A Hierarchical Distance-dependent Bayesian Model for Event Coreference Resolution

Bishan Yang **Claire Cardie**
Department of Computer Science
Cornell University
{bishan, cardie}@cs.cornell.edu

Peter Frazier
School of Operations Research
and Information Engineering
Cornell University
pf98@cornell.edu

Abstract

We present a novel hierarchical distance-dependent Bayesian model for event coreference resolution. While existing generative models for event coreference resolution are completely unsupervised, our model allows for the incorporation of pairwise distances between event mentions — information that is widely used in supervised coreference models to guide the generative clustering processing for better event clustering both within and across documents. We model the distances between event mentions using a feature-rich learnable distance function and encode them as Bayesian priors for nonparametric clustering. Experiments on the ECB+ corpus show that our model outperforms state-of-the-art methods for both within- and cross-document event coreference resolution.

1 Introduction

The task of *event coreference resolution* consists of identifying text snippets that describe events, and then clustering them such that all *event mentions* in the same partition refer to the same unique event. Event coreference resolution can be applied within a single document or across multiple documents and is crucial for many natural language processing tasks including topic detection and tracking, information extraction, question answering and textual entailment (Bejan and Harabagiu, 2010). More importantly, event coreference resolution is a necessary component in any reasonable, broadly applicable computational model of natural language understanding (Humphreys et al., 1997).

In comparison to entity coreference resolution (Ng, 2010), which deals with identifying and grouping noun phrases that refer to the same discourse entity, event coreference resolution has not been extensively studied. This is, in part, because events typically exhibit a more complex structure than entities: a single event can be described via multiple event mentions, and a single event mention can be associated with multiple *event arguments* that characterize the participants in the event as well as spatio-temporal information (Bejan and Harabagiu, 2010). Hence, the coreference decisions for event mentions usually require the interpretation of event mentions and their arguments in context. See, for example, Figure 1, in which five event mentions across two documents all refer to the same underlying event: *Plane bombs Yida camp*.

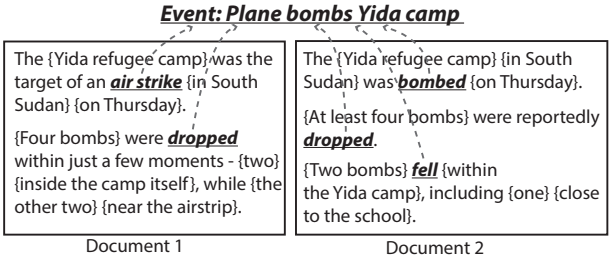

Figure 1: Examples of event coreference. Mutually coreferent event mentions are underlined and in boldface; participant and spatio-temporal information for the highlighted event is marked by curly brackets.

Most previous approaches to event coreference resolution (e.g., Ahn (2006), Chen et al. (2009)) operated by extending the supervised pairwise classi-

Transactions of the Association for Computational Linguistics, vol. 3, pp. 517–528, 2015. Action Editor: Hwee Tou Ng.
Submission batch: 4/2015; Revision batch: 7/2015; Published 9/2015.

fication model that is widely used in entity coreference resolution (e.g., Ng and Cardie (2002)). In this framework, pairwise distances between event mentions are modeled via event-related features (e.g., that indicate event argument compatibility), and agglomerative clustering is applied to greedily merge event mentions into clusters. A major drawback of this general approach is that it makes hard decisions on the merging and splitting of clusters based on heuristics derived from the pairwise distances. In addition, it only captures pairwise coreference decisions within a single document and can not account for signals that commonly appear across documents. More recently, Bejan and Harabagiu (2010; 2014) proposed several nonparametric Bayesian models for event coreference resolution that probabilistically infer event clusters both within a document and across multiple documents. Their method, however, is completely unsupervised, and thus can not encode any readily available supervisory information to guide the model toward better event clustering.

To address these limitations, we propose a novel Bayesian model for within- and cross-document event coreference resolution. It leverages supervised feature-rich modeling of pairwise coreference relations and generative modeling of cluster distributions, and thus allows for both probabilistic inference over event clusters and easy incorporation of pairwise linking preferences. Our model builds on the framework of the distance-dependent Chinese restaurant process (DDCRP) (Blei and Frazier, 2011), which was introduced to incorporate data dependencies into nonparametric clustering models. Here, however, we extend the DDCRP to allow the incorporation of feature-based, learnable distance functions as clustering priors, thus encouraging event mentions that are close in meaning to belong to the same cluster. In addition, we introduce to the DDCRP a representational hierarchy that allows event mentions to be grouped within a document and within-document event clusters to be grouped across documents.

To investigate the effectiveness of our approach, we conduct extensive experiments on the ECB+ corpus (Cybulska and Vossen, 2014b), an extension to EventCorefBank (ECB) (Bejan and Harabagiu, 2010) and the largest corpus available that contains event coreference annotations within and across documents. We show that integrating pairwise learning of event coreference relations with unsupervised hierarchical modeling of event clustering achieves promising improvements over state-of-the-art approaches for within- and cross-document event coreference resolution.

2 Related Work

Coreference resolution in general is a difficult natural language processing (NLP) task and typically requires sophisticated inferentially-based knowledge-intensive models (Kehler, 2002). Extensive work in the literature focuses on the problem of entity coreference resolution and many techniques have been developed, including rule-based deterministic models (e.g. Cardie and Wagstaff (1999), Raghunathan et al. (2010), Lee et al. (2011)) that traverse over mentions in certain orderings and make deterministic coreference decisions based on all available information at the time; supervised learning-based models (e.g. Stoyanov et al. (2009), Rahman and Ng (2011), Durrett and Klein (2013)) that make use of rich linguistic features and the annotated corpora to learn more powerful coreference functions; and finally, unsupervised models (e.g. Bhattacharya and Getoor (2006), Haghighi and Klein (2007, 2010)) that successfully apply generative modeling to the coreference resolution problem.

Event coreference resolution is a more complex task than entity coreference resolution (Humphreys et al., 1997) and also has been relatively less studied. Existing work has adapted similar ideas to those used in entity coreference. Humphreys et al. (1997) first proposed a deterministic clustering mechanism to group event mentions of prespecified types based on hard constraints. Later approaches (Ahn, 2006; Chen et al., 2009) applied learning-based pairwise classification decisions using event-specific features to infer event clustering. Bejan and Harabagiu (2010; 2014) proposed several unsupervised generative models for event mention clustering based on the hierarchical Dirichlet process (HDP) (Teh et al., 2006). Our approach is related to both supervised clustering and generative clustering approaches. It is a nonparametric Bayesian model in nature but encodes rich linguistic features in clustering priors. More recent work

modeled both entity and event information in event coreference. Lee et al. (2012) showed that iteratively merging entity and event clusters can boost the clustering performance. Liu et al. (2014) demonstrated the benefits of propagating information between event arguments and event mentions during a post-processing step. Other work modeled event coreference as a predicate argument alignment problem between pairs of sentences, and trained classifiers for making alignment decisions (Roth and Frank, 2012; Wolfe et al., 2015). Our model also leverages event argument information into the decisions of event coreference but incorporates it into Bayesian clustering priors.

Most existing coreference models, both for events and entities, focus on solving the within-document coreference problem. Cross-document coreference has attracted less attention due to lack of annotated corpora and the requirement for larger model capacity. Hierarchical models (Singh et al., 2010; Wick et al., 2012; Haghighi and Klein, 2007) have been popular choices for cross-document coreference as they can capture coreference at multiple levels of granularities. Our model is also hierarchical, capturing both within- and cross-document coreference.

Our model is also closely related to the distance-dependent Chinese Restaurant Process (DDCRP) (Blei and Frazier, 2011). The DDCRP is an infinite clustering model that can account for data dependencies (Ghosh et al., 2011; Socher et al., 2011). But it is a flat clustering model and thus cannot capture hierarchical structure that usually exists in large data collections. Very little work has explored the use of DDCRP in hierarchical clustering models. Kim and Oh (2011; Ghosh et al. (2011) combined a DDCRP with a standard CRP in a two-level hierarchy analogous to the HDP with restricted distance functions. Ghosh et al. (2014) proposed a two-level DDCRP with data-dependent distance-based priors at both levels. Our model is also a two-level DDCRP model but differs in that its distance function is learned using a feature-rich log-linear model. We also derive an effective Gibbs sampler for posterior inference.

Action	*bombs*
Participant	*Sudan, Yida refugee camp*
Time	*Thursday, Nov 10, 2011*
Location	*South Sudan*

Table 1: Mentions of event components

3 Problem Formulation

We adopt the terminology from ECB+ (Cybulska and Vossen, 2014b), a corpus that extends the widely used EventCorefBank (ECB (Bejan and Harabagiu, 2010)). An **event** is something that happens or a situation that occurs (Cybulska and Vossen, 2014a). It consists of four components: (1) an *Action*: what happens in the event; (2) *Participants*: who or what is involved; (3) a *Time*: when the event happens; and (4) a *Location*: where the event happens. We assume that each document in the corpus consists of a set of mentions — text spans — that describe event actions, their participants, times, and locations. Table 1 shows examples of these in the sentence "Sudan bombs Yida refugee camp in South Sudan on Thursday, Nov 10th, 2011."

In this paper, we also use the term **event mention** to refer to the mention of an event action, and **event arguments** to refer collectively to mentions of the participants, times and locations involved in the event. Event mentions are usually noun phrases or verb phrases that clearly describe events. Two event mentions are considered **coreferent** if they refer to the same actual event, i.e. a situation involving a particular combination of action, participants, time and location. Note that in text, not all event arguments are always present for an event mention; they may even be distributed over different sentences. Thus whether two event mentions are coreferential should be determined based on the context. For example, in Figure 1, the event mention *dropped* in DOCUMENT 1 corefers with *air strike* in the same document as they describe the same event, *Plane bombs Yida camp*, in the discourse context; it also corefers with *dropped* in DOCUMENT 2 based on the contexts of both documents.

The problem of event coreference resolution can be divided into two sub-problems: (1) **event extraction**: extracting event mentions and event arguments, and (2) **event clustering**: grouping event

mentions into clusters according to their coreference relations. We consider both within- and cross-document event coreference resolution and hypothesize that leveraging context information from multiple documents will improve both within- and cross-document coreference resolution. In the following, we first describe the event extraction step and then focus on the event clustering step.

4 Event Extraction

The goal of event extraction is to extract from a text all event mentions (actions) and event arguments (the associated participants, times and locations). One might expect that event actions could be extracted reasonably well by identifying verb groups; and event arguments, by applying semantic role labeling (SRL) to identify, for example, the *Agent* and *Patient* of each predicate. Unfortunately, most SRL systems only handle verbal predicates and so would miss event mentions described via noun phrases. In addition, SRL systems are not designed to capture event-specific arguments. Accordingly, we found that a state-of-the-art SRL system (SwiRL (Surdeanu et al., 2007)) extracted only 56% of the actions, 76% of participants, 65% of times and 13% of locations for events in a development set of ECB+ based on a head word matching evaluation measure. (We provide dataset details in Section 6.)

To produce higher recall, we adopt a supervised approach and train an event extractor using sentences from ECB+, which are annotated for event actions, participants, times and locations. Because these mentions vary widely in their length and grammatical type, we employ semi-Markov CRFs (Sarawagi and Cohen, 2004) using the loss-augmented objective of Yang and Cardie (2014) that provides more accurate detection of mention boundaries. We make use of a rich feature set that includes word-level features such as unigrams, bigrams, POS tags, WordNet hypernyms, synonyms and FrameNet semantic roles, and phrase-level features such as phrasal syntax (e.g., NP, VP) and phrasal embeddings (constructed by averaging word embeddings produced by word2vec (Mikolov et al., 2013)). Our experiments on the same (held-out) development data show that the semi-CRF-based extractor correctly identifies 95% of actions, 90% of participants, 94% of times and 74% of locations again based on head word matching.

Note that the semi-CRF extractor identifies event mentions and event arguments but not relationships among them, i.e. it does not associate arguments with an event mention. Lacking supervisory data in the ECB+ corpus for training an event action-argument relation detector, we assume that all event arguments identified by the semi-CRF extractor are related to all event mentions in the same sentence and then apply SRL-based heuristics to augment and further disambiguate intra-sentential action-argument relations (using the SwiRL SRL). More specifically, we link each verbal event mention to the participants that match its *ARG0*, *ARG1* or *ARG2* semantic role fillers; similarly, we associate with the event mention the time and locations that match its *AM-TMP* and *AM-LOC* role fillers, respectively. For each nominal event mention, we associate those participants that match the possessor of the mention since these were suggested in Lee et al. (2012) as playing the *ARG0* role for nominal predicates.

5 Event Clustering

Now we describe our proposed Bayesian model for event clustering. Our model is a hierarchical extension of the distance-dependent Chinese Restaurant Process (DDCRP). It first groups event mentions within a document to form within-document event cluster and then groups these event clusters across documents to form global clusters. The model can account for the similarity between event mentions during the clustering process, putting a bias toward clusters comprised of event mentions that are similar to each other based on the context. To capture event similarity, we use a log-linear model with rich syntactic and semantic features, and learn the feature weights using gold-standard data.

5.1 Distance-dependent Chinese Restaurant Process

The Distance-dependent Chinese Restaurant Process (DDCRP) is a generalization of the Chinese Restaurant process (CRP) that models distributions over partitions. In a CRP, the generative process can be described by imagining data points as customers

in a restaurant and the partitioning of data as tables at which the customers sit. The process randomly samples the table assignment for each customer sequentially: the probability of a customer sitting at an existing table is proportional to the number of customers already sitting at that table and the probability of sitting at a new table is proportional to a scaling parameter. For each customer sitting at the same table, an observation can be drawn from a distribution determined by the parameter associated with that table. Despite the sequential sampling process, the CRP makes the assumption of exchangeability: the permutation of the customer ordering does not change the probability of the partitions.

The exchangeability assumption may not be reasonable for clustering data that has clear interdependencies. The DDCRP allows the incorporation of data dependencies in infinite clustering, encouraging data points that are closer to each other to be grouped together. In the generative process, instead of directly sampling a table assignment for each customer, it samples a customer link, linking the customer to another customer or itself. The clustering can be uniquely constructed once the customer links are determined for all customers: two customers belong to the same cluster if and only if one can reach the other by traversing the customer links (treating these links as undirected).

More formally, consider a sequence of customers $1, ..., n$, and denote $\mathbf{a} = (a_1, ..., a_n)$ as the assignments of the customer links. $a_i \in \{1, \ldots, n\}$ is drawn from

$$p(a_i = j | F, \alpha) \propto \begin{cases} F(i,j), & j \neq i \\ \alpha, & j = i \end{cases} \quad (1)$$

where F is a distance function and $F(i,j)$ is a value that measures the distance between customer i and j. α is a scaling parameter, measuring self-affinity. For each customer, the observation is generated by the per-table parameters as in the CRP. A DDCRP is said to be *sequential* if $F(i,j) = 0$ when $i < j$, so customers may link only to themselves, and to previous customers.

5.2 A Hierarchical Extension of the DDCRP

We can model within-document coreference resolution using a sequential DDCRP. Imagining customers as event mentions and the restaurant as a document, each mention can either refer to an antecedent mention in the document or no other mentions, starting the description of a new event. However, the coreference relations may also exist across documents — the same event may be described in multiple documents. Thus it is ideal to have a two-level clustering model that can group event mentions within a document and further group them across documents. Therefore we propose a hierarchical extension of the DDCRP (HDDCRP) that employs a DDCRP twice: the first-level DDCRP links mentions based on within-document distances and the-second level DDCRP links the within-document clusters based on cross-document distances, forming larger clusters in the corpus.

The generative process of an HDDCRP can be described using the same "Chinese Restaurant" metaphor. Imagine a collection of documents as a collection of restaurants, and the event mentions in each document as customers entering a restaurant. The local (within-document) event clusters correspond to *tables*. The global (within-corpus) event clusters correspond to *menus* (tables that serve the same menu belong to the same cluster). The hidden variables are the customer links and the table links. Figure 2 shows a configuration of these variables and the corresponding clustering structure.

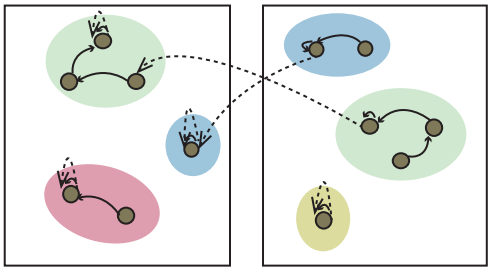

Figure 2: A cluster configuration generated by the HDDCRP. Each restaurant is represented by a rectangle. The small green circles represent customers. The ovals represent tables and the colors reflect the clustering. Each customer is assigned a customer link (a solid arrow), linking to itself or another customer in the same restaurant. The customer who first sits at the table is assigned a table link (a dashed arrow), linking to itself or another customer in a different restaurant, resulting in the linking of two tables.

More formally, the generative process for the HDDCRP can be described as follows:

1. For each restaurant $d \in \{1, ..., D\}$, for each

customer $i \in \{1, ..., n_d\}$, sample a customer link using a sequential DDCRP:

$$p(a_{i,d} = (j, d)) \propto \begin{cases} F_d(i, j), & j < i \\ \alpha_d, & j = i \\ 0, & j > i \end{cases} \quad (2)$$

2. For each restaurant $d \in \{1, ..., D\}$, for each table t, sample a table link for the customer (i, d) who first sits at t using a DDCRP:

$$p(c_{i,d} = (j, d')) \propto \begin{cases} F_0((i, d), (j, d')), & j \in \{1, ..., n_{d'}\}, d' \neq d \\ \alpha_0, & j = i, d' = d \end{cases} \quad (3)$$

3. Calculate clusters $\mathbf{z}(\mathbf{a}, \mathbf{c})$ by traversing all the customer links $\mathbf{a}$ and the table links $\mathbf{c}$. Two customers are in the same cluster if and only if there is a path from one to the other along the links, where we treat both table and customer links as undirected.

4. For each cluster $k \in \mathbf{z}(\mathbf{a}, \mathbf{c})$, sample parameters $\phi_k \sim G_0(\lambda)$.

5. For each customer i in cluster k, sample an observation $x_i \sim p(\cdot|\phi_{z_i})$ where $z_i = k$.

$F_{1:D}$ and F_0 are distance functions that map a pair of customers to a distance value. We will discuss them in detail in Section 5.4.

5.3 Posterior Inference with Gibbs Sampling

The central computation problem for the HDDCRP model is posterior inference — computing the conditional distribution of the hidden variables given the observations $p(\mathbf{a}, \mathbf{c}|\mathbf{x}, \alpha_0, F_0, \alpha_{1:D}, F_{1:D})$. The posterior is intractable due to a combinatorial number of possible link configurations. Thus we approximate the posterior using Markov Chain Monte Carlo (MCMC) sampling, and specifically using a Gibbs sampler.

In developing this Gibbs sampler, we first observe that the generative process is equivalent to one that, in step 2 samples a table link for *all* customers, and then in step 3, when calculating $\mathbf{z}(\mathbf{a}, \mathbf{c})$, includes only those table links $c_{i,d}$ originating at customers (i, d) that started a new table, i.e. that chose $a_{i,d} = (i, d)$.

The Gibbs sampler for the HDDCRP iteratively samples a customer link for each customer (i, d) from

$$p(a^*_{i,d}|\mathbf{a}_{-(i,d)}, \mathbf{c}, \mathbf{x}, \lambda) \propto p(a^*_{i,d}) H_a(\mathbf{x}, \mathbf{z}, \lambda) \quad (4)$$

where

$$H_a(\mathbf{x}, \mathbf{z}, \lambda) = \frac{p(\mathbf{x}|\mathbf{z}(\mathbf{a}_{-(i,d)} \cup a^*_{i,d}, \mathbf{c}, \lambda))}{p(\mathbf{x}|\mathbf{z}(\mathbf{a}_{-(i,d)}, \mathbf{c}), \lambda))}$$

After sampling all the customer links, it samples a table link for all customers (i, d) according to

$$p(c^*_{i,d}|\mathbf{a}, \mathbf{c}_{-(i,d)}, \mathbf{x}, \lambda) \propto p(c^*_{i,d}) H_c(\mathbf{x}, \mathbf{z}, \lambda) \quad (5)$$

where

$$H_c(\mathbf{x}, \mathbf{z}, \lambda) = \frac{p(\mathbf{x}|\mathbf{z}(\mathbf{a}, \mathbf{c}_{-(i,d)} \cup c^*_{i,d}, \lambda))}{p(\mathbf{x}|\mathbf{z}(\mathbf{a}, \mathbf{c}_{-(i,d)}), \lambda))}$$

For those customers (i, d) that did not start a new table, i.e. with $a_{i,d} \neq (i, d)$, the table link $c^*_{i,d}$ does not affect the clustering, and so $H_c(\mathbf{x}, \mathbf{z}, \lambda) = 1$ in this case.

Referring back to the event coreference example in 1, Figure 3 shows an example of variable configuration for the HDDCRP model and the corresponding coreference clusters.

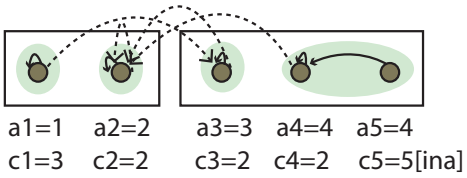

Figure 3: An example of event clustering and the corresponding variable assignments. The assignments of $\mathbf{a}$ induce tables, or within-document (WD) clusters, and the assignments of $\mathbf{c}$ induce menus, or cross-document (CD) clusters. [ina] denotes that the variable is inactive and will not affect the clustering.

In implementation, we can simplify the computations of both $H_a(\mathbf{x}, \mathbf{z}, \lambda)$ and $H_c(\mathbf{x}, \mathbf{z}, \lambda)$ by using the fact that the likelihood under clustering $\mathbf{z}(\mathbf{a}, \mathbf{c})$ can be factorized as

$$p(\mathbf{x}|\mathbf{z}(\mathbf{a}, \mathbf{c}), \lambda) = \prod_{k \in \mathbf{z}(\mathbf{a}, \mathbf{c})} p(\mathbf{x}_{\mathbf{z}=k}|\lambda)$$

where $\mathbf{x}_{\mathbf{z}=k}$ denotes all customers that belong to the global cluster k. $p(\mathbf{x}_{\mathbf{z}=k}|\lambda)$ is the marginal probability. It can be computed as

$$p(\mathbf{x}_{\mathbf{z}=k}|\lambda) = \int p(\phi|\lambda) \prod_{i \in \mathbf{z}=k} p(x_i|\phi) d\phi$$

where x_i is the observation associated with customer i. In our problem, the observation corresponds to the lemmatized words in the event mention. We model the observed word counts using cluster-specific multinomial distributions with symmetric Dirichlet priors.

5.4 Feature-based Distance Functions

The distance functions $F_{1:D}$ and F_0 encode the priors for the clustering distribution, preferring clustering data points that are closer to each other. We consider event mentions as the data points and encode the similarity (or compatibility) between event mentions as priors for event clustering. Specifically, we use a log-linear model to estimate the similarity between a pair of event mentions (x_i, x_j)

$$f_{\boldsymbol{\theta}}(x_i, x_j) \propto \exp\{\boldsymbol{\theta}^T \psi(x_i, x_j)\} \quad (6)$$

where ψ is a feature vector, containing a rich set of features based on event mentions i and j: (1) head word string match, (2) head POS pair, (3) cosine similarity between the head word embeddings (we use the pre-trained 300-dimensional word embeddings from word2vec[1]), (4) similarity between the words in the event mentions (based on term frequency (TF) vectors), (5) the Jaccard coefficient between the WordNet synonyms of the head words, and (6) similarity between the context words (a window of three words before and after each event mention). If both event mentions involve participants, we consider the similarity between the words in the participant mentions based on the TF vectors, similarly for the time mentions and the location mentions. If the SRL role information is available, we also consider the similarity between words in each SRL role, i.e. Arg0, Arg1, Arg2.

Training We train the parameter $\boldsymbol{\theta}$ using logistic regression with an L2 regularizer. We construct the training data by considering all ordered pairs of event mentions within a document, and also all pairs of event mentions across similar documents. To measure document similarity, we collect all mentions of events, participants, times and locations in each document and compute the cosine similarity between the TF vectors constructed from all the event-related mentions. We consider two documents to be similar if their TF-based similarity is above a threshold σ (we set it to 0.4 in our experiments).

After learning $\boldsymbol{\theta}$, we set the within-document distances as $F_d(i,j) = f_{\boldsymbol{\theta}}(x_i, x_j)$, and the across-document distances as $F_0((i,d),(j,d')) = w(d,d') f_{\boldsymbol{\theta}}(x_{i,d}, x_{j,d'})$, where $w(d,d') = \exp(\gamma sim(d,d'))$ captures document similarity where $sim(d,d')$ is the TF-based similarity between document d and d', and γ is a weight parameter. Higher γ leads to a higher effect of document-level similarities on the linking probabilities. We set $\gamma = 1$ in our experiments.

[1] https://code.google.com/p/word2vec/

	Train	Dev	Test	Total
# Documents	462	73	447	982
# Sentences	7,294	649	7,867	15,810
# Annotated event mentions	3,555	441	3,290	7,286
# Cross-document chains	687	47	486	1,220
# Within-document chains	2,499	316	2,137	4,952

Table 2: Statistics of the ECB+ corpus

6 Experiments

We conduct experiments using the ECB+ corpus (Cybulska and Vossen, 2014b), the largest available dataset with annotations of both within-document (WD) and cross-document (CD) event coreference resolution. It extends ECB 0.1 (Lee et al., 2012) and ECB (Bejan and Harabagiu, 2010) by adding event argument and argument type annotations as well as adding more news documents. The cross-document coreference annotations only exist in documents that describe the same seminal event (the event that triggers the topic of the document and has interconnections with the majority of events from its surrounding textual context (Bejan and Harabagiu, 2014)). We divide the dataset into a training set (topics 1-20), a development set (topics 21-23), and a test set (topics 24-43). Table 2 shows the statistics of the data.

We performed event coreference resolution on all possible event mentions that are expressed in the

documents. Using the event extraction method described in Section 4, we extracted 53,429 event mentions, 43,682 participant mentions, 5,791 time mentions and 3,836 location mentions in the test data, covering 93.5%, 89.0%, 95.0%, 72.8% of the annotated event mentions, participants, time and locations, respectively.

We evaluate both within- and cross-document event coreference resolution. As in previous work (Bejan and Harabagiu, 2010), we evaluate cross-document coreference resolution by merging all documents from the same seminal event into a meta-document and then evaluate the meta-document as in within-document coreference resolution. However, during inference time, we do not assume the knowledge of the mapping of documents to seminal events.

We consider three widely used coreference resolution metrics: (1) MUC (Vilain et al., 1995), which measures how many gold (predicted) cluster merging operations are needed to recover each predicted (gold) cluster; (2) B^3 (Bagga and Baldwin, 1998), which measures the proportion of overlap between the predicted and gold clusters for each mention and computes the average scores; and (3) CEAF (Luo, 2005) ($CEAF_e$), which measures the best alignment of the gold-standard and predicted clusters. We also consider the CoNLL F1, which is the average F1 of the above three measures. All the scores are computed using the latest version (v8.01) of the official CoNLL scorer (Pradhan et al., 2014).

6.1 Baselines

We compare our proposed HDDCRP model (HDDCRP) to five baselines:

- LEMMA: a heuristic method that groups all event mentions, either within or across documents, which have the same lemmatized head word. It is usually considered a strong baseline for event coreference resolution.

- AGGLOMERATIVE: a supervised clustering method for within-document event coreference (Chen et al., 2009). We extend it to within- and cross-document event coreference by performing single-link clustering in two phases: first grouping mentions within documents and then grouping within-document clusters to larger clusters across documents. We compute the pairwise-linkage scores using the log-linear model described in Section 5.4.

- HDP-LEX: an unsupervised Bayesian clustering model for within- and cross-document event coreference (Bejan and Harabagiu, 2010)[2]. It is a hierarchical Dirichlet process (HDP) model with the likelihood of all the lemmatized words observed in the event mentions. In general, the HDP can be formulated using a two-level sequential CRP. Our HDDCRP model is a two-level DDCRP that generalizes the HDP to allow data dependencies to be incorporated at both levels[3].

- DDCRP: a DDCRP model we develop for event coreference resolution. It applies the distance prior in Equation 1 to all pairs of event mentions in the corpus, ignoring the document boundaries. It uses the same likelihood function and the same log-linear model to learn the distance values as HDDCRP. But it has fewer link variables than HDDCRP and it does not distinguish between the within-document and cross-document link variables. For the same clustering structure, HDDCRP can generate more possible link configurations than DDCRP.

- HDDCRP*: a variant of the proposed HDDCRP that only incorporates the within-document dependencies but not the cross-document dependencies. The generative process of HDDCRP* is similar to the one described in Section 5.2, except that in step 2, for each table t, we sample

[2]We re-implement the proposed HDP-based models: the HDP_{1f}, HDP_{flat} (including HDP_{flat} (LF), (LF+WF), and (LF+WF+SF)) and HDP_{struct}, but found that the HDP_{flat} with lexical features (LF) performs the best in our experiments. We refer to it as HDP-LEX.

[3]Note that HDP-LEX is not a special case of HDDCRP because we define the table-level distance function as the distances between customers instead of between tables. In our model, the probability of linking a table t to another table s depends on the distance between the head customer at table t and all other customers who sit at table s. Defining the table-level distance function this way allows us to derive a tractable inference algorithm using Gibbs sampling.

a cluster assignment c_t according to

$$p(c_t = k) \propto \begin{cases} n_k, & k \leq K \\ \alpha_0, & k = K + 1 \end{cases}$$

where K is the number of existing clusters, n_k is the number of existing tables that belong to cluster k, α is the concentration parameter. And in step 3, the clusters $\mathbf{z}(\mathbf{a}, \mathbf{c})$ are constructed by traversing the customer links and looking up the cluster assignments for the obtained tables. We also use Gibbs sampling for inference.

6.2 Parameter settings

For all the Bayesian models, the reported results are averaged results over five MCMC runs, each for 500 iterations. We found that mixing happens before 500 iterations in all models by observing the joint log-likelihood. For the DDCRP, HDDCRP* and HDDCRP, we randomly initialized the link variables. Before initialization, we assume that each mention belongs to its own cluster. We assume mentions are ordered according to their appearance within a document, but we do not assume any particular ordering of documents. We also truncated the pairwise mention similarity to zero if it is below 0.5 as we found that it leads to better performance on the development set. We set $\alpha_1 = ... = \alpha_D = 0.5$, $\alpha_0 = 0.001$ for HDDCRP, $\alpha_0 = 1$ for HDDCRP*, $\alpha = 0.1$ for DDCRP, and $\lambda = 10^{-7}$. All the hyperparameters were set based on the development data.

6.3 Main Results

Table 3 shows the event coreference results. We can see that LEMMA-matching is a strong baseline for event coreference resolution. HDP-LEX provides noticeable improvements, suggesting the benefit of using an infinite mixture model for event clustering. AGGLOMERATIVE further improves the performance over HDP-LEX for WD resolution, however, it fails to improve CD resolution. We conjecture that this is due to the combination of ineffective thresholding and the prediction errors on the pairwise distances between mention pairs across documents. Overall, HDDCRP* outperforms all the baselines in CoNLL F1 for both WD and CD evaluation. The clear performance gains over HDP-LEX demonstrate that it is important to account for pairwise mention dependencies in the generative modeling of event clustering. The improvements over AGGLOMERATIVE indicate that it is more effective to model mention-pair dependencies as clustering priors than as heuristics for deterministic clustering.

Comparing among the HDDCRP-related models, we can see that HDDCRP clearly outperforms DDCRP, demonstrating the benefits of incorporating the hierarchy into the model. HDDCRP also performs better than HDDCRP* in WD CoNLL F1, indicating that incorporating cross-document information helps within-document clustering. We can also see that HDDCRP performs similarly to HDDCRP* in CD CoNLL F1 due to the lower B^3 F1, in particular, the decrease in B^3 recall. This is because applying the DDCRP prior at both within- and cross-document levels results in more conservative clustering and produces smaller clusters. This could be potentially improved by employing more accurate similarity priors.

To further understand the effect of modeling mention-pair dependencies, we analyze the impact of the features in the mention-pair similarity model. Table 4 lists the learned weights of some top features (sorted by weights). We can see that they mainly serve to discriminate event mentions based on the head word similarity (especially embedding-based similarity) and the context word similarity. Event argument information such as *SRL Arg1*, *SRL Arg0*, and *Participant* are also indicative of the coreferential relations.

6.4 Discussion

We found that HDDCRP corrects many errors made by the traditional agglomerative clustering model (AGGLOMERATIVE) and the unsupervised generative model (HDP-LEX). AGGLOMERATIVE easily suffers from error propagation as the errors made by the supervised distance learner cannot be corrected. HDP-LEX often mistakenly groups mentions together based on word co-occurrence statistics but not the apparent similarity features in the mentions. In contrast, HDDCRP avoids such errors by performing probabilistic modeling of clustering and making use of rich linguistic features trained on available annotated data. For example, HDDCRP correctly groups the event mention "unveiled" in *"Apple's Phil Schiller unveiled a revamped MacBook*

	MUC			B^3			$CEAF_e$			CoNLL
	P	R	F1	P	R	F1	P	R	F1	F1
	Cross-document Event Coreference Resolution (CD)									
LEMMA	75.1	55.4	63.8	71.7	39.6	51.0	36.2	61.1	45.5	53.4
HDP-LEX	75.5	63.5	69.0	65.6	43.7	52.5	34.8	60.2	44.1	55.2
AGGLOMERATIVE	78.3	59.2	67.4	73.2	40.2	51.9	30.2	65.6	41.4	53.6
DDCRP	79.6	58.2	67.1	78.1	39.6	52.6	31.8	**69.4**	43.6	54.4
HDDCRP*	77.5	66.4	71.5	69.0	**48.1**	**56.7**	38.2	63.0	47.6	58.6
HDDCRP	**80.3**	**67.1**	**73.1**	**78.5**	40.6	53.5	**38.6**	68.9	**49.5**	**58.7**
	Within-document Event Coreference Resolution (WD)									
LEMMA	60.9	30.2	40.4	78.9	57.3	66.4	63.6	69.0	66.2	57.7
HDP-LEX	50.0	39.1	43.9	74.7	67.6	71.0	66.2	71.4	68.7	61.2
AGGLOMERATIVE	61.9	39.2	48.0	80.7	67.6	73.5	65.6	76.0	70.4	63.9
DDCRP	71.2	36.4	48.2	85.4	64.9	73.8	61.8	76.1	68.2	63.4
HDDCRP*	58.1	**42.8**	49.3	78.4	**68.7**	73.2	**67.6**	74.5	70.9	64.5
HDDCRP	**74.3**	41.7	**53.4**	**85.6**	67.3	**75.4**	65.1	**79.8**	**71.7**	**66.8**

Table 3: Within- and cross-document coreference results on the ECB+ corpus

Pro today" together with the event mention "announced" in *"this notebook isn't the only laptop Apple announced for the MacBook Pro lineup today"*, while both HDP-LEX and AGGLOMERATIVE models fail to make such connection.

By looking further into the errors, we found that a lot of mistakes made by HDDCRP are due to the errors in event extraction and pairwise linkage prediction. The event extraction errors include false positive and false negative event mentions and event arguments, boundary errors for the extracted mentions, and argument association errors. The pairwise linking errors often come from the lack of semantic and world knowledge, and this applies to both event mentions and event arguments, especially for time and location arguments which are less likely to be repeatedly mentioned and in many cases require external knowledge to resolve their meanings, e.g., "*May 3, 2013*" is "*Friday*" and "*Mount Cook*" is "*New Zealand's highest peak*".

Features	Weight
Head Embedding sim	4.5
String match	2.77
Context sim	1.75
Synonym sim	1.56
TF sim	1.17
SRL Arg1 sim	1.10
SRL Arg0 sim	0.89
Participant sim	0.68

Table 4: Learned weights for selected features

7 Conclusion

In this paper we propose a novel Bayesian model for within- and cross-document event coreference resolution. It leverages the advantages of generative modeling of coreference resolution and feature-rich discriminative modeling of mention reference relations. We have shown its power in resolving event coreference by comparing it to a traditional agglomerative clustering approach and a state-of-the-art unsupervised generative clustering approach. It is worth noting that our model is general and can be easily applied to other clustering problems involving feature-rich objects and cluster sharing across data groups. While the model can effectively cluster objects of a single type, it would be interesting to extend it to allow joint clustering of objects of different types, e.g., events and entities.

Acknowledgments

We thank Cristian Danescu-Niculescu-Mizil, Igor Labutov, Lillian Lee, Moontae Lee, Jon Park, Chenhao Tan, and other Cornell NLP seminar participants and the reviewers for their helpful comments. This work was supported in part by NSF grant IIS-1314778 and DARPA DEFT Grant FA8750-13-2-0015. The third author was supported by

NSF CAREER CMMI-1254298, NSF IIS-1247696, AFOSR FA9550-12-1-0200, AFOSR FA9550-15-1-0038, and the ACSF AVF. 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, DARPA 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*, pages 1–8.

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 563–6.

Cosmin Adrian Bejan and Sanda Harabagiu. 2010. Unsupervised event coreference resolution with rich linguistic features. In *ACL*, pages 1412–1422.

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

Indrajit Bhattacharya and Lise Getoor. 2006. A latent Dirichlet model for unsupervised entity resolution. In *SDM*, volume 5, page 59.

David M. Blei and Peter I. Frazier. 2011. Distance dependent Chinese restaurant processes. *The Journal of Machine Learning Research*, 12:2461–2488.

Claire Cardie and Kiri Wagstaff. 1999. Noun phrase coreference as clustering. In *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 82–89.

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.

Agata Cybulska and Piek Vossen. 2014a. Guidelines for ECB+ annotation of events and their coreference. Technical report, NWR-2014-1, VU University Amsterdam.

Agata Cybulska and Piek Vossen. 2014b. Using a sledgehammer to crack a nut? lexical diversity and event coreference resolution. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC2014)*, pages 26–31.

Greg Durrett and Dan Klein. 2013. Easy victories and uphill battles in coreference resolution. In *EMNLP*, pages 1971–1982.

Soumya Ghosh, Andrei B. Ungureanu, Erik B. Sudderth, and David M. Blei. 2011. Spatial distance dependent Chinese restaurant processes for image segmentation. In *Advances in Neural Information Processing Systems*, pages 1476–1484.

Soumya Ghosh, Michalis Raptis, Leonid Sigal, and Erik B. Sudderth. 2014. Nonparametric clustering with distance dependent hierarchies.

Aria Haghighi and Dan Klein. 2007. Unsupervised coreference resolution in a nonparametric Bayesian model. In *ACL*, volume 45, page 848.

Aria Haghighi and Dan Klein. 2010. Coreference resolution in a modular, entity-centered model. In *NAACL*, pages 385–393.

Kevin Humphreys, Robert Gaizauskas, and Saliha Azzam. 1997. Event coreference for information extraction. In *Proceedings of a Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, pages 75–81.

Andrew Kehler. 2002. *Coherence, Reference, and the Theory of Grammar*. CSLI publications Stanford, CA.

Dongwoo Kim and Alice Oh. 2011. Accounting for data dependencies within a hierarchical Dirichlet process mixture model. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, pages 873–878.

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 28–34.

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.

Zhengzhong Liu, Jun Araki, Eduard Hovy, and Teruko Mitamura. 2014. Supervised within-document event coreference using information propagation. In *Proceedings of the International Conference on Language Resources and Evaluation*.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *EMNLP*, pages 25–32.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Proceedings of Workshop at ICLR*.

Vincent Ng and Claire Cardie. 2002. Improving machine learning approaches to coreference resolution. In *ACL*, pages 104–111.

Vincent Ng. 2010. Supervised noun phrase coreference research: The first fifteen years. In *ACL*, pages 1396–1411.

Sameer Pradhan, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, and Michael Strube. 2014. Scoring coreference partitions of predicted mentions: A reference implementation. In *ACL*, pages 22–27.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. 2010. A multi-pass sieve for coreference resolution. In *EMNLP*, pages 492–501.

Altaf Rahman and Vincent Ng. 2011. Coreference resolution with world knowledge. In *ACL*, pages 814–824.

Michael Roth and Anette Frank. 2012. Aligning predicate argument structures in monolingual comparable texts: A new corpus for a new task. In *SemEval*, pages 218–227.

Sunita Sarawagi and William W. Cohen. 2004. Semi-markov conditional random fields for information extraction. In *Advances in Neural Information Processing Systems*, pages 1185–1192.

Sameer Singh, Michael Wick, and Andrew McCallum. 2010. Distantly labeling data for large scale cross-document coreference. *arXiv:1005.4298*.

Richard Socher, Andrew L. Maas, and Christopher D. Manning. 2011. Spectral Chinese restaurant processes: Nonparametric clustering based on similarities. In *International Conference on Artificial Intelligence and Statistics*, pages 698–706.

Veselin Stoyanov, Nathan Gilbert, Claire Cardie, and Ellen Riloff. 2009. Conundrums in noun phrase coreference resolution: Making sense of the state-of-the-art. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 656–664.

Mihai Surdeanu, Lluís Màrquez, Xavier Carreras, and Pere R. Comas. 2007. Combination strategies for semantic role labeling. *Journal of Artificial Intelligence Research*, pages 105–151.

Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2006. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476).

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.

Michael Wick, Sameer Singh, and Andrew McCallum. 2012. A discriminative hierarchical model for fast coreference at large scale. In *ACL*, pages 379–388.

Travis Wolfe, Mark Dredze, and Benjamin Van Durme. 2015. Predicate argument alignment using a global coherence model. In *NAACL*, pages 11–20.

Bishan Yang and Claire Cardie. 2014. Joint modeling of opinion expression extraction and attribute classification. *Transactions of the Association for Computational Linguistics*, 2:505–516.