# Improving Event Coreference Resolution by Modeling Correlations between Event Coreference Chains and Document Topic Structures

**Prafulla Kumar Choubey and Ruihong Huang**
Department of Computer Science and Engineering
Texas A&M University
(prafulla.choubey, huangrh)@tamu.edu

## Abstract

This paper proposes a novel approach for event coreference resolution that models correlations between event coreference chains and document topical structures through an Integer Linear Programming 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 across KBP 2016 and 2017 datasets suggest that each of the structures contribute to improving event coreference resolution performance.

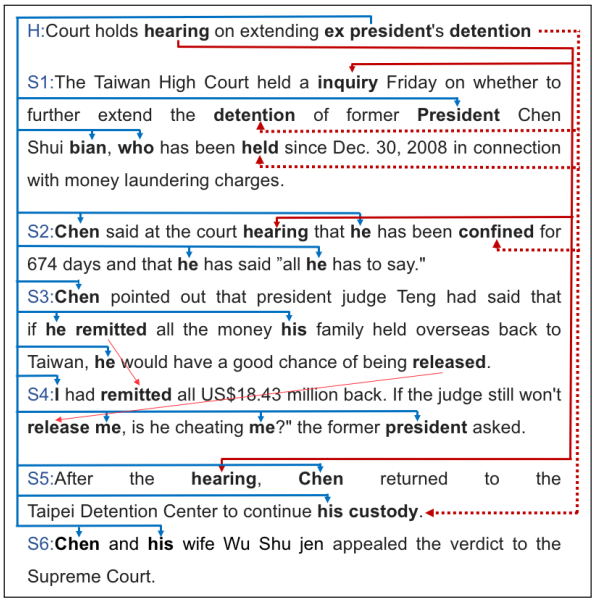


Figure 1: An example document to illustrate the characteristics of event (red) and entity (blue) coreference chains.

## 1 Introduction

Event coreference resolution aims to identify and link event mentions in a document that refer to the same real-world event, which is vital for identifying the skeleton of a story and text understanding and is beneficial to numerous other NLP applications such as question answering and summarization. In spite of its importance, compared to considerable research for resolving coreferential entity mentions, far less attention has been devoted to event coreference resolution. Event coreference resolution thus remained a challenging task and the best performance remained low.

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 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 se-

| 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 1: Percentages of adjacent (event vs. entity) mention pairs based on the number of sentences between two mentions.

ries 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 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 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 in 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, we 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 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 Integer Linear Programming (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.

## 2 Correlations between Event Coreference Chains and Document Topic Structures

We model four aspects of correlations.

**Correlations between Main Event Chains and Topic Transition Sentences:** the main events of a document, e.g., "hearing" and "detention" in this example 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 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).

**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.

**Genre-specific Distributional Patterns:** we model document level distributional patterns of coreferent event mentions that may be specific to a 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.

**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.

## 3 Related Work

Compared to entity coreference resolution (Lee et al., 2017; Clark and Manning, 2016a,b; Martschat and Strube, 2015; Lee et al., 2013), far less research was conducted for event coreference resolution. Most existing methods (Ahn, 2006; Chen et al., 2009; Cybulska and Vossen, 2015a,b) heavily rely on surface features, mainly event arguments (i.e., entities such as event participants, time, location, etc.) that were extracted from local contexts of two events, and determine that two events are coreferential if their arguments match. Often, a clustering algorithm, hierarchical Bayesian (Bejan and Harabagiu, 2010, 2014; Yang et al., 2015) or spectral clustering algorithms (Chen and Ji, 2009), is applied on top of a pairwise surface feature based classifier for inducing event clusters. However, identifying potential arguments, linking arguments to a proper event mention, and recognizing compatibilities between arguments are all error-prone (Lu et al., 2016). Joint event and entity coreference resolution (Lee et al., 2012), joint inferences of event detection and event coreference resolution (Lu and Ng, 2017), and iterative information propagation (Liu et al., 2014; Choubey and Huang, 2017a) have been proposed to mitigate argument mismatch issues.

However, such methods are incapable of handling more complex and subtle cases, such as partial event coreference with incompatible arguments (Choubey and Huang, 2017a) and cases lacking informative local contexts. Consequently, many event coreference links were missing and the resulted event chains are fragmented. The low performance of event coreference resolution limited its uses in downstream applications. (Choubey et al., 2018) shows that instead of human annotated event coreference relations, using system predicted relations resulted in a significant performance reduction in identifying the central event of a document. Moreover, the recent research by Moosavi and Strube (2017) found that the extensive use of lexical and surface features biases entity coreference resolvers towards seen mentions and do not generalize to unseen domains, and the finding can perfectly apply to event coreference resolution. Therefore, we propose to improve event coreference resolution by modeling

correlations between event coreferences and the overall topic structures of a document, which is more likely to yield robust and generalizable event coreference resolvers.

## 4 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 Integer Linear Programming (ILP) problem. Our baseline ILP system is defined over pairwise scores between event mentions obtained from a pairwise neural network-based coreference resolution classifier.

### 4.1 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[1] and prefix [2] 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 The Basic ILP for Event Coreference Resolution

Let $\lambda$ represents the set of all event mentions in a document, $\Lambda$ 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 1. Further we add constraints (equation 2) over each triplets of mentions to enforce transitivity (Denis et al., 2007; Finkel and Manning, 2008). This guarantees legal clustering by ensuring that $x_{ij} = x_{jk} = 1$ implies $x_{ik}$ = 1.

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

$$\neg x_{ij} + \neg x_{jk} \geq \neg x_{ik} \tag{2}$$

We then add constituent objective functions and constraints to the baseline ILP formulation to induce correlations between coreference chains and topical structures ($\Theta_T$), discourage fragmented chains ($\Theta_G$), encourage semantic associations among chains ($\Theta_C$), model genre-specific distributional patterns ($\Theta_D$) and discourage subevents from having coreferent mentions ($\Theta_S$). They are described in the following subsections.

#### 4.2.1 Modeling the Correlation between Main Event Chains and Topic Transition Sentences

As shown in the example Figure 1, main events are likely to have mentions appear in topic transition sentences. Therefore, We add the following objective function (equation 3) to the basic objective function and add the new constraint 4 in order to encourage coreferent event mentions to occur in topic transition sentences.

$$\Theta_T = \sum_{m,n \in \Omega} -log(s_{mn})w_{mn} - log(1-s_{mn})(\neg w_{mn}) \quad s.t.\; w_{mn} \in \{0,1\} \quad (n-m) \geq |S|/\theta_s \tag{3}$$

$$\sum_{i' \in \xi_m, j' \in \xi_n} x_{i'j'} \geq w_{mn} \tag{4}$$

Specifically, let $\omega$ represents the set of sentences in a document and $\Omega$ denotes the set of sentence pairs i.e. $\Omega = \{< m,n > \mid < m,n > \in \omega \times \omega$ 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 ($> 0.5$)

[1] 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

[2] re, in, at, tr, op

and are not near (with $|S|/\theta_s$or more sentences apart, in our experiments we set $\theta_s$ to 5), this objective function $\Theta_T$ tries to set the corresponding indicator variable $w_{mn}$ to 1. Then, we add constraint 4 to encourage coreferent event mentions to occur in topic transition sentences. Note that $\xi_m$ 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.

**Identifying Topic Transition Sentences Using Sentence Similarities:** First, we use the unsupervised method based on 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 embedding 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.

**Constraints for Avoiding Fragmented Partial Event Chains:** The above equations (3-4) 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 5) and the new constraints (equation 6 and 7):

$$\Theta_G = -\sum_{i,j \in \mu} \gamma_{ij} \tag{5}$$

$$\sigma_{ij} = \sum_{k<i} \neg x_{ki} \wedge \sum_{j<l} \neg x_{jl} \wedge x_{ij} \qquad \sigma_{ij} \in \{0,1\} \tag{6}$$

$$\begin{aligned} \Gamma_i &= \sum_{k,i \in \Lambda} x_{ki} + \sum_{i,j \in \Lambda} x_{ij} \\ M(1-y_{ij}) &\geq (\varphi[j] - \varphi[i]).\sigma_{ij} - \lceil 0.75\,(|S|) \rceil \\ \gamma_{ij} - \Gamma_i - \Gamma_j &\geq M.y_{ij} \\ \Gamma_i, \Gamma_j, \gamma_{ij} \in Z;\ \Gamma_i, \Gamma_j, \gamma_{ij} &\geq 0;\ y_{ij} \in \{0,1\} \end{aligned} \tag{7}$$

First, we define an indicator variable $\sigma_{ij}$ by equation 6 [3], 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 $\sigma_{ij}$ to 1 defines an event chain that starts from the event mention $i$ and ends at the event mention $j$.

Then with equation 7, variable $\sigma_{ij}$ 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, $\Gamma_i$ ($\Gamma_j$) 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 $\sigma_{ij}$ is a global chain. Given $\sigma_{i,j}$ is identified as a global chain, variable $\gamma_{ij}$ equals the sum of variables $\Gamma_i$ and $\Gamma_j$ and is used in the objective function $\Theta_G$ (equation 5) to encourage more coreferential mentions.

[3] Equation 6 can be implemented as

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

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

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.

#### 4.2.2 Cross-chain Inferences

As illustrated through Figure 1, semantically related events tend to have their mentions co-occur within the same sentence. So, we define the objective function (equation 8) and constraints (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.

$$\Theta_C = -\sum_{m,n \in \Omega} \Phi_{mn} \tag{8}$$

$$\Phi_{mn} = \sum_{i \in \xi_m, j \in \xi_n} x_{ij} \quad |\xi_m| > 1;\ |\xi_n| > 1;\ \Phi_{mn} \in Z;\ \Phi_{mn} \geq 0 \tag{9}$$

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

#### 4.2.3 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 10).

$$\Theta_D = -\sum_{i \in \xi_m, j \in \xi_n} x_{ij} + \sum_{k \in \xi_p, l \in \xi_q} x_{kl} \quad s.t.\ m, n < \lfloor \alpha |S| \rfloor;\ p, q > \lceil \beta |S| \rceil \quad \alpha \in [0, 1];\ \beta \in [0, 1] \tag{10}$$

Specifically, for the event pairs that belong to the first $\alpha$ (or the last $\beta$) sentences in a document, we add the negative (positive) sum of their indicator variables (x) in objective function $\Theta_D$.

The equation 10 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.

#### 4.2.4 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 11).

$$\Theta_S = \sum_{s \in \mathbb{S}} \Gamma_s \tag{11}$$

where $\mathbb{S}$ represents the set of subevents in a document. We define the objective function $\Theta_S$ as the sum of $\Gamma_s$, where $\Gamma_s$ equals the number of mentions that are coreferent to $s$. Then our goal is to minimize $\Theta_S$ 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.

### 4.3 The full ILP Model and the Parameters

The equations 3-11 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($\Theta$), defined in equation 12, that incorporates prior knowledge by means of hard or soft constraints.

$$\Theta = \kappa_B \Theta_B + \kappa_T \Theta_T + \kappa_G \Theta_G + \kappa_C \Theta_C + \kappa_D \Theta_D + \kappa_S \Theta_S \tag{12}$$

Here, all the $\kappa$ parameters are floating point constants. For the sake of simplicity, we set $\kappa_B$ and $\kappa_T$ to 1.0 and $\kappa_G = \kappa_C$. Then we estimate the parameters $\kappa_G(\kappa_C)$ and $\kappa_D$ 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, $\Theta_S$ aims to inhibit subevents from being included in coreference chains, we set a high value for $\kappa_S$ and found that, indeed, the performance

remained same for all the values of $\kappa_S$ in range [5.0,15.0]. In our final model, we keep $\kappa_S$ = 10.0. Also, we found that the performance is roughly invariant to the parameters $\kappa_G$ and $\kappa_C$ 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).

## 5 Evaluation

### 5.1 Experimental Setup

We trained our ILP system on the KBP 2015 (Ellis et al., 2015) English dataset and evaluated the system on KBP 2016 and KBP 2017 English datasets[4]. All the KBP corpora include documents from both discussion forum[5] and news articles. But as 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[6] from the KBP 2015 corpus and evaluate our system on news articles from the official KBP 2016 and 2017 evaluation corpora[7] 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.

We report the event coreference resolution results based on the version 1.8 of the official KBP 2017 scorer. The scorer employs four coreference scoring measures, namely $B^3$ (Bagga and Baldwin, 1998), $CEAF_e$ (Luo, 2005), MUC (Vilain et al., 1995) and BLANC (Recasens and Hovy, 2011) and the unweighted average of their F1 scores ($AVG_{F1}$).

### 5.2 Event Mention Identification

| | 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 2: F1 scores for event mention extraction on the KBP 2016 and 2017 corpus

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 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 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 sub-

[4]The ECB+ (Cybulska and Vossen, 2014) corpus is another commonly used dataset for evaluating event coreference resolution performance. But we determined that this corpus is not appropriate for evaluating our ILP model that explicitly focuses on using discourse level topic structures for event coreference resolution. Particularly, the ECB+ corpus was created to facilitate both cross-document and in-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 only annotated selectively in 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 approach. In addition, due to the selective annotation issue, in-document event coreference resolution with the ECB+ corpus is somewhat easier than with the KBP corpus, which partly explained the significant differences of published in-document event coreference resolution results on the two corpora.

[5]Each discussion forum document consists of a series of posts in an online discussion thread, which lacks coherent discourse structures as a regular document. Therefore, only news articles in the KBP corpora are appropriate for evaluating our approach.

[6]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 remaining 310 documents for training classifiers.

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

| 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 |
| Clustering | 46.97 | 41.95 | 18.79 | 26.88 | 33.65 | 46.51 | 40.21 | 23.10 | 25.08 | 33.72 |
| Basic ILP | 51.44 | 47.77 | 26.65 | 30.95 | 39.19 | 50.4 | 48.49 | 31.33 | 30.58 | 40.2 |
| +Topic structure | 51.44 | 47.94 | 28.86 | 31.87 | 40.03 | 50.39 | 48.23 | 33.08 | 31.26 | 40.74 |
| +Cross-chain | 51.09 | 47.53 | 31.27 | 33.07 | 40.74 | 50.39 | 47.67 | 35.15 | 31.88 | 41.27 |
| +Distribution | 51.06 | 48.28 | 33.53 | 33.63 | 41.62 | 50.42 | 48.67 | 37.52 | 32.08 | 42.17 |
| +Subevent | 51.67 | 49.1 | 34.08 | 34.08 | 42.23 | 50.35 | 48.61 | 37.24 | 31.94 | 42.04 |
| Joint learning | 50.16 | 48.59 | 32.41 | 32.72 | 40.97 | - | - | - | - | - |

Table 3: Results for 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). The file was obtained from the authors.

type 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.

Note that 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). But 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 document structures in the event coreference resolution task by overlooking any bias from upstream tasks.

### 5.3 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. This model has achieved the best event coreference resolution performance to date on the KBP 2016 corpus (Lu and Ng, 2017).

### 5.4 Our Systems

We gradually augment the ILP baseline with additional objective functions and constraints described in sub-sections 4.2.1, 4.2.2, 4.2.3 and 4.2.4. In all the systems below, we combine objective functions with their corresponding coefficients (as described in sub-section 4.3).
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 1, 2).
+Topic structure incorporates the topical structure and the characteristics of main event chains in baseline ILP system (Equations 1-5).
+Cross-chain adds constraints and objective function defined for cross-chain inference to the Topical structure system (Equations 1-8).
+Distribution further adds distributional patterns to the Cross-chain system (Equations 1-10).
+Subevent (Full) optimizes the objective function defined in equation 12 by considering all the constraints defined in 1-11, including constraints for modeling subevent structures.

### 5.5 Results and Analysis

Table 3 shows performance comparisons of our ILP systems with other event coreference resolu-

tion approaches including the recent joint learning approach (Lu and Ng, 2017) which is the best performing model 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 average F1 gain of 3.1 for KBP 2016 and 2.17 for KBP 2017. Most interestingly, we see over 28% improvement in MUC F1 score which directly evaluates the pairwise coreference link predictions. This implies that the document level structures, indeed, helps 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 on the KBP 2016 corpus (Lu and Ng, 2017) consistently using all the evaluation metrics, with an overall improvement of 1.21 based on the average F1 scores.

In Table 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.

### 5.6 Discussions on Generalizability

The correlations between event coreference chains and document topic structures are not specific to news articles and widely exist. Several main distributional characteristics of coreferent event mentions, including 1) main event coreference chains often have extended presence and have mentions scattered across segments, and 2) semantically correlated events often have their respective event mentions co-occur in a sentence, directly apply to other sources of texts such as clinical notes. But certain distributional characteristics are genre specific. For instance, while it is common to observe more coreferent event mentions early on in a news article, coreference chains in a clinical note often align well with pre-defined segments like the history of present illness, description of a visit and treatment plan. Thus, the objective functions and constraints defined in equations 1-8 can be directly applied for other domains as well, while other structures like segment-wise distributional patterns may require alteration based on domain-specific knowledge.

## 6 Conclusions and the Future Work

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 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. In the future, we will explore the use of additional discourse structures that correlate highly with event coreference chains. Moreover, we will extend this work to other domains such as biomedical domains.

## Acknowledgments

This work was partially supported by the National Science Foundation via NSF Award IIS-1755943. 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

David Ahn. 2006. The stages of event extraction. In *Proceedings of the Workshop on Annotating and Reasoning About Time and Events*. Association for Computational Linguistics, Stroudsburg, PA, USA, ARTE '06, pages 1–8. http://dl.acm.org/citation.cfm?id=1629235.1629236.

Jun Araki, Zhengzhong Liu, Eduard H Hovy, and Teruko Mitamura. 2014. Detecting subevent structure for event coreference resolution. In *LREC*. pages 4553–4558.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2016. A latent variable model approach to pmi-based word embeddings. *Transactions of the Association for Computational Linguistics* 4:385–399.

Allison Badgett and Ruihong Huang. 2016. Extracting subevents via an effective two-phase approach. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. pages 906–911.

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *The first international conference on language resources and evaluation workshop on linguistics coreference*. Granada, volume 1, pages 563–566.

Cosmin Adrian Bejan and Sanda Harabagiu. 2010. Unsupervised event coreference resolution with rich linguistic features. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 1412–1422.

Cosmin Adrian Bejan and Sanda Harabagiu. 2014. Unsupervised event coreference resolution. *Computational Linguistics* 40(2):311–347.

Zheng Chen and Heng Ji. 2009. Graph-based event coreference resolution. In *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing*. Association for Computational Linguistics, pages 54–57.

Zheng Chen, Heng Ji, and Robert Haralick. 2009. A pairwise event coreference model, feature impact and evaluation for event coreference resolution. In *Proceedings of the workshop on events in emerging text types*. Association for Computational Linguistics, pages 17–22.

Prafulla Kumar Choubey and Ruihong Huang. 2017a. 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.

Prafulla Kumar Choubey and Ruihong Huang. 2017b. Tamu at kbp 2017: Event nugget detection and coreference resolution. In *Proceedings of TAC KBP 2017 Workshop, National Institute of Standards and Technology*.

Prafulla Kumar Choubey, Kaushik Raju, and Ruihong Huang. 2018. The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, New Orleans, LA, USA.

Kevin Clark and Christopher D Manning. 2016a. Deep reinforcement learning for mention-ranking coreference models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. pages 2256–2262.

Kevin Clark and Christopher D Manning. 2016b. Improving coreference resolution by learning entity-level distributed representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. volume 1, pages 643–653.

Agata Cybulska and Piek Vossen. 2014. Using a sledgehammer to crack a nut? lexical diversity and event coreference resolution. In *LREC*. pages 4545–4552.

Agata Cybulska and Piek Vossen. 2015a. Translating granularity of event slots into features for event coreference resolution. In *Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*. pages 1–10.

A.K. Cybulska and P.T.J.M. Vossen. 2015b. Bag of events approach to event coreference resolution. supervised classification of event templates. *Lecture Notes in Computer Science* (9042). 978-3-319-18117-2.

Pascal Denis, Jason Baldridge, et al. 2007. Joint determination of anaphoricity and coreference resolution using integer programming. In *HLT-NAACL*. pages 236–243.

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 results. In *Proceedings of TAC KBP 2015 Workshop, National Institute of Standards and Technology*. pages 16–17.

Joe Ellis, Jeremy Getman, Neil Kuster, Zhiyi Song, Ann Bies, and Stephanie Strassel. 2016. 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*.

Jenny Rose Finkel and Christopher D Manning. 2008. Enforcing transitivity in coreference resolution. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*. Association for Computational Linguistics, pages 45–48.

Jeremy Getman, Joe Ellis, Zhiyi Song, Jennifer Tracey, and Stephanie Strassel. 2017. 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*.

Barbara J Grosz, Scott Weinstein, and Aravind K Joshi. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational linguistics* 21(2):203–225.

Aria Haghighi and Dan Klein. 2007. Unsupervised coreference resolution in a nonparametric bayesian model. In *Proceedings of the 45th annual meeting of the association of computational linguistics*. pages 848–855.

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

Heeyoung Lee, Marta Recasens, Angel Chang, Mihai Surdeanu, and Dan Jurafsky. 2012. 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*. Association for Computational Linguistics, pages 489–500.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. pages 188–197.

Zhengzhong Liu, Jun Araki, Eduard H Hovy, and Teruko Mitamura. 2014. Supervised within-document event coreference using information propagation. In *LREC*. pages 4539–4544.

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)*. volume 1, pages 90–101.

Jing Lu, Deepak Venugopal, Vibhav Gogate, and Vincent Ng. 2016. Joint inference for event coreference resolution. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. pages 3264–3275.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*. Association for Computational Linguistics, pages 25–32.

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

Sebastian Martschat and Michael Strube. 2015. Latent structures for coreference resolution. *Transactions of the Association of Computational Linguistics* 3(1):405–418.

Teruko Mitamura, Zhengzhong Liu, and Eduard Hovy. 2017. Events detection, coreference and sequencing: Whats next? overview of the tac kbp 2017 event track. In *Proceedings of TAC KBP 2017 Workshop, National Institute of Standards and Technology*.

Stuart Mitchell, Michael OSullivan, and Iain Dunning. 2011. 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* .

Nafise Sadat Moosavi and Michael Strube. 2017. Lexical features in coreference resolution: To be used with caution. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. volume 2, pages 14–19.

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.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. pages 1532–1543. http://www.aclweb.org/anthology/D14-1162.

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

Zhiyi Song, Ann Bies, Stephanie Strassel, Tom Riese, Justin Mott, Joe Ellis, Jonathan Wright, Seth Kulick, Neville Ryant, and Xiaoyi Ma. 2015. 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.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the 6th conference on Message understanding*. Association for Computational Linguistics, pages 45–52.

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

Bishan Yang, Claire Cardie, and Peter Frazier. 2015. A hierarchical distance-dependent bayesian model for event coreference resolution. *Transactions of the Association of Computational Linguistics* 3(1):517–528.