Discourse as a Function of Event: Profiling Discourse Structure in News Articles around the Main Event

Prafulla Kumar Choubey[1] **Aaron Lee**[1] **Ruihong Huang**[1] **Lu Wang**[2]
[1] Department of Computer Science and Engineering, Texas A&M University
[2] Khoury College of Computer Sciences, Northeastern University
(prafulla.choubey, aaronlee, huangrh)@tamu.edu
(luwang)@ccs.neu.edu

Abstract

Understanding discourse structures of news articles is vital to effectively contextualize the occurrence of a news event. To enable computational modeling of news structures, we apply an existing theory of functional discourse structure for news articles that revolves around the main event and create a human-annotated corpus of 802 documents spanning over four domains and three media sources. Next, we propose several document-level neural-network models to automatically construct news content structures. Finally, we demonstrate that incorporating system predicted news structures yields new state-of-the-art performance for event coreference resolution. The news documents we annotated are openly available and the annotations are publicly released for future research[1].

1 Introduction

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 paper, we focus on studying content structures specific to *news articles*, a broadly studied text genre for many NLP tasks and applications. 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 on computational modeling and automatic construction of news content structures. We introduce 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

[1]Dataset can be found at https://github.com/prafulla77/Discourse_Profiling

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5374–5386
July 5 - 10, 2020.

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.

We envision that the news discourse profiling dataset as well as the learnt computational systems are useful to many discourse level NLP tasks and applications. As an example, we analyze correlations between content structures and event coreference structures in news articles, and conduct experiments to incorporate system predicted sentence content types into an event coreference resolution system. Specifically, we analyze the lifespan and spread of event coreference chains over different content types, and design constraints to capture several prominent observations for event coreference resolution. Experimental results show that news discourse profiling enables consistent performance gains across all the evaluation metrics on two benchmark datasets, improving the previous best performance for the challenging task of event coreference resolution.

2 Related Work

Several well-studied discourse analysis tasks have been shown useful for many NLP applications. The RST (Mann and Thompson, 1988; Soricut and Marcu, 2003; Feng and Hirst, 2012; Ji and Eisenstein, 2014; Li et al., 2014a; Liu et al., 2019) and PDTB style (Prasad et al., 2008; Pitler and Nenkova, 2009; Lin et al., 2014; Rutherford and Xue, 2016; Qin et al., 2016; Xu et al., 2018) discourse parsing tasks identify discourse units that are logically connected with a predefined set of rhetorical relations, and have been shown useful for a range of NLP applications such as text quality assessment (Lin et al., 2011), sentiment analysis (Bhatia et al., 2015), text summarization (Louis et al., 2010), machine translation (Li et al., 2014b) and text categorization (Ji and Smith, 2017). Text segmentation (Hearst, 1994; Choi, 2000; Eisenstein and Barzilay, 2008; Koshorek et al., 2018) is another well studied discourse analysis task that aims to divide a text into a sequence of topically coherent segments and has been shown useful for text summarization (Barzilay and Lee, 2004), sentiment analysis (Sauper et al., 2010) and dialogue systems (Shi et al., 2019).

The news discourse profiling task is complementary to the well-established discourse analysis tasks and is likely to further benefit many NLP applications. First, it studies genre-specific discourse structures, while the aforementioned discourse analysis tasks study genre independent general discourse structures and thus fail to incorporate domain knowledge. Second, it focuses on understanding global content organization structures with the main event at the center, while the existing tasks focus on either understanding rhetorical aspects of discourse structures (RST and PDTB discourse parsing) or detecting shallow topic transition structures (text segmentation).

Genre-specific functional structures have been studied based on different attributes, but mostly for 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.

Past 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. Owing to the high specificity of the Dijk's theory, Yarlott et al. (2018) performed a pilot study for its computational feasibility and annotated a small dataset of 50 documents taken from the ACE Phase 2 corpus (Doddington et al., 2004). However, as mentioned in the paper, their annotators were given minimal training prior to annotations, consequently, the kappa inter-agreement (55%) between two annotators was not satisfactory. In addition, coverage of their annotated dataset on broad event domains and media sources was unclear. The only studies on functional structure of news article with sizable dataset 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.

3 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 topic. The original Van Dijk's theory was designed for

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 1: Examples for eight Fine-grained content types.

analyzing discourse functions of individual paragraphs w.r.t 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 contents, we decided to conduct sentence-level annotations instead to minimize disagreements between annotators. and allow consistent annotations[2].

Table 1 contains an example for each content type. Consistent with the theory presented by Van Dijk, the categories are theoretical and some of them may not occur in every news article.

3.1 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.

[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.

3.2 Context-informing Contents

Context-informing sentences provide information related to the actual situation in which 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 preconditions for the main event. They are restricted to events that have occurred very recently, within 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.

3.3 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 account of incidents of an unknown person especially aimed to exaggerate the situation. **Evaluation** (D3) introduces reactions from immediate participants, ex-

perts 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 on 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.

3.4 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.

3.5 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 a news.

4 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 annotator's agreement. 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 2 and 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 has 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

[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 corpus.

	M1	M2	C1	C2	D1	D2	D3	D4	N/A
Business	336(8.5)	40(1.0)	225(5.8)	1,041(26.6)	238(6.1)	70(1.8)	1,049(26.8)	545(13.9)	368(9.4)
Crime	374(10.4)	78(2.2)	271(7.5)	941(26.1)	510(14.2)	77(2.1)	816(22.7)	204(5.7)	328(9.1)
Disaster	407(10.6)	206(5.3)	223(5.8)	1,032(26.8)	139(3.6)	330(8.6)	741(19.2)	405(10.5)	368(9.5)
Politics	475 (10.4)	21(0.4)	218(4.8)	954(20.9)	228(5.0)	85(1.9)	1,492(32.7)	679(14.9)	414(9.1)

Table 2: Distribution of Content type labels across domains, with percentages shown within parentheses.

	M1	M2	C1	C2	D1	D2	D3	D4	N/A
NYT	492(8.4)	97(1.7)	342(5.8)	1401(24.0)	714(12.2)	197(3.4)	1876(32.1)	532(9.1)	197(3.3)
Xinhua	667(13.6)	95(1.9)	361(7.4)	1249(25.5)	214(4.4)	96(2.0)	953(19.5)	525(10.7)	736(15.0)
Reuters	624(8.4)	195(2.6)	391(5.1)	1837(24.8)	571(7.7)	316(4.3)	1867(25.2)	924(12.5)	686(9.3)
NYT_KBP	191(8.6)	42(1.9)	157(7.0)	519(23.3)	384(17.3)	47(2.1)	598(26.9)	148(6.7)	141(6.3)

Table 3: Distribution of Content type labels across media sources, with percentages shown within parentheses.

vs. not speech). In the created corpus, 5535 out of 18,155 sentences are labeled as speech.

5 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 a document encoding, and model associations between each sentence and the main topic of the document when determining content types for sentences. Shown in Figure 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 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.

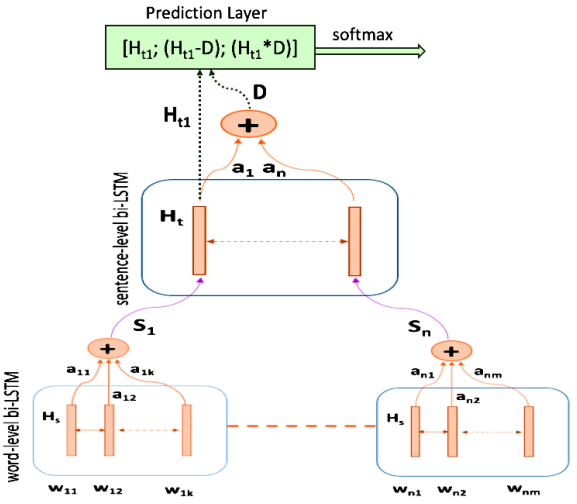

Figure 1: Neural-Network Architecture Incorporating Document Encoding for Content Type Classification

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

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 differ-

Models	M1	M2	C1	C2	D1	D2	D3	D4	Macro			Micro
	F1								P	R	F1	F1
Feature-based (SVM)	34.0	8.0	18.0	44.0	45.0	14.0	52.0	44.0	39.1	37.9	38.3	45.7
Basic Classifier	42.5	24.7	18.2	55.4	59.6	28.5	66.1	52.5	52.6	47.9	48.8(±0.8)	57.5(±0.6)
Document LSTM	49.3	27.3	20.2	57.0	63.6	45.8	**67.4**	55.6	56.6	52.6	53.2(±0.7)	60.2(±1.0)
+Headline	**49.8**	**30.0**	21.8	56.7	63.2	42.7	66.8	**58.7**	**57.3**	52.9	53.8(±0.7)	60.4(±1.0)
+Document encoding	49.6	27.9	**22.5**	**58.1**	**64.1**	**48.1**	**67.4**	57.6	56.9	**53.7**	**54.4(±0.8)**	**60.9(±0.7)**
CRF Fine-grained	47.7	26.4	22.2	56.0	63.3	45.2	66.4	55.2	55.4	52.9	52.9(±1.4)	59.4(±1.1)
CRF Coarse-grained	48.4	29.3	21.6	55.9	62.9	47.2	66.7	54.2	55.6	53.4	53.5(±0.9)	59.6(±0.7)

Table 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.

ences with the sentence encoding to obtain the final sentence representation that is used for predicting its sentence type.

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.

6 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 F1 score for each content type as well as micro and macro F1 scores.

6.1 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.

6.2 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.

6.3 Implementation Details

We set hidden states dimension 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

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

Table 5: Performance of different systems on Speech label classification task.

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.

6.4 Results and Analysis

Tables 4 and 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 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.

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 inverted pyramid, narrative, etc.), further implying the need to model those variations separately in future work.

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.

	M1	M2	C1	C2	D1	D2	D3	D4	N/A
M1	88.0	2.6	9.0	38.2	14.6	0.4	123.2	28.	2.0
M2	6.4	32.4	0.0	28.4	2.0	0.0	3.4	5.4	0.0
C1	13.6	0.6	15.2	27.8	15.2	0.2	25.4	12.0	6.0
C2	39.6	19.2	22.8	483.6	53.2	5.6	134.6	37.6	14.8
D1	3.0	0.0	8.8	54.8	125.4	5.8	41.2	4.2	7.8
D2	1.6	1.6	1.8	9.4	4.0	37.8	41.2	2.8	1.8
D3	6.8	0.0	6.0	82.6	20.4	12.0	586.6	58.2	5.4
D4	4.2	1.2	0.8	29.0	0.4	1.0	63.2	111.4	1.8
NA	1.2	0.0	0.0	1.6	0.6	0.0	3.4	0.0	158.2

Table 6: Confusion matrix for content-type classification based on prediction results of the model *Document LSTM+Document Encoding* on the dev set, averaged over 10 runs consistent with the results reported in Table 4.

We generated confusion matrix (Table 6) for content-type classification based on prediction results of the best performing model **Document LSTM + Document Encoding** on the dev set. Prediction errors mainly occur between Main Event (M1) and Current Context / Evaluation (C2/D3), between Previous Event (C1) and Current Context (C2), between Evaluation (D3) and Expectation (D4), and between Current Context (C2) and Historical Event / Evaluation (D1/D3).

7 Utilizing Content Structure to Improve Event Coreference Resolution

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

Table 7: Percentages of Singleton events in sentences of each content type.

We envision that news discourse profiling can be useful to many discourse level NLP tasks and applications. As an example, we investigate uses of news structures 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 7 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 sentences, but the likelihoods of

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

Table 8: Percentages of Sentences of each content type that contain a headline main event.

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

Table 9: Percentages of Intra-type events out of non-singleton events in sentences of each content type

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 8, around 58%[5] of main event sentences (M1) contain at least one headline event, in addition, 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 will die out within sentences of the same content type. Table 9 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.

Incorporating Content Structure for Event Coreference Resolution: We incorporate news functional structures for event coreference resolution by following the above analysis and implementing content structure informed constraints in an Integer Linear Programming (ILP) inference system to better identify singleton mentions, main event mentions and intra-type event mentions.

We use the *Document LSTM+Document encoding* classifier to predict sentence content types. In addition, we built a discourse-aware event singleton classifier, that resembles the sentence type classifier, to identify singleton event mentions in a document. Specifically, the singleton classifier combines document and sentence representations provided by the content type classifier with contextualized event word representations obtained from a separate word-level biLSTM layer with 512 hidden units. Then, the singleton classifier applies a two-layer feed forward neural network to identify event singletons, and the feed forward network has 3072-512-2 units.

We implement ILP constraints based on system predicted content types of sentences and singleton scores of event mentions. Detailed descriptions of ILP constraints we implemented and their equations are included in the appendix. The ILP formulation has been used in our previous work that yields the previous best system for event coreference resolution (Choubey and Huang, 2018), which aims to capture several specific document level distributional patterns of coreferential event mentions by simply using heuristics. For direct comparisons, we adopt the same experimental settings as in Choubey and Huang (2018), using KBP 2015 documents as the training data and using both KBP 2016 and KBP 2017 corpora for evaluation[6]. We retrained the sentence type classifier using 102 KBP 2015 documents annotated with content types, using 15 documents as the development set and the rest as the training data. We trained the event singleton classifier using the same train/dev split. In addition, we used the same event mentions and pairwise event coreference scores produced by a local pairwise classifier the same as in Choubey and Huang (2018)[7].

Experimental Results: We compare the content-

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

[6]All the KBP corpora include documents from both discussion forum and news articles. But as the goal of this study is to leverage discourse structures specific to news articles for improving event coreference resolution performance, we only evaluate the ILP system using news articles in the KBP corpora. This evaluation setting is consistent with our previous work Choubey and Huang (2018). For direct comparisons, the results reported for all the systems and baselines are based on news articles in the test datasets as well

[7]The classifier can be obtained from https://git.io/JeDw3

Model	KBP 2016					KBP 2017				
	B^3	$CEAF_e$	MUC	BLANC	AVG	B^3	$CEAF_e$	MUC	BLANC	AVG
Local classifier	51.47	47.96	26.29	30.82	39.13	50.24	48.47	30.81	29.94	39.87
+Content Structure	**52.78**	**49.7**	**34.62**	**34.49**	**42.9**	**51.68**	**50.57**	**37.8**	**33.39**	**43.36**
-Singletons	51.47	47.96	31.42	32.89	40.94	51.17	49.67	**38.01**	32.94	42.96
-Main Events	52.65	49.35	32.56	33.69	42.06	51.4	50.05	35.13	31.92	42.12
-Intra-type Events	52.62	49.63	32.97	34.07	42.32	51.62	50.45	37.54	**33.42**	43.26
Lu and Ng (2017)	50.16	48.59	32.41	32.72	40.97	-	-	-	-	-
Choubey and Huang (2018)	51.67	49.1	34.08	34.08	42.23	50.35	48.61	37.24	31.94	42.04

Table 10: Results for event coreference resolution systems on the benchmark datasets (KBP 2016 and 2017).

structure aware ILP system with a baseline system (the row *Local classifier*) that performs greedy merging of event mentions using local classifier predicted pairwise coreference scores as well as two most recent models for event coreference resolution, the heuristics-based ILP system (Choubey and Huang, 2018) and another recent system (Lu and Ng, 2017). We use the same evaluation method as in (Choubey and Huang, 2018) and evaluate event coreference resolution results directly without requiring event mention type match[8].

Table 10 shows experimental results. Event coreference resolution is a challenging task as shown by the small margins of performance gains achieved by recent systems. The ILP model constrained by system predicted content structures (the row *+Content Structure*) outperforms the pairwise classifier baseline system as well as the two most recent systems consistently across all the evaluation metrics over the two benchmark datasets. In particular, our ILP system outperforms the previous state-of-the-art, the heuristics-based ILP system Choubey and Huang, with average F1 gains of 0.67% and 1.32% on KBP 2016 and KBP 2017 corpora respectively. The superior performance shows that systematically identified content structures are more effective than heuristics in guiding event linking, and establishes the usefulness of the new discourse profiling task.

To further evaluate the importance of ILP constraints on Singletons, Main events and Intra-type events, we perform ablation experiments by removing each constraint from the full ILP model. Based on the results in Table 10, all the three types of constraints have noticeable impacts to coreference performance, and singletons and main events constraints contribute the most.

[8]The official KBP 2017 event coreference resolution scorer considers two event mentions coreferent if they strictly match on their event type and subtype, which requires building a high-performing event type identification system to enable an event coreference resolver to score well.

Intuitively, news content structures can help in identifying other event relations as well, such as temporal and causal relations, and thus disentangling complete event structures. For instance, events occurring in C1 (Previous Event) sentences are probable cause for the main event which in turn causes events in M2 (Consequence) sentences (the same rationale can be applied for temporal order).

8 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 feasibility of this task. We conducted experiments and demonstrated the usefulness of news discourse profiling for event coreference resolution. In the future, we will further improve the performance of news discourse profiling by investigating subgenres of news articles, and extensively explore its usage for various other NLP tasks and applications.

Acknowledgments

We thank our anonymous reviewers for providing insightful review comments. We gratefully acknowledge support from National Science Foundation via the awards IIS-1942918 and IIS-1755943. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: 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.

References

Daniela Baiamonte, Tommaso Caselli, and Irina Prodanof. 2016. Annotating content zones in news articles. *CLiC it*, page 40.

Regina Barzilay and Lillian Lee. 2004. Catching the Drift: Probabilistic Content Models, with Applications to Generation and Summarization.

Maria Becker, Michael Staniek, Vivi Nastase, Alexis Palmer, and Anette Frank. 2017. 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.

Allan Bell. 1998. The discourse structure of news stories. In *Approaches to media discourse*, pages 64–104.

Parminder Bhatia, Yangfeng Ji, and Jacob Eisenstein. 2015. Better document-level sentiment analysis from RST discourse parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2212–2218, Lisbon, Portugal. Association for Computational Linguistics.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Freddy YY Choi. 2000. Advances in domain independent linear text segmentation. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 26–33. Association for Computational Linguistics.

Prafulla Kumar Choubey and Ruihong Huang. 2018. 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.

Prafulla Kumar Choubey, Kaushik Raju, and Ruihong Huang. 2018. 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)*, volume 2, pages 340–345.

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

Zeyu Dai and Ruihong Huang. 2018a. 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.

Zeyu Dai and Ruihong Huang. 2018b. Improving implicit discourse relation classification by modeling inter-dependencies 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)*, volume 1, pages 141–151.

George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie M Strassel, and Ralph M Weischedel. 2004. The automatic content extraction (ace) program-tasks, data, and evaluation. In *Lrec*, volume 2, page 1.

Jacob Eisenstein and Regina Barzilay. 2008. Bayesian unsupervised topic segmentation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 334–343. Association for Computational Linguistics.

Joe Ellis, Jeremy Getman, Dana Fore, Neil Kuster, Zhiyi Song, Ann Bies, and Stephanie Strassel. 2015. Overview of linguistic resources for the tac kbp 2015 evaluations: Methodologies and reults. In *Proceedings of the TAC KBP 2015 Workshop*, pages 16–17.

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

Vanessa Wei Feng and Graeme Hirst. 2012. Text-level discourse parsing with rich linguistic features. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 60–68.

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

Marti A Hearst. 1994. Multi-paragraph segmentation of expository text. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 9–16. Association for Computational Linguistics.

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

Yangfeng Ji and Jacob Eisenstein. 2014. Representation learning for text-level discourse parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 13–24.

Yangfeng Ji and Jacob Eisenstein. 2015. One vector is not enough: Entity-augmented distributed semantics for discourse relations. *Transactions of the Association for Computational Linguistics*, 3:329–344.

Yangfeng Ji and Noah A. Smith. 2017. Neural discourse structure for text categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 996–1005, Vancouver, Canada. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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

Omri Koshorek, Adir Cohen, Noam Mor, Michael Rotman, and Jonathan Berant. 2018. Text segmentation as a supervised learning task. 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 469–473, New Orleans, Louisiana. Association for Computational Linguistics.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196.

Jiwei Li, Rumeng Li, and Eduard Hovy. 2014a. Recursive deep models for discourse parsing. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2061–2069.

Junyi Jessy Li, Marine Carpuat, and Ani Nenkova. 2014b. Assessing the discourse factors that influence the quality of machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 283–288.

Maria Liakata, Shyamasree Saha, Simon Dobnik, Colin Batchelor, and Dietrich Rebholz-Schuhmann. 2012. Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics*, 28(7):991–1000.

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

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2011. Automatically evaluating text coherence using discourse relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 997–1006. Association for Computational Linguistics.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2014. A pdtb-styled end-to-end discourse parser. *Natural Language Engineering*, 20(2):151–184.

Linlin Liu, Xiang Lin, Shafiq Joty, Simeng Han, and Lidong Bing. 2019. Hierarchical pointer net parsing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1007–1017, Hong Kong, China. Association for Computational Linguistics.

Annie Louis, Aravind Joshi, and Ani Nenkova. 2010. Discourse indicators for content selection in summarization. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 147–156. Association for Computational Linguistics.

Jing Lu and Vincent Ng. 2017. 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.

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

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

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

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. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. 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)*, volume 1, pages 2227–2237.

Emily Pitler and Ani Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text. In *ACL-IJCNLP*, pages 13–16.

R. Prasad, N. Dinesh, Lee A., E. Miltsakaki, L. Robaldo, Joshi A., and B. Webber. 2008. The Penn Discourse Treebank 2.0. In *lrec2008*.

Lianhui Qin, Zhisong Zhang, and Hai Zhao. 2016. A stacking gated neural architecture for implicit discourse relation classification. In *EMNLP*, pages 2263–2270.

Radim Řehůřek and Petr Sojka. 2010. 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. ELRA. http://is.muni.cz/publication/884893/en.

Attapol T Rutherford and Nianwen Xue. 2016. Robust non-explicit neural discourse parser in english and chinese. *ACL*, page 55.

Christina Sauper, Aria Haghighi, and Regina Barzilay. 2010. Incorporating Content Structure into Text Analysis Applications.

Weiyan Shi, Tiancheng Zhao, and Zhou Yu. 2019. Unsupervised dialog structure learning. 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 1797–1807, Minneapolis, Minnesota. Association for Computational Linguistics.

Radu Soricut and Daniel Marcu. 2003. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 228–235.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Simone Teufel, Jean Carletta, and Marc Moens. 1999. An annotation scheme for discourse-level argumentation in research articles. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, pages 110–117. Association for Computational Linguistics.

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

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

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

Anita Waard, Paul Buitelaar, and Thomas Eigner. 2009. Identifying the epistemic value of discourse segments in biology texts. *Proceedings of the Eighth International Conference on Computational Semantics:*, pages 351–354.

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

Yang Xu, Yu Hong, Huibin Ruan, Jianmin Yao, Min Zhang, and Guodong Zhou. 2018. Using active learning to expand training data for implicit discourse relation recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 725–731, Brussels, Belgium. Association for Computational Linguistics.

W Victor Yarlott, Cristina Cornelio, Tian Gao, and Mark Finlayson. 2018. Identifying the discourse function of news article paragraphs. In *Proceedings of the Workshop Events and Stories in the News 2018*, pages 25–33.

A 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. We formulate the baseline objective function that minimizes equation 1.

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

We then add constituent objective functions (equation 2) 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).

$$\Theta = \Theta_B + K_M\Theta_M + K_C\Theta_C + K_L\Theta_L + K_S\Theta_S + K_N\Theta_N \tag{2}$$

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.

A.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 3).

$$\Theta_S = \sum_{i\in\lambda, j\in\lambda, i\vee j\in S} x_{ij} \;;\; \Theta_N = - \sum_{i\in\lambda, j\in\lambda, i\wedge j\in N} x_{ij} \tag{3}$$

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.

A.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 tend to have coreferential event mentions within the same content type or sometimes in the main content. We model 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 4) that add the negative sum of their indicator variables to the main objective function.

$$\Theta_M = - \sum_{i\in\xi_H, j\in\xi_M} x_{ij} \tag{4}$$

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 5).

$$\Theta_C = - \sum_{i\in\xi_H, j\in\xi_R} x_{ij} \tag{5}$$

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) to penalize event chains that start in that non-main content type sentence but include event mentions from other non-main type sentences.

$$\Theta_L = \sum_{i\in\xi_T} Y_i$$
$$s.t. \quad \Gamma_i - Y_i \leq M\gamma_i \tag{6}$$
$$\Gamma_i = \sum_{i\in\xi_T, j\notin(\xi_M\cup\xi_T)} x_{ij} \;;\; \gamma_i = \sum_{k\notin\xi_T, i\in\xi_T} x_{ki}$$

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, 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.