t[i] &= W_{d1}(tanh(W_{d2}h_t[i] + b_{d2})) + b_{d1} \in R \\
A_t &= softmax(\alpha_t) \in R^n \\
D &= \sum_i A_t[i].H_t[i] \in R^{2d_{rnn}}
\end{aligned}
\tag{5.2}
$$

Modeling associations with the main topic: Sentence types are interpreted with respect to the main event. However, while the sentence-level biLSTM augments sentence representations with the local context, they may be still unaware of the main topic. Therefore, we compute element-wise products and differences between the document encoding and a sentence encoding to measure their correlations and further concatenate the products and differences with the sentence encoding to obtain the final sentence representation (t_i) that is used for predicting its sentence type.

$$
t_i = [h_t[i]; D - h_t[i]; D * h_t[i]] \tag{5.3}
$$

Predicting Sentence Types: First, we use a two-layer feed-forward neural network as a regular classifier to make local decisions for each sentence based on the final sentence representations. In addition, news articles are known to follow inverted pyramid [Bell, 1998] or other commonly used styles where the output labels are not independent. Therefore, we also use a linear chain CRF [Lafferty et al., 2001] layer on the output scores of the local classifier to model dependence among discourse labels.

5.4 Evaluation

We split 802 documents into training/dev/test sets of 502/100/200 documents. The training set includes 50 documents from each domain in Reuters and Xinhua, 9 documents from each domain in NYT, and 66 documents from KBP; the dev set includes 8 documents from each domain and source and 4 documents from KBP; and the test set includes 17 documents from each domain in Reuters and Xinhua, 8 documents from each domain in NYT and 32 documents from KBP. The dataset is released with the standard split we used in our experiments. For evaluation, we calculate the F1 score for each content type as well as micro and macro F1 scores.

5.4.1 Baseline Models

Feature-based (SVM) uses linear SVM classifier [Pedregosa et al., 2011] over features used by Yarlott et al. [2018], including bag of words, tf-idf and 100-dimensional paragraph vectors obtained through Doc2Vec [Le and Mikolov, 2014] implementation in Gensim [Řehůřek and Sojka, 2010]. Following Yarlott et al. [2018], we set minimum α to 0.01, minimum word count to 5 for Doc2Vec model and train it for 50 epochs. All three features are built on the entire training corpus and the value of C in SVM classifier is set to 10.

Basic Classifier uses only the word-level bi-LSTM with soft-attention to learn sentence representations followed by the local feed forward neural network classifier to make content type predictions.

5.4.2 Proposed Document-level Models

Document LSTM adds the sentence-level BiLSTM over sentence representations obtained from the word-level BiLSTM to enrich sentence representations with local contextual information.

+Document Encoding uses document encoding for modeling associations with the main topic and obtains the final sentence representations as described previously.

+Headline replaces document encoding with headline sentence encoding generated from the word-level biLSTM. Headline is known to be a strong predictor for the main event [Choubey et al., 2018].

CRF Fine-grained and **CRF Coarse-grained** adds a CRF layer to make content type predictions

for sentences which models dependencies among fine-grained (eight content types) and coarse-grained (main vs. context-informing vs. supportive contents) content types respectively.

5.4.3 Implementation Details

We set the dimension of the hidden state to 512 for both word-level and sentence-level biLSTMs in all our models. Similarly, we use two-layered feed-forward networks with 1024-512-1 units to calculate attention weights for both the BiLSTMs. The final classifier uses two-layer feed-forward networks with 3072-1024-9 units for predicting sentence types. All models are trained using Adam [Kingma and Ba, 2014] optimizer with the learning rate of 5e-5. For regularization, we use dropout [Srivastava et al., 2014] of 0.5 on the output activations of both BiLSTMs and all neural layers. Word embeddings are kept fixed during the training. All the neural model are trained for 15 epochs and we use the epoch yielding the best validation performance.

To alleviate the influence of randomness in neural model training and obtain stable experimental results, we run each neural model ten times with random seeds and report the average performance.

5.4.4 Results and Analysis

Tables 5.4 and 5.5 show the results from our experiments for content-type and speech label classification tasks. We see that a simple word-level biLSTM based *basic classifier* outperforms *features-based SVM* classifier [Yarlott et al., 2018] by 10.5% and 11.8% on macro and micro F1 scores respectively for content-type classification. Adding a sentence-level BiLSTM helps in modeling contextual continuum and improves performance by additional 4.4% on macro and 2.7% on micro F1 scores. Also, as content types are interpreted with respect to the main event, modeling associations between a sentence representation and the referred main topic representation using the headline or document embeddings improves averaged macro F1 score by 0.6% and 1.2% respectively. Empirically, the model using document embedding performs better than the one with headline embedding by 0.6% implying skewed headlining based on recency which is quite prevalent in news reporting.

Models	M1	M2	C1	C2	D1	D2	D3	D4
	F1 Scores							
Feature-based (SVM)	34.0	8.0	18.0	44.0	45.0	14.0	52.0	44.0
Basic Classifier	42.5	24.7	18.2	55.4	59.6	28.5	66.1	52.5
Document LSTM	49.3	27.3	20.2	57.0	63.6	45.8	**67.4**	55.6
+Headline	**49.8**	**30.0**	21.8	56.7	63.2	42.7	66.8	**58.7**
+Document encoding	49.6	27.9	**22.5**	**58.1**	**64.1**	**48.1**	**67.4**	57.6
CRF Fine-grained	47.7	26.4	22.2	56.0	63.3	45.2	66.4	55.2
CRF Coarse-grained	48.4	29.3	21.6	55.9	62.9	47.2	66.7	54.2

Models	Macro			Micro F1
	P	R	F1	
Feature-based (SVM)	39.1	37.9	38.3	45.7
Basic Classifier	52.6	47.9	48.8($\pm$0.8)	57.5($\pm$0.6)
Document LSTM	56.6	52.6	53.2($\pm$0.7)	60.2($\pm$1.0)
+Headline	**57.3**	52.9	53.8($\pm$0.7)	60.4($\pm$1.0)
+Document encoding	56.9	**53.7**	**54.4($\pm$0.8)**	**60.9($\pm$0.7)**
CRF Fine-grained	55.4	52.9	52.9($\pm$1.4)	59.4($\pm$1.1)
CRF Coarse-grained	55.6	53.4	53.5($\pm$0.9)	59.6($\pm$0.7)

Table 5.4: Performance of different systems on fine-grained discourse content type classification task. All results correspond to average of 10 training runs with random seeds. In addition, we report standard deviation for both macro and micro F1 scores. Reprinted with permission from Choubey et al. [2020].

Systems	P	R	F1
Feature-based (SVM)	61.0	71.0	69.0
Basic Classifier	**81.6**	80.7	81.2($\pm$0.4)
Document LSTM	80.7	**83.6**	**82.2($\pm$0.7)**

Table 5.5: Performance of different systems on speech label classification task. Reprinted with permission from Choubey et al. [2020].

We further aim to improve the performance by using CRF models to capture interdependencies among different content types, however, CRF models using both fine-grained and coarse-grained label transitions could not exceed a simple classifier model. The inferior performance of CRF models can be explained by variations in news content organization structures (such as the inverted pyramid, narrative, etc.), further implying the need to model those variations separately.

Similarly, for speech label classification task, word-level biLSTM model achieves 12.2% higher F1 score compared to the feature-based SVM classifier which is further improved by 1.0% with document-level biLSTM.

5.5 Conclusion

We have created the first broad-coverage corpus of news articles annotated with a theoretically grounded functional discourse structure. Our initial experiments using neural models ascertain the computational feasibility of this task.

6. IMPROVING EVENT COREFERENCE RESOLUTION BY INCORPORATING NEWS DISCOURSE STRUCTURES[1]

In chapter 4, we proposed the first holistic approach that captures several specific distributional patterns of coreferential event mentions using heuristics, to address the problem of severe distributional sparsity with event coreference resolution. The heuristics-based approach yielded clear empirical improvements over conventional clustering-based methods.

We further argue that many of the distributional patterns observed in chapter 4 can be explained and systematically identified by examining the functional structure of news discourse described in chapter 5. In order to verify this proposition, we associate event coreferences with news discourse structure in two ways. First, whether an event mention occurs as a singleton or a part of coreference chain depends on the nature of the event that is constrained by the discourse function of the sentence containing the event mention. For instance, the main event is often repeated several times in a news article while old historical events might only be mentioned once in sentences dedicated for describing historical backgrounds. Second, the number and locations of coreferential mentions are governed by the discourse role of an event and sentences containing the event as well, considering that the main event may appear in sentences focused on introducing the main event as well as sentences informing the general circumstances, while context events may only be mentioned in context-informing sentences.

We take the annotated KBP 2015 subset of NewsDiscourse corpus and analyze the dependency between event structure and document-level content structure. Further, we train the hierarchical content structure classifier, described in section 5.3, on the 2015 subset to identify discourse roles of sentences and another content-structure aware singleton classifier that distinguishes singletons from coreferential mentions. We then use predicted sentence content types and singleton predic-

[1]Reprinted with permission from "Discourse as a Function of Event: Profiling Discourse Structure in News Articles around the Main Event" by Prafulla Kumar Choubey, Aaron Lee, Ruihong Huang, and Lu Wang. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5374–5386, July 5 - 10, 2020. Copyright 2020 by Association for Computational Linguistics.

M1	M2	C1	C2	D1	D2	D3	D4
51%	**91%**	**79%**	**84%**	**86%**	**95%**	**84%**	**83%**

Table 6.1: Percentages of singleton events in sentences of each content type in the subset of KBP 2015 corpus with discourse profiling annotations. Reprinted with permission from Choubey et al. [2020].

tions along with pairwise coreference scores from a typical pairwise event coreference resolution classifier to jointly predict all event coreference clusters in an ILP formulation. Experimental results on the benchmark event coreference resolution datasets, the KBP 2016 [Ellis et al., 2016] and KBP 2017 [Getman et al., 2017] corpora, show that our systematic approach of identifying and incorporating content structure in an event coreference resolution system greatly improves the performance over the previous best heuristics-based approach.

6.1 Correlations between Event Coreferences and Content Structure

We investigate uses of news discourse profiling for event coreference resolution by analyzing 102 documents from the KBP 2015 corpus included in our *NewsDiscourse Corpus*. We analyze the lifespan and spread of event coreference chains over different content types. First, table 6.1 shows the percentage of events that are **singletons** out of all the events that appear in sentences of each content type. We can see that in contrast to main event sentences (M1), other types of sentences are more likely to contain singleton events.

We further analyze characteristics of non-singleton events to identify positions of their coreferential mentions and the spread of coreference chains in a document. Motivated by van Dijk's theory, we hypothesize that the **main events** appear in each type of sentence, but the likelihoods of seeing the main events in a sentence may vary depending on the sentence type. We consider events that appear in the news headline to approximate the main events of a news article. As shown in Table 6.2, around 58%[2] of main event sentences (M1) contain at least one headline event, in ad-

[2]While all the main event sentences are expected to mention some main event, we use headline events to approximate main events and headline events do not cover all the main events of a news article. As shown in our previous work [Choubey et al., 2018], identifying main events is a challenging task in its own right and main events do not always occur in the headline of a news article. In addition, event annotations in the KBP corpora only consider a limited set of event types, seven types specifically, therefore, if main events do not belong to those seven types, they

M1	M2	C1	C2	D1	D2	D3	D4
58%	15%	23%	15%	10%	9%	14%	14%

Table 6.2: Percentages of sentences of each content type that contain a headline main event in the subset of KBP 2015 corpus with discourse profiling annotations. Reprinted with permission from Choubey et al. [2020].

M1	M2	C1	C2	D1	D2	D3	D4
13%	0%	33%	49%	**69%**	**100%**	49%	13%

Table 6.3: Percentages of intra-type events out of non-singleton events in sentences of each content type in the subset of KBP 2015 corpus with discourse profiling annotations. Reprinted with permission from Choubey et al. [2020].

dition, context-informing sentences (C1+C2), especially sentences focusing on discussing recent pre-cursor events (C1), are more likely to mention headline events as well.

Other than the main events, we observe that many events have all of their coreferential mentions appear within sentences of the same content type. We call such events **intra-type events**. In other words, an *intra-type* event chain starts from a sentence of any type and dies out within sentences of the same content type. Table 6.3 shows the percentage of *intra-type* event chains out of all the event chains that begin in a certain type of sentence. We can see that non-main contents (e.g., content types C2-D3) are more likely to be self-contained from introducing to finishing describing an event. In particular, historical (D1) and anecdotal (D2) contents exhibit an even stronger tendency of having intra-type event repetitions compared to other non-main content types.

6.1.1 Comparisons with Heuristics from Chapter 4

We draw associations between event coreferences and the news discourse profiling and accordingly, present a systematic method for identifying and incorporating content structure for event coreference resolution. When compared to heuristics in chapter 4, we observe the following two correspondences.

are not annotated as events, which also contributes to the imperfect percentage of main event sentences containing a headline event.

- In chapter 4, we hypothesized that the main event often appears in topic transition sentences, but after noting that fine-grained topic segmentation within a document is difficult, we simply identified topic transition sentences as sentences that are mutually similar to each other by using a sentence similarity metric. Those topic transition sentences are akin to sentences of the main event content type, and here we present a systematic approach to identify the content type of any sentence (Section 6.2).

- Secondly, our heuristic rules distinguished local coreference chains from global coreference chains by using the stretch[3] of an event chain. But they failed to explain the nature of events featuring local coreference chains. With news discourse profiling, it becomes clear that local coreference chains correspond to events that are historical or anecdotal etc., that often appear within sentences of one non-main content type.

6.2 Content Structure-aware Singleton Classifier

We adjust the hierarchical discourse profiling system (Figure 5.1), to identify discourse roles of sentences as well as event singletons with shared model architecture (Figures 6.1). Given the enriched sentence representation (t_i) from equation 5.3, we first use a single-layer feed-forward neural network to obtain final sentence representations (T).

$$T_i = tanh(W_D t_i + b_D)) \in R^{d_{rnn}} \tag{6.1}$$

Identifying Discourse Role of Sentences we simply apply a linear classification layer over the final sentence representation from the shared model (T_i) to identify its content type.

Identifying Event Singletons we use a separate word-level biLSTM layer to encode event token along with its context in the sentence. We then combine the contextualized event representation with its final sentence representation (T_i) and document representation (D) and apply a two-layer feed-forward neural network to identify singletons. Let 'e' denotes the index of an event mention

[3]Defined as the number of sentences between the first and the last event mention based on their textual positions.

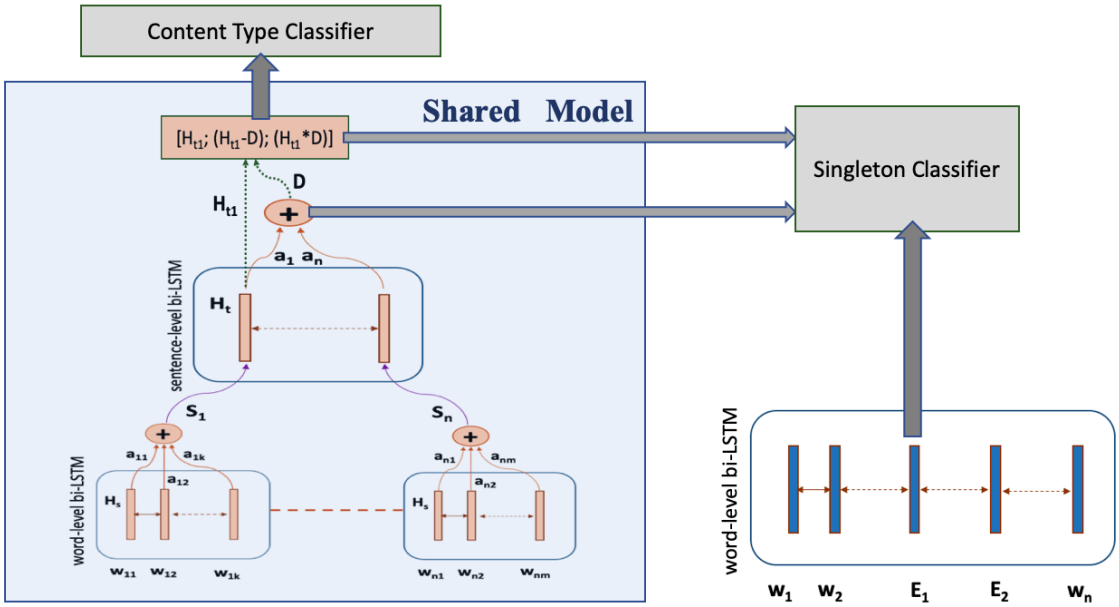

Figure 6.1: Neural-network model for discourse profiling and singleton event classification.

in sentence s_i, our singleton classifier is described as:

$$\begin{aligned} H_E &= biLSTM_{event}(s_i) \in R^{n_i \times 2d_{rnn}} \\ E &= [h_E[e]; D; T_i] \in R^{6d_{rnn}} \\ P &= W_{E1}(tanh(W_{E2}E + b_{E2})) + b_{E1} \in R \end{aligned} \tag{6.2}$$

6.3 ILP for Event Coreference Resolution

Let λ refers the set of all event mentions in a document and p_{ij} equals the score from the local pairwise classifier denoting event mentions 'i' and 'j' are coreferential. Similar to chapter 4, we formulate the baseline objective function that minimizes equation 6.3.

$$\begin{aligned} \Theta_B = \sum_{i \in \lambda, j \in \lambda} & -log(p_{ij})x_{ij} - log(1 - p_{ij})(\neg x_{ij}) \\ & s.t. \qquad x_{ij} \in \{0, 1\} \end{aligned} \tag{6.3}$$

We then add constituent objective functions (equation 6.4) and new constraints to the baseline objective to incorporate document-level content structure, including repetitions of headline events in main content (Θ_M) as well as in consequence, previous event and current context (Θ_C), intra-type

coreference chains in non-main contents (Θ_L) and exclusion of singletons from event coreferential chains (Θ_S) while reinforcing non-singletons to have more coreferential mentions (Θ_N).

$$\Theta = \Theta_B + K_M\Theta_M + K_C\Theta_C + K_L\Theta_L + K_S\Theta_S + K_N\Theta_N \tag{6.4}$$

The weighting parameters for all the constituent objective functions were obtained through grid search. We first preset all the values to 0.5 and then searched each parameter in the multiples of 0.5 over the range from 0.5 to 5. We found that the best performance was obtained for K_M=3.0, K_C=1.0, K_S=2.5, and K_N=0.5. Also, the best values for K_L are 0.5 for content types M2-C1 and 1.0 for content types C2-D8.

6.3.1 Infusing Singletons Score in the ILP Forumlation

Intuitively, coreferential event mentions and singletons are exclusive to each other. However, enforcing such mutual exclusion would be extremely unstable when both system predicted singletons and event coreference scores are imperfect. Therefore, we simply discourage singletons from being included in any coreference chains and encourage non-singletons to form more coreferential links in our model by adding two constituent objective functions Θ_S and Θ_N (equation 6.5).

$$\Theta_S = \sum_{i\in\lambda, j\in\lambda, i\vee j\in S} x_{ij} \;\; ; \;\; \Theta_N = -\sum_{i\in\lambda, j\in\lambda, i\wedge j\in N} x_{ij} \tag{6.5}$$

Where S and N are predicted singletons and non-singletons from content-structure aware singleton classifier. The relaxed Θ_S and Θ_N based implementation allows violations for predicted singletons when its pairwise coreference score with an event mention is high.

6.3.2 Incorporating Content Types in the ILP Forumlation

As evident from the analysis, main, consequence, previous event and current context content types favor coreferential event mentions with headline event. Furthermore, if an event chain starts in one of the C1-D4 content types, it tends to have coreferential event mentions within the same content type or sometimes in the main content. We model the above correlations between *main* and *non-main* content types and event coreference chains through their respective objective functions and constraints.

Main Events: for the event pairs with the first event mention from headline and the second one from main content sentences, we define a simple objective function (equation 6.6) that add the negative sum of their indicator variables to the main objective function.

$$\Theta_M = - \sum_{i \in \xi_H, j \in \xi_M} x_{ij} \tag{6.6}$$

Here, ξ_H and ξ_M indicate event mentions in headline and main content sentences respectively. By minimizing Θ_M in global objective function, our model encourages coreferential mentions between the headline and main content sentences.

Similarly, we define Θ_C that encourages coreferential mentions between the headline and sentences from consequence, previous event and current context content types (equation 6.7).

$$\Theta_C = - \sum_{i \in \xi_H, j \in \xi_R} x_{ij} \tag{6.7}$$

Here, ξ_R indicate event mentions in one of the consequence, previous event or current context content types.

Intra-type Events: for each non-main content type T, we define the objective function Θ_L and corresponding constraint (equation 6.8) to penalize event chains that start in that non-main content type sentence but include event mentions from other non-main type sentences.

$$\begin{aligned} \Theta_L &= \sum_{i \in \xi_T} Y_i \\ s.t. \quad & \Gamma_i - Y_i \leq M\gamma_i \\ \Gamma_i = \sum_{i \in \xi_T, j \notin (\xi_M \cup \xi_T)} x_{ij} \;\; ; \;\; & \gamma_i = \sum_{k \notin \xi_T, i \in \xi_T} x_{ki} \end{aligned} \tag{6.8}$$

First, we define an ILP variable Y_i for each event i in ξ_T, where ξ_T represents events in a non-main content type $T \in C1\text{-}D4$, and add that to the objective function Θ_L. Then, through the constraint in equation 6.8, we set the value of Y_i to Γ_i when λ_i is 0. Γ_i equals the number of subsequent coreferential event mentions of event i in sentences of other non-main types. γ_i equals the number of antecedent coreferential even mentions of event i in sentences of main or other non-main types. By minimizing Y_i in Θ_L, we discourage an event chain starting in a C1-D4 content type-sentence from forming coreferential links with subsequent event mentions in other non-main types.

6.3.3 Experimental Settings

We adopt the experimental settings used in chapter 4 for both development and evaluation of our ILP system. We use the same event mentions and pairwise event coreference scores for direct comparisons of all the results. In addition, we train our content type classifier on 102 documents annotated with the content types using 15 documents as the development set and the rest as training data. The singleton classifier was trained and tuned on training and evaluation documents from KBP 2015 [Ellis et al., 2015] respectively.

Model Implementation Details: We set hidden state dimensions of all biLSTMs (d^{rnn}) to 512 in both content type and singleton classifiers and use two-layered feed-forward networks with 1024-512-1 neurons to calculate attention weights for both word- and sentence-level BiLSTMs. Also, both models are trained using the dropout rate [Srivastava et al., 2014] of 0.5 over all BiLSTMs and neural layers. Adam [Kingma and Ba, 2014] optimizer with the learning rate of 3e-4 was used. Word embeddings are kept fixed during the training. Both models are trained for 30 epochs, and the final results correspond to the model with the best validation performance.

The weighting parameters for all the constituent objective functions are obtained through grid search. We first preset all the values to 0.5 and then search each parameter in the multiples of 0.5 over the range from 0.5 to 5. We found that the best performance was obtained for K_M=3.0, K_C=1.0, K_S=2.5, and K_N=0.5. Also, the best values for K_L are 0.5 for content types T2-T3 and 1.0 for content types T4-T8.

6.3.4 Our Systems

Full System: represents our content structure-guided model defined by the objective function in equation 6.4. It is optimized by considering all the constraints described in equations 6.3-6.5 and uses content type and event singleton classifiers that are fully trained from scratch.

- Structures: removes constraints and the constituent objective functions corresponding to that specific *structure* (singletons, main and non-main) from the full system.

Pretrained-Shared: uses the same ILP formulation as *Full System*. But it uses an event singleton

Model	B^3	$CEAF_e$	MUC	BLANC	AVG
	KBP 2016				
Heuristics	51.67	49.1	34.08	34.08	42.23
Full System	52.78	49.7	**34.62**	34.49	**42.9**
-Singletons	51.47	47.96	31.42	32.89	40.94
-Main	52.65	49.35	32.56	33.69	42.06
-Non-main	52.62	49.63	32.97	34.07	42.32
Pretrained-Shared	**52.82**	**49.82**	34.04	**34.58**	42.81
Local Singleton	52.72	49.77	31.68	33.33	41.88
	KBP 2017				
Heuristics	50.35	48.61	37.24	31.94	42.04
Full System	**51.68**	**50.57**	37.8	33.39	**43.36**
-Singletons	51.17	49.67	**38.01**	32.94	42.96
-Main	51.4	50.05	35.13	31.92	42.12
-Non-main	51.62	50.45	37.54	**33.42**	43.26
Pretrained-Shared	51.5	50.31	37.69	33.01	43.13
Local Singleton	51.3	50.07	36.66	32.4	42.61

Table 6.4: Results for event coreference resolution systems incorporating discourse-profiling structure on the benchmark evaluation datasets (KBP 2016 and 2017). Reprinted with permission from Choubey et al. [2020].

classifier with the shared model architecture pre-trained for content type prediction.

Local Singleton: another variant of the *Full System* which uses a singleton classifier trained on local features. Specifically, the singleton classifier replaces the final sentence and document representations from the shared model architecture with a local sentence representation obtained by summing hidden states of $biLSTM_{event}$ through scalar soft-attention weights.

6.3.5 Results and Analysis

As shown in Table 6.4, our content-structure aware model outperforms heuristics-based model consistently across all the evaluation metrics, with average F1 gains of 0.67% and 1.32% on KBP 2016 and KBP 2017 corpora respectively. In line with the heuristics-based model, the superior performance of our model comes from the significant improvement in the MUC F1 score. This implies that systematically identifying and incorporating both content-type and singleton structures is more competent in recognizing difficult coreferential mentions that are unidentifiable by the local

pairwise classifier as well as the previous heuristics-based system [Choubey and Huang, 2018].

To further evaluate the importance of Singletons, Main and Non-main structures, we perform ablation experiments by removing each structure from the full model. Based on the results in Table 6.4, structures pertaining to singletons and main-content type contribute the most to coreference performance. Specifically, the performance of *full model* drops by around 2% on KBP 2016 corpus when the singletons structure was removed. On the contrary, performance drop on KBP 2017 corpus was limited to 0.4% only. It is interesting to note that the event mention identification system has significantly lower precision on KBP 2016 (58.17%) compared to KBP 2017 (70.29%). We believe that the singleton classifier is capable of eliminating event mentions that are incorrectly identified by an event mention identification system.

The significant role of content-structure aware singleton classifier is also evident from the last two rows in Table 6.4, incorporating a *local singleton* classifier clearly hurt the performance of the system on the KBP 2017 corpus. This suggests the necessity of document-level cues for extracting event singletons. Lastly, the singleton classifier trained from the scratch used in *full model* allows the classifier to learn low-level features suitable for this task and thus has slightly better performance than the *Pretrained-Shared* model.

6.4 Conclusion

We have presented our approach of systematically identifying and incorporating content structures for event coreference resolution that employs all event, sentence, and document level representations to identify sentence content types and singleton events. The event coreference resolution system incorporating content structure constraints using an ILP based inference framework outperformed the heuristics-based model across all coreference scoring metrics on the KBP 2016 and KBP 2017 English corpora.

7. AUTOMATIC DATA ACQUISITION FOR EVENT COREFERENCE RESOLUTION[1]

In this chapter, we aim to improve the effectiveness of event coreference resolution systems by automatically acquiring coreferential event pairs from many documents requiring minimal supervision. Specifically, as noted in chapter 6, coreferential event mentions are associated with the discourse function of sentences in a news document. Next, we propose to use them to identify sentence pairs that are likely to contain coreferential event mentions as well as sentence pairs that are likely to contain non-coreferential event pairs. Consider the two example sentence pairs below, each pair having an event pair with synonymous trigger words.

(1): [People living in absolute poverty in rural areas of the eight regions and provinces **reduced** to 14.52 million from 30.76 million over the last decade.] [Yang admitted , however , that ethnic minority regions still lagged far behind the developed eastern regions and the government still faced serious challenges to **reduce** poverty.]

(2): [At least 30,000 war-displaced people camped in Angola's central province of Kwanza-sul are being **resettled** in productive areas, the official news agency angop reported here on Friday.] [The **resettlement** is being carried out jointly by the local municipal authorities of Seles, located in southern Kwanza-sul, and the charity organization German Agro Action, the news agency said.]

In example (1), the first sentence describes a historical event about the reduction in poverty during the last decade, while the second sentence projects the challenges of further reducing poverty in the coming years. Here, the two *reduce* events are non-overlapping in the temporal space and are non-coreferential. On the contrary, in example (2), both mentions for the event *resettle* refer to the same real-world event and can be so ascertained by knowing that both sentences describe the same main event in a news article. In general, we can recognize pairs of sentences in news

[1]Reprinted with permission from "Automatic Data Acquisition for Event Coreference Resolution" by Prafulla Kumar Choubey, and Ruihong Huang. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 1185–1196, April 19 - 23, 2021. Copyright 2021 by Association for Computational Linguistics.

articles that are likely to contain coreferential or non-coreferential event mention pairs by knowing the sentence's discourse function following news discourse profiling.

To ascertain our hypothesis, we first use the discourse profiling dataset introduced in chapter 5 and build a new improved model to identify the discourse role for each sentence in a news article. Then, we use multiple rules to capture the distributional correlation between event coreference chains and discourse roles of sentences and collect a diverse set of 9,210 coreferential and 232,135 non-coreferential event pairs[2]. To assess the reliability of the proposed data augmentation strategy, we perform manual validation on subsets of both coreferential and non-coreferential event pairs. Then, we train event coreference resolution systems using the acquired data alone or using the acquired data to augment a human-annotated training dataset.

We evaluate trained systems on two datasets, the news portion of the widely used benchmark evaluation corpus KBP 2017 as well as the news portion of the Richer Event Description (RED) corpus [O'Gorman et al., 2016]. Unlike the KBP corpora that only consider eight event types for event coreference annotations, the RED corpus comprehensively annotates all the event types that appear in a document and is arguably the only comprehensively annotated corpus of event coreference relations. Assuming the automatically acquired event coreference data is not available, we also train a supervised event coreference resolution system using the KBP 2015 corpus[3]. On the KBP 2017 corpus, the event coreference resolution system trained on the acquired data performs slightly worse than the system trained using the KBP 2015 corpus, the human-annotated in-domain training data. But, on the RED corpus, both the systems trained on either the annotated KBP 2015 corpus or the acquired data obtain roughly the same evaluation results. Further, the system trained on combined annotated KBP 2015 and automatically acquired data yields the best results on both the KBP 2017 dataset and the RED dataset.

Lastly, we evaluate all the trained systems on a different text genre, discussion forum articles from the KBP 2017 corpus, and found that all the systems obtain comparable results. Overall, the

[2]The acquired coreferential and non-coreferential event pairs are available at `https://github.com/prafulla77/Event-Coref-EACL-2021`

[3]We only use the news articles from KBP 2015 to train the supervised system.

performance gain of all the trained systems on discussion forum documents is marginal compared to a simple trigger word match baseline. Thus, increasing training data size does not improve the performance of an event coreference resolution system on a new text genre. We suspect that, for generalization across different text genres, we may require specialized learning algorithms, e.g., text style adaptation, which is not in the scope of this dissertation.

7.1 Event Coreference Data Acquisition

To acquire coreferential event-pairs without direct supervision, we first collect event trigger words along with their potential set of coreferential event mentions using The Paraphrase Database (PPDB 2.0) [Ganitkevitch et al., 2013, Pavlick et al., 2015][4]. Then, we use high precision rules informed by the functional news discourse structures [Teun A, 1986, Choubey et al., 2020] to identify seed coreferential and non-coreferential event pairs followed by a single bootstrapping iteration to collect additional non-coreferential event pairs.

7.1.1 Identifying Coreferential Event Trigger Words using The PPDB Database

We collect lexically diverse candidate coreferential event pairs using the paraphrases from PPDB-2.0-s-lexical [Pavlick et al., 2015] database. The corpus[5] contains 213,716 highest scoring lexical paraphrase pairs, each annotated with one of the equivalence, forward or reverse entailment, and contradiction relation classes. First, we extract all the *verb* paraphrase pairs as the potential event trigger words. While event mentions can take other part of speech types, we limit our paraphrase pairs to verbs to ensure high precision among the collected event trigger words. Additionally, many of the verb paraphrase pairs include nominalization (e.g., investing and investment), which adds to the syntactic diversity in the event pairs without compromising their quality. Then, among all verb paraphrase pairs, we filter out only three relation classes, namely *equivalence*, *reverse entailment* and *forward entailment*, as the potential coreferential event pairs. The forward and reverse entailment relations characterize hyponym and hypernym relations, which are

[4] A contemporary work by Meged et al. [2020] has also studied the potential correlation between coreferential event trigger words and predicate paraphrases.

[5] `http://nlpgrid.seas.upenn.edu/PPDB/eng/ppdb-2.0-tldr.gz`

not semantically equivalent but can often be coreferential and thus, add diversity to the pairs. Finally, we manually remove noisy event trigger words and cluster the remaining event pairs through pivoting, based on a common event trigger word shared between two paraphrase pairs[6]. Overall, we obtain 1023 clusters with an average of 3.375 event trigger words per cluster.

7.1.2 Post-Filtering Paraphrase-based Event Pairs using Functional News Discourse Structure

To generate the news discourse structure proposed, we first build a new discourse profiling system discussed in section 7.2. Note that the discourse profiling task classifies each sentence in a document into one of the eight content types where each content type describes the specific role of a sentence in describing the main event, context informing events, and other historical or future projected events. Among the eight content types, events described in *main event* sentences are central to the main news topic. They routinely appear in headline and sentences of other content types and consequently are more likely to form event coreference chains. On the contrary, events in the *historical event* content type are restricted to describing certain historical background and might only be mentioned once in the document. Additionally, events mentioned in *previous event* sentences tend to happen before those in *main event* and *consequence* sentences, and are unlikely to be coreferential with the events from the later two content types. Overall, content types provide cues for determining whether the events from a certain sentence possess coreferential event mentions and we leverage them to locate both coreferential and non-coreferential event pairs in a news article. Our event coreference data acquisition method works in two phases.

Rule-based Filtering to extract Coreferential and Non-coreferential Event Pairs: In the first phase, we extract both coreferential and non-coreferential event mention pairs based on their respective rules. Specifically, two event mentions from the headline or main event sentences with synonymous event trigger words are identified as coreferential event pairs. Considering that coreferential event mentions are very sparsely distributed, simple trigger-word matching is extremely noisy and damaging when used to train an event coreference classifier. However, narrowing coref-

[6]The processed event clusters are available at `https://git.io/JtnMf`

erential event mention pairs to synonymous event trigger words from main event sentences or headline significantly eliminates false coreferential event pairs. To get non-coreferential event pairs, we require both trigger words to be non-synonymous and belong to either the same sentence or two sentences of different non-main content types. Further, considering that events in historical event sentences tend to precede the main event by months and years, we identify non-synonymous event pairs with one mention in a historical event sentence and another mention in a main event sentence as non-coreferential. The latter rule allows us to also acquire non-coreferential event pairs with one event from main event sentences, adding to the overall diversity of the acquired dataset.

Distilling Non-coreferential Event Pairs with Synonymous Trigger Words: All the non-coreferential event pairs acquired in phase one have non-synonymous trigger words. However, we know that many of the synonymous words are non-coreferential. Therefore, to further diversify the acquired event coreference data, we use the second-phase bootstrapping to extract non-coreferential pairs with synonymous trigger words. We once again leverage the temporal separation between historical and other content types. We first identify synonymous event pairs that have one mention in a historical sentence and another mention in any non-historical sentence as candidate non-coreferential pairs. Then, we use an event coreference classifier trained on the dataset extracted in phase one to filter out high scoring non-coreferential event pairs (likelihood ≥ 0.9) from the candidate pairs.

7.1.3 Statistics of Acquired Coreference Data

We use Xinhua news articles[7] from the English Gigaword [Napoles et al., 2012] corpus to acquire coreferential and non-coreferential event pairs using the proposed methodology. We limit the number of coreferential and non-coreferential event pairs for each trigger word to 20 and 200, respectively, to ensure diversity and reduce repetitions of common event trigger words. We compare our acquired event pairs with the KBP 2015 corpus, which has 179 news documents annotated with eight event types and 38 event subtypes. It is the most widely used corpus for training a

[7]The discourse profiling system obtains the best performance on Xinhua news articles compared to NYT and Reuters

Data	# Coref	# Non-Coref
Rule-based (Phase I)	9210	226776
Distillation (Phase II)	0	5359
KBP 2015	4401	106383

Table 7.1: Number of coreferential and non-coreferential events pairs acquired through the proposed paraphrases with discourse profiling-based rules and the human annotated KBP 2015 corpus. Reprinted with permission from Choubey and Huang [2021].

Row	Data	Prec.	80% CI
1	Synonyms: Coref	49.0	45.3-52.6
2	Synonyms: Non-Coref	51.0	47.3-54.6
3	Phase I: Coref	83.0	80.3-85.6
4	Phase I: Non-Coref	99.3	98.6-100
5	Phase II: Non-Coref	93.0	90.0-96.0

Table 7.2: Precision (Prec.) and bootstrap 80% confidence interval (80% CI) score of precision for acquired event pairs based on human evaluation. Reprinted with permission from Choubey and Huang [2021].

within-document event coreference resolution system. Table 7.1 shows the number of event pairs obtained in the first and second phases of our data acquisition strategy and the human-annotated KBP 2015 corpus. Overall, the total number of extracted coreferential event pairs is more than twice the number of pairs in news documents from the KBP 2015 corpus. Note that we can increase the number of acquired pairs by expanding the synonymous event trigger word list or the unlabeled news article collection.

7.1.4 Manual Evaluation of Acquired Event Pairs

We randomly selected 300 event pairs from each of the coreferential and non-coreferential samples extracted in the first phase, 100 event pairs from non-coreferential samples distilled in the second phase, and 300 event pairs having synonymous event trigger words to evaluate the proposed data acquisition methodology. Then, we asked a human annotator to validate all the 1000 samples manually.

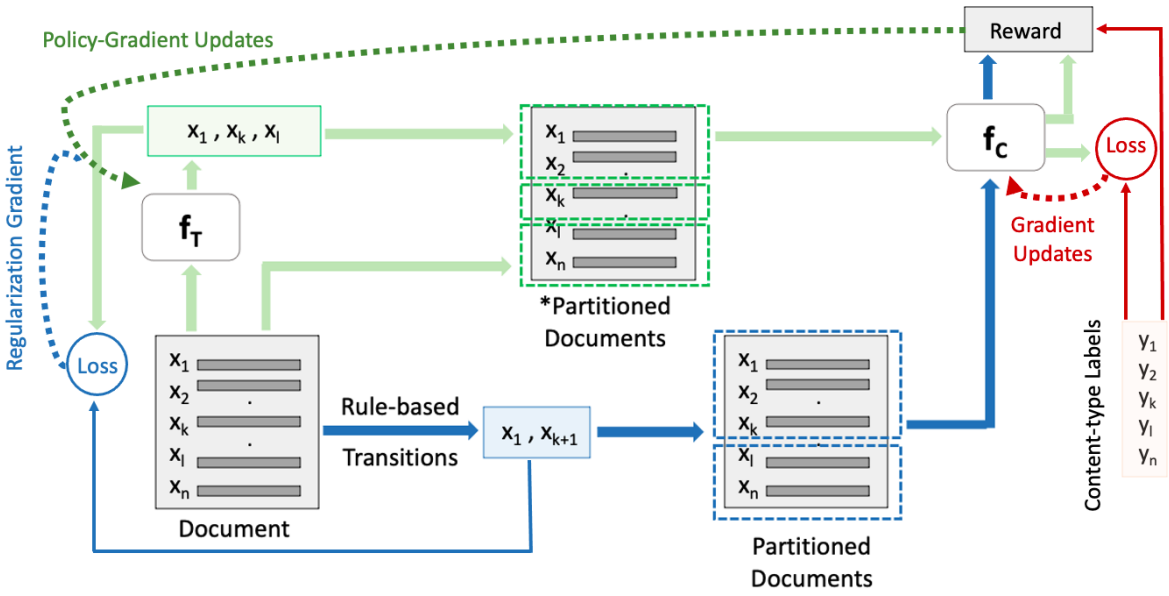

Figure 7.1: Neural-network architecture, including gradient flow paths, for incorporating document level content structures in a discourse profiling system.

Table 7.2 shows the precision and bootstrapped 80% confidence interval of precision for event pairs from each category. Rows 1 and 2 show that only 49% of synonymous event pairs are coreferential while the remaining are non-coreferential. By comparing rows 1 and 3, we can see that limiting coreferential event pairs to the synonymous event trigger words from the headline and main event sentences improves the precision from 49% to 83%. As shown in rows 4 and 5, our rules achieve high precision in identifying non-coreferential event pairs as well, achieving 99.3% for event pairs with non-synonymous trigger words acquired in the first phase and even 93% for event pairs with synonymous trigger words acquired in the second phase. Note that the high precision of non-coreferential event pair identification in both phases is partly due to the distributional sparsity of event coreference chains.

7.2 A New Improved Discourse profiling Model

Our new model tackles the discourse profiling task in a two step process. Given a news document $X : \{H, x_1, x_2, .., x_n\}$ comprising of headline H and n sentences with their content-type labels $Y : \{y_1, y_2, .., y_n\}$, we aim to learn a model $f : X \rightarrow Y$ that classifies each sentence x_i in

the document X to its content type y_i. In the first step, a latent function $f_T : X \rightarrow T \in \{0, 1\}^n$, a binary classifier, is used to identify transition sentences in the document. The transition sentences are used to partition documents into multiple subtopics. In the second step, a classification function $f_C : [X, T] \rightarrow Y$ combines the output of latent function f_T with the sentences X in document to perform final content-type classification (Figure 7.1). Overall, the model consists of two sets of sentence encoders, a biLSTM-based [Hochreiter and Schmidhuber, 1997] hierarchical encoder (§ 7.2.1) to obtain contextualized sentence representations used by both f_T and f_C, and two weighted bag-of-words and biLSTM-based sentence encoders to model local textual continuity (§ 7.2.2) that are exclusively used by f_T.

7.2.1 Learning Contextualized Sentence Representations

We use a hierarchical encoder, similar to chapter 5, to learn local context-aware sentence representations. Given a word sequence x_i represented by $\{w_{i1}, w_{i2}, .., w_{im}\}$, we first transform the sequence to contextualized word embeddings E_i using the pre-trained ELMo [Peters et al., 2018]. Then, we use a word-level biLSTM layer over E_i to obtain hidden state representations H_i and take their weighted average to obtain the local sentence embedding S^L. Weights for hidden states are obtained using a two-layered feed-forward neural network (eq. 7.1).

$$
\begin{aligned}
[E_{i1}, E_{i2}, .., E_{im}] &= ELMo([w_{i1}, w_{i2}, .., w_{im}]) \\
[H_{i1}, H_{i2}, .., H_{im}] &= biLSTM^L([E_{i1}, E_{i2}, .., E_{im}]) \\
\alpha_i[k] &= W_{\alpha 1}(tanh(W_{\alpha 2}E_{ik} + b_{\alpha 2})) + b_{\alpha 1} \in R \\
A_i &= softmax(\alpha_i) \in R^m \\
S_i^L &= \sum_k A_i[k]H_{ik} \in R^{2d_{rnn}}
\end{aligned}
\tag{7.1}
$$

Finally, we apply another sentence-level biLSTM over the sequence of local headline and sentence embeddings $\{H^L, S_1^L, S_2^L, .., S_n^L\}$ to obtain the contextualized sentences representations S^C that are later used in both the sub-modules f_T and f_C (eq. 7.2).

$$
[H^C, S_1^C, .., S_n^C] = biLSTM^C([H^L, S_1^L, .., S_n^L]) \tag{7.2}
$$

7.2.2 Modeling Local Continuity to Identify Transition Sentences

In addition to the contextualized sentence representations defined in § 7.2.1, we derive separate sentence representations that are especially useful for identifying transition sentences, considering that transition sentence identification is dependent on detecting breaks of local textual continuity different from the final task of discourse profiling that classifies sentences based on their functional discourse roles. Specifically, we learn two separate representations for every pair of adjacent sentences (E_{i-1}, E_i) in a document that aims to measure local coherence C_i^R and cohesion C_i^S. Both representations are learned using dual encoder architecture [Cho et al., 2019] that helps to preserve the asymmetry of coherence and cohesion properties.

Encoding Local Coherence: Coherence encoder uses weighted averaging over the word embeddings sequence to capture only high-level semantic information and measure relatedness between two adjacent sentences. First, we use a feed-forward neural network over word-embeddings to obtain mixture weights for all words in a sentence. Then, we sum all the word-embeddings with their respective weights and apply another single neural layer to obtain the weighted bag-of-words (W-BoW) representation, S_i^{wb}, for a sentence E_i (eq. 7.3).

$$\begin{aligned} \alpha_i[k] &= W_{b1}(tanh(W_{b2}E_{ik} + b_{b2})) + b_{b1} \in R \\ A_i &= softmax(\alpha_i) \in R^m \\ S_i^{wb} &= \sum_k A_i[k]E_{ik} \in R^{2d_{rnn}} \end{aligned} \tag{7.3}$$

We use two different encoders, defined by equations 7.3, to obtain W-BoW embeddings for each of the current (S_i^{wb}) and previous (S_{i-1}^{wb}) sentences. Then, we take the element-wise difference and product of the previous and current sentence embeddings to model the local coherence (eq. 7.4).

$$C_i^R = [S_i^{wb} * S_{i-1}^{wb};\ S_i^{wb} - S_{i-1}^{wb}] \in R^{4d_{rnn}} \tag{7.4}$$

Note that the ELMo representations, despite being static, are pre-contextualized and W-BoW is not an absolute bag-of-word representation. However, they were found empirically superior to the GloVe [Pennington et al., 2014] embeddings in our models.

Encoding Local Cohesion: Because cohesion is determined by explicit lexical and syntactic consistency [Todirascu et al., 2016], cohesion encoder incorporates the word order information when encoding a sentence. It follows the weighted-sum formulation used in word-level biLSTM encoder in the hierarchical model (eq. 7.1) to obtain sentence embedding. Then, similar to the coherence encoding, we use two different encoders to obtain cohesion embeddings for each of the current (S_i^{cs}) and previous (S_{i-1}^{cs}) sentences and take element-wise difference and product of the two sentence embeddings to model the local cohesion (eq. 7.5).

$$C_i^S = [S_i^{cs} * S_{i-1}^{cs};\ S_i^{cs} - S_{i-1}^{cs}] \in R^{4d_{rnn}} \tag{7.5}$$

7.2.3 Identifying Transition Sentences

In order to identify transition sentences, we use a two-layered feed-forward neural network that takes element-wise difference and product between contextualized sentence representations for the current S_i^C and preceding S_{i-1}^C sentences together with the local coherence C_i^R and cohesion C_i^S representations and outputs the transition-likelihood of current sentence. Overall, the feed-forward network (eq. 7.6) along with contextualized sentence encoder (eq. 7.1,7.2), and local coherence (eq. 7.4) and cohesion (eq. 7.5) encoders represent our f_T model which primarily measures content correlation between consecutive sentences and identify transition boundaries.

$$\begin{aligned} t_i &= [S_i^C * S_{i-1}^C;\ S_i^C - S_{i-1}^C; C_i^R; C_i^S] \in R^{12d_{rnn}} \\ T_i &= softmax(W_{t1}(tanh(W_{t2}t_i + b_{t2})) + b_{t1}) \in R^2 \end{aligned} \tag{7.6}$$

7.2.4 Discourse Profiling

Given the list of transition sentences $T_L : \{T_i | T_i \geq 0.5\}$ and contextualized sentence representations $[H^C, S_1^C, .., S_n^C]$, we use scalar soft-attentions (α_s) over sentence representation, as described in eq. 7.7, to learn local subtopic (T) and global document (D) representations. Finally, we combine sentence, local subtopic and document representations through element-wise product and differences (u_i) and use a two-layered feed-forward neural network to predict the labels. The networks defined in eq. 7.7 together with the contextualized sentence encoding network in equations 7.1 and 7.2 make the discourse profiling network f_C.

7.2.5 Learning f_T through Subtopic Structures-guided Critic

Our goal is to train the neural network-based transition sentence scorer f_T model using indirect supervision derived from the performance of f_C on the discourse profiling task. Intuitively, REINFORCE algorithm [Williams, 1992], which has shown success in a range of NLP tasks, offers a suitable mechanism to train our f_T model. However, the vanilla reinforce is known to suffer from the problem of high variance. Therefore, we propose a new variant of *actor-critic* [Konda and Tsitsiklis, 2000] model which defines critic through a known subtopic structure.

Specifically, we consider f_T as the actor network that gives the output score T indicating the likelihood that each sentence represents a subtopic boundary. Given, the output score $\pi(x_i; X; f_T)$, which indicates the likelihood that sentence x_i represents a subtopic boundary, we sample an action I_i according to categorical distribution to obtain transition sentences T_S (eq. 7.8).

$$\begin{aligned}
\alpha_s[i] &= W_{s1}(tanh(W_{s2}S_i^C + b_{s2})) + b_{s1} \in R \\
A_T &= softmax(\alpha_s[T_L[j] : T_L[j+1]) \in R^{T_L[j]-T_L[j+1]} \\
T &= \sum_{k=T_L[j]}^{T_L[j+1]} A_T[k].S^C[k] \in R^{2d_{rnn}} \\
A_s &= softmax(\alpha_s) \in R^n \\
D &= \sum_i A_s[i].H_s[i] \in R^{2d_{rnn}} \\
u_i &= [S_i^C - T; S_i^C * T; T - D; T * D] \in R^{8d_{rnn}} \\
\hat{y_i} &= softmax(W_{c1}(tanh(W_{c2}u_i + b_{c2})) + b_{c1}) \in R^9
\end{aligned} \tag{7.7}$$

$$\begin{aligned}
I_i &\sim Categorical(\pi(x_i; X; f_T)) \\
T_S &= \{x_i | I_i = 1\}
\end{aligned} \tag{7.8}$$

To calculate the reward, we use sampled transition sentences T_S to partition the news document and use eq. 7.7 to identify content-types $\hat{Y} : \{\hat{y_1}, .., \hat{y_n}\}$ for all the sentences. We calculate the average of micro and macro F1 scores of the predicted content types $\hat{Y}$ and use that as the reward R_A for our actor network. Following the same steps with reference transition sentences T_R that are derived from a known subtopic structure, we also obtain the reward R_C for our critic system and

use eq. 7.9 to derive training loss L_{RL} for training our transition scorer.

$$L_{RL} = (R_A - R_C)(\sum_{i \in T_S} -log(T_i^1) + \sum_{i \notin T_S} -log(T_i^0)) \tag{7.9}$$

At every iteration, the RL loss term forces the f_C model to perform at least as good as the model with known subtopic structure-based reference transition sentences T_R. The f_C model thus converges to parameters that obtain higher reward than its counterpart with reference transition sentences. For the f_T model, if it chooses good transition sentences T_S, that give a higher reward than T_R, it further increases the likelihood for T_S. When it chooses bad T_S, the negative reward ($R_A < R_C$) discourages the identified T_S. For the latter case, while the loss term knows the transition sentences that are undesirable, it does not know what transition sentences can increase the reward since the critic computes the exact reward values based on T_R that is independent of the actor model. Therefore, we add a regularization loss L_T, defined in eq. 7.10, with a small coefficient to the L_{RL} to encourage policy exploration towards the reference transition sentences. Overall, our combined loss for f_C and f_T constitutes average over the cross-entropy loss on discourse profiling task (L_C) and weighted average over L_{RL} and L_T as described in equation 7.10. Note that subtracting the critic's reward gives an unbiased estimate of the reward of actor in expectation.

$$\begin{aligned} L_C &= \sum_{i}^{n} \sum_{c \in labels} -y_i^c log(\hat{y_i^c}) \\ L_T &= \sum_{i} -log \frac{exp(T_R[i])}{\sum_{T_k \in T_R[i-1:]} exp(T_k)} \\ L_{full} &= 0.5(L_C + (1-\lambda)L_{RL} + \lambda L_T) \end{aligned} \tag{7.10}$$

7.2.6 Known Sub-topical Structure to Define the Critic

To define the critic, we consider inverted pyramid structure [Po¨ ttker, 2003], which is most often used in news media. It organizes the news content in decreasing order of relevance, placing the most relevant information at the top and then arranging the remaining details in decreasing order of relevance. While the inverted pyramid is a global content organization structure, we made a simplifying assumption that a document consists of smaller sequences of segments that locally follow the inverted pyramid structure. We identify a sentence as representing a transition boundary

if it breaks the non-increasing relevance order of preceding sentences, i.e. its relevance lies higher than its preceding sentences on the relevance scale. Given that the relevance order of sentences is not always aligned with their textual order, it provides an accessible proxy to define subtopical boundaries.

Specifically, since the eight discourse content types align with the relevance order of content in a document, with $M1$ being the most relevant and central to the document, followed by immediate consequences ($M2$) and causes ($C1$) and then the general context, opinions, and expectations ($C2, D3, D4$), it allows us to use the content types of sentences to extract transition sentences and partition a document into smaller subtopical segments. For instance, a main event sentence following context-informing or supportive contents will make the main event sentence a subtopic transition sentence. With the above rationales, we first identify the first sentence of a document as a transition sentence. Then, given a document and content labels $(x_i, y_i) \in X$, we identify new transition sentences x_i following the rules defined in Algorithm 2. Note that we dissociated historical ($D1$) and anecdotal ($D2$) content types from the relevance ordering as they are frequently used to set the tone for a news article or to highlight the main argument with personal experiences or historical events.

Algorithm 2 Rules: Identifying Transitions

1: $x_i \in \{M1\}$ and $x_{i-1} \in \{M2 - D4\}$

2: $x_i \in \{M2\}$ and $x_{i-1} \in \{C1 - D4\}$

3: $x_i \in \{C1\}$ and $x_{i-1} \in \{C2 - D4\}$

4: $x_i \in \{D1, D2\}$ and $x_{i-1} \notin \{D1, D2\}$

5: $x_i \notin \{D1, D2\}$ and $x_{i-1} \in \{D1, D2\}$

7.3 Event Coreference Resolution System

We design a neural network-based mention-pair classifier for event coreference resolution. We represent each event pair using 50 context words to the left and right of the first and second event

trigger words respectively, and with the maximum of 200 words in between the two event words[8].

Given the event context $(w_1, ., e_1, ., e_2, ., w_n)$, we first transform the context words sequence to word embeddings sequence $(b_{w1}, ., b_{e1}, ., b_{e2}, ., b_{wn})$ using the pre-trained Bert-Large-uncased model [Devlin et al., 2019]. Then, we model the semantic associations between two event mentions by measuring the similarity between their event embeddings (b_{e1}, b_{e2}) through element-wise product and difference. Further, we obtain context embedding (C) through $maxpool$ operation over the word embeddings sequence to model contextual cues. While the context provides important cues for identifying coreferential event mentions, it may not always be relevant for resolving coreference links. For instance, many event trigger word pairs such as (*"injuries"*, *"recommended"*) are extremely unlikely to exhibit coreferential relations irrespective of their context. Therefore, we use the similarity between event embeddings to control the context input and use them only in the scenarios where event trigger words are likely to possess coreferential link. To achieve so, we apply a linear neural layer over element-wise product and differences of two event mention embeddings followed by the $sigmoid$ activation, and multiply them with context embedding C. Finally, we concatenate the resulting set of embeddings and then use a three-layer feed-forward neural network classifier to score the coreference likelihood. The exact formulation of the coreference classifier is described in Eq. 7.11.

$$
\begin{aligned}
(b_{w1}.b_{e1}.b_{e2}.b_{wn}) &= BERT[(w_1.e_1.e_2.w_n)] \in R^{n\times 1024} \\
C &= maxpool(b_{w1}, ., b_{e1}, ., b_{e2}, ., b_{wn}) \in R^{1024} \\
s_1 &= sigmoid(W_1^s(b_{w1} \odot b_{w2}) + b_1^s) \in R^{1024} \\
s_2 &= sigmoid(W_2^s(b_{w1} - b_{w2}) + b_2^s) \in R^{1024} \\
R &= [b_{w1} \odot b_{w2}; b_{w1} - b_{w2}; s_1 \odot C; s_2 \odot C] \in R^{4096} \\
\hat{y_i} &= W_3(gelu(W_2(gelu(W_3 R + b_3)) + b_2)) + b_3 \in R
\end{aligned}
\tag{7.11}
$$

We train the model using binary cross-entropy loss. During inference, we use the best-first clustering approach, where we select the antecedent having the highest pairwise coreference score

[8]We take 100 context words to the right and left of the first and second event trigger words respectively when the number of context words in between them exceeds 200.

based on the coreference classifier, to build event chains.

7.4 Experiments

7.4.1 Datasets and Evaluation Setup

We use the news documents from the KBP 2016 for validation, and use news documents from KBP 2017 and RED corpora as well as discussion forum documents from the KBP 2017 corpus to evaluate the usefulness of our acquired data. KBP 2016, KBP 2017 and RED corpora contain 85, 83, and 30 news documents respectively, and KBP 2017 has 84 discussion forum documents. KBP corpora have been widely used for evaluating in-document event coreference resolution systems. We further evaluate our models on the RED corpus to examine systems' performance across different event types. KBP 2016 and 2017 corpora are annotated using a subset of 20 subtypes from 38 subtypes used in KBP 2015. On the contrary, RED documents are comprehensively annotated with event coreference relations with no restriction on event types or subtypes, thus, allowing us to evaluate coreference resolution performance on a broad range of events. Besides, we evaluate the performance of models across text genres by evaluating our models trained with news articles on KBP 2017 discussion forum documents.

7.4.2 Implementation Details

We use an ensemble of multi-layer feed-forward neural network classifiers to identify event mentions [Choubey and Huang, 2017a] for both news and discussion forum documents in KBP 2017 corpus. For the RED corpus, we use gold event mentions as that event extraction system can identify events from only eight event types annotated in KBP 2015 corpus. The coreference classifier uses a three-layer feed-forward neural network with 1024-512-1 units for scoring coreference likelihood. Two single-neural layers, used to transform element-wise dot product and difference between two event embeddings used for controlling context input, use 1024 units each. All hidden activations are followed by dropout with the rate of 0.1 for regularization [Srivastava et al., 2014]. All models are trained using AdamW optimizer [Loshchilov and Hutter, 2017, Kingma and Ba, 2014] with four different learning rates (1e-4, 5e-5, 1e-5, 5e-6) and for maximum of 100,000 up-

dates. We use the batch size of 16 and evaluate the model after every 5,000 steps. The epoch and learning rate yielding the best validation performance, average F1 score on KBP 2016 news documents, are used to obtain the final model. The BERT model is kept fixed during the training. All experiments are performed on NVIDIA GTX 2080 Ti 11GB using PyTorch 1.2.0+cu92 [Paszke et al., 2019] and HuggingFace Transformer libraries [Wolf et al., 2020].

7.4.3 Baseline Systems

Trigger Match (+Paraphrase): It links event mentions with the same trigger word (or are lexical paraphrases) as coreferential. Trigger match is a strong baseline for event coreference resolution.

DP-ILP: The full-model from chapter 6 that models correlations between event coreference chains and document topic structures.

7.4.4 Our Systems

KBP 2015, Paraphrase-based pairs, Post-Filtering Paraphrase pairs and KBP 2015+Post-Filtering Paraphrase pairs: The mention pair model, proposed in § 7.3, trained on different combinations of acquired and human-annotated datasets. *KBP 2015* is trained on event pairs from news documents in the KBP 2015 corpus. *Paraphrase-based pairs* is trained on paraphrase event pairs without rules-based filtering (§7.1.1). *Post-Filtering Paraphrase pairs* is trained on paraphrase event pairs that are filtered using rules defined over news discourse structure (§7.1.2). *KBP 2015+Post-Filtering Paraphrase pairs* is trained on aggregation of KBP 2015 and Post-Filtering Paraphrase event pairs.

Student Training: The mention pair model trained using the recently proposed self-training approach with a student network [Xie et al., 2020]. We first train a teacher mention pair model on the KBP 2015 corpus, then use the teacher model to annotate samples from unannotated news articles. We use the same set of event pairs from Xinhua articles in the Gigaword corpus, set the same upper bound of 20 coreferential and 200 non-coreferential pairs per event trigger word. Also, to allow fair comparisons, we selected only high scoring event pairs (likelihood ≥ 0.9) and collected 11,390 coreferential and 272,083 non-coreferential pairs. Finally, we train a new student network with the

combined KBP 2015 and teacher-annotated event pairs.

Masked Training: The mention pair model trained on all annotated and automatically acquired (or teacher annotated in case of student training model) event pairs. However, to limit the over-dependence on lexical features[9], we replace both the event trigger words with the *[MASK]* token for all acquired event pairs. Annotated event pairs from KBP 2015 are left unchanged.

7.4.5 Results and Analysis

The first segment in Table 7.3 shows the results for all models on KBP 2017 news articles corpus. The mention-pair model trained on KBP 2015 corpus using pre-trained language model and larger event context outperforms both local feature-based as well as the discourse-structure aware previous model, outperforming DP-ILP model by 0.94 points in average F1 score. The improvement is consistent across all metrics. Specifically, the used mention pair model gains MUC F1 score by 9.76 and 2.77 points over feature-based and discourse-aware systems, indicating that BERT-based embedding is more effective in resolving coreference links without exclusively modeling event-arguments or discourse-related features. The model trained on event pairs acquired following the proposed automatic strategy performs comparably to DP-ILP model, obtaining 0.68 points higher on MUC F1 than the DP-ILP. However, this model does worse than the equivalent model trained on KBP 2015 data, which can be explained by the related distribution of KBP 2015 and KBP 2017 datasets. Overall, training the model on KBP 2015 data combined with the acquired event pairs performs the best, outperforming both models trained on KBP 2015 only and the one trained with student training by 1.04 and 0.14 points respectively.

As shown in the second segment of Table 7.3, the improvement in the average F1 of the model trained on KBP 2015 over the trigger match baseline reduces to 2.3 points on the RED news articles corpus, compared to 5.69 points on KBP 2017 news articles. Mainly, RED annotates all event types while KBP has only 8 event types, and the change in event domains affects the overall performance gain of a model. The model trained on our Post-Filtering Paraphrase event pairs performs similarly to the one trained on KBP 2015, implying that the former generalizes similarly to the model trained

[9]All acquired event pairs are either synonyms or exhibit hypernym or hyponym relations

Model	b^3_{F1}	$ceafe_{F1}$	muc_R	muc_P	muc_{F1}	$blanc_{F1}$	AVG_{F1}
KBP 2017 News Articles							
Trigger Match	48.96	45.67	26.16	36.63	30.52	29.30	38.61
Trigger Match+Paraphrase	48.92	45.35	27.36	36.41	31.25	29.83	38.84
Feature-based Classifier	50.24	48.47	-	-	30.81	29.94	39.87
DP-ILP	51.68	50.57	-	-	37.8	33.39	43.36
KBP 2015	51.57	**50.90**	33.91	50.49	40.57	34.15	44.30
Paraphrase-based pairs	48.10	42.36	38.05	37.01	37.52	31.64	39.91
Post-Filtering PP	50.94	47.81	31.77	48.77	38.48	33.19	42.60
KBP 2015+Post-Filtering PP	**52.29**	50.50	35.24	**55.23**	43.03	35.53	45.34
Masked Training	52.10	50.72	36.31	53.02	43.10	35.51	**45.36**
Student Training	51.85	49.91	**38.18**	49.73	**43.20**	**35.83**	45.20
Masked Training	51.91	50.12	37.11	50.82	42.90	35.50	45.11
RED News Articles							
Trigger Match	88.07	84.21	42.63	35.14	38.52	64.34	68.78
Trigger Match+Paraphrase	87.18	83.09	47.16	33.87	39.43	64.88	68.65
KBP 2015	88.33	85.48	52.38	39.08	44.76	65.77	71.08
Paraphrase-based pairs	82.01	76.74	**68.02**	30.39	42.01	63.09	65.96
Post-Filtering PP	**89.25**	86.70	47.39	41.63	44.32	63.75	71.0
KBP 2015+Post-Filtering PP	**89.25**	**86.96**	56.00	**43.40**	48.91	66.74	72.96
Masked Training	89.16	86.90	58.04	43.31	**49.61**	**67.50**	**73.29**
Student Training	87.91	84.95	59.18	39.30	47.23	66.70	71.70
Masked Training	88.11	84.92	58.50	39.69	47.29	67.44	71.94
KBP 2017 Discussion Forum Documents							
Trigger Match	37.29	39.15	20.36	19.06	19.69	18.25	28.59
Trigger Match + Paraphrase	36.94	38.52	21.26	19.13	20.14	18.14	28.44
KBP 2015	38.11	38.67	25.33	**23.76**	24.52	**20.10**	**30.35**
Paraphrase-based pairs	35.58	34.30	28.65	21.37	24.48	19.19	28.39
Post-Filtering PP	**39.12**	**41.52**	17.34	20.75	18.89	18.81	29.59
KBP 2015+Post-Filtering PP	37.43	38.16	26.24	22.27	24.09	20.01	29.92
Masked Training	38.33	39.64	21.71	20.68	21.19	19.19	29.59
Student Training	36.80	36.68	**28.80**	22.68	**25.38**	20.08	29.73
Masked Training	37.06	38.01	22.77	20.00	21.29	17.51	28.47

Table 7.3: Results for event coreference resolution systems on the KBP 2017 and RED corpora. Feature-based classifier results are directly taken from Choubey and Huang [2018]. The results are statistically significant using bootstrap and permutation test [Dror et al., 2018] with $p<0.01$ between *Post-Filtering Paraphrase pairs* and *Paraphrase-based Pairs* and $p<0.002$ between *KBP 2015+Post-Filtering Paraphrase pairs+Masked Training* and *KBP 2015* models on both KBP 2017 and RED news articles test sets. Further, results for *KBP 2015+Post-Filtering Paraphrase pairs+Masked Training* are statistically significant compared to both *Student Training* and *Student Training+Masked Training* with $p<0.002$ on the RED news test set. Reprinted with permission from Choubey and Huang [2021].

on human-annotated data when applied to new data out of the training data distribution. Similar to the performance gain on KBP 2017 news articles, combining both KBP 2015 and acquired event pairs improves the average F1 on RED news articles, achieving the highest average F1 gain of 3.98 points against the trigger match baseline. Note that student training also improves performance on RED news articles. However, it is 1.26 points lower on average F1 score than the *KBP 2015+Post-Filtering Paraphrase pairs* model.

In the third segment of Table 7.3, we compare the performance of all models on a different text genre by evaluating them on the discussion forum documents from the KBP 2017 corpus. With shared event types, the model trained on KBP 2015 achieves the best result with 1.76 points improvement in the average F1 score over the lemma match baseline. The model trained using acquired event pairs, *Post-Filtering Paraphrase pairs*, achieves performance comparable to the model trained on KBP 2015. However, combining the KBP 2015 data with acquired event pairs (the model *KBP 2015+Post-Filtering Paraphrase pairs*) does not further improve the performance. Overall, we observe that none of the models obtain substantial performance improvement. The smaller improvements for all models on discussion forum documents, with the increased data size, also indicate the need for specialized learning algorithms to build a model that can generalize to a new text genre.

Post-Filtering Paraphrase Filtering and Masked Training: The model trained on Post-Filtering Paraphrase event pairs outperforms the one trained on paraphrase-based pairs by 2.69 and 5.04 average F1 points on KBP 2017 and RED news articles test sets respectively. Using news discourse structure-based rules to first constrain coreferential event paraphrase pairs within main sentences or headline and then add non-coreferential event paraphrase pairs from historical sentences inhibits the model from exclusively relying on lexical features. Further, masked training helps to completely circumvent any bias induced in a model by limiting coreferential event pairs to lexical paraphrases, which slightly improved the average F1 score.

Distributional Analysis of Predicted Coreferential Event Pairs across different Discourse Content Type Pairs: Finally, we analyze the distribution of predicted coreferential event pairs

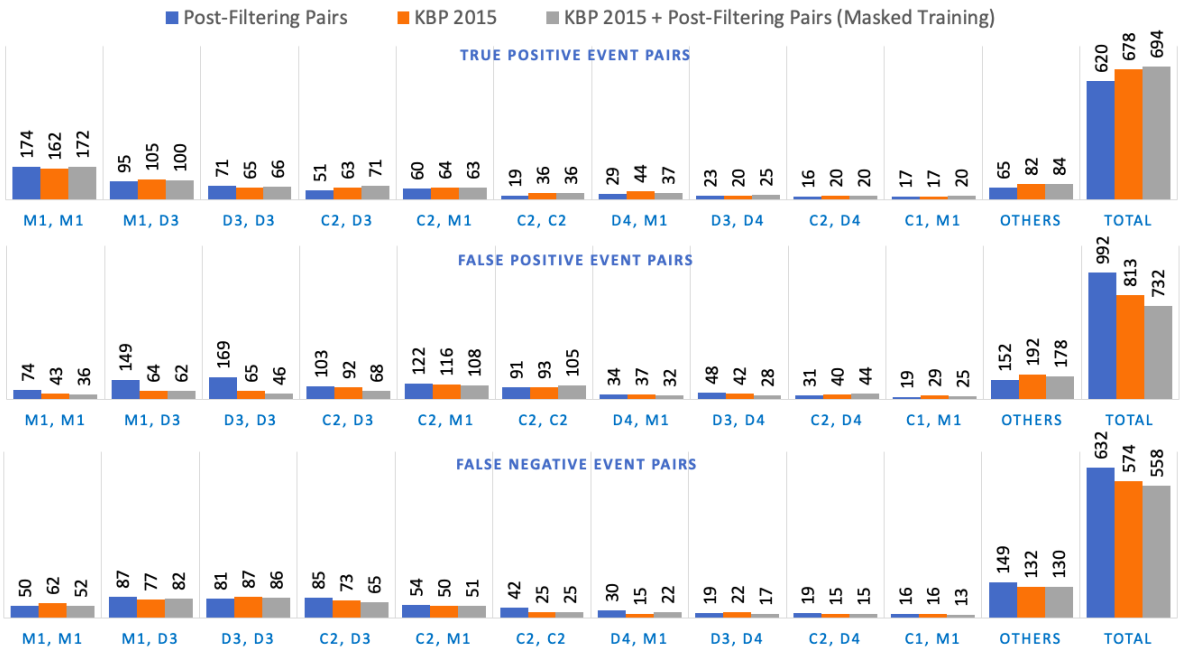

Figure 7.2: Distributions of predicted coreferential event pairs across different discourse content type pairs. Reprinted with permission from Choubey and Huang [2021].

across sentence pairs with different discourse content types on the validation dataset. We use the gold coreferential event pairs to identify the top 10 content type pairs of sentences that most frequently contain coreferential event mention pairs. Then, for the models trained on KBP 2015, Post-Filtering Paraphrase pairs and their combination with masked training, we report true-positive, false-positive, and false-negative predictions, shown in Figure 7.2. To ensure uniformity with rules used in §7.1.2, we merge the headline with main sentences.

Contrary to the rule that exclusively acquires coreferential event pairs from main sentences or headline, the classifier trained on acquired event pairs predicts coreferential event pairs across all discourse content type pairs. Notably, the Post-Filtering Pairs model predicted a comparable number of coreferential event pairs, 248, 244 and 240, in the (M1, M1), (M1, D3) and (D3, D3) content type pairs respectively. However, the number of true positives in (M1, M1) content pair is more than twice the number in either of the (M1, D3) or (D3, D3). This is expected given that the distribution of gold coreferential event pairs is normally skewed towards (M1, M1).

In comparison, models trained on KBP 2015 or combined KBP 2015 and Post-Filtering pairs

have lower false-positives while exhibiting similar distributions for true-positive predictions. Intuitively, despite second phase bootstrapping to include non-coreferential paraphrase pairs, the model trained solely on acquired event pairs focuses on lexical features more than the model trained on human-annotated corpus. On the other hand, masked training effectively overcomes excessive reliance on lexical cues and helps achieve a higher true positive rate without increasing false positives.

7.5 Conclusions and Future Work

We presented an automatic data acquisition strategy for event coreference resolution by mining the news discourse profiling structure. We performed both qualitative and empirical studies to determine the effectiveness of our proposed strategy. We found that the model trained on automatically acquired event pairs performs similarly to the model trained on human-annotated corpus when evaluated on the test set covering general event domains. Further, augmenting acquired event pairs to existing human-annotated data improves the performance of the model on both training-domain and broader domain test sets.

8. CONCLUSION AND FUTURE WORK

8.1 Research Summary

In this dissertation, through focused studies on news articles, we observed that discourse structures can be used to address the problems arising from both distributional and annotational sparsity of event coreference relations. The traditional approach of directly extracting and using fine-grained event information such as its temporal and spatial positions for event coreference resolution suffers from the problem of distributional sparsity. Knowing that the distributional sparsity arises by design where a concise and coherent discourse tends to avoid repetition of lesser relevant details when they are deducible from the context, we show that discourse cues can partially make up for the missing information. Secondly, to address the annotational sparsity, we have seen that augmenting news discourse profiling structure-based rules to the lemma match baseline can be used to automatically acquire event coreference resolution dataset that, when used to train a pairwise classifier, easily outperforms the lemma match model. Further, the acquired event coreference dataset, when used to augment existing human-annotated corpus to train the pairwise event coreference classifier, is empirically shown useful for improving event coreference resolution performance for news articles. When evaluated on discussion forum documents, we observe that acquired event pairs are not immediately useful, and it may require new training algorithms to improve the generalization capability of models trained on acquired event pairs on a new text genre. In § 8.2.1, I discuss future experiments to assess generalization capabilities of new models trained on acquired event pairs.

While most experiments in this dissertation are limited to the news genre, I believe that event is one of the most fundamental discourse units and genre-specific discourse structure will be beneficial for event coreference resolution for any new text genre. In § 8.2.2, I discuss a known theoretical discourse structure that may benefit event coreference resolution for discussion forum documents. Nonetheless, for most text genres, plenty of theoretical research on discourse structure has been

done. Through careful analyses, one should be able to identify an appropriate discourse structure that can benefit event coreference resolution.

Besides, one of the most significant contributions of this dissertation is the newly introduced task and dataset for news discourse profiling around the main event. I believe that news discourse profiling can benefit many challenging downstream NLP applications. For instance, it can help in building a comprehensive document-level event graph that identifies relation between every pair of event mentions. Secondly, news discourse profiling can benefit NLP applications that require sentence-level relevance ordering, such as document summarization or simplification. In sections 8.2.3, I briefly discuss potential approaches to leverage news discourse profiling for building document-level event graph and text summarization respectively.

8.2 Future Directions

8.2.1 Genre Adaptive Event Coreference Resolution System

In chapter 7, we observed that both human-annotated and automatically acquired datasets for event coreference resolution are only marginally effective on discussion forum documents. Unlike well-structured expository texts (e.g. news articles), discussion forum documents are hybrid of both expository and conversational styles. Secondly, discussion forums are asynchronously written by multiple participants resulting in highly non-linear content organization, very different from news discourse. Therefore, I believe that a pairwise classifier trained on news documents may only be able to use local event features when used on discussion forum documents. In addition, local features may not necessarily contribute proportionally to the final prediction for news and discussion forum documents.

There are several known ways that can be explored to leverage acquired event pairs for improving performance on discussion forum documents or any new text genre. One can try unsupervised text style transfer [Yang et al., 2018] to rephrase acquired event pairs to a new text genre style and use them to train the final model. Alternatively, one can use adversarial learning to learn domain invariant features [Ganin et al., 2016] on acquired event pairs followed by fine-tuning on human-

annotated data from the new genre, or use knowledge distillation techniques [Hinton et al., 2015] to train a student network that distills knowledge from multiple teacher models [Currey et al., 2020] trained on our acquired event pairs, genre-specific human-annotated data, and human-annotated data from other genres.

8.2.2 Discourse Act Categorization for Event Coreference Resolution in Discussion Forum

As a preliminary approach to incorporate discourse cues for resolving event coreference relations in discussion forum documents, one could investigate whether separating expository components from each comment and separately modeling its discourse structure helps in resolving local (comment-level) event coreference relations. A more systematic method could be to directly analyze the association between known discourse structure and event coreference relations. For instance, Zhang et al. [2017] proposed categorization of comments in discussion thread into predefined coarse discourse acts[1]. In addition, Zhang et al. [2017] proposed to potentially link some discourse acts with another (e.g. *answer* act relates to *question* act, *announcement* act is not related to any other act, etc.). The resulting discourse-act structure has the potential to benefit the event coreference resolution task. As an example, related *answer* and *question* acts are likely to contain coreferential event mentions. On the contrary, *announcement* comments may introduce new events that are not coreferential to any event from the preceding comments.

8.2.3 Downstream Applications of News Discourse Profiling

8.2.3.1 News Discourse Profiling and Event Relations

Intuitively, news discourse profiling can help in identifying other inter-sentence event relations, such as temporal and causal relations, and thus disentangling complete event structures. For instance, events occurring in *previous event* sentences are probable cause for the main event which in turn causes events in *Consequence* sentences (the same rationale can be applied for temporal order). Besides possible applications in identifying pair-wise event temporal and causal relations, one can leverage discourse profiling when building complex temporal dependency structures such

[1]The discourse act categories include question, answer, announcement, agreement, appreciation, disagreement, negative reaction, elaboration, and humor.

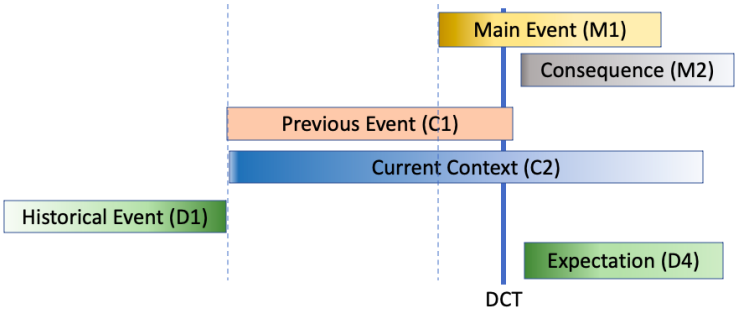

Figure 8.1: Temporal structures induced by different content types from the news discourse profiling. DCT: Document Creation Time.

as graph [Yao et al., 2020] or tree [Zhang and Xue, 2018]. For temporal dependency graph or tree, each event and time expression is referenced to only one time expression (or additionally an event). Thus, identified temporal relations represent the most salient relations that can potentially be used to infer additional temporal relations through transitivity or commonsense reasoning [Yao et al., 2020]. This makes identifying reference time expression and event more challenging, especially when they are mentioned across sentences, and one could explore discourse-level temporal cues for the same. Specifically, as shown in Figure 8.1, news discourse profiling induces different time frames relevant to a news story. For instance, mentions in historical sentences have temporal adjacency with other mentions in the historical sentences but are likely to be distant from mentions in other content types. Similarly, mentions in previous event sentences may have temporal adjacency with mentions from previous event, main event, or current-context sentences but are likely separated from mentions in historical, expectation, or consequence sentences. The induced time frames can help to locate reference time expressions or reference events that are mentioned across sentences.

8.2.3.2 News Discourse Profiling and Text Summarization

Given that content types are roughly ordered based on their relevance to the main news event, one can use them to perform text summarization. For instance, oracle summaries for extractive summarization generally contain the most relevant sentences while minimizing the repetitions.

Knowing that the main contents are the most relevant followed by context informing and distantly related, we can first pick the first main event sentence from a document as a summary. Consecutively, we can pick the first context informing sentence as the second sentence in summary followed by the first sentence from a distantly related content type. Apart from this simple baseline, one can also explore recent deep neural networks or reinforcement learning techniques to incorporate news discourse profiling in a text summarization system. Recently, Chen and Bansal [2018] proposed a policy network (extractor) to extract few salient sentences from a document followed by an abstractor network to generate summaries. We can explore some known or new techniques, such as knowledge distillation [Hinton et al., 2015], to use a discourse profiling model to guide the training of the extractor policy network.

REFERENCES

C. Adrian Bejan and S. Harabagiu. Unsupervised event coreference resolution. *Computational Linguistics*, 40(2):311–347, June 2014. doi: 10.1162/COLI_a_00174. URL https://www.aclweb.org/anthology/J14-2004.

J. Allan, J. G. Carbonell, G. Doddington, J. Yamron, and Y. Yang. Topic detection and tracking pilot study final report. Jun 2018. doi: 10.1184/R1/6626252.v1. URL https://kilthub.cmu.edu/articles/journal_contribution/Topic_Detection_and_Tracking_Pilot_Study_Final_Report/6626252/1.

J. Araki, Z. Liu, E. Hovy, and T. Mitamura. Detecting subevent structure for event coreference resolution. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/963_Paper.pdf.

S. Arora, Y. Li, Y. Liang, T. Ma, and A. Risteski. A latent variable model approach to PMI-based word embeddings. *Transactions of the Association for Computational Linguistics*, 4:385–399, 2016. doi: 10.1162/tacl_a_00106. URL https://www.aclweb.org/anthology/Q16-1028.

A. Badgett and R. Huang. Extracting subevents via an effective two-phase approach. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 906–911, Austin, Texas, Nov. 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1088. URL https://www.aclweb.org/anthology/D16-1088.

A. Bagga and B. Baldwin. Entity-based cross-document coreferencing using the vector space model. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 79–85, Montreal, Quebec, Canada, Aug. 1998. Association for Computational Linguistics. doi: 10.3115/980845.980859. URL https://www.aclweb.org/anthology/P98-1012.

D. Baiamonte, T. Caselli, and I. Prodanof. Annotating content zones in news articles. In *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*, 2016. Third Italian Conference on Computational Linguistics (CLiC-it 2016) ; Conference date: 01-01-2016 Through 01-01-2016.

S. Barhom, V. Shwartz, A. Eirew, M. Bugert, N. Reimers, and I. Dagan. Revisiting joint modeling of cross-document entity and event coreference resolution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4179–4189, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1409. URL https://www.aclweb.org/anthology/P19-1409.

M. Becker, M. Staniek, V. Nastase, A. Palmer, and A. Frank. Classifying semantic clause types: Modeling context and genre characteristics with recurrent neural networks and attention. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 230–240, Vancouver, Canada, Aug. 2017. Association for Computational Linguistics. doi: 10.18653/v1/S17-1027. URL https://www.aclweb.org/anthology/S17-1027.

C. Bejan and S. Harabagiu. Unsupervised event coreference resolution with rich linguistic features. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1412–1422, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P10-1143.

A. Bell. The discourse structure of news stories. In *Approaches to media discourse / edited by Allan Bell and Peter Garrett.*, chapter 3, pages 64–104. Blackwell, Oxford, 1998.

D. M. Bikel and V. Castelli. Event matching using the transitive closure of dependency relations. In *Proceedings of ACL-08: HLT, Short Papers*, pages 145–148, Columbus, Ohio, June 2008. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P08-2037.

S. Bird, E. Klein, and E. Loper. *Natural Language Processing with Python*. O'Reilly Media, Inc., 1st edition, 2009. ISBN 0596516495.

S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1):107–117, 1998. ISSN 0169-7552. doi: https://doi.org/10.1016/S0169-7552(98)00110-X. URL https://www.sciencedirect.com/science/article/pii/S016975529800110X. Proceedings of the Seventh International World Wide Web Conference.

C. Chen and V. Ng. SinoCoreferencer: An end-to-end Chinese event coreference resolver. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4532–4538, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/1144_Paper.pdf.

C. Chen and V. Ng. Joint inference over a lightly supervised information extraction pipeline: Towards event coreference resolution for resource-scarce languages. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 2913–2920. AAAI Press, 2016.

Y.-C. Chen and M. Bansal. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–686, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1063. URL https://www.aclweb.org/anthology/P18-1063.

Z. Chen and H. Ji. Graph-based event coreference resolution. In *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing (TextGraphs-4)*, pages 54–57, Suntec, Singapore, Aug. 2009. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W09-3208.

Z. Chen, H. Ji, and R. Haralick. A pairwise event coreference model, feature impact and evaluation for event coreference resolution. In *Proceedings of the Workshop on Events in Emerging Text Types*, pages 17–22, Borovets, Bulgaria, Sept. 2009. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W09-4303.

W. S. Cho, P. Zhang, Y. Zhang, X. Li, M. Galley, C. Brockett, M. Wang, and J. Gao. Towards

coherent and cohesive long-form text generation. In *Proceedings of the First Workshop on Narrative Understanding*, pages 1–11, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-2401. URL https://www.aclweb.org/anthology/W19-2401.

P. K. Choubey and R. Huang. Event coreference resolution by iteratively unfolding inter-dependencies among events. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2124–2133, 2017a.

P. K. Choubey and R. Huang. TAMU at KBP 2017: Event nugget detection and coreference resolution. *CoRR*, abs/1711.02162, 2017b. URL http://arxiv.org/abs/1711.02162.

P. K. Choubey and R. Huang. Improving event coreference resolution by modeling correlations between event coreference chains and document topic structures. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 485–495, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1045. URL https://www.aclweb.org/anthology/P18-1045.

P. K. Choubey and R. Huang. Automatic data acquisition for event coreference resolution. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1185–1196, Online, Apr. 2021. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2021.eacl-main.101.

P. K. Choubey, K. Raju, and R. Huang. Identifying the most dominant event in a news article by mining event coreference relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 340–345, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2055. URL https://www.aclweb.org/anthology/N18-2055.

P. K. Choubey, A. Lee, R. Huang, and L. Wang. Discourse as a function of event: Profiling discourse structure in news articles around the main event. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5374–5386, Online, July

2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.478. URL https://www.aclweb.org/anthology/2020.acl-main.478.

J. Cohen. Multiple regression as a general data-analytic system. *Psychological Bulletin*, 70:426–443, 1968.

A. Currey, P. Mathur, and G. Dinu. Distilling multiple domains for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4500–4511, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.364. URL https://www.aclweb.org/anthology/2020.emnlp-main.364.

A. Cybulska and P. Vossen. Using a sledgehammer to crack a nut? lexical diversity and event coreference resolution. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4545–4552, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/840_Paper.pdf.

Z. Dai and R. Huang. Building context-aware clause representations for situation entity type classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3305–3315, Brussels, Belgium, Oct.-Nov. 2018a. Association for Computational Linguistics. doi: 10.18653/v1/D18-1368. URL https://www.aclweb.org/anthology/D18-1368.

Z. Dai and R. Huang. Improving implicit discourse relation classification by modeling interdependencies of discourse units in a paragraph. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 141–151, New Orleans, Louisiana, June 2018b. Association for Computational Linguistics. doi: 10.18653/v1/N18-1013. URL https://www.aclweb.org/anthology/N18-1013.

M.-C. de Marneffe, A. N. Rafferty, and C. D. Manning. Finding contradictions in text. In *Proceedings of ACL-08: HLT*, pages 1039–1047, Columbus, Ohio, June 2008. Association for Compu-

tational Linguistics. URL https://www.aclweb.org/anthology/P08-1118.

N. Decker. The use of syntactic clues in discourse processing. In *23rd Annual Meeting of the Association for Computational Linguistics*, pages 315–323, Chicago, Illinois, USA, July 1985. Association for Computational Linguistics. doi: 10.3115/981210.981249. URL https://www.aclweb.org/anthology/P85-1039.

P. Denis and J. Baldridge. Joint determination of anaphoricity and coreference resolution using integer programming. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 236–243, Rochester, New York, Apr. 2007. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N07-1030.

P. Denis and J. Baldridge. Global joint models for coreference resolution and named entity classification. *Procesamiento del lenguaje natural*, 42, 2009.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

R. Dror, G. Baumer, S. Shlomov, and R. Reichart. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1128. URL https://www.aclweb.org/anthology/P18-1128.

J. Ellis, J. Getman, D. Fore, N. Kuster, Z. Song, A. Bies, and S. Strassel. Overview of linguistic resources for the tac kbp 2015 evaluations: Methodologies and results. In *Proceedings of TAC KBP 2015 Workshop, National Institute of Standards and Technology*, 2015.

J. Ellis, J. Getman, N. Kuster, Z. Song, A. Bies, and S. Strassel. Overview of linguistic resources

for the tac kbp 2016 evaluations: Methodologies and results. In *Proceedings of TAC KBP 2016 Workshop, National Institute of Standards and Technology*, 2016.

S. Evan. The new york times annotated corpus. *LDC2008T19. DVD. Philadelphia: Linguistic Data Consortium.*, 2008.

J. R. Finkel and C. D. Manning. Enforcing transitivity in coreference resolution. In *Proceedings of ACL-08: HLT, Short Papers*, pages 45–48, Columbus, Ohio, June 2008. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P08-2012.

A. Friedrich and A. Palmer. Situation entity annotation. In *Proceedings of LAW VIII-The 8th Linguistic Annotation Workshop*, pages 149–158, 2014.

F. Galton. Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246–263, 1886.

Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17(1): 2096–2030, Jan. 2016. ISSN 1532-4435.

J. Ganitkevitch, B. Van Durme, and C. Callison-Burch. PPDB: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 758–764, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N13-1092.

L. Gao, P. K. Choubey, and R. Huang. Modeling document-level causal structures for event causal relation identification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1808–1817, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1179. URL https://www.aclweb.org/anthology/N19-1179.

J. Getman, J. Ellis, Z. Song, J. Tracey, and S. Strassel. Overview of linguistic resources for the tac kbp 2017 evaluations: Methodologies and results. In *Proceedings of TAC KBP 2017 Workshop,*

National Institute of Standards and Technology, 2017.

J. E. Grimes. *The Thread of Discourse.* ERIC, 1972.

B. J. Grosz, A. K. Joshi, and S. Weinstein. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225, 1995. URL https://www.aclweb.org/anthology/J95-2003.

A. Haghighi and D. Klein. Unsupervised coreference resolution in a nonparametric Bayesian model. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 848–855, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P07-1107.

M. A. Hearst. Multi-paragraph segmentation expository text. In *32nd Annual Meeting of the Association for Computational Linguistics*, pages 9–16, Las Cruces, New Mexico, USA, June 1994. Association for Computational Linguistics. doi: 10.3115/981732.981734. URL https://www.aclweb.org/anthology/P94-1002.

G. Hinton, O. Vinyals, and J. Dean. Distilling the Knowledge in a Neural Network. *arXiv e-prints*, art. arXiv:1503.02531, Mar. 2015.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

K. Humphreys, R. Gaizauskas, and S. Azzam. Event coreference for information extraction. In *Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, 1997. URL https://www.aclweb.org/anthology/W97-1311.

H. Ji and R. Grishman. Knowledge base population: Successful approaches and challenges. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1148–1158, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P11-1115.

Y. Ji and J. Eisenstein. One vector is not enough: Entity-augmented distributed semantics for discourse relations. *Transactions of the Association for Computational Linguistics*, 3:329–344, 2015. doi: 10.1162/tacl_a_00142. URL https://www.aclweb.org/anthology/

Q15-1024.

D. Jurafsky and J. H. Martin. Speech and language processing: An introduction to speech recognition, computational linguistics and natural language processing. *Upper Saddle River, NJ: Prentice Hall*, 2008.

K. Kenyon-Dean, J. C. K. Cheung, and D. Precup. Resolving event coreference with supervised representation learning and clustering-oriented regularization. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 1–10, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/S18-2001. URL https://www.aclweb.org/anthology/S18-2001.

D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. *arXiv e-prints*, art. arXiv:1412.6980, Dec. 2014.

J. G. Kircz. Rhetorical structure of scientific articles: the case for argumentational analysis in information retrieval. *Journal of documentation*, 47(4):354–372, 1991.

V. R. Konda and J. N. Tsitsiklis. Actor-critic algorithms. In *Advances in neural information processing systems*, pages 1008–1014, 2000.

J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1558607781.

Q. Le and T. Mikolov. Distributed representations of sentences and documents. In E. P. Xing and T. Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1188–1196, Bejing, China, 22–24 Jun 2014. PMLR. URL http://proceedings.mlr.press/v32/le14.html.

H. Lee, M. Recasens, A. Chang, M. Surdeanu, and D. Jurafsky. Joint entity and event coreference resolution across documents. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 489–500, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL

https://www.aclweb.org/anthology/D12-1045.

K. Lee, L. He, M. Lewis, and L. Zettlemoyer. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1018. URL https://www.aclweb.org/anthology/D17-1018.

W. Li, M. Wu, Q. Lu, W. Xu, and C. Yuan. Extractive summarization using inter- and intra- event relevance. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, page 369–376, USA, 2006. Association for Computational Linguistics. doi: 10.3115/1220175.1220222. URL https://doi.org/10.3115/1220175.1220222.

M. Liakata, S. Saha, S. Dobnik, C. Batchelor, and D. Rebholz-Schuhmann. Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics*, 28(7):991–1000, 2012.

E. D. Liddy. The discourse-level structure of empirical abstracts: An exploratory study. *Information Processing & Management*, 27(1):55–81, 1991.

Z. Liu, J. Araki, E. Hovy, and T. Mitamura. Supervised within-document event coreference using information propagation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4539–4544, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/646_Paper.pdf.

I. Loshchilov and F. Hutter. Decoupled Weight Decay Regularization. *arXiv e-prints*, art. arXiv:1711.05101, Nov. 2017.

J. Lu and V. Ng. Learning antecedent structures for event coreference resolution. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 113–118, 2017. doi: 10.1109/ICMLA.2017.0-170.

J. Lu and V. Ng. Joint learning for event coreference resolution. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*,

pages 90–101, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1009. URL https://www.aclweb.org/anthology/P17-1009.

J. Lu and V. Ng. Event coreference resolution: A survey of two decades of research. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 5479–5486. International Joint Conferences on Artificial Intelligence Organization, 7 2018. doi: 10.24963/ijcai.2018/773. URL https://doi.org/10.24963/ijcai.2018/773.

J. Lu, D. Venugopal, V. Gogate, and V. Ng. Joint inference for event coreference resolution. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3264–3275, Osaka, Japan, Dec. 2016. The COLING 2016 Organizing Committee. URL https://www.aclweb.org/anthology/C16-1308.

X. Luo. On coreference resolution performance metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 25–32, Vancouver, British Columbia, Canada, Oct. 2005. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/H05-1004.

W. C. Mann and S. A. Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281, 1988.

C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60, 2014. URL http://www.aclweb.org/anthology/P/P14/P14-5010.

K. McConky, R. Nagi, M. Sudit, and W. Hughes. Improving event co-reference by context extraction and dynamic feature weighting. In *2012 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support*, pages 38–43. IEEE, 2012.

Y. Meged, A. Caciularu, V. Shwartz, and I. Dagan. Paraphrasing vs coreferring: Two sides of the same coin. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4897–4907, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.440. URL https://www.aclweb.org/anthology/

2020.findings-emnlp.440.

R. Mihalcea and P. Tarau. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W04-3252.

T. Mitamura, Z. Liu, and E. Hovy. Overview of tac kbp 2015 event nugget track. In *Text Analysis Conference*, 2015.

T. Mitamura, Z. Liu, and E. Hovy. Events detection, coreference and sequencing: What's next? overview of the tac kbp 2017 event track. In *Proceedings of TAC KBP 2017 Workshop, National Institute of Standards and Technology*, 2017.

S. Mitchell, M. OSullivan, and I. Dunning. Pulp: a linear programming toolkit for python. *The University of Auckland, Auckland, New Zealand, http://www. optimization-online. org/DB_FILE/2011/09/3178. pdf*, 2011.

Y. Mizuta, A. Korhonen, T. Mullen, and N. Collier. Zone analysis in biology articles as a basis for information extraction. *International journal of medical informatics*, 75(6):468–487, 2006.

N. S. Moosavi and M. Strube. Which coreference evaluation metric do you trust? a proposal for a link-based entity aware metric. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 632–642, Berlin, Germany, Aug. 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1060. URL https://www.aclweb.org/anthology/P16-1060.

C. Napoles, M. Gormley, and B. Van Durme. Annotated Gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction (AKBC-WEKEX)*, pages 95–100, Montréal, Canada, June 2012. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W12-3018.

S. Narayanan and S. Harabagiu. Question answering based on semantic structures. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 693–701, Geneva, Switzerland, aug 23–aug 27 2004. COLING. URL https:

//www.aclweb.org/anthology/C04-1100.

T. O'Gorman, K. Wright-Bettner, and M. Palmer. Richer event description: Integrating event coreference with temporal, causal and bridging annotation. *Computing News Storylines*, page 47, 2016.

Z. Pan and G. M. Kosicki. Framing analysis: An approach to news discourse. *Political communication*, 10(1):55–75, 1993.

A. Pandian, L. Mulaffer, K. Oflazer, and A. AlZeyara. Precision event coreference resolution using neural network classifiers. *Computación y Sistemas*, 24(1), 2020.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

E. Pavlick, P. Rastogi, J. Ganitkevitch, B. Van Durme, and C. Callison-Burch. PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 425–430, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-2070. URL https://www.aclweb.org/anthology/P15-2070.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

H. Peng, Y. Song, and D. Roth. Event detection and co-reference with minimal supervision. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 392–402, Austin, Texas, Nov. 2016. Association for Computational Linguistics. doi:

10.18653/v1/D16-1038. URL https://www.aclweb.org/anthology/D16-1038.

J. Pennington, R. Socher, and C. Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, Oct. 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL https://www.aclweb.org/anthology/D14-1162.

M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL https://www.aclweb.org/anthology/N18-1202.

H. Po¨ ttker. News and its communicative quality: The inverted pyramid—when and why did it appear? *Journalism Studies*, 4(4):501–511, 2003.

S. Pradhan, X. Luo, M. Recasens, E. Hovy, V. Ng, and M. Strube. Scoring coreference partitions of predicted mentions: A reference implementation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 30–35, Baltimore, Maryland, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-2006. URL https://www.aclweb.org/anthology/P14-2006.

R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and B. Webber. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May 2008. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2008/pdf/754_paper.pdf.

M. Recasens and E. Hovy. Blanc: Implementing the rand index for coreference evaluation. *Natural Language Engineering*, 17(4):485–510, 2011.

R. Řehůřek and P. Sojka. Software Framework for Topic Modelling with Large Corpora. In

Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, pages 45–50, Valletta, Malta, May 2010. ELRA. http://is.muni.cz/publication/884893/en.

H. Sahlani, M. Hourali, and B. Minaei-Bidgoli. Coreference resolution using semantic features and fully connected neural network in the persian language. *International Journal of Computational Intelligence Systems*, 13:1002–1013, 2020. ISSN 1875-6883. doi: https://doi.org/10.2991/ijcis.d.200706.002. URL https://doi.org/10.2991/ijcis.d.200706.002.

S. Sangeetha and M. Arock. Event coreference resolution using mincut based graph clustering. In *Proceedings of the Fourth International Workshop on Computer Networks & Communications*, pages 253–260, 2012.

Z. Song, A. Bies, S. Strassel, T. Riese, J. Mott, J. Ellis, J. Wright, S. Kulick, N. Ryant, and X. Ma. From light to rich ere: annotation of entities, relations, and events. In *Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 89–98, 2015.

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

S. Teufel, J. Carletta, and M. Moens. An annotation scheme for discourse-level argumentation in research articles. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 110–117, Bergen, Norway, June 1999. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/E99-1015.

V. D. Teun A. News schemata. *Studying writing: linguistic approaches*, 1:155–186, 1986.

A. Todirascu, T. François, D. Bernhard, N. Gala, and A.-L. Ligozat. Are cohesive features relevant for text readability evaluation? In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 987–997, Osaka, Japan, Dec. 2016. The COLING 2016 Organizing Committee. URL https://www.aclweb.org/anthology/C16-1094.

T. A. Van Dijk. News analysis. *Case Studies of International and National News in the Press. New Jersey: Lawrence*, 1988a.

T. A. Van Dijk. *News as discourse*. Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc., 1988b.

V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995. ISBN 0-387-94559-8.

M. Vilain, J. Burger, J. Aberdeen, D. Connolly, and L. Hirschman. A model-theoretic coreference scoring scheme. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*, 1995. URL https://www.aclweb.org/anthology/M95-1005.

A. Waard, P. Buitelaar, and T. Eigner. Identifying the epistemic value of discourse segments in biology texts. *Proceedings of the Eighth International Conference on Computational Semantics:*, pages 351–354, 01 2009. doi: 10.3115/1693756.1693802.

C. Walker, M. Strassel, M. Julie, and Kazuaki. Ace 2005 multilingual training corpus. In *Linguistic Data Consortium, LDC Catalog No.: LDC2006T06.*, 2006.

W. J. Wilbur, A. Rzhetsky, and H. Shatkay. New directions in biomedical text annotation: definitions, guidelines and corpus construction. *BMC bioinformatics*, 7(1):356, 2006.

R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush. Transformers: State-of-the-art natural language processing. pages 38–45, Oct. 2020. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

B. Yang, C. Cardie, and P. Frazier. A hierarchical distance-dependent Bayesian model for event

coreference resolution. *Transactions of the Association for Computational Linguistics*, 3:517–528, 2015. doi: 10.1162/tacl_a_00155. URL https://www.aclweb.org/anthology/Q15-1037.

Z. Yang, Z. Hu, C. Dyer, E. P. Xing, and T. Berg-Kirkpatrick. Unsupervised text style transfer using language models as discriminators. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/398475c83b47075e8897a083e97eb9f0-Paper.pdf.

J. Yao, H. Qiu, B. Min, and N. Xue. Annotating Temporal Dependency Graphs via Crowdsourcing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5368–5380, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.432. URL https://www.aclweb.org/anthology/2020.emnlp-main.432.

W. V. Yarlott, C. Cornelio, T. Gao, and M. Finlayson. Identifying the discourse function of news article paragraphs. In *Proceedings of the Workshop Events and Stories in the News 2018*, pages 25–33, Santa Fe, New Mexico, U.S.A, Aug. 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W18-4304.

D. Yu, X. Pan, B. Zhang, L. Huang, D. Lu, S. Whitehead, and H. Ji. Rpi blender tac-kbp2016 system description. In *TAC*, 2016.

A. X. Zhang, B. Culbertson, and P. Paritosh. Characterizing online discussion using coarse discourse sequences. In *Proceedings of the 11th International AAAI Conference on Weblogs and Social Media*, ICWSM '17, 2017.

Y. Zhang and N. Xue. Structured interpretation of temporal relations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://www.aclweb.org/anthology/L18-1490.

X. Zuo, Y. Chen, K. Liu, and J. Zhao. Event co-reference resolution via a multi-loss neural net-

work without using argument information. *Science China Information Sciences*, 62(11), Oct 2019. ISSN 1869-1919. doi: 10.1007/s11432-018-9833-1. URL http://dx.doi.org/10.1007/s11432-018-9833-1.

APPENDIX A

ANNOTATION GUIDELINES FOR THE DISCOURSE PROFILING

A.1 General Rules

- We label each sentence in a news article based on its relevance in describing the main story.
- The first category Main story (M) covers sentences that describe the main story, including main events and their real consequences.
- The second category Context-informing contents (C) covers sentences that directly explain the context of the main story
- The third category Distantly-related contents (D) covers sentences that provide further supporting context for the main story
- Lastly, we also have a category for not applicable sentences (N).

A.2 Main Story (M)

A.2.1 Main Event (M1)

- Sentences describing main event that directly relates to the major subject of an article. Main event is the most recent event (trigger event) that gave rise to a news report. Generally, main event should have happened already. However, main event can be a projected event since some news articles focus on events that are to happen soon.
- Statements rephrasing the main event made by entities that are directly related to the main event.

A.2.2 Consequences (M2)

- Sentences describing the real consequences of main events that often happen right after main events and are due to main events.

A.3 Context-informing Content (C)

A.3.1 Previous Event (C1)

- Sentences describing previous events that are specific, have occurred recently and inform the cause of main events. Note that sentences describing events that acted as the pre-condition for main event also belong to this category.

A.3.2 Current Context (C2)

- Sentences describing any general circumstances that inform the cause of main events.
- Sentences describing actual situation in which the main event took place. They should have temporal co-occurrence with the main event or talk about the ongoing situation.

A.4 Distantly-related Content (D)

A.4.1 Historical Event (D1)

- Sentences describing previous events that have not occurred recently, at least 2 months prior to the main event.
- Sentences describing events that occurred in previous years with no specification of time elapsed.

A.4.2 Anecdotal Event (D2)

- Sentences describing events that are anecdotal, such events may happen before or after main events. Anecdotal events are specific events with specific participants that are uncertain (may happen in future) or can't be verified (happened in past).
- If the statement was made in a private discussion (verbal or written) and is unverifiable.

A.4.3 Evaluation (D3)

- Sentences that are explicitly an opinion and comment on any events in the story.

A.4.4 Expectations (D4)

- Sentences describing expectations about the resolution or possible consequences of any events in the future.

A.5 N/A (N)

- Sentence that's not a part of the news article. For instance, "slideshow (images) and captions", "editing by XYZ", "visit news.org", etc.