# Automatic Data Acquisition for Event Coreference Resolution

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

## Abstract

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

## 1 Introduction

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

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

We aim to improve the effectiveness of event coreference resolution systems by automatically acquiring coreferential event pairs from many documents requiring minimal supervision. Specifically, coreferential event mentions are associated with discourse function of sentences in a news document (Choubey et al., 2020)[1]. We propose to use them to identify sentence pairs that are likely to contain coreferential event mentions as well as sentence pairs that are likely to contain non-coreferential event pairs. Consider the two example sentence pairs below, each pair having an event pair with synonymous trigger words.

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

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

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

[1] The discourse roles are roughly based on the Van Dijk's theory of news discourse (Teun A, 1986). It assigns discourse function to sentences in a news article, where the function is characterized by the operative role of sentence's content in describing the main event, context informing events, and other historical or future projected events

*Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1185–1196
April 19 - 23, 2021. ©2021 Association for Computational Linguistics

event in a news article. In general, we can recognize pairs of sentences in news articles that are likely to contain coreferential or non-coreferential event mention pairs by knowing the sentence's discourse function following Van Dijk's theory.

To ascertain our hypothesis, we first use the discourse profiling system and dataset introduced by Choubey et al. (2020) to identify the discourse role for each sentence in a news article. Then, we use multiple rules to capture the distributional correlation between event coreference chains and discourse roles of sentences and collect a diverse set of 9,210 coreferential and 232,135 non-coreferential event pairs[2]. To assess the reliability of the proposed data augmentation strategy, we perform manual validation on subsets of both coreferential and non-coreferential event pairs. Then, we train event coreference resolution systems using the acquired data alone or using the acquired data to augment a human-annotated training dataset.

We evaluate trained systems on two datasets, the news portion[3] of the widely used benchmark evaluation corpus KBP 2017 as well as the news portion[4] of the Richer Event Description (RED) corpus (O'Gorman et al., 2016). Unlike the KBP corpora that only consider eight event types for event coreference annotations, the RED corpus comprehensively annotates all the event types that appear in a document, and is arguably the only comprehensively annotated corpus of event coreference relations. Assuming the automatically acquired event coreference data is not available, we also train a supervised event coreference resolution system using the KBP 2015 corpus[5]. On the KBP 2017 corpus, the event coreference resolution system trained on the acquired data performs slightly worse than the system trained using the KBP 2015 corpus, the human-annotated in-domain training data. But, on the RED corpus, both the systems trained on either the annotated KBP 2015 corpus or the acquired data obtain roughly the same evaluation results. Further, the system trained on combined annotated KBP 2015 and automatically acquired data yields the best results on both the KBP 2017 dataset and the RED dataset.

Lastly, we evaluate all the trained systems on a different text genre, discussion forum articles from the KBP 2017 corpus, and found that all the systems obtain comparable results. Overall, the performance gain of all the trained systems on discussion forum documents is marginal compared to a simple trigger word match baseline. Thus, increasing training data size does not improve the performance of an event coreference resolution system on a new text genre. We suspect that, for generalization across different text genres, we may require specialized learning algorithms, e.g., text style adaptation, which is not in the scope of this work.

## 2 Related Work

The existing literature on supervised event coreference resolution primarily focuses on designing pairwise classifier based on the surface linguistic features such as lexical features comprising of lemma and part-of-speech tag similarity of event words (Bejan and Harabagiu, 2010; Lee et al., 2012; Liu et al., 2014; Yang et al., 2015; Lu et al., 2016; Cremisini and Finlayson, 2020), argument overlap (Chen et al., 2009; McConky et al., 2012; Sangeetha and Arock, 2012; Bejan and Harabagiu, 2014; Yang et al., 2015; Lu et al., 2016; Choubey and Huang, 2017), semantic similarity based on lexical resources such as wordnet (Bejan and Harabagiu, 2010; Liu et al., 2014; Yu et al., 2016) and word embeddings (Yang et al., 2015; Choubey and Huang, 2017; Kenyon-Dean et al., 2018; Barhom et al., 2019; Zuo et al., 2019; Pandian et al., 2020; Sahlani et al., 2020; Lu et al., 2020), and discourse features such as token and sentence distance (Liu et al., 2014; Cybulska and Vossen, 2015). The resulting classifier is used to cluster event mentions. The commonly used strategies include *agglomerative clustering* that selects the antecedent closest in mention distance that is classified as coreferent or the antecedent with the highest coreference likelihood (Chen et al., 2009; Chen and Ng, 2014), hierarchical bayesian (Yang et al., 2015) or spectral clustering algorithms (Chen and Ji, 2009). In this work, we use the pre-trained BERT model to extract both event and context features and use agglomerative clustering to form event coreference chains.

Supervised models suffer from a lack of human-

[2] The acquired coreferential and non-coreferential event pairs can be found at https://github.com/prafulla77/Event-Coref-EACL-2021

[3] All the KBP corpora include news articles as well as documents from discussion forums.

[4] In addition to news articles, the RED corpus contains several other types of documents, including news summaries, discussion forum posts, and web posts.

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

annotated event coreference data. To address the annotation scarcity problem, Peng et al. (2016) proposed to learn structured event representations on large amounts of text and use the similarity score between two event representations to form event coreference chains. Their model uses a small human-annotated event coreference dataset to find the appropriate similarity score threshold for linking two events. Unsupervised models based on probabilistic generative modeling have also been successfully used for event coreference resolution (Bejan and Harabagiu, 2010; Chen and Ng, 2015). However, both semi-supervised and unsupervised approaches have been found empirically lagging behind the supervised models (Lu and Ng, 2018).

The closest to our work are weakly-supervised and self-training methods that have been shown useful for many information extraction and classification tasks (Riloff, 1996; Riloff and Wiebe, 2003; Xie et al., 2019). But, to the best of our knowledge, we are the first to explore discourse-aware strategies to automatically label event coreference relations and use them exclusively or use them to augment existing human-annotated data for training event coreference resolution systems.

## 3 Event Coreference Data Acquisition

To acquire coreferential event-pairs without direct supervision, we first collect event trigger words along with their potential set of coreferential event mentions using The Paraphrase Database (PPDB 2.0) (Ganitkevitch et al., 2013; Pavlick et al., 2015)[6]. Then, we use high precision rules informed by the functional news discourse structures (Teun A, 1986; Choubey et al., 2020) to identify seed coreferential and non-coreferential event pairs followed by a single bootstrapping iteration to collect additional non-coreferential event pairs.

### 3.1 Identifying Coreferential Event Trigger Candidates using The PPDB Database

We collect lexically diverse candidate coreferential event pairs using the paraphrases from PPDB-2.0-s-lexical (Pavlick et al., 2015) database. The corpus[7] contains 213,716 highest scoring lexical paraphrase pairs, each annotated with one of the equivalence, forward or reverse entailment, and contradiction relation classes. First, we extract all the *verb* paraphrase pairs as the potential event trigger words. While event mentions can take other part of speech types, we limit our paraphrase pairs to verbs to ensure high precision among the collected event trigger words. Additionally, many of the verb paraphrase pairs include nominalization (e.g., investing and investment), which adds to the syntactic diversity in the event pairs without compromising their quality. Then, among all verb paraphrase pairs, we filter out only three relation classes, namely *equivalence*, *reverse entailment* and *forward entailment*, as the potential coreferential event pairs. The forward and reverse entailment relations characterize hyponym and hypernym relations, which are not semantically equivalent but can often be coreferential and thus, add diversity to the pairs. Finally, we manually remove noisy event trigger words and cluster the remaining event pairs through pivoting, based on a common event trigger word shared between two paraphrase pairs[8]. Overall, we obtain 1023 clusters with an average of 3.375 event trigger words per cluster.

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

To generate the news discourse structure proposed by Van Dijk (Teun A, 1986; Van Dijk, 1988a,b) and specify the discourse role of a sentence with respect to events in the document, we use the discourse profiling system proposed by Choubey et al. (2020). Note that the above discourse structure is functional (Webber and Joshi, 2012) and does not specify relations between two discourse units. Instead, it classifies each sentence in a document into one of the eight content types. Each content type describes the specific role of a sentence in describing the main event, context informing events, and other historical or future projected events.

The eight content types include *main event* (M1) sentences that describe the most newsworthy event of a news article. Sentences describing events that happen recently and act as triggers for the main event and events that are triggered by the main event constitute the *previous event* (C1) and *consequence* (M2) sentences respectively. The remaining context-informing events and states with temporal co-occurrence with the main event are

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

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

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

covered in *current context* (C2) sentences. In addition to the above four content types, a news article may contain sentences describing lesser relevant events such as *historical events* (D1) that temporally precedes the main event by months and years, *anecdotal events* (D2) that are unverifiable personal account of incidents, *evaluation* (D3) containing reactions from immediate participants, experts or known personalities and *expectation* (D4) that projects the possible consequences of the main event.

Among the eight content types, events described in *main event* sentences are central to the main news topic. They routinely appear in headline and sentences of other content types and consequently are more likely to form event coreference chains. On the contrary, events in the *historical event* content type are restricted to describing certain historical background and might only be mentioned once in the document. Additionally, events mentioned in *previous event* sentences tend to happen before those in *main event* and *consequence* sentences, and are unlikely to be coreferential with the events from the later two content types. Overall, content types provide cues for determining whether the events from a certain sentence possess coreferential event mentions and we leverage them to locate both coreferential and non-coreferential event pairs in a news article. Our event coreference data acquisition method works in two phases.

**Rule-based Filtering to extract Coreferential and Non-coreferential Event Pairs:** In the first phase, we extract both coreferential and non-coreferential event mention pairs based on their respective rules. Specifically, two event mentions from the headline or main event sentences with synonymous event trigger words are identified as coreferential event pairs. Considering that coreferential event mentions are very sparsely distributed, simple trigger-word matching is extremely noisy and damaging when used to train an event coreference classifier. However, narrowing coreferential event mention pairs to synonymous event trigger words from main event sentences or headline significantly eliminates false coreferential event pairs. To get non-coreferential event pairs, we require both trigger words to be non-synonymous and belong to either the same sentence or two sentences of different non-main content types. Further, considering that events in historical event sentences tend to precede the main event by months and years, we identify non-synonymous event pairs with one mention in a historical event sentence and another mention in a main event sentence as non-coreferential. The latter rule allows us to also acquire non-coreferential event pairs with one event from main event sentences, adding to the overall diversity of the acquired dataset.

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

### 3.3 Statistics of Acquired Coreference Data

We use Xinhua news articles[9] from the English Gigaword (Napoles et al., 2012) corpus to acquire coreferential and non-coreferential event pairs using the proposed methodology. We limit the number of coreferential and non-coreferential event pairs for each trigger word to 20 and 200, respectively, to ensure diversity and reduce repetitions of common event trigger words. We compare our acquired event pairs with the KBP 2015 corpus, which has 179 news documents annotated with eight event types and 38 event subtypes. It is the most widely used corpus for training a within-document event coreference resolution system. Table 1 shows the number of event pairs obtained in the first and second phases of our data acquisition strategy and the human-annotated KBP 2015 corpus. Overall, the total number of extracted coreferential event pairs is more than twice the number of pairs in news documents from the KBP 2015 corpus. Note that we can increase the number of acquired pairs by expanding the synonymous event

[9] The discourse profiling system (Choubey et al., 2020) obtains the best performance on Xinhua news articles compared to NYT and Reuters

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

Table 1: Number of coreferential and non-coreferential events pairs acquired through the proposed methodology and the human annotated KBP 2015 corpus.

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

Table 2: Precision (Prec.) and bootstrap 80% confidence interval (80% CI) score of precision for acquired event pairs based on human evaluation.

trigger word list or the unlabeled news article collection.

### 3.4 Manual Evaluation of Acquired Event Pairs

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

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

## 4 Event Coreference Resolution System

We design a neural network-based mention-pair classifier for event coreference resolution. We represent each event pair using 50 context words to the left and right of the first and second event trigger words respectively, and with the maximum of 200 words in between the two event words[10].

Given the event context $(w_1, ., e_1, ., e_2, ., w_n)$, we first transform the context words sequence to word embeddings sequence $(b_{w1}, ., b_{e1}, ., b_{e2}, ., b_{wn})$ using the pre-trained Bert-Large-uncased model (Devlin et al., 2019). Then, we model the semantic associations between two event mentions by measuring similarity between their event embeddings $(b_{e1}, b_{e2})$ through element-wise product and difference. Further, we obtain context embedding $(C)$ through $maxpool$ operation over the word embeddings sequence to model contextual cues. While the context provides important cues for identifying coreferential event mentions, it may not always be relevant for resolving coreference links. For instance, many event trigger word pairs such as (*"injuries"*, *"recommended"*) are extremely unlikely to exhibit coreferential relations irrespective of their context. Therefore, we use the similarity between event embeddings to control the context input and use them only in the scenarios where event trigger words are likely to possess coreferential link. To achieve so, we apply linear neural layer over element-wise product and differences of two event mention embeddings followed by the $sigmoid$ activation, and multiply them with context embedding $C$. Finally, we concatenate the resulting set of embeddings and then use a three-layer feed-forward neural network classifier to score the coreference likelihood. The exact formulation of the coreference classifier is described in Eq. 1.

$$\begin{aligned} (b_{w1}.b_{e1}.b_{e2}.b_{wn}) &= BERT[(w_1.e_1.e_2.w_n)] \in R^{n\times 1024} \\ C &= maxpool(b_{w1}, ., b_{e1}, ., b_{e2}, ., b_{wn}) \in R^{1024} \\ s_1 &= sigmoid(W_1^s(b_{w1} \odot b_{w2}) + b_1^s) \in R^{1024} \\ s_2 &= sigmoid(W_2^s(b_{w1} - b_{w2}) + b_2^s) \in R^{1024} \\ R &= [b_{w1} \odot b_{w2}; b_{w1} - b_{w2}; s_1 \odot C; s_2 \odot C] \in R^{4096} \\ \hat{y_i} &= W_3(gelu(W_2(gelu(W_3 R + b_3)) + b_2)) + b_3 \in R \end{aligned} \tag{1}$$

We train the model using binary cross-entropy

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

loss. During inference, we use the best-first clustering approach, where we select the antecedent having the highest pairwise coreference score based on the coreference classifier, to build event chains.

## 5 Experiments

### 5.1 Datasets and Evaluation Setup

We use the news documents from the KBP 2016 for validation, and use news documents from KBP 2017 and RED corpora as well as discussion forum documents from the KBP 2017 corpus to evaluate the usefulness of our acquired data[11]. KBP 2016, KBP 2017 and RED corpora contain 85, 83 and 30 news documents respectively, and KBP 2017 has 84 discussion forum documents. KBP corpora have been widely used for evaluating in-document event coreference resolution systems. We further evaluate our models on the RED corpus to examine systems' performance across different event types. KBP 2016 and 2017 corpora are annotated using a subset of 20 subtypes from 38 subtypes used in KBP 2015. On the contrary, RED documents are comprehensively annotated with event coreference relations with no restriction on event types or subtypes, thus, allowing us to evaluate coreference resolution performance on a broad range of events. Besides, we evaluate the performance of models across text genres by evaluating our models trained with news articles on KBP 2017 discussion forum documents.

Following previous work on event coreference resolution, we evaluate all the event coreference resolution systems using the official KBP 2017 scorer v1.8. The scorer employs four coreference scoring measures, namely $B^3$ (Bagga and Baldwin, 1998), CEAFe (Luo, 2005), MUC (Vilain et al., 1995) and BLANC (Recasens and Hovy, 2011) and the unweighted average of their F1 scores $AVG_{F1}$. In addition, since MUC directly evaluates pairwise coreference links, we also report MUC precision and recall scores.

### 5.2 Implementation Details

We use an ensemble of multi-layer feed-forward neural network classifiers to identify event mentions (Choubey and Huang, 2018) for both news and discussion forum documents in KBP 2017 corpus. For the RED corpus, we use gold event mentions as that event extraction system can identify events from only eight event types annotated in KBP 2015 corpus. The coreference classifier uses a three-layer feed-forward neural network with 1024-512-1 units for scoring coreference likelihood. Two single-neural layers, used to transform element-wise dot product and difference between two event embeddings used for controlling context input, use 1024 units each. All hidden activations are followed by dropout with the rate of 0.1 for regularization (Srivastava et al., 2014). All models are trained using AdamW optimizer (Loshchilov and Hutter, 2017; Kingma and Ba, 2014) with four different learning rates (1e-4, 5e-5, 1e-5, 5e-6) and for maximum of 100,000 updates. We use the batch size of 16 and evaluate the model after every 5,000 steps. The epoch and learning rate yielding the best validation performance, average F1 score on KBP 2016 news documents, are used to obtain the final model. Bert model is kept fixed during the training. All experiments are performed on NVIDIA GTX 2080 Ti 11GB using PyTorch 1.2.0+cu92 (Paszke et al., 2019) and HuggingFace Transformer libraries (Wolf et al., 2019).

### 5.3 Baseline Systems

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

**Feature-based Classifier**: The neural network classifier that uses GloVe (Pennington et al., 2014) based event trigger word embeddings and binary features indicating argument overlaps.

**Choubey and Huang (2018)**: It models correlations between event coreference chains and document topic structures through a heuristics-based ILP formulation and has achieved the best event coreference resolution performance to date on both KBP 2016 and KBP 2017 datasets.

### 5.4 Our Systems

**KBP 2015, Paraphrase-based pairs, Post-Filtering Paraphrase pairs and KBP 2015+Post-Filtering Paraphrase pairs**: The mention pair model, proposed in § 4, trained on different combinations of acquired and human-annotated datasets. *KBP 2015* is trained on event pairs from news docu-

[11] ECB+ (Cybulska and Vossen, 2014) is another popular dataset for evaluating event coreference resolution. However, documents in ECB+ are selectively annotated, comprising only of event mentions and within-document coreference chains that are relevant to cross-document event coreference chains. Since our data acquisition methodology is designed for collecting within-document event pairs, we decided to exclude evaluations on the ECB+ corpus.

ments in the KBP 2015 corpus. *Paraphrase-based pairs* is trained on paraphrase event pairs without rules-based filtering (§3.1). *Post-Filtering Paraphrase pairs* is trained on paraphrase event pairs that are filtered using rules defined over news discourse structure (§3.2). *KBP 2015+Post-Filtering Paraphrase pairs* is trained on aggregation of KBP 2015 and Post-Filtering Paraphrase event pairs.

**Student Training**: The mention pair model trained using the recently proposed self-training approach with a student network (Xie et al., 2019). We first train a teacher mention pair model on the KBP 2015 corpus, then use the teacher model to annotate samples from unannotated news articles. We use the same set of event pairs from Xinhua articles in the Gigaword corpus, set the same upper bound of 20 coreferential and 200 non-coreferential pairs per event trigger word. Also, to allow fair comparisons, we selected only high scoring event pairs (likelihood $\geq$ 0.9) and collected 11,390 coreferential and 272,083 non-coreferential pairs. Finally, we train a new student network with the combined KBP 2015 and teacher-annotated event pairs.

**Masked Training**: The mention pair model trained on all annotated and automatically acquired (or teacher annotated in case of student training model) event pairs. However, to limit the over-dependence on lexical features[12], we replace both the event trigger words with the *[MASK]* token for all acquired event pairs. Annotated event pairs from KBP 2015 are left unchanged.

### 5.5 Results and Analysis

The first segment in Table 3 shows the results for all models on KBP 2017 news articles corpus. The mention-pair model trained on KBP 2015 corpus using pre-trained language model and larger event context outperforms both local feature-based as well as the discourse-structure aware previous best model (Choubey and Huang, 2018), outperforming Choubey and Huang (2018) by 2.26 points in average F1 score. The improvement is consistent across all metrics. Specifically, the used mention pair model gains MUC F1 score by 9.76 and 3.33 points over feature-based and discourse aware systems, indicating that BERT-based embedding is more effective in resolving coreference links without exclusively modeling event-arguments or discourse-related features. The model trained on event pairs acquired following the proposed automatic strategy also outperforms Choubey and Huang (2018) by 1.24 and 0.56 points on MUC F1 and average F1 scores respectively. However, this model does worse than the equivalent model trained on KBP 2015 data, which can be explained by the related distribution of KBP 2015 and KBP 2017 datasets. Overall, training the model on KBP 2015 data combined with the acquired event pairs performs the best, outperforming both models trained on KBP 2015 only and the one trained with student training by 1.04 and 0.14 points respectively.

As shown in the second segment of Table 3, the improvement in the average F1 of the model trained on KBP 2015 over the trigger match baseline reduces to 2.3 points on the RED news articles corpus, compared to 5.69 points on KBP 2017 news articles. Mainly, RED annotates all event types while KBP has only 8 event types, and the change in event domains affects the overall performance gain of model. The model trained on our Post-Filtering Paraphrase event pairs performs similarly to the one trained on KBP 2015, implying that the former generalizes similarly to the model trained on human-annotated data when applied to new data out of the training data distribution. Similar to the performance gain on KBP 2017 news articles, combining both KBP 2015 and acquired event pairs improves the average F1 on RED news articles, achieving the highest average F1 gain of 3.98 points against the trigger match baseline. Note that student training also improves performance on RED news articles. However, it is 1.26 points lower on average F1 score than the *KBP 2015+Post-Filtering Paraphrase pairs* model.

In the third segment of Table 3, we compare the performance of all models on a different text genre by evaluating them on the discussion forum documents from the KBP 2017 corpus. With shared event types, the model trained on KBP 2015 achieves the best result with 1.76 points improvement in the average F1 score over the lemma match baseline. The model trained using acquired event pairs, *Post-Filtering Paraphrase pairs*, achieves performance comparable to the model trained on KBP 2015. However, combining the KBP 2015 data with acquired event pairs (the model *KBP 2015+Post-Filtering Paraphrase pairs*) does not further improve the performance. Overall, we observe that none of the models obtain substantial performance improvement. The smaller improvements

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

| Model | $b^3_{F1}$ | $ceafe_{F1}$ | $muc_R$ | $muc_P$ | $muc_{F1}$ | $blanc_{F1}$ | $AVG_{F1}$ |
|---|---|---|---|---|---|---|---|
| **KBP 2017 News Articles** | | | | | | | |
| Trigger Match | 48.96 | 45.67 | 26.16 | 36.63 | 30.52 | 29.30 | 38.61 |
| Trigger Match+Paraphrase | 48.92 | 45.35 | 27.36 | 36.41 | 31.25 | 29.83 | 38.84 |
| Feature-based Classifier | 50.24 | 48.47 | - | - | 30.81 | 29.94 | 39.87 |
| Choubey and Huang (2018) | 50.35 | 48.61 | - | - | 37.24 | 31.94 | 42.04 |
| KBP 2015 | 51.57 | **50.90** | 33.91 | 50.49 | 40.57 | 34.15 | 44.30 |
| Paraphrase-based pairs | 48.10 | 42.36 | 38.05 | 37.01 | 37.52 | 31.64 | 39.91 |
| Post-Filtering Paraphrase pairs | 50.94 | 47.81 | 31.77 | 48.77 | 38.48 | 33.19 | 42.60 |
| KBP 2015+Post-Filtering Paraphrase pairs | **52.29** | 50.50 | 35.24 | **55.23** | 43.03 | 35.53 | 45.34 |
| Masked Training | 52.10 | 50.72 | 36.31 | 53.02 | 43.10 | 35.51 | **45.36** |
| Student Training | 51.85 | 49.91 | **38.18** | 49.73 | **43.20** | **35.83** | 45.20 |
| Masked Training | 51.91 | 50.12 | 37.11 | 50.82 | 42.90 | 35.50 | 45.11 |
| **RED News Articles** | | | | | | | |
| Trigger Match | 88.07 | 84.21 | 42.63 | 35.14 | 38.52 | 64.34 | 68.78 |
| Trigger Match+Paraphrase | 87.18 | 83.09 | 47.16 | 33.87 | 39.43 | 64.88 | 68.65 |
| KBP 2015 | 88.33 | 85.48 | 52.38 | 39.08 | 44.76 | 65.77 | 71.08 |
| Paraphrase-based pairs | 82.01 | 76.74 | **68.02** | 30.39 | 42.01 | 63.09 | 65.96 |
| Post-Filtering Paraphrase pairs | **89.25** | 86.70 | 47.39 | 41.63 | 44.32 | 63.75 | 71.0 |
| KBP 2015+Post-Filtering Paraphrase pairs | **89.25** | **86.96** | 56.00 | **43.40** | 48.91 | 66.74 | 72.96 |
| Masked Training | 89.16 | 86.90 | 58.04 | 43.31 | **49.61** | **67.50** | **73.29** |
| Student Training | 87.91 | 84.95 | 59.18 | 39.30 | 47.23 | 66.70 | 71.70 |
| Masked Training | 88.11 | 84.92 | 58.50 | 39.69 | 47.29 | 67.44 | 71.94 |
| **KBP 2017 Discussion Forum Documents** | | | | | | | |
| Trigger Match | 37.29 | 39.15 | 20.36 | 19.06 | 19.69 | 18.25 | 28.59 |
| Trigger Match + Paraphrase | 36.94 | 38.52 | 21.26 | 19.13 | 20.14 | 18.14 | 28.44 |
| KBP 2015 | 38.11 | 38.67 | 25.33 | **23.76** | 24.52 | **20.10** | **30.35** |
| Paraphrase-based pairs | 35.58 | 34.30 | 28.65 | 21.37 | 24.48 | 19.19 | 28.39 |
| Post-Filtering Paraphrase pairs | **39.12** | **41.52** | 17.34 | 20.75 | 18.89 | 18.81 | 29.59 |
| KBP 2015+Post-Filtering Paraphrase pairs | 37.43 | 38.16 | 26.24 | 22.27 | 24.09 | 20.01 | 29.92 |
| Masked Training | 38.33 | 39.64 | 21.71 | 20.68 | 21.19 | 19.19 | 29.59 |
| Student Training | 36.80 | 36.68 | **28.80** | 22.68 | **25.38** | 20.08 | 29.73 |
| Masked Training | 37.06 | 38.01 | 22.77 | 20.00 | 21.29 | 17.51 | 28.47 |

Table 3: Results for event coreference resolution systems on the KBP 2017 and RED corpora. Feature-based Classifier results are directly taken from Choubey and Huang (2018). The results are statistically significant using bootstrap and permutation test (Dror et al., 2018) with $p<0.01$ between *Post-Filtering Paraphrase pairs* and *Paraphrase-based Pairs* and $p<0.002$ between *KBP 2015+Post-Filtering Paraphrase pairs+Masked Training* and *KBP 2015* models on both KBP 2017 and RED news articles test sets. Further, results for *KBP 2015+Post-Filtering Paraphrase pairs+Masked Training* are statistically significant compared to both *Student Training* and *Student Training+Masked Training* with $p<0.002$ on the RED news articles test set.

for all models on discussion forum documents, with the increased data size, also indicate the need for specialized learning algorithms to build a model that can generalize to a new text genre.

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

**Distributional Analysis of Predicted Coreferential Event Pairs across different Discourse Content Type Pairs:** Finally, we analyze the distribution of predicted coreferential event pairs across sentence pairs with different discourse content types on the validation dataset. We use the gold coreferential event pairs to identify the top 10 content type pairs of sentences that most frequently contain coreferential event mention pairs. Then, for the models trained on KBP 2015, Post-Filtering Paraphrase pairs and their combination with masked training, we report true-positive, false-positive, and false-negative predictions, shown in Figure 1. To ensure uniformity with rules used in

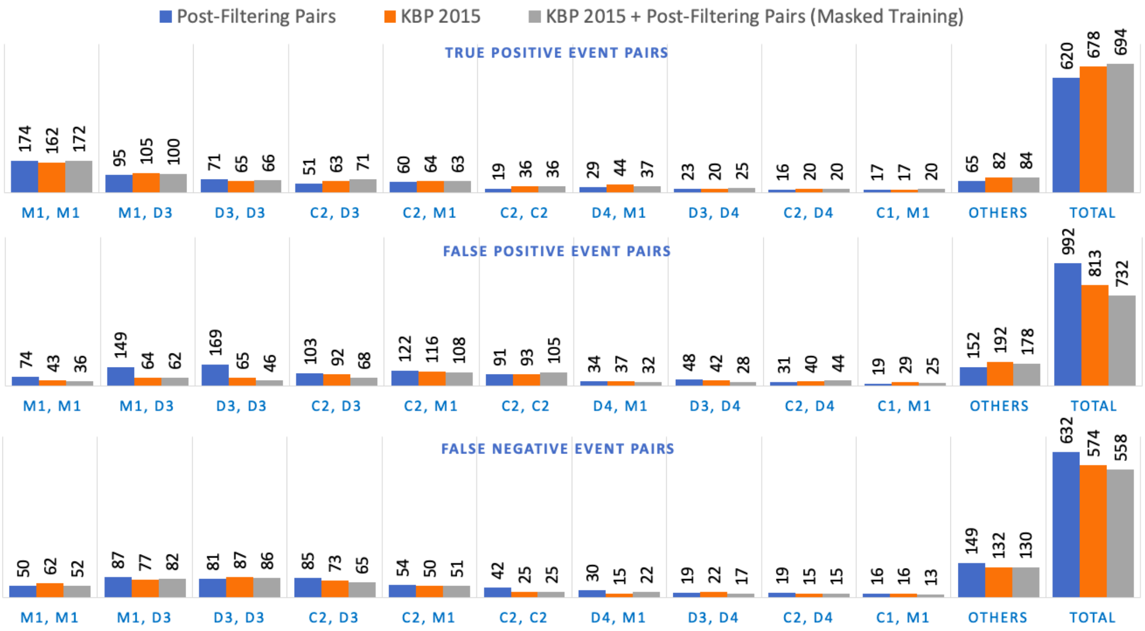

Figure 1: Distributions of Predicted Coreferential Event Pairs across different Discourse Content Type Pairs.

§3.2, we merge the headline with main sentences.

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

In comparison, models trained on KBP 2015 or combined KBP 2015 and Post-Filtering pairs have lower false-positives while exhibiting similar distributions for true-positive predictions. Intuitively, despite second phase bootstrapping to include non-coreferential paraphrase pairs, the model trained solely on acquired event pairs focuses on lexical features more than the model trained on human-annotated corpus. On the other hand, masked training effectively overcomes excessive reliance on lexical cues and helps achieve a higher true positive rate without increasing false positives.

## 6 Conclusions and Future Work

We presented an automatic data acquisition strategy for event coreference resolution by mining the functional news discourse structure. We performed both qualitative and empirical studies to determine the effectiveness of our proposed strategy. We found that the model trained on automatically acquired event pairs performs similarly to the model trained on human-annotated corpus when evaluated on the test set covering general event domains. Further, augmenting acquired event pairs to existing human-annotated data improves the performance of the model on both training-domain and broader domain test sets. For future work, we intend to develop new training algorithms to improve the generalization capability of models on a new text genre. Further, we plan to evaluate a similar event coreference data acquisition strategy for new languages.

## 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

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*, volume 1, pages 563–566. Granada.

Shany Barhom, Vered Shwartz, Alon Eirew, Michael Bugert, Nils Reimers, and Ido Dagan. 2019. Revisiting joint modeling of cross-document entity and event coreference resolution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4179–4189.

Cosmin 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*, pages 1412–1422.

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

Chen Chen and Vincent Ng. 2014. Sinocoreferencer: An end-to-end chinese event coreference resolver. In *LREC*, volume 2, page 3. Citeseer.

Chen Chen and Vincent Ng. 2015. Chinese zero pronoun resolution: A joint unsupervised discourse-aware model rivaling state-of-the-art resolvers. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 320–326, Beijing, China. Association for Computational Linguistics.

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*, pages 54–57. Association for Computational Linguistics.

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*, pages 17–22. Association for Computational Linguistics.

Prafulla Kumar Choubey and Ruihong Huang. 2017. 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. 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, Aaron Lee, Ruihong Huang, and Lu Wang. 2020. Discourse as a function of event: Profiling discourse structure in news articles around the main event. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5374–5386, Online. Association for Computational Linguistics.

Andres Cremisini and Mark Finlayson. 2020. New insights into cross-document event coreference: Systematic comparison and a simplified approach. In *Proceedings of the First Joint Workshop on Narrative Understanding, Storylines, and Events*, pages 1–10, Online. Association for Computational Linguistics.

Agata Cybulska and Piek Vossen. 2014. Using a sledgehammer to crack a nut? lexical diversity and event coreference resolution. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 4545–4552, Reykjavik, Iceland. European Languages Resources Association (ELRA).

Agata Cybulska and Piek Vossen. 2015. 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. Ppdb: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 758–764.

Kian Kenyon-Dean, Jackie Chi Kit Cheung, and Doina Precup. 2018. Resolving event coreference with supervised representation learning and clustering-oriented regularization. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 1–10.

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

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*, pages 489–500. Association for Computational Linguistics.

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

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Jing Lu and Vincent Ng. 2018. Event coreference resolution: A survey of two decades of research. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 5479–5486. International Joint Conferences on Artificial Intelligence Organization.

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.

Yaojie Lu, Hongyu Lin, Jialong Tang, Xianpei Han, and Le Sun. 2020. End-to-end neural event coreference resolution.

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

Katie McConky, Rakesh Nagi, Moises Sudit, and William Hughes. 2012. Improving event coreference by context extraction and dynamic feature weighting. In *2012 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support*, pages 38–43. IEEE.

Yehudit Meged, Avi Caciularu, Vered Shwartz, and Ido Dagan. 2020. Paraphrasing vs coreferring: Two sides of the same coin.

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

Tim O'Gorman, Kristin Wright-Bettner, and Martha Palmer. 2016. Richer event description: Integrating event coreference with temporal, causal and bridging annotation. In *Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016)*, pages 47–56, Austin, Texas. Association for Computational Linguistics.

Arun Pandian, Lamana Mulaffer, Kemal Oflazer, and Amna AlZeyara. 2020. Precision event coreference resolution using neural network classifiers. *Computación y Sistemas*, 24(1).

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2015. PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 425–430, Beijing, China. Association for Computational Linguistics.

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

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.

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

Ellen Riloff. 1996. Automatically generating extraction patterns from untagged text. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence - Volume 2*, AAAI'96, page 1044–1049. AAAI Press.

Ellen Riloff and Janyce Wiebe. 2003. Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 105–112.

Hossein Sahlani, Maryam Hourali, and Behrouz Minaei-Bidgoli. 2020. Coreference resolution using semantic features and fully connected neural network in the persian language. *International Journal of Computational Intelligence Systems*, 13:1002–1013.

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

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.

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.

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*, pages 45–52. Association for Computational Linguistics.

Bonnie Webber and Aravind Joshi. 2012. Discourse structure and computation: Past, present and future. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pages 42–54, Jeju Island, Korea. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. 2019. Self-training with noisy student improves imagenet classification.

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.

Dian Yu, Xiaoman Pan, Boliang Zhang, Lifu Huang, Di Lu, Spencer Whitehead, and Heng Ji. 2016. Rpi blender tac-kbp2016 system description. In *TAC*.

Xinyu Zuo, Yubo Chen, Kang Liu, and Jun Zhao. 2019. Event co-reference resolution via a multi-loss neural network without using argument information. *Science China Information Sciences*, 62(11).